AI Performance: MLPerf Inference on Cisco UCS X210c M8 Compute Node with Intel® Xeon® 6 Processor

October 2025

Contents

Introduction

Artificial Intelligence (AI) is transforming industries across the globe by enabling machines to learn, infer, and make decisions based on data. This document explores how organizations can implement AI with the help of Cisco® solutions focused on inferencing.

This document covers the product features, MLPerf Inference benchmarks, test configurations, and results to explain the artificial intelligence use cases best suited for enterprises looking to invest in a mainstream compute node.

We present the MLPerf™ v5.1 Data Center Inference results obtained on a Cisco UCS X210c M8 Compute Node powered by Intel Xeon 6 processor family.

Here we explore how the Cisco UCS X210c M8 Compute Node with Intel® Xeon® 6 Processors is a high-performance solution for data centers supporting deep learning and complex workloads, including databases and advanced analytics. Its capabilities support natural language processing, image classification, object detection, and many other AI-enhanced functions.

Executive summary

Generative AI is revolutionizing industries, enabling text-to-image generation, realistic voice synthesis, and even the creation of novel scientific materials. However, unleashing the full potential of these powerful models requires a robust and optimized infrastructure. Generative AI models typically require massive amounts of data and complex algorithms, leading to significant computational demands during inference. Challenges include:

High computational workloads: Inference often involves processing large amounts of data through complex neural networks, requiring high-performance computing resources.

Memory bandwidth demands: Large models often require substantial memory bandwidth to handle data transfer efficiently.

Latency requirements: Many applications require low-latency inference to ensure real-time responsiveness.

Technology overview

Cisco UCS X210c M8 Compute Node

The Cisco UCS X-Series Modular System simplifies your data center, adapting to the unpredictable needs of modern applications while also providing for traditional scale-out and enterprise workloads. It reduces the number of server types to maintain, helping to improve operational efficiency and agility as it also helps reduce complexity. Powered by the Cisco Intersight® cloud-operations platform, it shifts your thinking from administrative details to business outcomes—with hybrid-cloud infrastructure that is assembled from the cloud, shaped to your workloads, and continuously optimized.

The Cisco UCS X210c M8 Compute Node is the third generation of compute node to integrate into the Cisco UCS X-Series Modular System. It delivers performance, flexibility, and optimization for deployments in data centers, in the cloud, and at remote sites. This enterprise-class server offers market-leading performance, versatility, and density without compromise for workloads. Up to eight compute nodes can reside in the 7-Rack-Unit (7RU) Cisco UCS X9508 Chassis, offering one of the highest densities of compute, I/O, and storage per rack unit in the industry.

Figure 1. Cisco UCS X210c M8 Compute Node

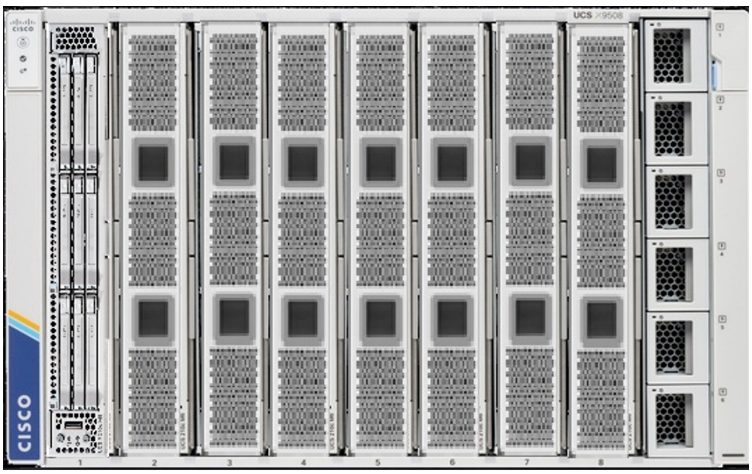

Figure 2. Cisco UCS X9508 Chassis

The Cisco UCS X210c M8 Compute node supports Intel Xeon 6 Processors so that you have the option to run inferencing in the data center or at the edge.

Intel Xeon 6 Processors are engineered to seamlessly handle demanding AI workloads, including inference and fine tuning on models containing up to 20 billion parameters, without an immediate need for additional hardware.

Intel Xeon processors are equipped with:

- Intel Advanced Matrix Extensions (Intel AMX) accelerator, an AI accelerator that is built into each core to significantly speed up deep-learning applications when 8-bit integer (INT8) or 16-bit float (bfloat16) datatypes are used
- Higher core frequency, larger last-level cache, and faster memory with DDR5, to speed up compute processing and memory access
- Improved cost-effectiveness that is provided by combining the latest-generation AI hardware with software optimizations, potentially lowering TCO by enabling the use of built-in accelerators to scale-out inferencing performance rather than relying on discrete accelerators, making generative AI more accessible and affordable
- DeepSpeed, which provides high-performance inference support for large transformer-based models with billions of parameters, through enablement of multi-CPU inferencing. It automatically partitions models across the specified number of CPUs and inserts necessary communications to run multi-CPU inferencing for the model

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.

- **Scalability and flexibility:** Cisco UCS is designed to scale the evolving needs of AI/ML workloads, ensuring that resources can be adjusted as required.
- **Performance optimization:** Cisco UCS offers high-performance computing power, making it well-suited for data-intensive AI/ML tasks. The use of CPUs such as Intel 6787P enhances performance for deep learning and other AI-related processes.
- **Diverse workload support:** Cisco UCS supports a wide range of AI/ML workloads, including deep learning, MLPerf inferencing, and other specialized applications.
- **Security and compliance:** Cisco's security solutions can be integrated with UCS to secure AI workloads and data, ensuring compliance with industry regulations.

Intel Xeon 6 AI capabilities

Intel Xeon 6 processor family offerings are differentiated with hyperthreaded cores featuring built-in matrix engines that accelerate compute-intensive AI, HPC, and data services workloads. All Intel Xeon 6 Processors, regardless of P-core or E-core focus, feature the same instruction sets, BIOS, and built-in I/O accelerators, including Intel QuickAssist Technology (Intel QAT), Intel Data Streaming Accelerator (Intel DSA), Intel In-Memory Analytics Accelerator (Intel IAA), and Intel Dynamic Load Balancer (Intel DLB).

Intel Xeon 6 Processors are designed to support many demanding AI use cases and expand on four generations of Intel's leadership in built-in AI with acceleration such as Intel Advanced Matrix Extensions (Intel AMX), which now supports int8, BF16, and FP16 (new) data types.

As a result, Intel Xeon 6 Processors help to meet Service Level Agreements (SLAs) for several AI models, ranging from object detection to midsize GenAI, while offering open standards; high performance; Reliability, Availability, and Serviceability (RAS) features; and support for additional accelerators as needed.

The Intel AMX matrix multiplication engine in each core boosts overall inferencing performance. With a focus on ease of use, Cisco technologies deliver exceptional CPU performance results with optimized BIOS settings that fully unleash the power of Intel's oneAPI Deep Neural Network Library (OneDNN) software that is fully integrated with both PyTorch and TensorFlow frameworks. The server configurations and the CPU specifications in the benchmark experiments are shown in Tables 1, 2, and 3, respectively.

Intel Xeon 6 Processors (as shown in Figure 1) excel at the complete spectrum of workloads, with a mainstream series that features a range of 8 to 86 cores in the mainstream offering, up to 176 PCIe 5.0 lanes for networking and storage add-in cards in dual CPU-based systems, and a single-socket offering with 136 PCIe lanes for single CPU-based systems.

The efficiency of all Intel Xeon 6 Processors is highlighted by their ability to provide scalable performance per watt as server utilization increases, delivering nearly linear power-performance consumption across the load line. For performance-demanding workloads, this means the platform efficiently uses power at high loads to help finish jobs fast.

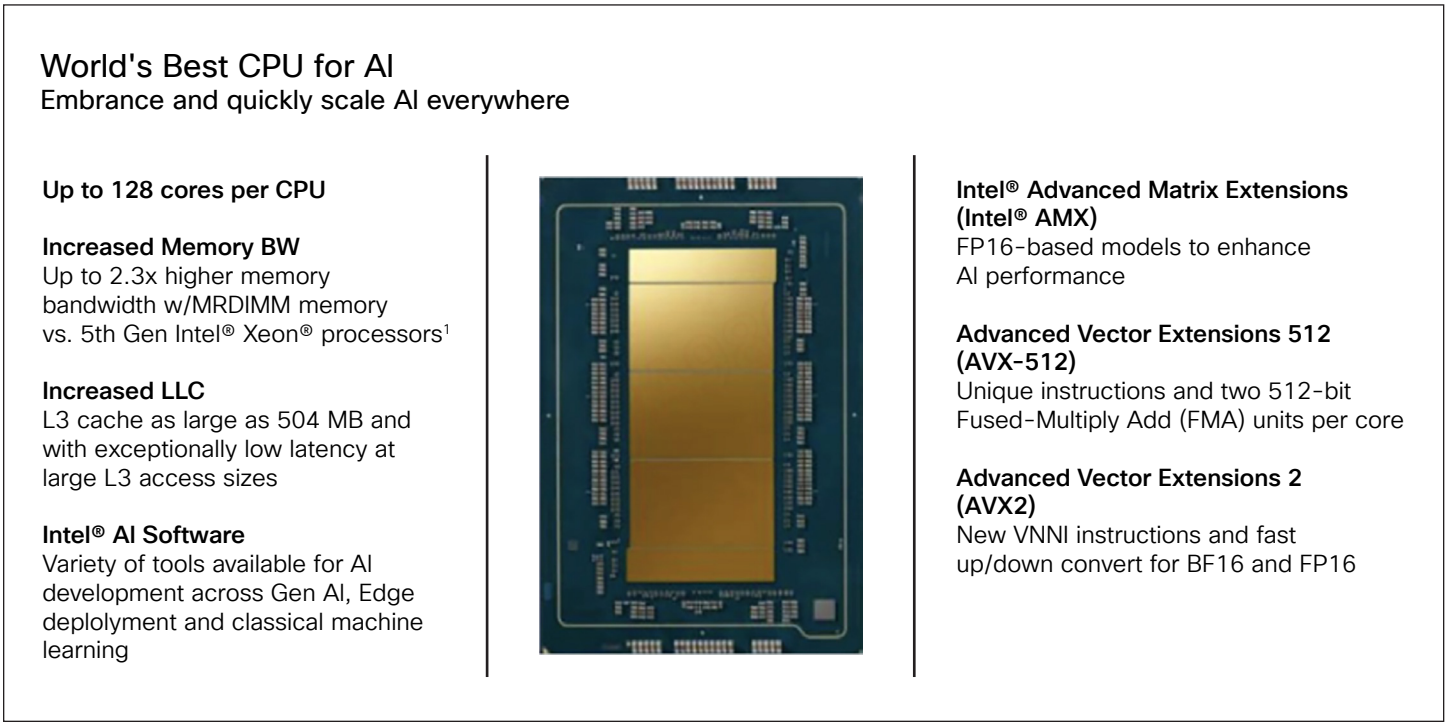

Figure 3. Features of the Intel Xeon 6 Processor

What is Intel AMX (Advanced Matrix Extensions)?

Intel AMX is a built-in accelerator that enables Intel Xeon 6 Processors to optimize Deep Learning (DL) training and inferencing workloads. With the high-speed matrix multiplications enabled by Intel AMX, Intel Xeon 6 Processors can quickly pivot between optimizing general computing and AI workloads.

Imagine an automobile that could excel at city driving and then quickly shift to deliver Formula 1 racing performance. Intel Xeon 6 Processors deliver this level of flexibility. Developers can code AI functionality to take advantage of the Intel AMX instruction set as well as code non-AI functionality to use the processor Instruction Set Architecture (ISA). Intel has integrated the oneAPI Deep Neural Network Library (oneDNN)—its oneAPI DL engine—into popular open-source tools for AI applications, including TensorFlow, PyTorch, PaddlePaddle, and Open Neural Network Exchange (ONNX).

Intel AMX architecture

Intel AMX architecture consists of two components, as shown in Figure 4:

Tiles consist of eight two-dimensional registers, each 1 kilobyte in size. They store large chunks of data.

Tile Matrix Multiplication (TMUL) is an accelerator engine attached to the tiles that perform matrix-multiply computations for AI.

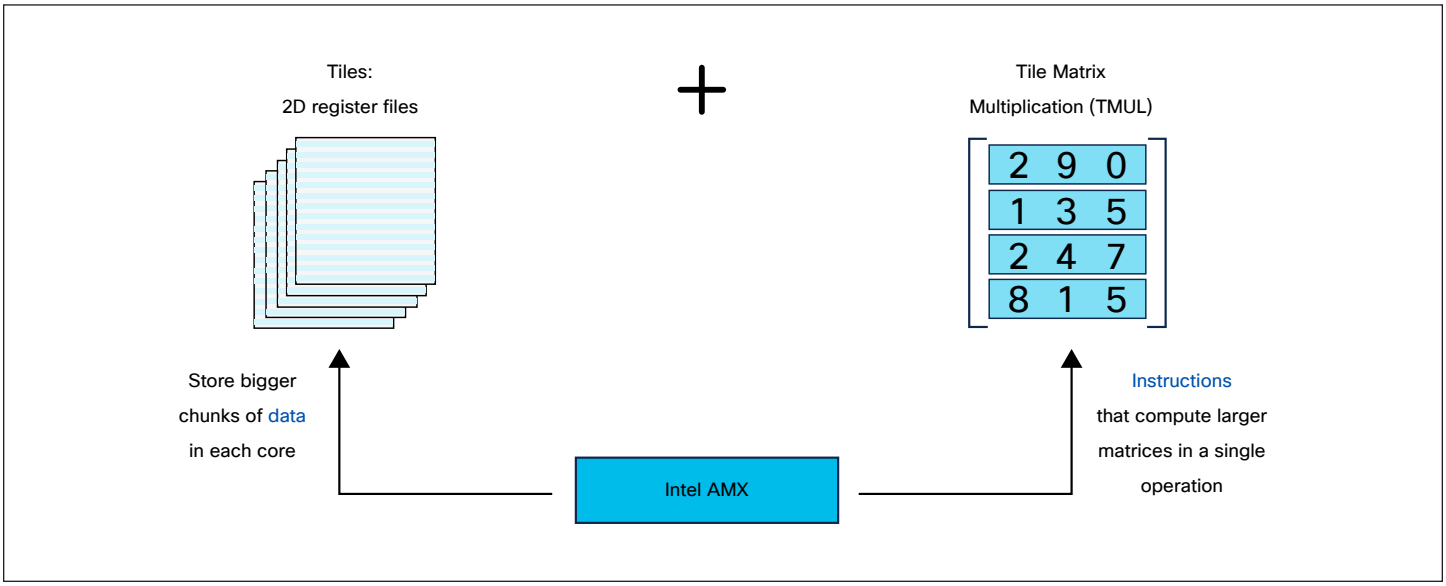

Figure 4. Intel AMX architecture consists of 2D register files (tiles) and TMU.

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 Inference

Inferencing refers to the process of using an existing machine-learning model to make predictions or decisions based on new data inputs. In other words, it involves applying a trained model to unseen data and generating outputs that can be used for various purposes such as classification, regression, or recommendation systems.

Inferencing is a critical component of many AI applications, including Natural Language Processing (NLP), computer vision, and predictive analytics.

MLPerf Inference is a data-center benchmark suite that 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 benchmarks measure the speed and efficiency of systems in performing this inferencing task. Typical performance metrics include throughput (measured in queries or tokens per second), latency (measured in milliseconds or seconds), and accuracy.

Typical performance metrics in MLPerf Inference

Throughput: the number of inference queries or tokens a system can process in a given time (for example, queries per second, tokens per second). It indicates the system's ability to handle a large volume of inference requests.

Latency: the time it takes for a system to complete a single inference query (for example, in milliseconds or seconds)

Accuracy: the correctness of the model's predictions. It ensures that the model is making reliable predictions. The **MLPerf Inference benchmark paper** provides a detailed description of the motivation and guiding principles behind the MLPerf Inference 5.1: Datacenter benchmark suite.

MLPerf Inference test configurations

For MLPerf Inference version 5.1 testing, we have used the hardware and software configurations shown in Table 1.

Table 1. Cisco UCS X210c M8 Compute Node configuration

Item	Specification
System name	Cisco UCS X210c M8 Compute Node
System type	Data center
Number of nodes	1
Host processor model	Intel Xeon 6 Processors
Processor name	Intel Xeon 6787P
Host processors per node	2
Host processor core count	86
# of threads	172
Host processor frequency	2.00 GHz, 3.80 GHz Intel Turbo Boost
Cache L3	336 MB
Memory type	DDR5 (6400MT/s)
Host memory capacity	2.04 TB
Host storage capacity	7.6 TB

Table 2. Software stack and system configuration

Item	Specification
OS	Ubuntu 22.04.4
Kernel	5.15.0-151-generic
CPU frequency governor	Performance
Framework	PyTorch for all models

Table 3. Data-center suite for MLPerf Inference 5.1 benchmarks

Source: MLCommons

Model	Dataset	Server latency constraints	QSL size	Quality
Dlrm-v2-99.9	1TB click logs	30 ms	204800	99% of FP32 and 99.9% of FP32 (AUC=80.25%)
Retinanet	OpenImages (800×800)	100 ms	64	99% of FP32 (0.3755 mAP)
RGAT	IGBH	NA	788379	99% of FP32 (72.86%)
Whisper	LibriSpeech	NA	1633	99% of FP32 and 99.9% of FP32 (WER=2.0671%)
Llama 3.1 8B	CNN Dailymail (v3.00, max_seq_len=2048)	Conversational: TTFT/TPOT: 2000 ms/100 ms. Interactive: TTFT/TPOT: 500 ms/30 ms.	13368	99% of FP32 and 99.9% of FP32 (rouge1=42.9865, rouge2=20.1235, rougeL=29.9881). Additionally, for both cases the total generation length of the texts should be more than 90% of the reference (gen_len=8167644)

Table 4. Server BIOS settings applied for MLPerf Inference 5.1 benchmarking

BIOS setting	Recommended value
Hyperthreading	Disabled
Turbo boost	Enabled
LLC prefetch	Enabled
CPU power and perf policy	Performance
NUMA-based cluster	SNC2
Hardware P state	Native (based on OS guidance)
Energy perf BIAS	OS controls EFB
Energy efficient turbo	Disable

Benchmark scenarios

The models are deployed in a variety of critical inference applications or use cases known as “scenarios,” where each scenario requires different metrics, demonstrating production environment performance in practice. Following is the description of each scenario. Table 5 shows the scenarios required for each data-center benchmark included in this submission.

Offline scenario: represents applications that process the input in batches of data available immediately and do not have latency constraints for the metric performance measured in samples per second.

Server scenario: represents deployment of online applications with random input queries. The metric performance is measured in Queries Per Second (QPS), subject to latency bound. The server scenario is more complicated in terms of latency constraints and input queries generation. This complexity is reflected in the throughput-degradation results compared to the offline scenario.

Each data-center benchmark requires the following scenarios:

Table 5. MLPerf Inference 5.1 benchmark scenarios

Source: MLCommons

Area	Task	Model	Required scenarios
Vision	Recommendation	Dlrm-v2-99.9	Server, offline
Vision	Object detection	Retinanet	Server, offline
Vision	Node classification	RGAT	Offline
Speech	Speech-to-text	Whisper	Offline
Language	Language processing	Llama-3.1-8B	Server, offline

MLPerf Inference performance results

Table 6. Summary of MLPerf Inference 5.1 benchmark performance results

Benchmark/model	Inferences/s	
	Offline	Server
DLRM-v2-99.9	12503.3	11801.67
Retinanet	501.263	400.42
RGAT	16102.2	NA
Whisper	1418.6	NA
Llama 3.1 8B	819.624	257.75

Intel® Xeon® 6787P Processor:

Granite Rapids is the code name for the Intel Xeon 6 Processor family.

The Intel Xeon 6787P is a Xeon 6 family processor designed for high-performance computing and AI workloads. It's part of the Intel Xeon 6 Processor family, offering 86 cores, such features as double data rate 5 synchronous (DDR5) memory and PCIe 5.0, and a TDP of 350W.

In this section, we have shown the MLPerf inference 5.1 results for Intel Xeon 6 processors.

DLRM-v2-99.9

The DLRM-v2-99.9 model, a highly accurate and optimized version of the deep learning recommendation model (DLRM), is designed to evaluate the performance of recommendation systems—such as those used for personalized ads or content suggestions—on high-performance hardware platforms.

Figure 5 shows the performance of the DLRM-v2-99.9 model tested on a Cisco UCS X210c M8 Compute Node with 2x Intel Xeon 6787P Processors.

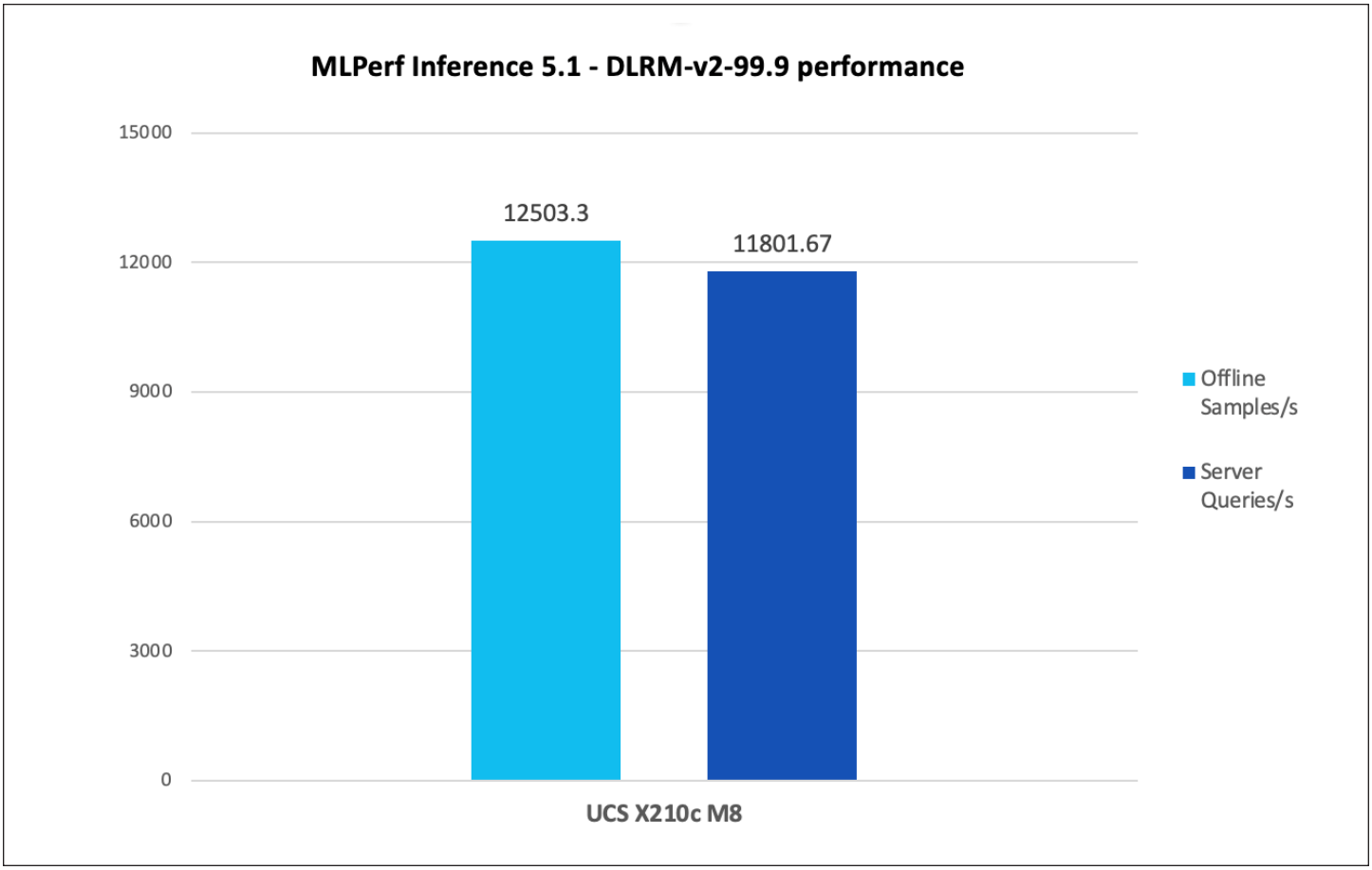

Figure 5. DLRM-v2-99.9 inference throughput in offline and server scenarios

Retinanet

Retinanet is a single-stage object-detection model known for its focus on addressing class imbalances using a novel focal-loss function. The “800x800” refers to the input image size, and the model is optimized for detecting small objects in high-resolution images.

Figure 6 shows the performance of the Retinanet model tested on a Cisco UCS X210c M8 Compute Node with 2x Intel Xeon 6787P Processors.

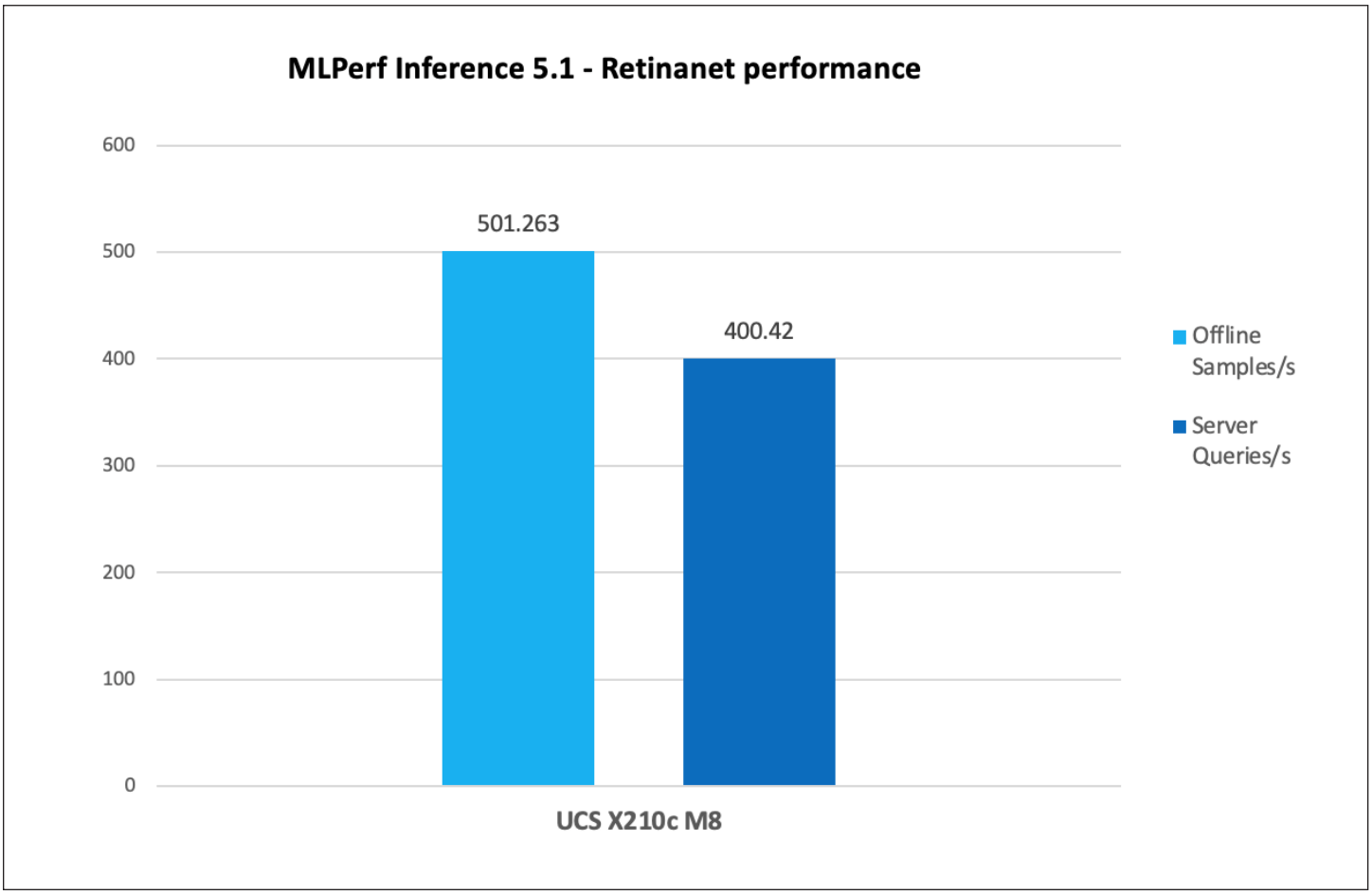

Figure 6. Retinanet inference throughput in offline and server scenarios

RGAT

RGAT (Relational Graph Attention Network) is a type of AI model designed to understand and process complex data structured as “graphs,” where different elements (nodes) are connected by various types of relationships (edges). It is especially useful for knowledge graphs and social network analysis.

Figure 7 shows the performance of the RGAT model tested on a Cisco UCS X210c M8 Compute Node with 2x Intel Xeon 6787P Processors.

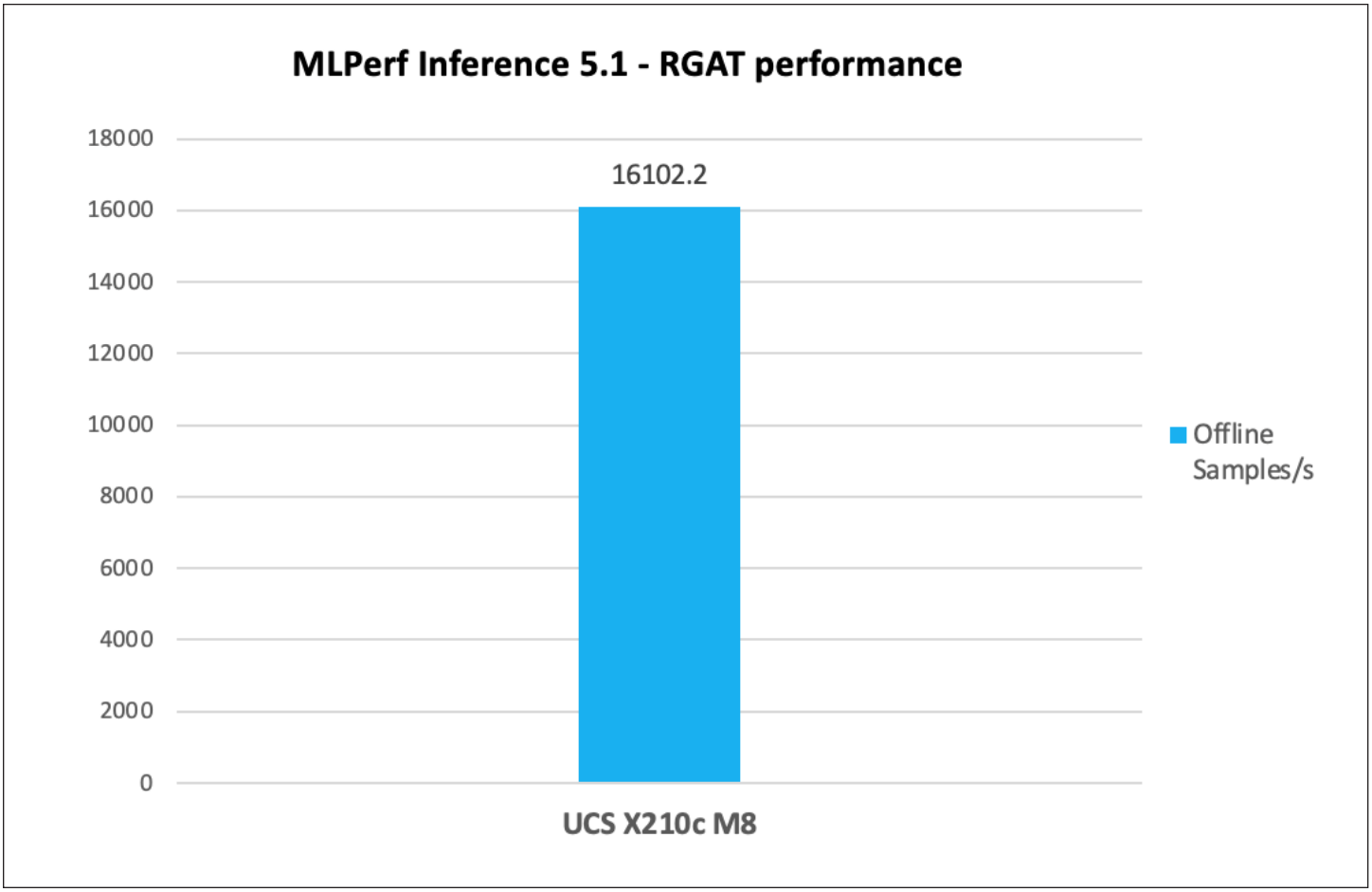

Figure 7. RGAT inference throughput in offline scenarios

Whisper

Whisper is an open-source artificial intelligence system developed by OpenAI. Its main purpose is Automatic Speech Recognition (ASR). It listens to speech, understands different languages, and turns spoken words into text and is useful for voice commands, subtitles, or transcribing conversations. Whisper has high-accuracy, multilingual, and translation capabilities.

Figure 8 shows the performance of the Whisper model tested on a Cisco UCS X210c M8 Compute Node with 2x Intel Xeon 6787P Processors.

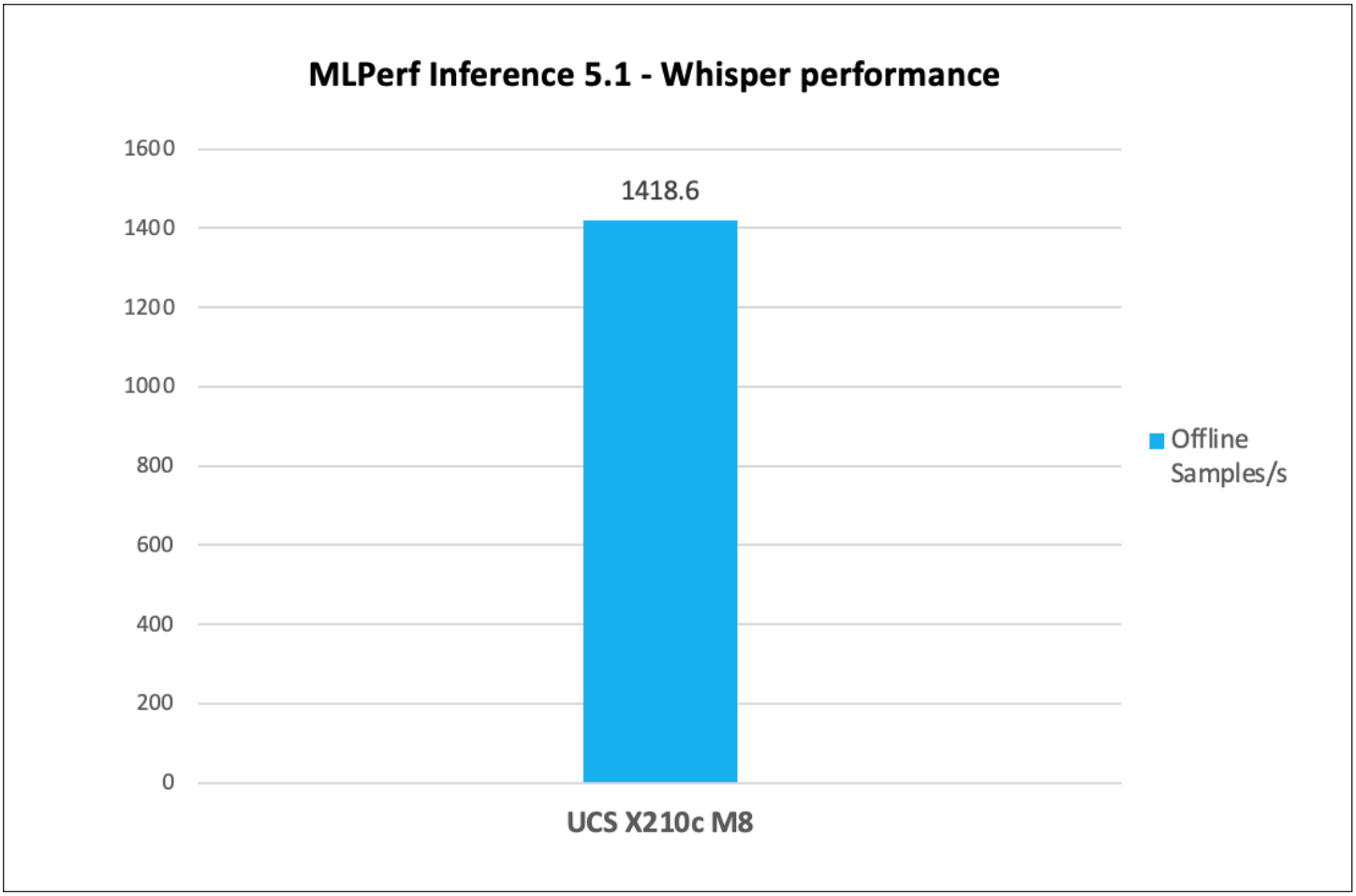

Figure 8. Whisper inference throughput in offline scenarios

Llama 3.1 8B

The Llama 3.1 8B model is a powerful, multilingual large language model that is optimized for dialogue use cases. With 8 billion parameters and trained on 15 trillion tokens, it can handle a wide range of tasks, from text generation to conversation, text summarization, text classification, sentiment analysis, and language translation requiring low-latency inferencing.

Figure 9 shows the performance of the Llama 3.1 8B model tested on a Cisco UCS X210c M8 Compute Node with 2x Intel Xeon 6787P Procesors.

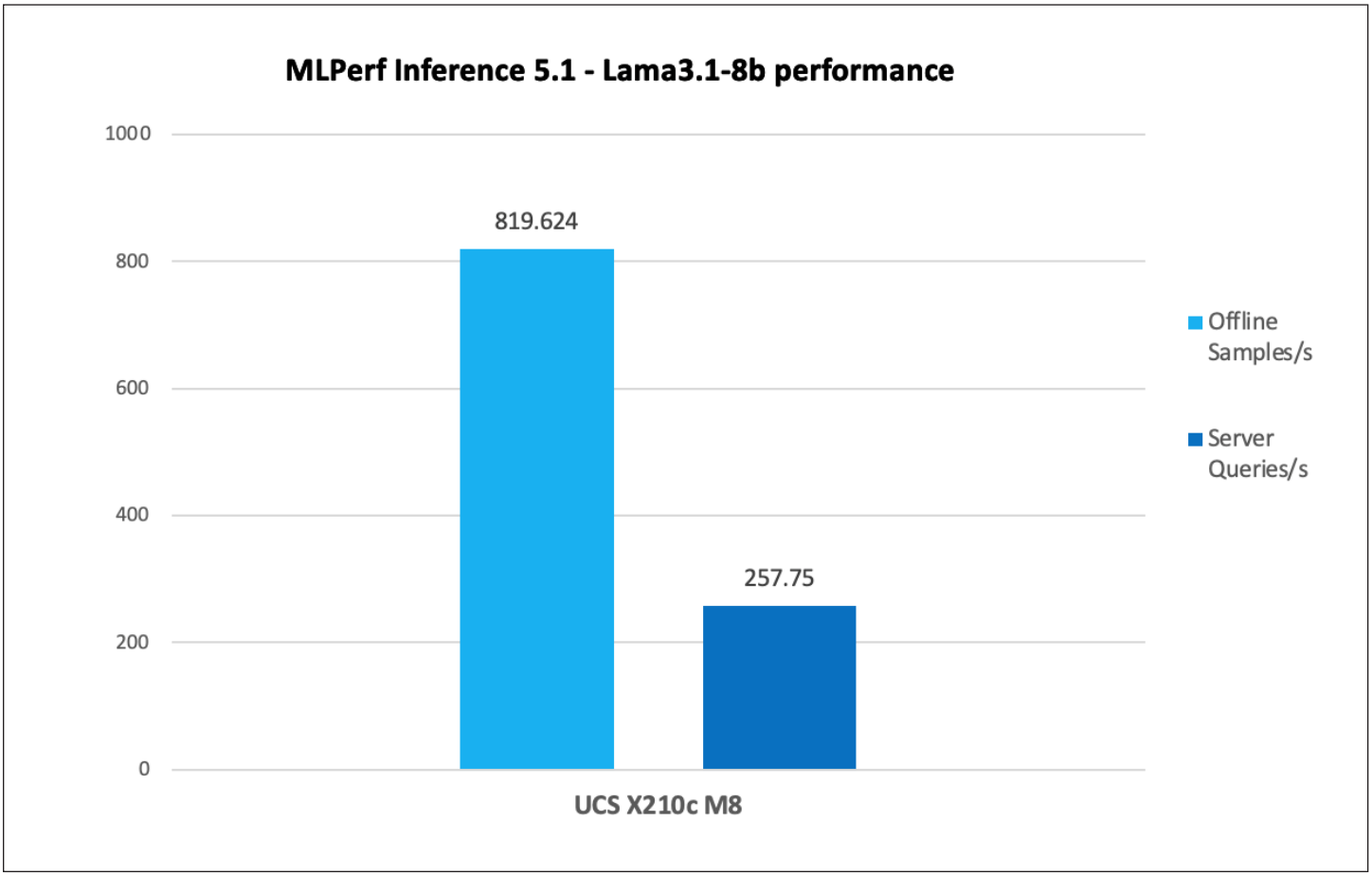

Figure 9. Llama 3.1 8B inference throughput in offline and server scenarios

Conclusion

The Cisco UCS X210c M8 Compute Node with Intel Xeon 6 P-core processors is designed for high-performance solution for data centers to support various machine learning and complex workloads, including databases and advanced analytics. Its capabilities are evident in supporting tasks such as large language models, natural language processing, object detection, speech translation, node classification, and recommendation systems. This white paper outlines a comprehensive roadmap for businesses aiming to leverage AI technology securely, scalably, and ethically.

References

For more information about Cisco UCS X210c M8 Compute Node specifications, the Intel Xeon 6 Processor family, and MLPerf Inference benchmark submission results, refer to the following links:

A specifications sheet for the Cisco UCS X210c M8 Compute Node is available at: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/x210cm8-specsheet.pdf

Cisco UCS X9508 Chassis specifications: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/x9508-specsheet.pdf

Intel Xeon 6 Processors specification: https://www.intel.com/content/www/us/en/products/details/processors/xeon/xeon6-p-cores.html

MLPerf Inference submission results: https://mlcommons.org/benchmarks/inference-datacenter/