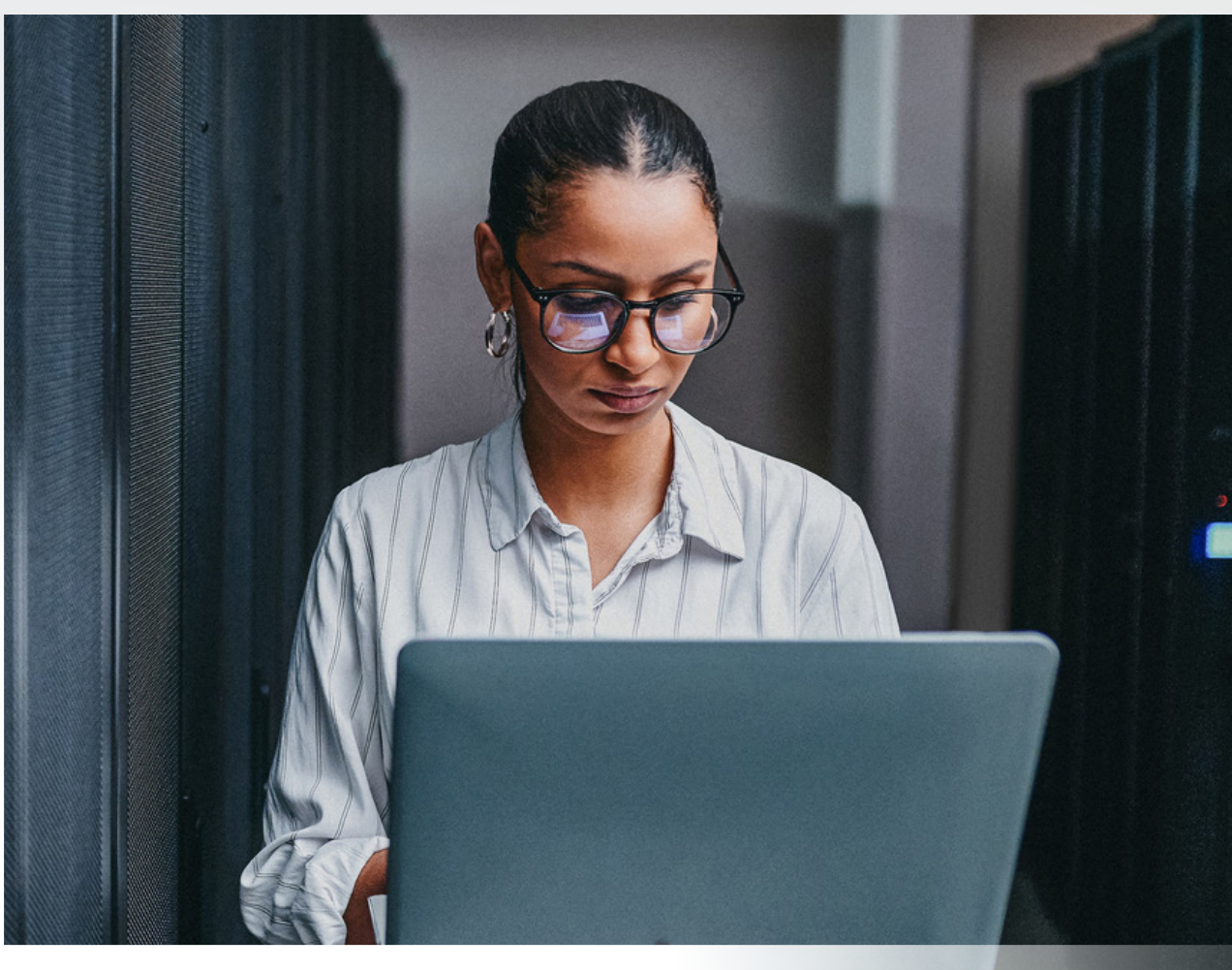

HPE domina la classifica: 13 posizioni al vertice per i server HPE ProLiant nei benchmark di inferenza dell'AI

Le prime submission dai server AI HPE per gli ambienti edge dominano i benchmark AI incentrati sui modelli Llama e i prodotti HPE ProLiant DL385 Gen11 dimostrano ottime prestazioni energetiche

La suite di benchmark MLPerf™ Inference: Datacenter, del consorzio ingegneristico MLCommons™ dedicato all'AI, definisce gli standard a livello di settore per una valutazione imparziale di varie prestazioni dell'AI e del machine learning su diverse piattaforme hardware.[1]

HPE ProLiant Compute DL380a Gen12 con otto GPU NVIDIA RTX PRO™ 6000 Blackwell Server Edition

Otto

Primi posti nei benchmark Llama[2]

29%

Superiorità rispetto al secondo classificato nel benchmark Llama3.1-8B Interactive[3]

Prestazioni superiori a quelle di Dell PowerEdge XE7745 in sei benchmark Llama[4]

N.1

Campione di benchmark: HPE ProLiant Compute DL380a Gen12 domina ancora una volta le classifiche dei benchmark Llama3.1-8B per i server

— MLPerf™ Datacenter v5.1 - Primo posto con 46.060,00 token/sec[5]

— MLPerf Datacenter v6.0 - Primo posto con 48.794,30 token/sec[6]

HPE ProLiant DL385 Gen11 con quattro GPU NVIDIA RTX PRO 6000 Blackwell Server Edition

Quattro

Primi posti nei benchmark Llama[7]

24%

Superiorità rispetto al secondo classificato nel benchmark Llama3.1-8B Interactive[8]

N.1

Primo posto nel benchmark Whisper per 2 anni consecutivi[9, 10]

Prestazioni energetiche

I server HPE ProLiant DL385 Gen11 hanno prodotto risultati quasi identici a quelli di Cisco UCS X215c M8, consumando il 25% in meno per le GPU[11]

Massime prestazioni nel benchmark Whisper

I modelli DL380a Gen12 e DL385 Gen11 hanno conquistato il primo posto nella conversione del parlato in testo scritto[9, 12]

Submission di HPE per la versione 6.0 dei test di benchmark

12

con dieci GPU NVIDIA RTX PRO 6000 Blackwell Server Edition[13]

4

con le GPU NVIDIA RTX PRO 4500 Blackwell Server Edition[14]

— Prime submission presentate dai server DL145 Gen11 per i test di benchmark MLPerf

2

per il modello di chatbot AI con 120 miliardi di parametri[15]

[1] MLPerf Inference: Datacenter v6.0 as of April 1, 2026. Retrieved from MLPerf Inference: Datacenter. See mlcommons.org for more information. Results verified by MLCommons Association.

[2] MLPerf Inference: Datacenter v6.0 Llama3.1-8B (Server, Offline, Interactive), Llama2-70B-99 (Server, Offline), and Llama2-70B-99.9 (Server, Offline) based on HPE ProLiant Compute DL380a Gen12 server utilizing Intel® Xeon® 6 processors and eight NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs (Submission ID 6.0-0047).

[3] 44.087,30 queries/sec (submission ID 6.0-0047) vs. 34.241 queries/sec (submission ID 6.0-0004), Retrieved from MLPerf Inference: Datacenter.

[4] MLPerf Inference: Datacenter v6.0 Llama3.1-8B, Llama2-70B-99, and Llama2-70B-99.9 (Server, Offline) (Submission IDs 6.0-6.0-0023 and 6.0-0047).

[5] MLPerf Inference: Datacenter v5.1 Llama3.1-8B Server benchmark (submission IDs 5.1-0011 and 5.1-0051).

[6] MLPerf Inference: Datacenter v6.0 Llama3.1-8B Server benchmark (submission IDs 6.0-0047, 6.0-0004, 6.0-0005, and 6.0-0023).

[7] MLPerf Inference: Datacenter v6.0 Llama3.1-8B (Offline, Interactive) and Llama2-70B-99.9 (Server, Offline) benchmarks based on a per-GPU basis of HPE ProLiant DL385 Gen11 server utilizing AMD EPYC processors and two NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs (submission ID 6.0-0052).

[8] MLPerf Inference: Datacenter v6.0 Llama3.1-8B Interactive benchmark based on per-GPU basis, TensorRT 10,14, AMD EPYC processors, and NVIDIA RTX PRO Blackwell Server Edition GPUs (600W) (submission IDs 6.0-0052 and 6.0-0004).

[9] MLPerf Inference: Datacenter v6.0 whisper benchmark based on a per-GPU basis and AMD EPYC processors (submission IDs 6.0-0052, 6.0-0004, 6.0-0023, 6.0-0014, and 6.0-0054).

[10] "HPE ProLiant Compute servers: Dominate AI inference benchmarks—Again!," HPE, 2025.

[11] MLPerf Inference: Datacenter v6.0 based on AMD EPYC processors (submission IDs 6.0-0014 and 6.0-0054).

[12] MLPerf Inference: Datacenter v6.0 whisper benchmark based on eight NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs (submission ID 6.0-0047).

[13] MLPerf Inference: Datacenter v6.0 (submission IDs 6.0-0044 and 6.0-0045).

[14] MLPerf Inference: Datacenter v6.0 (submission IDs 6.0-0043 and 6.0-0050).

[15] MLPerf Inference: Datacenter v6.0 gpt-oss-120b benchmark (Server, Offline) with AMD EPYC processors (submission IDs 6.0-0054).

Ulteriori informazioni alla pagina

HPE ProLiant Compute DL380a Gen12

HPE ProLiant DL385 Gen11

HPE ProLiant DL145 Gen11

Visita HPE.com

Avvia chat

© Copyright 2026 Hewlett Packard Enterprise Development LP. Le informazioni contenute nel presente documento sono soggette a modifica senza preavviso. Le uniche garanzie per i prodotti e i servizi Hewlett Packard Enterprise sono quelle espressamente indicate nelle dichiarazioni di garanzia esplicite che accompagnano tali prodotti e servizi. Nulla di quanto contenuto nel presente documento potrà essere interpretato come garanzia supplementare. Hewlett Packard Enterprise declina ogni responsabilità per eventuali omissioni ed errori tecnici o editoriali contenuti nel presente documento.

AMD è un marchio di Advanced Micro Devices, Inc. Intel Xeon è un marchio di Intel Corporation o di società controllate da Intel negli Stati Uniti e/o in altri Paesi. NVIDIA RTX, e TensorRT sono marchi e/o marchi registrati di NVIDIA Corporation negli Stati Uniti e in altri Paesi. MLPERF™ e MLCOMMONS™ sono marchi e marchi di servizio di MLCommons Association negli Stati Uniti e in altri Paesi. Tutti i marchi di terzi sono di proprietà dei rispettivi titolari.

a00158089ITE

HEWLETT PACKARD ENTERPRISE

hpe.com