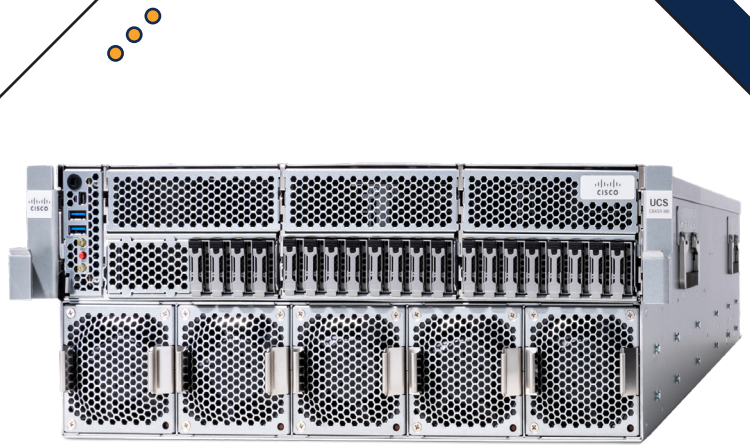
Cisco UCS C845A M8 Rack Server



Benefits

Optimized for AI use cases

Built on NVIDIA MGX modular reference design, the Cisco UCS C845A M8 Rack Server provides the accelerated computing power and NVIDIA AI Enterprise software necessary to handle the most challenging AI workloads.

Adaptable design

With flexible options for two, four, six or eight NVIDIA PCIe GPUs, you can begin with a smaller setup and expand as your needs grow. The modularity of the NVIDIA MGX design supports a wide variety of use cases.

Consistent management

Manage your AI infrastructure seamlessly with Cisco Intersight®, an operations platform that helps IT teams see, control, and automate their Cisco UCS®, converged, and hyperconverged infrastructure throughout its lifecycle—wherever it is—from one place.

Figure 1. Cisco UCS C845A M8 Rack Server

Overview

The Cisco UCS C845A M8 Rack Server is a highly scalable, flexible, and customizable two to eight GPU system based on the NVIDIA MGX reference design for accelerated computing. Together with NVIDIA AI Enterprise software, it is designed to deliver high performance across multiple AI workloads.

The versatility of the UCS C845A M8 makes it ideal for a variety of use cases, including:

- GenAI fine-tuning and inference: provides a foundation of knowledge and language patterns, enabling faster adaptation to specific tasks and domains using significantly less data compared to traditional models.
- High-Performance Computing (HPC): supports complex computations needed for simulations and large-scale data processing.
- Data analytics and visualization: utilizes advanced analytics tools to extract insights from vast datasets, facilitating data-driven decision-making.
- Design and simulation: supports 3D content creation and photorealistic simulation workloads such as digital twins, multi-user design collaboration, and Extended Reality (XR).
- Language processing: enables servers to understand and interpret human language using complex algorithms, allowing for text analysis, sentiment analysis, machine translation, and human-like text generation.
- Conversational AI: utilizes technologies such as natural language processing and machine learning to comprehend human language, discern intent, maintain context within a conversation, and generate human-like responses.
- Graphics and rendering: processes large amounts of data swiftly to generate complex visual images, making rendering faster than relying on CPUs.
- Virtual desktop: harnesses the power of accelerated computing to boost AI-enhanced virtualized workloads and deliver high-performance workstation instances to remote users with Virtual Desktop Infrastructure (VDI). Ideal for remote workloads with CAD, video editing, 3D modeling, and AI use cases.

What it offers

The Cisco UCS C845A M8 Rack Server, built on the NVIDIA MGX reference design for accelerating computing, brings AI capabilities to mainstream enterprise PCIe servers. Its adaptable configuration addresses a variety of data-center workloads, from demanding Generative AI use cases to more mainstream graphics-accelerated VDI solutions.

Configurations:

- Two 5th Gen AMD EPYC CPUs in a 4RU form factor.
- 2, 4, 6 or 8x NVIDIA H200 NVL/H100 NVL/L40S GPUs.
- 5x PCIe x16 FHHL slots and 8 x PCIe x16 GPU slots.
- 4x single-port 400G NVIDIA ConnectX-7 Smart NICs or NVIDIA BlueField-3 DPUs to scale out. One dual port 200G NIC or DPU to scale up for north-south traffic.
- Up to 20x E1.S NVMe SSDs for high-speed local storage.

Software:

Systems equipped with NVIDIA H100 NVL or H200 NVL come with a 5-year license for NVIDIA AI Enterprise, a cloud-native software platform that streamlines development and deployment of production-grade AI solutions, including AI agents, Generative AI, computer vision, speech AI, and more. Easy-to-use microservices optimize model performance with enterprise-grade security, support, and stability, ensuring a smooth transition from prototype to production for enterprises that run their businesses on AI.

Management:

The Cisco UCS C845A M8 Rack Server is managed by Cisco Intersight, a cloud-delivered IT operations platform that helps your IT operations team see, control, and automate the Cisco UCS infrastructure throughout its lifecycle—wherever it is—from one place.

By using Intersight, you can operate with consistency and control, strengthen your security posture, and increase energy efficiency to drive innovation and growth.

Learn more

For additional information about the Cisco UCS C885A M8 Rack Server, refer to the **data sheet**.

For information about our data center solutions for AI visit **https://www.cisco.com/site/us/en/solutions/artificial-intelligence/infrastructure/index.html**.

C45-4892551-00 02/25