Whitepaper

Rechenzentren

# Aus der Datenflut Erkenntnisse gewinnen

**Angesichts stetig wachsender Datenvolumen stehen Unternehmen vor der Herausforderung, daraus aussagekräftige Erkenntnisse für die Optimierung ihres Geschäftsbetriebs zu gewinnen. Erfahren Sie, wie die Einführung von Künstliche-Intelligenz- und Analytics-Projekten gelingt.**

## Kurzfassung

Riesige Datenmengen fließen heutzutage durch Unternehmen und Behörden. Die Herausforderung besteht darin, echten Nutzen daraus zu ziehen – unter Berücksichtigung ihrer Fragmentierung, ihres Umfangs und Speicherorts.

Künstliche Intelligenz (KI) und Analytics spielen eine wichtige Rolle bei der Gewinnung von Erkenntnissen in größerem Umfang, vorausgesetzt, dass die Daten zuerst konsolidiert und bereinigt werden.

In diesem Whitepaper wird erklärt, wie sich KI und Analytics in vier Schritten einführen lassen. Als erster Schritt wird ermittelt, wie diese Technologien dem Unternehmen helfen können.

Der nächste Schritt besteht darin, die Daten zu sammeln, zu konsolidieren und zu bereinigen. Unternehmen benötigen Speichermedien, die die Anforderungen ihrer Analytics- und KI-Workloads hinsichtlich Performance und Kapazität erfüllen.

Der dritte Schritt ist die Entwicklung der Lösung. Dazu gehört auch die Wahl der geeigneten Hardware. Die skalierbaren Intel® Xeon® Prozessoren verfügen über eine integrierte KI-Beschleunigung und ermöglichen Unternehmen durch den Rückgriff auf eine vertraute Architektur und vorhandenes Know-how die Einführung von KI. Dieser Artikel erläutert einige der Software-Lösungen, Frameworks und Toolkits, die Intel bereitstellt, um die Entwicklung und Performance von KI- und Analytics-Lösungen zu beschleunigen.

Der letzte Schritt ist der Einsatz der Lösung. Dieses Paper stellt einige der Ergebnisse vor, die Unternehmen durch den Einsatz ihrer Lösungen auf Intel® Architektur in der Cloud und am Edge erzielen konnten.

## Von Daten zu Erkenntnissen

Viele Unternehmen und Behörden sind von der anfallenden Datenmenge überwältigt. Weniger als 10 % der Daten sind praktisch verwertbar[1], wodurch der Großteil der Daten keinen Mehrwert für das Unternehmen bietet.

Die Datenvolumen werden außerdem kräftig wachsen. Juniper Research prognostiziert, dass es bis 2022 50 Milliarden vernetzte IoT-Sensoren und -Geräte geben wird.[2] Laut Schätzungen von IDC werden bis zum Jahr 2025 jährlich 175 Zettabytes an Daten generiert, erfasst und repliziert werden.[3] Das ist eine massive Steigerung im Vergleich zu den 33 Zettabytes des Jahres 2018 (ein Zettabyte entspricht einer Billion Gigabytes bzw. einer Milliarde Terabytes).

Daten haben enormes Potenzial, ein Unternehmen zu transformieren und einen verborgenen Mehrwert zu erschließen. Sie können dazu genutzt werden, die Service- und Produktqualität zu steigern, die Kosten zu optimieren und die Planungsgenauigkeit zu verbessern. Um dies erreichen zu können, müssen Unternehmen zuvor einige Hürden überwinden:

- **Konsolidierung von „Datenpfützen“**. In deutlichem Gegensatz zu „Datenseen“, in denen Daten aus dem gesamten Unternehmen zusammenfließen, sind Datenpfützen Datenbestände, die ausschließlich einer bestimmten Geschäftsfunktion dienen. Die Geschäftsprozesse innerhalb dieser Funktion mögen zwar hochoptimiert sein, aber diese Daten zu anderen Geschäftsfunktionen zu transferieren, ist mit hohem Arbeitsaufwand

**Inhaltsverzeichnis**

verbunden. Solche Datensilos erschweren es, optimale Entscheidungen für das Unternehmen als Ganzes zu treffen. Das Marketing benötigt beispielsweise Lager- und Produktionsdaten, um entscheiden zu können, ob und wann eine Werbekampagne für ein bestimmtes Produkt gestartet werden soll.

- **Einsatz von Automatisierung, Analytics und KI**. Ausgehend von einem sehr hohen Ausgangsniveau wachsen die Datenvolumen weiterhin. Unternehmen benötigen deshalb eine skalierbare Lösung, um alle Daten erfassen und verarbeiten zu können. KI- und Analytics-Tools können dazu genutzt werden, Rohdaten in praktisch umsetzbare Erkenntnisse zu verwandeln. Automatisierung kann zur weiteren Rationalisierung des Unternehmens genutzt werden, indem sie es ermöglicht, dass Routine-Entscheidungen sofort getroffen und umgesetzt werden können.
- **Verarbeitung am Edge**. Mit zunehmenden Datenvolumen steigt auch der Druck auf die Bandbreite und das Netzwerk. Eine Lösungsmöglichkeit ist es, einen Teil der Daten dort zu verarbeiten, wo sie sich befinden. Das erspart die Übertragung der Daten und senkt die Latenz. Manche Daten werden immer in der Cloud liegen und manche zentralen Daten werden immer erforderlich sein, um unternehmensweite Erkenntnisse zu liefern. Daten von Sensoren können jedoch oft lokal verarbeitet werden. So sind kürzere Reaktionszeiten möglich und es werden weniger Daten über das Netzwerk gesendet. In der Fertigung kann Computer Vision beispielsweise für die Qualitätskontrolle genutzt werden, wobei der Bildstrom lokal verarbeitet wird. Das ermöglicht ein zeitnäheres Eingreifen, falls ein Problem auftrifft, und verringert die Abhängigkeit von der Netzwerk- und Cloud-Verfügbarkeit. Gleichzeitig könnten für vorausschauende Instandhaltung eingesetzte Sensordaten zur längerfristigen Analyse an die Cloud geschickt werden. Dabei werden Netzwerkkapazitäten je nach Verfügbarkeit genutzt.

Da Daten überall verfügbar sind und die Intelligenz dort verteilt ist, wo sie benötigt wird, besteht die Gefahr, dass sich die Komplexität der IT erhöht. Unternehmen benötigen eine einzelne IT-Architektur, die sich vom Edge bis zur Cloud skalieren lässt und neue sowie sich weiterentwickelnde KI-Workloads unterstützt.

## Definition künstlicher Intelligenz (KI)

KI ist die Fähigkeit von Maschinen, ohne explizite Programmierung aus Erfahrung zu lernen und Funktionen auszuüben, die typischerweise mit dem menschlichen Verstand in Verbindung gebracht werden. Es gibt vier Hauptkategorien von KI:

- **Überwachtes Lernen**, bei dem gekennzeichnete Daten verwendet werden, um den Modellen eine nützliche Funktion beizubringen;
- **Unüberwachtes Lernen**, bei dem Erkenntnisse aus ungekennzeichneten Daten gewonnen werden;
- **Halbüberwachtes Lernen**, das eine Kombination aus gekennzeichneten und ungekennzeichneten Daten nutzt; sowie
- **Bestärkendes Lernen**, bei dem der Algorithmus nach der Versuch-und-Irrtum-Methode lernt, die bestmögliche Entscheidung zu treffen.

Die Art und Weise, wie diese Lernmethoden implementiert werden, lässt sich in folgende Kategorien einteilen:

- **Maschinelles Lernen**, bei dem sich Algorithmen im Laufe der Zeit durch Integration von mehr Daten verbessern, aber möglicherweise Funktionen definiert werden müssen, mit denen der Algorithmus arbeitet; oder
- **Deep Learning**, das geschichtete neuronale Netze für die Lösung von Problemen nutzt, bei denen es schwierig ist, Funktionen zu definieren, wie zum Beispiel Computer Vision und Spracherkennung. Das neuronale Netz passt sich automatisch den Trainingsdaten an, mit denen es „gefüttert" wird, und findet heraus, was für den Mustervergleich relevant ist.

Maschinelles Lernen lässt sich für Anwendungen wie Klassifikation, Clusterbildung und Entscheidungsbäume nutzen. Deep Learning eignet sich für Bild- und Sprachverarbeitung, Empfehlungssysteme und maschinelle Übersetzung.

## Der Weg zu Analytics und künstlicher Intelligenz (KI)

Die Implementierung von Analytics und KI lässt sich in vier Hauptphasen unterteilen:

- Die **Findungsphase**, zu der die Planung und Vorbereitung gehören;
- Die **Datenphase**, in der die Daten gesammelt und für die Verarbeitung standardisiert werden;
- Die **Entwicklungsphase**, in der die Hardware-Architektur ausgewählt und die Lösung entwickelt wird; sowie
- Die **Einsatzphase**, in der die Lösung in Betrieb genommen wird und anfängt, nützlich zu sein.

### Findungsphase

Der erste Schritt auf dem Weg zu KI besteht darin, herauszufinden, welche Anwendungen von Nutzen für die Funktionen und Geschäftsbereiche eines Unternehmens sein könnten. Bahnbrechende KI-Technologien verleiten einen schnell dazu, sich auf das zu stürzen, was technisch möglich ist, anstatt sich auf das zu konzentrieren, was am nützlichsten ist.

Einer unserer Kunden erstellte mittels internem Brainstorming eine Auswahlliste von vielversprechenden Möglichkeiten, schaute sich genau an, was andere getan hatten und untersuchte die mehr als 100 Lösungen im Intel® AI Builders Program, einem Ökosystem von KI-Lösungsanbietern. Zur Priorisierung der vielversprechendsten Möglichkeiten schätzte das Unternehmen die Kosten der einzelnen Lösungen und stellte diese dem potenziellen geschäftlichen Wert grafisch gegenüber. Dabei zeigte sich, dass die Kosten mit der Komplexität korrelieren. Die Erstellung eines neuen Deep-Learning-Modells ist kostspieliger als die Nutzung eines existierenden, was wiederum teurer als ein einfacher Ansatz des maschinellen Lernens ist. Die siegreiche Lösung, von der man sich die höchste Rentabilität erwartete, war eine Anwendung zur Automatisierung der industriellen Fehlererkennung unter Wasser, die Deep-Learning-Bilderkennung nutzt.

Ein wichtiger Teil dieser Phase ist es, alle möglichen ethischen Aspekte und potenziellen unerwünschten Nebenwirkungen zu erwägen und ein Team mit Fachkompetenz und technischem Know-how zusammenzustellen. Intel bietet kostenloses Online-Training im Rahmen des Intel® AI Developer Program, um Entwickler bei der Weiterbildung im Bereich KI zu unterstützen.

### Datenphase

Es wurde bereits erläutert, welche Herausforderungen sich aus in Silos gespeicherten Daten ergeben. Sie können zu fragmentierten Lösungen und zu Schwierigkeiten bei der Integration neu hereinkommender Daten führen. Weitere Probleme können auch noch hinsichtlich Konsistenz, Genauigkeit und Duplizierung der Daten auftauchen.

Die Anwendung von maschinellem Lernen und Deep Learning oder auch bloß die Erstellung von Unternehmens-Dashboards kann äußerst schwierig sein, wenn die Daten über Geschäftsbereiche hinweg fragmentiert sind. Es ist zwar möglich, Analytics und KI innerhalb eines Datensilos laufen zu lassen, aber eine Konsolidierung der Daten in eine kleinere Anzahl von Datenseen kann vorteilhaft sein. Eine Konsolidierung erleichtert nicht nur KI und Analytics, sondern kann auch die Qualität der Daten verbessern. Das lässt sich erreichen, wenn man einen einzelnen genauen Eintrag für jedes Datensubjekt hat und Inkonsistenzen zwischen den Silos vermeidet. Sie kann außerdem Nutzer in die Lage versetzen, effektiver zu arbeiten, indem sie Zugang zu vollständigeren Daten erhalten. Dadurch können sie besser verstehen, wie sich ihre Arbeit auf die anderen Abteilungen auswirkt.

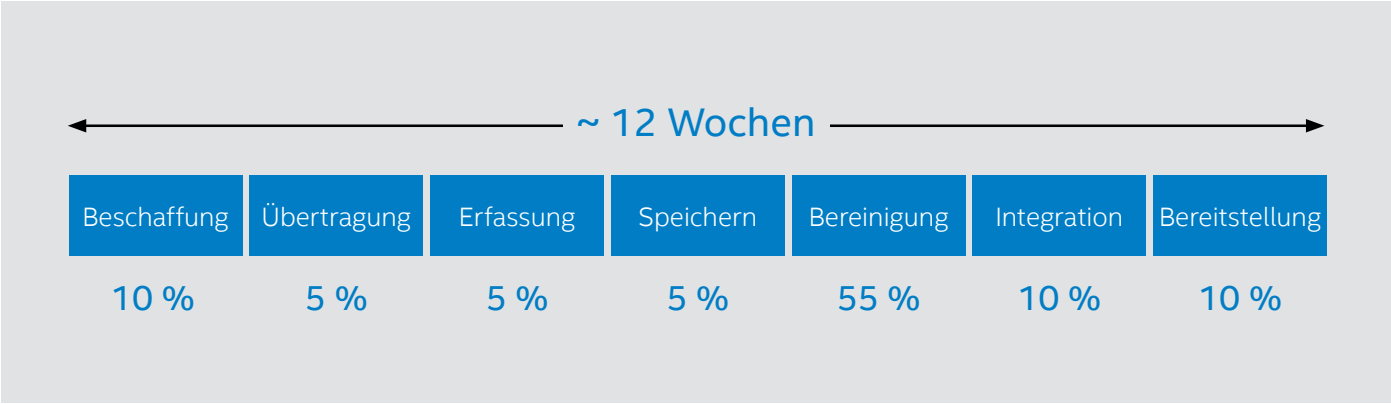


**Abbildung 1:** Bei einem KI-Projekt eines Intel Kunden benötigte Zeit, um Daten für die Verarbeitung zu beschaffen und aufzubereiten.

Bei Intel verfolgen wir die Strategie, unsere Datenpfützen in relevante Datenseen zusammenzufassen, und diese Daten allen unseren Geschäftsprozessen zugänglich zu machen. Wir haben 2017 ein sogenanntes Corporate Data Office (CDO) ins Leben gerufen, das eine „Data-First"-Einstellung in unserem gesamten Unternehmen fördern soll. Seine Aufgabe besteht darin, zu verbessern, wie wir mit unseren 315 Petabytes an Datenbeständen umgehen. Das CDO hat bereits einen Geschäftswert von 1,25 Milliarden US-Dollar erwirtschaftet – durch den Einsatz von Advanced Analytics für Vertrieb und Marketing, Produktdesign und -verbesserung und anderes.

Der zuvor erwähnte Kunde führte ein 12-wöchiges Programm zur Erfassung, Speicherung und Bereinigung seiner Daten durch. Streaming-Daten wurden von Drohnen gesammelt und auf CPUs verarbeitet. Das Unternehmen konnte aus einem großen Angebot an CPU-kompatiblen Tools wie Kafka*, Sqoop*, MQTT*, WS*, REST* und Flume* wählen. Blockbasierter Datenspeicher bot hohe Performance und Kubernetes* wurde für das Ressourcenmanagement eingesetzt. Bei größeren Clustern (mehr als 2000 Knoten) oder speziellen Anforderungen an das Ressourcenmanagement kann ein anderer Scheduler zum Einsatz kommen wie beispielsweise Slurm* für High-Performance-Computing (HPC) oder YARN* für Big Data. Eine Datenbereinigung ist häufig erforderlich, um sicherzustellen, dass die Daten vollständig und unbeschädigt sind. Außerdem kann es nötig sein, dass die Daten in eine für die Anwendung geeignete Größe, Struktur und Detailgenauigkeit umgewandelt werden müssen. Zu den Software-Tools, mit denen sich Daten auf der CPU bereinigen lassen, gehören Hadoop MapReduce*, Apache Storm* und Beam*.

Die Datenbereinigung nimmt manchmal viel Zeit in Anspruch. Wie Abbildung 1 zeigt, hat unser Kunde 55 % seiner Zeit mit dieser Tätigkeit verbracht. Andere Projekte wenden möglicherweise mehr Zeit in anderen Bereichen auf, aber der Prozess selbst wird in etwa gleich sein.

Die Konsolidierung der Daten und die dadurch möglichen leistungsintensiven Aktivitäten bedeuten eine Belastung für die Datenspeicherarchitektur. Mit dem persistenten Intel® Optane™ Speicher hat Intel eine bahnbrechende Innovation entwickelt, die Kapazitäten von Datenspeicher mit nahezu der Geschwindigkeit von Arbeitsspeicher bietet. Dadurch wurde ein komplett neuer Tier in der Speicherhierarchie geschaffen (siehe Abbildung 2). Persistenter Intel® Optane™ Speicher ermöglicht es, dass mehr Daten näher am Prozessor gespeichert werden, was die Datenanalyse beschleunigt und das zu niedrigeren Kosten pro Bit als bei DRAM. Unterstützung für persistenten Intel® Optane™ Speicher wurde bei der 2. Generation der skalierbaren Intel® Xeon® Prozessoren eingeführt.

Die neuen persistenten Intel® Optane™ Speichermodule der Produktreihe 200 bieten Gesamtspeicherkapazitäten von bis zu 4,5 TB pro Sockel. Außerdem besitzt die neueste Generation im Vergleich zur Vorgängergeneration eine durchschnittlich 25 % höhere Speicherbandbreite[4], wodurch die Anwendungsperformance gesteigert wird.

So wie DRAM ist persistenter Speicher byteweise adressierbar. Dadurch ist ein schnellerer Zugriff auf Daten als bei Flash-Speicher möglich. Im Vergleich zu einer konventionellen NAND-SSD bietet persistenter Intel® Optane™ Speicher der Produktreihe 200 einen über 225 Mal schnelleren Zugriff auf die Daten.[5]

Persistenter Speicher löst außerdem eines der von großen Datenmengen aufgeworfenen Probleme: Neustartzeiten. Gemeinsam mit Intel und HPE führte T-Systems einen Proof of Concept durch, der auf der vorherigen Generation des persistenten Speichers basierte. Diese Studie ergab, dass persistenter Speicher bei einem HPE Superdome Flex* Server die Neustartzeiten für große SAP HANA*-Instanzen um das 13,7-fache beschleunigte.[6] Während die typische Neustartzeit von SAP HANA* mehrere Stunden beträgt, wurde sie durch den Einsatz von persistentem Speicher auf unter 15 Minuten gedrückt. Dadurch verursachen Failovers und Neustarts zu Wartungszwecken nur mehr minimale Ausfallzeiten. T-Systems stellte fest, dass die Vorteile mit der Größe der Datenbank zunehmen.

Die Siemens AG erzielte einen ähnlichen Erfolg, der wiederum auf der vorherigen Generation des persistenten Speichers basierte. Siemens beschleunigte mit persistentem Intel® Optane™ Speicher das Laden seiner Datenbankdaten bei einem Neustart um das mehr als 15-fache.[7] Das Unternehmen betreibt eine der größten und komplexesten SAP HANA*-Datenbanken auf der ganzen Welt, die auf einer zweistelligen Anzahl von Servern läuft.

Für Daten mit weniger anspruchsvollen Anforderungen an die Latenz liefern Intel® Solid-State-Laufwerke (Intel® SSDs) die benötigte Performance. Die Intel® Optane™ SSDs der Produktreihe DC liefern bahnbrechende Performance bei Kapazitäten von 375 GB bis 1,5 TB. Die neuen Intel® 3D-NAND-SSDs der Produktreihen D7-P5500 und D7-P5600 bieten bis zu 40 %[8] weniger Latenz und bis zu 33 %[9] mehr Performance als die vorherige Generation.

Der Einsatz von Flash-Speicher senkt die Gesamtbetriebskosten (TCO; Total Cost of Ownership) im Vergleich zu proprietären Datenspeichersystemen, wodurch Unternehmen im Rahmen ihrer Budgets große Datenmengen bewältigen können. SysEleven ist ein Anbieter von zuverlässiger, effizienter und sicherer Infrastructure as a Service (IaaS) für Public- und Private-Cloud-Kunden. Anstatt auf eine teure proprietäre Shared-Storage-Lösung zu setzen, entschied sich SysEleven für Software-definierten Datenspeicher, um zwei rein auf Flash-Speicher basierte Speichertypen in einer Lösung zu vereinen: 50 Intel® SSDs der Produktreihe DC P4610 und 300 Intel® SSDs der Produktreihe DC P4500. So konnten die TCO um bis zu 75 % gesenkt werden.[10]

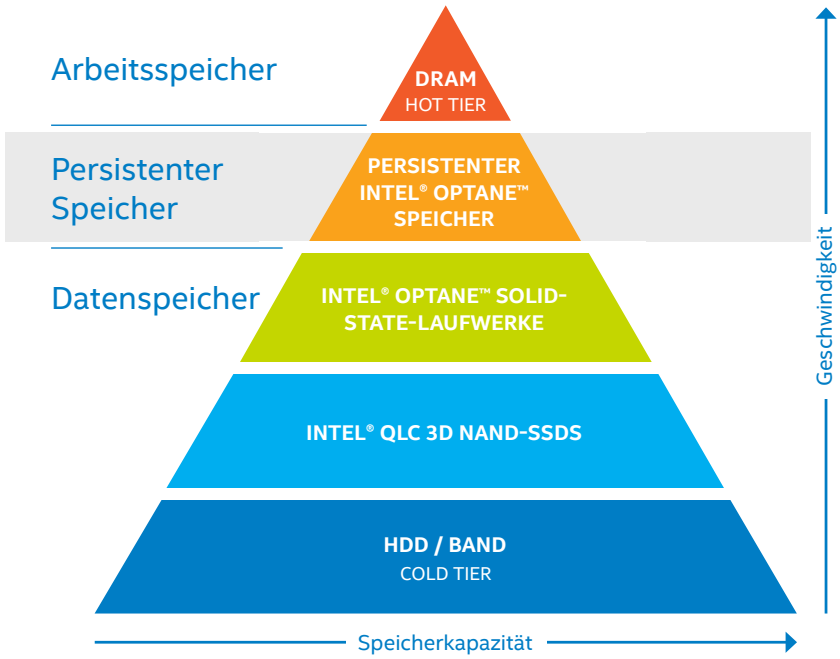


**Abbildung 2:** Persistenter Intel® Optane™ Speicher ist in der Speicherpyramide zwischen DRAM und SSDs angesiedelt und bietet die Geschwindigkeit von Arbeitsspeicher mit der Kapazität von Massenspeicher.

## Persistenter Speicher erhöht Dichte

Persistenter Intel® Optane™ Speicher beschleunigt nicht nur datenintensive Anwendungen, sondern erhöht auch die Dichte von virtuellen Maschinen (VMs), Containern und Anwendungen auf Servern. Das trägt dazu bei, die Gesamtbetriebskosten zu senken. Gleichzeitig werden auch die Folgenkosten im Rechenzentrum gesenkt, die dann entstehen, wenn mehr Server als nötig im Einsatz sind.

Der Online-Gaming-Markt bietet zwei aufschlussreiche Fallstudien zu diesem Thema. Geschwindigkeit ist hier wichtig, damit sichergestellt ist, dass die Spieler ein großartiges Gaming-Erlebnis genießen können. Aber die Branche ist auch kostensensibel.

GPORTAL bietet weltweit Game-Hosting-Services an. Das Unternehmen investierte in eine moderne Rechenzentrumsinfrastruktur, unter anderem in skalierbare Intel® Xeon® Prozessoren, persistenten Intel® Optane™ Speicher sowie Dell EMC PowerEdge* Server. Dadurch **konnte GPORTAL die Anzahl der Minecraft*-Spielinstanzen auf einem Server mehr als verdoppeln** – auf 500 Instanzen pro Server.[11] Erreicht wurde das, ohne eine maximale CPU-Auslastung von 60 % zu überschreiten. Das verbesserte die Ressourcenauslastung und senkte die Gesamtbetriebskosten für das Game-Hosting, wodurch GPORTAL zum ersten Mal in der Unternehmensgeschichte den Servermietpreis senken konnte.

Nitrado ist ein führender Anbieter von Gaming-Servern und Applikations-Hosting-Services. Durch Aufrüstung mit persistentem Intel® Optane™ Speicher **konnte die Anzahl der auf einem der Server von Nitrado laufenden Minecraft*-Instanzen um 175 % gesteigert werden**. Gleichzeitig wurde auch die CPU-Auslastung von 40 % auf 85 % gesteigert.[12] Die Server-Performance blieb hoch. Aufgrund der positiven Ergebnisse bietet Nitrado seinen Kunden nun Server an, die mit persistentem Intel® Optane™ Speicher ausgestattet sind.

### Entwicklungsphase

Nach der Aufbereitung der Daten wird es Zeit, die Lösung zu entwickeln. Der erste Schritt dieses Prozesses besteht in der Beschaffung der benötigten Rechenressourcen. Die KI- und Deep-Learning-Anforderungen der meisten Unternehmen lassen sich am besten durch CPUs erfüllen. Wenn GPUs zum Einsatz kommen, werden sich diese zeitweise im Leerlauf befinden, da sie keine allgemeinen Computing-Workloads durchführen können und es innerhalb des Unternehmens keinen ausreichenden Bedarf an Deep Learning gibt. Außerdem fallen bei der Beschaffung, Verwaltung und Entwicklung einer eigenen Computing-Architektur zusätzliche Kosten an.

Skalierbare Intel® Xeon® Prozessoren werden häufig für die Inferenz eingesetzt (die Analyse- und Entscheidungsfindungsphase der KI). Sie können auch für das Training eingesetzt werden (der Teil des Prozesses, bei dem das KI-Modell erstellt wird, typischerweise ein kleinerer Teil des Projekts als die Inferenz).

Bei Facebook basiert die Inferenz auf CPUs und wird für den Ranking-Algorithmus, das Foto-Tagging, die Fototextgenerierung, die Suche, die Sprachübersetzung, die Spam-Kennzeichnung und die Spracherkennung eingesetzt.[13] Außerdem werden CPUs für das Training des Ranking-Algorithmus und der Spam-Kennzeichnung und zusammen mit GPUs für das Training des Foto-Tagging-Modells genutzt.[14]

Unternehmen und Behörden, die noch vor der Einführung von KI stehen, können mit der Intel® Xeon® Architektur beginnen, die ihnen bereits vertraut ist. Zunächst sollte abgeklärt werden, ob es freie Kapazitäten in der bestehenden Infrastruktur gibt, die für einen Proof of Concept oder sogar für längerfristige Einsätze genutzt werden können. Viele Unternehmen werden feststellen, dass sie keine Beschleuniger benötigen und all ihre KI-Anforderungen mit der vertrauten Intel® Xeon® Architektur abdecken können.

Skalierbare Intel® Xeon® Prozessoren besitzen mehrere Funktionsmerkmale, die sie für KI-Workloads prädestinieren. Dazu gehört unter anderem die integrierte KI-Beschleunigung:

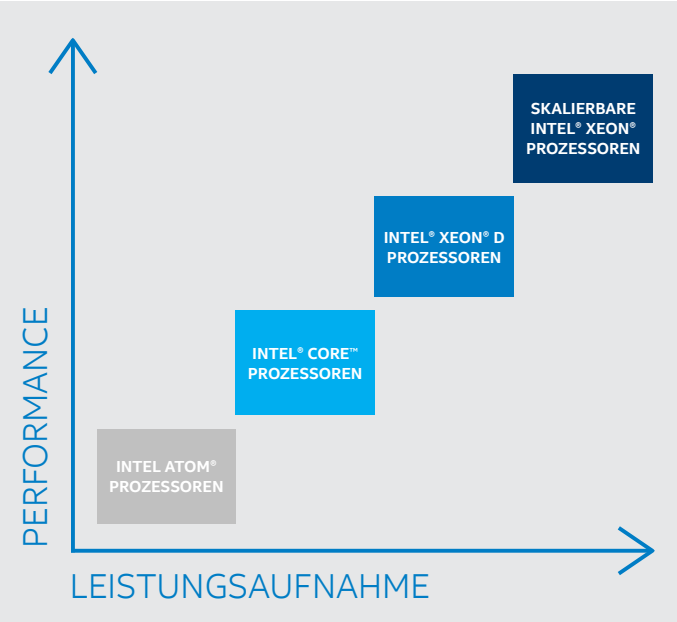


**Abbildung 3:** Intel bietet ein vollständiges Portfolio von Prozessoren, die die Anforderungen hinsichtlich Leistungsaufnahme und Performance für Einsätze am Edge erfüllen.

- Intel® Advanced Vector Extensions 512 (Intel® AVX-512) ermöglicht 512 Bit breite Vektor-Operationen. So können mehr Daten gleichzeitig verarbeitet werden, was KI- und Deep-Learning-Workloads potenziell beschleunigt. Verfügbar ist Intel® AVX-512 bei skalierbaren Intel® Xeon® Prozessoren.
- Intel® Deep Learning Boost (Intel® DL Boost) bietet neue Prozessorinstruktionen zur Beschleunigung von auf Intel® AVX-512 basierendem Deep Learning. Die neuesten Intel® Xeon® Platinum 9200 Prozessoren mit Intel® DL Boost ermöglichen im Vergleich zu einem Intel® Xeon® Platinum 8180 Prozessor aus dem Jahr 2017 einen 30 mal so hohen Inferenz-Durchsatz, wenn Caffe ResNet-50* für die Bilderkennung genutzt wird.[15]
- Eine geringere numerische Präzision lässt sich dazu nutzen, Deep-Learning-Operationen mit wenig oder gar keinem Genauigkeitsverlust zu beschleunigen. Skalierbare Intel® Xeon® Prozessoren unterstützen 8-Bit-Integer-Daten (INT8), die sich schneller als die früher üblichen 32-Bit-Gleitkommazahlen (FP32) verarbeiten lassen. Skalierbare Intel® Xeon® Prozessoren der 3. Generation verfügen über um 16-Bit-Gleitkommazahlen (BF16) ergänztes Intel® DL Boost. Sie bieten dadurch eine ähnliche Genauigkeit wie FP32 und eine schnellere Verarbeitung bei minimalen Softwareänderungen.
- Die Intel® Speed Select-Technologie (Intel® SST) ermöglicht eine bessere Kontrolle der CPU-Performance zur Optimierung der TCO. Durch den Einsatz von Intel® SST – Performance Profile lässt sich die CPU mit mehreren Konfigurationen auf wechselnde Workloads abstimmen. Mit Intel® SST – Base Frequency lässt sich die Grundtaktfrequenz für die wichtigsten Workloads zu den kritischsten Zeiten erhöhen.
- Application Device Queues (ADQ) sind eine Funktion, die verfügbar ist, wenn skalierbare Intel® Xeon® Prozessoren gemeinsam mit der Intel® Ethernet 800 Reihe genutzt werden. ADQs ermöglichen es, dass der Anwendungsverkehr in eine Reihe dedizierter Prozessor-Warteschlangen gefiltert wird. Das senkt die Latenz, erhöht den Durchsatz und sorgt bei äußerst Performance-abhängigen Anwendungen für eine verbesserte Leistungsstabilität.

Um die Beschaffung und den Einsatz von Servern für KI-Inferenz-Workloads zu vereinfachen, bieten Intel® Select Lösungen für KI-Inferenz eine einsatzbereite Plattformlösung für Inferenz mit niedriger Latenz und hohem Durchsatz, die auf den skalierbaren Intel® Xeon® Prozessoren basiert.

Für das Edge bietet Intel ein vollständiges Portfolio an Prozessoren, die auf einer gemeinsamen Architektur basieren. Diese Auswahl an Prozessoren ermöglicht es Unternehmen, ein optimales Verhältnis zwischen Performance und Stromverbrauch zu erreichen, um Inferenz dort durchzuführen, wo die Daten erzeugt werden (siehe Abbildung 3). Unternehmen, die auf Intel setzen, können vom Edge bis zur Cloud eine einzige Prozessorarchitektur und einen einzigen Software-Stack nutzen.

Manche Unternehmen erreichen irgendwann den Punkt, wo ihre Anwendungen so stark gewachsen sind, dass sich Beschleuniger aufgrund der Performance-Anforderungen lohnen. Der Einsatz von Intel® Field Programmable Gate Arrays (Intel® FPGAs) wie dem Intel® Stratix® 10 NX bietet die benötigte Beschleunigung innerhalb desselben Intel Software-Ökosystems. Der Intel® Stratix 10 NX ist der erste KI-optimierte FPGA von Intel für KI-Beschleunigung mit hoher Bandbreite und niedriger Latenz. Er bietet um bis zu 15 Mal so viel INT8-Rechenleistung für KI-Workloads als das Vorgängermodell Intel® Stratix 10 MX.[16]

Was den Workload selbst betrifft, ist es wichtig, das für diese Aufgabe richtige Tool zu wählen (siehe Tabelle 1). Planungsentscheidungen können mit Hilfe von Analytics-Lösungen getroffen werden, die die vergangene Nachfrage untersuchen, ohne dass dazu KI nötig wäre. Eine Prognose der Produktionsausbeute erfordert möglicherweise eine Machine-Learning-Lösung, die dazu in der Lage ist, alle Variablen zu korrelieren, die die Ausbeute beeinflussen und schwanken lassen könnten. Fehlererkennung kann eine Deep-Learning-Bilderkennungslösung nutzen, die mit einer großen Datenbank mit Bildern von guten und fehlerhaften Produkten trainiert wurde. Die Effizienz eines Roboterarms lässt sich möglicherweise durch eine Deep-Learning-Lösung verbessern, die bestärkendes Lernen nutzt. Eine solche Lösung kann es der Maschine ermöglichen, ihr Deep-Learning-Modell anhand des positiven oder negativen Feedbacks zu verbessern.

Intel bietet ein Software-Ökosystem, um die Entwicklung von KI- und Deep-Learning-Lösungen zu beschleunigen und effizienter zu gestalten. Dazu gehören die Intel® oneAPI Toolkits, die Tools zur Bereitstellung von Anwendungen auf verschiedenen Architekturen wie CPUs, GPUs, FPGAs und anderen Beschleunigern bieten. Die Intel® Distribution des OpenVINO™ Toolkit beschleunigt KI- und Deep-Learning-Anwendungen durch Nutzung jeglicher vorhandenen Beschleuniger. Das kann zum Beispiel die Intel® Movidius™ Vision Processing Unit sein, die Computer Vision am Edge beschleunigt. Zu den für Intel® Architektur optimierten Frameworks und Bibliotheken gehören ONNX Runtime*, PyTorch* und TensorFlow*. Über 100 Topologien (Typen von neuronalen Netzen, die für verschiedene Arten von Problemen geeignet sind) wurden für Intel® Xeon® Prozessoren optimiert. Intel bietet einen sogenannten Modell Zoo, der optimierte vortrainierte Modelle, Beispiel-Scripts und Tutorials für viele beliebte Open-Source-Machine-Learning-Modelle umfasst.

Analytics Zoo* ist eine vereinheitlichte Analytics- und KI-Plattform, die Apache Spark*, TensorFlow*, Keras* und BigDL* in einer integrierten Datenpipeline zusammenführt. Sie kann dazu genutzt werden, KI-Modelle auf tausende von Knoten für verteiltes Training oder verteilte Inferenz zu skalieren. Die Intel® Distribution für Python verbessert die Performance von Python* auf Intel® Architektur, wobei minimale oder gar keine Code-Änderungen erforderlich sind. Die beliebten Bibliotheken NumPy*, SciPy* und scikit-learn* werden von den Intel® Performance Libraries beschleunigt, die Teil der Intel® Distribution für Python sind.

| FRAGESTELLUNG | EMPFOHLENES TOOL |
|---|---|
| Welche Stückzahl sollte idealerweise produziert werden? | **Analytics**, um Angebot und Nachfrage der Vergangenheit zu verstehen. |
| Wie hoch wird die Ausbeute sein? | **Maschinelles Lernen**, um die mit der Ausbeute korrelierenden Variablen zu identifizieren. |
| Welche Teile haben sichtbare Fehler? | **Deep Learning**, um Fehler auf Bildern zu identifizieren. |
| Kann ein Roboterarm lernen, besser zu werden? | **Deep Learning,** um zu lernen und sich durch Feedback anzupassen. |

**Tabelle 1:** Welcher Ansatz eignet sich am besten?

Zur Unterstützung der Entwicklung und des Testens von Lösungen bietet die Intel® DevCloud kostenlosen Cloud-Zugang zu Intel® Hardware und Software. Das kann hilfreich sein, egal ob man einen Einsatz im Rechenzentrum oder am Edge plant oder die Beschleunigung mittels FPGA ausprobieren möchte.

Nochmals zurück zum zuvor erwähnten Kunden, der Drohnen zur Fehlererkennung unter Wasser einsetzt. Zu Beginn experimentierte das Unternehmen mit verschiedenen Deep-Learning-Topologien und der Abstimmung der Hyperparameter, die die Struktur des neuronalen Netzes definieren. Das Training war ein iterativer Prozess, der eine kontinuierliche Abstimmung der Hyperparameter und Bearbeitung der Eingabedaten beinhaltete. Zur Bestätigung, dass das trainierte Modell funktioniert, wurde es an einem Kontrolldatensatz getestet, um festzustellen, wie genau seine Inferenz ist. Für die Entwicklung des Modells, das Testen und die Dokumentationsphase benötigte der Kunde ebenfalls ungefähr 12 Wochen.

**Einsatzphase**
Der letzte Schritt ist der Einsatz der Lösung. Unternehmen, die Intel® Architektur einsetzen, können sich entscheiden, wo die Bereitstellung erfolgt – am Edge, vor Ort oder in der Cloud. Skalierbare Intel® Xeon® Prozessoren finden breite Verwendung in der Cloud und hardwarebeschleunigte Modelle von Azure Machine Learning bieten Zugang zu tiefen neuronalen Netzen, die von Intel® FPGAs beschleunigt werden.

## Die Vorteile von Analytics und künstlicher Intelligenz (KI) erkennen

Auch bei Intel selbst wurden KI und Analytics eingeführt und es zeigte sich, welche Vorteile das bringt. KI wurde zur Optimierung der Lieferkette eingesetzt, wodurch erhebliche Einsparungen erzielt wurden. Durch Optimierung des Lagerbestands mittels automatisierter Prognosen konnte die Planungsperiode von sechs Monaten auf eine Woche reduziert werden und ließen sich Einsparungen von 58 Millionen US-Dollar erzielen. Dank der Nutzung von Big Data, Dashboards und maschinellem Lernen konnten allein während des Pilotprojekts 23 Millionen US-Dollar an Stücklistenkosten gespart werden. Bei der Produktvalidierung konnten durch den Einsatz von KI 50 % mehr Probleme identifiziert werden – bei halben Kosten.[17]

Kunden von Intel profitieren erheblich von KI und Analytics, die auf Intel® Prozessoren basieren. AccuRad entschied sich für skalierbare Intel® Xeon® Prozessoren, um seine Bildgebungstechnik in der Cloud einzusetzen. Dadurch können Daten verschiedener medizinischer Geräte leichter gesammelt und verarbeitet werden. AccuRad entwickelte die @iMAGES Core Engine*, um Cloud-Computing zur Verarbeitung und Analyse von medizinischem Bildmaterial zu nutzen. „Unser System nutzte ursprünglich eine GPU-Karte für das Rendering und eine andere GPU-Karte für KI-Computing, während für die Verarbeitung von Geschäftsdaten ein Allzweckprozessor verwendet wurde. Die Kosten waren hoch und die Wartung war aufwändig. Nun müssen wir nur noch skalierbare Intel® Xeon® Prozessoren einsetzen und alles ist erledigt“, so Yedong Huang, Gründer von AccuRad. „Darum planen wir jetzt, alle bisher auf verschiedene Hardware-Plattformen verteilten Aufgaben auf die auf Intel® Architektur basierende Plattform zu migrieren.“

Intels Optimierungen für KI-Frameworks wie Caffe* und TensorFlow* steigern die Leistungsfähigkeit des KI-gestützten Diagnosesystems von AccuRad noch weiter. Bei dem für Intel® Technologien optimierten RFCN-Modell konnte die Leistung beim Zurechtschneiden und Zusammenführen um beinahe 30 % verbessert werden. Eine weitere Leistungssteigerung von 40 bis 50 %[18] wird durch die Optimierung der Multithreading-Implementierungslösung OpenMP* ermöglicht. Die Testdaten des Ersteinsatzes zeigen, dass der Intel® Xeon® Gold 6148 Prozessor bei der Ausführung eines RFCN-Modells mit Daten einer einzelnen Thoraxaufnahme die Bearbeitungszeit im Vergleich zu einer Standard-GPU um 10 %[19] reduziert.

Audi setzte im Rahmen eines Proof of Concept ein Machine-Learning-Modell am Edge ein, um den Qualitätskontrollprozess für die Schweißnähte seiner Fahrzeuge zu verbessern. Es basierte auf Intel® Xeon® Prozessoren, lässt sich aber problemlos in jede Richtung skalieren. Die Lösung lässt sich von Intel® Core™ Prozessoren auf Intel® Xeon® E Prozessoren und skalierbare Intel® Xeon® Prozessoren skalieren – ohne Änderungen an der Software. Audi hatte einen arbeitsintensiven manuellen Prozess, bei dem ein Fahrzeug pro Tag mittels Ultraschallsonden von einem Team aus 18 Ingenieuren inspiziert wurde. Dank maschinellem Lernen ist das Unternehmen nun in der Lage, 5000 Schweißnähte pro Fahrzeug zu überprüfen und das Ergebnis jeder Schweißnaht innerhalb von 18 Millisekunden zu ermitteln. Im Werk in Neckarsulm wurden die Arbeitskosten für die Inspektion bereits um 30 bis 50 % reduziert, während die Qualitätskontrolle stark verbessert werden konnte.

## Fazit

Intel bietet verschiedene Technologien, die von einem robusten Ökosystem von Software und Partnern unterstützt werden, um mit Analytics und KI beeindruckende Ergebnisse zu erzielen. In diesem Artikel wurde gezeigt, wie Kunden von Intel aus dem Einsatz von Analytics- und KI-Workloads auf Intel® Architektur echte Geschäftsvorteile ziehen konnten. Wird der oben skizzierte, aus vier Schritten bestehende Prozess befolgt, sollte es möglich sein, dem Beispiel dieser Kunden zu folgen.

Finden Sie die passende Lösung für Ihr Unternehmen. Wenden Sie sich bitte an Ihren Ansprechpartner bei Intel oder **besuchen Sie Intel online**.

## Weitere Informationen

- **Skalierbare Intel® Xeon® Prozessoren**
- **Intel® Edge-Computing-Lösungen**
- **Persistenter Intel® Optane™ Speicher**
- **Intel® AI Builders Ökosystem**
- **Intel® AI Developer Program**

intel®

[1] InformationWeek, 17. Juli 2019, „What's Holding Back Edge Computing for Enterprises" von Kelly Herrell

[2] Juniper Research, Juni 2018, „IoT Connections to Grow 140% to Hit 50 Billion by 2022 as Edge Computing Accelerates ROI", Pressemitteilung

[3] IDC, November 2018, „The Digitization of the World from the Edge to Core" von David Reinsel, John Gantz und John Rydning

[4] Ausgangskonfiguration: 1 Knoten, 1 x Intel® Xeon® 8280L Prozessor @ 2,7 GHz (28 Kerne) auf Neon City mit einzelnem persistenten Speichermodul (6 x 32 GB DRAM; 1 x {128 GB, 256 GB, 512 GB} persistentes Intel® Optane™ Speichermodul der Produktreihe 100 mit 15 W), ucode Rev.: 04002F00 auf Fedora* 29 Kernel 5.1.18-200.fc29.x86_64 und MLC Version 3.8 im App-Direct-Modus. Quelle: 2020ww18_CPX_BPS_DI. Von Intel am 27. April 2020 getestet Neue Konfiguration: 1 Knoten, 1 x Intel® Xeon® CPX6 Vorserien-Prozessor @ 2,9 GHz (28 Kerne) auf Cooper City mit einem einzelnen persistenten Speichermodul (6 x 32 GB DRAM; 1 x {128 GB, 256 GB, 512 GB} persistentes Intel® Optane™ Speichermodul der Produktreihe 200 mit 15 W), Vorserien-ucode auf Fedora* 29 Kernel 5.1.18-200.fc29.x86_64 und MLC Version 3.8 im App-Direct-Modus. Quelle: 2020ww18_CPX_BPS_BG. Von Intel am 31. März 2020 getestet.

[5] Die Leselatenz von persistentem Intel® Optane™ Speicher beträgt im Leerlauf 340 Nanosekunden. Die Leselatenz von Intel® SSDs der Produktreihe DC P4610 mit TLC-3D-NAND-Technologie beträgt im Leerlauf 77 Mikrosekunden.

[6] Die Tests von T-Systems wurden am 18. März 2019 durchgeführt. Referenzkonfiguration – Hardware: HPE Superdome Flex* Server mit 4 x CPU-Sockel (Intel® Xeon® Platinum Beta 8276M Prozessor @ 2,2 GHz; Arbeitsspeicher = 4 x 6 x 256 GB persistenter Intel® Optane™ DC Speicher (6 TB) – DEAKTIVIERT und 4 x 6 x 64 GB DDR4-Speicher (1,5 TB) für eine Gesamtspeicherkonfiguration von 1,5 TB. Software: Datenbank: SAP-S/4HANA* Datenbank mit 4 TB im App-Direct-Modus; Betriebssystem: Standard SUSE Linux Enterprise Server* 12 Servicepack 4, Microcode = 0xb00002e, Kernel = Linux 4.12.14-95.16, auf Standard NetApp cDOT* basierender Datenspeicher für Persistenz; Neustartzeit einer SAP HANA* 2.0 SPS4 Rev. 40 Installation mit BW-Benchmark-Workload: 10.248 Sekunden (ca. 2,85 Stunden). Proof-of-Concept-Konfiguration – HPE Superdome Flex* Server mit 4 x CPU-Sockel (Intel® Xeon® Platinum Beta 8276M Prozessor @ 2,2 GHz; Arbeitsspeicher = 4 x 6 x 256 GB persistenter Intel® Optane™ DC Speicher (6 TB) und 4 x 6 x 64 GB DDR4-Speicher (1,5 TB) für eine Gesamtspeicherkonfiguration von 7,5 TB. Software: SAP-S/4HANA* Datenbank mit 4 TB im App-Direct-Modus; Betriebssystem: Standard SUSE Linux Enterprise Server* 12 Servicepack 4, Microcode = 0xb00002e, Kernel = Linux 4.12.14-95.16, auf Standard NetApp CDOT* basierender Datenspeicher für Persistenz; Neustartzeit einer SAP HANA* 2.0 SPS4 Rev. 40 Installation mit BW-Benchmark-Workload: 748 Sekunden (ca. 12,47 Minuten).

[7] Die angeführten Daten stammen aus einer internen Verifizierung und Testreihe von Siemens vom April 2019. Wenn Sie weitere Einzelheiten über diese Tests erfahren möchten, wenden Sie sich bitte an Siemens.

[8] Quelle: Intel. Verglichen wurden die Datenblattwerte für die Latenz bei einer Warteschlangentiefe von 1 bei wahlfreien Schreibzugriffen mit 4-KB-Blöcken von Intel® SSDs der Produktreihe D7-P5500 mit 7,68 TB und Intel® SSDs der Produktreihe DC P4510 mit 8 TB, beide auf PCIe* 3.1. Die gemessene Latenz betrug 15 µs für die Produktreihe D7-P5500 bzw. 25 µs für die Produktreihe DC P4510. Die Leistung beider Laufwerke wurde mit FIO Linux CentOS* 7.2 Kernel 4.8.6 mit einer Übertragungsgröße von 4 KB (4096 Bytes) bei einer Warteschlangentiefe von 1 (1 Worker) gemessen. Die Messungen erfolgten auf einer vollen LBA-Laufwerksspanne (Logical Block Addressing), sobald der Workload einen stabilen Status erreicht hatte, einschließlich aller zum normalen Betrieb und für die Datenzuverlässigkeit notwendigen Hintergrundaktivitäten. Der Energiesparmodus war auf PM0 gesetzt. Unterschiede in der Hardware, Software oder Konfiguration des Systems können die tatsächliche Leistung beeinflussen. Intel geht davon aus, dass bei der Messung der Daten mehrerer Laufwerke ein gewisses Maß an Abweichungen besteht.

[9] Quelle: Intel. Verglichen wurden die Datenblattwerte für die Latenz bei einer Warteschlangentiefe von 256 bei wahlfreien Schreibzugriffen mit 4-KB-Blöcken von Intel® SSDs der Produktreihe D7-P5500 mit 7,68 TB und Intel® SSDs der Produktreihe DC P4510 mit 8 TB, beide auf PCIe* 3.1. Die gemessene Performance betrug 854.000 IOPS für die Produktreihe D7-P5500 bzw. 641800 für die Produktreihe DC P4510. Die Leistung beider Laufwerke wurde mit FIO Linux CentOS* 7.2 Kernel 4.8.6 mit einer Übertragungsgröße von 4 KB (4096 Bytes) bei einer Warteschlangentiefe von 64 (4 Worker) gemessen. Die Messungen erfolgten auf einer vollen LBA-Laufwerksspanne (Logical Block Addressing), sobald der Workload einen stabilen Status erreicht hatte, einschließlich aller zum normalen Betrieb und für die Datenzuverlässigkeit notwendigen Hintergrundaktivitäten. Der Energiesparmodus war auf PM0 gesetzt. Unterschiede in der Hardware, Software oder Konfiguration des Systems können die tatsächliche Leistung beeinflussen. Intel geht davon aus, dass bei der Messung der Daten mehrerer Laufwerke ein gewisses Maß an Abweichungen besteht.

[10] Die angenommenen Kosteneinsparungen basieren auf Schätzung von SysEleven vom 1. Februar 2020. Das neue, rein auf Flash-Speicher basierte Shared-Storage-System ersetzt mehrere proprietäre Datenspeichersysteme, die vom OEM nicht mehr unterstützt werden. Die Entscheidung von SysEleven, in eine Software-definierte Storage-Lösung auf Basis handelsüblicher Intel® Hardware zu investieren, stützte sich sowohl auf Preis- als auch auf Leistungsaspekte. Schätzungen zufolge würden die Gesamtbetriebskosten (TCO) bei dieser Lösung innerhalb von fünf Jahren um 75 % geringer ausfallen als bei einer aktualisierten, rein auf Flash-Speicher basierten Version der vorhandenen Storage-Lösung (einschließlich Support- und Lizenzkosten).

[11] Ausgangskonfiguration: Dell EMC PowerEdge* R640 Server; 2 x Intel® Xeon® Gold 6154 Prozessor @ 3,0 GHz (18 Kerne/36 Threads); 768 GB DDR4; BIOS = 2.3.10; Betriebssystem = Linux; Ergebnisse: 180 Minecraft*-Spielinstanzen. DUT-Konfiguration: Dell EMC PowerEdge* R640 Server; 2 x Intel® Xeon® Platinum 8268 Prozessor @ 2,9 GHz (24 Kerne/48 Threads); 12 x 32 GB DDR4 + 12 x 128 GB persistente Intel® Optane™ DC Speichermodule; BIOS = 2.3.10; Betriebssystem = Linux; Ergebnisse: 500 Minecraft*-Spielinstanzen. Die Tests von GPORTAL wurden am 5. Dezember 2019 durchgeführt. Die Leistungsergebnisse basieren auf Tests, die zu dem in den Konfigurationen angegebenen Datum durchgeführt wurden, und berücksichtigen möglicherweise nicht alle öffentlich erhältlichen Sicherheitsupdates. Weitere Einzelheiten finden Sie in den veröffentlichten Konfigurationsdaten. Kein Produkt und keine Komponente bietet absolute Sicherheit.

[12] Die Tests von Nitrado wurden am 7. Februar 2019 durchgeführt. Konfiguration nur mit DRAM: Dual-Sockel Intel® Xeon® Gold 6148 Prozessor (8 x 64 GB DDR4-2666 DRAM), insgesamt installierter Arbeitsspeicher = 512 GB. Verfügbarer Systemspeicher = 512 GB. Anzahl der Minecraft*-Instanzen: 182. CPU-Auslastung: 40 %. Konfiguration mit DRAM + persistentem Intel® Optane™ DC Speicher: Dual-Sockel Intel® Xeon® Gold 6252 Prozessor (12 x 128 GB (1,5 TB) persistenter Intel® Optane™ DC Speicher und 12 x 16 GB (192 GB) DDR4-2600 DRAM), insgesamt installierter Arbeitsspeicher = 1.692 GB. Verfügbarer Systemspeicher = 1.536 GB. Anzahl der Minecraft*-Instanzen: 500. CPU-Auslastung: 85 %. Die endgültigen Ergebnisse wurden aus den Testdaten von Nitrado hochgerechnet.

[13] https://research.fb.com/wp-content/uploads/2017/12/hpca-2018-facebook.pdf

[14] https://research.fb.com/wp-content/uploads/2017/12/hpca-2018-facebook.pdf

[15] **1-fache Inferenzdurchsatz-Verbesserung im Juli 2017 (Ausgangskonfiguration):** Getestet von Intel am 11. Juli 2017. Plattform: Dual-Sockel Intel® Xeon® Platinum 8180 Prozessor @ 2,5 GHz (28 Kerne), HT deaktiviert, Turbo deaktiviert, Scaling-Governor festgelegt auf „Performance“ über intel_pstate-Treiber, 384 GB DDR4 2666-ECC RAM. CentOS* Linux, Release 7.3.1611 (Core), Linux-Kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD der Produktreihe DC S3700 (800 GB, 2,5", SATA 6,0 Gbit/s, 25-nm-Technologie, MLC). Leistung gemessen mit: Umgebungsvariablen: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU-Frequenz festgelegt mit: cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe*: (http://github.com/intel/caffe/), Revision f96b759f71b2281835f690af267158b82b150b5c. Inferenz gemessen mit „caffe time --forward_only“-Befehl, Training gemessen mit „caffe time“-Befehl. Für die „ConvNet“-Topologien wurde ein Test-Datensatz verwendet. Für andere Topologien wurden Daten im lokalen Datenspeicher gespeichert und vor dem Training im Arbeitsspeicher zwischengespeichert. Topologie-Spezifikation von https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50) und https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet-Benchmarks; die Dateien wurden aktualisiert und verwenden das neuere prototxt-Format von Caffe*, sind aber funktional gesehen äquivalent). Intel C++ Compiler-Version 17.0.2 20170213, Intel® MKL Small Libraries, Version 2018.0.20170425. Caffe* ausgeführt mit „numactl -l“.

**30-fache Verbesserung des Inferenzdurchsatzes mit Cascade-Lake-AP im Vergleich zur Referenzkonfiguration:** Von Intel am 26.2.2019 getestet. Plattform: Dual-Sockel Intel® Xeon® Platinum 9282 Prozessor (56 Kerne) auf Dragon Rock, HT aktiviert, Turbo aktiviert, insgesamt 768 GB Arbeitsspeicher (24 Steckplätze / 32 GB / 2933 MHz), BIOS: SE5C620.86B. 0D.01.0241.112020180249, CentOS* 7 Kernel 3.10.0-957.5.1.el7. x86_64, Deep-Learning-Framework: Intel® Optimierung für Caffe* Version: https://github.com/intel/caffe d554cbf1, ICC 2019.2.187, MKL-DNN-Version: 0.17 (Commit-Hash: 830a10059a018cd-2634d94195140cf2d8790a75a, Modell https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS = 64, keine Datenebene – synthetische Daten: 3x224x224, 56 Instanzen / Dual-Sockel, Datentyp: INT8 im Vergleich mit: Von Intel getestet am 11. Juli 2017: Dual-Sockel Intel® Xeon® Platinum 8180 Prozessor @ 2,5 GHz (28 Kerne), HT deaktiviert, Turbo deaktiviert, Scaling-Governor festgelegt auf „Performance“ über intel_pstate-Treiber, 384 GB DDR4 2666-ECC RAM. CentOS* Linux, Release 7.3.1611 (Core), Linux-Kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD der Produktreihe DC S3700 (800 GB, 2,5", SATA 6,0 Gbit/s, 25-nm-Technologie, MLC). **Leistung gemessen mit**: Umgebungsvariablen: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU-Frequenz festgelegt mit: cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe*: (http://github.com/intel/caffe/), Revision f96b759f71b2281835f690af267158b82b150b5c. Inferenz gemessen mit „caffe time --forward_only“-Befehl, Training gemessen mit „caffe time“-Befehl. Für die „ConvNet“-Topologien wurde ein Test-Datensatz verwendet. Für andere Topologien wurden Daten im lokalen Datenspeicher gespeichert und vor dem Training im Arbeitsspeicher zwischengespeichert. Topologie-Spezifikation von https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50), Intel® C++ Compiler-Version 17.0.2 20170213, Intel® MKL Small Libraries, Version 2018.0.20170425. Caffe* ausgeführt mit „numactl -l“.

[16] Auf Basis interner Schätzwerte von Intel. Siehe https://www.intel.de/content/www/de/de/products/programmable/fpga/stratix-10/nx.html

[17] Weitere Informationen über diese Ergebnisse finden Sie im 2018-2019 Intel IT Annual Performance Report

[18] Die angeführten Daten stammen aus internen Performancetests von AccuRad, die von Intel unterstützt wurden

[19] Die angeführten Daten stammen aus internen Performancetests von AccuRad, die von Intel unterstützt wurden

In Leistungstests verwendete Software und Workloads können speziell für die Leistungseigenschaften von Intel® Mikroprozessoren optimiert worden sein.

Leistungstests wie SYSmark* und MobileMark* werden mit spezifischen Computersystemen, Komponenten, Softwareprogrammen, Operationen und Funktionen durchgeführt. Jede Veränderung bei einem dieser Faktoren kann abweichende Ergebnisse zur Folge haben. Als Unterstützung für eine umfassende Bewertung Ihrer geplanten Anschaffung sollten Sie zusätzliche Informationen und Leistungstests heranziehen – auch im Hinblick auf die Leistung des betreffenden Produkts in Verbindung mit anderen Produkten. Ausführlichere Informationen finden Sie unter: https://www.intel.de/benchmarks.

Die Leistungswerte basieren auf Tests, die mit den in den Konfigurationen angegebenen Daten durchgeführt wurden und spiegeln möglicherweise nicht alle öffentlich verfügbaren Updates wider. Kein Produkt und keine Komponente bietet absolute Sicherheit.

Weitere Informationen über die Leistungs- und Optimierungsoptionen bei Intel Softwareprodukten finden Sie unter https://software.intel.com/articles/optimization-notice.

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

Intel® Technologien können entsprechend geeignete Hardware, Software oder die Aktivierung von Diensten erfordern.

Kosten und Ergebnisse können variieren.