GPU Card Installation

This appendix contains the following topics:

- **Overview of Server Firmware Requirements**, page D-1
- **GPU Population Rules**, page D-2
- **Requirement For All Supported GPUs: Memory-Mapped I/O Greater than 4 GB**, page D-2
- **Installing a GPU Card**, page D-3
- **Using NVIDIA GRID License Server For M-Series GPUs**, page D-6
- **Installing Drivers to Support GPU Cards**, page D-16

### Overview of Server Firmware Requirements

Table D-1 lists the minimum server firmware versions for the supported GPU cards.

<table>
<thead>
<tr>
<th>GPU</th>
<th>Cisco IMC/BIOS Minimum Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA GRID K1</td>
<td>1.5(6)</td>
</tr>
<tr>
<td>NVIDIA GRID K2</td>
<td>1.5(6)</td>
</tr>
<tr>
<td>NVIDIA Tesla K10</td>
<td>1.5(7)</td>
</tr>
<tr>
<td>NVIDIA Tesla K20</td>
<td>1.5(7)</td>
</tr>
<tr>
<td>NVIDIA Tesla K20X</td>
<td>1.5(7)</td>
</tr>
<tr>
<td>NVIDIA Tesla K40</td>
<td>1.5(7)</td>
</tr>
<tr>
<td>NVIDIA Tesla K80</td>
<td>2.0(6)</td>
</tr>
<tr>
<td>NVIDIA Tesla M60</td>
<td>2.0(9)</td>
</tr>
<tr>
<td>NVIDIA Tesla M10</td>
<td>3.0(3)</td>
</tr>
<tr>
<td>NVIDIA Tesla P100 12GB</td>
<td>3.0(3)</td>
</tr>
<tr>
<td>NVIDIA Tesla P100 16GB</td>
<td>3.0(3)</td>
</tr>
<tr>
<td>AMD FirePro S7150 X2</td>
<td>3.0(3)</td>
</tr>
</tbody>
</table>
GPU Population Rules

- The GPU cards are double-wide and cover two slots. A GPU card in slot 2 also covers slot 1. A GPU card in slot 7 also covers slot 6.
- NVIDIA K-Series and M-Series GPUs can support only less-than 1 TB memory in the server.
- NVIDIA P-Series GPUs can support 1 TB or more memory in the server.
- AMD FirePro S7150 X2 can support only less-than 1 TB memory in the server.
- Table D-2 shows the rules for populating NVIDIA GPU cards in the server.

Table D-2 | GPU Population Rules
--- | --- | ---
**Single GPU** | **Dual GPU** | **Mix Two Different GPUs**
PCIE 2 or PCIE 7 | PCIE 2 and PCIE 7 | Not supported at this time

1. If you have a GPU card in slot 7, you cannot also have a Qlogic QLE2562 HBA card in slot 9.

**Requirement For All Supported GPUs: Memory-Mapped I/O Greater than 4 GB**

All supported GPU cards require enablement of the BIOS setting that allows greater than 4 GB of memory-mapped I/O (MMIO).

**Standalone Server**

If the server is used in standalone mode, this BIOS setting is enabled by default:

Advanced > PCI Configuration > Memory Mapped I/O Above 4 GB [Enabled]

If you need to change this from a different setting, enter the BIOS Setup Utility by pressing F2 when prompted during bootup.

**Cisco UCS Manager Controlled Server**

If the server is integrated with Cisco UCS Manager and controlled by a service profile, this setting is enabled by default in the service profile when a GPU is present.

To change this setting manually, use the following procedure.

**Step 1** Refer to the Cisco UCS Manager configuration guide (GUI or CLI) for your release for instructions on configuring service profiles:

Cisco UCS Manager Configuration Guides

**Step 2** Refer to the chapter on Configuring Server-Related Policies > Configuring BIOS Settings.

**Step 3** In the section of your profile for PCI Configuration BIOS Settings, set Memory Mapped I/O Above 4 GB Config to one of the following:

- **Disabled**—Does not map 64-bit PCI devices to 64 GB or greater address space.
- **Enabled**—Maps I/O of 64-bit PCI devices to 64 GB or greater address space.
- **Platform Default**—The policy uses the value for this attribute contained in the BIOS defaults for the server. Use this only if you know that the server BIOS is set to use the default enabled setting for this item.

**Step 4** Reboot the server.
Note
Cisco UCS Manager pushes BIOS configuration changes through a BIOS policy or default BIOS settings to the Cisco Integrated Management Controller (CIMC) buffer. These changes remain in the buffer and do not take effect until the server is rebooted.

Installing a GPU Card

Double-Wide GPU Power Cables

In the table below, the cable that is used with the GPU is listed. It is also indicated whether the cable is included in the GPU BOM or must be ordered separately.
Separate = Cable must be ordered separately when the ordering tool prompts you.
Included = Cable is included with the GPU; no additional action is needed

<table>
<thead>
<tr>
<th>Double-Wide GPU</th>
<th>GPU Power Cable</th>
<th>Cable Included When the GPU Card is Ordered With a System Order?</th>
<th>Cable Included When the GPU Card is Ordered as a Spare?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nvidia GRID K1</td>
<td>UCSC-AUXCBL8-EX</td>
<td>Included</td>
<td>Separate</td>
</tr>
<tr>
<td>Nvidia GRID K2</td>
<td>UCSC-AUXCBL8-EX</td>
<td>Included</td>
<td>Separate</td>
</tr>
<tr>
<td>Nvidia Tesla K10</td>
<td>UCSC-AUXCBL8-EX</td>
<td>Included</td>
<td>Separate</td>
</tr>
<tr>
<td>Nvidia Tesla K20</td>
<td>UCSC-AUXCBL8-EX</td>
<td>Included</td>
<td>Separate</td>
</tr>
<tr>
<td>Nvidia Tesla K20X</td>
<td>UCSC-AUXCBL8-EX</td>
<td>Included</td>
<td>Separate</td>
</tr>
<tr>
<td>Nvidia Tesla K40</td>
<td>UCSC-AUXCBL8-EX</td>
<td>Included</td>
<td>Separate</td>
</tr>
<tr>
<td>Nvidia Tesla K80</td>
<td>UCSC-300W-460M4</td>
<td>Included</td>
<td>Separate</td>
</tr>
<tr>
<td>Nvidia Tesla M60</td>
<td>UCSC-300W-460M4</td>
<td>Included</td>
<td>Separate</td>
</tr>
<tr>
<td>Nvidia Tesla M10</td>
<td>UCSC-AUXCBL8-EX</td>
<td>Included</td>
<td>Separate</td>
</tr>
<tr>
<td>Nvidia Tesla P100 12GB</td>
<td>UCSC-300W-460M4</td>
<td>Included</td>
<td>Separate</td>
</tr>
<tr>
<td>Nvidia Tesla P100 16GB</td>
<td>UCSC-300W-460M4</td>
<td>Included</td>
<td>Separate</td>
</tr>
<tr>
<td>AMD FirePro S7150 X2</td>
<td>UCSC-300W-460AMD</td>
<td>Included</td>
<td>Separate</td>
</tr>
</tbody>
</table>

Procedure

Step 1
Remove a GPU card (or a blank filler panel) from the PCIe riser assembly:

a. Shut down and power off the server as described in Shutting Down and Powering Off the Server, page 3-8.

b. Slide the server out the front of the rack far enough so that you can remove the top cover.

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

c. Remove the top cover as described in Removing or Replacing the Server Top Cover, page 3-10.
Appendix D      GPU Card Installation

Step 2
Install a GPU card:

a. Connect the white connector of the included power cable to the white power-connector on the riser. The power connector is adjacent to PCIE 2 on riser 1 or PCIE 6 on riser 2 (see Figure 3-22).

For a list of cables that go with supported double-wide GPUs, see Double-Wide GPU Power Cables, page D-3.

Caution
Do not reverse the GPU power cable. Connect the black connector on the cable to the black connector on the GPU card. Connect the white connector on the cable to the white connector on the PCIe riser. Never connect a 6-pin cable connector to an 8-slot receptacle.

Note
Your GPU card might be shipped with two power cables: a straight cable and a Y-cable. The straight cable is used for connecting power to the GPU card in this server; do not use the Y-cable, which is used for connecting the GPU card in external devices only.

Note
If your cable has a 10-pin connector, the white 10-pin connector on the cable connects to a white 8-pin socket on the riser. The cable connector is keyed so that when you insert it, only 8 pins are used. Two pins of the cable connector sit outside the riser socket (see Figure D-1.)

Figure D-1 High-Power GPU Power Cable (UCSC-300W-460M4) 10-Pin Connector

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>White 10-pin cable connector</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Two pins on cable connector that are outside the socket on riser (not used)</td>
<td></td>
</tr>
</tbody>
</table>

b. Connect the black connector of the power cable to the black connector on the GPU card.

c. Align the new GPU card with the empty socket on the PCIe riser assembly (PCIE 2 on riser 1 or PCIE 7 on riser 2).
d. Push down evenly on both ends of the card until it is fully seated in the socket.  
   Ensure that the card’s rear-panel tab sits flat against the PCIe riser rear-panel opening.

e. Close the hinged card-tab retainer and press it down until it clicks and locks in place.

f. Close the hinged blue plastic card cover.

g. Align the riser so that its connector is over the motherboard socket.

h. With the blue plastic retaining latch fully open (vertical), lower the riser into the chassis alignment channels until its connector makes contact with the motherboard socket.

i. Close the retaining latch until it is flat to fully engage the riser with the motherboard socket.

j. Replace the top cover.

k. Replace the server in the rack, replace cables, and then power on the server by pressing the Power button.

Step 3  If you installed a Tesla M60 or M10 GPU, continue with Using NVIDIA GRID License Server For M-Series GPUs, page D-6.
Using NVIDIA GRID License Server For M-Series GPUs

Installation Overview

This section applies to the following GPUs:

- NVIDIA Tesla M10
- NVIDIA Tesla M60

Use the topics in this appendix in the following order:

1. Install the hardware.
   - Installing a GPU Card, page D-3
2. Register your product activation keys with NVIDIA.
   - NVIDIA GRID License Server Overview, page D-7
   - Registering Your Product Activation Keys With NVIDIA, page D-8
3. Download the GRID software suite.
   - Downloading the GRID Software Suite, page D-8
4. Install the GRID License Server software to a host.
   - Installing NVIDIA GRID License Server Software, page D-9
5. Generate licenses on the NVIDIA Licensing Portal and download them.
   - Installing GRID Licenses From the NVIDIA Licensing Portal to the License Server, page D-11
6. Manage your GRID licenses.
   - Managing GRID Licenses, page D-12
7. Decide whether to use the GPU in compute mode or graphics mode.
   - Switching Between Compute Mode and Graphics Mode, page D-14
NVIDIA GRID License Server Overview

The NVIDIA Tesla M-Series GPU combines Tesla and GRID functionality when the licensed GRID features such as GRID vGPU and GRID Virtual Workstation are enabled. These features are enabled during OS boot by borrowing a software license that is served over the network from the NVIDIA GRID License Server virtual appliance. The license is returned to the license server when the OS shuts down.

You obtain the licenses that are served by the GRID License Server from NVIDIA’s Licensing Portal as downloadable license files, which you install into the GRID License Server via its management interface (see Figure D-2).

![Figure D-2 GRID Licensing Architecture](image)

There are three editions of GRID licenses, which enable three different classes of GRID features. The GRID software automatically selects the license edition based on the features that you are using (see Table D-3).

<table>
<thead>
<tr>
<th>GRID License Edition</th>
<th>GRID Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA GRID Virtual Applications</td>
<td>• Designed for PC-level applications and server-based desktops.</td>
</tr>
<tr>
<td>GRID Virtual PC</td>
<td>• Designed for virtual desktops that can use PC Windows applications, browsers, and high-definition video.</td>
</tr>
<tr>
<td>GRID Virtual Workstation</td>
<td>• Designed for remote, professional graphics applications with full performance on any device.</td>
</tr>
</tbody>
</table>
Registering Your Product Activation Keys With NVIDIA

After your order is processed, NVIDIA sends you a Welcome email that contains your product activation keys (PAKs) and a list of the types and quantities of licenses that you purchased.

Step 1  Select the Log In link, or the Register link if you do not already have an account. The NVIDIA Software Licensing Center > License Key Registration dialog opens.

Step 2  Complete the License Key Registration form and then click Submit My Registration Information. The NVIDIA Software Licensing Center > Product Information Software dialog opens.

Step 3  If you have additional PAKs, click Register Additional Keys. For each additional key, complete the form on the License Key Registration dialog and then click Submit My Registration Information.

Step 4  Agree to the terms and conditions and set a password when prompted.

Downloading the GRID Software Suite

Step 1  Return to the NVIDIA Software Licensing Center > Product Information Software dialog.

Step 2  Click the Current Releases tab.

Step 3  Click the NVIDIA GRID link to access the Product Download dialog. This dialog includes download links for:
- NVIDIA License Manager software
- The gpumodeswitch utility
- The host driver software

Step 4  Use the links to download the software.
Installing NVIDIA GRID License Server Software

For full installation instructions and troubleshooting, refer to the NVIDIA GRID License Server User Guide. Also refer to the NVIDIA GRID License Server Release Notes for the latest information about your release.

http://www.nvidia.com

Platform Requirements for NVIDIA GRID License Server

- The hosting platform can be a physical or a virtual machine. NVIDIA recommends using a host that is dedicated only to running the License Server.
- The hosting platform must run a supported Windows OS.
- The hosting platform must have a constant IP address.
- The hosting platform must have at least one constant Ethernet MAC address.
- The hosting platform’s date and time must be set accurately.

Installing on Windows

The License Server requires a Java runtime environment and an Apache Tomcat installation. Apache Tomcat is installed when you use the NVIDIA installation wizard for Windows.

Step 1

Note: Install the 32-bit Java Runtime Environment, regardless of whether your platform is Windows 32-bit or 64-bit.

Step 2
Create a server interface:
- On the NVIDIA Software Licensing Center dialog, click Grid Licensing > Create License Server.
- On the Create Server dialog, fill in your desired server details.
- Save the .bin file that is generated onto your license server for installation.

Step 3
Unzip the NVIDIA License Server installer Zip file that you downloaded previously and run setup.exe.

Step 4
Accept the EULA for the NVIDIA License Server software and the Apache Tomcat software. Tomcat is installed automatically during the License Server installation.

Step 5
Use the installer wizard to step through the installation.

Note: On the Choose Firewall Options dialog, select the ports to be opened in the firewall. NVIDIA recommends that you use the default setting, which opens port 7070 but leaves port 8080 closed.

Step 6
Verify the installation. Open a web browser on the License Server host and connect to the URL http://localhost:8080/licserver. If the installation was successful, you see the NVIDIA License Client Manager interface.
Installing on Linux

The License Server requires a Java runtime environment and an Apache Tomcat installation. You must install both separately before installing the License Server on Linux.

**Step 1**
Verify that Java was installed with your Linux installation. Use the following command:

`java -version`

If no Java version is displayed, use your Linux package manager to install with the following command:

`sudo yum install java`

**Step 2**
Use your Linux package manager to install the tomcat and tomcat-webapps packages.

a. Use the following command to install Tomcat:

`sudo yum install java`

b. Enable the Tomcat service for automatic startup on boot:

`sudo systemctl enable tomcat.service`

c. Start the Tomcat service:

`sudo systemctl start tomcat.service`

d. Verify that the Tomcat service is operational. Open a web browser on the License Server host and connect to the URL http://localhost:8080. If the installation was successful, you see the Tomcat webapp.

**Step 3**
Install the License Server:

a. Unpack the License Server tar file using the following command:

`tar xzf NVIDIA-linux-2015.09-0001.tgz`

b. Run the unpacked setup binary as root:

`sudo ./setup.bin`

c. Accept the EULA and then continue with the installation wizard to finish the installation.

**Note**
On the Choose Firewall Options dialog, select the ports to be opened in the firewall. NVIDIA recommends that you use the default setting, which opens port 7070 but leaves port 8080 closed.

**Step 4**
Verify the installation. Open a web browser on the License Server host and connect to the URL http://localhost:8080/licserver. If the installation was successful, you see the NVIDIA License Client Manager interface.
Installing GRID Licenses From the NVIDIA Licensing Portal to the License Server

Accessing the GRID License Server Management Interface

Open a web browser on the License Server host and access the URL http://localhost:8080/licserver.

If you configured the License Server host’s firewall to permit remote access to the License Server, the management interface is accessible from remote machines at the URL http://hostname:8080/licserver.

Reading Your License Server’s MAC Address

Your License Server’s Ethernet MAC address is used as an identifier when registering the License Server with NVIDIA’s Licensing Portal.

Step 1
Access the GRID License Server Management Interface in a browser. See Accessing the GRID License Server Management Interface, page D-11.

Step 2
In the left-side License Server panel, select Configuration. The License Server Configuration panel opens. Next to Server host ID, a pull-down menu lists the possible Ethernet MAC addresses.

Step 3
Select your License Server’s MAC address from the Server host ID pull-down.

Note
It is important to use the same Ethernet ID consistently to identify the server when generating licenses on NVIDIA’s Licensing Portal. NVIDIA recommends that you select one entry for a primary, non-removable Ethernet interface on the platform.

Installing Licenses From the Licensing Portal

Step 1
Access the GRID License Server Management Interface in a browser. See Accessing the GRID License Server Management Interface, page D-11.

Step 2
In the left-side License Server panel, select Configuration. The License Server Configuration panel opens.

Step 3
Use the License Server Configuration menu to install the .bin file that you generated earlier.

a. Click Choose File.

b. Browse to the license .bin file that you want to install and click Open.

c. Click Upload.

The license file is installed on your License Server. When installation is complete, you see the confirmation message, “Successfully applied license file to license server.”
Viewing Available Licenses

Use the following procedure to view which licenses are installed and available, along with their properties.

**Step 1**  Access the GRID License Server Management Interface in a browser. See Accessing the GRID License Server Management Interface, page D-11.

**Step 2**  In the left-side License Server panel, select Licensed Feature Usage.

**Step 3**  Click on a feature in the Feature column to see detailed information about the current usage of that feature.

Viewing Current License Usage

Use the following procedure to view information about which licenses are currently in-use and borrowed from the server.

**Step 1**  Access the GRID License Server Management Interface in a browser. See Accessing the GRID License Server Management Interface, page D-11.

**Step 2**  In the left-side License Server panel, select Licensed Clients.

**Step 3**  To view detailed information about a single licensed client, click on its Client ID in the list.

Managing GRID Licenses

Features that require GRID licensing run at reduced capability until a GRID license is acquired.

Acquiring a GRID License on Windows

To acquire a GRID license on Windows, use the following procedure.

**Step 1**  Open the NVIDIA Control Panel using one of the following methods:
- Right-click on the Windows desktop and select NVIDIA Control Panel from the menu.
- Open Windows Control Panel and double-click the NVIDIA Control Panel icon.

**Step 2**  In the NVIDIA Control Panel left-pane under Licensing, select Manage License.

The Manage License task pane opens and shows the current license edition being used. The GRID software automatically selects the license edition based on the features that you are using. The default is Tesla (unlicensed).

**Step 3**  If you want to acquire a license for GRID Virtual Workstation, under License Edition, select GRID Virtual Workstation.

**Step 4**  In the License Server field, enter the address of your local GRID License Server.

The address can be a domain name or an IP address.

**Step 5**  In the Port Number field, enter your port number of leave it set to the default used by the server, which is 7070.

**Step 6**  Select Apply.
The system requests the appropriate license edition from your configured License Server. After a license is successfully acquired, the features of that license edition are enabled.


**Note**  
After you configure licensing settings in the NVIDIA Control Panel, the settings persist across reboots.

### Acquiring a GRID License on Linux

To acquire a GRID license on Linux, use the following procedure.

**Step 1**  
Edit the configuration file /etc/nvidia/gridd.conf:  

```bash  
sudo vi /etc/nvidia/gridd.conf  
```

**Step 2**  
Edit the `ServerUrl` line with the address of your local GRID License Server.  
The address can be a domain name or an IP address. See the example file below.

**Step 3**  
Append the port number (default 7070) to the end of the address with a colon. See the example file below.

**Step 4**  
Edit the `FeatureType` line with the integer for the license type. See the example file below.

- GRID vGPU = 1
- GRID Virtual Workstation = 2

**Step 5**  
Restart the nvidia-gridd service.  

```bash  
sudo service nvidia-gridd restart  
```

The service automatically acquires the license edition that you specified in the `FeatureType` line. You can confirm this in `/var/log/messages`.


**Note**  
After you configure licensing settings in `gridd.conf`, the settings persist across reboots.

**Sample configuration file:**

```bash  
# /etc/nvidia/gridd.conf - Configuration file for NVIDIA Grid Daemon  
# Description: Set License Server URL  
# Data type: string  
# Format: "<address>:<port>"  
ServerUrl=10.31.20.45:7070

# Description: Set Feature to be enabled  
# Data type: integer  
# Possible values:  
# 1 => for GRID vGPU  
# 2 => for GRID Virtual Workstation  
FeatureType=1
```
Switching Between Compute Mode and Graphics Mode

Overview of GPU Modes

The NVIDIA Tesla M60 GPU is shipped in compute mode, which is optimized for high-performance compute (HPC) applications. However, while compute mode is best for HPC usage, it can cause compatibility issues with OS and hypervisors if you use the GPU primarily as a graphics device.

The mode is determined at power-on, from settings stored in the GPU’s non-volatile memory. You can use the command-line tool `gpumodeswitch` to toggle the GPU between compute mode and graphics mode. Table D-4 and Table D-5 compare the compute mode and graphic mode default settings.

### Table D-4 Compute Mode Default Settings

<table>
<thead>
<tr>
<th>Setting</th>
<th>Value</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classcode</td>
<td>3D Controller</td>
<td>This classcode tells the OS that the GPU is not intended for use as the primary display device.</td>
</tr>
<tr>
<td>Memory base address register (BAR)</td>
<td>8 GB</td>
<td>A large BAR is exposed for direct access to the frame buffer from the CPU and PCIe devices.</td>
</tr>
<tr>
<td>I/O base BAR</td>
<td>Disabled</td>
<td>The GPU does not consume any legacy I/O resources when used as a non-display device.</td>
</tr>
<tr>
<td>Error-correcting code (ECC) protection</td>
<td>Enabled</td>
<td>ECC is enabled on the GPU frame buffer to protect against single- and multi-bit memory errors.</td>
</tr>
</tbody>
</table>

### Table D-5 Graphic Mode Default Settings

<table>
<thead>
<tr>
<th>Setting</th>
<th>Value</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classcode</td>
<td>VGA Controller</td>
<td>This classcode tells the OS that the GPU can function as the primary display device.</td>
</tr>
<tr>
<td>Memory base address register (BAR)</td>
<td>256 MB</td>
<td>A smaller BAR is exposed for direct access to the frame buffer from the CPU and PCIe devices.</td>
</tr>
<tr>
<td>I/O base BAR</td>
<td>Enabled</td>
<td>The GPU exposes an I/O BAR to claim the resources required to operate as a VGA controller.</td>
</tr>
<tr>
<td>Error-correcting code (ECC) protection</td>
<td>Disabled</td>
<td>ECC protection is disabled.</td>
</tr>
</tbody>
</table>
Using gpumodeswitch

The command line utility gpumodeswitch can be run in the following environments:

- Windows 64-bit command prompt (requires administrator permissions)
- Linux 32/64-bit shell (including Citrix XenServer dom0) (requires root permissions)

**Note**
Consult NVIDIA product release notes for the latest information on compatibility with compute and graphic modes.

The gpumodeswitch utility supports the following commands:

- **--listgpumodes**
  This command writes information to a log file named listgpumodes.txt in the current working directory.

- **--gpumode graphics**
  Switches to graphics mode. Switches mode of all supported GPUs in the server unless you specify otherwise when prompted.

- **--gpumode compute**
  Switches to compute mode. Switches mode of all supported GPUs in the server unless you specify otherwise when prompted.

**Note**
After you switch GPU mode, reboot the server to ensure that the modified resources of the GPU are correctly accounted for by any OS or hypervisor running on the server.
Installing Drivers to Support GPU Cards

After you install the hardware, you must update to the correct level of server BIOS and then install GPU drivers and other software in this order:

- 1. Updating the Server BIOS, page D-16
- 2. Updating the GPU Drivers, page D-16

1. Updating the Server BIOS

Install the latest server BIOS by using the Host Upgrade Utility for the Cisco UCS C460 M4 server.

Note: You must do this procedure before you update the GPU drivers.

Step 1 Navigate to the following URL: http://www.cisco.com/cisco/software/navigator.html.
Step 2 Click Servers–Unified Computing in the middle column.
Step 3 Click Cisco UCS C-Series Rack-Mount Standalone Server Software in the right-hand column.
Step 4 Click the name of your model of server in the right-hand column.
Step 5 Click Unified Computing System (UCS) Server Firmware.
Step 6 Click the release number.
Step 7 Click Download Now to download the ucs-server platform-huu-version_number.iso file.
Step 8 Verify the information on the next page, and then click Proceed With Download.
Step 9 Continue through the subsequent screens to accept the license agreement and browse to a location where you want to save the file.
Step 10 Use the Host Upgrade Utility to update the server BIOS.
The user guides for the Host Upgrade Utility are at Utility User Guides.

2. Updating the GPU Drivers

After you update the server BIOS, you can install GPU drivers to your hypervisor virtual machine.

Step 1 Install your hypervisor software on a computer. Refer to your hypervisor documentation for the installation instructions.
Step 2 Create a virtual machine in your hypervisor. Refer to your hypervisor documentation for instructions.
Step 3 Install the GPU drivers to the virtual machine. Download the drivers from either:
   - AMD: http://support.amd.com/en-us/download
Step 4 Restart the server.
Step 5  Check that the virtual machine is able to recognize the GPU card. In Windows, use the Device Manager and look under Display Adapters.