GPU Installation

This appendix contains configuration rules and installation procedures for the supported GPU cards.

Server Firmware Requirements

The following table lists the minimum server firmware versions for the supported GPU cards.

GPU Card	Cisco IMC/BIOS Minimum Version Required
NVIDIA Tesla A10 24GB	4.2(1)
NVIDIA Tesla A100 40GB	4.2(1)
Nvidia A16 PCIe FHFL	4.2(1g)

GPU Card Configuration Rules

Note the following rules when populating a server with GPU cards.

Caution When using NVIDIA Tesla A10 or A100 GPU cards in this server, there are special temperature requirements. See Installing a Double-Wide GPU Card, on page 4.

- Use the HX power calculator at the following link to determine the power needed based on your server configuration: http://ucspowercalc.cisco.com
- Up to two double-wide GPU cards are supported in PCIe riser 1, slot 2 and in PCIe riser 2, slot 5.

Note Double-wide GPU cards are not supported in all PCIe riser options. Double-wide GPU cards are supported only in the following riser options:

- PCIe riser 1 with Riser 1A (HX-RIS1A-240M6)
- PCIe riser 2 with Riser 2A (HX-RIS2A-240M6)
- PCIe riser 3 with Riser 3A (HX-RIS3A-240M6) or Riser 3C (HX-RIS3C-240M6)

Note Only with slot 7.

- A double-wide GPU card installed in slot 2 also covers slot 4; a double wide GPU card installed in slot 5 also covers slot 6.
- Do not mix different brands or models of GPU cards in the server.
- You can install a GPU card and a Cisco HX VIC in the same riser. When you install a GPU card in slot 2, NCSI support in riser 1 automatically moves to slot 1. When you install a GPU card in slot 5, NCSI support in riser 2 automatically moves to slot 4.
- Some GPUs have a limitation on whether they can support 1 TB or more memory in the server.

Requirement For All 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).

In standalone mode, BIOS setting is enabled by default:

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

Step 1 If you need to change this setting, enter the BIOS Setup Utility by pressing **F2** when prompted during bootup.

Step 2 Browse to **Advanced** > **PCI Configuration** > **Memory Mapped I/O Above 4 GB**.

Installing a Single-Wide GPU Card

Use the following procedure to install or replace the following supported single-wide GPU cards:

- NVIDIA Tesla A10 24GB

Step 1 Shut down and remove power from the server as described in Shutting Down and Removing Power From the Server.

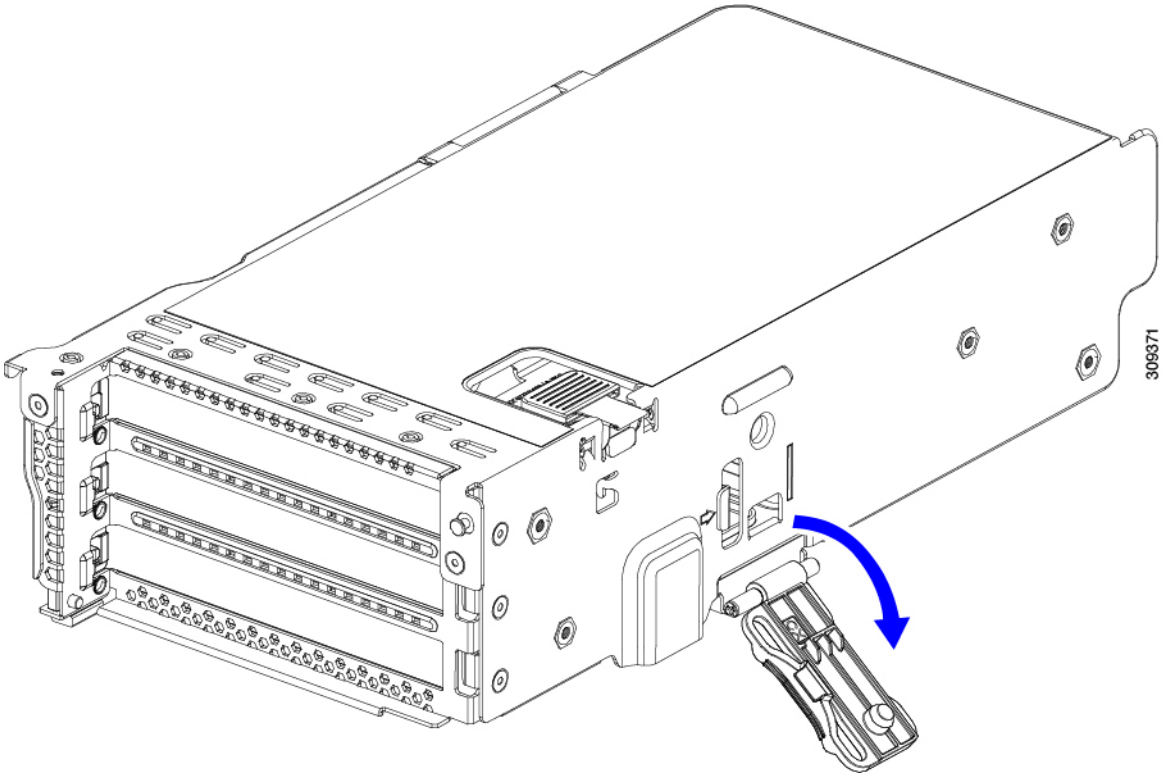
Step 2 Slide the server out the front of the rack far enough so that you can remove the top cover. You might have to detach cables from the rear panel to provide clearance.

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

Step 3 Remove the top cover from the server as described in Removing the Server Top Cover.

Step 4 Remove the single-wide GPU card that you are replacing:

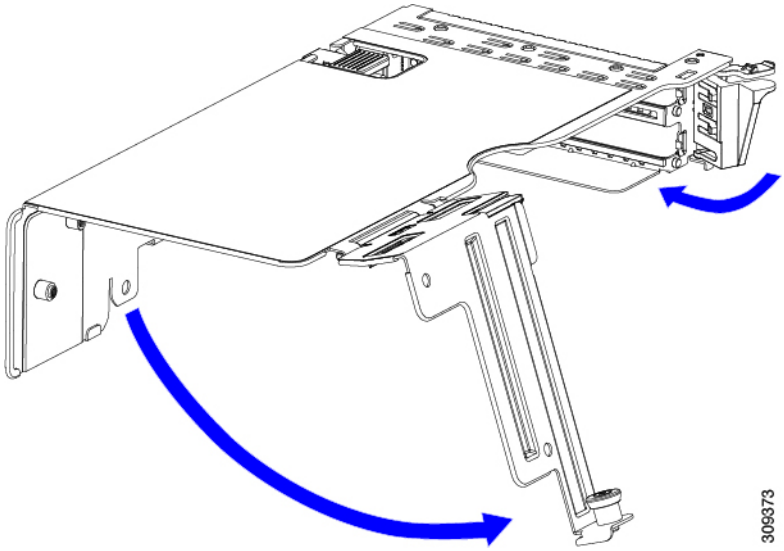
a) Use two hands to flip up and grasp the blue riser handle and the blue finger grip area on the front edge of the riser, and then lift straight up.



b) On the bottom of the riser, push the release latch that holds the securing plate, and then swing the hinged securing plate open.

c) Open the hinged card-tab retainer that secures the rear-panel tab of the card.

Figure 1: PCIe Riser Card Securing Mechanisms



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

d) Pull evenly on both ends of the single-wide GPU card to remove it from the socket on the PCIe riser.

If the riser has no card, remove the blanking panel from the rear opening of the riser.

Step 5 Install a new single-wide GPU card:

a) With the hinged card-tab retainer open, align the new single-wide GPU card with the empty socket on the PCIe riser.
b) Push down evenly on both ends of the card until it is fully seated in the socket.
c) Ensure that the card's rear panel tab sits flat against the riser rear-panel opening and then close the hinged card-tab retainer over the card's rear-panel tab.
d) Swing the hinged securing plate closed on the bottom of the riser. Ensure that the clip on the plate clicks into the locked position.
e) Position the PCIe riser over its socket on the motherboard and over the chassis alignment channels.
f) Carefully push down on both ends of the PCIe riser to fully engage its connector with the sockets on the motherboard.

Step 6 Replace the top cover to the server.

Step 7 Replace the server in the rack, replace cables, and then fully power on the server by pressing the Power button.

Step 8 Optional: Continue with Installing Drivers to Support the GPU Cards, on page 15.

Note If you installed an NVIDIA Tesla M-series or P-Series GPU, you must install GRID licenses to use the GRID features. See Using NVIDIA GRID License Server For P-Series and T-Series GPUs, on page 8.

Installing a Double-Wide GPU Card

Use the following procedure to install or replace the following supported double-wide GPU card:

- NVIDIA Tesla A100 40GB

Note When using NVIDIA Tesla A10 or A100 GPU cards in this server, there are special temperature requirements, as described in the table below.

Table 1: Cisco HyperFlex Operating Temperature Requirements For Double-Wide GPU Cards

GPU Card	Maximum Server Operating Temperature (Air Inlet Temperature)
NVIDIA Tesla A10	30° C (86.0° F)
NVIDIA Tesla A100	30° C (86.0° F)

Note **For NVIDIA GPUs:** The NVIDIA 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 (such as the Magma chassis).

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 2: Double-Wide GPU Required Power Cables

GPU	GPU Power Cable	Cable Included When the GPU Card is Ordered With a System Order?	Cable Included When the GPU Card is Ordered as a Spare?
NVIDIA Tesla A10 24GB	HX-P100CBL-240M5	Included	Separate
NVIDIA Tesla A100 40GB	HX-P100CBL-240M5	Included	Separate

Step 1 Shut down and remove power from the server as described in Shutting Down Using the Power Button.

Step 2 Slide the server out the front of the rack far enough so that you can remove the top cover. You might have to detach cables from the rear panel to provide clearance.

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

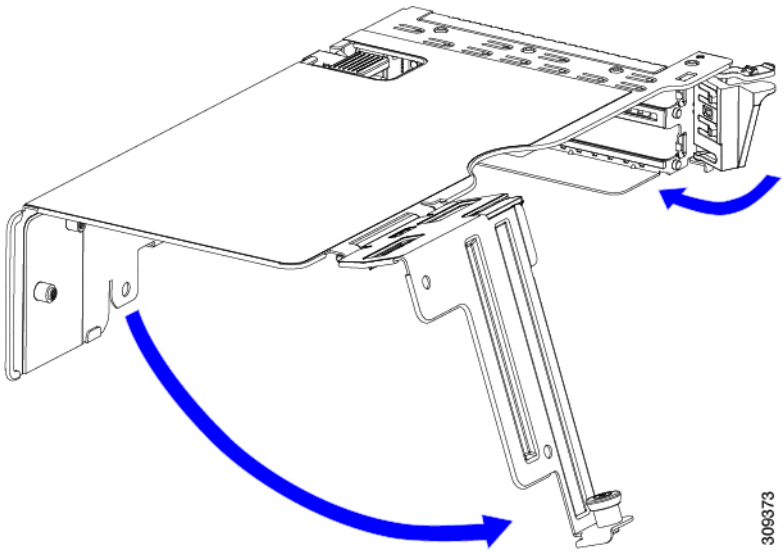
Step 3 Remove the top cover from the server as described in Removing the Server Top Cover.

Step 4 Remove an existing GPU card:

a) Use two hands to grasp the metal bracket of the PCIe riser and lift straight up to disengage its connector from the socket on the motherboard. Set the riser on an antistatic surface.

b) On the bottom of the riser, press down on the clip that holds the securing plate.
c) Swing open the hinged securing plate to provide access.
d) Open the hinged plastic retainer that secures the rear-panel tab of the card.
e) Disconnect the GPU card's power cable from the power connector on the PCIe riser.
f) Pull evenly on both ends of the GPU card to remove it from the socket on the PCIe riser.

Figure 2: PCIe Riser Card Securing Mechanisms



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

Step 5 Install a new GPU card:

Note Observe the configuration rules for this server, as described in GPU Card Configuration Rules, on page 1.

a) Align the GPU card with the socket on the riser, and then gently push the card's edge connector into the socket. Press evenly on both corners of the card to avoid damaging the connector.
b) Connect the GPU power cable. The straight power cable connectors are color-coded. Connect the cable's black connector into the black connector on the GPU card and the cable's white connector into the white GPU POWER connector on the PCIe riser.

Caution Do not reverse the straight power cable. Connect the *black* connector on the cable to the *black* connector on the GPU card. Connect the *white* connector on the cable to the *white* connector on the PCIe riser.

c) Close the card-tab retainer over the end of the card.
d) Swing the hinged securing plate closed on the bottom of the riser. Ensure that the clip on the plate clicks into the locked position.
e) Position the PCIe riser over its socket on the motherboard and over the chassis alignment channels.
f) Carefully push down on both ends of the PCIe riser to fully engage its connector with the sockets on the motherboard.

At the same time, align the GPU front support bracket (on the front end of the GPU card) with the securing latch that is on the server's air baffle.

Step 6 Insert the GPU front support bracket into the latch that is on the air baffle:

a) Pinch the latch release tab and hinge the latch toward the front of the server.
b) Hinge the latch back down so that its lip closes over the edge of the GPU front support bracket.
c) Ensure that the latch release tab clicks and locks the latch in place.

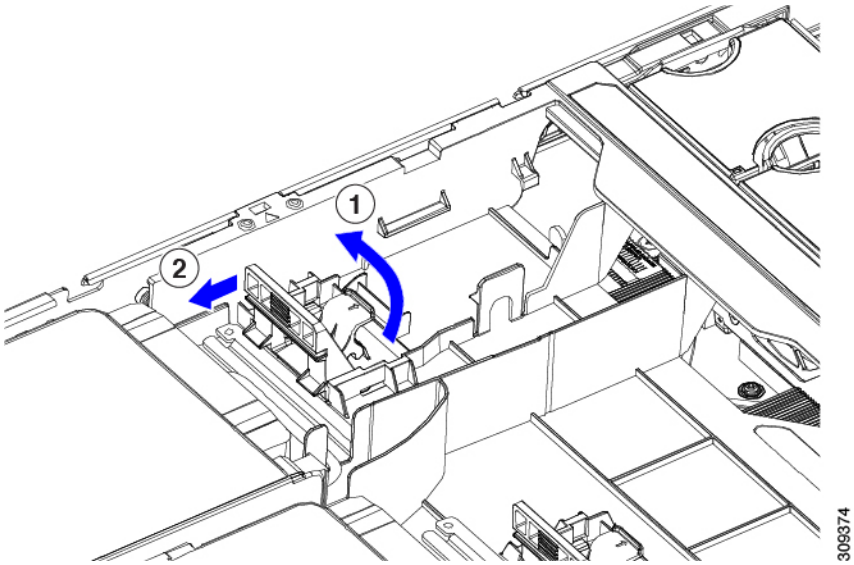
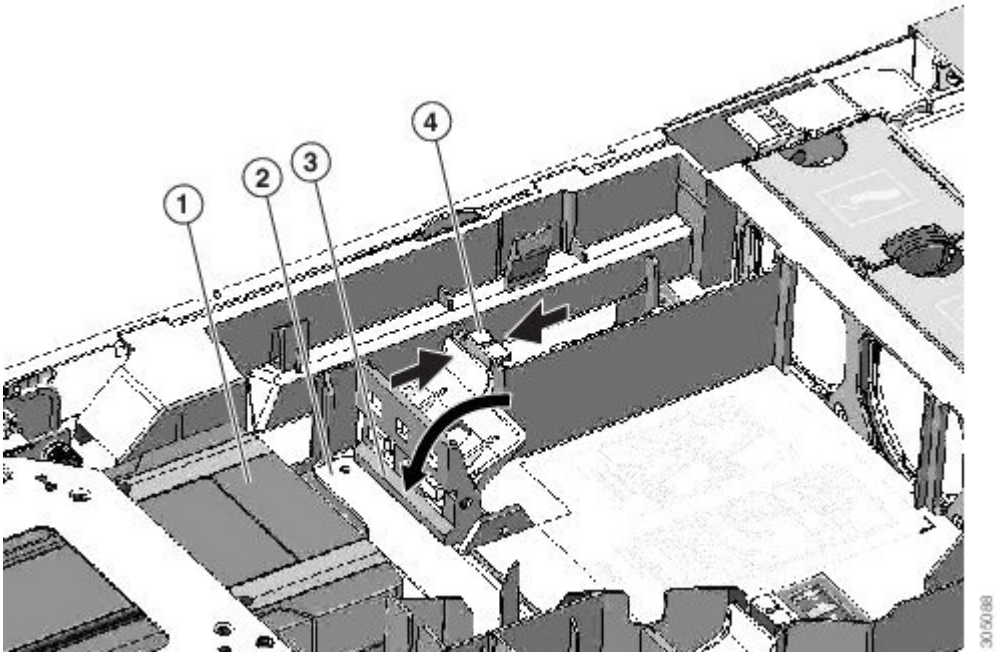


Figure 3: GPU Front Support Bracket Inserted to Securing latch on Air Baffle



1	Front end of GPU card	**3**	Lip on securing latch
2	GPU front support bracket	**4**	Securing latch release tab

Step 7 Replace the top cover to the server.

Step 8 Replace the server in the rack, replace cables, and then fully power on the server by pressing the Power button.

Step 9 Optional: Continue with Installing Drivers to Support the GPU Cards, on page 15.

Note If you installed an NVIDIA Tesla M-series GPU, you must install GRID licenses to use the GRID features. See Using NVIDIA GRID License Server For P-Series and T-Series GPUs, on page 8.

Using NVIDIA GRID License Server For P-Series and T-Series GPUs

This section applies to NVIDIA Tesla P-Series and T-Series GPUs.

Use the topics in this section in the following order when obtaining and using NVIDIA GRID licenses.

1. Familiarize yourself with the NVIDIA GRID License Server.

   NVIDIA GRID License Server Overview, on page 8

2. Register your product activation keys with NVIDIA.

   Registering Your Product Activation Keys With NVIDIA, on page 9

3. Download the GRID software suite.

   Downloading the GRID Software Suite, on page 10

4. Install the GRID License Server software to a host.

   Installing GRID Licenses From the NVIDIA Licensing Portal to the License Server, on page 12

5. Generate licenses on the NVIDIA Licensing Portal and download them.

   Installing Licenses From the Licensing Portal, on page 12

6. Manage your GRID licenses.

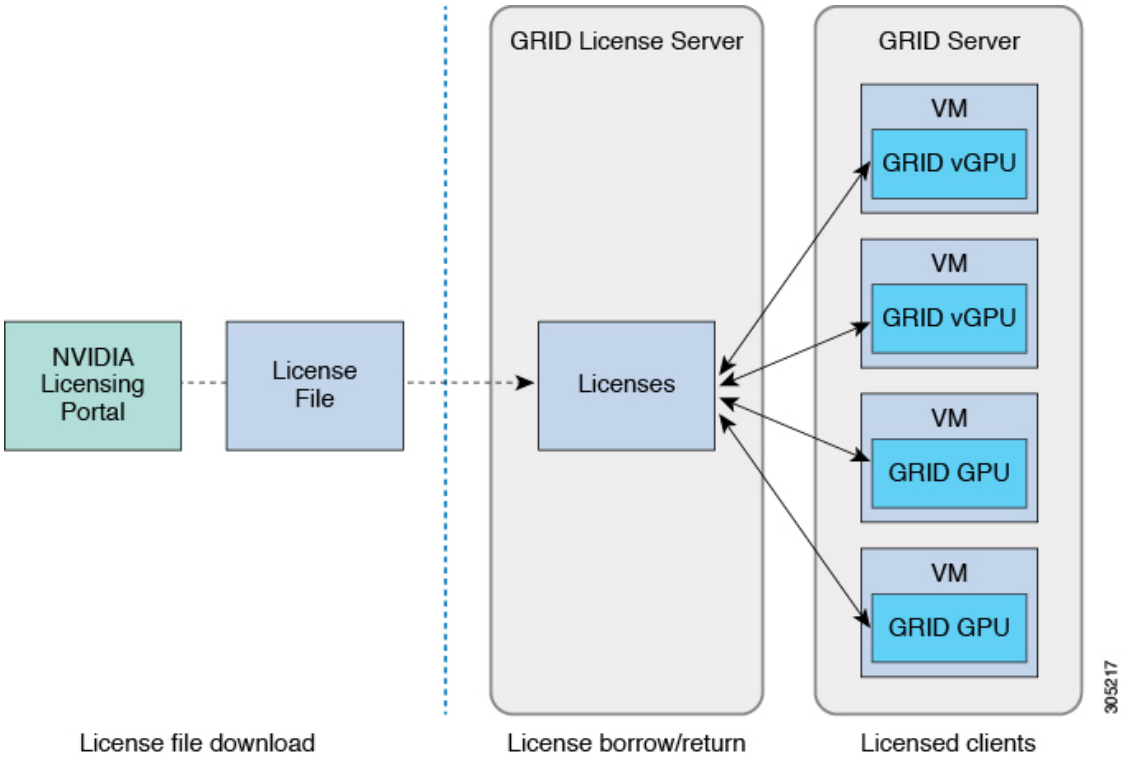
   Managing GRID Licenses , on page 13

NVIDIA GRID License Server Overview

The NVIDIA M-Series GPUs combine 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.

Figure 4: NVIDIA GRID Licensing Architecture



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.

GRID License Edition	GRID Feature
GRID Virtual GPU (vGPU)	Virtual GPUs for business desktop computing
GRID Virtual Workstation	Virtual GPUs for midrange workstation computing
GRID Virtual Workstation – Extended	Virtual GPUs for high-end workstation computing Workstation graphics on GPU pass-through

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 GRID License Server 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 Download and install the latest Java 32-bit runtime environment from https://www.oracle.com/downloads/index.html.

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:

a) On the NVIDIA Software Licensing Center dialog, click **Grid Licensing > Create License Server**.
b) On the Create Server dialog, fill in your desired server details.
c) 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 GRID License Server 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 tomcat
```

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 xfz 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.

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.

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 GRID 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.

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.

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

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

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

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

Step 2 Edit the Server URL 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.

```
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 the NVIDIA Control Panel, the settings persist across reboots.

Sample configuration file:

```
# /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=2
```

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`

  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 the 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. Update the server BIOS.
2. Update the GPU drivers.

1. Updating the Server BIOS

Install the latest BIOS by using the Host Upgrade Utility.

Note You must do this procedure before you update the NVIDIA 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 Card 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:

- NVIDIA Enterprise Portal for GRID hypervisor downloads (requires NVIDIA login): https://nvidia.flexnetoperations.com/
- NVIDIA public driver area: http://www.nvidia.com/Download/index.aspx
- 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.