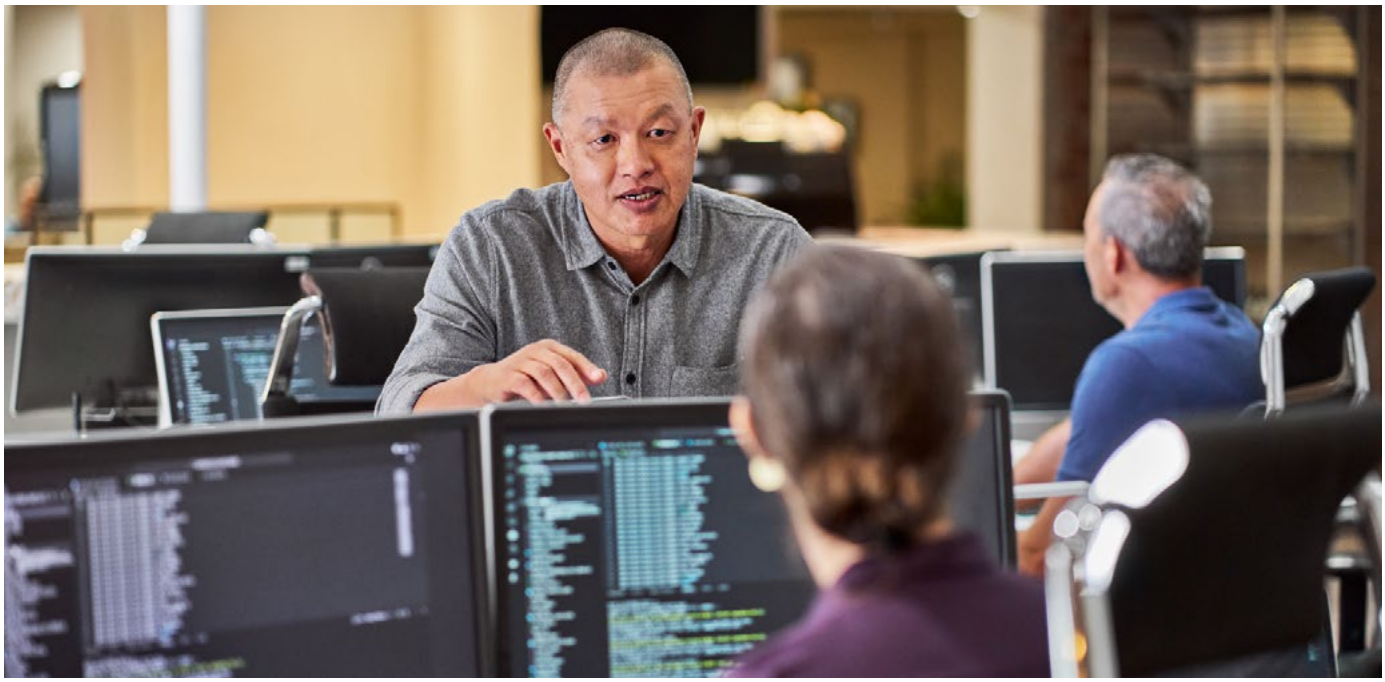
Lösungsübersicht

ULTRASKALIERBARE GPU-BESCHLEUNIGUNG FÜR UNTERNEHMENS-KI

HPE ProLiant Compute DL380a Gen12 mit NVIDIA H200 NVL Tensor Core GPUs, Teil des Portfolios von NVIDIA AI Computing by HPE



NVIDIA GPU-Unterstützung

Der HPE ProLiant Compute DL380a Gen12 fördert die schnelle Einführung generativer KI und unterstützt die beiden folgenden GPUs:

— **NVIDIA RTX PRO™ 6000**
Die NVIDIA RTX PRO™ 6000 Blackwell Server Edition ist die ultimative Rechenzentrums-GPU für KI und Visual Computing. Sie liefert bahnbrechende Beschleunigung für anspruchsvollste Unternehmens-Workloads, von agentenbasierter und physikalischer KI bis zu wissenschaftlichem Computing, Grafiken und Videoanwendungen. Sie wurde für Workloads optimiert, die die Rechendichte und den Maßstab erfordern, die bei einer Bereitstellung im Rechenzentrum gegeben sind, und bietet ein passiv gekühltes Thermal Design sowie 96 GB ultraschnellen GDDR7-Arbeitsspeicher.

— **NVIDIA® H200 NVL**
Die NVIDIA H200 Tensor Core GPU optimiert generative KI- und High-Performance-Computing-(HPC-)Workloads mit bahnbrechender Leistung und Arbeitsspeicherkapazität. Als erste GPU mit HBM3e unterstützt der größere und schnellere Arbeitsspeicher der H200 die schnelle Einführung generativer KI und LLMs und optimiert das wissenschaftliche Computing für HPC-Workloads.

KI im Unternehmensmaßstab nutzen

Es ist zu erwarten, dass das Aufkommen der künstlichen Intelligenz (KI) eine jede Branche grundlegend verändern wird. Unternehmen benötigen spezielle Umgebungen, in denen KI-Anwendungen in großem Umfang entwickelt, bereitgestellt und verwaltet werden. Dies stellt einen grundlegenden Wandel in der Art und Weise dar, wie diese Technologien in Unternehmen konzipiert, entwickelt und eingesetzt werden. Mit einem unternehmensweiten Ansatz für die Entwicklung und Verwaltung von KI können Unternehmen neue Möglichkeiten erschließen. Der Einsatz von KI in großem Maßstab erfordert jedoch einige Überlegungen zur Sicherstellung eines optimalen Geschäftswerts.

Skalieren Sie Rechenressourcen passend zur Phase des Lebenszyklus Ihrer KI

Logischerweise können die Rechenleistungsanforderungen in den verschiedenen Phasen des KI-Lebenszyklus – wie Modelltraining, Feinabstimmung und Inferenz – erheblich variieren. Beim Modelltraining geht es darum, ein KI-Modell von Grund auf neu zu erstellen, was umfangreiche Daten- und Rechenressourcen erfordert. Die Feinabstimmung hingegen erfordert weniger Ressourcen, da sie bereits bestehende Modelle an spezifische Aufgaben und Anforderungen anpasst. Bei KI-Inferenzen schließlich kommen die KI-Modelle zum Einsatz und werden genutzt, um wertvolle Erkenntnisse zu gewinnen. Für Inferenzen werden weniger Rechenressourcen benötigt, aber Unternehmen, die sie einsetzen möchten, müssen die Größe des Modells, die Reaktionszeit und die Anzahl der gleichzeitigen Benutzer berücksichtigen, um eine optimale Unternehmensleistung zu gewährleisten.

Unternehmen können die Leistung von Large Language Models (LLMs) mit Retrieval Augmented Generation (RAG) verbessern, indem sie eine Informationsabrufkomponente mit Textgenerierungsfunktionen kombinieren. Dadurch werden relevante Daten abgerufen, die zusätzlichen Kontext für ein besseres Verständnis der Eingaben und eine präzisere Antwortgenerierung liefern. Durch die Trennung der Abruf- und Generierungskomponenten kann RAG bei großen Datensätzen und komplexen Abfragen effektiv skaliert werden.

HPE ProLiant Compute DL380a Gen12 – Ultraskalierbare GPU-Beschleunigung für optimale Feinabstimmung und Inferenzleistung

Für Unternehmen, die die beste Leistung bei der Feinabstimmung und Inferenz mit RAG zur Verbesserung der Leistung von LLMs bieten möchten, eignet sich die neueste Ergänzung der HPE ProLiant-Serverfamilie besonders gut, um die Leistung von GenAI für Unternehmen nutzbar zu machen.

Der HPE ProLiant Compute DL380a Gen12, Teil des NVIDIA AI Computing by HPE Portfolios, verfügt über eine ultraskalierbare Architektur, um KI-Leistung der nächsten Generation für Ihre Unternehmensanforderungen bereitzustellen. Mit bis zu zwei Intel® Xeon® 6-Prozessoren und bis zu acht NVIDIA® H200 NVL Tensor Core GPUs – GPUs, die die weltweit leistungsstärkste GPU-Architektur nutzen, um KI- und HPC-Workloads (High Performance Computing) zu beschleunigen – bietet der HPE ProLiant Compute DL380a Gen12:

— **Beispiellose Leistung** mit 18-facher Leistung im Vergleich zum A100 und bis zu 2-facher Inferenzgeschwindigkeit im Vergleich zum NVIDIA H100 NVL für Inferenzen

— **Unternehmenstaugliche Zuverlässigkeit** mit fortschrittlichem Energiemanagement, das sechs dedizierte und redundante Stromversorgungen für die GPUs bereitstellt und so für effizientere und zuverlässigere Leistung sorgt

— **Branchenführende Sicherheitsinnovationen** zum Schutz Ihrer Infrastruktur, Workloads und Daten



HPE ProLiant Compute DL380a Gen12

Verbessern Sie die Kerndichte und Leistung pro Watt und sorgen Sie gleichzeitig für einen hohen Durchsatz.

Intel Xeon 6-Prozessoren bieten eine neue Klasse von Efficient Cores (E-Cores) mit hoher Kerndichte und deutlichen Vorteilen für Cloud-Workloads. Durch den Einsatz dieser Intel® Prozessoren der neuesten Generation werden Ihre Energiekosten gesenkt, die Nachhaltigkeit gefördert und die Rack-Dichte verbessert. So können Sie mehr aus Ihrer Rechenzentrumsinfrastruktur herausholen – und gleichzeitig die Kapazität für neue Workloads erhöhen. Integrierte Beschleuniger steigern die Leistung und Effizienz bei der Ausführung bestimmter Workloads zusätzlich.

Erweiterte Verwaltung und Überwachung

Zur Vereinfachung der Verwaltung des HPE ProLiant Compute DL380a Gen12 sind erweiterte Verwaltungs- und Überwachungsfunktionen enthalten:

— Mit **HPE iLO 7** können Sie Ihre Server von jedem Ort der Welt aus sicher anzeigen, konfigurieren und aktualisieren.

— **HPE OneView** Infrastrukturmanagement arbeitet innerhalb des Rechenzentrums, um den IT-Betrieb über Server, Speicher und Netzwerke hinweg zu automatisieren und zu optimieren.

— **HPE Compute Ops Management** bietet ein intuitives Cloud-Betriebserlebnis, mit dem Sie Ihre verteilte Rechenumgebung nahtlos überwachen und verwalten können – und das unabhängig davon, wo sich Ihre Recheninfrastruktur befindet.



Beschleunigen Sie die KI-Produktion mit HPE Private Cloud AI

Beschleunigen Sie den KI-Erfolg mit HPE Private Cloud AI, der branchenweit ersten schlüsselfertigen Full-Stack-Private-Cloud für KI, die zum NVIDIA AI Computing by HPE Portfolio von HPE gehört. Sie bietet KI- und IT-Teams leistungsstarke Tools zum Experimentieren und Operationalisieren von KI, wobei Ihre Daten privat und sicher bleiben, und nutzt marktübliche Software-Tools von NVIDIA, HPE und Open-Source.

HPE Private Cloud AI wird auf der HPE GreenLake-Cloud bereitgestellt und basiert auf validierten Designs, die durch KI-optimierte Rechenleistung, Speicherung und Netzwerktechnik von HPE und NVIDIA unterstützt werden. Sie können mit einem einzelnen, auf einem kleinen Modell basierenden Inferenzpilotprojekt beginnen und dieses auf vielfältige Anwendungsfälle, höheren Durchsatz sowie RAG- oder LLM-Feinabstimmung in einer einzigen Lösung hochskalieren. Erweitern Sie einfach Ihre Infrastruktur ohne neue Software, ohne Integrationsaufwand und mit geringerem Bedarf an Fachwissen.

HPE Private Cloud AI bietet alles, was Unternehmen an der Cloud schätzen: Self-Service, moderne Entwicklungstools, schnelle Skalierbarkeit und günstige Abonnements – alles in Ihrer privaten Umgebung. Sie können klein anfangen und die Technologie und Investitionen nahtlos skalieren, wenn sich Ihre Anwendungsfälle weiterentwickeln. Und mit unseren fachkundigen Dienstleistungen können wir Ihnen Hinweise geben, wo Sie ansetzen sollten.

HPE.com besuchen

Weitere Informationen

HPE.com/ProLiant/DL380a-gen12

NVIDIA AI Computing by HPE

Jetzt chatten

© Copyright 2025 Hewlett Packard Enterprise Development LP. Die hier enthaltenen Informationen können sich jederzeit ohne vorherige Ankündigung ändern. 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 hierin enthaltene technische oder redaktionelle Fehler oder Auslassungen.

Intel, Intel Xeon und das Intel-Logo sind Marken der Intel Corporation oder ihrer Tochtergesellschaften in den USA und/oder anderen Ländern. NVLink, NVIDIA und das NVIDIA-Logo sind Marken und/oder eingetragene Marken der NVIDIA Corporation in den USA und anderen Ländern. Alle Marken von Dritten sind Eigentum der jeweiligen Rechteinhaber.

a00138852DEE, Rev. 3

HEWLETT PACKARD ENTERPRISE

hpe.com