

Configure Cisco UCS Rack and Blade Servers with NVIDIA GRID 2.0 for VMware Horizon 7 on VMware vSphere 6.0

What You Will Learn

Using the increased processing power of today's Cisco UCS® B-Series Blade Servers and C-Series Rack Servers, applications with demanding graphics requirements are now being virtualized. To enhance the capability to deliver these high-performance and graphics-intensive applications, Cisco offers support for the NVIDIA GRID M6, M60, and M10 cards in the Cisco Unified Computing System™ (Cisco UCS) portfolio of PCI Express (PCIe) or mezzanine form-factor cards for the Cisco UCS B-Series Blade Servers and C-Series Rack Servers.

With the addition of the new graphics processing capabilities, the engineering, design, imaging, and marketing departments of organizations can now experience the benefits that desktop virtualization brings to the applications they use. Users of Microsoft Windows 10 and Office 2016 or later versions can benefit from the new NVIDIA M10 high-density graphics card, deployable on Cisco UCS C240 M4 Rack Servers, currently in standalone mode only.

Note: Support for Cisco UCS managed configuration of NVIDIA M10 cards will be added in Cisco UCS Manager Release 3.1(3a) in the near future.

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

The PCIe graphics cards in the Cisco UCS C-Series offer these benefits:

- Support for full-length, full-power NVIDIA GRID cards in a 2-rack-unit (2RU) form factor
- 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 C240 M4 Rack Servers with two NVIDIA GRID cards than with the 2-slot, 2.5-inch equivalent rack unit: the HP ProLiant WS460c Gen9 Graphics Server Blade with the GRID card in a second slot

The modular LAN-on-motherboard (mLOM) form-factor NVIDIA graphics card in the Cisco UCS B-Series 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. 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.

The purpose of this document is to help our partners and customers integrate NVIDIA GRID 2.0 graphics processing cards, Cisco UCS B200 M4 servers, Cisco UCS C240 M4 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.

Our objective here is to provide the reader with specific methods for integrating Cisco UCS servers with NVIDIA GRID M6, M60, 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 Graphic 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 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 PCs with an onboard graphics processor.

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

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

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

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

vGPU Profiles

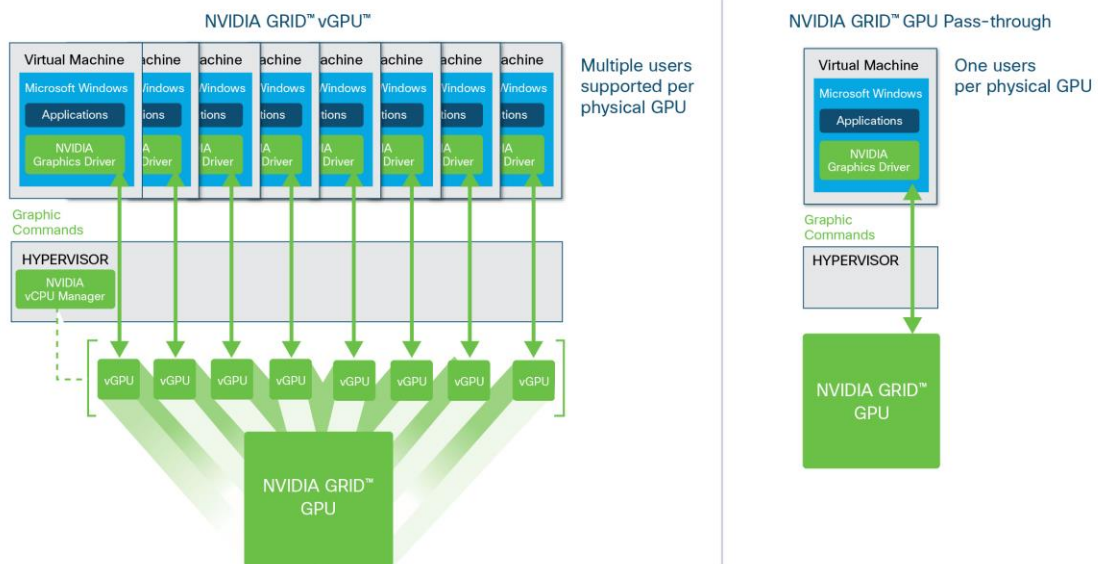
In any given enterprise, the needs of individual users vary widely. One of the main benefits of the GRID vGPU is the flexibility to use various vGPU profiles designed to serve the needs of different classes of end users.

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

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

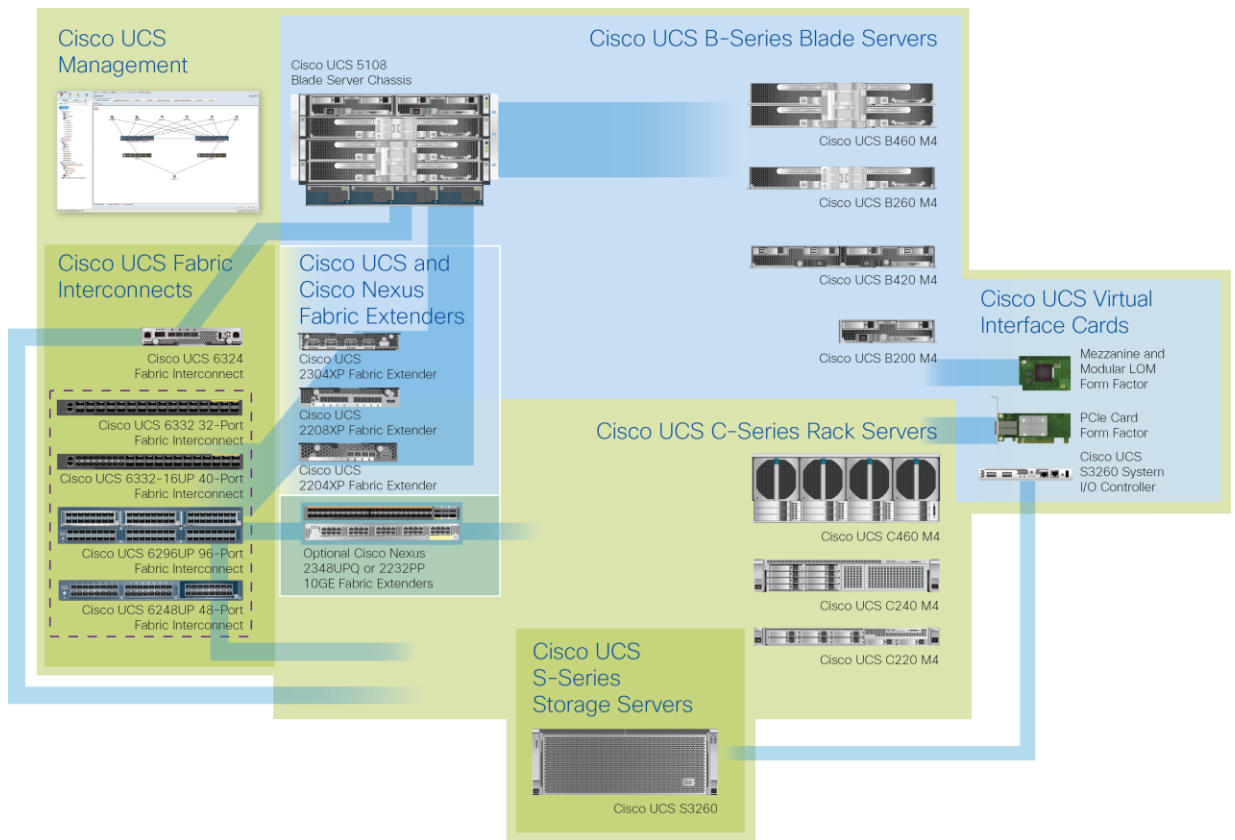
Figure 1. NVIDIA GRID vGPU GPU System Architecture



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 10 Gigabit Ethernet unified network fabric with enterprise-class, x86-architecture servers. It is an integrated, scalable, multi-chassis 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, 10-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.

Cisco UCS Mini

Cisco UCS Mini is incorporated into this solution to manage the Cisco UCS C240 M4 server and the NVIDIA GRID cards. In addition, Cisco UCS Mini hosts the virtual infrastructure components such as the domain controllers and desktop broker using Cisco UCS B200 M4 Blade Servers. Cisco UCS Mini is an optional part of this reference architecture. Another choice for managing the Cisco UCS and NVIDIA equipment is the Cisco UCS 6200 Series Fabric Interconnects.

Cisco UCS Mini is designed for customers who need fewer servers but still want the robust management capabilities provided by Cisco UCS Manager. This solution delivers servers, storage, and 10 Gigabit networking in an easy-to-deploy, compact form factor (Figure 3).

Figure 3. Cisco UCS Mini



Cisco UCS Mini consists of the following components:

- **Cisco UCS 5108 Blade Server Chassis:** A chassis can accommodate up to eight half-width Cisco UCS B200 M4 Blade Servers.
- **Cisco UCS B200 M4 Blade Server:** Delivering performance, versatility, and density without compromise, the Cisco UCS B200 M4 addresses a broad set of workloads.
- **Cisco UCS 6324 Fabric Interconnect:** The Cisco UCS 6324 provides the same unified server and networking capabilities as the top-of-rack Cisco UCS 6200 Series Fabric Interconnects embedded in the Cisco UCS 5108 Blade Server Chassis.
- **Cisco UCS Manager:** Cisco UCS Manager provides unified, embedded management of all software and hardware components in a Cisco UCS Mini solution.

Cisco UCS Fabric Interconnect

Cisco UCS 6300 Series Fabric Interconnects provide the management and communication backbone for the Cisco UCS B-Series Blade Servers, 5100 Series Blade Server Chassis, and C-Series Rack Servers. In addition, the firmware for the NVIDIA GRID cards can be managed using both the 6300 and 6200 Series Fabric Interconnects, an exclusive feature of the Cisco UCS portfolio.

The chassis, blades, and rack-mount servers that are attached to the interconnects are part of a single highly available management domain. By supporting unified fabric, the fabric interconnects provide the flexibility to support LAN and storage connectivity for all blades in the domain at configuration time.

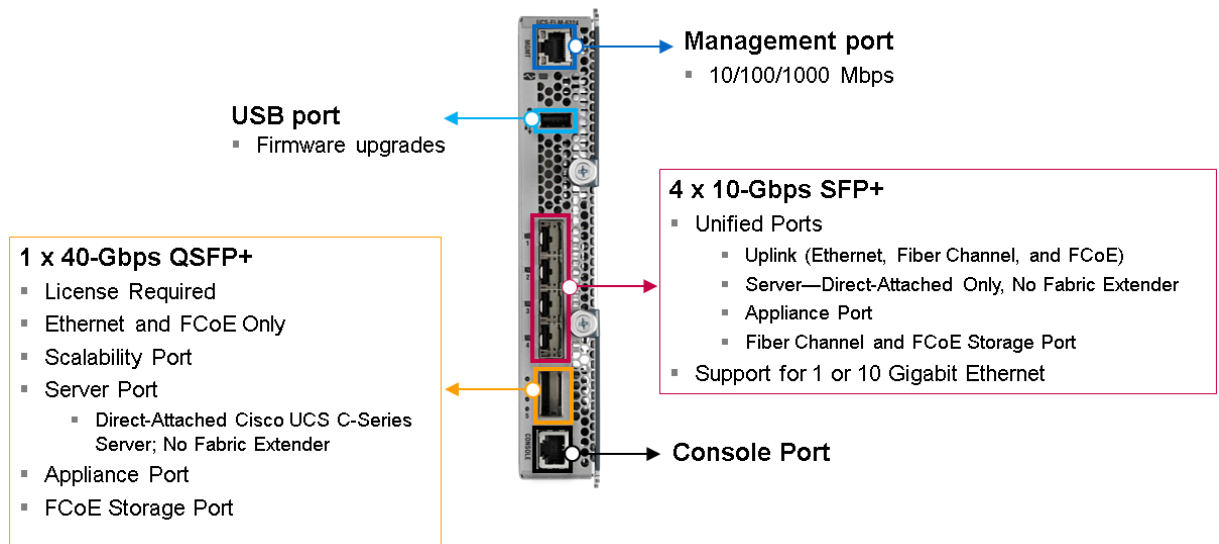
Typically deployed in redundant pairs, the 6300 Series Fabric Interconnects deliver uniform access to both networks and storage to help create a fully virtualized environment.

Cisco UCS 6324 Fabric Interconnect

The Cisco UCS 6324 Fabric Interconnect (Figure 4) offers several major features and benefits that reduce TCO, including:

- Bandwidth of up to 500 Gbps
- Ports capable of line-rate, low-latency, lossless 1 and 10 Gigabit Ethernet; FCoE; and 8-, 4-, and 2-Gbps Fibre Channel
- Centralized management with Cisco UCS Manager software
- Cisco UCS 5108 Blade Server Chassis capabilities for cooling and serviceability
- Quad Enhanced Small Form-Factor Pluggable (QSFP+) port for rack-mount server connectivity

Figure 4. Cisco UCS 6324 Fabric Interconnect



Cisco UCS C-Series Rack Servers

Cisco UCS C-Series Rack Servers keep pace with Intel® Xeon® processor innovation by offering the latest processors with an increase in processor frequency and improved security and availability features. With the increased performance provided by the Intel Xeon processor E5-2600 and E5-2600 v2 product families, 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, these servers enable organizations to deploy systems incrementally—using as many or as few servers as needed—on a schedule that best meets the organization’s timing and budget. C-Series servers offer investment protection through the capability to deploy them either as standalone servers or as part of Cisco UCS.

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

Cisco UCS C240 M4 Rack Server

The Cisco UCS C240 M4 Rack Server (Figures 5 and 6 and Table 1) is designed for both performance and expandability over a wide range of storage-intensive infrastructure workloads, from big data to collaboration.

The enterprise-class Cisco UCS C240 M4 server extends the capabilities of the Cisco UCS portfolio in a 2RU form factor with the addition of the Intel Xeon processor E5-2600 v4 and v3 product family, which delivers a superb combination of performance, flexibility, and efficiency

The enterprise-class Cisco UCS C240 M4 server extends the capabilities of the Cisco UCS portfolio in a 2RU form factor. Based on the Intel Xeon processor E5-2600 v4 and v3 series, it delivers an outstanding combination of performance, flexibility, and efficiency. In addition, the C240 M4 offers outstanding levels of internal memory and storage expandability with exceptional performance. It delivers:

- Up to 24 DDR4 DIMMs at speeds up to 2400 MHz for improved performance and lower power consumption
- Up to 6 PCIe 3.0 slots (4 full-height, full-length)
- Up to 24 small-form-factor (SFF) drives or 12 large-form-factor (LFF) drives, plus two (optional) internal SATA boot drives
- Support for 12-Gbps SAS drives
- An mLOM slot for installing a next-generation Cisco virtual interface card (VIC) or third-party network interface card (NIC) without consuming a PCIe slot
- Two 1 Gigabit Ethernet embedded LOM ports
- Support for up to 2 double-wide NVIDIA graphics processing units (GPUs), providing a graphics-rich experience to more virtual users
- Excellent reliability, availability, and serviceability (RAS) features with tool-free CPU insertion, easy-to-use latching lid, hot-swappable and hot-pluggable components, and redundant Cisco Flexible Flash (FlexFlash) Secure Digital (SD) cards.

The C240 M4 also increases performance and customer choice over many types of storage-intensive applications such as:

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

The C240 M4 can be deployed as a standalone server or as part of Cisco UCS. Cisco UCS unifies computing, networking, management, virtualization, and storage access into a single integrated architecture that enables end-to-end server visibility, management, and control in both bare-metal and virtualized environments. Within a Cisco UCS deployment, the C240 M4 takes advantage of Cisco's standards-based unified computing innovations, which significantly reduce customers' TCO and increase business agility.

- For more information about the Cisco UCS C240 M4 Rack Server, see <http://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c240-m4-rack-server/index.html>.

Figure 5. Cisco UCS C240 M4 Rack Server Front View

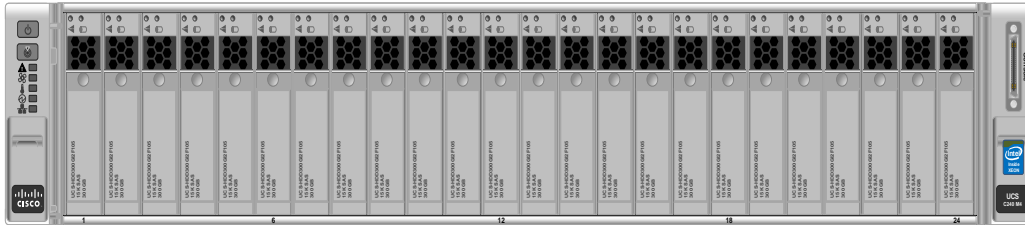


Figure 6. Cisco UCS C240 M4 Rack Server Rear View

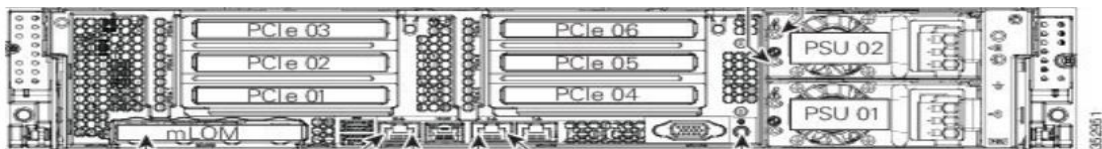


Table 1. Cisco UCS C240 M4 PCIe Slots

PCIe Slot	Length	Lane
1	¾	x8
2	Full	x16
3	Full	x8
4	¾	x8
5	Full	x16
6	Full	x8

Cisco UCS VIC 1227

The Cisco UCS VIC 1227 (Figure 7) is a dual-port Enhanced Small Form-Factor Pluggable (SFP+) 10-Gbps Ethernet and FCoE-capable PCIe mLOM adapter installed in the Cisco UCS C-series Rack Servers. The mLOM slot can be used to install a Cisco VIC without consuming a PCIe slot, which provides greater I/O expandability. It incorporates next-generation converged network adapter (CNA) technology from Cisco, providing investment protection for future feature releases. The card enables a policy-based, stateless, agile server infrastructure that can present up to 256 PCIe standards-compliant interfaces to the host that can be dynamically configured as either network interface cards (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 <http://www.cisco.com/c/en/us/products/interfaces-modules/ucs-virtual-interface-card-1227/index.html>.

Figure 7. Cisco UCS VIC 1227 CNA



Cisco UCS B200 M4 Blade Server

The enterprise-class Cisco UCS B200 M4 Blade Server (Figure 8) extends the capabilities of the Cisco UCS portfolio in a half-width blade form factor. The B200 M4 uses the power of the latest Intel Xeon processor E5-2600 v4 and v3 series CPUs with up to 768 GB of RAM (using 32-GB DIMMs), two solid-state disks (SSDs) or hard-disk drives (HDDs), and throughput of up to 80 Gbps. The B200 M4 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) for up to 768 GB of total memory capacity (Cisco UCS B200 M4 configured with two CPUs using 32-GB DIMMs). It supports one connector for the Cisco UCS VIC 1340 or 1240 adapter, which provides Ethernet and FCoE. A second mezzanine card slot also is available, which can be used for the NVIDIA M6 graphics cards.

- For more information, see <http://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-b200-m4-blade-server/index.html>.

Figure 8. Cisco UCS B200 M4 Blade Server Front View



Cisco UCS VIC 1340

The Cisco UCS VIC 1340 (Figure 9) is a 2-port 40-Gbps Ethernet or dual 4 x 10-Gbps Ethernet, FCoE-capable mLOM designed exclusively for the M4 generation of Cisco UCS B-Series Blade Servers. When used in combination with an optional port expander, the VIC 1340 is enabled for two ports of 40-Gbps Ethernet. The VIC 1340 enables a policy-based, stateless, agile server infrastructure that can present more than 256 PCIe standards-compliant interfaces to the host that can be dynamically configured as either NICs or HBAs. In addition, the VIC 1340 supports Cisco Virtual Machine Fabric Extender (VM-FEX) technology, which extends the Cisco UCS fabric interconnect ports to virtual machines, simplifying server virtualization deployment and management.

For more information, see <http://www.cisco.com/c/en/us/products/interfaces-modules/ucs-virtual-interface-card-1340/index.html>.

Figure 9. Cisco UCS VIC 1340



NVIDIA GRID Cards

For desktop virtualization applications, the NVIDIA Tesla M6, M10, and M60 cards are an optimal choice for high-performance graphics (Table 2).

Table 2. Technical Specifications for NVIDIA GRID Cards



Technical Specifications for NVIDIA GRID Cards			
Number of GPUs	Single high-end Maxwell	Quad midlevel Maxwell	Dual high-end Maxwell
NVIDIA Compute Unified Device Architecture (CUDA) Cores	1536	2560 (640 per GPU)	4096 (2048 per GPU)
Memory Size	8-GB GDDR5 (8 GB per GPU)	32-GB GDDR5	16-GB GDDR5 (8 GB per GPU)
Maximum Number of vGPU Instances	16	64	32
Power	100 watts (W; 75W optimal)	225W	240W or 300W (225W optimal)
Form Factor	MXM (blade servers) P	PCIe 3.0 dual slot (rack servers)	PCIe 3.0 dual slot (rack servers)
Cooling Solution	Bare board	Passive	Active and passive
H.264 1080p30 Streams 2	18	28	36
Maximum Number of Users per Board	16	64 (16 per GPU)	32 (16 per GPU)
Virtualization Use Case	Blade optimized	User-density optimized	Performance optimized

NVIDIA GRID 2.0 Technology

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

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

NVIDIA GRID 2.0 GPU

The NVIDIA GRID solution runs on top of award-winning, [NVIDIA Maxwell-powered GPUs](#). These GPUs come in two server form factors: the NVIDIA Tesla [M6](#) for blade servers and converged infrastructure, and the NVIDIA Tesla [M10](#) and [M60](#) for rack and tower servers.

NVIDIA GRID 2.0 License Requirements

GRID 2.0 requires concurrent user licenses and an on-premises NVIDIA license server to manage the licenses. When the guest OS boots up, it contacts the NVIDIA license server and consumes one concurrent license. When the guest OS shuts down, the license is returned to the pool.

GRID 2.0 also requires the purchase of a 1:1 ratio of concurrent licenses to NVIDIA Support, Update, and Maintenance Subscription (SUMS) instances.

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

- Virtual workstation
- Virtual PC
- Virtual applications

For complete details about GRID 2.0 license requirements, see <https://images.nvidia.com/content/grid/pdf/GRID-Licensing-Guide.pdf>.

VMware vSphere 6.0

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 6.0 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 6.0 platform includes these features:

- Computing
 - **Increased scalability:** vSphere 6.0 supports larger maximum configuration sizes. Virtual machines support up to 128 virtual CPUs (vCPUs) and 4 TB of virtual RAM (vRAM). Hosts support up to 480 CPUs and 12 TB of RAM, 1024 virtual machines per host, and 64 nodes per cluster.
 - **Expanded support:** Get expanded support for the latest x86 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.
 - **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 now enables data centers physically located in New York and London to migrate live workloads between one another.
 - **Replication-assisted vMotion:** Customers with active-active replication set up between two sites can perform more efficient vMotion migration, resulting in huge savings in time and resources, with up to 95 percent more efficient migration depending on the amount of data moved.
 - **Fault tolerance (up to 4 vCPUs):** Get expanded support for software-based fault tolerance for workloads with up to four vCPUs.
- Management
 - **Content library:** This centralized repository provides simple and effective management for content, including virtual machine templates, ISO images, and scripts. With 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.

Graphics Acceleration in VMware Horizon 7

New 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

others. 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 VMware Virtual GPU Solution.

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 are 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 only copy and paste of 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 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.

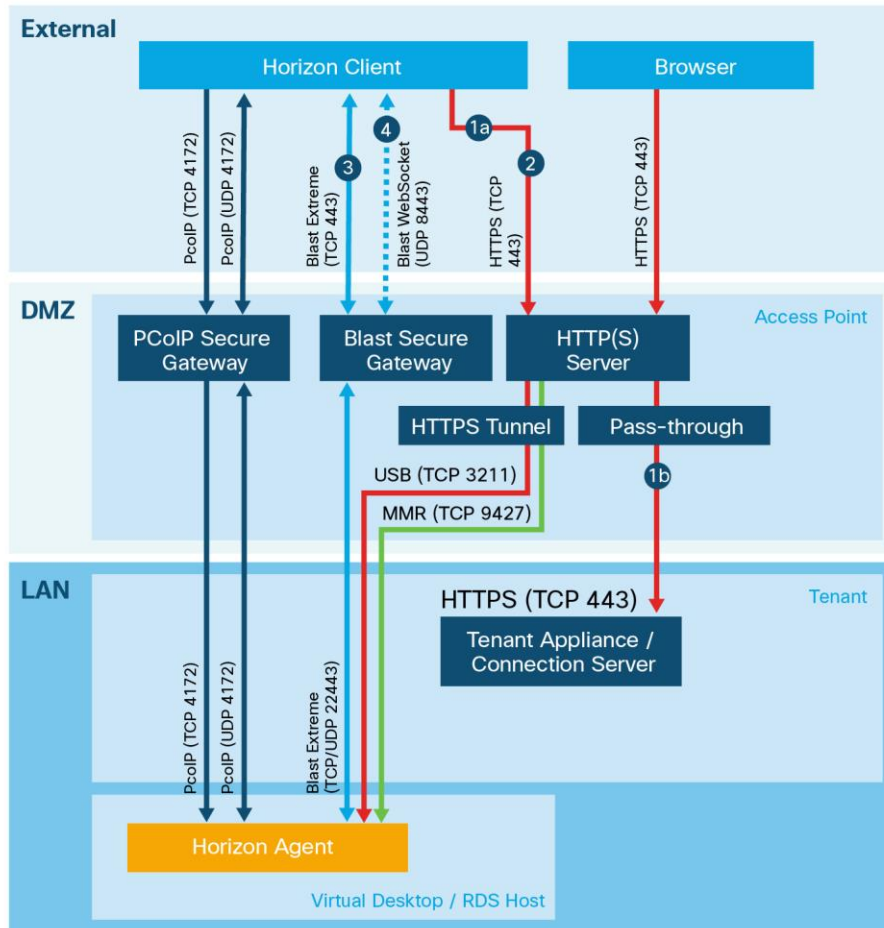
As shown in Figure 10, Blast Extreme provides an enhanced remote session experience introduced with Horizon for Linux desktops, Horizon 7, and Horizon Desktop as a Service (DaaS). In this case, the connection flow from the Horizon Client differs from the flow for PC over IP (PCoIP).

-
- 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 through 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 10. 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. Selecting 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](#).

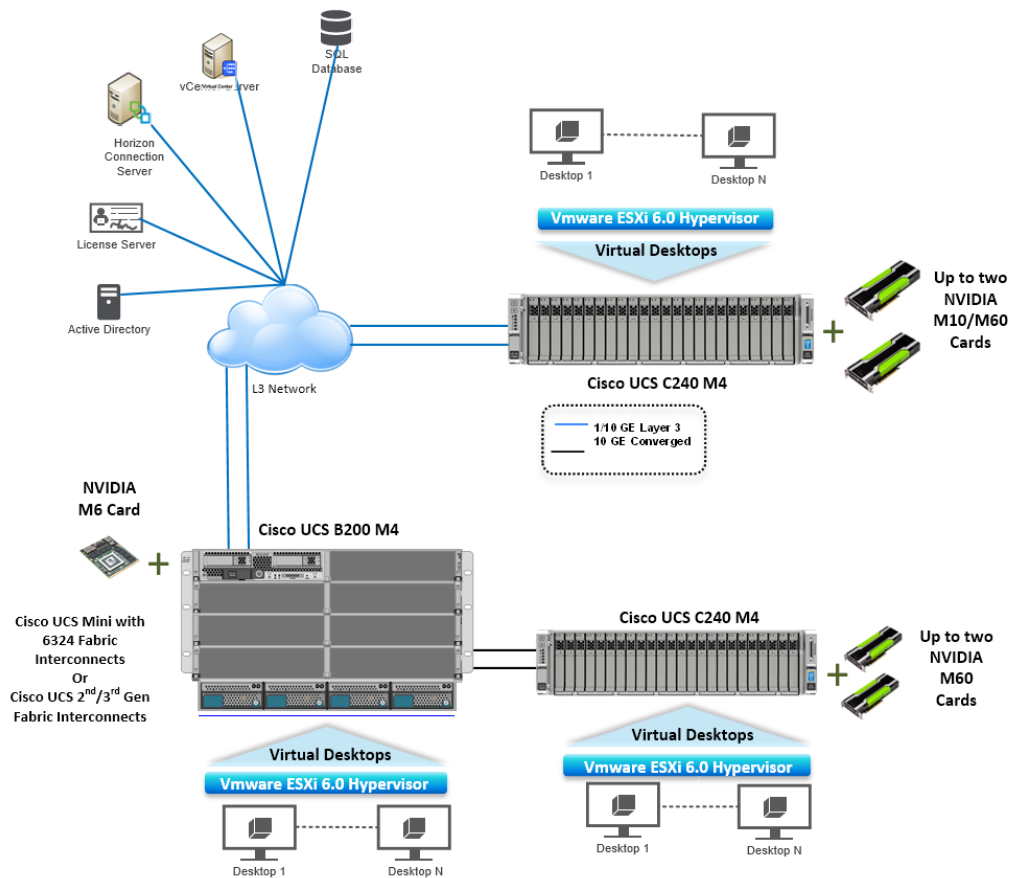
Note: If you configure a pool to use Blast Extreme and do not allow users to choose a protocol, View Connection Server automatically allows PCoIP connections from PCoIP zero clients and older (earlier than Release 4.0) Horizon Clients.

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 11 provides an overview of the solution configuration.

Figure 11. Reference Architecture



The hardware components in the solution are:

- Cisco UCS C240 M4 Rack Server (two Intel Xeon processor E5-2690 v4 CPUs at 2.60 GHz) with 512 GB of memory (32 GB x 16 DIMMs at 2400 MHz)
- Cisco UCS B200 M4 Blade Server (two Intel Xeon E5-2690 v4 CPUs at 2.60 GHz) with 512 GB of memory (32 GB x 16 DIMMs at 2400 MHz)
- Cisco UCS VIC 1227 mLOM (Cisco UCS C240 M4)
- Cisco UCS VIC 1340 mLOM (Cisco UCS B200 M4)
- Two Cisco UCS 6324 fabric interconnects in Cisco UCS Mini or UCS second- or third-generation fabric interconnects
- Twelve 600-GB SAS disks at 10,000 rpm
- NVIDIA Tesla M10, M6, and M60 cards
- Two Cisco Nexus® 9372 Switches (optional access switches)

The software components of the solution are:

- Cisco UCS Firmware Release 3.1(2e)
- VMware ESXi 6.0 (4192238) for VDI hosts
- VMware Horizon 7
- Microsoft Windows 10 64-bit
- Microsoft Server 2012 R2
- NVIDIA GRID 2.0 software and licenses:
 - NVIDIA-vGPU-VMware_ESXi_6.0_Host_Driver_367.64-1OEM.600.0.0.2494585
 - 369.71_grid_win10_server2016_64bit_international

Configure Cisco UCS

This section describes the Cisco UCS configuration.

Install NVIDIA Tesla GPU Card on Cisco UCS C240 M4

Install the M10 or M60 GPU card on the Cisco UCS C240 M4 server. Table 3 lists the minimum firmware required for the GPU cards.

Table 3. Minimum Server Firmware Versions Required for GPU Cards

Cisco Integrated Management Controller (IMC)	BIOS Minimum Version
NVIDIA Tesla M10	Release 2.0(13c)
NVIDIA Tesla M60	Release 2.0(9)

Note: The NVIDIA Tesla M10 currently is supported on the standalone server only.

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. Table 4 lists rules for populating the Cisco UCS C240 M4 with NVIDIA GPUs. Figure 12 shows a one-GPU installation, and Figure 13 shows a two-GPU installation.

Table 4. NVIDIA GPU Population Rules for Cisco UCS C240 M4 Rack Server

Single GPU	Dual GPU
Riser 1A, slot 2 or Riser 2, slot 5	Riser 1A, slot 2 and Riser 2, slot 5

Figure 12. One-GPU Scenario

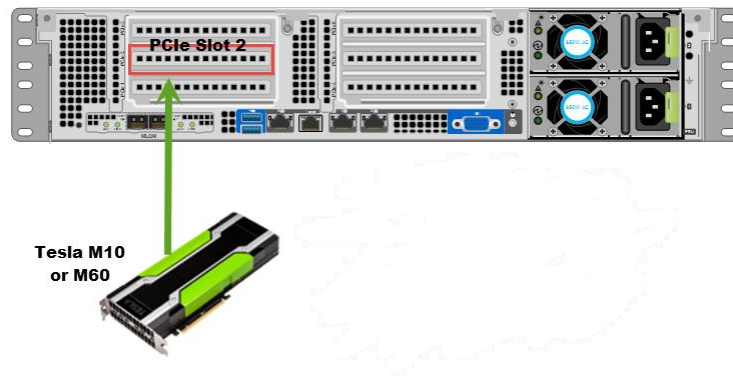
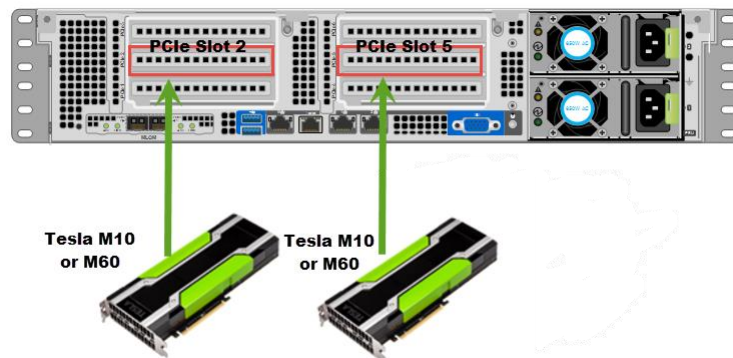


Figure 13. Two-GPU Scenario



For more information, see

http://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/c/hw/C240M4/install/C240M4/gpu.html.

Install NVIDIA Tesla GPU Card on Cisco UCS B200 M4

Install the M6 GPU card on the Cisco UCS B200 M4 server. Table 5 lists the minimum firmware required for the GPU card. Figure 14 shows the card in the server.

Table 5. Minimum Server Firmware Versions Required for GPU Card

Cisco Integrated Management Controller (IMC)	BIOS Minimum Version
NVIDIA Tesla M6	Release 2.0(13c)

Before installing the NVIDIA M6 GPU, do the following:

- Remove any adapter card, such as a Cisco UCS VIC 1380 or 1280 or port extender card from mLOM slot 2. You cannot use any other card in slot 2 when the NVIDIA M6 GPU is installed.
- Upgrade your Cisco UCS system to a version of Cisco UCS Manager that supports this card. Refer to the latest version of the release notes for Cisco UCS software at the following URL for information about supported hardware: <http://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-manager/products-release-notes-list.html>.

Figure 14. Cisco UCS B200 M4 Blade Server

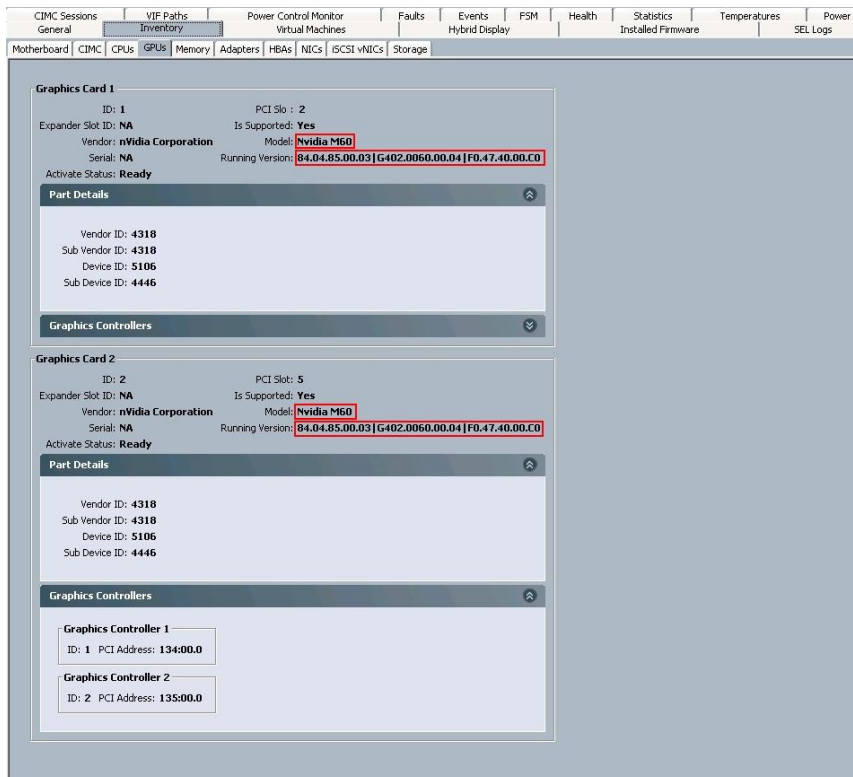


Configure the GPU Card

Follow these steps to configure the GPU card.

1. After the NVIDIA M60 GPU cards are physically installed and the Cisco UCS C240 M4 Rack Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 15, PCIe slots 2 and 5 are used with two GRID M60 cards.

Figure 15. NVIDIA GRID Cards Inventory Displayed in Cisco UCS Manager



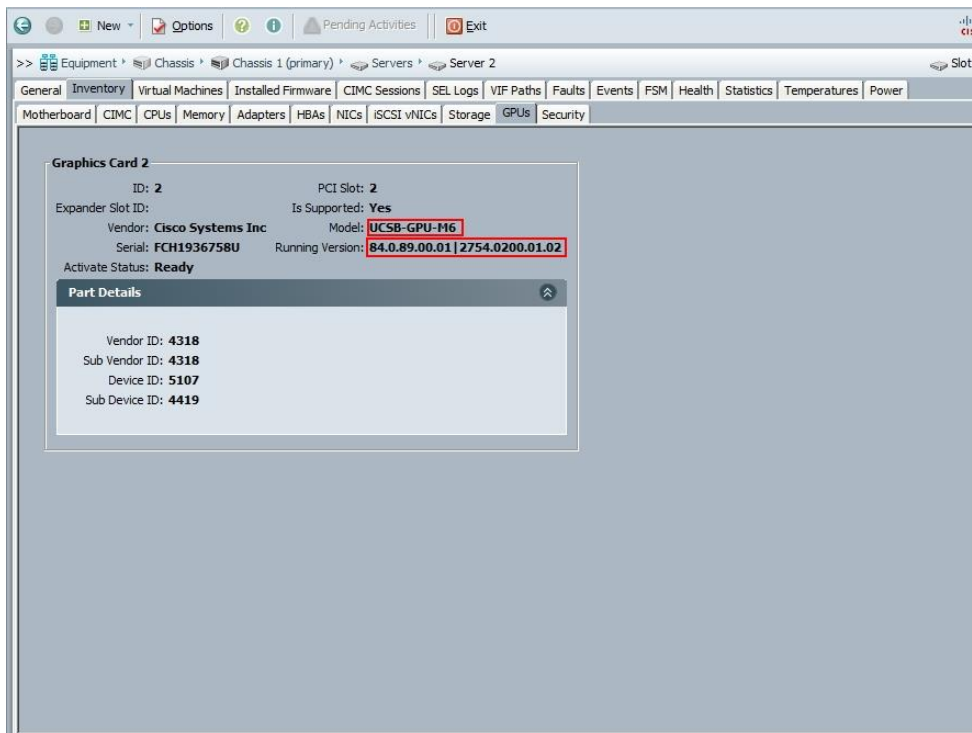
- After the NVIDIA M10 GPU cards are physically installed, log in to the standalone Cisco UCS C240 M4 Rack Server IMC and view the cards by choosing Inventory > PCI Adapters. As shown in Figure 16, PCIe slots 2 and 5 are used with two GRID M10 cards.

Figure 16. NVIDIA GRID Cards Inventory Displayed in Cisco Integrated Management Controller



- After the NVIDIA M6 GPU card is physically installed and the Cisco UCS B200 M4 Blade Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 17, PCIe slot 2 is used with the GRID M6 card.

Figure 17. NVIDIA GRID Cards Inventory Displayed in Cisco UCS Manager



You can use Cisco UCS Manager to perform firmware upgrades to the NVIDIA GPU cards in managed Cisco UCS C240 M4 servers.

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

Install the NVIDIA GRID License Server

This section summarizes the installation and configuration process for the GRID 2.0 License Server.




The NVIDIA GRID vGPU is a licensed feature on Tesla M6, M10, and M60. A software license is required to use the full vGPU features on a guest virtual machine. An NVIDIA license server with the appropriate licenses is required.

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

Three packages are required for VMware ESXi host setup, as shown in Figure 18:

- The GRID license server installer
- The NVIDIA GRID Manager software, which is installed on VMware vSphere ESXi; the NVIDIA drivers and software that are installed in Microsoft Windows are also in this folder
- The GPU Mode Switch utility, which changes the cards from the default Compute mode to Graphics mode

Figure 18. Software Required for NVIDIA GRID 2.0 Setup on the VMware ESXi Host

Name	Date modified	Type	Size
 NVIDIA-ls-windows-2015.12-0001.zip	1/30/2017 9:39 AM	WinRAR ZIP archive	140,643 KB
 NVIDIA-GRID-vSphere-6.0-367.64-369.71.zip	1/30/2017 9:30 AM	WinRAR ZIP archive	1,024,211 KB
 NVIDIA-gpumodeswitch-2016-04.zip	1/30/2017 9:30 AM	WinRAR ZIP archive	98,933 KB




Install the GRID 2.0 License Server

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

The GRID 2.0 license server requires Java Version 7 or later. Go to Java.com and install the latest version.

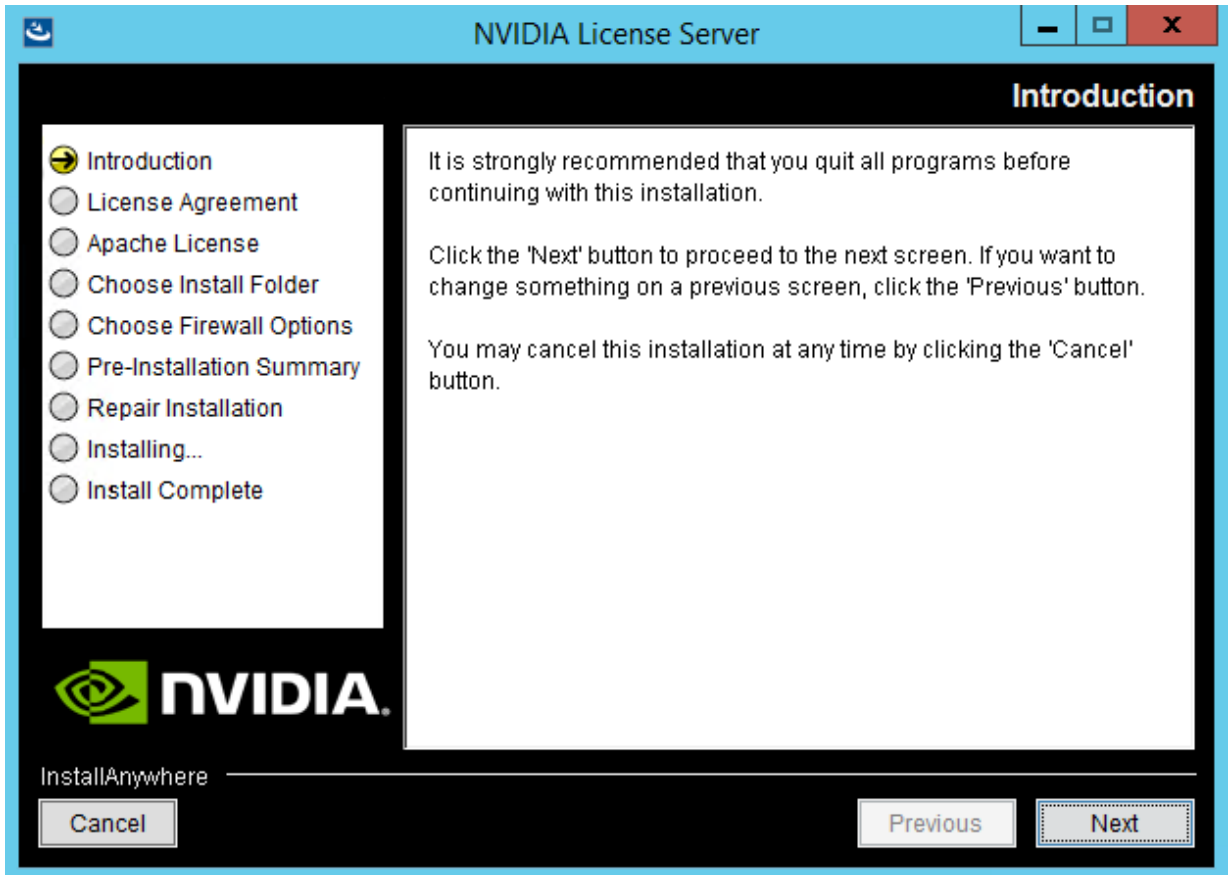
1. Extract and open the NVIDIA-ls-windows-2015.12-0001 folder. Run setup.exe (Figure 19).

Figure 19. Run setup.exe

Name	Date modified	Type	Size
 GRID License Server Release Notes.pdf	12/18/2015 11:15 AM	Adobe Acrobat Doc...	640 KB
 GRID License Server User Guide.pdf	12/18/2015 11:15 AM	Adobe Acrobat Doc...	2,511 KB
 setup.exe	12/18/2015 11:17 AM	Application	138,486 KB

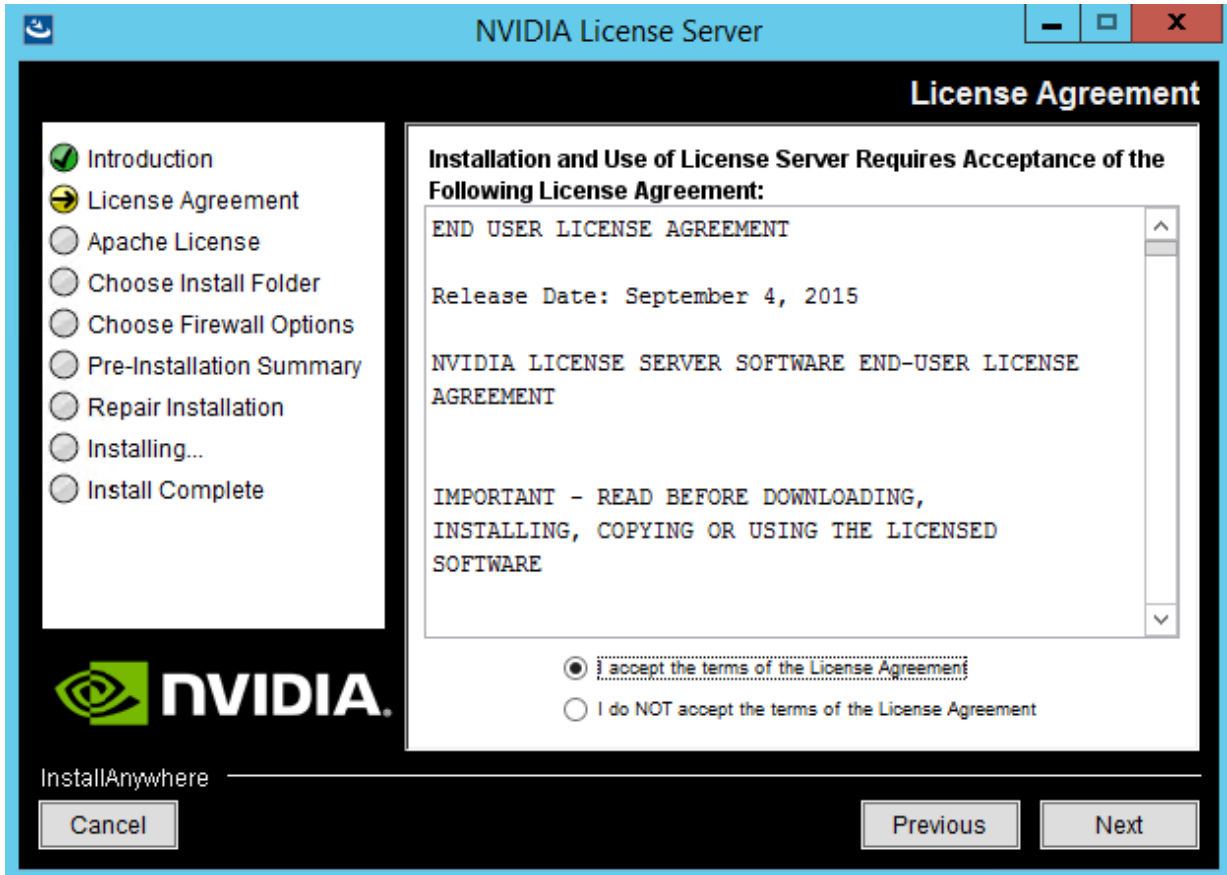
2. Click Next (Figure 20).

Figure 20. NVIDIA License Server



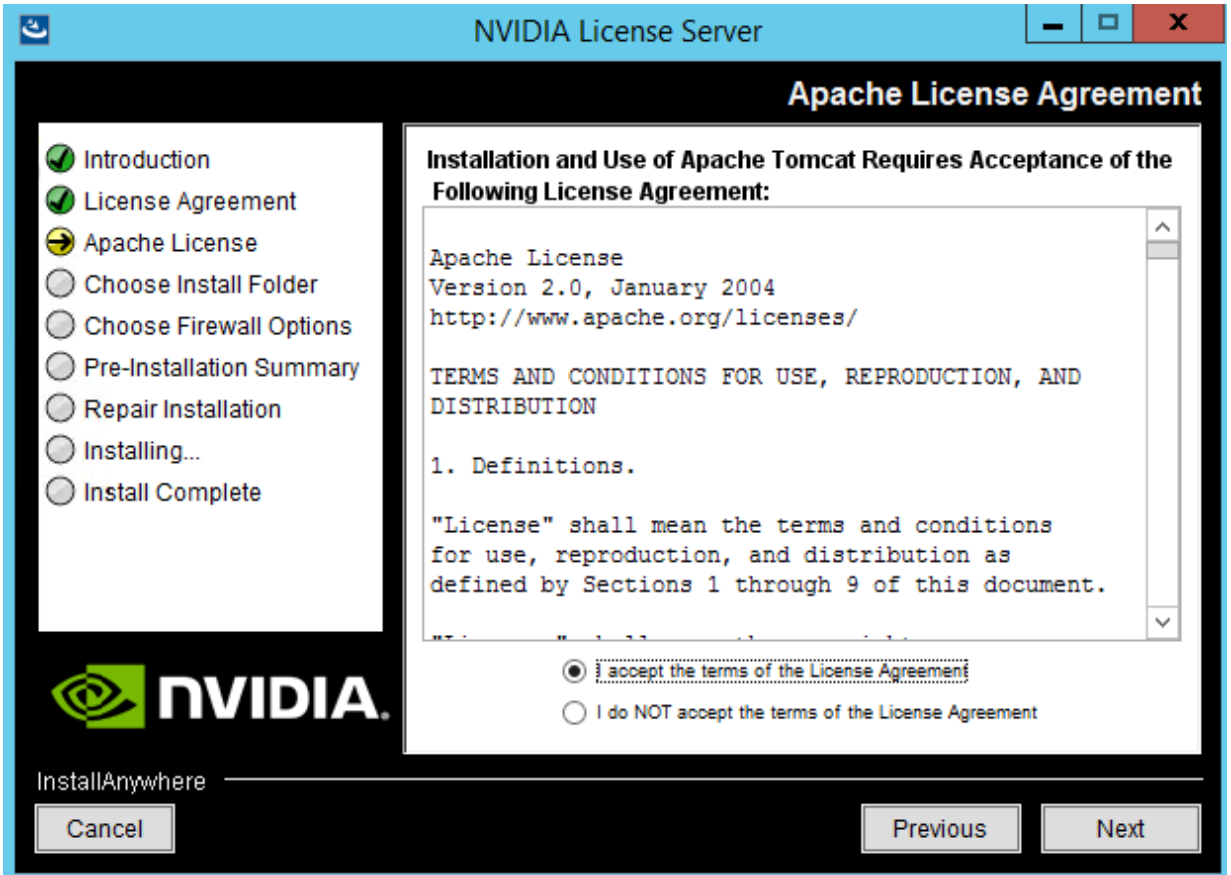
3. Accept the license agreement and click Next (Figure 21).

Figure 21. NVIDIA License Agreement



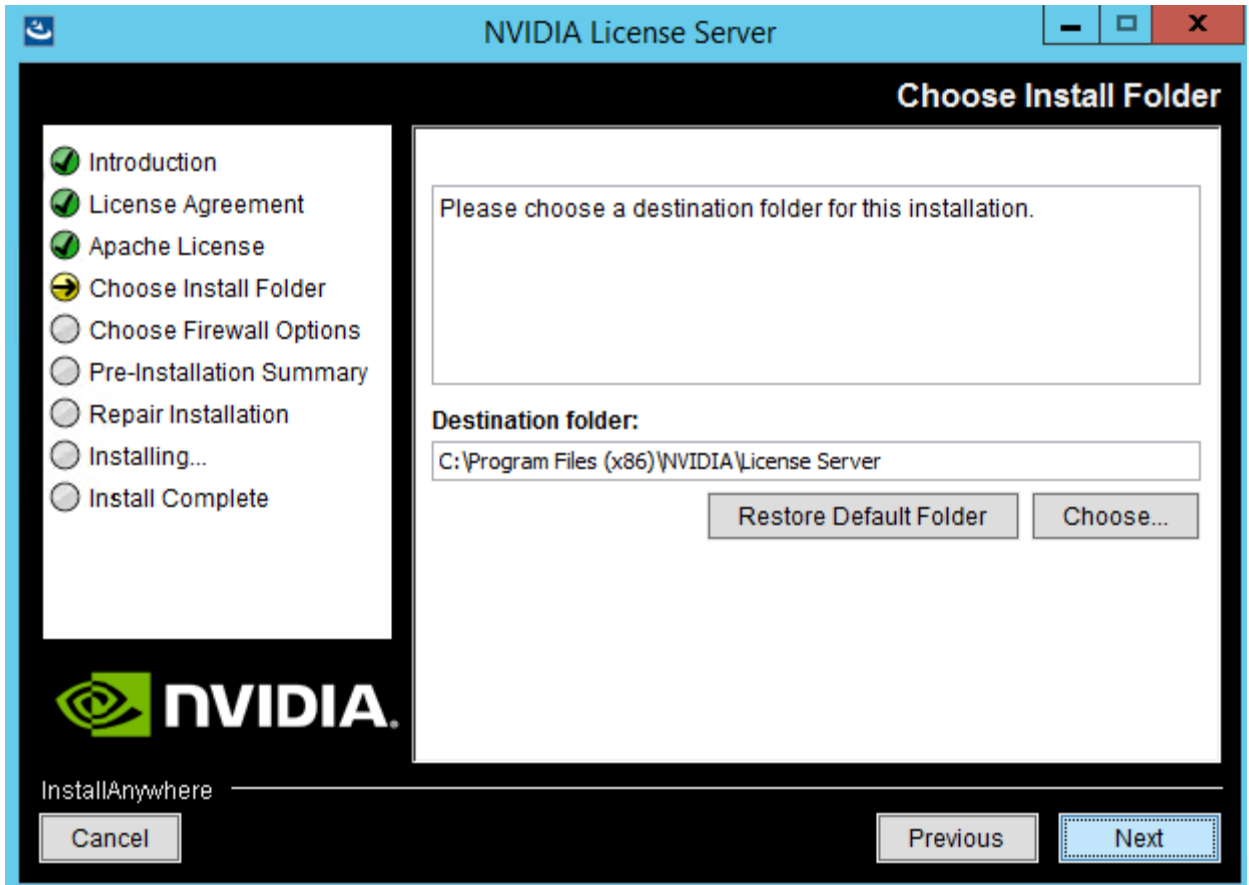
4. Accept the Apache license agreement and click Next (Figure 22).

Figure 22. NVIDIA License Agreement



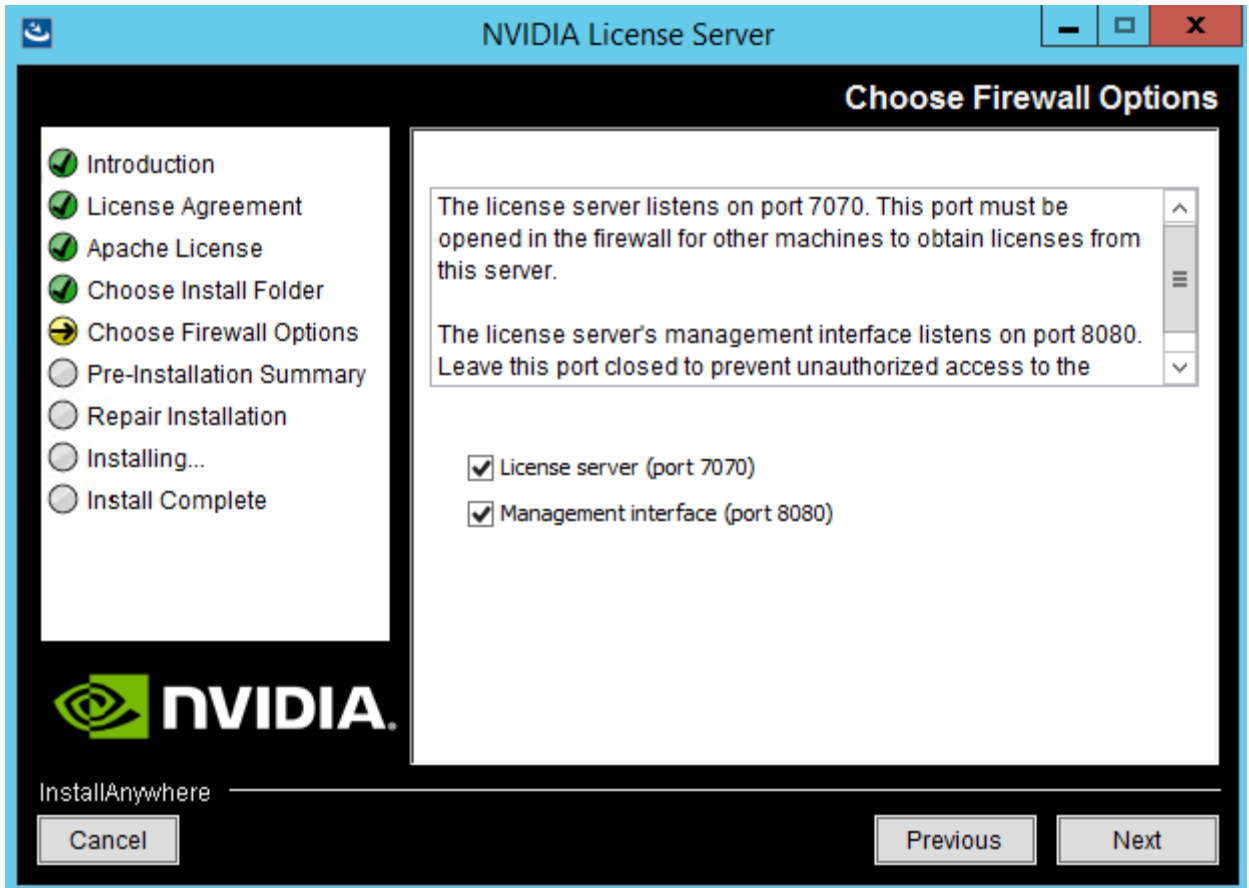
5. Choose the desired installation folder and click Next (Figure 23).

Figure 23. Choosing a Destination Folder



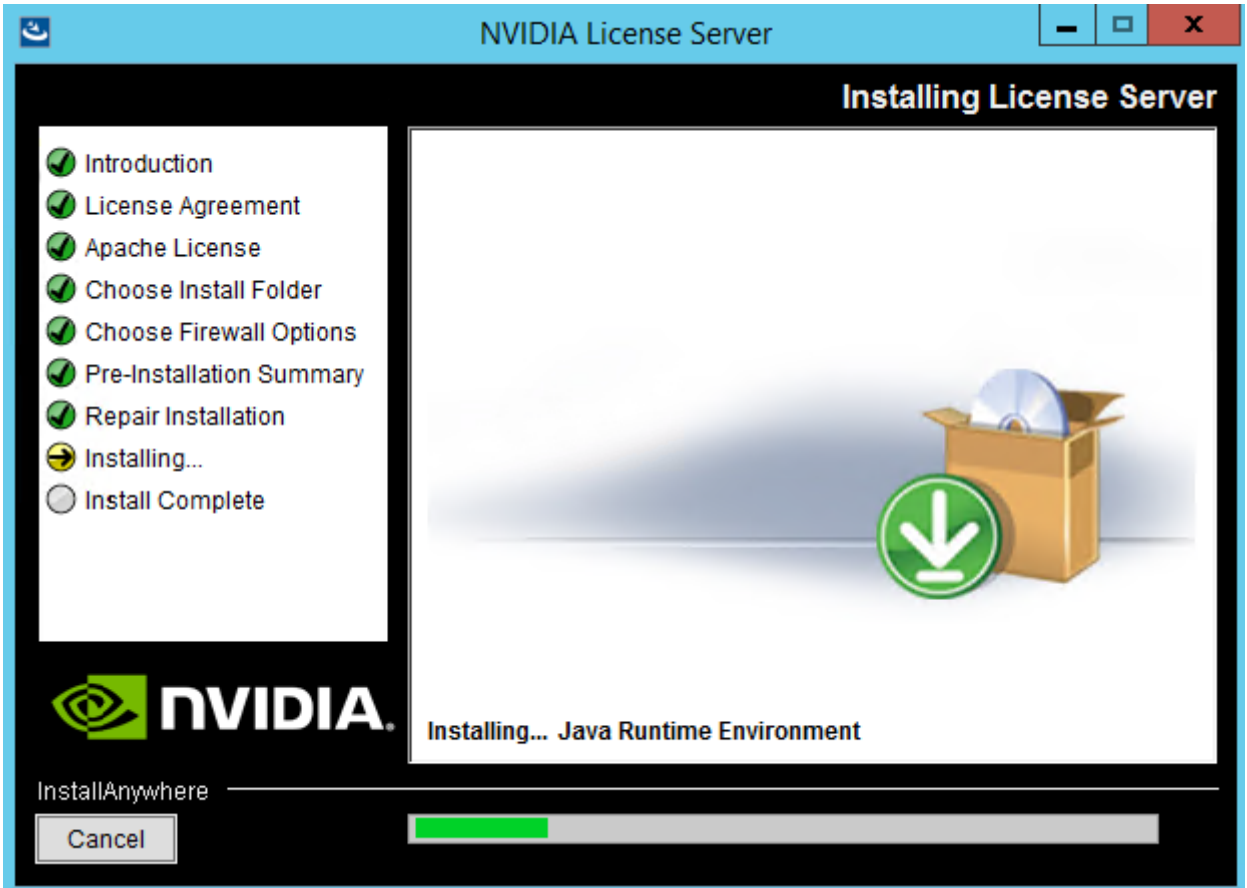
6. The license server listens on port 7070. This port must be opened in the firewall for other machines to obtain licenses from this server. Select the "License server (port 7070)" option.
7. The license server's management interface listens on port 8080. If you want the administration page accessible from other machines, you will need to open up port 8080. Select the "Management interface (port 8080)" option.
8. Click Next (Figure 24).

Figure 24. Setting Firewall Options



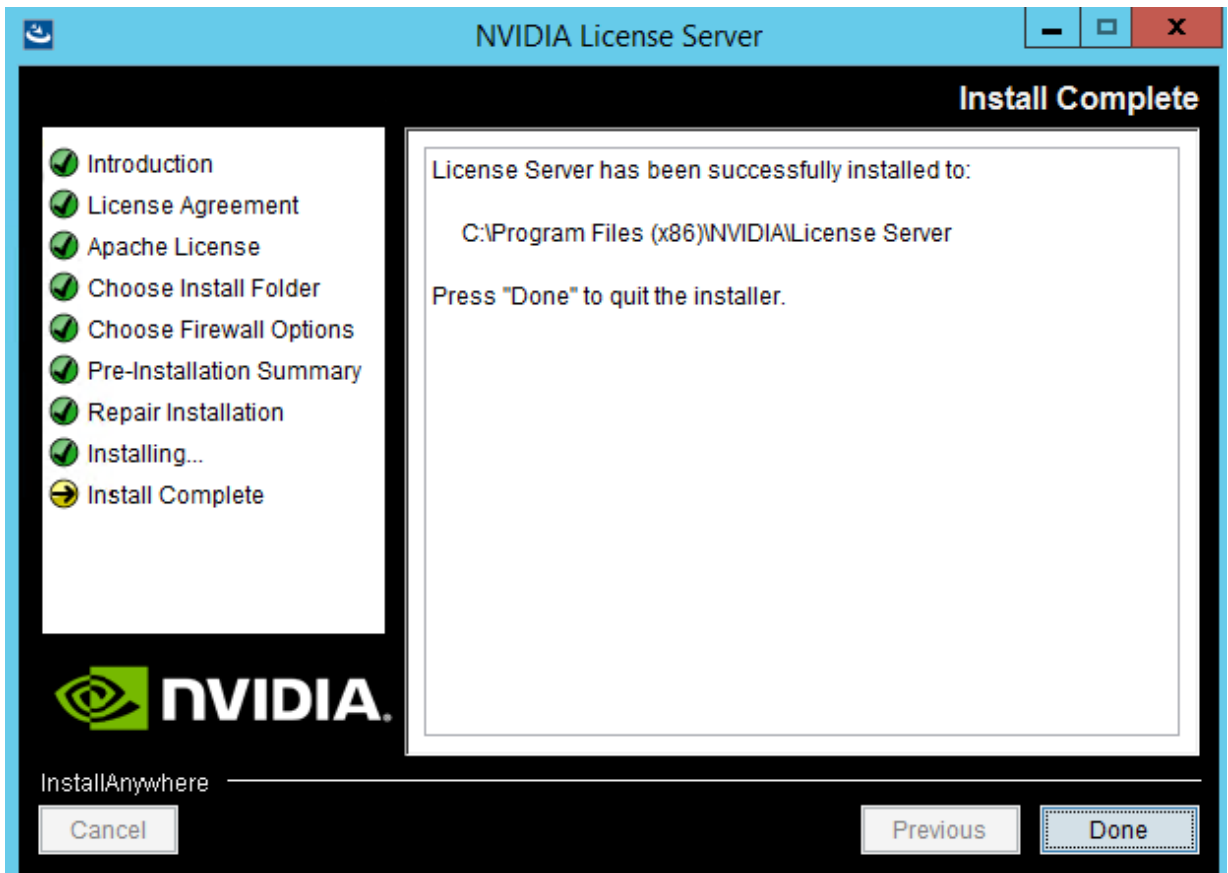
9. The Pre-installation Summary and Repair Installation options automatically progresses without user input (Figure 25).

Figure 25. Installing the License Server



10. When the installation process is complete, click Done (Figure 26).

Figure 26. Installation Complete



Configure the NVIDIA GRID 2.0 License Server

Now configure the NVIDIA Grid license server.

1. Log in to the license server site with the credentials set up during the registration process at nvidia.com/grideval. A license file is generated from <https://nvidia.flexnetoperations.com>.
2. After you are logged in, click Create License Server.
3. Specify the fields as shown in Figure 27. 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 27. Creating the License Server

The screenshot shows a web browser window with the URL `https://nvidia.flexnetoperations.com/control/nvda/createServer.lfs`. The page features the NVIDIA logo and a navigation breadcrumb: `HOME > NVIDIA SOFTWARE LICENSING CENTER > CREATE SERVER`. On the left, there is a sidebar menu with sections: **Software & Services** (Home, Product Search, Order History, Search Line Items, Recent Product Releases, Register Additional Keys), **Iray Licensing** (Search Licenses, View Licenses By Host, View Licenses Generated by User), and **Grid Licensing** (Search License Servers, **Create License Server**). The main content area is titled **Create Server** and includes the instruction: "To register an FNE license server to your account, provide the ID, ID type, and additional information below." The form contains the following fields: **License Server ID*** (text input with placeholder "<enter MAC Address of license server here>"), **ID Type** (dropdown menu with "ETHERNET" selected), **Alias** (text input with placeholder "<friendly name of your choosing>"), and **Site Name** (text input with placeholder "<friendly name of your site>"). A **Create** button is located below the Site Name field.

4. Click the Search License Servers node.
5. Click your license server ID (Figure 28).

Figure 28. Selecting the License Server ID

The screenshot shows the NVIDIA Software Licensing Center interface. The main heading is "Search Servers". On the left, there is a sidebar with sections: "Software & Services" (Home, Product Search, Order History, Search Line Items, Recent Product Releases, Register Additional Keys), "Iray Licensing" (Search Licenses, View Licenses By Host, View Licenses Generated by User), and "Grid Licensing" (Search License Servers, Create License Server). The "Search License Servers" option is highlighted. The main content area contains a search form with the following fields: "License Server ID" (text input), "ID Type" (dropdown menu), and "Activation Code" (text input). There are also "Alias" and "Site Name" labels on the right side of the form. A "Filter" button is located below the form. Below the form, there is a pagination control showing "1 to 1 of 1" and "Entries per page: 25". A table below the pagination shows one search result with columns "License Server ID" and "ID Type". The "ID Type" value is "ETHERNET".

- Click Map Add-Ons and choose the number of license “units” out of your total pool to allocate to this license server (Figure 29).

Figure 29. Choosing the Number of License Units from the Pool

View Server

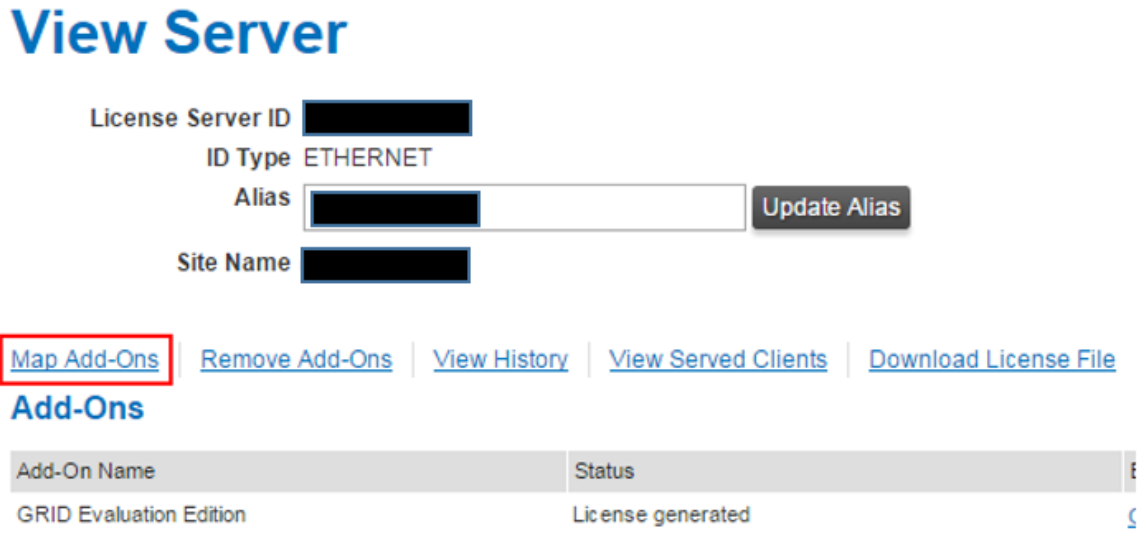
License Server ID [REDACTED]
 ID Type ETHERNET
 Alias [REDACTED] [Update Alias](#)
 Site Name [REDACTED]

[Map Add-Ons](#) | [Remove Add-Ons](#) | [View History](#) | [View Served Clients](#) | [Download License File](#)

Add-Ons

Add-On Name	Status
GRID Evaluation Edition	License generated

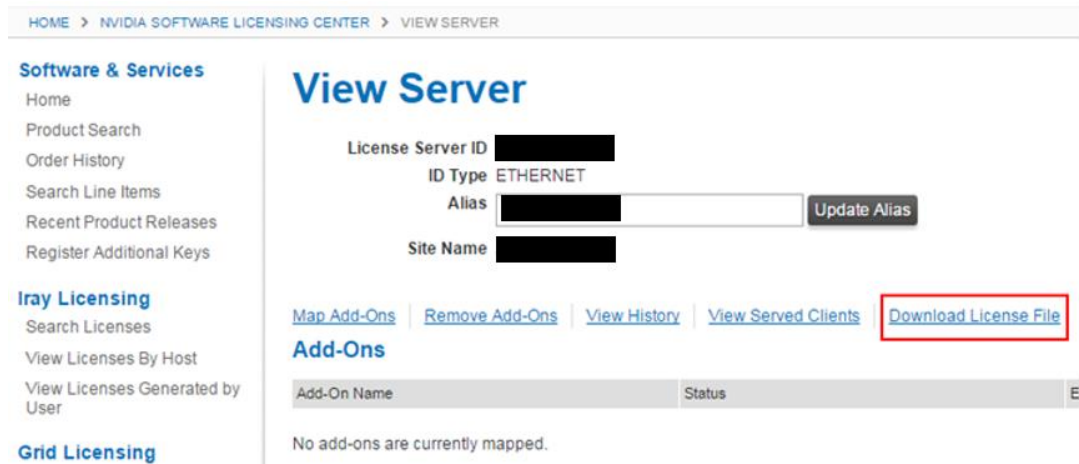
Figure 30. After the add-ons are mapped, the interface will look like Figure 30, showing 128 units mapped, for example.



7. Click Download License File and save the .bin file to your license server (Figure 31).

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

Figure 31. 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 32). The features are available (Figure 33).

Figure 32. License File Successfully Applied

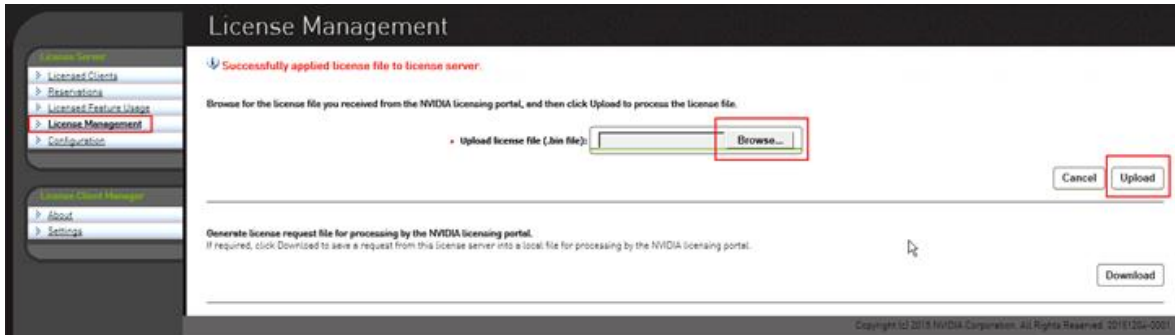
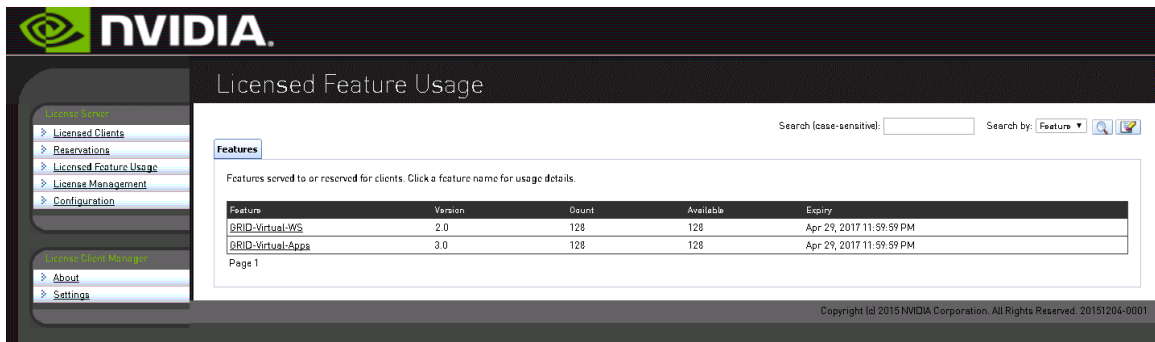
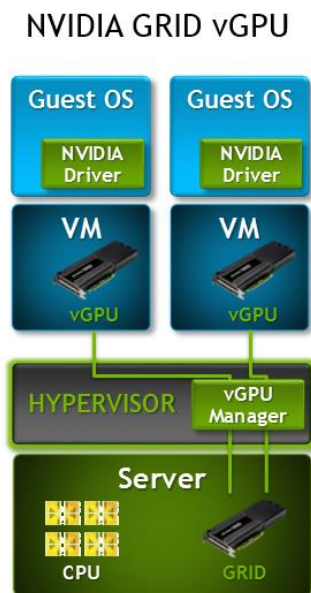


Figure 33. NVIDIA License Server



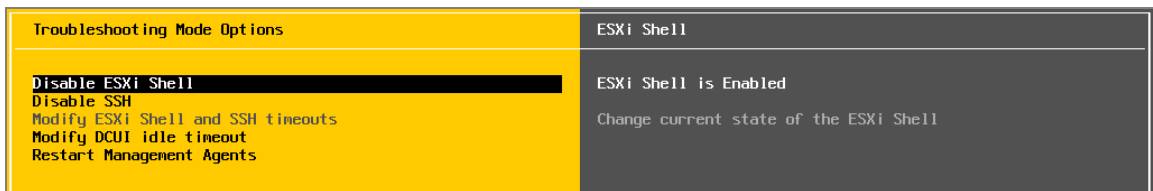
Install NVIDIA GRID Software on the VMware ESX Host and Microsoft Windows Virtual Machine
 This section summarizes the installation process for configuring an ESXi host and virtual machine for vGPU support. Figure 34 shows the components used for vGPU support.

Figure 34. NVIDIA GRID vGPU Components



1. Download the NVIDIA GRID GPU driver pack for VMware vSphere ESXi 6.0.
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 35).

Figure 35. 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.

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

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

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

The command should return output similar to that shown in Figure 36.

Figure 36. VMware ESX SSH Console Connection for vGPU Driver Installation

```
[root@esx:~]# esxcli software vib install -v /tmp/NVIDIA-vGPU-VMware_ESXi_6.0_Host_Driver_367.64-10EM.600.0.0.2494585.vib
Installation Result
Message: Operation finished successfully.
Reboot Required: false
VIBs Installed: NVIDIA_bootbank_NVIDIA-vGPU-VMware_ESXi_6.0_Host_Driver_367.64-10EM.600.0.0.2494585
VIBs Removed:
VIBs Skipped:
```

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

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

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

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

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

The command should return output similar to that shown in Figure 37.

Figure 37. VMware ESX SSH Console Connection for Driver Verification

```
[root@esx:~]# esxcli software vib list | grep -i nvidia
NVIDIA-VMware_ESXi_6.0_GpuModeSwitch_Driver 1.0-10EM.600.0.0.2494585 NVIDIA VMwareAccepted 2017-02-16
NVIDIA-vGPU-VMware_ESXi_6.0_Host_Driver 367.64-10EM.600.0.0.2494585 NVIDIA VMwareAccepted 2017-02-16
```


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

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

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

```
nvidia-smi
```

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

Figure 38. VMware ESX SSH Console Connection for GPU M60 Card Detection

```
[root@ ~] nvidia-smi
Mon Feb 20 17:23:51 2017

+-----+
| NVIDIA-SMI 367.64                Driver Version: 367.64          |
+-----+-----+
| 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 M60      On          | 0000:0E:00.0  Off  |      0%      Off  |
| N/A   36C   P8      24W / 150W | 19MiB / 8191MiB |           Default |
+-----+-----+
|  1   Tesla M60      On          | 0000:0F:00.0  Off  |      0%      Off  |
| N/A   34C   P8      24W / 150W | 19MiB / 8191MiB |           Default |
+-----+-----+
|  2   Tesla M60      On          | 0000:86:00.0  Off  |      0%      Off  |
| N/A   40C   P8      23W / 150W | 19MiB / 8191MiB |           Default |
+-----+-----+
|  3   Tesla M60      On          | 0000:87:00.0  Off  |      0%      Off  |
| N/A   36C   P8      23W / 150W | 19MiB / 8191MiB |           Default |
+-----+-----+

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

Figure 39. VMware ESX SSH Console Connection for GPU M10 Card Detection

```
[root@B200-M6:~] nvidia-smi
Mon Feb 6 19:27:07 2017

+-----+
| NVIDIA-SMI 367.64                Driver Version: 367.64          |
+-----+-----+
| 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          | 0000:0E:00.0  Off  |      N/A           |
| N/A   48C   P8     11W /  53W |  13MiB /  8191MiB |      0%      Default |
+-----+-----+
|  1  Tesla M10      On          | 0000:0F:00.0  Off  |      N/A           |
| N/A   47C   P8     11W /  53W |  13MiB /  8191MiB |      0%      Default |
+-----+-----+
|  2  Tesla M10      On          | 0000:10:00.0  Off  |      N/A           |
| N/A   36C   P8     11W /  53W |  13MiB /  8191MiB |      0%      Default |
+-----+-----+
|  3  Tesla M10      On          | 0000:11:00.0  Off  |      N/A           |
| N/A   38C   P8     11W /  53W |  13MiB /  8191MiB |      0%      Default |
+-----+-----+
|  4  Tesla M10      On          | 0000:86:00.0  Off  |      N/A           |
| N/A   45C   P8     10W /  53W |  13MiB /  8191MiB |      0%      Default |
+-----+-----+
|  5  Tesla M10      On          | 0000:87:00.0  Off  |      N/A           |
| N/A   44C   P8     10W /  53W |  13MiB /  8191MiB |      0%      Default |
+-----+-----+
|  6  Tesla M10      On          | 0000:88:00.0  Off  |      N/A           |
| N/A   33C   P8     10W /  53W |  13MiB /  8191MiB |      0%      Default |
+-----+-----+
|  7  Tesla M10      On          | 0000:89:00.0  Off  |      N/A           |
| N/A   37C   P8     10W /  53W |  13MiB /  8191MiB |      0%      Default |
+-----+-----+

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

Figure 40. VMware ESX SSH Console Connection for GPU M6 Card Detection

```
[root@B200-M6:~] nvidia-smi
Thu Feb 16 14:19:37 2017

+-----+
| NVIDIA-SMI 367.64                Driver Version: 367.64          |
+-----+-----+
| 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 M6      On          | 0000:81:00.0  Off  |      Off           |
| N/A   39C   P8     16W / 100W |  18MiB /  8191MiB |      0%      Default |
+-----+-----+

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

[root@B200-M6:~] █
```

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`.

- By default, the M6 and M60 cards use Compute mode. They will need to be switched to Graphics mode, which is required for vGPU support. You will need to download the gpumodeswitch utility from the NVIDIA website. The example here uses the boot ISO file, which loads a Linux environment with the [gpumodeswitch utility](#) already loaded (Figure 41).

Figure 41. Downloading the gpumodeswitch Utility

Name ^	Type	Compressed size	Password
gpumodeswitch	File	766 KB	No
gpumodeswitch	Application	618 KB	No
gpumodeswitch	Virtual CloneDrive	47,289 KB	No
gpumodeswitch	Compressed (zipped) Folder	47,268 KB	No
GRID gpumodeswitch User Guide	Firefox HTML Document	691 KB	No
LICENSES	Text Document	19 KB	No
nvfish64.sys	System file	8 KB	No

- Mount the ISO file through the Cisco UCS Manager Kernel-based Virtual Machine (KVM) and reboot the host.
- When the Linux shell loads, enter the following command (Figure 42):

```
gpumodeswitch -gpumode graphics
```

Figure 42. Installing the gpumodeswitch Utility

```
# gpumodeswitch --gpumode graphics

NVIDIA GPU Mode Switch Utility Version 1.02
Copyright (C) 2015, NVIDIA Corporation. All Rights Reserved.

Update GPU Mode of all adapters to "graphics"?
Press 'y' to confirm or 'n' to choose adapters or any other key to abort:
```

Alternatively, you can install gpumodeswitch vib on the ESXi host and enter the following command (Figure 43):

```
gpumodeswitch -gpumode graphics -auto
```

Figure 43. Installing the gpumodeswitch Utility on the VMware ESXi Host

```
[root@esx1:~] gpumodeswitch --gpumode graphics --auto

NVIDIA GPU Mode Switch Utility Version 1.23.0
Copyright (C) 2015, NVIDIA Corporation. All Rights Reserved.

Tesla M6          (10DE,13F3,10DE,1143) H:--:NRM S:00,B:81,PCI,D:00,F:00
Adapter: Tesla M6          (10DE,13F3,10DE,1143) H:--:NRM S:00,B:81,PCI,D:00,F:00

Identifying EEPROM...
EEPROM ID (EF,3013) : WBond W25X40A 2.7-3.6V 4096Kx1S, page

Programming UPR setting for requested mode..
License image updated successfully.

Programming ECC setting for requested mode..
The display may go *BLANK* on and off for up to 10 seconds or more during the update process
depending on your display adapter and output device.

Identifying EEPROM...
EEPROM ID (EF,3013) : WBond W25X40A 2.7-3.6V 4096Kx1S, page
NOTE: Preserving straps from original image.
Clearing original firmware image...
Storing updated firmware image...
.....
Verifying update...
Update successful.

Firmware image has been updated from version 84.04.89.00.01 to 84.04.89.00.01.

A reboot is required for the update to take effect.

InfoROM image updated successfully.
```

NVIDIA Tesla M60, M10, and M6 Profile Specifications

The Tesla M6 card has a single physical GPU, and the Tesla M60 and M10 cards each implement multiple physical GPUs. Each physical GPU can support several different types of virtual GPU. Each type of vGPU has a fixed amount of frame buffer space, a fixed number of supported display heads, and a fixed maximum resolution, and each is targeted at a different class of workload. Table 6 lists the vGPU types supported by GRID GPUs.

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

Table 6. User Profile Specifications for NVIDIA Tesla Cards

NVIDIA GRID Card	Physical GPUs	GRID vGPU	Intended Use Case	Frame Buffer (MB)	Virtual Display Heads	Maximum Resolution per Display Head	Maximum GPUs per GPU	Maximum vGPUs per Board
Tesla M60	2	M60-8Q	Designer	8192	4	4096 x 2160	1	2
Tesla M60	2	M60-4Q	Designer	4096	4	4096 x 2160	2	4
Tesla M60	2	M60-2Q	Designer	2048	4	4096 x 2160	4	8
Tesla M60	2	M60-1Q	<ul style="list-style-type: none"> Power user Designer 	1024	2	4096 x 2160	8	16
Tesla M60	2	M60-0Q	<ul style="list-style-type: none"> Power user Designer 	512	2	2560 x 1600	16	32
Tesla M60	2	M60-1B	Power user	1024	4	2560 x 1600	8	16
Tesla M60	2	M60-0B	Power user	512	2	2560 x 1600	16	32
Tesla M60	2	M60-8A	Virtual application user	8192	1	1280 x 1024	1	2
Tesla M60	2	M60-4A	Virtual application user	4096	1	1280 x 1024	2	4
Tesla M60	2	M60-2A	Virtual application user	2048	1	1280 x 1024	4	8

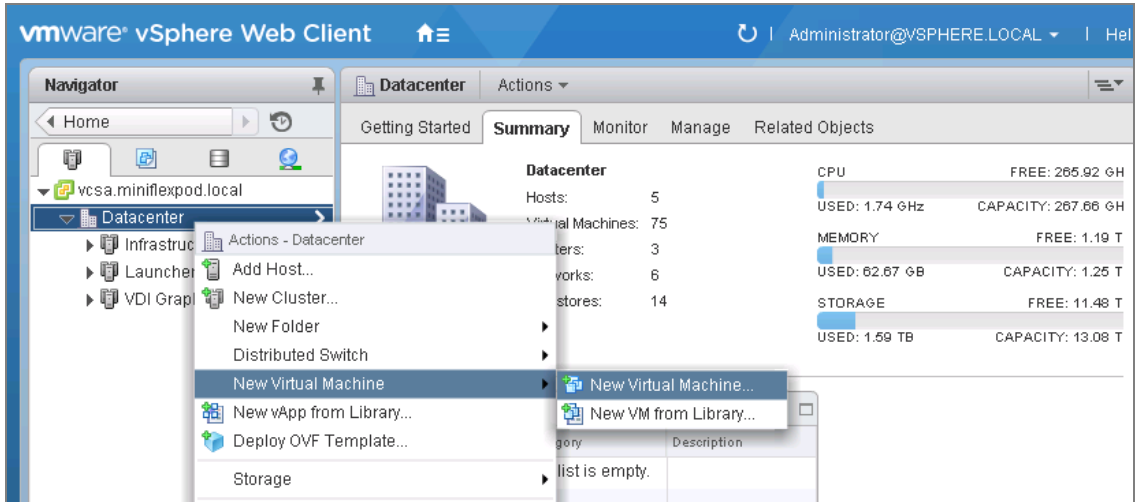
NVIDIA GRID Card	Physical GPUs	GRID vGPU	Intended Use Case	Frame Buffer (MB)	Virtual Display Heads	Maximum Resolution per Display Head	Maximum GPUs per GPU	Maximum vGPUs per Board
Tesla M60	2	M60-1A	Virtual application user	1024	1	1280 × 1024	8	16
Tesla M10	4	M10-8Q	Designer	8192	4	4096 × 2160	1	4
Tesla M10	4	M10-4Q	Designer	4096	4	4096 × 2160	2	8
Tesla M10	4	M10-2Q	Designer	2048	4	4096 × 2160	4	16
Tesla M10	4	M10-1Q	<ul style="list-style-type: none"> • Power user • Designer 	1024	2	4096 × 2160	8	32
Tesla M10	4	M10-0Q	<ul style="list-style-type: none"> • Power user • Designer 	512	2	2560 × 1600	16	64
Tesla M10	4	M10-1B	Power user	1024	4	2560 × 1600	8	32
Tesla M10	4	M10-0B	Power user	512	2	2560 × 1600	16	64
Tesla M10	4	M10-8A	Virtual application user	8192	1	1280 × 1024	1	4
Tesla M10	4	M10-4A	Virtual application user	4096	1	1280 × 1024	2	8
Tesla M10	4	M10-2A	Virtual application user	2048	1	1280 × 1024	4	16
Tesla M10	4	M10-1A	Virtual application user	1024	1	1280 × 1024	8	32
Tesla M6	1	M6-8Q	Designer	8192	4	4096 × 2160	1	1
Tesla M6	1	M6-4Q	Designer	4096	4	4096 × 2160	2	2
Tesla M6	1	M6-2Q	Designer	2048	4	4096 × 2160	4	4
Tesla M6	1	M6-1Q	<ul style="list-style-type: none"> • Power user • Designer 	1024	2	4096 × 2160	8	8
Tesla M6	1	M6-0Q	<ul style="list-style-type: none"> • Power user • Designer 	512	2	2560 × 1600	16	16
Tesla M6	1	M6-1B	Power user	1024	4	2560 × 1600	8	8
Tesla M6	1	M6-0B	Power user	512	2	2560 × 1600	16	16
Tesla M6	1	M6-8A	Virtual application user	8192	1	1280 × 1024	1	1
Tesla M6	1	M6-4A	Virtual application user	4096	1	1280 × 1024	2	2
Tesla M6	1	M6-2A	Virtual application user	2048	1	1280 × 1024	4	4
Tesla M6	1	M6-1A	Virtual application user	1024	1	1280 × 1024	8	8

Prepare a Virtual Machine for vGPU Support

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

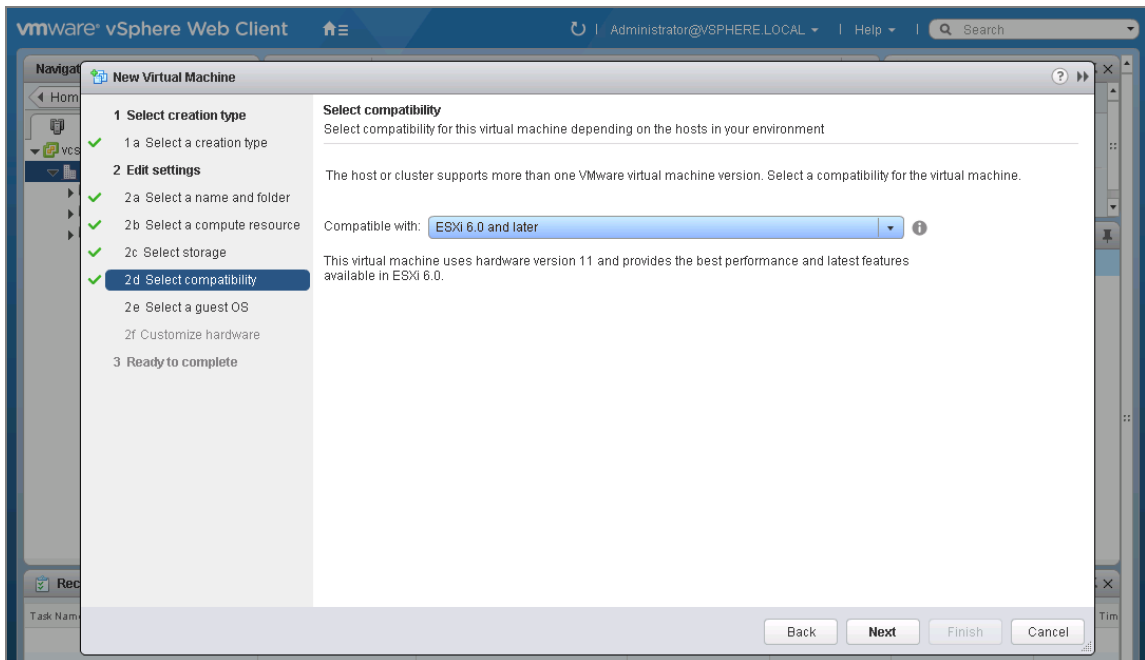
1. Using the vSphere Web Client, create a new virtual machine. To do this, right-click a host or cluster and choose New Virtual Machine. Work through the New Virtual Machine wizard. Unless another configuration is specified, select the configuration settings appropriate for your environment (Figure 44).

Figure 44. Creating a New Virtual Machine in VMware vSphere Web Client



2. Choose "ESXi 6.0 and later" from the "Compatible with" drop-down menu to use the latest features, including the mapping of shared PCI devices, which is required for the vGPU feature (Figure 45).

Figure 45. Selecting Virtual Machine Hardware Version 11



3. In customizing the hardware of the new virtual machine, add a new shared PCI device, select the appropriate GPU profile, and reserve all virtual machine memory (Figures 46 and 47).

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

Figure 46. Adding a Shared PCI Device to the Virtual Machine to Attach the GPU Profile

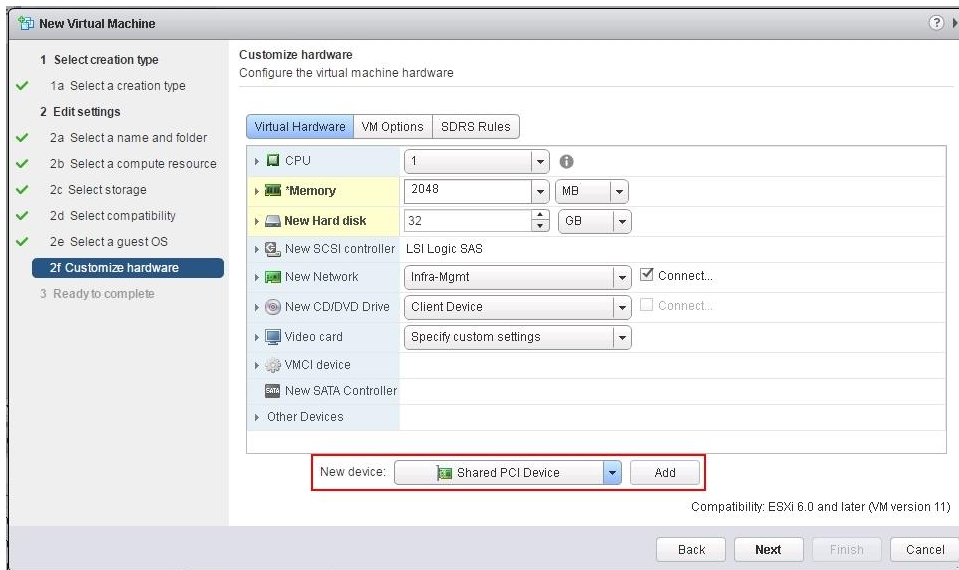
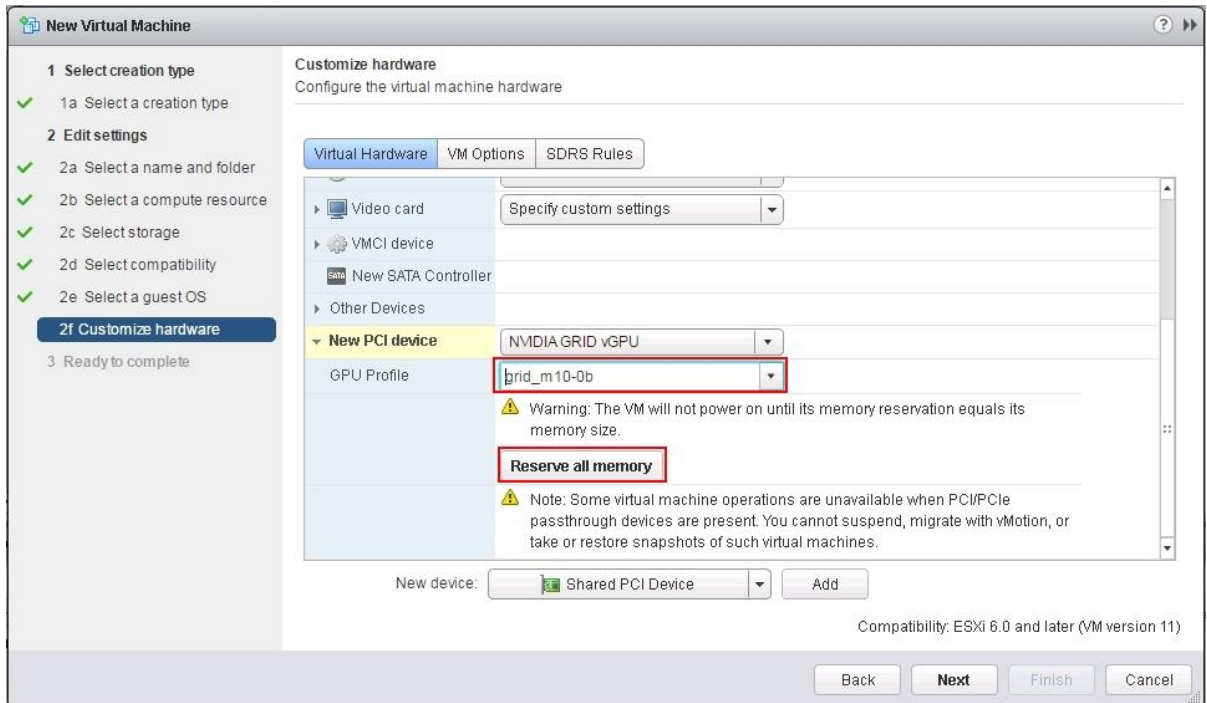


Figure 47. Attaching the GPU Profile to a Shared PCI Device and Reserving All Memory



4. 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.

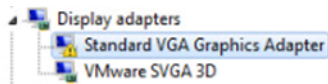
- d. Choose “Allow remote connections to this computer” on the Windows System Properties menu.
- e. Install VMware Horizon Agent with appropriate settings. Enable the remote desktop capability if prompted to do so.
- f. Install Horizon Direct Connection agent.
- g. Optimize the Windows OS. [VMware OSOT](#), 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 48).

Figure 48. Device Manager Before the NVIDIA Driver Is Installed



1. Copy the Windows drivers from the NVIDIA GRID vGPU driver pack downloaded earlier to the primary virtual machine
2. Copy the 32- or 64-bit NVIDIA Windows driver from the vGPU driver pack to the desktop virtual machine and run setup.exe (Figure 49).

Figure 49. NVIDIA Driver Pack

Name	Date modified	Type	Size
367.64-369.71-nvidia-grid-licensing-guide.pdf	11/15/2016 6:19 PM	Adobe Acrobat Doc...	1,729 KB
367.64-369.71-nvidia-grid-vgpu-release-notes-vmware-vsphere.pdf	12/15/2016 9:12 AM	Adobe Acrobat Doc...	1,589 KB
367.64-369.71-nvidia-grid-vgpu-user-guide.pdf	12/15/2016 9:12 AM	Adobe Acrobat Doc...	6,335 KB
369.71_grid_win8_win7_32bit_international.exe	11/14/2016 9:03 PM	Application	184,917 KB
369.71_grid_win8_win7_server2012R2_server2008R2_64bit_international.exe	11/14/2016 9:04 PM	Application	255,095 KB
369.71_grid_win10_32bit_international.exe	11/14/2016 9:03 PM	Application	186,615 KB
369.71_grid_win10_server2016_64bit_international.exe	11/14/2016 9:03 PM	Application	258,721 KB

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

3. Install the graphics drivers using the Express option (Figure 50). After the installation has been completed successfully (Figure 51), restart the virtual machine.

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

Figure 50. Selecting the Express Installation Option

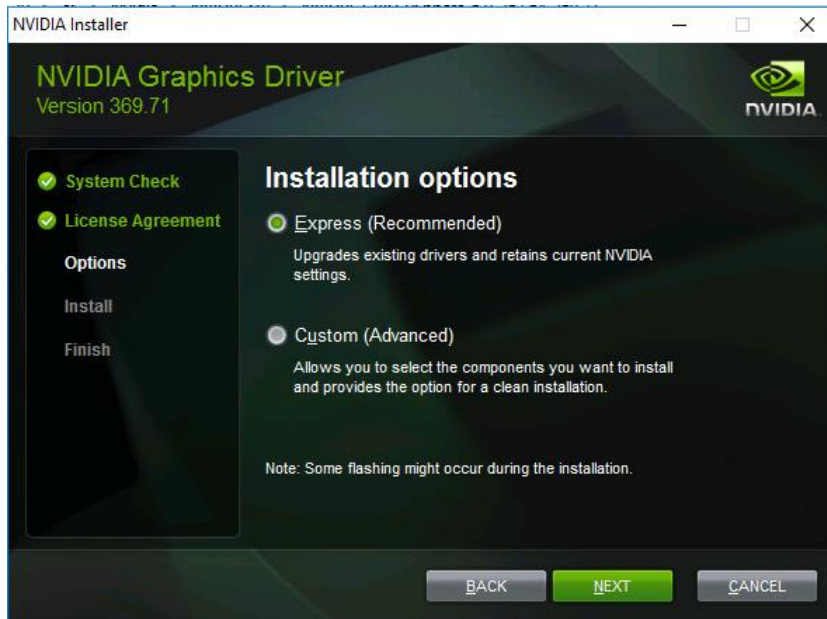
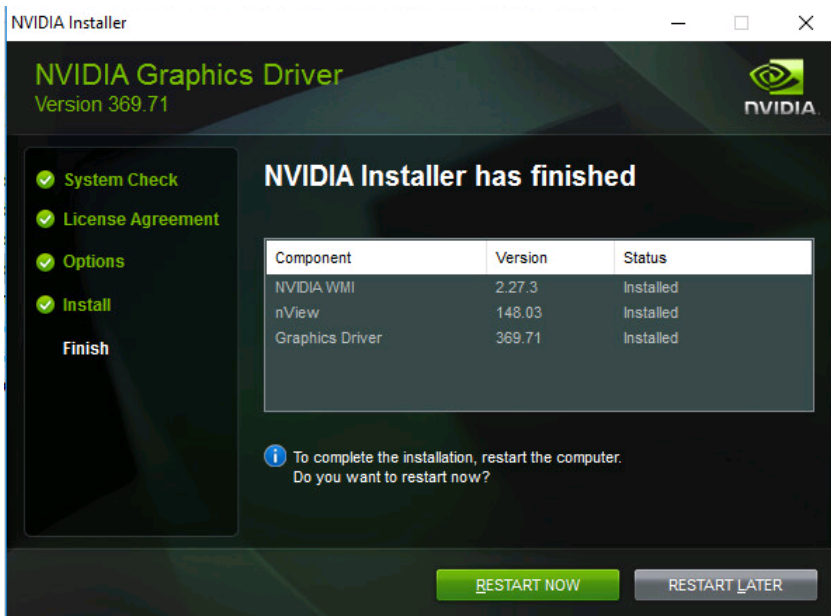


Figure 51. Express Installation Complete

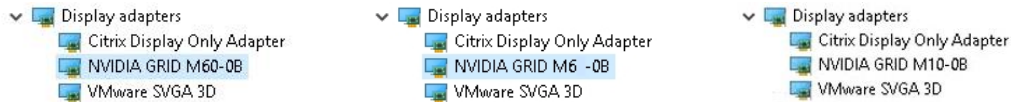


Verify That Applications Are Ready to Support vGPU

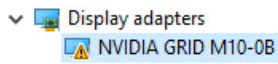
Validate the successful installation of the graphics drivers and the vGPU device.

Open Windows Device Manager and expand the Display Adapter section. The device will reflect chosen profile (Figure 52).

Figure 52. 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 primary 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.

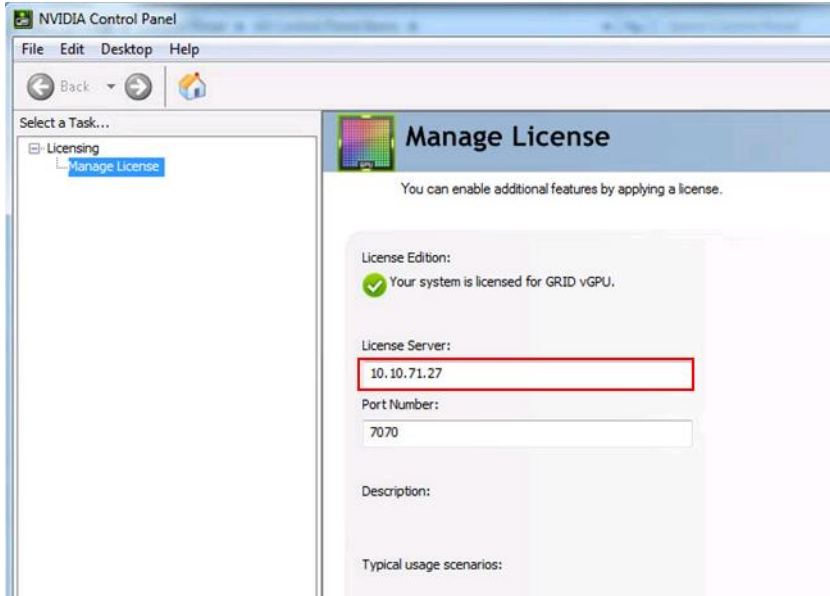
1. In the Microsoft Windows Control Panel, double-click NVIDIA Control Panel (Figure 53).

Figure 53. Choosing the NVIDIA Control Panel



2. Select Manage License from the left pane and enter your license server address and port (Figure 54).

Figure 54. Managing Your License



3. Select Apply.

Verify vGPU Deployment

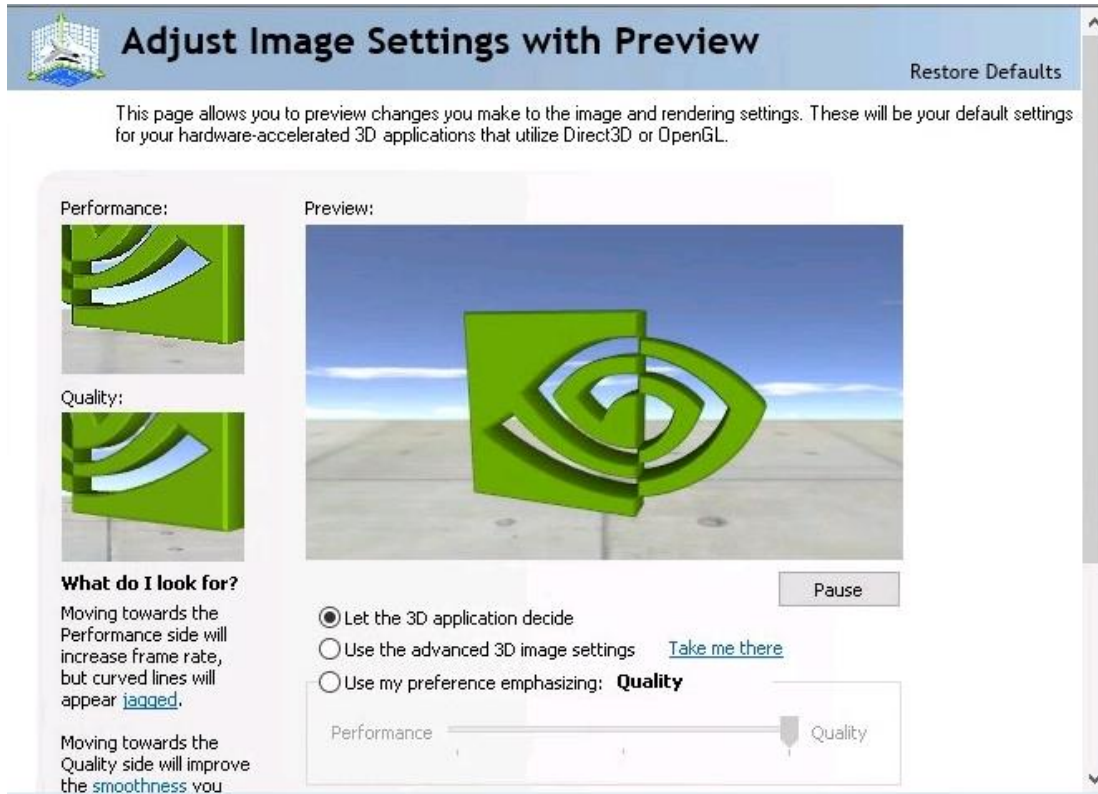
After the desktops are provisioned, use the following steps to verify vGPU deployment in the VMware Horizon environment.

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 55).

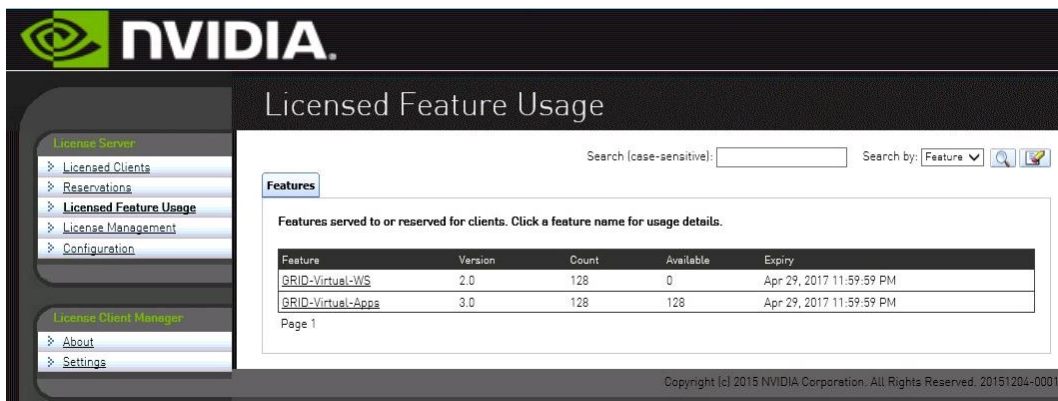
Figure 55. NVIDIA Control Panel



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 56).

Figure 56. NVIDIA License Server: Licensed Feature Usage



To view the details, select Licensed Clients in the left pane (Figure 57).

Figure 57. NVIDIA License Server: Licensed Clients

Search (case-sensitive): Search by: Client ID

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

Client ID	Client Alias	Client Type	Licensed Features	License Reservations
005056B29A8E	gpuxd-22.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B23BB2	gpuxd-109.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B2E1B3	gpuxd-27.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B278F1	gpuxd-100.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B29014	gpuxd-2.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B2A93B	gpuxd-1.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B2B081	gpuxd-4.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B2BFC9	gpuxd-10.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B278C8	gpuxd-5.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B2D3E8	gpuxd-9.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B2405D	gpuxd-18.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B2E7A4	gpuxd-24.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B26FE3	gpuxd-26.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B299B3	gpuxd-33.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B2DA32	gpuxd-36.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B29AC0	gpuxd-35.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B2A81C	gpuxd-32.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B2B175	gpuxd-44.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B2A6B1	gpuxd-53.vdilab-vs.local	GRID-Virtual-WS (1)		
005056B21C12	gpuxd-34.vdilab-vs.local	GRID-Virtual-WS (1)		

Verify the NVIDIA Configuration on the Host

To obtain a hostwide overview of the NVIDIA GPUs, enter the `nvidia-smi` command without any arguments (Figure 58).

Figure 58. The nvidia-smi Command Output from the Host with Two NVIDIA Tesla M10 Cards and 128 Microsoft Windows 10 Desktops with M10-0B vGPU Profile

```
[root@HV-GPUHost01:~] nvidia-smi
Thu Feb 16 02:20:46 2017

+-----+
| NVIDIA-SMI 367.64                  Driver Version: 367.64          |
+-----+-----+
| 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           | 0000:0E:00.0 | Off  | N/A |
| N/A  37C      P8           11W /  53W | 7476MiB / 8191MiB | 0%      Default |
+-----+-----+-----+-----+-----+-----+
| 1   Tesla M10   On           | 0000:0F:00.0 | Off  | N/A |
| N/A  37C      P8           11W /  53W | 7476MiB / 8191MiB | 1%      Default |
+-----+-----+-----+-----+-----+-----+
| 2   Tesla M10   On           | 0000:10:00.0 | Off  | N/A |
| N/A  33C      P8           11W /  53W | 7476MiB / 8191MiB | 0%      Default |
+-----+-----+-----+-----+-----+-----+
| 3   Tesla M10   On           | 0000:11:00.0 | Off  | N/A |
| N/A  33C      P8           11W /  53W | 7476MiB / 8191MiB | 0%      Default |
+-----+-----+-----+-----+-----+-----+
| 4   Tesla M10   On           | 0000:86:00.0 | Off  | N/A |
| N/A  36C      P8           10W /  53W | 7476MiB / 8191MiB | 0%      Default |
+-----+-----+-----+-----+-----+-----+
| 5   Tesla M10   On           | 0000:87:00.0 | Off  | N/A |
| N/A  35C      P8           10W /  53W | 7476MiB / 8191MiB | 0%      Default |
+-----+-----+-----+-----+-----+-----+
| 6   Tesla M10   On           | 0000:88:00.0 | Off  | N/A |
| N/A  31C      P8           10W /  53W | 7476MiB / 8191MiB | 0%      Default |
+-----+-----+-----+-----+-----+-----+
| 7   Tesla M10   On           | 0000:89:00.0 | Off  | N/A |
| N/A  33C      P8           10W /  53W | 7476MiB / 8191MiB | 0%      Default |
+-----+-----+-----+-----+-----+-----+

Processes:
+-----+-----+-----+-----+-----+-----+
| GPU  PID  Type  Process name      GPU Memory |
|      |      |      |                   | Usage     |
+-----+-----+-----+-----+-----+-----+
| 0    40630 C+G   win10gpu-002      464MiB |
| 0    40775 C+G   win10gpu-010      464MiB |
| 0    41357 C+G   win10gpu-015      464MiB |
| 0    41650 C+G   win10gpu-028      464MiB |
| 0    41836 C+G   win10gpu-037      464MiB |
| 0    41837 C+G   win10gpu-036      464MiB |
| 0    42222 C+G   win10gpu-053      464MiB |
| 0    42225 C+G   win10gpu-051      464MiB |
| 0    43353 C+G   win10gpu-069      464MiB |
| 0    43483 C+G   win10gpu-074      464MiB |
| 0    43723 C+G   win10gpu-083      464MiB |
| 0    43724 C+G   win10gpu-081      464MiB |
| 0    44089 C+G   win10gpu-103      464MiB |
| 0    44090 C+G   win10gpu-102      464MiB |
| 0    44317 C+G   win10gpu-117      464MiB |
| 0    44576 C+G   win10gpu-121      464MiB |
| 1    40648 C+G   win10gpu-008      464MiB |
| 1    41293 C+G   win10gpu-014      464MiB |
| 1    41413 C+G   win10gpu-019      464MiB |
| 1    41414 C+G   win10gpu-020      464MiB |
| 1    41415 C+G   win10gpu-018      464MiB |
| 1    42014 C+G   win10gpu-042      464MiB |
| 1    42224 C+G   win10gpu-052      464MiB |
| 1    42329 C+G   win10gpu-058      464MiB |
+-----+-----+-----+-----+-----+-----+

```

```

1 42329 C+G win10gpu-058 464M1B |
1 42330 C+G win10gpu-060 464M1B |
1 43527 C+G win10gpu-076 464M1B |
1 43767 C+G win10gpu-084 464M1B |
1 43909 C+G win10gpu-091 464M1B |
1 43910 C+G win10gpu-093 464M1B |
1 43911 C+G win10gpu-092 464M1B |
1 44316 C+G win10gpu-115 464M1B |
1 44578 C+G win10gpu-122 464M1B |
2 40665 C+G win10gpu-007 464M1B |
2 41294 C+G win10gpu-012 464M1B |
2 41534 C+G win10gpu-023 464M1B |
2 41838 C+G win10gpu-035 464M1B |
2 42071 C+G win10gpu-047 464M1B |
2 42072 C+G win10gpu-046 464M1B |
2 42073 C+G win10gpu-041 464M1B |
2 43247 C+G win10gpu-063 464M1B |
2 43249 C+G win10gpu-061 464M1B |
2 43526 C+G win10gpu-075 464M1B |
2 43768 C+G win10gpu-087 464M1B |
2 43954 C+G win10gpu-094 464M1B |
2 44091 C+G win10gpu-101 464M1B |
2 44136 C+G win10gpu-107 464M1B |
2 44376 C+G win10gpu-120 464M1B |
2 44577 C+G win10gpu-123 464M1B |
3 40666 C+G win10gpu-005 464M1B |
3 41295 C+G win10gpu-013 464M1B |
3 41535 C+G win10gpu-022 464M1B |
3 41536 C+G win10gpu-024 464M1B |
3 41537 C+G win10gpu-021 464M1B |
3 42116 C+G win10gpu-048 464M1B |
3 42117 C+G win10gpu-049 464M1B |
3 43248 C+G win10gpu-062 464M1B |
3 43355 C+G win10gpu-068 464M1B |
3 43528 C+G win10gpu-071 464M1B |
3 43529 C+G win10gpu-077 464M1B |
3 43955 C+G win10gpu-097 464M1B |
3 43956 C+G win10gpu-095 464M1B |
3 44194 C+G win10gpu-110 464M1B |
3 44195 C+G win10gpu-109 464M1B |
3 44977 C+G win10gpu-125 464M1B |
4 40668 C+G win10gpu-003 464M1B |
4 41596 C+G win10gpu-025 464M1B |
4 41598 C+G win10gpu-027 464M1B |
4 41883 C+G win10gpu-040 464M1B |
4 41884 C+G win10gpu-038 464M1B |
4 42284 C+G win10gpu-055 464M1B |
4 43143 C+G win10gpu-031 464M1B |
4 43144 C+G win10gpu-011 464M1B |
4 43354 C+G win10gpu-070 464M1B |
4 43589 C+G win10gpu-079 464M1B |
4 43769 C+G win10gpu-085 464M1B |
4 43957 C+G win10gpu-096 464M1B |
4 44134 C+G win10gpu-104 464M1B |
4 44196 C+G win10gpu-108 464M1B |
4 44374 C+G win10gpu-118 464M1B |
4 44979 C+G win10gpu-127 464M1B |
5 40667 C+G win10gpu-001 464M1B |
5 41356 C+G win10gpu-017 464M1B |
5 41597 C+G win10gpu-026 464M1B |
5 41771 C+G win10gpu-032 464M1B |

```



```

4 42284 C+G win10gpu-055 464MiB |
4 43143 C+G win10gpu-031 464MiB |
4 43144 C+G win10gpu-011 464MiB |
4 43354 C+G win10gpu-070 464MiB |
4 43589 C+G win10gpu-079 464MiB |
4 43769 C+G win10gpu-085 464MiB |
4 43957 C+G win10gpu-096 464MiB |
4 44134 C+G win10gpu-104 464MiB |
4 44196 C+G win10gpu-108 464MiB |
4 44374 C+G win10gpu-118 464MiB |
4 44979 C+G win10gpu-127 464MiB |
5 40667 C+G win10gpu-001 464MiB |
5 41356 C+G win10gpu-017 464MiB |
5 41597 C+G win10gpu-026 464MiB |
5 41771 C+G win10gpu-032 464MiB |
5 41882 C+G win10gpu-039 464MiB |
5 42118 C+G win10gpu-050 464MiB |
5 42283 C+G win10gpu-056 464MiB |
5 43293 C+G win10gpu-064 464MiB |
5 43481 C+G win10gpu-072 464MiB |
5 43588 C+G win10gpu-080 464MiB |
5 43770 C+G win10gpu-086 464MiB |
5 44014 C+G win10gpu-100 464MiB |
5 44135 C+G win10gpu-105 464MiB |
5 44271 C+G win10gpu-113 464MiB |
5 44375 C+G win10gpu-119 464MiB |
5 44978 C+G win10gpu-124 464MiB |
6 40669 C+G win10gpu-006 464MiB |
6 40670 C+G win10gpu-004 464MiB |
6 41649 C+G win10gpu-030 464MiB |
6 41773 C+G win10gpu-033 464MiB |
6 42011 C+G win10gpu-043 464MiB |
6 42223 C+G win10gpu-054 464MiB |
6 42282 C+G win10gpu-057 464MiB |
6 43294 C+G win10gpu-066 464MiB |
6 43482 C+G win10gpu-073 464MiB |
6 43590 C+G win10gpu-078 464MiB |
6 43827 C+G win10gpu-088 464MiB |
6 44015 C+G win10gpu-099 464MiB |
6 44016 C+G win10gpu-098 464MiB |
6 44269 C+G win10gpu-111 464MiB |
6 44270 C+G win10gpu-112 464MiB |
6 44980 C+G win10gpu-126 464MiB |
7 40760 C+G win10gpu-009 464MiB |
7 41358 C+G win10gpu-016 464MiB |
7 41648 C+G win10gpu-029 464MiB |
7 41774 C+G win10gpu-034 464MiB |
7 42012 C+G win10gpu-044 464MiB |
7 42013 C+G win10gpu-045 464MiB |
7 42328 C+G win10gpu-059 464MiB |
7 43722 C+G win10gpu-082 464MiB |
7 43828 C+G win10gpu-089 464MiB |
7 43829 C+G win10gpu-090 464MiB |
7 44137 C+G win10gpu-106 464MiB |
7 44314 C+G win10gpu-114 464MiB |
7 44315 C+G win10gpu-116 464MiB |
7 45037 C+G win10gpu-128 464MiB |
7 46007 C+G win10gpu-065 464MiB |
7 46008 C+G win10gpu-067 464MiB |
+-----+
[root@HV-GPUHost01:~]

```

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 primary 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 Remote Display Protocol, choose VMware Blast (Figure 59).

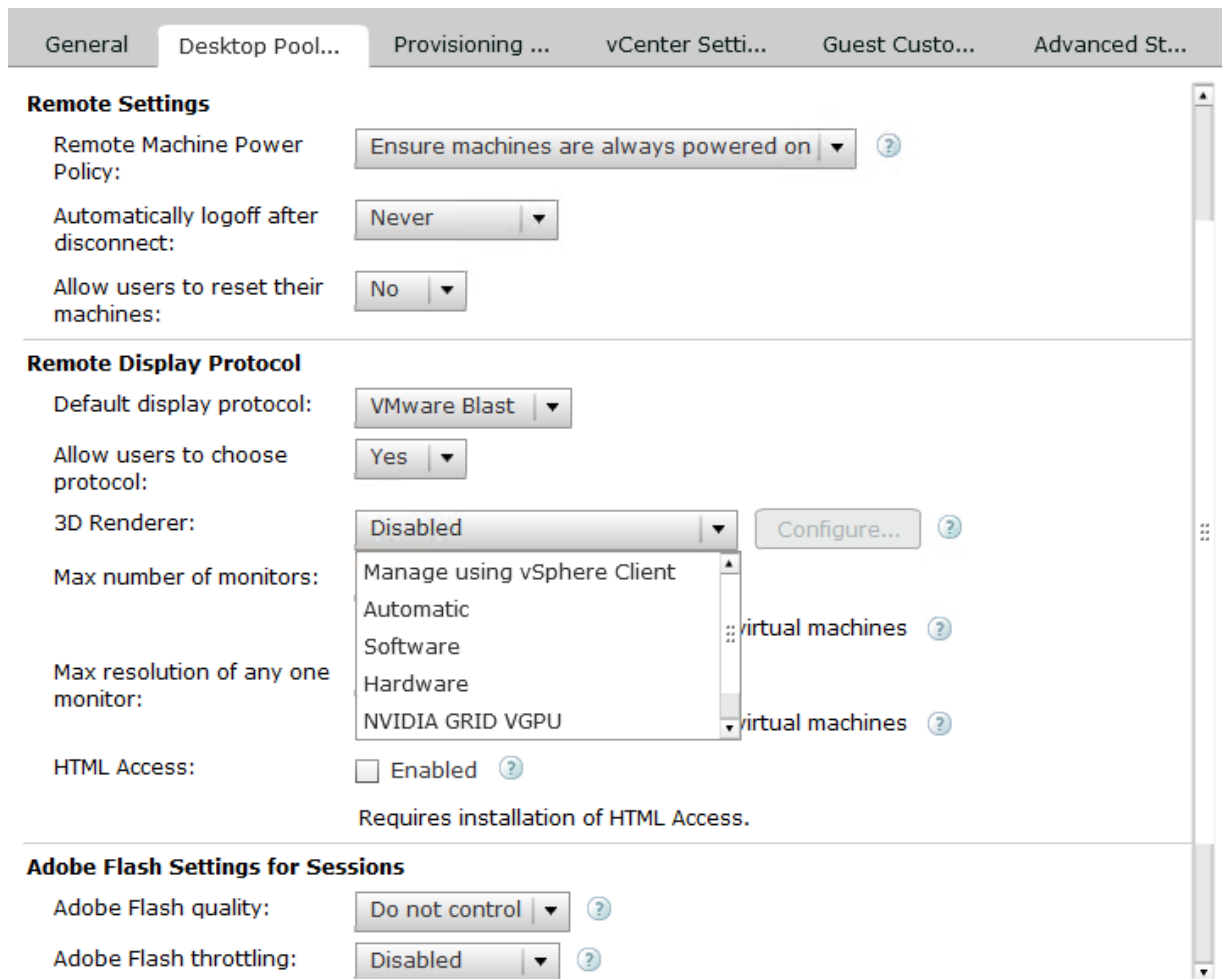
Figure 59. Configuring the Remote Display Protocol

The screenshot shows the configuration interface for a Horizon 7 desktop pool, specifically the 'Remote Display Protocol' settings. The interface has a top navigation bar with tabs: 'General', 'Desktop Pool...', 'Provisioning ...', 'vCenter Setti...', 'Guest Custo...', and 'Advanced St...'. The 'Desktop Pool...' tab is active. Below the navigation bar, there are three main sections: 'Remote Settings', 'Remote Display Protocol', and 'Adobe Flash Settings for Sessions'.
Remote Settings:
- Remote Machine Power Policy: 'Ensure machines are always powered on' (dropdown menu with a help icon).
- Automatically logoff after disconnect: 'Never' (dropdown menu).
- Allow users to reset their machines: 'No' (dropdown menu).
Remote Display Protocol:
- Default display protocol: 'VMware Blast' (dropdown menu).
- Allow users to choose protocol: A dropdown menu is open, showing options: 'Microsoft RDP', 'PCoIP', and 'VMware Blast' (highlighted).
- 3D Renderer: 'VMware Blast' (dropdown menu) with a 'Configure...' button and a help icon.
- Max number of monitors: '2' (dropdown menu with a help icon).
- Max resolution of any one monitor: '1920x1200' (dropdown menu with a help icon).
- HTML Access: 'Enabled' (checkbox with a help icon).
Adobe Flash Settings for Sessions:
- Adobe Flash quality: 'Do not control' (dropdown menu with a help icon).
- Adobe Flash throttling: 'Disabled' (dropdown menu with a help icon).

Select the option for 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 (Figure 60).

Figure 60. Configuring Rendering Settings



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

Use the OpenGL Software Accelerator

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

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

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

Try the OpenGL Software Accelerator in the following cases:

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

Conclusion

The combination of Cisco UCS Manager, Cisco UCS C240 M4 Rack Servers and B200 M4 Blade Servers and NVIDIA Tesla cards running on VMware vSphere 6.0 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](#)
- [NVIDIA](#)
- 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:
 - http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2107948

-
- http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2109712
 - http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2033434



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)