AI Performance: MLPerf Inference on Cisco UCS C880A M8 Rack Servers with NVIDIA B300 SXM GPUs

April 2026

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.

Based on the NVIDIA HGX platform, the Cisco UCS® C880A M8 Rack Server is a high-density, air-cooled rack server designed to power the most demanding Artificial Intelligence (AI) and High-Performance Computing (HPC) workloads. It integrates the NVIDIA HGX platform with eight NVIDIA HGX B300 (SXM) GPUs and is powered by two Intel® Xeon® 6th Gen processors, making it ideal for real-time Large Language Model (LLM) inference, next-level inference performance, and large-volume data processing.

The UCS C880A M8 supports customers across the entire AI stack, from large-scale model inference and fine-tuning to real-time inferencing and large-volume data processing. This server expands the Cisco UCS–dense AI server portfolio, offering a powerful solution for enterprises across various industries, including service providers, financial services, manufacturing, healthcare, life sciences, and automotive. With its advanced architecture, the UCS C880A M8 provides industry-leading performance, scalability, and enterprise manageability, making it ideal for compute-intensive AI use cases such as large-scale AI model inference, fine tuning, and inferencing.

To help demonstrate the AI performance capacity of the new Cisco UCS C880A M8 Rack Server, MLPerf Benchmarking performance testing for Inference 6.0 was conducted by Cisco using NVIDIA HGX B300 (SXM) GPUs as detailed later in this document.

Scope of this document

For the MLPerf Benchmarking performance testing for Inference 6.0: Datacenter focuses on evaluating performance using 8x NVIDIA B300 SXM GPUs configured on a Cisco UCS C880A M8 Rack Server. The inference benchmark results were collected for various datasets to help understand the performance benefits of the UCS C880A M8 server with NVIDIA B300 GPUs for inference workloads. This white paper highlights performance data for MLPerf Inference 6.0 on selected datasets to provide a quick understanding of the Cisco UCS C880A M8 Rack Server's performance in this context.

This aligns with Cisco's approach to showcasing how their UCS C880A M8 server, equipped with advanced NVIDIA B300 SXM GPUs and Intel Xeon 6th-Gen CPUs, delivers high throughput and efficiency for AI inference workloads, including large language model inference and other AI-native data center applications.

Key points include:

- Performance evaluation using 8x NVIDIA B300 SXM GPUs on the Cisco UCS C880A M8 Rack Server.
- Collection of inference benchmark results across various datasets.

The data serves to illustrate the performance benefits of this server and GPU configuration for inference workloads.

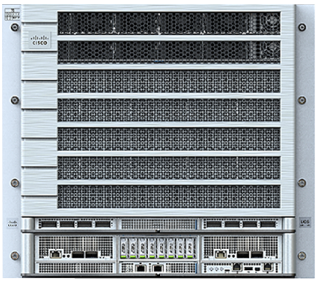
The white paper provides a concise overview of performance for selected datasets to aid quick understanding.

This summary reflects the scope as described in the relevant Cisco documentation and blog content about MLPerf Benchmarking and the UCS C880A M8 platform with NVIDIA B300 SXM GPUs.

Product overview

Based on the NVIDIA HGX platform, the Cisco UCS C880A M8 Rack Server is a high-density, air-cooled rack server designed to power the most demanding Artificial Intelligence (AI) and High-Performance Computing (HPC) workloads. It integrates the NVIDIA HGX platform with eight NVIDIA B300 SXM GPUs and is powered by two Intel Xeon 6th Gen Processors, making it ideal for real-time Large Language Model (LLM) inference, next-level inference performance, and large-volume data processing. The UCS C880A M8 supports customers across the entire AI stack, from large-scale model inference and fine-tuning to real-time inferencing and large-volume data processing. This server expands the Cisco UCS—dense AI server portfolio, offering a powerful solution for enterprises across various industries, including service providers, financial services, manufacturing, healthcare, life sciences, and automotive. With its advanced architecture, the C880A M8 ensures unparalleled performance, scalability, and enterprise manageability, making it ideal for compute-intensive AI use cases such as large-scale AI model inference, fine tuning, and inferencing.

UCS C880A Dense GPU Server Specifications



Product Specifications	
Form Factor	• HGX 10RU 19" Rack Server
Compute + Memory	• 2x 6th Gen Intel Xeon CPUs (Select SKUs for AI and HPC workloads) • Up to 32x DDR5 RDIMMS
Storage	• 2x M.2 SATA Boot Drives with HW RAID Controller (Boot) • Up to 8x PCIe Gen5 x4 E1.S NVMe SSDs (Data)
GPUs	• 8x NVIDIA HGX B300 NVL8 air-cooled GPUs
Network Cards	• **E-W:** Integrated ConnectX-8 • **N-S:** 4x PCIe Gen5 x16 FHHL slots, 1x OCP TSFF Gen5 x8
Cooling	• 20 Hot swappable FANs
Physical I/O	• 1 USB 3 type A, 1 mDP, 1 ID Button, 1 System Power Button, 1 Reset Button, 1 USB type C (for debugging), 1 RJ45 (OOB mgmt.) 1 RJ45 (LOM port)
Power Supply	• 12x 50V 3.2kW (N+N redundancy)

Figure 1. Cisco UCS C880A M8 Rack Server views, with product specifications

Refer to the data sheet for the Cisco UCS C880A M8 Rack Server.

Accelerated compute

A typical AI journey starts with inference GenAI models with large amounts of data to build the model intelligence. For this important stage, the new Cisco UCS C880A M8 Rack Server is a powerhouse designed to tackle the most demanding AI-Inference tasks. The UCS C880A 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.

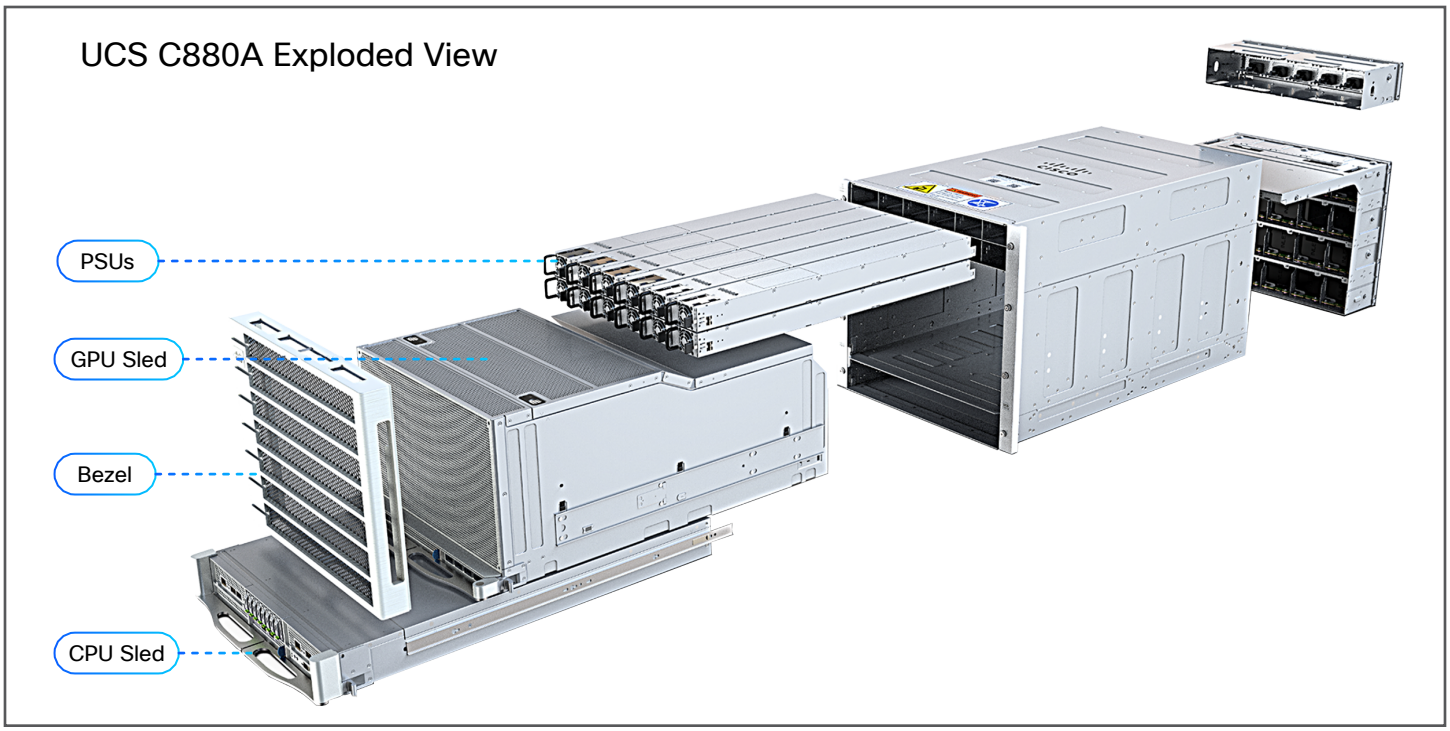


Figure 2. Exploded view of the Cisco UCS C880A Rack Server

Prominent features

Unleashing AI potential with NVIDIA HGX B300 SXM GPUs

The Cisco UCS C880A M8 Rack Server stands out by integrating the cutting-edge NVIDIA HGX platform with eight NVIDIA B300 (SXM) GPUs. This powerful GPU configuration is at the heart of its capability to deliver next-level performance for the most demanding AI workloads, including large-scale AI model inference, fine tuning, and real-time inferencing. The B300 GPUs provide immense parallel processing capabilities and high-speed GPU interconnects, which are critical for accelerating complex deep learning models and large language models. This integration ensures that enterprises can achieve higher token throughput and improve the economics of their AI operations, enabling profitable scaling of LLM and agentic workloads.

Comprehensive enterprise AI manageability

The Cisco UCS C880A M8 Rack Server is designed for enterprise readiness. In a future release, the UCS C880A M8 will enable management through Cisco Intersight®.

Cisco Intersight provides a cloud-based management platform that simplifies server lifecycle management, offering capabilities such as power operations, extensive monitoring metrics, server configuration management, and firmware bundle release management. This centralized control and observability streamlines AI infrastructure operations, reduces complexity, and ensures consistent policy enforcement across the data center.

Purpose-built for AI and HPC workloads

Beyond raw power, the Cisco UCS C880A M8 Rack Server is architected specifically to meet the unique demands of AI and HPC. Its design supports real-time large language model Inference, enabling rapid deployment and responsiveness for AI-driven applications. It also excels in next-level inference performance, significantly reducing the time required to train complex AI models. Furthermore, its capacity for large-volume data processing makes it an ideal platform for data-science and big-data analytics, including GPU-accelerated ETL processes. This specialized design ensures that organizations can build, optimize, and utilize AI models efficiently, accelerating business growth with scalable and high-performance solutions.

Scalable network fabric for AI connectivity

Network fabric: Cisco Nexus 9000 Series Switches and Nexus Dashboard

In distributed inference, training and fine-tuning, the network fabric plays a crucial role in providing high-bandwidth, low-latency communication to interconnect dense GPU servers such as the Cisco UCS C885A M8 Rack Server and the Cisco UCS C845A M8 Rack Server. The Cisco Nexus® 9000 Series is designed to meet these demanding requirements, serving as the high-performance foundation for both the leaf and spine layers of the backend and frontend fabrics in the architecture.

The Cisco AI PODs architecture leverages the following key platforms:

- **Cisco Nexus 9332D-GX2B:** a 1RU, 32-port 400GbE switch based on Cisco Cloud Scale technology, ideally suited for leaf role.
- **Cisco Nexus 9364D-GX2A:** a 2RU, 64-port 400GbE switch based on Cisco Cloud Scale technology, ideally suited for larger leaf or spine roles.
- **Cisco Nexus 9364E-SG2:** a 2RU, 64-port 800GbE (or 128 x 400GbE ports) switch based on Cisco® Silicon One® technology. Designed for next-generation fabrics, it is available in QSFP-DD and OSFP form factors with dual-port transceivers for 400GbE connectivity, making it suitable for both leaf and spine roles.

All of these Nexus switches provide the port density, switching capacity, and advanced features necessary for AI/ML workloads, including support for RDMA over Converged Ethernet (RoCE), hardware-accelerated telemetry, and advanced load-balancing mechanisms.

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

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.

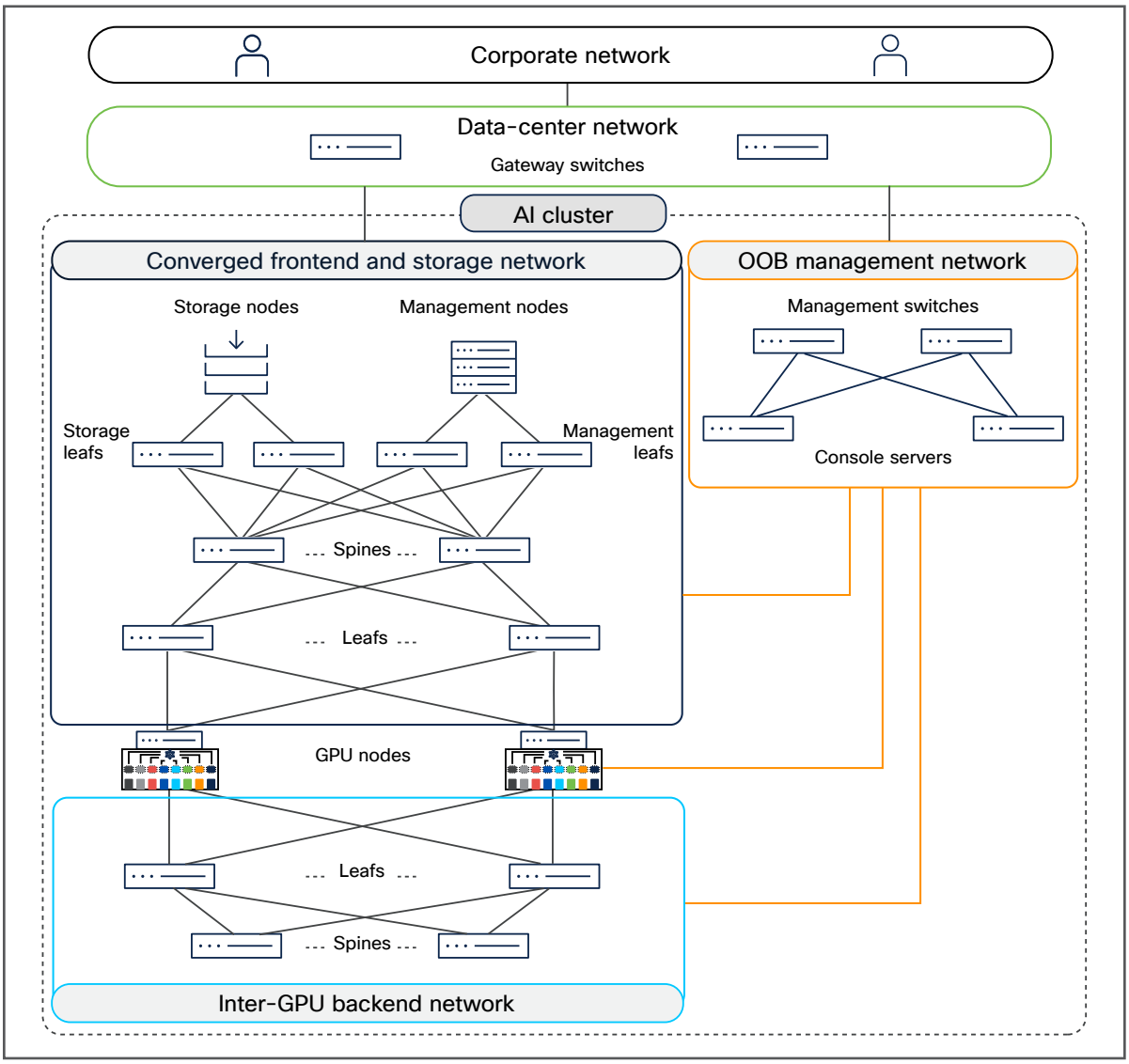


Figure 3. AI-cluster network design

Refer to the Cisco Nexus 9000 Series Switches for AI Clusters White Paper.

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.

MLPerf overview

MLPerf is a benchmark suite designed to evaluate the performance of machine-learning software, hardware, and services. It is developed by MLCommons, a consortium of AI leaders from academia, research labs, and industry. The primary goal of MLPerf is to provide an objective and standardized yardstick for assessing machine-learning platforms and frameworks.

MLPerf includes multiple benchmarks, notably:

- **MLPerf Training:** measures the time required to train machine-learning models to a specified accuracy level.
- **MLPerf Inference:** Datacenter: measures how quickly a trained neural network can perform inference tasks on new data.

MLPerf 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 summary of the current benchmarks and metrics: https://mlcommons.org/benchmarks/inference-datacenter/.

This 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 6.0 performance testing covered in this document, the Cisco UCS C880A M8 Rack Server was configured with:

- 8x NVIDIA B300 SXM GPUs

MLPerf Inference performance results

MLPerf Inference benchmarks

The MLPerf Inference models given in Table 1 were configured on a Cisco UCS C880A M8 Rack Server and tested for performance.

Table 1. MLPerf Inference 6.0 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.
Llama3.1-405B	language/llama3-405b	Designed for advanced AI tasks such as reasoning, coding, research assistance, and building AI agents.
deepseek-r1	language/deepseek-r1	Designed to excel at step-by-step reasoning, mathematics, coding, and complex problem solving, such as specialized reasoning models.
Wan2.2-T2V-A14B-Diffusers	text_to_video	Converts natural language prompts into short videos using a diffusion-based generative architecture.
GPT-OSS	language/gpt-oss-120b	Open-source or open-weight implementations inspired by the GPT. Used to make powerful language models accessible to developers and researchers without requiring closed commercial APIs.

MLPerf Inference 6.0 performance data

As part of the MLPerf Inference 6.0 submission, Cisco has tested most of the datasets listed in Table 1 on the Cisco UCS C880A M8 Rack Server and submitted the results to MLCommons. The results are published on MLCommons results page: https://mlcommons.org/benchmarks/inference-datacenter/.

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

Performance data for NVIDIA B300 SXM GPU

Llama2-70b

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

Figure 4 shows the performance of the Llama2-70b model, with an accuracy of 99, tested on a Cisco UCS C880A M8 Rack Server with 8x NVIDIA B300 SXM GPUs.

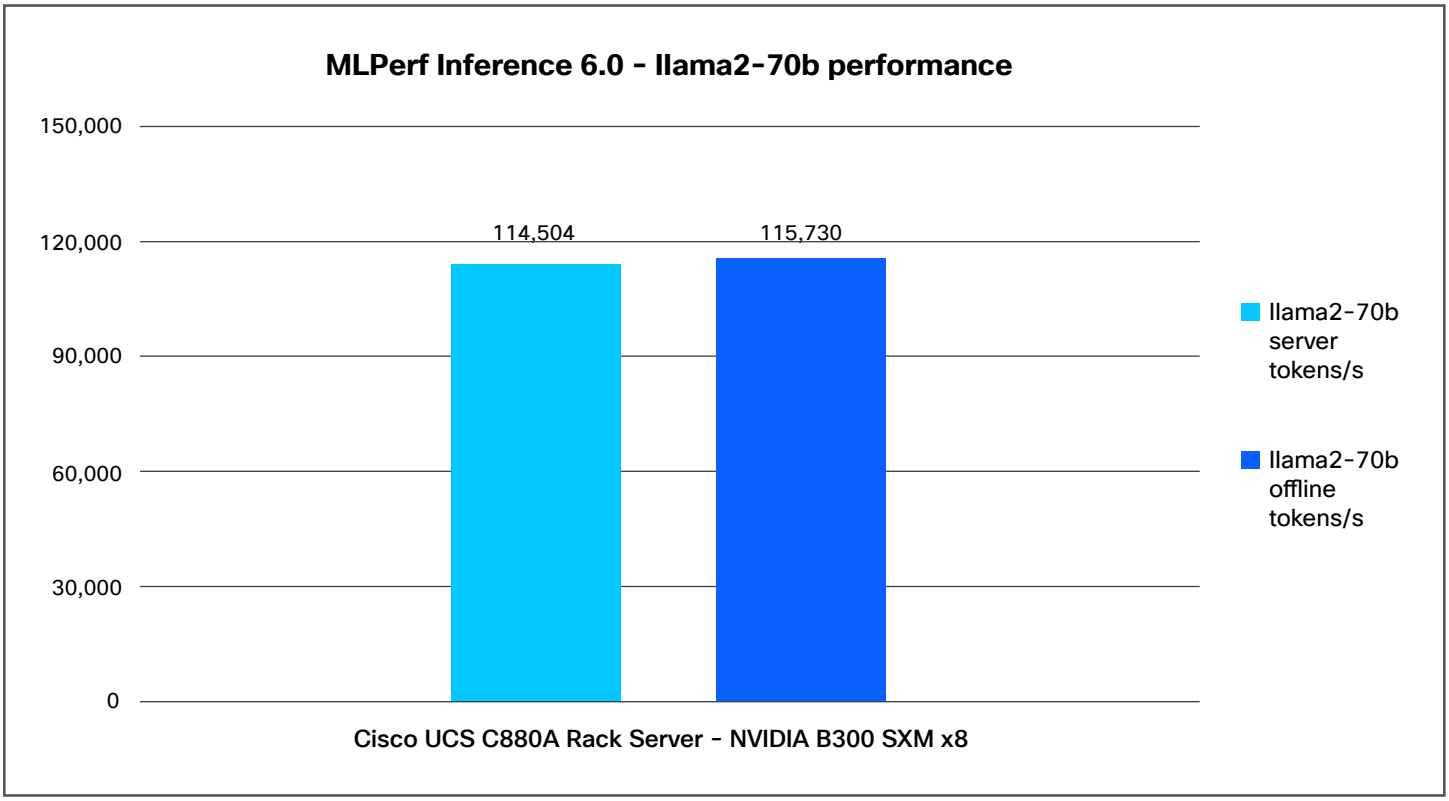


Figure 4. Llama2-70b performance data on a Cisco UCS C880A M8 Rack Server with NVIDIA B300 SXM GPUs

Llama3.1-405B

Llama 3.1 405B is the largest and most powerful open-weight large language model released by Meta as part of the Llama 3.1 family. It is designed for advanced AI tasks such as reasoning, coding, research assistance, and building AI agents.

Figure 5 shows the performance of the Llama3.1-405b model tested on a Cisco UCS C880A M8 Rack Server with 8x NVIDIA B300 SXM GPUs.

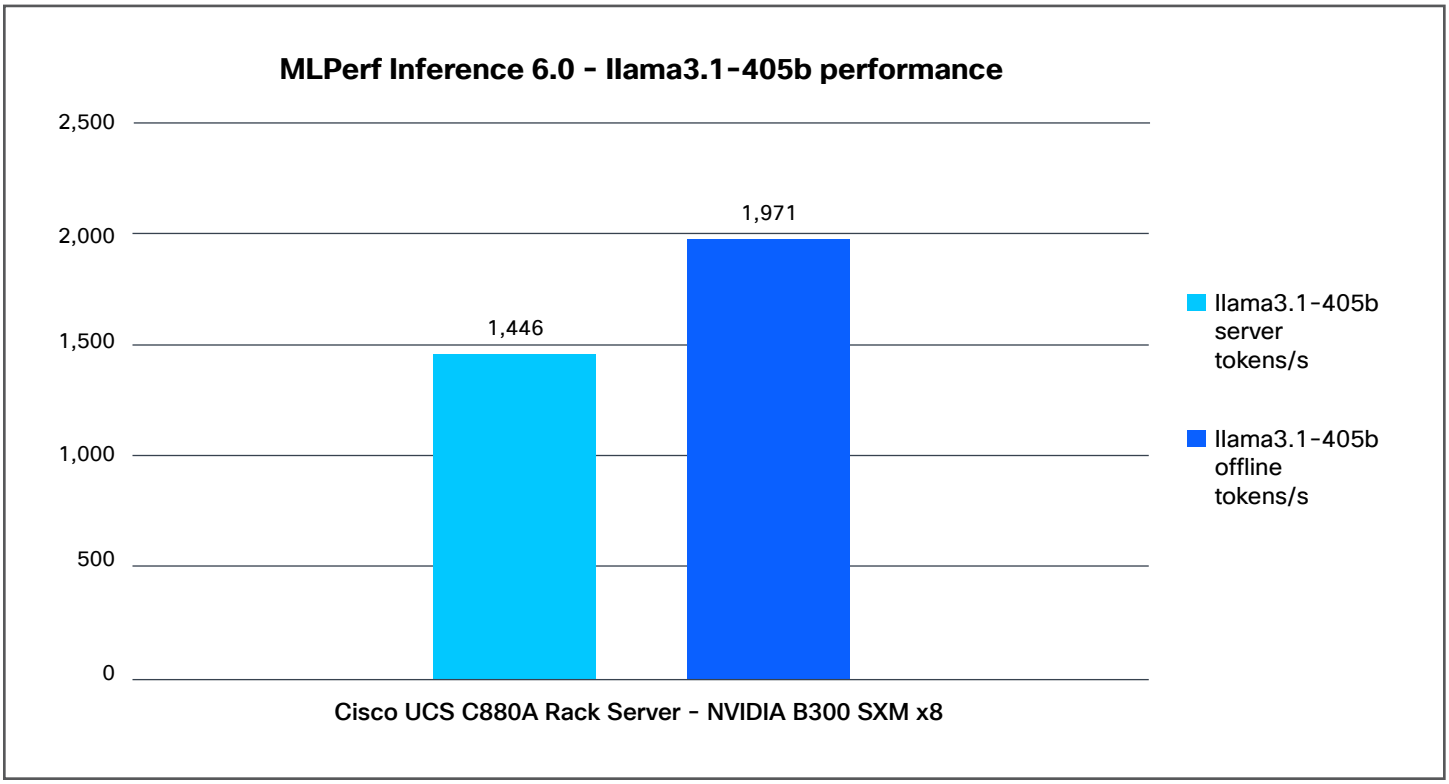


Figure 5. Llama3.1-405b performance data on a Cisco UCS C880A M8 Rack Server with NVIDIA B300 SXM GPUs

Deepseek-r1

Deepseek-r1 is an advanced reasoning-focused large language model designed to excel at step-by-step reasoning, mathematics, coding, and complex problem solving.

Figure 6 shows the performance of the Deepseek-r1 model, tested on a Cisco UCS C880A M8 Rack Server with 8x NVIDIA B300 SXM GPUs.

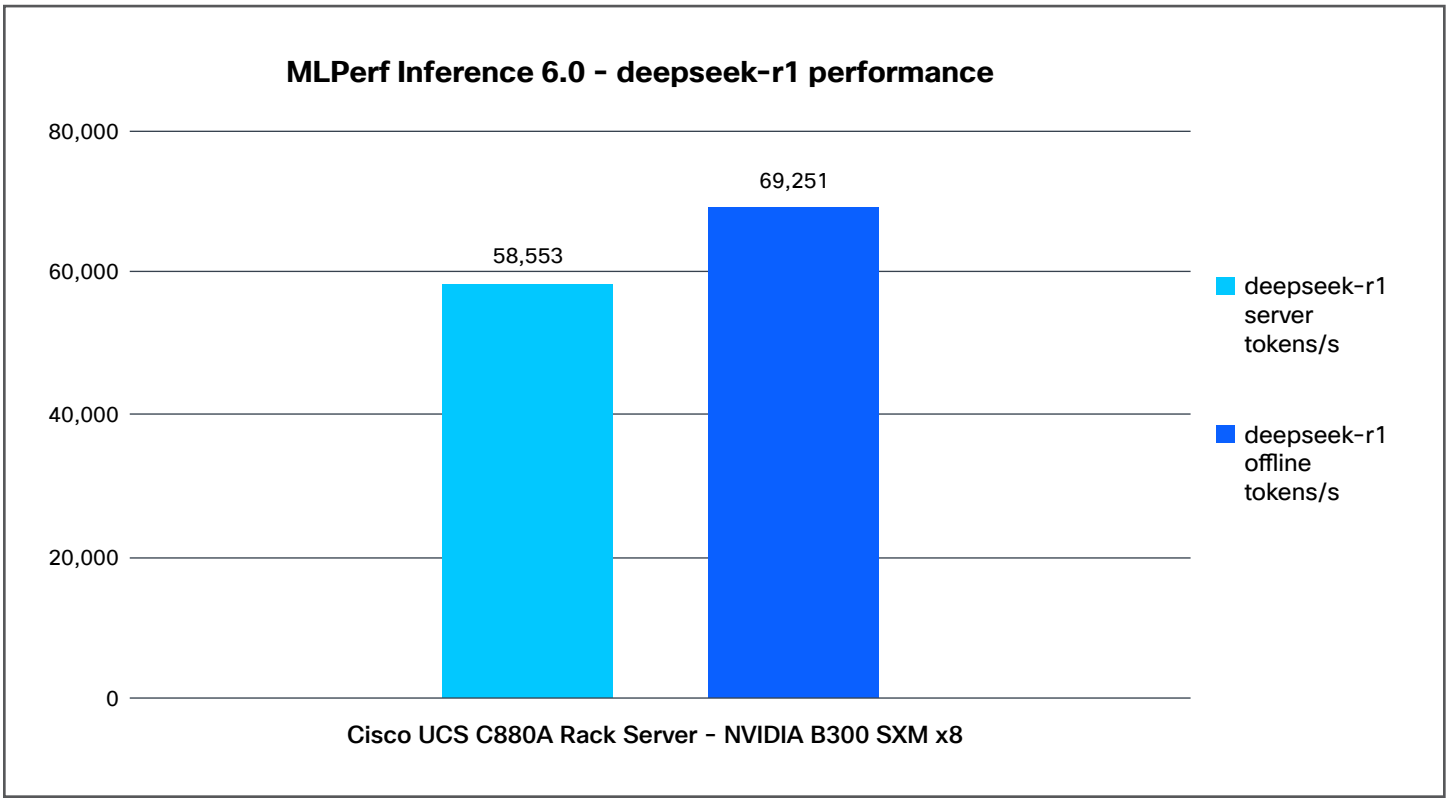


Figure 6. Deepseek-r1 performance data on a Cisco UCS C880A M8 Rack Server with NVIDIA B300 SXM GPUs

wan-2.2-t2v-a14b

Wan2.2-t2v-a14b is a large-scale text-to-video generative AI model which converts natural language prompts into short videos using a diffusion-based generative architecture. The model is part of the wan 2.2 series, designed for high-quality AI video generation.

Testing for wan2.2-t2v-a14b is done with two scenarios, Offline and SingleStream. Graphs for both the results are given below.

Figure 7 shows the latency of the wan-2.2-t2v-a14b model tested for a SingleStream scenario on a Cisco UCS C880A M8 Rack Server with 8x NVIDIA B300 SXM GPUs.

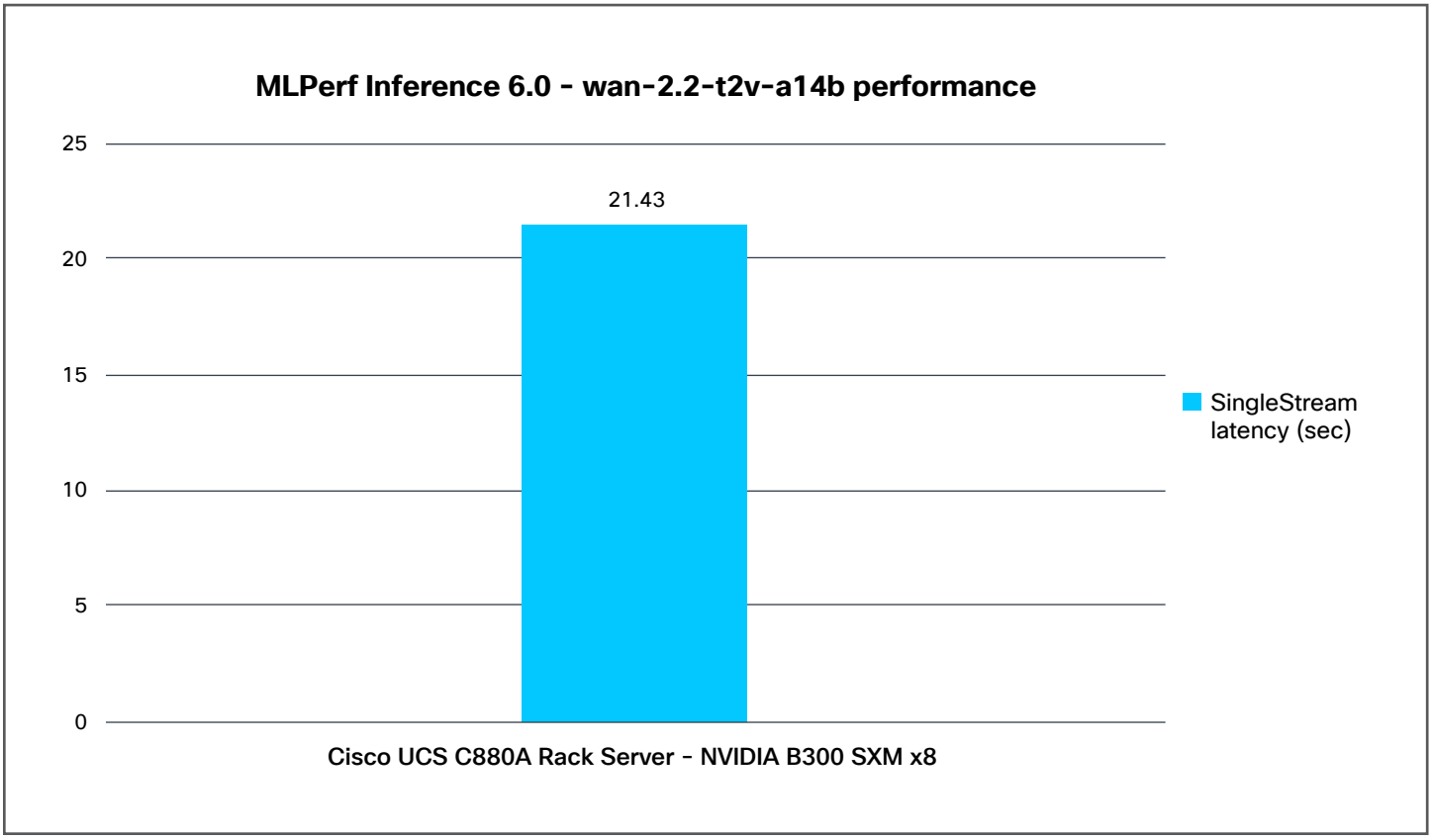


Figure 7. Wan-2.2-t2v-a14b SingleStream latency on a Cisco UCS C880A M8 Rack Server with NVIDIA B300 SXM GPUs

Figure 8 shows the performance of the wan-2.2-t2v-a14b model tested for an Offline scenario on a Cisco UCS C880A M8 Rack Server with 8x NVIDIA B300 SXM GPUs.

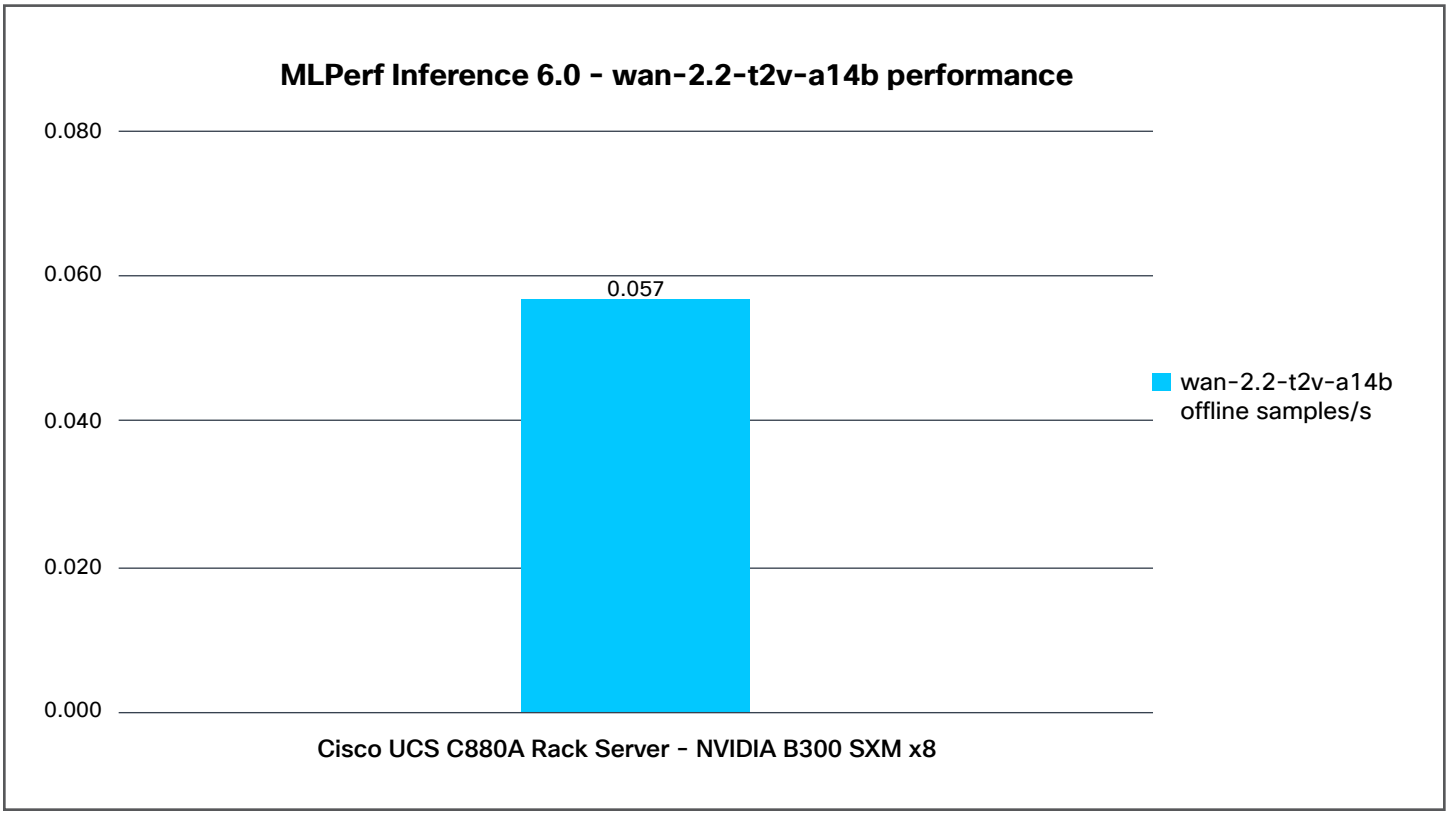


Figure 8. Wan-2.2-t2v-a14b Offline performance data on a Cisco UCS C880A M8 Rack Server with NVIDIA B300 SXM GPUs

gpt-oss-120b

Gpt-oss generally refers to open-source or open-weight implementations inspired by the GPT (Generative Pre-trained Transformer) architecture. The term is often used by the AI community to describe projects that aim to provide GPT-like language models with openly available code and/or weights.

Figure 9 shows the performance of the gpt-oss-120b model tested on a Cisco UCS C880A M8 Rack Server with 8x NVIDIA B300 SXM GPUs.

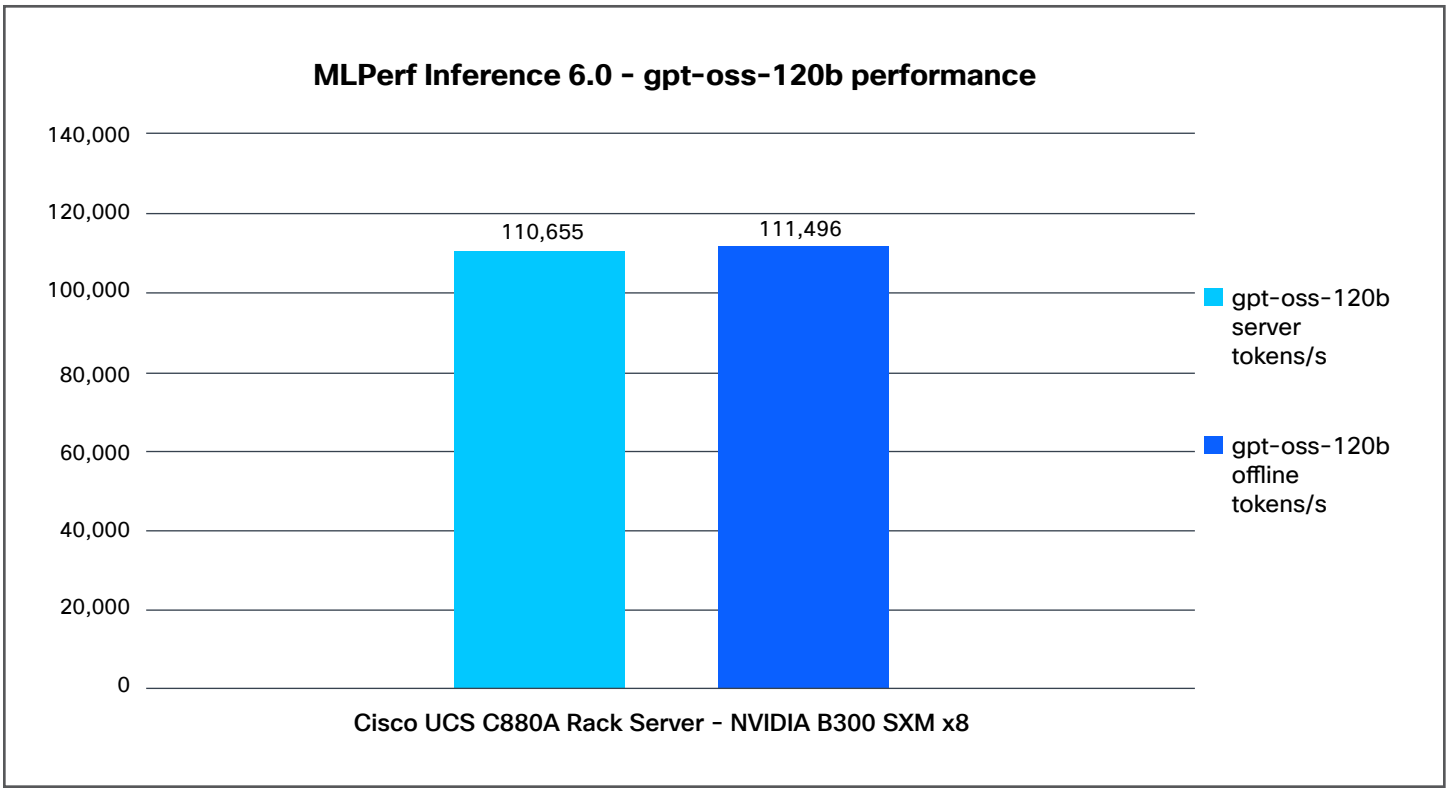


Figure 9. Gpt-oss-120b performance data on a Cisco UCS C880A M8 Rack Server with NVIDIA B300 SXM GPUs

Performance summary

Built on the NVIDIA HGX platform, the Cisco UCS C880A M8 Rack Server provides the accelerated computing power required to tackle the most demanding AI workloads. Its robust performance and streamlined deployment enable faster outcomes for your AI projects. Cisco, in collaboration with NVIDIA, successfully submitted MLPerf Inference 6.0 results that enhance performance and efficiency across various inference workloads, including summarization LLM (language), Text generation LLM (language), Reasoning LLMs (language), generative image (text to video).

With MLPerf Inference 6.0, Cisco continues to set the pace in AI infrastructure performance with the Cisco UCS C880A M8 Rack Server, powered by 8x NVIDIA B300 SXM GPUs, delivering industry-leading results across multiple leading benchmarks.

- The Cisco UCS C880A M8 Rack Server achieved #1 performance in the Llama2-70B benchmark for both Offline and Server scenarios and also secured #1 in the Llama3.1-405B Offline scenario. These results highlight the platform's exceptional ability to handle demanding large language model workloads with speed and efficiency.
- Cisco also demonstrated strong leadership in Deepseek-r1, where the UCS C880A M8 ranked #1 in the Server scenario and #2 in the Offline scenario, reinforcing its strength across diverse AI inference environments.
- For Gpt-oss-120b, the Cisco UCS C880A M8 Rack Server once again captured #1 in both Offline and Server scenarios. In addition, for the Wan-2.2-t2v-a14b test, Cisco delivered the best SingleStream latency compared to other vendors, underscoring its advantage in responsiveness and real-time AI performance.

The test results demonstrate substantial performance improvements across all primary metrics.

Table 2. Performance Summary table for 8 x NVIDIA B300 GPUs

Metric	Llama2-70b (Offline)	Llama3.1-405b (Offline)	Deepseek-r1 (Offline)	GPT-OSS-120b (Offline)
Tokens/sec	115730	1971.11	69251.4	111496
Samples/sec	423.839	3.09594	18.4735	85.2161
Concurrent Users	3072	2048	512	896
Perf/Watt	10.33	0.17	6.089	9.93

Note: From Table 2, the above metric, **Performance per Watt**, is a critical benchmark for evaluating the energy efficiency of AI inference and Large Language Model (LLM) deployments.

In the context of modern data centers and Cisco's infrastructure solutions (such as Cisco UCS servers), this calculation helps organizations balance computational power with sustainability goals.

Breakdown of the formula:

1. **Performance (Tokens/sec):** This represents the throughput of the model—how quickly the system can generate text. High throughput is essential for real-time applications and user experience.
2. **Total Server Consumption (Watts):** This measures the actual power drawn by the hardware (CPU, GPU, memory, and cooling fans) while processing the workload.
3. **The Result:** A higher value indicates a more efficient system that can process more data using less electricity.

Together, these benchmark results showcase the Cisco UCS C880A M8 Rack Server as a powerful, high-performance platform designed to lead the next generation of AI workloads.

Conclusion

The MLPerf Inference 6.0 results validate that the Cisco UCS C880A M8 Rack Server is not merely an incremental upgrade, but a foundational element for robust, enterprise-grade AI. By combining the raw power of the NVIDIA HGX architecture with Cisco's proven reliability, the UCS C880A M8 provides the high-density performance required to tackle the most demanding AI and HPC workloads.

This white paper has provided the technical evidence required for architects and decision-makers to confidently integrate the Cisco UCS C880A Rack Server into their AI strategy. When paired with the scale-out reliability of Cisco Silicon One networking and the modular management of the Cisco UCS ecosystem, the UCS C880A M8 ensures peak performance, operational agility, and a simplified path to deployment in an increasingly AI-centric data center.

As enterprises continue to scale their AI initiatives, the ability to rely on a validated, full-stack infrastructure becomes a critical competitive advantage. With the UCS C880A M8, Cisco remains committed to delivering the high-performance building blocks necessary to turn complex AI potential into measurable business outcomes.

Appendix: Test environment

Table 3 details the properties of the Cisco UCS C880A Rack Server under test environment conditions.

Table 3. Server properties

Description	Value
Product name	Cisco UCS C880A M8 Rack Server
CPU	2x Intel Xeon 6th Gen 6776P Processor
Number of cores	64
Number of threads	128
Total memory	4 TB
Memory DIMMs (16)	32x 128GB DDR5 RDIMM
Memory speed	6400 MHz
Network adapter	• 8x GPU-board integrated NVIDIA ConnectX-8 Infiniband SuperNICS mart host channel adapter • 2x NVIDIA ConnectX-7 smart host channel adapter (2x200G) • 1x Intel X710-T2L OCP
GPU controllers	• NVIDIA B300 SXM 8-GPU
SFF NVMe SSDs	• Up to 8x PCIe Gen5 x4 E1.S NVMe SSD

Note: Platform-default BIOS settings were applied during the MLPerf Inference validation.

For more information

For additional information on the server, refer to: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c880a-m8-rack-server-spec-sheet.pdf.

Data sheet: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c880a-m8-rack-server-ds.html.

Cisco AI-Ready Data Center Infrastructure: https://blogs.cisco.com/datacenter.

Cisco AI PODs: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ai-pods-aag.html.

Cisco AI-Native Infrastructure for Data Center: https://www.cisco.com/site/us/en/solutions/artificial-intelligence/infrastructure/index.html.