Cisco UCS C880A M8 Rack Server

Contents

Value statement

The Cisco UCS C880A M8 accelerates advanced AI and High-Performance Computing (HPC) workloads in every data center with next-generation NVIDIA HGX B300 NVL8 GPUs.

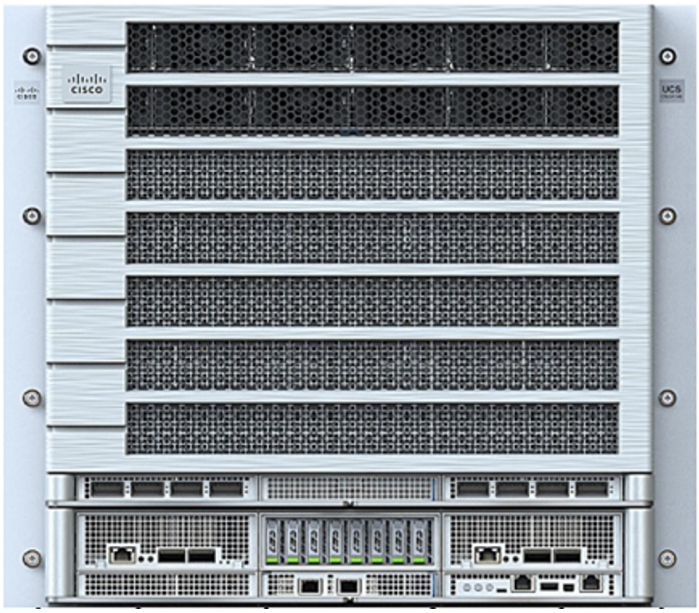

Figure 1. UCS C880A M8 Front facing

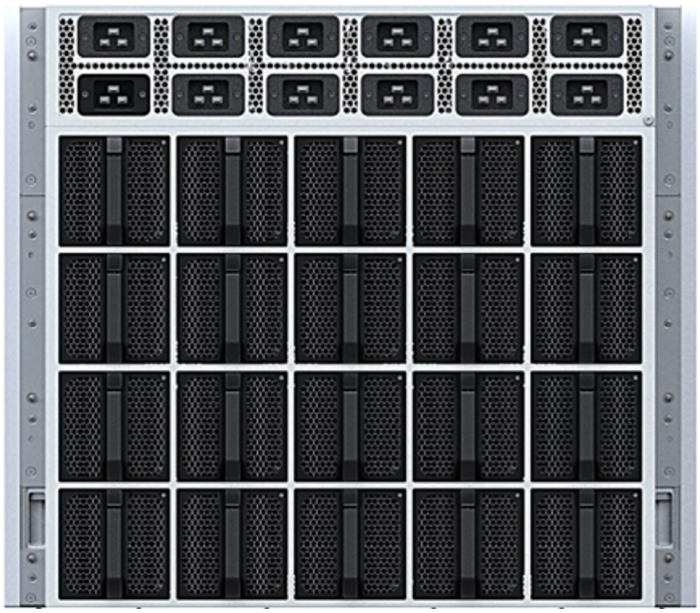

Figure 2. UCS C880A M8 Rear Facing

Product overview

Based on the NVIDIA HGX platform, the Cisco UCS C880A M8 Rack Server is a high-density, air-cooled rack server designed to power the most demanding Artificial Intelligence (AI) and High-Performance Computing (HPC) workloads. It integrates the NVIDIA HGX platform with eight NVIDIA HGX B300 (SXM) GPUs and is powered by two Intel® Xeon® 6th Gen Processors, making it ideal for real-time Large Language Model (LLM) inference, next-level training performance, and large-volume data processing. The C880A M8 supports customers across the entire AI stack, from large-scale model training and fine-tuning to real-time inferencing and large-volume data processing. It integrates seamlessly into Cisco's AI strategy, connecting and protecting the AI era by providing robust compute infrastructure. This server expands the Cisco UCS® dense AI server portfolio, offering a powerful solution for enterprises across various industries, including service providers, financial services, manufacturing, healthcare, life sciences, and automotive. With its advanced architecture, the C880A M8 ensures unparalleled performance, scalability, and enterprise manageability, making it ideal for compute-intensive AI use cases such as large-scale AI model training, fine tuning, and inferencing.

Prominent feature

Unleashing AI potential with NVIDIA HGX B300

The Cisco UCS C880A M8 Rack Server stands out by integrating the cutting-edge NVIDIA HGX platform with eight NVIDIA B300 (SXM) GPUs. This powerful GPU configuration is at the heart of its capability to deliver next-level performance for the most demanding AI workloads, including large-scale AI model training, fine tuning, and real-time inferencing. The B300 GPUs provide immense parallel processing capabilities and high-speed GPU interconnects, which are critical for accelerating complex deep learning models and large language models. This integration ensures that enterprises can achieve higher token throughput and improve the economics of their AI operations, enabling profitable scaling of LLM and agentic workloads.

Comprehensive enterprise AI manageability

The Cisco UCS C880A M8 Rack Server is designed for enterprise readiness. In a future release, the C880A M8 will enable management through Cisco Intersight.

Cisco Intersight provides a cloud-based management platform that simplifies server lifecycle management, offering capabilities such as power operations, extensive monitoring metrics, server configuration management, and firmware bundle release management. This centralized control and observability streamlines AI infrastructure operations, reduces complexity, and ensures consistent policy enforcement across the data center.

Purpose-built for AI and HPC workloads

Beyond raw power, the Cisco UCS C880A M8 Rack Server is architected specifically to meet the unique demands of AI and HPC. Its design supports real-time large language model Inference, enabling rapid deployment and responsiveness for AI-driven applications. It also excels in next-level training performance, significantly reducing the time required to train complex AI models. Furthermore, its capacity for large-volume data processing makes it an ideal platform for data-science and big-data analytics, including GPU-accelerated ETL processes. This specialized design ensures that organizations can build, optimize, and utilize AI models efficiently, accelerating business growth with scalable and high-performance solutions.

Platform support

The Cisco UCS C880A M8 is a dedicated rack server platform designed to host and accelerate AI and HPC workloads. It supports various operating systems and virtualization platforms typically used in data center environments for AI/HPC deployments. Specific software stack compatibility includes NVIDIA AI Enterprise and NVIDIA NIM (NVIDIA Inference Microservices) for AI application deployment and optimization.

Features and benefits

Table 1. Summary of features and benefits of the Cisco UCS C880A M8 Rack Server

Feature	Benefit
NVIDIA HGX with 8 NVIDIA B300 (SXM) GPUs	Leverages NVIDIA's flagship supercomputing GPUs to deliver unparalleled processing power, crucial for accelerating AI model training, fine tuning, and inferencing
Two Intel Xeon 6th Gen processors	High-frequency, high-throughput CPUs optimized to complement GPU acceleration—ideal for feeding data into training and inference pipelines without bottlenecks
Board-integrated NVIDIA ConnectX-8 NICs (E/W traffic)	Embedded east/west 800G networking fabric delivers ultra-low latency, high-bandwidth inter-GPU, and inter-server communication for scalable AI training
Upto 8 X E1.S NVMe SSDs	High-performance local NVMe storage provides ultra-low-latency data caching and fast checkpointing for AI-model training.
Hot-swappable, redundant power supplies	Enterprise-class resiliency with redundant, easily serviceable power supplies designed to minimize downtime and maximize availability
Real-time Large Language Model (LLM) inference	Enables rapid and efficient deployment of LLMs, supporting real-time applications and services that require immediate responses
Next-level training performance	Significantly reduces the time required for training large and complex AI models, allowing for faster iteration and development cycles
Large-volume data processing	Designed to handle massive datasets, facilitating accelerated Extract, Transform, and Load (ETL) processes and GPU-accelerated big-data analytics
Future releases: enterprise AI manageability with Cisco Intersight®	Cisco Intersight will provide centralized cloud-based management for the server in an upcoming release, offering such capabilities as power operations, monitoring, server configuration management, and firmware updates, simplifying operational tasks and ensuring consistent control
Validated solutions for AI	Part of Cisco's strategy to offer validated AI solutions that encompass compute, network, storage, and software, ensuring reliable and optimized performance for AI factories
Scalable infrastructure	Optimized for high-density GPU platforms, delivering predictable performance across AI factories and allowing for flexible expansion inside Cisco AI PODs to accommodate growing AI demands

Product specifications

Table 2. Cisco UCS C880A M8 Rack Server key specifications (NVIDIA HGX B300 GPU-based configurations)

Component	Specification
Form factor	10RU 19” rack server (based on NVIDIA’s HGX reference architecture)
Processors	2x Intel Xeon 6th Gen 6776P or 2x Intel Xeon 6th Gen 6767P
Memory	32x 64GB DDR5 RDIMM or 32x 96GB DDR5 RDIMM or 32x 128GB DDR5 RDIMM
GPU	8x Nvidia HGX B300 NVL8
Boot Drive	2x 960GB M.2 NVMe SSD with RAID controller
Internal Storage	Up to 8x PCIe Gen5 x4 E1.S NVMe SSD
E-W Networking	8x GPU-board integrated ConnectX-8
N-S Networking	2x CX-7 2x200G (crypto) or 2x B3220 2x200G (crypto) or 2x B3240 2x400G (crypto) 1x OCP TFF Gen5 x8
Power Supply	12x 80PLUS 54V 3.2kW MCRPS hot-swappable redundant PSUs (N+N)
Management	Cisco BMC
Hardware and Software Interoperability	See the Cisco Hardware and Software Interoperability List for a complete listing of supported operating systems and peripheral options

Ordering information

Table 3. Ordering information

Part #	Product description
UCSC-880A-M8-B301	2x Intel Xeon 6776P 2.3 GHz (Max Turbo 3.9 GHz) CPUs, 8x NVIDIA HGX B300 SXM GPUs, 32x 96GB up to 4000 MT/s DIMMs, 2x 960GB M.2 SATA Boot Drive, 2x E1.S 3.84TB NVMe SSD Data drives, 8x NVIDIA ConnectX-8 (GPU board integrated) for East/West N/W, 2x NVIDIA ConnectX-7 (2x200G) crypto-enabled for North/South N/W, 1x Intel X710-T2L OCP
UCSC-880A-M8-B302	2x Intel Xeon 6776P 2.3 GHz (Max Turbo 3.9 GHz) CPUs, 8x NVIDIA HGX B300 SXM GPUs, 32x 96GB up to 4000 MT/s DIMMs, 2x 960GB M.2 SATA Boot Drive, 2x E1.S 3.84TB NVMe SSD Data Drives, 8x NVIDIA ConnectX-8 (GPU board integrated) for East/West N/W, 2x NVIDIA B3220 (2x200G) crypto-enabled for North/South N/W, 1x Intel X710-T2L OCP
UCSC-880A-M8-B303	2x Intel Xeon 6776P 2.3 GHz (Max Turbo 3.9 GHz) CPUs, 8x NVIDIA HGX B300 SXM GPUs, 32x 128GB up to 4000 MT/s DIMMs, 2x 960GB M.2 SATA Boot Drive, 2x E1.S 3.84TB NVMe SSD Data Drives, 8x NVIDIA ConnectX-8 (GPU board integrated) for East/West N/W, 2x NVIDIA ConnectX-7 (2x200G) crypto-enabled for North/South N/W, 1x Intel X710-T2L OCP
UCSC-880A-M8-B304	2x Intel Xeon 6776P 2.3 GHz (Max Turbo 3.9 GHz) CPUs, 8x NVIDIA HGX B300 SXM GPUs, 32x 128GB up to 4000 MT/s DIMMs, 2x 960GB M.2 SATA Boot Drive, 2x E1.S 3.84TB NVMe SSD Data Drives, 8x NVIDIA ConnectX-8 (GPU board integrated) for East/West N/W, 2x NVIDIA B3220 (2x200G) crypto-enabled for North/South N/W, 1x Intel X710-T2L OCP
UCSC-880A-M8-B305	2x Intel Xeon 6776P 2.3 GHz (Max Turbo 3.9 GHz) CPUs, 8x NVIDIA HGX B300 SXM GPUs, 32x 96GB up to 4000 MT/s DIMMs, 2x 960GB M.2 SATA Boot Drive, 2x E1.S 3.84TB NVMe SSD Data Drives, 8x NVIDIA ConnectX-8 (GPU board integrated) for East/West N/W, 2x NVIDIA B3240 (2x400G) crypto-enabled for North/South N/W, 1x Intel X710-T2L OCP
UCSC-880A-M8-B306	2x Intel Xeon 6776P 2.3 GHz (Max Turbo 3.9 GHz) CPUs, 8x NVIDIA HGX B300 SXM GPUs, 32x 128GB up to 4000 MT/s DIMMs, 2x 960GB M.2 SATA Boot Drive, 2x E1.S 3.84TB NVMe SSD Data Drives, 8x NVIDIA ConnectX-8 (GPU board integrated) for East/West N/W, 2x NVIDIA B3240 (2x400G) crypto-enabled for North/South N/W, 1x Intel X710-T2L OCP
UCSC-880A-M8-B307	2x Intel Xeon 6776P 2.3 GHz (Max Turbo 3.9 GHz) CPUs, 8x NVIDIA HGX B300 SXM GPUs, 32x 64GB up to 4000 MT/s DIMMs, 2x 960GB M.2 SATA Boot Drive, 2x E1.S 3.84TB NVMe SSD Data Drives, 8x NVIDIA ConnectX-8 (GPU board integrated) for East/West N/W, 2x NVIDIA ConnectX-7 (2x200G) crypto-enabled for North/South N/W, 1x Intel X710-T2L OCP
UCSC-880A-M8-B308	2x Intel Xeon 6767P 2.3 GHz (Max Turbo 3.9 GHz) CPUs, 8x NVIDIA HGX B300 SXM GPUs, 32x 96GB up to 4000 MT/s DIMMs, 2x 960GB M.2 SATA Boot Drive, 2x E1.S 3.84TB NVMe SSD Data Drives, 8x NVIDIA ConnectX-8 (GPU board integrated) for East/West N/W, 2x NVIDIA ConnectX-7 (2x200G) crypto-enabled for North/South N/W, 1x Intel X710-T2L OCP

Warranty information

Cisco UCS C885A M8 Rack Servers have a three-year Next-Business-Day (NBD) hardware warranty and 90-day software warranty.

Cisco Support

Augmenting the Cisco UCS warranty is Cisco Success Tracks. Success Tracks add the best of both digital and human intelligence to your support experience. For more detailed information on the ST deliverables, please refer to the description here.

Product sustainability

Information about Cisco's Environmental, Social, and Governance (ESG) initiatives and performance is provided in Cisco's CSR and sustainability reporting.

Table 2. Cisco environmental sustainability information

Sustainability Topic		Reference
General	Information on product-material-content laws and regulations	Materials
	Information on electronic waste laws and regulations, including our products, batteries, and packaging	WEEE Compliance
	Information on product takeback and reuse program	Cisco Takeback and Reuse Program
	Sustainability Inquiries	Contact: csr_inquiries@cisco.com
Material	Product packaging weight and materials	Contact: environment@cisco.com

Cisco and partner services

Cisco and our industry-leading partners deliver services that accelerate your transition to Cisco UCS® solutions for AI and high-performance computing. Cisco Unified Computing Services™ can help you create an agile infrastructure, accelerate time to value, reduce costs and risks and maintain availability during deployment and migration. After deployment, our services can help you improve performance, availability, and resiliency as your business needs evolve, and help you further mitigate risk.

For more information, visit https://www.cisco.com/go/unifiedcomputingservices.

Cisco Capital

Flexible payment options make it easier than ever to get the Cisco® technology you need.

Cisco Capital® delivers leading-edge payment solutions, allowing you to stay focused on what's most important—your business. We can help you drive business outcomes, accelerate innovation and digital transformation, and adapt to market dynamics faster with flexible payment options tailored to your specific business needs. Reduce the total cost of ownership, conserve capital, and accelerate growth. We help you realize the full benefits of Cisco technology today, and in the future, and pay for it in the way that best suits your business requirements. Whether you are looking for a pay-as-you consume model, or need to bundle Cisco hardware, software, services, subscriptions, and third-party solutions, learn more about how Cisco Capital can help.

Learn more

Visit the Dense GPU – HGX and OAM webpage.

C78-5455801-00 09/25