AI Performance: MLPerf Inference on Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL and L40S GPUs

October 2025

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® 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 5.1 was conducted by Cisco, using both NVIDIA H200 NVL and L40S PCIe 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.

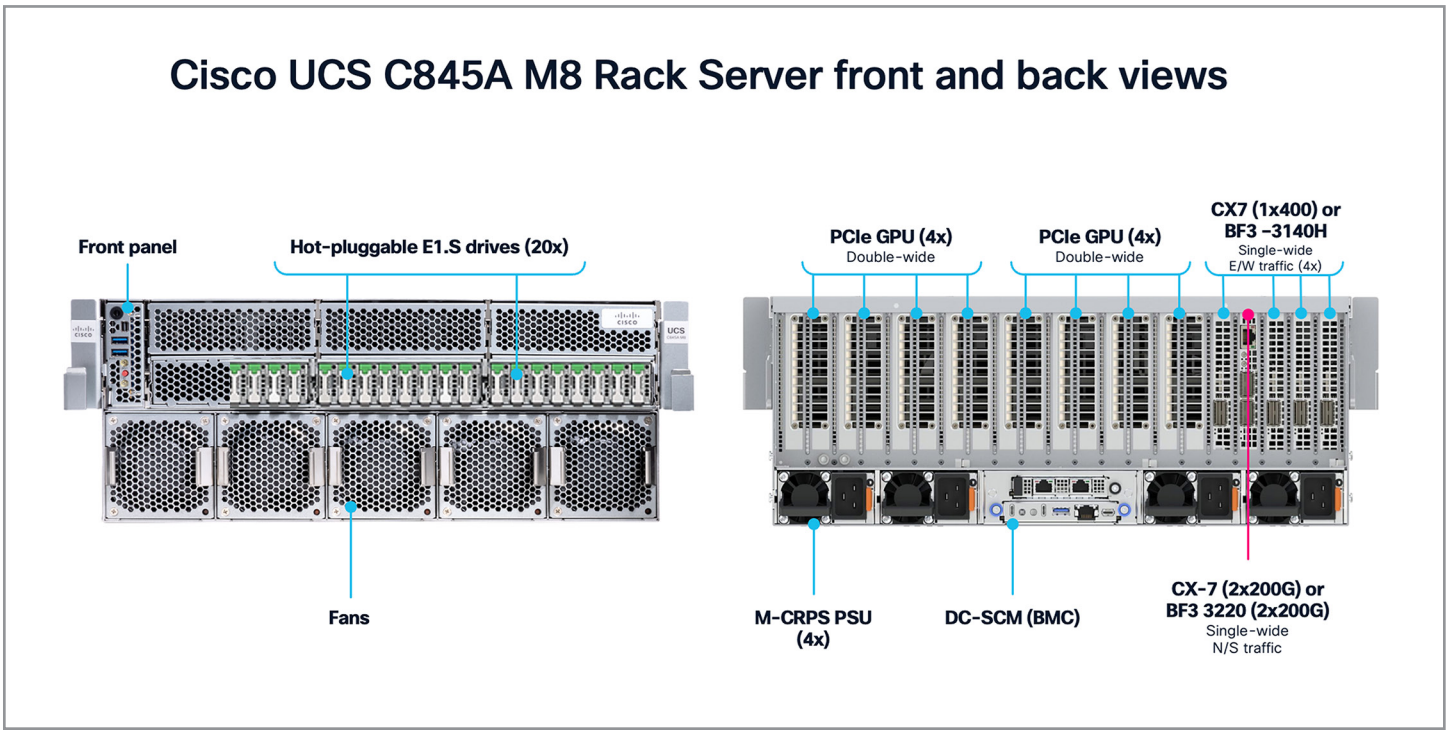

Figure 1. Cisco UCS C845A M8 Rack Server front and back views

Scalable network fabric for AI connectivity

To train GenAI models, clusters of these powerful servers often work in unison, generating an immense flow of data that necessitates a network fabric capable of handling high bandwidth with minimal latency. This is where the newly released Cisco Nexus® 9364E-SG2 Switches shine. Their high-density 800G aggregation ensures smooth data flow between servers, while advanced congestion management and large buffer sizes minimize packet drops—keeping latency low and training performance high. The Nexus 9364E-SG2 Switches serve as a cornerstone for a highly scalable network infrastructure, allowing AI clusters to expand seamlessly as organizational needs grow.

The new Cisco Nexus 9364E-SG2 Switch provides 800G aggregation for AI connectivity

Figure 2. Cisco Nexus 9364E-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.

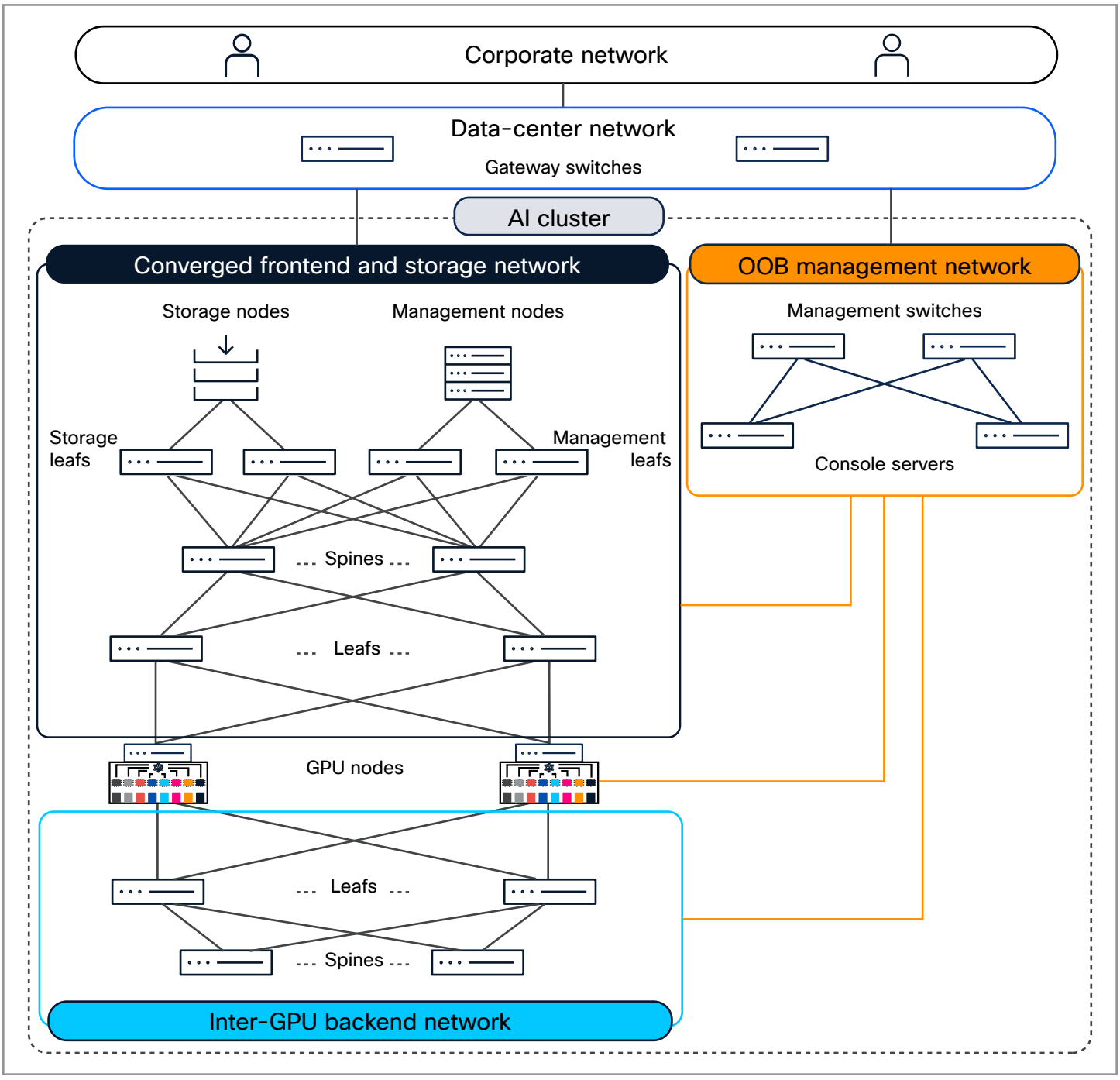

Figure 3. AI-cluster network design

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

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

AMD Instinct MI210 accelerators

- **High-Performance Computing (HPC):** accelerating scientific research, simulations, and complex modeling
- **Energy-efficient AI/ML workloads:** supporting deep-learning training and inferencing with a focus on power efficiency
- **Large-scale data analytics:** speeding up data processing and analytics for enterprise applications
- **Hybrid AI inference:** providing balanced compute and efficiency for inferencing tasks in diverse environments
- **Specialized simulation workloads:** enhancing performance for engineering, manufacturing, and bioinformatics simulations

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 the MLPerf Benchmarking performance testing for Inference 5.1: Datacenter, performance was evaluated using 8x NVIDIA H200 NVL and 8x NVIDIA L40S PCIe GPUs configured on a Cisco UCS C845A M8 Rack Server, and Inference benchmark results were collected for various datasets. This data will help in understanding the performance benefits of the UCS C845A M8 server using these PCIe GPUs for inference workloads. Performance data for MLPerf Inference 5.1 is highlighted in this white paper for selected datasets, to provide a quick understanding of the performance of the Cisco UCS C845A M8 Rack Server.

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, or AMD MI210 accelerators GPUs.Thanks to the sophistication of the MGX design, more “next-generation” NVIDIA and AMD GPUs are planned for introduction on this platform as they become available.
- 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.

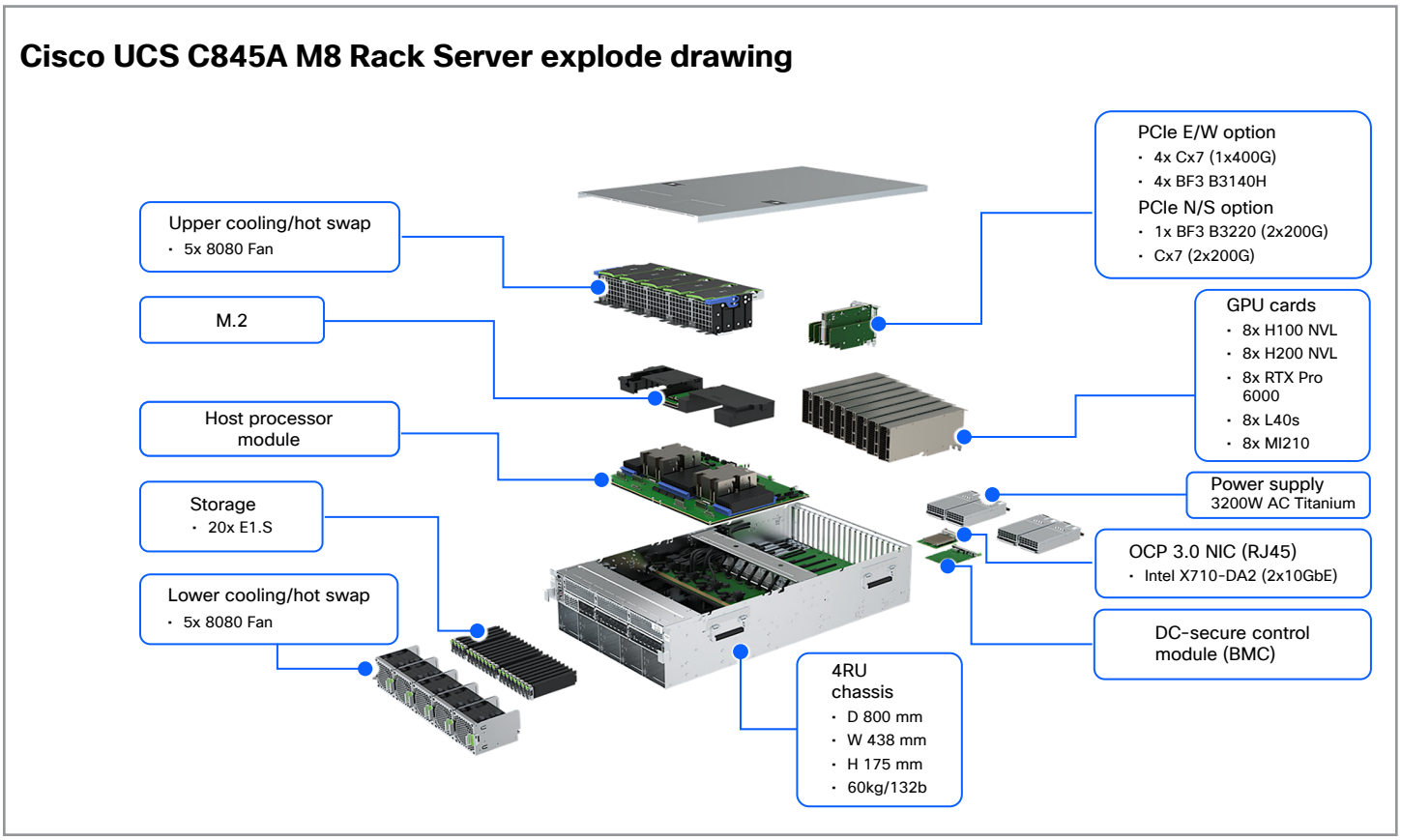

Figure 4. Components of the Cisco UCS C845A Rack Server

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.

MLPerf overview

MLPerf is a benchmark suite that evaluates the performance of machine-learning software, hardware, and services. The benchmarks are developed by MLCommons, a consortium of AI leaders from academia, research labs, and industry. The goal of MLPerf is to provide an objective yardstick for evaluating machine-learning platforms and frameworks.

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/.

Test configuration

For the MLPerf Inference 5.1 performance testing covered in this document, the following two Cisco UCS C845A M8 Rack Server configurations were used:

- 8x NVIDIA H200 NVL PCIe GPUs
- 8x NVIDIA L40S PCIe 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 5.1 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

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

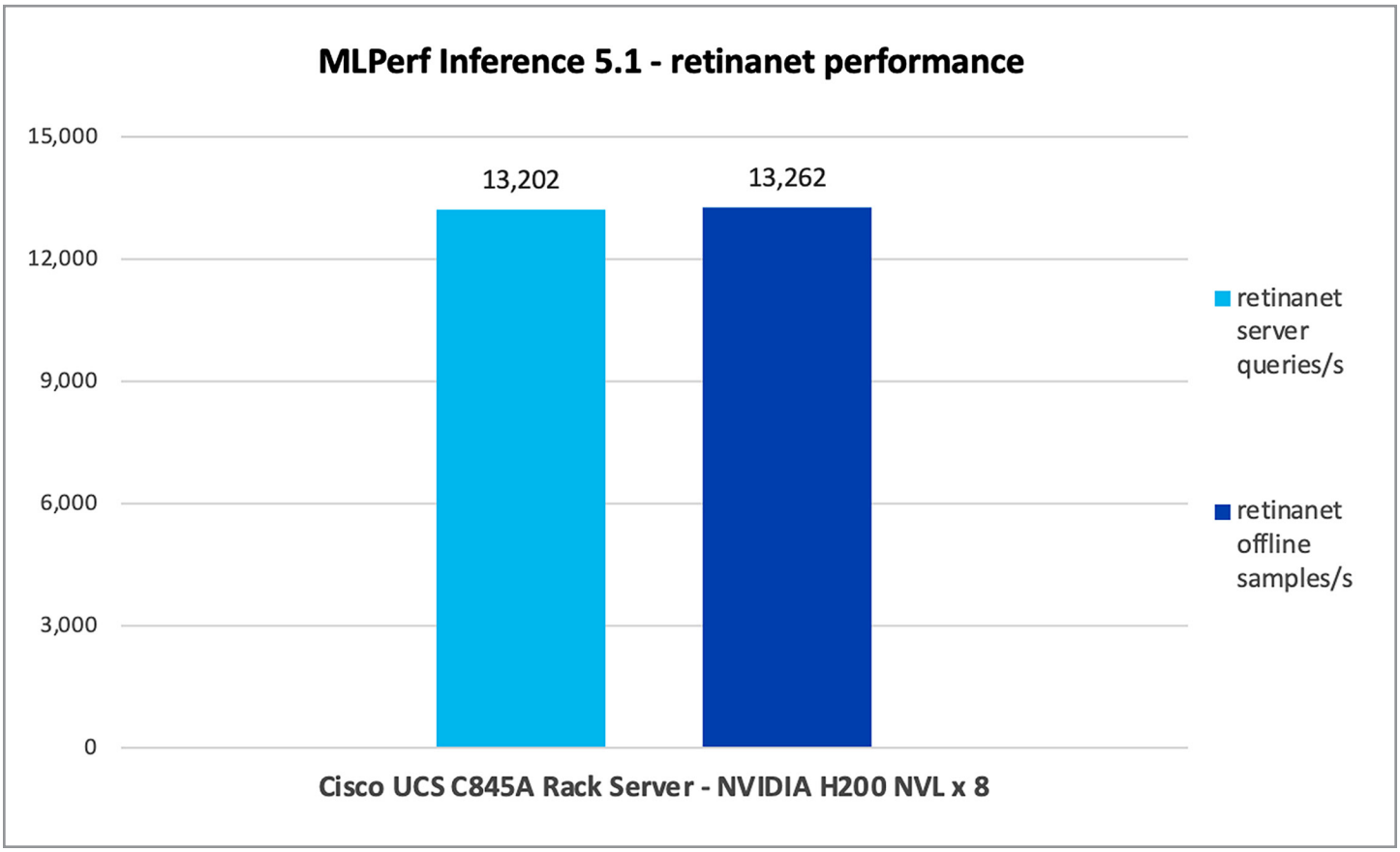

Figure 5. Retinanet performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

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.

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.

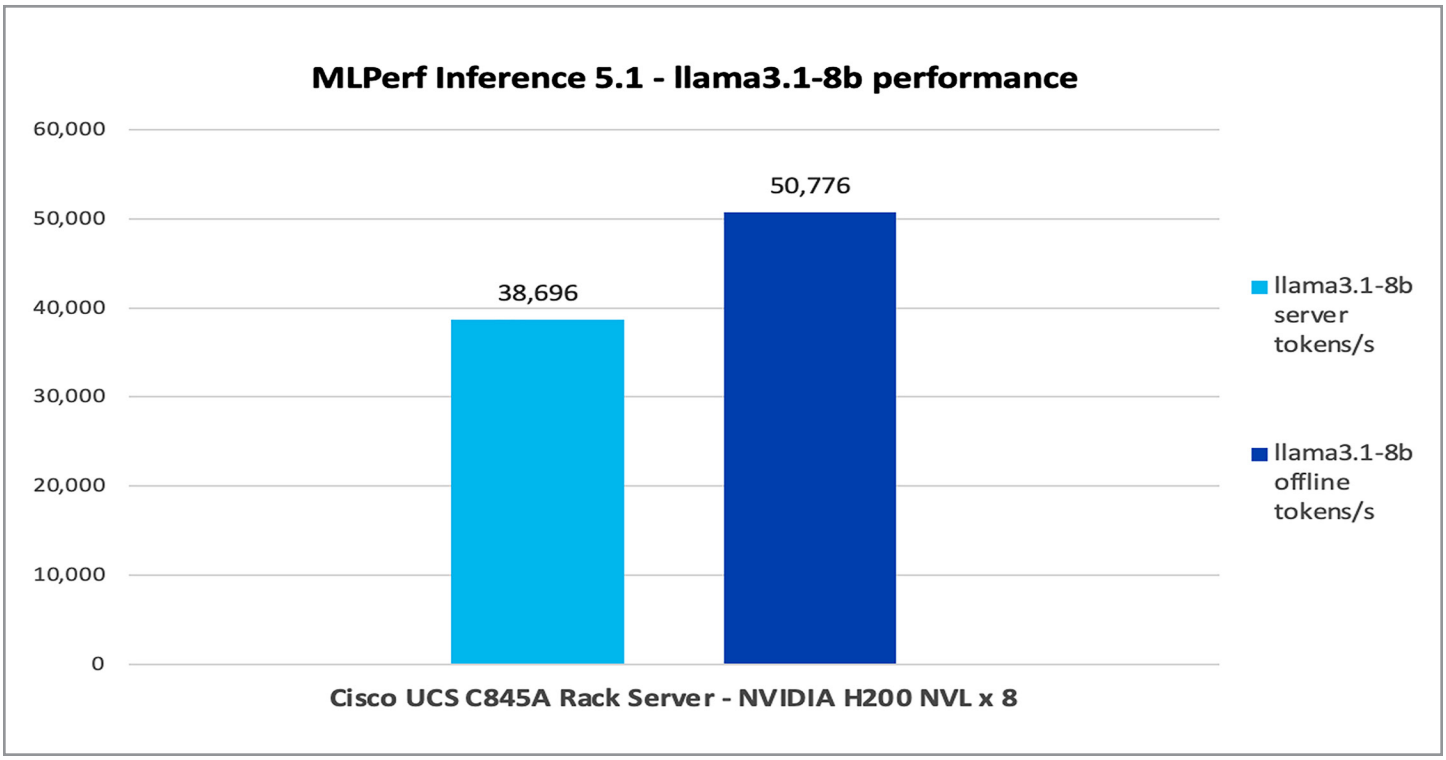

Figure 6. Llama3.1-8b performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

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.

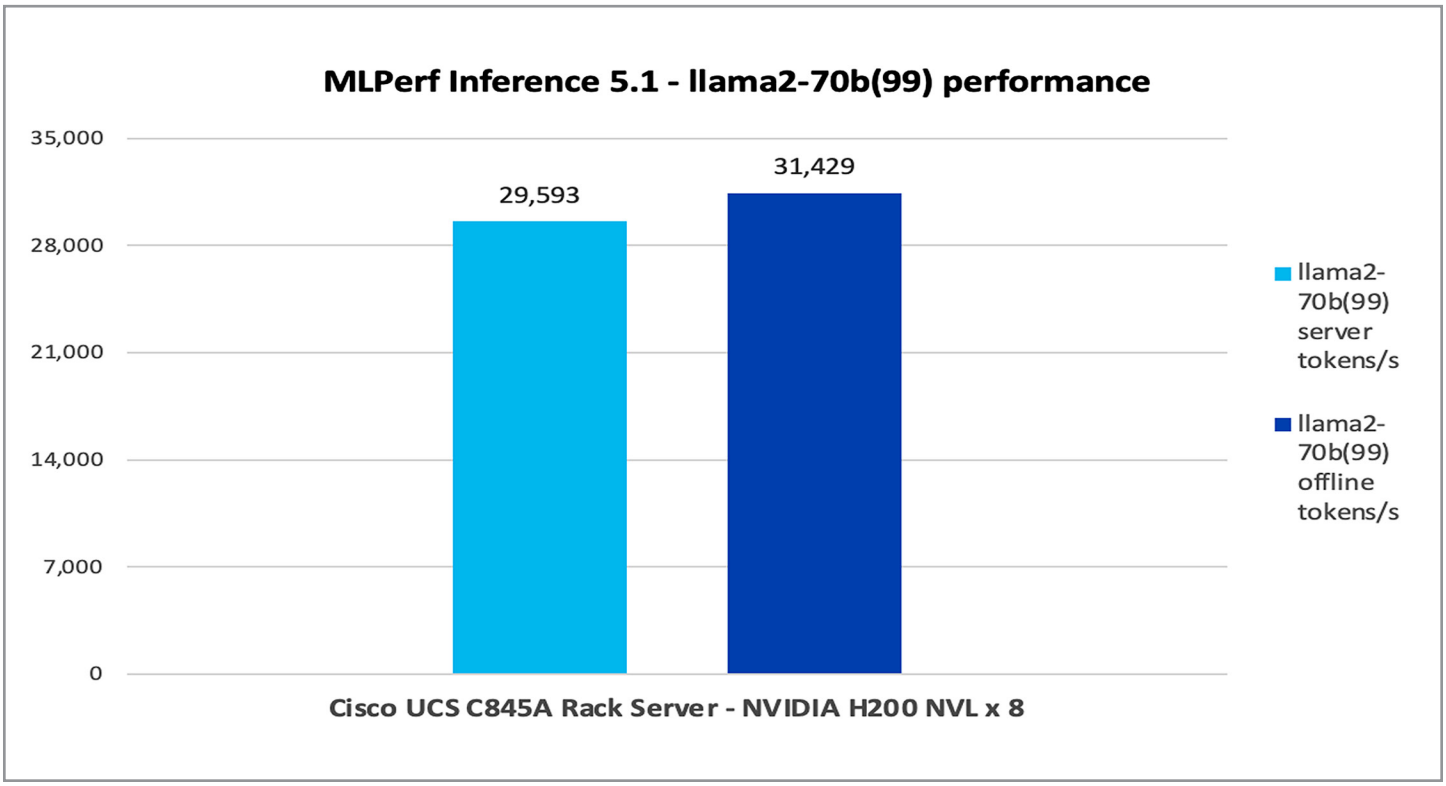

Figure 7. Llama2-70b performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

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.

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.

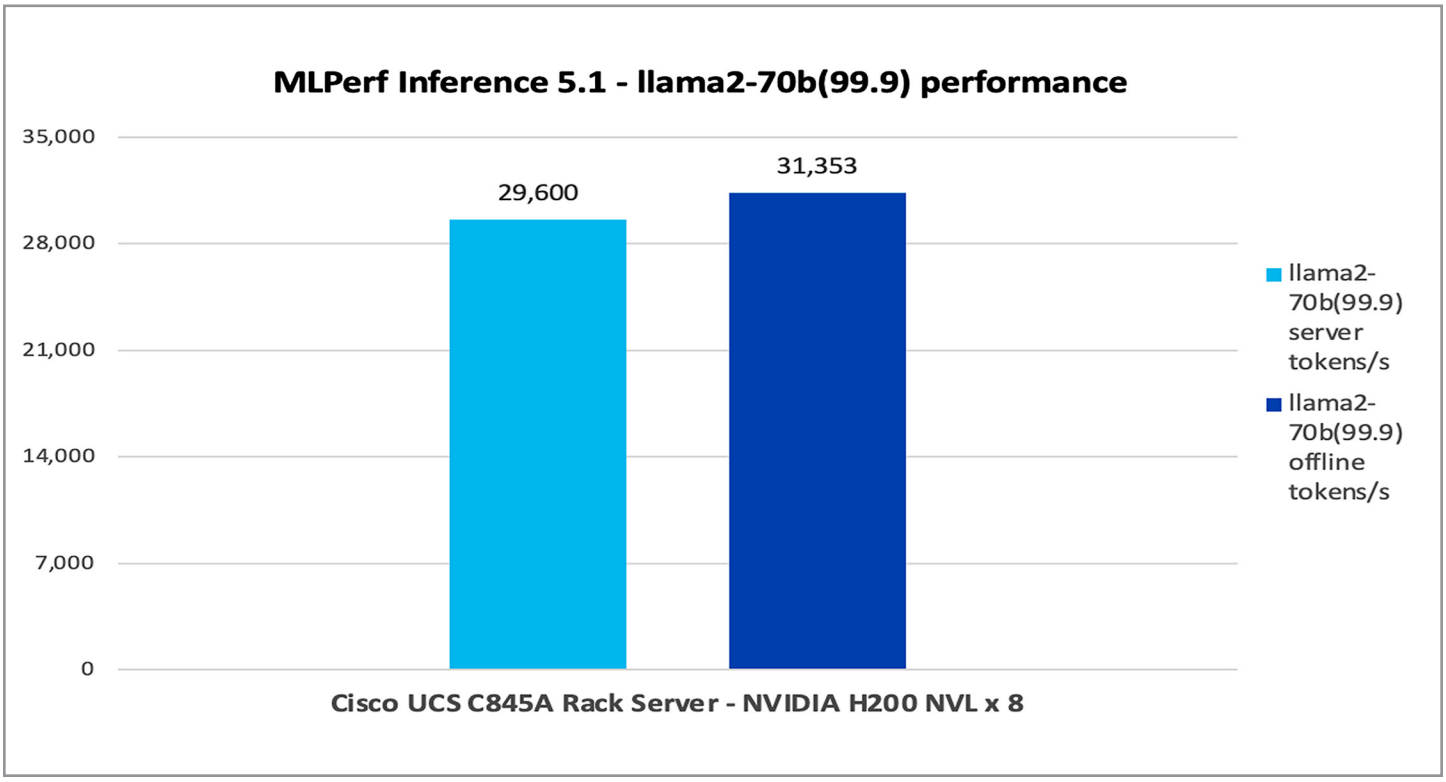

Figure 8. Llama2-70b (99.9) performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

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.

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.

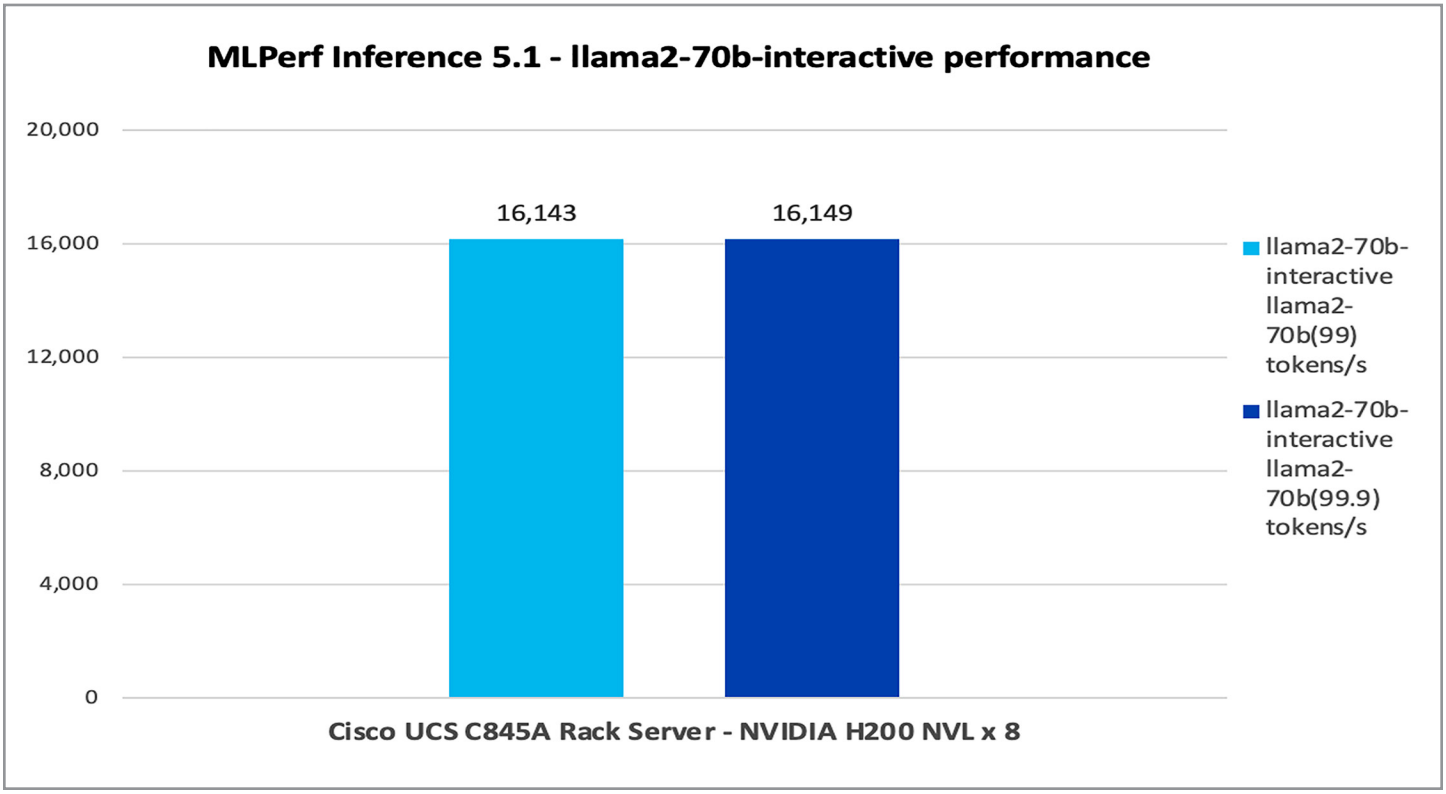

Figure 9. Llama2-70b Interactive performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

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.

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.

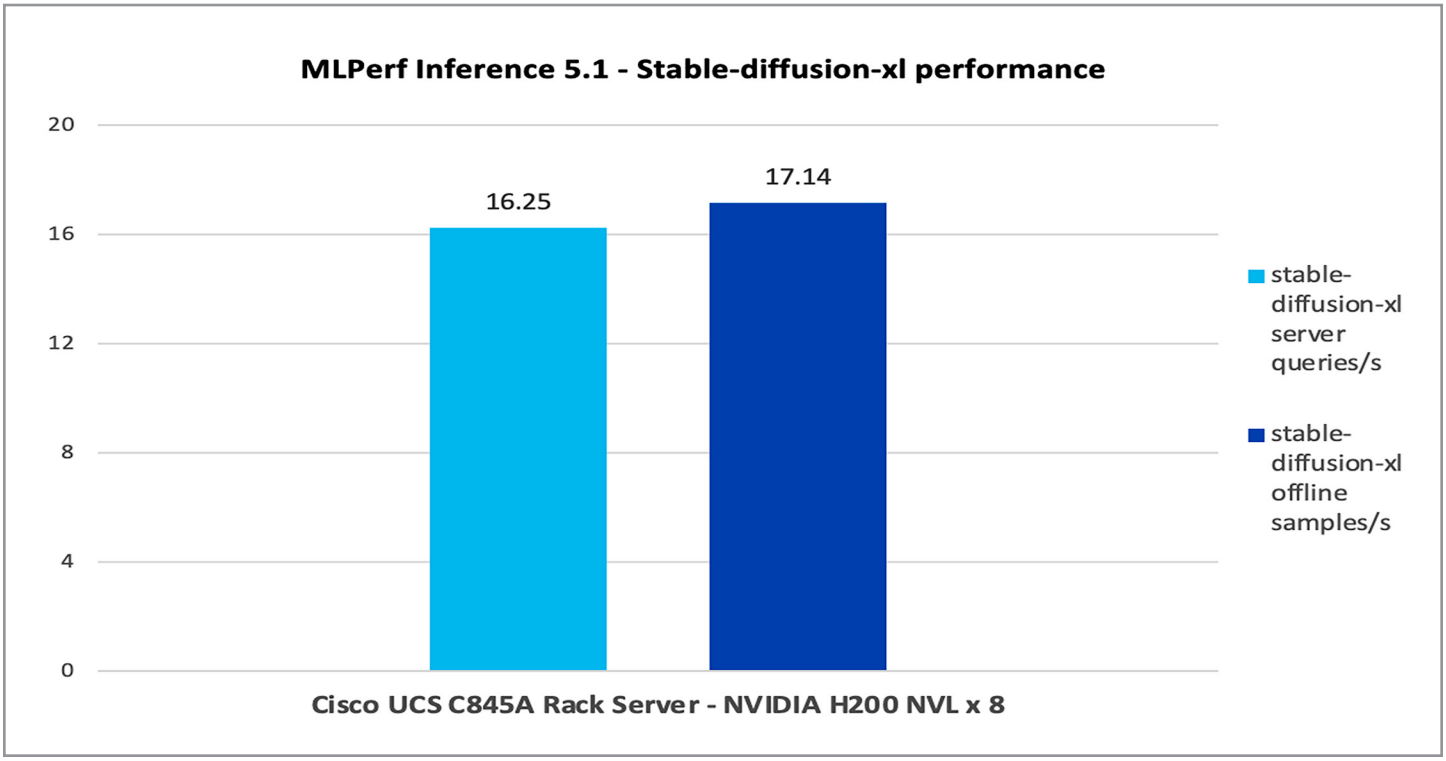

Figure 10. Stable Diffusion XL performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

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.

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.

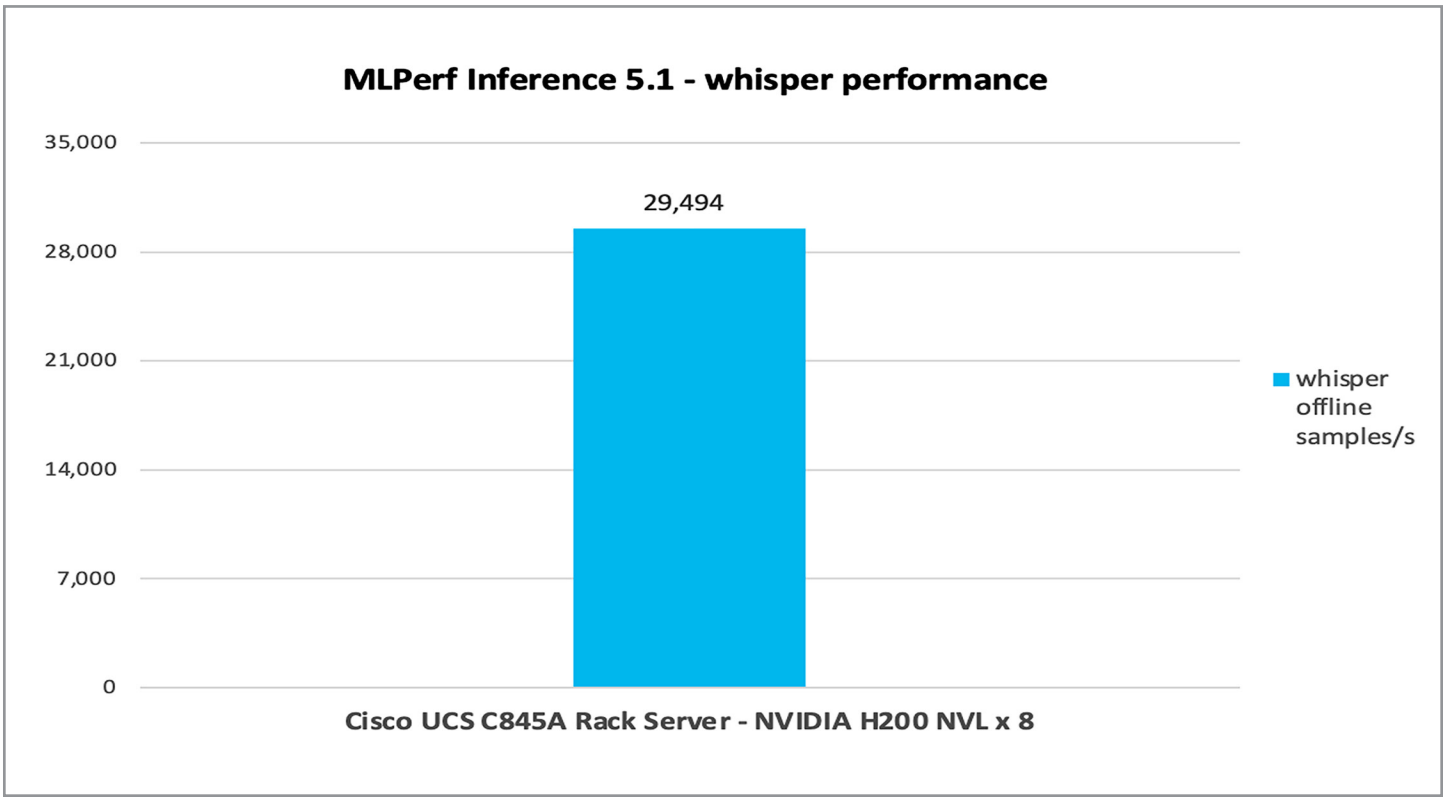

Figure 11. Whisper performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

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.

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.

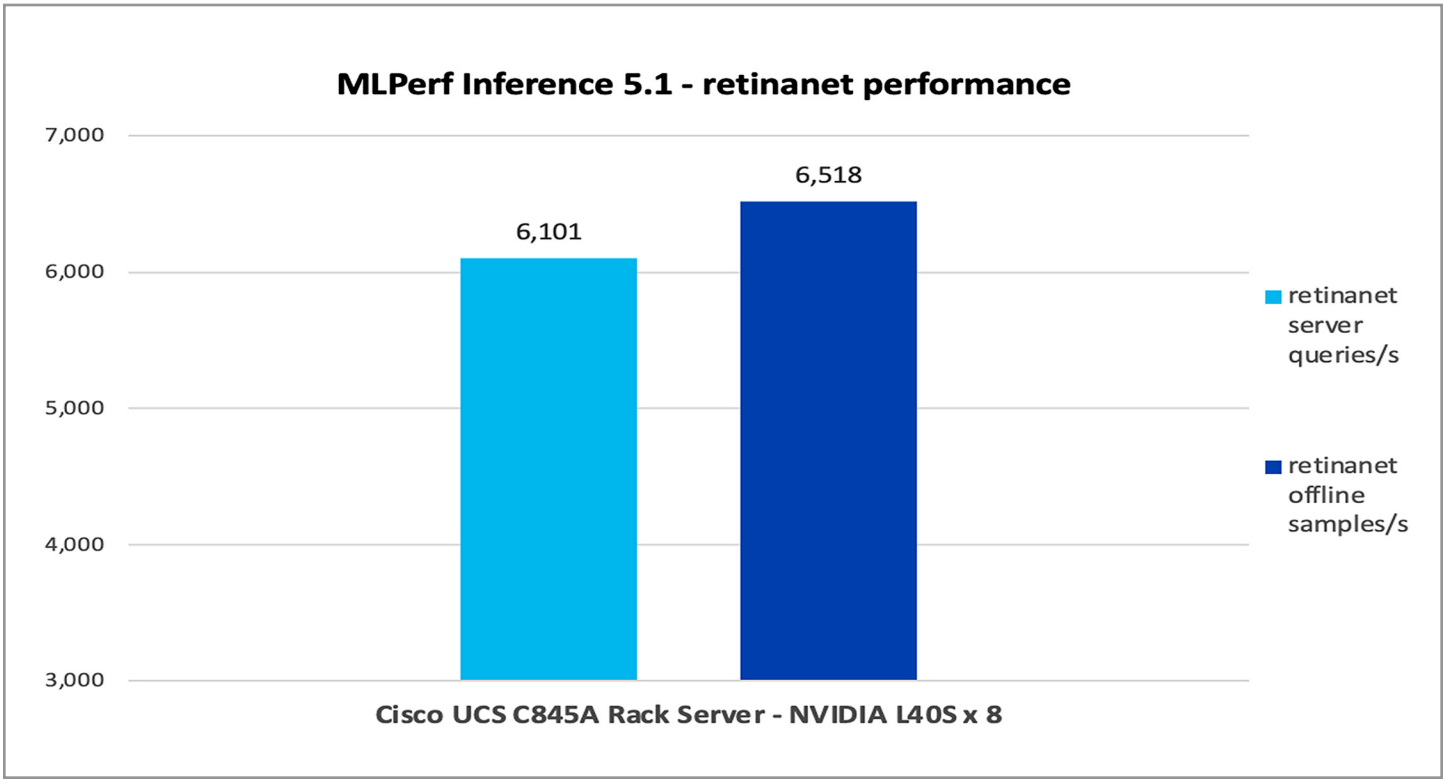

Figure 12. Retinanet performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA L40S GPUs

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.

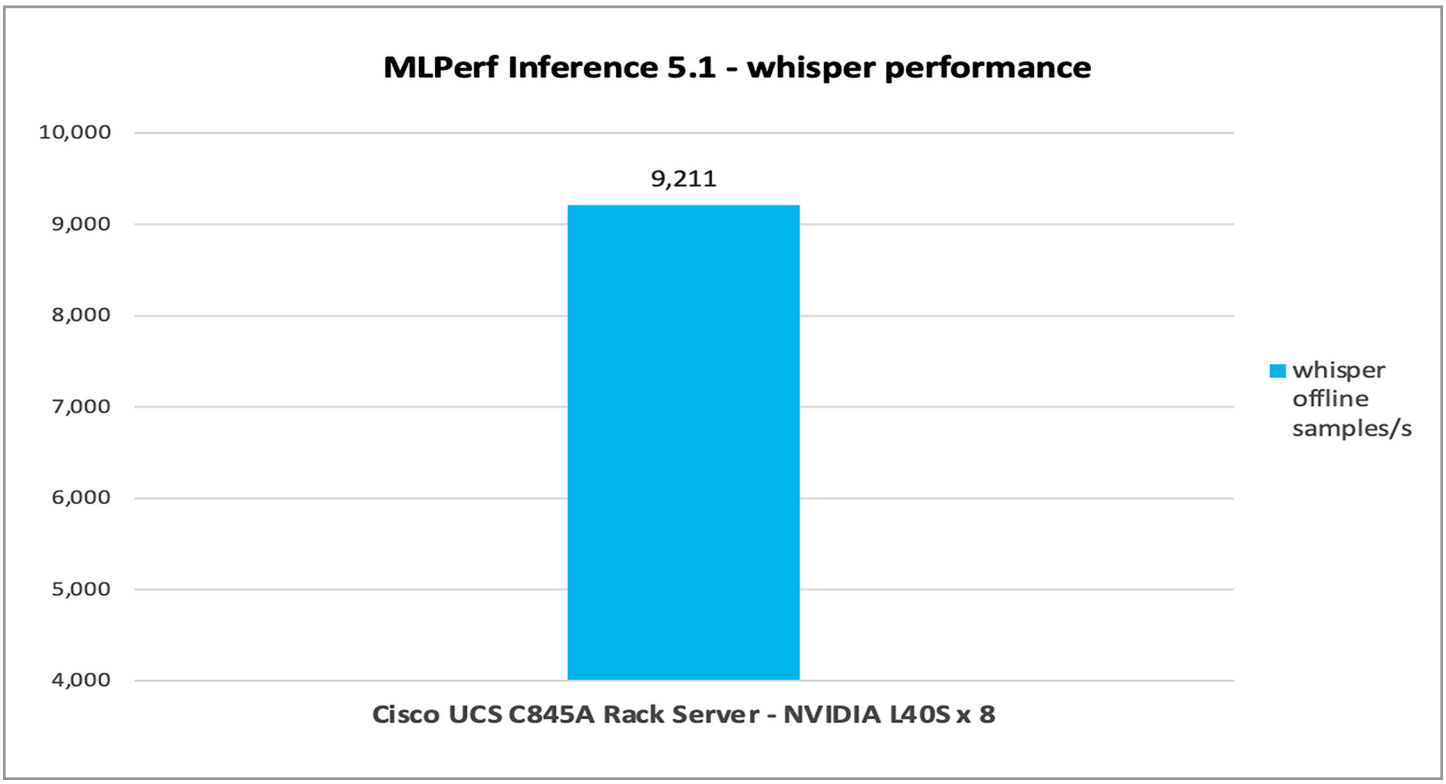

Figure 13. Whisper performance data on a Cisco UCS C845A M8 Rack Server with NVIDIA L40S GPUs

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 5.1 Inference 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 5.1:

- 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.

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

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.