White paper
Cisco public

AI Performance: MLPerf Inference on Cisco UCS C240 M8 with Intel® Xeon® 6 Processor

November 2025

© 2025 Cisco and/or its affiliates. All rights reserved.

Contents

Introduction 3

Executive summary 3

Technology overview 3

Key benefits of AI + Cisco UCS 5

Intel® Xeon® 6 AI capabilities 5

MLPerf overview 7

Typical Performance Metrics in MLPerf Inference 8

Benchmark scenarios 10

MLPerf Inference Performance Results 10

Conclusion 16

References 17

© 2025 Cisco and/or its affiliates. All rights reserved.

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 solutions with the help of Cisco®, focusing on inferencing.

This document covers the product features, MLPerf benchmark, test configuration, and results to explain the Artificial Intelligence use cases best suited for enterprises looking to invest in a mainstream rack server.

We present the MLPerf™ Data Center Inference results obtained on a Cisco UCSC C240 M8 powered by Intel® Xeon® 6 processor family.

Here we explore how the Cisco UCS C240M8 server with Intel® Xeon® 6 processor family 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.

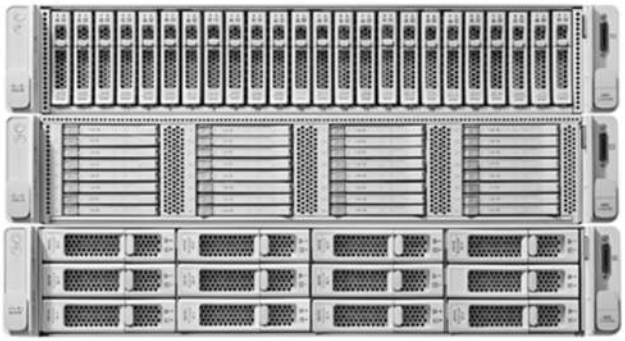
Technology overview

Cisco UCS C240 M8 Rack Server

The Cisco UCS C240 M8 Small Form-Factor Pluggable (SFF) Rack Server extends the capabilities of Cisco's 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.

The Cisco UCS C240 M8 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.

© 2025 Cisco and/or its affiliates. All rights reserved.



The Cisco UCS C240 M8 rack server 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, 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 speed up compute processing and memory access.
- Improved cost-effectiveness is provided by combining the latest-generation AI hardware with software optimizations, which potentially lowers 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 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.

© 2025 Cisco and/or its affiliates. All rights reserved.

Key benefits of AI + Cisco UCS

Cisco Unified Computing System (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 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 like 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, Xeon 6 processor helps 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.

Intel AMX matrix multiplication engine in each core to boost overall inferencing performance. With a focus on ease of use, Cisco Technologies delivers 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.

© 2025 Cisco and/or its affiliates. All rights reserved.

Intel® Xeon® 6 processor (as shown in Figure 1) excel at the complete spectrum of workloads, with a mainstream series that features a range of eight 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 a remarkable 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.

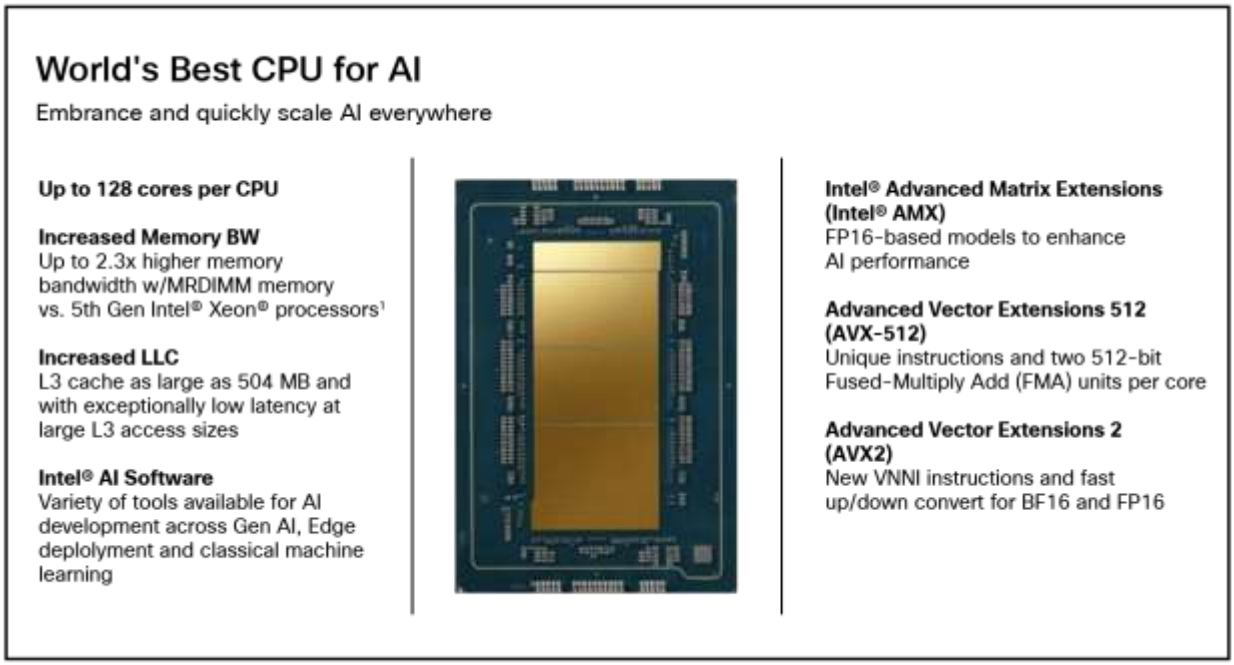


Figure 1.
Intel® Xeon® 6 processor.

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 Scalable 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 Scalable 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).

© 2025 Cisco and/or its affiliates. All rights reserved.

AMX architecture

Intel AMX architecture consists of two components, as shown in Figure 2:

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.

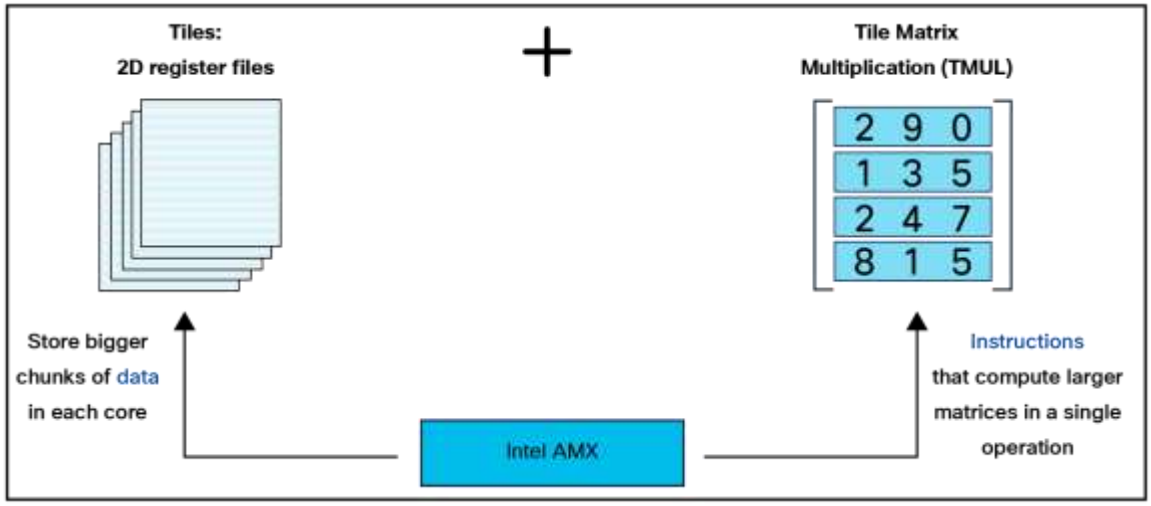


Figure 2.
Intel AMX architecture consists of 2D register files (tiles) and TMU

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 datacenter 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.

© 2025 Cisco and/or its affiliates. All rights reserved.

Typical Performance Metrics in MLPerf Inference

Throughput: The number of inference queries or tokens a system can process in a given time (e.g., 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 (e.g., 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: Datacenter benchmark suite.

MLPerf Inference test configuration

For MLPerf testing, we have used the following hardware and software configurations:

Table 1. Cisco UCSC C240 M8 Server Configuration

Item	Specification
System name	UCSC C240 M8
System type	Data Center
Number of nodes	1
Host processor model	6th Generation Intel® Xeon® Scalable Processors
Processor name	6787P
Host processors per node	2
Host processor core count	86
# of threads	172
Host processor frequency	2.00 GHz, 3.80 GHz Turbo Boost
Cache L3	336 MB
Memory type	DDR5 (6400MT/s)
Host memory capacity	1.04 TB
Host storage capacity	6.8 TB

© 2025 Cisco and/or its affiliates. All rights reserved.

Table 2. Software stack and system configuration

Item	Specification
OS	Ubuntu 24.04.1
Kernel	5.15.0-131-generic
CPU frequency governor	Performance
Framework	PyTorch, (INT4 for GPT-J, and INT8 for all other models)

Table 3. Data Center Suite Benchmarks

Source: MLCommons

Model	Dataset	Server Latency Constraints	QSL Size	Quality
ResNet50-v1.5	ImageNet (224×224)	15 ms	1024	99% of FP32 (76.46%)
Retinanet	OpenImages (800×800)	100 ms	64	99% of FP32 (0.3755 mAP)
3D UNET	KITS 2019 (602x512x512)	N/A	16	99% of FP32 and 99.9% of FP32 (0.86330 mean DICE score)
GPT-J 6B	CNN-DailyMail News Text Summarization	20 seconds	13368	99.9% or 99% of the original FP32 ROUGE 1 – 42.9865 ROUGE 2 – 20.1235 ROUGE L – 29.9881
Whisper	LibriSpeech	NA	1633	99% of FP32 and 99.9% of FP32 (WER=2.0671%)

Table 4. Server BIOS settings applied for MLPerf benchmarking

BIOS Setting	Recommended value
Hyperthreading	Disabled
Turbo Boost	Enabled
CPU Performance	Enterprise
LLC Prefetch	Disable
Processor EPP Profile	Performance
Sub-NUMA Clustering	SNC2
Enhanced CPU Performance	Auto
Power performance tuning	BIOS
Energy Efficient Turbo	Disable

© 2025 Cisco and/or its affiliates. All rights reserved.

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. Data Center Suite Benchmark Scenarios

Source: MLCommons

Area	Task	Model	Required Scenarios
Vision	Image classification	ResNet50-v1.5	Server, offline
Vision	Object detection	Retinanet	Server, offline
Vision	Medical imaging	3D UNET	Offline
Language	Summarization	GPT-J 6B	Server, offline
Speech	Speech-to-text	Whisper	Offline

MLPerf Inference Performance Results

Table 6. Summary of MLPerf v5.0 Inference performance results

Benchmark / Model	Inferences/s	
	Offline	**Server**
ResNet50-v1.5	31559.2	27988
Retinanet	476.59	350.92
GPT-J 6B	315.39	168.69
3D-UNET	2.26	NA

© 2025 Cisco and/or its affiliates. All rights reserved.

Comparing performances of different generations of CPUs

The benchmark results are based on Cisco UCS C240 M7 server with 2x Intel® Xeon® EMR 8792+ CPUs vs Cisco UCS C240 M8 with Intel® Xeon® 6 2x 6787P CPUs along with optimized software stacks.

Intel® Xeon® EMR 8792+ CPU:

The Intel® Xeon® EMR 8792+ CPU is Intel® Xeon® 5 Processor family and features a 12-core design with a base clock speed of 2.40 GHz and a turbo boost of 3.10 GHz. It supports DDR5 memory at speeds up to 5600 MT/s and has a Thermal Design Power (TDP) of 125W.

Intel® Xeon® 6 6787P CPU:

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

The Intel® Xeon® 6787 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, features like Double Data Rate 5 Synchronous (DDR5) memory, PCIe 5.0, and has a TDP of 350W.

In this section, we show the performance in the comparing mode so the improvement between different generations of CPUs can be easily observed.

Resnet50

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

Figure 3 shows the performance comparison of the Resnet50 model tested on UCS C240 M8 server with 2x 6787P CPUs vs UCS C240 M7 server with 2x 8592+ CPUs.

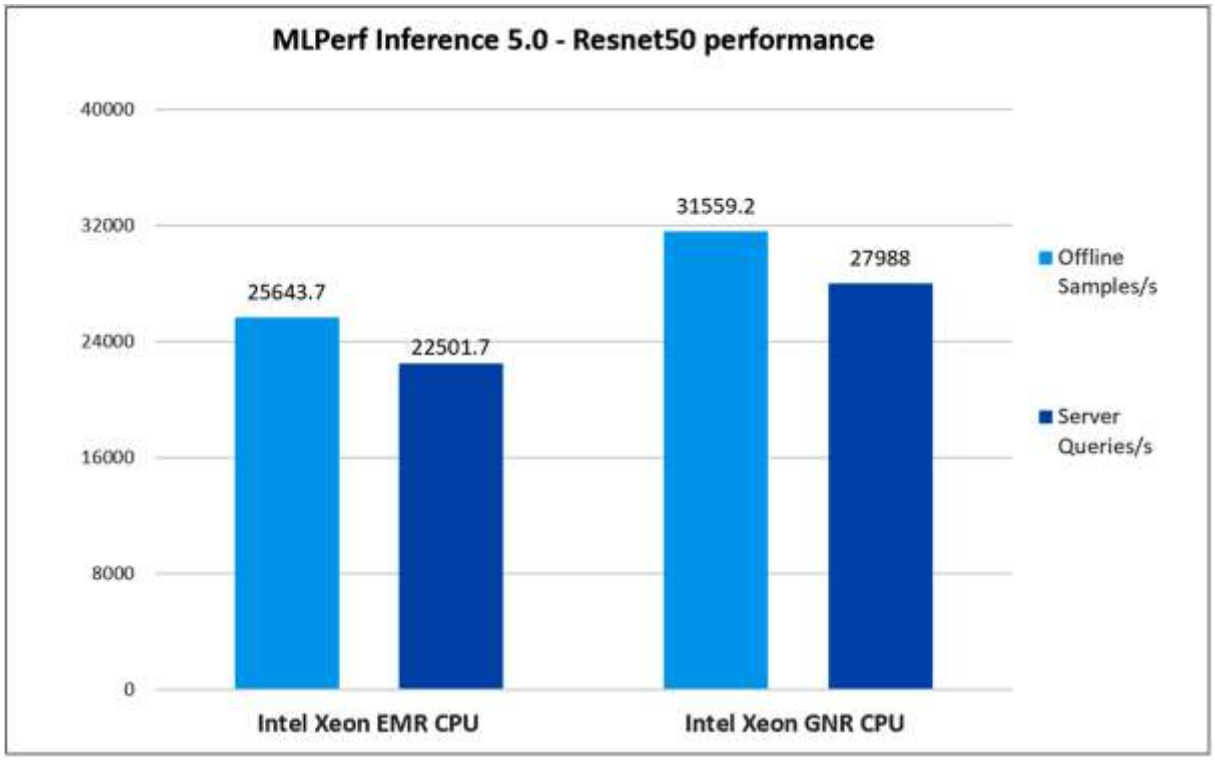


Figure 3.
Resnet50 inference throughput in Offline and Server scenarios

© 2025 Cisco and/or its affiliates. All rights reserved.

Retinanet

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

Figure 4 shows the performance of the Retinanet model tested on UCS C240 M8 server with 2x 6787P CPUs vs UCS C240 M7 server with 2x 8592+ CPUs.

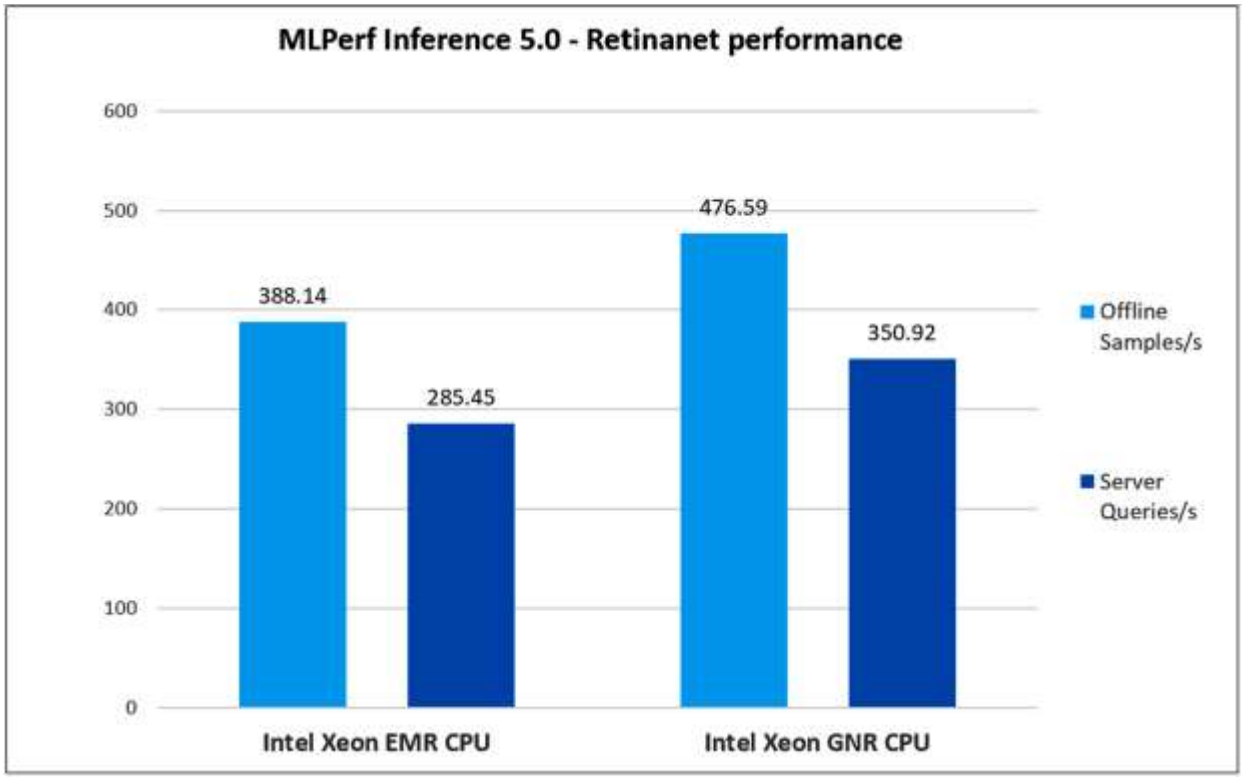


Figure 4.
Retinanet inference throughput

GPTJ

GPTJ is an open-source transformer-based language model developed by EleutherAI. It has 6 billion parameters and is designed for tasks such as text generation, translation, summarization, and other natural language processing tasks. Intel MLPerf used vLLM (inference serving framework) for the GPT-J benchmark, it would be good to mention vellum as part the Technology overview section. vLLM is very popular inference framework in the industry and is used in the deployments.

Refer for more details : https://docs.vllm.ai/en/stable/getting_started/installation/cpu.html.

© 2025 Cisco and/or its affiliates. All rights reserved.

Figure 5 shows the performance of the GPTJ model tested on UCS C240 M8 server with 2x 6787P CPUs vs UCS C240 M7 server with 2x 8792+ CPUs.

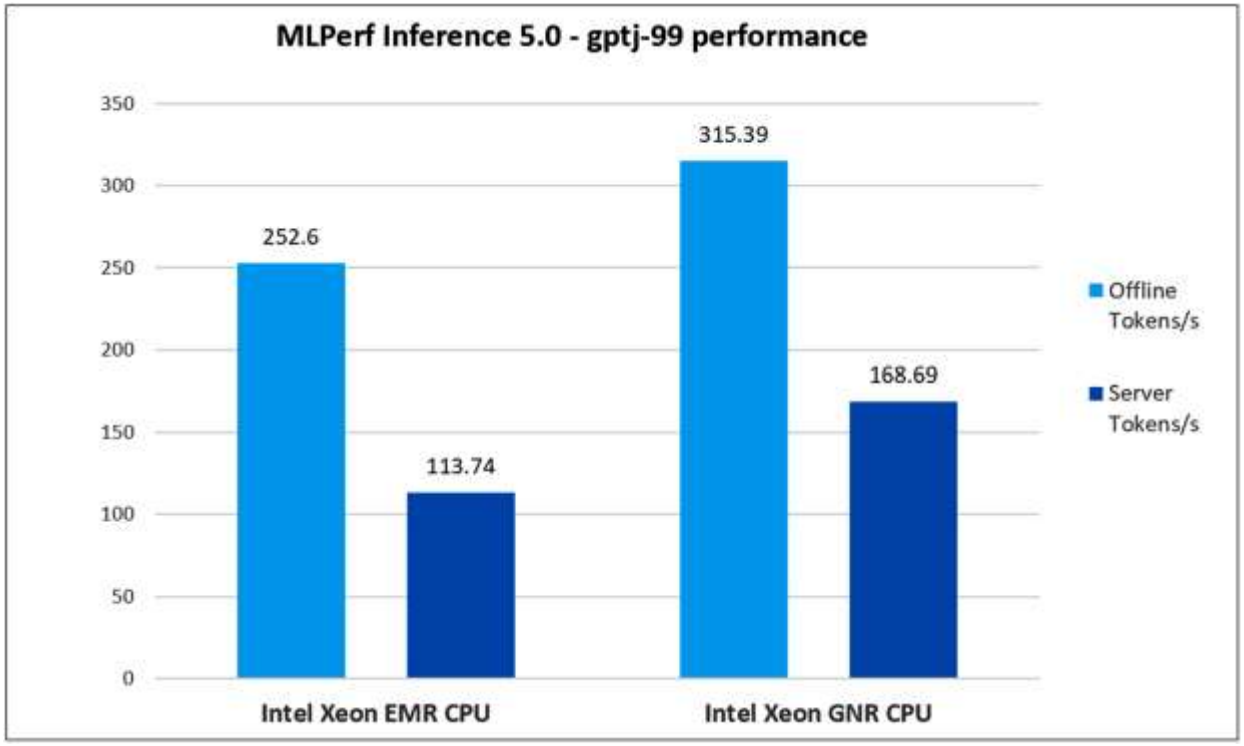


Figure 5.
gptj-99 inference throughput in Offline and Server scenarios

3D-unet

The 3D U-Net is an extension of the original U-Net architecture designed for processing three-dimensional volumetric data, such as CT or MRI scans. It employs a U-shaped architecture with an encoder (contractive path) and a decoder (expanding path), using 3D convolutions to analyze volumetric context.

Figure 6 shows the performance of the 3D-Unet model tested on UCS C240 M8 server with 2x 6787P CPUs vs UCS C240 M7 server with 2x 8792+ CPUs.

© 2025 Cisco and/or its affiliates. All rights reserved.

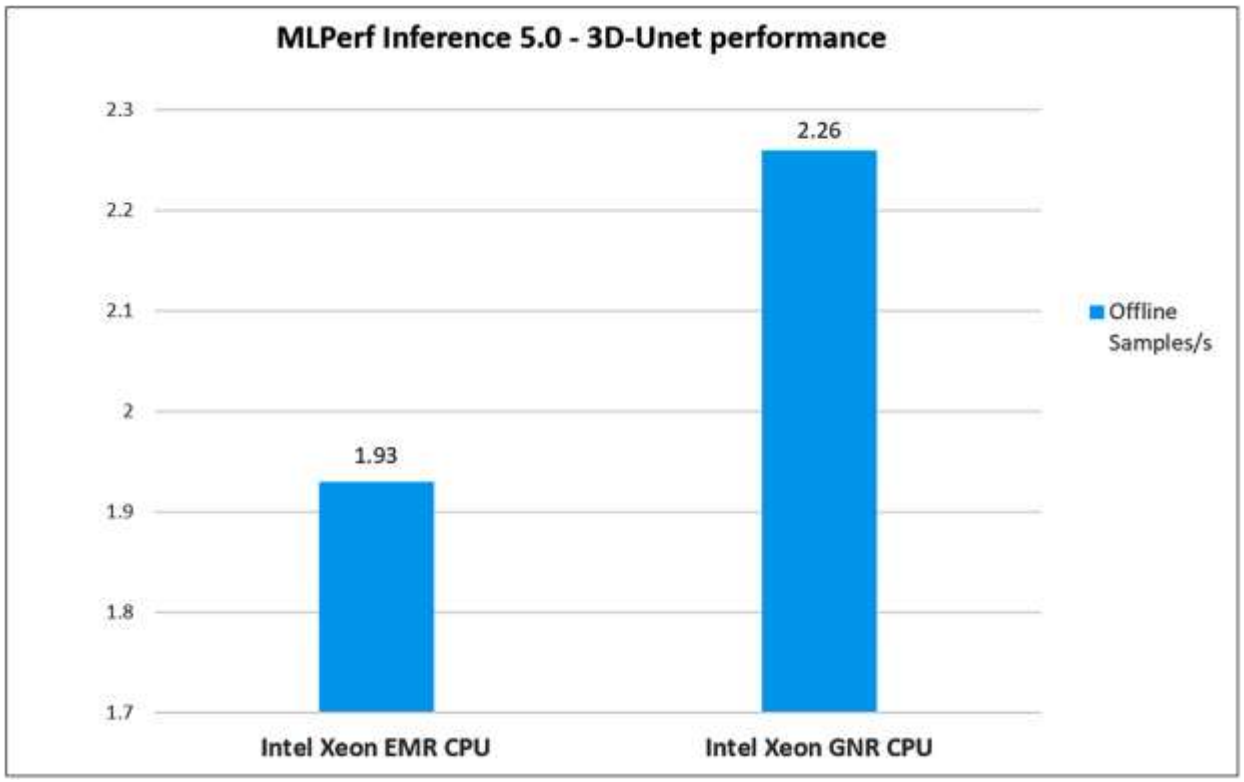


Figure 6.
3D-unet inference throughput in Offline scenario

Table 7. Summary of MLPerf 5.1 Inference performance results

Benchmark / Model	Inferences/s	
	Offline	Server
Retinanet	498.421	400.42
Whisper	1403.84	NA

© 2025 Cisco and/or its affiliates. All rights reserved.

Retinanet

Figure 7 shows the performance of the Retinanet model tested on a Cisco UCS C240 M8 Compute Node with 2x Intel Xeon GNR 6787P Processors.

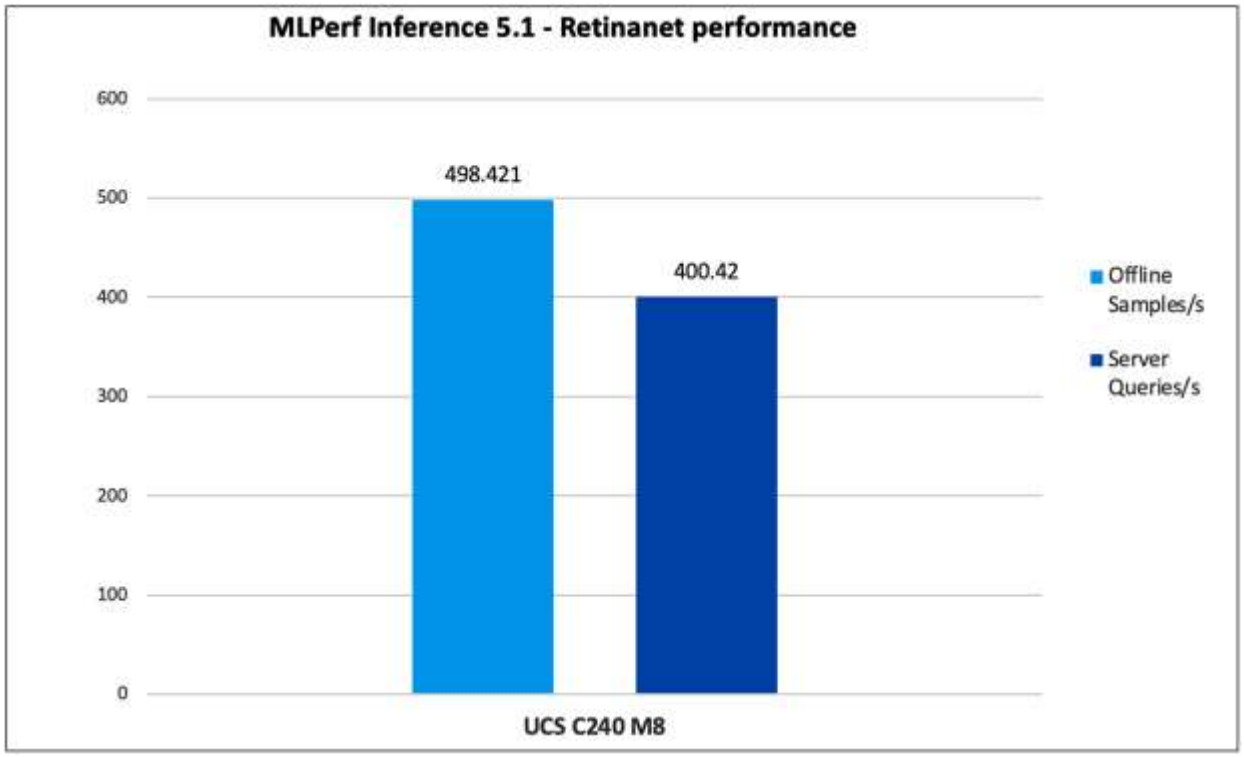


Figure 7.
Retinanet inference throughput in offline and server scenarios

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.

© 2025 Cisco and/or its affiliates. All rights reserved.

Figure 8 shows the performance of the Whisper model tested on a Cisco UCS C240 M8 Compute Node with 2x Intel Xeon 6787P Processors.

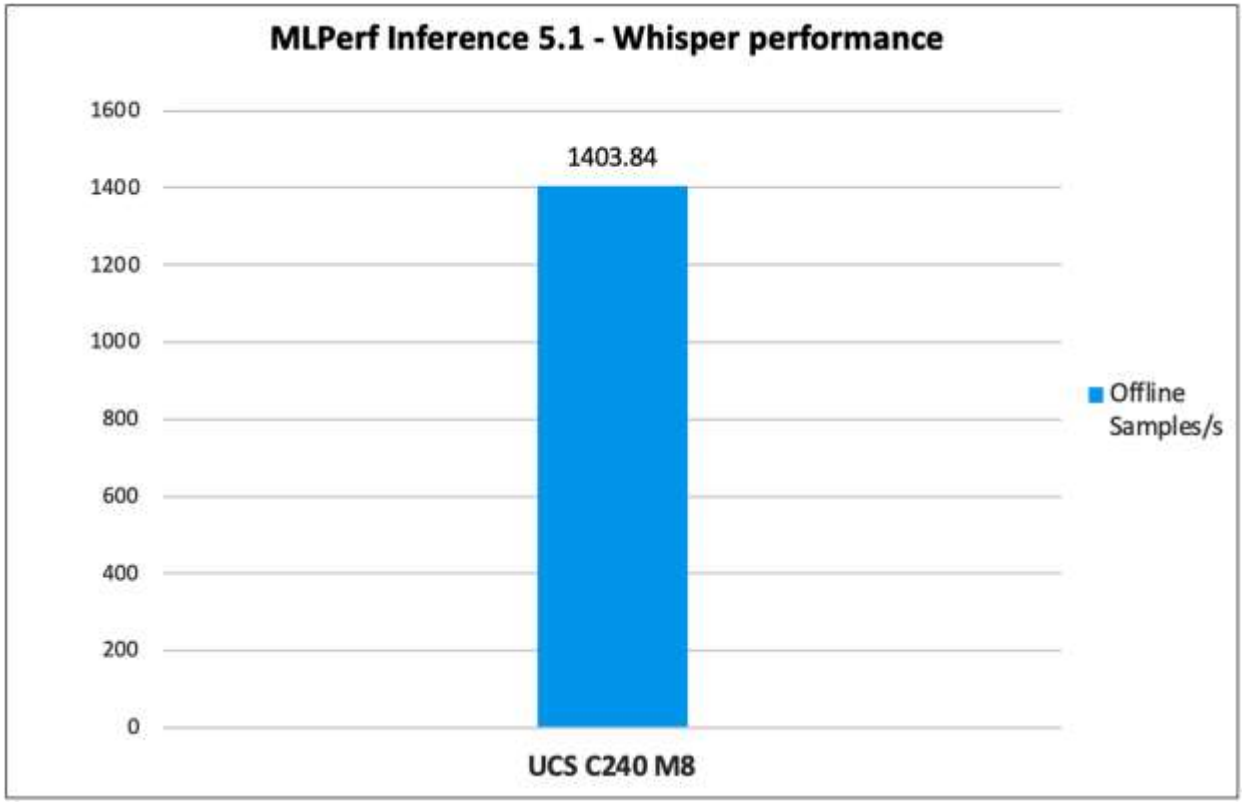


Figure 8.
Whisper inference throughput in offline scenarios

Conclusion

The Cisco UCS C240 M8 server with Intel® Xeon® 6 P-core processor is designed for high-performance solution for data centers to support various deep learning and complex workloads, including databases and advanced analytics. Its capabilities are evident in supporting tasks such as natural language processing, image classification, object detection, medical imaging, and recommendation systems. This white paper outlines a comprehensive roadmap for businesses aiming to leverage AI technology securely, scalable, and ethically.

© 2025 Cisco and/or its affiliates. All rights reserved.

References

For more information about Cisco UCS C240 M8 server specifications, Intel® Xeon® 6 processors family and MLPERF inference benchmark submission results, refer to the following links:

A specifications sheet for the C240 M8 is available at: 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

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/

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

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

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

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

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: https://www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

Printed in USA C11-5143751-01 11/25

© 2025 Cisco and/or its affiliates. All rights reserved.