HPE ProLiant Compute Gen12 erzielt mehrere Weltrekorde bei KI-Inferenz-Benchmarks

Platzierung als Nr. 1 bei mehr als 23 MLPerf™ Benchmark-Modellen für HPE ProLiant Compute DL384 Gen12 und HPE ProLiant Compute DL380a Gen12 Server

HPE ProLiant Compute DL384 Gen12 erzielte 13 neue Weltrekorde bei MLPerf Inference: Datacenter v5.0[1]

Nr. 1

Stable Diffusion XL (SDXL)

Erweiterte Bildgenerierungsmodell, das qualitativ hochwertige, detaillierte Bilder aus Textbeschreibungen erstellt

Nr. 1

Lama2-70B-99

Hochpräzise Variante von Lama 2-70B, die für verbesserte Genauigkeit und Zuverlässigkeit in KI-Argumentationsaufgaben optimiert wurde

Nr. 1

Lama2-70B-99.9

Ultrapräzise Version von Llama 2-70B, die für nahezu perfekte Konsistenz und minimale Fehler in komplexen KI-Anwendungen entwickelt wurde

1.

Mixtral-8x7b

Hochwertiges sparsames Mixture-of-Experts-Modell, bekannt für seine außergewöhnliche Leistung und schnelle Inferenzgeschwindigkeiten

MLPerf Inference: Datacenter v5.0-Ergebnisse auf dem HPE ProLiant Compute DL384 Gen12-Server[1]

Nr. 1

Überlegene KI-Leistung: Hochleistungsserver mit NVIDIA® GH200 NVL2 Beschleuniger[2]

Benchmark-Tests	Server[3]	Offline[4]
Stable Diffusion XL (SDXL)	4,34	5,02
Lama2-70b-99	8674,58	9362,85
Llama2-70b–99.9	8674,58	9362,85
Mixtral-8x7B	15570,80	16703,40

In einer bahnbrechenden Premiere für unsere MLPerf-Ergebnisse liefert der HPE ProLiant Compute DL384 Gen12 Server mit Dual-Socket NVIDIA GH200 Grace Hopper™ Superchip 144 GB die doppelte Leistung unseres Single-Socket-Setups mit linearer Skalierung.

Der HPE ProLiant Compute DL384 Gen12 Server bietet Rechenzentrumsinferenz mit geringer Latenz und skalierbaren 1P- und 2P-Systemen und verdoppelt in den meisten Szenarien die KI-Leistung nahezu.

HPE ProLiant Compute DL380a Gen12 Server ist Spitzenreiter bei vier Benchmarks[1]

Nr. 1

Bildklassifizierung[5]

ResNet50 Server und Offline Benchmarks

Nr. 1

Objekterkennung[6]

Retinanet-Server-Benchmark

Nr. 1

LLM-Zusammenfassung[7]

GPT-J 99 Server-Benchmark

Nr. 1

LLM-Zusammenfassung[8]

GPT-J 99.9 Server-Benchmark

Nr. 1

Überlegene Leistung im Vergleich zu anderen Anbietern[1]
HPE ProLiant Compute DL380a Gen12 Server

DLRM-v2-99 – Offline[9]

57%

Überragende Leistung

ResNet50 – Server[5]

15%

Besser als der Server mit der nächstbesten Leistung und 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).

HPE.com besuchen

Weitere Informationen finden Sie unter

HPE.com/de/de/hpe-proliant-servers

Jetzt chatten

© Copyright 2025 Hewlett Packard Enterprise Development LP. Die hier enthaltenen Informationen können jederzeit ohne vorherige Ankündigung geändert werden. Neben der gesetzlichen Gewährleistung gilt für Produkte und Services von Hewlett Packard Enterprise (HPE) ausschließlich die Herstellergarantie, die in den Garantieerklärungen für die jeweiligen Produkte und Services explizit genannt wird. Aus dem vorliegenden Dokument sind keine weiterreichenden Garantieansprüche abzuleiten. Hewlett Packard Enterprise haftet nicht für technische oder redaktionelle Fehler oder Auslassungen.

Intel Xeon ist eine Handelsmarke der Intel Corporation oder ihrer Tochtergesellschaften in den USA und/oder anderen Ländern. NVIDIA ist eine Marke und/oder eingetragene Marke der NVIDIA Corporation in den USA und anderen Ländern. MLCOMMONS™ und MLPERF™ sind Marken und Dienstleistungsmarken der MLCommons Association in den USA und anderen Ländern. Alle Drittanbietermarken sind Eigentum ihrer jeweiligen Inhaber.

a00146614DEE, Rev. 1

HEWLETT PACKARD ENTERPRISE

hpe.com