



# HPE ProLiant Compute servers: Dominate AI inference benchmarks—Again!



Eight #1 rankings across MLPerf benchmarks, including a new leader: HPE ProLiant DL385 Gen11<sup>1</sup>

HPE ProLiant Compute DL380a Gen12—outstanding performance per GPU on MLPerf™ Datacenter v5.1

#1



DLRM-v2-99—Server → 65,021 queries/second per GPU<sup>2</sup>

#1



DLRM-v2-99.9—Server → 41,357 queries/second per GPU<sup>2</sup>

#1

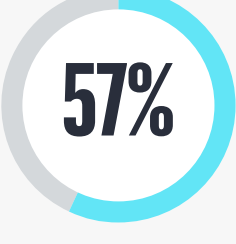
Benchmark champion

HPE ProLiant Compute DL380a Gen12 dominates DLRM and Llama benchmarks again!

DLRM-v2-99

MLPerf Datacenter v5.0

Offline



better than the next best submission<sup>3</sup>

MLPerf Server v5.1

Server



better than the next best submission<sup>4</sup>

Llama

MLPerf Datacenter v5.0, Llama2-70b-99 and Llama2-70b-99.9—Offline

#1

with 3,655.89 tokens/sec<sup>5</sup>

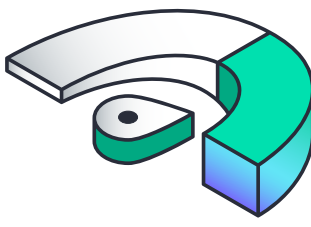
MLPerf Datacenter v5.1, Llama3.1-8b—Server

#1

with 46,060.00 tokens/sec<sup>6</sup>

## NEW WHISPER LLM BENCHMARK

HPE ProLiant DL385 Gen11 server with NVIDIA® H200 NVL 141 GB GPUs achieves #1 spot for performance with 3,962.78 samples/second/GPU<sup>7</sup>



HPE ProLiant Compute DL380a Gen12 with NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs

#1

First and only OEM, supplier or vendor to submit results<sup>8</sup>

Benchmark tests	Server (queries/sec)	Offline (samples/sec)
Llama2-70B-99	26,005.90	26,205.80
Llama2-70B-99.9	26,001.00	26,205.30
Llama3.1-8B	46,060.00	46,841.80
Mixtral-8x7B	30,143.30	32,738.80

HPE was the **only vendor** to submit the following additional results:

- DLRM-v2-99—the only server utilizing the NVIDIA GH200 NVL2 accelerator (in [HPE ProLiant Compute DL384 Gen12](#))—161,030 queries per second (Server) and 174,456 samples per second<sup>9</sup>
- MLPerf Edge v5.1 RetinaNet—[HPE ProLiant ML30 Gen11](#) utilizing a single NVIDIA RTX 4000 Ada 20GB GPU—258.352 Samples/second (Offline) and 29.96 ms (Multistream)<sup>10</sup>



<sup>1</sup> MLPerf Inference: Datacenter v5.1 as of September 9, 2025. Retrieved from [mlcommons.org/benchmarks/inference-datacenter/](https://mlcommons.org/benchmarks/inference-datacenter/). See [mlcommons.org](https://mlcommons.org) for more information. Results verified by MLCommons™ Association.

<sup>2</sup> MLPerf Inference: Datacenter v5.1 DLRM-v2-99 and DLRM-v2-99.9 Server benchmarks based on HPE ProLiant DL380a Gen12 Server utilizing Intel® Xeon® 6787P processors and ten NVIDIA H200-NVL-141GB GPUs (submission ID 5.1-0050).

<sup>3</sup> “[HPE ProLiant Compute Gen12 achieves multiple world records in AI inference benchmarks](#),” HPE, 2025.

<sup>4</sup> MLPerf Inference: Datacenter v5.1 DLRM-v2-99 Server benchmark based on systems utilizing Intel Xeon 6787P processors and NVIDIA H200-NVL-141GB GPUs (submission IDs 5.1-0050 and 5.1-0080).

<sup>5</sup> MLPerf Inference: Datacenter v5.0 Llama2-70B-99 Offline and llama2-70b-99.9 Offline benchmarks (submission IDs 5.0-0018 and 5.0-0046).

<sup>6</sup> MLPerf Inference: Datacenter v5.1 Llama 3.1-8b Server benchmark (submission IDs 5.1-0011 and 5.1-0051).

<sup>7</sup> MLPerf Inference: Datacenter v5.1 (submission IDs 5.1-0023 and 5.1-0052). HPE ProLiant DL385 Gen11 delivered 7,925.55 samples/second with 2 NVIDIA H200-NVL-141GB GPUs in the Whisper benchmark. This results in 3,962.78 samples/second per GPU.

<sup>8</sup> MLPerf Inference: Datacenter v5.1 (submission ID 5.1-0051).

<sup>9</sup> MLPerf Inference: Datacenter v5.1 (submission ID 5.1-0053).

<sup>10</sup> MLPerf Inference: Edge v5.1 (submission ID 5.1-0054)

Visit [HPE.com](https://www.hpe.com)

Learn more at

[HPE ProLiant Compute DL380a Gen12](#)

[HPE ProLiant DL385 Gen11](#)

[HPE ProLiant ML30 Gen11](#)

[HPE ProLiant Compute DL384 Gen12](#)

[Chat now](#)

© Copyright 2025 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

Intel Xeon is the trademark of Intel Corporation or its subsidiaries in the U.S. and/or other countries. NVIDIA and NVIDIA RTX are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. MLPerf™ and MLCOMMONS™ are trademarks and service marks of MLCommons Association in the United States and other countries. All third-party marks are property of their respective owners.

a0015117ENW

HEWLETT PACKARD ENTERPRISE

[hpe.com](https://www.hpe.com)