Cisco UCS C885A M8

Overview

The acceleration of AI is fundamentally changing our world and creating new growth drivers for organizations, such as improving productivity and business efficiency while achieving sustainability goals. Scaling infrastructure for AI workloads is more important than ever to realize the benefits of these new AI initiatives. IT departments are being asked to step in and modernize their data center infrastructure to accommodate these new demanding workloads.

AI projects go through different phases: training your model, fine-tuning it, and then deploying the model to end users. Each phase has different infrastructure requirements. Training is the most compute-intensive phase, and Large Language Model (LLM), deep learning, Natural Language Processing (NLP), and digital twins require significant accelerated compute.

Built on the NVIDIA HGX platform, the Cisco Unified Computing System (Cisco UCS) C885A M8 rack server delivers the accelerated compute needed to address the most demanding AI workloads. With its powerful performance and simplified deployment, it helps you achieve faster results from your AI initiatives.

Benefits

AI-Ready

Built on NVIDIA HGX architecture, and with 8 high-performance GPUs, the Cisco® UCS C885A M8 delivers the accelerated compute power needed for the most demanding AI workloads.

Scalable

Scale your AI workloads across a cluster of Cisco UCS® C885A M8 servers to address deep learning, large Language Model Training (LLM), model fine-tuning, large model inferencing, and Retrieval-Augmented Generation (RAG).

Consistent Management

Avoid silos of AI infrastructure by managing your AI servers with the same tool as your regular workloads.

What it does

The Cisco UCS C885A M8 is a high-density GPU server designed for demanding AI workloads, offering powerful performance for model training, deep learning, and inference. Built on the NVIDIA HGX platform, it can scale out to deliver clusters of computing power that will bring your most ambitious AI projects to life.

Configuration

The UCS C885A M8 offers a choice of 8 NVIDIA HGX H100 or H200 Tensor Core GPUs, or 8 AMD MI300X OAM GPUs to deliver massive, accelerated computational performance in a single server. Additionally, it includes one NVIDIA ConnectX-7 NIC or NVIDIA BlueField-3 SuperNIC per GPU to scale AI model training across a cluster of dense GPU servers. Each server also includes NVIDIA BlueField-3 DPUs to accelerate GPU access to data. It is equipped with two 4th or 5th Gen AMD EPYC™ processors.

Management

The Cisco UCS C885A M8 is managed by Cisco Intersight®, the cloud-delivered IT operations platform that helps your IT operations team see, control, and automate 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, 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.

C45-4757152-00 10/24