White paper
Cisco public

AI Performance: MLPerf Inference on Cisco UCS C885A M8 HGX platform with NVIDIA H100 and H200 GPUs

May 2025

Contents

Executive summary

With generative AI 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.

To this end, Cisco recently introduced the Cisco UCS C885A M8, a high-density GPU server designed for demanding AI workloads, offering powerful performance for model training, deep learning, and inference. Built on the NVIDIA HGX platform, it can scale out to deliver clusters of computing power that will bring your most ambitious AI projects to life. Each server includes NVIDIA network interface cards (NICs) or SuperNICs to accelerate AI networking performance, as well as NVIDIA BlueField-3 data processing units (DPUs) to accelerate GPU access to data and enable robust, zero-trust security. The new Cisco UCS C885A M8 is Cisco's first entry into its dedicated AI server portfolio and its first eight-way accelerated computing system built on the NVIDIA HGX platform.

To help demonstrate the AI performance capacity of the new Cisco UCS C885A M8 Server, MLPerf Benchmarking performance testing for Inference 5.0 was conducted by Cisco, using both NVIDIA H100 and H200 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 C885A M8 Server is a powerhouse designed to tackle the most demanding AI training tasks. With its high-density configuration of NVIDIA H100 and H200 Tensor Core GPUs, coupled with the efficiency of NVIDIA HGX architecture, the UCS C885A 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.

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 Switch shines. Its 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 serves 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

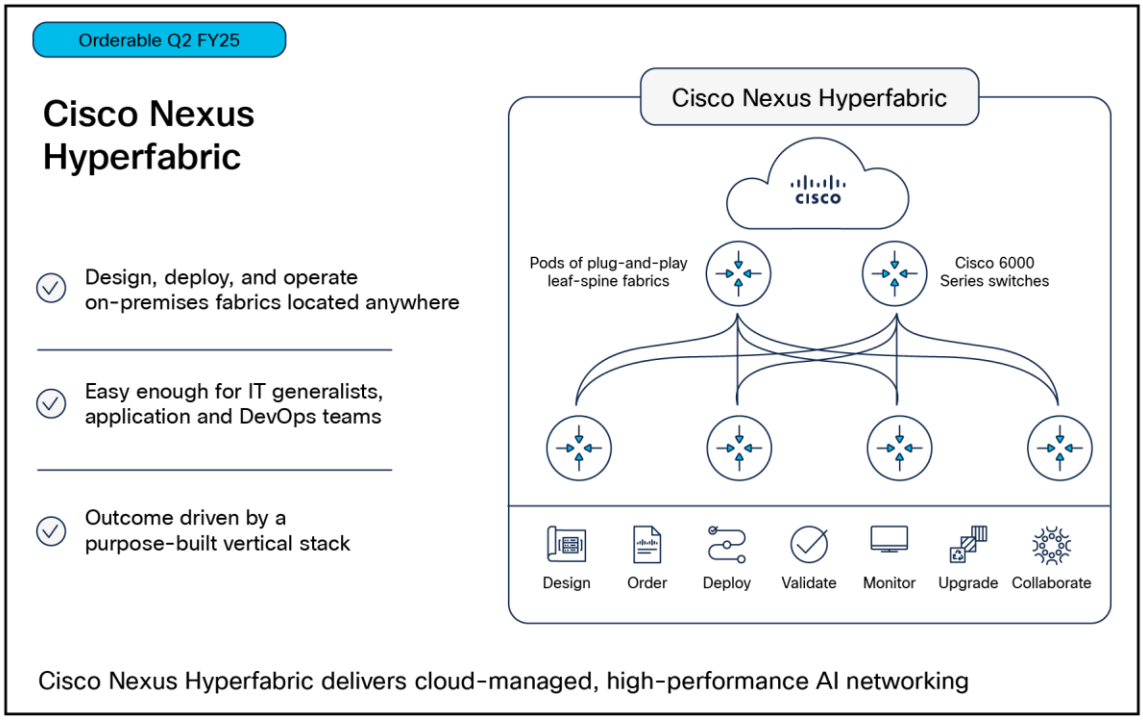
Purchasing simplicity

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. No matter where you fall on the spectrum of use cases mentioned at the beginning of this whitepaper, 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, with more use cases planned for release over the next few months.

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.

Cloud Deployed and Cloud Managed

To help organizations modernize their data center operations and enable AI use cases, we further simplify infrastructure deployment and management with Cisco Nexus Hyperfabric, a fabric-as-a-service solution announced earlier this year at Cisco Live. Cisco Nexus Hyperfabric features a cloud-managed controller that simplifies the design, deployment, and management of the network fabric for consistent performance and operational ease. The hardware-accelerated performance of Cisco Nexus Hyperfabric, with its inherent high bandwidth and low latency, optimizes AI inferencing, enabling fast response times and efficient resource utilization for demanding, real-time AI applications. Furthermore, Cisco Nexus Hyperfabric's comprehensive monitoring and analytics capabilities provide real-time visibility into network performance, allowing for proactive issue identification and resolution to maintain a smooth and reliable inferencing environment.



By providing a seamless continuum of solutions, from powerful training servers and high-performance networking to simplified inference deployments, we are enabling enterprises to accelerate their AI initiatives, unlock the full potential of their data, and drive meaningful innovation.

Introduction

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

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

Benefits of Cisco UCS Servers

AI-Ready

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

Scalable

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

Consistent Management

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

Scope of this document

For the MLPerf Benchmarking performance testing for Inference 5.0, performance was evaluated using 8 x NVIDIA H100 and 8 x NVIDIA H200 GPUs configuration on Cisco UCS C885A M8 server. This is the standard configuration on UCS C885A M8 server, and Inference benchmark results are collected for various datasets. This data will help in understanding the performance benefits of UCS C885A M8 server for Inference workloads. Performance data for MLPerf Inference 5.0 is highlighted in this white paper for selected datasets to have a brief understanding of performance on Cisco UCS C885A M8 server.

Product overview

- Built on the NVIDIA HGX platform, the Cisco UCS C885A M8 Rack Server offers a choice of 8 NVIDIA HGX H200 or H100 Tensor Core GPUs to deliver massive, accelerated computational performance in a single server, as well as one NVIDIA ConnectX-7 NIC or NVIDIA BlueField-3 SuperNIC per GPU to scale AI model training across a cluster of dense GPU servers.
- The server is managed by Cisco Intersight, which can help reduce your Total Cost of Ownership (TCO) and increase your business agility.

  Note: Initially, the local server management interface will handle configuration and management, while Cisco Intersight will provide inventory capabilities through an integrated Device Connector. Full management operations and configurations through Cisco Intersight will be introduced shortly thereafter in a subsequent phase.
- The server is offered in fixed configurations that are optimized for intensive AI and HPC workloads.

Figure 1.
Front view of server

A specifications sheet for the C885A M8 is available at: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c885a-m8-ds.html

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:** 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 below.

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

MLPerf Inference

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

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

Test configuration

For the MLPerf Inference 5.0 performance testing covered in this document, following two Cisco UCS C885A M8 configurations were used:

- 8 x NVIDIA H100 SXM GPUs
- 8 x NVIDIA H200 SXM GPUs

MLPerf Inference Performance Results

MLPerf Inference Benchmarks

The below mentioned MLPerf Inference models were configured on Cisco UCS C885A M8 server and tested for performance.

Table 1. MLPerf Inference 5.0 models

Model	Reference Implementation Model	Description
resnet50-v1.5	vision/classification_and_detection	Resnet is deep convolutional neural network (CNN) used for image classification tasks
retinanet 800x800	vision/classification_and_detection	Single-stage object detection model optimized for detecting small objects in high resolution images
dlrm-v2	recommendation/dlrm_v2	Advanced recommendation model designed to handle large-scale, high-dimensional data
3d-unet	vision/medical_imaging/3d-unet-kits19	Convolutional neural network designed for 3D medical imaging segmentation
gpt-j	language/gpt-j	Open source transformer-based language model, it is designed for natural language processing (NLP) tasks
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 question answering
llama3.1-405b	language/llama3-405b	Highly advanced language model with 405 billion parameters aimed at performing complex NLP tasks at scale
mixtral-8x7b	language/mixtral-8x7b	Mixture of experts model that combines multiple specialized sub-models, each with 7 billion parameters, to optimize performance on a range of tasks

Model	Reference Implementation Model	Description
rgat	graph/rgat	Graph-based neural network model that uses attention mechanisms to learn from relational data

MLPerf Inference 5.0 Performance Data

As part of the MLPerf Inference 5.0 submission, Cisco has tested most of the datasets mentioned in Table 1 on Cisco UCS C885A M8 server and submitted the results to MLCommons. The results are published on MLCommons results page: https://mlcommons.org/benchmarks/inference-datacenter/

Note: The below graph includes unverified MLPerf Inference 5.0 results collected after the MLPerf submission deadline. For such data, there is a note added "Result not verified by MLCommons Association."

MLPerf inference results are measured in both Offline and Server scenarios. The Offline scenario focuses on maximum throughput, while the Server scenario measures both throughput and latency, ensuring a certain percentage of requests are served within a specified latency threshold.

Resnet50

ResNet50-v1.5 is a deep convolutional neural network (CNN) used for image classification tasks. It improves upon the original ResNet50 by introducing optimizations that enhance its performance and training efficiency while maintaining its residual learning architecture.

Figure 2 shows the performance of the Resnet50 model tested on UCS C885A M8 server with 8 x 100 and H200 GPUs.

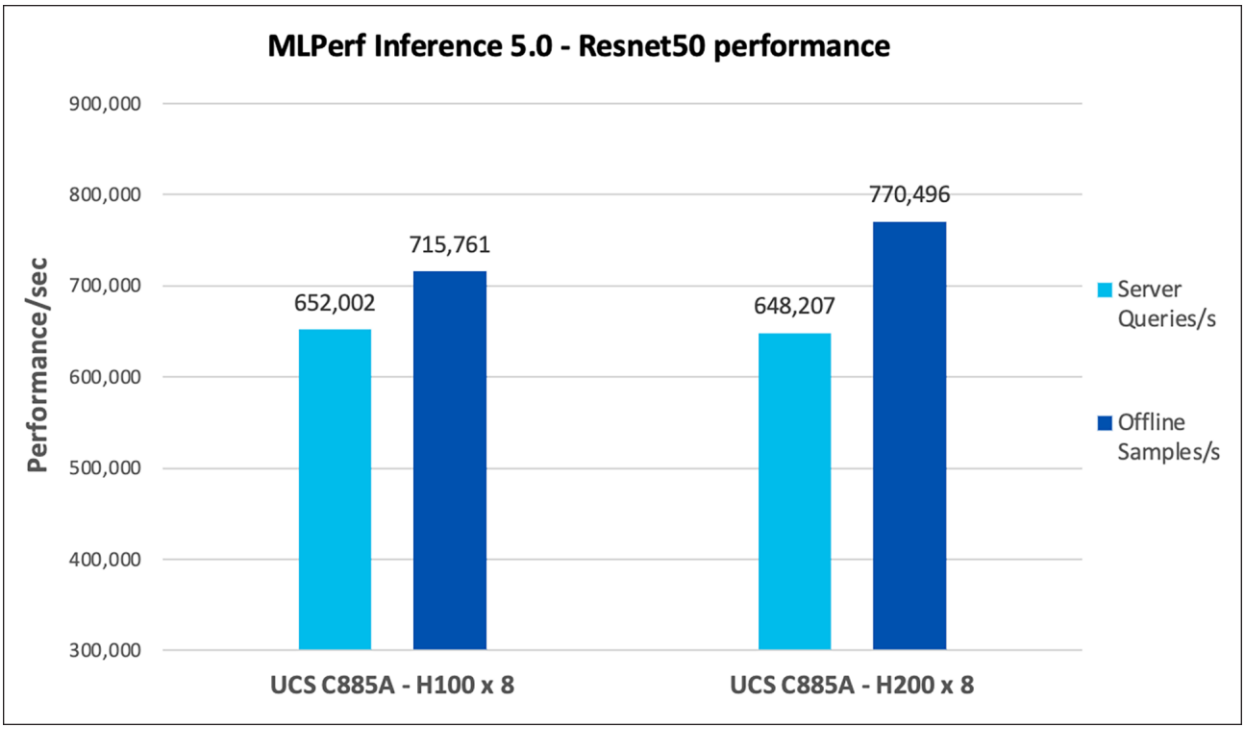


Figure 2.
Resnet50 performance data on Cisco UCS C885A M8 server with NVIDIA H100 and H200 GPUs

Retinanet

Retinanet is a single-stage object detection model known for its focus on addressing class imbalance 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 3 shows the performance of the Retinanet model tested on UCS C885A M8 server with 8 x H100 and H200 GPUs.

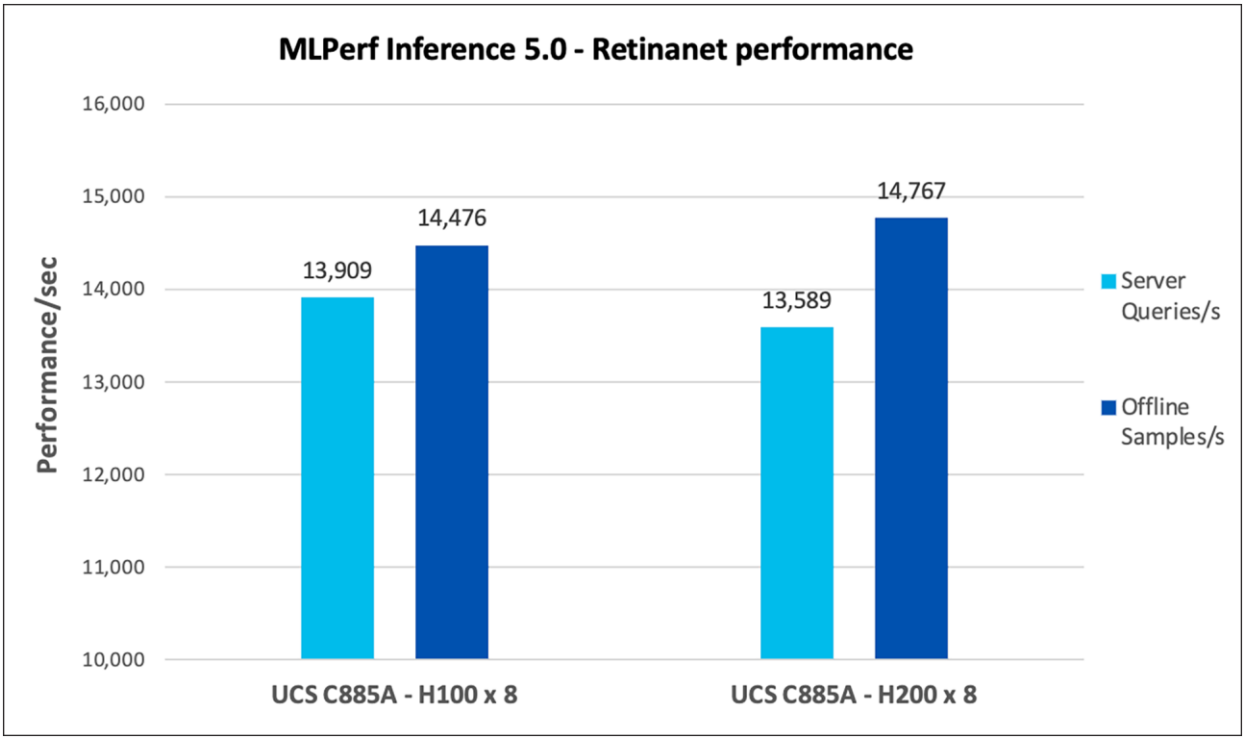


Figure 3.
Retinanet performance data on Cisco UCS C885A M8 server with NVIDIA H100 and H200 GPUs

Note: For H200 Retinanet performance data, result is not verified by MLCommons Association as the results are collected after the MLPerf submission deadline.

GPTJ

GPT-J is an open-source transformer-based language model developed by EleutherAI. It has 6 billion parameters and is designed for tasks such as text generation, translation, summarization, and other natural language processing tasks.

Figure 4 shows the performance of the GPTJ model tested on UCS C885A M8 server with 8 x H100 and H200 GPUs.

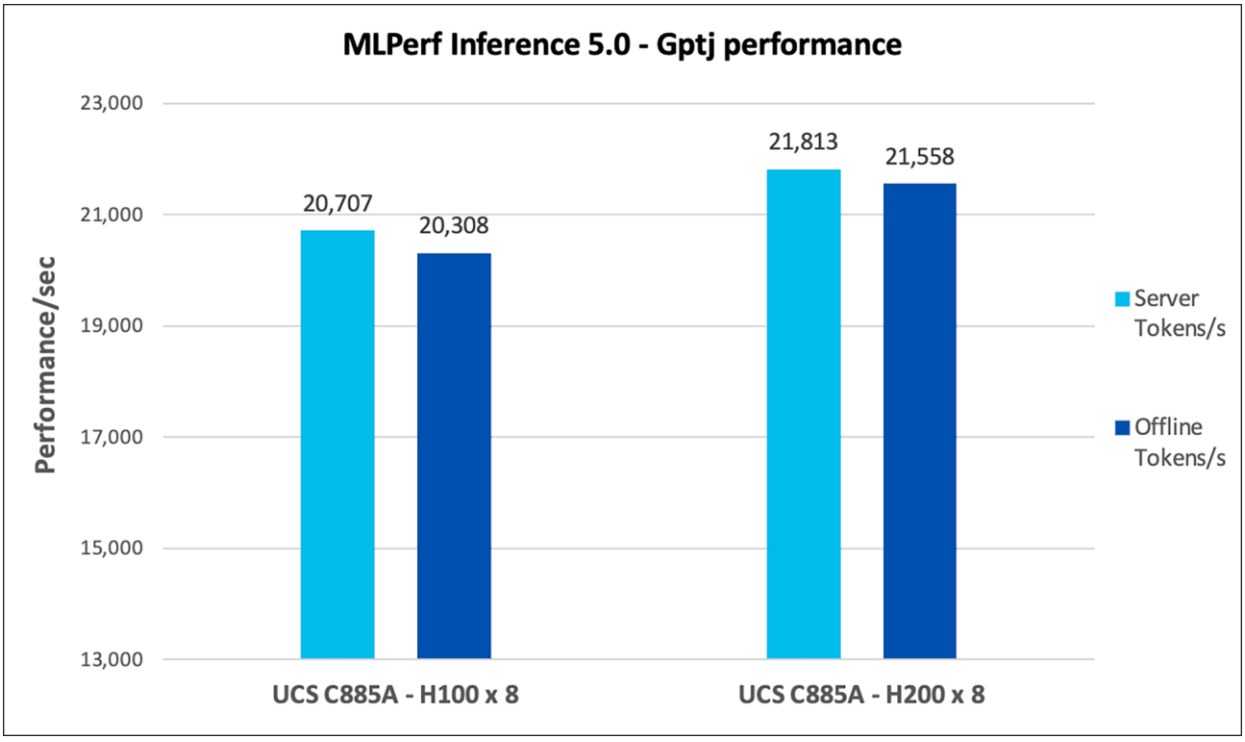


Figure 4.
GPTJ performance data on Cisco UCS C885A M8 server with NVIDIA H100 and H200 GPUs

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 like image synthesis, art generation, and image editing.

Figure 5 shows the performance of the Stable-Diffusion-XL model tested on UCS C885A M8 server with 8 x H100 and H200 GPUs.

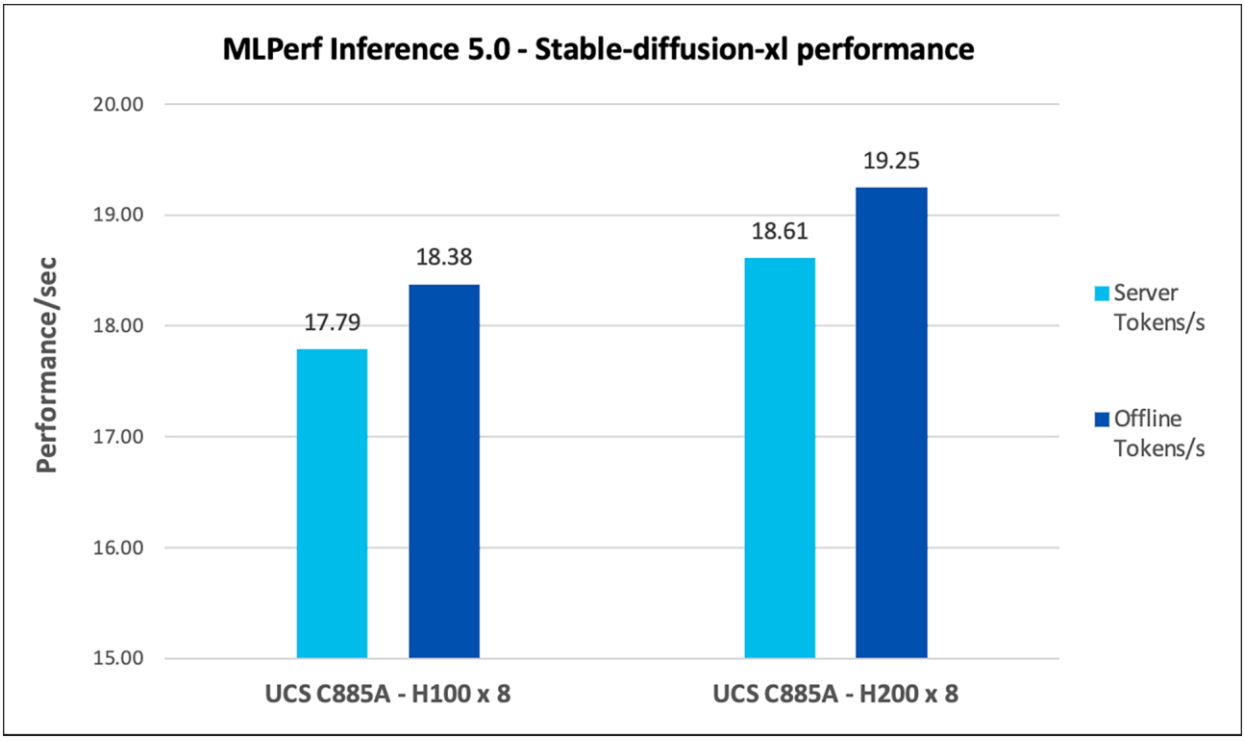


Figure 5.
Stable-Diffusion-XL performance data on Cisco UCS C885A M8 server with NVIDIA H100 and H200 GPUs

Note: For H200 Stable-diffusion-xl performance data, result is not verified by MLCommons Association as the results are collected after the MLPerf submission deadline.

Llama2-70B

Llama2-70B is a large language model from Meta, with 70 billion parameters. It is designed for various natural language processing tasks like text generation, summarization, translation, and question answering.

Figure 6 shows the performance of the Llama2-70B model tested on UCS C885A M8 server with 8 x H100 and H200 GPUs

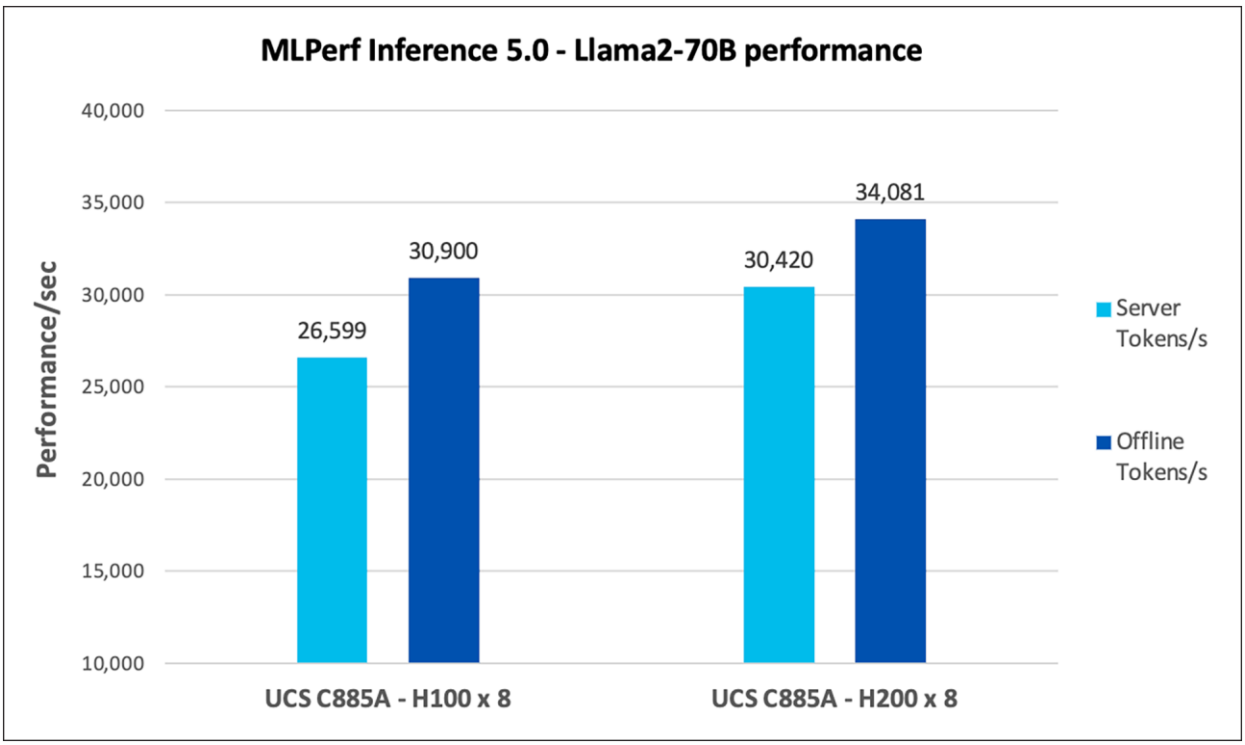


Figure 6.
Llama2-70B performance data on Cisco UCS C885A M8 server with NVIDIA H100 and H200 GPUs

Llama2-70B-Interactive

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

Figure 7 shows the performance of the Llama2-70B-Interactive model tested on UCS C885A M8 server with 8 x H100 and H200 GPUs

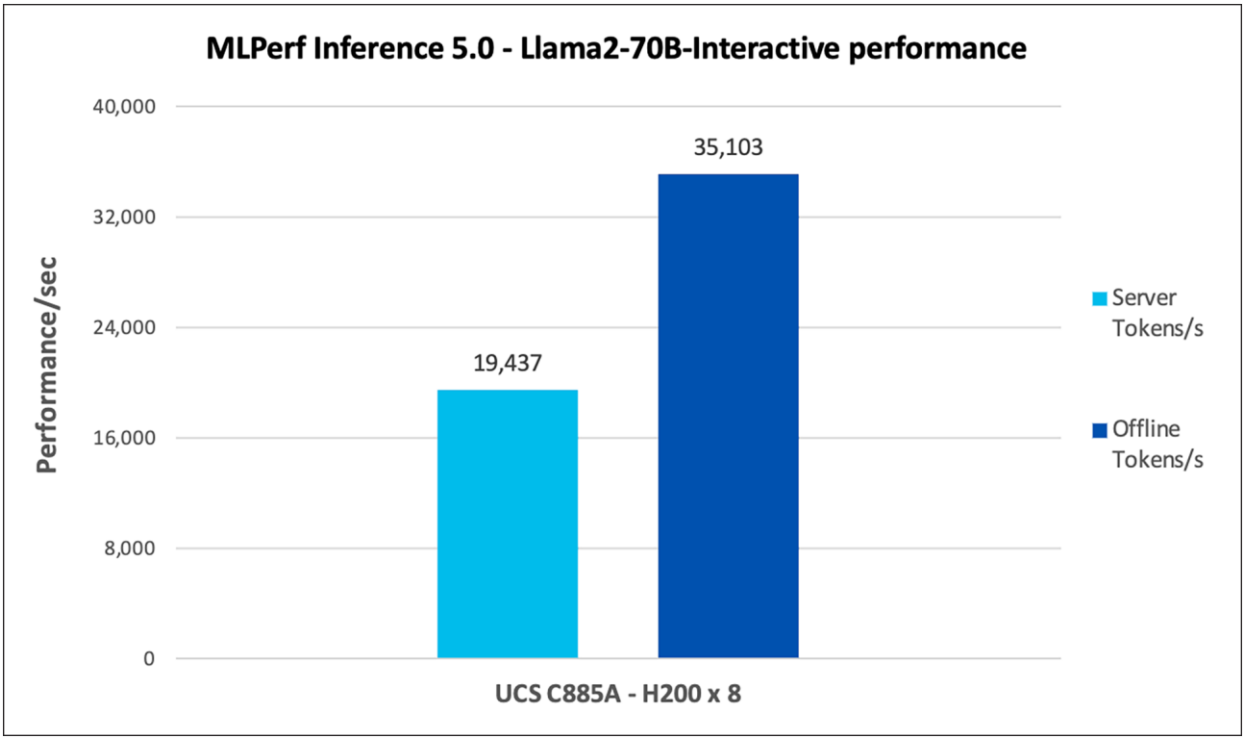


Figure 7.
Llama2-70B-Interactive performance data on Cisco UCS C885A M8 server with NVIDIA H100 and H200 GPUs

Note: For H100 Llama2-70B performance data, result is not verified by MLCommons Association as the results are collected after the MLPerf submission deadline.

Llama3.1-405b

Llama3.1-405B (speculative) would be a highly advanced language model with 405 billion parameters, likely an iteration of the Llama family, aimed at performing complex NLP tasks at scale, such as multi-turn dialogue, reasoning, and multi-modal processing.

Figure 8 shows the performance of the Llama2-70b model tested on UCS C885A M8 server with 8 x H100 and H200 GPUs.

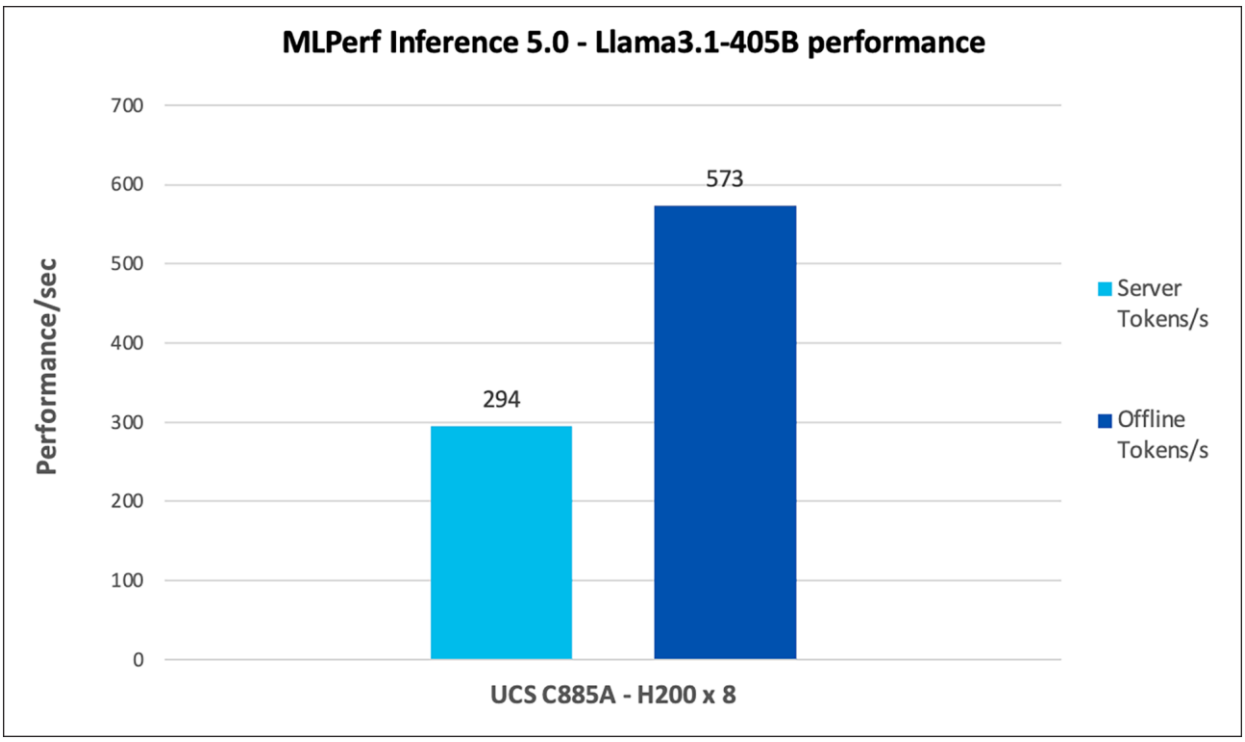


Figure 8.
Llama3.1-405b performance data on Cisco UCS C885A M8 server with Nvidia H100 and H200 GPUs

Mixtral-8x7B

Mixtral-8x7B is a mixture of experts model that combines multiple specialized “expert” sub-models, each with 7 billion parameters, to optimize performance on a range of tasks. It uses a gating mechanism to activate relevant experts during training and inference.

Figure 9 shows the performance of the Mixtral-8x7B model tested on UCS C885A M8 server with 8 x H100 and H200 GPUs.

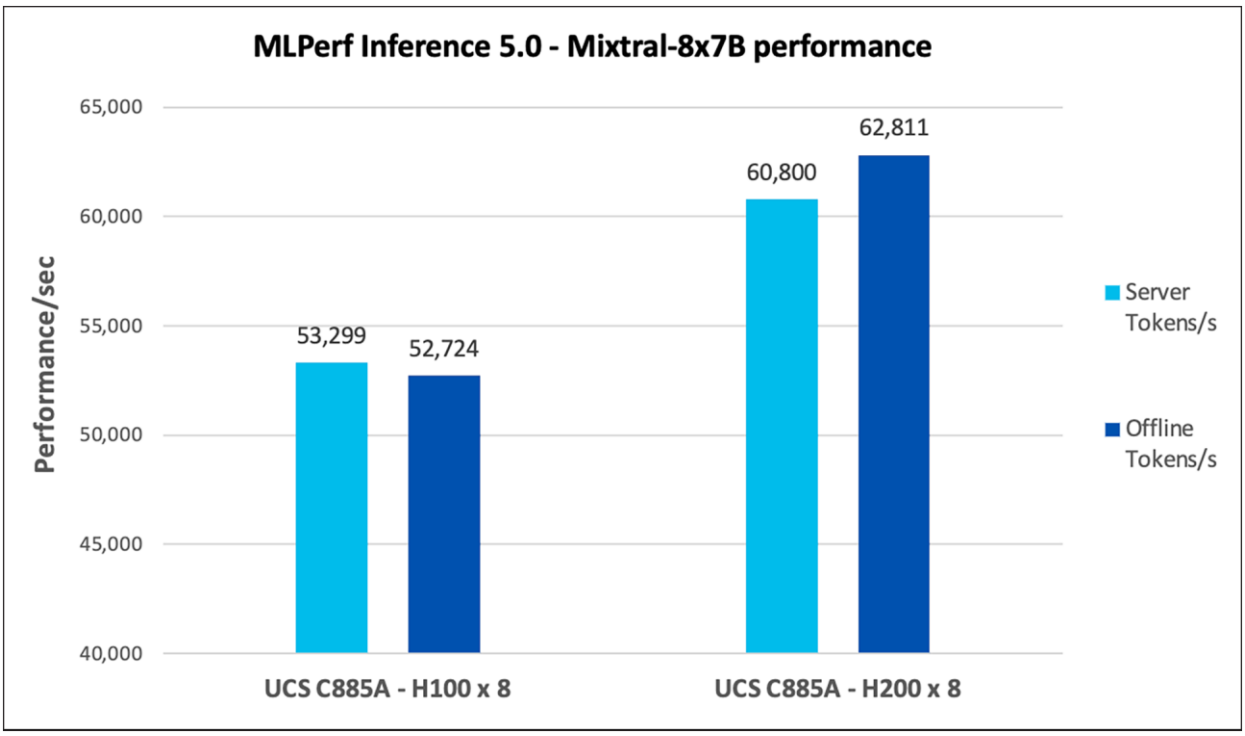


Figure 9.
Mixtral-8x7B performance data on Cisco UCS C885A M8 server with Nvidia H100 and H200 GPUs

Note: For H200 Mixtral-8x7B performance data, the result is not verified by MLCommons Association as the results are collected after the MLPerf submission deadline.

RGAT

Relational Graph Attention Network (RGAT) is a graph-based neural network model that uses attention mechanisms to learn from relational data. It is used for tasks like graph classification, link prediction, and node classification, where the relationships between entities are key.

Figure 10 shows the performance of the RGAT model tested on UCS C885A M8 server with 8 x H100 and H200 GPUs and this applicable only for Offline test scenario.

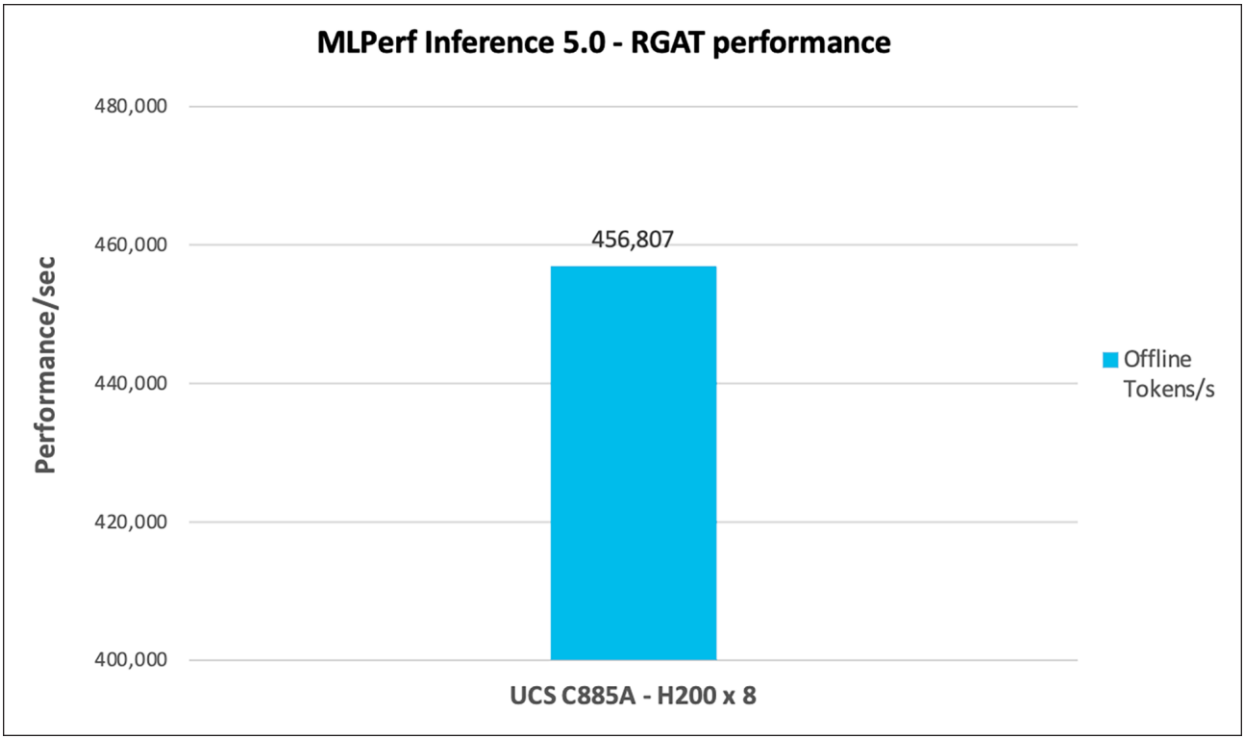


Figure 10.
RGAT performance data on Cisco UCS C885A M8 server with NVIDIA H100 and H200 GPUs

Performance summary

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

Cisco successfully submitted MLPerf 5.0 Inference results in partnership with NVIDIA to enhance performance and efficiency, optimizing various inference workloads such as Large language model (Language), Natural language processing (Language), Image Generation (Image), Generative image (Text to Image), Image classification (Vision), Object detection (Vision), Medical image segmentation (Vision), and Recommendation (Commerce).

The results were exceptional AI performance across Cisco UCS platforms for MLPerf Inference 5.0:

- The Cisco UCS C885A M8 platform with 8x NVIDIA H200 SXM GPUs emerged as the leader, securing first position for Llama-3.1-405B and second position for Llama2-70B-interactive and Gptj models.
- The Cisco UCS C885A M8 platform with 8x NVIDIA H100 SXM GPUs emerged as the leader, securing first position for Stable-diffusion-xl and Mixtral-8x7B models.

Appendix: Test environment

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

Table 2. Server properties

Name	Value
Product names	Cisco UCS C885A M8
CPUs	CPU: 2 x 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	• 8 x BlueField-3 E-series SuperNIC 400GbE/NDR • 2 x NIC cards
GPU controllers	• NVIDIA HGX H100 8-GPU • NVIDIA HGX H200 8-GPU
SFF NVMe SSDs	• 16 x 1.9 TB 2.5-inch high performance high endurance NVMe SSD

Note: For the Server BIOS settings, system default values were applied.

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

Table 3. Server BIOS settings

Name	Value
SMT control	Auto
NUMA nodes per socket	NPS4
IOMMU	Enabled
Core Performance Boost	Auto
Determinism enable	Power
APBDIS	1
Global C-state control	Disabled
DF C-states	Auto
Power Profile Selection	High Performance Mode

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-c885a-m8-aag.html

Data sheet: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c885a-m8-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

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Printed in USA C11-5143711-00 05/25