

Integrate Cisco UCS M5 Servers with NVIDIA GRID 6.2 on VMware vSphere 6.7 and Horizon 7.5



New: Four-socket Cisco UCS® C480 M5 Rack Servers with six NVIDIA Tesla P40 cards, two-socket Cisco UCS C220 M5 Rack Servers with two NVIDIA Tesla P4 cards, and two-socket Cisco UCS C40 M5 Rack Servers with six NVIDIA Tesla P4 cards have been added.

Contents

What you will learn	4
Why use NVIDIA GRID vGPU for graphics deployments on VMware Horizon.....	5
NVIDIA GRID vGPU profiles	6
Cisco Unified Computing System.....	7
Cisco UCS Manager	8
Cisco UCS 6300 Series Fabric Interconnects.....	8
Cisco UCS C-Series Rack Servers.....	9
Cisco UCS C240 M5 Rack Server	9
Cisco UCS C220 M5 Rack Server	10
Cisco UCS C480 M5 Rack Server	11
Cisco UCS Virtual Interface Card 1387	12
Cisco UCS B200 M5 Blade Server	13
Cisco UCS Virtual Interface Card 1340	13
NVIDIA Tesla cards	14
NVIDIA GRID vGPU.....	15
NVIDIA GRID.....	15
NVIDIA vGPU license requirements.....	15
VMware vSphere 6.7.....	15
Graphics acceleration in VMware Horizon 7.5.....	17
GPU acceleration for Microsoft Windows desktops	17
Enhanced graphics with VMware Horizon 7 with Blast 3D	18
GPU acceleration for Microsoft Windows Server	20
GPU sharing for VMware Horizon remote desktop session host workloads.....	20
Solution configuration	22
Configure Cisco UCS	23
Create BIOS policy	23
Create graphics card policy.....	24
Install NVIDIA Tesla GPU card on Cisco UCS M5	24
Install NVIDIA Tesla GPU card on Cisco UCS B200 M5.....	26
Configure the GPU card	26
NVIDIA P6 installation	26
NVIDIA M10 installation	27
NVIDIA P40 installation	27
NVIDIA P4 installation	28
Install NVIDIA GRID License Server.....	29
Install GRID License Server.....	30
Configure NVIDIA GRID 6.2 License Server.....	34
Install NVIDIA GRID software on the VMware ESXi host.....	37
NVIDIA Tesla P4, P6, P40, and M10 profile specifications	41
VMware ESXi host configuration for vDGA (pass-through) or vGPU (Virtual GPU).....	42
Install the NVIDIA vGPU software driver	46
Verify that applications are ready to support NVIDIA vGPU.....	48

Configure the virtual machine for an NVIDIA GRID vGPU license	49
Verify that the NVIDIA driver is running on the desktop.....	50
Verify NVIDIA license acquisition by desktops	50
Verify the NVIDIA configuration on the host	51
Additional configurations.....	53
Install and upgrade NVIDIA drivers.....	53
Create the VMware Horizon 7 pool.....	53
Use the VMware vSphere 6.7 Performance tab to monitor GPU use	55
Use GPU acceleration for Microsoft Windows Server DirectX, Direct3D, and WPF rendering	56
Use the OpenGL Software Accelerator.....	56
Conclusion	57
For more information.....	57

What you will learn

The Cisco Unified Computing System™ (Cisco UCS®) recently added fifth-generation (M5) blade and rack servers based on the Intel® Xeon® Processor Scalable Family architecture. Nearly concurrently, NVIDIA launched new hardware and software designed specifically to leverage the new server architectures.

Using the increased processing power of the new Cisco UCS B-Series Blade Servers and C-Series Rack Servers, applications with the most demanding graphics requirements are being virtualized. To enhance the capability to deliver these high-performance and graphics-intensive applications in Virtual Desktop Infrastructure (VDI), Cisco offers support for the NVIDIA GRID P4, P6, P40, P100, V100, and M10 cards in the Cisco UCS portfolio of PCI Express (PCIe) and mezzanine form-factor cards for the Cisco UCS C-Series Rack Servers and B-Series Blade Servers respectively.

With the addition of the new graphics processing capabilities, the engineering, design, imaging, and marketing departments of organizations can now experience the benefits that desktop virtualization brings to the applications they use at higher user densities per server.

New in this document is the inclusion of the Cisco UCS C480 M5, a four-socket four-rack-unit (4RU) server with 12 PCI slots that can support up to six NVIDIA Tesla P40 graphics cards, providing high user density per rack unit for demanding graphics applications. Users of Microsoft Windows 10 and Office 2016 and later versions can benefit from the new NVIDIA M10 high-density graphics card, deployable on Cisco UCS C240 M5 and C480 M5 Rack Servers.

Also, new in this document is the inclusion of the Cisco UCS C220 M5 and C240 M5 servers with NVIDIA Tesla P4 GPU cards. The Cisco UCS C220 M5 two-socket 1RU server offers 2 PCI slots and can support up to two NVIDIA Tesla P4 graphics cards. The C240 M5 two-socket 2RU server offers 6 PCI slots and can support up to six NVIDIA Tesla P4 graphics processing cards.

This new graphics capabilities helps enable organizations to centralize their graphics workloads and data in the data center. This capability greatly benefits organizations that need to be able to shift work or collaborate geographically. Until now, graphics files have been too large to move, and the files have had to be local to the person using them to be usable.

The PCIe graphics cards in the Cisco UCS rack servers offer these benefits:

- Support for full-length, full-power NVIDIA GRID cards in a 2RU or 4RU form factor
- Support for a mezzanine form-factor adapter graphics processing unit (GPU) card in half-width and full-width blade servers.
- Cisco UCS Manager integration for management of the servers and NVIDIA GRID cards
- End-to-end integration with Cisco UCS management solutions, including Cisco UCS Central Software and Cisco UCS Director
- More efficient use of rack space with Cisco UCS blade and rack servers with two NVIDIA GRID cards than with the 2-slot, 2.5-inch equivalent rack unit: the HP ProLiant WS460c Gen9 Graphics Server Blade with the GRID card in a second slot
- Support for up to 144 high-performance GPU-supported graphics workstation users on a single rack server.

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

- Cisco UCS Manager integration for management of the servers and the NVIDIA GRID GPU card
- End-to-end integration with Cisco UCS management solutions, including Cisco UCS Central Software and Cisco UCS Director

An important element of this document's design is VMware's support for the NVIDIA GRID Virtual Graphics Processing Unit (vGPU) feature in VMware vSphere 6.7. Prior versions of vSphere supported only virtual direct graphics acceleration (vDGA) and virtual shared graphics acceleration (vSGA), so support for vGPU in vSphere 6 greatly expands the range of deployment scenarios using the most versatile and efficient configuration of the GRID cards.

vSphere 6.7 further improves the support and capabilities introduced for GPUs through VMware's collaboration with NVIDIA. It virtualizes NVIDIA GPUs even for non-VDI use cases and for computing that is other than general purpose, such as artificial intelligence, machine learning, big data, and more. With enhancements to NVIDIA GRID vGPU technology in vSphere 6.7, administrators can simply suspend and resume workloads running on GPUs instead of having to power off those virtual machines. This feature is especially valuable when migration of the workload or virtual machine is required during maintenance or other operations. It enables better lifecycle management of the underlying host and significantly reduces disruption for end users.

The purpose of this document is to help our partners and customers integrate NVIDIA P40 graphics processing cards and Cisco UCS C480 M5 rack servers on VMware vSphere and VMware Horizon in vGPU mode.

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

The objective here is to provide the reader with specific methods for integrating Cisco UCS servers with NVIDIA GRID P4, P6, P40, V100, and M10 cards with VMware vSphere and Horizon products so that the servers, hypervisor, and virtual desktops are ready for installation of graphics applications.

Why use NVIDIA GRID vGPU for graphics deployments on VMware Horizon

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

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

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

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

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

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

NVIDIA GRID vGPU profiles

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

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

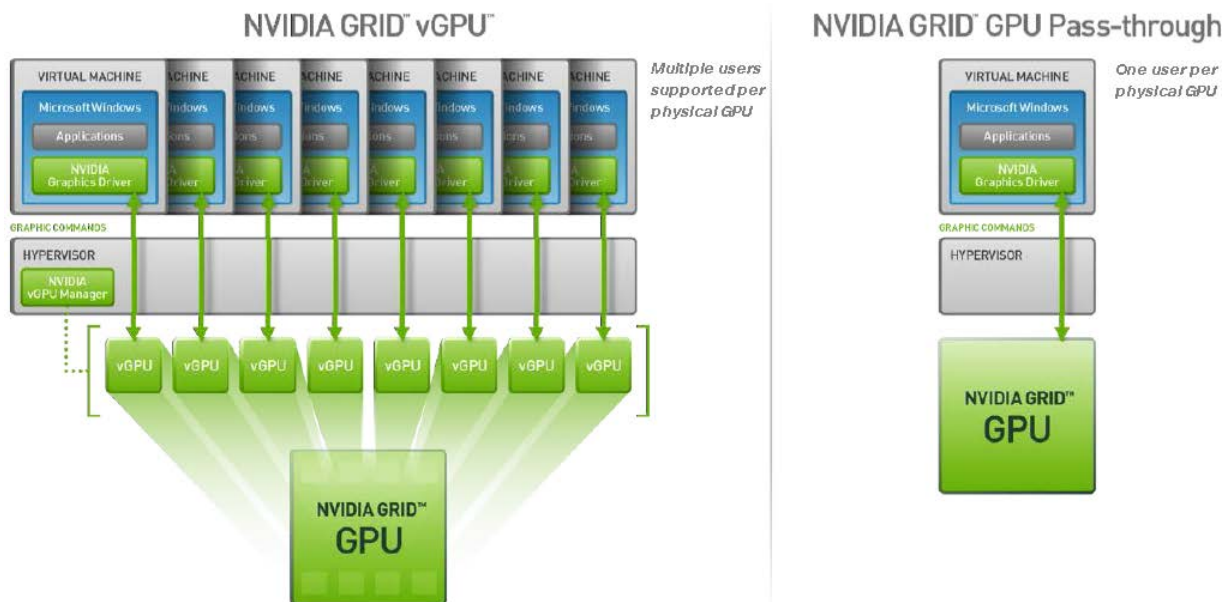
For knowledge workers, the main areas of importance include office productivity applications, a robust web experience, and fluid video playback. Knowledge workers have the least-intensive graphics demands, but they expect the same smooth, fluid experience that exists natively on today’s graphics-accelerated devices such as desktop PCs, notebooks, tablets, and smartphones.

Power users are users who need to run more demanding office applications, such as office productivity software, image editing software such as Adobe Photoshop, mainstream computer-aided design (CAD) software such as Autodesk AutoCAD, and product lifecycle management (PLM) applications. These applications are more demanding and require additional graphics resources with full support for APIs such as OpenGL and Direct3D.

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

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

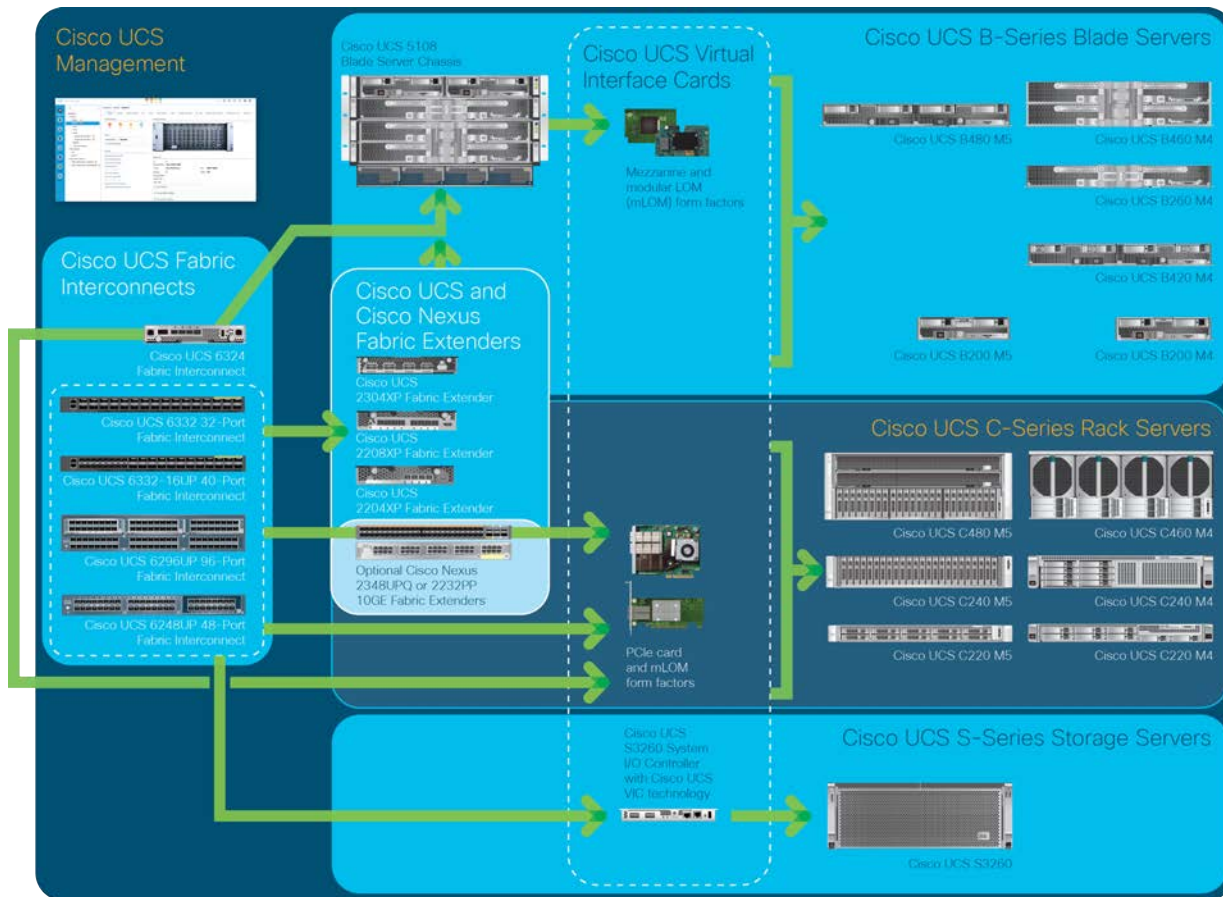
Figure 1. NVIDIA GRID vGPU GPU system architecture



Cisco Unified Computing System

Cisco UCS is a next-generation data center platform that unites computing, networking, and storage access. The platform, optimized for virtual environments, is designed using open industry-standard technologies and aims to reduce total cost of ownership (TCO) and increase business agility. The system integrates a low-latency, lossless 40 Gigabit Ethernet unified network fabric with enterprise-class, x86-architecture servers. It is an integrated, scalable, multichassis platform in which all resources participate in a unified management domain (Figure 2).

Figure 2. Cisco UCS components



The main components of Cisco UCS are:

- **Computing:** The system is based on an entirely new class of computing system that incorporates blade servers and modular servers based on Intel processors.
- **Network:** The system is integrated onto a low-latency, lossless, 40-Gbps unified network fabric. This network foundation consolidates LANs, SANs, and high-performance computing (HPC) networks, which are separate networks today. The unified fabric lowers costs by reducing the number of network adapters, switches, and cables and by decreasing power and cooling requirements.
- **Virtualization:** The system unleashes the full potential of virtualization by enhancing the scalability, performance, and operational control of virtual environments. Cisco security, policy enforcement, and diagnostic features are now extended into virtualized environments to better support changing business and IT requirements.

- **Storage access:** The system provides consolidated access to local storage, SAN storage, and network-attached storage (NAS) over the unified fabric. With storage access unified, Cisco UCS can access storage over Ethernet, Fibre Channel, Fibre Channel over Ethernet (FCoE), and Small Computer System Interface over IP (iSCSI) protocols. This capability provides customers with choice for storage access and investment protection. In addition, server administrators can preassign storage-access policies for system connectivity to storage resources, simplifying storage connectivity and management and helping increase productivity.
- **Management:** Cisco UCS uniquely integrates all system components, enabling the entire solution to be managed as a single entity by Cisco UCS Manager. The manager has an intuitive GUI, a command-line interface (CLI), and a robust API for managing all system configuration processes and operations.

Cisco UCS is designed to deliver:

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

Cisco UCS Manager

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

Cisco UCS 6300 Series Fabric Interconnects

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

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

Features and capabilities include the following:

- Bandwidth of up to 2.56 Tbps of full-duplex throughput
- Thirty-two 40-Gbps Quad Enhanced Small Form-Factor Pluggable (QSFP+) ports in 1RU
- Support for four 10-Gbps breakout cables
- Ports capable of line-rate, low-latency, lossless 40 Gigabit Ethernet and FCoE and 320-Gbps bandwidth per chassis.
- Centralized unified management with [Cisco UCS Manager](#)
- Efficient cooling and serviceability

Figure 3. Cisco UCS 6332 Fabric Interconnect

Front View



Rear View



Cisco UCS C-Series Rack Servers

Cisco UCS C-Series Rack Servers keep pace with Intel Xeon processor innovation by offering the latest processors with increased processor frequency and improved security and availability features. With the increased performance provided by the [Intel Xeon Scalable Family 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 Cisco UCS managed configuration, these servers enable organizations to deploy systems incrementally—using as many or as few servers as needed—on a schedule that best meets the organization’s timing and budget. C-Series servers offer investment protection through the capability to deploy them either as standalone servers or as part of Cisco UCS.

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

Cisco UCS C240 M5 Rack Server

UCS C240 M5 small-form-factor (SFF) server (Figures 4 and 5) 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 capacity points of up to 128 GB, up to 6 PCIe 3.0 slots (Table 1), and up to 26 internal SFF drives. The C240 M5 SFF server also includes one dedicated internal slot for a 12-Gbps SAS storage controller card.

The C240 M5 server includes a dedicated internal mLOM slot for installation of a Cisco virtual interface card (VIC) or third-party network interface card (NIC) without consuming a PCI slot, in addition to two 10GBASE-T Intel x550 LOM ports embedded on the motherboard. The Cisco UCS C240 M5 server can be used as a standalone server or as part of Cisco UCS, which unifies computing, networking, management, virtualization, and storage access into a single integrated architecture enabling end-to-end server visibility, management, and control in both bare-metal and virtualized environments.

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

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

Figure 4. Cisco UCS C240 M5 Rack Server (front view)



Figure 5. Cisco UCS C240 M4 Rack Server (rear view)



Table 1. Cisco UCS C240 M5 PCIe slots

Part number	Description
UCSC-PCI-1-C240M5	Riser 1. Includes 3 PCIe slots (x8, x16, and x8). Slots 1 and 2 are controlled with CPU1; slot 3 is controlled with CPU2 Note: M10, P40, P100, and V100 GPU will be installed in slot 2 (x16).
UCSC-PCI-1B-240M5	Riser 1B. Includes 3 PCIe slots (x8, x8, and x8). All slots are controlled with CPU1.
UCSC-PCI-2A-240M5	Riser 2A. Includes 3 PCIe slots (x16, x16, and x8) and supports a GPU. Note: M10, P40, P100, and V100 GPU will be installed in slot 5 (x16).
UCSC-PCI-2B-240M5	Riser 2B. Includes 3 PCIe slots (x8, x16, and x8) plus 1 Non-Volatile Memory Express (NVMe) connector (controls 2 rear SFF NVMe drives) and supports a GPU. Note: M10, P40, P100, and V100 GPU will be installed in slot 5 (x16).

Cisco UCS C220 M5 Rack Server

The Cisco UCS C220 M5 SFF server (Figures 6 and 7) extends the capabilities of the Cisco UCS portfolio in a 1RU form factor with the addition of the Intel Xeon Scalable Family Processors, 24 DIMM slots for 2666-MHz DIMMs and capacity points of up to 128 GB, 2 PCIe 3.0 slots, and up to 10 SAS and SATA hard-disk drives (HDDs) or solid-state drives (SSDs). The C220 M5 SFF server also includes one dedicated internal slot for a 12-Gbps SAS storage controller card.

The C220 M5 server included one dedicated internal mLOM slot for installation of a Cisco VIC or third-party NIC without consuming a PCI slot, in addition to two 10GBASE-T Intel x550 LOM ports embedded on the motherboard.

The Cisco UCS C220 M5 server can be used as a standalone server or as part of Cisco UCS, which unifies computing, networking, management, virtualization, and storage access into a single integrated architecture enabling end-to-end server visibility, management, and control in both bare-metal and virtualized environments.

For more information about the Cisco UCS C220 M5 Rack Server, see

<https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/c220m5-sff-specsheet.pdf>

Figure 6. Cisco UCS C220 M5 Rack Server (front view)



Figure 7. Cisco UCS C220 M5 Rack Server (rear view)



Cisco UCS C480 M5 Rack Server

The high-performance Cisco UCS C480 M5 Rack Server (Figures 8 and 9) is a 4RU server supporting the Intel Xeon Scalable Family Processors, with up to 6 terabytes (TB) of double-data-rate 4 (DDR4) memory in 48 slots, and up to 32 SFF hot-swappable SAS, SATA, and PCIe NVMe disk drives. Twelve PCIe expansion slots support Cisco UCS C-Series network adapters, storage controllers, and up to six GPUs, with additional I/O provided by two 10GBASE-T LOM ports and one 1 Gigabit Ethernet dedicated out-of-band (OOB) management port. A separate PCIe slot is reserved inside the chassis for a RAID controller card.

For more information about the Cisco UCS C480 M5 Rack Server, see

<https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/c480-m5-high-performance-specsheet.pdf>

Figure 8. Cisco UCS C480 M5 Rack Server (front view)



Figure 9. Cisco UCS C480 M5 Rack Server (rear view)



Figure 10 shows the PCIe slot expansion configuration.

Figure 10. Cisco UCS C480 M5 PCIe slot expansion configuration

- Slot 1: CPU1 controlled, Gen-3 x16, FL, FH, GPU, NCSI, VIC primary
- Slot 2: CPU1 controlled, Gen-3 x16, FL, FH, GPU, NCSI, VIC secondary
- Slot 3: CPU3 controlled, Gen-3 x8, FL, FH, NCSI, VIC
- Slot 4: CPU3 controlled, Gen-3 x16, FL, FH, GPU, NCSI, VIC
- Slot 5: CPU2 controlled, Gen-3 x8, FL, FH, NCSI, VIC
- Slot 6: CPU3 controlled, Gen-3 x16, FL, FH, GPU, NCSI, VIC
- Slot 7: CPU4 controlled, Gen-3 x8, FL, FH, NCSI, VIC
- Slot 8: CPU2 controlled, Gen-3 x16, FL, FH, GPU, NCSI, VIC
- Slot 9: CPU2 controlled, Gen-3 x8, FL, FH
- Slot 10: CPU2 controlled, Gen-3 x16, FL, FH, GPU
- Slot 11: CPU4 controlled, Gen-3 x8, FL, FH
- Slot 12: CPU4 controlled, Gen-3 x8, FL, FH

Note: CPUs must be installed as indicated to support specific slots. The CPU installation options are either CPU1 and CPU2 or CPU1, CPU2, CPU3, and CPU4.

Cisco UCS Virtual Interface Card 1387

The Cisco UCS VIC 1387 (Figure 11) is a dual-port Enhanced Small Form-Factor Pluggable (SFP+) 40 Gigabit 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 11. 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 12) addresses a broad set of workloads, including IT and web infrastructure and distributed databases. The enterprise-class Cisco UCS B200 M5 server extends the capabilities of the Cisco UCS portfolio in a half-width blade form factor. The Cisco UCS B200 M5 harnesses the power of the latest Intel Xeon Scalable Family Processors with up to 3072 GB of RAM (using 128-GB DIMMs), two SSDs or HDDs, and up to 80-Gbps throughput connectivity

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

The UCS 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 Gigabit Ethernet interface or a dual 4 x 10 Gigabit Ethernet port-channeled 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 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 12. Cisco UCS B200 M5 Blade Server (front view)



Cisco UCS Virtual Interface Card 1340

The Cisco UCS VIC 1340 (Figure 13) is a two-port 40 Gigabit Ethernet or dual 4 x 10 Gigabit Ethernet, FCoE-capable mLOM designed exclusively for the M4 generation of Cisco UCS B-Series Blade Servers. When used in combination with an optional port expander, the VIC 1340 is enabled for two ports of 40 Gigabit 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 13. Cisco UCS VIC 1340



NVIDIA Tesla cards

For desktop virtualization applications, the NVIDIA Tesla P4, P6, M10, and P40 cards are an optimal choice for high-performance graphics (Table 2).

Table 2. Technical specifications for NVIDIA Tesla cards



	P6	P4	M10	P40
Number of GPUs	Single mid-range Pascal	Single mid-range Pascal	Quad midrange Maxwell	Single high-end Pascal
NVIDIA Compute Unified Device Architecture (CUDA) cores	2048	2560	2560 (640 per GPU)	3840
Memory size	16-GB GDDR5	8-GB GDDR5	32-GB GDDR5 (8 GB per GPU)	24-GB GDDR5
Maximum number of vGPU instances	16	8	64	24
Power	75 watts (W)	75W	225W	250W
Form factor	Mobile PCI Express Module (MXM) blade servers, with x16 lanes	PCIe 3.0 dual-slot low-profile rack server	PCIe 3.0 dual slot (rack servers), x16 lanes	PCIe 3.0 dual-slot rack servers, with x16 lanes
Cooling solution	Bare board	Passive	Passive	Passive
H.264 1080p30 streams	24	24	28	24
Maximum number of users per board	16 (with 1-Gbps profile)	8 (with 1-Gbps profile)	32 (with 1-G profile)	24 (with 1-Gbps profile)
Virtualization use case	Blade optimized	Inferencing	User-density Optimized	Performance optimized

NVIDIA GRID vGPU

NVIDIA GRID vGPU is the industry's most advanced technology for sharing virtual 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

The NVIDIA GRID solution runs on top of award-winning, [NVIDIA Maxwell/Pascal-powered GPUs](#). These GPUs come in two server form factors: the NVIDIA Tesla [P6](#) for blade servers and converged infrastructure, and the NVIDIA Tesla [M10, P4, P40, P100](#), and [V100](#) for rack and tower servers.

For a list of supported hypervisors, guest OSs, and new features and functions, please refer to the [NVIDIA GRID software release notes](#).

NVIDIA vGPU license requirements

GRID Virtual PC, Quadro Virtual Datacenter Workstation (vDWS), and GRID Virtual Applications are available on a per-concurrent user (CCU) basis. A CCU license is required for every user who is accessing or using the software at any given time, whether or not an active connection to the virtualized desktop or session is maintained.

Note: The CCU model counts licenses based on active user virtual machines. If the virtual machine is active and the NVIDIA vGPU software is running, then this counts as one CCU. A virtual GPU CCU is independent of the connection to the virtual machine.

An NVIDIA vGPU software edition can be purchased either as a perpetual license with an annual Support Updates and Maintenance Subscription (SUMS), or as an annual subscription. The first year of SUMS is required with the purchase of a perpetual license and can then be purchased as a yearly subscription. For annual licenses, SUMS is bundled into the annual license cost.

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

- GRID Virtual Applications
- GRID Virtual PC
- Quadro vDWS

For complete details about GRID software license requirements, see <http://images.nvidia.com/content/grid/pdf/161207-GRID-Packaging-and-Licensing-Guide.pdf>.

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 VDS improves performance and scalability with multicast traffic.
 - Multiple TCP/IP stacks for VMware vMotion: Implement a dedicated networking stack for vMotion traffic, simplifying IP address management with a dedicated default gateway for vMotion traffic.
- Availability
 - vMotion enhancements: Perform nondisruptive live migration of workloads across virtual switches and vCenter Servers and over distances with a round-trip time (RTT) of up to 100 milliseconds (ms). This support for dramatically longer RTT—a 10x increase in the supported time—for long-distance vMotion 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 nongraphical 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 VMware vSphere Hypervisor.

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

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

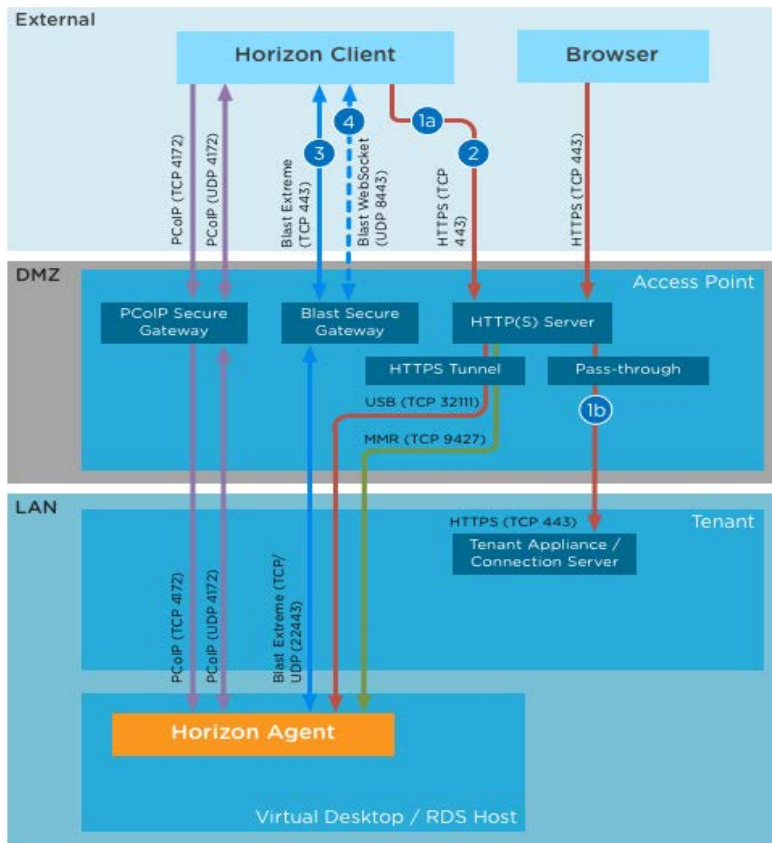
As shown in Figure 14, Blast Extreme provides an enhanced remote session experience introduced with Horizon for Linux desktops, Horizon 7, and Horizon Desktop as a Service (DaaS). In this case, the connection flow from the Horizon Client differs from the flow for PC over IP (PCoIP).

- 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 14. 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. By moving OpenGL, DirectX, Direct3D, and Windows Presentation Foundation (WPF) rendering to the server's GPU, the server's CPU is not slowed by graphics rendering. Additionally, the server can process more graphics because the workload is split between the CPU and the GPU.

GPU sharing for 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 VMware 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 a VMware 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](#).

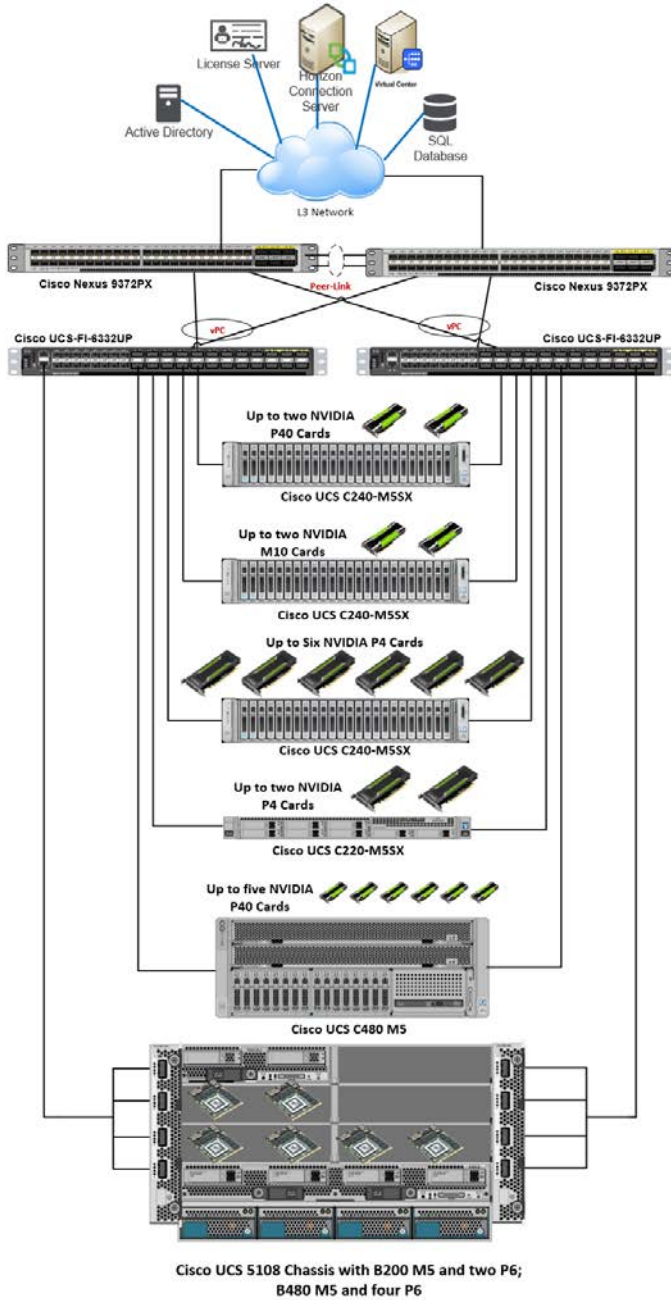
Note: If you configure a pool to use Blast Extreme and do not allow users to choose a protocol, View Connection Server automatically allows PCoIP connections from PCoIP zero clients and older (earlier than Release 4.0) Horizon Clients.

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 15 provides an overview of the solution configuration.

Figure 15. Reference architecture



The hardware components in the solution are:

- Cisco UCS C240 M5 Rack Server (two Intel Xeon Scalable Family Processor Platinum 8176 CPUs at 2.10 GHz) with 768 GB of memory (32 GB x 24 DIMMs at 2666 MHz)
- Cisco UCS C220 M5 Rack Server (two Intel Xeon Scalable Family Processor Platinum 8176 CPUs at 2.10 GHz) with 768 GB of memory (32 GB x 24 DIMMs at 2666 MHz)
- Cisco UCS B200 M5 Blade Server (two Intel Xeon Scalable Family Processor Platinum 8176 CPUs at 2.10 GHz) with 768 GB of memory (32 GB x 24 DIMMs at 2666 MHz)
- Cisco UCS VIC 1387 mLOM (Cisco UCS C240 M5 and C220 M5)
- Cisco UCS VIC 1385 PCIe card for UCS Managed C480 M5*
- Cisco UCS VIC 1340 mLOM (Cisco UCS B200 M5)
- Two Cisco UCS 6332 third-generation fabric interconnects
- NVIDIA Tesla M10, P4, P6, and P40 cards
- Two Cisco Nexus® 9372 Switches (optional access switches)

* Cisco UCS C480 M5 in UCS Managed configuration, up to five NVidia Tesla GPU cards are supported as slot 1 is consumed by Cisco VIC for management and data traffic

The software components of the solution are:

- Cisco UCS Firmware Release 3.2(3g)
- VMware ESXi 6.7 (8169922) for VDI hosts
- VMware Horizon 7.5
- Microsoft Windows 10 64-bit
- Microsoft Server 2016
- NVIDIA GRID 6.2 software and licenses:
 - NVIDIA-VMware_ESXi_6.5_Host_Driver-390.72-1OEM.650.0.0.4598673.vib
 - 391.81_grid_win10_server2016_64bit_international.exe

Configure Cisco UCS

This section describes the Cisco UCS configuration.

Create BIOS policy

Default BIOS policy configured in Cisco UCS M5 Blade and Rack Server contains correct BIOS configuration required to support NVidia Tesla GPU card.

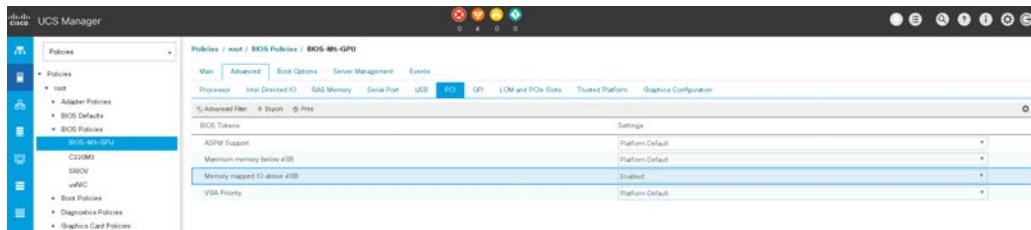
Configuration below is for reference or setting up stand-alone Cisco UCS Rack Server with custom BIOS policy.

Create a new BIOS policy.

1. Right-click BIOS Policy.
2. On the Advanced tab for the new BIOS policy, click PCI.
3. Select settings (Figure 16):

- For “Memory mapped IO above 4GB,” select Enabled.

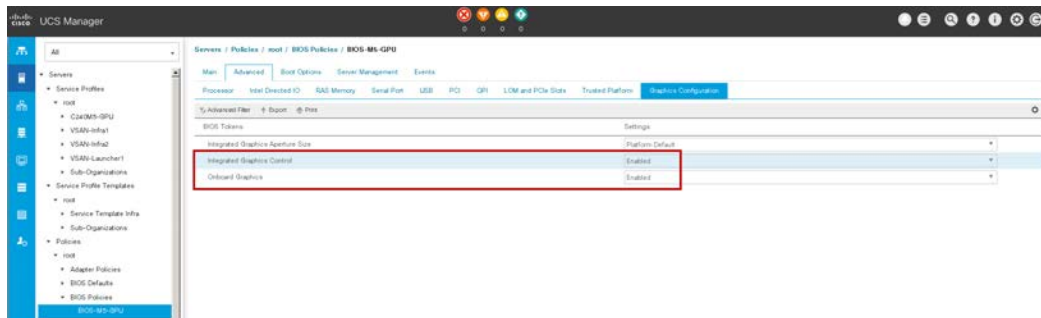
Figure 16. PCI setting for BIOS policy: Enable MMIO above 4 GB



Select graphics configuration settings in the BIOS policy (Figure 17):

- For Integrated Graphics Control, select Enabled.
- For Onboard Graphics, select Enabled.

Figure 17. PCI BIOS policy configuration



Create graphics card policy

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

- For VDI deployment, select Graphics mode (Figure 18).

Figure 18. Graphics card policy



Install NVIDIA Tesla GPU card on Cisco UCS M5

Install the M10, P4, P6, or P40 GPU card on the Cisco UCS M5 rack and blade servers as described here.

The rules for mixing NVIDIA GPU cards are as follows:

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

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

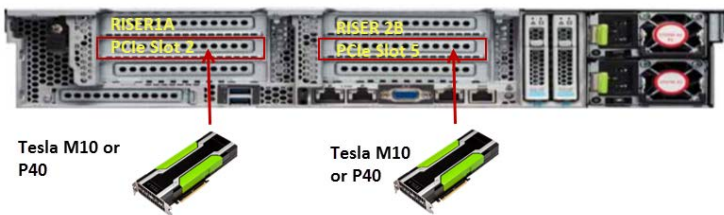
The rules for configuring the server with GPUs differ, depending on the server version and other factors. Figures 19 to 23 show GPU installation on the M5 rack server.

1. Install the NVIDIA Tesla GPU card on the C240 M5 (Figures 19 and 20).

Figure 19. NVIDIA Tesla P4 installation on Cisco UCS C240 M5



Figure 20. NVIDIA Tesla M10 or P40 installation on Cisco UCS C240 M5



2. Install the NVIDIA Tesla GPU card on the Cisco UCS C220 M5 (Figure 21).

Figure 21. NVIDIA Tesla P4 installation on Cisco UCS C220 M5



3. Install the NVIDIA Tesla GPU card on the Cisco UCS C480 M5 (Figure 22).

Figure 22. NVIDIA Tesla P40 installation on Cisco UCS C480 M5



Note: In a Cisco UCS managed C480 M5 server, slot 1 will be consumed by the Cisco VIC, and only five NVIDIA Tesla P40 GPUs can be installed in slots 2, 4, 6, 8, and 10. To configure six GPU cards, the Cisco UCS C480 M5 server must be in standalone mode.

Install NVIDIA Tesla GPU card on Cisco UCS B200 M5

Before installing the NVIDIA P6 GPU, do the following:

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

Figure 23 shows the B200 M5 with two P6 GPU cards.

Figure 23. Cisco UCS B200 M5 Blade Server with two P6 GPU cards



Configure the GPU card

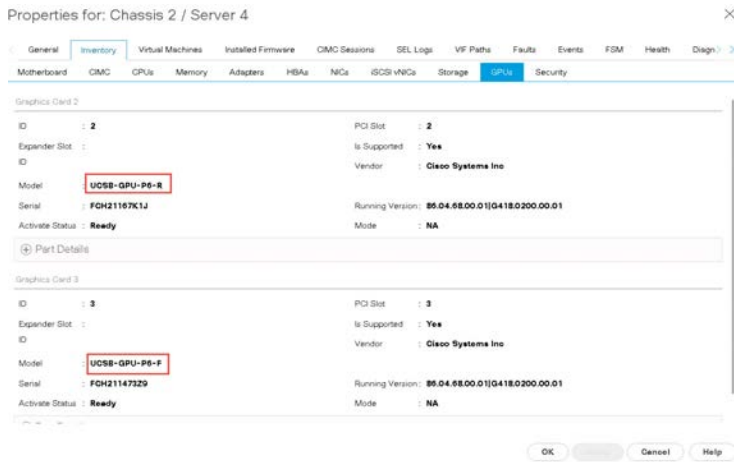
Follow these steps to configure the GPU card.

NVIDIA P6 installation

After the NVIDIA P6 GPU cards are physically installed and the Cisco UCS B200 M5 Blade Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 24, PCIe slots 2 and 3 are used with two GRID P6 cards.

Note: As highlighted in the red boxes in Figure 24, different part numbers are used for the front and rear GPU cards.

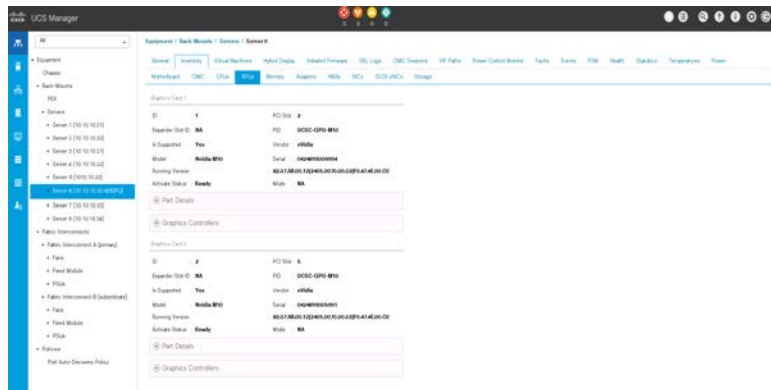
Figure 24. NVIDIA GRID P6 card inventory displayed in Cisco UCS Manager



NVIDIA M10 installation

After the NVIDIA M10 GPU cards are physically installed and the Cisco UCS C240 M5 Rack Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 25, PCIe slots 2 and 5 are used with two GRID M10 cards.

Figure 25. NVIDIA GRID M10 card inventory displayed in Cisco UCS Manager



NVIDIA P40 installation

After the NVIDIA P40 GPU card is physically installed in the Cisco UCS C480 M5 Rack Server in stand-alone mode, Login to CIMC and browse to Inventory > PCI Adapters. As shown in Figure 26, PCIe slots 1, 2, 4, 6, 8 and 10 are configured with six GRID P40 card.

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 27, PCIe slots 2 and 5 are used with the two GRID P40 card.

Figure 26. NVIDIA GRID P40 card inventory displayed in the Cisco Integrated Management Controller (IMC) for Cisco UCS C480 M5

Slot ID	Product Name	Option ROM Status	Firmware Version	Vendor ID	Sub Vendor ID	Device ID	Sub Device ID
1	nVidia P40 PASCAL_PASSIVE_250W_FF 3.0_24GB	Not Loaded	N/A	0x10de	0x10de	0x1038	0x1109
10	nVidia P40 PASCAL_PASSIVE_250W_FF 3.0_24GB	Not Loaded	N/A	0x10de	0x10de	0x1038	0x1109
11	QLogic Serial bus controller	Not Loaded	N/A	0x1077	0x1077	0x2031	0x0249
2	nVidia P40 PASCAL_PASSIVE_250W_FF 3.0_24GB	Not Loaded	N/A	0x10de	0x10de	0x1038	0x1109
4	nVidia P40 PASCAL_PASSIVE_250W_FF 3.0_24GB	Not Loaded	N/A	0x10de	0x10de	0x1038	0x1109
6	nVidia P40 PASCAL_PASSIVE_250W_FF 3.0_24GB	Not Loaded	N/A	0x10de	0x10de	0x1038	0x1109
8	nVidia P40 PASCAL_PASSIVE_250W_FF 3.0_24GB	Not Loaded	N/A	0x10de	0x10de	0x1038	0x1109
L	Cisco(R) LOM X550-T2	Loaded	0x8000AF5-1.812.1	0x8086	0x1137	0x1583	0x0145

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

Equipment / Rack Mounts / Servers / Server6

General | Inventory | Virtual Machines | Hybrid Display | Installed Firmware | SEL Logs | CIMC Sessions | VIF Paths

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

Graphics Card 2

ID	: 2	PCI Slot	: 6
Expander Slot ID	: NA	Is Supported	: Yes
Vendor	: nVidia	Model	: Nvidia P40
Serial	: 0321017007344	Running	: 80.02.23.00.01 G010.0200.00.03
Version			
Activate Status	: Ready	Mode	: NA

Part Details

Graphics Controllers

Graphics Card 1

ID	: 1	PCI Slot	: 2
Expander Slot ID	: NA	Is Supported	: Yes
Vendor	: nVidia	Model	: Nvidia P40
Serial	: 0321517008182	Running	: 80.02.23.00.01 G010.0200.00.03
Version			
Activate Status	: Ready	Mode	: NA

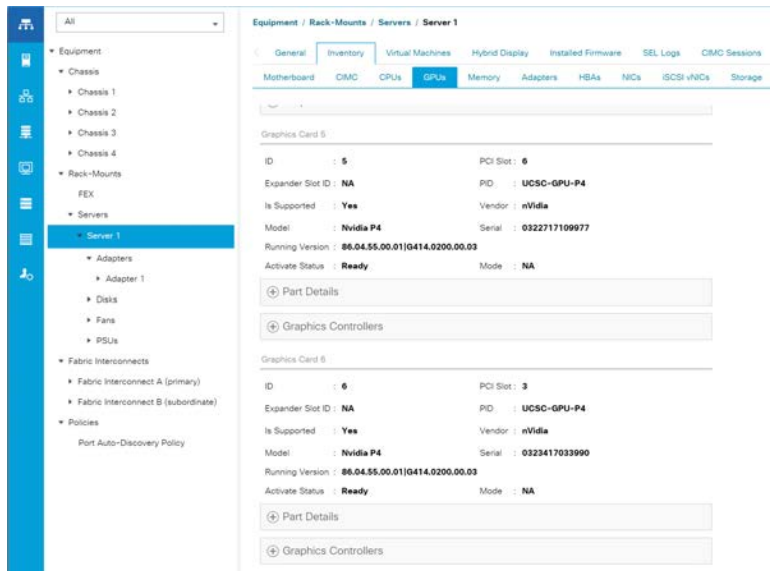
Part Details

Graphics Controllers

NVIDIA P4 installation

After the NVIDIA P4 GPU card is physically installed and the Cisco UCS C220 M5 or C240 M5 Rack Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 28, PCIe slots 1 through 6 in the C240 M5 are used with six GRID P4 cards or two GRID P4 cards in the C220 M5.

Figure 28. NVIDIA GRID P4 card inventory displayed in Cisco UCS Manager



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

Install NVIDIA GRID License Server

This section summarizes the installation and configuration process for the GRID 5.0 License Server.

The NVIDIA GRID vGPU is a licensed feature on Tesla P4, P6, P40, and M10 cards. 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/vgpu-evaluation.html#utm_source=shorturl&utm_medium=referrer&utm_campaign=grideval.

Three packages are required for VMware ESXi host setup, as shown in Figure 29:

- GRID License Server installer
- NVIDIA GRID Manager software, which is installed on VMware vSphere ESXi; the NVIDIA drivers and software that are installed in Microsoft Windows are also in this folder
- GPU Mode Switch utility, which changes the cards from the default Compute mode to Graphics mode (Ofor a Maxwell-based GPU)

Figure 29. 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 GRID License Server

The steps shown here use the Microsoft Windows version of the license server installed on Windows Server 2016. A Linux version of the license server is also available.

The GRID License Server requires Java Version 7 or later. Go to Java.com and install the latest version.

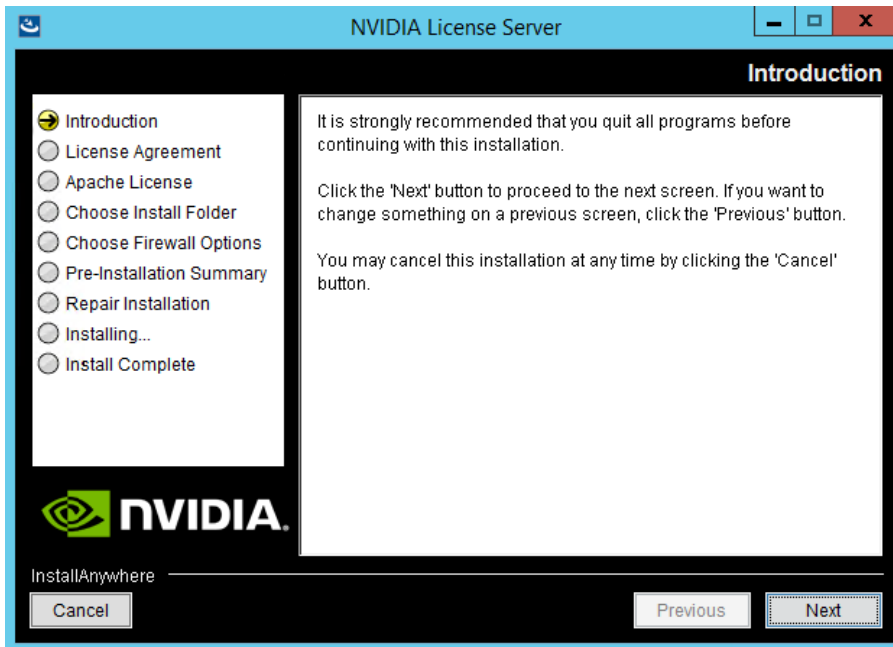
1. Extract and open the NVIDIA-Is-windows-2018.06.0.24304595 folder. Run setup.exe (Figure 30).

Figure 30. Run setup.exe

Name	Date modified	Type	Size
grid-license-server-release-notes	8/2/2018 5:53 PM	Chrome HTML Docu...	1,626 KB
grid-license-server-user-guide	8/2/2018 5:53 PM	Chrome HTML Docu...	3,587 KB
grid-licensing-user-guide	8/2/2018 5:53 PM	Chrome HTML Docu...	2,046 KB
grid-software-quick-start-guide	8/2/2018 5:53 PM	Chrome HTML Docu...	3,526 KB
setup	8/2/2018 5:53 PM	Application	238,574 KB

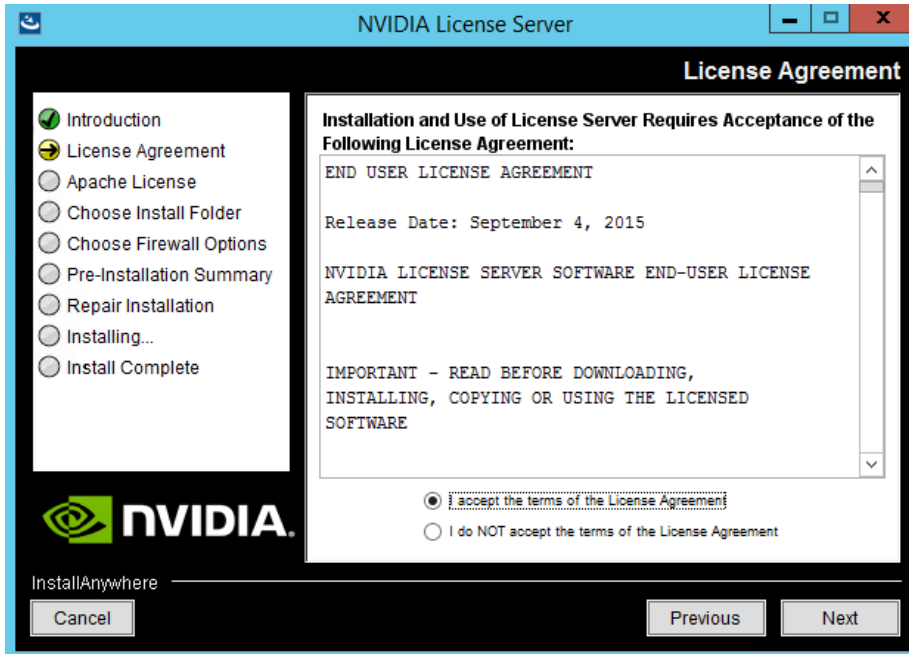
2. Click Next (Figure 31).

Figure 31. NVIDIA License Server



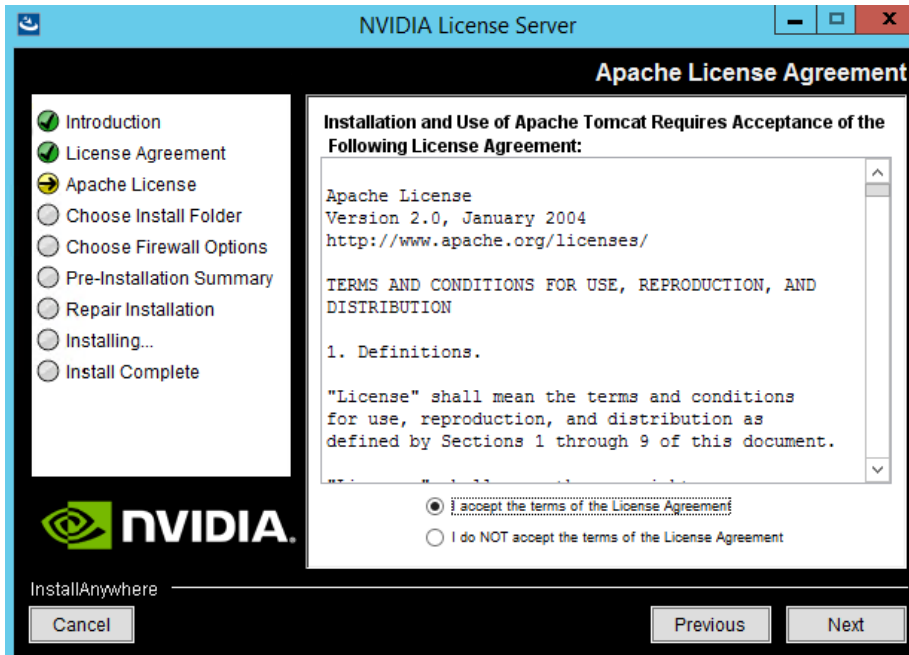
3. Accept the NVIDIA License Agreement and click Next (Figure 32).

Figure 32. NVIDIA License Agreement

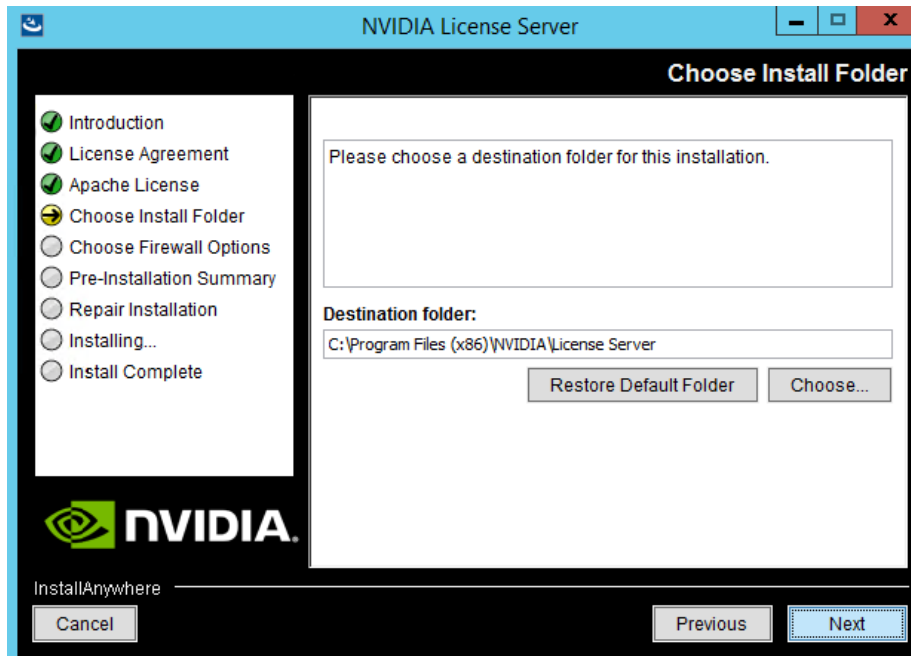


4. Accept the Apache License Agreement and click Next (Figure 33).

Figure 33. Apache License Agreement

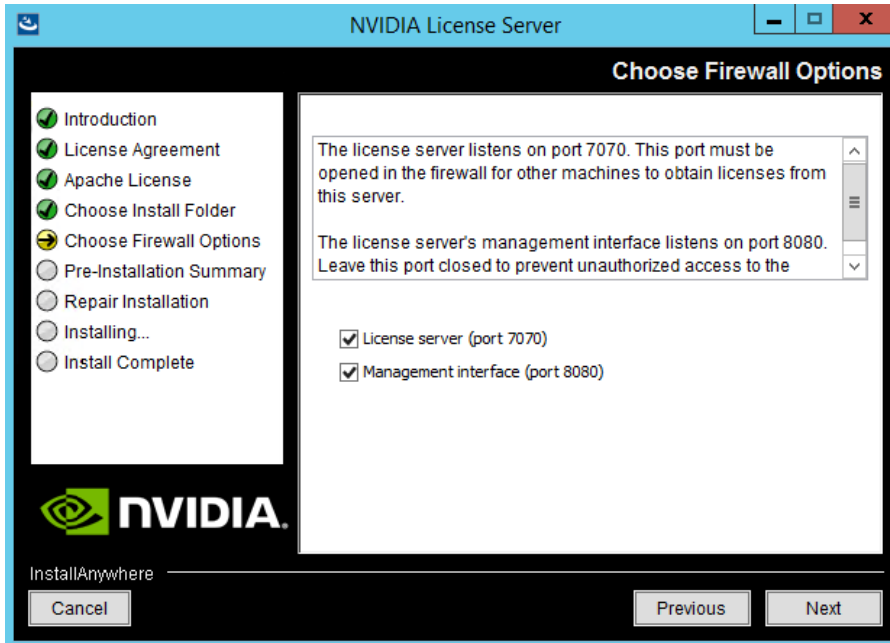


5. Choose the desired installation folder and click Next (Figure 34).

Figure 34. Choosing a destination folder

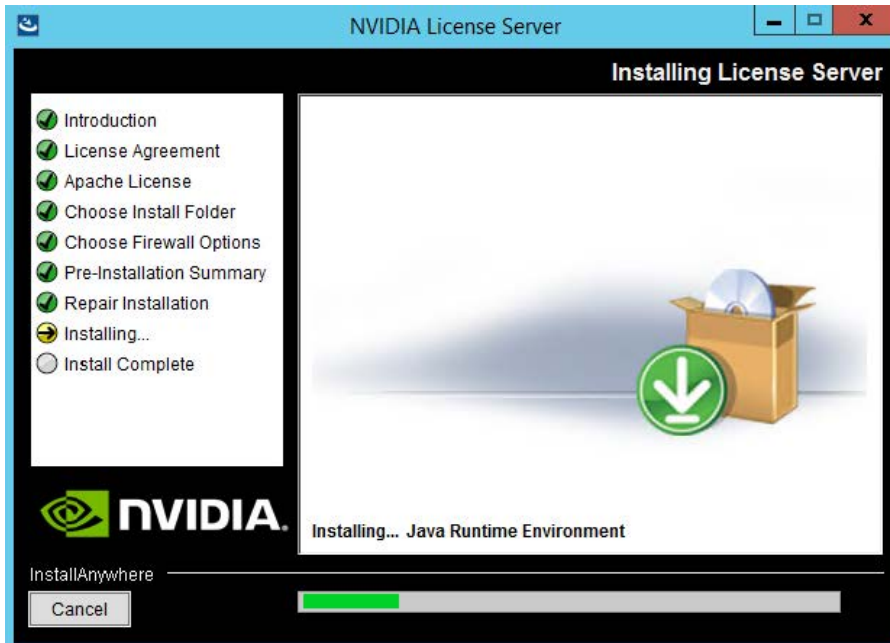
6. The license server listens on port 7070. This port must be opened in the firewall for other machines to obtain licenses from this server. Select the “License server (port 7070)” option.
7. The license server’s management interface listens on port 8080. If you want the administration page accessible from other machines, you will need to open up port 8080. Select the “Management interface (port 8080)” option.
8. Click Next (Figure 35).

Figure 35. Setting firewall options



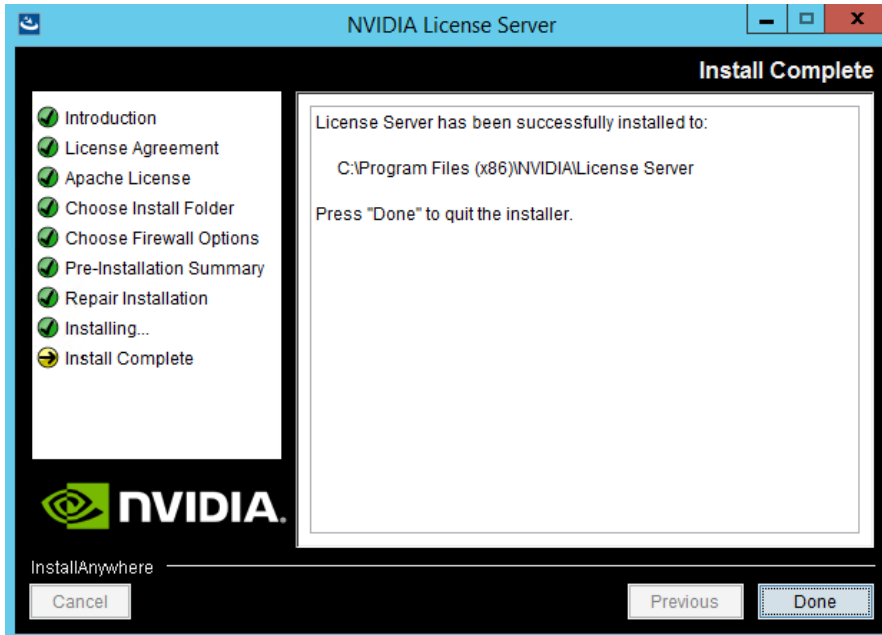
9. The Pre-installation Summary and Repair Installation options automatically progress without user input (Figure 36).

Figure 36. Installing the license server



10. When the installation process is complete, click Done (Figure 37).

Figure 37. Installation complete

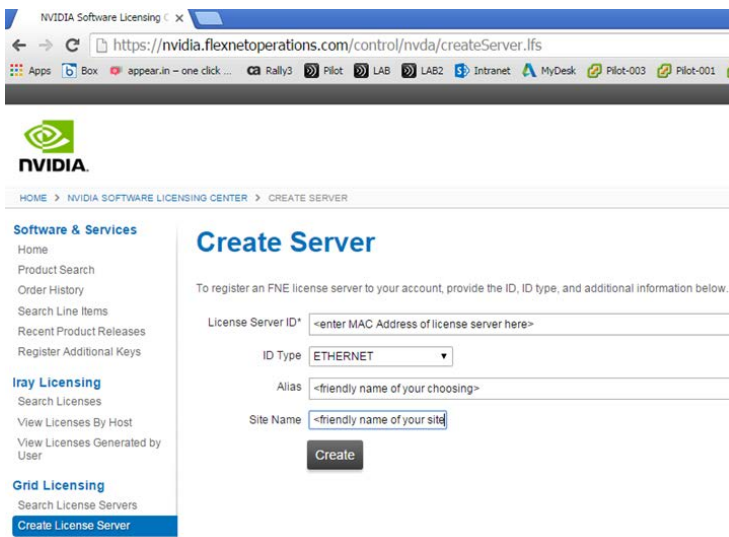


Configure NVIDIA GRID 6.2 License Server

Now configure the NVIDIA GRID License Server.

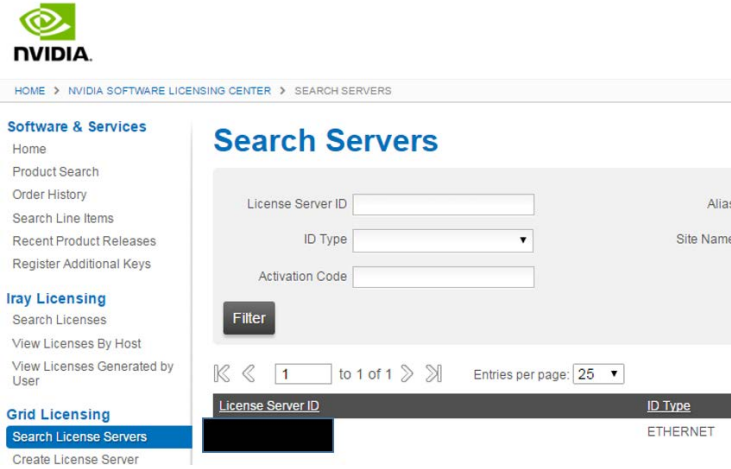
1. Log in to the license server site with the credentials set up during the registration process at nvidia.com/grideval. A license file is generated from <https://nvidia.flexnetoperations.com>.
2. After you are logged in, click Create License Server.
3. Specify the fields as shown in Figure 38. In the License Server ID field, enter the MAC address of your local license server’s NIC. Leave the ID Type set to Ethernet. For the Alias and Site Name, choose user-friendly names. Then click Create.

Figure 38. Creating the license server



4. Click the Search License Servers node.
5. Click your license server ID (Figure 39).

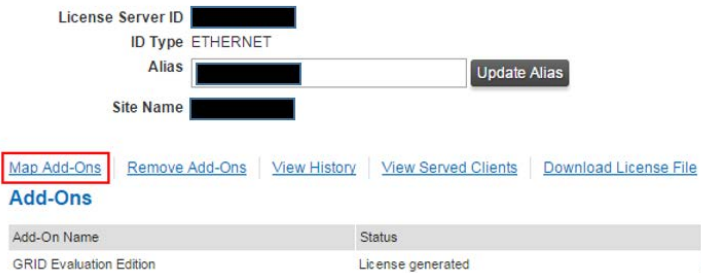
Figure 39. Selecting the license server ID



6. Click Map Add-Ons and choose the number of license units out of your total pool to allocate to this license server (Figure 40).

Figure 40. Choosing the number of license units from the pool

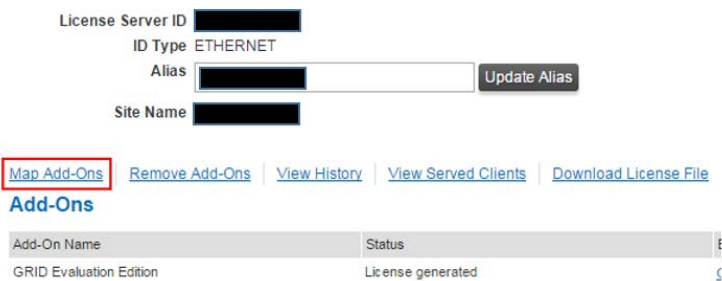
View Server



After the add-ons are mapped, the interface will look like Figure 41, showing 128 units mapped, for example.

Figure 41. Display of mapped units

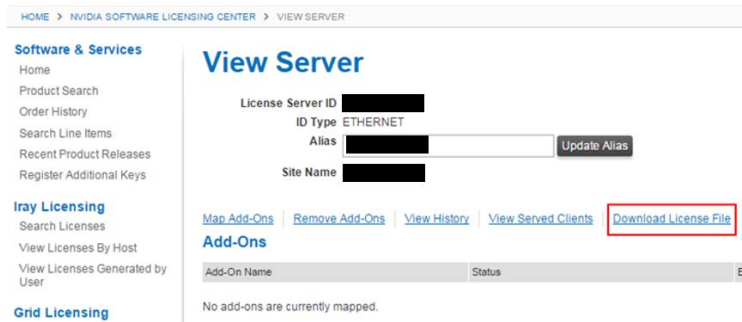
View Server



7. Click Download License File and save the .bin file to your license server (Figure 42).

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

Figure 42. 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 43). The features are available (Figure 44).

Figure 43. License file successfully applied

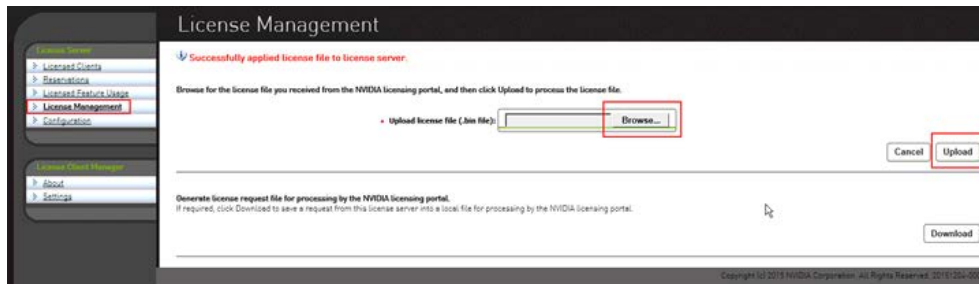
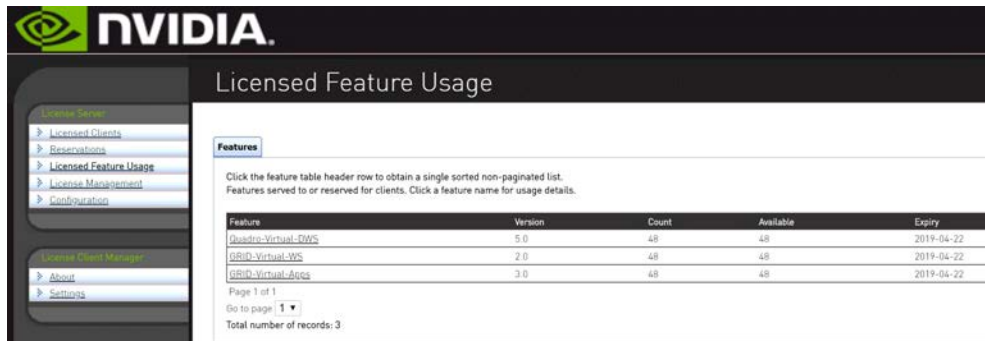


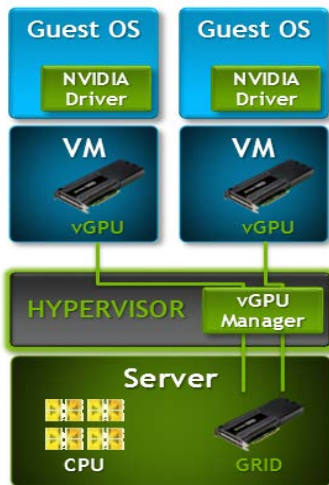
Figure 44. NVIDIA License Server with features available for use



Install NVIDIA GRID software on the VMware ESXi host

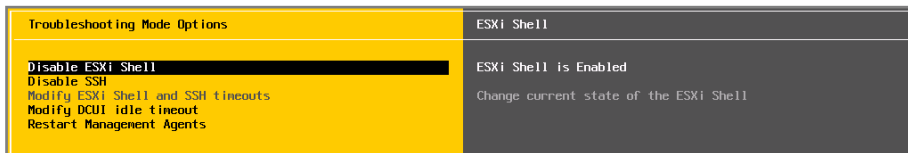
This section summarizes the installation process for configuring an ESXi host and virtual machine for vGPU support. Figure 45 shows the components used for vGPU support.

Figure 45. NVIDIA GRID vGPU components



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 46).

Figure 46. VMware ESXi Configuration Console



3. Upload the NVIDIA driver (vSphere Installation Bundle [VIB] file) to the /tmp directory on the ESXi host using a tool such as WinSCP. (Shared storage is preferred if you are installing drivers on multiple servers or using the VMware Update Manager.)
4. Log in as root to the vSphere console through SSH using a tool such as Putty.
5. 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
```
6. 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:
```

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

7. 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
```

8. 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.

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

```
nvidia-smi
```

The command should return output similar to that shown in Figures 47, 48, 49, or 50, depending on the cards used in your environment.

Figure 47. VMware ESX SSH console report for GPU P4 card detection on Cisco UCS C240 M5 Rack Server

```
[root@C240M5-P4:~] nvidia-smi
Fri Aug 3 04:16:04 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   33C   P8     11W /  75W |  21MiB /  8191MiB |      0%      Default  |
+-----+-----+
|  1   Tesla P4     On         | 00000000:5E:00.0 Off  |    Off                |
| N/A   32C   P8     11W /  75W |  21MiB /  8191MiB |      0%      Default  |
+-----+-----+
|  2   Tesla P4     On         | 00000000:86:00.0 Off  |    Off                |
| N/A   31C   P8     10W /  75W |  21MiB /  8191MiB |      0%      Default  |
+-----+-----+
|  3   Tesla P4     On         | 00000000:AF:00.0 Off  |    Off                |
| N/A   34C   P8     11W /  75W |  21MiB /  8191MiB |      0%      Default  |
+-----+-----+
|  4   Tesla P4     On         | 00000000:D8:00.0 Off  |    Off                |
| N/A   31C   P8     11W /  75W |  21MiB /  8191MiB |      0%      Default  |
+-----+-----+
|  5   Tesla P4     On         | 00000000:D9:00.0 Off  |    Off                |
| N/A   29C   P8     11W /  75W |  21MiB /  8191MiB |      0%      Default  |
+-----+-----+

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

[root@C240M5-P4:~] █
```

Figure 48. VMware ESX SSH console report for GPU P40 card detection on Cisco UCS C240 M5 Rack Server

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

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

[root@C240-M5:~] █
```

Figure 49. VMware ESX SSH console report for GPU M10 card detection on Cisco UCS C240 M5 Rack Server

```

GPU Name Persistence-M Bus-Id Disp.A Volatile Uncorr. ECC
Fan Temp Perf Pwr:Usage/Cap Memory-Usage GPU-Util Compute M.
-----
 0 Tesla M10 On | 0000:60:00.0 Off | N/A
N/A 29C P8 10W / 53W | 18MiB / 8191MiB | 0% Default |
-----
 1 Tesla M10 On | 0000:61:00.0 Off | N/A
N/A 30C P8 10W / 53W | 18MiB / 8191MiB | 0% Default |
-----
 2 Tesla M10 On | 0000:62:00.0 Off | N/A
N/A 26C P8 10W / 53W | 18MiB / 8191MiB | 0% Default |
-----
 3 Tesla M10 On | 0000:63:00.0 Off | N/A
N/A 26C P8 10W / 53W | 18MiB / 8191MiB | 0% Default |
-----
 4 Tesla M10 On | 0000:88:00.0 Off | N/A
N/A 27C P8 10W / 53W | 18MiB / 8191MiB | 0% Default |
-----
 5 Tesla M10 On | 0000:89:00.0 Off | N/A
N/A 28C P8 10W / 53W | 18MiB / 8191MiB | 0% Default |
-----
 6 Tesla M10 On | 0000:8A:00.0 Off | N/A
N/A 25C P8 10W / 53W | 18MiB / 8191MiB | 0% Default |
-----
 7 Tesla M10 On | 0000:8B:00.0 Off | N/A
N/A 24C P8 10W / 53W | 18MiB / 8191MiB | 0% Default |
-----

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

Figure 50. VMware ESX SSH console report for GPU P6 card detection on Cisco UCS B200 M5 Blade Server

```

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

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

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

[root@M5:~] █
    
```

Note: The NVIDIA system management interface (SMI) also allows GPU monitoring using the following command (this command adds a loop, automatically refreshing the display): `nvidia-smi -l`.

NVIDIA Tesla P4, P6, P40, and M10 profile specifications

The Tesla P4, P6, and P40 cards each have a single physical GPU, and the Tesla M10 card has multiple physical GPUs. Each physical GPU can support several 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 designed for a different class of workload. Table 3 lists the vGPU types supported by GRID GPUs.

For more information, see <https://docs.nvidia.com/grid/latest/pdf/grid-vgpu-user-guide.pdf>.

Table 3. User profile specifications for NVIDIA Tesla cards

vGPU profile options			
End-user profile	GRID Virtual Application profiles	GRID Virtual PC profiles	GRID Virtual Workstation profiles
1 GB	P6-1A M10-1A P40-1A P4-1A	P6-1B M10-1B P40-1B P4-1B	P6-1Q M10-1Q P40-1Q P4-1Q
2 GB	P6-2A M10-2A P40-2A P4-2A	P6-2B P6-2B4 P4-2B P4-2B4 P40-2B P40-2B4	P6-2Q M10-2Q P40-2Q P4-2Q
3 GB	P40-3A	N/A	P40-3Q
4 GB	P6-4A M10-4A P40-4A P4-4A	N/A	P6-4Q M10-4Q P40-4Q P4-4Q
6GB	P40-6A	N/A	P40-6Q
8 GB	P6-8A M10-8A P40-8 P4-8A	N/A	P6-8Q M10-8Q P40-8Q P4-8Q
12 GB	P40-12A	N/A	P40-12Q
16 GB	P6-16A	N/A	P6-16Q
24 GB	P40-24A	N/A	P40-24Q
Pass-through	N/A	N/A	N/A

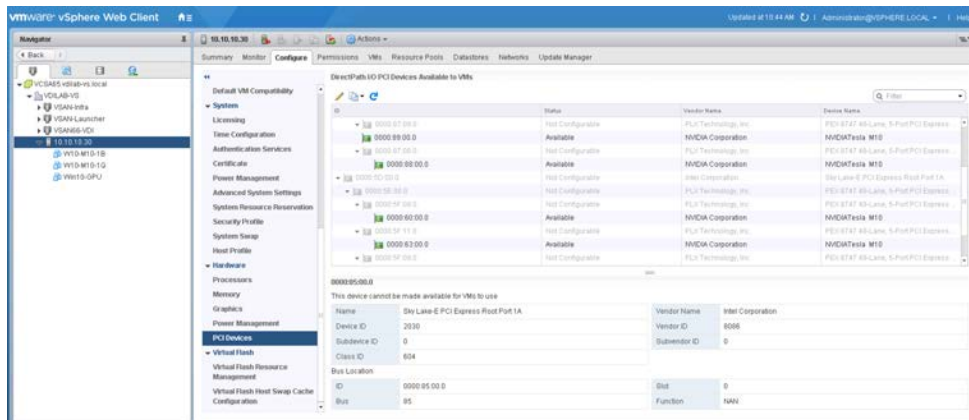
Note: GRID software release v6.2 supports a new type of vGPU profile, PXX-2B4, which supports four displays with 2560 × 1600 resolution.

VMware ESXi host configuration for vDGA (pass-through) or vGPU (Virtual GPU)

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

1. To configure pass-through mode for the NVIDIA GPU, navigate to vSphere Web Client > host Settings > PCI Devices and make GPUs available for pass-through (Figure 51).

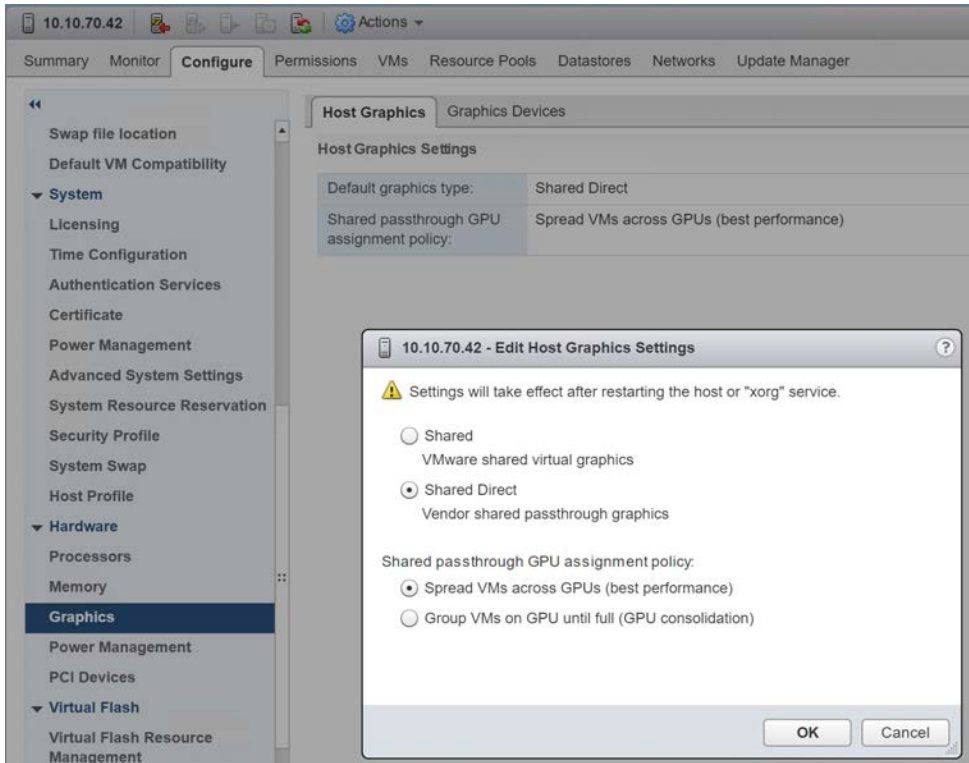
Figure 51. Make GPUs available for pass-through



2. To configure Virtual GPU on the ESXi, follow the steps outlined earlier in the document to install the NVIDIA GRID License Server, the NVIDIA GRID vGPU software on the ESXi host, and the switch mode utility for a Maxwell-based NVIDIA GPU card.
 - a. Select ESXi host and click the Configure tab. From the list of options on the left, select Graphics. Click Edit Host Graphics Settings and select the following settings (Figure 52):
 - Shared direct: Vendor shared passthrough graphics
 - Spread VMs across GPUs (best performance)

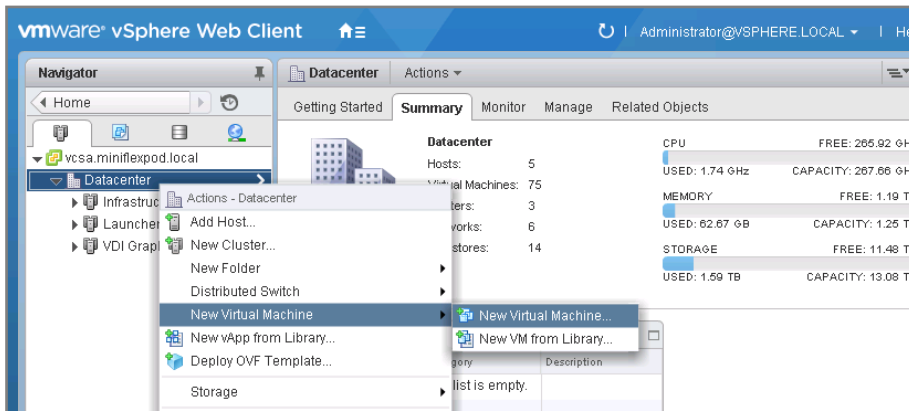
Note: A reboot is required to make the changes take effect.

Figure 52. Editing host graphics settings



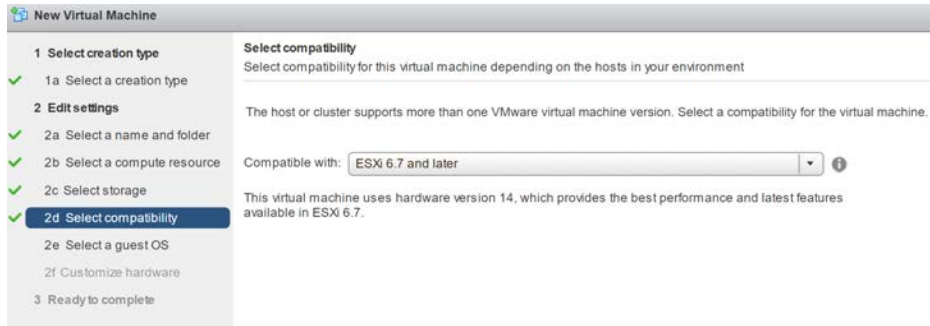
- b. Using the vSphere Web Client or HTML5 interface (vSphere 6.7), create a new virtual machine. To do this, right-click a host or cluster and choose New Virtual Machine. Work through the New Virtual Machine wizard. Unless another configuration is specified, select the configuration settings appropriate for your environment (Figure 53).

Figure 53. Creating a new virtual machine in VMware vSphere Web Client



- c. Choose "ESXi 6.7 and later" from the "Compatible with" drop-down menu to use the latest features, including the mapping of shared PCI devices, which is required for the vGPU feature (Figure 54). "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.

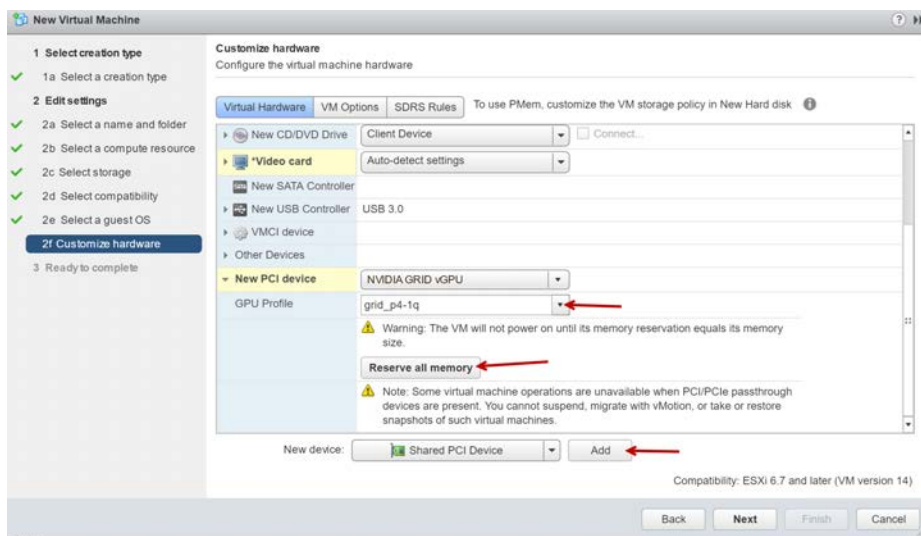
Figure 54. Selecting the virtual machine version and compatibility



- d. 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 55).

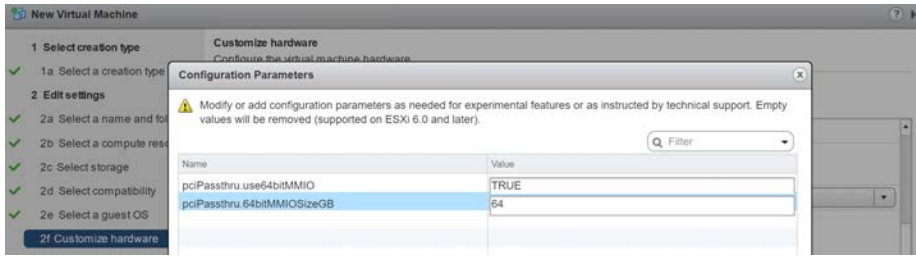
Note: vGPU allocation to a virtual machine requires the reservation of all guest memory. Click “Reserve all memory.”

Figure 55. Adding a shared PCI device to the virtual machine to attach the GPU profile



- e. Open the VM Options tab, As shown in Figure 56, from the drop-down menu under Advanced, choose Configuration Parameters (*) and click Edit Configuration. Add names as shown here:
 - For pciPassthru.use64bitMIMO, set the value to TRUE.
 - For pciPassthru.64bitMMIOSizeGB, set the value to 64.

Figure 56. Setting configuration parameters

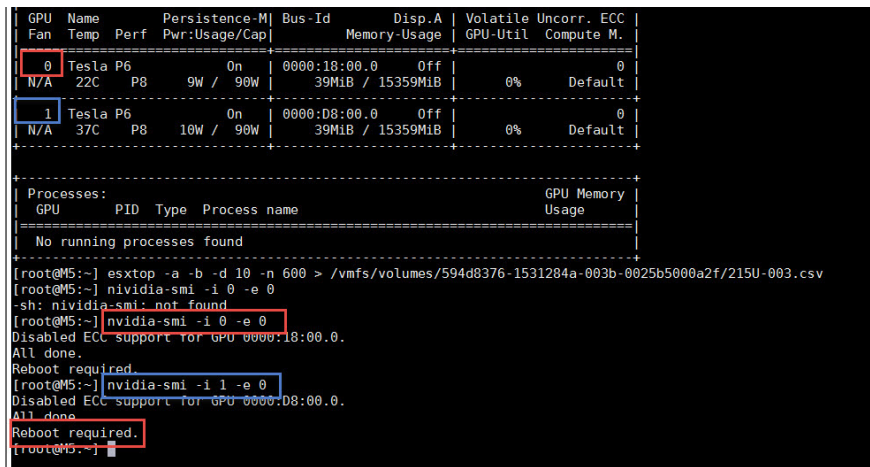


Note: A virtual machine with vGPU assigned will not start if ECC is enabled. As a workaround, disable ECC by entering the following commands (Figure 57):

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

Note: Use `-i` to target a specific GPU. If two cards are installed in a server, run the command twice as shown here, using 0 and 1 for the two GPU cards.

Figure 57. Disabling ECC



- f. Install and configure Microsoft Windows on the virtual machine:
 - a. Configure the virtual machine with the appropriate amount of vCPU and RAM according to the GPU profile selected.
 - b. Install VMware Tools.
 - c. Join the virtual machine to the Microsoft Active Directory domain.
 - d. Choose “Allow remote connections to this computer” on the Windows System Properties menu.
 - e. Install VMware Horizon Agent with appropriate settings. Enable the remote desktop capability if prompted to do so.
 - f. Install Horizon Direct Connection Agent.

g. 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. Because most Windows system services are enabled by default, the optimization tool can be used to easily disable unnecessary services and features to improve performance.

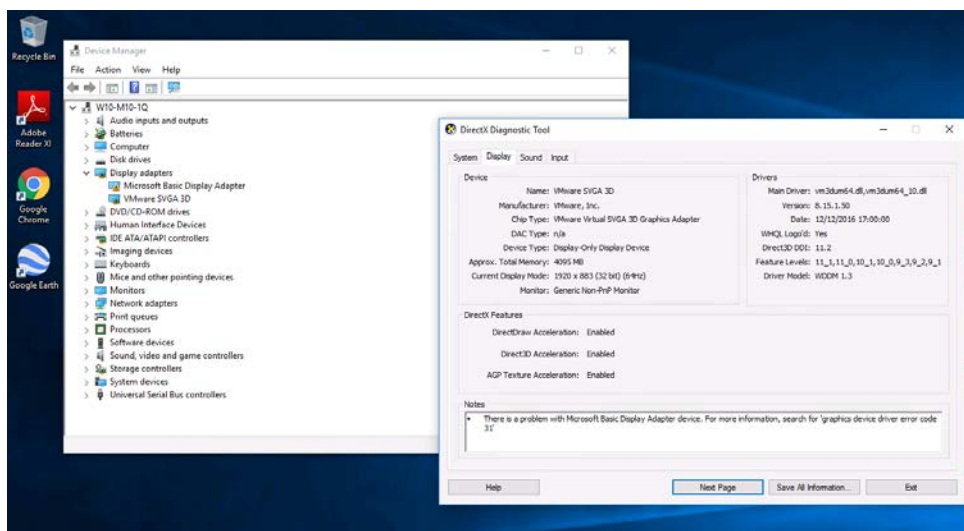
h. Restart the Windows OS when prompted to do so.

Install the NVIDIA vGPU software driver

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

Before the NVIDIA driver is installed on the guest virtual machine, the Device Manager shows the standard VGA graphics adapter (Figure 58).

Figure 58. 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 59).

Figure 59. NVIDIA driver pack

Name	Date modified	Type	Size
390.72-390.75-391.81-grid-gpumodeswitch-user-guide	8/2/2018 5:53 PM	Chrome HTML Docu...	1,633 KB
390.72-390.75-391.81-grid-license-server-release-notes	8/2/2018 5:53 PM	Chrome HTML Docu...	1,626 KB
390.72-390.75-391.81-grid-license-server-user-guide	8/2/2018 5:53 PM	Chrome HTML Docu...	3,591 KB
390.72-390.75-391.81-grid-licensing-user-guide	8/2/2018 5:53 PM	Chrome HTML Docu...	2,046 KB
390.72-390.75-391.81-grid-software-quick-start-guide	8/2/2018 5:53 PM	Chrome HTML Docu...	3,525 KB
390.72-390.75-391.81-grid-vgpu-release-notes-vmware-v...	8/2/2018 5:53 PM	Chrome HTML Docu...	1,706 KB
390.72-390.75-391.81-grid-vgpu-user-guide	8/2/2018 5:53 PM	Chrome HTML Docu...	5,841 KB
391.81_grid_win8_win7_32bit_international	8/2/2018 5:53 PM	Application	187,949 KB
391.81_grid_win8_win7_server2012R2_server2008R2_64bit...	8/2/2018 5:53 PM	Application	266,800 KB
391.81_grid_win10_32bit_international	8/2/2018 5:53 PM	Application	208,005 KB
391.81_grid_win10_server2016_64bit_international	8/2/2018 5:53 PM	Application	313,465 KB
NVIDIA-Linux-x86_64-390.75-grid.run	8/2/2018 5:53 PM	RUN File	86,055 KB
NVIDIA-VMware_ESXi_6.7_Host_Driver-390.72-10EM.670...	8/2/2018 5:53 PM	VIB File	25,265 KB
NVIDIA-VMware_ESXi_6.7_Host_Driver-390.72-10EM.670...	8/2/2018 5:53 PM	Compressed (zipped)...	24,520 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.

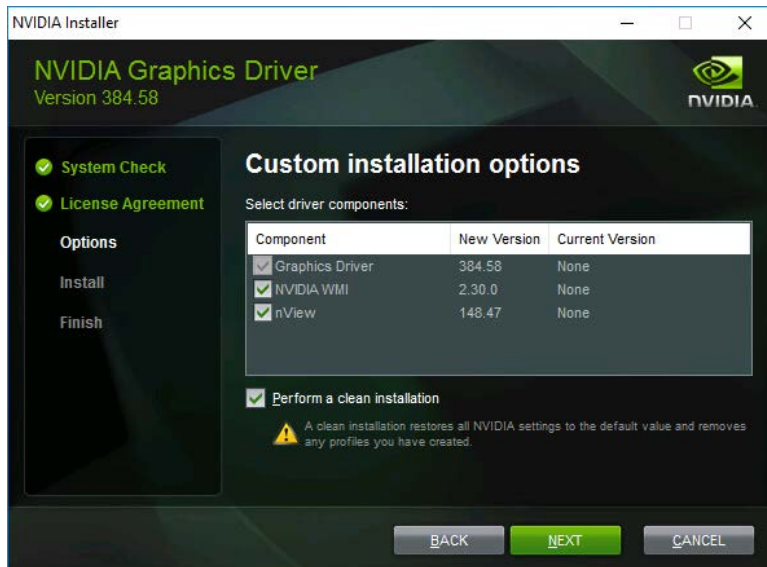
3. Install the graphics drivers using the Express/Custom option (Figure 60). After the installation has been completed successfully (Figure 61), restart the virtual machine.

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

Figure 60. Select the Express/Custom installation option



Figure 61. Components to be installed during the NVIDIA graphics driver installation process.

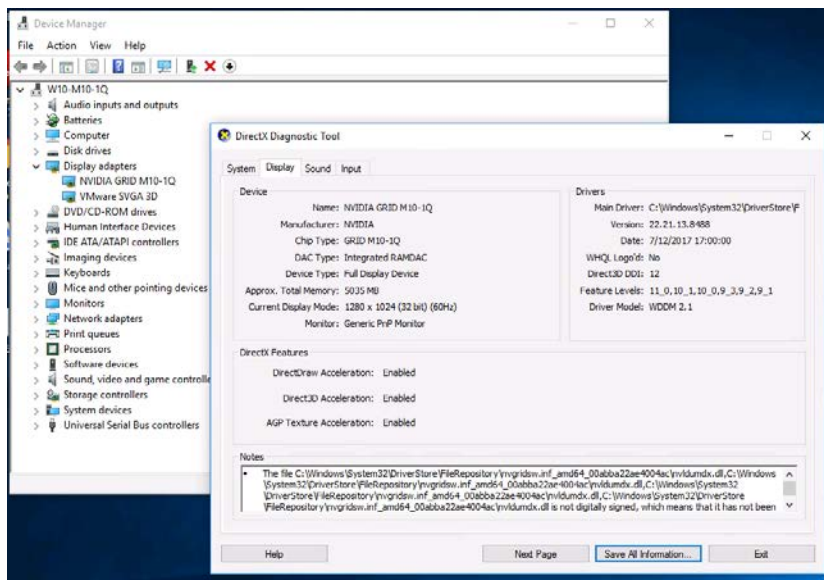


Verify that applications are ready to support NVIDIA vGPU

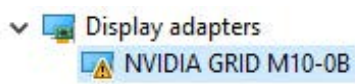
Verify the successful installation of the graphics drivers and the vGPU device.

Open Windows Device Manager and expand the Display Adapter section. The device will reflect the chosen profile (Figure 62).

Figure 62. Validating the driver installation



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



The following are the most likely the reasons:

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

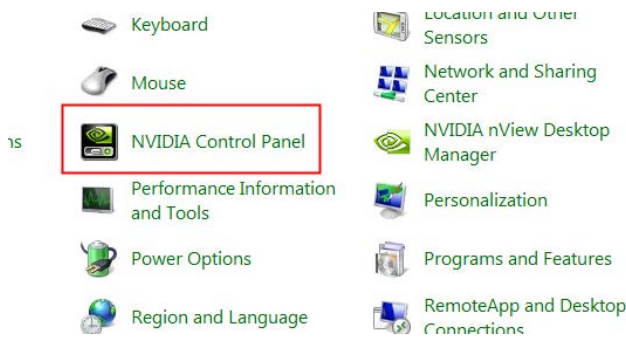
Configure the virtual machine for an NVIDIA GRID vGPU license

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

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

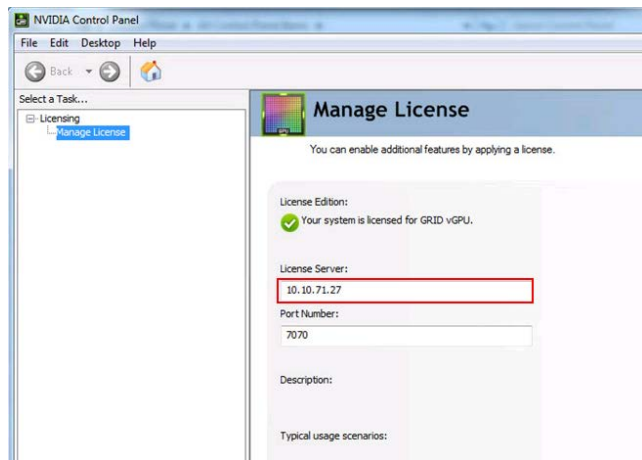
1. In the Microsoft Windows Control Panel, double-click NVIDIA Control Panel (Figure 63).

Figure 63. Choosing the NVIDIA Control Panel



2. In the left pane, select Manage License and enter your license server address and port (Figure 64).

Figure 64. Managing your license



3. Select Apply.

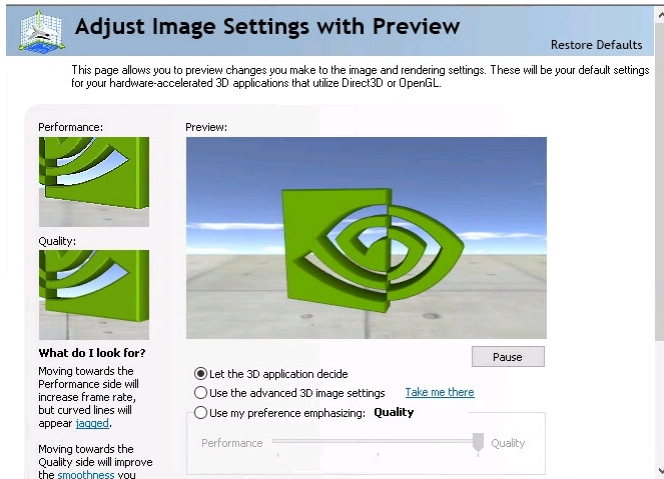
After the desktops are provisioned, verify vGPU deployment in the VMware Horizon environment.

Verify that the NVIDIA driver is running on the desktop

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

1. Right-click the desktop. In the menu, choose NVIDIA Control Panel to open the control panel.
2. In the control panel, select System Information to see the vGPU that the virtual machine is using, the vGPU’s capabilities, and the NVIDIA driver version that is loaded (Figure 65).

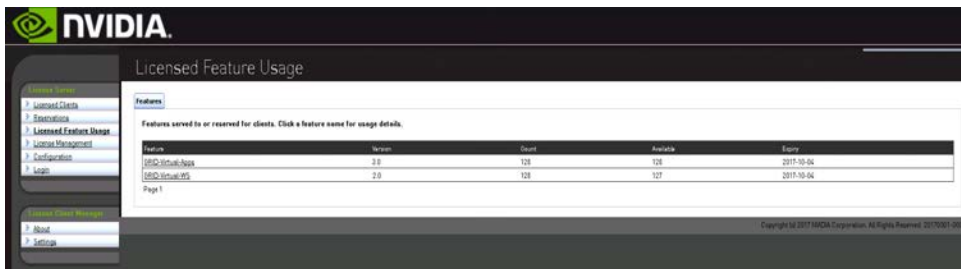
Figure 65. NVIDIA Control Panel



Verify NVIDIA license acquisition by desktops

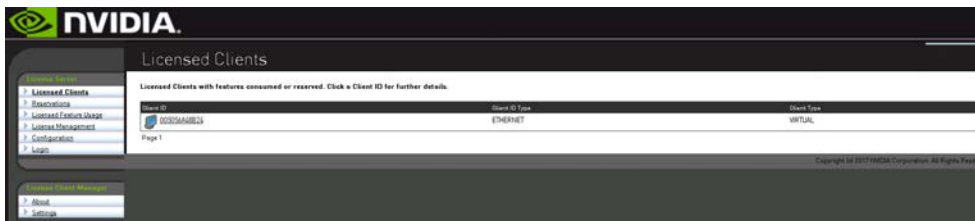
A license is obtained before the user logs on to the virtual machine after the virtual machine is fully booted (Figure 66).

Figure 66. NVIDIA License Server: Licensed Feature Usage



To view the details, select Licensed Clients in the left pane (Figure 67).

Figure 67. NVIDIA License Server: Licensed Clients



Verify the NVIDIA configuration on the host

To obtain a hostwide overview of the NVIDIA GPUs, enter the nvidia-smi or vSphere Web Client interface (Figures 68 through 70).

Figure 68. Six NVIDIA Tesla P4 GPU cards installed with 48 P4-1Q profiles attached to virtual machines provisioned through VMware Horizon, with 9 virtual machines per P4 GPU card

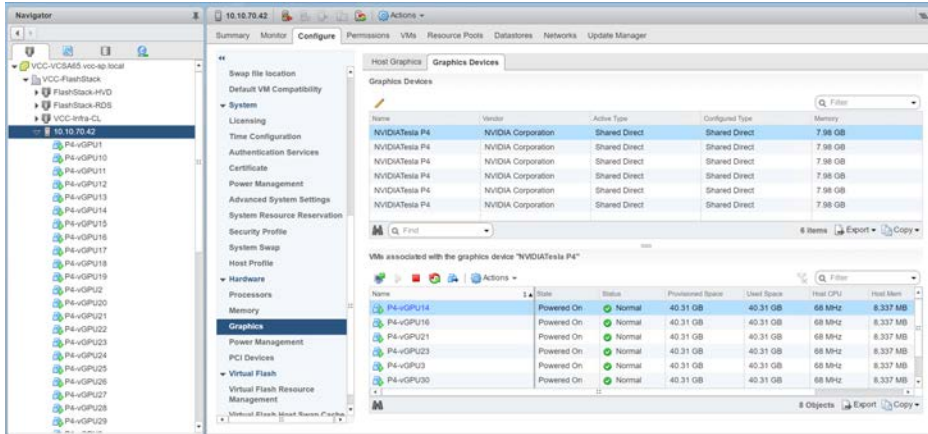


Figure 69. The nvidia-smi command output from the ESXi host with 2 NVIDIA P40 cards and 48 Microsoft Windows 10 desktops with the P40-1B vGPU profile

```

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

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

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

Figure 70. The nvidia-smi command Output from the host with 2 NVIDIA P6 cards and 32 Microsoft Windows 10 Desktops with the P6-1B vGPU profile

```

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

```


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

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.

Create the VMware Horizon 7 pool

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

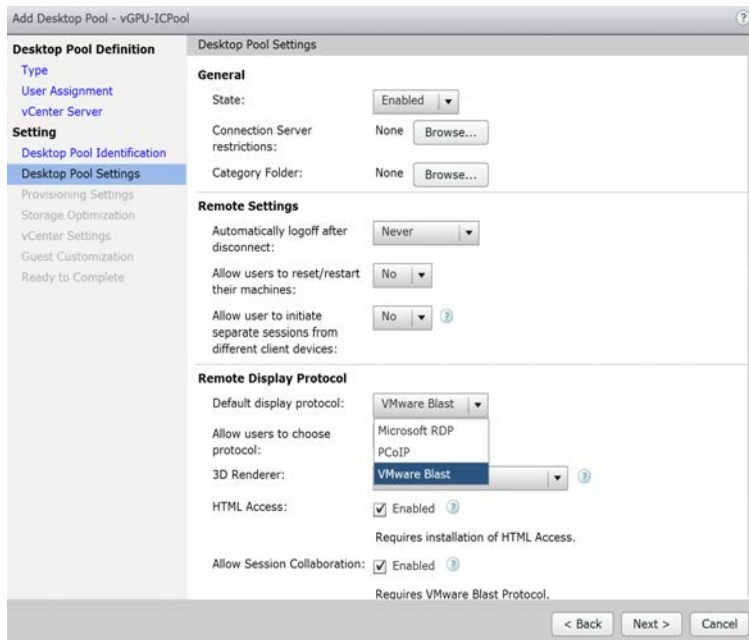
The desktop pool created as part of the solution verification is based on persistent desktops. The virtual machines are deployed as full clones from the master image template.

Pro tip: vDGA desktops require a GPU to be assigned to the virtual machine as a pass-through PCI device. To create a vDGA desktop template, the GPU must be assigned to the template virtual machine to allow driver installation. Be sure to remove the

GPU from the virtual machine's hardware configuration before converting the virtual machine to a template to avoid deployment problems. After the final desktop virtual machines are cloned, you must manually add the GPU to each virtual machine hardware configuration.

In creating the Horizon 7 desktop pool, for the Microsoft Remote Display Protocol (RDP), choose VMware Blast (Figure 71).

Figure 71. Configuring RDP



The screenshot shows the 'Add Desktop Pool - vGPU-ICPool' configuration window. The left sidebar lists navigation options: Desktop Pool Definition (Type, User Assignment, vCenter Server), Setting (Desktop Pool Identification, Desktop Pool Settings), Provisioning Settings, Storage Optimization, vCenter Settings, Guest Customization, and Ready to Complete. The main area is titled 'Desktop Pool Settings' and is divided into three sections: General, Remote Settings, and Remote Display Protocol. In the Remote Display Protocol section, the 'Default display protocol' is set to 'VMware Blast', and the '3D Renderer' is also set to 'VMware Blast'. The 'Allow users to choose protocol' dropdown is also set to 'VMware Blast'. Other settings include 'State: Enabled', 'Connection Server restrictions: None', 'Category Folder: None', 'Automatically logoff after disconnect: Never', 'Allow users to reset/restart their machines: No', 'Allow user to initiate separate sessions from different client devices: No', 'HTML Access: Enabled', and 'Allow Session Collaboration: Enabled'. The bottom of the window has '< Back', 'Next >', and 'Cancel' buttons.

Select the option for the 3D renderer based on your deployment scenario: vDGA, vGPU, or vSGA.

Also select the amount of vRAM to be configured for hardware and software rendering. An option is now available for an NVIDIA GRID vGPU-based 3D renderer (Figures 72 and 73).

Figure 72. Configuring rendering settings

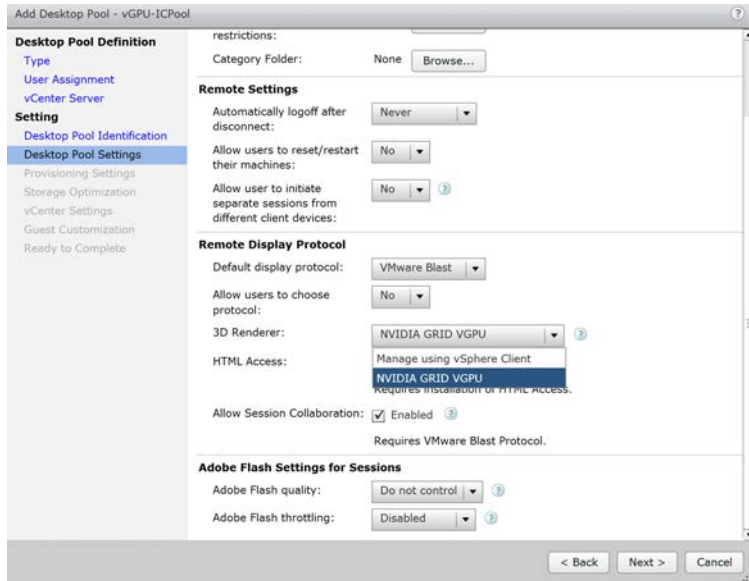
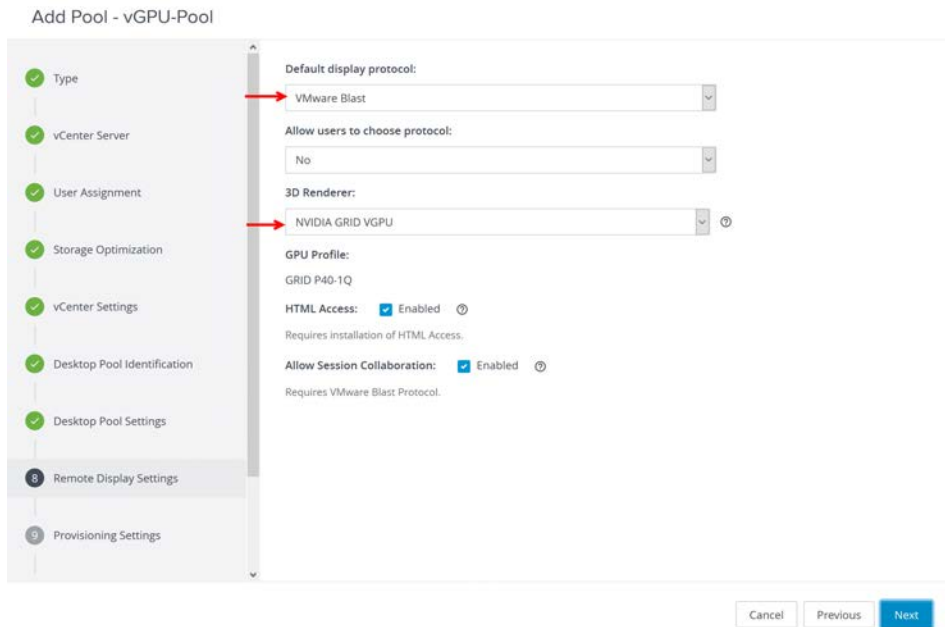


Figure 73. VMware Horizon console when creating an Instant Clone desktop pool with an NVIDIA vGPU profile attached to a Microsoft Windows 10 master image



Use the VMware vSphere 6.7 Performance tab to monitor GPU use

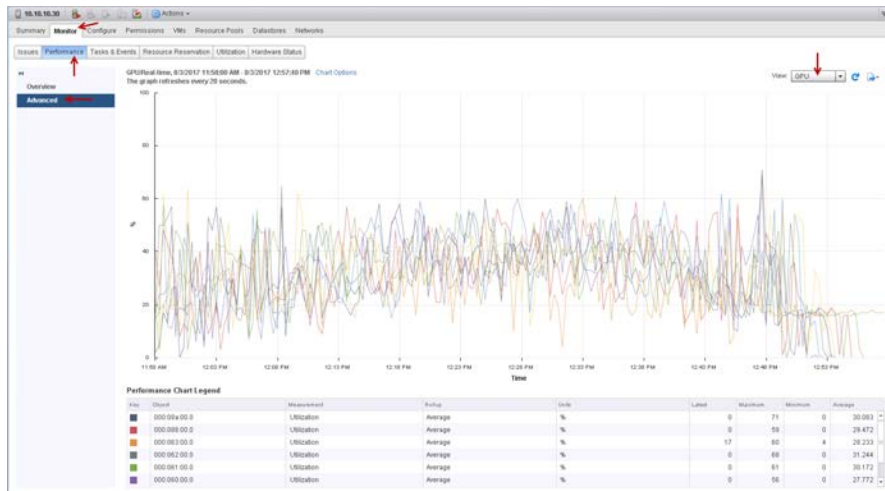
New with vSphere 6.5 and later, you can monitor NVIDIA GPU use through the vSphere Web Client.

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 74 shows an example of GPU use when a graphics workload was running on a server with six NVidia Tesla P4 card configured with the vGPU profile P4-1Q on 48 Windows 10 64-bit dedicated assigned Full-clone desktop virtual machines.

Figure 74. vSphere 6.7 GPU performance monitoring example



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) 9ex, 10, or 11.

Use the OpenGL Software Accelerator

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

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

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

Try the OpenGL Software Accelerator in the following cases:

- If the performance of OpenGL applications running in virtual machines is a concern, try using the OpenGL accelerator. For some applications, the accelerator outperforms the Microsoft OpenGL software rasterizer that is included with Windows because the OpenGL accelerator uses SSE4.1 and AVX. The OpenGL accelerator also supports applications using OpenGL versions up to Version 2.1.
- For applications running on a workstation, first try the default version of OpenGL support provided by the workstation's graphics adapter. If the graphics card is the latest version, in most cases it will deliver the best performance. If the graphics card is an earlier version or does not deliver satisfactory performance, then try the OpenGL Software Accelerator.

- 3D OpenGL applications that are not adequately delivered using CPU-based software rasterization may benefit from OpenGL GPU hardware acceleration. This feature can be used on bare-metal devices and virtual machines.

Conclusion

The combination of Cisco UCS Manager; Cisco UCS C220 M5, Cisco UCS C240 M5, and C480 M5 Rack Servers; Cisco UCS B200 M5 Blade Servers; and NVIDIA Tesla cards running on VMware vSphere 6.5 and Horizon 7 provides a high-performance platform for virtualizing graphics-intensive applications.

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

For more information

- Cisco UCS C-Series Rack Servers and B-Series Blade Servers: <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)