

AI Performance: MLPerf Inference on Cisco UCS C240 M8 Server with NVIDIA RTX Pro 4500 GPUs

September 2026



Executive summary

As generative AI transitions from experimental proof-of-concept to a critical production engine, the pressure on IT infrastructure has shifted from simple availability to extreme performance density. Organizations now require a foundation that balances high-compute throughput with the operational flexibility to scale seamlessly. The Cisco UCS® C240 M8 Rack Server is engineered to meet these rigorous demands, offering a modular, high-performance architecture specifically tuned for modern AI workflows. This white paper presents a comprehensive technical evaluation of the C240 M8 server, utilizing the latest MLPerf™ Inference benchmark results to demonstrate its performance superiority. We provide a deep dive into how this platform optimizes the data-to-insight pipeline, serving as the strategic cornerstone for enterprises seeking to operationalize AI at scale.

The Cisco UCS C240 M8 Rack Server was subjected to comprehensive testing in the MLPerf Inference v6.1 Datacenter for Closed Division.

Key findings include:

- The UCS C240 M8 demonstrates exceptional adaptability, supporting high-performance GPU configurations, including the NVIDIA RTX Pro 4500.
- MLPerf results confirm that the UCS C240 M8 delivers consistent, high-throughput inference performance, effectively minimizing latency for complex AI workloads.

To validate the AI performance capabilities of the new Cisco UCS C240 M8 Rack Server, Cisco conducted MLPerf Inference v6.1 Datacenter for Closed Division benchmarking using NVIDIA RTX Pro 4500 GPUs; the detailed results are presented in this document.

Scope of this document

This technical white paper delivers a granular AI inference performance analysis of the Cisco UCS C240 M8 Rack Server, validated through rigorous MLPerf Inference: Datacenter benchmark for Closed Division. We provide the empirical metrics and architectural insights necessary for engineers and designers to optimize their infrastructure strategies for the high-concurrency demands of modern, enterprise-scale AI.

The scope of this document encompasses the following key areas:

- **Hardware architecture:** an exploration of how the Cisco UCS C240 M8 Rack Server modular design – specifically its ability to decouple compute resources from GPU resources, which enhances overall performance efficiency and scalability.
- **Performance validation:** detailed analysis of inference throughput and latency metrics using 5x NVIDIA RTX Pro 4500 GPU configurations
- **Methodological framework:** an overview of the MLPerf Inference testing environment, including the software stack and dataset selection that supports the transparency and reproducibility of the results

By presenting these benchmarks, this document aims to demonstrate the platform's capacity to deliver consistent, high-performance results across diverse AI workloads, providing the technical evidence required to integrate the Cisco UCS C240 M8 Rack Server into modern, AI-centric data center architectures.

Product overview

Cisco UCS C240 M8 Rack Server

The introduction of the Cisco UCS C240 M8 Small Form-Factor (SFF) pluggable rack server extends the capabilities of the Cisco Unified Computing System™ portfolio in a 2U form factor with the Intel® Xeon® 6 scalable processors designed for high-performance computing with optimal power efficiency and balanced performance to boost your data-center productivity, supporting up to 5 NVIDIA RTX PRO 4500 Blackwell Server Edition GPUs.

The Cisco UCS C240 M8 Rack Server is equipped with two Intel Xeon 6 processors and offers standardization that easily integrates into existing environments. These servers offer ample amounts of performance, memory, and Peripheral Component Interconnect express (PCIe) slots to make it the ideal solution for both enterprise and scalable infrastructures.

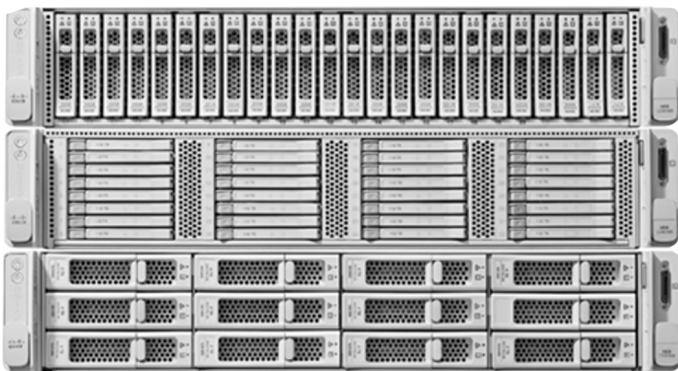


Figure 1. Cisco UCS C240 M8 Rack Server

Key benefits of AI + Cisco UCS

Cisco Unified Computing System (Cisco UCS) provides a robust platform for AI and Machine Learning (ML) workloads, offering a scalable and adaptable solution for various applications. By integrating AI with Cisco UCS, organizations can enhance resource utilization, accelerate insights, and maximize the value of AI investments.

- **High-density GPU support:** The UCS C240 M8 is engineered to support the latest high-performance accelerators, including the NVIDIA RTX Pro 4500. It provides the necessary bandwidth and thermal capacity to drive these GPUs at peak performance, supporting up to five PCIe Gen5 GPUs.
- **Scalability and flexibility:** Cisco UCS is designed to scale with evolving AI and machine learning workloads, allowing resources to be adjusted as needed to meet demand.
- **Performance optimization:** Cisco UCS delivers high-performance computing power suitable for data-intensive AI tasks, leveraging advanced CPUs such as Intel Xeon processors to accelerate deep learning and AI processes.
- **Diverse workload support:** The platform supports a wide range of AI and ML workloads, including deep learning, MLPerf inferencing, and specialized AI applications.
- **Unified and simplified management:** Cisco UCS offers a unified management platform with 100 percent programmability, enabling automation of server provisioning, configuration, and lifecycle management through Cisco Intersight®. This reduces deployment time and operational complexity.
- **Security and compliance:** Integrated security features such as role-based access control, secure boot, and data encryption help protect AI workloads and ensure compliance with industry regulations.



- **High-performance connectivity:** Cisco UCS supports high bandwidth and low-latency networking, enabling fast data transfer and consistent performance for AI workloads.
- **Enhanced resource utilization:** Integration with AI orchestration platforms like Run:ai on Cisco UCS optimizes GPU scheduling and resource allocation, improving utilization and reducing costs.

This overview positions the Cisco UCS C240 M8 Rack Server with NVIDIA RTX PRO 4500 GPUs as a versatile, high-performance server platform optimized for enterprise AI workloads including inference, big data analytics, and high-performance computing, combining powerful GPU acceleration with flexible management and operational efficiency.

MLPerf benchmark 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 Inference: Datacenter

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

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

The MLPerf Inference Benchmark paper, linked to the URL above, provides a detailed description of the motivation and guiding principles behind the MLPerf Inference: Datacenter benchmark suite.

Test configuration

For the MLPerf Inference performance testing covered in this document, the following Cisco UCS C240 M8 Rack Server configurations were used:

- 5x NVIDIA RTX Pro 4500 GPUs

MLPerf Inference models validated

The MLPerf Inference Datacenter models for Closed Division that were used for validating (see Table 1) were configured on a Cisco UCS C240 M8 Rack Server and tested for performance.

Table 1. MLPerf Inference 6.1 models

Model	Reference implementation model	Description
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

Llama3.1-8b (large language model):

Llama 3.1-8b is a state-of-the-art, compact, and powerful Large Language Model (LLM) with impressive capabilities in text generation, translation, and question answering. It is designed to deliver high-performance natural language processing with exceptional computational efficiency. With 8 billion parameters, this model represents a strategic balance between sophisticated reasoning capabilities and the low-latency requirements of modern production environments.

Llama 3.1-8b is engineered to provide high-speed inference, making it an ideal candidate for real-time applications where rapid response times are critical. Despite its smaller parameter count compared to larger models like the 70b variant, Llama 3.1-8b utilizes an advanced transformer architecture that excels in the following attributes:

- Due to its smaller memory footprint, Llama 3.1-8b can achieve significantly higher tokens-per-second generation, enabling massive concurrency in server-side inference scenarios.
- The model's architecture is optimized for minimal Time-To-First-Token (TTFT), making it highly effective for interactive AI assistants, real-time summarization, and automated customer support interfaces.
- The 8b (eight-billion) parameter scale allows for deployment on a wider range of hardware configurations, including edge-to-core deployments, without sacrificing the linguistic nuance and contextual accuracy expected of modern LLMs.

Llama 3.1-8b is a critical component of the modern AI ecosystem, offering a powerful combination of speed, accuracy, and efficiency. By benchmarking this model, this document illustrates the UCS C240 M8's capability to deliver high-performance, scalable inference solutions that meet the rigorous demands of enterprise AI deployments. Whether deployed for real-time interaction or high-volume data processing, the

Cisco® infrastructure ecosystem provides the robust foundation necessary to maximize the performance of Llama 3.1-8b at scale.

Whisper:

OpenAI's Whisper is a foundational Automatic Speech Recognition (ASR) model that has redefined the capabilities of speech-to-text processing. Trained on a massive, diverse dataset of 680,000 hours of multilingual, multitask supervised audio data collected from the web, Whisper is engineered to perform with high accuracy across a wide spectrum of acoustic environments.

The model is designed as a multitask transformer, enabling it to perform several key functions within a single inference pipeline:

- Support for 99 different languages, allowing for global-scale deployment
- Integrated capability to translate non-English audio directly into English
- Automatic detection of the source language, streamlining the processing of mixed-language datasets

In the context of this white paper, Whisper serves as a critical benchmark for compute-intensive inference workloads. Unlike LLMs that are primarily memory-bound, Whisper's inferencing involves complex signal processing and sequence-to-sequence modeling that places unique demands on the underlying AI infrastructure.

Whisper represents a significant advancement in the accessibility and accuracy of speech-to-text technology. By benchmarking Whisper on the Cisco UCS C240 M8 Rack Server, the tested results in this document provides technical evidence of the platform's versatility. It demonstrates that the Cisco infrastructure ecosystem is not only optimized for the memory-heavy demands of LLMs, but it is equally capable of delivering the high-throughput, low-latency performance required for complex, real-time ASR workloads.

MLPerf Inference benchmarking methodology

As part of the MLPerf Inference submission, Cisco conducted rigorous performance testing across the datasets outlined in Table 1, utilizing the Cisco UCS C240 M8 Rack Server with RTX Pro 4500 GPUs. These results, which have been formally submitted to and published by MLCommons, provide an objective, third-party-verified assessment of our infrastructure's performance capabilities. The complete dataset and official results are available for review on the MLPerf Inference: Datacenter results page.

To provide a holistic view of performance across varying enterprise requirements, MLPerf evaluates systems using two distinct operational scenarios:

- **Offline scenario (throughput-centric):** This scenario measures the system's maximum processing capacity by allowing the server to ingest all input data at once. It is the primary metric for assessing raw throughput in batch-processing environments, such as large-scale data analysis or asynchronous model training, where latency is secondary to total volume.
- **Server scenario (latency-sensitive):** This scenario simulates real-world production environments by processing requests as they arrive. It measures both throughput and latency, ensuring that the system maintains a consistent response time within a strictly defined latency threshold. This scenario is critical for evaluating the performance of interactive AI applications, such as real-time generative AI assistants or automated decision-making engines, where user experience is directly tied to response speed.

Note: Certain of the performance graphs presented below include preliminary results obtained after the official MLPerf submission deadline; these data points have not been formally verified by MLCommons. For such graphs, there is an added note: "Result not verified by MLCommons Association."

Performance data for NVIDIA RTX Pro 4500 GPU

The integration of the NVIDIA RTX Pro 4500 GPU within the Cisco UCS C240 M8 Rack Server delivers a versatile and balanced solution tailored for enterprise AI inference and generative AI fine-tuning. The RTX Pro 4500 is designed for department-scale AI and intelligent automation, supporting multi-workload environments such as AI inference, AI video, data processing, and computer vision, all within a compact 2U server form factor that balances performance, power, and cooling requirements effectively.

Performance benchmarking and operational insights

Validated through MLPerf Inference v6.1 Datacenter for Closed Division benchmarking, the Cisco UCS C240 M8 Rack Server configured with 5x NVIDIA RTX Pro 4500 GPUs demonstrates consistent and reliable performance across various datasets. Key technical observations include:

- **Optimized throughput:** The UCS C240 M8 server architecture provides the necessary Gen5 bandwidth to ensure that the RTX Pro 4500 GPUs are not bottlenecked, allowing for high-efficiency inference across batch-processing (offline) and real-time (server) scenarios.
- **Thermal and power stability:** Benefits of using RTX Pro 4500 GPUs in UCS C240 M8 servers include advanced thermal management. By leveraging the server's ability to support up to 600W per card, the RTX Pro 4500 operates within its optimal performance envelope, avoiding the throttling common in space-constrained rack servers.
- **Balanced resource utilization:** The configuration excels in scenarios where a balance between compute density and power efficiency is required. This makes it a preferred choice for organizations deploying diverse AI models, including computer vision, natural language processing, and predictive analytics.
- **Unified management:** By managing the RTX Pro 4500 configurations through Cisco Intersight, administrators gain granular visibility into GPU utilization and health, ensuring that the infrastructure remains optimized for consistent, high-performance output.



Performance data of the llama3.1-8b model:

Figure 2 shows the performance of the llama3.1-8b model, tested on a Cisco UCS C240 M8 Rack Server with 5x NVIDIA RTX Pro 4500 GPUs.

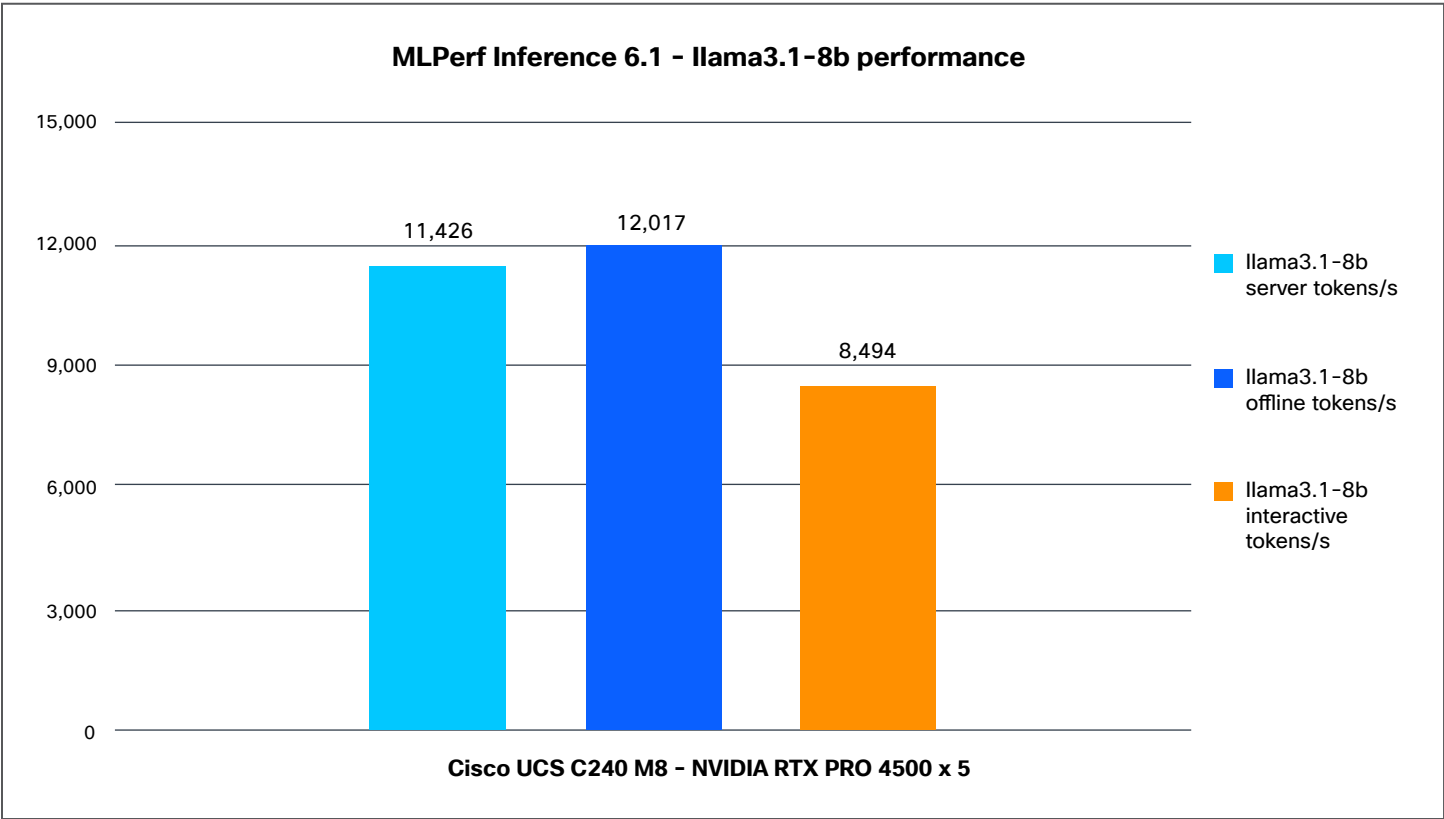


Figure 2. Llama3.1-8b performance data on a Cisco UCS C240 M8 Rack Server with 5x NVIDIA RTX Pro 4500 GPUs



Figure 3 shows the scaling performance of the llama3.1-8b model, tested on a Cisco UCS C240 M8 Rack Server with 5x NVIDIA RTX Pro 4500 GPUs.

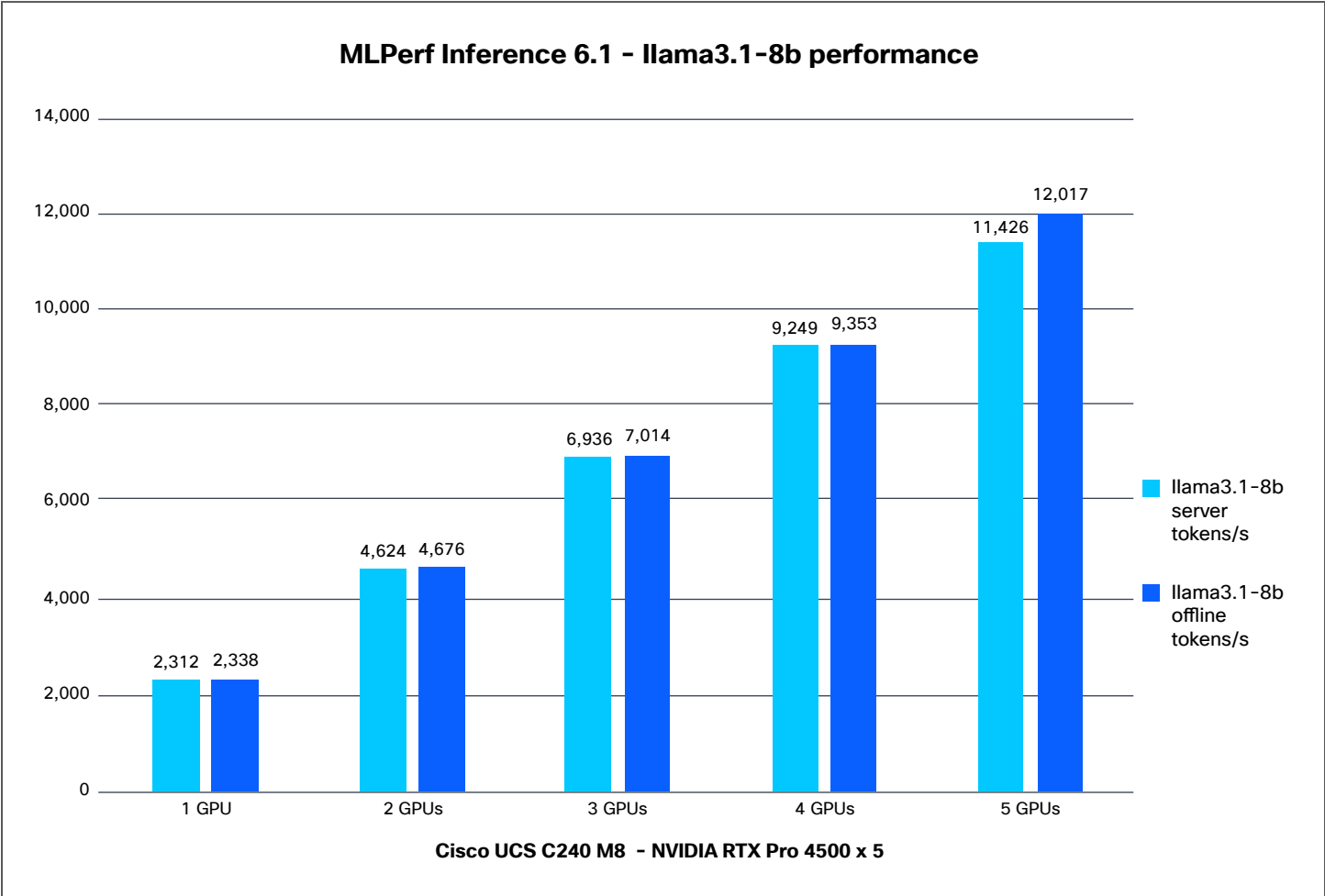


Figure 3. Llama3.1-8b inference scaling performance data of Cisco UCS C240 M8 Rack Server with 5x NVIDIA RTX Pro 4500 GPUs



Performance data of the Whisper model:

Figure 4 shows the performance of the Whisper model tested on a Cisco UCS C240 M8 Rack Server with 5x NVIDIA RTX Pro 4500 GPUs.

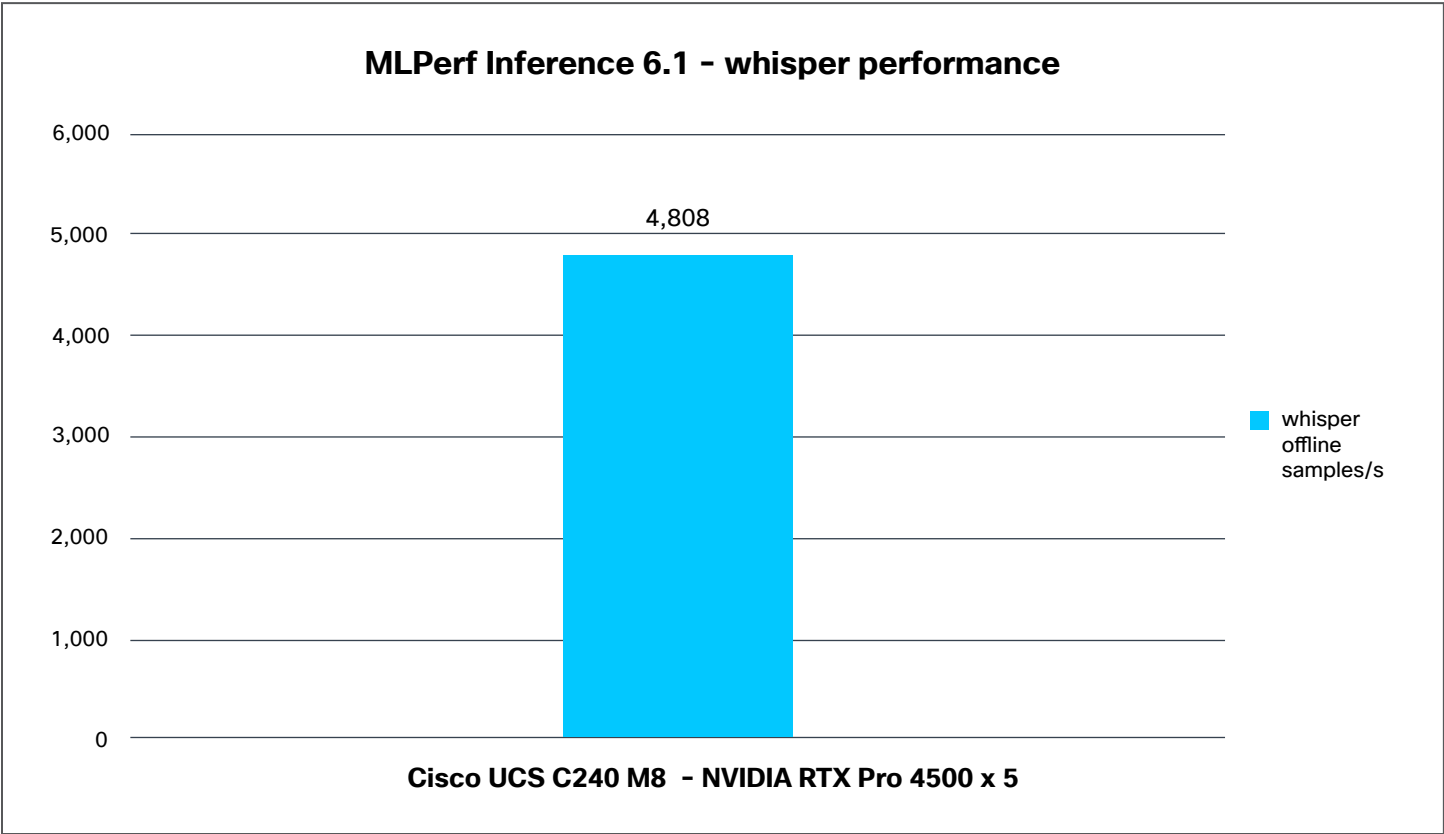


Figure 4. Whisper performance data of Cisco UCS C240 M8 Rack Server with 5x NVIDIA RTX Pro 4500 GPUs



Figure 5 shows the scaling performance of the Whisper model, tested on a Cisco UCS C240 M8 Rack Server with 5x NVIDIA RTX Pro 4500 GPUs.

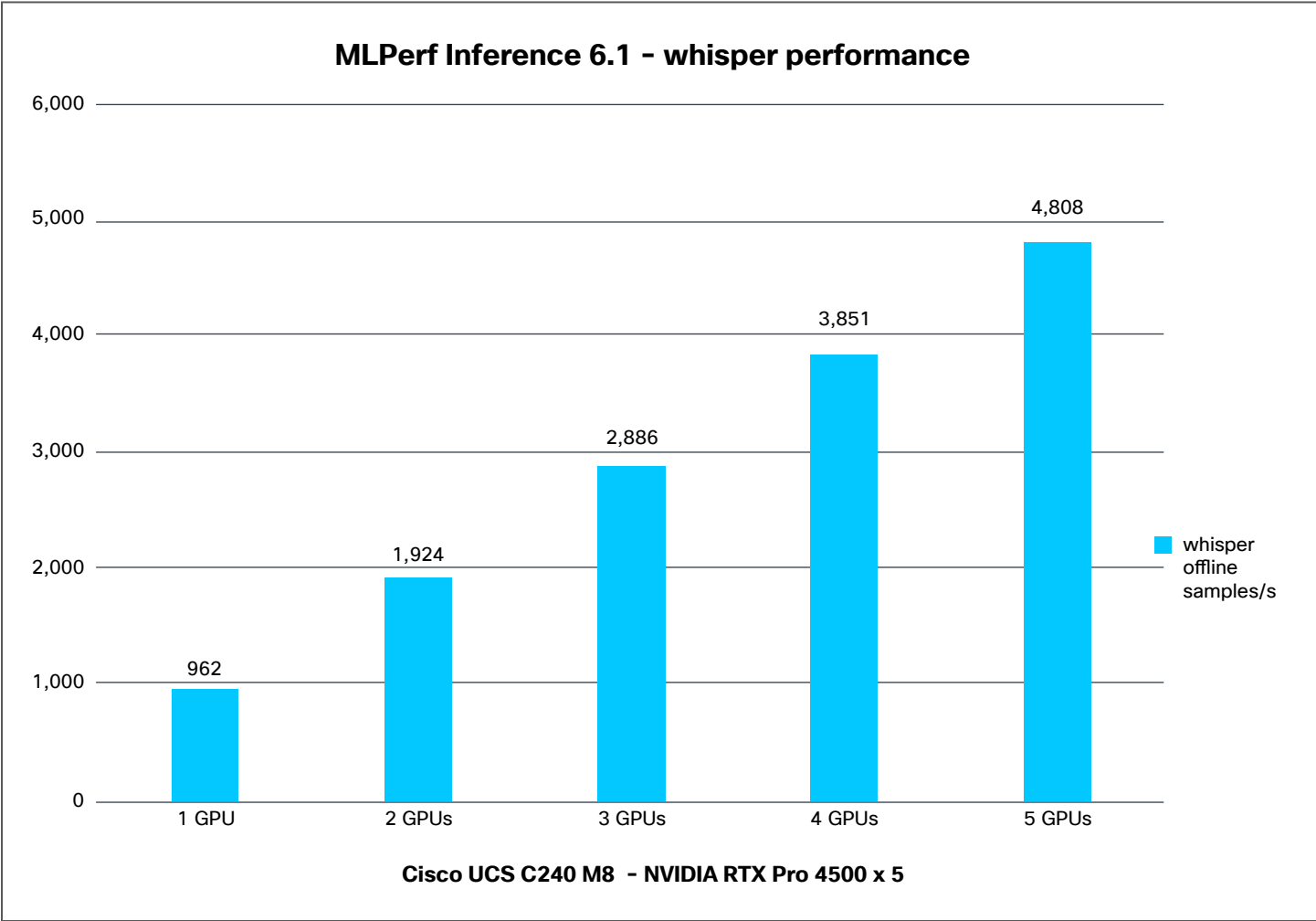


Figure 5. Whisper scaling performance data on a Cisco UCS C240 M8 Rack Server with 5x NVIDIA RTX Pro 4500 GPUs



Performance summary

The Cisco UCS C240 M8 serves as the high-performance backbone for enterprise AI, combining modular density with the raw computational power of the latest NVIDIA GPU accelerators. This architecture streamlines the transition from experimentation to production-scale operations, ensuring that organizations can deliver the speed, reliability, and efficiency required for today’s most demanding AI workloads

In partnership with NVIDIA, Cisco submitted comprehensive results for MLPerf Inference v6.1 Datacenter for Closed Division. These results validate the platform’s ability to deliver industry-leading performance across a diverse range of generative AI workloads, including LLMs, speech-to-text processing, and generative image synthesis. The benchmark data confirms that the Cisco UCS C240 M8 Rack Server maintains exceptional throughput and latency characteristics, even under the intensive computational demands of modern AI models.

Key performance achievements

The MLPerf Inference results highlight the leadership position of the Cisco UCS C240 M8 Rack Server in key AI inference categories:

- **Llama 3.1-8b leadership:** The Cisco UCS C240 M8 Rack Server, configured with 5x NVIDIA RTX Pro 4500 GPUs, achieved the top-ranked position for the llama 3.1-8b model. This result underscores the platform’s ability to handle memory-intensive, high-concurrency LLM inference with superior efficiency.
- **Whisper ASR leadership:** The Cisco UCS C240 M8 Rack Server, configured with 5x NVIDIA RTX Pro 4500 GPUs, secured the first position ranking for the Whisper speech recognition model. This demonstrates the server’s versatility in managing complex, real-time signal processing and sequence-to-sequence inference tasks.

Table 2. Performance summary table for the Cisco UCS C240 M8 Rack Server with 5x NVIDIA RTX Pro 4500 GPUs

Benchmark/model	Inferences/s		
	Offline	Server	Interactive
llama3.1-8b	12,017 tokens/sec	11,426 tokens/sec	8494 tokens/sec
Whisper	4808 samples/sec	NA	NA

Conclusion

The MLPerf Inference results confirm that the Cisco UCS C240 M8 Rack Server are a critical foundation for enterprise-grade AI, delivering high performance for AI-accelerated workloads optimized with NVIDIA RTX Pro 4500 GPUs. This 2U rack server offers the flexibility and density necessary to meet the evolving demands of generative AI and inference tasks.

The white paper provides the technical validation that data center architects and decision-makers need to confidently incorporate the UCS C240 M8 into their AI strategies. When integrated with the broader Cisco AI infrastructure ecosystem, the server enables organizations to sustain peak performance, optimize resource use, and achieve the operational agility required for AI-centric data centers.

Appendix: Test environment

For MLPerf Inference testing, we have used the hardware configurations given in Table 3.

Table 3. Server properties

Description	Specification
Product name	Cisco UCS C240 M8 Rack Server
Product type	Data center
Number of nodes	1
Processor name	2x Intel Xeon 6747P Processor
Host processor core count	86
# of threads	172
Host processor frequency	2.00 GHz, 3.80 GHz Turbo Boost
Total memory	1 TB
Memory DIMMs	16x 64GB DIMMs
Memory type	DDR5 (6400MT/s)
Network adapter	Cisco VIC 15237 mLOM
Host storage capacity	7.6TB 2.5in U.3 Micron 7500 NVMe High Perf High Endurance
GPU controllers	5 x NVIDIA RTX Pro 4500 PCIe GPUs

Note: We configured platform-default BIOS settings during the testing.

For more information

For additional information on the Cisco UCS C240 M8 Rack Server:

- <https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c240-m8-rack-server-aag.html>.
- <https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c240-m8-sff-rack-server.pdf>.

For published MLPerf Inference results: <https://mlcommons.org/benchmarks/inference-datacenter/>.