Häufig gestellte Fragen zur Auswahl einer KI- und Analyseinfrastruktur mit skalierbaren Intel® Xeon® Prozessoren der 3. Generation

1. Welche Infrastrukturressourcen wirken sich auf die KI- und Analyseleistung aus?

Viele glauben, dass Rechnerressourcen die einzigen wichtigen Infrastrukturressourcen für KI und Analysen sind. Und diese Ressourcen sind für schnellere Erkenntnisse von entscheidender Bedeutung.

Aber sehr oft werden andere Ressourcen, die einen großen Einfluss auf die KI- und Analyseleistung haben, ignoriert, beispielsweise Arbeits- und Datenspeicherkapazität, Speicherleistung und Netzwerkleistung.

Diese Infrastrukturressourcen sind von entscheidender Bedeutung, da Rechnerressourcen auf schnellen Datenzugriff angewiesen sind und die Leistung durch im Leerlauf befindliche Kerne beeinträchtigt wird.

Aus diesem Grund hat Intel einen umfassenden End-to-End-Ansatz für KI- und Analyseinfrastrukturinnovationen gewählt. Vorteile dieser Innovationen

- Erhebliche Erhöhung der Speicherkapazität, sodass sich mehr Daten näher am Prozessor befinden können
- Zahlreiche SSD-Optionen, die eine optimale Balance zwischen bahnbrechender Speicherleistung und kosteneffizienter Kapazität ermöglichen
- Beschleunigung der I/O-Vorgänge zu und von Servern mit Netzwerklösungen, die die Leistung verbessern, Latenz reduzieren und Netzwerkaufgaben von CPUs auslagern

2. Was ist KI-Training und -Inferenzierung?

Künstliche Intelligenz oder KI umfasst eine breite Palette von Techniken zur Ableitung von nutzbarem Wissen und verwertbaren Erkenntnissen.

Deep Learning ist eine Untergruppe von KI, die mithilfe gewaltiger Datenmengen vielschichtige neuronale Netzwerke erstellt. Diese neuronalen Netzwerke werden dann verwendet, um Muster autonom zu erkennen und Aktionen zu optimieren.

KI-Training dient zur Erstellung eines neuronalen Netzwerks. KI-Inferenzierung ist die autonome Erkennung, die das trainierte neuronale Netzwerk nutzt.

KI-Training findet in der Regel in einem Rechenzentrum oder in der Cloud statt, da es enorme Datenmengen und MPP-Leistung (Massively Parallel Processing) erfordert.

KI-Inferenzierungen werden meist in Edge-Anwendungen eingesetzt, die für optimale Ergebnisse Erkennungen und Aktionen nahezu in Echtzeit erfordern.

Häufige Anwendungsfälle der KI-Inferenzierungen sind computergestütztes Sehen, Bilderkennung, Spracherkennung, Erkennung von anomalen Verhaltensweisen und mehr.

Diese Techniken werden in vielen Bereichen praktisch eingesetzt, z. B.:

- Intelligente digitale Werbedisplays: Mithilfe von computergestütztem Sehen kann erkannt werden, ob Kunden auf Inhalte digitaler Werbedisplays achten. Diese Informationen werden mit anonymen demografischen Informationen kombiniert, um Inhalte zu messen und anzupassen.

- Chatbots: Spracherkennung ermöglicht persönliche Unterstützung, Selbsthilfe und interaktive digitale Erlebnisse.
- Erkennung von Finanzbetrug: Anomales Finanzverhalten wird erkannt, um illegale oder nicht autorisierte Finanztransaktionen proaktiv zu verhindern.
- Schutz vor Cyber-Bedrohungen: Anomales Netzwerk- oder Anwendungsverhalten wird erkannt, um böswillige Angriffe auf sensible Systeme, Anwendungen oder Daten zu verhindern.

3. Was sind die Vorteile einer CPU-basierten Infrastruktur für KI?

Eine wichtige Überlegung bei der Auswahl Ihrer KI-Infrastruktur ist, ob Sie sich für eine GPU- oder CPU-basierte Lösung entscheiden.

Zu Beginn der Verbreitung von KI galten GPUs als erste Wahl für viele KI-Aufgaben, einschließlich KI-Training und -Inferenzierung, da sie gegenüber Allzweck-CPUs eine höhere Leistung aufwiesen.

Die Bereitstellung einer GPU-basierten Infrastruktur für KI-Workloads bedeutete jedoch im Allgemeinen die Bereitstellung und Verwaltung einer isolierten Infrastruktur für KI-Anforderungen, die von der allgemeinen Infrastruktur getrennt war, die für viele andere Workloads konzipiert wurde.

Die neuesten CPU- und Plattforminnovationen von Intel haben KI-Workloads wie z. B. Training und Inferenzierungen erheblich beschleunigt, sodass die IT-Abteilung eine gemeinsame Infrastruktur für KI- und datenzentrierte Workloads sowie eine Vielzahl anderer geschäftsfördernder Anwendungen bereitstellen, betreiben und unterstützen kann.

Dies bedeutet eine verbesserte Auslastung und Effizienz von IT-Ressourcen, geringere Investitions- und Betriebskosten sowie weniger Komplexität im Rechenzentrum.

4. Warum sind skalierbare Intel® Xeon® Prozessoren der 3. Generation eine großartige Wahl für KI-Training und -Inferenzierung?

Intel® Deep Learning Boost (Intel® DL Boost) wurde in skalierbaren Intel® Xeon® Prozessoren der 2. Generation eingeführt, um die Leistung von KI-Inferenzierungen deutlich zu verbessern. Unternehmen können damit anspruchsvolle Inferenzierungs-Workloads in ihrer allgemeinen Infrastruktur ausführen, ohne deren Leistung zu beeinträchtigen.

Mit den skalierbaren Intel Xeon Prozessoren der 3. Generation baut Intel seine Marktführerschaft im Bereich CPU-basierte KI-Beschleunigung durch die Integration von bfloat16-Support in Intel DL Boost aus.

Dies sind die ersten Allzweck-CPUs mit bfloat16-Support, was die KI-Trainingsleistung weiter beschleunigt. Aber neu und noch spannender ist, wie bfloat16 die KI-Trainingsleistung deutlich beschleunigt.

Gängige KI-Frameworks und -Bibliotheken wurden von Intel Softwareingenieuren leistungsoptimiert, um die neuesten Intel Deep Learning Boost-Funktionen zu nutzen, um die KI-Leistung auf skalierbaren Intel Xeon Prozessoren der 3. Generation zu maximieren.

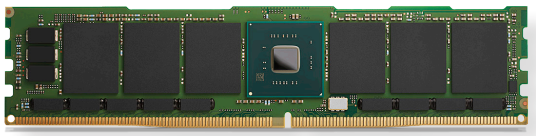
Jetzt können Sie mehr KI-Workloads auf Ihrer Allzweck-IT-Infrastruktur mit Intel Technik ausführen, um IT-Kosten und -Komplexität zu reduzieren und gleichzeitig Auslastung und Effizienz zu verbessern.

5. Wie verbessert persistenter Intel® Optane™ Speicher der Produktreihe 200 KI und Analysen?

Was fällt Ihnen beim Gedanken an KI- und Analyse-Workloads sofort ein, wenn Sie die Infrastrukturanforderungen berücksichtigen? Die meisten denken dabei an die Rechenleistung, die diese Workloads benötigen. Sie auch?

Was KI und Analysen jedoch wirklich von anderen leistungsabhängigen Workloads unterscheidet, ist, wie datenintensiv sie sind. Sie erfordern große Datenmengen. Die Geschwindigkeit, mit der der Prozessor auf diese Daten zugreifen kann, hat einen enormen Einfluss auf die Gesamtleistung.

Bis vor kurzem war DRAM die einzige Wahl für Systemspeicher. Obwohl DRAM hervorragende Leistung liefert, konnte es nicht mit dem Moore'schen Gesetz in Bezug auf Dichte, Kapazität und Kosten mithalten. Somit konnte es nicht mehr mit der Verarbeitungsleistung mithalten.



Und hier kommt der persistente Intel Optane Speicher der Produktreihe 200 ins Spiel: Diese bahnbrechende Speicherinnovation wird von skalierbaren Intel Xeon Prozessoren der 3. Generation unterstützt und erweitert die Systemspeicherkapazität auf bis zu 4,5 TB pro Prozessorsockel[1].

Die erweiterte Kapazität bedeutet, dass Sie viel größere Datenmengen für Analysen nutzen können. Und im Gegensatz zu DRAM können Daten dauerhaft gespeichert werden, was im Vergleich zu herkömmlichen NAND-SSDs einen bis zu 225-mal schnelleren[2] Lesezugriff ermöglicht.

Anwendungen wie SAP nutzen die höhere Kapazität und Persistenz, um Analysen zu beschleunigen und neue Möglichkeiten für In-Memory-Datenbanken zu bieten. Die Persistenz ermöglicht auch deutlich schnellere Neustarts nach einer planmäßigen Wartung.

1) 6 x persistenter Intel Optane Speicher mit 512 GB (3.072 GB) + 6 x 256-GB-DDR4-DRAM (1.536 GB) = 4.608 GB Gesamtspeicher pro Sockel.

2) Leselatenz des persistenten Intel Optane Speichers im Leerlauf: 340 Nanosekunden. Leselatenz des TLC NAND SSD der Intel® Produktreihe Intel® SSD DC P4610 im Leerlauf: 77 Mikrosekunden.

Die Funktionsmerkmale und Vorteile von Intel Technik hängen von der Systemkonfiguration ab und können geeignete Hardware, Software oder die Aktivierung von Diensten erfordern. Die Leistung kann je nach Systemkonfiguration unterschiedlich ausfallen. Kein Produkt und keine Komponente bietet absolute Sicherheit.