Produktbeschreibung

Beschleunigte KI
Skalierbare Intel® Xeon® Prozessoren

Integrierte Beschleuniger in skalierbaren Intel® Xeon® Prozessoren steigern Leistung der gesamten KI-Pipeline

70 %
der KI-Inferenz in Rechenzentren **laufen auf Intel® Xeon® Prozessoren**[1]

9 von 10
Unternehmens-anwendungen werden bis 2025 **KI nutzen**[2]

KI umfasst ein breites Spektrum von Workloads und Anwendungsfällen – von der Datenanalyse und klassischem maschinellen Lernen bis hin zur Sprachverarbeitung und Bilderkennung. Skalierbare Intel® Xeon® Prozessoren kombinieren flexible Rechenleistung für die gesamte KI-Pipeline mit integrierten Beschleunigern für spezifische KI-Workloads in den Bereichen Data Science, Modell-Training und Deep-Learning-Inferenz.

KI ist mehr als Deep Learning und wird noch umfangreicher werden

KI befindet sich in einem Frühstadium und entwickelt sich in jeder Hinsicht rasant weiter. Klassische Machine-Learning-Algorithmen und Deep-Learning-Modelle werden zu grundlegenden Bausteinen für das Geschäftsleben – von zentralen Unternehmensanwendungen bis zu automatisierten Telefonzentralen. Bis KI zur Nutzung in großem Maßstab bereit ist, muss ein langer Entwicklungsprozess durchlaufen werden – von Data Science über das Training und die Validierung bis hin zum Einsatz. Jeder Schritt hat seine eigenen Entwicklungs-Toolchains, Frameworks und Workloads. Sie alle führen zu spezifischen Engpässen und stellen unterschiedliche Anforderungen an die Rechenressourcen. Skalierbare Intel® Xeon® Prozessoren verfügen über integrierte Beschleunigung, die diese Hürden überwinden kann und die KI-Leistung umfassend steigert.

KI ist eigentlich nur Mathematik. Ziemlich viel Mathematik

Jede KI-Aufgabe und -Operation basiert auf Unmengen mathematischer Berechnungen. Viele Data-Science-Operationen – wie die Modellierung von Daten und Machine-Learning-Algorithmen – beruhen auf Statistik, Algebra und komplexer Vektormathematik. Deep-Learning-KI erfordert eine Fülle an Matrixmultiplikationen. Alle diese KI-Anwendungen sind Brute-Force-Operationen, die große Datenbestände und umfangreiche Verarbeitungsressourcen wie CPUs, GPUs, FPGAs und Workload-spezifische, maßgefertigte ASICs umfassen.

Intel® Advanced Vector Extensions 512 (Intel® AVX-512) – der mathematische Trick, der KI beschleunigt

Intel® Xeon® Kerne können für das Hashing der SSL-Verschlüsselungen von Websites, die Verarbeitung riesiger Datenbanken und Simulationen für die pharmazeutische Forschung, das Chipdesign oder Formel-1-Motoren genutzt werden. Sie sind leistungsstarke Allrounder, aber beim Deep-Learning-Training – einer Teilmenge der gesamten KI-Pipeline – kommt ihre Geschwindigkeit normalerweise nicht an die von dedizierten Beschleunigern heran. Das liegt daran, dass CPUs Anweisungen sequentiell verarbeiten, eine Berechnung nach der anderen. Andere Prozessortypen können Anweisungen parallel verarbeiten, also mehrere Berechnungen gleichzeitig durchführen.

Intel® AVX-512 überwindet die architektonisch bedingten Einschränkungen einer CPU, indem es mehr Anweisungen in jeden Taktzyklus packt. So kann die CPU sozusagen tricksen und ähnlich wie ein Parallelprozessor arbeiten.

Erfolgsstorys von Kunden – Praxistaugliche Beschleunigung auf skalierbaren Intel® Xeon® Prozessoren

Tencent Cloud bietet Sprachsynthese in Echtzeit mit skalierbaren Intel® Xeon® Prozessoren der 3. Generation.

Mehr Details erhalten ›

BeeKeeperAI entwickelt klinische KI-Algorithmen, die den Datenschutz wahren.

Story lesen ›

Komplizierte CPU-Befehle, einfache Strategie: Intelligenter arbeiten, in jedem Taktzyklus mehr erledigen

Die Erweiterungen in Intel® AVX-512 sind Befehlssätze, die der CPU vorgeben, was sie tun soll – und wie. Ihre Arbeitsweise ist sehr komplex, aber die grundlegende Logik von Intel® AVX-512 ist recht einfach. Als erstes werden mehrere Schritte zu weniger Anweisungen zusammengefasst, soweit dies möglich ist. Als zweites wird die CPU dabei unterstützt, mehr Anweisungen pro Taktzyklus zu erledigen.

Weniger Schritte bedeuten schnellere Verarbeitung

Mathematik kann sehr intelligent sein – und sehr elegant. Intel® AVX-512 verwendet viel intelligente, schöne Mathematik, um gängige Rechenoperationen zu verdichten und zu weniger Schritten zu kombinieren. Ein einfaches Beispiel: Man könnte eine CPU anweisen, 3 x 3 x 3 x 3 x 3 zu berechnen, was fünf Taktzyklen dauern würde. Oder man könnte einen Befehl für 3^5 geben, den die CPU in einem Taktzyklus ausführen kann. Intel® AVX-512 nutzt diese Logik und wendet sie auf Hunderte von Workload-spezifischen Operationen an, darunter einige der schwierigsten Operationen im Bereich KI.

Mit acht zählt es sich viel schneller als mit eins

Das „512" in AVX-512 bezieht sich auf die zweite Art und Weise, wie diese Befehle die Anzahl der Bits, die der CPU zur Verfügung stehen, mit jedem Taktzyklus erhöhen. Vor 40 Jahren war ein 16-Bit-PC ziemlich beeindruckend. Bald übernahmen dann 32-Bit-Rechner. Heutzutage laufen Smartphones mit 64 Bit. Die Bitanzahl bezieht sich auf die Anzahl der Register – die Speicherbereiche, in denen die CPU Daten aufbewahrt –, die die CPU pro Taktzyklus adressieren kann. Intel® AVX-512 erweitert die Anzahl der Register auf – unschwer zu erraten – 512 Bits. Wenn eine Anwendung Intel® AVX-512 nutzt, läuft sie bis zu 8 mal schneller als die 64-Bit-Basisgeschwindigkeit der CPU, indem einfach die Anzahl der Register erweitert wird. Es ist, als würde man anstatt mit eins, zwei, drei … 8, 16, 24 bis 96 zählen.

Intel® Deep Learning Boost (Intel® DL Boost) – intelligentere Mathematik für neuronale Netze

Deep-Learning-KI nutzt eine Fülle an Matrixmultiplikationen, um Modelle neuronaler Netze zu trainieren und diese Modelle mithilfe einer Methode namens Inferenz auf reale Aufgaben anzuwenden. Bei der Inferenz vergleicht ein Computer eingehende Daten (z. B. ein Audiosignal, das Sprache enthält) mit einem Modell (in diesem Fall einem Spracherkennungsmodell) und schließt daraus auf die Bedeutung der Daten. Inferenz kommt bei der Objekterkennung, der Bildsegmentierung, der Texterkennung und praktisch jeder anderen KI-Aufgabe im Bereich Deep Learning zum Einsatz.

Das Trainieren von Deep-Learning-Modellen kann Stunden oder Tage an Rechenleistung in Anspruch nehmen. Deep-Learning-Inferenz kann Bruchteile von Sekunden bis Minuten dauern – je nachdem, wie komplex ein Modell ist und wie präzise die Ergebnisse sein müssen. Wenn man das Trainieren oder die Inferenz auf Rechenzentrumsebene skaliert, werden die Zeit-, Energie- und Leistungsbudgets immens.

Intel® DL Boost nutzt verschiedene Intel® AVX-512 Befehle zur Beschleunigung von Deep-Learning-Workloads. Es kombiniert drei Anweisungen zu einem einzigen Befehl von Vector Neural Network Instructions (VNNI), was die Anzahl der Anweisungen pro Taktzyklus reduziert. Intel® DL Boost beschleunigt Deep-Learning-Workloads zudem mit INT8-Präzision.

Bevorstehende Neuerungen werden die KI-Leistung noch weiter steigern

Die skalierbaren ® Xeon® Prozessoren der 4. Generation werden über einen integrierten Beschleuniger für Matrixmultiplikationen verfügen, die das Herzstück von Deep-Learning-Workloads sind. Intel® Advanced Matrix Extensions (Intel® AMX) kombiniert einen neuen Befehlssatz, der große Matrizen in eine einzige Anweisung verwandelt, mit zweidimensionalen Registerdateien, die größere Datenmengen für jeden Kern speichern.

Intel® Xeon® Prozessoren sorgen praktisch automatisch für schnellere KI

KI-Beschleunigung ist bei skalierbaren Intel® Xeon® Prozessoren in die Befehlssatzarchitektur (Instruction Set Architecture, ISA) der CPU integriert. Sie ist also für jede Software bereit und verfügbar, die zu ihrer Nutzung fähig ist. Wir erwarten nicht, dass Datenwissenschaftler:innen und KI-Entwickler:innen ihre Tools umschreiben und für Intel® AVX-512 neu kompilieren, sondern erledigen das für sie.

Die Software-Ingenieur:innen von Intel optimieren ständig als Open Source verfügbare KI-Toolchains und stellen diese Optimierungen der Community zur Verfügung. So wird zum Beispiel TensorFlow* 2.9 standardmäßig mit den Optimierungen der Intel® oneAPI Deep Neural Network Library (Intel® oneDNN) ausgeliefert. Laden Sie die neueste Version herunter, und TensorFlow* wird automatisch die Vorteile der Optimierungen von Intel nutzen.

Für andere Anwendungen in der KI-Pipeline können Datenwissenschaftler:innen und Entwickler:innen als Open Source verfügbare Distributionen, Bibliotheken und Entwicklungsumgebungen von Intel kostenlos herunterladen. Diese nutzen alle integrierten Beschleuniger in unserer ISA für skalierbare Intel® Xeon® Prozessoren der 3. Generation.

Im Grunde genommen ist schnellere KI auf Hardware von Intel sehr einfach: Sie laden die Intel Version der Tools herunter, die Sie bereits verwenden, und schon kann es losgehen.

Weitere Informationen

KI und Deep Learning auf skalierbaren Intel® Xeon® Prozessoren ›

Intel® AVX-512 ›

Intel® Deep Learning Boost ›

Intel® AI Analytics Toolkit ›

Vorteile der Software-Optimierung für Anwendungen in der KI-Pipeline

~38–200x schnelleres scikit-learn* mit der Intel® Erweiterung für scikit-learn[3]

~90x schnelleres pandas* mit der Intel® Distribution von Modin[3]

Bis zu **3x** schnelleres TensorFlow* bei Nutzung von Intel® oneDNN[3]

KI-Beschleunigung bei skalierbaren Intel® Xeon® Prozessoren der 3. Generation

Geschwindigkeitssteigerungen für Deep-Learning-KI-Workloads

Bis zu **1,74x** **höherer INT8-Batch-Inferenz-Durchsatz** bei BERT-Large SQuAD mit Intel® DL Boost auf skalierbaren Intel® Xeon® Prozessoren der 3. Generation im Vergleich zur vorherigen Generation[4]

Bis zu **1,59x** **höherer INT8-Echtzeit-Inferenz-Durchsatz** mit Intel® DL Boost auf skalierbaren Intel® Xeon® Prozessoren der 3. Generation im Vergleich zur vorherigen Generation[5]

Bis zu **4,5x** **mehr Bilder pro Sekunde bei INT8-[6] und bis zu 6x mehr Bilder pro Sekunde bei BF16-Objekterkennung[7]** (SSD-ResNet-34) bei Nutzung von Intel® AMX bei der kommenden 4. Generation der skalierbaren Intel® Xeon® Prozessoren

Beginnen Sie noch heute mit der Beschleunigung von KI-Workloads – in der Cloud oder auf Ihrer eigenen Infrastruktur – mit Optimierungen von Intel für KI und maschinelles Lernen.

Mehr erfahren ›

intel XEON

1. Basierend auf der Intel Marktmodellierung der weltweit installierten Basis von Rechenzentrumsservern mit KI-Inferenz-Workloads im Dezember 2021.
2. „IDC FutureScape: Worldwide IT Industry 2020 Predictions", Oktober 2019. Doc #US45599219. idc.com/getdoc.jsp?containerId=US45599219.
3. „One-Line Code Changes to Boost pandas, scikit-learn, and TensorFlow Performance", Juli 2021. intel.com/content/www/us/en/developer/articles/technical/code-changes-boost-pandas-scikit-learn-tensorflow.html
4. Siehe [123] unter intel.com/3gen-xeon-config. Die Ergebnisse können von Fall zu Fall abweichen.
5. Siehe [122] unter intel.com/3gen-xeon-config. Die Ergebnisse können von Fall zu Fall abweichen.
6. Siehe [41] unter edc.intel.com/content/www/us/en/products/performance/benchmarks/vision-2022/. Die Ergebnisse können von Fall zu Fall abweichen.
7. Siehe [42] unter edc.intel.com/content/www/us/en/products/performance/benchmarks/vision-2022/. Die Ergebnisse können von Fall zu Fall abweichen.

0922/MP/CMD/PDF