DELL Technologies

AMD
together we advance_

2024 | Whitepaper

KI zur Unterstützung von Unternehmen: Das Zeitalter der Wahl ist gekommen

Inhaltsverzeichnis

DELL Technologies

AMD
together we advance_

2024 | Whitepaper

Die Chance, ganze Branchen dank KI zu transformieren

Heute gibt es keine größere Chance, Ihr Unternehmen dank KI für die Zukunft der Innovation zu transformieren. Laut von Accenture Vision Technology 2023 erhobenen Daten sind 98 % der Führungskräfte weltweit der Meinung, dass KI-basierte Modelle in den nächsten drei bis fünf Jahren eine wichtige Rolle in den Strategien ihres Unternehmens spielen werden.[1]

KI ist für Unternehmen in Branchen wie Einzelhandel, Gesundheitswesen und Finanzdienstleistungen mittlerweile unglaublich nützlich geworden, da sie in der Lage ist, die Effizienz von Aufgaben zu steigern, Innovationen voranzutreiben und Entscheidungsprozesse zu verbessern. Trotz dieser Vorteile gibt es jedoch immer noch eine wahrgenommene Einstiegshürde bei der Integration von KI, da sich einige weit verbreitete Missverständnisse hartnäckig halten.

Sie benötigen ein KI-Entwicklungsteam, um loszulegen:

Fachwissen im Bereich Data Science ist zwar nach wie vor wertvoll, um fortschrittliche KI-Lösungen zu entwickeln und die zugrunde liegenden Prinzipien zu verstehen, jedoch keine Voraussetzung mehr. Es gibt eine Vielzahl von nutzerfreundlichen KI-Tools, Plattformen wie Hugging Face und aufgabenspezifischen Modellen, die einen Großteil der Komplexität bei der Entwicklung von KI-Lösungen beseitigen.

Sie müssen zig Millionen für Hardware ausgeben, um Ergebnisse zu erzielen:

Dieser Irrglaube missachtet die Vielfalt der heute verfügbaren KI-Ressourcen enorm. Obwohl diese allgemein bekannten Ressourcen oft leistungsstark sind und gut unterstützt werden, sind sie nicht immer die am besten geeignete oder kostengünstigste Wahl für jedes Unternehmen.

Sie müssen sich abmühen, um Accelerators zu beschaffen:

Accelerators liefern zwar bei hohen KI-Workloads eine hervorragende Leistung, aber Unternehmen benötigen möglicherweise gar nicht so viel Rechenleistung für ihre KI-Anwendungen. Es ist auch einfach nicht realistisch, einen übermäßig langen Zeitraum lang zu warten, um Zugang zu marktführenden Accelerators zu erhalten. In vielen Fällen können KI-optimierte CPUs bereits die Leistung und Effizienz liefern, die für die Erstellung von KI-gestützten Analysen und Entscheidungen in Echtzeit erforderlich sind – und sind dabei eine viel kostengünstigere und anpassungsfähigere Lösung.

1 Accenture, 30. März 2023, „Accenture Technology Vision 2023: Generative AI to Usher in a Bold New Future for Business, Merging Physical and Digital Worlds“, https://newsroom.accenture.com/news/2023/accenture-technology-vision-2023-generative-ai-to-usher-in-a-bold-new-future-for-business-merging-physical-and-digital-worlds

Zum Glück entwickelt sich die KI-Landschaft ständig weiter. Gemeinsam räumen **Dell** und **AMD** mit diesen Mythen auf, indem sie KI-Technologien und -Tools einem breiteren Nutzendenkreis zugänglich machen – mit einer durchgängigen Infrastruktur, die auf die heutigen KI-Anforderungen ausgelegt ist.

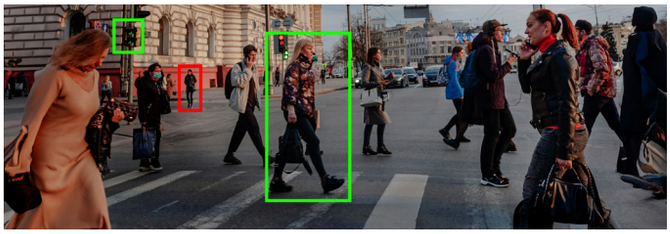
Steigen Sie mit einem bereits optimierten Modell, einem zuverlässigen Software-Stack und einem vielseitigen Hardwaresystem ein, die alle dank der Partnerschaft von Dell und AMD frei verfügbar sind. Der Zugang zu immer knapper werdenden Accelerators, einer beträchtlichen Gruppe qualifizierter KI-Fachkräften oder Ressourcen für die Bereitstellung riesiger Cloud-Cluster ist keine Voraussetzung mehr für die Nutzung von KI.

Die Zusammenarbeit von **Dell** und **AMD** bietet ein einheitliches Ökosystem aus Hardware und Software, mit dem die Entwicklung umfassende KI-Lösungen erstellen kann, die unkompliziertes und effizientes Transfer Learning, Feinabstimmung und Inferencing umfassen. Dank der Unterstützung von **Hugging Face** verfügen wir jetzt über ein wachsendes Portfolio an Modellen, die auf Dell PowerEdge-Servern mit AMD EPYC™-Prozessoren oder AMD Instinct™ MI300X Accelerators ausgeführt werden, sodass Entwicklungsfachkräfte Optimierungen vornehmen, Transfer Learning anwenden und Inferencing bereitstellen können. Die Investitionen in AMD ROCm™ und AMD ZenDNN™ sowie die Partnerschaften mit den Runtime-Frameworks PyTorch, Tensorflow und ONNX läuten die Demokratisierung der KI-Anwendungsentwicklung ein. Das folgende Stapeldiagramm zeigt die Komponenten, die das einheitliche KI-Ökosystem von Dell und AMD bilden.

KI in der Industrie

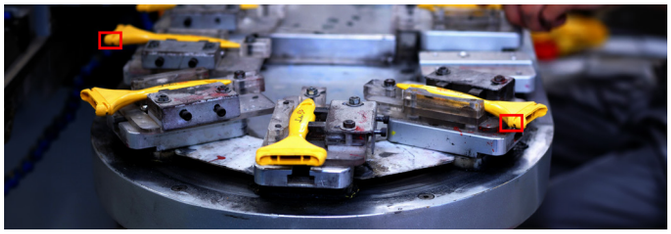
Mit der Diversifizierung der Ressourcen und dem Schwerpunkt auf Open-Source-Innovationen hält KI in vielen verschiedenen Branchen Einzug, darunter Kundenservice, Finanzen und Bankwesen, Gesundheitswesen und Einzelhandel, um nur einige zu nennen. In diesen Branchen ermöglicht KI Unternehmen, das Potenzial ihrer eigenen Daten zu erschließen und ihre KI-Workflows neu zu gestalten, indem sie die folgenden Schlüsselfunktionen in Angriff nehmen: Datenanalyse, Automatisierung, Personalisierung und vorausschauende Analysen. AMD ROCm- und ZenDNN-Bibliotheken beschleunigen diese KI-Workflows zusätzlich, um Ergebnisse nahezu in Echtzeit zu liefern.

Schauen Sie sich im Folgenden an, wie genau KI verschiedene Branchen beeinflusst.



Automobilindustrie

KI wird für die Objekterkennung, Fahrspurverfolgung und Entscheidungsfindung in autonomen Fahrzeugen verwendet. KI kann auch vorhersagen, wann eine Fahrzeugkomponente wahrscheinlich ausfallen wird, was eine proaktive Wartung ermöglicht und Ausfallzeiten reduziert.



Fertigung und Industrie

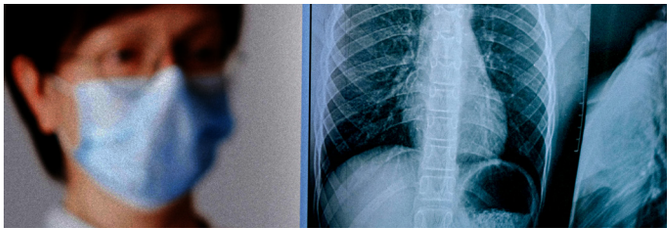
KI kann in der Fertigung und Industrie für vorausschauende Wartung, Qualitätskontrolle, Prozessoptimierung und Lieferkettenmanagement eingesetzt werden, um die Effizienz zu verbessern und Betriebszeit zu maximieren.

Einzelhandel

KI kann das Kundenverhalten analysieren, um personalisierte Produktempfehlungen bereitzustellen und so die Kundenbindung und den Umsatz zu verbessern. Sie ermöglicht auch eine Optimierung der Lagerbestände, indem sie die Nachfrage vorhersagt und Überbestände oder Fehlbestände minimiert.

Finanzdienstleister

Im Finanz- und Bankwesen kann KI zur Betrugserkennung, Risikobewertung, zum Kundenservice und zur Investitionsanalyse eingesetzt werden, was zu verbesserter Sicherheit und fundierterer Entscheidungsfindung führt.



Gesundheitswesen

KI kann im Gesundheitswesen für eine Vielzahl von Anwendungen eingesetzt werden, darunter medizinische Bildanalysen, Krankheitsdiagnosen, personalisierte Behandlungsplanung und Arzneimittelforschung, und so die Patientenergebnisse verbessern und Kosten senken.

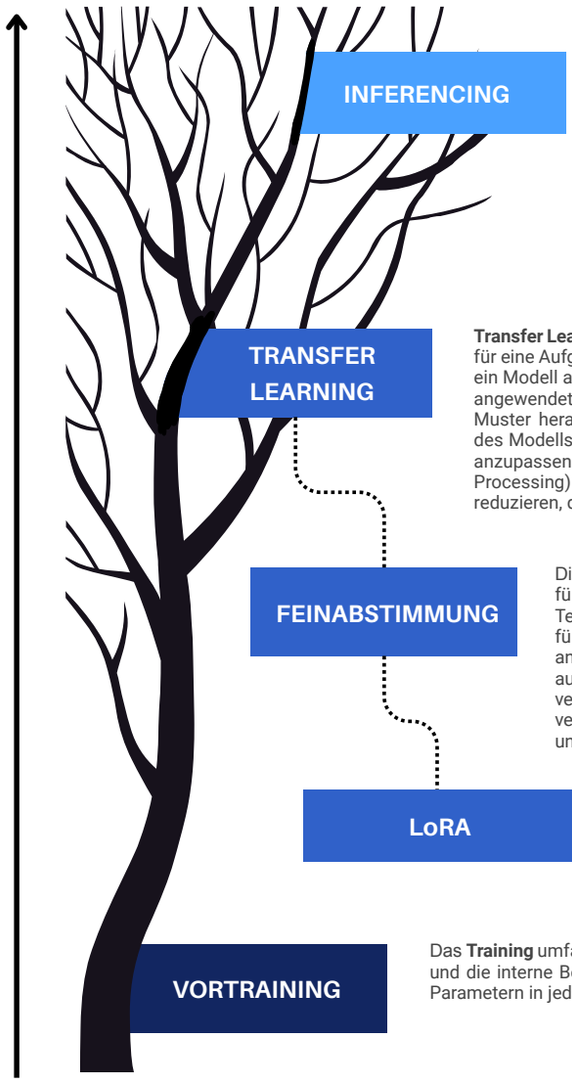
Service-Automatisierung

KI-gestützte Chatbots sind in der Lage, Kundenanfragen zu bearbeiten und Support zu leisten, wodurch der Bedarf an menschlichem Eingreifen reduziert wird. KI kann auch sich wiederholende Aufgaben wie die Dateneingabe oder Dokumentenverarbeitung automatisieren, um die Effizienz zu verbessern und Fehler zu reduzieren.

Worauf IT-Führungskräfte mit Entscheidungsbefugnis achten müssen

ERSTE SCHRITTE: AUFSCHLÜSSELUNG DER KI

Bevor wir diese Anwendungsfälle eingehender betrachten, werfen wir einen genaueren Blick auf den KI-Lebenszyklus. Der KI-Lebenszyklus (künstliche Intelligenz) bezieht sich auf die Phasen der Entwicklung, Bereitstellung und Wartung eines KI-Systems. Auch wenn die spezifischen Methoden und die Terminologie variieren können, umfasst ein typischer KI-Lebenszyklus immer eine Trainings- und Inferencing-Phase für das Modell.



Inferencing ist der Prozess, bei dem ein trainiertes KI-Modell verwendet wird, um Prognosen, Klassifizierungen oder Entscheidungen basierend auf neuen, zuvor noch nie gesehenen Daten zu treffen. In dieser Phase des KI-Lebenszyklus wird das trainierte Modell bereitgestellt und in einer Produktionsumgebung verwendet, um die spezifische Aufgabe auszuführen, für die es entwickelt wurde. Während des Inferencings erfasst das Modell Eingabedaten, verarbeitet sie mithilfe der gelernten Muster und Beziehungen aus der Trainingsphase und erzeugt eine Ausgabe, z. B. eine Prognose, eine Klassifizierung oder Text in GenAI-Workloads. Diese Ausgabe kann dann für verschiedene Anwendungen verwendet werden, z. B. zum Empfehlen von Produkten für NutzerInnen, zum Aufspüren von Anomalien in einem System oder zum Erkennen von Objekten in einem Bild. Im Rahmen der Verarbeitung natürlicher Sprache verbessert die Retrieval-Augmented Generation (RAG) die Genauigkeit bei der Bereitstellung von Modellen, indem sie ihnen Zugriff auf externe Wissensquellen ermöglicht, sodass sie Eingaben in natürlicher Sprache in einem breiteren Spektrum von Kontexten besser verstehen und darauf reagieren können.

Transfer Learning, auch als „Re-Training" bezeichnet, ist eine Technik des maschinellen Lernens, bei der ein Modell, das für eine Aufgabe trainiert wurde, für eine zweite verwandte Aufgabe wiederverwendet wird. Dabei soll das Wissen, das ein Modell aus einer Aufgabe mit vielen verfügbaren beschrifteten Trainingsdaten gelernt hat, in einer neuen Aufgabe angewendet werden, die nicht viele Daten umfasst. Anstatt den Lernprozess von Grund auf neu zu beginnen, werden Muster herangezogen, die beim Lösen einer verwandten Aufgabe erlernt wurden. Dadurch werden die Parameter des Modells, das für die zugehörige Aufgabe trainiert wurde, effektiv aktualisiert, um es besser an die neue Aufgabe anzupassen. Transfer Learning wird hauptsächlich bei Computer-Vision- und NLP-Aufgaben (Natural Language Processing) wie der Stimmungsanalyse eingesetzt, um die normalerweise enorme Menge an Rechenleistung zu reduzieren, die für diese Aufgaben erforderlich ist.

Die **Feinabstimmung** im KI-Lebenszyklus bezieht sich auf den Prozess, bei dem ein vorab trainiertes Modell für maschinelles Lernen mit einem neuen Datenvolumen oder einer neuen Aufgabe weiter trainiert wird. Diese Technik wird häufig beim Transfer Learning verwendet, bei dem ein Modell, das mit einem großen Datensatz für eine bestimmte Aufgabe trainiert wurde, an eine verwandte Aufgabe oder ein kleineres Datenvolumen angepasst wird. Die Feinabstimmung ermöglicht die Übertragung von Wissen vom vortrainierten Modell auf die neue Aufgabe oder das neue Datenvolumen, was häufig zu einer schnelleren Konvergenz und verbesserten Leistung im Vergleich zum Training eines Modells von Grund auf führt. Sie kommt häufig in verschiedenen KI-Anwendungen zum Einsatz, darunter Computer Vision, Verarbeitung natürlicher Sprache und Spracherkennung.

LoRA (Low Rank Adaptation) ist eine neue Technik zur Feinabstimmung von Deep-Learning-Modellen, die die Anzahl der trainierbaren Parameter reduziert und einen effizienten Aufgabenwechsel ermöglicht, wodurch der Feinabstimmungsprozess weniger rechenintensiv wird und besser mit gängiger Hardware bewältigt werden kann.

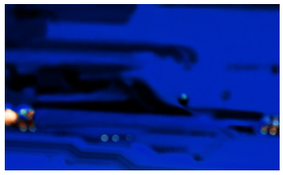
Das **Training** umfasst die Auswahl des geeigneten Algorithmus oder der Modellarchitektur, das Tuning der Modellparameter und die interne Bewertung der Leistung des Modells. Je nach Größe des zu trainierenden Modells kann das Update von Parametern in jeder Iteration des Trainingsprozesses extrem daten- und rechenintensiv sein.

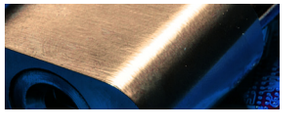
Abbildung 1: Der KI-Lebenszyklus

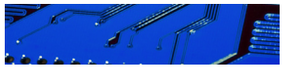
WICHTIGE ENTSCHEIDUNGEN

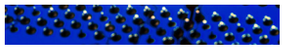
| Leistung

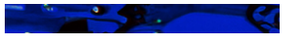
In vielen dieser realen Anwendungen ist eine Entscheidungsfindung in Echtzeit oder nahezu in Echtzeit für den Erfolg unabdingbar. So müssen beispielsweise betrügerische Aktivitäten bei Finanztransaktionen oder Versicherungsansprüchen zeitnah erkannt werden, um finanzielle Verluste zu vermeiden und Unternehmensressourcen zu schützen. In einem Fertigungsszenario müssen Fehler in der Montagelinie oder in den Fabrikbedingungen zur Qualitätssicherung dynamisch überwacht werden. Effektiv muss der Prozessor, der die Inferencing-Workload verarbeitet, für die schnelle und effiziente Verarbeitung eingehender Datenströme optimiert werden. Dell PowerEdge-Server in Kombination mit AMD EPYC-Prozessoren sind eine vielseitige Kombination, die sich gut für die Verarbeitung von Edge-Inferencing-Workloads sowie für Aufgaben im Zusammenhang mit High-Performance-Computing, Cloud-Computing und Big Data Analytics eignet.











| Datensicherheit

Datensicherheit ist entscheidend für den Erfolg von KI-Systemen, insbesondere solcher, die generative KI nutzen, und ein wichtiges Anliegen für technologisch führende Unternehmen, die KI in ihre Abläufe integrieren möchten. KI-Systeme basieren in der Regel auf riesigen Datenmengen, die sensible und vertrauliche Informationen wie persönliche Daten, Finanzdaten oder proprietäre Informationen umfassen können. Der Schutz dieser Daten ist entscheidend, um unbefugten Zugriff oder Datendiebstahl zu verhindern und die Präzision, Zuverlässigkeit und Konsistenz von KI-Modellen und -Prognosen zu gewährleisten.

Confidential Computing ist eine Technologie, die die Datenverarbeitung in einer Secure Enclave erleichtert und sie vor unbefugtem Zugriff oder Manipulation durch Unbefugte, einschließlich des Cloud-Anbieters und anderen NutzerInnen, schützt.[2] Verschlüsselung und andere Sicherheitsmaßnahmen werden verwendet, um die Daten während der Verarbeitung zu isolieren. AMD Infinity Guard, eine Zusammenstellung ausgeklügelter Sicherheitsfunktionen, die in AMD EPYC-Prozessoren integriert sind, unterstützt Confidential Computing durch den Einsatz von Secure Encrypted Virtualization (SEV), bei der virtuelle Maschinen (VMs) mit einem Schlüssel verschlüsselt werden, der nur dem Prozessor bekannt ist. Diese Services zielen darauf ab, hardwarebasierte vertrauenswürdige Ausführungsumgebungen mit AMD SEV-Secure Nested Paging (SEV-SNP) bereitzustellen, wodurch der Gastschutz verbessert wird, um externe Bedrohungen abzuwehren.

Gemeinsames Lernen ist eine weitere Methode zur Aufrechterhaltung der Datensicherheit. Dabei wird ein zentrales Modell über dezentrale Geräte oder Server hinweg trainiert.[3] Anstatt alle Daten an einen zentralen Ort zu übertragen, trainiert jedes Gerät das Modell lokal und nur die Modellaktualisierungen werden freigegeben. Dieser Ansatz wahrt den Datenschutz und ermöglicht kollaboratives Lernen ohne die Weitergabe von Rohdaten. Die Federated-AI-Plattform von Dell Technologies ermöglicht die Ausführung von Rechenprozessen, KI- und ML-Algorithmen auf Datenvolumen am Netzwerk-Edge, während sie erfasst werden. Dabei werden nur mathematische Modelle, Metadaten und Abfrageergebnisse über das Netzwerk an andere Edge-Geräte, Rechenzentren oder die Cloud weitergegeben. Dieser Austausch verbessert die Ergebnisse, da nahezu in Echtzeit verwertbare Erkenntnisse aus großen, verteilten Datenvolumen extrahiert werden können, ohne dass die Daten und geistiges Eigentum offengelegt werden.

2 Advanced Micro Devices, Inc., 30. August 2023, „AMD shares the technical details of technology Powering Innovative Confidential Computing Leadership Cloud Offerings“, https://www.AMD.com/en/newsroom/press-releases/2023-8-30-AMD-shares-the-technical-details-of-technology-pow.html
Advanced Micro Devices, Inc., 2021, Lösungsübersicht „Data Center Solutions, Confidential Computing“, https://www.AMD.com/content/dam/AMD/en/documents/EPYC-business-docs/solution-briefs/confidential-computing-solution-brief.pdf

3 Analytics Vidhya, Dezember 2023, „Federated Learning: A Beginner’s Guide“, https://www.analyticsvidhya.com/blog/2021/05/federated-learning-a-beginners-guide/#:~:text=Federated%20learning%20works%20by%20training,learning%20without%20sharing%20raw%20data
Dell Technologies, 2021, Lösungsübersicht „A federated learning platform for real-time artificial intelligence“, https://www.Delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/dt-sb-analytics-anywhere.pdf

SKALIERUNG IHRER LÖSUNG

| Gleichgewicht zwischen Kosten und Innovation

Das richtige Gleichgewicht zwischen Kosten und Innovation stellt sicher, dass KI-Lösungen nicht nur finanziell machbar, sondern auch wirkungsvoll sind und sowohl für Unternehmen als auch für NutzerInnen einen echten Mehrwert schaffen. Für dieses Gleichgewicht müssen Sie Hardware identifizieren, die sowohl Ihre Anwendungsfälle löst als auch sich einfach in die vorhandene Infrastruktur integrieren lässt. Auf dem Markt für moderne KI-Hardware verursacht die gestiegene Nachfrage nach Accelerators aus verschiedenen Branchen neben Beschränkungen der Produktionskapazitäten, logistischen Herausforderungen und Halbleiterknappheit Engpässe bei Accelerators bei.

CPUs sind jedoch in den meisten Rechenzentren bereits eine Standardkomponente, sodass sie sich im Vergleich zu vollständig neuer Accelerator-Hardware einfacher und kosteneffizienter integrieren lassen. KI-optimierte CPUs können vorhandene Software und Tools nutzen, wodurch der Umrüstungs- und Schulungsbedarf sinkt. CPUs bieten außerdem mehr Flexibilität und Effizienz für eine Vielzahl von Aufgaben jenseits von KI, sodass Ressourcen innerhalb des Rechenzentrums vielseitiger genutzt werden können. Die Aktualisierung Ihres Rechenzentrums mit Dell PowerEdge-Servern mit AMD EPYC-Prozessoren unterstützt Ihre vorhandenen Workloads und ist gleichzeitig bereit für mehr Innovation und Effizienz durch KI.

| Einfachheit und Flexibilität

Für den Aufbau von KI-Lösungen, die auf lange Sicht effektiv, anpassungsfähig und skalierbar sind, muss Ihr KI-System einfach und flexibel sein. Der Zugang zu einer Suite von Software-Frameworks und Optimierungen, die Ihre Hardware ergänzen, verbessert die Leistung, ohne zusätzlichen Zeit- und Arbeitsaufwand für die plattformübergreifende Integration. Diese Eigenschaften sind besonders für die Bewältigung gemischter KI-Workloads wichtig, die eine Kombination verschiedener Arten von KI-Aufgaben wie Training, Inferencing und Datenverarbeitung umfassen.

AMD und Dell Technologies bewältigen gemischte KI-Workloads durch eine Kombination aus Hardware- und Softwarelösungen. AMD EPYC-Prozessoren bieten eine hohe Rechenleistung mit Funktionen wie simultanem Multithreading (SMT) und einer hohen Core-Anzahl, die eine effiziente parallele Verarbeitung für KI-Workloads gestattet. Diese Prozessoren sind für KI-Aufgaben optimiert und liefern eine starke Leistung sowohl für Trainings- als auch für Inferencing-Workloads. Dell PowerEdge-Server, die mit AMD EPYC-Prozessoren ausgestattet sind, bilden eine skalierbare und flexible Plattform für die Bereitstellung von KI-Workloads. Darüber hinaus umfasst die Dell OpenManage-Softwaresuite Managementtools zur Optimierung der Ressourcenzuweisung und des Performancemonitorings für gemischte KI-Workloads.

AMD stellt außerdem das Unified Inference Frontend (UIF) bereit, das die leistungsoptimierten Versionen aller heutigen Software-Stacks nutzt und auf die AMD ZenDNN-Bibliothek für AMD EPYC-Prozessoren, den Open-Source-AMD ROCm-Stack für AMD Instinct Accelerators sowie einen Software-Stack für adaptive AMD-SoCs zurückgreift. AMD ROCm ist außerdem für die Zusammenarbeit mit einer Vielzahl von AMD CPUs und Accelerators konzipiert, einschließlich professioneller und privater Produkte.

| Sicherstellen der Erklärbarkeit

Erklärbare KI spielt eine zentrale Rolle bei der Einhaltung der Transparenz, Vertrauenswürdigkeit und Effektivität von KI-Anwendungen. Erklärbare KI bietet Einblicke in die Art und Weise, wie KI-Modelle Entscheidungen treffen, und beleuchtet die zugrunde liegenden Faktoren und Begründungsprozesse. Diese Transparenz ist entscheidend, um das Vertrauen der Stakeholder zu gewinnen, insbesondere in sensiblen Bereichen wie dem Gesundheitswesen, dem Finanzwesen und der Strafjustiz, in denen sich Entscheidungen direkt auf das Leben von Einzelpersonen auswirken.

Human-in-the-Loop: KI-Systeme nutzen menschliche Intelligenz, um die KI-Leistung zu verbessern und algorithmische Voreingenommenheit zu mindern. Durch die Integration menschlicher Aufsicht können diese Systeme komplexe und mehrdeutige Situationen effektiver bewältigen und sicherstellen, dass KI-Lösungen mit ethischen und sozialen Normen in Einklang stehen. Darüber hinaus ermöglicht die menschliche Beteiligung eine kontinuierliche Verfeinerung und Anpassung von KI-Modellen auf der Grundlage von realem Feedback, was iterative Verbesserungen und die langfristige Zuverlässigkeit fördert. Diese Ansätze sind unerlässlich für den Aufbau verantwortungsvoller, rechenschaftspflichtiger und inklusiver KI-Systeme, die den besten Interessen der Gesellschaft dienen.

Reale Szenarien

Scalers AI arbeitete mit Dell und AMD zusammen, um die Funktionen von Dell PowerEdge-Servern mit AMD-Prozessoren zu demonstrieren. Erfahren Sie, wie diese Technologien für Schulungen, Transfer Learning und Inferencing in Einzelhandels- und Gesundheitsszenarien genutzt werden.

EINZELHANDEL

Scalers AI hat die Retail Inventory Management Reference Solution entwickelt – ein System zum Überwachen und Managen der Lagerbestände in Einzelhandelsregalen durch die Implementierung eines KI-Modells zur Objekterkennung. Diese Referenzlösung nutzt das SSD_MobileNet_V2-Modell zur Identifizierung und Erkennung von Produkten in den Ladenregalen und ermöglicht letztendlich eine automatische Bestandszählung und eine präzise Überwachung der Lagerbestände. Das Modell wurde anhand des SKU110K-Bilddatensatzes, der 23.000 Bilder von Roboflow umfasst, einem Transfer Learning unterzogen. Durch die Nutzung von Computer-Vision- und maschinellen Lernalgorithmen kann das System erkennen, wenn Artikel zur Neige gehen oder nicht vorrätig sind, und das Filialpersonal warnen, um die Lagerbestände rechtzeitig aufzufüllen oder nachzubestellen.

Diese Lösung nutzt den Dell PowerEdge R7615-Server mit dem AMD EPYC 9354P-Prozessor mit 32 Cores.

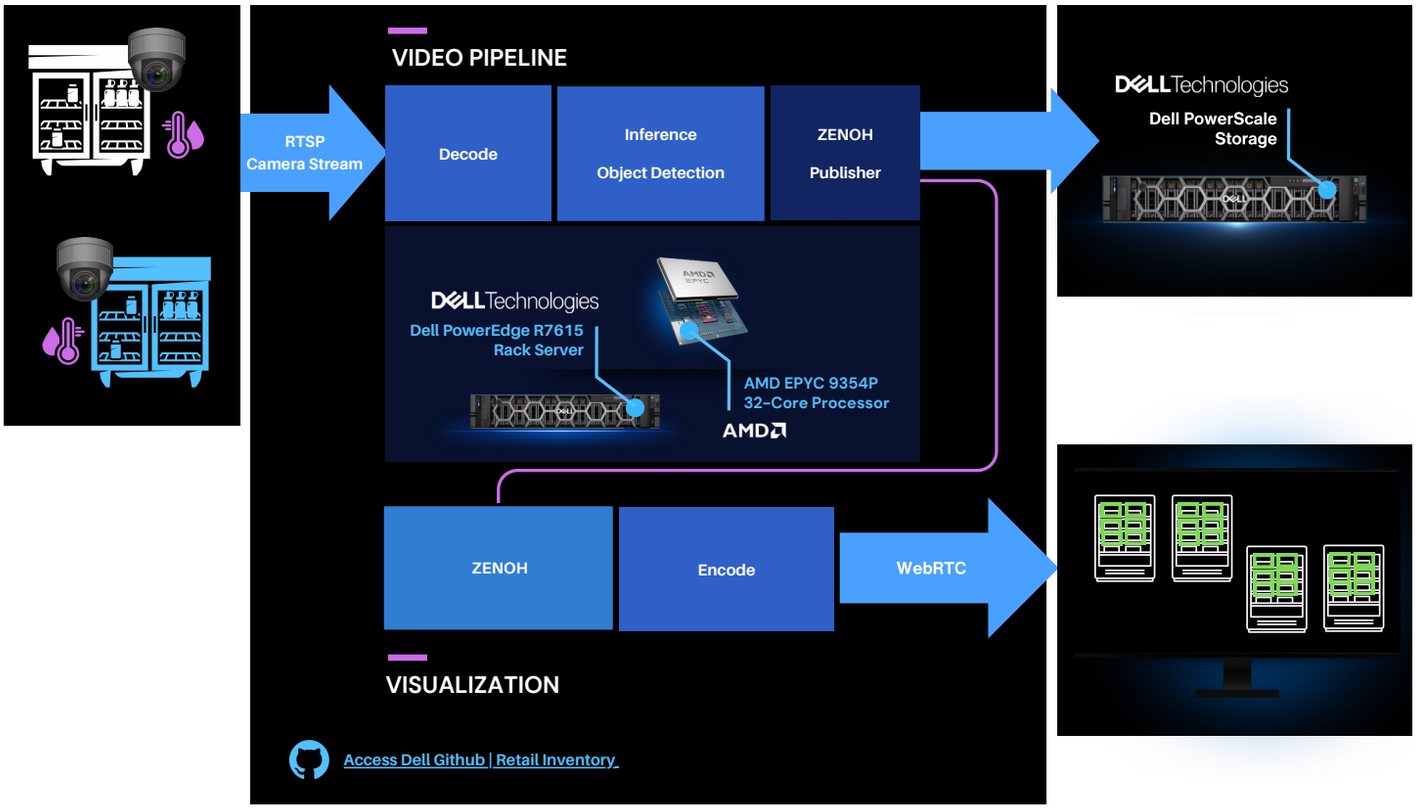


Abbildung 2: Architekturdiagramm der Referenzlösung für das Bestandsmanagement im Einzelhandel

GESUNDHEITSWESEN

Die KI-gestützte medizinische Bildgebung ist äußerst wertvoll. Sie kann das Gesundheitswesen verbessern, indem sie die diagnostische Genauigkeit und Effizienz optimiert und medizinischem Fachpersonal präzise Einblicke in Erkrankungen bietet, die mit bloßem Auge möglicherweise nur schwer zu erkennen sind. Durch die automatisierte Analyse medizinischer Bilder beschleunigt KI die Diagnose, was schnellere Behandlungsentscheidungen ermöglicht und letztendlich die Behandlungsergebnisse verbessert.

Scalers AI nutzte die Funktionen des Dell PowerEdge R7625-Servers, der mit AMD EPYC 9554-Prozessoren mit 64 Cores ausgestattet ist, um eine KI-gestützte medizinische Bildgebungslösung für die Erkennung von Lungenentzündungen zu entwickeln. Durch den Einsatz fortschrittlicher Algorithmen und Techniken des maschinellen Lernens zur Analyse medizinischer Bilder, wie z. B. Röntgenaufnahmen oder CT-Scans, trägt die Lösung dazu bei, die Geschwindigkeit und Genauigkeit der Diagnose von Lungenentzündungen bei PatientInnen zu erhöhen. Letztendlich entsteht so eine zusätzliche Ebene der computergestützten Überprüfung, die medizinisches Fachpersonal bei der effizienteren Handhabung großer Mengen von Bilddaten unterstützen kann.

Diese Referenzlösung nutzt das ResNet50-Modell zur Analyse von Röntgenbildern des Brustkorbs, die aus dem Datenvolumen des NIH Clinical Center stammen. Ihr Hauptziel besteht darin, das Vorhandensein oder Fehlen einer Lungenentzündung zu erkennen, indem im Wesentlichen eine binäre Klassifizierung erfolgt. Das Modell wurde mit dem Xray DICOM-Datenvolumen des NIH Clinical Center-Datenvolumens trainiert, wobei Transfer Learning mit der ResNet50-Architektur durchgeführt wurde.

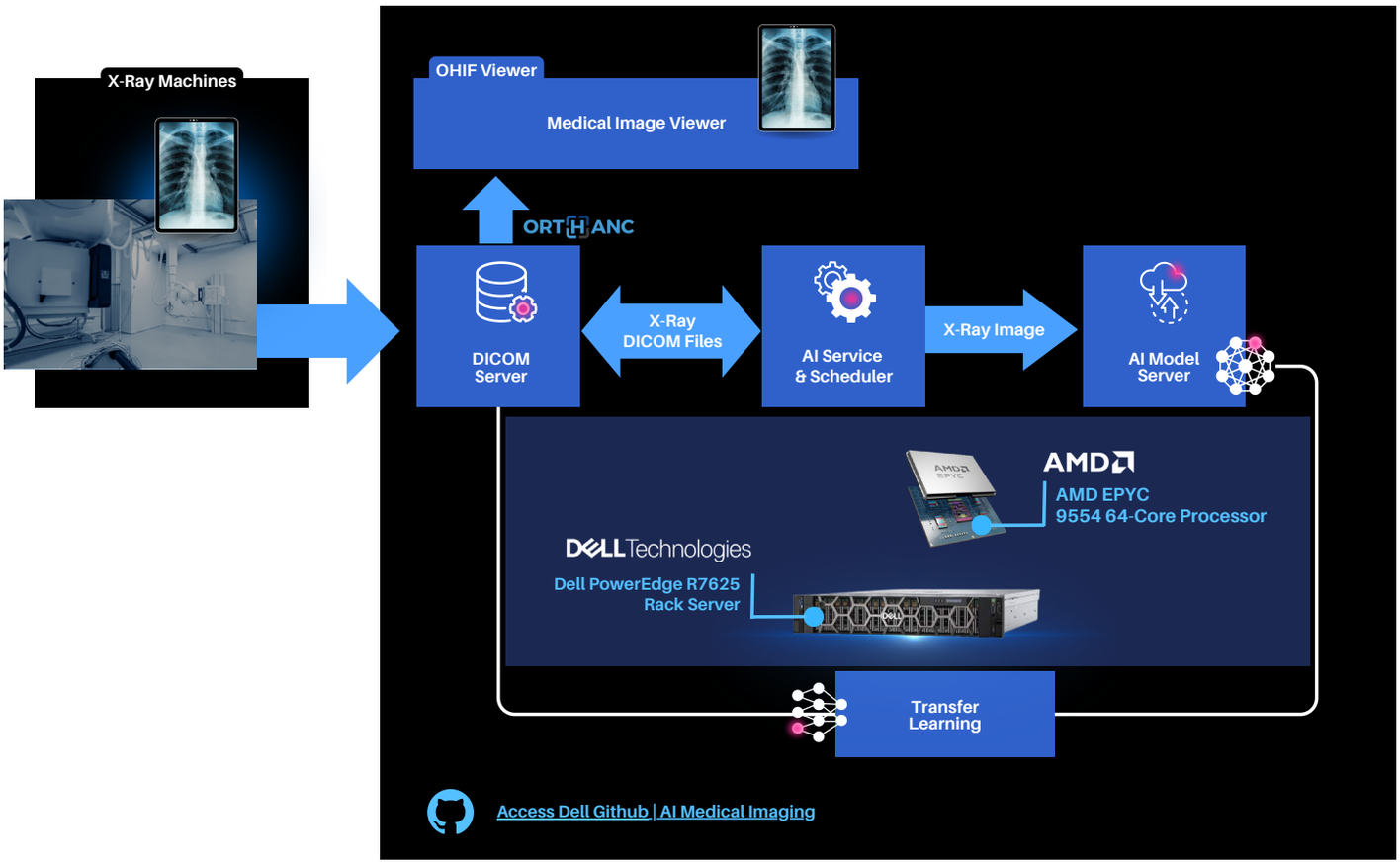
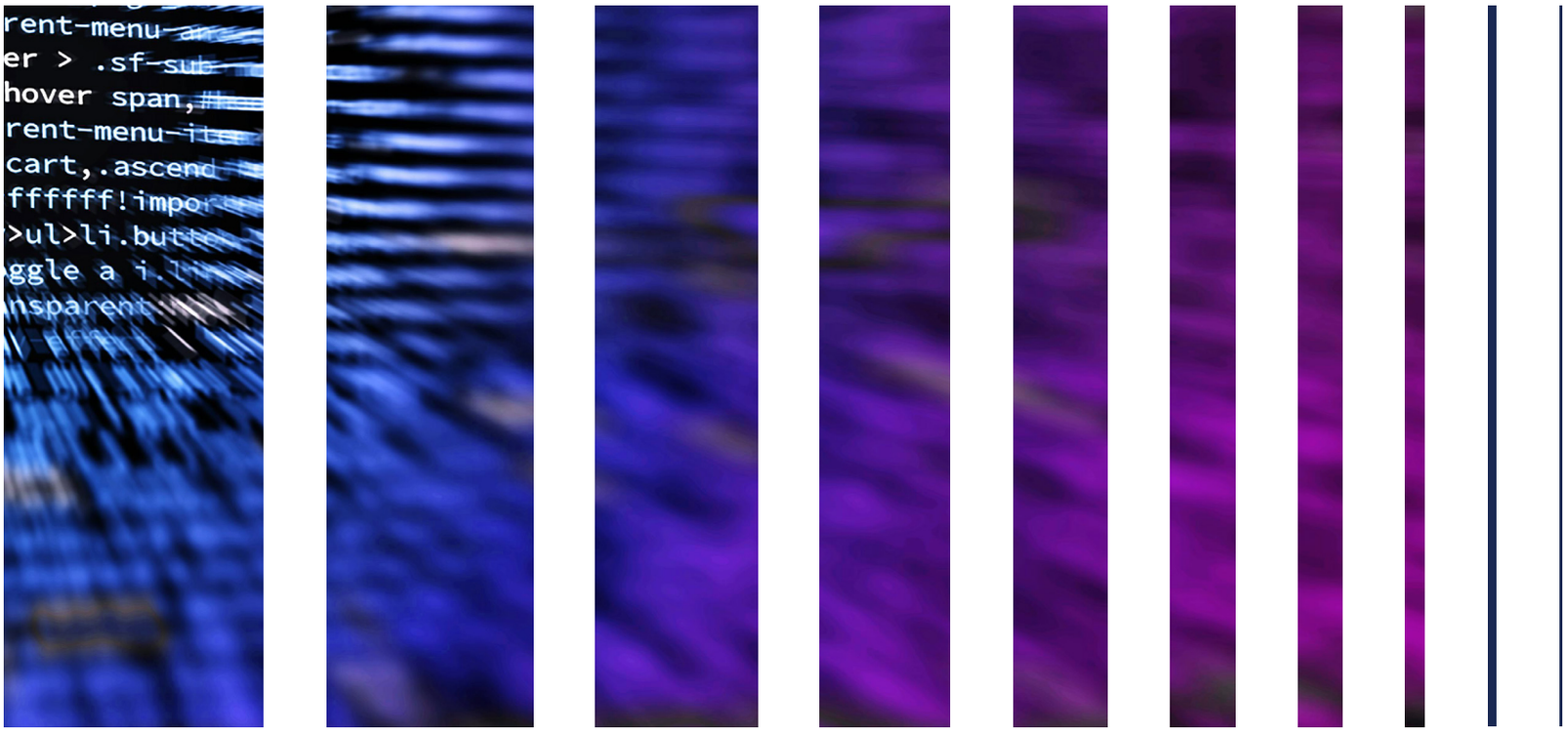


Abbildung 3: Architekturdiagramm der medizinischen KI-Bildgebungslösung

Unsere Lösungen

KI IST FÜR ALLE DA: DELL UND AMD DEMOKRATISIEREN KI

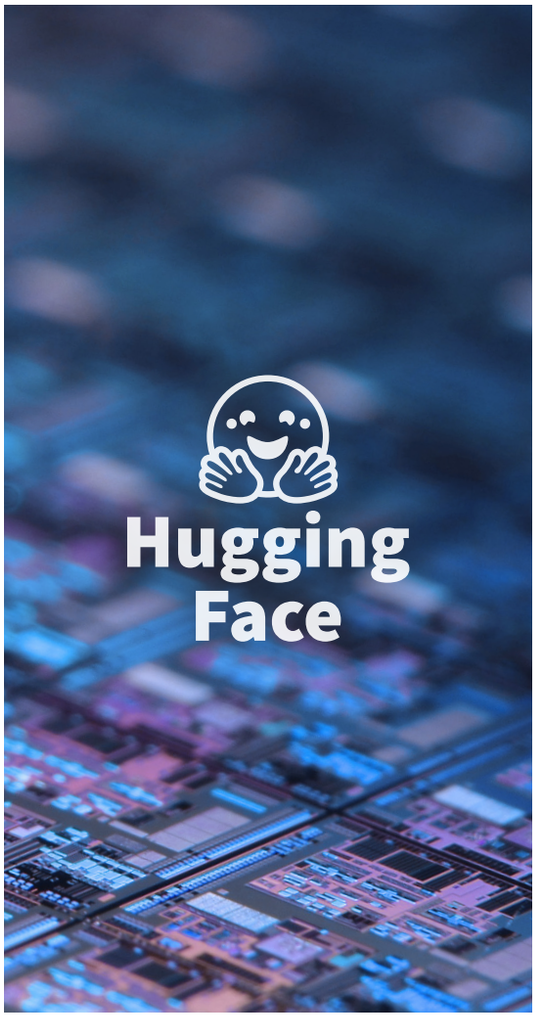
Diese Zusammenarbeit legt den Grundstein für die Demokratisierung der KI, die für die Förderung von Innovationen und Inklusion im KI-Ökosystem unerlässlich ist. Dazu versetzen Dell und AMD Einzelpersonen und Unternehmen in die Lage, KI zu nutzen und einzigartige Herausforderungen in ihren jeweiligen Bereichen mit einer zugänglichen Suite leistungsstarker Server zu lösen, die mit modernsten CPU- und Accelerator-Technologien von AMD ausgestattet sind. Dell PowerEdge-Server mit AMD Instinct MI300X Accelerators sind in der Lage, große KI-Workloads wie Training und Feinabstimmung großer Sprachmodelle (LLMs) zu verarbeiten, während Dell PowerEdge-Server, die mit AMD EPYC-Prozessoren ausgestattet sind, eine hervorragende Verarbeitung von Edge-Inferencing-Workloads bieten. Zusätzlich zur zugrunde liegenden Hardwareplattform bietet AMD auch die ZenDNN-Softwarebibliothek zur Optimierung des Deep-Learning-Inferencings mit AMD-CPUs sowie die AMD ROCm-Softwarebibliothek zur Verbesserung der Trainings-, Feintuning- und Inferencing-Funktionen mit AMD Instinct Accelerators. All diese Optionen sind nahtlos im Unified Inferencing Model (UIF) von AMD miteinander verknüpft. Mit diesem Modell können NutzerInnen End-to-End-KI-Lösungen erstellen und dabei flexibel Software-Frameworks, Softwareoptimierungen und Hardwareplattformen auswählen.

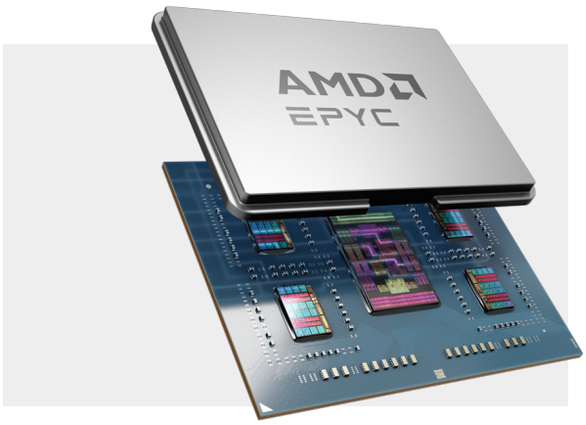


ZUSAMMENARBEIT MIT HUGGING FACE

Unternehmen, die KI einführen möchten, können zum Einstieg bereits bestehende Modelle oder KI-Workflows nutzen, die auf ihre spezifischen Anforderungen zugeschnitten sind. Diese werden direkt von Hugging Face bereitgestellt, einer Open-Source-Plattform für Data Science und maschinelles Lernen. AMD ist eine Zusammenarbeit mit Hugging Face eingegangen, mit dem gemeinsamen Ziel, eine erstklassige Transformer-Leistung über AMD-spezifische Softwareoptimierungen für Softwarebibliotheken und Frameworks anzubieten, die bereits nahtlos in AMD-Plattformen integriert sind. Hugging Face arbeitet aktiv mit dem Entwicklungsteam von AMD zusammen, um wichtige Modelle für Spitzenleistung zu optimieren, AMD ROCm in die Transformer-Bibliothek zu integrieren und die speziell für AMD-Plattformen entwickelte Bibliothek Optimum-AMD zu verbessern. So können Hugging Face-NutzerInnen diese Technologien mit minimalen Codeänderungen verwenden.

Dell Technologies hat sich kürzlich auch mit Hugging Face zusammengetan, um Unternehmen den Prozess der Entwicklung, Feinabstimmung und Anwendung ihrer eigenen Open-Source-Modelle für generative KI (Gen KI) mithilfe der Hugging Face-Community zu vereinfachen – und das alles auf branchenführenden Infrastrukturprodukten und -services von Dell. Auf der Hugging Face-Plattform wird derzeit ein neues Dell Portal entwickelt, das nutzerdefinierte, dedizierte Container und Skripte umfasst. Es soll NutzerInnen bei der sicheren und mühelosen Bereitstellung von Open-Source-Modellen unterstützen, die auf Hugging Face verfügbar sind, und zwar mithilfe von Servern und Daten-Storage-Systemen von Dell. Unternehmen können jetzt die Ressourcen von Hugging Face voll ausschöpfen, um Modelle direkt auf Dell PowerEdge-Servern mit AMD-Prozessoren bereitzustellen und End-to-End-KI-Lösungen mit ihren eigenen proprietären Daten zu entwickeln.



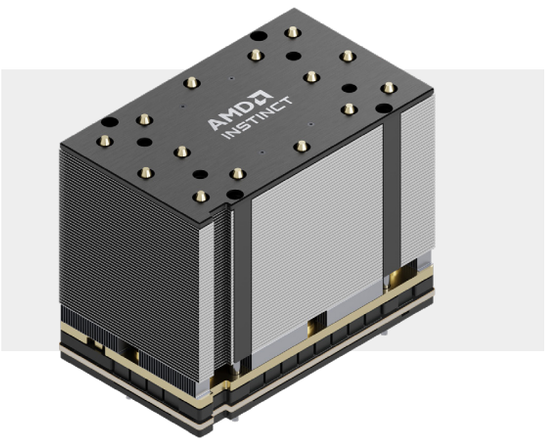


AMD EPYC-PROZESSOREN

AMD bietet mit seinen AMD EPYC-Prozessoren die technologischen Fortschritte, die für moderne cloudbasierte Rechenzentren erforderlich sind. Bei diesen Prozessoren handelt es sich um ein System-on-Chip (SoC), das von Grund auf neu entwickelt wurde, um die Anforderungen aktueller und zukünftiger Rechenzentren effizient zu erfüllen. Die AMD EPYC-Prozessoren der Serie 9000 statten das Rechenzentrum mit bis zu 128 Cores, 256 Threads, 12 Speicherkanälen, die bis zu 6 TB Speicher pro Sockel unterstützen, und 128 PCIe Gen5-Lanes aus. Hinzu kommt die branchenweit bahnbrechende hardwareintegrierte x86-Serversicherheitslösung. Durch die Integration wesentlicher Rechen-, Arbeitsspeicher-, E/A- und Sicherheitsressourcen in das SoC bieten AMD EPYC-Prozessoren erstklassige Leistung und ermöglichen niedrigere Gesamtbetriebskosten (TCO).

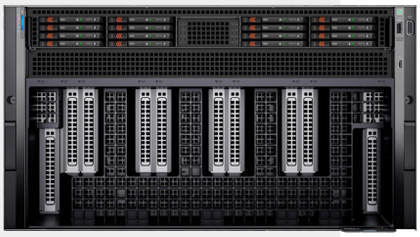
AMD INSTINCT MI300X ACCELERATOR

Der AMD Instinct MI300X Accelerator basiert auf der hochmodernen AMD CDNA 3-Architektur und bietet branchenführende Effizienz und Performance für die intensivsten KI- und HPC-Anwendungen. Er ist mit 304 High-Performance-Computing-Einheiten ausgestattet und verfügt über KI-spezifische Funktionen wie die Unterstützung neuer Datentypen und die Dekodierung von Fotos und Videos sowie einen beispiellosen HBM3-Speicher von 192 GB auf einem einzigen Accelerator.



AMD ROCM 6-OPEN-SOURCE-SOFTWAREPLATTFORM

Die Open-Source-Softwareplattform AMD ROCm 6 ist für die Maximierung der Performance von High-Performance Computing (HPC) und KI-Workloads mit AMD Instinct MI300X Accelerators optimiert. Außerdem wird die Unterstützung für AMD Instinct MI300X Accelerators erweitert, um die Kompatibilität mit branchenüblichen Software-Frameworks zu gewährleisten. Die AMD ROCm-Plattform umfasst eine Vielzahl von Treibern, Entwicklungstools und APIs, die die Accelerator-Programmierung von der Kernel-Ebene bis hin zu Endnutzeranwendungen erleichtern und auf Ihre spezifischen Anforderungen zugeschnitten werden können. AMD ROCm eignet sich besonders für Anwendungen in den Bereichen High-Performance Computing (HPC), künstliche Intelligenz (KI) und wissenschaftliches Computing. Darüber hinaus bietet die AMD ROCm-Plattform Unterstützung für Multi-Accelerator-Computing, einschließlich Remote Direct Memory Access (RDMA) für die Server-Node-Kommunikation.



DELL POWEREDGE SERVERPORTFOLIO

Die Investition von Dell in AMD leistet einen wichtigen Beitrag zur Demokratisierung von KI auf dem Markt, wie die vier Serverplattformen mit EPYC und der Dell PowerEdge XE9680-Rack-Server als Spitzenmodell mit AMD Instinct MI300X Accelerators belegen. Die Dell PowerEdge-Server der neuesten Generation mit AMD EPYC-Prozessoren verbessern Ihre geschäftliche Agilität und beschleunigen die Markteinführung, da sie transformative Workloads wie Datenbanken und Analysen, Virtualisierung, Software Defined Storage, VDI (virtuelle Desktopinfrastruktur), Containerisierung, High Performance Computing (HPC), KI und maschinelles Lernen (ML) unterstützen. Die Rack-Server mit einem Sockel (eine CPU) bieten ein kosteneffizientes Gleichgewicht zwischen Leistung und Storage-Kapazität und sind darauf ausgelegt, nahtlos mit Ihrem Unternehmen zu wachsen, während die Rack-Server mit zwei Sockeln (zwei CPUs) anspruchsvollere Workloads mit einer Vielzahl von Funktionen bewältigen.

Der Dell PowerEdge XE9680-Rack-Server ist auf leistungsstarke Datenverarbeitung ausgelegt und wurde speziell für KI-Aufgaben entwickelt. Er unterstützt acht Accelerators, die sich ideal für das Training von ML (maschinelles Lernen)/Deep Learning (DL) und Inferencing-Workloads eignen, insbesondere für das Training großer Sprachmodelle (LLMs). Ausgestattet mit acht MI300X Accelerators mit jeweils 192 GB High Bandwidth Memory (HBM3) mit 5,3 TB/s, was eine HBM3-Gesamtkapazität von 1,5 TB pro Server und einer FP16-Leistung von über 21 Petaflops ergibt, ist der Dell PowerEdge XE9680-Rack-Server mit AMD Instinct MI300X Accelerators mehr als bereit, den Zugriff auf generative KI für Unternehmen weiter auszubauen. Auf diese Weise lassen sich größere Modelle trainieren, die Stellfläche im Rechenzentrum minimieren, die Gesamtbetriebskosten senken und Wettbewerbsvorteile erzielen.

Zusammenfassung

Das rasante Innovationstempo, das durch KI begünstigt wird, revolutioniert Rechenzentrums-Workloads schneller als jede andere technologische Transformation. Um diese technologischen Fortschritte zu unterstützen, arbeiten Dell und AMD gemeinsam an einem inklusiveren, innovativeren und ethisch besser entwickelten KI-Ökosystem, das Entwickler aus allen Branchen dazu ermutigt, an Open-Source-Ressourcen zusammenzuarbeiten und die Gen-KI-Innovation von heute voranzutreiben. Unabhängig davon, ob Ihre KI-Lösung Ihre Performanceanforderungen auf AMD EPYC-Prozessoren oder auf Servern mit AMD Instinct Accelerators erfüllt, bieten wir Ihnen die Flexibilität, Ihre KI-Workloads auf unseren Hardwareplattformen auszuführen, sodass Sie das Beste aus dem Angebot von Dell und AMD herausholen können.

REFERENZEN

Bilder von AMD: AMD.com, AMD Partner Resource Library, **https://www.amd.com/en/partner/resources/resource-library.html**

Bilder von Dell: **Dell.com**