Cisco UCS C240-M3 Rack Server with NVIDIA GRID Cards On VMware Horizon View 5.3

July 2014
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive Summary</td>
<td>3</td>
</tr>
<tr>
<td><strong>Cisco Unified Computing System</strong></td>
<td>3</td>
</tr>
<tr>
<td>Cisco UCS Manager</td>
<td>5</td>
</tr>
<tr>
<td>Cisco UCS Fabric Interconnect</td>
<td>5</td>
</tr>
<tr>
<td>Cisco UCS Fabric Extenders</td>
<td>6</td>
</tr>
<tr>
<td>Cisco UCS 2232PP Fabric Extender</td>
<td>6</td>
</tr>
<tr>
<td>Cisco UCS C-Series Rack Servers</td>
<td>6</td>
</tr>
<tr>
<td>Cisco UCS C240 M3 Rack Server</td>
<td>7</td>
</tr>
<tr>
<td>Cisco VIC 1225–10GE Option</td>
<td>8</td>
</tr>
<tr>
<td>Emulex OneConnect OCE11102-F CNA</td>
<td>9</td>
</tr>
<tr>
<td><strong>NVIDIA GRID Cards</strong></td>
<td>10</td>
</tr>
<tr>
<td>Network Delivered GPU Acceleration for Virtual Desktops</td>
<td>10</td>
</tr>
<tr>
<td>Graphics Accelerated Virtual Desktops and Applications</td>
<td>10</td>
</tr>
<tr>
<td><strong>VMware vSphere 5.5</strong></td>
<td>12</td>
</tr>
<tr>
<td>VMware ESXi 5.5 Hypervisor</td>
<td>12</td>
</tr>
<tr>
<td><strong>Graphics Acceleration in Horizon View 5.3</strong></td>
<td>13</td>
</tr>
<tr>
<td>Difference Between Soft 3D, vSGA and vDGA</td>
<td>13</td>
</tr>
<tr>
<td>Soft 3D—Software-Based 3D Rendering</td>
<td>14</td>
</tr>
<tr>
<td>vSGA—Virtual Shared Graphics Acceleration</td>
<td>14</td>
</tr>
<tr>
<td>vDGA—Virtual Dedicated Graphics Acceleration</td>
<td>15</td>
</tr>
<tr>
<td>Software Requirement for vSGA and vDGA</td>
<td>15</td>
</tr>
<tr>
<td><strong>Solution Configuration</strong></td>
<td>16</td>
</tr>
<tr>
<td>UCS Configuration</td>
<td>17</td>
</tr>
<tr>
<td>Installing NVIDIA GRID or Tesla GPU Card on C240 M3:</td>
<td>17</td>
</tr>
<tr>
<td>Base UCS System Configuration</td>
<td>19</td>
</tr>
<tr>
<td>vSGA configuration</td>
<td>21</td>
</tr>
<tr>
<td>vDGA Configuration</td>
<td>21</td>
</tr>
<tr>
<td>Enable Device Pass-through</td>
<td>21</td>
</tr>
<tr>
<td>Enable Virtual Machine for vDGA configuration</td>
<td>23</td>
</tr>
<tr>
<td>Configure 3D rendering using VMware vSphere:</td>
<td>28</td>
</tr>
<tr>
<td>Configure 3D rendering using Horizon View</td>
<td>30</td>
</tr>
<tr>
<td>Screen Resolution</td>
<td>31</td>
</tr>
<tr>
<td>Performance Tuning Tips</td>
<td>32</td>
</tr>
<tr>
<td><strong>Conclusion</strong></td>
<td>33</td>
</tr>
<tr>
<td><strong>References</strong></td>
<td>33</td>
</tr>
</tbody>
</table>
Executive Summary

With the increased processor power of today’s Cisco UCS B-Series and Cisco UCS C-Series servers, applications with demanding graphics components are now being considered for virtualization. Enhancing the capability to deliver these high-performance, graphics-rich applications are the addition of GRID K1 and K2 cards from NVIDIA to the UCS portfolio of PCIe cards for our C-Series rack servers.

With the addition of the new graphics processing capabilities, inclusion of the engineering, design, imaging and marketing departments of organizations can now enjoy the benefits that desktop virtualization brings to this space.

This new graphics capability gives organizations the ability to centralize their graphics workloads and data in the datacenter. This provides a very large benefit to organizations that need to be able to work-shift geographically. Until now, that has not been possible because the graphics files are too large to move and the files must be local to the person using them to be usable.

The benefit of PCIe graphics cards in Cisco UCS C-Series servers is four fold:

- Support for full length, full power NVIDIA GRID cards in a 2U form factor
- Cisco UCS Manager integration for management of the servers and GRID cards
- End to end integration with Cisco UCS management suite, including Cisco UCS Central and Cisco UCS Director
- Cisco UCS C240 M3s with two NVIDIA GRID K1 or K2 cards provide more efficient rack space than the two slot, 2.5 equivalent rack unit, HP WS460c workstation blade with the NVIDIA GRID card in a second chassis slot

The purpose of this white paper is to help partners and customers integrate NVIDIA GRID graphics processor cards with Cisco UCS C240 M3 rack servers on VMware vSphere and VMware Horizon in both Virtual Dedicated Graphics Acceleration (vDGA) and Virtual Shared Graphics Acceleration (vSGA) modes.

We believe that our partners, NVIDIA and VMware, are in the best position to provide lists of applications that are supported by the card, their hypervisor and their desktop broker in vDGA and vSGA modes.

The objective of this white paper is to provide the reader with specific methods to integrate Cisco UCS C240 M3 rack servers with NVIDIA GRID K1 and/or K2 cards with VMware products so that the servers, the hypervisor and the desktop broker are ready for graphic application installation.

Cisco Unified Computing System

The Cisco Unified Computing System is a next-generation data center platform that unites compute, network, 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.
The main components of Cisco Unified Computing System are:

- **Computing**—The system is based on an entirely new class of computing system that incorporates blade servers based on Intel Xeon E5-2600/4600 and E7-2800 Series 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 networks which are separate networks today. The unified fabric lowers costs by reducing the number of network adapters, switches, and cables, and by decreasing the power and cooling requirements.

- **Virtualization**—The system unleashes the full potential of virtualization by enhancing the scalability, performance, and operational control of virtual environments. Cisco security, policy enforcement, and diagnostic features are now extended into virtualized environments to better support changing business and IT requirements.

- **Storage access**—The system provides consolidated access to both SAN storage and Network Attached Storage (NAS) over the unified fabric. By unifying the storage access the Cisco Unified Computing System can access storage over Ethernet, Fibre Channel, Fibre Channel over Ethernet (FCoE), and iSCSI.
provides customers with choice for storage access and investment protection. In addition, the server administrators can pre-assign storage-access policies for system connectivity to storage resources, simplifying storage connectivity, and management for increased productivity.

- **Management**—The system uniquely integrates all system components which enable the entire solution to be managed as a single entity by the Cisco UCS Manager. The Cisco UCS Manager has an intuitive graphical user interface (GUI), a command-line interface (CLI), and a robust application programming interface (API) to manage all system configuration and operations.

The Cisco Unified Computing System is designed to deliver:

- A reduced Total Cost of Ownership and increased business agility.
- Increased IT staff productivity through just-in-time provisioning and mobility support.
- A cohesive, integrated system which 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 the Cisco Unified Computing System through an intuitive GUI, a command line interface (CLI), or an XML API. The Cisco UCS Manager provides unified management domain with centralized management capabilities and controls multiple chassis and thousands of virtual machines.

**Cisco UCS Fabric Interconnect**

The Cisco® UCS 6200 Series Fabric Interconnect is a core part of the Cisco Unified Computing System, providing both network connectivity and management capabilities for the system. The Cisco UCS 6200 Series offers line-rate, low-latency, lossless 10 Gigabit Ethernet, Fibre Channel over Ethernet (FCoE) and Fibre Channel functions.

The Cisco UCS 6200 Series provides the management and communication backbone for the Cisco UCS B-Series Blade Servers and Cisco UCS 5100 Series Blade Server Chassis. All chassis, and therefore all blades, attached to the Cisco UCS 6200 Series Fabric Interconnects become part of a single, highly available management domain. In addition, by supporting unified fabric, the Cisco UCS 6200 Series provides both the LAN and SAN connectivity for all blades within its domain.

From a networking perspective, the Cisco UCS 6200 Series uses a cut-through architecture, supporting deterministic, low-latency, line-rate 10 Gigabit Ethernet on all ports, 1Tb switching capacity, 160 Gbps bandwidth per chassis, independent of packet size and enabled services. The product family supports Cisco low-latency, lossless 10 Gigabit Ethernet unified network fabric capabilities, which increase the reliability, efficiency, and scalability of Ethernet networks. The Fabric Interconnect supports multiple traffic classes over a lossless Ethernet fabric from a blade server through an interconnect. Significant TCO savings come from an FCoE-optimized server design in which network interface cards (NICs), host bus adapters (HBAs), cables, and switches can be consolidated.
Cisco UCS Fabric Extenders

Fabric Extenders are zero-management, low-cost, low-power consuming devices that distribute the system's connectivity and management planes into rack and blade chassis to scale the system without complexity. Designed never to lose a packet, Cisco fabric extenders eliminate the need for top-of-rack Ethernet and Fibre Channel switches and management modules, dramatically reducing infrastructure cost per server.

Cisco UCS 2232PP Fabric Extender

The Cisco Nexus® 2000 Series Fabric Extenders comprise a category of data center products designed to simplify data center access architecture and operations. The Cisco Nexus 2000 Series uses the Cisco® Fabric Extender architecture to provide a highly scalable unified server-access platform across a range of 100 Megabit Ethernet, Gigabit Ethernet, 10 Gigabit Ethernet, unified fabric, copper and fiber connectivity, rack, and blade server environments. The platform is ideal to support today's traditional Gigabit Ethernet while allowing transparent migration to 10 Gigabit Ethernet, virtual machine-aware unified fabric technologies.

The Cisco Nexus 2000 Series Fabric Extenders behave as remote line cards for a parent Cisco Nexus switch or Fabric Interconnect. The fabric extenders are essentially extensions of the parent Cisco UCS Fabric Interconnect switch fabric, with the fabric extenders and the parent Cisco Nexus switch together forming a distributed modular system. This architecture enables physical topologies with the flexibility and benefits of both top-of-rack (ToR) and end-of-row (EoR) deployments.

Today's data centers must have massive scalability to manage the combination of an increasing number of servers and a higher demand for bandwidth from each server. The Cisco Nexus 2000 Series increases the scalability of the access layer to accommodate both sets of demands without increasing management points within the network.

Cisco UCS C-Series Rack Servers

Cisco UCS C-Series Rack-Mount 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, Cisco UCS C-Series servers offer an improved price-to-performance ratio; extend Cisco Unified Computing System 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 the Cisco Unified Computing System, 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. Cisco UCS C-Series servers offer investment protection through the capability to deploy them either as standalone servers or as part of the Cisco Unified Computing System.

One compelling reason that many organizations prefer rack-mount servers is the wide range of I/O options available in the form of PCI Express (PCIe) adapters. Cisco UCS C-Series servers supports spectrum of I/O options, which includes interfaces supported by Cisco as well as adapters from third parties.

Cisco UCS C240 M3 Rack Server

The Cisco UCS C240 M3 Rack Server (Figure X1) 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 M3 server further extends the capabilities of the Cisco UCS portfolio in a 2RU form factor with the addition of the Intel® Xeon processor E5-2600 and E5-2600 v2 product families, which deliver an outstanding combination of performance, flexibility, and efficiency gains. The Cisco UCS C240 M3 offers up to two Intel Xeon processor E5-2600 or E5-2600 v2 processors, 24 DIMM slots, 24 disk drives, and four 1 Gigabit Ethernet LAN-on-motherboard (LOM) ports to provide exceptional levels of internal memory and storage expandability and exceptional performance.

The Cisco UCS C240 M3 interfaces with the Cisco UCS Virtual Interface Card. The Cisco UCS Virtual Interface Card is a virtualization-optimized Fibre Channel over Ethernet (FCoE) PCI Express (PCIe) 2.0 x8 10-Gbps adapter designed for use with Cisco UCS C-Series Rack Servers. The VIC is a dual-port 10 Gigabit Ethernet PCIe adapter that can support up to 256 PCIe standards-compliant virtual interfaces, which can be dynamically configured so that both their interface type (network interface card [NIC] or host bus adapter [HBA]) and identity (MAC address and worldwide name [WWN]) are established using just-in-time provisioning. In addition, the Cisco UCS VIC 1225 can support network interface virtualization and Cisco® Data Center Virtual Machine Fabric Extender (VM-FEX) technology. An additional five PCIe slots are made available for certified third party PCIe cards. The server is equipped to handle 24 on board SAS drives or SSDs along with shared storage solutions offered by our partners.

Cisco UCS C240 M3 server's disk configuration delivers balanced performance and expandability to best meet individual workload requirements. With up to 12 LFF (Large Form Factor) or 24 SFF (Small Form Factor) internal drives, the Cisco UCS C240 M3 optionally offers 10,000-RPM and 15,000-RPM SAS drives to deliver a high number of I/O operations per second for transactional workloads such as database management systems. In addition, high-capacity SATA drives provide an economical, large-capacity solution. Superfast SSDs are a third option for workloads that demand extremely fast access to smaller amounts of data. A choice of RAID controller options also helps increase disk performance and reliability.

The Cisco UCS C240 M3 further 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
Cisco VIC 1225–10GE Option
A Cisco UCS Virtual Interface Card (VIC) 1225 (Figure 3) is a dual-port Enhanced Small Form-Factor Pluggable (SFP+) 10 Gigabit Ethernet and Fibre Channel over Ethernet (FCoE)-capable PCI Express (PCIe) card designed exclusively for Cisco UCS C-Series Rack Servers. With its half-height design, the card preserves full-height slots in servers for third-party adapters certified by Cisco. It incorporates next-generation converged network adapter (CNA) technology. 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). In addition, the Cisco UCS VIC 1225 supports Cisco Data Center Virtual Machine Fabric Extender (VM-FEX) technology, which extends the Cisco UCS fabric interconnect ports to virtual machines, simplifying server virtualization deployment.

Table 1. Cisco UCS C240 M3 PCIe slot specification

<table>
<thead>
<tr>
<th>PCIe Slot</th>
<th>Length</th>
<th>Lane</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>half</td>
<td>x8</td>
</tr>
<tr>
<td>2</td>
<td>half</td>
<td>x16</td>
</tr>
<tr>
<td>3</td>
<td>half</td>
<td>x8</td>
</tr>
<tr>
<td>4</td>
<td>3/4</td>
<td>x8</td>
</tr>
<tr>
<td>5</td>
<td>3/4</td>
<td>x16</td>
</tr>
</tbody>
</table>
Due to current limitation of UCS C240M3 platform; if intend to use two Nvidia GPU cards installed on C240 M3 then Cisco VIC 1225 has to be replace with Emulex 3rd Gen CNA card. For this study we used OCE11102-F CNA.

Emulex OneConnect OCE11102-F CNA

The Emulex OneConnect® OCE11102-F is a dual-port 10Gb Ethernet (10GbE) PCIe 2.0 x8 adapter that consolidates network and storage traffic with high performance CPU offloads for Fibre Channel over Ethernet (FCoE) and Internet Small Computer System Interface (iSCSI) protocols.

A member of the Emulex OneConnect Universal Converged Network Adapter (UCNA) family, the OCE11102-F adapter supports a common infrastructure for networking and storage, reducing capital expenditures (CAPEX) for adapters, switches and cables, and operational expenditures (OPEX) for power, cooling and IT administration.

The OneConnect OCE11102-F supports protocol offloads for FCoE, iSCSI, TCP/IP and TCP Chimney to provide maximum bandwidth with minimum use of CPU resources. For virtualized server deployments, OneConnect
protocol offloads enable more virtual machines (VMs) per server, providing greater cost saving to optimize return on investment.

Figure 8. Emulex CNA OCE11102-F

NVIDIA GRID Cards

Network Delivered GPU Acceleration for Virtual Desktops
The NVIDIA GRID portfolio of technologies leverages the power of the GPU and the world's best graphics applications to deliver GPU-accelerated applications and games over the network to any user. NVIDIA GRID GPUs are based on the NVIDIA Kepler™ GPU architecture, delivering fast, reliable, energy-efficient performance. This architecture's virtualization capabilities lets multiple users simultaneously share GPUs with ultra-fast streaming display capability that eliminates lag, making a remote data center feel like it's next door. NVIDIA GRID software is a complete stack of GPU virtualization, remoting and session-management libraries that allows multiple users to experience graphics-intensive desktops, applications and games using GPUs. This enables exceptional capture, efficient compression, fast streaming, and low-latency display of high-performance enterprise applications. See more at: http://www.nvidia.com/object/nvidia-grid.html

Graphics Accelerated Virtual Desktops and Applications
NVIDIA GRID™ technology offers the ability to offload graphics processing from the CPU to the GPU in virtualized environments, allowing the data center manager to deliver true PC graphics-rich experiences to more users for the first time.
Benefits of NVIDIA GRID for IT:

- Leverage industry-leading virtualization solutions, including Citrix, Microsoft, and VMware
- Add your most graphics-intensive users to your virtual solutions
- Improve the productivity of all users

Benefits of NVIDIA GRID for users:

- Highly responsive windows and rich multimedia experiences
- Access to all critical applications, including the most 3D-intensive
- Access from anywhere, on any device

See more at:


Table 2. NVIDIA GRID K1 and GRID K2 Comparison

<table>
<thead>
<tr>
<th></th>
<th>GRID K1</th>
<th>GRID K2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of GPUs</td>
<td>4 x entry Kepler GPUs</td>
<td>2 x high-end Kepler GPUs</td>
</tr>
<tr>
<td>Total NVIDIA CUDA cores</td>
<td>768</td>
<td>3072</td>
</tr>
<tr>
<td>Total memory size</td>
<td>16 GB DDR3</td>
<td>8 GB GDDR5</td>
</tr>
<tr>
<td>Max power</td>
<td>130 W</td>
<td>225 W</td>
</tr>
<tr>
<td>Board length</td>
<td>10.5</td>
<td>10.5</td>
</tr>
<tr>
<td>Board height</td>
<td>4.4</td>
<td>4.4</td>
</tr>
<tr>
<td>Board width</td>
<td>Dual slot</td>
<td>Dual slot</td>
</tr>
<tr>
<td>Display IO</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Aux power</td>
<td>6-pin connector</td>
<td>8-pin connector</td>
</tr>
<tr>
<td>PCIe</td>
<td>X16</td>
<td>X16</td>
</tr>
<tr>
<td>PCIe generation</td>
<td>Gen3 (Gen2 compatible)</td>
<td>Gen3 (Gen2 compatible)</td>
</tr>
<tr>
<td>Cooling solution</td>
<td>Passive</td>
<td>Passive</td>
</tr>
<tr>
<td>Technical Specifications</td>
<td>GRID K1 Board Specifications</td>
<td>GRID K2 Board Specifications</td>
</tr>
</tbody>
</table>

1. NVIDIA GRID™ vGPU™ is only supported on compatible versions of Citrix XenServer. Consult Citrix for compatibility.
2. Only compatible with VMware vSphere Hypervisor. Consult VMware for compatibility.
VMware vSphere 5.5

VMware, Inc. provides virtualization software. VMware’s enterprise software hypervisors for servers—VMware vSphere ESX, VMware 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 which provides central management, complete control and visibility into clusters, hosts, virtual machines, storage, networking and other critical elements of your virtual infrastructure.

VMware ESXi 5.5 Hypervisor

ESXi 5.5 is a "bare-metal" hypervisor, so it installs directly on top of the physical server and partitions it into multiple virtual machines that can run simultaneously, sharing the physical resources of the underlying server. VMware introduced ESXi in 2007 to deliver industry-leading performance and scalability while setting a new bar for reliability, security and hypervisor management efficiency.

Due to its ultra-thin architecture with less than 100MB of code-base disk footprint, ESXi delivers industry-leading performance and scalability plus:

- **Improved Reliability and Security**—with fewer lines of code and independence from general purpose OS, ESXi drastically reduces the risk of bugs or security vulnerabilities and makes it easier to secure your hypervisor layer.

- **Streamlined Deployment and Configuration**—ESXi has far fewer configuration items than ESX, greatly simplifying deployment and configuration and making it easier to maintain consistency.

- **Higher Management Efficiency**—The API-based, partner integration model of ESXi eliminates the need to install and manage third party management agents. You can automate routine tasks by leveraging remote command line scripting environments such as vCLI or PowerCLI.

- **Simplified Hypervisor Patching and Updating**—Due to its smaller size and fewer components, ESXi requires far fewer patches than ESX, shortening service windows and reducing security vulnerabilities.
Graphics Acceleration in Horizon View 5.3

3D graphics capabilities in VMware Horizon View further expand the target user base and potential use cases IT can deliver with virtual desktops. In addition, 3D augments the virtual desktop user interface by enabling a more graphically rich user experience.

**Figure 10.** Virtual Desktop User Segmentation

![Virtual Desktop User Segmentation](image)

**Difference Between Soft 3D, vSGA and vDGA**

**Table 3.** Graphics-Driver Comparison

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Soft 3D</strong></td>
<td>Software 3D Renderer</td>
<td>Support for software-accelerated 3D graphics is provided via a VMware WDDM (Windows Display Driver Model) 1.1–compliant driver without any physical GPUs being installed in the ESXi host</td>
</tr>
<tr>
<td>vSGA</td>
<td>Virtual Shared Graphics Acceleration</td>
<td>Multiple virtual machines leverage physical GPUs installed locally in the VMware ESXi™ host(s) to provide hardware-accelerated 3D graphics to multiple virtual desktops</td>
</tr>
<tr>
<td>vDGA</td>
<td>Virtual Dedicated Graphics Acceleration</td>
<td>Only one virtual machine is mapped to a single physical GPU installed in the ESXi host to provide high-end, hardware-accelerated workstation graphics where a discrete GPU is needed</td>
</tr>
</tbody>
</table>
Soft 3D—Software-Based 3D Rendering:
The Soft 3D renderer is based on a VMware WDDM 1.1–compliant driver and is installed with VMware Tools™ onto Windows 7 virtual desktops. Soft 3D is differentiated from vSGA and vDGA in that it does not require any physical GPUs to be installed in the ESXi host.

The VMware Soft 3D graphics driver provides support for DirectX 9.0c and OpenGL 2.1. The driver is supported on Windows 7 for 2D and 3D and is used for both Soft 3D and vSGA. Virtual Dedicated Graphics Acceleration (vDGA) configurations do not use the VMware Soft 3D driver; instead, they use the native graphics-card driver installed directly in the guest OS.

One of the benefits of VMware Soft 3D for both software 2D and 3D and vSGA implementations is that a virtual machine can dynamically switch between software and hardware acceleration, without your having to reconfigure. Additionally, sharing this driver allows the use of VMware high-availability technologies such as VMware vSphere® vMotion®. Finally, having a single driver greatly simplifies image management and deployment.

Note: If you are dynamically moving from hardware 3D rendering to software 3D rendering, you may notice a performance drop in applications running in the virtual machine. However, if you are moving in the reverse direction (software to hardware), you should notice an improvement in performance.

vSGA—Virtual Shared Graphics Acceleration
To provide hardware-accelerated 3D graphics, vSGA allows multiple virtual machines to leverage physical GPUs installed locally in the ESXi hosts. This differs from Soft 3D in that there are physical GPUs installed in the host server. These GPUs are shared across multiple virtual machines, unlike vDGA, where each virtual machine is directly mapped to a single GPU.

The maximum amount of video memory that can be assigned per virtual machine is 512MB. However, video-memory allocation is evenly divided: Half the video memory is reserved on the hardware GPU, while the other half is reserved via host RAM. (Take this into consideration when sizing your ESXi host RAM.)

You can use this rule to calculate basic consolidation ratios. For example, the NVIDIA GRID K1 card has 16GB of GPU RAM. If all virtual machines are configured with 512MB of video memory, half of which (256MB) is reserved on the GPU; you can calculate that a maximum of 64 virtual machines can run on that specific GPU at any given time.

The ESXi host reserves GPU hardware resources on a first-come, first-served basis as virtual machines are powered on. If all GPU hardware resources are already reserved, additional virtual machines will be unable to power on if they are explicitly set to use hardware 3D rendering. If the virtual machines are set to Automatic, they will be powered on using software 3D rendering.

vSGA is limited by the amount of memory on the installed boards. VMware ESXi assigns a virtual machine to a particular graphics device during power on. The assignment is based on graphics-memory reservation that occurs in a round-robin fashion. The current policy is to reserve one-half of the virtual machine’s VRAM size with a minimum of 128MB. This means a graphics device with a 4GB memory can accept at most 32 virtual machines with minimum reservation. Once a graphics device reaches its reservation maximum, no more virtual machines will be assigned to it until another virtual machine leaves the GPU. This can occur by virtual machine power off, suspend, or vSphere vMotion moving it to another host.
vDGA—Virtual Dedicated Graphics Acceleration

Virtual Dedicated Graphics Acceleration (vDGA) is a graphics-acceleration capability provided by VMware ESXi for delivering high-end workstation graphics for use cases where a discrete GPU is needed. This graphics-acceleration method dedicates a single GPU to a single virtual machine for high performance.

**Note:** Some graphics cards can have multiple GPUs on them.

If you are using vDGA, graphics adapters installed in the underlying host are assigned to virtual machines using VMware vSphere DirectPath I/O. Assigning a discrete GPU to the virtual machine dedicates the entire GPU to it.

For vDGA, the number of 3D-enabled virtual machines is limited by the number of GPUs in the server. A UCS C240 M3 can accommodate two cards. The NVIDIA GRID K2 card has two GPUs. If an administrator installs a GRID K2 card into each available slot on the UCS C240 M3, there will be a total of 4 GPUs in the system. This will be the total number of vDGA-enabled virtual machines that the server can support.

**Note:** Both vSGA and vDGA can support maximum of eight GPU cards per ESXi host.

Software Requirement for vSGA and vDGA

<table>
<thead>
<tr>
<th>Product</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMware vSphere Hypervisor</td>
<td>vSGA and vDGA: ESXi 5.1 U1 or ESXi 5.5 (ESXi 5.5 recommended)</td>
</tr>
<tr>
<td>VMware Horizon View</td>
<td>vSGA: VMware Horizon View 5.2 or later (VMware Horizon View 5.3 recommended)</td>
</tr>
<tr>
<td></td>
<td>vDGA: VMware Horizon View 5.3</td>
</tr>
<tr>
<td>Display Protocol</td>
<td>vSGA and vDGA: PCoIP with a maximum of two display monitors</td>
</tr>
<tr>
<td>NVIDIA Drivers</td>
<td>vSGA: NVIDIA drivers for VMware vSphere ESXi5.5 version 319.65</td>
</tr>
<tr>
<td></td>
<td>vDGA: Tesla/GRID desktop driver version 333.11</td>
</tr>
<tr>
<td></td>
<td><strong>Note:</strong> These drivers are supplied and supported by NVIDIA. Both drivers can be downloaded from the NVIDIA Drivers Download page.</td>
</tr>
<tr>
<td>Guest Operating System</td>
<td>vSGA: Windows 7 32- or 64-bit</td>
</tr>
<tr>
<td></td>
<td>vDGA: Windows 7 64-bit</td>
</tr>
</tbody>
</table>
Solution Configuration

Figure 11. Reference Architecture

Hardware Components:
- Cisco UCS C240-M3 Rack Server (2 Intel Xeon processor E5-2680 v2 @ 2.70 GHz) with 256GB of memory (16 GB X 16 DIMMS @ 1866 MHz), hypervisor host
- Cisco UCS VIC1225 CNA per rack server - 1 GPU scenario
- 2 x Cisco Nexus 2232PP Fabric Extenders – 2 GPU scenario
- 2 x Cisco UCS 6248UP Fabric Interconnects
- 12 x 600GB SAS disks @ 10000 RPM
- 1 or 2 NVIDIA GRID K1/K2 Cards
- Emulex OneConnect OCE11102-F Gen 3 CNA card 10Gbps per rack server - 2 GPU scenario

Software components:
- Cisco UCS firmware 2.2(2C)
- VMware ESXi 5.5 for virtual desktop infrastructure (VDI) Hosts
- VMware Horizon View 5.3
- Microsoft Windows 7 SP1 32 bit, 2 virtual CPU, and 2 GB of memory (For vSGA)
- Microsoft Windows 7 SP1 64 bit, 4 virtual CPU, and 8 GB of memory (For vDGA)
UCS Configuration

Installing NVIDIA GRID or Tesla GPU Card on C240 M3:

Prerequisite:

Table 4. Minimum server firmware versions for the GPU cards

<table>
<thead>
<tr>
<th>GPU</th>
<th>Minimum BIOS Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA GRID K1</td>
<td>1.5(1)</td>
</tr>
<tr>
<td>NVIDIA GRID K</td>
<td>1.5(1)</td>
</tr>
<tr>
<td>NVIDIA Tesla K20</td>
<td>1.5(3)</td>
</tr>
<tr>
<td>NVIDIA Tesla K20X</td>
<td>1.5(3)</td>
</tr>
</tbody>
</table>

NVIDIA GPU Card Configuration Rules:

- Mixing different GPU cards in the same server is not supported.
- All GPU cards require two CPUs and two 1200 W power supplies in the server.
- It is not possible to use dual NVIDIA GPU cards and a Cisco virtual interface card (VIC) at the same time. This is because dual NVIDIA GPUs must be installed in slots 2 and 5 of the server, and a Cisco VIC must be installed in either slot 2 or slot 5. If you require two GPU cards and 10-Gb Ethernet connectivity, you must choose a different supported adapter that can be used in a different slot. For supported adapters, see the Technical Specifications Sheet for the Cisco UCS C240 M3 server (Small Form Factor or Large Form Factor) at: http://www.cisco.com/en/US/products/ps10493/products_data_sheets_list.html

Figure 12. One GPU scenario
Figure 13.  Two GPU scenario

- Slot 3 & 4 are available for CNA and/or RAID Controllers
- Dual wire Management with LoM ports

GRID K1/K2 or Tesla K20/K20X

GRID K1/K2 or Tesla K20/K20X

Note:  As show in the figure for two GPU scenario, we used an Emulex OneConnect OCe11102-F Gen 3 CNA card.

Follow link given below for physical configuration of GRID cards in Riser cards slot 2 and 5:
Base UCS System Configuration:
Please follow below given link for physical connectivity guide and best practices for UCS C-Series server integration with Cisco UCSM.


1. Once server is discovered in UCSM; Select server → Inventory → GPUs.

As shown the screenshot below; PCIe slot 2 and 5 are utilized with two GRID K2 cards with running firmware version 80.04.60.00.30 | 2055.0552.01.08

2. Create host firmware policy by selecting Server node on UCSM. Select Policies → Host Firmware Packages. Right click and select Create Host Firmware Package.
3. Select radio button for simple configuration of Host firmware package, select Rack package as 2.2.2C.

4. Click OK.

5. Apply this host firmware package in Service Profile Template/Service Profiles firmware policy. Once completed upgrading firmware for server GPUs are running on firmware version selected.
**Note:** Virtual Machine Hardware Version 9 or later is required for vSGA/vDGA configuration. Virtual machines with a hardware version of 9 or later can have their settings managed only via the vSphere Web Client.

**vSGA configuration:**
Download NVIDIA driver for vSphere ESXi 5.5 from link:

Extract file downloaded and upload NVIDIA-VMware_ESXi_5.5_Host_Driver_319.65-1OEM.550.0.0.1331820.vib file to a datastore on ESXi (Preferable shared storage if installing drivers on multiple servers) or using VMware Update Manager. See VMware documentation for installing patch or host extension on ESXi.

**Note:** ESXi host must be in maintenance mode to install vib module.

Using ESX cmd-line: esxcli software vib install -v /vmfs/volumes/datastore/ NVIDIA-VMware_ESXi_5.5_Host_Driver_319.65-1OEM.550.0.0.1331820.vib

Verify driver installation using cmd-line: esxcli software vib list | grep NVIDIA

```
- # esxcli software vib list | grep NVIDIA
NVIDIA-VMware ESXi 5.5 Host Driver 319.65-1OEM.550.0.0.1331820 NVIDIA VMwareתקבל 2015-10-23
```

**vDGA Configuration:**
When you deploy vDGA, it uses the graphics driver from the GPU vendor rather than the virtual machine’s vSGA 3D driver. To provide frame-buffer access, vDGA uses an interface between the remote protocol and graphics driver.

**vDGA Does Not Support Live VMware vSphere vMotion Capabilities:** Live VMware vSphere vMotion is not supported with vDGA. Bypassing the virtualization layer, vDGA uses vSphere DirectPath I/O to allow direct access to the GPU card. By enabling direct pass-through from the virtual machine to the PCI device installed on the host, the virtual machine is effectively locked to that specific host.

If a user needs to move a vDGA-enabled virtual machine to a different host, they should power off the virtual machine, use vSphere vMotion to migrate it to another host that has a GPU card installed, and re-enable pass-through to the specific PCI device on that host. Only then should the user power on the virtual machine.

**Enable Device Pass-through**
Select ESXi host → configuration → Advanced Settings → Configure Pass-through

![Image of NVIDIA GRID Cards](image.png)

Check box against NVIDIA GRID Cards.
Reboot ESXi host after making changes for pass-through configuration.

Devices enabled for pass-through configuration of dedicated graphics deployment will show as green after reboot. If it doesn’t show up as green than do take troubleshoot steps based on VMware documentation.

**Note:** vDGA Does Not Support Live VMware vSphere vMotion Capabilities:

Live VMware vSphere vMotion is not supported with vDGA. Bypassing the virtualization layer, vDGA uses vSphere DirectPath I/O to allow direct access to the GPU card. By enabling direct pass-through from the virtual machine to the PCI device installed on the host, the virtual machine is effectively locked to that specific host.

If a user needs to move a vDGA-enabled virtual machine to a different host, they should power off the virtual machine, use vSphere vMotion to migrate it to another host that has a GPU card installed, and re-enable pass-through to the specific PCI device on that host. Only then should the user power on the virtual machine.
Enable Virtual Machine for vDGA configuration:
1. Upgrade Virtual Machine Hardware Version:
   In virtual machine edit settings options select Upgrade Virtual Hardware

2. Reserve all guest memory:
   In virtual machine edit settings options select Resources tab.
   Click on check box for Reserve all guest memory.

3. Adjust pciHole.start:
Note: This is required only if the virtual machine has more than 2GB of configured memory. For virtual machines that have more than 2GB of configured memory, add the following parameter to the .vmx file of the virtual machine (you can add this at the end of the file):

```
pciHole.start = "2048"
```

4. Add PCI device:
   a. In virtual machine edit settings; select Hardware tab.

   Click on add; select PCI device. Click Next.

   ![Add PCI device screenshot]

   b. Select PCI device from the drop down list. Click Next.

   ![Select PCI device screenshot]

   Note: Only one virtual machine can be powered on if same PCI device is added to multiple VMs.
c. Click finish on ready to complete screen.
5. Install latest NVIDIA windows 7 or windows 8 desktop drivers on virtual machine. All NVIDIA drivers can be found at link given below:

**Note:** Prior installing NVIDIA drivers do windows optimization based on VMware best practices.


a. Accept license terms and agreement click next.

b. Select radio button for custom installation. Click Next.

c. Select check box for each option. Click Next.

6. Install VMware View Agent. Reboot when its prompted.

7. Enable the proprietary NVIDIA capture APIs.
   a. After the virtual machine has rebooted, enable the proprietary NVIDIA capture APIs by running: “C:\Program Files\Common Files\VMware\Teradici PCoIP Server\ MontereyEnable.exe” -enable
   
   b. After the process is complete, restart the virtual machine.
8. Check that the virtual machine is using the NVIDIA GPU and driver.

In order to activate the NVIDIA display adapter, you must connect to the virtual machine for the first time via PCoIP in full screen mode from the endpoint (at native resolution), or the virtual machine will use the Soft3D display adapter. Virtual Dedicated Graphics Acceleration (vDGA) does not work through the vSphere console session.

After the virtual machine has rebooted and you have connected via PCoIP in full screen, check to ensure that the GPU is active by viewing the display information in DXDiag.exe:

a. Click the Start menu.

b. Type dxdiag and click Enter after DxDiag shows up in the list, or click on it in the list.

c. After DxDiag launches, check the Display tab to verify that the virtual machine is using the NVIDIA GPU and driver.

Configure 3D rendering using VMware vSphere:

When configuring vSGA/vDGA via the VMware vSphere Web interface, there are four 3D rendering options:

- **Automatic** uses hardware acceleration if there is a capable and available hardware GPU in the host in which the virtual machine is running. However, if hardware GPU is not available, the Automatic option uses software 3D rendering for any 3D tasks. This allows the virtual machine to be started on, or migrated (via vSphere vMotion) to, any host (VMware vSphere version 5.0 or higher) and use the best graphics solution available on that host.

- **Software** only uses vSphere software 3D rendering, even if there is an available hardware GPU in the host in which the virtual machine is running. This option will not provide the performance benefits that hardware
3D acceleration offers. However, this configuration both allows the virtual machine to run on any host (vSphere5.0 or higher) and allows you to block virtual machines from using a hardware GPU in a host.

- **Hardware** only uses hardware-accelerated GPUs. If hardware GPU is not present in a host, the virtual machine will either not start or you will not be able to live-migrate it via vSphere vMotion to that host. As long as the host that the virtual machine is being moved to has a capable and available hardware GPU, vSphere vMotion is possible with this specification. This setting can be used to guarantee that a virtual machine will always use hardware 3D rendering when a GPU is available; but that, in turn, limits the virtual machine to host with hardware GPUs.

- **Disabled** does not use 3D rendering at all (software or hardware) and overrides vSphere 3D settings to disable 3D. Use this setting to ensure that Horizon View desktop pools with non-graphical workloads do not use unnecessary resources, like sharing hardware GPU when running on the same cluster as Horizon View desktops with heavier graphics workloads.
Configure 3D rendering using Horizon View

To configure vSGA/vDGA in Horizon View 5.3 Pool Settings, there are five 3D rendering options: Manage using vSphere Client, Automatic, Software, Hardware, and Disabled.

- **Manage using vSphere Client** will not make any change to the 3D settings of the individual virtual machines in that pool. This allows individual virtual machines to have different settings set through vSphere. This setting will most likely be used during testing or for manual desktop pools.

- **Automatic** uses hardware acceleration if there is a capable and available hardware GPU in the host in which the virtual machine is running. However, if hardware GPU is not available, it uses software 3D rendering for any 3D tasks. This allows the virtual machine to be started on, or migrated (via vSphere vMotion) to, any host (vSphere version 5.0 or higher) and use the best solution available on that host.

- **Software** only uses vSphere software 3D rendering, even if there is an available hardware GPU in the host in which the virtual machine is running. This will not provide the performance benefits that hardware 3D acceleration offers. However, it both allows the virtual machine to run on any host (vSphere 5.0 or higher) and allows you to block virtual machines from using hardware GPU in a host.

- **Hardware** only uses hardware-accelerated GPUs. If hardware GPU is not present in a host, the virtual machine will either not start or you will not be able to live-migrate it via vSphere vMotion to that host. As long as the host that the virtual machine is being moved to has a capable and available hardware GPU, vSphere vMotion is possible with this specification. This setting can be used to guarantee that a virtual machine will always use hardware 3D rendering when a GPU is available; but that, in turn, limits the virtual machine to host with hardware GPUs.

- **Disabled** does not use 3D rendering at all (software or hardware) and overrides vSphere 3D settings to disable 3D. Use this setting to ensure that Horizon View desktop pools with non-graphical workloads do not use unnecessary resources, like sharing hardware GPU when running on the same cluster as Horizon View desktops with heavier graphics workloads.

As shown in the screenshot below; For this study we selected PCoIP as default display protocol, do not allow users to choose display protocol, select Hardware from the drop down menu for 3D renderer, click on configure to select VRAM (Video Memory) size.

**Table 5.** Minimum and maximum VRAM size for both vSGA and vDGA

<table>
<thead>
<tr>
<th></th>
<th>Soft 3D (Software 3D)</th>
<th>vSGA (Hardware 3D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>1.18MB</td>
<td>64MB</td>
</tr>
<tr>
<td>Default</td>
<td>64MB</td>
<td>96MB</td>
</tr>
<tr>
<td>Maximum</td>
<td>512MB</td>
<td>512MB</td>
</tr>
</tbody>
</table>

A few key notes about vRAM setting that require attention:

1. Whenever 3D Renderer setting changes, it reverts the amount of video memory to the 96MB default. Make sure you change the video memory to the appropriate number after you change this setting.

2. VRAM settings that you configure in Horizon View Administrator take precedence over VRAM settings that are configured for the virtual machines in vSphere Client or vSphere Web Client. Select the Manage using vSphere Client option to prevent this.
3. If you are using the Manage using vSphere Client option, VMware recommends that you use the Web Client to configure the virtual machines, rather than the traditional vSphere Client. This is because the traditional vSphere Client does not display the various rendering options; it will display only Enable/Disable 3D support.

4. After making VRAM changes to the Horizon View pool, there may be a short delay (sometimes a couple of minutes) before the message “Reconfiguring virtual machine” settings appears in the vCenter console. It is important to wait for this process to complete before power cycling the virtual machines.

Screen Resolution
If you enable the 3D Renderer setting, configure the Max number of monitors setting for one or two monitors. You cannot select more than two monitors. The Max resolution of any one monitor setting is 1920 x 1200 pixels. You cannot configure this value higher.

**Note:** You must power off and on existing virtual machines for the 3D Renderer setting to take effect. Restarting or rebooting a virtual machine does not cause the setting to take effect.

**Note:** For this case study we used “Hardware” based 3D rendering using 512MB vRAM per virtual machine for vSGA, on the other side for vDGA we used “Manage using vSphere Client” and 512MB vRAM.
Performance Tuning Tips
This section offers tips to improve the performance of both vSGA and vDGA.

1. Configuring Adequate Virtual Machine Resources
   Unlike traditional VDI desktops, desktops using high-end 3D capabilities must be provisioned with more vCPUs and memory. Make sure your desktop virtual machines meet the memory and CPU requirements for the applications you use. The minimum requirements VMware recommends for 3D workloads are two vCPUs and 4GB RAM.

2. Optimizing PCoIP
   Occasionally, PCoIP custom configurations can contribute to poor performance. By default, PCoIP is set to allow a maximum of 30fps. Some applications require significantly more than that. If you notice that the frame rate of an application is lower than expected, reconfigure the PCoIP GPO to allow a maximum of 120fps.
   With PCoIP, another option is to enable Disable Build-To-Lossless. This reduces the overall amount of PCoIP traffic, which, in turn, reduces the load placed on both the virtual machine and endpoint.

3. Enabling Relative Mouse
   If you are using an application or game in which the cursor is moving uncontrollably, enabling the relative-mouse feature may improve mouse control.
   Relative Mouse is a new feature in Windows Horizon View Client that changes the way client mouse movement is tracked and sent to the server via PCoIP. Traditionally, PCoIP uses absolute coordinates. Absolute mouse events allow the client to render the cursor locally, which is a significant optimization for high-latency environments. However, not all applications work well when using the absolute mouse. Two notable classes of applications, CAD applications and 3D games, rely on relative mouse events to function correctly.
   With the introduction of vSGA and vDGA, VMware expects the requirements for relative mouse to increase rapidly as CAD and 3D games become more heavily used in Horizon View environments.
   The end user can enable Relative Mouse manually. To manually enable relative mouse, right-click the Horizon View Client Shade at the top of the screen and select Relative Mouse. A check mark appears next to Relative Mouse.

Note: The Windows Horizon View Client is required to enable relative mouse. As of now, this feature is not available through any other Horizon View Clients or zero clients. Relative Mouse must be selected on each and every connection. As of now, there is no option to enable this by default.

4. Virtual Machines Using VMXNET3
   For desktop virtual machines using VMXNET3 Ethernet adapters, you can significantly improve peak video-playback performance of your Horizon View desktop by following these steps, which are recommended by Microsoft for virtual machines:
   1. Start the Registry Editor (Regedit32.exe).
   2. Locate the following key in the registry:
      HKLM\System\CurrentControlSet\Services\Afd\Parameters
   3. In the Edit menu, click Add Value and add the following registry value:
      Value Name: FastSendDatagramThreshold
      Data Type: REG_DWORD
      Value: 1500
4. Quit the Registry Editor.

**Note:** A reboot of the desktop virtual machine is required after changing this registry setting. If this setting does not exist, create it as a DWORD value. Further information on this setting can be found on the Microsoft Support Web site.

5. **Workaround for CAD Performance Issue**

VMware has experienced a performance issue when users deploy CATIA. Occasionally, when working with CAD models (when turning and spinning), you may find that objects move irregularly and with a delay. However, the objects themselves are displayed clearly, without blurring.

The workaround in this case is to disable the MaxAppFrameRate registry entry. The registry key can be found at:

HKLM\Software\VMware, Inc.\VMware SVGA DevTap\MaxAppFrameRate

Change this registry setting to:

dword:00000000

**Note:** If this registry key does not exist, it defaults to 30.

**Important:** This change can negatively affect other applications. Use with caution and only if you are experiencing the symptoms mentioned above.

**Conclusion**

The combination of Cisco UCS Manager, UCS C240 M3 rack servers, NVIDIA GRID and Tesla cards, VMware vSphere ESXi 5.5, and VMware Horizon 6 provides a high performance platform for virtualizing graphics intensive applications.

By following this white paper, 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.

**References**

**Cisco UCS C-Series Rack-Mount Servers**


**NVIDIA references**

- [http://www.nvidia.com/content/grid/pdf/GRID_K1_BD-06633-001_v02.pdf](http://www.nvidia.com/content/grid/pdf/GRID_K1_BD-06633-001_v02.pdf)

**Emulex CNA cards**

**VMware Horizon View 5.3 Reference Documents**

- [https://www.vmware.com/support/pubs/view_pubs.html](https://www.vmware.com/support/pubs/view_pubs.html)

**View 5 with PCoIP Network Optimization Guide**


**Virtual Desktop—Windows 7 Optimization Guide:**


**VMware vSphere ESXi and vCenter Server 5 Documentation:**

- [http://www.vmware.com/pdf/Perf_Best_Practices_vSphere5.5.pdf](http://www.vmware.com/pdf/Perf_Best_Practices_vSphere5.5.pdf)