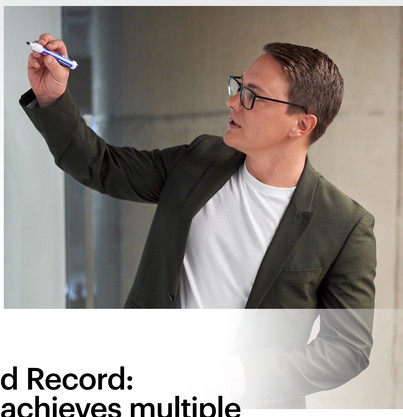

World Record: HPE achieves multiple #1 performance results for AI inference benchmarks

#1 on more than 10 MLPerf Benchmark models — HPE ProLiant Compute DL384 Gen12 and HPE ProLiant DL380a Gen11 servers

As part of MLCommons™, an AI engineering consortium, MLPerf™ Inference: Data center benchmarks set industry-wide standards for fairly assessing and evaluating diverse AI/ML performance across different hardware platforms.[1]

HPE ProLiant Compute DL384 Gen12 Server — outstanding performance per GPU on MLPerf Inference: Data center benchmark

#1

Stable Diffusion XL (SDXL)[3]

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

#1

DLRM-v2-99[2, 3]

A deep learning recommendation model (DLRM) designed for high-accuracy prediction tasks

#1

DLRM-v2-99.9[2, 3]

An advanced DLRM optimized to achieve 99.9% accuracy in prediction tasks

#1

Mixtral-8x7B[2, 3]

An efficient AI model that outperforms larger models such as Llama 2 70B using fewer parameters

MLPerf Inference: Data center v4.1 results on HPE ProLiant Compute DL384 Gen12 Server[4]

#1

Best server performance with a single accelerator

Benchmark tests	Server[2]	Offline[3]
SDXL		2.31
Mixtral-8x7b	7450.72	8063.02
DLRM-v2-99	81,009.60	87,052.70
DLRM-v2-99.9	51,014.20	53,611.90

HPE ProLiant Compute DL384 Gen12 Server is an ideal solution for low-latency data center inference.

Hewlett Packard Enterprise is the first to submit performance results with the NVIDIA® GH200 NVL with 144GB HBM3e memory.

HPE ProLiant DL380a Gen11 Server top performer on 4 benchmarks

#1

Image classification[1, 5]

Resnet50 Server benchmark

#1

Object detection[1, 6]

Retinanet Server benchmark

#1

Speech-to-text[1, 7]

RNNT Server benchmark

#1

Large Language Model[1, 8]

Llama 2 70B benchmark

#1

LLM inference on HPE ProLiant DL380a Gen11 Server

Superior performance

33%

better than the next top-performing server with 94 GB GPUs[1, 8]

44%

better than the next top-performing server with 80 GB GPUs[1, 9]

[1] MLPerf™ Inference: Data center v4.1 and v4.0 as of August 28, 2024. Retrieved from mlcommons.org/benchmarks/inference-datacenter/. See mlcommons.org for more information. Results verified by MLCommons Association.

[2] Server: Scenario representing low-latency inference applications. Mixtral-8x7b: tokens per second; DLRM-v2-99 and DLRM-v2-99.9: queries per second.

[3] Offline: Scenario representing high-batch size inference applications. SDXL: Samples per second. Mixtral-8x7b: tokens per second; DLRM-v2-99 and DLRM-v2-99.9: samples per second.

[4] Based on results for NVIDIA GH200 NVL Grace Hopper Superchip with 144 GB HBM3e memory compared with all other GH200 systems (Submission ID 4.1-0072).

[5] MLPerf Inference: Data center v4.1 Resnet50 Server benchmark based on HPE ProLiant DL380a Gen11 Server utilizing Intel® Xeon® Gold 6530 processors and four NVIDIA H100-NVL-94GB GPUs (Submission ID 4.1-0032).

[6] MLPerf Inference: Data center v4.1 Retinanet Server benchmark based on HPE ProLiant DL380a Gen11 Server utilizing Intel Xeon Gold 6530 processors and four NVIDIA H100-NVL-94GB GPUs (Submission ID 4.1-0032).

[7] MLPerf Inference: Data center v4.0 RNNT Server benchmark based on HPE ProLiant DL380a Gen11 Server utilizing Intel® Xeon® Platinum 8468 processors and four NVIDIA H100-PCIe-80GB GPUs (Submission ID 4.0-0048).

[8] MLPerf Inference: Data center v4.1 Llama 2 70B benchmark based on HPE ProLiant DL380a Gen11 Server utilizing Intel Xeon Gold 6530 processors and four NVIDIA H100-NVL-94 GB GPUs (Submission ID 4.1-0032).

[9] MLPerf Inference: Data center v4.0 Llama 2 70B benchmark based on HPE ProLiant DL380a Gen11 Server utilizing Intel Xeon Platinum 8468 and four NVIDIA H100-PCIe-80GB GPUs (Submission ID 4.0-0048).

Visit HPE.com

Learn more at

HPE ProLiant Compute DL384 Gen12

HPE ProLiant Compute DL380a Gen11

Chat now

a00142423ENW, Rev. 1

HEWLETT PACKARD ENTERPRISE

hpe.com