Overview

This chapter contains the following topics:

Overview

The Cisco UCS X440p PCIe node (UCSX-440P) is the first PCIe node that is part of the Cisco UCS X-Fabric system.

The PCIe node pairs with Cisco UCS X-Series compute nodes and X-Fabric modules to support workloads that require GPUs. The node is designed to work with its related hardware to simplify adding, removing, or upgrading GPUs on compute nodes. For information about the hardware, see Required Hardware, on page 2.

The PCIe node supports PCIe connectivity for a variety of GPU form factors.

- GPUs:

  - PCIe connectivity for either:

    - Two x16 FHFL (full-height, full length) or HHFL (half-height, full length) dual slot PCIe cards, one per riser cage

    - Four x8 HHHL (half-height, half length) single slot PCIe card, two per riser cage

Note Each PCIe node must be configured with the same type of GPU. Per vendor limitation, only the same GPUs are supported on a PCIe node, either two NVIDIA A16, two NVIDIA A40, two NVIDIA L40, two NVIDIA A100, two NVIDIA H100, or two Intel Flex 170 GPUs in a PCIe node (Riser Type A), or four NVIDIA L4, NVIDIA T4, or Intel Flex 140 GPUs in a PCIe node (Riser Type B).

- Host connection between the PCIe mezzanine (MEZZ) and the PCIe node is supported through a PCIe Gen 4 (2 x16) connector in the rear MEZZ slot.

- Riser options, a maximum of two risers is supported in each PCIe node. Each riser type can contain a specific type of GPU:

- Riser Type A supports 1x16 PCIe connectivity for FHFL and HHFL GPUs

- Riser Type B supports 2x8 PCIe connectivity for HHHL GPUs

Each PCIe node must have the same type of riser, either two Type A or two Type B risers. You cannot mix and match riser types in the same PCIe node.

Required Hardware

The Cisco UCS X440p PCIe Node is part of an integrated system to provide GPU acceleration for Cisco UCS compute nodes. For a complete system, the PCIe node requires the following hardware components.

- An X-Series server chassis, such as the Cisco UCS X9508 Server Chassis, that contains the paired compute and PCIe Nodes and the Cisco X-Fabric Modules. For more information, go to the Cisco UCS X9508 Server Chassis Installation Guide.

- Two X-Fabric Modules to connect Cisco UCS Compute Nodes to Cisco UCS X440p PCIe Nodes in the X-Series Server chassis. For more information, go to the Cisco UCS X9508 Server Chassis Installation Guide.

- Compute Nodes, such as the Cisco UCS X210c M6 Compute Node, to pair with the PCIe Nodes. For more information, go to the Cisco UCS X210c M6 Compute Node Installation and Service Note.

- UCS PCIe mezzanine (MEZZ) card for Fabric connectivity, which is a rear mezzanine card required on each paired UCS Compute Node. For more information, go to the Cisco UCS X210c M6 Compute Node Installation and Service Note

- Cisco UCS X440p PCIe nodes themselves to pair with the compute nodes.

- Risers, which are sized for either full height, full length (FHFL) or half-height, half length (HHHL) GPUs. Risers are part of each PCIe Node. For more information, see Riser Cage Options, on page 4.

- GPUs which provide the GPU processing offload and acceleration. For more information, see Supported GPUs, on page 6

If you require hardware to support or expand your PCIe node support, see Obtaining Hardware.

Front Panel

The Cisco UCS X440p PCIe node occupies an entire slot in the Cisco UCS X-Series server chassis. The node is front loading, so it is inserted into, and removed from, the front of the server chassis.

The following image shows the PCIe node front panel.
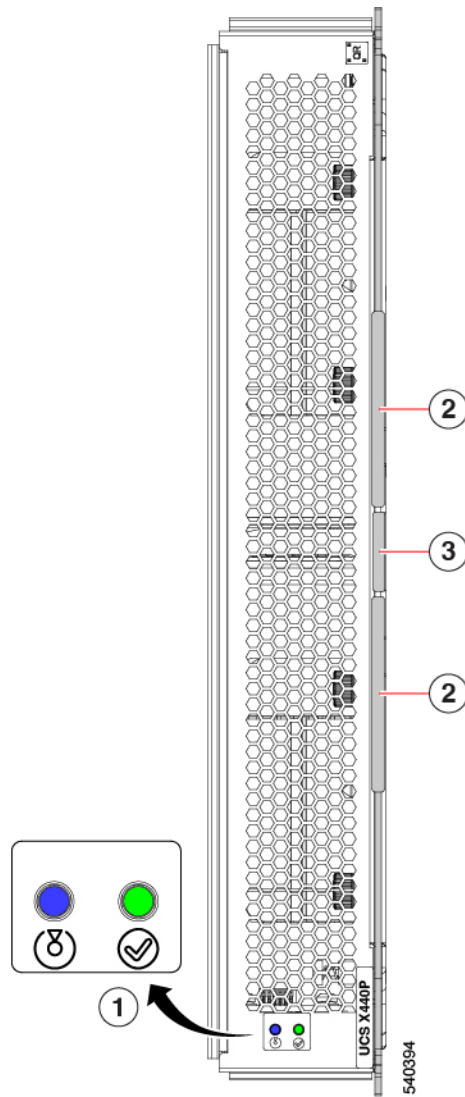
1	PCIe node LED cluster	2	PCIe node Ejector handles, 2
	See LEDs, on page 3.		
3	PCIe node Ejector Button	-	

LEDs

The PCIe node front panel has the following status LEDs.

Table 1: PCIe node LEDs

LED	Color	Description
PCIe node Locator LED	Off	Locator not enabled.
	Blinking Blue (1 Hz)	**Note** The PCIe node Locator LED is synced with its paired compute node (host). When the compute node locator LED is turned on, the paired slot PCIe node locator LED is also turned on. Locates a selected PCIe node—If the LED is not blinking, the PCIe node is not selected. You can activate the LED in UCS Intersight, which enables toggling the LED on and off.
PCIe node Health	Off	PCIe node is not powered on.
	Solid Green	Host is receiving power.
	Blinking Green (1 Hz)	Host is powered off. You can safely remove the PCIe node.
	Solid Amber	Fault condition exists, such as a configuration or system or device inventory problem.
	Blinking Amber (1 Hz)	Severe fault, such as an insufficient power condition.

Riser Cage Options

GPUs are contained in risers that mount onto the PCIe node sheet metal. Power and signaling are supported by cables that connect each riser cage to the PCIe node rear mezzanine PCBA.

- Riser Type A: Supports FHFL or HHFL GPU cards through a single x16 FHFL dual-slot PCIe connector.

- Riser Type B: Supports HHHL GPU cards through two x16 PCIe connectors, which the PCIe node supports as two x8 PCIe connectors.

If your PCIe node is not fully populated, any empty riser cage must contain PCIe slot blanks. For example, if you have a PCIe node with two HHHL GPUs in one riser, but no GPUs in the second riser, the second riser must contain two Cisco PCIe blanks.

Regardless of the riser type, any empty GPU card slot in a riser must be filled with a PCIe slot blank.

Slot Numbering

Risers and GPU slots have set numbering to identify the correct locations in the PCIe node.

- For hardware, slot numbering consists of `Riser Number Riser Type/Slot`. So, for example, `1A/1` indicates riser number 1, riser type A, slot 1.

- For Cisco management software, such as Cisco Intersight Managed Mode (IMM), the slot numbering consists of **Riser** *[number/type]*-**Slot** *[number]*. So, for example, `RISER1A-SLOT1` indicates riser number 1, riser type A, slot 1.
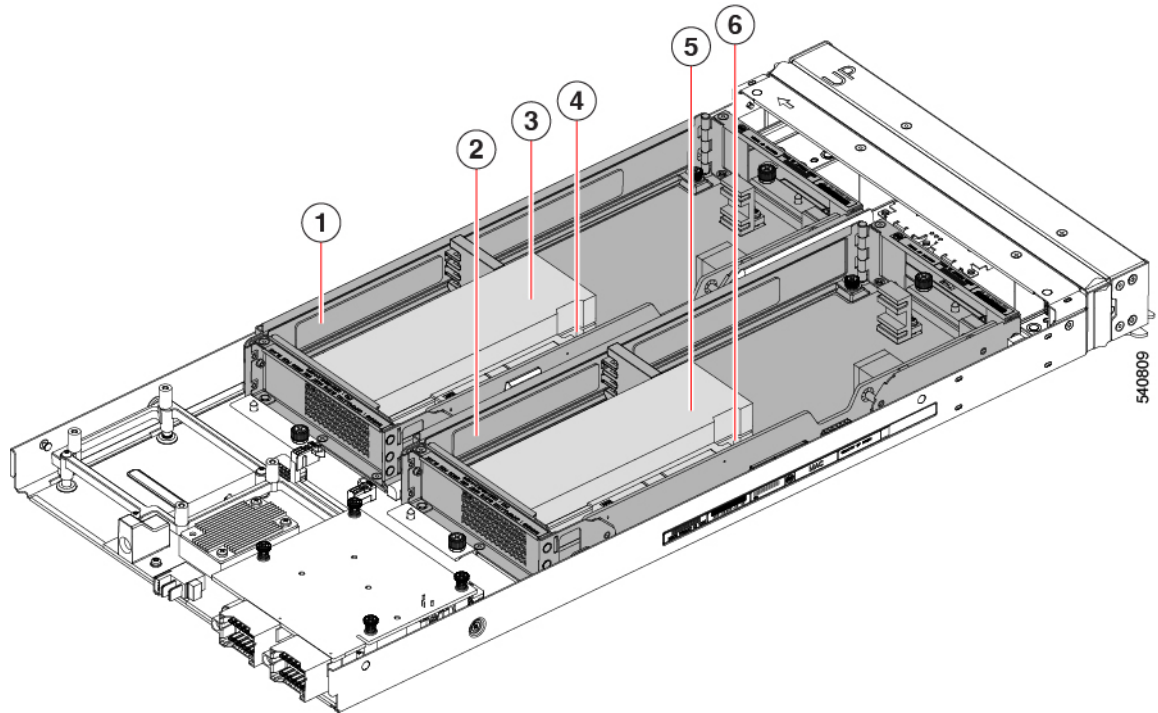
GPU slot numbering differs depending on the type of riser.

Figure 1: Riser and Slot Numbering, Riser Type A (UCSX-RISA-440P)
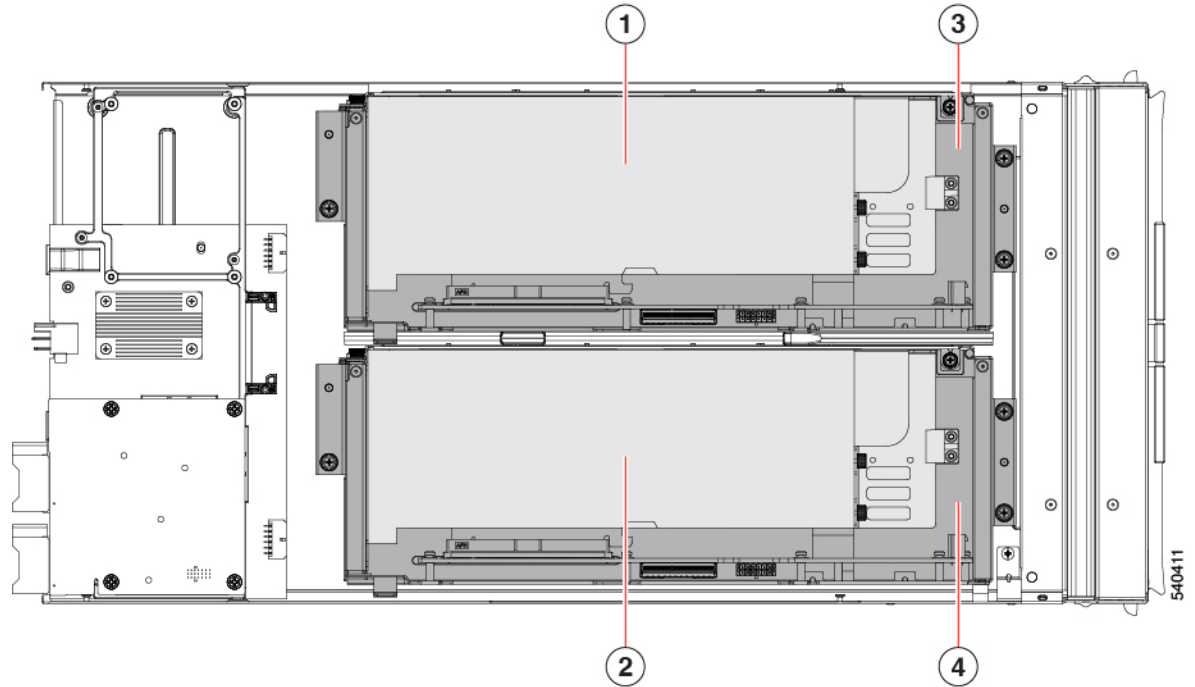
1	Slot 1A/1 for FHFL GPU Riser1A is controlled by CPU1 on the Cisco UCS compute node paired with the PCIe node	2	Slot 2A/2 for FHFL GPU Riser2A is controlled by CPU2 on the Cisco UCS compute node paired with the PCIe node
3	Riser 1 location	4	Riser 2 location

Figure 2: Riser and Slot Numbering, Riser Type B (UCSX-RISB-440P)

1	Riser 1 location	2	Riser 2 location
3	Slot 1B/2 for HHHL GPU Riser1B is controlled by CPU1 on the Cisco UCS compute node paired with the PCIe node	4	Slot 1B/1 for HHHL GPU (underneath Slot 1B/2) Riser1B is controlled by CPU1 on the Cisco UCS compute node paired with the PCIe node
5	Slot 2B/4 for HHHL GPU Riser2B is controlled by CPU2 on the Cisco UCS compute node paired with the PCIe node	6	Slot 2B/3 for HHHL GPU (underneath Slot 2B/4) Riser2B is controlled by CPU2 on the Cisco UCS compute node paired with the PCIe node

Supported GPUs

The following tables show the models and form factors of GPU supported by the Cisco UCS X440p PCIe node.

The Supported Risers column lists the slot and riser type that can accept the GPU. For example, `Riser1A` and `Riser 2A` indicate that the GPU can be installed in riser number 1, riser type A and riser number 2, riser type A. For more information, see Slot Numbering, on page 4.

Table 2: Full-Height, Full Length (FHFL) GPUs

GPU	Cisco PID	Supported Risers
NVIDIA A16 PCIe 250W 4X16 GB	UCSX-GPU-A16-D	Riser Type A only. Riser 1A (Gen 4) and Riser 2A (Gen 4)
NVIDIA Tesla A40 RTX, Passive, 300W, 48 GB	UCSX-GPU-A40-D	Riser Type A only. Riser 1A (Gen 4) and Riser 2A (Gen 4)
NVIDIA L40 300W, 48GB	UCSX-GPU-L40	Riser Type A only. Riser 1A (Gen 4) and Riser 2A (Gen 4)
NVIDIA Tesla A100, Passive, 300W, 80 GB	UCSX-GPU-A100-80-D	Riser Type A only. Riser 1A (Gen 4) and Riser 2A (Gen 4)
NVIDIA Tesla H100, Passive, 350W, 80 GB	UCSX-GPU-H100-80	Riser Type A only. Riser 1A (Gen 4) and Riser 2A (Gen 4)

Table 3: Half-Height, Full-Length (HHFL) GPU

GPU	Cisco PID	Supported Risers
Intel GPU Flex 170, Gen4 x16, 150W PCIe	UCSX-GPU-FLEX170	Riser Type A only. Riser 1A (Gen 4) and Riser 2A (Gen 4)

Table 4: Half-Height, Half-Length (HHHL) GPUs

GPU	Cisco PID	Supported Risers
NVIDIA L4 Tensor Core, 70W, 24 GB	UCSX-GPU-L4	Riser Type B only. Riser 1B (Gen 4) and Riser 2B (Gen 4)
NVIDIA T4 PCIe 75W 16 GB	UCSX-GPU-T4-16	Riser Type B only. Riser 1B (Gen 4) and Riser 2B (Gen 4)
Intel GPU Flex 140, Gen4x8, 75W PCIe	UCSX-GPU-FLEX140	Riser Type B only. Riser 1B (Gen 4) and Riser 2B (Gen 4)