AI Performance: MLPerf Inference on Cisco UCS X580p PCIe Node with NVIDIA H200 NVL and RTX Pro 6000 GPUs

April 2026

Executive summary

As organizations transition from AI experimentation to large-scale production, the demand for infrastructure that balances high-performance compute with operational flexibility has never been greater. The Cisco UCS® X580p PCIe Node represents a significant evolution in modular infrastructure, specifically engineered to address the rigorous demands of generative AI fine-tuning and inference. This white paper details the performance capabilities of the UCS X580p, validated through the latest MLPerf Inference benchmarking results.

The Cisco UCS X580p PCIe Node was subjected to comprehensive testing in the MLPerf Inference: Datacenter for Closed division.

Key findings include:

- The UCS X580p demonstrates exceptional adaptability, supporting high-performance GPU configurations, including the NVIDIA H200 NVL and RTX Pro 6000.
- MLPerf results confirm that the UCS X580p delivers consistent, high-throughput inference performance, effectively minimizing latency for complex AI workloads.
- As a core component of the Cisco UCS X-Series, the UCS X580p integrates into existing fabric-based architectures and thus helps to ensure that AI inference performance is supported by high-bandwidth, low-latency networking.

To validate the AI performance capabilities of the new Cisco UCS X580p PCIe node, Cisco conducted MLPerf Inference v6.0 Datacenter for Closed Division benchmarking using NVIDIA H200 NVL and RTX Pro 6000 GPUs; the detailed results are presented in this document.

Scope of this document

This white paper provides a comprehensive technical evaluation of the AI inference performance of the Cisco UCS X580p PCIe Node, validated through the MLPerf Inference: Datacenter benchmark for Closed Division. The primary objective of this document is to provide architects, data-center engineers, and IT decision-makers with empirical performance data to support infrastructure planning and deployment strategies for enterprise-grade AI.

The scope of this document encompasses the following key areas:

- **Hardware architecture:** An exploration of how the Cisco UCS X580p PCIe node's modular design—specifically its ability to decouple compute from GPU resources—impacts performance efficiency and scalability.
- **Performance validation:** Detailed analysis of inference throughput and latency metrics using 4x NVIDIA H200 NVL and 4x NVIDIA RTX Pro 6000 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 X580p PCIe Node into modern, AI-centric data center architectures.

Product overview

The introduction of the Cisco UCS X580p PCIe Node and the Cisco UCS X9516 X-Fabric Module marks the next evolution in blade computing. The existing Cisco UCS X9508 Chassis supports the new Cisco UCS X580p PCIe Nodes (UCSX-580P) and Cisco UCS X9516 (UCSX-F-X9516) X-Fabric Modules paired with Cisco UCS X210c M8 or Cisco UCS X215c M8 compute nodes. The servers will need the new PCIe Gen5 mezz card (UCSX-V5-PCIME) to communicate with the UCS X9516 X-Fabric modules. UCS X580p supports the NVIDIA L40S, H200-NVL, or the RTX 6000 GPUs. Given that the GPUs are populated side by side in each cage, the X580p also supports the two-way NVLink bridges attached to the NVIDIA H200-NVL GPUs. The UCS X9516 also supports NVIDIA's ConnectX-7 SmartNICs. Customers may choose either a 2x200GbE CX-7 or a 1x400GbE CX-7.

The Cisco UCS X580p Compute Node and Cisco UCS X9516 X-Fabric Module represent a paradigm shift in blade computing, enabling high-performance workloads, including NVLink-bridged GPU configurations and dedicated east/west fabrics, to be deployed in a modular, efficient blade form-factor. This innovation adopts a disaggregated infrastructure model, decoupling GPUs and SmartNICs from the motherboard to facilitate dynamic resource allocation. The Cisco UCS X580p PCIe Node allows you to add up to four GPUs to Cisco UCS X210c and X215c compute nodes with Cisco UCS X-Fabric Technology. Now you can easily and independently manage the different lifecycles of CPU and GPU components.

Cisco UCS X580P PCIe Node:

The Cisco UCS X580p PCIe Node is a high-performance, modular compute extension designed to meet the rigorous demands of modern AI, machine learning, and high-performance computing (HPC) workloads. As a core component of the Cisco UCS X-Series, the UCS X580p extends the platform's modularity, enabling organizations to deploy specialized GPU-accelerated resources within a unified, fabric-based architecture.

The UCS X580p represents a significant shift in data-center design by adopting a disaggregated infrastructure approach. By decoupling GPUs and SmartNICs from the host server's motherboard, the UCS X580p allows for dynamic resource provisioning. This architecture enables IT teams to assign compute, networking, and acceleration resources on demand, ensuring that infrastructure can scale in lockstep with evolving workload requirements throughout the server lifecycle.

Figure 1. Cisco UCS X580p PCIe Node

Key technical capabilities:

- **High-density GPU support:** The UCS X580p is engineered to support the latest high-performance accelerators, including the NVIDIA H200 NVL and RTX Pro 6000. It provides the necessary bandwidth and thermal capacity to drive these GPUs at peak performance, supporting up to four PCIe Gen5 GPUs per node.
- **Advanced Connectivity:** The node supports NVIDIA's NVLink two-way bridge across any slot within the same PCIe zone, facilitating high-speed, low-latency communication between GPUs. This is critical for large-scale generative AI fine-tuning and inference tasks.
- **Thermal and power efficiency:** Designed for high-density environments, the UCS X580p supports air-cooled GPUs with thermal envelopes of up to 600W per card. Its advanced power delivery system eliminates the need for chassis-level power upgrades and prevents the performance throttling often associated with traditional rack-server power constraints.
- **Unified management:** Fully integrated into the Cisco Intersight® platform, the UCS X580p utilizes policy-based connectivity. Administrators can assign specific GPU and SmartNIC resources to any server in the chassis through software, simplifying complex configuration tasks and reducing operational overhead.

Integration with the Cisco UCS ecosystem:

The UCS X580p operates within the unified Cisco UCS X-Series environment, leveraging the high-bandwidth, low-latency Cisco UCS fabric. When combined with Cisco® Silicon One®-based networking and the broader Cisco UCS X-Series ecosystem, the UCS X580p provides a validated, full-stack infrastructure. This integration ensures that AI inference workloads are supported by a reliable, secure, and highly scalable foundation, enabling organizations to transition from AI experimentation to large-scale, production-ready deployments with confidence.

Cisco UCS X9516 X-Fabric Module:

The Cisco UCS X9516 X-Fabric Module serves as the high-speed connectivity backbone of the Cisco UCS X-Series modular system. Engineered to provide massive, low-latency bandwidth, the UCS X9516 is the critical component that enables the Cisco UCS X-Series to function as a disaggregated, fabric-based infrastructure, allowing compute nodes and PCIe resources to communicate with unprecedented efficiency.

The UCS X9516 module transforms the traditional chassis backplane into a high-performance fabric. By providing dedicated PCIe connectivity between the server nodes and the PCIe nodes (such as the Cisco UCS X580p), the UCS X9516 eliminates the physical constraints of traditional server architectures. This design allows for the flexible, policy-driven assignment of hardware resources, enabling a single server node to access remote GPUs and SmartNICs as if they were locally installed.

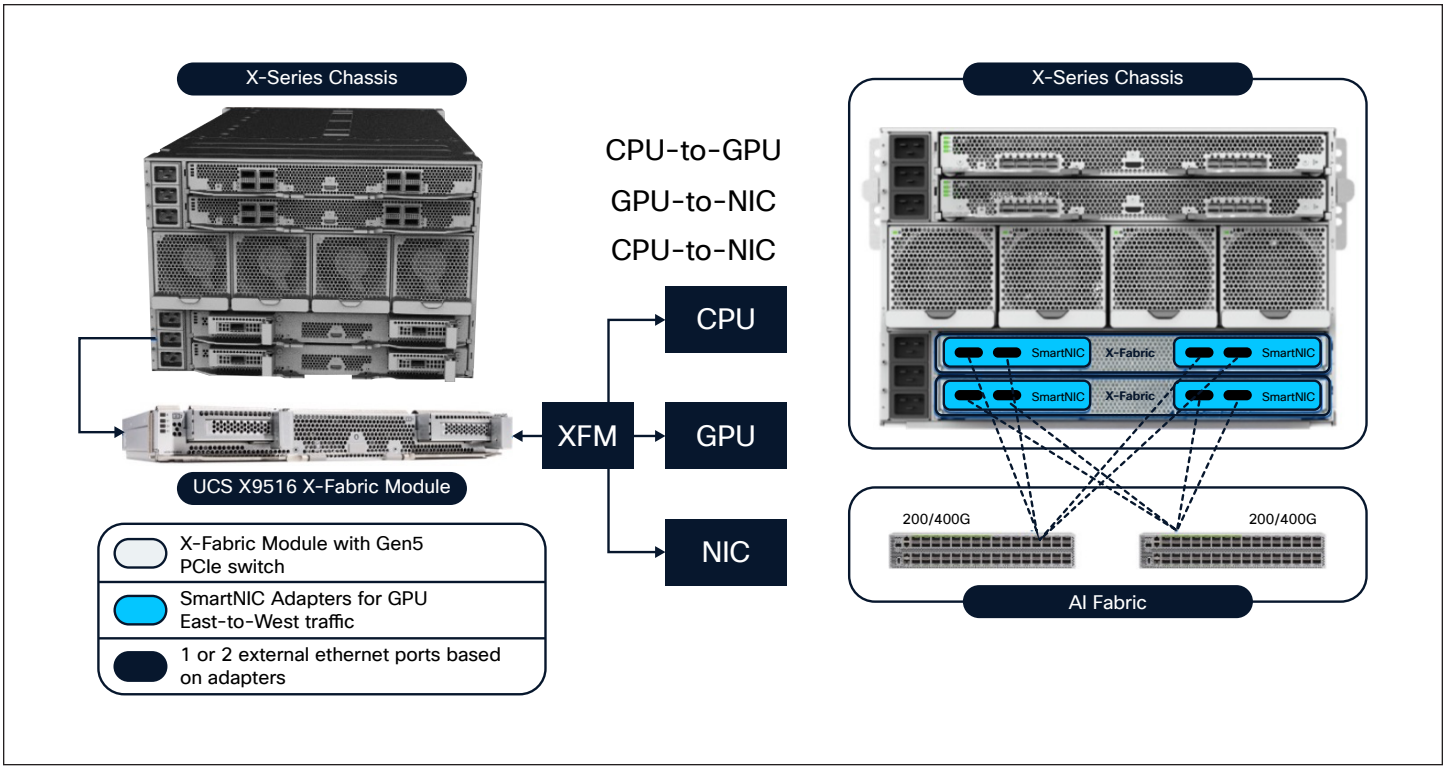

Figure 2. Cisco UCS X9516 X-Fabric Module

Key technical capabilities:

- **High-bandwidth PCIe fabric:** The UCS X9516 supports high-speed PCIe Gen5 connectivity, ensuring that data-intensive AI and HPC workloads experience minimal latency. This is essential for maintaining the high throughput required by modern generative AI models and large-scale inference tasks.
- **Disaggregated resource pooling:** The module facilitates the disaggregation of infrastructure, allowing IT administrators to pool GPU and networking resources. This architecture enables the dynamic scaling of hardware, where resources can be provisioned or reallocated to different server nodes based on real-time workload demands.
- **Seamless integration with Cisco Intersight:** The UCS X9516 is fully managed through the Cisco Intersight platform. Through Intersight PCIe connectivity policies, administrators can define the fabric topology, assigning specific PCIe resources to compute nodes without manual cabling or physical intervention.
- **Scalability and future-proofing:** The modular design of the UCS X9516 ensures that the fabric can support evolving hardware generations. By providing a robust, high-speed interconnect, the module allows organizations to upgrade compute and acceleration components independently, extending the lifecycle of the chassis and reducing the need for costly infrastructure refreshes.

Strategic value for the enterprise:

The Cisco UCS X9516 X-Fabric Module is the engine that drives the agility of the Cisco UCS X-Series. Its integration into the X-Series chassis provides several strategic advantages:

- **Optimized resource efficiency:** By enabling the sharing of high-performance GPUs and SmartNICs across multiple nodes, the UCS X9516 reduces the Total Cost of Ownership (TCO) by maximizing the utilization of expensive hardware assets.
- **Simplified data-center operations:** The fabric-based approach replaces complex, siloed cabling with a unified, software-defined connectivity layer. This streamlines deployment, simplifies maintenance, and reduces the potential for human error.
- **Enhanced performance for AI workloads:** The low-latency fabric is optimized for the high-speed data transfers required by AI inference and fine-tuning. It ensures that the compute nodes are never bottlenecked by the interconnect, allowing for peak performance across the entire AI stack.

Scalable network fabric for AI connectivity

In the era of large-scale AI and machine learning, the network fabric is no longer a peripheral component; it is the backbone of the entire AI infrastructure. As compute clusters scale to thousands of GPUs, the interconnect must provide ultra-low latency, high bandwidth, and lossless data transmission to prevent GPU starvation. The Cisco Nexus® 9000 Series Switches serve as the foundational building blocks for this high-performance fabric, specifically engineered to meet the stringent demands of AI/ML workloads.

The role of Cisco Nexus 9000 in AI fabrics

The Cisco Nexus 9000 Series provide the high-density, non-blocking switching fabric required to interconnect compute nodes and storage clusters. By supporting high-speed interfaces—ranging from 100G to 800G—the Nexus 9000 Series ensure that the massive data movement associated with model training and inference occurs with minimal latency.

Key architectural capabilities for AI

- **Lossless Ethernet through RoCEv2:** AI workloads rely heavily on remote direct memory access (RDMA). The Nexus 9000 Series provide full support for **RoCEv2 (RDMA over converged Ethernet)**, enabling direct memory-to-memory data transfer between GPUs across the network. This bypasses the CPU, significantly reducing latency and jitter.
- **Advanced congestion management:** To maintain a lossless fabric, the Nexus 9000 Series utilize sophisticated congestion control mechanisms, including **Priority Flow Control (PFC)** and **Explicit Congestion Notification (ECN)**. These features work in tandem to prevent packet loss and manage traffic bursts, ensuring that AI model synchronization remains stable under heavy load.
- **Leaf-spine architecture:** The Nexus 9000 Series are designed for high-radix, non-blocking leaf-spine topologies. This architecture provides predictable, low-latency paths between any two endpoints in the fabric, allowing organizations to scale their AI clusters linearly from a single rack to massive, multi-pod deployments without performance degradation.
- **Deep buffer and high throughput:** For AI applications that involve large-scale dataset ingestion and checkpointing, the Nexus 9000 Series' deep-buffer architecture ensures that traffic spikes do not lead to dropped packets, maintaining the integrity of the data stream.

Integration with Cisco UCS X-Series

The synergy between the Cisco UCS X-Series (compute) and the Nexus 9000 (fabric) creates a seamless, full-stack AI infrastructure. The UCS X-Series connects to the Nexus fabric through high-bandwidth uplinks, ensuring that the compute nodes are integrated directly into the high-speed network. This integration allows for:

- **Unified management:** simplified visibility across the entire compute and network stack through Cisco Intersight.
- **Optimized traffic flows:** intelligent traffic steering that helps ensure that AI-specific workloads receive the necessary bandwidth and priority across the fabric.

Strategic value for AI deployments

Cisco Nexus 9000 Series Switches provide the reliability and scalability required to transform AI potential into production reality.

A scalable, high-performance network fabric is the critical enabler of modern AI. By leveraging Cisco Nexus 9000 Series Switches, organizations can build a robust, lossless, and highly efficient network that removes bottlenecks and empowers compute nodes to operate at peak efficiency. Whether supporting real-time inference or massive model training, the Nexus 9000 Series provide the deterministic performance and architectural flexibility required for the AI-centric data center.

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

Figure 3. Cisco Nexus 9364E-SG2 switch for AI connectivity

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

https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/cisco_ai_pod_for_training_design.html.

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 two Cisco UCS X580p PCIe Node configurations were used:

- 4x NVIDIA H200-NVL PCIe GPUs
- 4x NVIDIA RTX Pro 6000 PCIe 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 X580p PCIe Node and tested for performance.

Table 1. MLPerf Inference models

Model	Reference implementation model	Description
Llama2-70b	language/llama2-70b	Large language model with 70 billion parameters. It is designed for Natural Language Processing (NLP) tasks and answering questions
Llama3.1-8b	language/llama3.1-8b	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
Mixtral-7x8b	LLM – Text generation (Question Answering, Math and Code Generation)	High-performance Sparse Mixture-of-Experts (SMoE) language model

Llama2-70b (large language model):

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. This dense transformer-based model is engineered to handle complex Natural Language Processing (NLP) tasks with high precision and contextual understanding.

Key capabilities include:

- Exceptional performance in text generation, creative writing, and nuanced conversational interaction.
- Advanced capabilities in summarization, multi-lingual translation, and complex question-answering.
- The ability to maintain coherence across long-form inputs, making it suitable for enterprise-grade documentation analysis and automated customer support.

In the context of this white paper, Llama 2-70b serves as a critical proxy for modern, compute-intensive generative AI workloads. Because the model requires substantial resources for weight loading and activation processing, it provides a definitive test of the Cisco UCS X580p PCIe Node's ability to manage high-throughput inference. By benchmarking Llama 2-70b, this document demonstrates how the Cisco infrastructure ecosystem effectively balances the memory-intensive demands of large-scale models with the operational agility required for production-ready AI deployments.

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 X580p'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 inference 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 X580p PCIe Node, 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 is equally capable of delivering the high-throughput, low-latency performance required for complex, real-time WASR workloads.

Mixtral-7x8b:

The Mixtral-8x7b model, developed by Mistral AI, represents a transformative shift in LLM architecture. Unlike traditional dense transformer models, Mixtral-8x7b utilizes a Mixture-of-Experts (MoE) architecture. This sparse model features a total of 47 billion parameters, yet it only activates approximately 13 billion parameters per token. This design allows the model to achieve the performance levels of much larger dense models while maintaining the inference speed and computational efficiency of a significantly smaller model.

The MoE approach introduces unique computational requirements that distinguish it from standard dense LLMs. Key characteristics include:

- Through a routing mechanism, the model dynamically selects the most relevant "expert" sub-networks for each input token. This results in high-quality reasoning and contextual understanding without the massive compute overhead typically required by dense 47 billion+ parameter models.
- While the active parameter count is 13 billion, the total model weights (47 billion) must reside in GPU memory. This makes Mixtral-8x7b highly memory-bandwidth-intensive, because the system must rapidly fetch weights for the various experts during the inference process.
- The expert-routing layer adds a unique dimension to the inference pipeline, requiring the underlying infrastructure to handle frequent, small-batch memory access patterns efficiently.

In the context of this white paper, Mixtral-8x7b serves as a critical benchmark for evaluating sparse-activation inference on the Cisco UCS X580p PCIe Node. Mixtral-8x7b is a benchmark for the next wave of efficient, high-performance generative AI. By benchmarking this model on the Cisco UCS X580p PCIe Node, this document provides empirical evidence of the platform's ability to handle the complex, memory-intensive demands of sparse-activation models. This validation confirms that the Cisco infrastructure ecosystem is uniquely positioned to empower enterprises to deploy sophisticated, MoE-based AI solutions with the reliability, speed, and agility required for modern, AI-centric data centers.

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 X580P PCIe Node. 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 the Cisco UCS X580p PCIe Node with NVIDIA H200 NVL PCIe GPUs

The combination of the Cisco UCS X580p PCIe Node and the NVIDIA H200 NVL GPU represents a significant milestone in high-performance AI infrastructure. Validated through the MLPerf Inference v6.0 (Datacenter) benchmark suite, this configuration demonstrates a capability to deliver massive throughput and low-latency inference for the most demanding generative AI and LLM workloads.

The MLPerf Inference v6.0 results highlight the operational efficiency of the UCS X580p when configured with 4x NVIDIA H200 NVL GPUs. Key performance observations include:

- **High-throughput inference:** The UCS X580p architecture effectively manages the high-bandwidth requirements of the H200 NVL, enabling superior performance in offline scenarios where total system throughput is the primary metric.
- **Latency consistency:** In server-side scenarios, the UCS X580p maintains strict adherence to latency thresholds. By leveraging the high-speed PCIe Gen5 fabric and the optimized data path of the Cisco UCS X-Series, the system ensures that inference requests are processed within the tight time windows required for real-time AI applications.
- **NVLink advantage:** The support for NVIDIA's NVLink two-way bridge within the UCS X580p node is a critical performance driver. By enabling high-speed, direct GPU-to-GPU communication, the system minimizes data movement overhead, which is essential for models such as Llama 2-70b that require significant inter-GPU synchronization during inference.

Performance data of the Llama3.1-8b model:

Figure 4 shows the performance of the Llama3.1-8b model, tested on a Cisco UCS X580p PCIe Node with 4x NVIDIA H200 NVL GPUs.

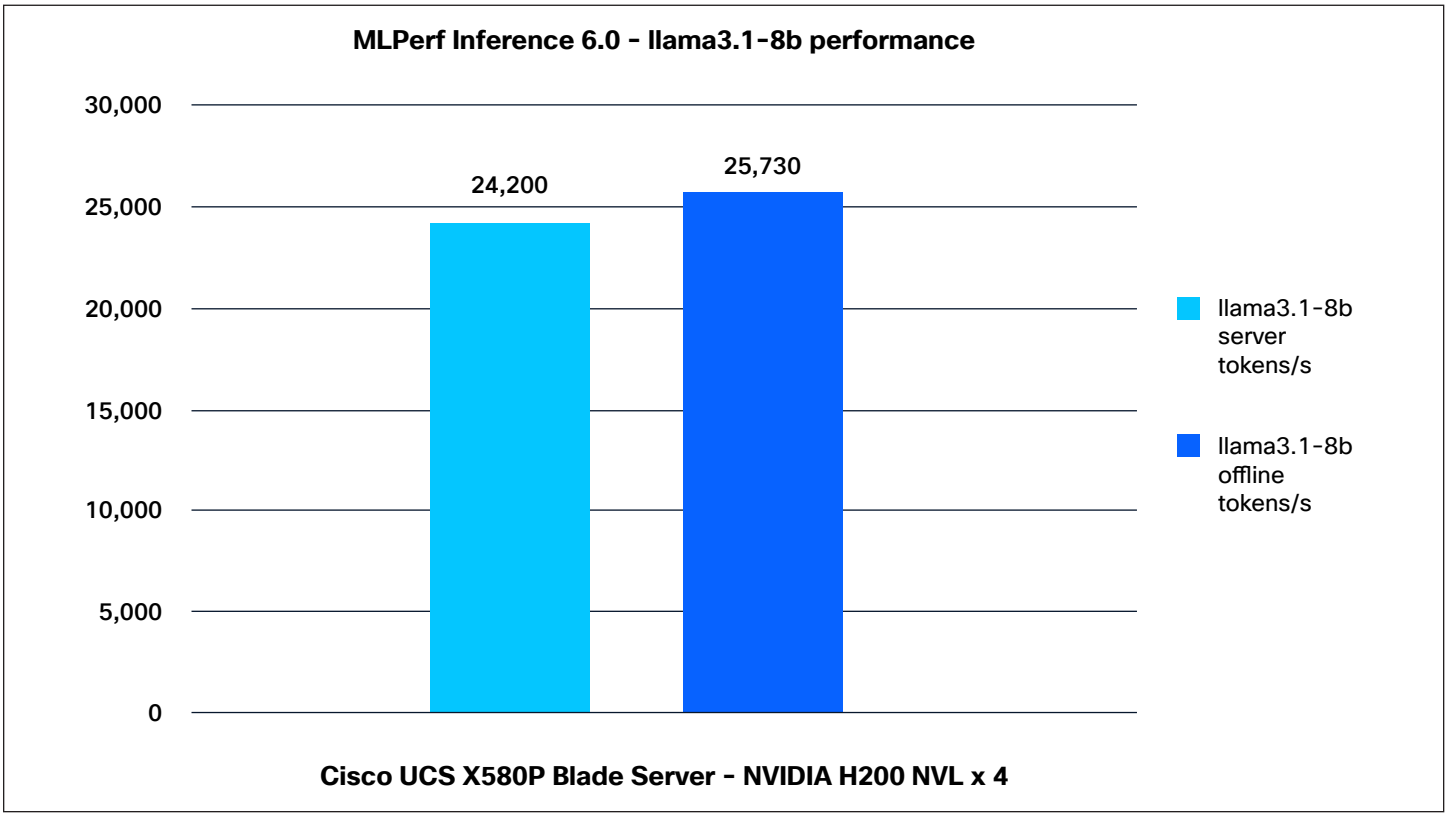

Figure 4. Llama3.1-8b performance data on a Cisco UCS X580p PCIe Node with NVIDIA H200 NVL GPUs

Performance data of the Whisper model:

Figure 5 shows the performance of the Whisper model tested on a Cisco UCS X580p PCIe Node with 4x NVIDIA H200 NVL GPUs.

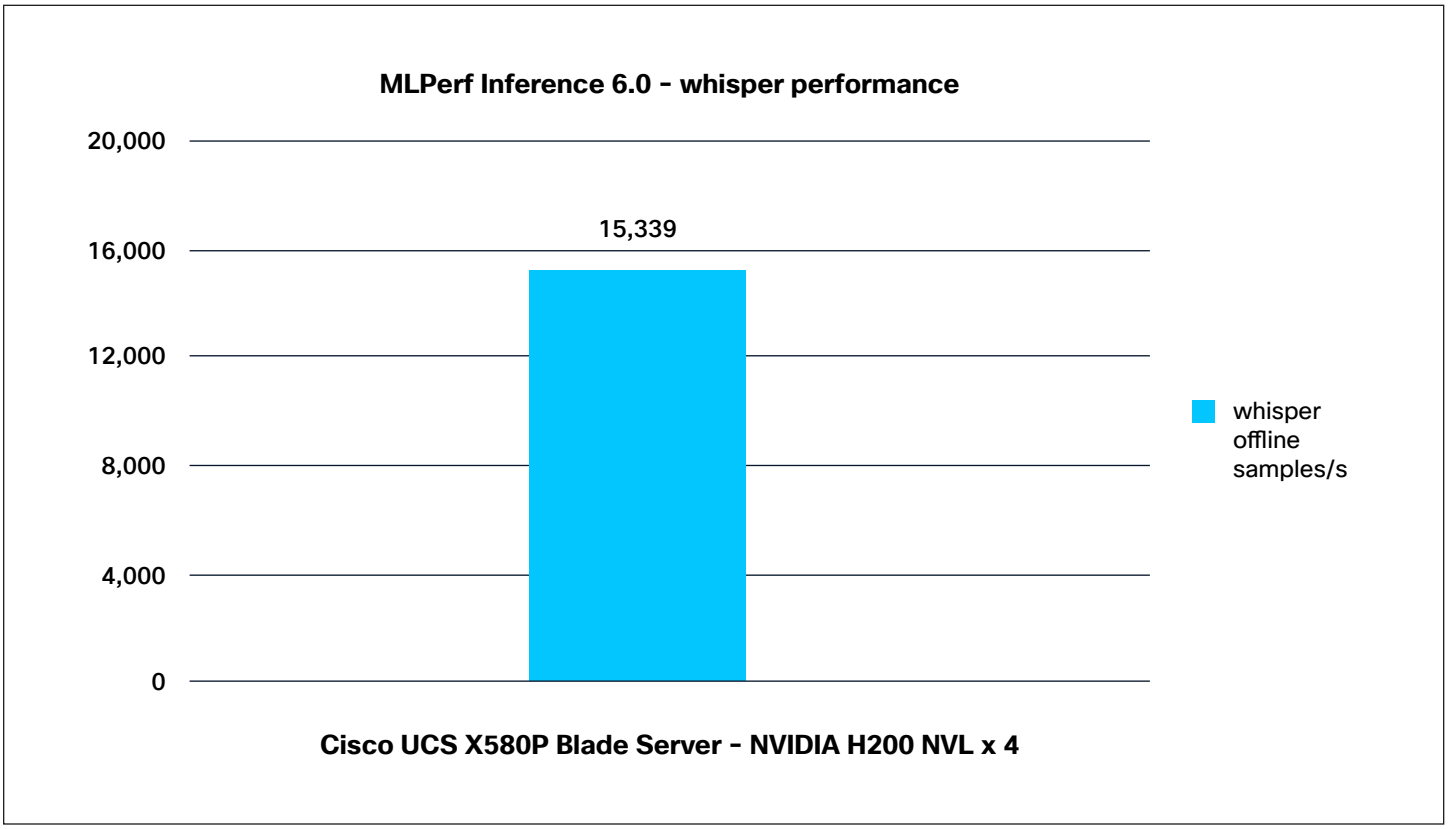

Figure 5. Whisper performance data on a Cisco UCS X580p PCIe Node with NVIDIA H200 NVL GPUs

Performance data of the Mixtral-8x7b model:

Figure 6 shows the performance of the Mixtral-8x7b model tested on a Cisco UCS X580P PCIe Node with 4x NVIDIA H200 NVL GPUs.

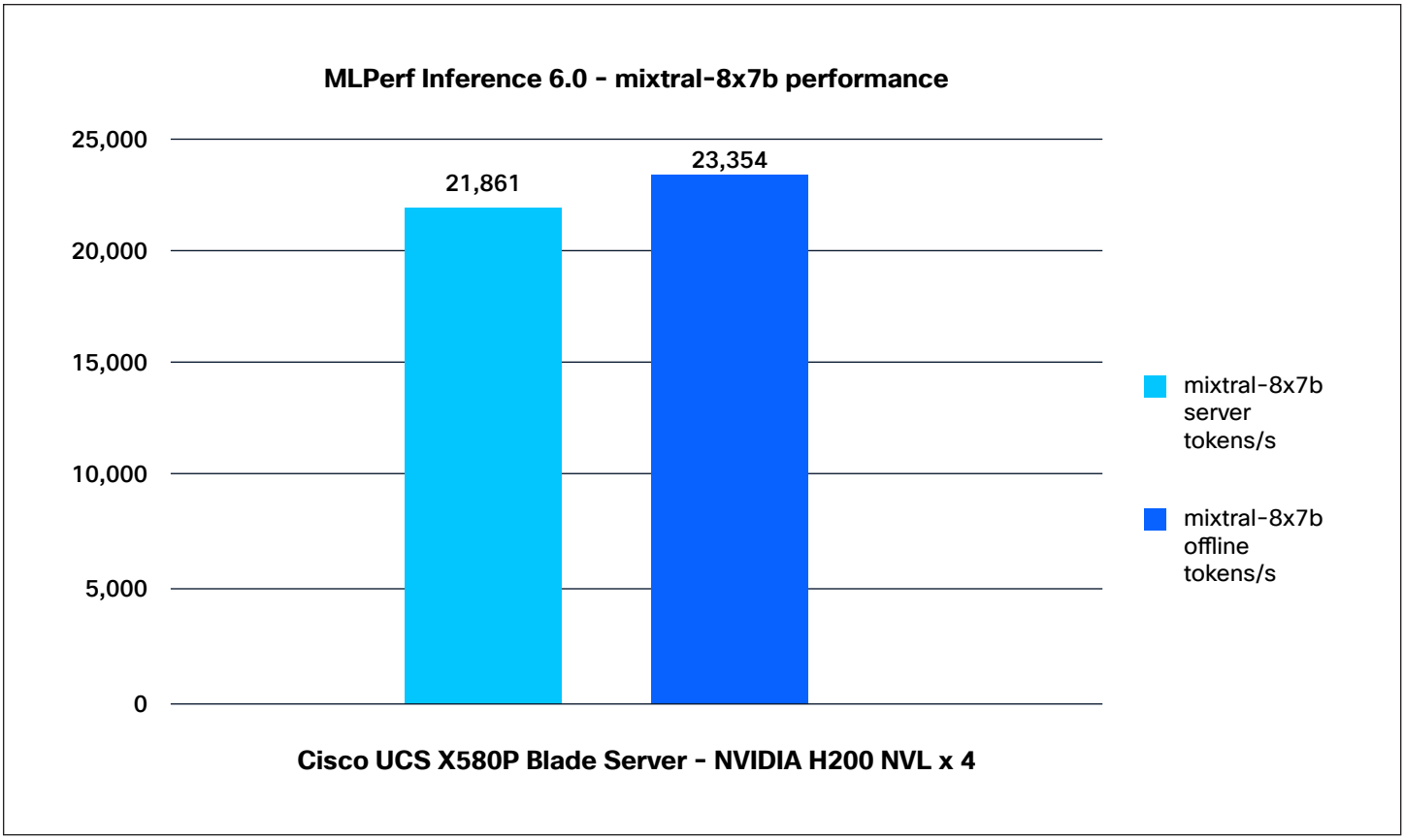

Figure 6. Mixtral-8x7b performance data on a Cisco UCS X580p PCIe Node with NVIDIA H200 NVL GPUs

Note: For Mixtral-8x7b performance, NVIDIA H200 NVL GPU is validated using MLPerf Inference benchmark v5.1 release, the result has not been verified by MLCommons.

Performance data for NVIDIA RTX Pro 6000 PCIe GPUs

The integration of the NVIDIA RTX Pro 6000 GPU within the Cisco UCS X580p PCIe Node provides a highly versatile, balanced solution for enterprise AI inference and generative AI fine-tuning. While the NVIDIA H200 NVL targets extreme-scale LLM workloads, the NVIDIA RTX Pro 6000 offers an optimized performance-per-watt profile, making it an ideal choice for a wide range of production-grade AI applications that require high throughput without the thermal footprint of flagship data-center accelerators.

Performance benchmarking and operational insights.

Validated through MLPerf Inference v5.1 (Datacenter) benchmarking, the Cisco UCS X580p PCIe Node—configured with 4x NVIDIA RTX Pro 6000 GPUs—demonstrates consistent and reliable performance across various datasets. Key technical observations include:

- **Optimized throughput:** The UCS X580p architecture provides the necessary PCIe Gen5 bandwidth to ensure that the RTX Pro 6000 GPUs are not bottlenecked, allowing for high-efficiency inference across batch-processing (offline) and real-time (server) scenarios.
- **Thermal and power stability:** The RTX Pro 6000 benefits from the UCS X580p's advanced thermal management. By leveraging the node's ability to support up to 600W per card, the RTX Pro 6000 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 6000 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 Llama2-70b model:

Figure 7 shows the performance of the Llama2-70b model, tested on a Cisco UCS X580P PCIe Node with 4x NVIDIA RTX Pro 6000 GPUs.

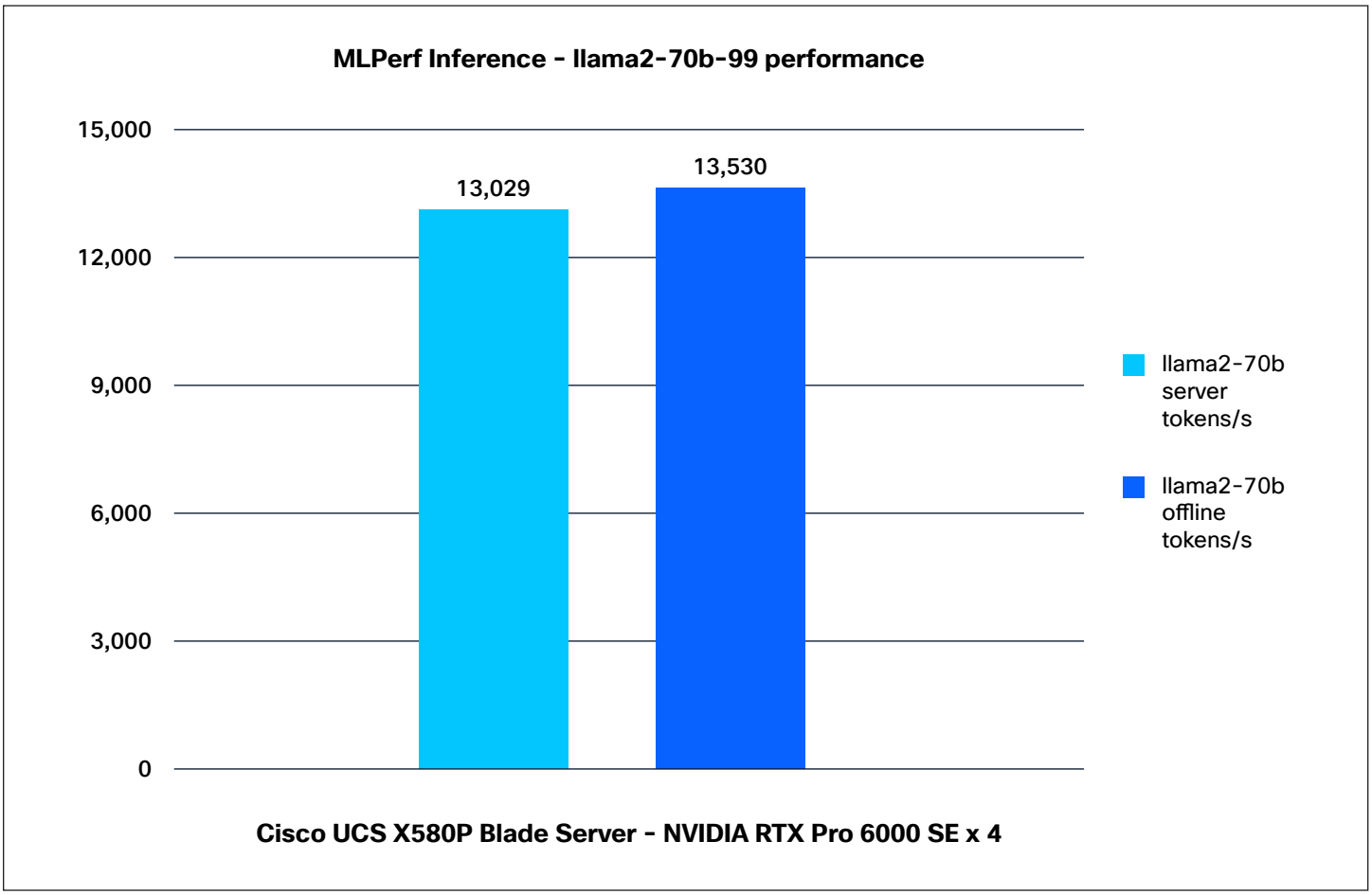

Figure 7. Llama2-70b performance data on a Cisco UCS X580p PCIe Node with NVIDIA RTX Pro 6000 GPUs

Performance data of the Llama3.1-8b model:

Figure 8 shows the performance of the Llama3.1-8b model, tested on a Cisco UCS X580P PCIe Node with 4x NVIDIA RTX Pro 6000 GPUs.

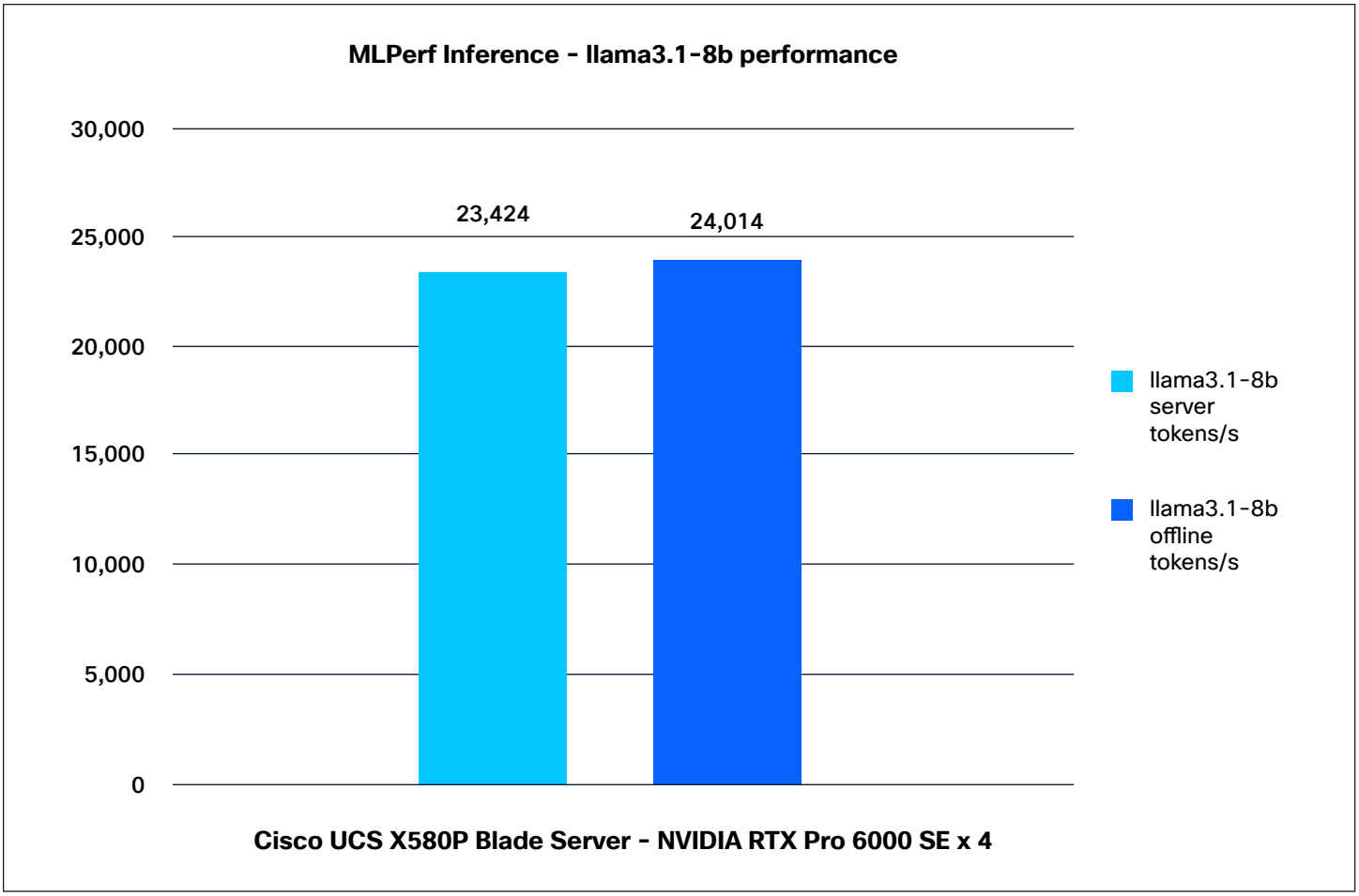

Figure 8. Llama3.1-8b performance data on a Cisco UCS X580p PCIe node with NVIDIA RTX Pro 6000 GPUs

Performance data of the Whisper model:

Figure 9 shows the performance of the Whisper model tested on a Cisco UCS X580P PCIe Node with 4x NVIDIA RTX Pro 6000 GPUs.

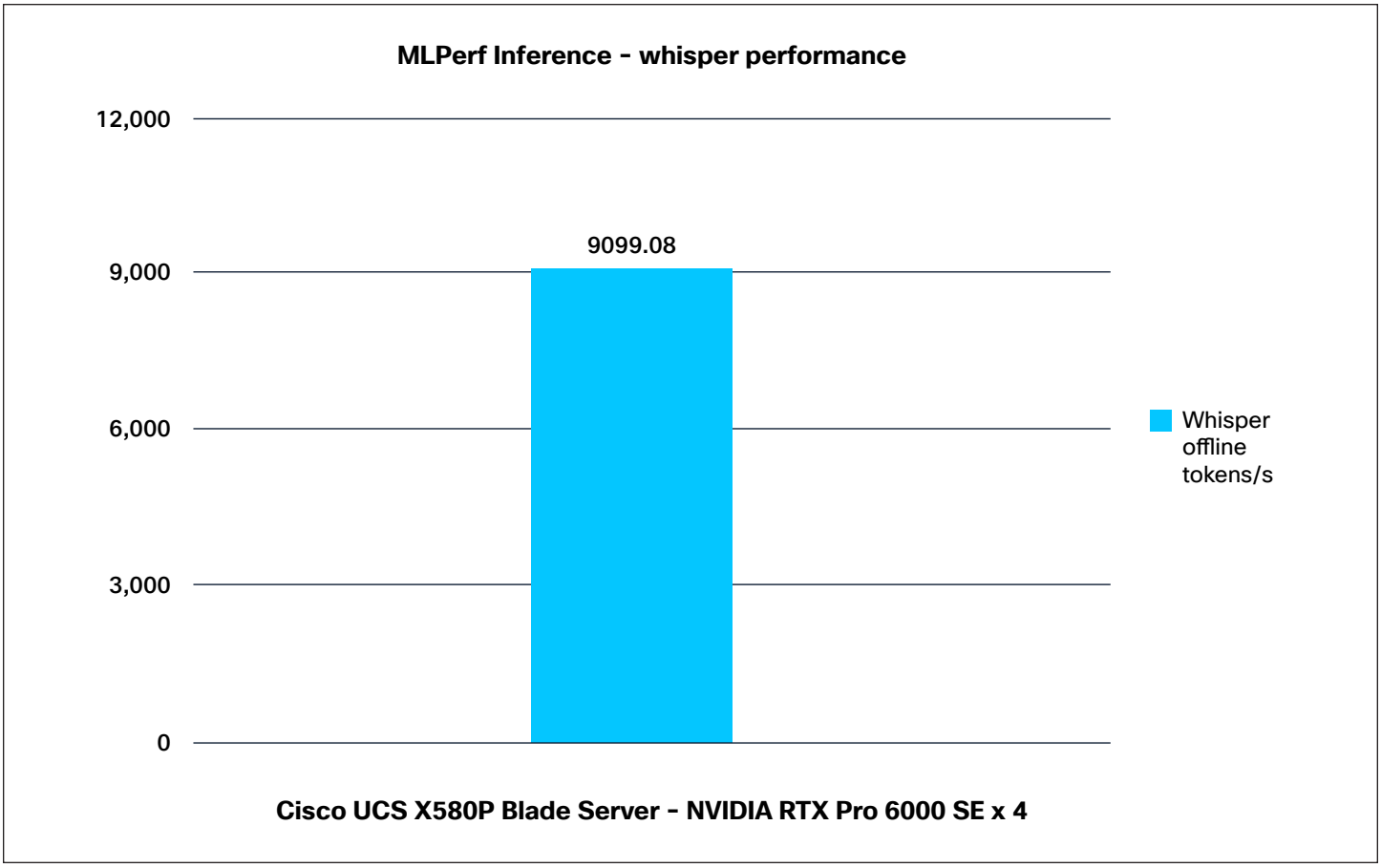

Figure 9. Whisper performance data on a Cisco UCS X580p PCIe node with NVIDIA RTX Pro 6000 GPUs

Performance summary

The Cisco UCS X580p PCIe Node, powered by the NVIDIA HGX architecture, provides a high-performance foundation engineered to accelerate the most demanding enterprise AI workloads. By combining modular density with the raw power of NVIDIA's latest GPU accelerators, the UCS X580p enables organizations to transition from AI experimentation to large-scale production with unprecedented speed and operational efficiency.

In partnership with NVIDIA, Cisco submitted comprehensive results for the MLPerf Inference: Datacenter benchmark suite (using versions v5.1 and v6.0).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 X580p PCIe Node 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 X580p PCIe Node in key AI inference categories:

- **Llama 3.1-8b leadership:** The Cisco UCS X580p PCIe Node, configured with 4x NVIDIA H200 NVL 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 X580p PCIe Node, configured with 4x NVIDIA RTX Pro 6000 GPUs, secured the first-position ranking for the Whisper speech recognition model. This demonstrates the node's versatility in managing complex, real-time signal processing and sequence-to-sequence inference tasks.

Appendix: Test environment

Table 2 details the properties of the Cisco UCS X580p PCIe Node managed by Cisco UCS X210c M8 and Cisco UCS X215c M8 compute nodes under test environment conditions.

Table 2. Server properties

Description	Compute node	Compute node
Product name	Cisco UCS X210c M8 Compute Node	Cisco UCS X215c M8 Compute Nodeex et unt eosam et ellor sitia demporesequi
CPU	2x Intel® Xeon® 6736P	2x AMD EPYC 9845 64-Core Processor
Number of cores	36	160
Number of threads	72	320
Total memory	1 TB	1.5 TB
Memory DIMMs	64 GB x 16 DIMMs	64 GB x 24 DIMMs
Memory speed	6400 MHz	6400 MHz
Network adapter	1x Cisco VIC 15420 MLOM 1x Mellanox CX7 NIC	1x Cisco VIC 15420 MLOM 1x Mellanox CX7 NIC
GPU controllers	4 x NVIDIA H200 NVL PCIe GPUs	4 x NVIDIA RTX Pro 6000 PCIe GPUs
SFF NVMe SSDs	6.4 TB 2.5in U.3 Micron 7450 NVMe High Perf High Endurance	6.4 TB 2.5in U.3 Micron 7450 NVMe High Perf High Endurance

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

Conclusion

The MLPerf Inference results validate that the Cisco UCS X580p PCIe Node is not merely an incremental upgrade, but a foundational element for robust, enterprise-grade AI. By extending the modularity of the Cisco UCS X-Series to support high-performance GPU configurations — including the NVIDIA H200 NVL and RTX Pro 6000 — the UCS X580p provides the flexibility and density required to adapt to the rapidly shifting landscape of generative AI and inference workloads.

This white paper has provided the technical evidence required for architects and decision-makers to confidently integrate the UCS X580p into their AI strategy. When coupled with the broader Cisco AI infrastructure ecosystem, the UCS X580p ensures that your organization can maintain peak performance, optimize resource utilization, and achieve the operational agility necessary to thrive in an increasingly AI-centric data center.

As you scale your AI initiatives, the Cisco UCS X580p PCIe Node stands as a testament to Cisco's commitment to delivering validated, high-performance solutions that turn complex AI potential into tangible, scalable business outcomes. With the UCS X580p, you are not just deploying hardware; you are building a future-ready foundation for innovation.

For more information

- For additional information on the Cisco UCS X580p PCIe Node and the Cisco UCS X9516 X-Fabric Module, refer to these data sheets:

  https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/ucs-x580p-pcie-node-ds.html.

  https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/ucs-9516-x-fabric-module-ds.html.
- Also refer to these spec sheets:
  https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/x580p-specsheet.pdf.

  https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/ucs-9516-x-fabric-module-ds.html.
- Cisco AI-Ready Data Center Infrastructure.
- Cisco AI PODs.
- For published MLPerf Inference results:
  https://mlcommons.org/benchmarks/inference-datacenter/.

C11-6039751-00 04/26