GPU Card Installation

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

Node Firmware Requirements

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

GPU Card	Cisco IMC/BIOS Minimum Version Required
NVIDIA Tesla A10 (HX-GPU-A10 or HX-GPU-A10=)	4.2(1)
NVIDIA Tesla A100 40GB (HX-GPU-A100=)	4.2(1)

GPU Card Configuration Rules

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

- The HX-C240-M6SX and HX-C240-M6N servers support a "GPU Ready" config which presets the node to accept a GPU at a later date.

  This configuration sets the node with a with low-profile heatsink (HX-HSLP-M6) and GPU air duct (HX-ADGPU-240M6) so that only the GPU is required when you choose to expand the node's compute power. The GPU Ready config has the following considerations:

  - The node must follow the same temperature limits as a node with a GPU installed, even if the node does not currently have a GPU installed. By following the temperature limits even when the GPU is not present, you ensure correct operation when the GPU is installed later.

    - This configuration requires a low-profile heatsink and a GPU air duct. If you are ordering the GPU Ready configuration, you must select the GPU air duct PID to enable GPU ready configuration when ordering the node through the Cisco online ordering and configuration tool. Follow the additional rules displayed in the tool.

- All GPU cards must be procured from Cisco because of a unique SBIOS ID that is required by CIMC and UCSM.
- Do not mix different brands or models of GPU cards in the node.
- NVIDIA A10 Single Wide GPUs (HX-GPU-A10 or HX-GPU-A10) are supported:
    - This GPU can be populated in Riser 1A slots 2 (x16) and 3 (x8), Riser 2 slots 5 (x16) and 6 (x8), and Riser 3C slot 7 (x16).
    - Each node can support a maximum of five of these GPUs.
- NVIDIA A100 Double Wide GPUs (HX-GPU-A100) are supported:
    - This GPU requires x16 support.
    - This GPU can be populated in Riser 1A slot 2 (x16), Riser 2A slot 5 (x16), and Riser 3C slot 7 (x16).
    - Each node can support three of these GPUs.
- GPUs are not supported in Riser 1B or Riser 3B. Riser 3B cannot mechanically accept a GPU.
- The HX-C240M6-S and HX-C240M6-S servers support one full-height, full-length, double-wide GPU (PCIe slot 7 only) in Riser 3C.
- The HX-C240-M6N and HX-C240-M6N servers do not support any GPU in slot 7 of Riser 3C.
- HX managed servers are discoverable only if a PCIe VIC card is installed in slot 1 or slot 4 or an mLOM VIC card is installed in the mLOM slot. If you install double-width GPUs, they must be located in slots 2, 5, or 7. Therefore, if two GPUs are installed, UCSM managed servers are discoverable only if you install a VIC in slot 1, slot 4, or the mLOM slot. The node can support 2 PCIe VICs and 1 mLOM VIC along with 2 or 3 GPUs.
- Use the UCS power calculator at the following link to determine the power needed based on your node configuration: http://ucspowercalc.cisco.com

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).

- 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 setting, enter the BIOS Setup Utility by pressing **F2** when prompted during bootup.

- If the server is integrated with Cisco UCS Manager and is 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 IO Above 4GB 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 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 Single-Wide GPU Card

Use the following procedure to install or replace an NVIDIA Tesla A10 Passive, 150W, 24GB, single-wide GPU or dual A10 GPUs.

With Cisco IMC version 4.2(1) and later, the node can support up to five NVIDIA Tesla A10 GPUs.

Ordering PID	Description
HX-RIS1A-240M6	Riser 1A
HX-RIS2A-240M6	Riser 2A
HX-RIS3C-240M6	Riser 3C

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

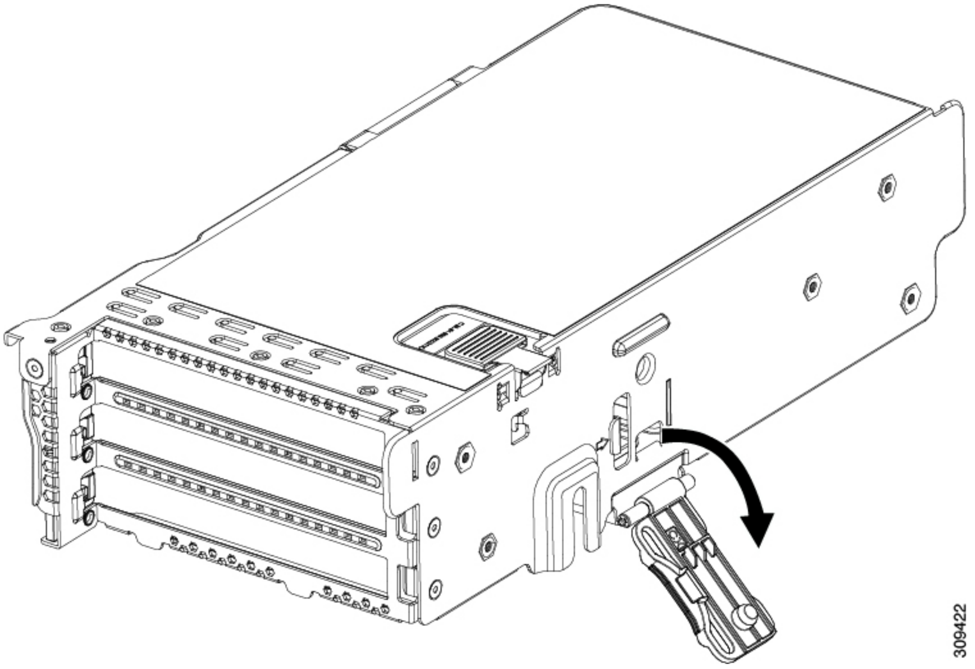
Step 2 Slide the node out the front of the rack far enough so that you can remove the top cover. You might have to detach cables from the rear panel to provide clearance.

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

Step 3 Remove the top cover from the node as described in Removing the Node Top Cover.

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

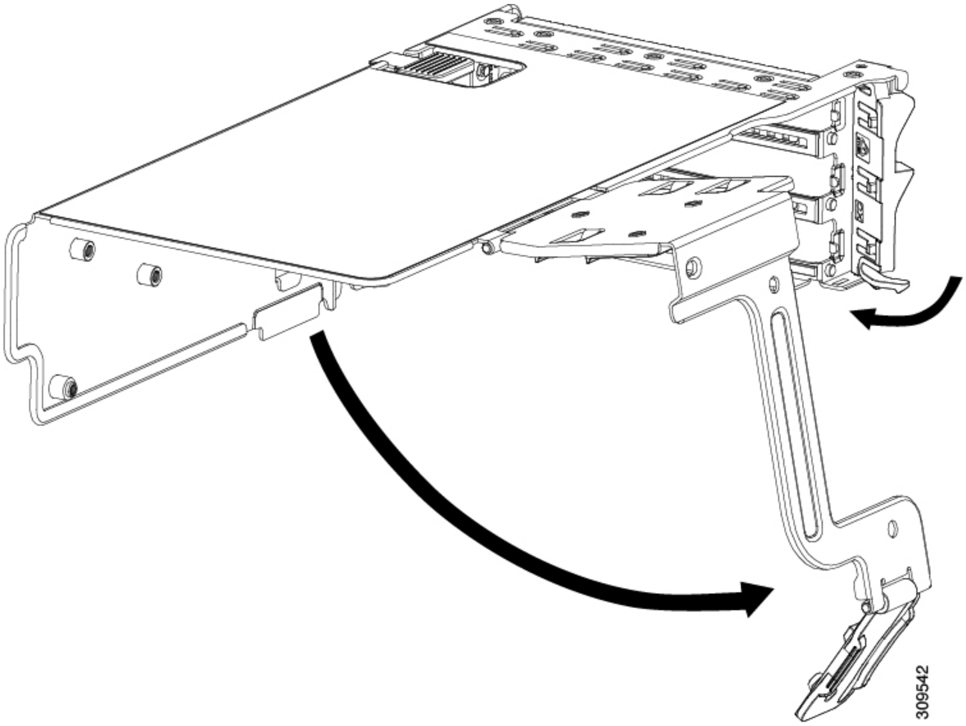
a) Use two hands to flip up and grasp the blue riser handle and the blue finger grip area on the front edge of the riser, and then lift straight up.



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) Disconnect the Y cable (UCSC-CBL-240M6) from the GPU ends (the black 8-pin connectors).

If you are inserting another GPU, you can leave the opposite end of the cable (the white connector) attached. If you are not inserting another GPU, you can leave the cable in place or completely remove it.

e) 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.
g) Attach the Y cable to the new GPU by connecting the white end to the motherboard and one of the black ends to each GPU.

Note If the node has only one GPU in the riser, one of the black GPU cables will be unused. You can let it dangle in place.

Step 6 Replace the top cover to the node.

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

Step 8 Optional: Continue with Installing Drivers to Support the GPU Cards, on page 18.

Installing a Double-Wide GPU Card

Use the following procedure to install or replace the NVIDIA Tesla A100 250W 40GB Double-Wide GPU.

With Cisco IMC version 4.2(1) and later, the server can support up to three NVIDIA Tesla A100 GPUs.

Server	Component	Recommended T Ambient Spec, Normal
All C240 24 SFF	PCIe	35 C
	Rear HDD	30 C
	GPU A100	
	GPU A10	
C240 12 LFF	Rear HDD	30 C
	GPU A100	
	GPU A10	28 C

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.

The supported NVIDIA GPU requires a C240 M5 NVIDIA Cable (HX-P100CBL-240M5).

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

Step 2 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.

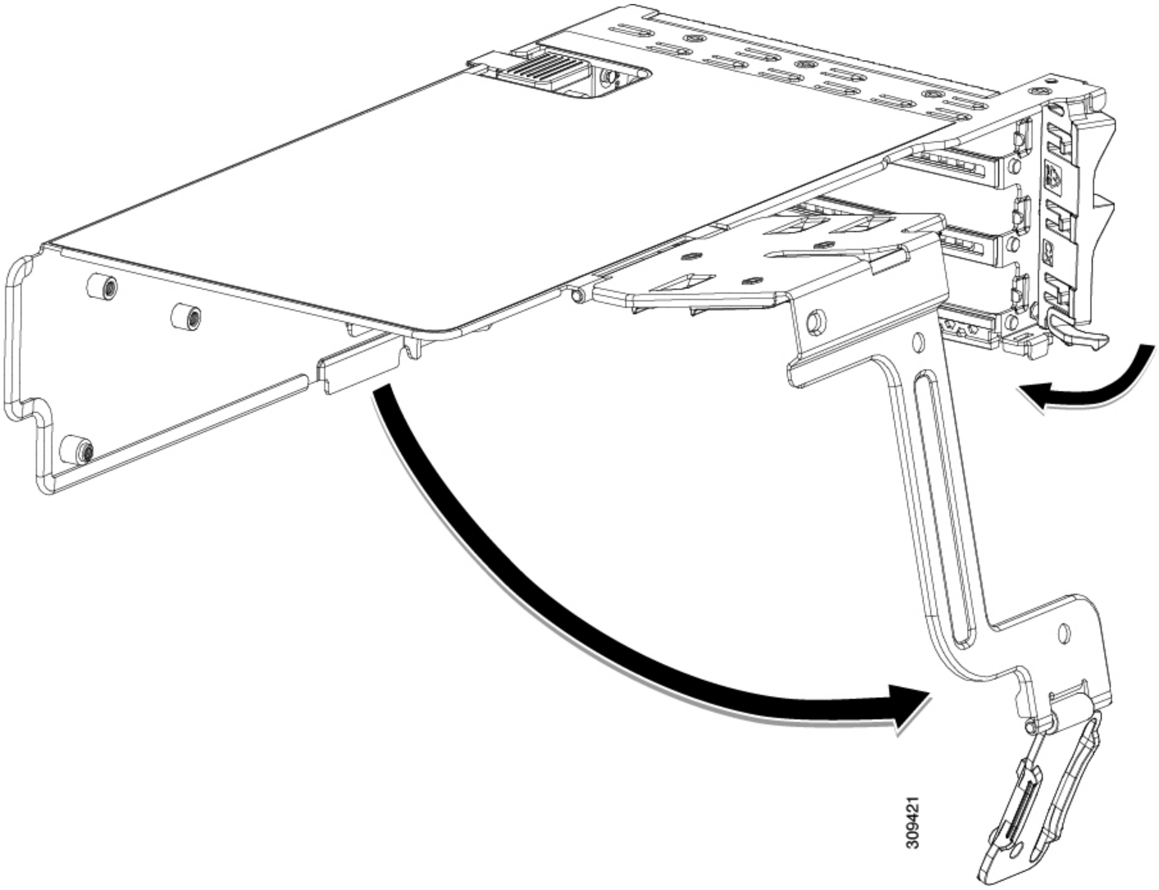
Caution If you cannot safely view and access the component, remove the node from the rack.

Step 3 Remove the top cover from the node as described in Removing the Node Top Cover.

Step 4 Remove an existing GPU card:

a) Disconnect any existing cable from the GPU card.
b) 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.
c) On the bottom of the riser, press down on the clip that holds the securing plate.
d) Swing open the hinged securing plate to provide access.
e) Open the hinged plastic retainer that secures the rear-panel tab of the card.
f) Disconnect the GPU card's power cable from the power connector on the PCIe riser.
g) 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 node, 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.

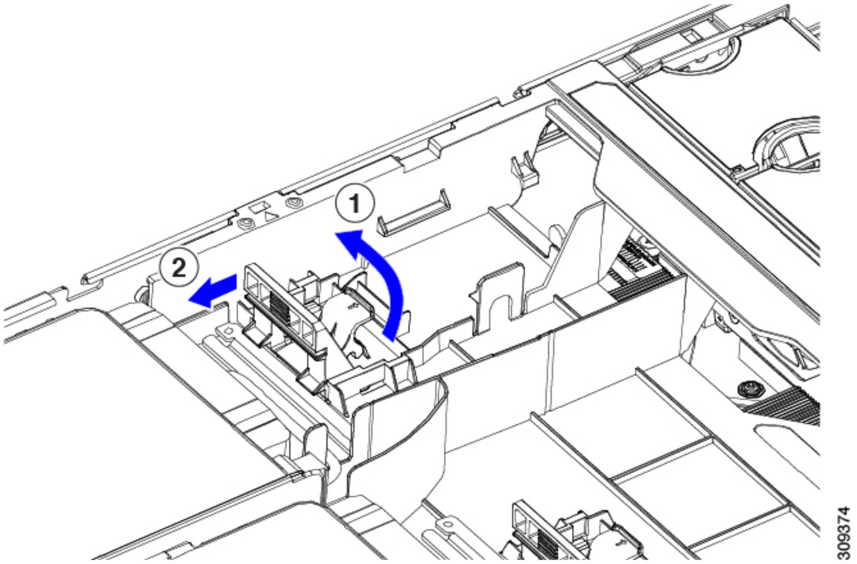
At the same time, align the GPU front support bracket (on the front end of the GPU card) with the securing latch that is on the node's air baffle.

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 node.

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.



Step 7 Replace the top cover to the node.

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

Step 9 Optional: Continue with Installing Drivers to Support the GPU Cards, on page 18.

Replacing a Heatsink

For GPUs, the correct heatsink is the low-profile heatsink (HX-HSLP-M6), which has 4 T30 Torx screws on the main heatsink, and 2 Phillips-head screws on the extended heatsink. High profile heatsinks (HX-HSHP-240M6) cannot be used on a GPU.

Use the following procedures to replace the heatsink on a GPU.

- Removing a Heat Sink, on page 8
- Installing a Heatsink, on page 12

Removing a Heat Sink

Use this procedure to remove a low-profile heatsink (HX-HSLP-M6) from a GPU.

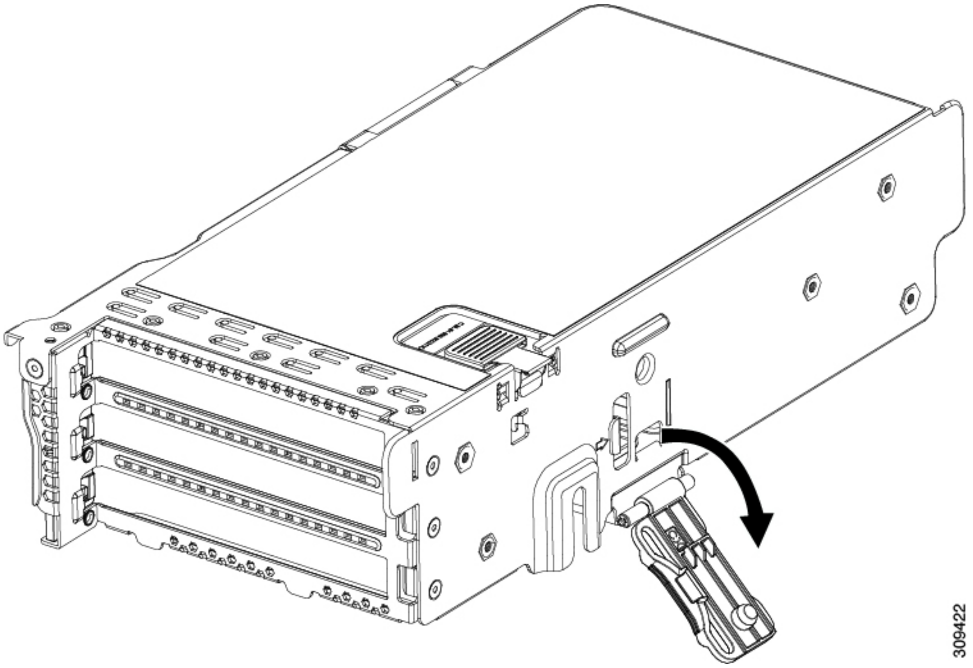
Step 1 Remove the server top cover.

Step 2 Remove the air duct.

Step 3 Choose the appropriate GPU to remove.

- For a single-wide GPU

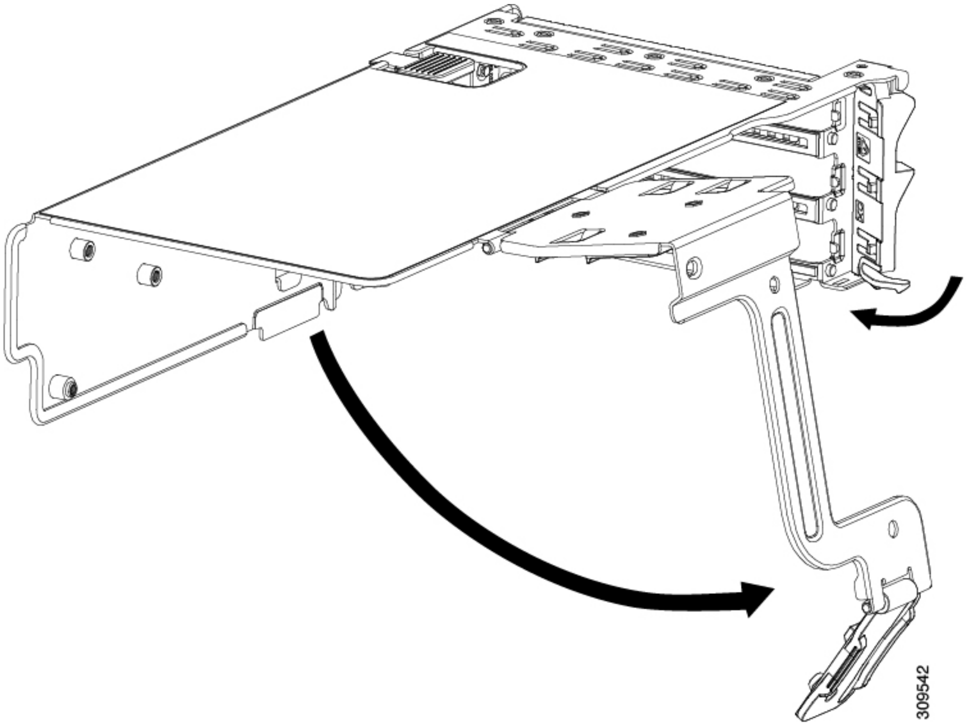
a) Use two hands to flip up and grasp the blue riser handle and the blue finger grip area on the front edge of the riser, and then lift straight up.



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 3: PCIe Riser Card Securing Mechanisms

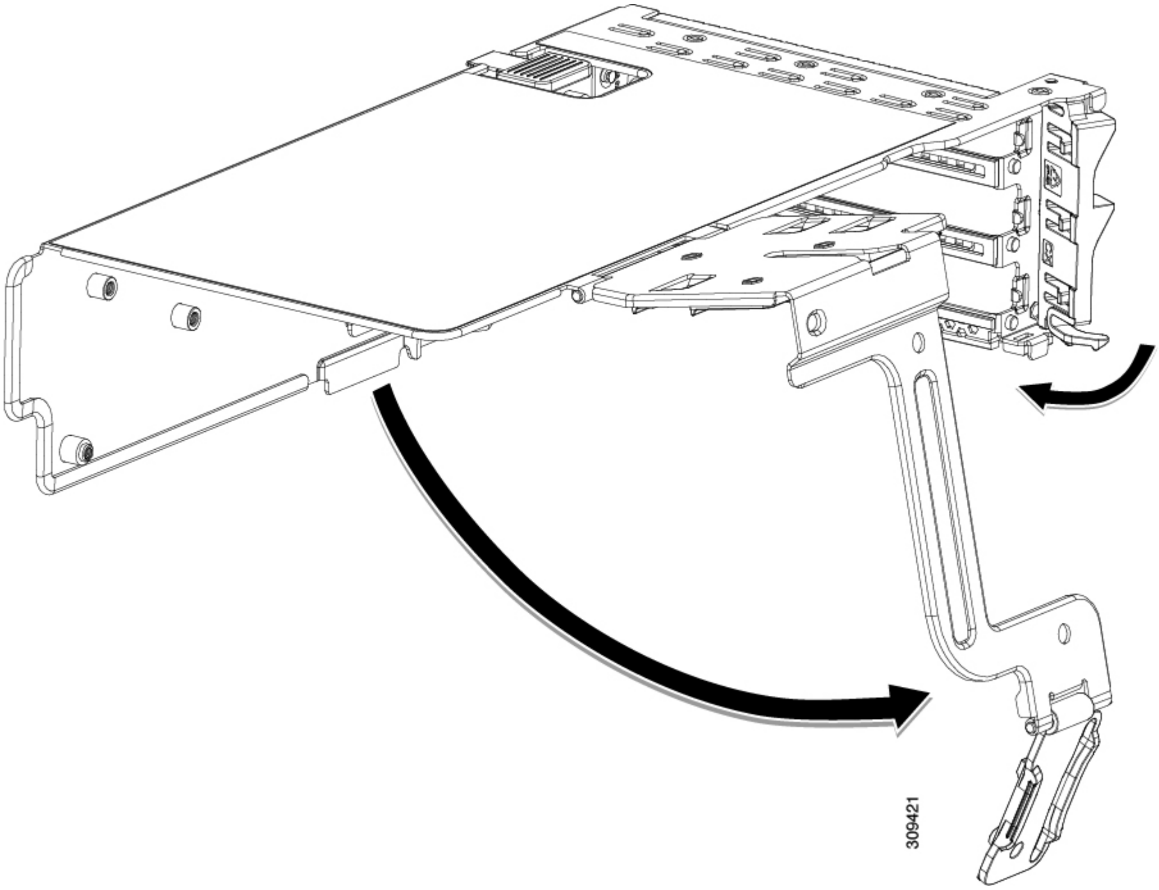


1	Release latch on hinged securing plate	**3**	Hinged card-tab retainer
2	Hinged securing plate	**-**	

- For a double-wide GPU:

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 4: PCIe Riser Card Securing Mechanisms

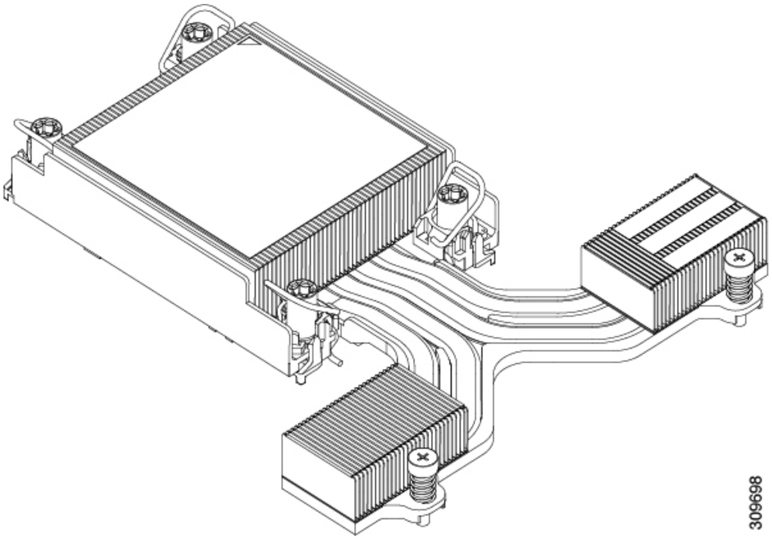


1	Release latch on hinged securing plate	**3**	Hinged card-tab retainer
2	Hinged securing plate	**-**	

Step 4 Remove the CPU.

a) Using a #2 Phillips screwdriver, loosen the two Phillips head screws for the extended heatsink.

b) Using a T30 Torx driver, loosen the four Torx securing nuts.



c) Push the rotating wires towards each other to move them to the unlocked position.

Caution Make sure that the rotating wires are as far inward as possible. When fully unlocked, the bottom of the rotating wire disengages and allows the removal of the CPU assembly. If the rotating wires are not fully in the unlocked position, you can feel resistance when attempting to remove the CPU assembly.

d) Grasp the CPU and heatsink along the edge of the carrier and lift the CPU and heatsink off of the motherboard.

Caution While lifting the CPU assembly, make sure not to bend the heatsink fins. Also, if you feel any resistance when lifting the CPU assembly, verify that the rotating wires are completely in the unlocked position.

Step 5 Remove the heatsink from the GPU.

What to do next

Install a low profile heatsink (HX-HSLP-M6) onto the GPU. See Installing a Heatsink, on page 12.

Installing a Heatsink

Use this procedure to install a low-profile heatsink (HX-HSLP-M6) on a GPU.

Step 1 Apply new TIM, if needed.

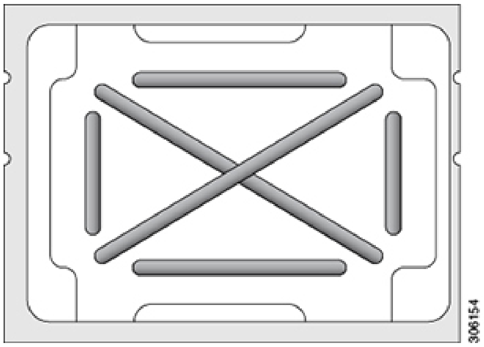
Note The heatsink must have new TIM on the heatsink-to-CPU surface to ensure proper cooling and performance.

- If you are installing a new heatsink, it is shipped with a pre-applied pad of TIM. Go to step 2 below.
- If you are reusing a heatsink, you must remove the old TIM from the heatsink and then apply new TIM to the CPU surface from the supplied syringe. Continue with step **a** below.

a) Apply the Bottle #1 cleaning solution that is included with the heatsink cleaning kit (UCSX-HSCK=), as well as the spare CPU package, to the old TIM on the heatsink and let it soak for a least 15 seconds.

b) Wipe all of the TIM off the heatsink using the soft cloth that is included with the heatsink cleaning kit. Be careful to avoid scratching the heatsink surface.
c) Completely clean the bottom surface of the heatsink using Bottle #2 to prepare the heatsink for installation.
d) Using the syringe of TIM provided with the new CPU (HX-CPU-TIM=), apply 1.5 cubic centimeters (1.5 ml) of thermal interface material to the top of the CPU. Use the pattern shown in the following figure to ensure even coverage.

Figure 5: Thermal Interface Material Application Pattern



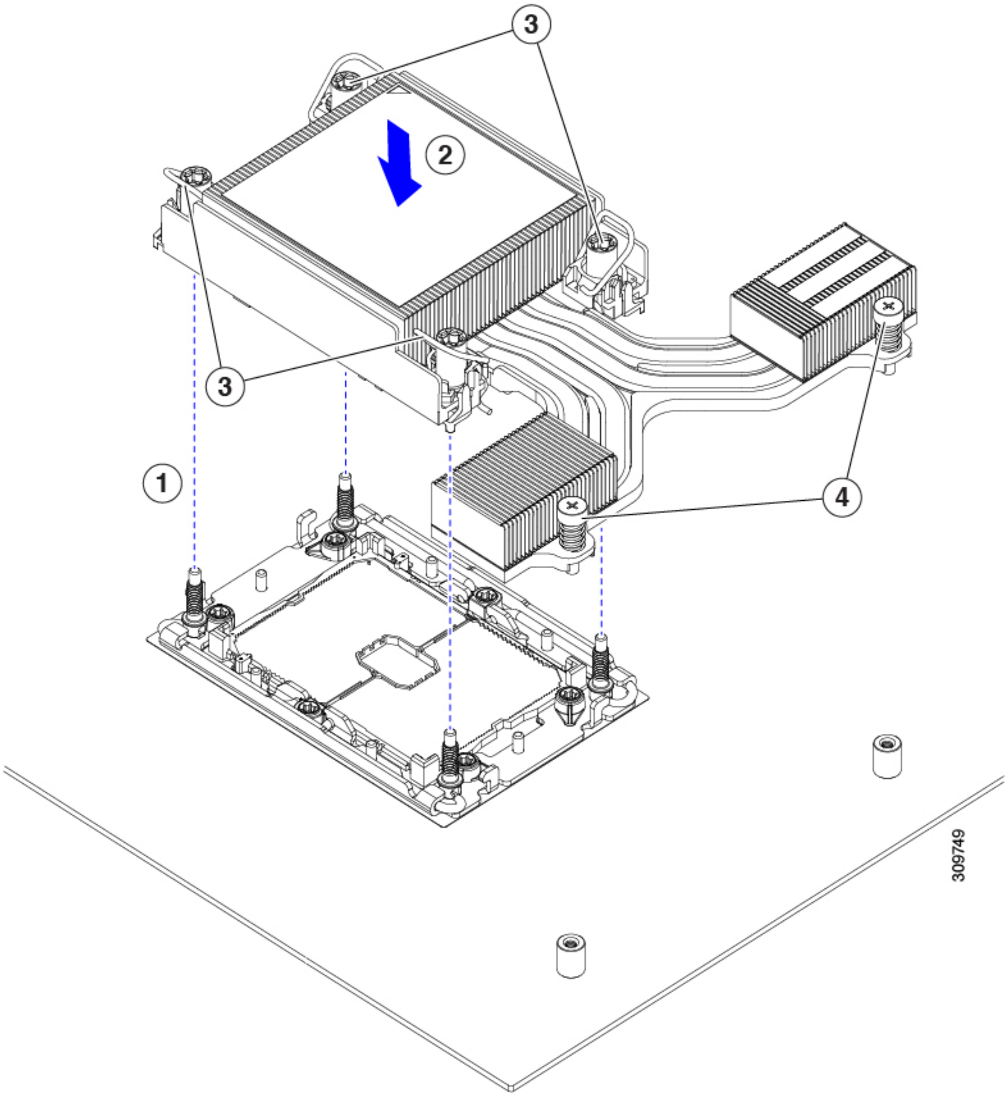
Caution Use only the correct heatsink for your CPU. CPU 1 uses heatsink HXB-HS-M6-R and CPU 2 uses heatsink HXB-HS-M6-F.

Step 2 Install the heatsink.

a) Push the rotating wires to the unlocked position so that they do not obstruct installation.
b) Grasp the heatsink by the fins, align the pin 1 location on the heatsink with the pin 1 location on the GPU (2 in the following image), then seat the heatsink onto the CPU socket.
c) Holding the CPU assembly level, orient it as shown and lower it onto the CPU socket.
d) Push the rotating wires away from each other to lock the CPU assembly into the CPU socket.

Caution Make sure that you close the rotating wires completely before using the Torx driver to tighten the securing nuts.

e) Set the T30 Torx driver to 12 in-lb of torque and tighten the 4 securing nuts to secure the CPU to the motherboard (3) first.
f) Set the torque driver to 6 in-lb of torque and tighten the two Phillips head screws for the extended heatsink (4).



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 Node 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 Node Management Interface in a browser.

Step 2 In the left-side License Node 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 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 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 Cisco HX C240 M6 node BIOS by using the Host Upgrade Utility for the Cisco HX C240 M6 node.

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 HX C-Series Rack-Mount Standalone Node Software** in the right-hand column.

Step 4 Click the name of your model of node in the right-hand column.

Step 5 Click **Cisco HyperFlex Firmware**.

Step 6 Click the release number.

Step 7 Click **Download Now** to download the `hx-`*`node 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 node 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 node.

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.