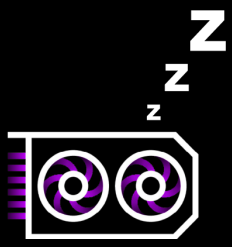
Versorgen Sie datenhungrige GPUs mit blitzschneller Leistung

Tipps für die optimale Nutzung Ihre GPUs

Bewältigen Sie die Herausforderungen Ihrer Arbeits- und Datenspeicher, um eine Unter- oder Überbeanspruchung teurer GPUs zu vermeiden.

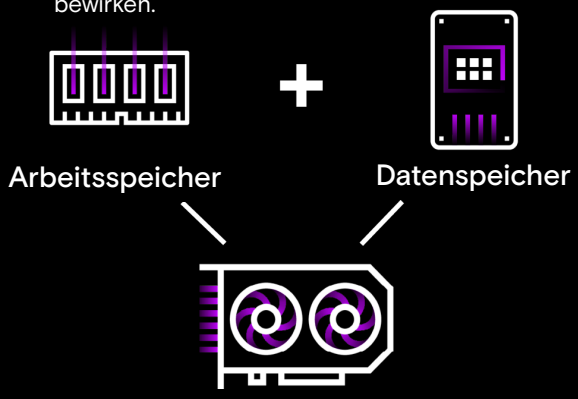
Sie stecken jede Menge Ressourcen in GPUs, und das aus gutem Grund. GPUs sind essenziell für KI, ML, GNN und weitere Innovationen, die unsere Datennutzung grundlegend verändern werden. Wie bei allen großen Investition möchten Sie in jedem Fall die bestmöglichen Renditen erzielen. Betrachten Sie Ihr System, und fragen Sie sich: „Setze ich meine GPUs optimal ein?“



- Die Unterbeanspruchung von GPUs bedeutet verschwendete Ressourcen. Denn die GPUs sind untätig, obwohl sie hart arbeiten sollten. In diesem Fall amortisiert sich die Investition nicht, und Sie verschwenden Energie, Platz und potenzielle Leistung.

- Eine Überbeanspruchung von GPUs liegt vor, wenn Sie Ihren GPUs zu viel abverlangen. Ihr Stromverbrauch steigt, die GPUs sind durch Überhitzung gefährdet, und die Leistung leidet unter Engpässen, langsamen Verarbeitungszeiten und verringerter Effizienz.

Arbeits- und Datenspeicher spielen eine entscheidende Rolle für eine optimale GPU-Nutzung, insbesondere bei ressourcenintensiven Aufgaben wie AI, ML und GNN. In diesen Fällen können Arbeits- und Datenspeicher mit hoher Leistung viel bewirken.



Optimieren Sie Ihren Speicher für KI-Workloads

KI ist auf schnelle, skalierbare GPU-Architekturen angewiesen, aber die richtige Dimensionierung der Speicherinfrastruktur kann ebenso wichtig sein, um Ihre Ziele zu erreichen. Durch die Wahl der richtigen Speichergröße können Sie sicherstellen, dass die GPU-Leerlaufzeit reduziert wird, während der Stromverbrauch erheblich sinkt. Auf diese Weise wird der Einsatz größerer KI-Anwendungen innerhalb eines gegebenen Strombudgets möglich.

Die Micron 9550 SSD ist ein bahnbrechendes, leistungsstarkes Speichermedium, das starke Leistung, niedrige Latenz und Energieeffizienz bietet, um GPUs bei anspruchsvollen Workloads im Rechenzentrum auf optimalem Niveau zu halten.

- Ideal für KI und andere Hochleistungsanwendungen
- Reduziert den Stromverbrauch signifikant1
- Controller-ASIC, NAND der 8. Generation und DRAM von Micron entwickelt

Nutzen Sie NVIDIA Magnum IO GPUDirect Speicher, um eine direkte Verbindung zwischen GPUs und SSDs herzustellen.

Die direkte Verbindung zwischen GPUs und SSDs[2] ist eine Methode zur Latenzverringerung und Verbesserung der Datenübertragungsgeschwindigkeit zwischen dem Speicher und den GPUs. Eine beliebte Methode für diese Aufgabe ist die Verwendung der NVIDIA Magnum IO GPUDirect-Speicherlösung.

Aufgrund des GPUDirect-Speichers hat die SSD-Leistung einen großen Einfluss. Da die GPU-Daten direkt von der SSD abgerufen werden können, sind hohe IOPS und Durchsätze unbedingt notwendig. Sie stellen sicher, dass die GPU nicht untätig ist, während sie auf Daten wartet.

Die Micron 9550 NVMe SSD erfüllt diese Aufgabe hervorragend[3], da ihre starke Leistung eine bessere GPU-Auslastung gewährleistet, insbesondere bei Workloads wie KI, ML und GNN.

Im Vergleich zu anderen Hochleistungs-SSDs erhöht die Micron 9550 die System-Bandbreite dramatisch und senkt den Stromverbrauch bei verschiedenen Trainings-Workloads unter Verwendung von GPUDirect Storage signifikant[4].

34 %
schnellerer Durchsatz

76 %
mehr Energieeffizienz

81 %
weniger Stromverbrauch

33 %
schnellere Training-Workloads

60 %
mehr SSD-Leistung

29 %
weniger Stromverbrauch durch das Trainingssystem

Machen Sie NVMe-Speicher zu einer dritten Speicherklasse für „langsamen“ Speicher

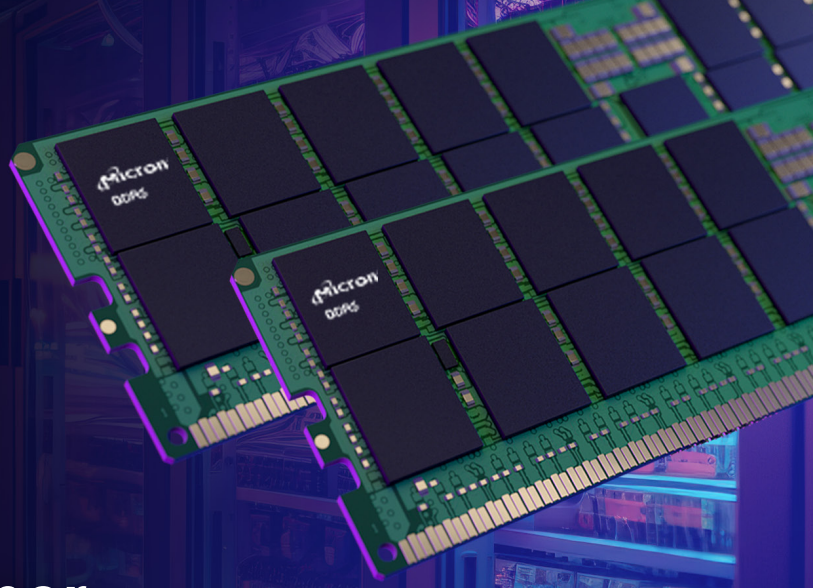
Eine Methode zum Trainieren großer Modelle besteht darin, so viel Hochgeschwindigkeitsspeicher wie möglich auf der GPU zu installieren und möglichst viel System-DRAM einzusetzen. Passt ein Modell nicht zu diesem „HBM + DRAM“-Ansatz, kann es über mehrere GPU-Systeme parallelisiert werden.

Aufgrund der geringeren GPU-Auslastung und des weniger effizienten Datenflusses über Netzwerk- und Systemverbindungen ist parallelisiertes Training über mehrere Server jedoch extrem kostenintensiv. Und es kann dabei leicht zu Engpässen kommen.

Um diese Probleme zu überwinden, kann NVMe-Speicher als dritte Speicherklasse für „langsamen“ Speicher verwendet werden. Dies kann durch die Verwendung von Big Accelerator Memory (BAM) mit GPU Initiated Direct Storage (GIDS) erreicht werden, um den NVMe-Treiber neu zu konfigurieren und zu optimieren, um die Daten- und Steuerpfade zur GPU zu behandeln.

Der BaM-Software-Stack stützt sich auf die geringe Latenz, den hohen Durchsatz, die große Dichte und die hohe Ausdauer von NVMe-SSDs als Speichererweiterung. Diese Anforderungen erfüllt die Micron 9550 NVMe SSD hervorragend.

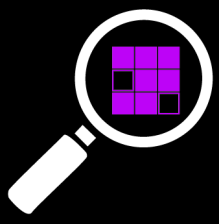
BaM- und GIDS-Test zeigen, dass die Micron NVMe SSD ein schnelleres GNN-Training sowie höhere SSD-Leistung, weniger Systemleistung und stabile Skalierungsergebnisse ermöglicht[5].



Aufrechterhalten hoher Durchsätze mit DDR5-Speicher

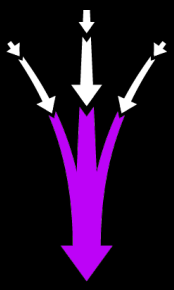
DDR5 bietet im Vergleich zu DDR4[6] größere Bandbreiten sowie eine verbesserte Zuverlässigkeit, Verfügbarkeit und Skalierung. Es sorgt dafür, dass Daten reibungslos fließen, damit GPUs kontinuierlich versorgt werden und mit maximaler Effizienz arbeiten können.

Für KI/ML-Workloads ist DDR5 unerlässlich, um ein volles Serverpotenzial zu erreichen. Testdaten[7] zeigen, dass DDR5 massive Leistungsverbesserungen gegenüber DDR4 für diese Arten von Workloads bietet.



Bildklassifizierung

- 7,3-mal schnellere Bildklassifizierung
- 40 % mehr dauerhafte Speicherbandbreite



Empfehlung Durchsatz

- 4,3-mal höherer Empfehlungsdurchsatz
- 200 % mehr Speicherbandbreite für Empfehlungen

Natürliche Sprache

- 4,9-fache Steigerung bei der Verarbeitung natürlicher Sprache
- 55 % mehr Speicherbandbreite für höheren Durchsatz

Erhalten Sie Expertentipps zur Optimierung Ihrer GPU

Wir arbeiten eng mit unseren Kunden an den technischen Standorten weltweit zusammen, um Prozesse zu optimieren und die Belastung Ihrer technischen Teams zu verringern. Die Experten von Microsoft führen fundierte Tests an der Serverarchitektur durch, um spezielle Lösungen zu entwickeln, mit denen sich die GPUs auf optimalen Niveaus halten lassen, während gleichzeitig der Stromverbrauch reduziert und die Gesamteffizienz verbessert wird.

Mehr erfahren auf micronpg.com/datacenter

1. Basierend auf Testergebnissen von Micron bei der Auslagerung von KI-Training, wobei die direkte Datenübertragungsrate von SSD zu GPU mit einem 1 TB-Datensatz unter Verwendung standardmäßiger KI-Leistungsbenchmarks gemessen wurde.
2. Siehe https://developer.nvidia.com/gpudirect-storage für weitere Einzelheiten zu den IO-Pfadunterschieden.
3. Beachten Sie den Leitfaden zur NVIDIA GPUDirect Speicherlösung für weitere Informationen zu GDS.
4. Die internen Analysen von Micron zu KI-Training-Workloads machen deutlich, dass die IO-Größen je nach Modell und Datenformat unterschiedlich sind. Deshalb betrachtet dieses Dokument anhand zweier Beispiele kleine (4 KB), mittlere (128 KB) und große (1 MB) Übertragungsgrößen.
5. Die Werte entsprechen den während der Tests beobachteten Maximalwerten. Wettbewerbsfähige PCIe Gen5 SSDs von den Top 10 PCIe SSD-Lieferanten, die im Forward Insights Analystenbericht „SSD Supplier Status Q1/24 Mai 2024“ aufgeführt sind.
6. Mit einer Startdatenrate von 6400 MT/s überträgt DDR5 zweimal (100 %) mehr Daten als DDR4, das nur eine maximale Standarddatenrate von 3200 MT/s bietet. Die vom JEDEC-Komitee prognostizierte Geschwindigkeit von 8800 MT/s ist 2,75-mal höher als die maximale Standardgeschwindigkeit von DDR4, die 3200 MT/s beträgt.
7. Das Data Center Workload Engineering (DCWE)-Team von Micron hat zusammen mit Supermicro und Intel Tests und Validierungen durchgeführt, um eine ideale CPU-gestützte und für KI-Inferenz-Workloads optimierte Plattform zu bestimmen. Die von Micron durchgeführten Workload-Test konzentrierten sich auf Inferenz-Benchmarking, das auf MLPerf (Machine Learning Performance) basiert. Es wurde gemessen, wie schnell Systeme die Modelle in einem Anwendungsszenario mit der NLP-Lösung BERT (Bidirectional Encoder Representations from Transformers), DLRM (Deep Learning Recommendation Model) und der Bildklassifizierung mittels ResNet ausführen. Die tatsächlichen Ergebnisse können abweichen. Mehr erfahren: Micron Server DDR5 AI Use Case Test Results eBook (EN) (micronpg.com)