Produktbeschreibung

Intel® AI Engines
Skalierbare Intel® Xeon® Prozessoren der 5. Generation

Skalierbare Intel® Xeon® Prozessoren der 5. Generation und Intel® AI Engines steigern Leistung der gesamten KI-Pipeline

65 %
der KI-Inferenz in Rechenzentren **laufen auf Intel® Xeon® Prozessoren**[1]

Bis zu
14 x höhere
Inferenzleistung bei Echtzeit-Objekterkennung (SSD-ResNet34*) auf Intel® Xeon® Prozessoren der 5. Generation mit Intel® AMX BF16 im Vergleich zu Intel® Xeon® Prozessoren der 3. Generation[2]

Bis zu
9,9 x höhere
Inferenzleistung bei natürlicher Sprachverarbeitung (BERT-Large*) und 7,7 x höhere Leistung/Watt auf Intel® Xeon® Prozessoren der 5. Generation mit Intel® AMX BF16 im Vergleich zu Intel® Xeon® Prozessoren der 3. Generation[3]

Bis zu
8,7 x höhere
Batch-Inferenzleistung bei Empfehlungssystem (DLRM *) und 6,2 x höhere Leistung/Watt auf Intel® Xeon® Prozessoren der 5. im Vergleich zu Intel® Xeon® Prozessoren der 3. Generation[4]

KI umfasst eine große Bandbreite von Workloads und Anwendungsfällen, von der Datenvorverarbeitung und klassischem maschinellen Lernen (ML) bis hin zu Deep-Learning-Algorithmen wie natürliche Sprachverarbeitung und Bilderkennung. Skalierbare Intel® Xeon® Prozessoren bieten eine hohe Rechenleistung für die gesamte KI-Pipeline. Sie umfassen integrierte Beschleuniger, die für bestimmte KI-Workloads in den Bereichen maschinelles Lernen, Datenanalyse und Deep Learning (DL) optimiert sind.

Integrierte Leistung für KI im gesamten Unternehmen

KI ist allgegenwärtig und wird von unterschiedlichsten und kritischen Workloads genutzt. Klassisches ML und DL werden zu grundlegenden Bausteinen für das Geschäftsleben – von zentralen Unternehmensanwendungen bis zu automatisierten Telefonzentralen. Für die Nutzung von KI in großem Maßstab ist ein langer Entwicklungsprozess nötig – von der Datenvorverarbeitung über das Training 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 Beschleuniger, die dafür genutzt werden können, die gesamte Pipeline sofort ohne weitere Anpassungen laufen zu lassen und die KI-Leistung umfassend zu steigern.

Intel® Accelerator Engines sind spezialisierte integrierte Beschleuniger, die die anspruchsvollsten neuen Workloads unterstützen

Die skalierbaren Intel® Xeon® Prozessoren der 5. Generation bringen Spitzenleistung bei allgemeinen Rechenaufgaben und werden weiterhin die Grundlage für die Unterstützung vieler kritischer KI-Workloads von heute bilden. Diese Prozessoren verfügen über Intel® Advanced Matrix Extensions (Intel® AMX), einen integrierten KI-Beschleuniger, der dafür konzipiert ist, Deep-Learning-Inferenz und -Training auf der CPU schneller zu machen. In vielen Fällen werden damit die zusätzlichen Kosten und die Komplexität eines diskreten Beschleunigers vermieden. Die neueste Generation der Intel® Xeon® Prozessoren eignet sich hervorragend für Large Language Models (LLMs; große Sprachmodelle) mit weniger als 20 Milliarden Parametern – wodurch in der Regel die SLAs der Kunden erfüllt werden.[5] Intel® AMX bietet auch beim Transfer Learning und bei der Feinabstimmung ausgezeichnete Leistung. Dadurch benötigt das Training von Modellen nur 4 Minuten – und nicht Stunden oder Tage –, ohne dass dafür zusätzliche Hardware nötig wäre. 65 % der KI-Inferenz in Rechenzentren laufen auf Intel® Xeon® Prozessoren. Deshalb können Kunden ihre bestehende Architektur für Allzweck-KI nutzen, anstatt sich mit der Komplexität des Umstiegs auf eine GPU-Infrastruktur auseinandersetzen zu müssen.

Skalierbare Intel® Xeon® Prozessoren der 5. Generation und Intel® AI Engines bilden die Innovation der Zukunft

Ganz gleich, ob sie für Workloads lokal, in der Cloud oder am Edge eingesetzt werden, Intel® Xeon® Prozessoren mit integrierten Intel® Accelerator Engines helfen Unternehmen dabei, neue Ziele zu erreichen. Sie bieten eine Reihe von Vorteilen wie stärkeren Datenschutz und bessere Auslastung der Infrastruktur.

Erfolgsstorys von Kunden: Reale Beschleunigung auf skalierbaren Intel® Xeon® Prozessoren

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

Mehr Details erhalten ›

Gunpowder nutzt Google Cloud C3*-Instanzen mit skalierbaren Intel® Xeon® CPUs der 4. Generation, um die Rendering-Leistung zu steigern.

Lesen Sie die Story ›

Intel® Accelerator Engines tragen auch dazu bei, die virtuelle und physische CPU-Auslastung zu verbessern und die Lizenzkosten pro Prozessorkern zu senken. Vor allem sorgen diese integrierten Beschleuniger für eine höhere Anwendungsleistung, niedrigere Kosten und mehr Effizienz auf Plattformebene.

Beschleunigtes Deep Learning mit Intel® Advanced Matrix Extensions

Intel® AMX ist die neueste Verbesserung von Intel für Deep-Learning-Training und -Inferenz auf skalierbaren Intel® Xeon® Prozessoren der 5. Generation. Intel® AMX ist ideal für Workloads wie natürliche Sprachverarbeitung, Empfehlungssysteme und Bilderkennung. Damit erzielen Kunden eine bis zu 7,2 x höhere Inferenzleistung bei Echtzeit-Objektklassifizierung und eine bis zu 5,3 x höhere Leistung/Watt auf Intel® Xeon® Prozessoren der 5. Generation mit Intel® AMX BF16 im Vergleich zu Intel® Xeon® Prozessoren der 3. Generation.[6]

Intel® AMX beschleunigt zudem die Workloads für KI-Modelle und ermöglicht mehr Kunden, ihre SLAs auf den Plattformen zu erfüllen, die sie bereits betreiben. Skalierbare Intel® Xeon® Prozessoren der 5. Generation bieten höhere Turbo-Taktfrequenzen für Workloads mit einer Affinität für Vektor- und Matrixoperationen wie High-Performance Computing (HPC) und KI, da sie über 5 Turbo Ratios verfügen.

Intel® AMX verbessert die Leistung von Matrix-Multiplikations-Operationen durch einen höheren Durchsatz (OPs/Zyklus) im Vergleich zu Intel® Advanced Vector Extensions 512 (Intel® AVX-512) auf CPU-Kernen.[7] Das führt zu einer schnelleren Fertigstellung von Deep-Learning-Training-Workloads und ermöglicht es mehr Kunden, ihre SLAs auf den bereits von ihnen genutzten Plattformen zu erfüllen.

Unterstützung von natürlicher Sprachverarbeitung und generativer KI

Skalierbare Intel® Xeon® Prozessoren der 5. Generation mit Intel® AMX sorgen für einen großen Leistungssprung bei natürlicher Sprachverarbeitung – und das ohne zusätzliche Hardware. Intel Bibliotheken sind in TensorFlow* und PyTorch* integriert und auch dafür optimiert. Dadurch profitieren Entwickler:innen ohne weitere Anpassungen von der integrierten KI-Beschleunigung. Entwickler:innen können zudem Code von unterschiedlichen Hardware-Umgebungen einfach migrieren – ein Prozess, der ansonsten zeitaufwändig und teuer sein kann.

Durch die Beschleunigung von Deep-Learning-Inferenz und -Training helfen die skalierbaren Intel® Xeon® Prozessoren der 5. Generation mit Intel® AMX dabei, SLAs zu erfüllen und gleichzeitig die Gesamtbetriebskosten (Total Cost of Ownership, TCO) zu optimieren. Das geschieht zum Beispiel bei einem Deep-Learning-basierten Empfehlungssystem, das Echtzeitdaten über das Nutzerverhalten und zusätzliche Hintergrundinformationen wie Zeit und Standort berücksichtigt.

Auf den Intel® Xeon® Prozessoren der 5. Generation laufen auch generative KI-Modelle, die menschenbezogene Inhalte imitieren und so Large Language Models und Text-zu-Bild-Generierung unterstützen. Bei intensiveren generativen KI-Aufgaben können der spezialisierte KI-Beschleuniger Intel® Gaudi®, die Intel® Data Center GPU und andere Hardwarekomponenten zur Erweiterung der CPU-Fähigkeiten genutzt werden.

Intel® AVX-512 für schnelleres ML

Intel® Xeon® Prozessoren 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.

Intel® AVX-512 wurde über zahlreiche Generationen hinweg verbessert. Dadurch gelingt es den skalierbaren Intel® Xeon® Prozessoren, mehr Anweisungen in jeden Taktzyklus zu packen und die Leistung von parallel verarbeitenden Anwendungen zu verbessern. Die Befehlssatzarchitektur (Instruction Set Architecture, ISA) von Intel® AVX-512 beinhaltet Erweiterungen, die die Leistung verschiedenster Workloads in den Bereichen KI, HPC, Netzwerk und Storage steigern.

Die Turboleistung erhöht sich bei der neuen Prozessorgeneration, da diese vier anstatt fünf Turbo Ratios bietet. Das steigert die Turbo-Taktfrequenzen für bestimmte HPC- und KI-Workloads, die Intel® AMX and Intel® AVX-512 nutzen.

Weniger Schritte bedeuten schnellere Verarbeitung

Mathematik kann sehr intelligent sein – und sehr elegant. Intel® AVX-512 auf skalierbaren Intel® Xeon® Prozessoren der 5. Generation verwendet viel intelligente, schöne Mathematik, um gängige Rechenoperationen zu verdichten und zu weniger Schritten zu kombinieren. Hier ist 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 35 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. Wie der Name bereits andeutet, erweitert Intel® AVX-512 die Anzahl der Register auf 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 … mit 8, 16, 24 bis 96 zählen.

Engines, die für leistungsstärkere KI weniger Strom brauchen

Da skalierbare Intel® Xeon® Prozessoren mit Intel® AI Engines weniger Hardware-Ressourcen benötigen, bieten sie eine leistungsstärkere und energieeffizientere Lösung für das Ausführen von KI-Workloads.

Skalierbare Intel® Xeon® Prozessoren mit integrierten Beschleuniger-Engines ermöglichen auch verbesserte Workload-Ergebnisse wie niedrigere TCO und eine höhere Kapitalrendite (Return on Investment, ROI) bei den anspruchsvollen KI-Workloads von heute.

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. Das bedeutet, dass sie für jede Software bereit und verfügbar ist, die sie nutzen kann. 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. Nach dem Herunterladen der neuesten Version wird TensorFlow* 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. Warum sollten Datenwissenschaftler:innen und KI-Entwickler:innen ihre Tools umschreiben und für Intel® AVX-512 neu kompilieren, wenn sie das für sich erledigen lassen können?

Unternehmen müssen heute mehr Workload-Leistung aus ihrer Infrastruktur herausholen, was ihnen mit höherer Energieeffizienz und zu niedrigeren Kosten gelingt. Die in die skalierbaren Intel® Xeon® Prozessoren integrierten spezialisierten Intel® AI Engines helfen dabei, das Maximum aus den KI-Workloads herauszuholen, die für ein Unternehmen am wichtigsten sind.

Erfahren Sie mehr darüber, was skalierbare Intel® Xeon® Prozessoren mit integrierten Beschleuniger-Engines bei den wichtigsten KI-Workloads Ihres Unternehmens leisten können.

Weitere Informationen

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

Intel® AVX-512 ›

Intel® AI Analytics Toolkit ›

Entwickeln auf Intel® Hardware und Software ›

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 ML.

Mehr erfahren ›

---

intel XEON

1. Basierend auf einer Marktmodellierung von Intel für die weltweit installierte Basis von Rechenzentrumsservern, auf denen KI-Inferenz-Workloads laufen, Stand Dezember 2022.
2. Siehe [A21] unter intel.com/processorclaims: Skalierbare Intel Xeon Prozessoren der 5. Generation. Die Ergebnisse können von Fall zu Fall abweichen.
3. Siehe [A19] unter intel.com/processorclaims: Skalierbare Intel® Xeon® Prozessoren der 5. Generation. Die Ergebnisse können von Fall zu Fall abweichen.
4. Siehe [A20] unter intel.com/processorclaims: Skalierbare Intel® Xeon® Prozessoren der 5. Generation. Die Ergebnisse können von Fall zu Fall abweichen.
5. Basierend auf einer internen Modellierung von Intel, Stand Dezember 2023.
6. Siehe [A22] unter intel.com/processorclaims: Skalierbare Intel® Xeon® Prozessoren der 5. Generation. Die Ergebnisse können von Fall zu Fall abweichen.
7. https://edc.intel.com/content/www/de/de/products/performance/benchmarks/vision-2022/, Session-Benchmark #41 und #42. Die Ergebnisse können von Fall zu Fall abweichen.

Hinweise und Disclaimer

Die Leistung variiert je nach Nutzung, Konfiguration und anderen Faktoren. Weitere Informationen finden Sie auf der Performance-Index-Website.

Die Leistungsergebnisse basieren auf Tests, die an den in den Konfigurationen angegebenen Daten durchgeführt wurden, und berücksichtigen möglicherweise nicht alle öffentlich verfügbaren Sicherheitsupdates. Konfigurationsdetails finden Sie im Backup. Kein Produkt und keine Komponente kann absolute Sicherheit bieten.

Ihre Kosten und Ergebnisse können variieren.

Weitere Informationen zu Workloads und Konfigurationen finden Sie unter „5th Generation Intel® Xeon® Scalable Processors“ auf www.intel.com/processorclaims.
Die Ergebnisse können von Fall zu Fall abweichen.

Intel® Technik kann entsprechend geeignete Hardware, Software oder die Aktivierung von Diensten erfordern.

Intel hat keinen Einfluss auf und keine Aufsicht über die Daten Dritter. Sie sollten andere Quellen heranziehen, um die Richtigkeit zu überprüfen.

Die Verfügbarkeit von Beschleunigern variiert je nach SKU. Weitere Produktdetails finden Sie auf der Seite Intel Produktspezifikationen.

Intel® Advanced Vector Extensions (Intel® AVX) bietet einen höheren Durchsatz bei bestimmten Prozessoroperationen. Bedingt durch veränderliche Charakteristika bei der Leistungsaufnahme kann die Verwendung von Intel® AVX-Befehlen folgende Auswirkungen haben: a) einige Teile arbeiten mit einer geringeren als der Nennfrequenz und b) einige Teile mit Intel® Turbo-Boost-Technik 2.0 erreichen keine bzw. nicht die maximale Turbo-Taktfrequenz. Die Leistung kann je nach Hardware, Software und Systemkonfiguration unterschiedlich ausfallen. Mehr erfahren Sie unter https://www.intel.de/content/www/de/de/architecture-and-technology/turbo-boost/intel-turbo-boost-technology.html.

Intel verpflichtet sich zur Achtung der Menschenrechte und der Vermeidung der Mittäterschaft bei Menschenrechtsverletzungen. Weitere Informationen finden Sie in den Globalen Menschenrechtsprinzipien von Intel. Die Produkte und Software von Intel sind ausschließlich für die Nutzung in Anwendungen vorgesehen, die keine Verletzung international anerkannter Menschenrechte darstellen oder zu einer Verletzung dieser Rechte beitragen.

Intel® Technik kann geeignete Hardware, Software oder die Aktivierung von Diensten erfordern.