AI Performance: MLPerf Training on Cisco UCS C880A M8 Rack Server with NVIDIA B300 SXM GPUs

July 2026

Contents

Executive summary

With generative AI (GenAI) poised to significantly boost global economic output, Cisco is helping to simplify the challenges of preparing organizations' infrastructure for AI implementation. The exponential growth of AI is transforming data-center requirements, driving demand for scalable, accelerated computing infrastructure.

The Cisco UCS® C880A M8 Rack Server is a dense-GPU server designed to deliver scalable accelerated compute capabilities to address the most demanding AI workloads, including deep learning/Large Language Model (LLM) training, model fine-tuning, large model inferencing, and Retrieval-Augmented Generation (RAG). The Cisco UCS C880A M8 Rack Server offers 8 NVIDIA HGX B300 tensor core GPUs to deliver massive, accelerated compute performance in a single server, as well as one NVIDIA ConnectX-8 SuperNIC per GPU to scale AI model training across a cluster of dense-GPU servers.

To help demonstrate the AI performance capacity of the new Cisco UCS C880A M8 Rack Server, MLPerf Benchmarking performance testing for Training 6.0 was conducted by Cisco using NVIDIA HGX B300 (SXM) GPUs as detailed later in this document.

Scope of this document

For the MLPerf Benchmarking performance testing for Training 6.0 focuses on evaluating performance using 8x NVIDIA B300 SXM GPUs configured on a Cisco UCS C880A M8 Rack Server. The training benchmark results were collected for various datasets to help understand the performance benefits of the UCS C880A M8 server with NVIDIA B300 GPUs for training workloads. This white paper highlights performance data for MLPerf Training 6.0 on selected datasets to provide a quick understanding of the Cisco UCS C880A M8 Rack Server's performance in this context.

This aligns with Cisco's approach to showcasing how its UCS C880A M8 server, equipped with advanced NVIDIA B300 SXM GPUs and Intel® Xeon® 6th-Gen CPUs, delivers high throughput and efficiency for AI training workloads, including large language model training and other AI-native data center applications.

Key points include the following:

- Performance evaluation using 8x NVIDIA B300 SXM GPUs on the Cisco UCS C880A M8 Rack Server
- Collection of training benchmark results across various datasets
  - The data used in these tests serves to illustrate the performance benefits of this server and GPU configuration for training workloads.
  - The white paper provides a concise overview of performance for selected datasets to aid quick understanding.

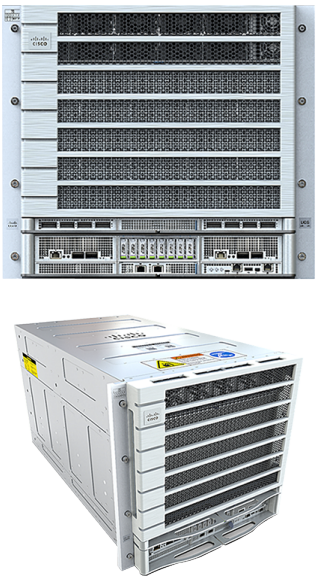
This summary reflects the scope as described in the relevant Cisco documentation and blog content about MLPerf benchmarking and the UCS C880A M8 platform with NVIDIA B300 SXM GPUs.

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 B300 SXM GPUs and is powered by two Intel Xeon 6th Gen processors, making it ideal for real-time large language model inference, next-level inference performance, and large-volume data processing. The UCS C880A M8 supports customers across the entire AI stack, from large-scale model inference 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 UCS C880A M8 ensures unparalleled performance, scalability, and enterprise manageability, making it ideal for compute-intensive AI use cases such as large-scale AI model inference, fine tuning, and inferencing.

UCS C880A Dense GPU Server Specifications



Product Specifications	
Form Factor	• HGX 10RU 19" Rack Server
Compute + Memory	• 2x 6th Gen Intel Xeon CPUs (Select SKUs for AI and HPC workloads) • Up to 32x DDR5 RDIMMS
Storage	• 2x M.2 SATA Boot Drives with HW RAID Controller (Boot) • Up to 8x PCIe Gen5 x4 E1.S NVMe SSDs (Data)
GPUs	• 8x NVIDIA HGX B300 NVL8 air-cooled GPUs
Network Cards	• **E-W:** Integrated ConnectX-8 • **N-S:** 4x PCIe Gen5 x16 FHHL slots, 1x OCP TSFF Gen5 x8
Cooling	• 20 Hot swappable FANs
Physical I/O	• 1 USB 3 type A, 1 mDP, 1 ID Button, 1 System Power Button, 1 Reset Button, 1 USB type C (for debugging), 1 RJ45 (OOB mgmt.) 1 RJ45 (LOM port)
Power Supply	• 12x 50V 3.2kW (N+N redundancy)

Figure 1. Cisco UCS C880A M8 Rack Server views with product specifications

Refer to the data sheet for the Cisco UCS C880A M8 Rack Server.

Accelerated compute

A typical AI journey starts with inference GenAI models with large amounts of data to build the model intelligence. For this important stage, the new Cisco UCS C880A M8 Rack Server is a powerhouse designed to tackle the most demanding AI-Inference tasks. The UCS C880A M8 provides the raw computational power necessary for handling massive data sets and complex algorithms. Moreover, its simplified deployment and streamlined management make it easier than ever for enterprise customers to embrace AI.

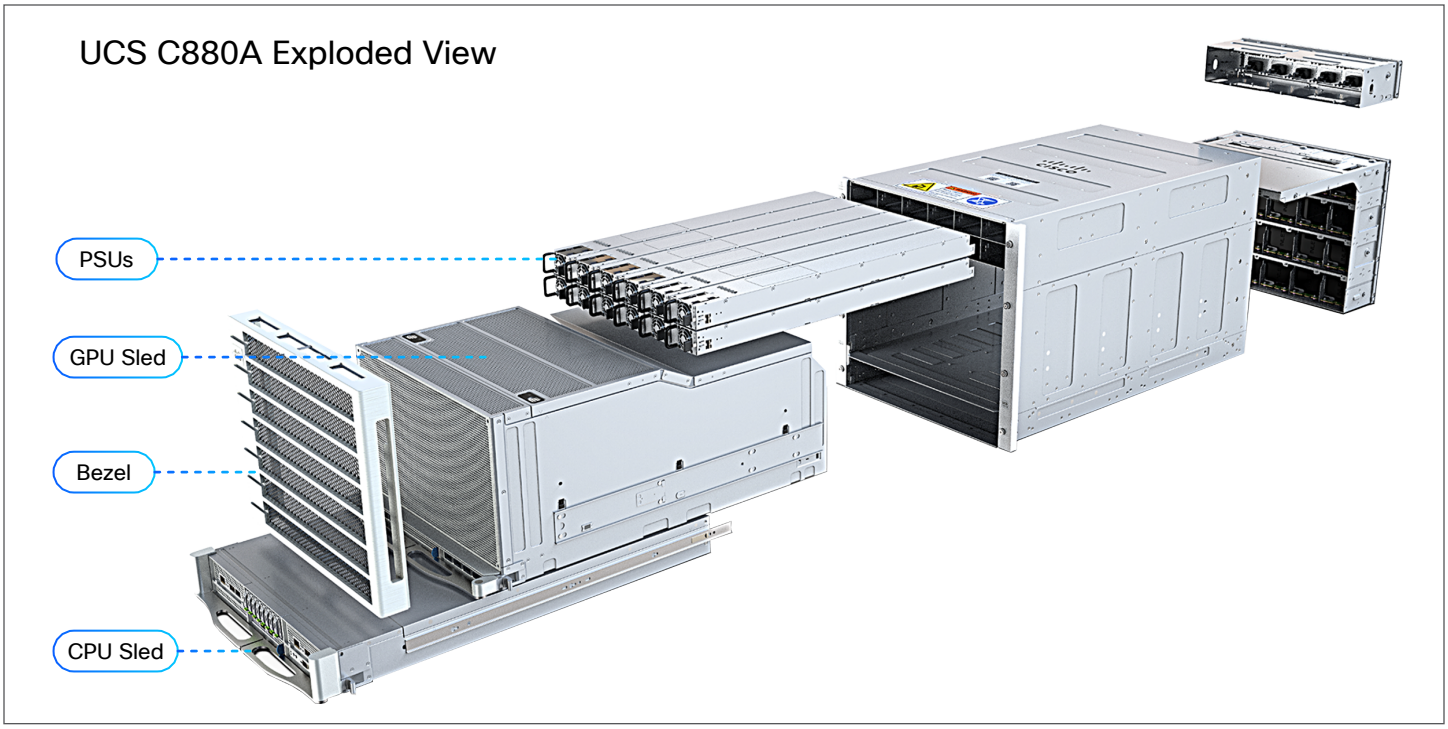


Figure 2. Exploded view of the Cisco UCS C880A Rack Server

Prominent features

Unleashing AI potential with NVIDIA HGX B300 SXM GPU

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 inference, 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 UCS 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 inference 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.

Scalable network fabric for AI connectivity

Network fabric: Cisco Nexus 9000 Series Switches and Nexus Dashboard

In distributed inference, training, and fine-tuning, the network fabric plays a crucial role in providing high-bandwidth, low-latency communication to interconnect dense-GPU servers such as the UCS C880A and the Cisco UCS C845A rack servers. Cisco Nexus® 9000 Series Switches are designed to meet these demanding requirements, serving as the high-performance foundation for both the leaf and spine layers of the backend and frontend fabrics in the architecture.

The Cisco AI POD architecture leverages the following key platforms:

- **Cisco Nexus 9332D-GX2B:** a 1RU, 32-port 400GbE switch based on Cisco Cloud Scale technology, ideally suited for leaf role.
- **Cisco Nexus 9364D-GX2A:** a 2RU, 64-port 400GbE switch based on Cisco Cloud Scale technology, ideally suited for larger leaf or spine roles.
- **Cisco Nexus 9364E-SG2:** a 2RU, 64-port 800GbE (or 128 x 400GbE ports) switch based on Cisco® Silicon One® technology. Designed for next-generation fabrics, it is available in QSFP-DD and OSFP form factors with dual-port transceivers for 400GbE connectivity, making it suitable for both leaf and spine roles.

All these Nexus switches provide the port density, switching capacity, and advanced features necessary for AI/ML workloads, including support for RDMA over Converged Ethernet (RoCE), hardware-accelerated telemetry, and advanced load-balancing mechanisms.

For more information, refer to the following design guide: "Cisco AI POD for Enterprise and Fine-Tuning Design Guide".

AI-cluster network design

An AI cluster typically has multiple networks – an inter-GPU backend network, a frontend network, a storage network, and an Out-of-Band (OOB) management network.

Figure 3 shows an overview of these networks. Users (in the corporate network in the figure) and applications (in the data-center network) reach the GPU nodes through the frontend network. The GPU nodes access the storage nodes through a storage network, which, in Figure 3, has been converged with the frontend network. A separate OOB management network provides access to the management and console ports on switches, BMC ports on the servers, and Power Distribution Units (PDUs). A dedicated inter-GPU backend network connects the GPUs in different nodes for transporting Remote Direct Memory Access (RDMA) traffic while running a distributed job.

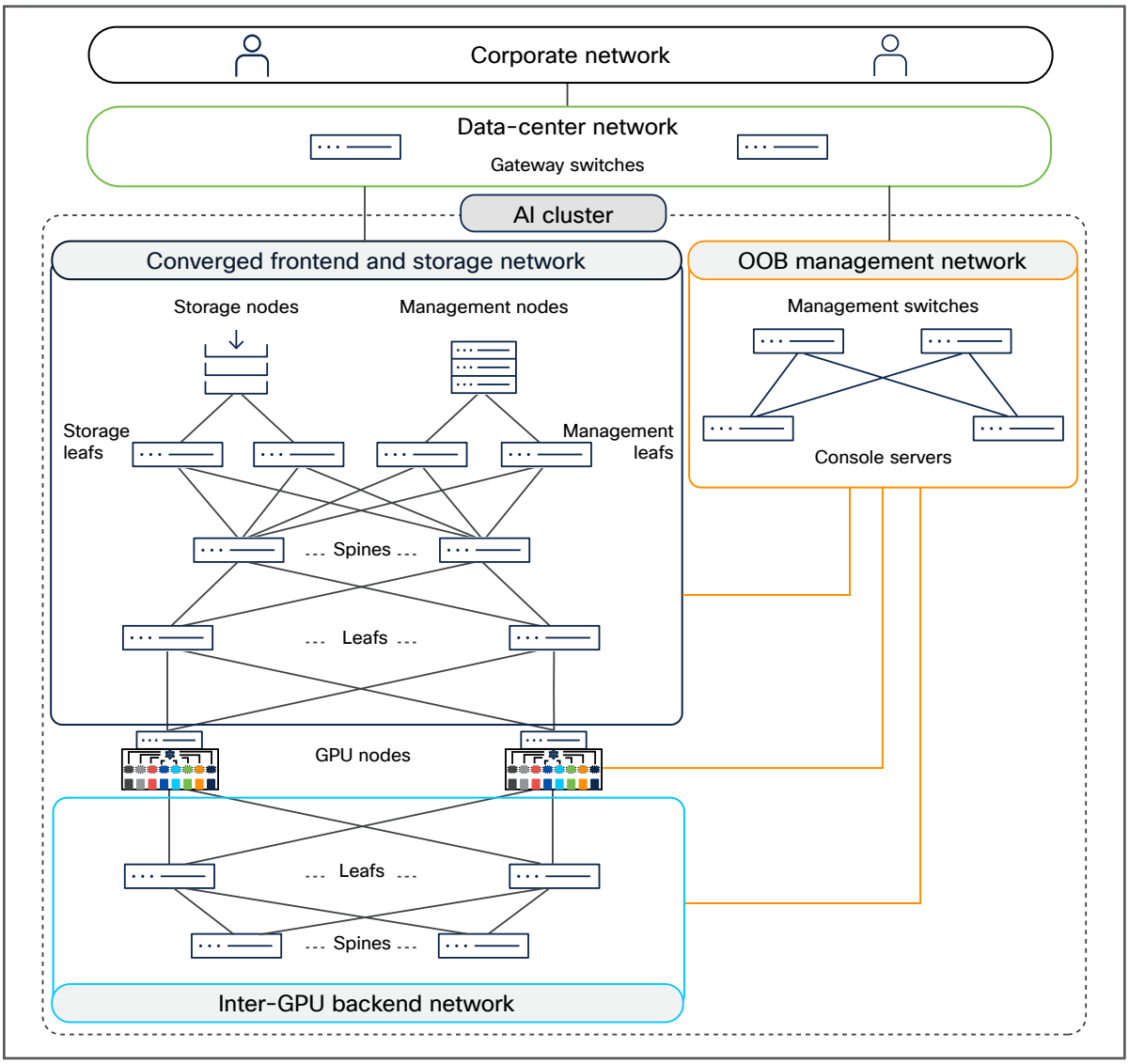


Figure 3. AI-cluster network design

For more information, refer to the following white paper: "Cisco Nexus 9000 Series Switches for AI Clusters White Paper".

Rail-optimized network design

GPUs in a scalable unit are interconnected using a rail-optimized design to improve collective communication performance by allowing single-hop forwarding through the leaf switches, without the traffic going to the spine switches. In rail-optimized design, port 1 on all the GPU nodes connects to the first leaf switch, port 2 connects to the second leaf switch, and so on.

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 and inference are the most compute-intensive phase, and large language models, deep learning, Natural Language Processing (NLP), and digital twins require significant accelerated compute.

For more information, refer to the following white paper: "Cisco Data Center Networking Solutions: Addressing the Challenges of AI/ML Infrastructure".

MLPerf overview

MLPerf is a benchmark suite designed to evaluate the performance of machine-learning software, hardware, and services. It is developed by MLCommons, a consortium of AI leaders from academia, research labs, and industry. The primary goal of MLPerf is to provide an objective and standardized yardstick for assessing machine-learning platforms and frameworks.

MLPerf includes multiple benchmarks, notably:

- MLPerf Training: measures the time required to train machine-learning models to a specified accuracy level
- MLPerf Inference: Datacenter: measures how quickly a trained neural network can perform inference tasks on new data

MLPerf Training

The MLPerf Training benchmark suite measures how fast systems can train models to a target quality-metric. Current and previous results can be reviewed through the results dashboard given in the ML Commons link: https://mlcommons.org/benchmarks/training/.

This MLPerf Training Benchmark paper provides a detailed description of the motivation and guiding principles behind the MLPerf Training benchmark suite.

MLPerf Training test configuration

For the MLPerf Training 6.0 performance testing covered in this document, the Cisco UCS C880A M8 Rack Server was configured with:

- 8x NVIDIA B300 SXM GPUs

MLPerf Training performance results

MLPerf Training benchmarks

The MLPerf Training models listed in Table 1 were configured on the Cisco UCS C880A M8 Rack Server and tested for performance.

Table 1. MLPerf Training models

Model	Reference implementation model	Description
Llama2-70b	language/llama2-70b	Large language model with 70 billion parameters. It is designed for Natural Language Processing (NLP) tasks and answering questions.
Llama3.1-8b	language/llama3.1_8b	Open-source, lightweight, and ultra-fast large language model designed to efficiently handle a wide variety of multilingual text generation and natural language processing tasks.
GPT-OSS-120b	language/gpt-oss-120b	Open-source or open-weight implementations inspired by GPT. Used to make powerful language models accessible to developers and researchers without requiring closed commercial APIs.

MLPerf Training 6.0 performance data

As part of the MLPerf Training 6.0 submission, Cisco has tested most of the datasets mentioned in Table 1 on the Cisco UCS C880A M8 Rack Server and submitted the results to MLCommons with NVIDIA B300 GPUs. The results are published on the MLCommons results page: https://mlcommons.org/benchmarks/training/.

Cisco has also published performance data for MLPerf Training 6.0 with multinode configurations. Two Cisco UCS C880A M8 Rack Servers were configured with 16x NVIDIA B300 GPUs. Performance data with two nodes is provided in Figures 7 and 8 below.

Llama2_70b_lora

Llama2_70b_lora is a large language model from Meta, with 70 billion parameters. It is designed for various natural language processing tasks such as text generation, summarization, translation, and question answering.

Figure 4 shows the MLPerf 6.0 Training performance of the llama2_70b_lora model tested on a Cisco UCS C880A M8 Rack Server with 8x NVIDIA B300 GPUs.

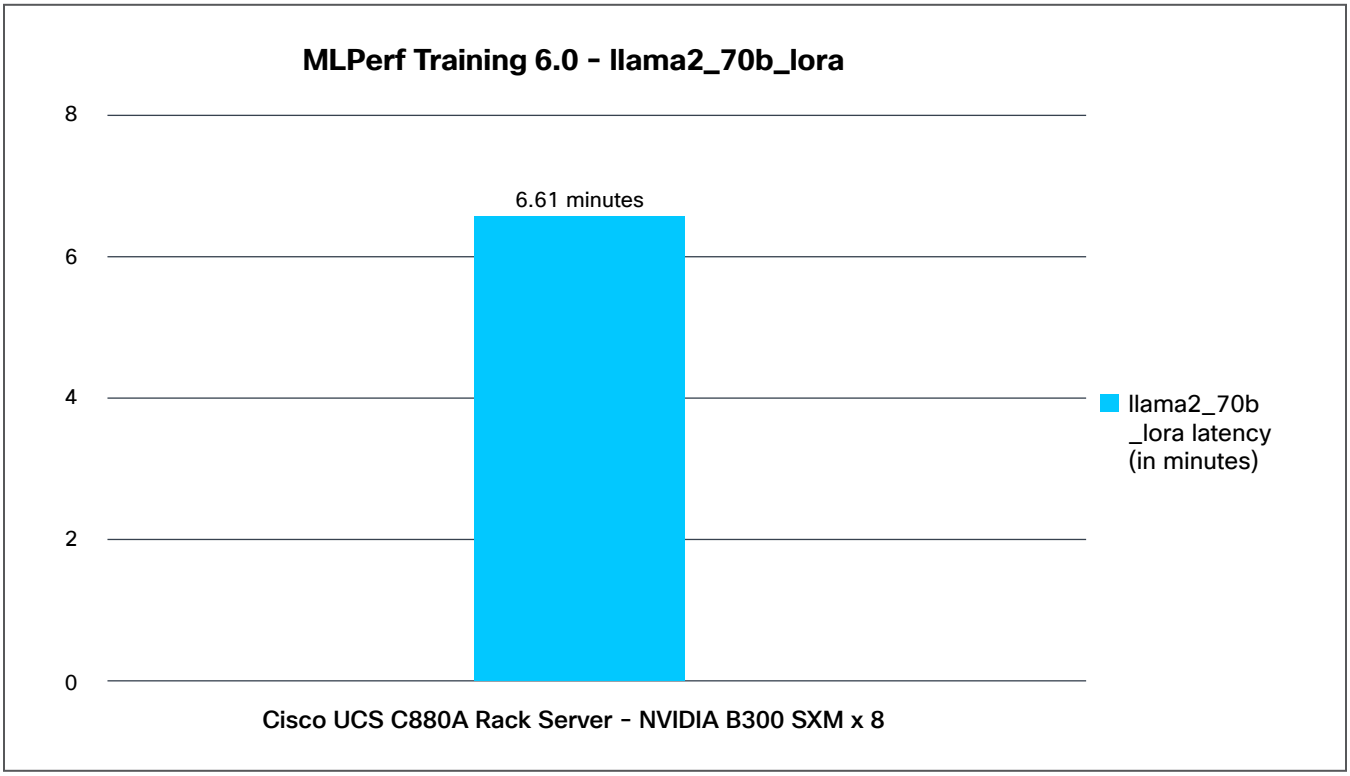


Figure 4. Llama2_70b_lora performance data on a Cisco UCS C880A M8 Rack Server with 8 x NVIDIA B300 GPUs

Llama3.1_8b

Llama 3.1_8b is an open-source, lightweight, and ultra-fast large language model developed by Meta AI. It contains 8 billion parameters and is designed to efficiently handle a wide variety of multilingual text generation and natural language processing tasks.

Figure 5 shows the MLPerf Training 6.0 performance of the Retinanet model tested on Cisco UCS C880A M8 Rack Server with 8x NVIDIA B300 GPUs.

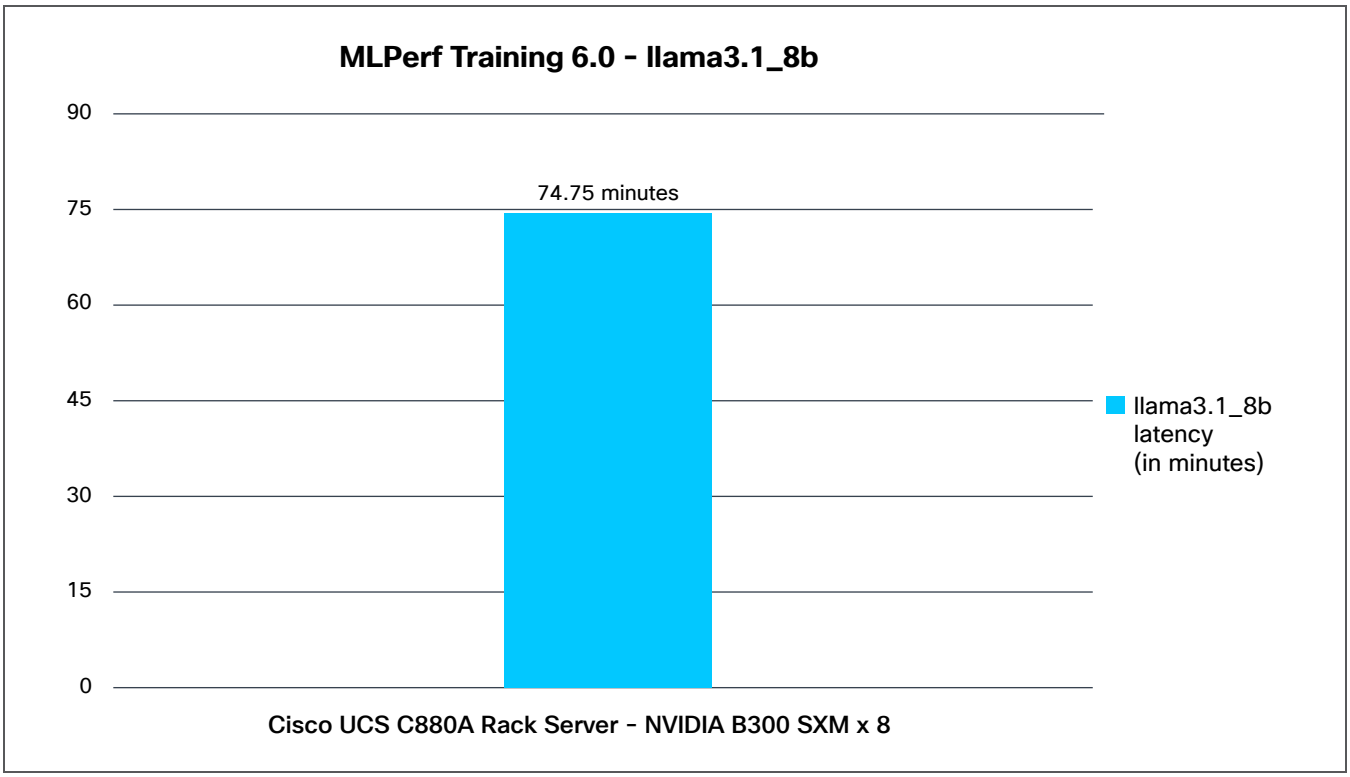


Figure 5. Llama3.1_8b performance data on a Cisco UCS C880A M8 Rack Server with 8 x NVIDIA B300 GPUs

Gpt-oss-120b

Gpt-oss-120b generally refers to open-source or open-weight implementations inspired by the GPT (Generative Pre-trained Transformer) architecture. The term is often used by the AI community to describe projects that aim to provide GPT-like language models with openly available code and/or weights.

Figure 6 shows the MLPerf Training 6.0 performance of the gpt-oss-20b model tested on a Cisco UCS C880A M8 Rack Server with 8x NVIDIA B300 GPUs.

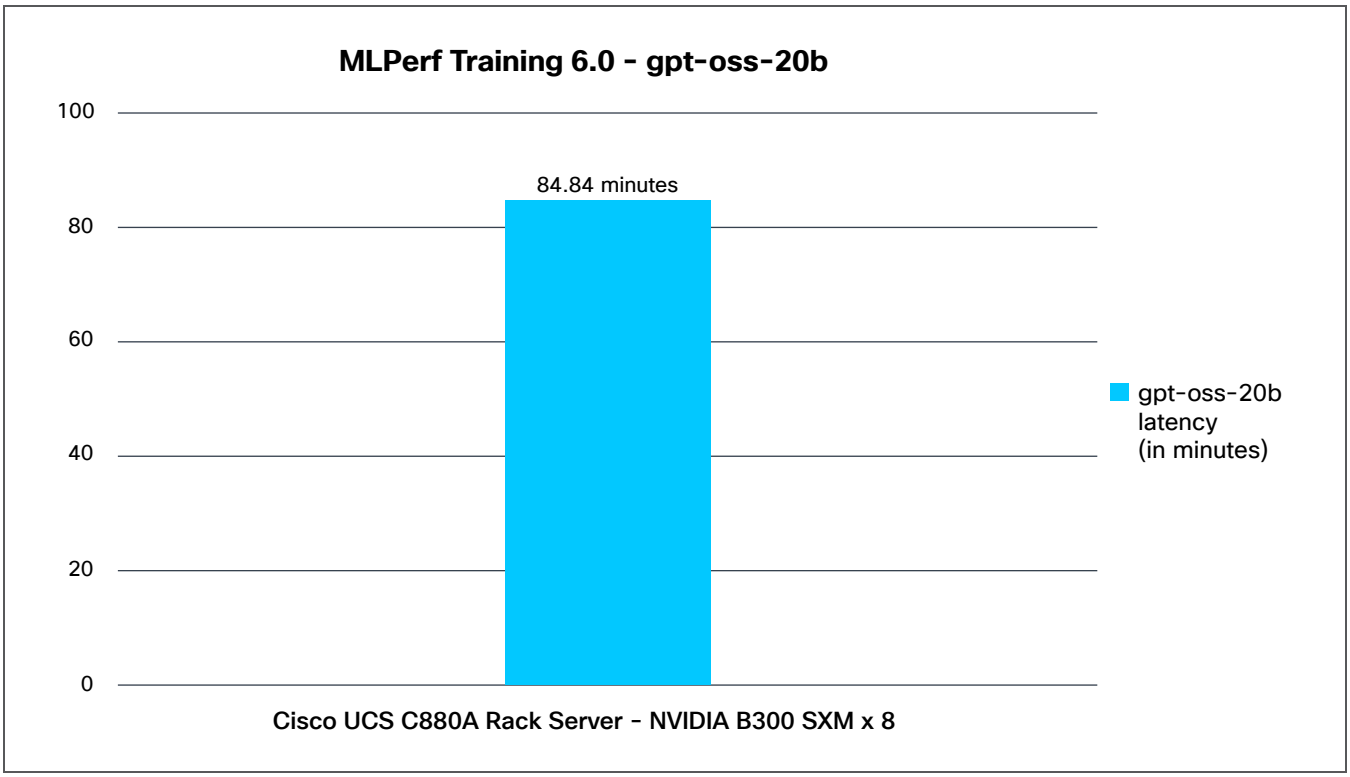


Figure 6. Gpt-oss-120b performance data on a Cisco UCS C880A M8 Rack Server with 8 x NVIDIA B300 GPUs

MLPerf Training 6.0 multinode performance data

MLPerf Training multinode testing evaluates how efficiently systems can train machine-learning models across multiple interconnected computing nodes. This benchmarking suite, developed by MLCommons, aims to provide standardized metrics for comparing the performance of various hardware, software, and services in the context of distributed machine learning.

The benchmarks are continuously evolving to include new and emerging AI workloads, such as generative AI (GenAI), and MLPerf results highlight the importance of dedicated low-latency interconnects between GPUs in multi-GPU systems for optimal distributed deep-learning training. Training models on multiple nodes introduces complexities, primarily due to communication overhead between nodes. To achieve efficient scaling, several technologies and optimizations are employed, such as RDMA (remote direct memory access), that are crucial for optimizing cross-node GPU-to-GPU communication and distributing training jobs efficiently. Distributed training frameworks and libraries such as NCCL (NVIDIA Collective Communications Library) are commonly used for distributed training and efficient communication across GPUs and nodes.

Llama2_70b_lora

Llama2_70b_lora is a large language model from Meta, with 70 billion parameters. It is designed for various natural language processing tasks such as text generation, summarization, translation, and question answering.

Figure 7 shows the single-node and multi-node configuration for MLPerf 6.0 Training performance of the llama2_70b_lora model tested on a Cisco UCS C880A M8 Rack Server with 8x and 16x NVIDIA B300 GPUs.

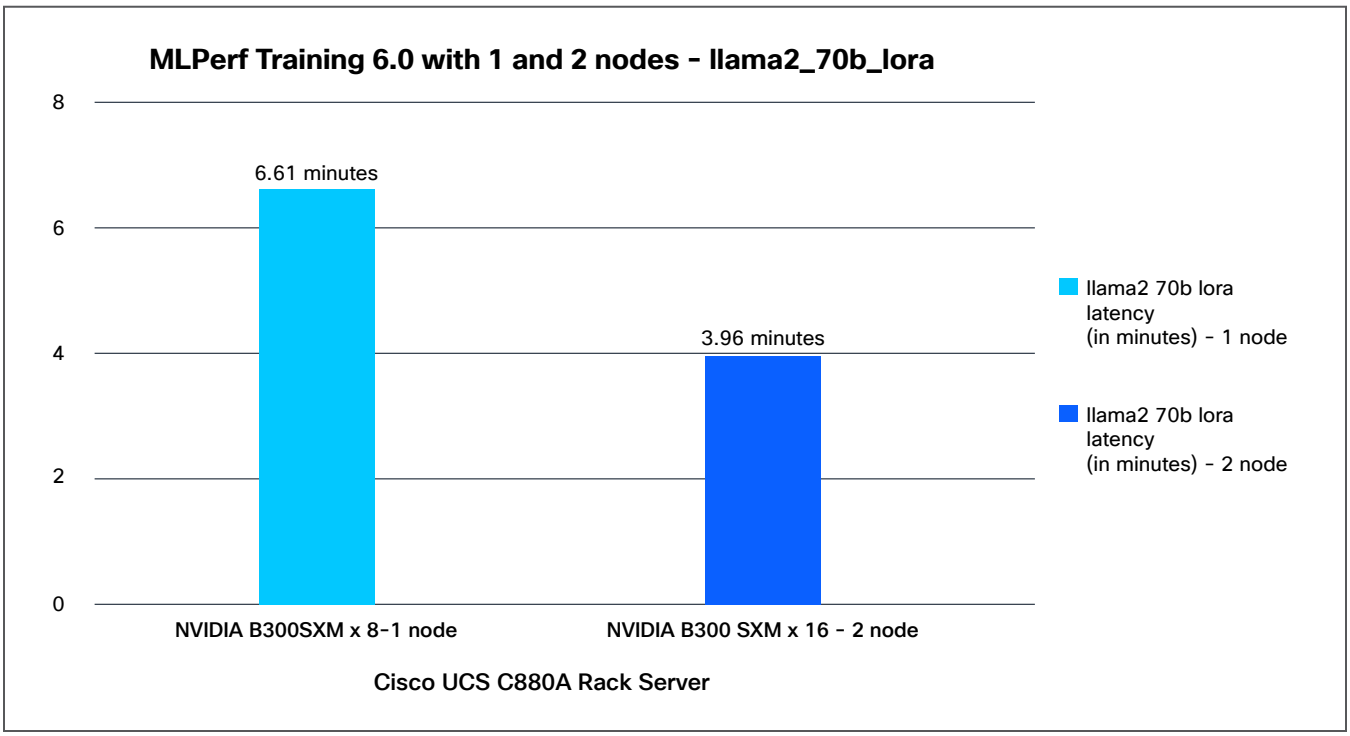


Figure 7. Multinode llama2_70b_lora performance data on a Cisco UCS C880A M8 Rack server with 8x and 16x NVIDIA B300 GPUs

Llama3.1_8b

Developed by Meta AI, llama 3.1_8b is a lightweight, high-performance open-source large language model. With 8 billion parameters, it is engineered for efficient multilingual text generation and a broad range of natural language processing tasks.

Figure 8 shows MLPerf 6.0 Training performance for the llama 3.1_8b model, evaluated across both single-node and multinode configurations on the Cisco UCS C880A M8 Rack Server, utilizing 8x and 16x NVIDIA B300 GPUs.

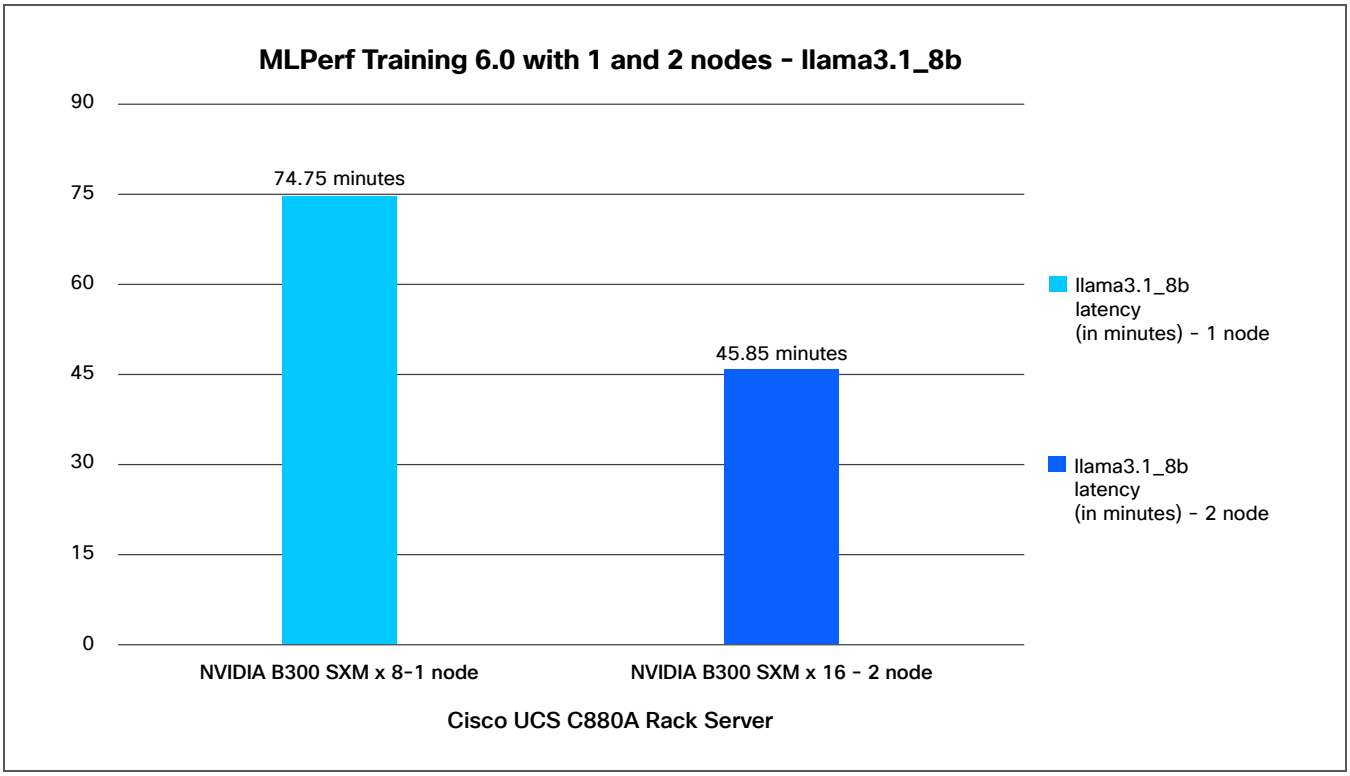


Figure 8. Multinode llama3.1_8b performance data on a Cisco UCS C880A M8 Rack Server with 8x and 16x NVIDIA B300 GPUs

Performance summary

The Cisco UCS C880A M8 Rack Server, powered by the NVIDIA HGX platform, provides the high-performance compute necessary for the most demanding AI workloads. By combining robust performance with simplified deployment, the platform enables organizations to accelerate time-to-value for their AI initiatives.

Cisco's commitment to AI excellence is further demonstrated through its collaborative MLPerf Training submissions with NVIDIA. These benchmarks validate optimized performance and efficiency across a wide spectrum of AI applications, including large language models, natural language processing, image classification, object detection, and graph classification.

The Cisco UCS C880A M8 platform has demonstrated industry-leading AI performance in the MLPerf Training 6.0 benchmark. Key highlights include:

- Llama2_70b_lora: delivered leadership performance in both single-node and multinode configurations utilizing 8x and 16x NVIDIA B300 SXM GPUs
- Llama3.1_8b: achieved top-tier results in a multinode configuration equipped with 16x NVIDIA B300 SXM GPUs.

These results underscore the exceptional capabilities of the Cisco UCS portfolio for demanding AI training workloads.

Appendix: Test environment

Table 2 lists the details of the server under test-environment conditions.

Table 2. Server properties

Description	Value
Product name	Cisco UCS C880A M8 Rack Server
CPU	2x Intel Xeon 6th Gen 6776P Processor
Number of cores	64
Number of threads	128
Total memory	4 TB
Memory DIMMs (16)	32x 128GB DDR5 RDIMM
Memory speed	6400 MHz
Network adapter	• 8x GPU-board integrated NVIDIA ConnectX-8 • 2x NVIDIA ConnectX-7 (2x200G) • 1x Intel X710-T2L OCP
GPU controllers	NVIDIA B300 SXM 8-GPU
SFF NVMe SSDs	Up to 8x PCIe Gen5 x4 E1.S NVMe SSD

Note: Platform-default BIOS settings were applied during the MLPerf Training validation.

For more information

For additional information on the Cisco UCS C880A M8 Rack Server, refer to: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c880a-m8-rack-server-spec-sheet.pdf.

Cisco UCS C880A M8 Rack Server Data Sheet: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c880a-m8-rack-server-ds.html.

Cisco AI-Ready Data Center Infrastructure: https://blogs.cisco.com/datacenter.

Cisco AI PODs At-a-Glance: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ai-pods-aag.html.

Cisco AI-Native Infrastructure for Data Center: https://www.cisco.com/site/us/en/solutions/artificial-intelligence/infrastructure/index.html.

C11-6470201-00 07/26