AI Performance: MLPerf Inference on Cisco UCS C845A M8 Rack Server with NVIDIA GPUs

October 2026



© 2026 Cisco and/or its affiliates. All rights reserved.

Contents

Executive summary 3
Scope of this document 8
Product overview 8
MLPerf overview 10
MLPerf Training 10
MLPerf Inference: Datacenter 10
Test configuration 11
MLPerf Inference performance results 11
Performance data for NVIDIA H200 NVL PCIe GPU 12
Performance data for NVIDIA L40S PCIe GPU 20
Performance data for NVIDIA RTX PRO 6000 PCIe GPU 22
Performance data for NVIDIA RTX PRO 4500 PCIe GPU 26
Performance summary 29
Appendix: Test environment 30
For more information 31

© 2026 Cisco and/or its affiliates. All rights reserved.

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® C845A M8 Rack Server is a highly scalable, flexible, and customizable AI system based on the NVIDIA MGX reference design for accelerated computing. With support for Two (2) to Eight (8) NVIDIA or AMD PCIe GPUs and NVIDIA AI Enterprise software, it delivers high performance for a wide range of AI workloads – including generative AI fine-tuning, Retrieval-Augmented Generation (RAG), and inference.

The Cisco UCS C845A M8 Rack Server is designed to address the most demanding AI workloads. Now an integral component of Cisco AI PODs (Cisco Validated Designs for AI), the Cisco UCS C845A M8 provides a robust foundation for modern AI infrastructure, allowing organizations to easily scale their AI capabilities with confidence.

With support for 2, 4, 6, or 8 NVIDIA GPUs—including the NVIDIA RTX PRO 6000 Blackwell, NVIDIA H200 NVL, and NVIDIA L40S, and also support for AMD Instinct MI210 accelerators, this system offers unparalleled flexibility to meet the diverse needs of enterprises. Leveraging the sophistication of the MGX modular reference design, this platform is also future-ready, with next-generation NVIDIA GPUs expected to seamlessly integrate as they become available.

To help demonstrate the AI performance capacity of the new Cisco UCS C845A M8 Rack Server, MLPerf Benchmarking performance testing for Inference Datacenter was conducted by Cisco, using NVIDIA H200 NVL, L40S PCIe, RTX PRO 6000 and RTX PRO 4500 GPUs, as detailed later in this document.

Accelerated compute

A typical AI journey starts with training GenAI models with large amounts of data to build the model intelligence. For this important stage, the new Cisco UCS C845A M8 Rack Server is a powerhouse designed to tackle the most demanding AI-training tasks. With its high-density configuration of NVIDIA H200 NVL and L40S Tensor Core GPUs, coupled with the efficiency of NVIDIA MGX architecture, the UCS C845A 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.

© 2026 Cisco and/or its affiliates. All rights reserved.

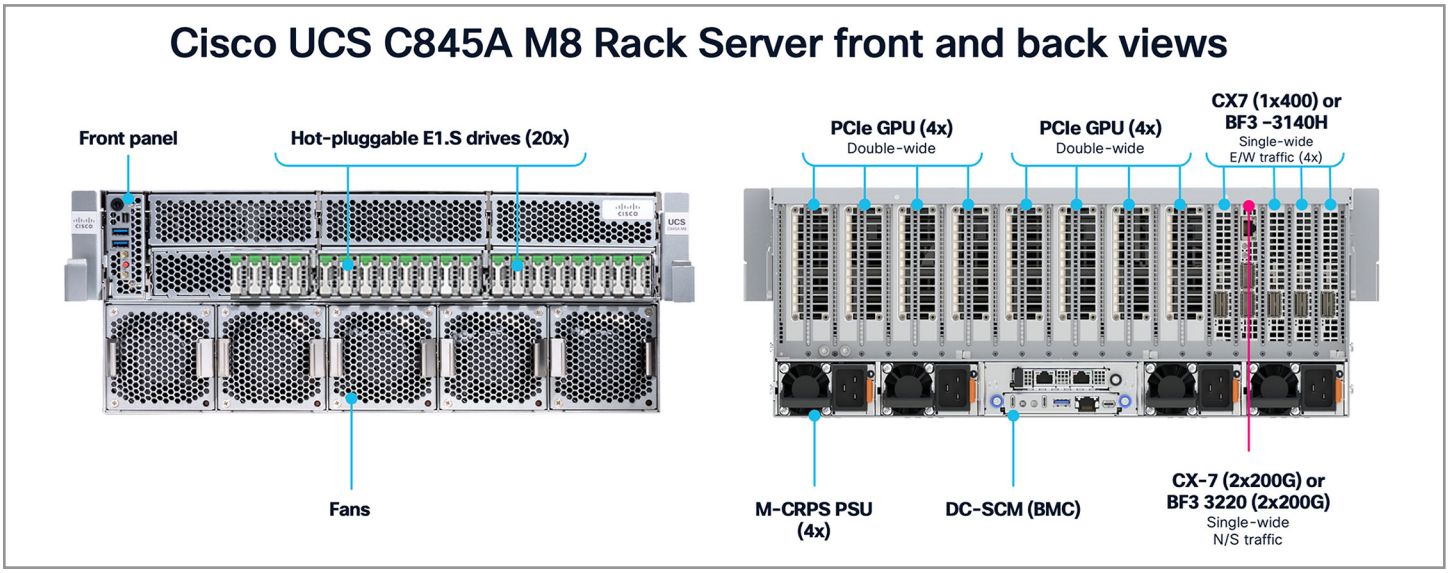


Figure 1. Cisco UCS C845A M8 Rack Server front and back views

Scalable network fabric for AI connectivity

Network Fabric: Cisco N9000 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 like the UCS C885A and C845A. The Cisco N9000 series is 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 AI POD architecture leverages the following key platforms:

- **Cisco N9332D-GX2B:** A 1RU, 32-port 400GbE switch based on Cisco Cloud Scale technology, ideally suited for leaf role.
- **Cisco N9364D-GX2A:** A 2RU, 64-port 400GbE switch based on Cisco Cloud Scale technology, ideally suited for larger leaf or spine roles.
- **Cisco N9364E-SG2** is 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 following design guide: “Cisco AI POD architecture design guide for AI workloads, including inference, training and fine-tuning.

© 2026 Cisco and/or its affiliates. All rights reserved.



The new Cisco N9364E-SG2 Switch provides 800G aggregation for AI connectivity

Figure 2. Cisco N9364E-SG2 Switch for AI connectivity

https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/nexus-9000-series-switches-ai-clusters-wp.html.

Purchasing simplicity

Once these powerful models are trained, you need infrastructure deployed for inferencing to provide actual value, often across a distributed landscape of data centers and edge locations. We have greatly simplified this process with new Cisco AI PODs that accelerate deployment of the entire AI infrastructure stack itself. AI PODs are designed to offer a plug-and-play experience with NVIDIA-accelerated computing. The pre-sized and pre-validated bundles of infrastructure eliminate the guesswork from deploying edge-inferencing, large-scale clusters, and other AI inferencing solutions.

Our goal is to enable customers to confidently deploy AI PODs with predictability around performance, scalability, cost, and outcomes, while shortening time to production-ready inferencing with a full stack of infrastructure, software, and AI toolsets. AI PODs include NVIDIA AI Enterprise, an end-to-end, cloud-native software platform that accelerates data-science pipelines and streamlines AI development and deployment. Managed through Cisco Intersight®, AI PODs provide centralized control and automation, simplifying everything from configuration to day-to-day operations, with more use cases to come.

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.

© 2026 Cisco and/or its affiliates. All rights reserved.

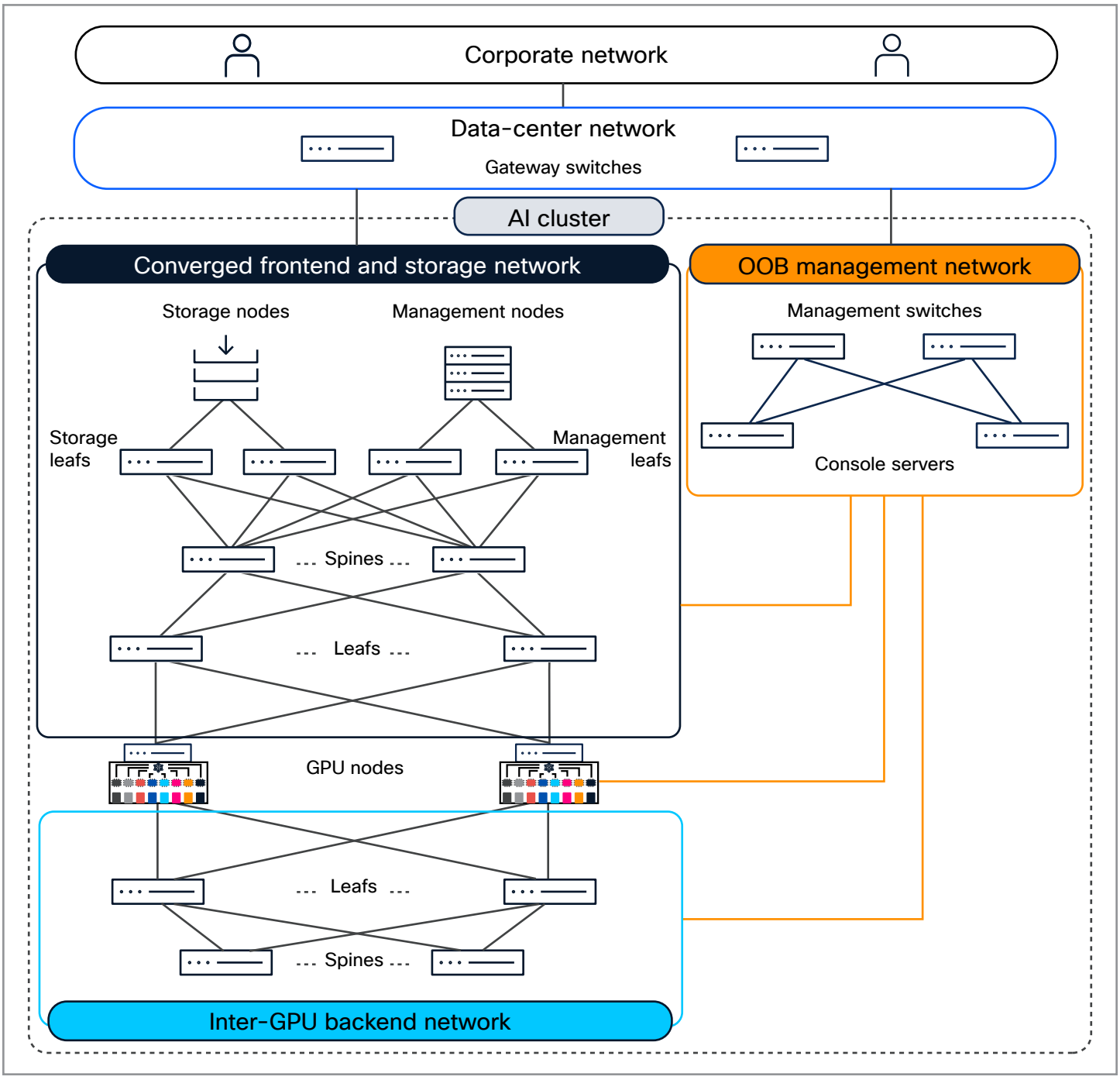


Figure 3. AI-cluster network design

https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/nexus-9000-series-switches-ai-clusters-wp.html.

© 2026 Cisco and/or its affiliates. All rights reserved.

Rail-optimized network design

GPUs in a scalable unit are interconnected using 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 is the most compute-intensive phase, and Large Language Models (LLMs), deep learning, Natural Language Processing (NLP), and digital twins require significant accelerated compute.

https://www.cisco.com/c/en/us/td/docs/dcn/whitepapers/cisco-addressing-ai-ml-network-challenges.html.

What use cases does the Cisco UCS C845A M8 Rack Server address?

The Cisco UCS C845A M8 Rack Server, a highly scalable and customizable server integrated into Cisco AI PODs, is engineered to drive a multitude of AI workloads. Its flexible GPU configurations enable it to address the most demanding AI challenges, including large deep learning, Large Language Model (LLM) training, model fine-tuning, large model inferencing, and Retrieval-Augmented Generation (RAG).

The platform's versatility is further enhanced by its support for various GPUs, each optimized for specific market needs:

NVIDIA H200 NVL:

- High-performance LLM inference
- Generative AI training and fine tuning

NVIDIA RTX PRO 6000 Blackwell:

- Agentic and physical AI: powering autonomous systems, smart factories, and robotics for real-time decision making
- Advanced scientific computing and rendering: accelerating complex simulations, medical imaging, and engineering analysis across hybrid environments
- High-fidelity 3D graphics and video: driving content creation, post-production, and immersive VR/AR experiences
- Hybrid AI/ML workflows: enabling seamless training, inferencing, and AI-assisted graphics across on-premises and cloud
- Edge-to-core AI applications: supporting real-time AI at the edge with centralized management and cloud integration

© 2026 Cisco and/or its affiliates. All rights reserved.

NVIDIA RTX PRO 4500 Blackwell Server Edition:

- **Department-scale AI and intelligent automation:** powering agentic workflows, industrial analytics, and real-time decision support in compact servers.
- **Mainstream scientific computing and simulation:** accelerating engineering analysis, digital twins, and data-heavy modeling with strong performance per watt.
- **Professional visualization and content pipelines:** enabling high-quality 3D visualization, design review, and video workflows for distributed teams.
- **Hybrid AI/ML workflows:** supporting efficient fine-tuning, inference, and AI-assisted graphics across on-prem and cloud environments.
- **Edge-to-core AI deployments:** delivering low-latency AI at remote sites with low power budget requirements, centralized policy, monitoring, and lifecycle management.

NVIDIA L40S:

- Generative AI foundation model fine-tuning
- Deployment of intelligent chatbots and search tools
- Language processing and conversational AI
- Graphics, rendering, and NVIDIA Omniverse applications
- Virtual desktop infrastructure

Scope of this document

For MLPerf Benchmark performance testing:

- Inference 5.1 and 6.1: Datacenter performance was measured on a Cisco UCS C845A M8 Rack Server equipped with 8 NVIDIA H200 NVL GPUs, 8 NVIDIA L40S PCIe GPUs and 8 NVIDIA RTX PRO 6000 and RTX PRO 4500 PCIe GPUs. Inference benchmark results were gathered across multiple datasets.
- These results help illustrate the inference performance advantages of the Cisco UCS C845A M8 Rack Server when deployed with these PCIe GPU configurations. This white paper highlights selected MLPerf Inference 5.1 and 6.1 performance results to provide a concise overview of the server's inference capabilities.

Product overview

Built on the NVIDIA MGX modular reference design, the Cisco UCS C845A M8 Rack Server is a flexible, scalable, and customizable AI system capable of growing as your AI needs grow. Configure with 2, 4, 6, or 8 PCIe GPUs to address a multitude of workloads ranging from generative AI, graphics and rendering, to virtual desktop infrastructure.

- UCS C845A M8 servers can be configured with two to eight NVIDIA GPUs. Depending on the configuration, customers can choose between the PCIe-based NVIDIA H200 NVL, L40S, RTXP6000 and RTXP4500 GPUs. Thanks to the sophistication of the MGX design, more "next-generation" NVIDIA GPUs are planned for introduction on this platform as they become available.

© 2026 Cisco and/or its affiliates. All rights reserved.

- With a compute node powered by AMD's new high-end EPYC Turin (5th Gen) CPUs, designed specifically for AI workloads, the UCS C845A M8 provides a no-compromise solution for CPU or GPU performance required to avoid bottlenecks within an AI server. Another benefit is the capability to configure the server with NVIDIA ConnectX-7 SmartNIC adapters and/or NVIDIA BlueField-3 DPUs to handle data traffic in and out of the server.

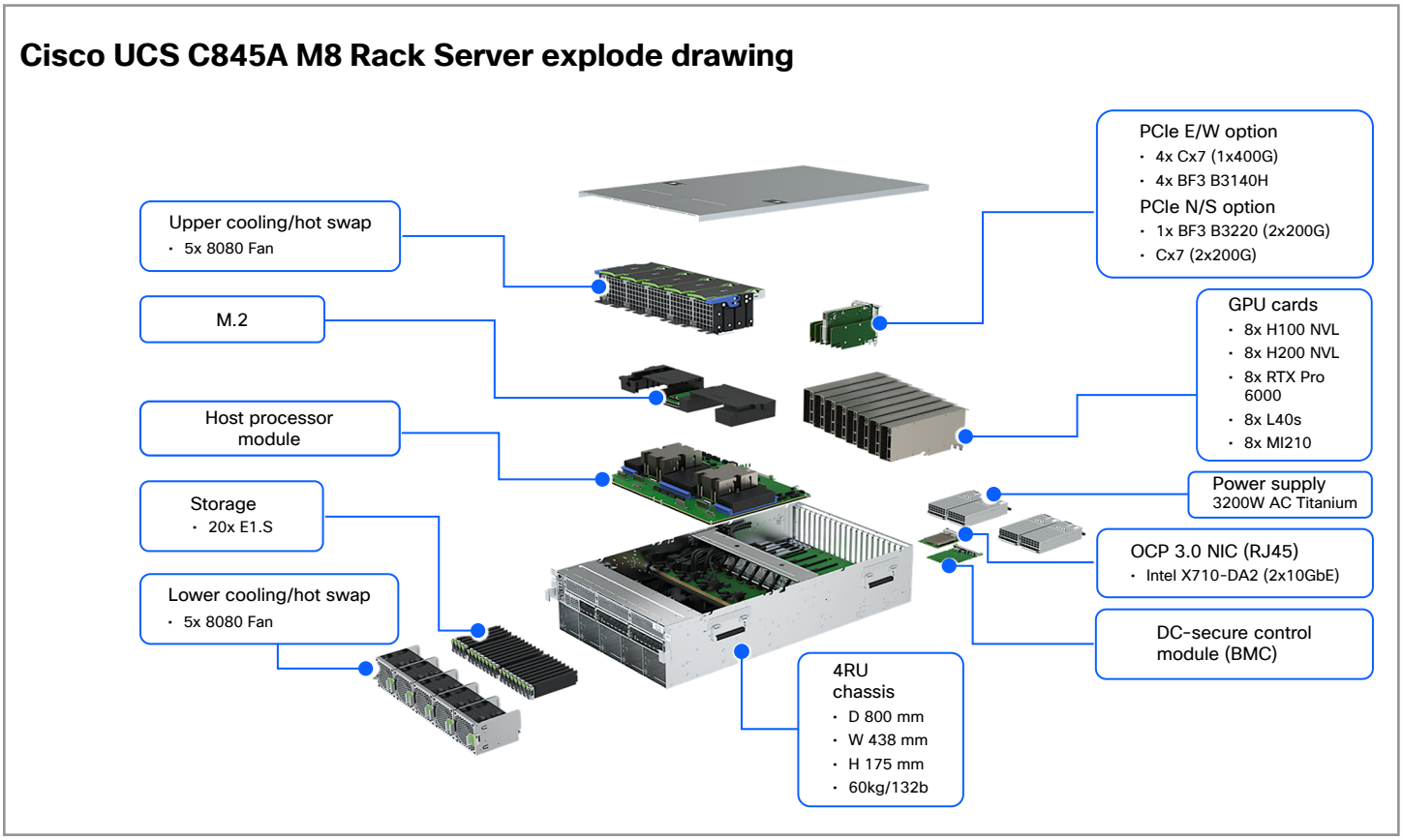


Figure 4. Components of the Cisco UCS C845A Rack Server

A specifications sheet for the Cisco UCS C845A M8 Rack Server is available at:

https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c845a-m8-rack-server-spec-sheet.pdf.

© 2026 Cisco and/or its affiliates. All rights reserved.

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 has multiple benchmarks, including:

- **MLPerf Training:** measures the time it takes to train machine-learning models to a target level of accuracy.
- **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 below mlcommons link.

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

https://mlcommons.org/benchmarks/training/.

MLPerf Inference: Datacenter

The MLPerf Inference: Datacenter benchmark suite measures how fast systems can process inputs and produce results using a trained model. Below mlcommons link gives summary of the current benchmarks and metrics.

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

https://mlcommons.org/benchmarks/inference-datacenter/.

© 2026 Cisco and/or its affiliates. All rights reserved.

Test configuration

For the MLPerf Inference performance testing covered in this document, the following configurations were used with Cisco UCS C845A M8 Rack Server:

- 8x NVIDIA H200 NVL PCIe GPUs
- 8x NVIDIA L40S PCIe GPUs
- 8x NVIDIA RTX PRO 6000 GPUs
- 8x NVIDIA RTX PRO 4500 GPUs

MLPerf Inference performance results

MLPerf Inference benchmarks

The MLPerf Inference models given in Table 1 were configured on a Cisco UCS C845A M8 Rack Server and tested for performance.

Table 1. MLPerf Inference models

Model	Reference implementation model	Description
Retinanet 800x800	vision/classification_and_detection	Single-stage object detection model optimized for detecting small objects in high-resolution images
Stable Diffusion XL	text_to_image	Generative model for creating high-quality images from text prompts
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	Multilingual Large Language Models (LLMs) with a collection of pretrained and instruction tuned generative models
Whisper	speech2text	Designed to enable not only transcriptions but also such tasks as language identification, phrase-level timestamps, and speech translation from other languages into English

© 2026 Cisco and/or its affiliates. All rights reserved.

Model	Reference implementation model	Description
Mixtral	language/mixtral-8x7b	The Mixtral-8x7B dataset refers to a mixture of experts model composed of multiple specialized sub-models, each having 7 billion parameters. This model uses a gating mechanism to activate relevant experts during training and inference to optimize performance across various tasks.
GPT-OSS	language/gpt-oss-120b	Open-source or open-weight implementations inspired by the GPT. Used to make powerful language models accessible to developers and researchers without requiring closed commercial APIs.

MLPerf Inference 5.1 performance data

As part of the MLPerf Inference 5.1 submission, Cisco has tested most of the datasets listed in Table 1 on the Cisco UCS C845A M8 Rack Server and submitted the results to MLCommons. The results are published on MLCommons results page: https://mlcommons.org/benchmarks/inference-datacenter/.

Note: The graph below includes unverified MLPerf Inference 5.1 results collected after the MLPerf submission deadline. For such data, there is an added note: “Result not verified by MLCommons Association.”

MLPerf Inference results are measured in both offline and server scenarios. The offline scenario focuses on maximum throughput, whereas the server scenario measures both throughput and latency, ensuring that a certain percentage of requests are served within a specified latency threshold.

Performance data for NVIDIA H200 NVL PCIe GPU

Retinanet

Retinanet is a single-stage object-detection model known for its focus on addressing class imbalances using a novel focal-loss function. The “800x800” refers to the input image size, and the model is optimized for detecting small objects in high-resolution images.

Figure 5 shows the performance of the Retinanet model tested on UCS C845A M8 Rack Server with NVIDIA 8x H200 NVL GPUs.

© 2026 Cisco and/or its affiliates. All rights reserved.

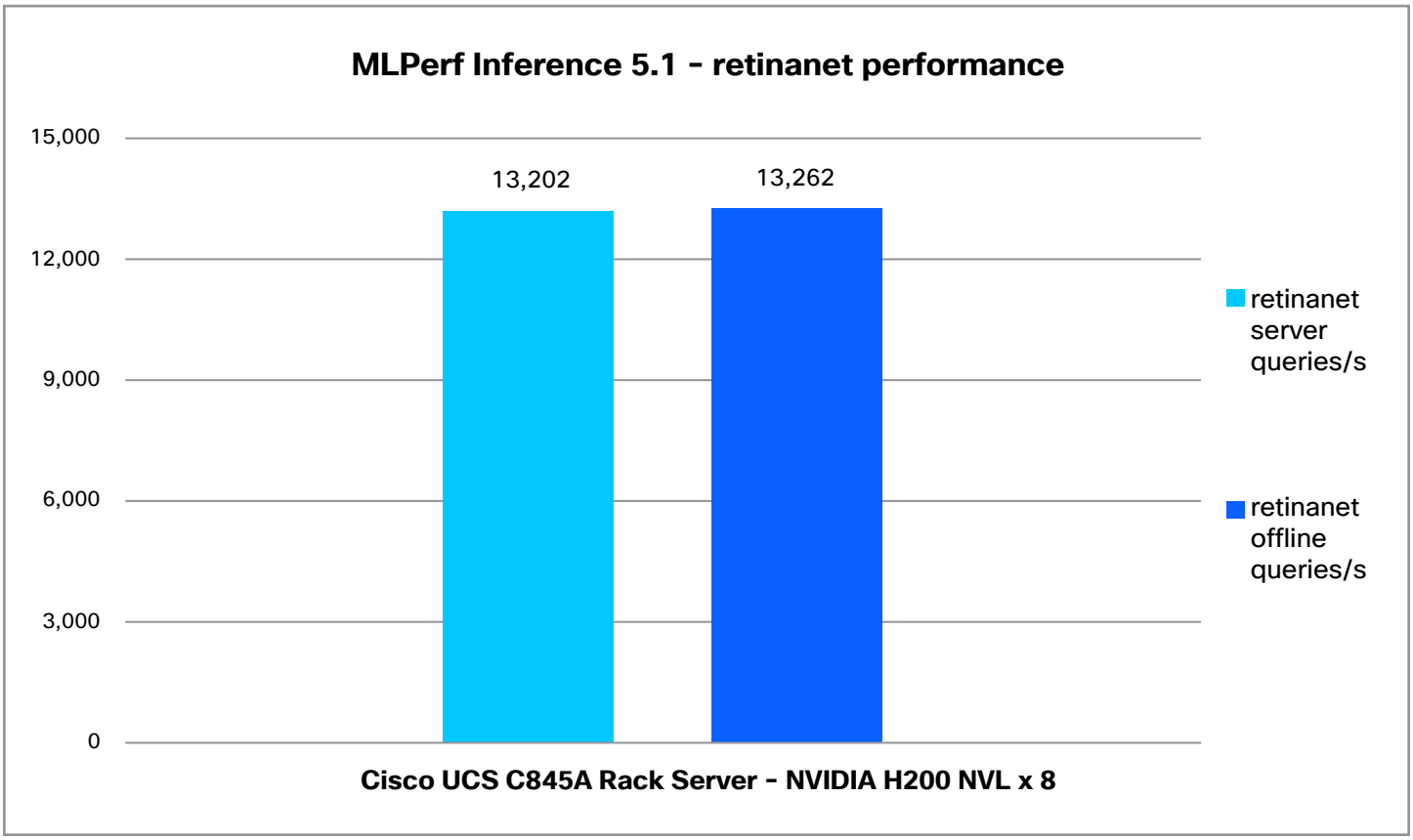


Figure 5. Retinanet performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

Llama3.1-8b

Llama3.1-8b is a powerful Large Language Model (LLM) with impressive capabilities in text generation, translation, and question answering. However, using cutting-edge LLMs often requires cloud resources. This tutorial empowers you to run the 8b version of Meta Llama3.1 directly on your local machine, giving you more control and privacy over your AI interactions.

© 2026 Cisco and/or its affiliates. All rights reserved.

Figure 6 shows the performance of the Llama3.1-8b model tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA H200 NVL GPUs.

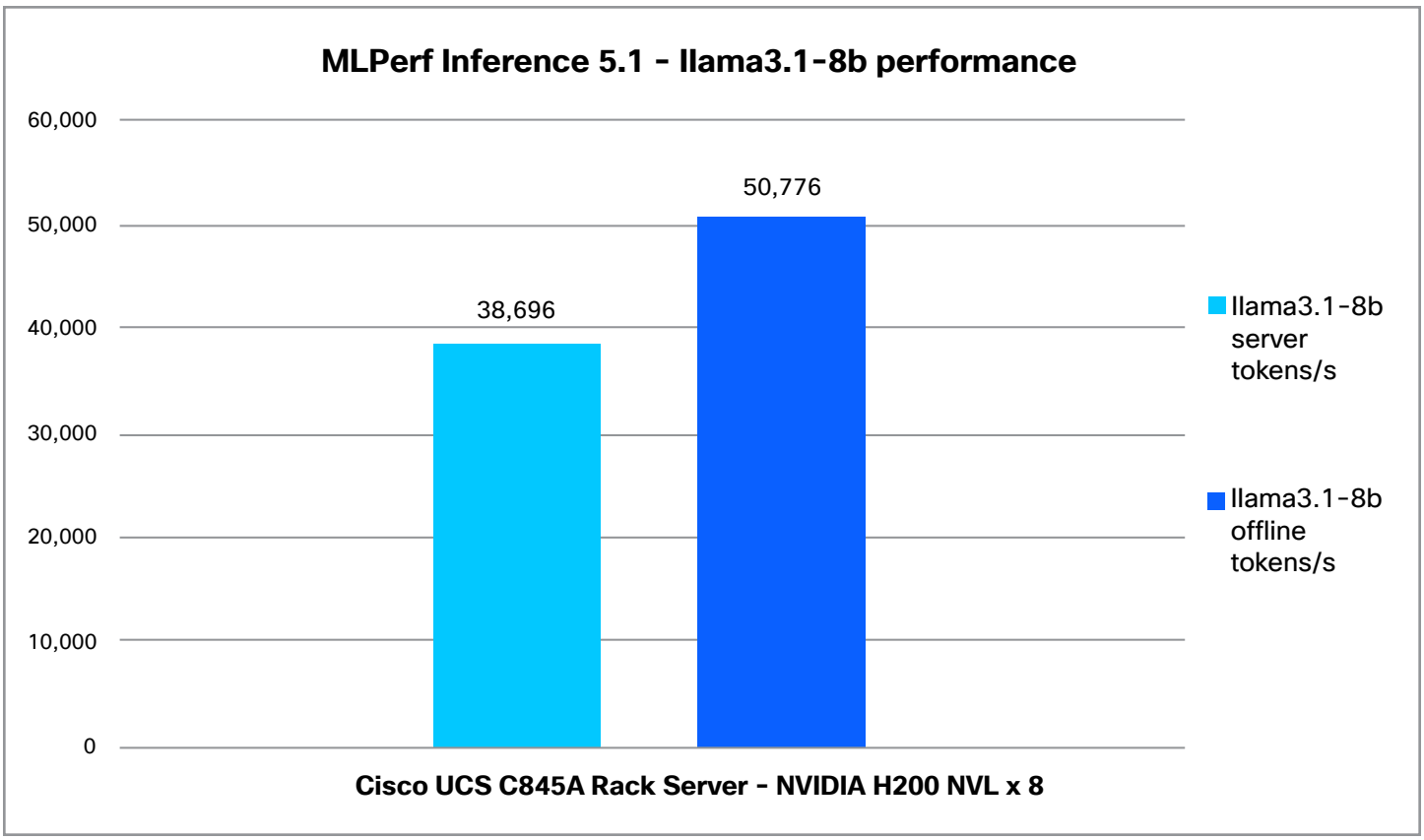


Figure 6. Llama3.1-8b performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

© 2026 Cisco and/or its affiliates. All rights reserved.

Llama2-70b (99)

Llama2-70b 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 answering questions.

Figure 7 shows the performance of the Llama2-70b model, with an accuracy of 99, tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA H200 NVL GPUs.

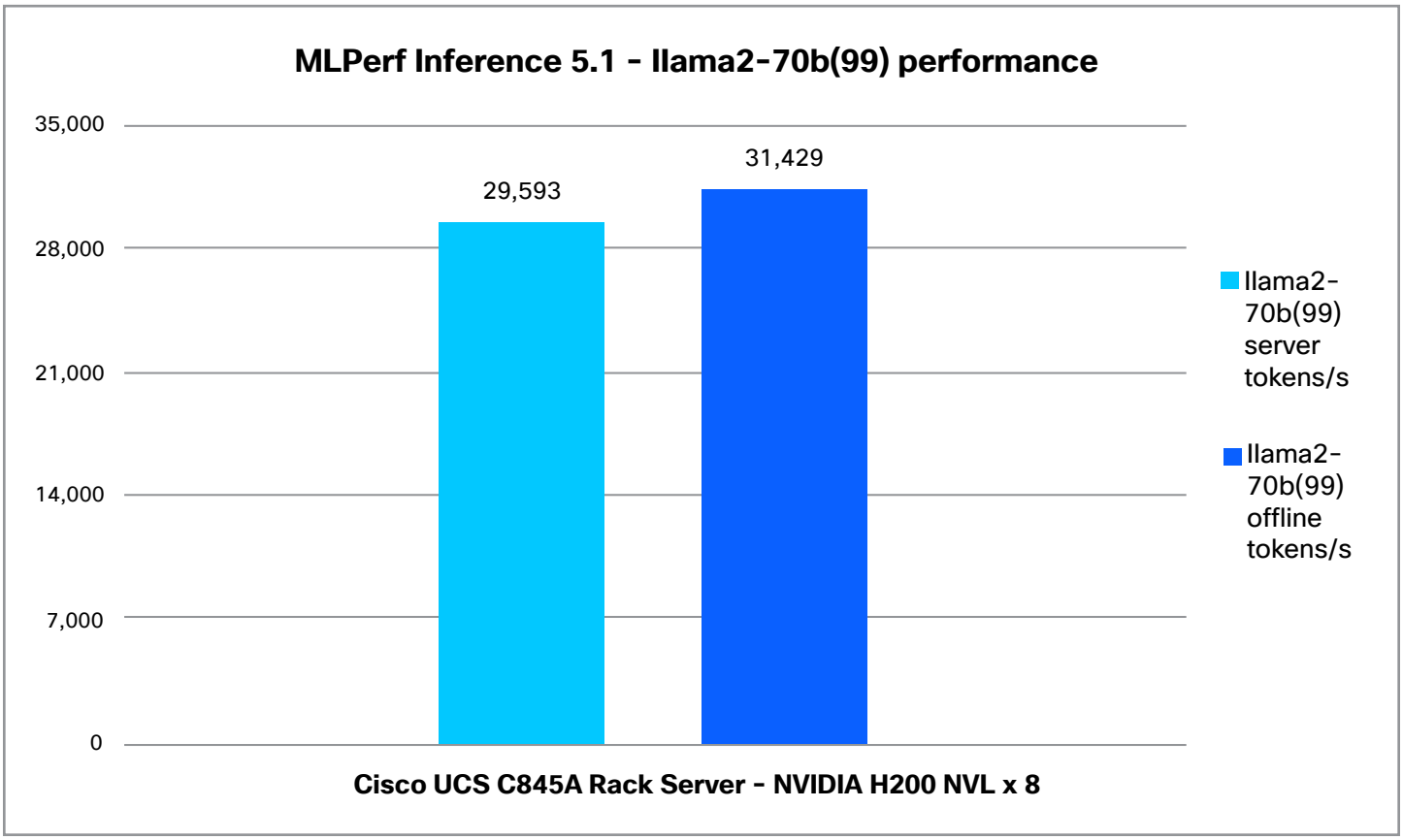


Figure 7. Llama2-70b performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

Note: For Llama2-70b performance data, the results have not been verified by MLCommons Association because the results were collected after the MLPerf submission deadline.

© 2026 Cisco and/or its affiliates. All rights reserved.

Llama2-70b (99.9)

Llama2-70b (99.9) 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 answering questions.

Figure 8 shows the performance of the Llama2-70b model, with a high accuracy of 99.9, tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA H200 NVL GPUs.

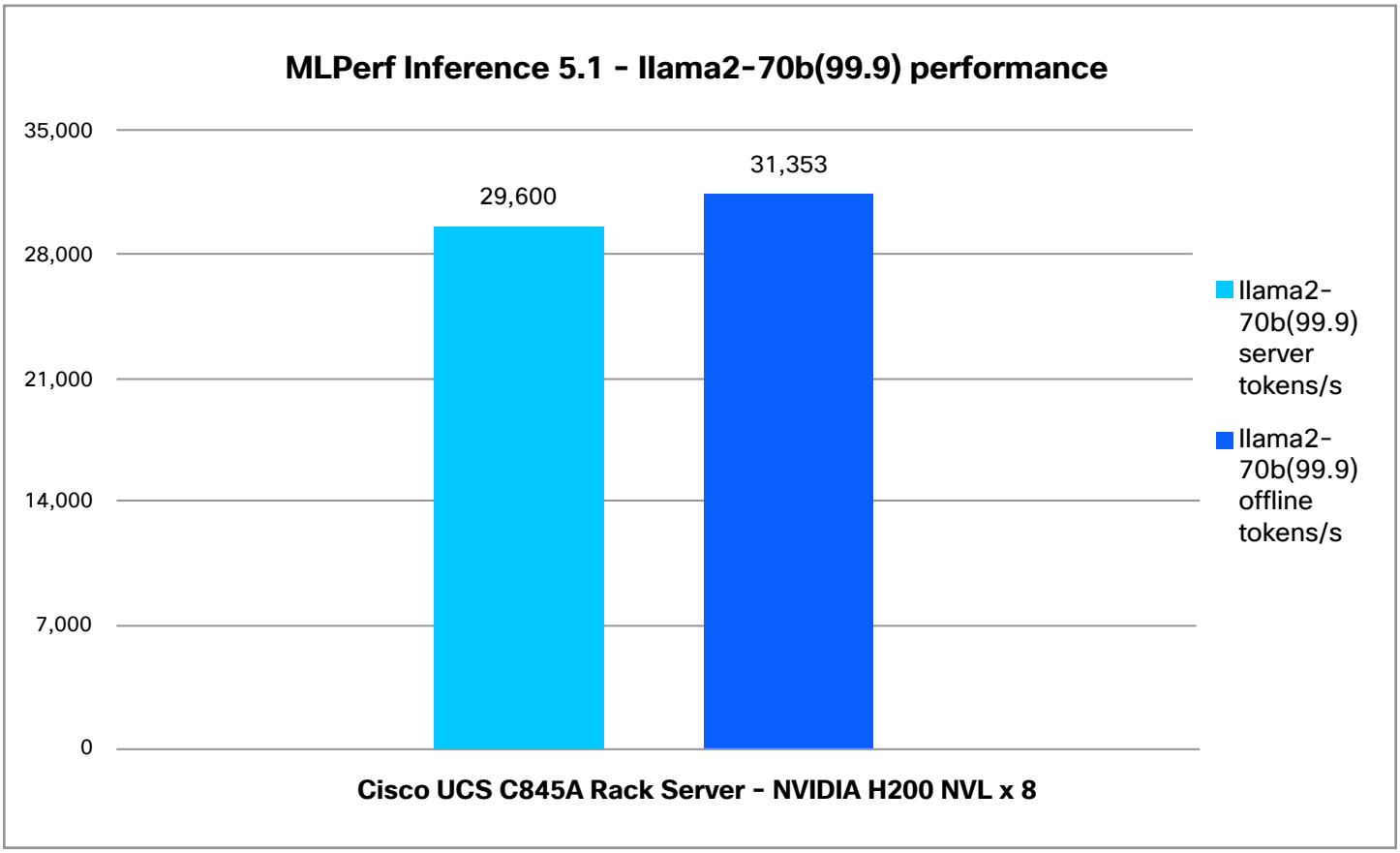


Figure 8. Llama2-70b (99.9) performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

Note: For Llama2-70b high-accuracy (99.9) performance data, the results have not been verified by MLCommons Association because the results were collected after the MLPerf submission deadline.

© 2026 Cisco and/or its affiliates. All rights reserved.

Llama2-70b Interactive

This model is essentially identical to the existing Llama2-70b workload, but for the tighter latency constraints in a server scenario.

Figure 9 shows the performance of the Llama2-70b Interactive model for 99 and 99.9 accuracy tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA H200 NVL GPUs.

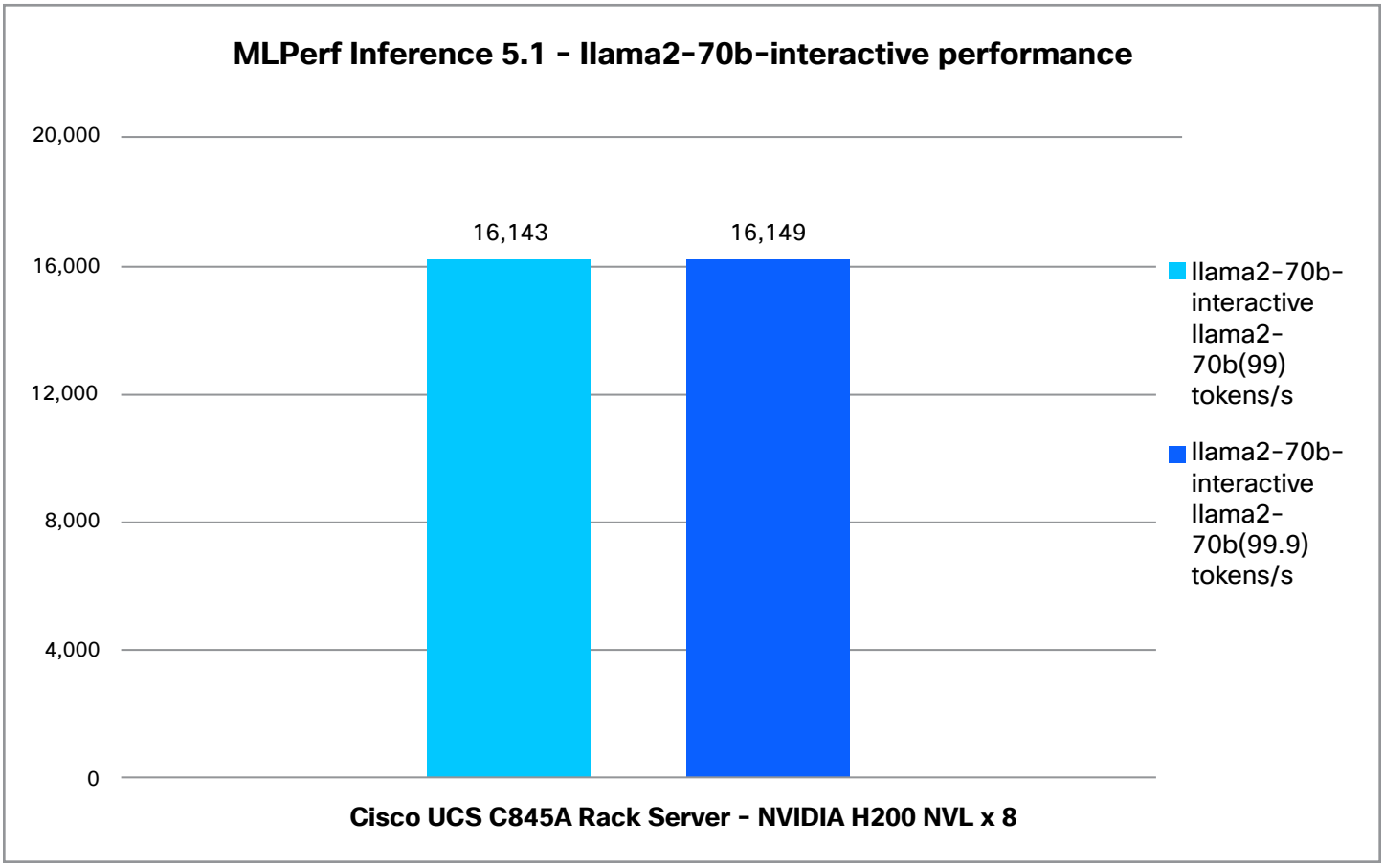


Figure 9. Llama2-70b Interactive performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

Note: For NVIDIA H200 NVL Llama2-70b Interactive performance data, the results have not been verified by MLCommons Association because the results were collected after the MLPerf submission deadline.

© 2026 Cisco and/or its affiliates. All rights reserved.

Stable Diffusion XL

Stable Diffusion XL is a generative model for creating high-quality images from text prompts. It is an advanced version of Stable Diffusion, offering larger models and better image quality, used for tasks such as image synthesis, art generation, and image editing.

Figure 10 shows the performance of the Stable Diffusion XL model tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA H200 NVL GPUs.

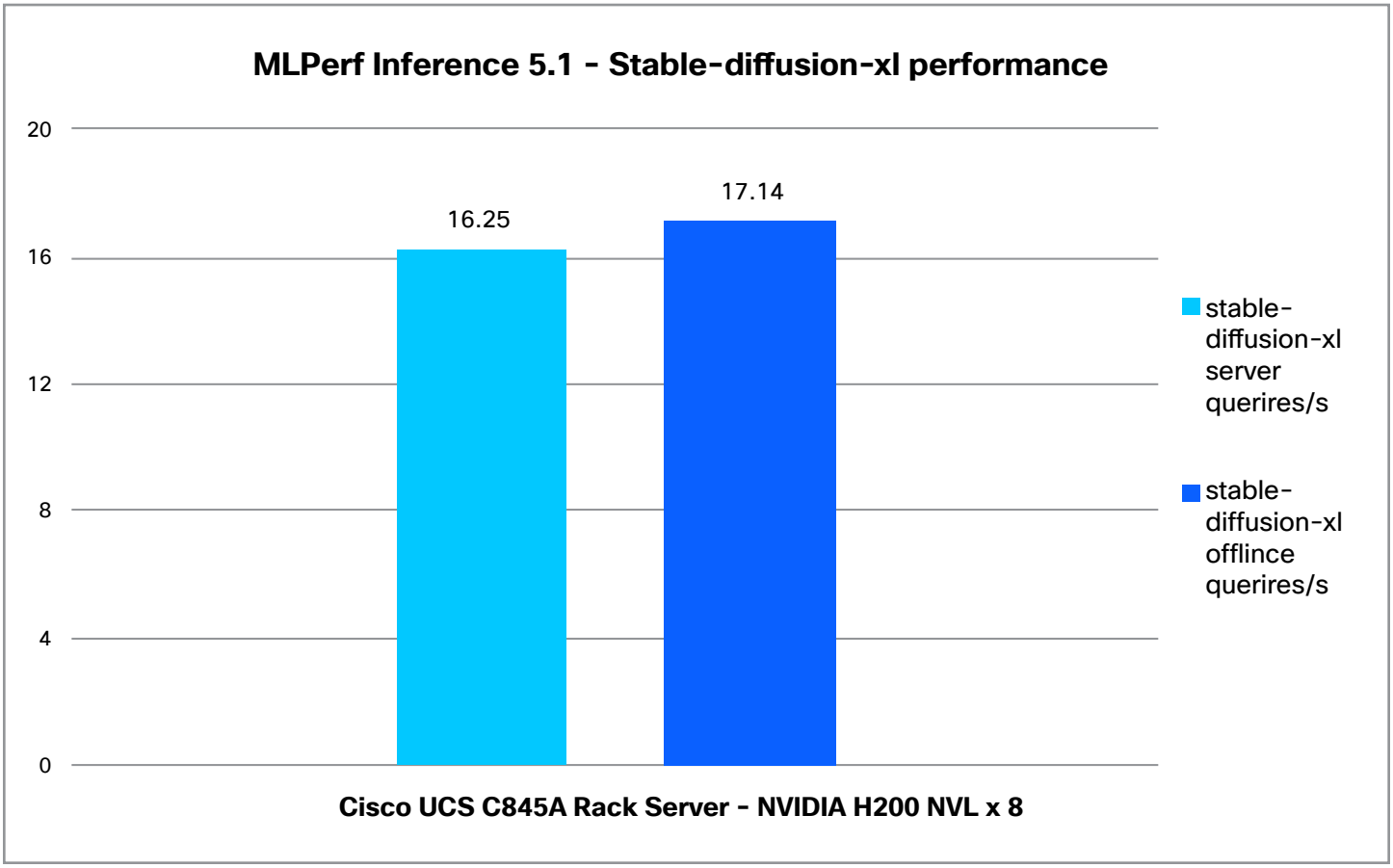


Figure 10. Stable Diffusion XL performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

Note: For NVIDIA H200 NVL Stable Diffusion XL performance data, the results have not been verified by MLCommons Association because the results were collected after the MLPerf submission deadline.

© 2026 Cisco and/or its affiliates. All rights reserved.

Whisper

Whisper is an automatic speech recognition model trained on 680,000 hours of multilingual data collected from the web. As per OpenAI, this model is robust on accents, background noise, and technical language. In addition, it supports transcription of 99 different languages and translation from those languages into English.

Figure 11 shows the performance of the Whisper model tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA H200 NVL GPUs.

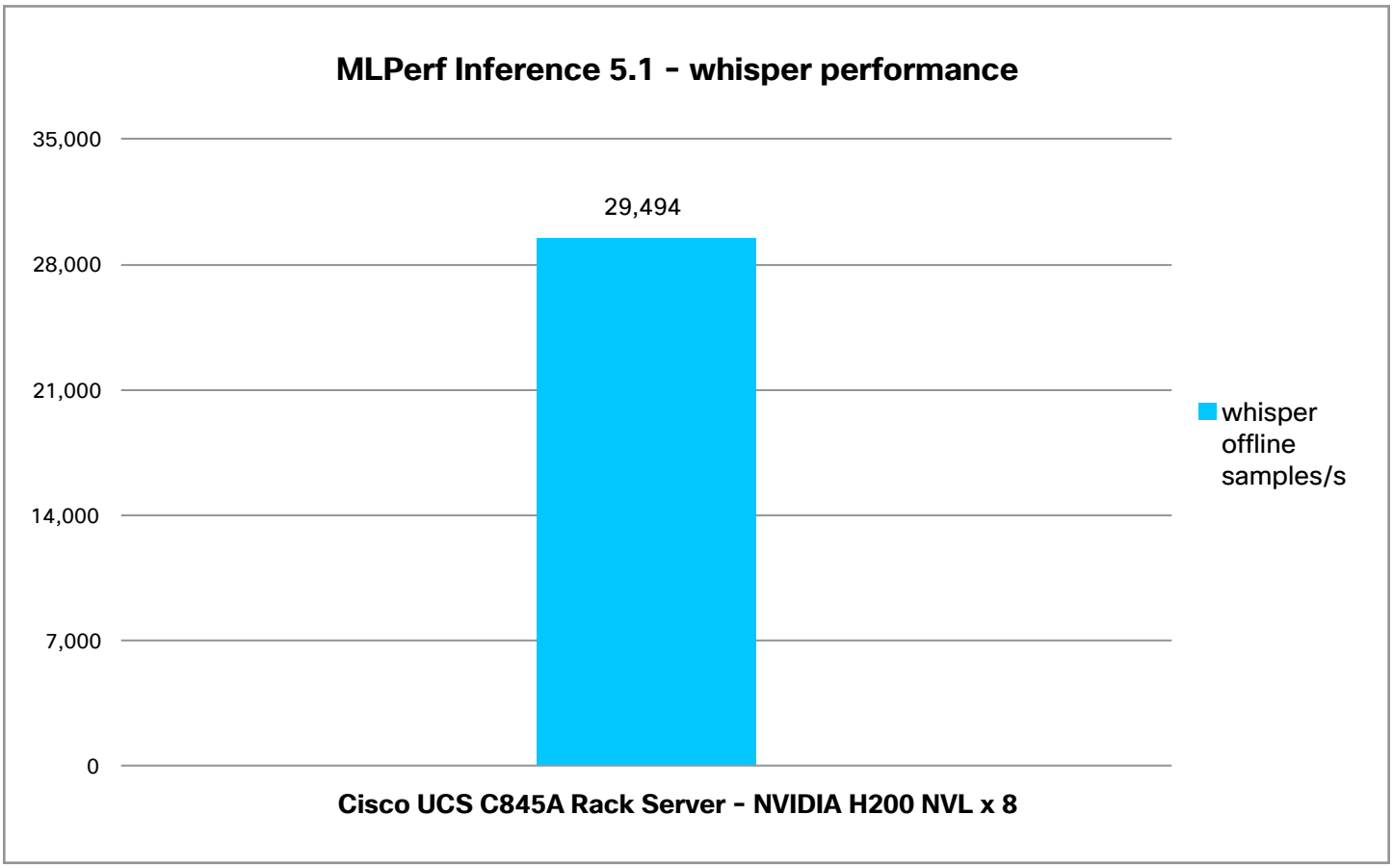


Figure 11. Whisper performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

Note: For NVIDIA H200 NVL Whisper performance data, the results have not been verified by MLCommons Association because the results were collected after the MLPerf submission deadline.

© 2026 Cisco and/or its affiliates. All rights reserved.

Performance data for NVIDIA L40S PCIe GPU

Retinanet

Retinanet is a single-stage object detection model known for its focus on addressing class imbalances using a novel focal-loss function. The “800x800” refers to the input image size, and the model is optimized for detecting small objects in high-resolution images.

Figure 12 shows the performance of the Retinanet model tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA L40S GPUs.

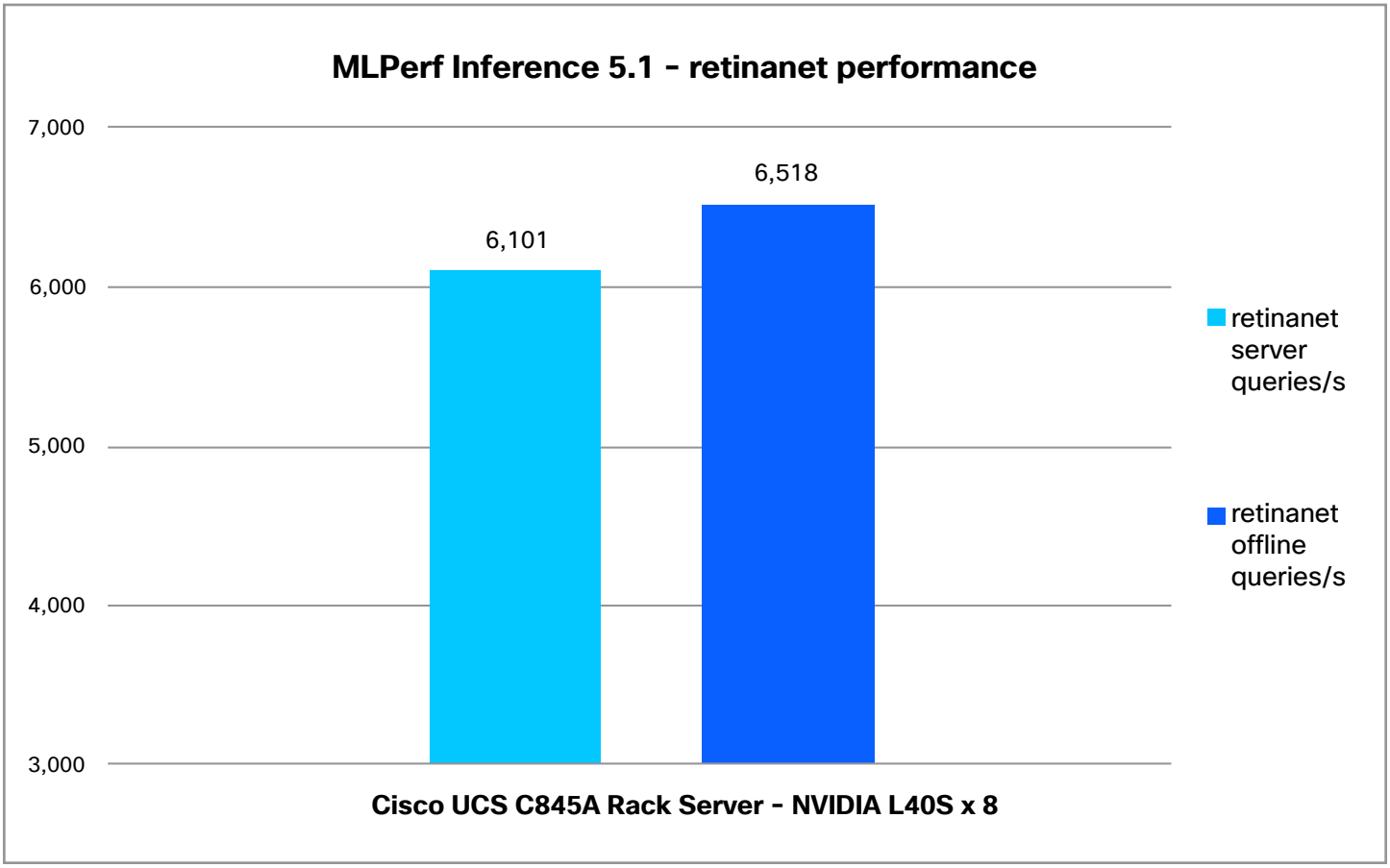


Figure 12. Retinanet performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA L40S GPUs

© 2026 Cisco and/or its affiliates. All rights reserved.

Whisper

Whisper is an automatic speech recognition model trained on 680,000 hours of multilingual data collected from the web. As per OpenAI, this model is robust on accents, background noise, and technical language. In addition, it supports the transcription of 99 different languages and translation from those languages into English.

Figure 13 shows the performance of the Whisper model tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA L40S GPUs.

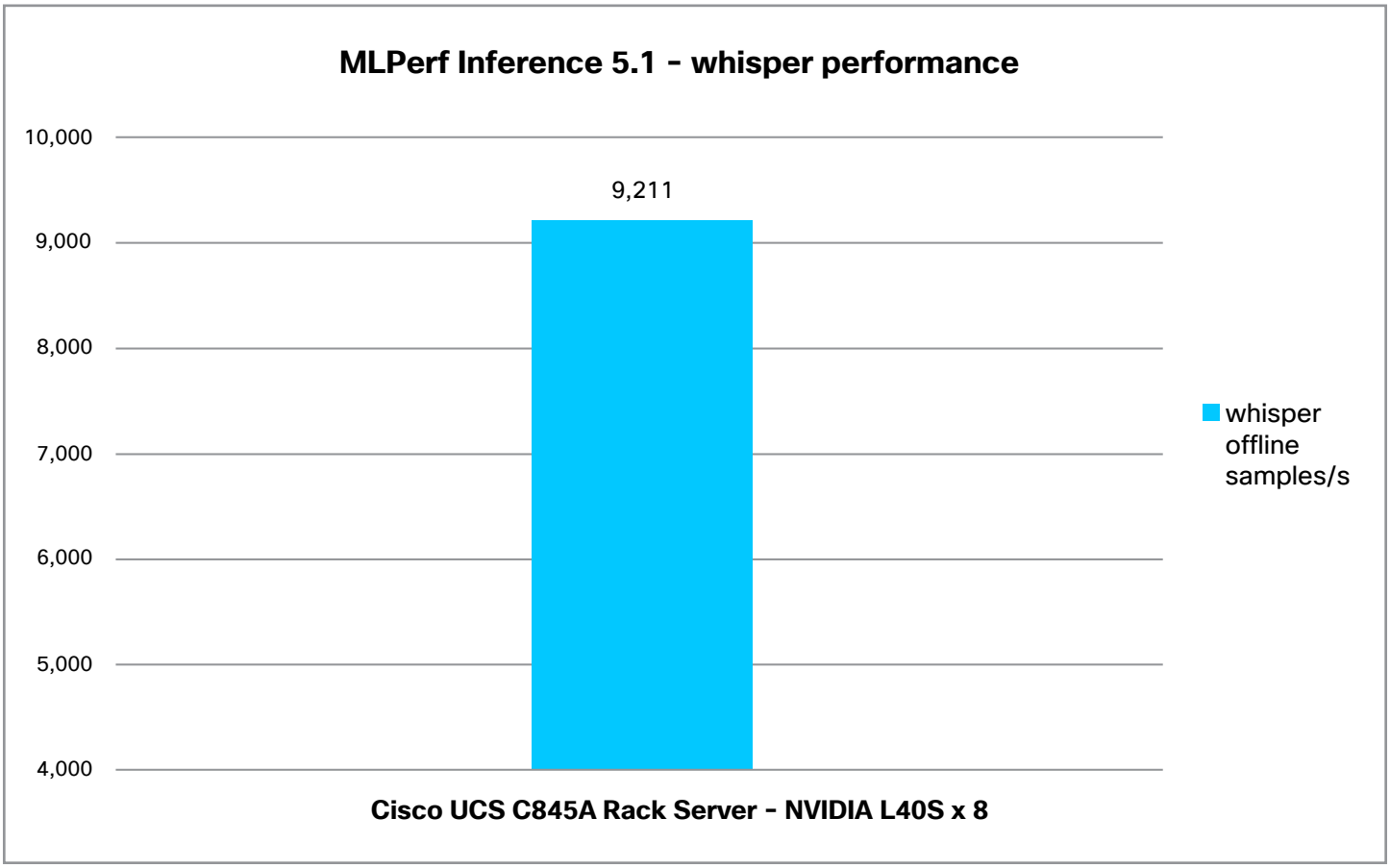


Figure 13. Whisper performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA L40S GPUs

© 2026 Cisco and/or its affiliates. All rights reserved.

MLPerf Inference 6.1 performance data

As part of the MLPerf Inference 6.1 submission, Cisco has tested selected datasets listed in Table 1 on the Cisco UCS C845A M8 Rack Server and submitted the results to MLCommons. The results are published on MLCommons results page: https://mlcommons.org/benchmarks/inference-datacenter/.

MLPerf Inference results are measured in both offline and server scenarios. The offline scenario focuses on maximum throughput, whereas the server scenario measures both throughput and latency, ensuring that a certain percentage of requests are served within a specified latency threshold.

Performance data for NVIDIA RTX PRO 6000 PCIe GPU

Llama3.1-8b

Llama3.1-8b is a powerful Large Language Model (LLM) with impressive capabilities in text generation, translation, and question answering. However, using cutting-edge LLMs often requires cloud resources. This tutorial empowers you to run the 8b version of Meta Llama3.1 directly on your local machine, giving you more control and privacy over your AI interactions.

Figure 14 shows the performance of the Llama3.1-8b model tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA RTX PRO 6000 GPUs.

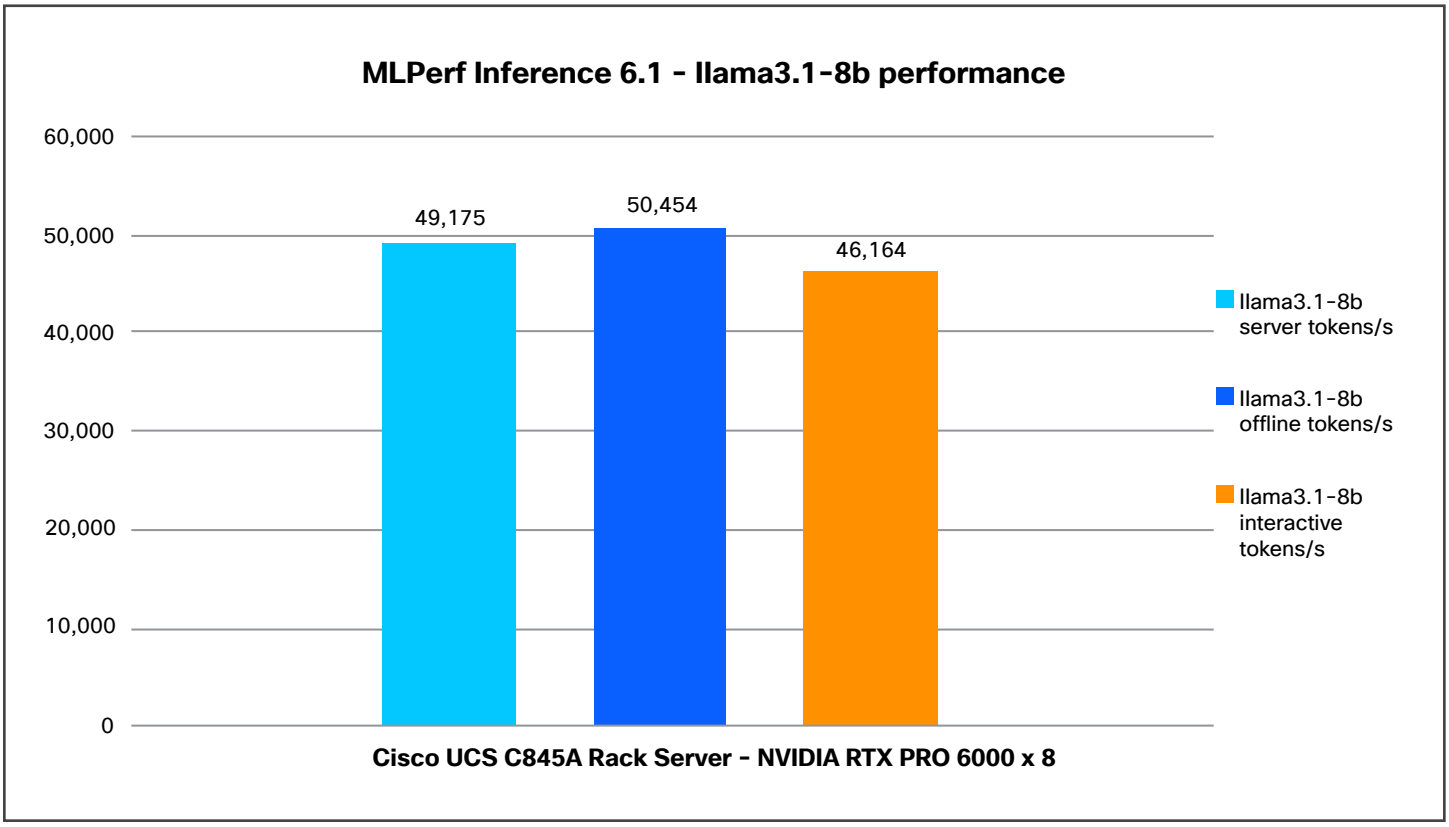


Figure 14. Llama3.1-8b performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA RTX PRO 6000 GPUs

© 2026 Cisco and/or its affiliates. All rights reserved.

Llama2-70b

Llama2-70b 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 answering questions.

Figure 15 shows the performance of the Llama2-70b model tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA RTX PRO 6000 GPUs.

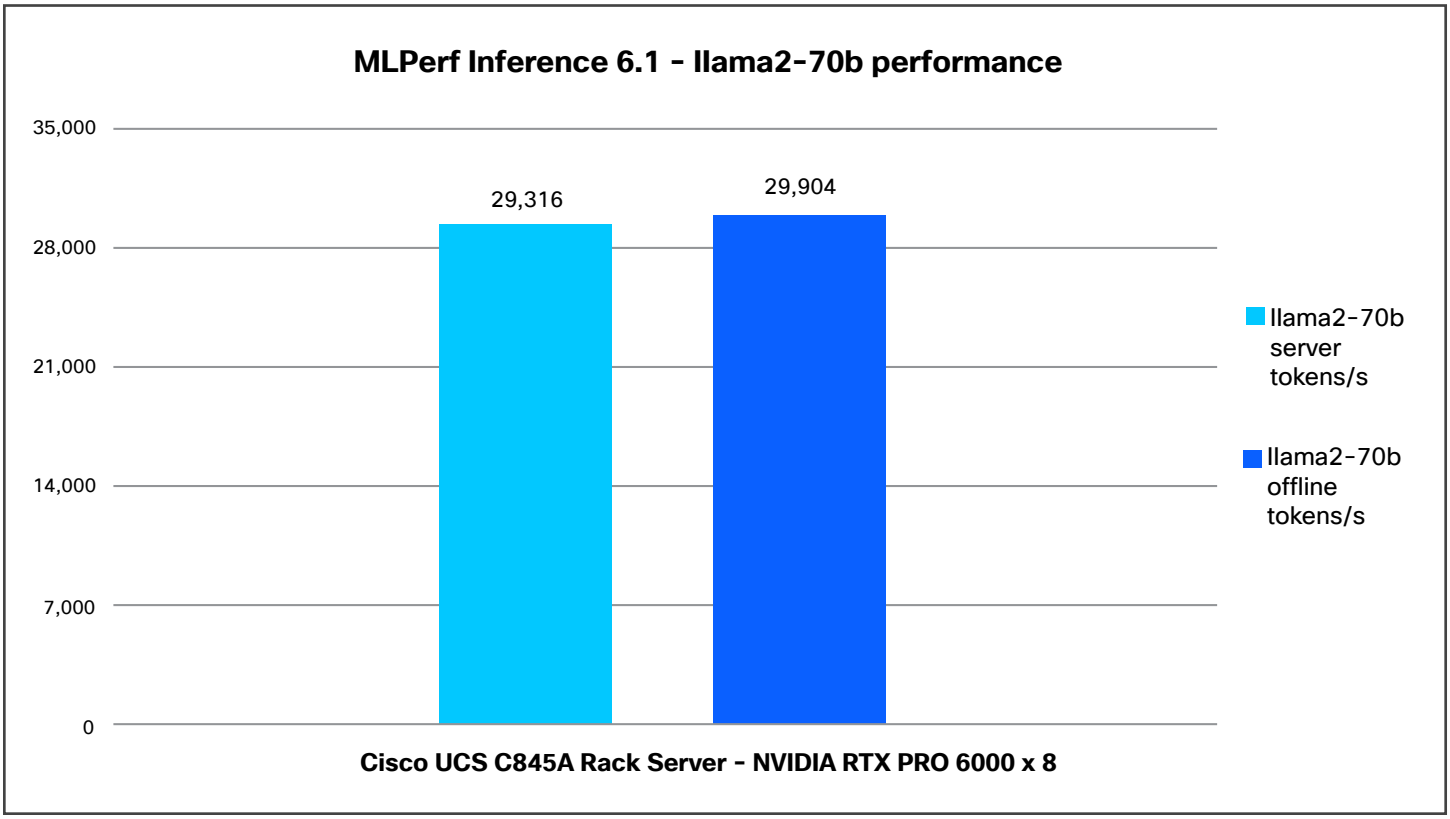


Figure 15. Llama2-70b performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA RTX PRO 6000 GPUs

© 2026 Cisco and/or its affiliates. All rights reserved.

gpt-oss-120b

Gpt-oss 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 16 shows the performance of the gpt-oss-120b model tested on a Cisco UCS C880A M8 Rack Server with 8x NVIDIA RTX PRO 6000 GPUs.

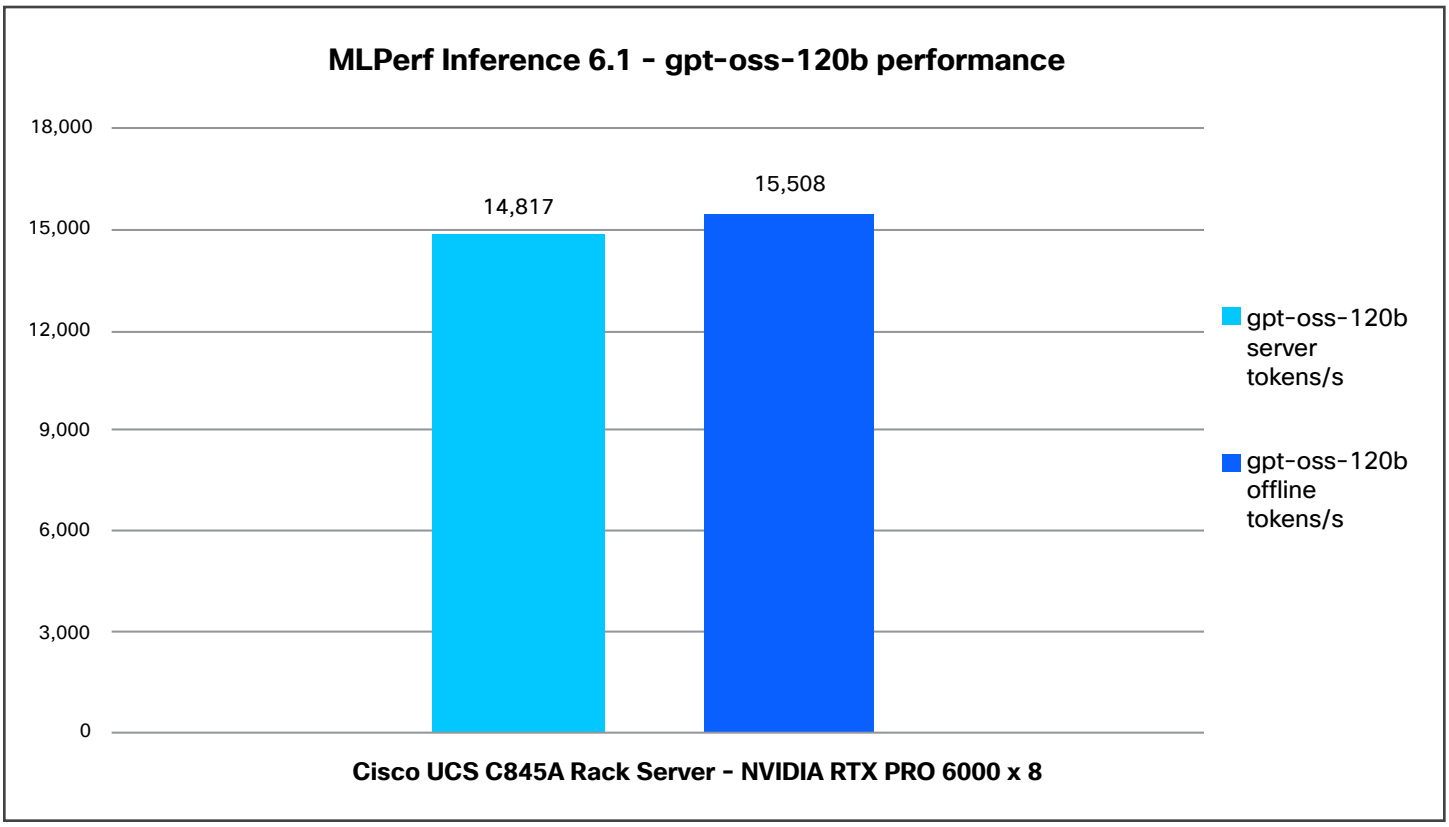


Figure 16. Gpt-oss-120b performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA RTX PRO 6000 GPUs

© 2026 Cisco and/or its affiliates. All rights reserved.

Whisper

Whisper is an automatic speech recognition model trained on 680,000 hours of multilingual data collected from the web. As per OpenAI, this model is robust on accents, background noise, and technical language. In addition, it supports the transcription of 99 different languages and translation from those languages into English.

Figure 17 shows the performance of the Whisper model tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA RTX PRO 6000 GPUs.

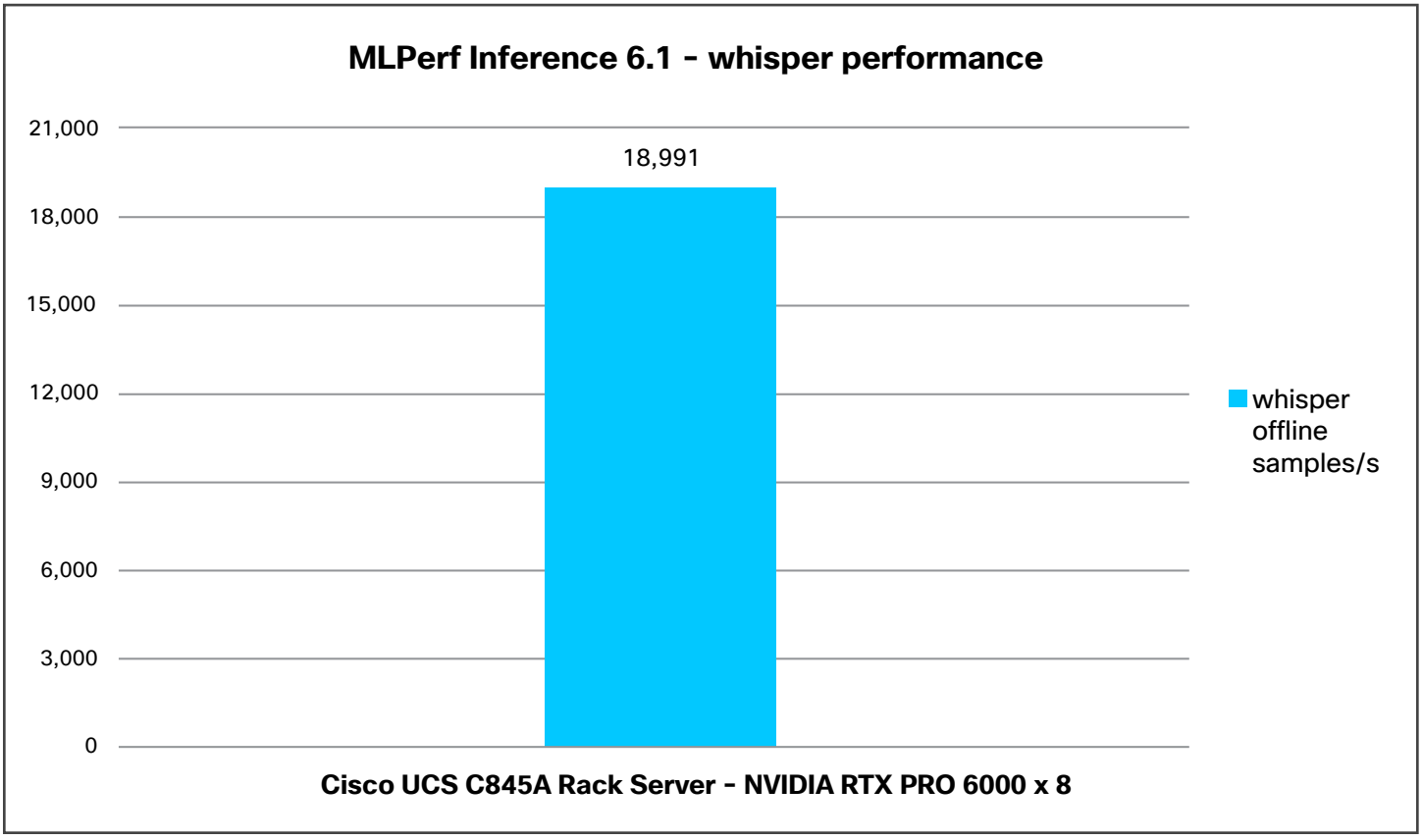


Figure 17. Whisper performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA RTX PRO 6000 GPUs

© 2026 Cisco and/or its affiliates. All rights reserved.

Performance data for NVIDIA RTX PRO 4500 PCIe GPU

Llama3.1-8b

Llama3.1-8b is a powerful Large Language Model (LLM) with impressive capabilities in text generation, translation, and question answering. However, using cutting-edge LLMs often requires cloud resources. This tutorial empowers you to run the 8b version of Meta Llama3.1 directly on your local machine, giving you more control and privacy over your AI interactions.

Figure 18 shows the performance of the Llama3.1-8b model tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA RTX PRO 4500 GPUs.

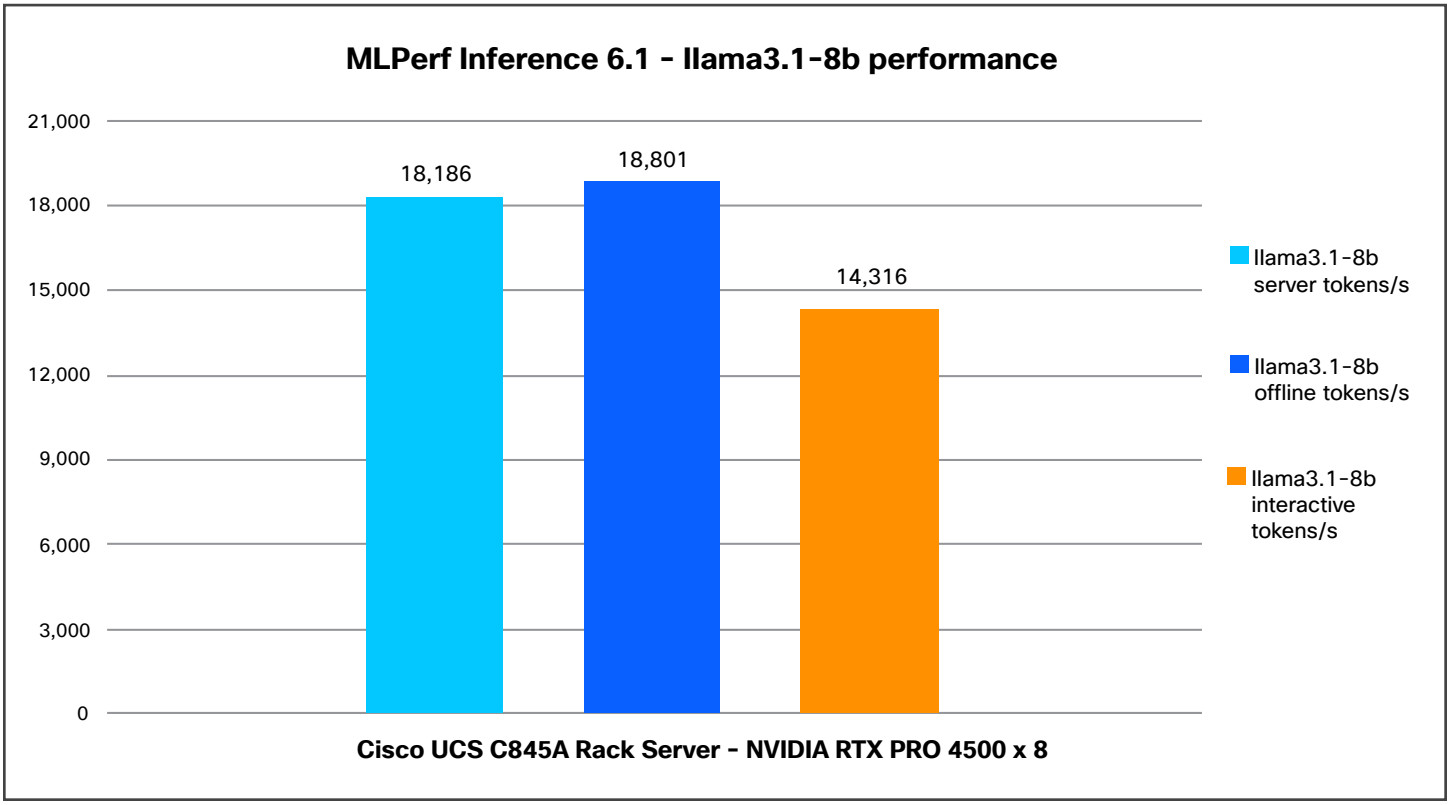


Figure 18. Llama3.1-8b performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA RTX PRO 4500 GPUs

© 2026 Cisco and/or its affiliates. All rights reserved.

Llama2-70b

Llama2-70b 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 answering questions.

Figure 19 shows the performance of the Llama2-70b model tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA RTX PRO 4500 GPUs.

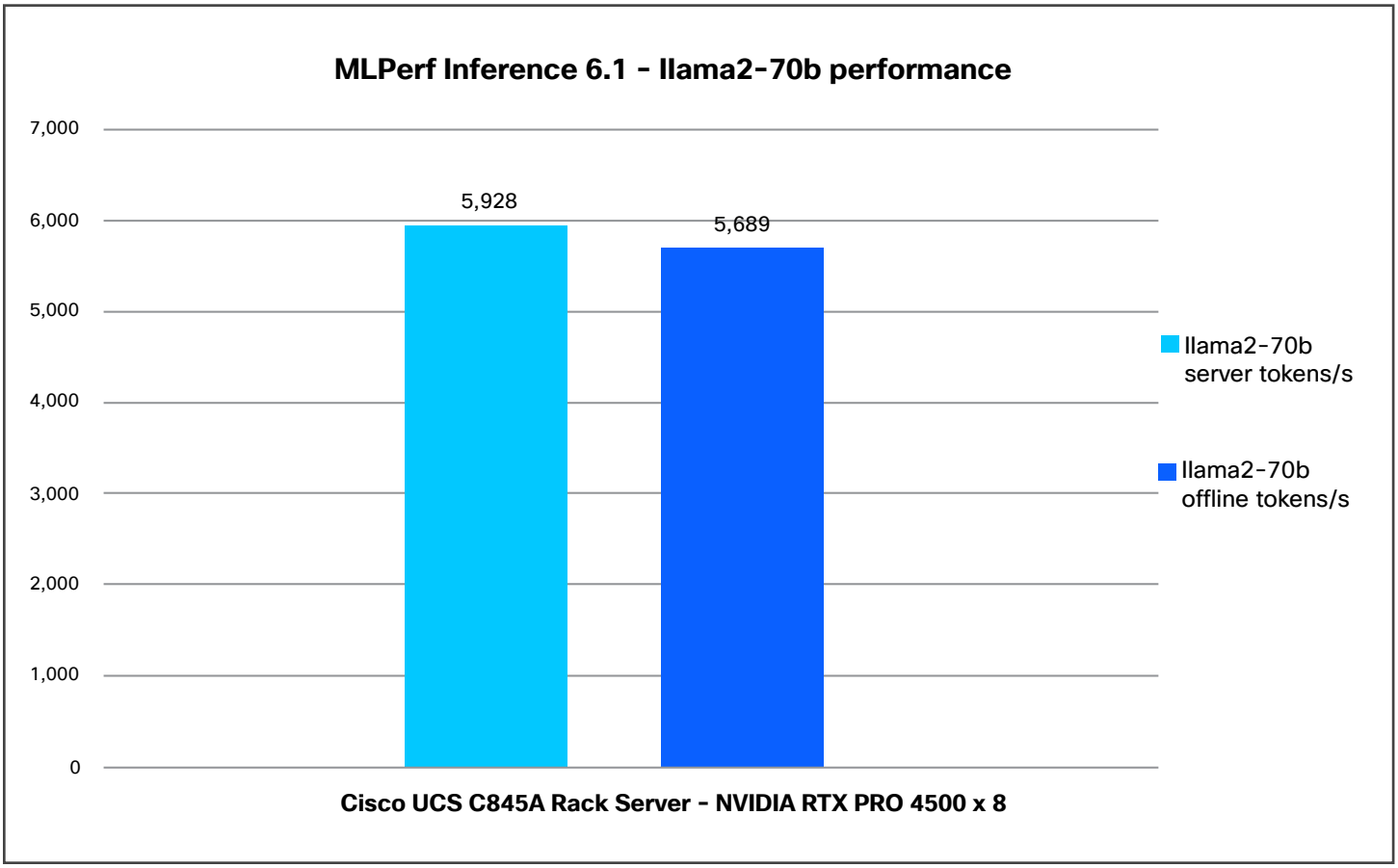


Figure 19. Llama2-70b performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA RTX PRO 4500 GPUs

Note: For NVIDIA RTX PRO 4500 Llama2-70b performance data, the results have not been verified by MLCommons Association because the results were collected after the MLPerf submission deadline.

© 2026 Cisco and/or its affiliates. All rights reserved.

Whisper

Whisper is an automatic speech recognition model trained on 680,000 hours of multilingual data collected from the web. As per OpenAI, this model is robust on accents, background noise, and technical language. In addition, it supports the transcription of 99 different languages and translation from those languages into English.

Figure 20 shows the performance of the Whisper model tested on a Cisco UCS C845A M8 Rack Server with 8x NVIDIA RTX PRO 4500 GPUs.

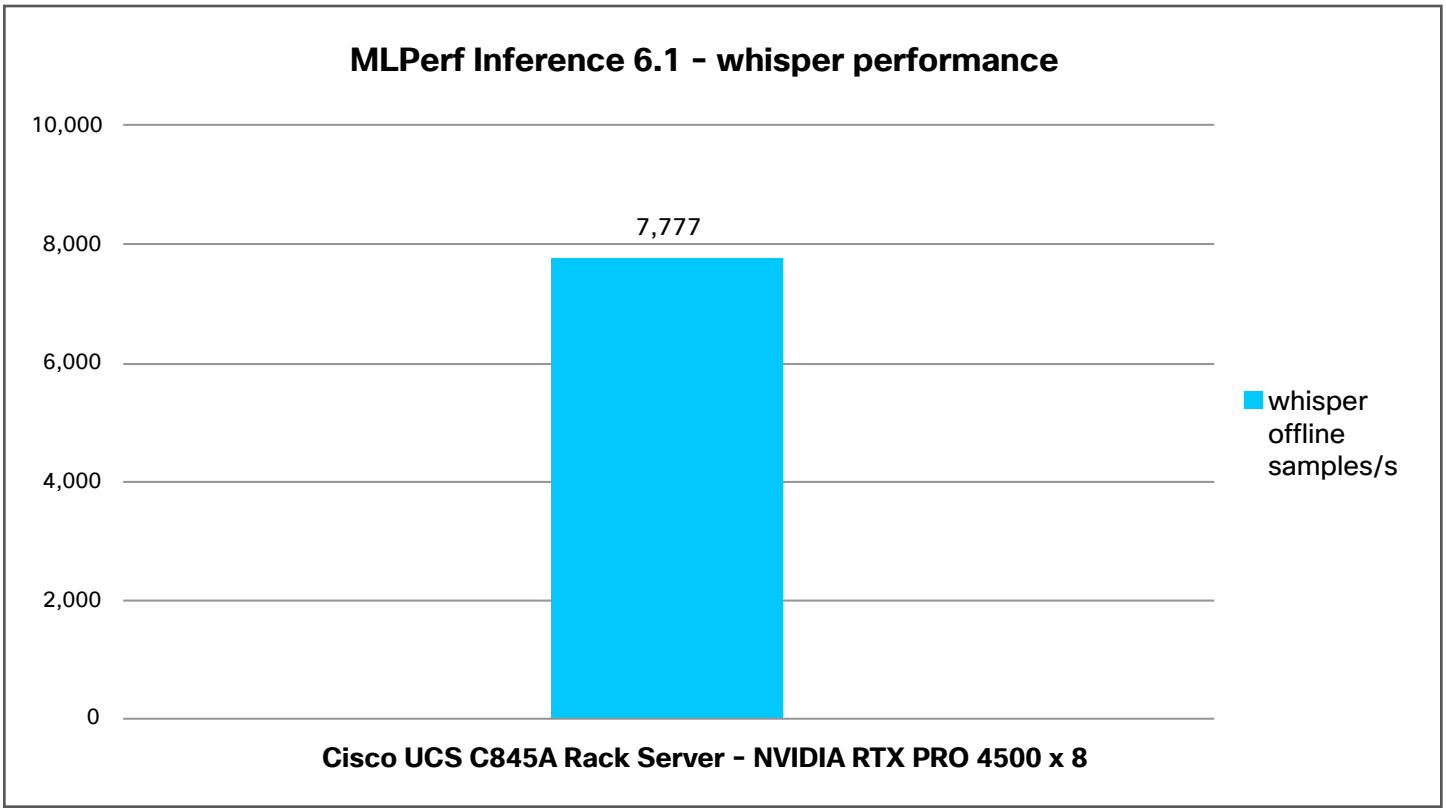


Figure 20. Whisper performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA RTX PRO 4500 GPUs

© 2026 Cisco and/or its affiliates. All rights reserved.

Performance summary

Built on the NVIDIA MGX platform, the Cisco UCS C845A 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.

Cisco successfully submitted MLPerf Inference Datacenter results in partnership with NVIDIA to enhance performance and efficiency, optimizing various inference workloads such as large language models (language), natural language processing (language), image generation (image), generative image (text to image), and object detection (vision).

The results were exceptional AI performance across Cisco UCS platforms for MLPerf Inference Datacenter:

- The Cisco UCS C845A M8 platform with 8x NVIDIA H200 NVL GPUs emerged as the leader, securing first position for the Llama3.1-8b model.
- The Cisco UCS C845A M8 platform with 8x NVIDIA H200 NVL GPUs emerged as the leader, securing first position for the Retinanet model.
- The Cisco UCS C845A M8 platform with 8x NVIDIA RTX PRO 6000 GPUs emerged as the leader, securing first position for the Llama2-70b model for Server scenario
- The Cisco UCS C845A M8 platform with 8x NVIDIA RTX PRO 4500 GPUs emerged as the leader, securing first position for the Llama3.1-8b model.
- The Cisco UCS C845A M8 platform with 8x NVIDIA RTX PRO 4500 GPUs emerged as the leader, securing first position for the Whisper model.

© 2026 Cisco and/or its affiliates. All rights reserved.

Appendix: Test environment

Table 2 details the properties of the Cisco UCS C845A Rack Server under test environment conditions.

Table 2. Server properties

Description	Value
Product name	Cisco UCS C845A M8 Rack Server
CPU	2x AMD EPYC 9575 64-Core Processor
Number of cores	64
Number of threads	128
Total memory	2.3 TB
Memory DIMMs (16)	96 GB x 24 DIMMs
Memory speed	6400 MHz
Network adapter	8x NVIDIA BlueField-3 E-series SuperNIC 400GbE/NDR 2x NIC cards
GPU controllers	NVIDIA H200 NVL PCIe 8-GPU NVIDIA L40S PCIe 8-GPU
SFF NVMe SSDs	16x 1.9 TB 2.5-inch high-performance, high-endurance NVMe SSD

© 2026 Cisco and/or its affiliates. All rights reserved.

Table 3 lists the server BIOS settings applied for MLPerf testing.

Table 3. Server BIOS settings

BIOS Settings	Value
SMT mode	Auto
NUMA nodes per socket	NPS2
IOMMU	Auto
Core performance boost	Auto
Determinism slider	Power
DRAM refresh rate	Platform default
L1 stream HW prefetcher	Enable
L2 stream HW prefetcher	Enable
AVX512	Enable
3-link xGMI max speed	Platform default
Streaming Stores Control	Auto

Note: The rest of the BIOS settings are platform default values.

For more information

- For additional information on the server, refer to: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c845a-m8-rack-server-aag.html.
- Data sheet: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c845a-m8-rack-server-ds.html.
- Cisco AI-Ready Data Center Infrastructure: https://blogs.cisco.com/datacenter/power-your-genai-ambitions-with-new-cisco-ai-ready-data-center-infrastructure.
- Cisco AI PODs: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/ai-infrastructure-pods-inferencing-aag.html.
- Cisco AI-Native Infrastructure for Data Center: https://www.cisco.com/site/us/en/solutions/artificial-intelligence/infrastructure/index.html.

© 2026 Cisco and/or its affiliates. All rights reserved. Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)