Proven AI performance—HPE ProLiant Compute DL380a Gen12

No. 1 ranking on 10 MLCommons[1] benchmarks

HPE

About the MLPerf Inference: Datacenter v5.0 benchmarks

MLPerf Inference: Datacenter v5.0 benchmarks measure the speed, accuracy, and efficiency of data center AI and ML systems in processing inputs and generating results using trained models, a crucial metric for optimizing performance.

Using these metrics, engineers can make reliable assessments of a system's ability to handle cutting-edge AI workloads so they can design reliable, high-performing, and efficient AI products and services that provide valuable insights to users.

Why benchmarks matter

With the rapid expansion of enterprise artificial intelligence (AI) and machine learning (ML) workloads, identifying the optimal performance and efficiency of the underlying hardware for specific use cases is essential, but also quite challenging. Open-source, unbiased benchmarks enable fair and reproducible comparisons of systems and provide a clear, data-driven approach to assessing their ability to handle cutting-edge AI workloads.

MLCommons™ — a collaboration between AI leaders from academia, research labs, and industry providers — has established the MLPerf™ Inference: Datacenter benchmark suite as the trusted standard for evaluating AI and ML systems. The results empower engineers to strike a balance between harnessing the full potential of AI while optimizing for performance and efficiency.

The HPE ProLiant Compute DL380a Gen12 server has achieved an industry-leading 10 world-record MLPerf Inference: Datacenter v5.0 benchmark results, setting a new standard for enterprise-grade AI solutions and highlighting our ability to deliver exceptional performance, scalability, and efficiency for AI workloads. These results translate into deeper insights, faster innovation cycles, and a competitive edge in the race to operationalize AI and ML to drive growth and transformation. Competitive benchmark results are proof points for a server's capability to run desired workloads at the level of business needs.

Mapping benchmarks to enterprise AI use cases

Generative AI (GenAI) and large language models (LLMs)

GenAI and LLMs handle content generation tasks with unprecedented speed and accuracy, unleashing new efficiencies, enhancing decision-making, and delivering personalized experiences at scale. Related MLPerf benchmarks include **GPT-J**, which tests how well LLMs perform on tasks such as text generation, **Mixtral 8x7b**, which assesses versatility across tasks such as reasoning, coding, and answering questions, and **Llama2-70B**, which measures general natural language processing (NLP) performance for applications such as virtual assistants.

Computer vision

Computer vision accelerates tasks such as image recognition, object detection, and pattern analysis to drive innovation across workloads such as medical imaging, autonomous vehicles, and quality control. Related MLPerf benchmarks include **Resnet50**, which tests image classification performance, and **Retinanet**, which evaluates object detection capabilities, focusing on identifying and localizing objects within images.

Recommendation engines

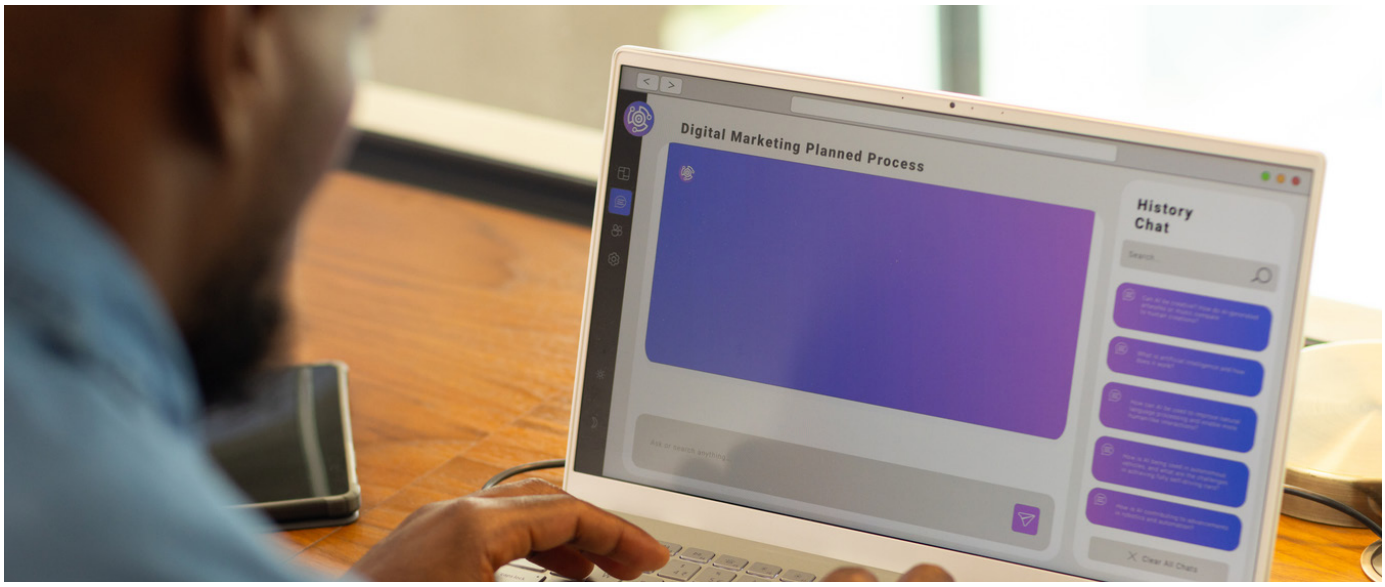
AI-driven recommendation engines are pivotal for delivering personalized experiences that enhance customer satisfaction and drive business growth by suggesting relevant products, content, or services.

Scenario descriptions

In MLCommons' MLPerf Inference: Datacenter benchmarks, the **Server** and **Offline** scenarios evaluate different aspects of system performance.

Server: This scenario represents low-latency inference applications and simulates real-time applications by sending inference requests to the system under test (SUT) following a Poisson distribution. This assesses the system's ability to handle unpredictable, concurrent queries, measuring the maximum throughput it can sustain under latency constraints.

Offline: This scenario represents high-batch size inference applications and measures how many inference queries a system can process when all inputs are available up front. This scenario reflects batch processing tasks, focusing on total throughput without latency constraints.



The HPE ProLiant Compute DL380a Gen12—A perfect scale-up solution for entry-level AI workloads

The rapid advancements in AI have introduced new challenges in finding the right infrastructure solutions to support demanding AI workloads such as LLMs and natural language processing (NLP), computer vision, and recommendation engines. These applications require a delicate balance of computational power, memory capacity, high-speed data processing, and storage flexibility to manage vast and complex datasets efficiently. Additionally, helping ensure scalability and reliability under heavy workloads is crucial to maintain performance and deliver timely results. Navigating these requirements to identify a single, versatile infrastructure solution can be daunting.

The HPE ProLiant Compute DL380a Gen12, designed to address the rigorous demands of these contemporary AI workloads, stands out with its support for up to ten double-wide GPUs while accommodating a variety of high performance NVIDIA GPUs such as the H200 NVL, H100, L40S, L4, and RTX PRO™ 6000 Server Edition. This flexibility in GPU support provides the necessary acceleration to handle intricate computations and large-scale data processing efficiently. Furthermore, it offers powerful scalability within a single node with its support of multiple high performance GPUs, which enables organizations to scale compute capacity without the complexity of distributed systems.

In addition to GPUs, the HPE ProLiant Compute DL380a Gen12 is powered by Intel® Xeon® 6 CPUs, offering up to 144 cores per processor. This extensive processing capability is crucial for running complex algorithms and managing substantial datasets seamlessly. The server also features a memory capacity of up to 8 TB, which is essential for high-demand operations and large datasets, helping ensure smooth and efficient performance. Storage needs are met with support for up to 8 SFF or 16 EDSFF drives, providing versatile and high-capacity options to manage the extensive data requirements of AI workloads.

The HPE ProLiant Compute DL380a Gen12 also emphasizes reliability with advanced power management features that enhance GPU reliability. The server includes six dedicated and redundant power supplies for GPUs, helping ensure consistent performance even under heavy loads. This robust design helps maximize uptime and provides the stability required for mission-critical AI applications.

All in all, the HPE ProLiant Compute DL380a Gen12 is a comprehensive solution that meets the diverse needs of next-generation AI workloads, as evidenced by its recent results in MLPerf Inference: Datacenter v5.0 benchmark tests.

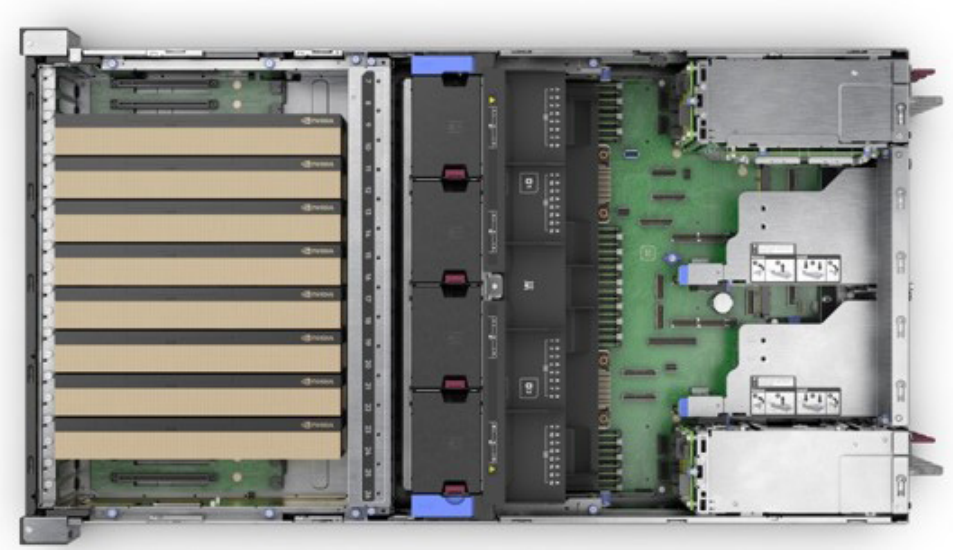


Figure 1. Top-down view of the HPE ProLiant Compute DL380a Gen12 showing eight double-wide GPUs

GenAI and LLMs test results

GenAI technologies such as NLP and LLMs are revolutionizing how machines understand and generate human language, powering everything from chatbots to advanced content creation. The MLPerf Inference: Datacenter v5.0 benchmark suite includes specific NLP and LLM tests — namely GPT-J and Llama2 — to evaluate infrastructure readiness for these workloads. Stated differently, these benchmarks indicate how well a system can handle LLM inference tasks such as text summarization and Q&A. The HPE ProLiant Compute DL380a Gen12 delivered outstanding performance in both, showcasing its unprecedented speed and accuracy in handling content generation tasks and delivering personalized experiences at scale.

In the Llama2-70B Offline benchmarks — which involved servers utilizing 8 NVIDIA® L40S GPUs — the HPE ProLiant Compute DL380a Gen12, powered by Intel Xeon 6740E processors, outperformed the Dell PowerEdge XE7745 equipped with AMD EPYC 9965 processors. The HPE ProLiant DL380a Gen12 achieved an impressive 3,656 tokens[2] per second,[3] surpassing the Dell PowerEdge XE7745's result of 3,482 tokens per second.[4]

Additionally, as highlighted in the following table, the DL380a Gen12 demonstrated superior performance in GPT-J 99 and GPT-J 99.9 benchmarks, reinforcing its versatility and efficiency in handling LLM inference tasks. Notably, these benchmark results leveraged both NVIDIA H200 NVL GPUs and L40S GPUs, exemplifying the optimized synergy between hardware and software in delivering cutting-edge AI solutions.

Table 1. MLPerf Inference: Datacenter v5.0 GPT-J 99 and 99.99 Server benchmark results (with 8 NVIDIA GPUs)

	GPT-J 99.0 Server scenario tokens/sec (with NVIDIA L40S GPUs)	GPT-J 99.0 Server scenario tokens/sec (with NVIDIA H200 NVL GPUs)
HPE ProLiant Compute DL380a Gen12	6,821[5]	18,065[6]
Dell PowerEdge XE7745	6,213[4]	17,975[7]

In addition to GPT-J and Llama2, the MLPerf Inference: Datacenter v5.0 benchmark suite also includes the Mixtral 8x7b test, which is an excellent measure for assessing versatility across various tasks such as reasoning, coding, and answering questions. The only submission for this test using a combination of CPUs and PCIe-based GPUs was the HPE ProLiant Compute DL380a Gen12, which utilized H200 NVL GPUs. The test results speak loudly with impressive results ranging from 50,124 to 52,206 tokens per second.[6]

#1

— Resnet50 Server

- HPE ProLiant Compute DL380a Gen12 with 8 H200 NVL GPUs — 620,188 queries/second
- HPE ProLiant Compute DL380a Gen12 with 8 L40S GPUs—344,052 queries/second[5]

— Resnet50 Offline

- HPE ProLiant Compute DL380a Gen12 with 8 H200 NVL GPUs — 686,884 queries/second[6]

Figure 2. HPE ProLiant Compute DL380a Gen12

Image classification and object detection test results

Computer vision is another critical AI workload that involves analyzing and interpreting visual data from the world, enabling applications such as facial recognition, autonomous vehicles, and medical imaging. This domain necessitates powerful GPUs, high-throughput data processing capabilities, and reliable storage to manage large volumes of image and video data. The MLPerf Inference: Datacenter v5.0 benchmark suite includes tests for computer vision — specifically Resnet50 and Retinanet to assess image classification performance and evaluate object detection capabilities, respectively. The HPE ProLiant Compute DL380a Gen12 is well-suited for these tasks due to its support for multiple GPUs, high-speed networking options, and flexible storage configurations, delivering outstanding performance in both benchmarks.

The MLCommons Resnet50 benchmark test is designed to evaluate the performance of systems in image classification tasks, providing a critical measure of how well hardware can handle computer vision workloads. Organizations often look at these results to gauge the efficiency and speed of different systems in processing large volumes of visual data. In one such test, the Resnet50 Server benchmark, the HPE ProLiant Compute DL380a Gen12 with NVIDIA H200 NVL GPUs took first place and achieved 15% more queries per second than the next highest submission (620,188[6] vs. 540,125[8]). Even more impressive is the fact that the HPE ProLiant Compute DL380a secured first place in three Resnet50 benchmark tests.[5, 6, 8]

The MLCommons Retinanet Server benchmark test focuses on assessing system performance in object detection tasks, providing a vital benchmark for understanding how well different hardware configurations can manage sophisticated computer vision applications. Organizations often examine these results to gauge the efficiency and speed of different systems in processing large volumes of visual data, particularly in applications that require precise object localization and identification. In one such test, the HPE ProLiant Compute DL380a Gen12 with 8 NVIDIA L40S GPUs outperformed the Dell PowerEdge XE7745, achieving 6,095 queries per second.[5] Additionally, when comparing systems with Intel Xeon processors and 8 NVIDIA H200 NVL GPUs, the DL380a was again the top performer, surpassing the Supermicro SYS-522GA-NRT[9] with an impressive 11,988.40 queries per second.[6]

57%

better than next best submission for the DLRM-v2 Offline benchmark

Deep learning recommendation model test results

Recommendation engines are essential for personalized user experiences in various sectors, including e-commerce, streaming services, and social media. These engines rely on vast amounts of user data and sophisticated algorithms to provide accurate and timely recommendations. Efficient data processing, high I/O throughput, and scalable storage are critical for their performance. The HPE ProLiant Compute DL380a excels in processing power, memory expansion, and storage versatility, thus making it an ideal solution for running recommendation engines. Its ability to manage large datasets and perform complex algorithms helps ensure that users receive personalized and relevant content swiftly and effectively. This was clearly demonstrated by its results in the MLCommons DLRM-v2 Offline benchmark test. The HPE ProLiant Compute DL380a Gen12 was the only server to submit results with both NVIDIA H100 and L40S GPUs and also achieved the No. 1 spot, surpassing the competition by an astounding 57%.[8, 10]

Conclusion

The MLPerf benchmark results highlight the remarkable capabilities of the HPE ProLiant Compute DL380a Gen12, showcasing its outstanding performance and adaptability across a wide spectrum of AI and machine learning workloads. This achievement not only reinforces the system's ability to tackle complex and resource-intensive tasks but also positions it as a reliable and future-ready solution for organizations navigating the rapidly advancing AI landscape. By delivering consistent results across diverse use cases, the HPE ProLiant Compute DL380a Gen12 proves itself as an essential tool for meeting the ever-evolving demands of the AI-driven era.

[1] MLPerf Inference: Datacenter v5.0 results retrieved from mlcommons.org/benchmarks/inference-datacenter/ as of April 2025. See mlcommons.org for more information. All results verified by MLCommons Association.

[2] In AI and LLMs, a token is a unit of text the model processes. It can be as short as one character (like "a") or as long as one word (like "apple"). Most words are 1–2 tokens long. Models such as ChatGPT and Llama2-70B break input and output text into tokens to understand, generate, and benchmark performance.

[3] Benchmark Suite Results: MLPerf Inference: Datacenter, Submission ID 5.0-0046.

[4] Benchmark Suite Results: MLPerf Inference: Datacenter, Submission ID 5.0-0018.

[5] Benchmark Suite Results: MLPerf Inference: Datacenter, Submission ID 5.0-0046.

[6] Benchmark Suite Results: MLPerf Inference: Datacenter, Submission ID 5.0-0043.

[7] Benchmark Suite Results: MLPerf Inference: Datacenter, Submission ID 5.0-0017.

[8] Benchmark Suite Results: MLPerf Inference: Datacenter, Submission ID 5.0-0004.

[9] Benchmark Suite Results: MLPerf Inference: Datacenter, Submission ID 5.0-0071.

[10] Benchmark Suite Results: MLPerf Inference: Datacenter, Submission ID 5.0-0003.

Learn more at

HPE ProLiant Compute DL380a Gen12

Visit HPE.com

Chat now

a00147573ENW, Rev. 1

HEWLETT PACKARD ENTERPRISE

hpe.com