APPENDIX D

GPU Card Installation

This appendix contains the following topics:

- GPU Card Configuration Rules, page D-1
- Requirement For All Supported GPUs: Memory-Mapped I/O Greater than 4 GB, page D-2
- Installing a GRID K1 or K2 GPU Card, page D-3
- Installing a Tesla M60 GPU Card and 300 W GPU Conversion Kit, page D-5
- Installing Drivers to Support the NVIDIA GPU Cards, page D-23

GPU Card Configuration Rules

Observe the following rules when installing a GPU in the node:

Caution When using NVIDIA GPU cards, you must preserve at least 10 mm of space between nodes to ensure adequate air flow. When using GPU cards, the operating temperature range is 32° to 95° F (0° to 35° C).

- Only one GPU card per node is supported. The GPU card must be installed in PCIe riser 2, slot 5.
- All GPU cards require two CPUs in the node.
- NVIDIA GRID K1 and GRID K2 GPUs require two power supplies in the node (minimum 1200 W, recommended 1400 W).
- NVIDIA Tesla M60 GPUs require two 1400 W power supplies in the node.
- NVIDIA GPUs can support only less-than 1 TB of memory in the node. Therefore, do not install more than fourteen 64-GB DIMMs when using an NVIDIA GPU card in this node.

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 Node

If the node 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 Node

If the node is integrated with Cisco UCS Manager and controlled by a service profile, this setting is not enabled by default in the service profile. You must enable it with a BIOS policy in your service profile.

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 IO Above 4GB Config` to one of the following:

- **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 node. Use this only if you know that the node BIOS is set to use the default enabled setting for this item.

Step 4 Reboot the node.

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 node is rebooted.

Installing a GRID K1 or K2 GPU Card

Use this section to install or replace a GPU card that draws less than 300 W power. Use this section to install the following GPUs:

- GRID K1
- GRID K2

Step 1 Put the node in Cisco HX Maintenance mode as described in Shutting Down the Node Through vSphere With Cisco HX Maintenance Mode, page 3-10.

Step 2 Shut down the node as described in Shutting Down the Node, page 3-8.

Step 3 Decommission the node as described in Decommissioning the Node Using Cisco UCS Manager, page 3-11.

Caution After a node is shut down to standby power, electric current is still present in the node. To completely remove power, you must disconnect all power cords from the power supplies in the node.

Step 4 Disconnect all power cables from the power supplies.

Step 5 Slide the node 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 node from the rack.

Step 6 Remove the top cover as described in Removing and Replacing the Node Top Cover, page 3-14.

Step 7 Remove PCIe riser 2 and any existing GPU card in slot 5:

a. Lift straight up on both ends of the riser to disengage its circuit board from the socket on the motherboard. Set the riser on an antistatic mat.

b. On the bottom of the riser, loosen the single thumbscrew that holds the securing plate. See Figure D-1.

c. Swing open the securing plate and remove it from the riser to provide access.

d. Swing open the card-tab retainer that secures the back-panel tab of the card (see Figure D-1).

e. Pull evenly on both ends of the GPU card to disengage it from the socket on the PCIe riser (or remove a blanking panel) and then set the card aside.

Step 8 Install your GPU card into riser 2, PCIe slot 5. See Figure 3-17 for the riser and slot locations.

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 card power cable (UCSC-GPUCBL-240M4) into the GPU card and into the GPU POWER connector on the PCIe riser (see Figure D-1).

c. Return the securing plate to the riser. Insert the two hinge-tabs into the two slots on the riser, and then swing the securing plate closed.

d. Tighten the single thumbscrew that holds the securing plate.

e. Close the card-tab retainer (see Figure D-1).

f. Position the PCIe riser over its socket on the motherboard and over its alignment features in the chassis (see Figure 3-16).

g. Carefully push down on both ends of the PCIe riser to fully engage its circuit board connector with the socket on the motherboard.

Step 9 Replace the top cover.

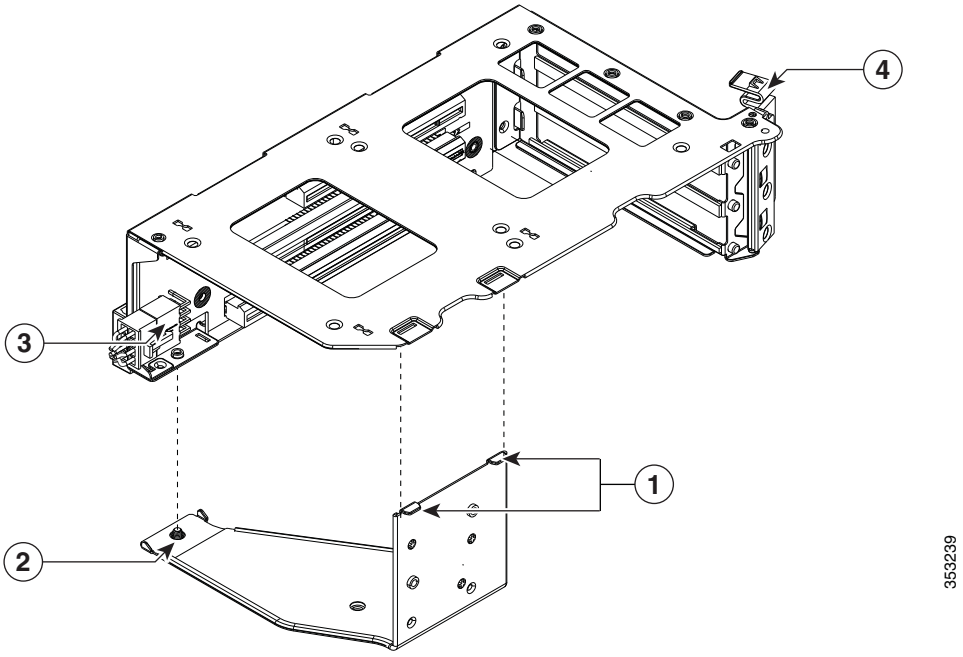
Step 10 Replace the node in the rack, replace power cables, and then power on the node by pressing the **Power** button.

Step 11 Recommission the node as described in Recommissioning the Node Using Cisco UCS Manager, page 3-12.

Step 12 Associate the node to its service profile as described in Associating a Service Profile With an HX Node, page 3-12.

Step 13 After ESXi reboot, exit HX Maintenance mode as described in Exiting HX Maintenance Mode, page 3-13.

Figure D-1 PCIe Riser Securing Features



1	Securing plate hinge-tabs	**3**	GPU card power connector
2	Securing plate thumbscrew (knob not visible on underside of plate)	**4**	Card-tab retainer in open position

Installing a Tesla M60 GPU Card and 300 W GPU Conversion Kit

Installation Overview

When installing an NVIDIA M60 GPU and the GRID software, use the topics in this section in the following order:

1. Install the hardware.
   - Installing the NVIDIA M60 Hardware, page D-5
2. Register your product activation keys with NVIDIA.
   - NVIDIA GRID License Server Overview, page D-13
   - Registering Your Product Activation Keys With NVIDIA, page D-14
3. Download the GRID software suite.
   - Downloading the GRID Software Suite, page D-14
4. Install the GRID License Server software to a host.
   - Installing NVIDIA GRID License Server Software, page D-15
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-17
6. Manage your GRID licenses.
   - Managing GRID Licenses, page D-19
7. Decide whether to use the GPU in compute mode or graphics mode.
   - Switching Between Compute Mode and Graphics Mode, page D-21

Installing the NVIDIA M60 Hardware

The NVIDIA Tesla M60 GPU requires a hardware conversion kit.

300 W GPU Card Conversion Kit

Caution Do not operate the node with the 300W GPU kit installed, but no GPU card installed. The kit has been designed to provide adequate airflow for cooling only when at least one GPU card is installed.

The contents of the conversion kit are as follows:

- Replacement fan-module fan cage
- CPU cleaning kit
- Low-profile CPU heatsinks (two)
- Replacement air baffle (includes base, bridge, and filler panel)
- 300 W GPU card front support bracket
- 300 W GPU card straight power cable

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 node; do not use the Y-cable, which is used for connecting the GPU card to external devices only.

Installing the NVIDIA M60 GPU Card

Step 1 Put the node in Cisco HX Maintenance mode as described in Shutting Down the Node Through vSphere With Cisco HX Maintenance Mode, page 3-10.

Step 2 Shut down the node as described in Shutting Down the Node, page 3-8.

Step 3 Decommission the node as described in Decommissioning the Node Using Cisco UCS Manager, page 3-11.

Caution After a node is shut down to standby power, electric current is still present in the node. To completely remove power, you must disconnect all power cords from the power supplies in the node.

Step 4 Disconnect all power cables from the power supplies.

Step 5 Slide the node 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 node from the rack.

Step 6 Remove the top cover as described in Removing and Replacing the Node Top Cover, page 3-14.

Step 7 Remove the plastic air-baffle that covers the CPUs and DIMMs.

Step 8 Remove the existing node fan cage (see Figure D-2):

- **a.** Open the plastic locking-lever at each end of the existing fan cage to the upright 90-degree position.
- **b.** Lift the existing fan cage with fan modules from the node. Set the fan cage with fan modules aside.

Step 9 Install the new empty node fan cage from the conversion kit:

- **a.** Open the plastic locking-lever at each end of the new fan cage to the upright 90-degree position.
- **b.** Set the new fan cage into the guides on the chassis walls and then lower the cage evenly.
- **c.** Close the plastic locking-lever at each end of the fan cage to the flat, locked position.

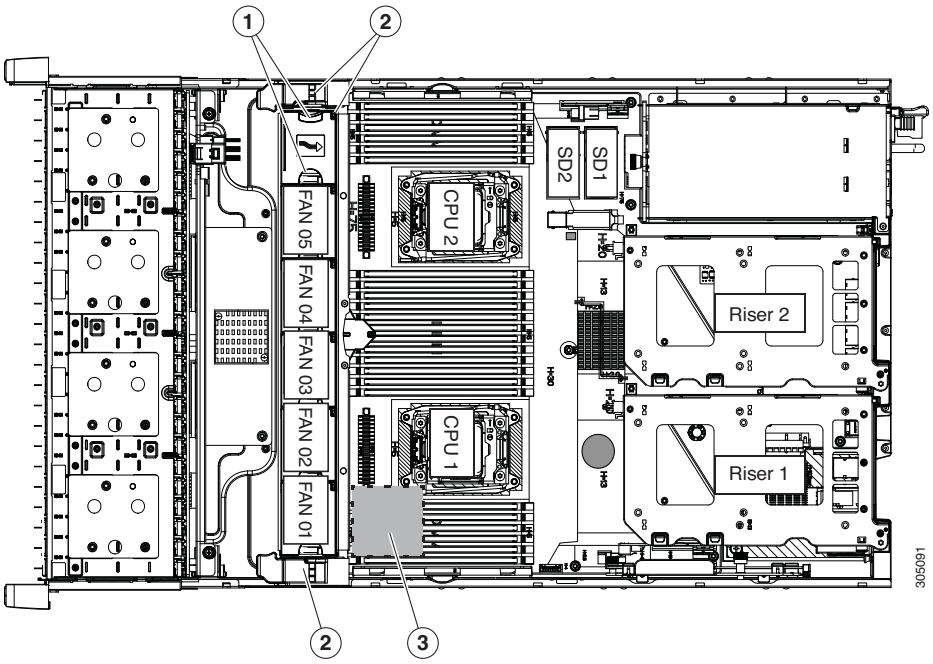
Step 10 Move the six fan modules from the old fan cage to the new fan cage that you just installed:

- **a.** Pinch the two finger latches on each fan module together, then lift up on the module to remove it from the cage (see Figure D-2).
- **b.** Set the fan module in an open slot in the new fan cage, aligning the connector on the bottom of the fan module with the connector on the motherboard.

Note The arrow label on the top of the fan module, which indicates the direction of airflow, should point toward the rear of the node.

- **c.** Press down gently on the fan module until the latch clicks and locks in place.
- **d.** Repeat until you have moved all fan modules into the new fan cage.

Figure D-2 Fan Cage and Fan Modules



1	Finger latches (on each fan module)	**3**	SuperCap power module position on removable air baffle (air baffle not shown)
2	Fan cage plastic locking-levers		

Step 11 Remove the existing heatsink from each CPU.

a. Use a Number 2 Phillips-head screwdriver to loosen the four captive screws that secure the heatsink.

Note Alternate loosening each screw evenly to avoid damaging the heatsink or CPU.

b. Lift the heatsink off of the CPU and set it aside.

Step 12 Use the heatsink cleaning kit that comes with the conversion kit to clean the existing thermal grease from the top surface of each CPU.

Step 13 Install the low-profile replacement heatsinks that come with the conversion kit:

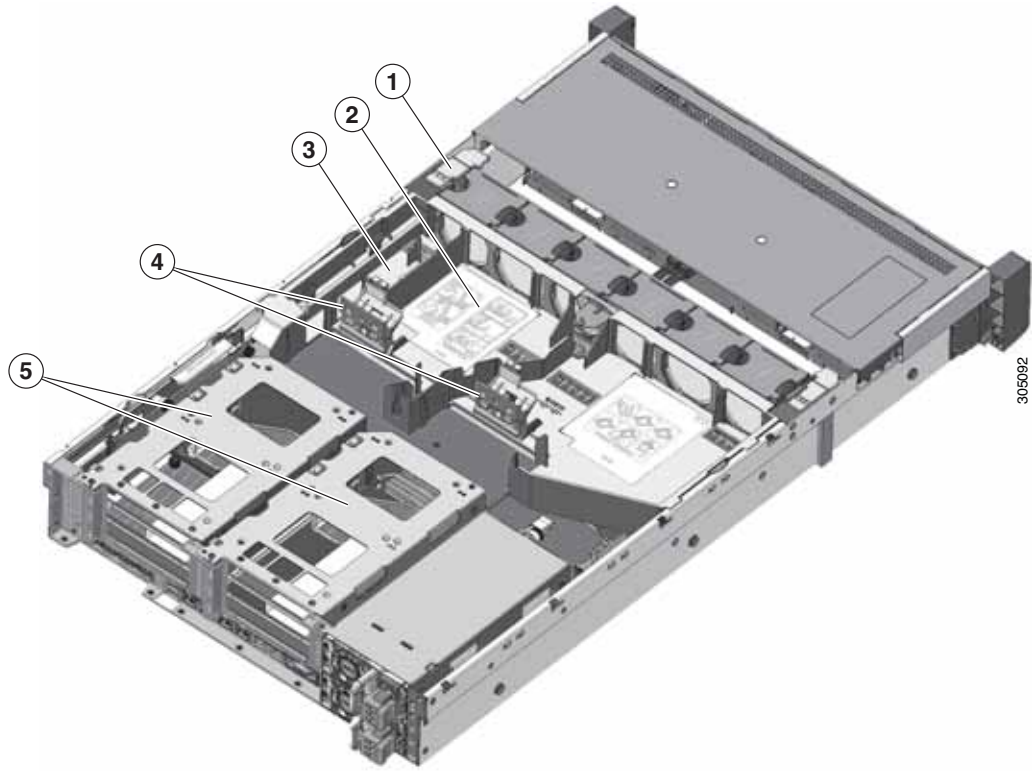
a. Remove the protective tape from the pre-applied pad of thermal grease that is on the underside of the new heatsink.

b. Align the four heatsink captive screws with the motherboard standoffs, and then use a Number 2 Phillips-head screwdriver to tighten the captive screws evenly.

Note Alternate tightening each screw evenly to avoid damaging the heatsink or CPU.

Step 14 Set the base of the replacement air baffle into the node (see Figure D-3).

Figure D-3 Air Baffle Base

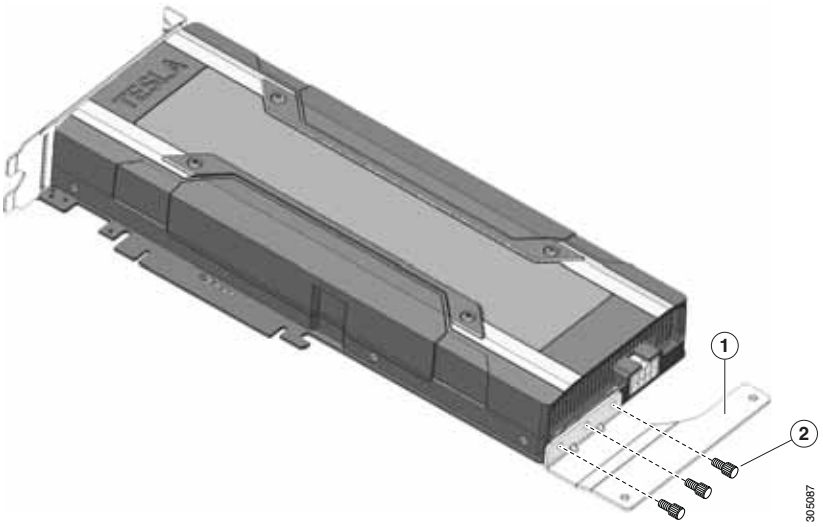


1	Fan cage	**4**	GPU front support bracket latches
2	Air baffle base	**5**	PCIe risers
3	Holder for RAID battery on air baffle		

Step 15 Install the NVIDIA M60 GPU card front support bracket to the GPU card (see Figure D-4):

- **a.** Remove the three thumbscrews from the front end of the GPU card.
- **b.** Set the three holes in the GPU card front support bracket over the three screw holes.
- **c.** Insert and tighten the three thumbscrews into the three screw holes to secure the front bracket to the GPU card.

Figure D-4 Front Support Bracket



1	Front support bracket	**2**	Thumbscrews (three)

Step 16 Remove PCIe riser 2 from the node.

- **a.** Grasp the top of the riser and lift straight up on both ends to disengage its circuit board from the socket on the motherboard.
- **b.** Set the riser on an antistatic surface.
- **c.** On the bottom of the riser, loosen the single thumbscrew that holds the securing plate (see Figure D-1).
- **d.** Swing open the securing plate and remove it from the riser to provide access.
- **e.** Swing open the card-tab retainer (see Figure D-1).

Step 17 Install the M60 GPU card into PCIe riser 2, slot 5. See Figure 3-17 for the riser and slot locations.

- **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. The straight-cable connectors are color-coded. Connect the GPU card power cable *black* connector into *black* connector on the GPU card and the *white* connector into the *white* GPU POWER connector on the PCIe riser (see Figure D-1).

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.

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 node; do not use the Y-cable, which is used for connecting the GPU card in external devices only.

c. Close the card-tab retainer (see Figure D-1).

d. Return the securing plate to the riser. Insert the two hinge-tabs into the two slots on the riser, and then swing the securing plate closed.

e. Tighten the single thumbscrew that holds the securing plate.

Step 18 Install the PCIe riser with GPU card back into the node:

a. Position the PCIe riser over its socket on the motherboard and over its alignment features in the chassis (see Figure 3-16).

b. Carefully push down on both ends of the PCIe riser to fully engage its circuit board connector with the socket on the motherboard.

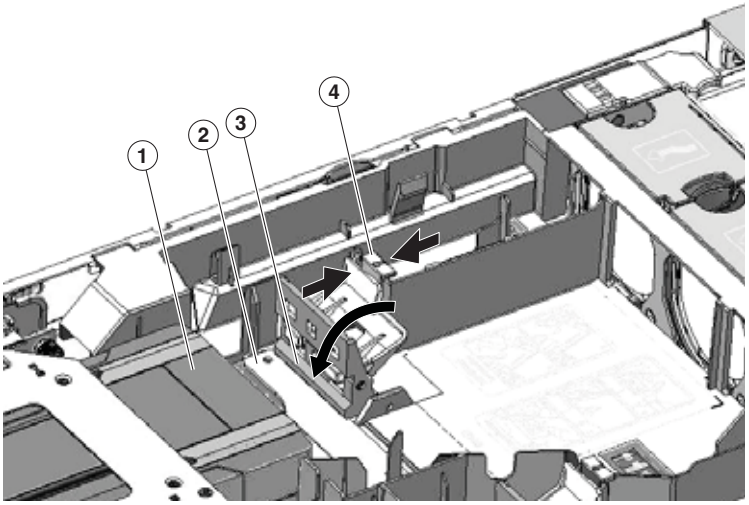
Step 19 Insert the front support bracket that you installed to the card in Step 15 into the latch that is on the air baffle base that you installed in Step 14.

a. Pinch the latch release tab (see Figure D-5) and hinge the latch toward the front of the node.

b. Hinge the latch back down so that its lip closes over the front edge of the support bracket that is attached to the GPU card (see Figure D-5).

c. Ensure that the latch release tab clicks and locks the latch in place.

Figure D-5 GPU Front Support Bracket Inserted to Air Baffle Latch

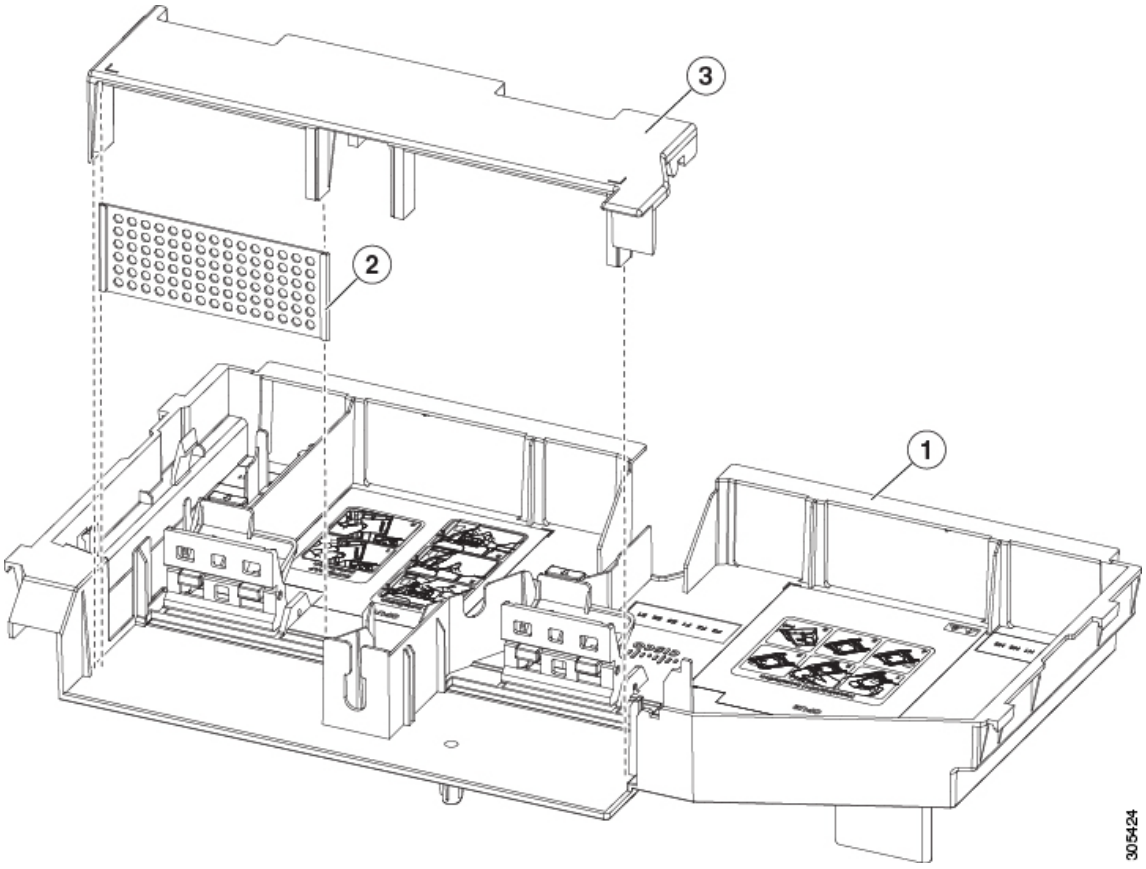


1	Front end of GPU	**3**	Lip on support bracket latch
2	Front support bracket attached to GPU	**4**	Latch release tab

Step 20 Install the filler panel to the air baffle base as shown in Figure D-6. Slide the filler panel edges into the slots as shown. The filler panel shunts more airflow to PCIe riser 2 where the M60 GPU is installed.

Step 21 Install the bridge to the air baffle. Ridges on the bridge legs slide into slots on the air baffle base (see Figure D-6).

Figure D-6 Air Baffle Filler Panel and Bridge



1	Air baffle base	**3**	Bridge
2	Filler panel		

Step 22 Replace the top cover to the node.

Step 23 Replace the node in the rack, replace power cables, and then power on the node by pressing the **Power** button.

Step 24 Recommission the node as described in Recommissioning the Node Using Cisco UCS Manager, page 3-12.

Step 25 Associate the node to its service profile as described in Associating a Service Profile With an HX Node, page 3-12.

Step 26 After ESXi reboot, exit HX Maintenance mode as described in Exiting HX Maintenance Mode, page 3-13.

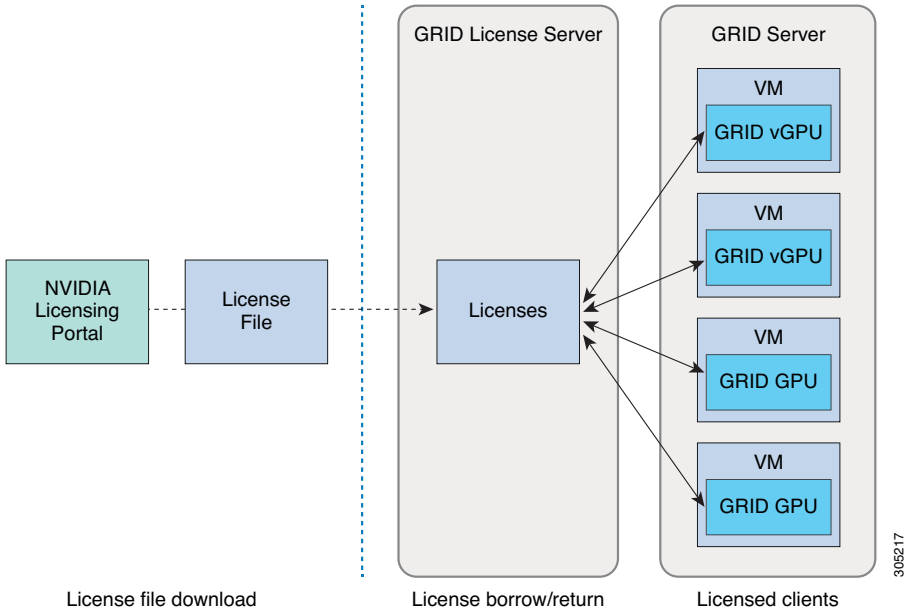
Step 27 To set up licenses for the M60 GPU, continue with NVIDIA GRID License Server Overview, page D-13.

NVIDIA GRID License Server Overview

The NVIDIA Tesla M60 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-7).

Figure D-7 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 (see Table D-1).

Table D-1 GRID Licensing Editions

GRID License Edition	GRID Features
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 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 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 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. See Accessing the GRID License Server Management Interface, page D-17.

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-17.

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-17.

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-17.

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:

```
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.

```
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:

```
# /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-2 and Table D-3 compare the compute mode and graphic mode default settings.

Table D-2 Compute Mode Default Settings

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

Table D-3 Graphic Mode Default Settings

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

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 the NVIDIA GPU Cards

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

- 1. Updating the Server BIOS Firmware, page D-23
- 2. Activating the Server BIOS Firmware, page D-24
- 3. Updating the GPU Drivers, page D-24

1. Updating the Server BIOS Firmware

Install the latest Cisco server BIOS for your blade server by using Cisco UCS Manager.

Note You must perform this procedure before you update the NVIDIA drivers.

Caution Do not remove the hardware that contains the endpoint or perform any maintenance on it until the update process completes. If the hardware is removed or otherwise unavailable due to maintenance, the firmware update fails. This failure might corrupt the backup partition. You cannot update the firmware on an endpoint with a corrupted backup partition.

Step 1 In the UCS Manager Navigation pane, click Equipment.

Step 2 On the Equipment tab, expand Equipment > Chassis > Chassis Number > Servers.

Step 3 Click the Name of the server for which you want to update the BIOS firmware.

Step 4 On the Properties page in the Inventory tab, click Motherboard.

Step 5 In the Actions area, click Update BIOS Firmware.

Step 6 In the Update Firmware dialog box, do the following:

a. From the Firmware Version drop-down list, select the firmware version to which you want to update the endpoint.

b. Click OK.

Cisco UCS Manager copies the selected firmware package to the backup memory slot, where it remains until you activate it.

Step 7 (Optional) Monitor the status of the update in the Update Status field.

The update process can take several minutes. Do not activate the firmware until the firmware package you selected displays in the Backup Version field in the BIOS area of the Inventory tab.

2. Activating the Server BIOS Firmware

Step 1 In the Navigation pane, click Equipment.

Step 2 On the Equipment tab, expand Equipment > Chassis > Chassis Number > Servers.

Step 3 Click the Name of the server for which you want to activate the BIOS firmware.

Step 4 On the Properties page in the Inventory tab, click Motherboard.

Step 5 In the Actions area, click Activate BIOS Firmware.

Step 6 In the Activate Firmware dialog box, do the following:

a. Select the appropriate server BIOS version from the Version To Be Activated drop-down list.

b. If you want to set only the start-up version and not change the version running on the server, check Set Startup Version Only.

If you configure Set Startup Version Only, the activated firmware moves into the pending-next-reboot state and the server is not immediately rebooted. The activated firmware does not become the running version of firmware until the server is rebooted.

c. Click OK.

3. 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:

- 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.