HPE ProLiant Compute

HPE ProLiant Compute Gen12 achieves multiple world records in AI inference benchmarks

Achieving #1 rankings across more than 23 MLPerf™ Benchmark models for HPE ProLiant Compute DL384 Gen12 and HPE ProLiant Compute DL380a Gen12 servers

HPE ProLiant Compute DL384 Gen12 delivered 13 new world records on MLPerf Inference: Datacenter v5.0[1]

#1

Stable Diffusion XL (SDXL)

Advanced image generation model that produces high-quality, detailed images from text descriptions

#1

Llama2-70B-99

High-precision variant of Llama 2-70B, optimized for enhanced accuracy and reliability in AI reasoning tasks

#1

Llama2-70B-99.9

Ultra-high-precision version of Llama 2-70B, designed for near-perfect consistency and minimal errors in complex AI applications

#1

Mixtral-8x7B

High-quality sparse mixture of experts model known for its exceptional performance and fast inference speeds

MLPerf Inference: Datacenter v5.0 benchmark results on HPE ProLiant Compute DL384 Gen12 server[1]

#1

Superior AI power: High performance server with the NVIDIA® GH200 NVL2 accelerator[2]

Benchmark tests	Server[3]	Offline[4]
Stable Diffusion XL (SDXL)	4.34	5.02
Llama2-70b–99	8674.58	9362.85
Llama2-70b–99.9	8674.58	9362.85
Mixtral-8x7B	15570.80	16703.40

In a groundbreaking first for our MLPerf results, the HPE ProLiant Compute DL384 Gen12 server with dual-socket NVIDIA GH200 Grace Hopper™ Superchip 144 GB delivers 2x the performance of our single-socket setup with linear scaling.

HPE ProLiant Compute DL384 Gen12 server delivers low-latency datacenter inference with scalable 1P and 2P systems, nearly doubling AI performance in most scenarios.

HPE ProLiant Compute DL380a Gen12 Server top performer on four benchmarks[1]

#1

Image classification[5]

ResNet50 Server and Offline benchmarks

#1

Object detection[6]

Retinanet Server benchmark

#1

LLM summarization[7]

GPT-J 99 Server benchmark

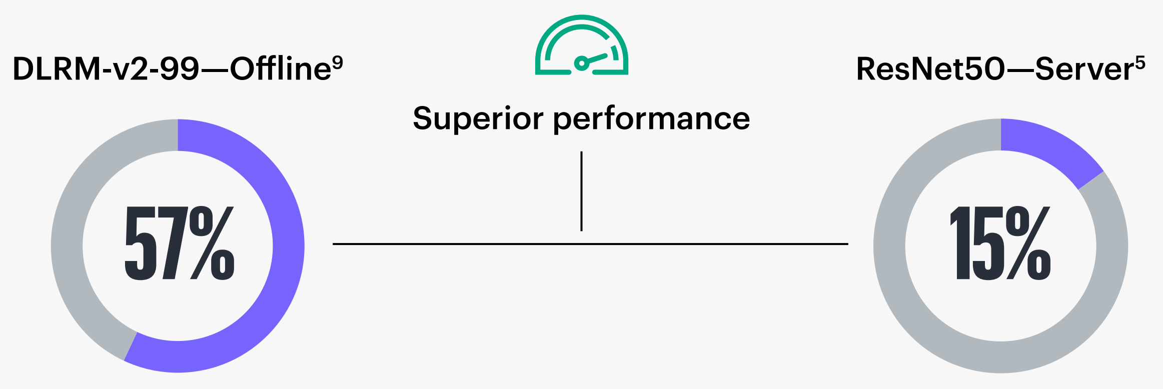
#1

LLM summarization[8]

GPT-J 99.9 Server benchmark

#1

Superior performance over other vendors[1] HPE ProLiant Compute DL380a Gen12 server



Better than the next top-performing server with 141 GB GPUs

[1] MLPerf Inference: Datacenter v5.0 as of April 2, 2025. Retrieved from mlcommons.org/benchmarks/inference-datacenter. See mlcommons.org for more information. Results verified by MLCommons™ Association.

[2] Based on results for NVIDIA GH200 NVL Grace Hopper Superchip with 144 GB HBM3E memory compared with all other GH200 systems (Submission ID 5.0-0038).

[3] Server: Scenario representing low-latency inference applications. Queries per second (SDXL); tokens per second (Llama2-70b-99, Llama2-70b-99.9, and Mixtral-8x7b).

[4] Offline: Scenario representing high-batch size inference applications. Samples per second (SDXL); tokens per second (Mixtral-8x7b, Llama2-70b-99 and Llama2-70b-99.9).

[5] MLPerf Inference: Datacenter v5.0 ResNet50 Server and Offline benchmarks based on HPE ProLiant DL380a Gen12 Server utilizing Intel® Xeon® 6740E processors and eight NVIDIA H200-NVL-141GB GPUs (submission ID 5.0-0043).

[6] MLPerf Inference: Datacenter v5.0 Retinanet Server benchmark based on HPE ProLiant DL380a Gen12 Server utilizing Intel Xeon 6740E processors and four NVIDIA L40S GPUs (submission ID 5.0-0045).

[7] MLPerf Inference: Datacenter v5.0 GPT-J-99 Server benchmark based on HPE ProLiant DL380a Gen12 Server utilizing Intel Xeon 6740E processors and four NVIDIA L40S GPUs (submission ID 5.0-0045).

[8] MLPerf Inference: Datacenter v5.0 GPT-J-99.9 Server benchmark based on HPE ProLiant DL380a Gen12 Server utilizing Intel Xeon 6740E processors and four NVIDIA L40S GPUs (submission ID 5.0-0045).

[9] MLPerf Inference: Datacenter v5.0 DLRM-v2-99 Offline benchmark based on HPE ProLiant DL380a Gen12 Server utilizing Intel Xeon 6740E processors and eight NVIDIA H200-NVL-141GB GPUs (submission ID 5.0-0043).

Visit HPE.com

Learn more at

HPE.com/ProLiant

Chat now

a00146614ENW, Rev. 1

HEWLETT PACKARD ENTERPRISE

hpe.com

HPE