

3D Graphics Visualization with VMware and NVIDIA on Cisco UCS

Understand the performance of the VMware ESXi hypervisor and VMware Horizon with NVIDIA Tesla P4, P6, and P40 solution on Cisco UCS C240 M5 Rack Servers and B200 M5 Blade Servers.



November 2018

Contents

- What you will learn 4
- vGPU Profiles 4
- Cisco Unified Computing System 5
 - Cisco UCS Manager 7
 - Cisco UCS 6332 Fabric Interconnect 7
 - Cisco UCS C-Series Rack Servers 7
 - Cisco UCS C240 M5 Rack Server 8
 - Cisco UCS VIC 1387 10
 - Cisco UCS B200 M5 Blade Server 11
 - Cisco UCS VIC 1340 11
- NVIDIA Tesla graphics cards 12
- NVIDIA GRID 12
- NVIDIA GRID 6.2 GPU 12
 - NVIDIA GRID 6.2 license requirements 13
- VMware vSphere 6.7 13
 - Graphics acceleration in VMware Horizon 7.5 14
 - GPU acceleration for Microsoft Windows desktops 15
 - Enhanced graphics with VMware Horizon 7 with Blast 3D 16
 - GPU acceleration for Microsoft Windows Server 17
 - GPU sharing for VMware Horizon remote desktop session host workloads 17
- Solution configuration 19
- Configure Cisco UCS 21
 - Create BIOS policy 21
 - Create graphics card policy 21
 - Install the NVIDIA Tesla GPU card on the Cisco UCS B200 M5 22
 - Install the NVIDIA Tesla GPU card on the Cisco UCS C240 M5 26
 - Configure the GPU card 29
- Install NVIDIA GRID software on the VMware ESXi host 31
- Modify GPU allocation policy 33
- Install and configure the NVIDIA GRID license server 34
 - Install the NVIDIA GRID 6.2 license server 34
 - Configure the NVIDIA GRID 6.2 license server 39
- NVIDIA Tesla P6, P40, and P4 profile specifications 42
- Create virtual desktops with vGPU support 42
 - Create the base image for virtual desktops 42
 - Install and configure Microsoft Windows on the virtual machine 46
 - Install the NVIDIA vGPU software driver 46
 - Verify that the virtual machine is ready to support the vGPU 48
 - Configure the virtual machine for an NVIDIA GRID vGPU license 49
 - Create a vGPU-enabled desktop pool with VMware Horizon 7.5 50



- Verify vGPU deployment51**
 - Verify that the NVIDIA driver is running on the desktop 51
 - Verify NVIDIA license acquisition by desktops 52
 - Use the VMware vSphere 6.7 Performance tab to monitor GPU use 53
- SPECviewperf 13 benchmark results.....53**
 - NVIDIA Tesla P4 test results 55
 - NVIDIA Tesla P40 test results 55
 - NVIDIA Tesla P6 test results 56
 - Host CPU utilization test results 57
 - Host GPU utilization test results 59
- Live Migration of vGPU-enabled virtual machines with VMware vMotion 61**
- Additional configurations..... 65**
 - Install and upgrade NVIDIA drivers 65
 - Use VMware Horizon Performance Tracker 65
 - Optimize VMware Blast with the GPO bundle to achieve optimal end-user experience..... 66
 - Use GPU acceleration for Microsoft Windows Server DirectX, Direct3D, and WPF rendering 66
- Conclusion..... 67**
- For more information..... 67**

What you will learn

Using the increased processing power of today's Cisco UCS® B-Series Blade Servers and C-Series Rack Servers, applications with demanding graphics requirements are now being virtualized. To enhance the capability to deliver these high-performance and graphics-intensive applications in Virtual Client Computing (VCC), Cisco offers support for the NVIDIA Tesla P4, P6 and P40 graphics cards in the Cisco Unified Computing System™ (Cisco UCS) portfolio of Mobile PCI Express (PCIe) Module (MXM) mezzanine form-factor and PCI Express (PCIe) cards for the B-Series Blade Servers and C-Series Rack Servers respectively.

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

A major focus of this document is the Cisco® data center infrastructure and VMware support for the NVIDIA Virtual Graphics Processing Unit (vGPU), including the capability of VMware ESXi to suspend and resume (ESXi Version 6.7 and GRID Software Version 6.2) or, using VMware vMotion (ESXi Version 6.7 Update 1 and GRID Software Version 7), move vGPU-enabled virtual machines, reducing user downtime.

The purpose of this document is to help our partners and customers integrate NVIDIA GRID 6.2 software and NVIDIA Tesla graphics cards and Cisco UCS B200 M5 Blade Servers and C240 M5 Rack Servers with the VMware ESXi 6.7 hypervisor and VMware Horizon 7.5 with Microsoft Windows 10 virtual machines in vGPU mode.

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

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

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

vGPU Profiles

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

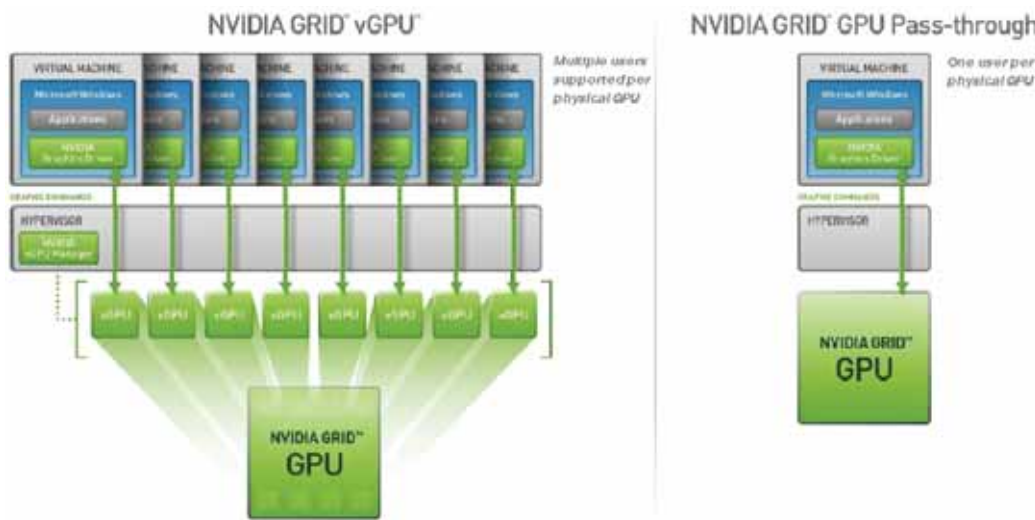
Although the needs of end users can be diverse, for simplicity users can be grouped into the following categories: knowledge workers, designers, and power users.

- For knowledge workers, the main areas of importance include office productivity applications, a robust web experience, and fluid video playback. Knowledge workers have the least-intensive graphics demands, but they expect the same smooth, fluid experience that exists natively on today's graphics-accelerated devices such as desktop PCs, notebooks, tablets, and smartphones.
- Power users are users who need to run more demanding office applications, such as office productivity software, image editing software such as Adobe Photoshop, mainstream computer-aided design (CAD) software such as Autodesk AutoCAD, and product lifecycle management (PLM) applications. These applications are more demanding and require additional graphics resources with full support for APIs such as OpenGL and Direct3D.

- Designers are users in an organization who run demanding professional applications such as high-end CAD software and professional digital content creation (DCC) tools. Examples include Autodesk Inventor, PTC Creo, Autodesk Revit, and Adobe Premiere. Historically, designers have used desktop workstations and have been a difficult group to incorporate into virtual deployments because of their need for high-end graphics and the certification requirements of professional CAD and DCC software.

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

Figure 1. NVIDIA GRID vGPU GPU system architecture



Cisco Unified Computing System

The main components of Cisco UCS are:

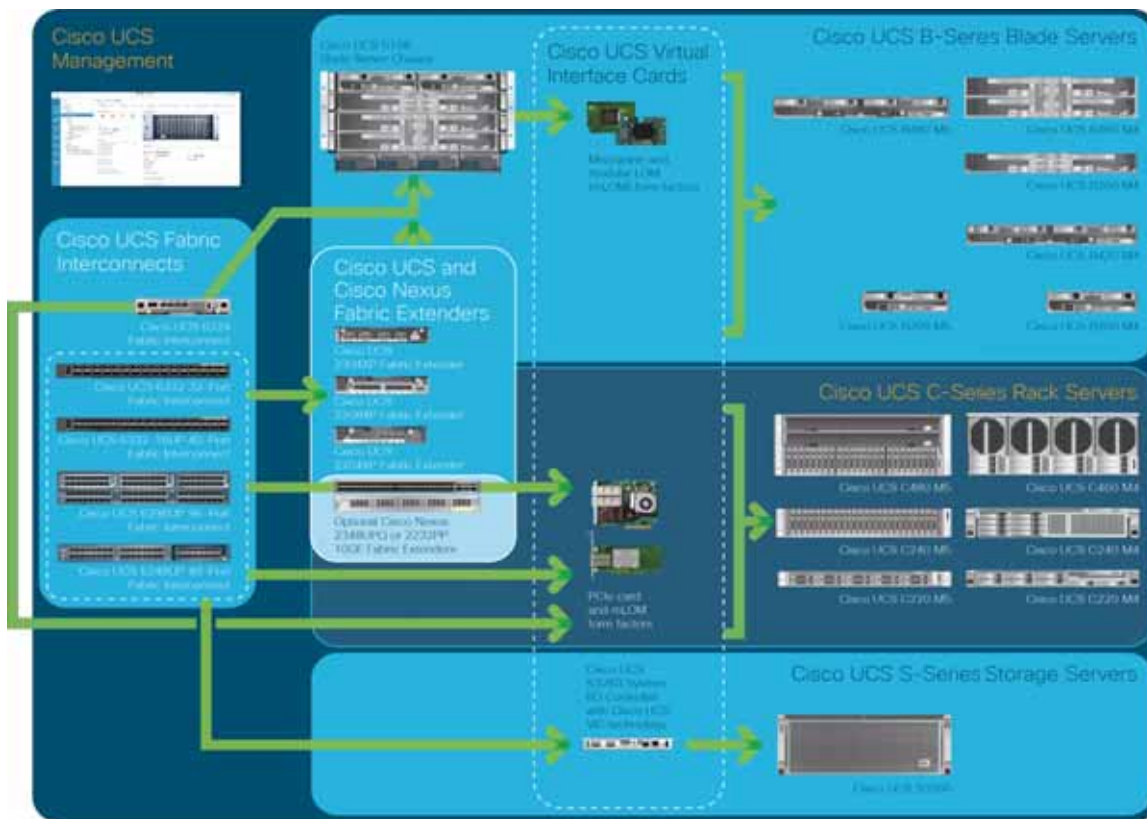
Compute: The system is based on an entirely new class of computing system that incorporates blade servers based on Intel® Xeon® Scalable Family processors.

- **Network:** The system is integrated on a low-latency, lossless, 40-Gbps unified network fabric. This network foundation consolidates LANs, SANs, and high-performance computing (HPC) networks, which are separate networks today. The unified fabric lowers costs by reducing the number of network adapters, switches, and cables needed and by decreasing the power and cooling requirements.
- **Virtualization:** The system unleashes the full potential of virtualization by enhancing the scalability, performance, and operational control of virtual environments. Cisco security, policy enforcement, and diagnostic features are now extended into virtualized environments to better support changing business and IT requirements.
- **Storage access:** The system provides consolidated access to local storage, SAN storage, and network-attached storage (NAS) over the unified fabric. With storage access unified, Cisco UCS can access storage over Ethernet, Fibre Channel, Fibre Channel over Ethernet (FCoE), and Small Computer System Interface over IP (iSCSI) protocols. This capability provides customers with a choice for storage access and investment protection. In addition, server administrators can preassign storage-access policies for system connectivity to storage resources, simplifying storage connectivity and management and helping increase productivity.

- **Management:** Cisco UCS uniquely integrates all system components, enabling the entire solution to be managed as a single entity by Cisco UCS Manager. The manager has an intuitive GUI, a command-line interface (CLI), and a robust API for managing all system configuration processes and operations.

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

Figure 2. Cisco Data center overview



Cisco UCS is designed to deliver:

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

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

Cisco UCS Manager

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

Cisco UCS 6332 Fabric Interconnect

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

Because they support unified fabric, Cisco UCS 6300 Series Fabric Interconnects provide both LAN and SAN connectivity for all servers within their domains.

Features and capabilities include:

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

Figure 3. Cisco UCS 6332 Fabric Interconnect

Front View



Rear View



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

Cisco UCS C-Series Rack Servers

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

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

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

Cisco UCS C240 M5 Rack Server

The Cisco UCS C240 M5 Rack Server (Figure 4, Figure 5, and Table 1) is designed for both performance and expandability over a wide range of storage-intensive infrastructure workloads, from big data to collaboration.

The Cisco UCS C240 M5 small-form-factor (SFF) server extends the capabilities of the Cisco UCS portfolio in a 2RU form factor with the addition of the Intel Xeon Scalable family processors, 24 DIMM slots for 2666-MHz DDR4 DIMMs and up to 128-GB capacity points, up to 6 PCIe 3.0 slots, and up to 26 internal SFF drives. The C240 M5 SFF server also includes one dedicated internal slot for a 12-Gbps SAS storage controller card. The C240 M5 server includes a dedicated internal modular LAN on motherboard (mLOM) slot for installation of a Cisco virtual interface card (VIC) or third-party network interface card (NIC), without consuming a PCI slot, in addition to 2 x 10GBASE-T Intel x550 LOM ports (embedded on the motherboard).

In addition, the C240 M5 offers outstanding levels of internal memory and storage expandability with exceptional performance. It delivers:

- Up to 24 DDR4 DIMMs at speeds of up to 2666 MHz for improved performance and lower power consumption
- One or two Intel Xeon Scalable family processor
- Up to 6 PCIe 3.0 slots (4 full-height, full-length for GPU)
- Six hot-swappable fans for front-to-rear cooling
- 24 SFF front-facing SAS/SATA hard disk drives (HDDs) or SAS/SATA solid state disks (SSDs)
- Optionally, up to two front-facing SFF Non-Volatile Memory Express (NVMe) PCIe SSDs (replacing SAS/SATA drives); these drives must be placed in front drive bays 1 and 2 only and are controlled from Riser 2 option C
- Optionally, up to two SFF, rear-facing SAS/SATA HDDs or SSDs, or up to two rear-facing SFF NVMe PCIe SSDs, with rear-facing SFF NVMe drives connected from Riser 2, Option B or C; 12-Gbps SAS drives are also supported
- The dedicated mLOM slot on the motherboard can flexibly accommodate the following cards:
 - Cisco VICs
 - Quad-port Intel i350 1GbE RJ45 mLOM NIC
 - Two 1 Gigabit Ethernet embedded LOM ports
- Support for up to 2 double-wide NVIDIA GPUs, providing a graphics-rich experience to more virtual users
- Excellent reliability, availability, and serviceability (RAS) features with tool-free CPU insertion, easy-to-use latching lid, hot-swappable and hot-pluggable components
- One slot for a micro-SD card on PCIe Riser 1 (Option 1 and 1B); the micro-SD card serves as a dedicated local resource for utilities such as the Cisco Host Upgrade Utility (HUU), and images can be pulled from a file share (Network File System [NFS] or Common Internet File System [CIFS]) and uploaded to the cards for future use
- A mini-storage module connector on the motherboard that supports either:
 - An SD card module with two SD card slots; mixing of different capacity SD cards is not supported
 - An M.2 module with two SATA M.2 SSD slots; mixing of different capacity M.2 modules is not supported

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

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

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

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

For more information about the Cisco UCS C240 M5 Rack Server, see

<https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/c240m5-sff-specsheet.pdf>.

Figure 4. Cisco UCS C240 M5 Rack Server



Figure 5. Cisco UCS C240 M4 Rack Server rear view



Table 1. Cisco UCS C240 M4 PCIe slots

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

Cisco UCS VIC 1387

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

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

Figure 6. Cisco UCS VIC 1387 CNA

Cisco UCS B200 M5 Blade Server

Delivering performance, versatility and density without compromise, the Cisco UCS B200 M5 Blade Server (Figure 7) addresses the broadest set of workloads, from IT and web infrastructure to distributed database workloads. The enterprise-class Cisco UCS B200 M5 blade server extends the capabilities of the Cisco UCS portfolio in a half-width blade form factor. The B200 M5 harnesses the power of the latest Intel Xeon Scalable processor family CPUs, with up to 3072 GB of RAM (using 128-GB DIMMs), two SSDs or HDDs, and connectivity with throughput of up to 80 Gbps.

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

The B200 M5 has one rear mezzanine adapter slot, which can be configured with a Cisco UCS port expander card for the VIC. This hardware option enables an additional four ports of the VIC 1340, bringing the total capability of the VIC 1340 to a dual native 40-Gbps interface or a dual 4 x 10 Gigabit Ethernet port-channel interface, respectively. Alternatively the same rear mezzanine adapter slot can be configured with an NVIDIA P6 GPU.

The B200 M5 has one front mezzanine slot. The B200 M5 can be ordered with or without the front mezzanine card. The front mezzanine card can accommodate a storage controller or an NVIDIA P6 GPU.

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

Figure 7. Cisco UCS B200 M5 Blade Server front view



Cisco UCS VIC 1340

The Cisco UCS VIC 1340 (Figure 8) is a 2-port 40-Gbps Ethernet or dual 4 x 10-Gbps Ethernet and FCoE-capable mLOM designed exclusively for the M4 generation of Cisco UCS B-Series Blade Servers. When used in combination with an optional port expander, the VIC 1340 is enabled for two ports of 40-Gbps Ethernet. The VIC 1340 enables a policy-based, stateless, agile server infrastructure that can present more than 256 PCIe standards-compliant interfaces to the host that can be dynamically configured as either NICs or HBAs. In addition, the VIC 1340 supports Cisco Virtual Machine Fabric Extender (VM-FEX) technology, which extends the Cisco UCS fabric interconnect ports to virtual machines, simplifying server virtualization deployment and management.

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

Figure 8. Cisco UCS VIC 1340



NVIDIA Tesla graphics cards

For desktop virtualization applications, the NVIDIA Tesla P4, P6, and P40 cards are excellent choices for most high-performance graphics applications. Table 2 lists the technical specifications.

Table 2. Technical specifications for NVIDIA GRID cards



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

NVIDIA GRID

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

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

NVIDIA GRID 6.2 GPU

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

NVIDIA GRID 6.2 license requirements

NVIDIA GRID 6.2 requires concurrent user licenses and an on-premises NVIDIA license server to manage the licenses. When the guest OS boots, it contacts the NVIDIA license server and consumes one concurrent license. When the guest OS shuts down, the license is returned to the pool.

GRID 6.2 also requires the purchase of a 1:1 ratio of concurrent licenses to NVIDIA Support, Update, and Maintenance Subscription (SUMS) instances.

The following NVIDIA GRID products are available as licensed products on NVIDIA Tesla GPUs:

- Virtual workstation
- Virtual PC
- Virtual applications

For complete details about GRID 6.2 license requirements, see the [NVIDIA documentation](#).

VMware vSphere 6.7

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

vSphere 6.7 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 platform includes these features:

- Computing
 - **Increased scalability:** vSphere 6.7 supports larger maximum configuration sizes. Virtual machines support up to 128 virtual CPUs (vCPUs) and 6128 GB of virtual RAM (vRAM). Hosts support up to 768 CPUs and 16 TB of RAM, 1024 virtual machines per host, and 64 hosts per cluster.
 - **Expanded support:** Get expanded support for the latest x86 chip sets, devices, drivers, and guest operating systems. For a complete list of guest operating systems supported, see the [VMware Compatibility Guide](#).
 - **Outstanding graphics:** The NVIDIA GRID vGPU delivers the full benefits of NVIDIA hardware-accelerated graphics to virtualized solutions.
 - **Suspend-resume support:** Suspend-resume support is provided for virtual machines that are configured with vGPU.
 - **Instant cloning:** Technology built in to vSphere 6.0 lays the foundation for rapid cloning and deployment of virtual machines—up to 10 times faster than what is possible today.
- Storage
 - **Transformation of virtual machine storage:** vSphere Virtual Volumes enable your external storage arrays to become virtual machine aware. Storage policy-based management (SPBM) enables common management across storage tiers and dynamic storage class-of-service (CoS) automation. Together these features enable exact combinations of data services (such as clones and snapshots) to be instantiated more efficiently on a per-virtual machine basis.
- Network
 - **Network I/O control:** New support for per-virtual machine VMware Distributed Virtual Switch (DVS) bandwidth reservation helps ensure isolation and enforce limits on bandwidth.

- **Multicast snooping:** Support for Internet Group Management Protocol (IGMP) snooping for IPv4 packets and Multicast Listener Discovery (MLD) snooping for IPv6 packets in VMware vSphere Distributed Virtual Switch (DVS) improves performance and scalability with multicast traffic.
- **Multiple TCP/IP stacks for VMware vMotion:** Implement a dedicated networking stack for vMotion traffic, simplifying IP address management with a dedicated default gateway for vMotion traffic.
- Availability
 - **vMotion enhancements:** Perform nondisruptive live migration of workloads across virtual switches and vCenter Servers and over distances with a round-trip time (RTT) of up to 100 milliseconds (ms). This support for dramatically longer RTT—a 10x increase in the supported time—for long-distance vMotion enables data centers physically located in New York and London now to migrate live workloads between one another.
 - **Replication-assisted vMotion:** Customers with active-active replication set up between two sites can perform more efficient vMotion migration, resulting in huge savings in time and resources, with up to 95 percent more efficient migration depending on the amount of data moved.
 - **Fault tolerance:** Get expanded support for software-based fault tolerance for workloads with up to 8 vCPUs, 16 virtual disks, 128 GB of RAM, and a 2-TB disk size.
- Management
 - **Content library:** This centralized repository provides simple and effective management for content, including virtual machine templates, ISO images, and scripts. With the vSphere content library, you can now store and manage content from a central location and share content through a publish-and-subscribe model.
 - **Cloning and migration across vCenter:** Copy and move virtual machines between hosts on different vCenter Servers in a single action.
 - **Enhanced user interface:** vSphere Web Client is more responsive, more intuitive, and simpler than ever before.

For more information about vSphere 6.7 maximum configurations, please refer to [vSphere 6.7 Configuration Maximums](#).

Graphics acceleration in VMware Horizon 7.5

Now with [VMware Horizon 7](#) and NVIDIA GRID, you can significantly improve latency, bandwidth, and frames per second while decreasing CPU utilization and increasing the number of users per host by using NVIDIA Blast Extreme Acceleration.

[VMware's new Blast Extreme protocol](#) was built from the start to deliver a remarkable user experience through the LAN or WAN by using H.264 as the default video codec. The video codec is a very important element in delivering remarkable user experiences because it affects many factors: latency, bandwidth, frames per second (FPS), and more. Moving to H.264 as the primary video codec also allows VMware to use millions of H.264-enabled access devices to offload the encode-decode process from the CPU to dedicated H.264 engines on NVIDIA GPUs. This feature is available with NVIDIA GRID.

Examples of 3D professional applications include:

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

Blast Extreme 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.

Blast Extreme provides GPU acceleration for Microsoft Windows desktops and Microsoft Windows Server. When used with VMware vSphere 6 and NVIDIA GRID GPUs, Blast Extreme provides vGPU acceleration for Windows desktops. For more information, see <https://techzone.vmware.com/sites/default/files/vmware-horizon-7-view-blast-extreme-display-protocol.pdf>.

GPU acceleration for Microsoft Windows desktops

With VMware Blast Extreme, you can deliver graphics-intensive applications as part of hosted desktops or applications on desktop OS machines. Blast Extreme supports physical host computers (including desktop, blade, and rack workstations) and GPU pass-through and GPU virtualization technologies offered by 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 those deployed on non-virtualized operating systems.

VMware Blast Extreme offers the following features:

- Users outside the corporate firewall can use this protocol with your company's virtual private network (VPN), or users can make secure, encrypted connections to a security server or access-point appliance in the corporate DMZ.
- Advanced Encryption Standard (AES) 128-bit encryption is supported and is turned on by default. You can, however, change the encryption key cipher to AES-256.
- You can make connections from all types of client devices.
- Optimization controls help you reduce bandwidth use on the LAN and WAN.
- 32-bit color is supported for virtual displays.
- ClearType fonts are supported.
- You can use audio redirection with dynamic audio quality adjustment for the LAN and WAN.
- Real-time audio and video is supported for webcams and microphones on some client types.
- You can copy and paste text and, on some clients, images between the client operating system and a remote application or desktop. Other client types support copy and paste of only plain text. You cannot copy and paste system objects such as folders and files between systems.
- Multiple monitors are supported for some client types. On some clients, you can use up to four monitors with a resolution of up to 2560 x 1600 pixels per display, or up to three monitors with a resolution of 4K (3840 x 2160 pixels) for Microsoft Windows 7 remote desktops with Aero disabled. Pivot display and autofit are also supported.
- When the 3D feature is enabled, up to two monitors are supported with a resolution of up to 1920 x 1200 pixels, or one monitor with a resolution of 4K (3840 x 2160 pixels).
- USB redirection is supported for some client types.

- Multimedia redirection (MMR) is supported for some Windows client operating systems and some remote desktop operating systems (with Horizon Agent installed).

Enhanced graphics with VMware Horizon 7 with Blast 3D

Horizon with Blast 3D breaks the restraints of the physical workstation. Virtual desktops now deliver immersive 2D and 3D graphics smoothly rendered on any device, accessible from any location. Power users and designers can collaborate with global teams in real time, and organizations can increase workforce productivity, save costs, and expand user capabilities.

With a portfolio of solutions, including software- and hardware-based graphics-acceleration technologies, VMware Horizon provides a full-spectrum approach to enhancing the user experience and accelerating application responsiveness. Take advantage of Soft-3D, vSGA, vDGA, and NVIDIA GRID vGPU to deliver the right level of user experience and performance for every use case in your organization with secure, immersive 3D graphics from the cloud.

Power users and designers get the same graphics experience that they expect from dedicated hardware, delivered securely and cost effectively and with improved collaboration workflow. Enable dispersed teams to collaborate on large graphics data sets in real time from the cloud. Provide greater security for mission-critical data. Protect intellectual property and improve security by centralizing data files.

Deploy with confidence. A growing portfolio of leading independent software vendor (ISV) certifications, including certifications from ESRI, PTC, and Siemens, helps ensure that users get the same graphics performance and experience as from their physical PCs and workstations.

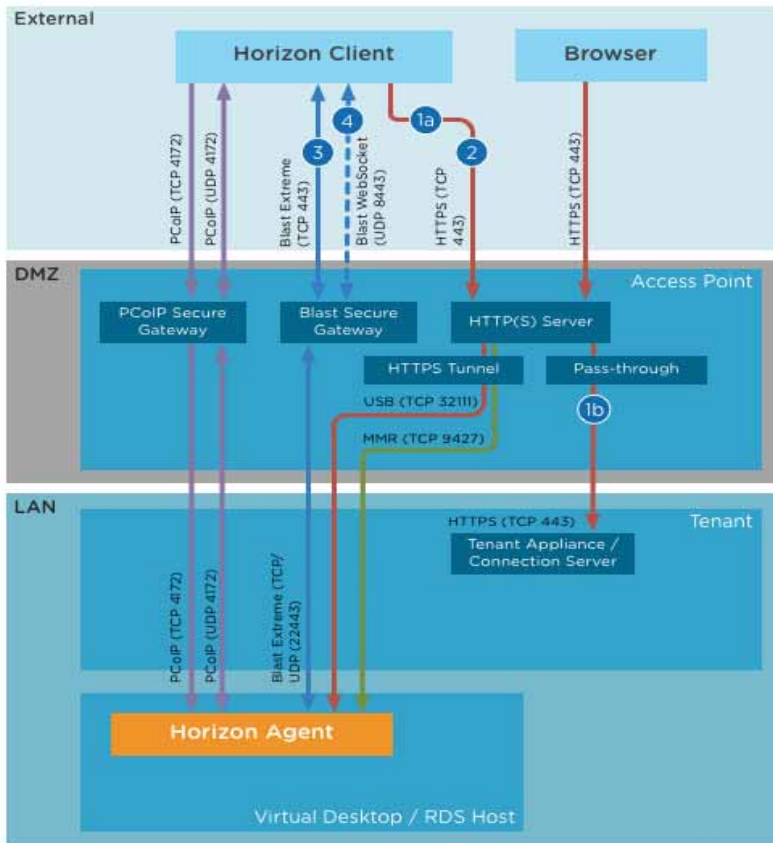
As shown in Figure 9, Blast Extreme provides an enhanced remote session experience introduced with Horizon for Linux desktops, Horizon 7, and Horizon Desktop as a Service (DaaS). In this case, the connection flow from the Horizon Client differs from the flow for PC over IP (PCoIP).

- The Horizon Client sends authentication credentials using the XML API over HTTPS to the external URL on an access-point appliance or a security server. This process typically uses a load-balancer virtual IP address.
- HTTPS authentication data is passed from the access point to the tenant appliance (Horizon DaaS). In the case of a security server, the server will use Apache JServ Protocol 13 (AJP13)-forwarded traffic, which is protected by IP Security (IPsec), from the security server to a paired connection server. Any entitled desktop pools are returned to the client.

Note: If multiple access-point appliances are used, which is often the case, a load-balancer virtual IP address will be used to load-balance the access-point appliances. Security servers use a different approach, with each security server paired with a connection server. No such pairing exists for access points.

- The user selects a desktop or application, and a session handshake occurs over HTTPS (TCP 443) to the access point or security server.
- A secure WebSocket connection is established (TCP 443) for the session data between the Horizon Client and the access point or security server.
- The Blast Secure Gateway service (for the access point or security server) will attempt to establish a User Datagram Protocol (UDP) WebSocket connection on port 443. This approach is preferred, but if this fails because, for example, a firewall is blocking it, then the initial WebSocket TCP 443 connection will be used.

Figure 9. VMware Blast Extreme process flow



GPU acceleration for Microsoft Windows Server

VMware Blast Extreme 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 VMware Horizon remote desktop session host workloads

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

- Sharing can be used on 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 VMware 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 vSphere 6. For RDS hosts, applications in application pools and applications running on RDS desktops both can display 3D graphics.

The following 3D graphics options are available:

- With vDGA, you allocate an entire GPU to a single machine for maximum performance. The RDS host must be in a manual farm.
- With NVIDIA GRID vGPU, each graphics card can support multiple RDS hosts, and the RDS hosts must be in a manual farm. If an ESXi host has multiple physical GPUs, you can also configure the way that the ESXi host assigns virtual machines to the GPUs. By default, the ESXi host assigns virtual machines to the physical GPU with the fewest virtual machines already assigned. This approach is called performance mode. You can also choose consolidation mode, in which the ESXi host assigns virtual machines to the same physical GPU until the maximum number of virtual machines is reached before placing virtual machines on the next physical GPU.
- To configure consolidation mode, edit the `/etc/vmware/config` file on the ESXi host and add the following entry:
vGPU.consolidation = "true"
- 3D graphics is supported only when you use the PCoIP or VMware Blast protocol. Therefore, the farm must use PCoIP or VMware Blast as the default protocol, and users must not be allowed to choose the protocol.
- Configuration of 3D graphics for RDS hosts in the VMware View Administrator is not required. Selection of the option 3D Remote Desktop Session Host (RDSH) when you install Horizon Agent is sufficient. By default, this option is not selected, and 3D graphics is disabled.

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/>.

VMware recommends Blast Extreme for most use cases. It is required for connections to Linux desktops and for HTML access. Linux desktops use the JPG or PNG codec, and HTML access uses the JPG or PNG codec except for Chrome browsers, which can be configured to use the H.264 codec. For a detailed description of these codecs, see [Codecs Used by Blast Extreme](#).

The only end users who should continue to use PCoIP rather than Blast Extreme are users of zero-client devices that are specifically manufactured to support PCoIP. For a list of zero and thin clients that support Blast Extreme, see the [VMware Compatibility Guide](#).

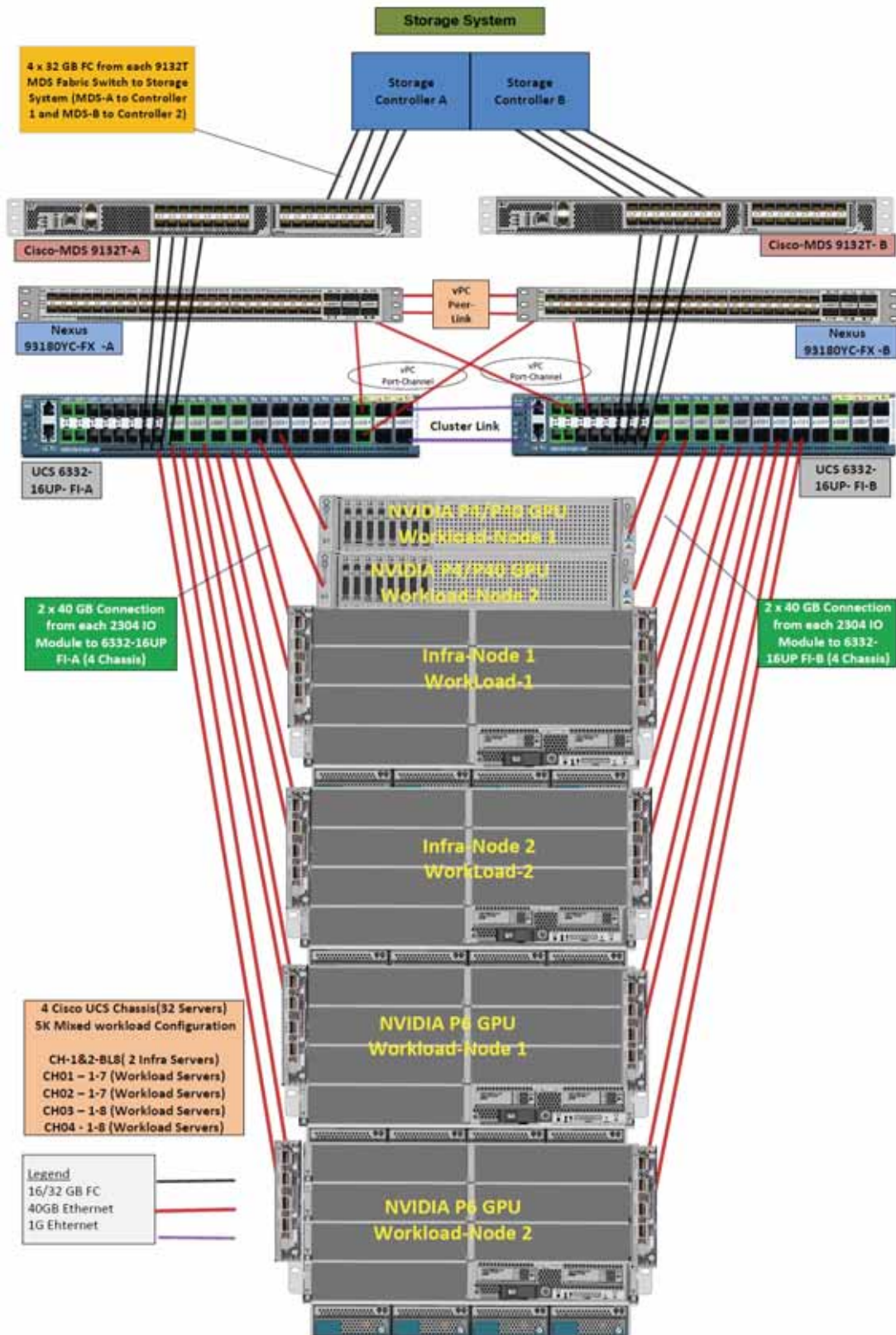
Note: If you configure a pool to use Blast Extreme and do not allow users to choose a protocol, View Connection Server automatically allows PCoIP connections from PCoIP zero clients and older (earlier than Release 4.0) Horizon Clients.

When used in an NVIDIA GRID vGPU solution, Blast Extreme outperforms PCoIP for 3D rendering in graphics-intensive applications, and it can enable hardware encoding in addition to hardware decoding. For a performance comparison of PCoIP and Blast Extreme, see the blog post [VMware Horizon Blast Extreme Acceleration with NVIDIA GRID](#).

Solution configuration

Figure 10 provides an overview of the physical connectivity configuration of the Cisco UCS data center solution. The solution is described in a detail in the Cisco Validated Design. [See the Cisco VDI landing page for details](#). This architecture was used to verify the performance of NVIDIA Tesla graphic cards using SPECviewperf 13, VMware Horizon with the Blast display protocol, and the VMware ESXi hypervisor, and to create this document.

Figure 10. Cabling diagram for a Cisco data center with Cisco UCS



The hardware components in the solution are as follows:

- Cisco UCS B200 M5 Blade Servers with Intel Xeon Silver 4114 2.20-GHz 10-core processors and 192-GB 2666-MHz RAM for infrastructure
- Cisco UCS B200 M5 Blade Servers with Intel Xeon Gold 6140 2.30-GHz 18-core processors, 768-GB 2666-MHz RAM, and two NVIDIA Tesla P6 GPUs for graphics accelerated virtual client computing workloads
- Cisco UCS C240 M5 Rack Server with Intel Xeon Gold 6140 2.30-GHz 18-core processors, 768-GB 2666-MHz RAM, and six NVIDIA Tesla P4 or P40 GPUs for graphics accelerated virtual client computing workloads
- Cisco UCS VIC 1387 mLOM (Cisco UCS C240 M5)
- Cisco UCS VIC 1340 mLOM (Cisco UCS B200 M5)
- Cisco Nexus® 93180YC-FX Switches in Cisco NX-OS mode for Layer 2 communications
- Cisco MDS 9132T 32G Multilayer Fabric Switch for Fibre Channel connectivity (optional)
- Customer's choice of storage system

The software components of the solution are as follows:

- Cisco UCS Firmware Release 4.0(1a)
- VMware vSphere ESXi 6.7 for VDI hosts
- VMware Horizon 7.5
- Microsoft Windows 10 64-bit
- Microsoft Server 2016
- SPECviewperf 13 graphics benchmark software and commercial license (optional.)
- NVIDIA GRID 6.2 software and licenses:
 - NVIDIA-VMware_ESXi_6.7_Host_Driver-390.72-1OEM.670.0.0.8169922.vib

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 the tabs that need to be modified to achieve optimal BIOS configuration.

Refer to [Performance Tuning Guide for Cisco UCS M5 Servers](#).

Create graphics card policy

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

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

Figure 11. Graphics card policy



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

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

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

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

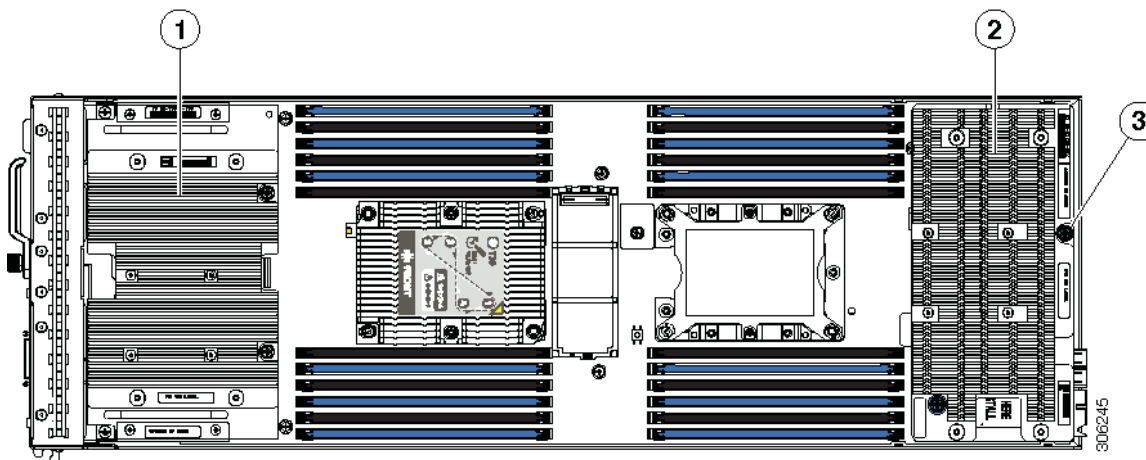
- The UCSB-GPU-P6-F card can be installed only in the front mezzanine slot of the server.

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

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

Figure 12 shows the installed NVIDIA P6 GPU in the front and rear mezzanine slots.

Figure 12. NVIDIA GPU installed in the front and rear mezzanine slots

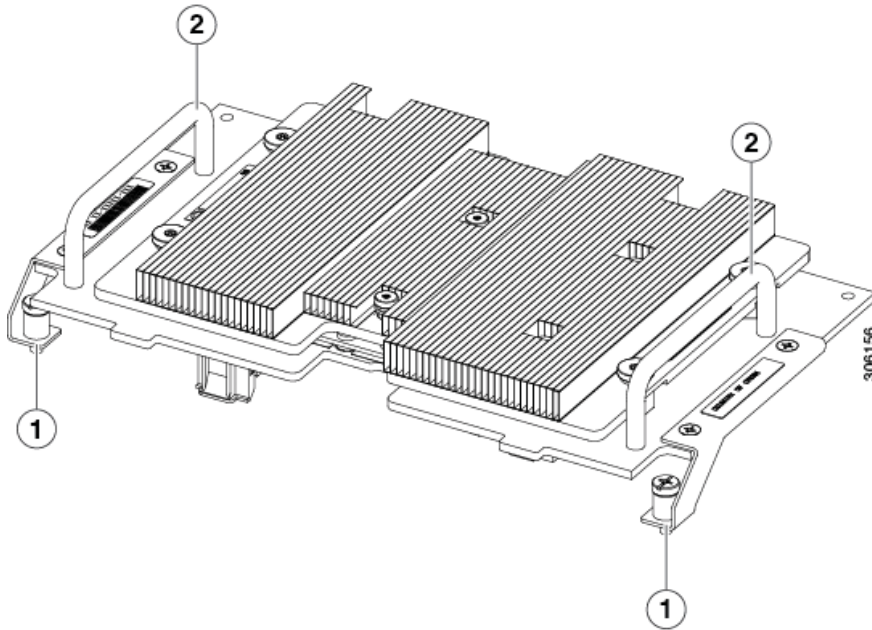


1	Front GPU	2	Rear GPU
3	Custom standoff screw	-	

Installing an NVIDIA GPU card in the front of the server

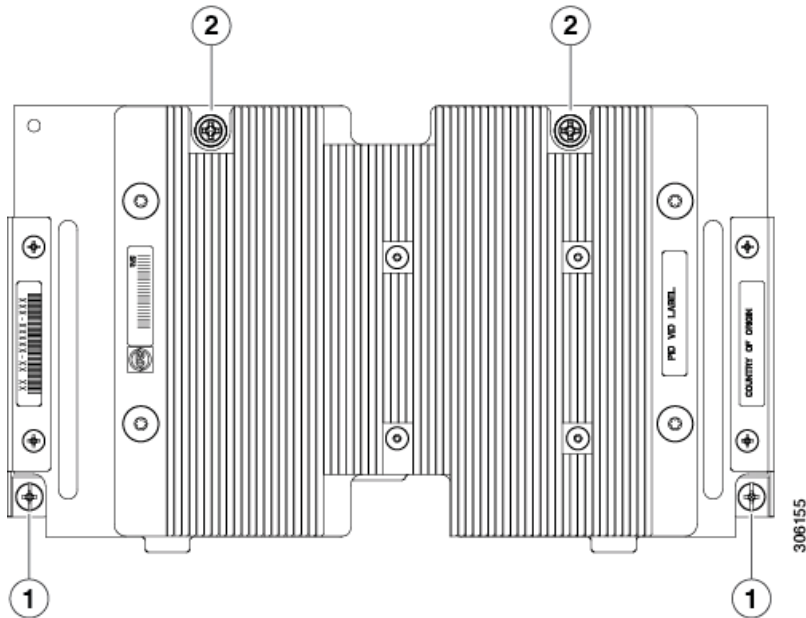
Figure 13 shows the front NVIDIA P6 GPU (UCSB-GPU-P6-F), and Figure 14 shows the top view.

Figure 13. NVIDIA P6 GPU that installs in the front of the server



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

Figure 14. Top view of the NVIDIA P6 GPU in the front of the server



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

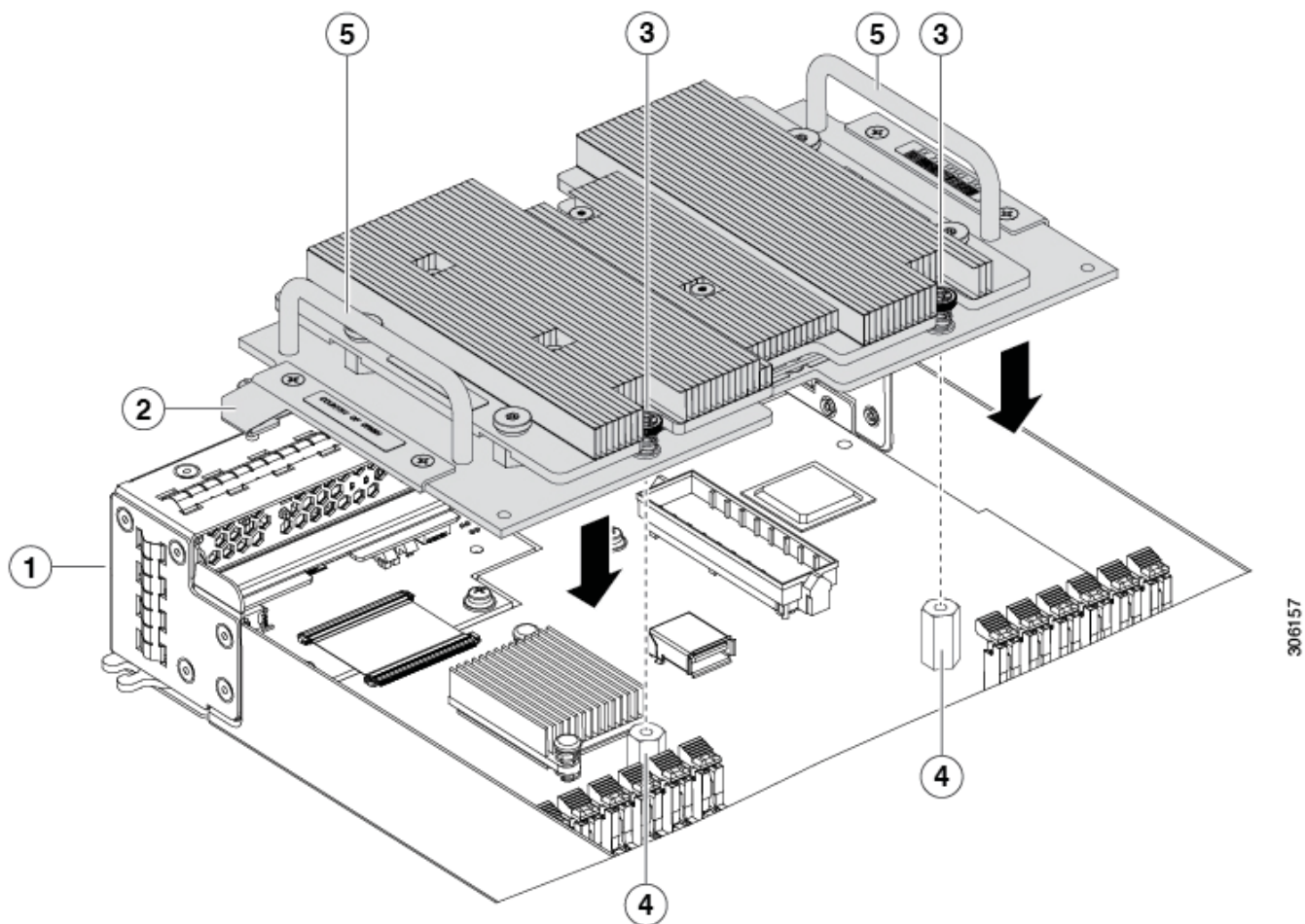
To install the NVIDIA P6 GPU, follow the steps presented here.

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

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

1. Position the GPU in the correct orientation relative to the front of the server (number 1) as shown in Figure 15.
2. Install the GPU in the server. Press down on the handles (number 5) to firmly secure the GPU.
3. Tighten the thumb screws (number 3) at the back of the GPU with the standoffs (number 4) on the motherboard.
4. Tighten the thumb screws on the legs (number 2) of the motherboard.
5. Install the drive blanking panels.

Figure 15. Installing the NVIDIA GPU in the front of the server

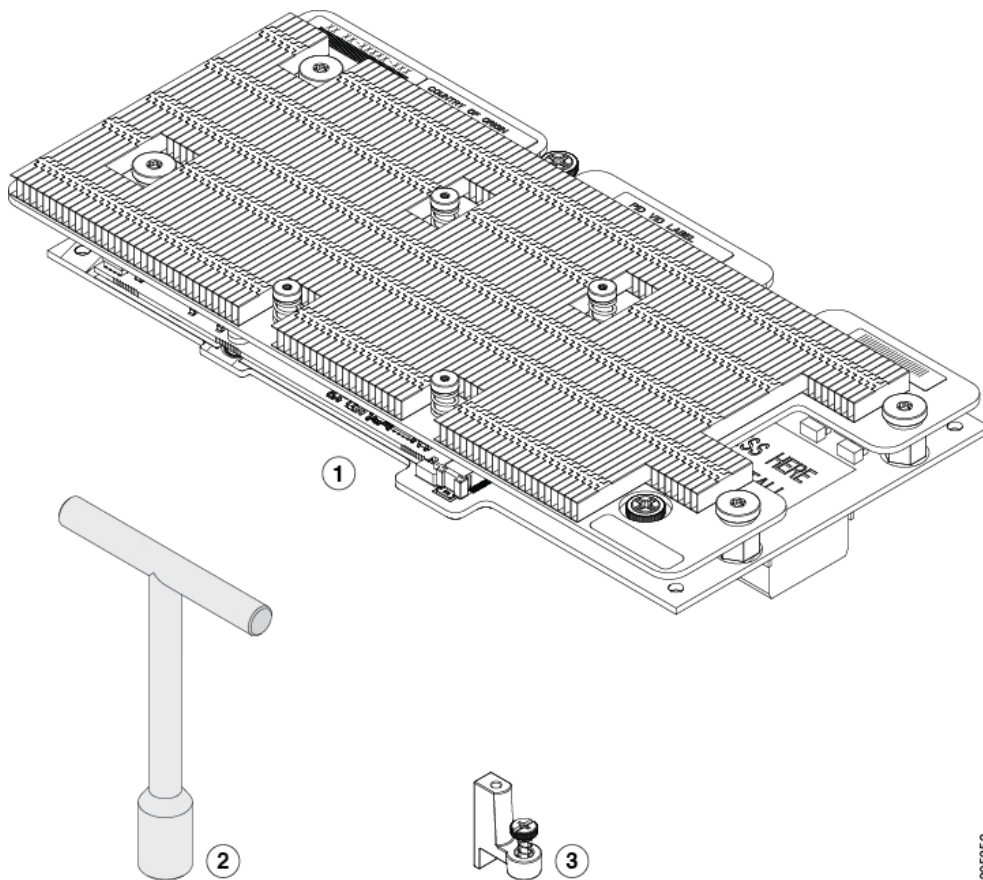


1	Front of the server	2	Leg with thumb screw that attaches to the motherboard
3	Thumbscrew to attach to standoff below	4	Standoff on the motherboard
5	Handle to press down on to firmly install the GPU	-	

Installing an NVIDIA GPU card in the rear of the server

If you are installing the UCSB-GPU-P6-R on a server in the field, the option kit comes with the GPU itself (CPU and heat sink), a T-shaped installation wrench, and a custom standoff to support and attach the GPU to the motherboard. Figure 16 shows the three components of the option kit.

Figure 16. NVIDIA P6 GPU (UCSB-GPU-P6-R) option kit



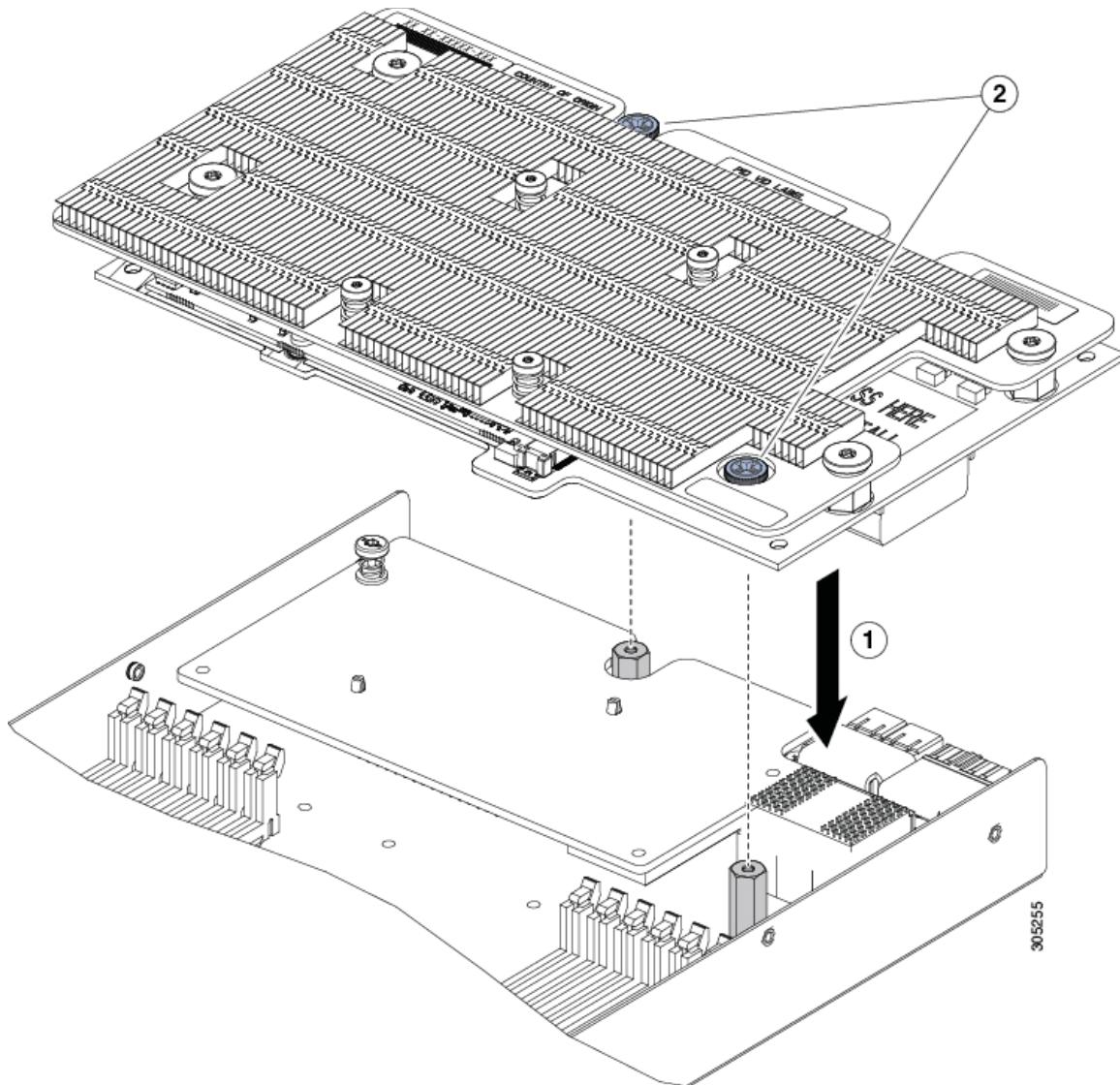
1	NVIDIA P6 GPU (CPU and heatsink)	2	T-shaped wrench
3	Custom standoff	-	

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

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

1. Use the T-shaped wrench that comes with the GPU to remove the existing standoff at the back end of the motherboard.
2. Install the custom standoff in the same location at the back end of the motherboard (Figure 17).
3. Position the GPU over the connector on the motherboard and align all the captive screws with the standoff posts (number 1).
4. Tighten the captive screws (number 2).

Figure 17. Installing the NVIDIA P6 GPU in the rear mezzanine slot



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

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

Physically installing an NVIDIA Tesla P4 card

Use the following procedure to install NVIDIA Tesla P4:

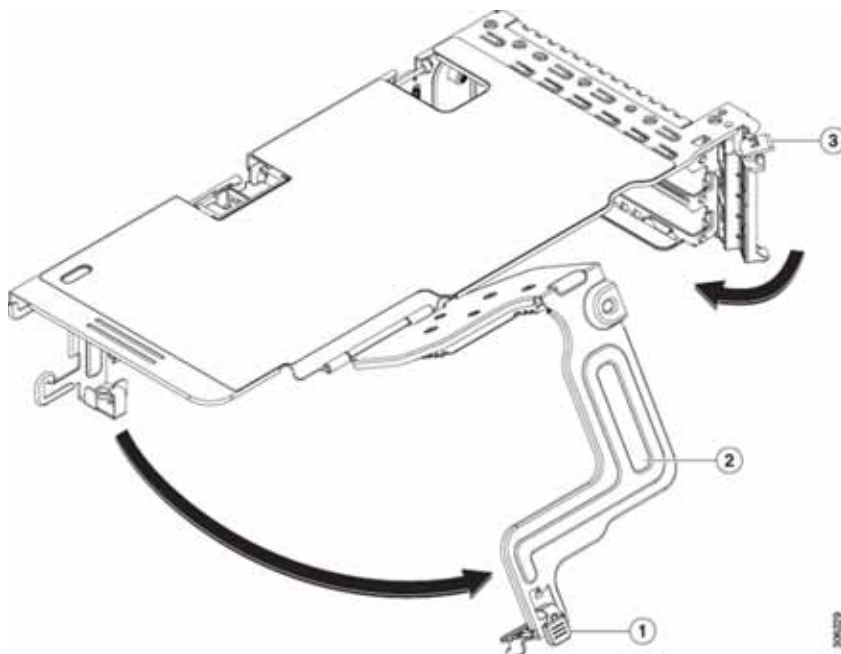
Note: This server can support up to six single-wide NVIDIA Tesla P4 GPU cards. These half-height, half-length (HHHL) GPU cards are supported in all PCIe slots.

1. Shut down and remove power from the server.
2. Slide the server out the front of the rack far enough so that you can remove the top cover. You may have to detach cables from the rear panel to provide clearance.
3. Remove the top cover from the server.
4. Install a new single-wide GPU card (Figure 18):

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

- a. With the hinged card-tab retainer open, align the new single-wide GPU card with the empty socket on the PCIe riser.
 - b. Push down evenly on both ends of the card until it is fully seated in the socket.
 - c. Verify that the card’s rear panel tab sits flat against the riser rear-panel opening and then close the hinged card-tab retainer over the card’s rear-panel tab.
 - d. Swing the hinged securing plate closed on the bottom of the riser. Verify that the clip on the plate clicks into the locked position.
 - e. Position the PCIe riser over its socket on the motherboard and over the chassis alignment channels.
 - f. Carefully push down on both ends of the PCIe riser to fully engage its connector with the sockets on the motherboard.
5. Replace the top cover to the server.
 6. Replace the server in the rack, replace cables, and then fully power on the server by pressing the Power button.

Figure 18. PCIe riser card securing mechanism



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

Installing a double-wide GPU card

Use the following procedure to install NVIDIA Tesla P40 card:

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

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

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

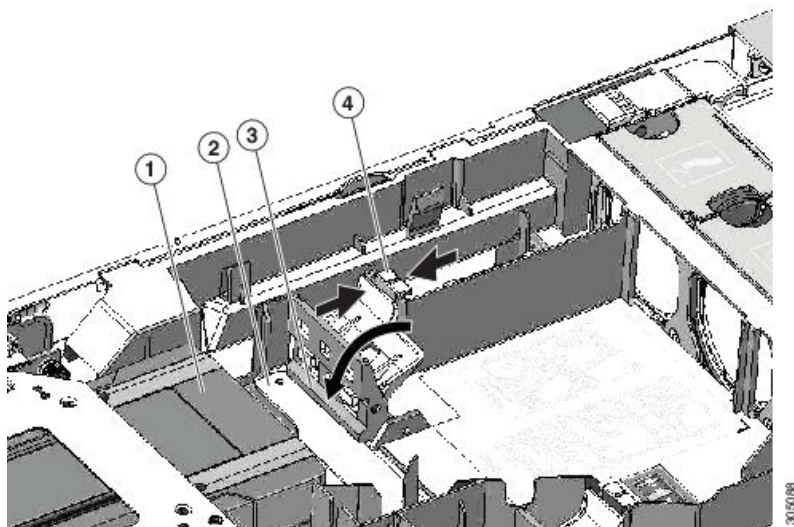
Note: Observe the configuration rules for this server, as described in [GPU Card Configuration Rules](#).

- 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. Verify that the clip on the plate clicks into the locked position.
- e. Position the PCIe riser over its socket on the motherboard and over the chassis alignment channels.
- f. Carefully push down on both ends of the PCIe riser to fully engage its connector with the sockets on the motherboard.
- 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 19):
 - Pinch the latch release tab and hinge the latch toward the front of the server.
 - Hinge the latch back down so that its lip closes over the edge of the GPU front support bracket.
 - Verify that the latch release tab clicks and locks the latch in place.

Figure 19. 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

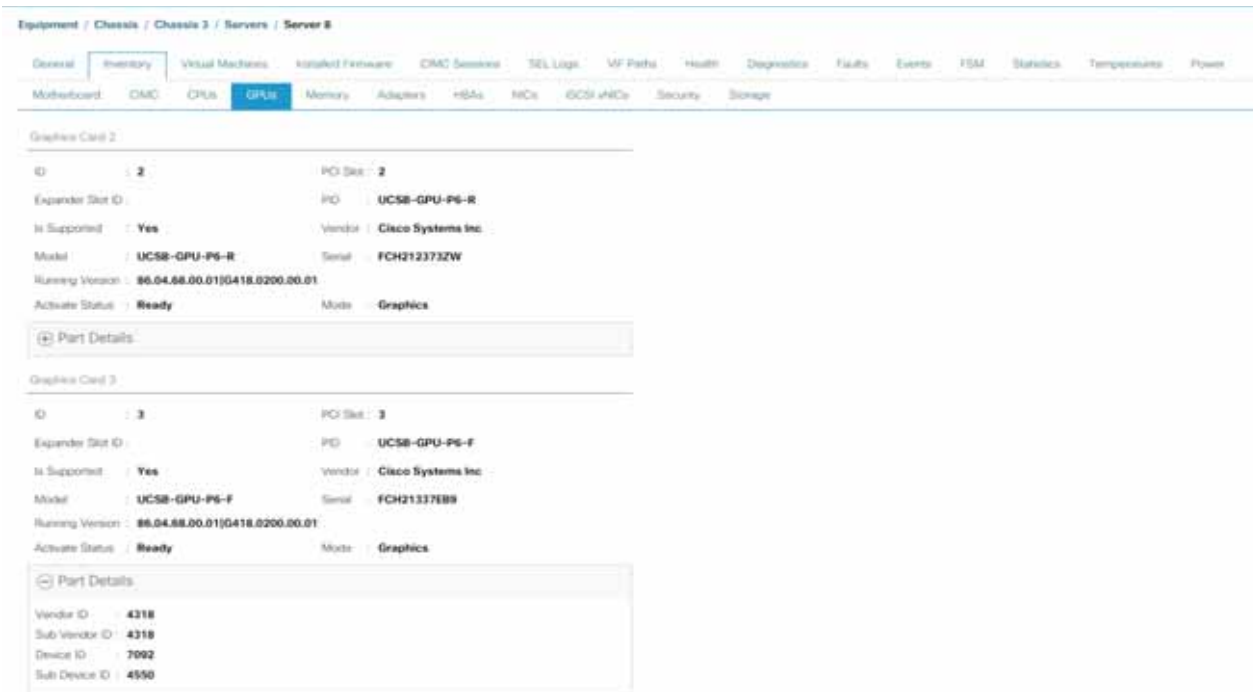
5. Replace the top cover to the server.
6. Replace the server in the rack, replace cables, and then fully power on the server by pressing the Power button.

Configure the GPU card

Follow these steps to configure the GPU card:

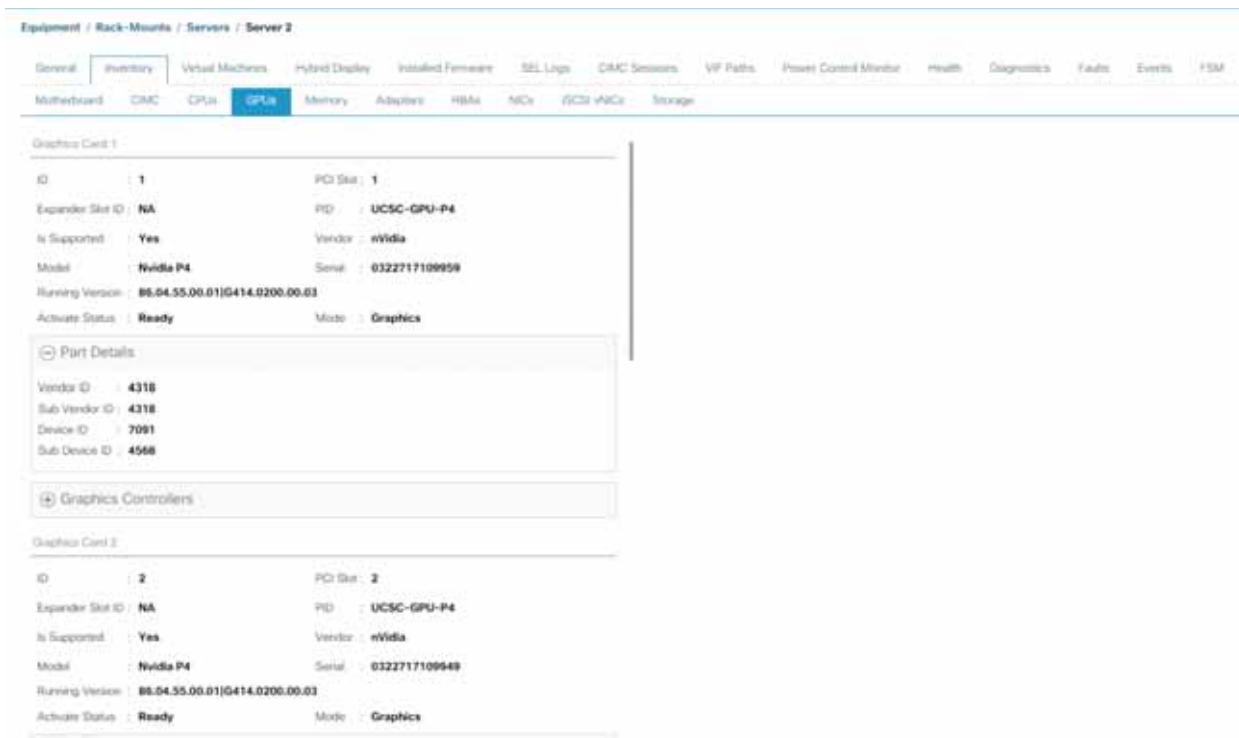
1. After the NVIDIA P6 GPU cards are physically installed and the Cisco UCS B200 M5 Blade Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 20, PCIe slots 2 and 3 are used with two GRID P6 cards.

Figure 20. NVIDIA GRID P6 card inventory displayed in Cisco UCS Manager



2. After the NVIDIA P4 GPU cards are physically installed and the Cisco UCS C240 M5 Rack Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 21, PCIe slots 1 through 6 are used with six GRID P4 cards.

Figure 21. NVIDIA GRID P4 card inventory displayed in Cisco UCS Manager



Equipment / Rack-Mounts / Servers / Server 2

General | **Inventory** | Virtual Machines | Hybrid Display | Installed Firmware | SBL Logs | CIMC Sensors | VIF Paths | Power Control Monitor | Health | Diagnostics | Faults | Events | FSM

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

Graphics Card 1

ID	1	PCI Slot	1
Expander Slot ID	NA	PID	UCSC-GPU-P4
Is Supported	Yes	Vendor	nVidia
Model	Nvidia P4	Serial	032271710959
Running Version	B6.04.55.00.01 G414.0200.00.03		
Activate Status	Ready	Mode	Graphics

Part Details

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

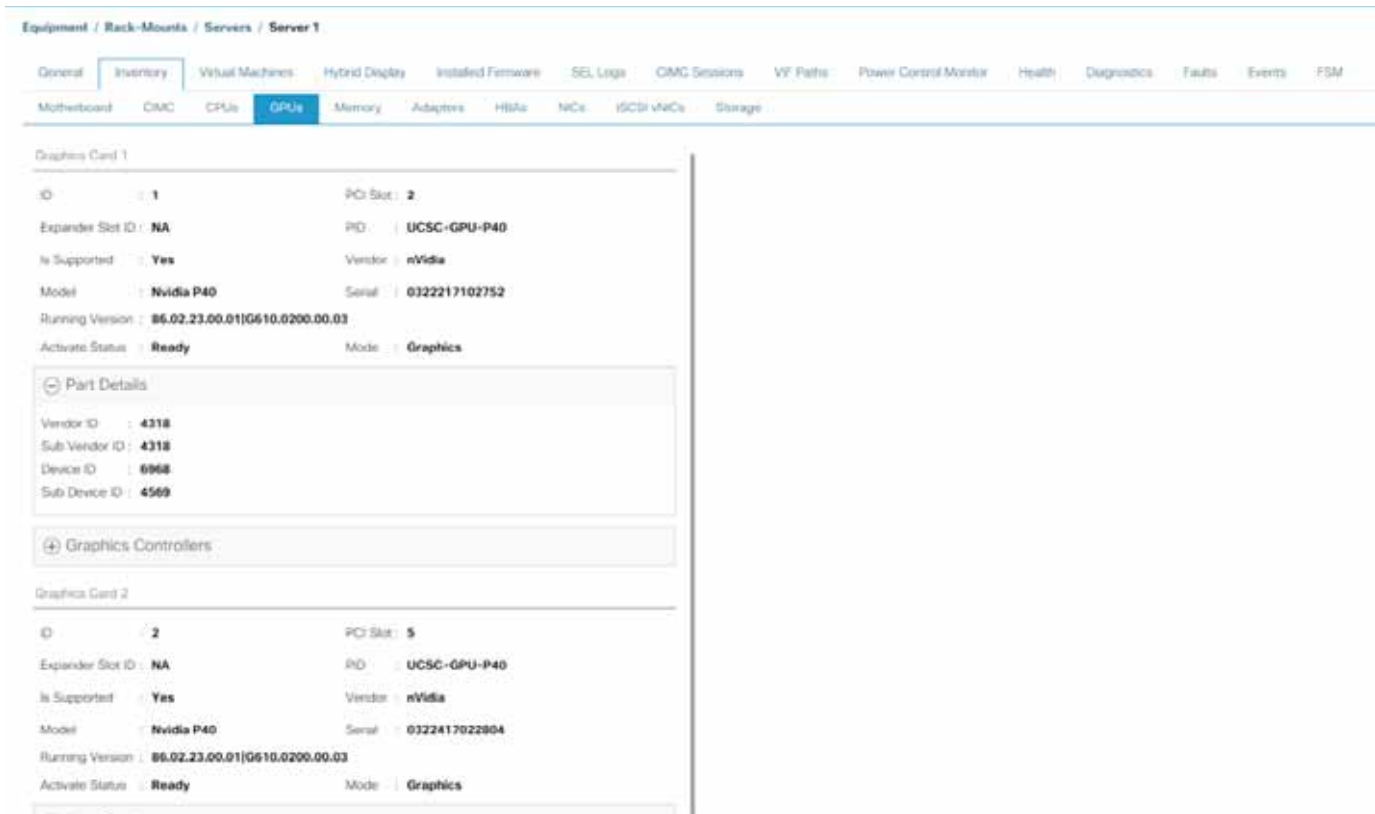
Graphics Controllers

Graphics Card 2

ID	2	PCI Slot	2
Expander Slot ID	NA	PID	UCSC-GPU-P4
Is Supported	Yes	Vendor	nVidia
Model	Nvidia P4	Serial	032271710959
Running Version	B6.04.55.00.01 G414.0200.00.03		
Activate Status	Ready	Mode	Graphics

- 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 22, PCIe slot 2 and slot 5 are used with the two GRID P40 cards.

Figure 22. NVIDIA GRID P40 card inventory displayed in Cisco UCS Manager



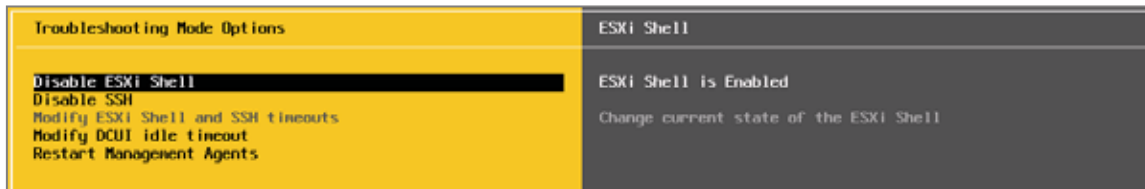
You can use Cisco UCS Manager to perform firmware upgrades to the NVIDIA GPU cards in managed Cisco UCS C240 M5 servers.

Install NVIDIA GRID software on the VMware ESXi host

This section summarizes the installation process for configuring an ESXi host and virtual machine for vGPU support.

1. Download the NVIDIA GRID GPU driver pack for VMware vSphere ESXi 6.7.
2. Enable the ESXi shell and the Secure Shell (SSH) protocol on the vSphere host from the Troubleshooting Mode Options menu of the vSphere Configuration Console (Figure 23).

Figure 23. VMware ESXi Configuration Console



1. Upload the NVIDIA driver (vSphere Installation Bundle [VIB] file) to the /tmp directory on the ESXi host using a tool such as WinSCP. (Shared storage is preferred if you are installing drivers on multiple servers or using the VMware Update Manager.)
2. Log in as root to the vSphere console through SSH using a tool such as Putty.
3. The ESXi host must be in maintenance mode for you to install the VIB module. To place the host in maintenance mode, use this command:

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


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

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

The command should return output similar to that shown here:

```
[root@C240M5-GPU:~] esxcli software vib install -v /tmp/ NVIDIA-VMware_ESXi_6.7_Host_Driver-390.72-1OEM.670.0.0.8169922.vib --no-sig-check
```

Installation Result

Message: Operation finished successfully.

Reboot Required: false

VIBs Installed: NVIDIA-VMware_ESXi_6.7_Host_Driver-390.72-1OEM.670.0.0.8169922.vib

VIBs Removed:

VIBs Skipped:

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

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

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

```
#reboot
```

3. After the host reboots successfully, verify that the kernel module has loaded successfully by entering the following command:

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

The command should return output similar to that shown here:

```
[root@C240M5-GPU:~] esxcli software vib list | grep -i NVidia
```

```
NVIDIA-VMware_ESXi_6.7_Host_Driver 390.72-1OEM.670.0.0.8169922
```

```
NVIDIA
```

```
VMwareAccepted 2018-08-03
```

See the VMware knowledge base article for information about removing any existing NVIDIA drivers before installing new drivers:

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

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 Figure 24.

Figure 24. List of GPUs

```

root@C240M5-GPU02:~] nvidia-smi
Tue Oct 30 19:27:08 2018
-----+-----+-----+-----+-----+-----+-----+-----+
NVIDIA-SMI 390.72                Driver Version: 390.72
-----+-----+-----+-----+-----+-----+-----+-----+
GPU  Name          Persistence-M| Bus-Id          Disp.A | Volatile Uncorr. ECC |
Fan  Temp   Perf    Pwr:Usage/Cap| Memory-Usage | GPU-Util  Compute M. |
-----+-----+-----+-----+-----+-----+-----+-----+
  0   Tesla P4             On          | 00000000:19:00.0 Off  |             Off      |
N/A   37C    P8      11W / 75W | 8156MiB / 8191MiB |          0%   Default |
-----+-----+-----+-----+-----+-----+-----+-----+
  1   Tesla P4             On          | 00000000:5E:00.0 Off  |             Off      |
N/A   36C    P8      11W / 75W | 28MiB / 8191MiB  |          0%   Default |
-----+-----+-----+-----+-----+-----+-----+-----+
  2   Tesla P4             On          | 00000000:86:00.0 Off  |             Off      |
N/A   35C    P8      11W / 75W | 28MiB / 8191MiB  |          0%   Default |
-----+-----+-----+-----+-----+-----+-----+-----+
  3   Tesla P4             On          | 00000000:AF:00.0 Off  |             Off      |
N/A   38C    P8      11W / 75W | 28MiB / 8191MiB  |          0%   Default |
-----+-----+-----+-----+-----+-----+-----+-----+
  4   Tesla P4             On          | 00000000:D8:00.0 Off  |             Off      |
N/A   35C    P8      11W / 75W | 28MiB / 8191MiB  |          0%   Default |
-----+-----+-----+-----+-----+-----+-----+-----+
  5   Tesla P4             On          | 00000000:D9:00.0 Off  |             Off      |
N/A   34C    P8      11W / 75W | 28MiB / 8191MiB  |          0%   Default |
-----+-----+-----+-----+-----+-----+-----+-----+

```

5. Repeat the process for all the hosts in the pool.

Modify GPU allocation policy

VMware vSphere supports the breadth-first and depth-first GPU allocation policies for vGPU-enabled virtual machines.

- Breadth-first allocation policy attempts to reduce the number of vGPUs running on each physical GPU. Newly created vGPUs are placed on the physical GPU that can support the new vGPU and that has the fewest vGPUs already resident on it. This policy generally leads to higher performance because it attempts to reduce the sharing of physical GPUs, but it may artificially limit the total number of vGPUs that can run.
- Depth-first allocation policy attempts to increase the number of vGPUs running on each physical GPU. Newly created vGPUs are placed on the physical GPU that can support the new vGPU and that has the most vGPUs already resident on it. This policy generally leads to higher density of vGPUs, particularly when different types of vGPUs are being run, but may result in lower performance because it attempts to increase the sharing of physical GPUs.

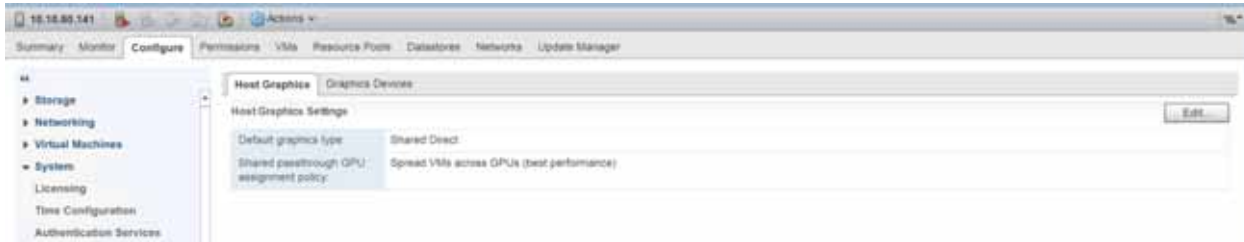
Note: VMware vSphere ESXi uses the breadth-first allocation policy by default.

Before using the vSphere Web Client to change the allocation scheme, verify that the ESXi host is running and that all virtual machines on the host are powered off.

1. Log in to vCenter Server by using the vSphere Web Client.
2. In the navigation tree, select your ESXi host and click the Configure tab.
3. From the menu, choose Graphics and then click the Host Graphics tab.
4. On the Host Graphics tab, click Edit.
5. In the Edit Host Graphics Settings dialog box that opens, select the options as shown in Figure 25; then click OK.
 - a. Select Shared Direct, if this option is not already selected.

b. Select Spread VMs across GPU until full.

Figure 25. Virtual machine distribution policy: Breadth first



6. To change the virtual machine allocation policy to depth-first, click the Edit button, select the “Group VMs on GPU until full” option, and then click OK (Figure 26).

Figure 26. Virtual machine distribution policy: Depth first



Install and configure the NVIDIA GRID license server

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

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

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

The following packages are required to set up the VMware environment (Figure 27):

- NVIDIA GRID license server installer
- NVIDIA GRID Manager software, which is installed on the VMware ESXi Hypervisor
- NVIDIA drivers and software that are installed in Microsoft Windows

Figure 27. Software required for NVIDIA GRID 6.2 setup on the VMware ESXi host

NVIDIA-ls-windows-2018.06.0.24304595.zip	8/6/2018 1:14 PM	Compressed (zipped)...	248,221 KB
NVIDIA-GRID-vSphere-6.7-390.72-390.75-391.8...	8/6/2018 1:15 PM	Compressed (zipped)...	1,128,788 KB
NVIDIA-GRID-VMware-vROps-1.0.zip	8/6/2018 1:14 PM	Compressed (zipped)...	17,437 KB
NVIDIA-gpumodeswitch-2016-04.zip	8/6/2018 1:14 PM	Compressed (zipped)...	98,782 KB

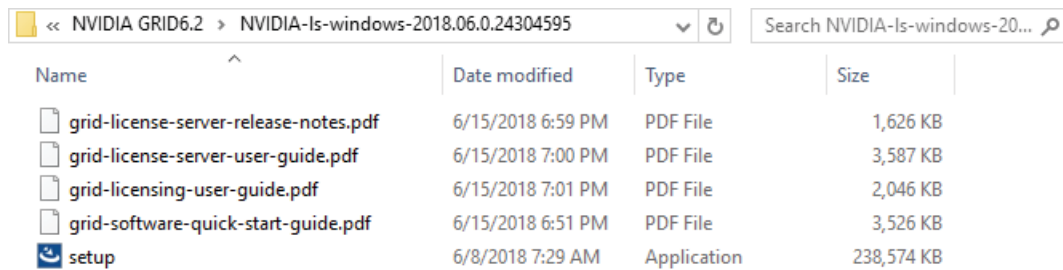
Install the NVIDIA GRID 6.2 license server

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

The GRID 6.2 license server requires Java Version 7 or later. Go to Java.com and install the latest version.

1. Extract and open the NVIDIA-Is-windows- $\$$ version folder. Run setup.exe (Figure 28).

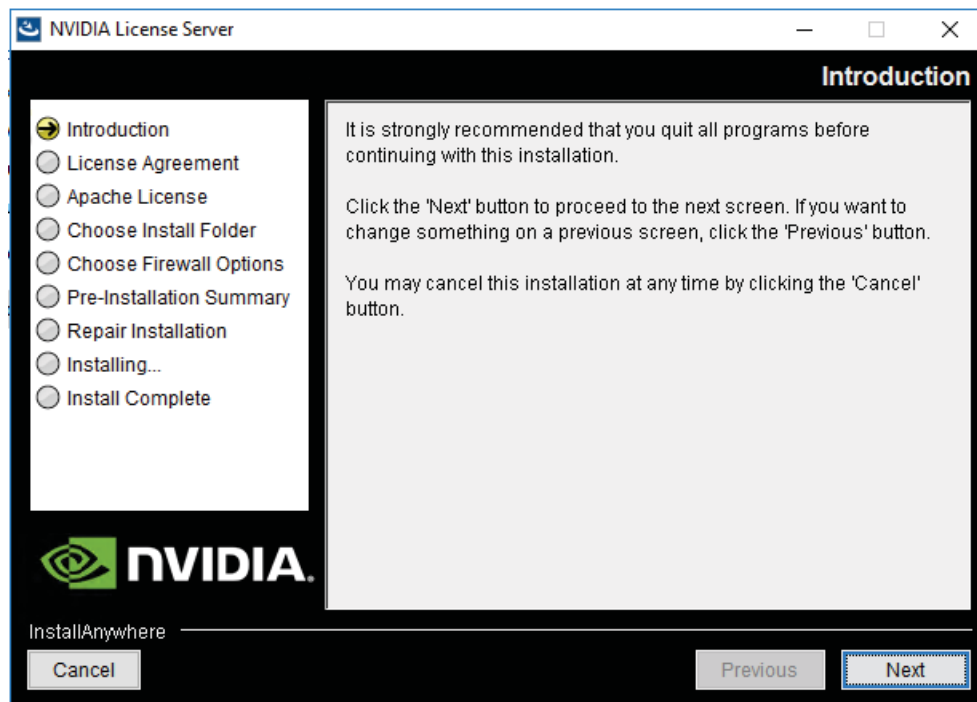
Figure 28. Run setup.exe



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

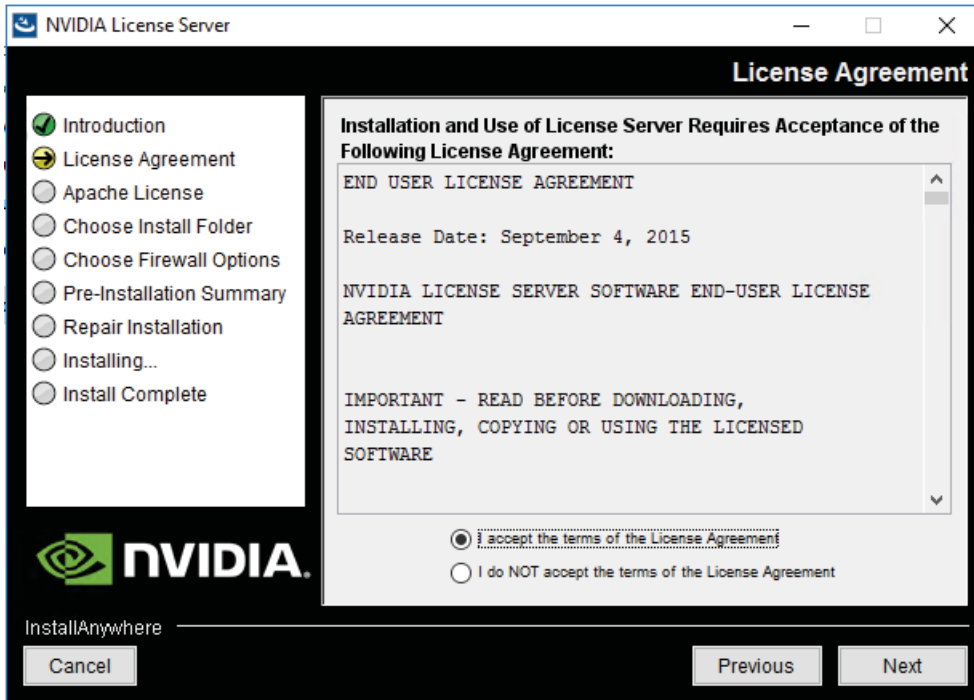
2. Click Next (Figure 29).

Figure 29. NVIDIA License Server page



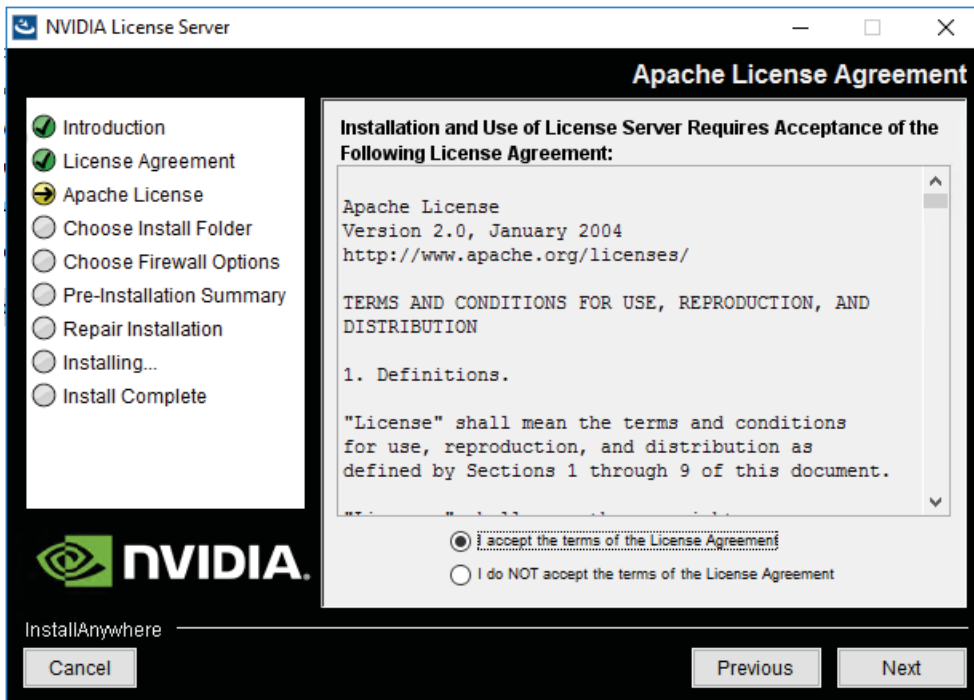
3. Accept the license agreement and click Next (Figure 30).

Figure 30. NVIDIA License Agreement page

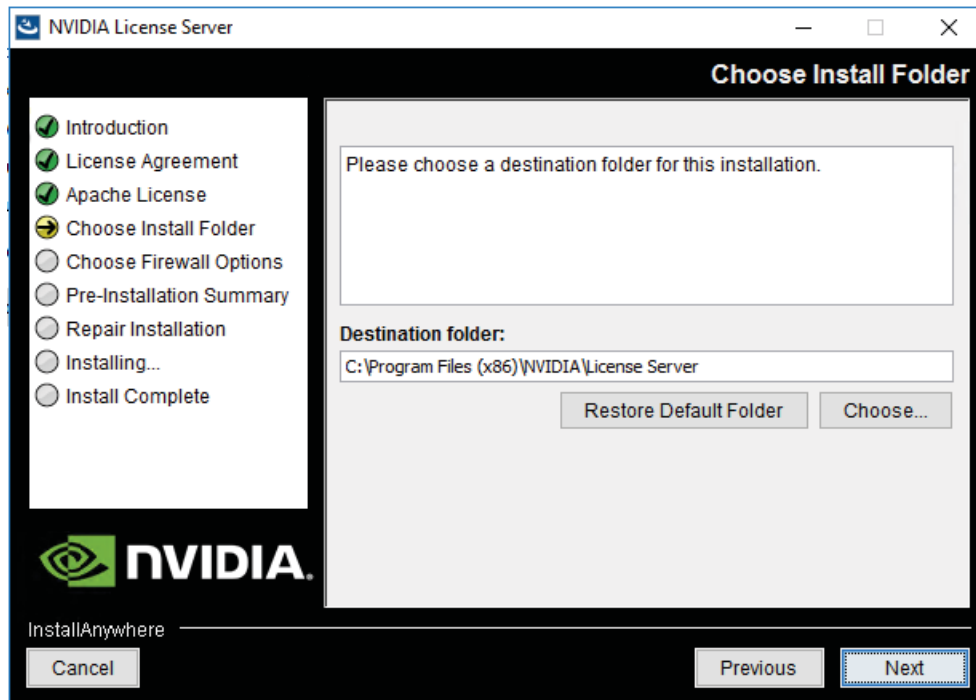


4. Accept the Apache license agreement and click Next (Figure 31).

Figure 31. Apache License Agreement page

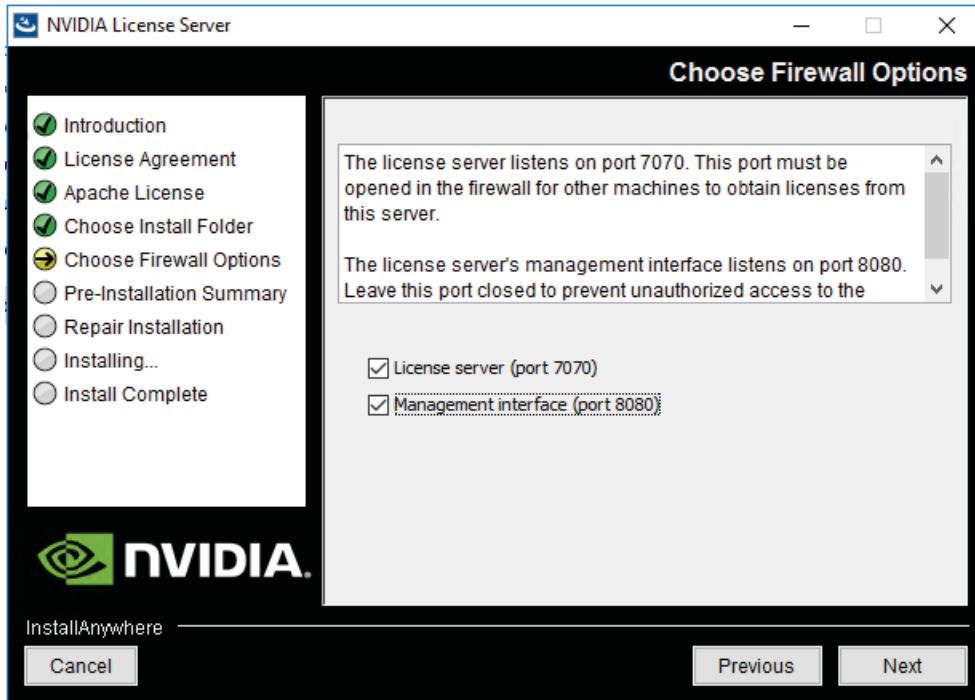


5. Choose the desired installation folder and click Next (Figure 32).

Figure 32. Choosing a destination folder

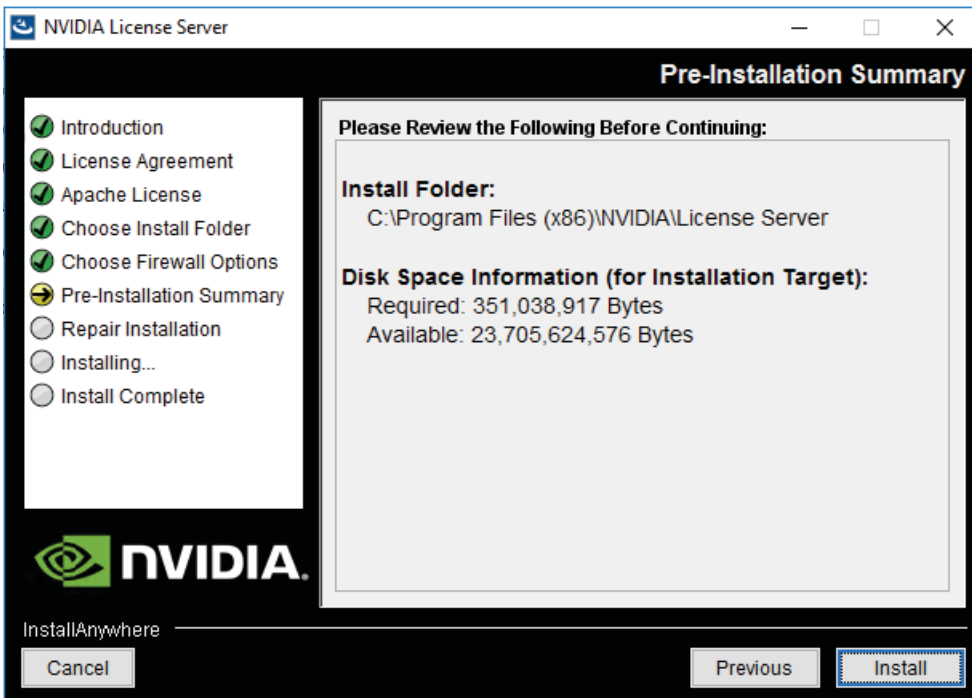
6. The license server listens on port 7070. This port must be opened in the firewall for other machines to obtain licenses from this server. Select the “License server (port 7070)” option.
7. The license server’s management interface listens on port 8080. If you want the administration page accessible from other machines, you need to open port 8080. Select the “Management interface (port 8080)” option.
8. Click Next (Figure 33).

Figure 33. Setting firewall options



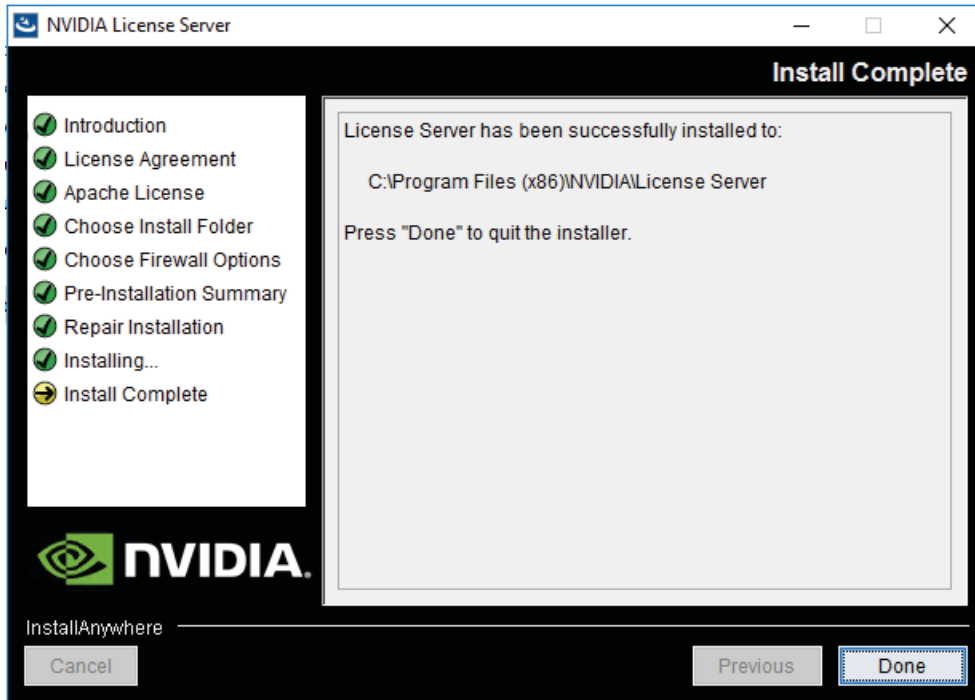
9. The Pre-installation Summary and Repair Installation options automatically progress without user input (Figure 34).

Figure 34. Installing the license server



10. When the installation process is complete, click Done (Figure 35).

Figure 35. Installation complete



Configure the NVIDIA GRID 6.2 license server

Now configure the NVIDIA GRID license server.

1. Log in to the license server site with the credentials set up during the registration process at nvidia.com/grideval. A license file is generated from <https://nvidia.flexnetoperations.com>.
2. After you are logged in, click Register License Server.
3. Specify the fields as shown in Figure 36. In the License Server ID field, enter the MAC address of your local license server’s NIC. Leave ID Type set to Ethernet. For Alias and Site Name, choose user-friendly names. Then click Create.

Figure 36. Registering the license server



4. Click the Search License Servers node.
5. Click your license server ID (Figure 37).

Figure 37. Selecting the license server ID



6. Click Map Add-Ons, choose the number of license units from your total pool to allocate to this license server, and click Map Add-Ons (Figure 38 and Figure 39).

Figure 38. Choosing the number of license units



Figure 39. Mapped add-ons after successful mapping



7. Click Download License File and save the .bin file to your license server (Figure 40).

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 40. Saving the .bin file



8. On the local license server, browse to `http://<FQDN>:8080/licserver` to display the License Server Configuration page.
9. Click License Management in the left pane.
10. Click Browse to locate your recently download .bin license file. Select the .bin file and click OK.
11. Click Upload. The message “Successfully applied license file to license server” should appear on the screen (Figure 41). The features are now available (Figure 42).

Figure 41. License file successfully applied

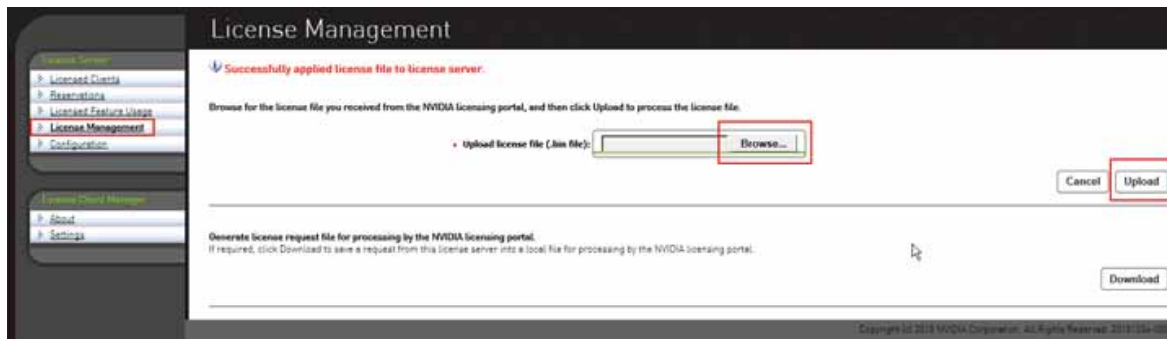
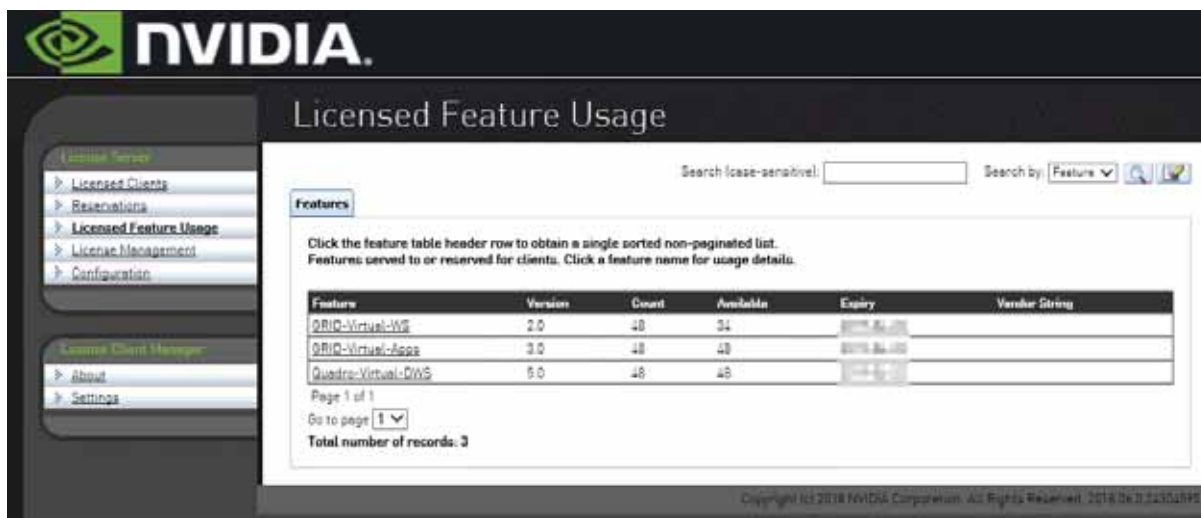


Figure 42. NVIDIA Licensed Feature Usage page



NVIDIA Tesla P6, P40, and P4 profile specifications

The Tesla P6, P4, and P40 cards have a single physical GPU. Each physical GPU can support several different types of vGPU. Each type of vGPU has a fixed amount of frame buffer space, a fixed number of supported display heads, and a fixed maximum resolution, and each is targeted at a different class of workload. Table 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 GRID options			
End-user profile	GRID virtual application profiles	GRID virtual PC profiles	NVIDIA Quadro Virtual Data Center Workstation (Quadro vDWS) profiles
1 GB	P6-1A P4-1A P40-1A	P6-1B P4-1B P40-1B	P6-1Q P4-1Q P40-1Q
2 GB	P6-2A P4-2A P40-2A	P6-2B P4-2B P40-2B	P6-2Q P4-2Q P40-2Q
3 GB	P40-3A	–	P40-3Q
4 GB	P6-4A P4-4A P40-4A	–	P6-4Q P4-4Q P40-4Q
6GB	P40-6A	–	P40-6Q
8 GB	P6-8A P4-8A P40-8A	–	P6-8Q P4-8Q P40-8Q
12 GB	P40-12A	–	P40-12Q
16 GB	P6-16A	–	P6-16Q
24 GB	P40-24A	–	P40-24Q

Create virtual desktops with vGPU support

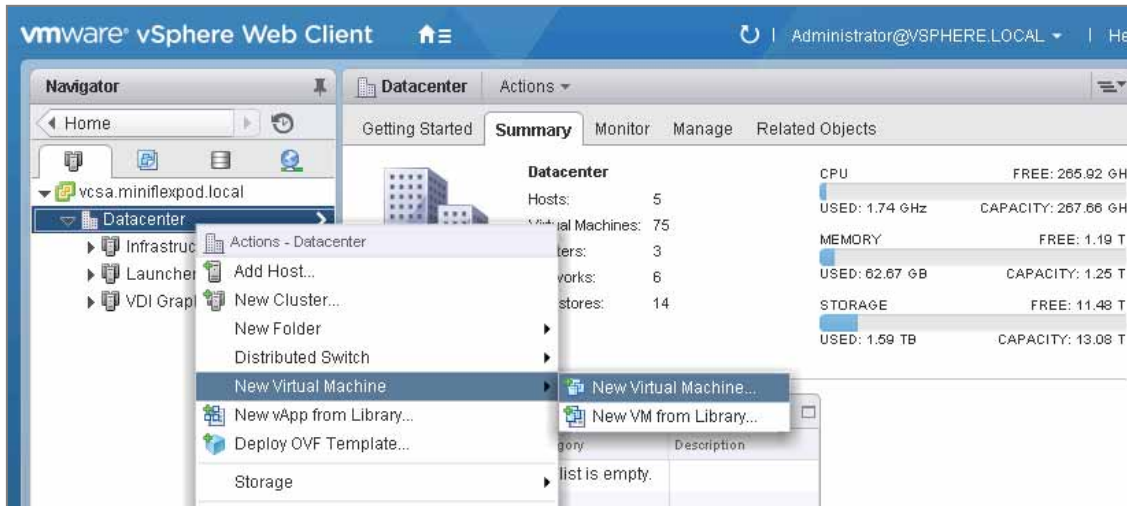
Now create virtual desktops with vGPU support.

Create the base image for virtual desktops

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

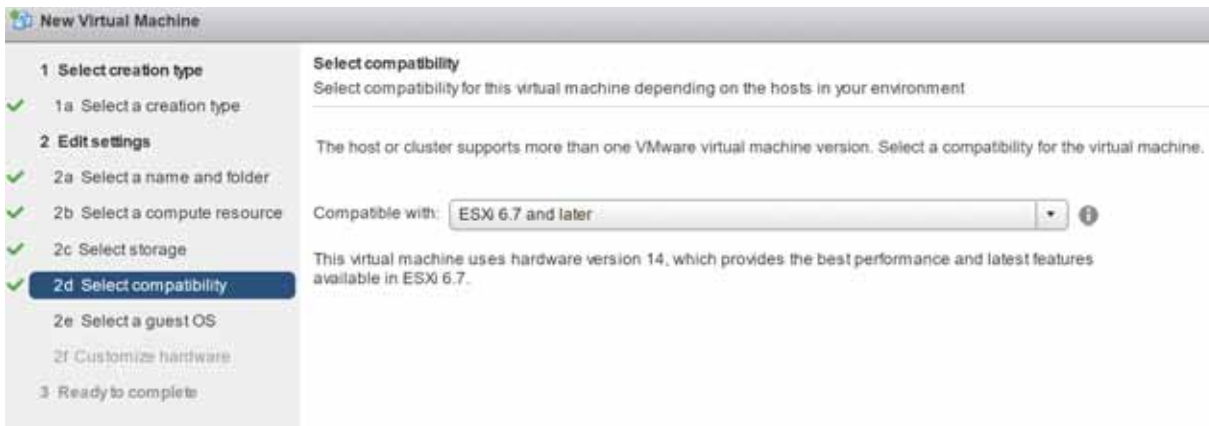
1. Using the vSphere Web Client or HTML5 interface (vSphere 6.7), create a new virtual machine. To do this, right-click a host or cluster and choose New Virtual Machine. Work through the New Virtual Machine wizard. Unless another configuration is specified, select the configuration settings appropriate for your environment (Figure 43).

Figure 43. Creating a new virtual machine in VMware vCenter Web Client



2. 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. “ESXi 6.7 and later” is used for this study, which provides the latest features available in ESXi 6.7 and virtual machine hardware Release 14 (Figure 44).

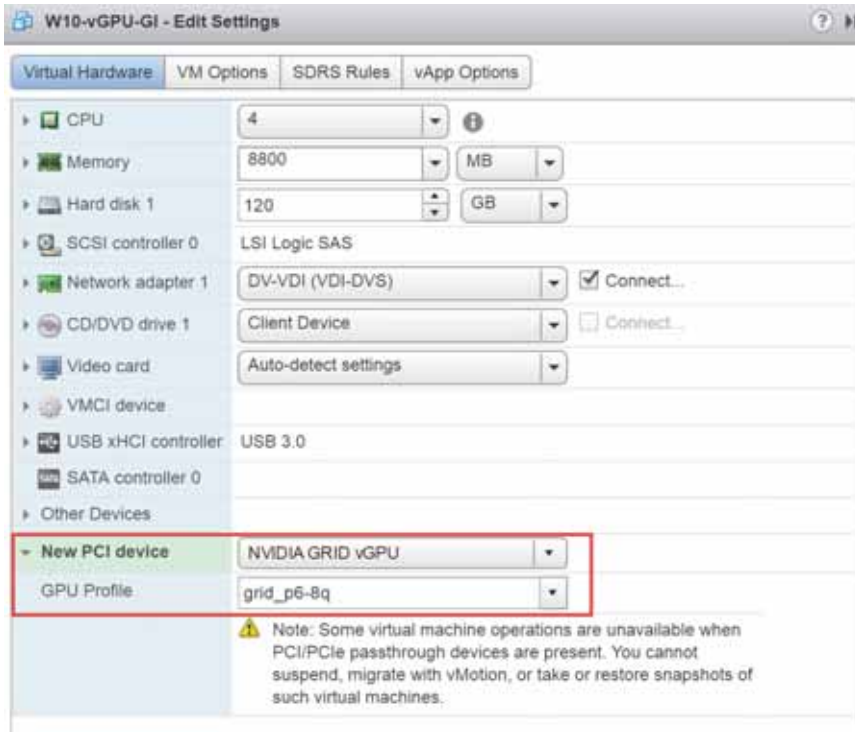
Figure 44. Selecting the virtual machine version and compatibility



3. In customizing the hardware of the new virtual machine, add a new shared PCI device, select the appropriate GPU profile, and reserve all virtual machine memory (Figure 45).

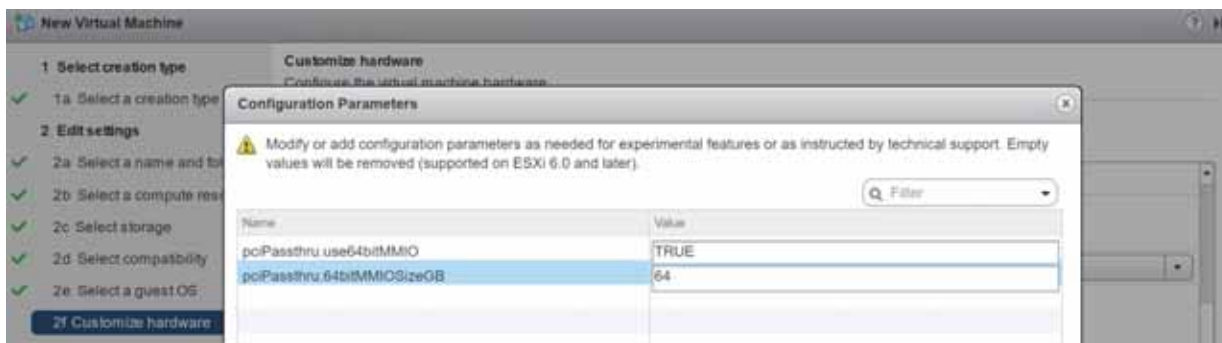
Note: vGPU allocation to a virtual machine requires the reservation of all guest memory. Click “Reserve all memory.”

Figure 45. Add a shared PCI device to a virtual machine to enable the vGPU profile



4. Open the VM Options tab, under Advanced choose Configuration Parameters, and click Edit Configuration. Add these values (Figure 46):
 - For pciPassthru.use64bitMMIO, set the value to TRUE.
 - For pciPassthru.64bitMMIOSizeGB, set the value to 64.

Figure 46. Setting configuration parameters



5. A virtual machine with vGPU assigned will not start if error-checking code (ECC) is enabled. As a workaround, disable ECC with the command `nvidia-smi -i x -e 0` (Figure 47, Figure 48, and Figure 49).

Note: Use option `-i` to target a specific GPU. If two cards are installed in a server, run the command twice as shown in Figure 47, where 0 and 1 each represent a different GPU card.

Figure 47. Disable ECC on NVIDIA Tesla GPU

```
[root@C240M5-GPU02:~] nvidia-smi -i 0 -e 0
Disabled ECC support for GPU 00000000:19:00.0.
All done.
Reboot required.
[root@C240M5-GPU02:~] nvidia-smi -i 1 -e 0
Disabled ECC support for GPU 00000000:5E:00.0.
All done.
Reboot required.
[root@C240M5-GPU02:~] nvidia-smi -i 2 -e 0
Disabled ECC support for GPU 00000000:86:00.0.
All done.
Reboot required.
[root@C240M5-GPU02:~] nvidia-smi -i 3 -e 0
Disabled ECC support for GPU 00000000:AF:00.0.
All done.
Reboot required.
[root@C240M5-GPU02:~] nvidia-smi -i 4 -e 0
Disabled ECC support for GPU 00000000:D8:00.0.
All done.
Reboot required.
[root@C240M5-GPU02:~] nvidia-smi -i 5 -e 0
Disabled ECC support for GPU 00000000:D9:00.0.
All done.
Reboot required.
```

Figure 48. Before disabling ECC on NVIDIA Tesla GPU: ECC value is 0, as shown in the nvidia-smi command

```
[root@C240M5-GPU02:~] nvidia-smi
Thu Oct 11 22:00:32 2018

+-----+
| NVIDIA-SMI 390.72                Driver Version: 390.72          |
+-----+-----+
| GPU  Name      Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp      Perf         Pwr:Usage/Cap| 20MiB / 7679MiB | GPU-Util  Compute M. |
+-----+-----+-----+-----+-----+-----+
| 0   Tesla P4   On          | 00000000:19:00.0 Off  |      0%      Default  |
| N/A   36C      P8         11W / 75W | 20MiB / 7679MiB |           |
+-----+-----+-----+-----+-----+-----+
| 1   Tesla P4   On          | 00000000:5E:00.0 Off  |      0%      Default  |
| N/A   35C      P8         11W / 75W | 20MiB / 7679MiB |           |
+-----+-----+-----+-----+-----+-----+
| 2   Tesla P4   On          | 00000000:86:00.0 Off  |      0%      Default  |
| N/A   34C      P8         11W / 75W | 20MiB / 7679MiB |           |
+-----+-----+-----+-----+-----+-----+
| 3   Tesla P4   On          | 00000000:AF:00.0 Off  |      0%      Default  |
| N/A   37C      P8         11W / 75W | 20MiB / 7679MiB |           |
+-----+-----+-----+-----+-----+-----+
| 4   Tesla P4   On          | 00000000:D8:00.0 Off  |      0%      Default  |
| N/A   34C      P8         11W / 75W | 20MiB / 7679MiB |           |
+-----+-----+-----+-----+-----+-----+
| 5   Tesla P4   On          | 00000000:D9:00.0 Off  |      0%      Default  |
| N/A   32C      P8         11W / 75W | 20MiB / 7679MiB |           |
+-----+-----+-----+-----+-----+-----+

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

Figure 49. After disabling ECC on NVIDIA Tesla GPU: ECC value is Off, as shown in the `nvidia-smi` command

```
[root@C240M5-GPU02:~] nvidia-smi
Thu Oct 11 22:00:32 2018

+-----+
| NVIDIA-SMI 390.72                Driver Version: 390.72          |
+-----+-----+
| GPU  Name           Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf    Pwr:Usage/Cap|  Memory-Usage | GPU-Util  Compute M. |
+-----+-----+
|  0   Tesla P4             On         | 00000000:19:00:0 Off |             Off      |
| N/A   36C    P8     11W / 75W |  20MiB / 7679MiB |    0%      Default   |
+-----+-----+
|  1   Tesla P4             On         | 00000000:5E:00:0 Off |             Off      |
| N/A   35C    P8     11W / 75W |  20MiB / 7679MiB |    0%      Default   |
+-----+-----+
|  2   Tesla P4             On         | 00000000:86:00:0 Off |             Off      |
| N/A   34C    P8     11W / 75W |  20MiB / 7679MiB |    0%      Default   |
+-----+-----+
|  3   Tesla P4             On         | 00000000:AF:00:0 Off |             Off      |
| N/A   37C    P8     11W / 75W |  20MiB / 7679MiB |    0%      Default   |
+-----+-----+
|  4   Tesla P4             On         | 00000000:D8:00:0 Off |             Off      |
| N/A   34C    P8     11W / 75W |  20MiB / 7679MiB |    0%      Default   |
+-----+-----+
|  5   Tesla P4             On         | 00000000:D9:00:0 Off |             Off      |
| N/A   32C    P8     11W / 75W |  20MiB / 7679MiB |    0%      Default   |
+-----+-----+
```

Install and configure Microsoft Windows on the virtual machine

Set up Microsoft Windows on the virtual machine.

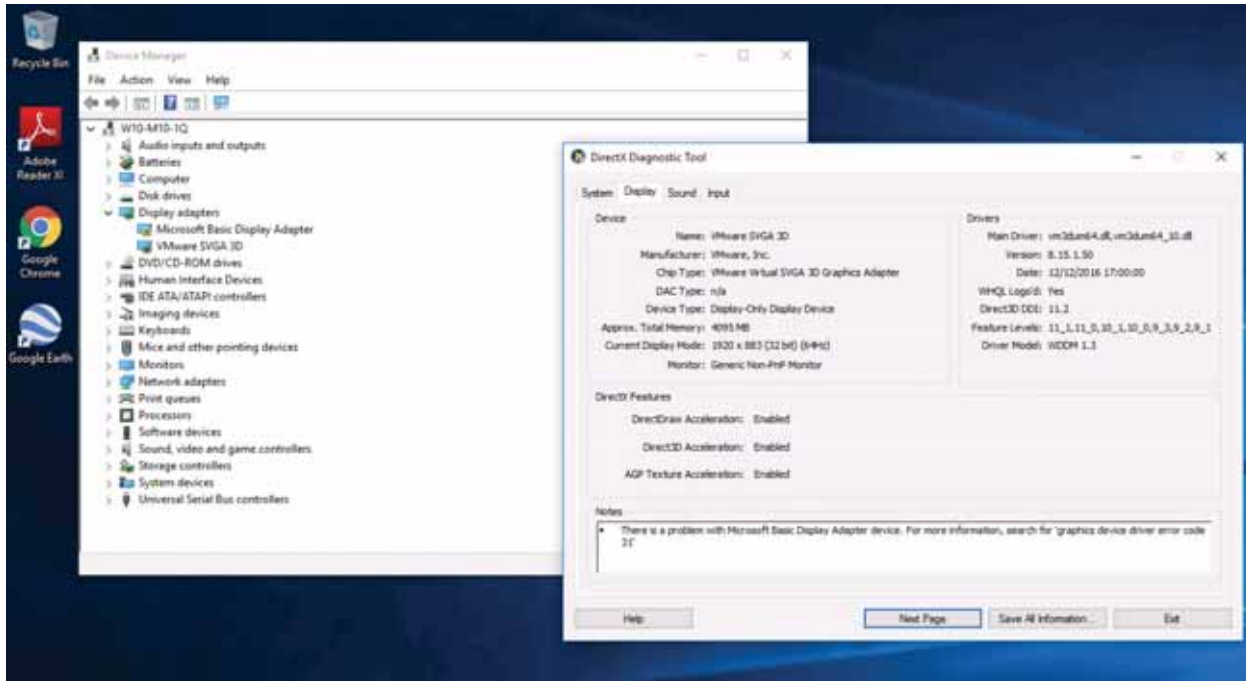
1. Configure the virtual machine with the appropriate amount of vCPU and RAM according to the GPU profile selected.
2. Install VMware Tools.
3. Join the virtual machine to the Microsoft Active Directory domain.
4. Choose “Allow remote connections to this computer” from the Windows System Properties menu.
5. Install VMware Horizon Agent with appropriate settings. Enable the remote desktop capability if prompted to do so.
6. Install Horizon Direct Connection Agent.
7. Optimize the Windows OS. [VMware OSOT](#), the optimization tool, includes customizable templates to enable or disable Windows system services and features using VMware recommendations and best practices across multiple systems. Most Windows system services are enabled by default, but you can use the optimization tool to easily disable unnecessary services and features to improve performance.
8. Restart the Windows OS when prompted to do so.

Install the NVIDIA vGPU software driver

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

Before the NVIDIA driver is installed on the guest virtual machine, the Device Manager shows the Microsoft Basic Display adapter installed (Figure 50).

Figure 50. Device Manager before the NVIDIA driver is installed



1. Copy the Windows drivers from the NVIDIA GRID vGPU driver pack downloaded earlier to the master virtual machine.
2. Copy the 32- or 64-bit NVIDIA Windows driver from the vGPU driver pack to the desktop virtual machine and run setup.exe (Figure 51).

Figure 51. NVIDIA driver pack

Local Disk - OS (C:) > NVIDIA > 391.81

Name	Date modified	Type	Size
Display.Driver	10/30/2018 5:27 PM	File folder	
Display.NView	10/30/2018 5:27 PM	File folder	
NV12	10/30/2018 5:27 PM	File folder	
NVWMI	10/30/2018 5:27 PM	File folder	
EULA	3/23/2016 9:43 PM	Text Document	48 KB
license	6/28/2018 9:14 AM	Text Document	27 KB
setup.cfg	6/28/2018 9:14 AM	CFG File	10 KB
setup	7/4/2018 4:19 AM	Application	425 KB

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

3. Install the graphics drivers using the Express Option (Figure 52). After the installation is completed successfully, click Close (Figure 53) and restart the virtual machine.

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

Figure 52. Select the Express or Custom installation option



Figure 53. Components installed during the NVIDIA graphics driver installation process

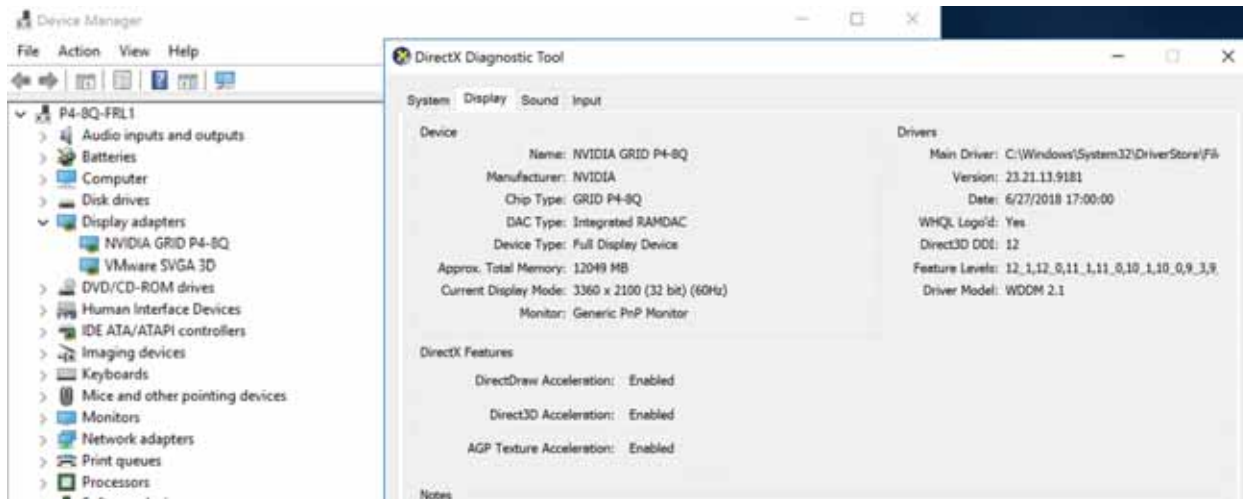


Verify that the virtual machine is ready to support the vGPU

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

1. Open Windows Device Manager.
2. Expand the Display Adapter section. The device will reflect the chosen profile (Figure 54).

Figure 54. Verifying the driver installation: Device Manager



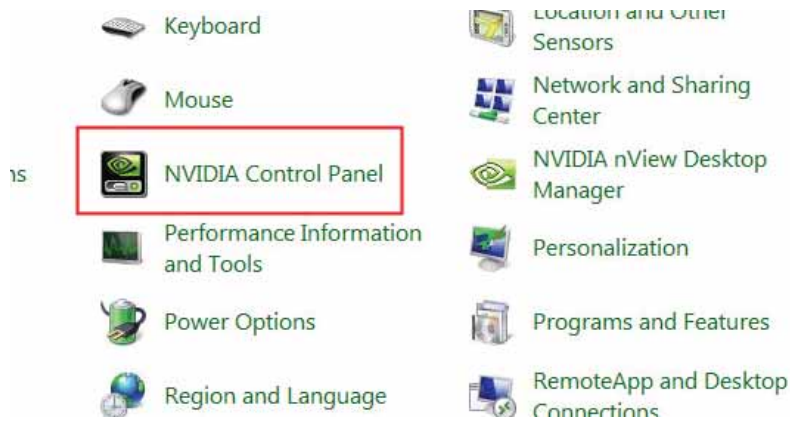
Configure the virtual machine for an NVIDIA GRID vGPU license

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

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

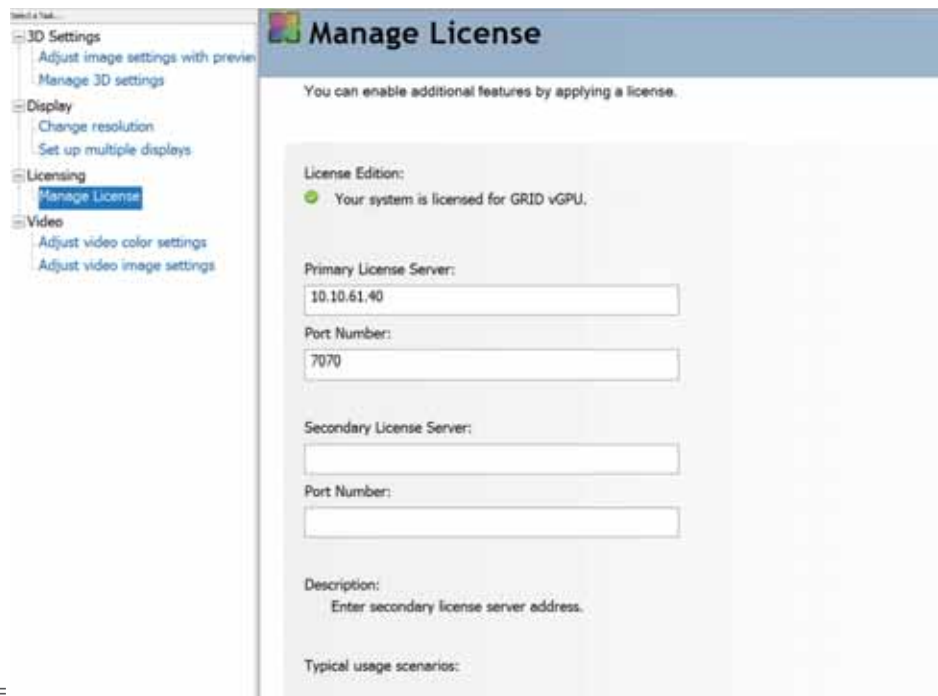
1. In the Microsoft Windows Control Panel, double-click NVIDIA Control Panel (Figure 55).

Figure 55. Choosing the NVIDIA control panel



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

Figure 56. Managing your license



Create a vGPU-enabled desktop pool with VMware Horizon 7.5

Each Horizon desktop pool configuration depends on the specific use case.

The desktop pool created as part of the solution verification is based on persistent desktops. The virtual machines are deployed as full clones from the master image template.

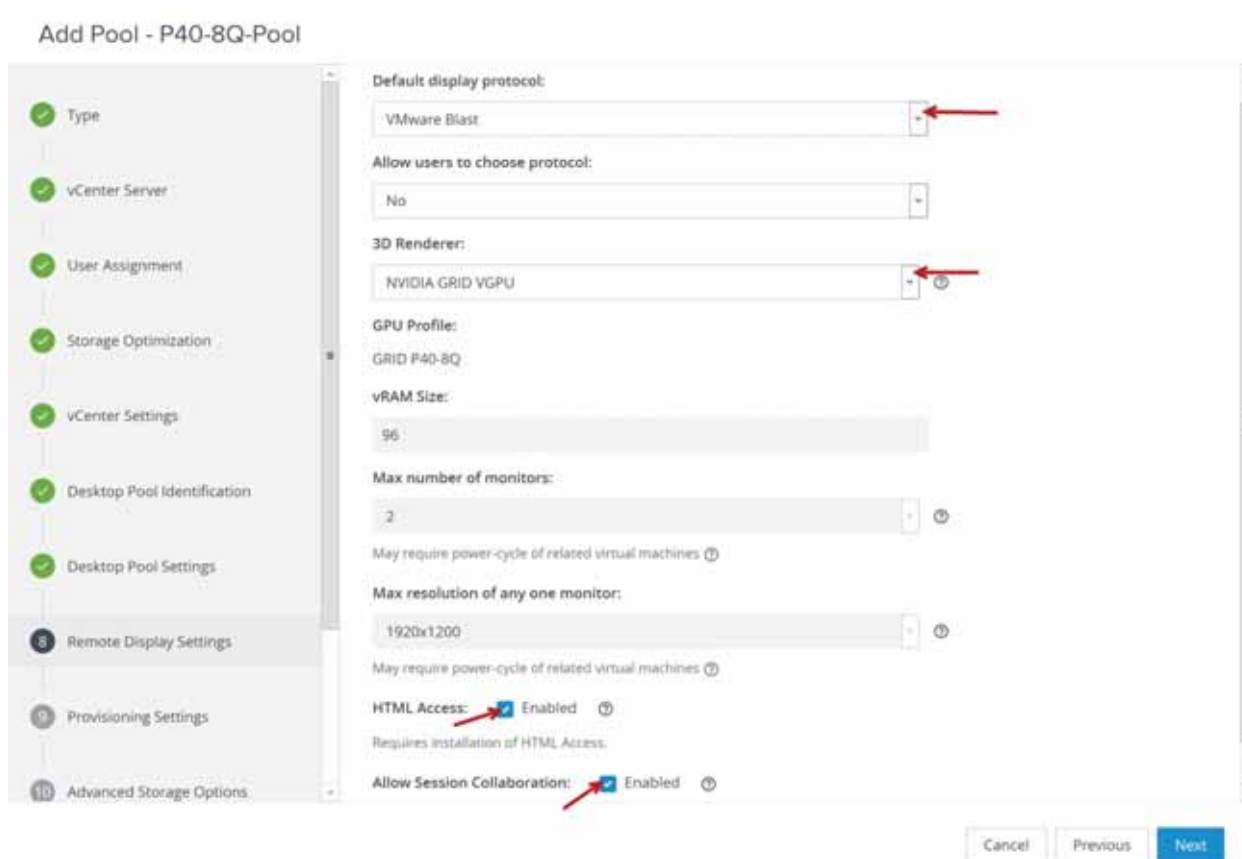
Follow these steps to create the Horizon 7 desktop pool:

1. For the Remote Display Protocol (RDP), choose VMware Blast.
2. In the available features for Horizon, select the check boxes for HTML Access and Allow Session Collaboration.

Note: The VMware Blast display protocol is required for the Allow Session Collaboration feature. If any other protocol is selected, the option will be dimmed.

3. Select the 3D Renderer option according to your deployment scenario. For this document, NVIDIA GRID vGPU is selected, as shown in Figure 57.

Figure 57. VMware Horizon console when creating a desktop pool with an NVIDIA vGPU profile attached to a Microsoft Windows 10 master image



Verify vGPU deployment

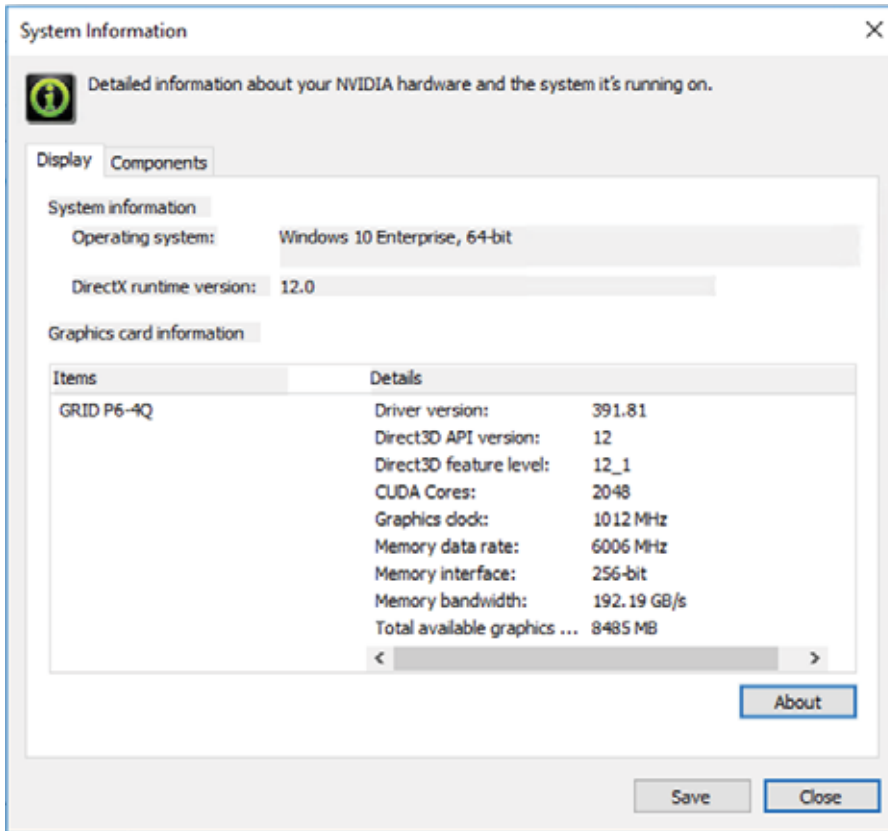
After the desktops are provisioned, use the following steps to verify vGPU deployment.

Verify that the NVIDIA driver is running on the desktop

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

1. Right-click the desktop. In the menu, choose NVIDIA Control Panel to open the control panel.
2. In the control panel, select System Information to see the vGPU that the virtual machine is using, the vGPU's capabilities, and the NVIDIA driver version that is loaded (Figure 58).

Figure 58. NVIDIA control panel: System Information page



Verify NVIDIA license acquisition by desktops

A license is obtained after the virtual machine is fully booted and before the user logs on to the virtual machine (Figure 59).

Figure 59. NVIDIA license server: Licensed Feature Usage page



To view the details, select Licensed Clients in the left pane (Figure 60).

Figure 60. NVIDIA license server: Licensed Clients page



Use the VMware vSphere 6.7 Performance tab to monitor GPU use

New with vSphere 6.5 and later, you can monitor NVIDIA GPU use through the vSphere Web Client (Figure 61).

1. Navigate to the ESXi host and open the Monitor tab.
2. Select the Performance tab and click Advanced.
3. From the drop-down menu at the right, for View select GPU.

Figure 61. vSphere 6.7 GPU performance monitoring example



SPECviewperf 13 benchmark results

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

SPECviewperf 13 has the following viewsets:

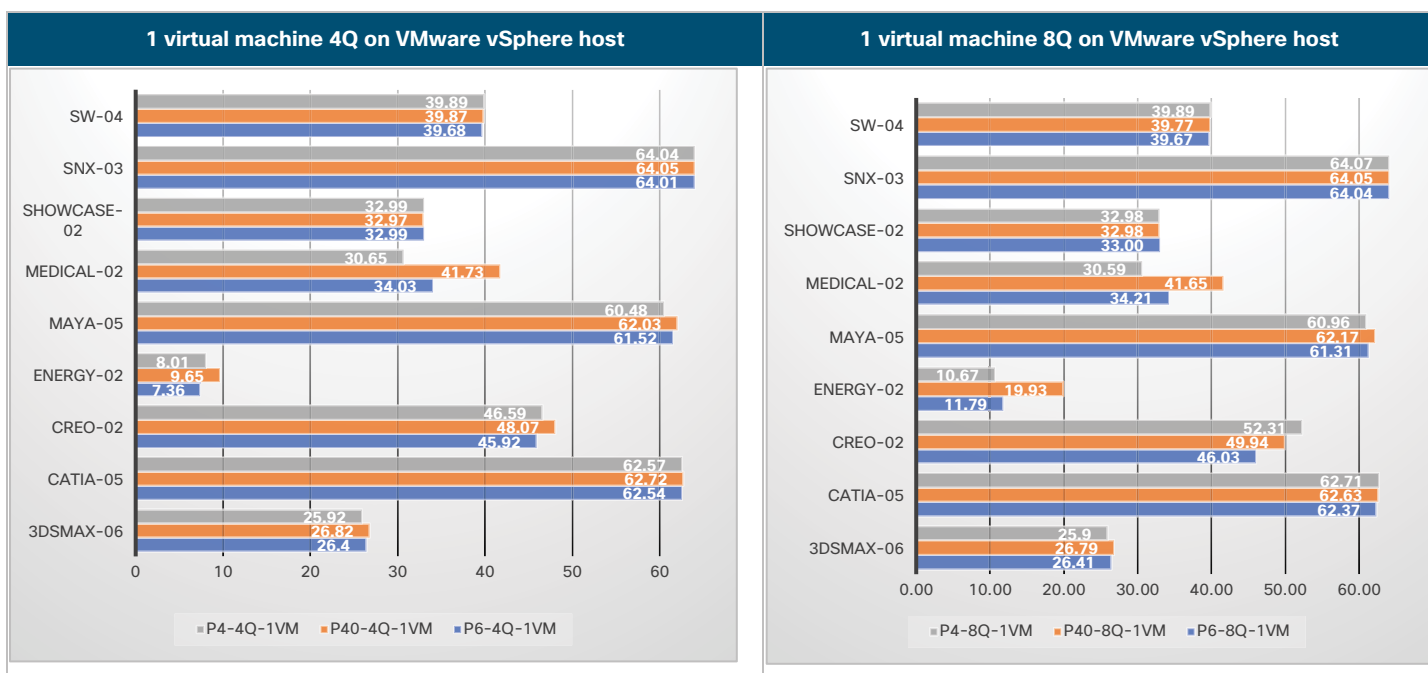
- 3ds Max (3dsmax-06)
- CATIA (catia-05)
- Creo (creo-02)
- Energy (energy-02)
- Maya (maya-05)

- Medical (medical-02)
- Showcase (showcase-02)
- Siemens NX (snx-03)
- Solidworks (sw-04)

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

Figure 62 compares all three graphics cards studied over the nine benchmark applications with NVIDIA GRID vDWS 4Q and 8Q profiles to show the impact of increased video frame buffer size by card and by application. The tests reported here used the default settings for the NVIDIA driver for the frame rate limiter (on), with the best-effort GPU scheduler. The intent was to get a clean benchmark comparison of the three Tesla cards with the default settings. In some situations and for some applications, these settings can be changed to enhance the end-user experience. The tests used the minimum 4Q profile required by the Energy application in the test suite. The 8Q profile was used for comparison testing to show the effect of additional frame buffer allocation without changing the host’s CPU or memory configuration. In addition, a frame rate policy for Horizon administration is available that can be configured as advanced settings using group policy. Download [VMware Horizon GPO Bundle](#).

Figure 62. SPECviewperf results for a single virtual machine comparison: P4, P6, and P40 with vDWS 4Q (left) and 8Q (right)



For the NVIDIA Tesla P4, P6, and P40 GPU cards running in Cisco UCS C240 M5 and B200 M5 servers, increasing the frame buffer size from 4 GB to 8 GB had only a small effect on the composite frame rates on most of the applications in the benchmark. However, there was a measurable improvement in frame rate with more frame buffer in the Energy (P40) and Creo (P4) applications.

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

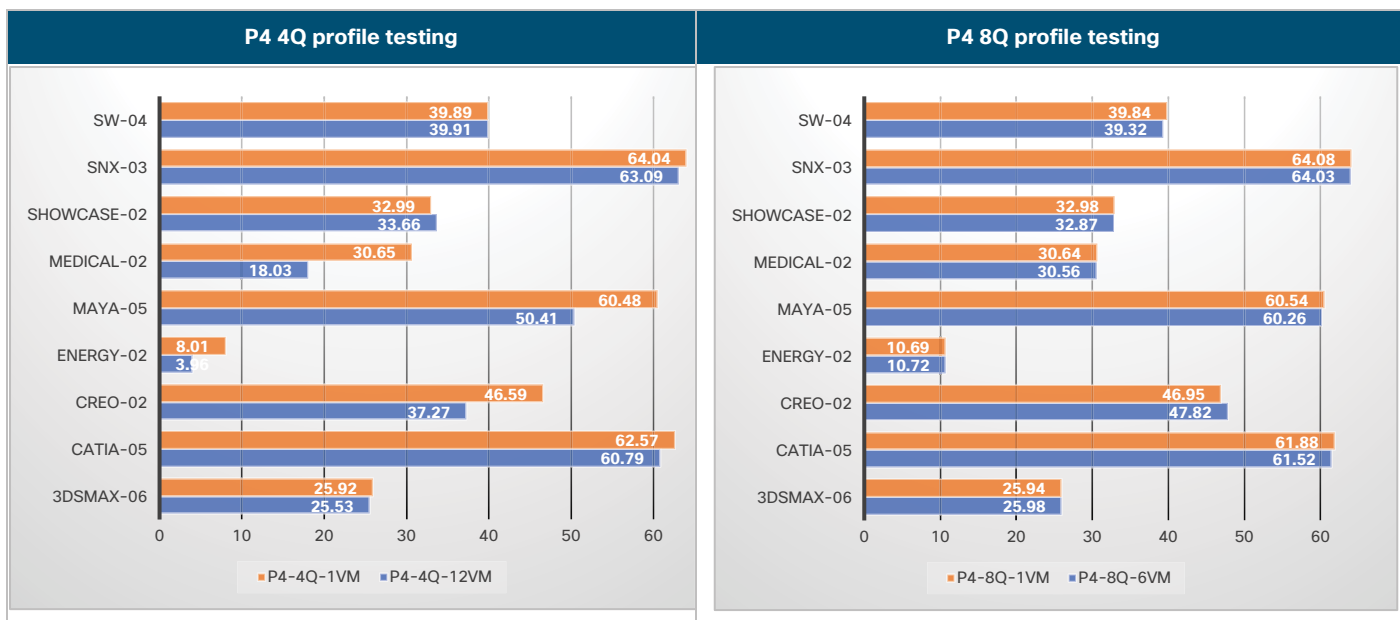
NVIDIA Tesla P4 test results

Figure 63 shows the results for the NVIDIA Tesla P4 8-GB card running each profile size. The Cisco UCS C240 host contained six P4 cards, for a total of 48 GB of frame buffer space across the server. The testing continues to use the default frame rate limiter setting (On) and the default scheduler (Best Effort.)

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

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

Figure 63. SPECviewperf results for P4 4Q and 8Q profile tests: Single virtual machine versus maximum host density based on profile size



This test reveals that using a P4 card with twelve 4-GB profiles across the six cards in the server has a measurable negative impact on application frame rate for the Medical, Maya, Energy, and Creo applications.

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

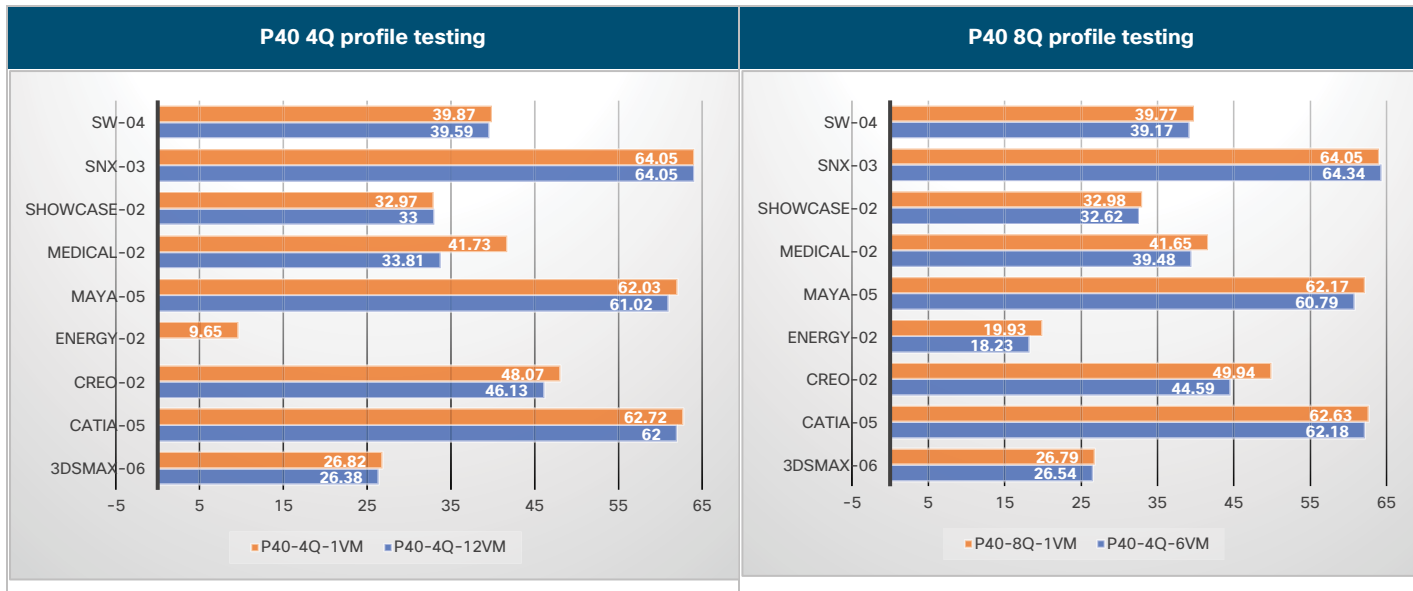
NVIDIA Tesla P40 test results

Figure 64 shows the results for the NVIDIA Tesla P40 24-GB card running the 4-GB and 8-GB profile sizes. The Cisco UCS C240 host contained two P40 cards, for a total of 48 GB of frame buffer space across the server.

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

The right chart shows the P40 performance with one virtual machine running an 8-GB profile (orange) and six virtual machines running 8-GB profiles (blue) on the same host (two test cycles).

Figure 64. SPECviewperf results for P40 4Q and 8Q profile tests: Single virtual machine versus maximum host density based on profile size



For the 4-GB frame buffer tests, the impact of running six virtual machines on each Tesla P40 card versus a single virtual machine on one card is pronounced for the Medical application only. For the other eight applications in the benchmark, the performance of a single virtual machine versus that of six virtual machines is negligible.

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

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

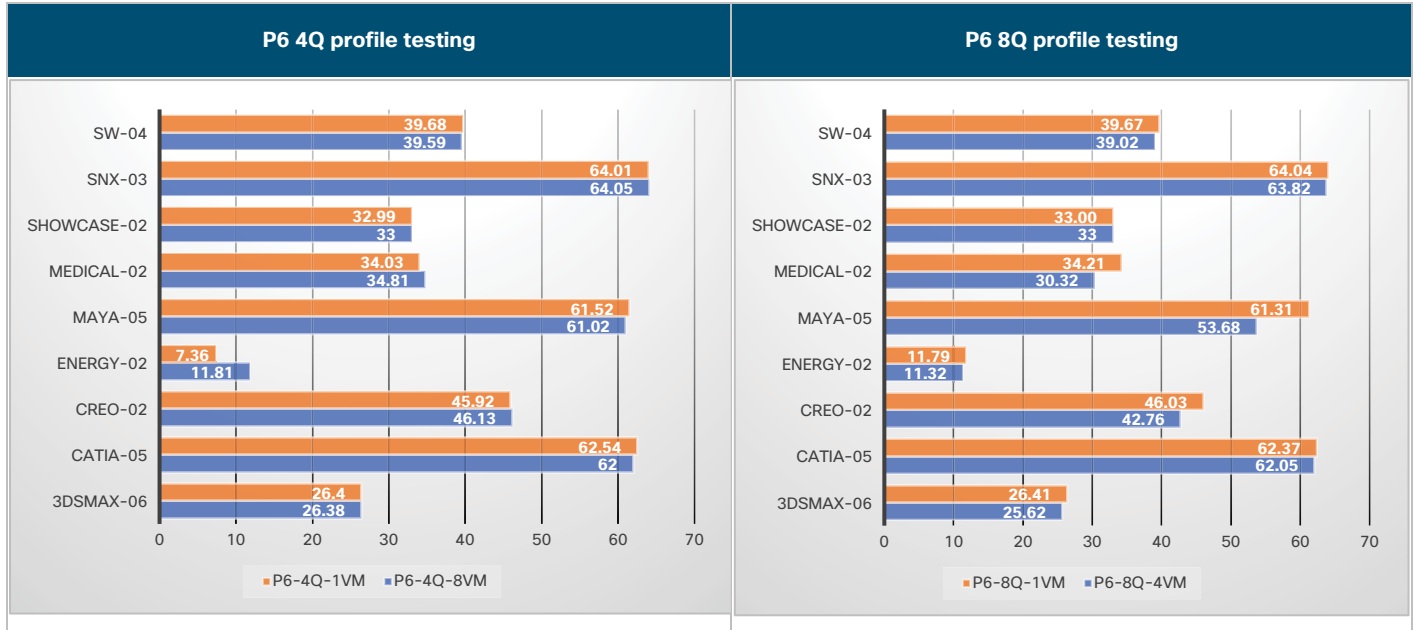
NVIDIA Tesla P6 test results

Figure 65 shows the results for the NVIDIA Tesla P6 16-GB card running the 4-GB and 8-GB profile sizes. The Cisco B200 M5 host contained two P6 cards, for a total of 32-GB of frame buffer space across the server.

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

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

Figure 65. SPECviewperf results for P6 4Q and 8Q profile tests: Single virtual machine versus maximum host density based on profile size



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

Host CPU utilization test results

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

Figure 66 presents the data for the Tesla P4 4Q and 8Q profiles.

Figure 67 presents the data for the Tesla P40 4Q and 8Q profiles.

Figure 68 presents the data for the Tesla P6 4Q and 8Q profiles.

Figure 66. Cisco UCS C240 CPU utilization results for SPECviewperf P-4 4Q and 8Q profile tests: Single virtual machine versus maximum density

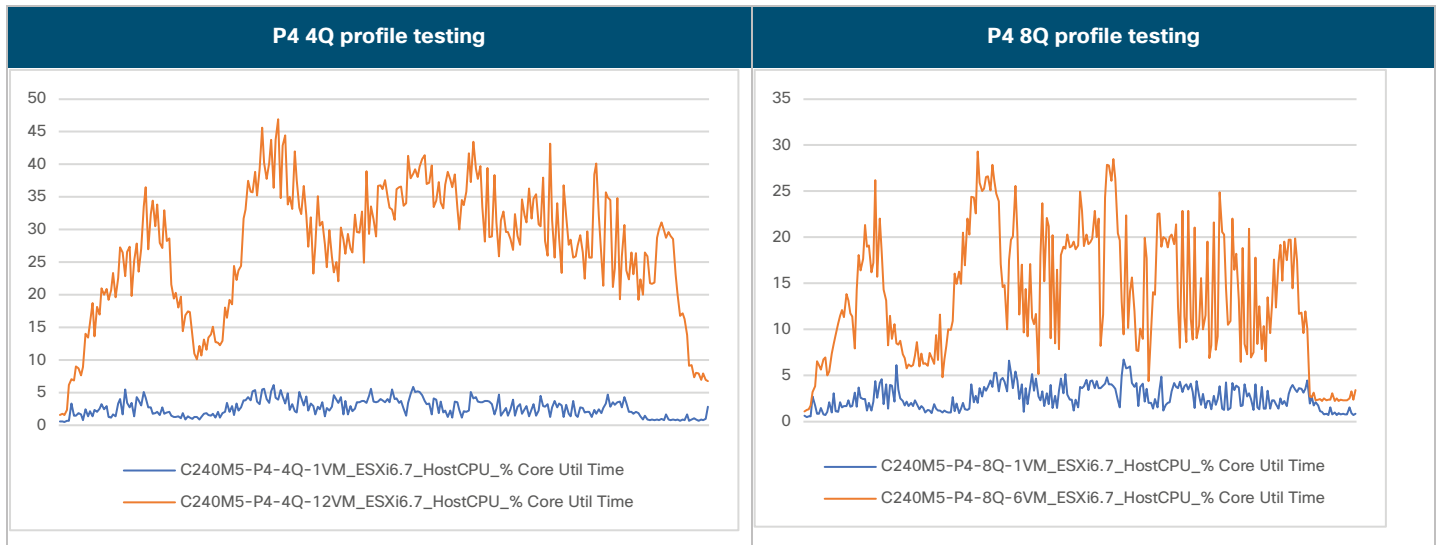


Figure 67. Cisco UCS C240 CPU utilization results for SPECviewperf P40 4Q and 8Q profile tests: Single virtual machine versus maximum density

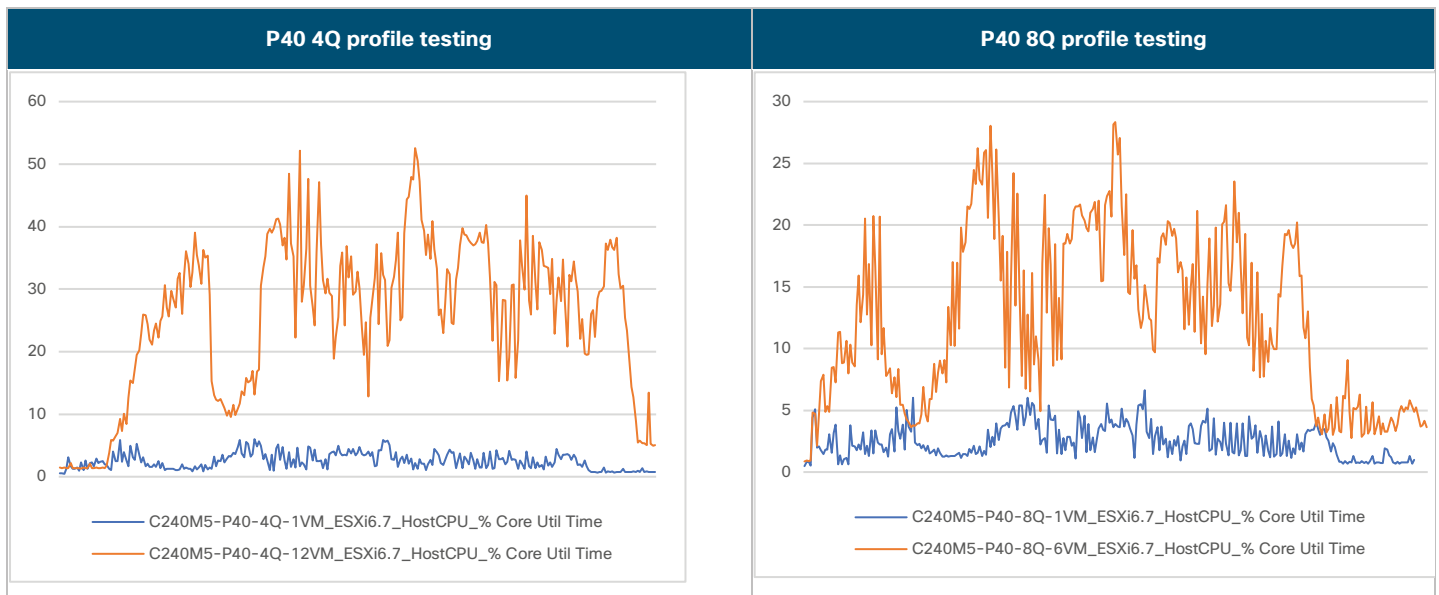
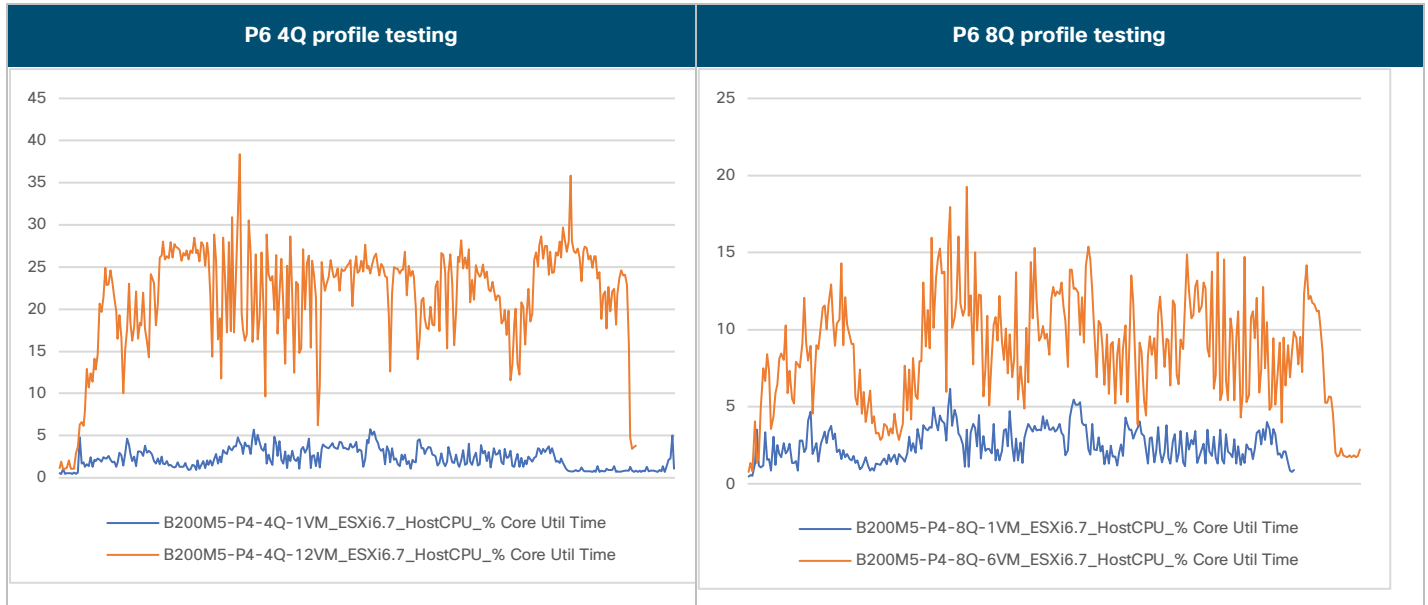


Figure 68. Cisco UCS B200 CPU utilization results for SPECviewperf P6 4Q and 8Q profile tests: Single virtual machine versus maximum density



Host GPU utilization test results

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

Figure 69 presents the data for the Tesla P4 4Q and 8Q profiles.

Figure 70 presents the data for the Tesla P40 4Q and 8Q profiles.

Figure 71 presents the data for the Tesla P6 4Q and 8Q profiles.

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

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

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

Figure 69. NVIDIA P4 results for SPECviewperf P4 4Q and 8Q profile tests: Single virtual machine versus maximum density

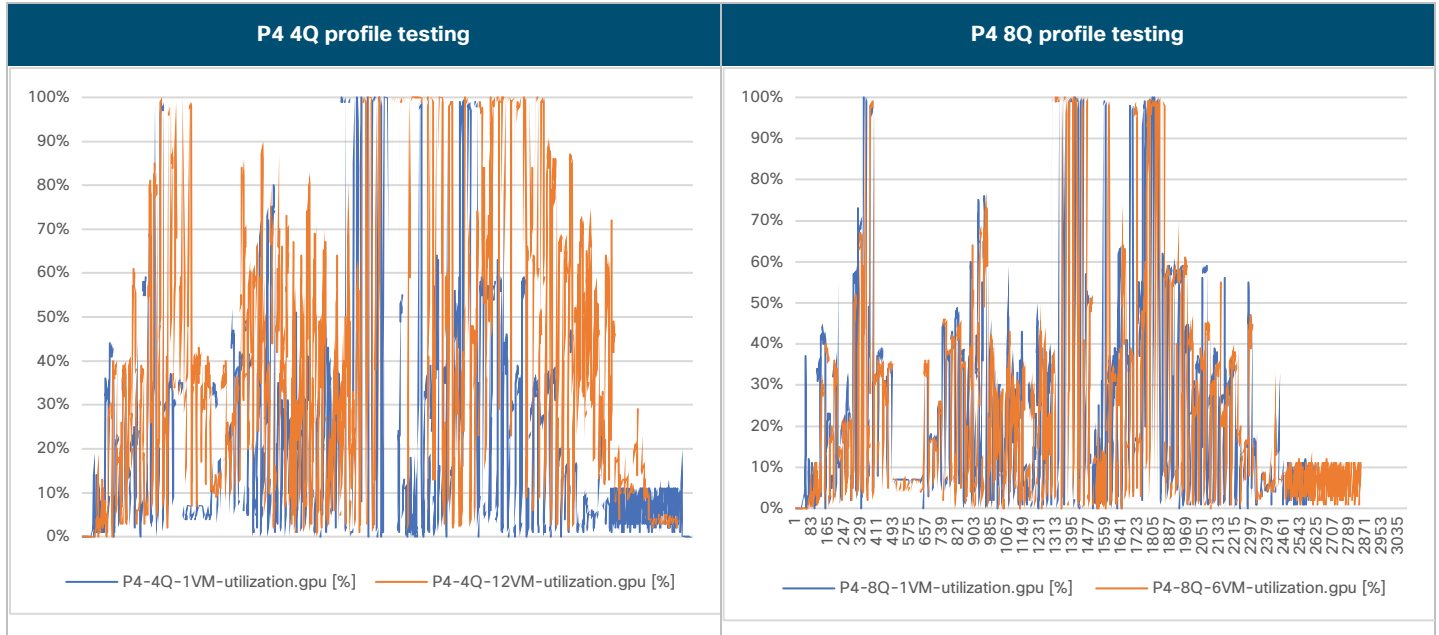


Figure 70. NVIDIA P40 results for SPECviewperf P40 4Q and 8Q profile tests: Single virtual machine versus maximum density

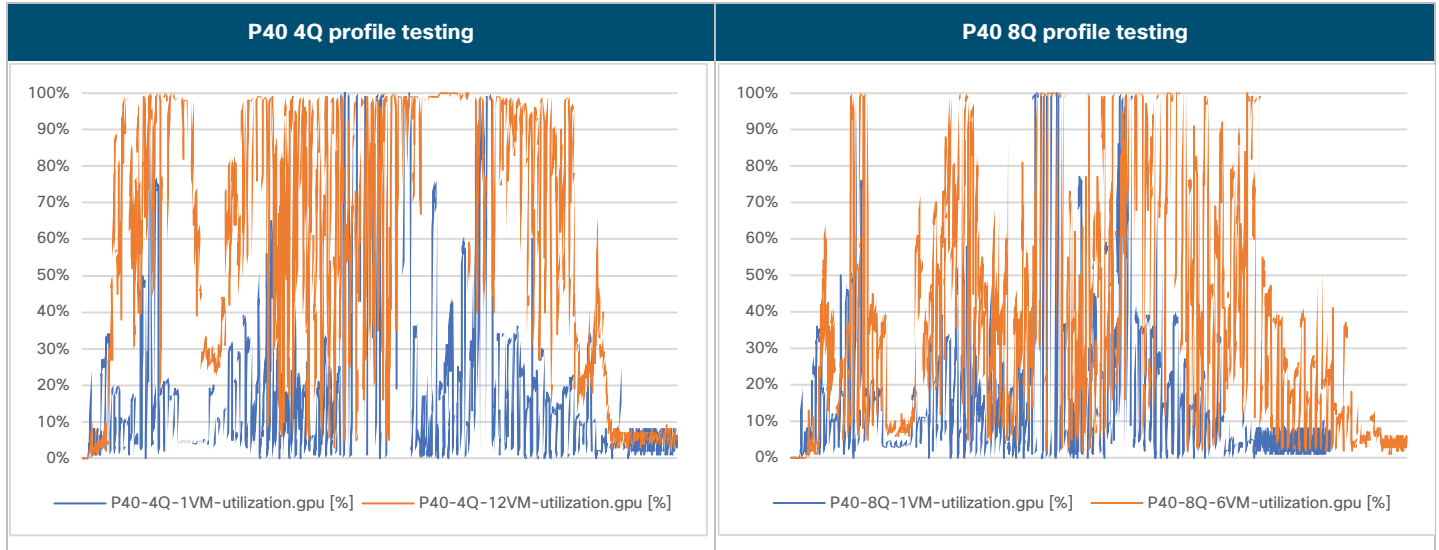
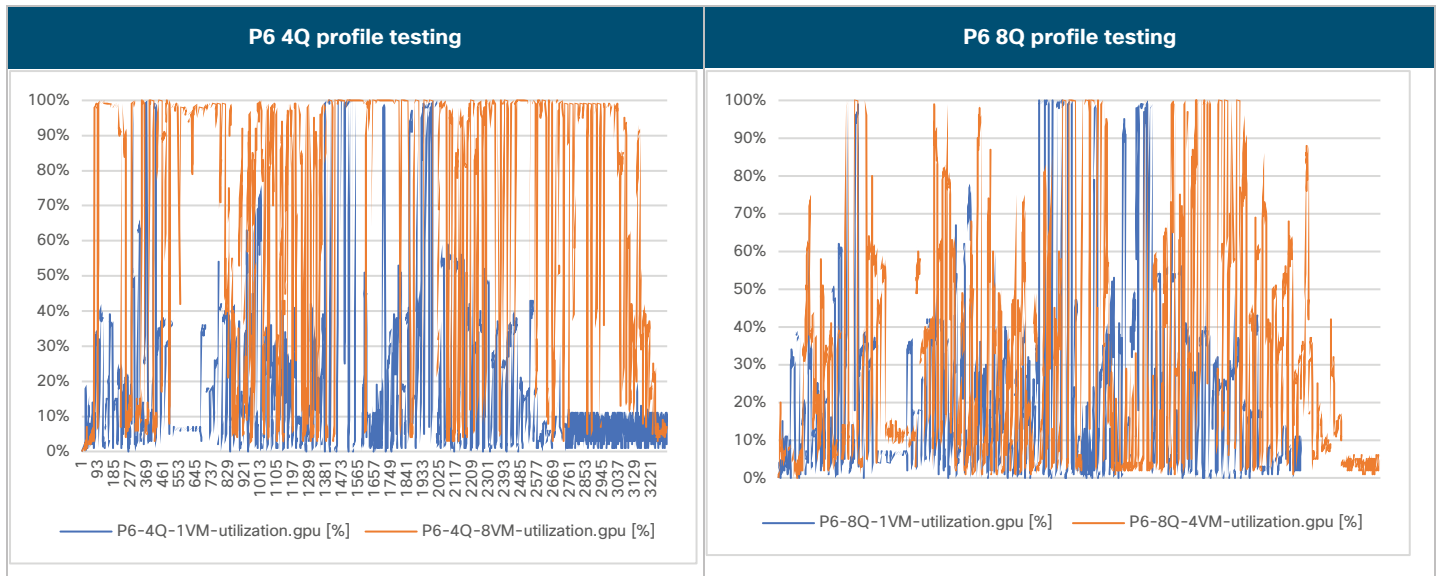


Figure 71. NVIDIA P6 results for SPECviewperf P6 4Q and 8Q profile tests: Single virtual machine versus maximum density



Live Migration of vGPU-enabled virtual machines with VMware vMotion

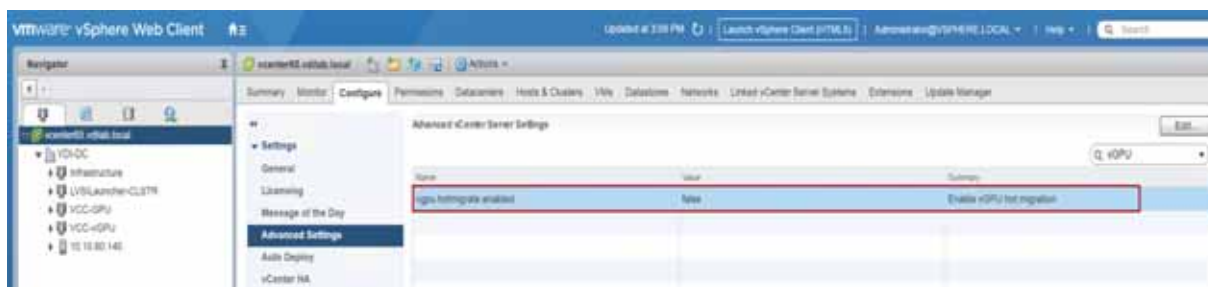
Following the introduction of Suspend and Resume capabilities in vSphere 6.7, VMware and NVIDIA are introducing VMware vSphere vMotion with NVIDIA vGPU-enabled virtual machines. Virtual machines with vGPU vMotion capabilities help infrastructure administrators avoid downtime and create high-availability deployments. This capability allows administrators to perform maintenance operations on the underlying GPU-accelerated VDI or computing infrastructure while helping ensure that end users and their applications are not disrupted.

With vSphere 6.7 Update 1, VMware and NVIDIA have collaborated to significantly enhance the operational flexibility and utilization of virtual infrastructure accelerated with NVIDIA vGPU solutions, including Quadro Virtual Data Center Workstation (Quadro vDWS), GRID Virtual PC (vPC), and GRID Virtual Applications (vApps).

For more information about support and restrictions, refer to the [VMware ESXi Installation and Setup Guide](#).

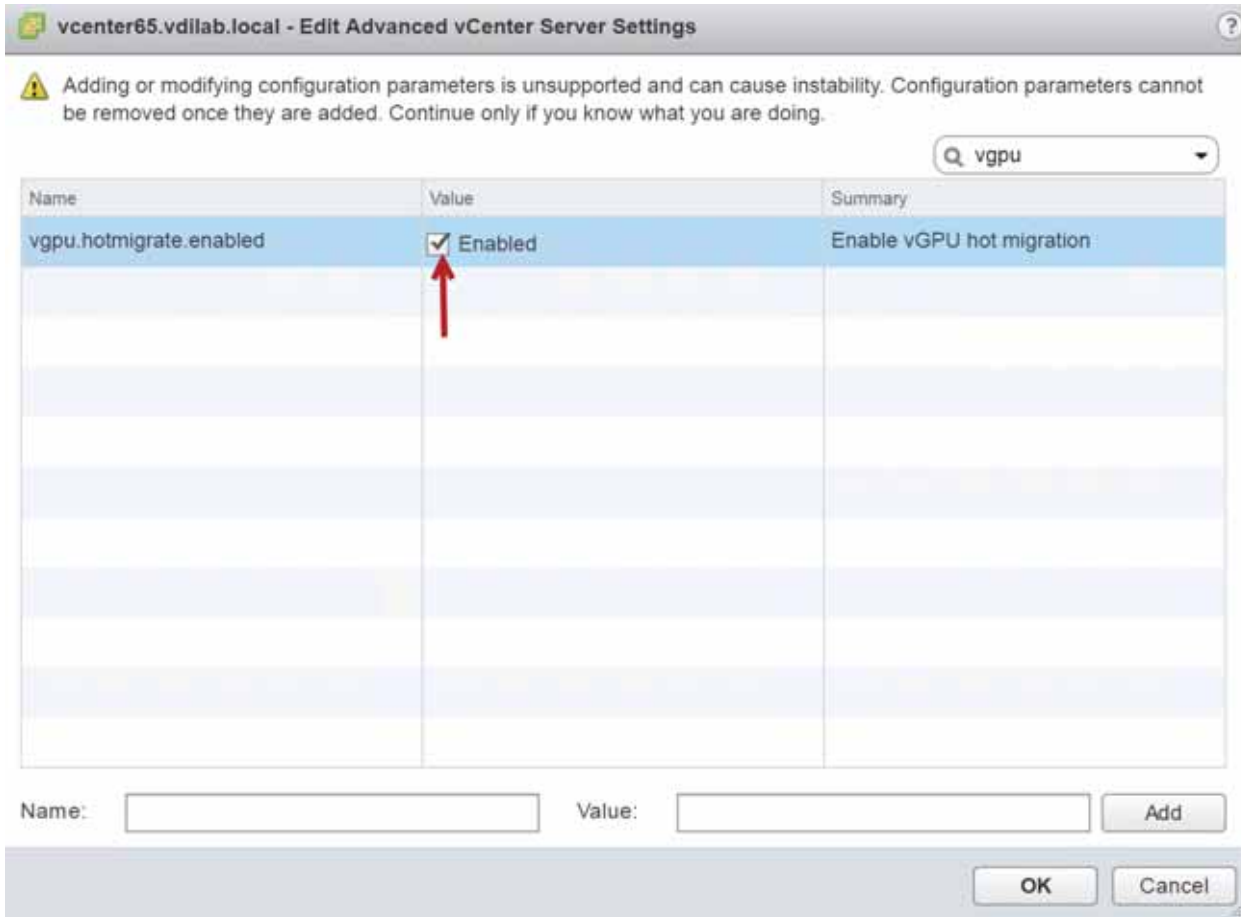
1. To initiate vMotion movement of the vGPU-enabled virtual machines, enable vGPU vMotion on the vCenter Web Client by setting `vgpu.hotmigrate.enabled` to true, as shown in the following steps.
 - a. On the vCenter server Web Client, navigate to the Configure tab. From the Settings, menu, choose Advanced Settings. In the search bar, filter on vGPU, as shown in Figure 72. By default, `vgpu.hotmigrate.enabled` is set to false.

Figure 72. VMware vCenter Web Client Advanced Settings



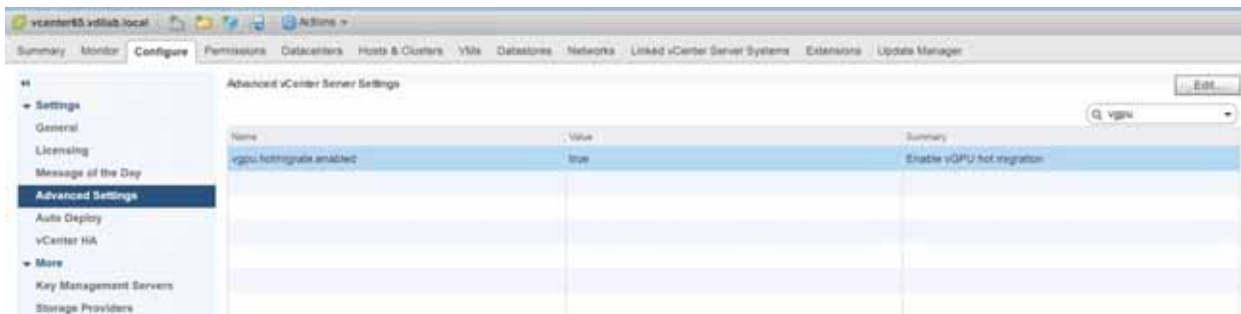
- b. Click the Edit button in the top-right corner. Filter the Advanced Setting options on the value **vgpu**. Select the checkbox to enable `vgpu.hotmigrate.enabled` on vGPU-enabled virtual machines, as shown in Figure 73. Click OK.

Figure 73. VMware vCenter Web Client: Enable vgpu.hotmigrate.enabled



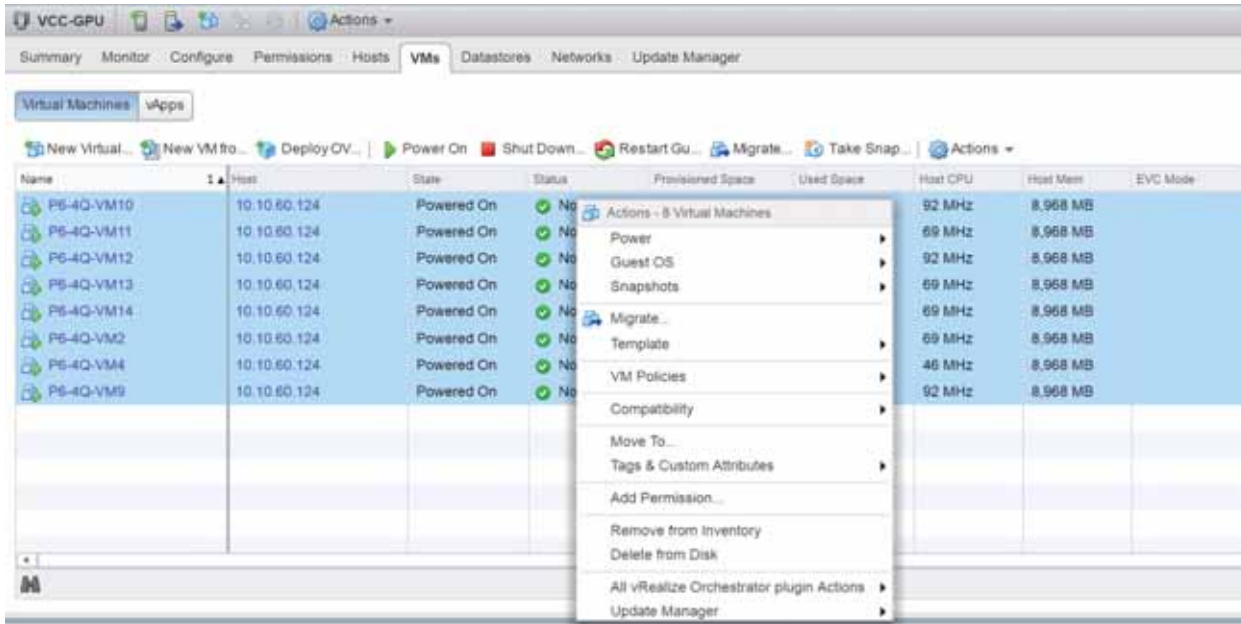
c. Verify that vgpu.hotmigrate.enabled is now set to true (Figure 74).

Figure 74. VMware vCenter Web Client: vgpu.hotmigrate.enabled value set to true



2. Click Apply to make the changes take effect immediately. Now vGPU-enabled virtual machines can be migrated on the ESXi host with the same NVIDIA Tesla GPU configured with the vGPU configuration as discussed in the section Install NVIDIA GRID software on the VMware ESXi host.
3. Select the virtual machines that need to be live-migrated to another ESXi host. Right-click and choose Migrate (Figure 75).

Figure 75. VMware vCenter Web Client: Select the virtual machines and choose Migrate

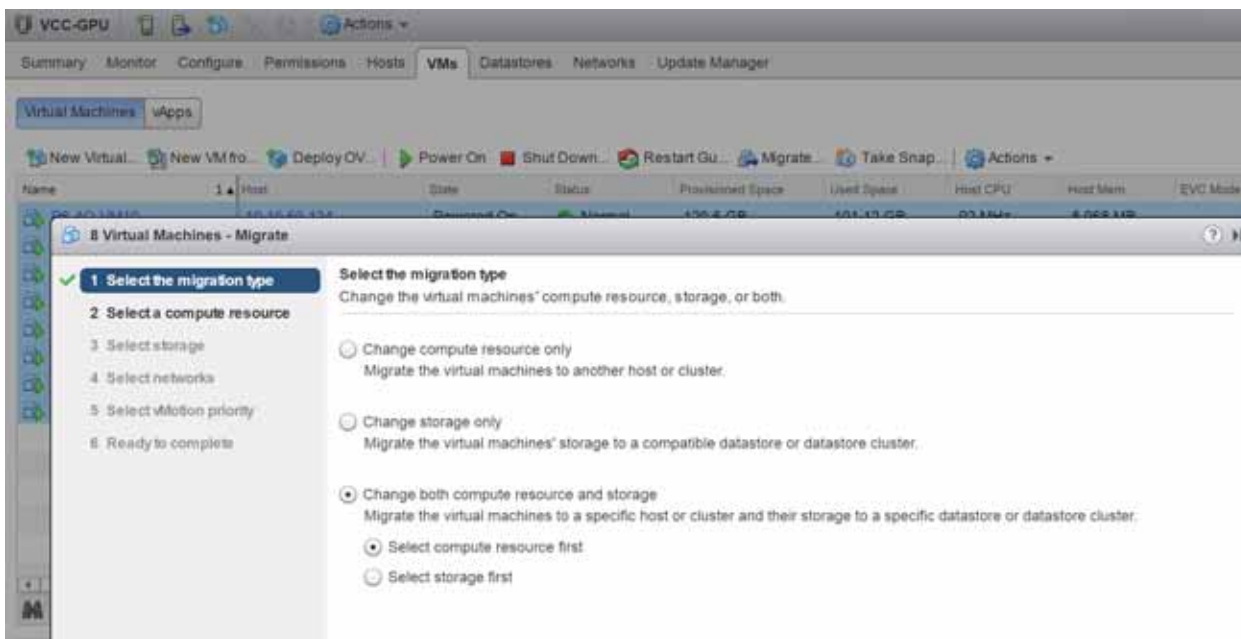


4. Three options are available for virtual machine migration (Figure 76):

- Change compute resource only
- Change storage only
- Change both compute resource and storage

Select an option. This document used the “Changed both compute resource and storage” option.

Figure 76. VMware vCenter Web Client: Select the migration type



5. Follow the prompts for host and storage selection to complete the migration of the virtual machines (Figure 77, Figure 78, and Figure 79).
6. For vMotion movement of virtual machines with vGPU enabled, make sure that the destination host is configured with the vGPU resources required to complete the vMotion operation successfully. You will see a warning about compatibility issues, as shown in Figure 77.

Figure 77. VMware vCenter Web Client: vGPU-enabled virtual machine vMotion compatibility warning

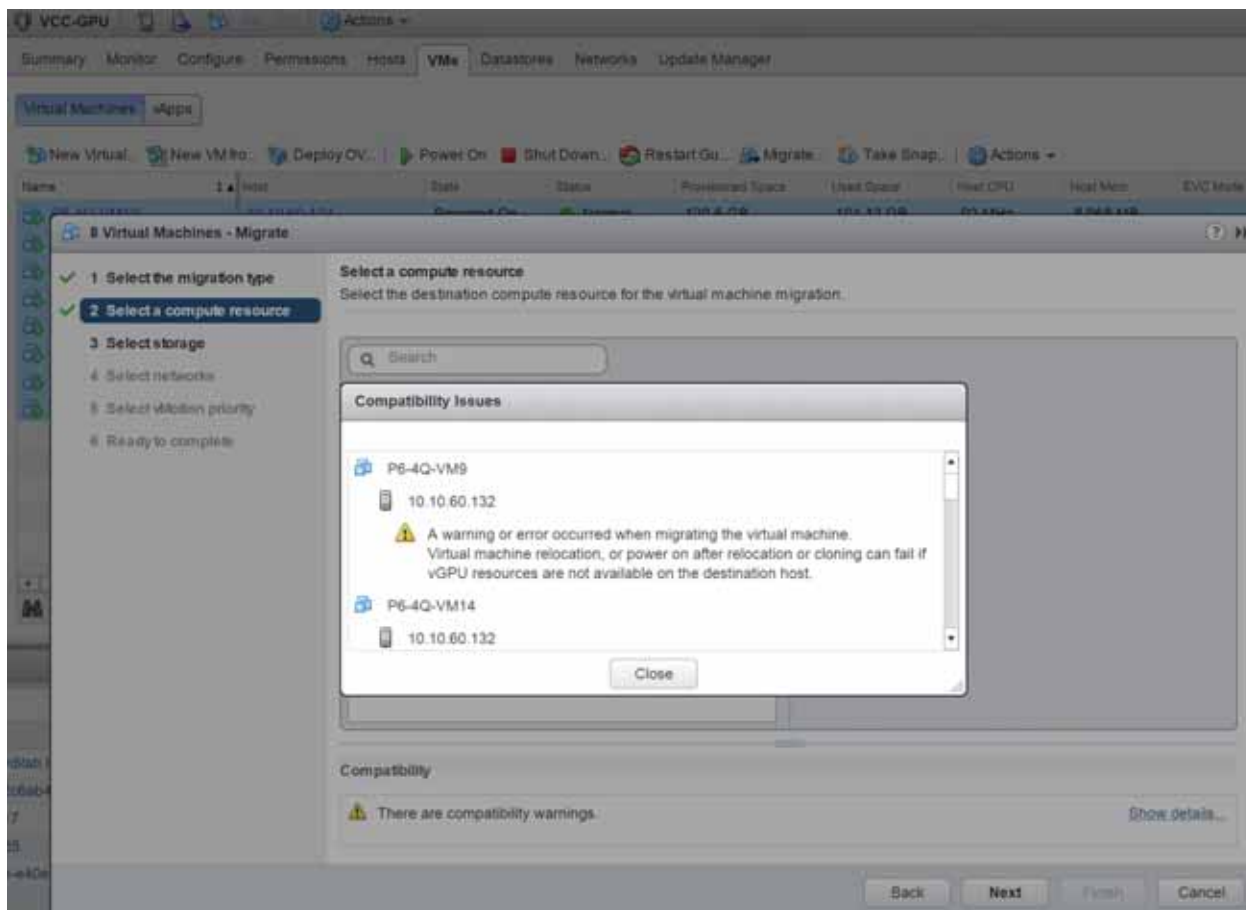


Figure 78. VMware vCenter Web Client Events window: vGPU-enabled virtual machine vMotion operation in progress

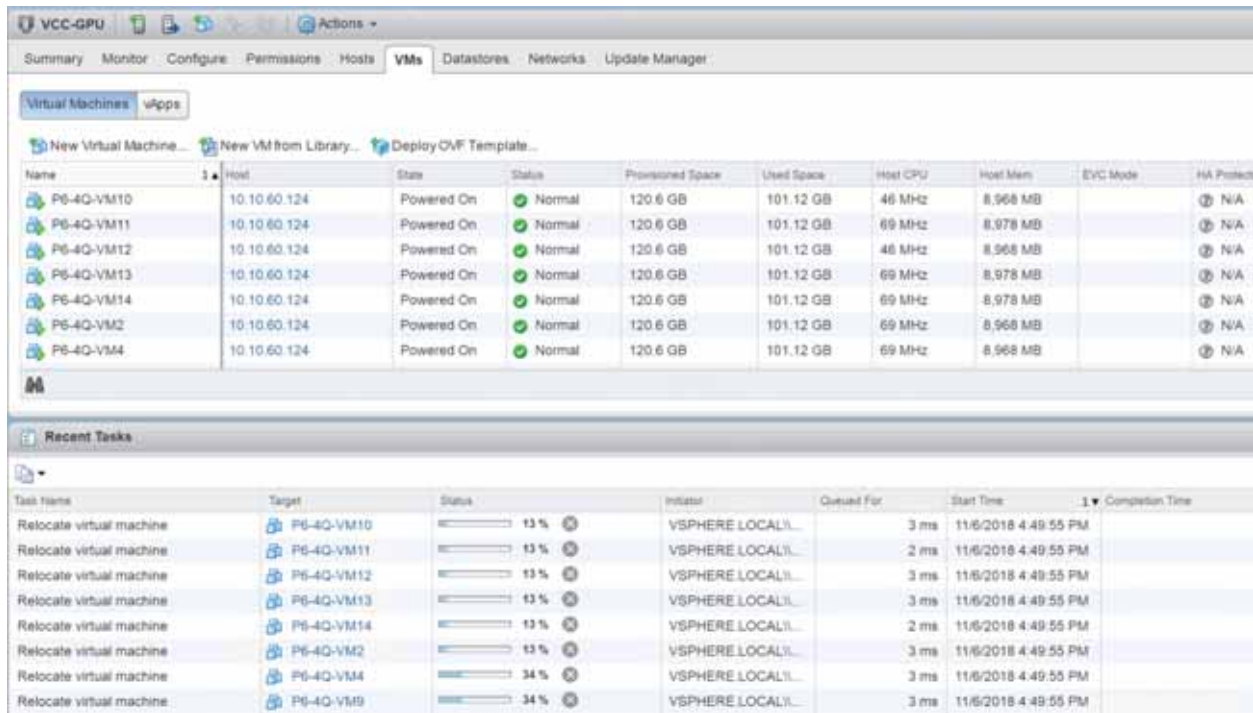
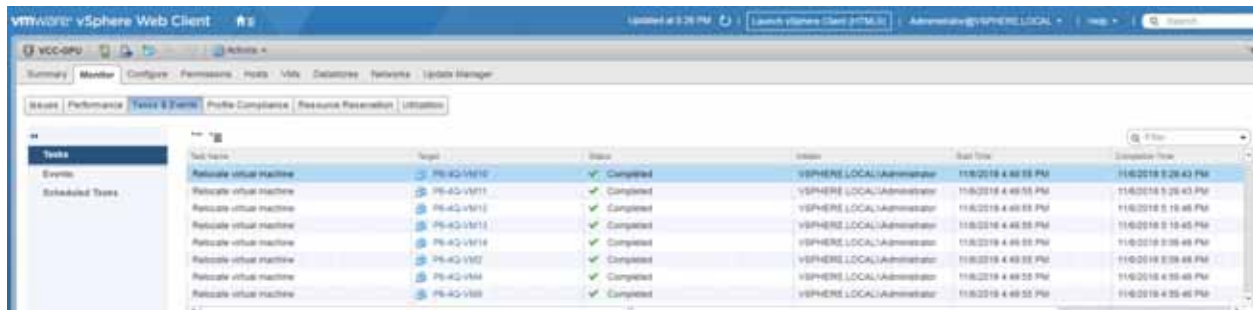


Figure 79. VMware vCenter Web Client Events window: vGPU-enabled virtual machine vMotion operation completion



Additional configurations

This section presents additional configuration options.

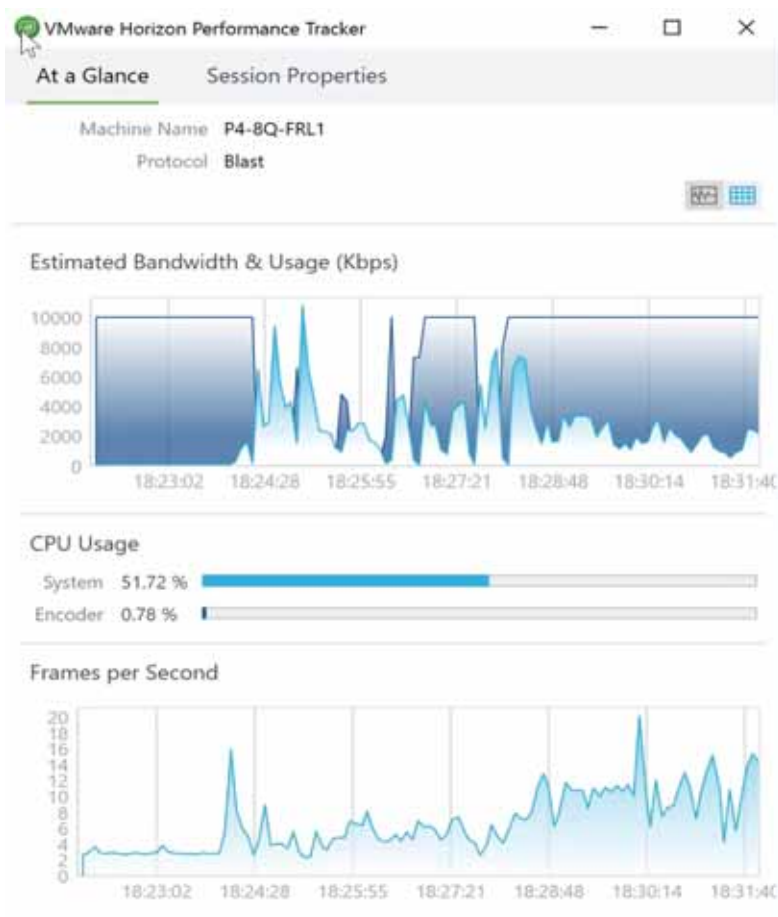
Install and upgrade NVIDIA drivers

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 VMware Horizon Performance Tracker

VMware Horizon 7.5 and newer versions include Horizon Performance Tracker as part of the agent installation process for the master image (Figure 80). The tracker can show the user performance of the remote session. It can also be provisioned as a published application.

Figure 80. Example of VMware Horizon Performance Tracker



For information about the processes that the Horizon Performance Tracker can help track and the system requirements for running the tracker, see <https://docs.vmware.com/en/VMware-Horizon-7/7.5/horizon-administration/GUID-E90DA3B8-6ADF-4130-B4C7-89E93CFC9D24.html>.

Optimize VMware Blast with the GPO bundle to achieve optimal end-user experience

Follow the optimization guide for the VMware Blast Extreme display protocol to achieve the best performance with Blast Extreme in low-bandwidth, high-latency situations. See <https://techzone.vmware.com/resource/blast-extreme-display-protocol-vmware-horizon-7#section8>.

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

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

Use the OpenGL Software Accelerator

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

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

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

Try the OpenGL Software Accelerator in the following cases:

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

Conclusion

The combination of Cisco data center hardware, Cisco UCS Manager, Cisco UCS C240 M5 Rack Servers and B200 M5 Blade Servers, the customer's choice of storage array, and NVIDIA Tesla cards running NVIDIA GRID 6.2 on VMware vSphere 6.7 and VMware Horizon 7.5 provides high-performance platforms for virtualizing real-world enterprise 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.

Additional performance tuning by application, which is outside the scope of this document, can enhance the end-user experience for each of the applications in the SPECViewPerf 13 test suite. We encourage customers and partners to categorize users of their graphics applications and develop sizing and tuning optimizations to give each user a great experience.

For more information

- Cisco UCS C-Series Rack Servers and B-Series Blade Servers:
 - <https://www.cisco.com/c/en/us/products/servers-unified-computing/index.html>
- NVIDIA:
 - <https://www.nvidia.com/en-us/design-visualization/solutions/virtualization/>
 - <https://www.nvidia.com/en-us/design-visualization/vmware/>
- VMware Horizon 7:
 - https://www.vmware.com/support/pubs/view_pubs.html
 - <http://www.vmware.com/products/horizon/vgpu-blast-performance.html>
 - <https://blogs.nvidia.com/blog/2016/02/09/nvidia-grid-blast-extreme-vmware-horizon/>
 - <http://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/products/horizon/grid-vgpu-deployment-guide.pdf>
 - <https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/vmware-horizon-7-view-blast-extreme-display-protocol.pdf>

- Microsoft Windows and VMware optimization guides for virtual desktops:
 - <http://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/whitepaper/vmware-view-optimizationguidewindows7-en-white-paper.pdf>
 - <http://www.vmware.com/techpapers/2010/optimization-guide-for-windows-7-and-windows-8-vir-10157.html>
 - <https://labs.vmware.com/flings/vmware-os-optimization-tool>
- VMware vSphere ESXi and vCenter Server 6.7:
 - <https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/products/vsphere/vmware-whats-new-in-vsphere-whitepaper.pdf>
 - <https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/products/vsphere/vmware-vsphere-67-datasheet.pdf>

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)