DELL Technologies

AMD
together we advance_

2024 | Whitepaper

Ondernemingen optimaliseren met AI: het keuzetijdperk binnentreden

Inhoudsopgave

DELL Technologies

AMD
together we advance_

2024 | Whitepaper

De gelegenheid om sectoren te transformeren met AI

Tegenwoordig ligt de grootste kans om uw bedrijf te transformeren voor de toekomst in een innovatie met behulp van AI. Uit data die voor 2023 is verzameld door Accenture Vision Technology blijkt dat 98% van de leidinggevenden wereldwijd het erover eens is dat AI-basismodellen de komende drie tot vijf jaar een belangrijke rol zullen spelen in de strategieën van hun organisatie.[1]

AI is ongelooflijk nuttig geworden voor bedrijven op gebieden zoals de detailhandel, gezondheidszorg en financiële dienstverlening vanwege het vermogen van AI om de efficiëntie van taken te verbeteren, innovatie te stimuleren en besluitvormingsprocessen te verbeteren. Ondanks de voordelen wordt er echter nog steeds een toetredingsdrempel ervaren als het gaat om het integreren van AI, vanwege enkele veelvoorkomende misvattingen.

Je hebt een team van AI-ontwikkelaars nodig om aan de slag te gaan:

Hoewel expertise op het gebied van datawetenschap nog steeds waardevol is voor het ontwikkelen van geavanceerde AI-oplossingen en het begrijpen van de onderliggende principes, is dat niet langer een vereiste. Er is een aanwas van gebruiksvriendelijke AI-tools, platforms zoals Hugging Face en taakspecifieke modellen die door middel van abstractie een groot deel van de complexiteit die gepaard gaat met het ontwikkelen van AI-oplossingen eruit filteren.

Je moet tientallen miljoenen aan hardware uitgeven om resultaten te behalen:

Deze misvatting onderschat de enorme diversiteit van AI-bronnen die tegenwoordig beschikbaar zijn. Hoewel de bekendste bronnen vaak krachtig zijn, met goede ondersteuning, zijn ze mogelijk niet altijd de meest geschikte of rendabele keuze voor elk bedrijf.

Je moet onvermoeibaar werken om versnellers aan te schaffen:

Hoewel versnellers uitblinken in zware AI-workloads, hebben bedrijven mogelijk niet zoveel rekenkracht nodig voor hun AI-applicaties. Te lang wachten om toegang te krijgen tot marktleidende versnellers is bovendien ook niet realistisch. In veel gevallen kunnen AI-geoptimaliseerde CPU's prima de daadwerkelijk vereiste prestaties en efficiëntie leveren om AI-ondersteunde analyses en beslissingen in realtime te produceren, en zijn ze een veel rendabelere en flexibelere oplossing.

[1] Accenture, 30 maart 2023, "Accenture Technology Vision 2023: Generative AI to Usher in a Bold New Future for Business, Merging Physical and Digital Worlds" (Generatieve AI zal een gedurfde nieuwe toekomst voor bedrijven inluiden, waarbij fysieke en digitale werelden worden samengevoegd), 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

Gelukkig evolueert het AI-landschap. Samen werken **Dell** en **AMD** aan het doorprikken van deze mythes, door AI-technologieën en -tools toegankelijk te maken voor een breder scala aan gebruikers met een end-to-end infrastructuur die is ontworpen om de huidige AI-eisen te ondersteunen.

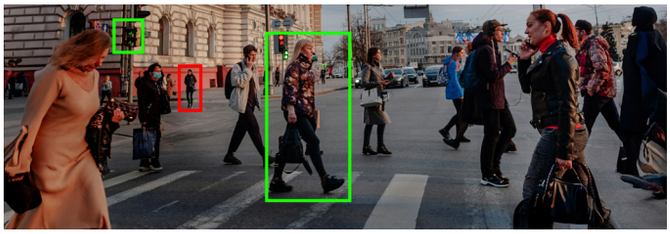
Zo kunt u aan de slag gaan met een reeds geoptimaliseerd model, een betrouwbare softwarestack en een veelzijdig hardwaresysteem, die allemaal vrijelijk beschikbaar zijn via de samenwerking tussen Dell en AMD. Het is niet langer een vereiste voor het gebruik van AI om toegang te hebben tot steeds schaarser wordende versnellers, een substantiële groep ervaren AI-engineers of resources voor het implementeren van enorme cloudclusters.

Dankzij de samenwerking tussen **Dell** en **AMD** bieden we een uniform ecosysteem van hardware en software, ontworpen om ontwikkelaars in staat te stellen end-to-end AI-oplossingen te creëren waarin transfer learning, fine-tuning en inferentie eenvoudig en efficiënt zijn geïntegreerd. Met ondersteuning van **Hugging Face** beschikken we nu over een groeiend portfolio van modellen die draaien op Dell PowerEdge servers met AMD EPYC™ processors of AMD Instinct™ MI300X versnellers, zodat ontwikkelaars kunnen verfijnen, transfer learning kunnen toepassen en implementeren voor inferentie. De investeringen in AMD ROCm™ en AMD ZenDNN™, evenals partnerschappen met PyTorch-, Tensorflow- en ONNX Runtime-frameworks, zijn de fundamentele enablers waardoor Applied AI-ontwikkelaars de democratisering van AI kunnen ervaren. Het onderstaande stackdiagram geeft een overzicht van de componenten die samen het uniforme AI-ecosysteem van Dell en AMD vormen.

AI in de sector

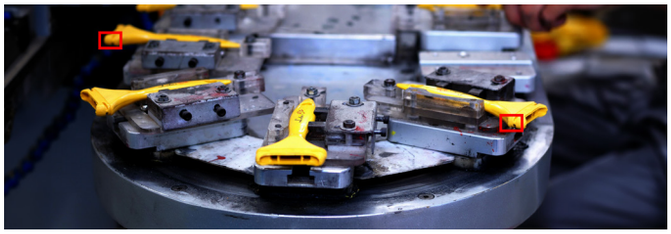
Met de diversificatie van resources en de nadruk op open-source innovatie, dringt AI door in veel verschillende sectoren, waaronder klantenservice, financiën, bankwezen, gezondheidszorg en detailhandel om er maar een paar te noemen. In al deze sectoren biedt AI organisaties ook de kans om het potentieel van hun eigen bedrijfseigen data te ontgrendelen en hun AI-workflows opnieuw vorm te geven door de volgende belangrijke mogelijkheden aan te pakken: data-analyse, automatisering, personalisatie en voorspellende analyse. Daarnaast versnellen AMD ROCm- en ZenDNN-bibliotheken deze AI-workflows om resultaten in bijna realtime te leveren.

Bekijk hieronder hoe AI verschillende sectoren precies beïnvloedt.



Auto-industrie

AI wordt gebruikt voor objectdetectie, het volgen van rijstroken en besluitvorming in autonome voertuigen. AI kan ook voorspellen wanneer een voertuigonderdeel waarschijnlijk defect raakt, waardoor proactief onderhoud mogelijk is en downtime wordt verminderd.



Productie en industrie

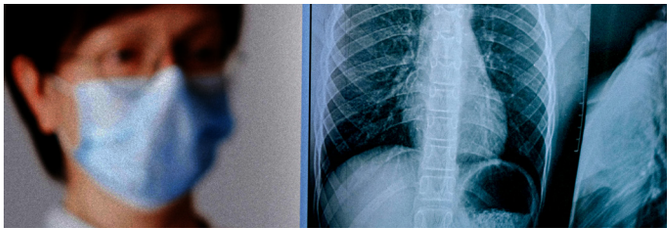
AI kan worden gebruikt in productie en industrie voor voorspellend onderhoud, kwaliteitscontrole, procesoptimalisatie en beheer van leveringsketen, wat leidt tot verbeterde efficiëntie en minder downtime.

Detailhandel

AI kan klantgedrag analyseren om gepersonaliseerde productaanbevelingen te doen, waardoor de klantbetrokkenheid en omzet worden verbeterd. Het kan ook voorraadniveaus optimaliseren door de vraag te voorspellen en voorraadoverschotten of -tekorten tot een minimum te beperken.

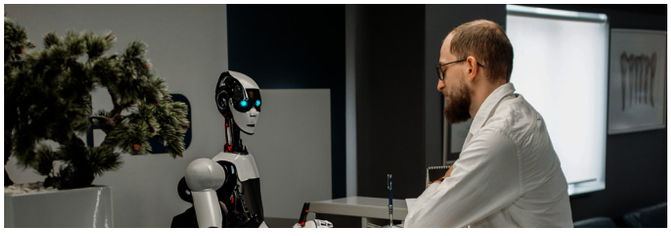
Financiële services

AI kan in de financiële sector en het bankwezen worden gebruikt voor fraudeopsporing, risicobeoordeling, klantenservice en investeringsanalyse, wat leidt tot verbeterde beveiliging en beter geïnformeerde besluitvorming.



Medisch

AI kan in de gezondheidszorg op verschillende manieren worden toegepast, bijvoorbeeld voor medische beeldanalyse, diagnosticering van ziekten, gepersonaliseerde behandeltrajecten en het ontdekken van geneesmiddelen, wat leidt tot verbeterde patiëntresultaten en lagere kosten.



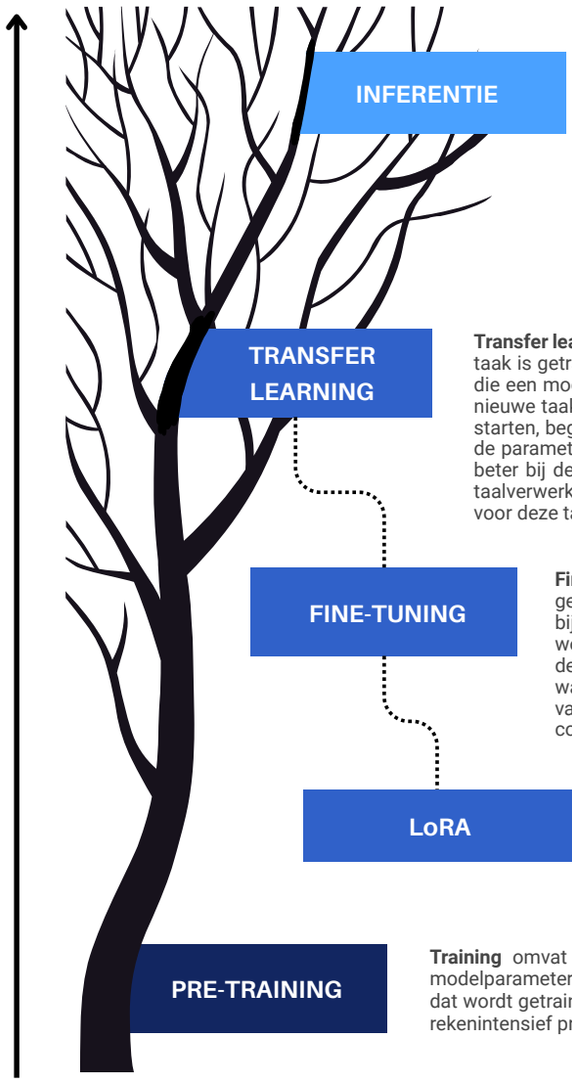
Serviceautomatisering

Door AI aangedreven chatbots kunnen vragen van klanten afhandelen en ondersteuning bieden, waardoor er minder menselijke tussenkomst nodig is. AI kan ook terugkerende taken zoals data-invoer of documentverwerking automatiseren, waardoor de efficiëntie wordt verbeterd en fouten worden verminderd.

Waar IT-besluitvormers rekening mee moeten houden

AAN DE SLAG: AI NADER BEKEKEN

Voordat we door deze gebruiksscenario's navigeren, gaan we eerst dieper in op de AI-levenscyclus. De levenscyclus van AI (kunstmatige intelligentie) verwijst naar de fasen waaruit het proces van ontwikkelen, implementeren en onderhouden van een AI-systeem bestaat. Hoewel specifieke methodieken en terminologie kunnen variëren, omvat een typische AI-levenscyclus altijd modeltraining en 'inferentie'.



Inferentie is het proces waarbij een getraind AI-model wordt gebruikt om voorspellingen te doen, classificaties te maken of beslissingen te nemen op basis van nieuwe, ongeziene data. In deze fase van de AI-levenscyclus wordt het getrainde model geïmplementeerd en gebruikt in een productieomgeving om de specifieke taak uit te voeren waarvoor het is ontworpen. Tijdens deze deductiefase neemt het model invoerdata, verwerkt deze met behulp van de geleerde patronen en relaties uit de trainingsfase en produceert het een uitvoer, zoals een voorspelling, classificatie of tekst in GenAI-workloads. Deze output kan vervolgens worden gebruikt voor verschillende toepassingen, zoals het aanbevelen van producten aan een gebruiker, het detecteren van afwijkingen in een systeem of het herkennen van objecten in een afbeelding. Bij natuurlijke taalverwerking verbetert RAG (retrieval-augmented generation) de nauwkeurigheid bij het implementeren van modellen door deze modellen toegang te geven tot en gebruik te laten maken van externe kennisbronnen, waardoor ze natuurlijke taalinvoer in een breder scala van contexten beter kunnen begrijpen en erop kunnen reageren.

Transfer learning, ook wel re-training genoemd, is een machine learning-techniek waarbij een model dat op één taak is getraind, opnieuw wordt ingezet voor een tweede gerelateerde taak. Het globale idee is om de kennis die een model heeft opgedaan bij een taak met veel beschikbare gelabelde trainingsdata te gebruiken in een nieuwe taak waarvoor niet veel data beschikbaar is. In plaats van het leerproces van meet af aan weer op te starten, beginnen we met de patronen die zijn aangeleerd door een gerelateerde taak op te lossen, waardoor de parameters van het model dat voor de gerelateerde taak is getraind, effectief worden bijgewerkt zodat ze beter bij de nieuwe taak passen. Transfer learning wordt meestal gebruikt bij computervisie- en natuurlijke taalverwerkingstaken zoals sentimentanalyse om de normaal gesproken enorme hoeveelheid rekenkracht die voor deze taken nodig is, te verminderen.

Fine-tuning in de AI-levenscyclus verwijst naar het proces van het verder trainen van een vooraf getraind machine learning-model op een nieuwe dataset of taak. Deze techniek wordt vaak gebruikt bij transfer learning, waarbij een model dat is getraind op een grote dataset voor een specifieke taak, wordt aangepast aan een gerelateerde taak of een kleinere dataset. Deze fijnafstemming maakt de overdracht van kennis mogelijk van het vooraf getrainde model naar de nieuwe taak of dataset, wat vaak resulteert in snellere convergentie en verbeterde prestaties in vergelijking met het trainen van een model vanaf nul. Het wordt veel gebruikt in verschillende AI-toepassingen, waaronder computervisie, natuurlijke taalverwerking en spraakherkenning.

LoRA (Low Rank Adaptation) is een nieuwe techniek voor het verfijnen van deep learning-modellen die werkt door het aantal aan te leren parameters te verminderen en efficiënte taakwisseling mogelijk te maken, waardoor het fijnafstemmingsproces uiteindelijk minder rekenintensief en toegankelijker wordt met algemeen beschikbare hardware.

Training omvat het selecteren van het juiste algoritme of de juiste modelarchitectuur, het afstemmen van modelparameters en het intern evalueren van de prestaties van het model. Afhankelijk van de grootte van het model dat wordt getraind, kan het bijwerken van parameters in elke iteratie van het trainingsproces een extreem data- en rekenintensief proces zijn.

Afbeelding 1: De AI-levenscyclus

CRUCIALE KEUZES

| Prestaties

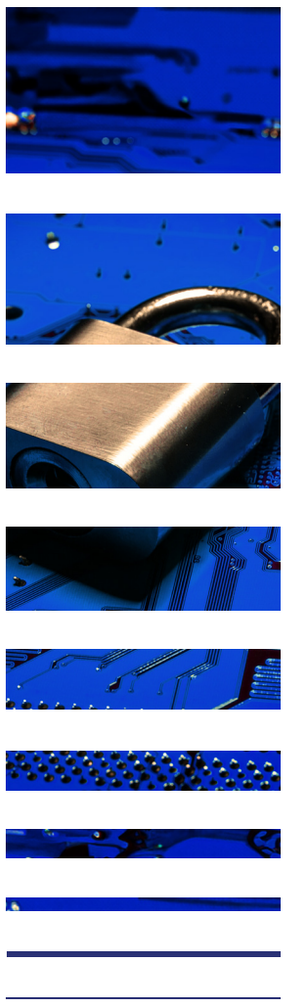
In veel van deze toepassingen in de echte wereld is realtime of near-realtime besluitvorming van cruciaal belang voor het slagen ervan. Frauduleuze activiteiten bij financiële transacties of verzekeringsclaims moeten bijvoorbeeld onmiddellijk worden geïdentificeerd om financiële verliezen te voorkomen en bedrijfsmiddelen te beschermen. In een productiescenario moeten defecten in de productielijn of fabrieksomstandigheden dynamisch worden bewaakt voor kwaliteitsborging. In feite moet de processor die uw inferentieworkload afhandelt, worden geoptimaliseerd voor het snel en efficiënt verwerken van inkomende datastromen. Dell PowerEdge servers in combinatie met AMD EPYC-processors vormen een veelzijdige combinatie, zeer geschikt voor het verwerken van edge-inferentieworkloads, evenals taken met betrekking tot high-performance computing, cloudcomputing en big data analytics.

| Databeveiliging

Databeveiliging is cruciaal voor het succes van AI-systemen, met name systemen die gebruikmaken van generatieve AI, en is een belangrijk aandachtspunt voor technologieleiders die AI in hun activiteiten willen integreren. AI-systemen zijn doorgaans afhankelijk van enorme hoeveelheden data, waaronder gevoelige en vertrouwelijke informatie zoals persoonlijke gegevens, financiële data of bedrijfseigen data. Het beschermen van deze data is van cruciaal belang om ongeoorloofde toegang of datadiefstal te voorkomen en om de precisie, betrouwbaarheid en consistentie van AI-modellen en -voorspellingen te waarborgen.

Confidential computing is een technologie die dataverwerking in een beveiligde enclave mogelijk maakt en deze beschermt tegen ongeoorloofde toegang of manipulatie door onbevoegde partijen, waaronder de cloudprovider en andere gebruikers.[2] Versleuteling en andere beveiligingsmaatregelen worden gebruikt om de data tijdens de verwerking te isoleren. De AMD Infinity Guard, een verzameling geavanceerde beveiligingsfuncties die zijn geïntegreerd in AMD EPYC-processors, ondersteunt confidential computing door gebruik te maken van Secure Encrypted Virtualization (SEV), waarbij virtuele machines (VM's) worden versleuteld met behulp van een sleutel die alleen bekend is bij de processor. Deze services zijn gericht op het bieden van op hardware gebaseerde vertrouwde uitvoeringsomgevingen met behulp van AMD SEV-Secure Nested Paging (SEV-SNP), die de gastbeveiliging verbetert om te helpen beschermen tegen externe bedreigingen.

Federated learning is een andere methode om de databeveiliging te handhaven. Het traint een centraal model op gedecentraliseerde apparaten of servers.[3] In plaats van alle data naar een centrale locatie over te brengen, traint elk apparaat het model lokaal en worden alleen de modelupdates gedeeld. Deze aanpak beschermt de privacy en maakt samenwerkend leren mogelijk zonder onbewerkte data te delen. Met het federatieve AI-platform van Dell Technologies kunnen rekenprocessen, AI en ML-algoritmen worden uitgevoerd op datasets aan de netwerkrand terwijl ze worden verzameld, waarbij alleen wiskundige modellen, metadata en queryresultaten via het netwerk worden gedeeld met andere edge-apparaten, datacenters of de cloud. Deze uitwisseling verbetert de resultaten doordat er bijna in real time bruikbare inzichten uit grote, gedistribueerde datasets kunnen worden gehaald zonder dat de data en eventueel intellectueel eigendom worden prijsgegeven.



2 Advanced Micro Devices, Inc. 2023, August 30, "AMD shares the technical details of technology Powering Innovative Confidential Computing Leadership Cloud Offerings" (AMD deelt de technische details van technologie voor innovatief cloudaanbod van leiders in vertrouwelijke dataverwerking), 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, "Data Center Solutions, Confidential Computing" Solution Brief (Oplossingen voor datacenters, vertrouwelijke dataverwerking, beknopt oplossingsoverzicht), https://www.AMD.com/content/dam/AMD/en/documents/EPYC-business-docs/solution-briefs/confidential-computing-solution-brief.pdf

3 Analytics Vidhya, 2023, Dec, "Federated Learning: A Beginner's Guide" (Federatief leren: een handleiding voor beginners), 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, "A federated learning platform for real-time artificial intelligence" Solution Brief (Een federatief leerplatform voor realtime kunstmatige intelligentie, Beknopt oplossingsoverzicht) https://www.Delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/dt-sb-analytics-anywhere.pdf

UW OPLOSSING OPSCHALEN

| Kosten en innovatie in balans brengen

Het vinden van de juiste balans tussen kosten en innovatie zorgt ervoor dat AI-oplossingen niet alleen financieel haalbaar zijn, maar ook impact hebben, waardoor echte waarde wordt gegenereerd voor zowel bedrijven als gebruikers. Een cruciaal onderdeel bij het vinden van deze balans ligt in het identificeren van hardware die zowel uw gebruiksscenario's oplost als gemakkelijk kan worden geïntegreerd in de bestaande infrastructuur. In de moderne AI-hardwaremarkt dragen de toegenomen vraag naar versnellers vanuit verschillende industrieën, bovenop beperkingen van de productiecapaciteit, logistieke uitdagingen en tekorten aan halfgeleiders, allemaal bij aan versnellertekorten.

CPU's vormen echter al een standaardcomponent in de meeste datacenters, waardoor integratie eenvoudiger en kosteneffectiever is in vergelijking met het toevoegen van volledig nieuwe versnellerhardware. AI-geoptimaliseerde CPU's kunnen gebruikmaken van bestaande software en tooling, waardoor uitgebreide retooling of retraining minder nodig is. CPU's bieden ook meer flexibiliteit en efficiëntie voor een breed bereik aan taken naast AI, waardoor een veelzijdiger gebruik van resources binnen het datacenter mogelijk is. Door uw datacenter te vernieuwen met Dell PowerEdge servers met AMD EPYC-processors wordt het vervullen van uw bestaande workloads ondersteund, terwijl u klaar blijft voor vooruitgang naar meer innovatie en efficiëntie aangedreven door AI.

| Eenvoud en flexibiliteit

Eenvoud en flexibiliteit van uw AI-systeem zijn essentieel voor het bouwen van AI-oplossingen die effectief, aanpasbaar en schaalbaar zijn op de lange termijn. Door toegang te hebben tot een suite van softwareframeworks en optimalisaties die uw hardware aanvullen, verbetert u de prestaties zonder extra tijd en moeite te besteden aan platformonafhankelijke integratie. Deze eigenschappen zijn vooral belangrijk voor het aanpakken van gemengde AI-workloads, die een combinatie van verschillende typen AI-taken omvatten, zoals training, inferentie en dataverwerking.

AMD en Dell Technologies pakken gemengde AI-workloads aan met een combinatie van hardware- en softwareoplossingen. AMD EPYC-processors bieden krachtige rekenkracht, met functies zoals gelijktijdige multithreading (SMT) en een hoog aantal cores, waardoor efficiënte parallelle verwerking voor AI-workloads mogelijk is. Deze processors zijn geoptimaliseerd voor AI-taken en bieden krachtige prestaties voor zowel trainings- als inferentieworkloads. Dell PowerEdge servers, uitgerust met AMD EPYC-processors, bieden een schaalbaar en flexibel platform voor het implementeren van AI-workloads. Daarnaast biedt de Dell OpenManage softwaresuite beheertools voor het optimaliseren van resourcetoewijzing en prestatiecontrole voor workloads met gemengde AI.

AMD biedt ook de Unified Inference Frontend (UIF), die gebruikmaakt van de prestatieverbeterde versies van elk van de huidige softwarestacks en gebruikmaakt van de AMD ZenDNN-bibliotheek voor AMD EPYC-processors, de open-source AMD ROCm-stack voor AMD Instinct Accelerators, evenals een softwarestack voor AMD adaptieve SoC's. MD ROCm is ook ontworpen om te werken met een breed scala aan AMD CPU's en versnellers, waaronder zowel professionele als consumentenproducten.

| Zorgen voor verklaarbaarheid

Verklaarbare AI speelt een cruciale rol bij het waarborgen van transparantie, betrouwbaarheid en effectiviteit in kunstmatige-intelligentietoepassingen. Verklaarbare AI geeft inzicht in hoe AI-modellen beslissingen nemen, en werpt licht op de onderliggende factoren en redeneerprocessen. Deze transparantie is cruciaal om het vertrouwen van belanghebbenden te winnen, vooral in gevoelige domeinen zoals gezondheidszorg, financiën en strafrecht, waar beslissingen rechtstreeks van invloed zijn op het leven van individuen.

Human-in-the-loop AI-systemen maken gebruik van menselijke intelligentie om AI-prestaties te verbeteren en algoritmische vooroordelen te beperken. Door menselijk toezicht te integreren, kunnen deze systemen complexe en dubbelzinnige situaties effectiever afhandelen en ervoor zorgen dat AI-oplossingen aansluiten bij ethische en sociale normen. Bovendien maakt menselijke betrokkenheid voortdurende verfijning en aanpassing van AI-modellen mogelijk op basis van feedback uit de praktijk, wat iteratieve verbetering en betrouwbaarheid op lange termijn bevordert. Deze benaderingen zijn essentieel voor het bouwen van verantwoorde, verantwoordelijke en inclusieve AI-systemen die de belangen van de samenleving dienen.

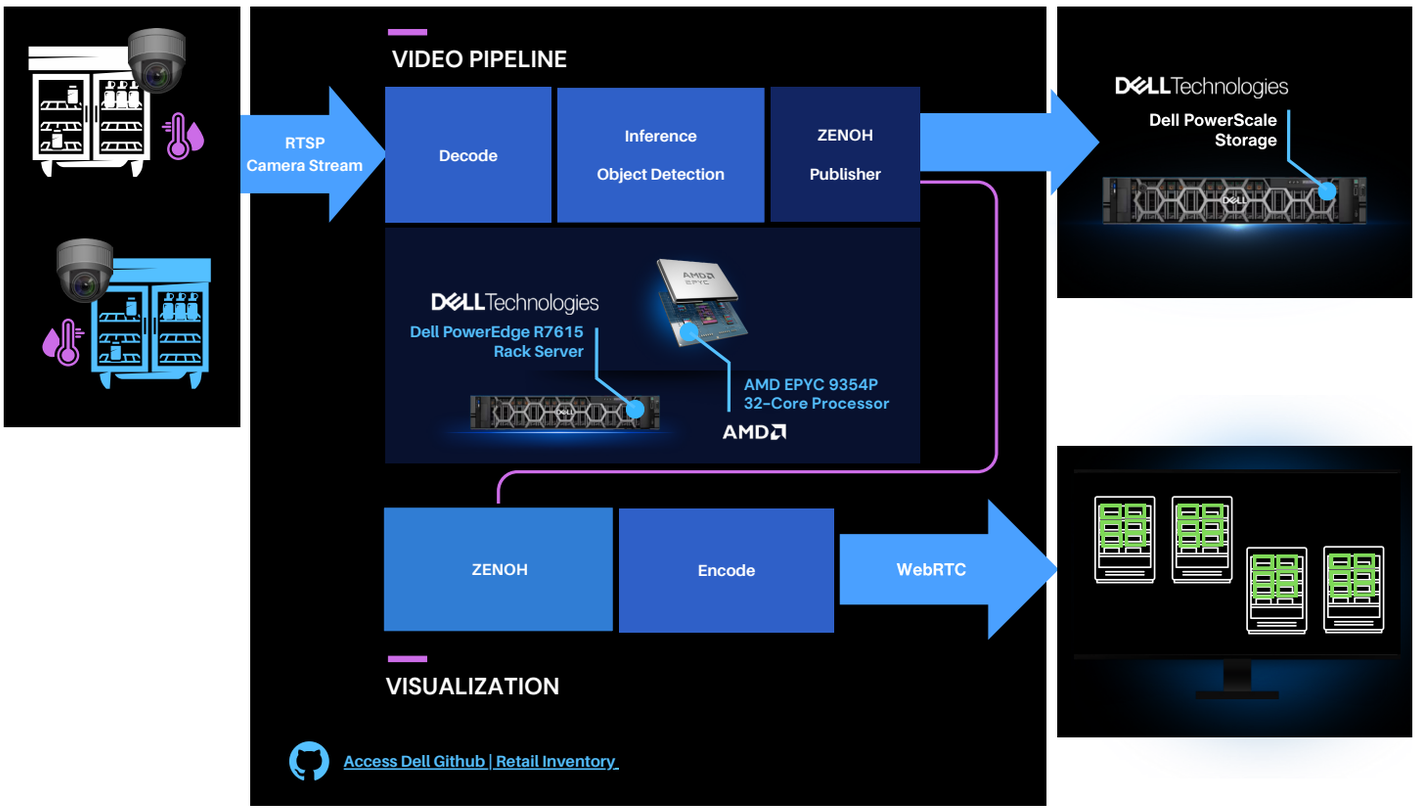
Scenario's uit de praktijk

Scalers AI werkte samen met Dell en AMD om de mogelijkheden van Dell PowerEdge servers met AMD-processors te demonstreren. Bekijk hoe deze technologieën worden ingezet voor training, transfer learning en inferentie in detailhandel- en gezondheidszorgscenario's.

DETAILHANDEL

Scalers AI heeft de referentieoplossing Retail Inventory Management gebouwd, een systeem dat is ontworpen voor het bewaken en beheren van voorraadniveaus in de winkelschappen door de implementatie van een AI-model voor objectdetectie. Deze referentieoplossing maakt gebruik van het SSD_MobileNet_V2-model voor het identificeren en herkennen van producten in de winkelschappen, waardoor uiteindelijk automatische voorraadtellingen en nauwkeurige bewaking van voorraadniveaus mogelijk worden. Het model onderging transfer learning met behulp van de SKU110K beelddataset, bestaande uit 23.000 afbeeldingen van Roboflow. Door gebruik te maken van computervisie en machine learning-algoritmen kan het systeem detecteren wanneer artikelen bijna op zijn of niet meer op voorraad zijn, waardoor het winkelpersoneel wordt gewaarschuwd voor tijdige aanvulling of noodzaak tot bijbestellen.

Deze oplossing maakt gebruik van de Dell PowerEdge R7615 server met de AMD EPYC 9354P processor met 32 cores.



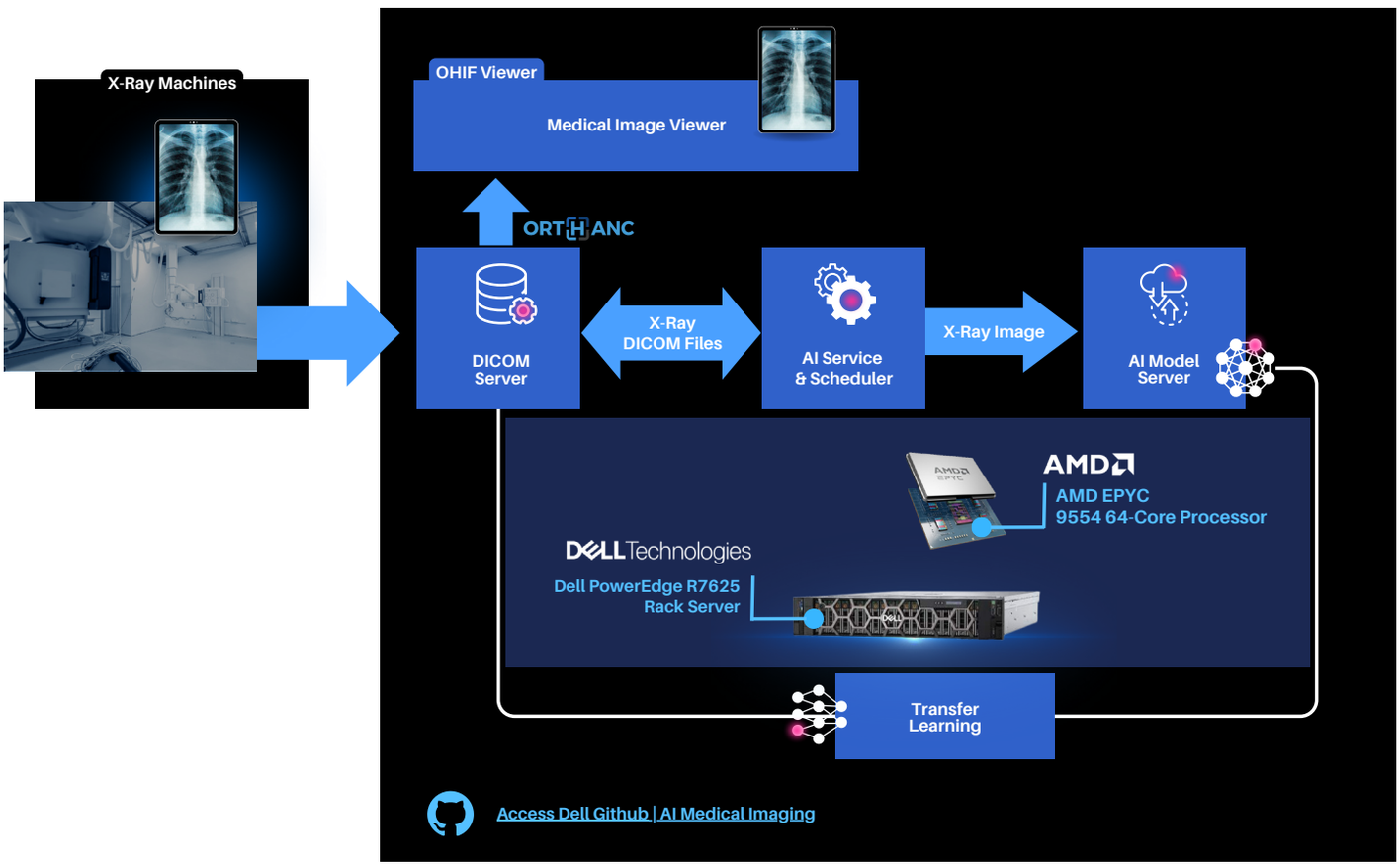
Afbeelding 2: Architectuurdiagram van de referentieoplossing Retail Inventory Management

GEZONDHEIDSZORG

Door AI aangedreven medische beeldvorming is enorm waardevol vanwege het vermogen om de gezondheidszorg te verbeteren. Het verbetert de diagnostische nauwkeurigheid en efficiëntie en geeft zorgverleners nauwkeurig inzicht in aandoeningen die mogelijk met het blote oog lastig te detecteren zijn. Door de analyse van medische beelden te automatiseren, verkort AI de tijd die nodig is voor de diagnose, waardoor snellere behandelingsbeslissingen mogelijk worden en uiteindelijk de resultaten voor de patiënt worden verbeterd.

Scalers AI maakt gebruik van de mogelijkheden van de Dell PowerEdge R7625 server die is uitgerust met AMD EPYC 9554 64-Core processors om een door AI aangedreven oplossing voor medische beeldvorming te creëren voor de detectie van longontsteking. Met behulp van geavanceerde algoritmen en machine learning-technieken voor het analyseren van medische beelden, zoals röntgenfoto's of CT-scans, helpt deze oplossing de snelheid en nauwkeurigheid van de diagnose van longontsteking bij patiënten te verhogen. Uiteindelijk introduceert dit een extra laag van computerondersteunde beoordeling, waardoor zorgverleners in staat worden gesteld om grote hoeveelheden beeldvormingsdata efficiënter te verwerken.

Deze referentieoplossing maakt gebruik van het ResNet50-model om röntgenfoto's van de borstkas te analyseren die zijn verkregen uit de dataset van het NIH Clinical Center. Het primaire doel is om de aan- of afwezigheid van longontsteking te detecteren. Het voert dus in wezen een binaire classificatie uit. Het model is getraind met behulp van de Xray DICOM-dataset uit de dataset van het NIH Clinical Center, waarbij transfer learning met de ResNet50-architectuur betrokken is.

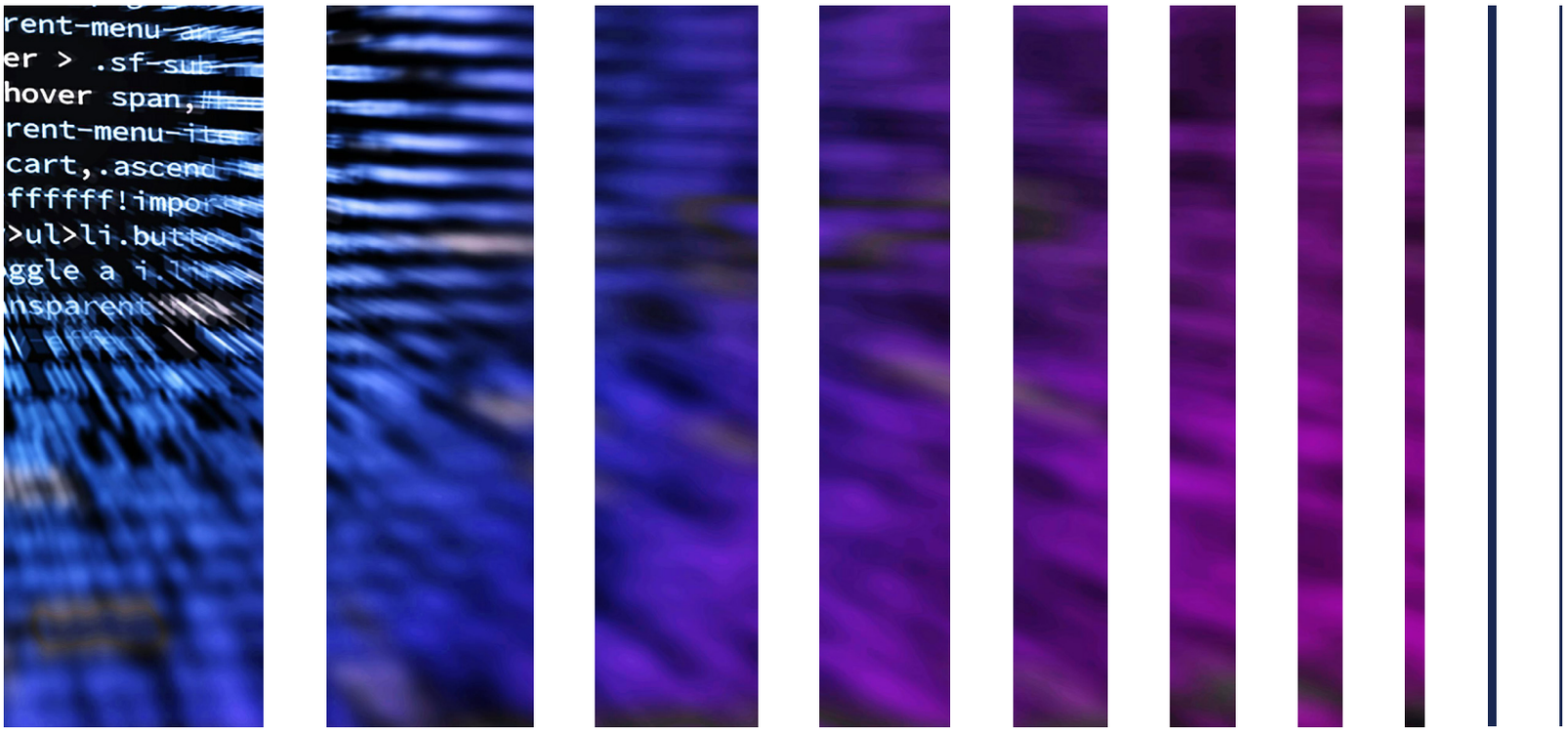


Afbeelding 3: Architectuurdiagram van de medische AI-imagingoplossing

Onze oplossingen

AI IS VOOR IEDEREEN: DELL EN AMD DEMOCRATISEREN AI

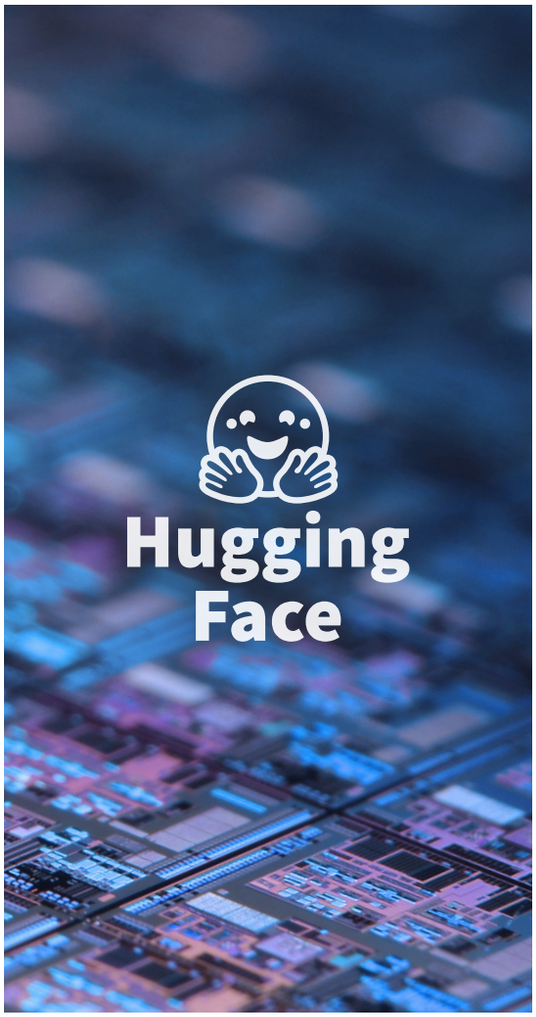
Deze samenwerking legt de basis voor de democratisering van AI, wat essentieel is voor het stimuleren van innovatie en het bevorderen van inclusiviteit in het AI-ecosysteem. Dell en AMD bereiken dit resultaat door individuen en organisaties in staat te stellen gebruik te maken van AI en unieke uitdagingen in hun respectievelijke vakgebieden op te lossen met een toegankelijke suite van krachtige servers die zijn uitgerust met geavanceerde AMD CPU- en versnellertechnologieën. Dell PowerEdge servers met de AMD Instinct MI300X versnellers kunnen grote AI-workloads verwerken, zoals training en het verfijnen van grote taalmodellen (LLM's), terwijl Dell PowerEdge servers die zijn uitgerust met AMD EPYC-processors uitblinken in het verwerken van edge-inferentieworkloads. Naast het onderliggende hardwareplatform biedt AMD ook de ZenDNN-softwarebibliotheek voor de optimalisatie van deep learning-inferentie op AMD CPU's, evenals de AMD ROC-softwarebibliotheek om de training, fine-tuning en inferentiemogelijkheden op AMD Instinct Accelerators te verbeteren. Al deze opties zijn naadloos met elkaar verbonden in het Unified Inferencing Model (UIF) van AMD, waarmee gebruikers end-to-end AI-oplossingen kunnen bouwen, met flexibiliteit in de keuze van softwareframeworks, softwareoptimalisaties en hardwareplatformkeuzes.

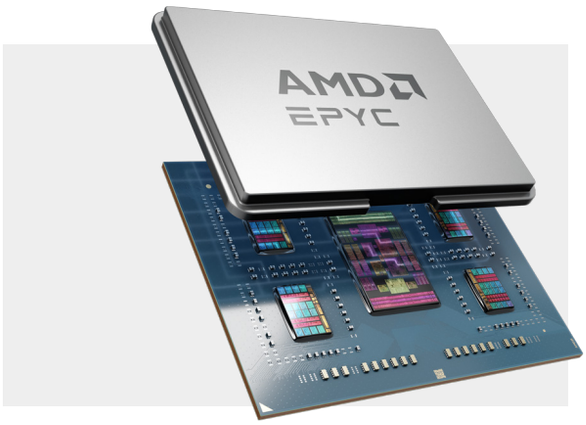


SAMENWERKING MET HUGGING FACE

Bedrijven die graag AI willen implementeren, kunnen beginnen met het gebruik van reeds bestaande modellen of AI-workflows die zijn afgestemd op hun specifieke behoeften, rechtstreeks vanuit Hugging Face, een opensourceplatform dat is gewijd aan datawetenschap en machine learning. AMD is een samenwerking aangegaan met Hugging Face, met het gedeelde doel om eersteklas transformatorprestaties te leveren door AMD-specifieke softwareoptimalisaties toe te voegen aan softwarebibliotheken en frameworks die nu al naadloos integreren met AMD-platforms. Hugging Face werkt actief samen met het technische team van AMD om belangrijke modellen te optimaliseren voor topprestaties, AMD ROCm op te nemen in de Transformers-bibliotheek en Optimum-AMD te verbeteren, een bibliotheek die speciaal is ontworpen voor AMD-platforms, om Hugging Face-gebruikers te helpen ze te gebruiken met minimale codewijzigingen.

Dell Technologies heeft onlangs ook de krachten gebundeld met Hugging Face om het voor ondernemingen eenvoudiger te maken om hun eigen open-source generatieve AI-modellen (Gen AI) te ontwikkelen, te verfijnen en toe te passen met behulp van de Hugging Face-community, allemaal op toonaangevende infrastructuurproducten en -services van Dell. Er wordt een nieuwe Dell portal ontwikkeld op het Hugging Face-platform, met aangepaste, speciale containers en scripts om gebruikers te helpen bij het veilig en moeiteloos implementeren van opensourcemodellen die beschikbaar zijn op Hugging Face met behulp van de servers en datastoragesystemen van Dell. Bedrijven kunnen nu optimaal profiteren van de bronnen van Hugging Face om modellen rechtstreeks te implementeren op Dell PowerEdge servers met AMD-processors en end-to-end AI-oplossingen te bouwen met behulp van hun eigen bedrijfseigen data.



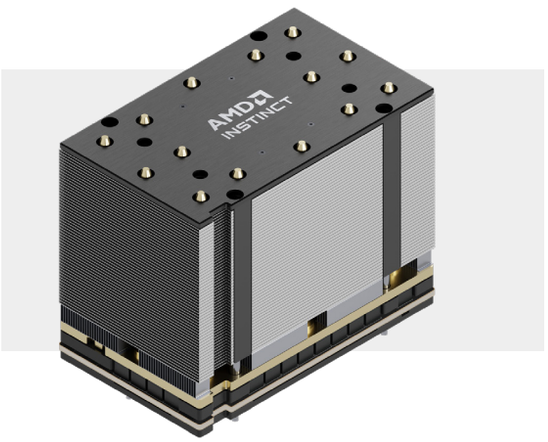


AMD EPYC-PROCESSORS

AMD biedt de technologische vooruitgang die nodig is voor moderne cloudgebaseerde datacenters via hun AMD EPYC-processors. Deze processors zijn een systeem-op-chip (SoC) die helemaal opnieuw zijn ontworpen om efficiënt te voldoen aan de eisen van huidige en toekomstige datacenters. De AMD EPYC-processors uit de 9000-serie rusten het datacenter uit met maximaal 128 cores, 256 threads, 12 geheugenkanalen die tot 6 TB geheugen per socket ondersteunen en 128 PCIe Gen5-lanes. Dit wordt gecombineerd met de hardware-geïntegreerde x86-serverbeveiligingsoplossing die baanbrekend is in de branche. Door essentiële resources voor rekenkracht, geheugen, I/O en beveiliging in de SoC te integreren, leveren AMD EPYC-processors topprestaties en een lagere TCO (Total Cost of Ownership).

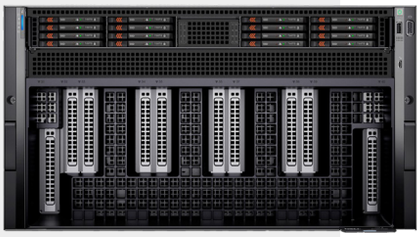
AMD INSTINCT MI300X VERSNELLER

De AMD Instinct MI300X versneller is gebouwd op de geavanceerde AMD CDNA 3 architectuur en biedt toonaangevende efficiëntie en prestaties voor de meest intensieve AI- en HPC-applicaties. Het is uitgerust met 304 krachtige rekeneenheden en beschikt over AI-specifieke functies zoals ondersteuning voor nieuwe datatypen en foto- en videodecodering, evenals een ongeëvenaard HBM3-geheugen van 192 GB op een enkele versneller.



AMD ROCM 6 OPEN-SOURCE SOFTWAREPLATFORM

Het AMD ROCm 6 open-source softwareplatform is geoptimaliseerd om de prestaties van high-performance computing (HPC) en AI-workloads van AMD Instinct MI300X versnellers te maximaliseren. Het breidt ook de ondersteuning voor AMD Instinct MI300X-versnellers uit, wat compatibiliteit met softwareframeworks uit de industrie garandeert. Het AMD ROCm-platform bevat een verscheidenheid aan drivers, ontwikkeltools en API's die het programmeren van versnellers vergemakkelijken, van kernelniveau tot applicaties voor eindgebruikers, en die kunnen worden aangepast aan uw specifieke vereisten. AMD ROCm is vooral geschikt voor toepassingen op het gebied van high-performance computing (HPC), kunstmatige intelligentie (AI) en wetenschappelijke computing. Bovendien biedt het AMD ROC-platform ondersteuning voor multi-accelerator computing, waaronder remote direct memory access (RDMA) voor server-knooppuntcommunicatie.



PORTFOLIO DELL POWEREDGE SERVERS

De investering van Dell in AMD creëert een cruciale keuze in de markt om AI te democratiseren, zoals blijkt uit hun vier serverplatforms met EPYC en hun vlaggenschip Dell PowerEdge XE9680-rackserver met AMD Instinct MI300X versnellers. De nieuwste generatie Dell PowerEdge servers aangedreven door AMD EPYC-processors verbeteren de zakelijke flexibiliteit en de time-to-market, met de mogelijkheid om transformatieve workloads zoals databases en analytics, virtualisatie, softwaregedefinieerde storage, virtuele desktopinfrastructuur (VDI), containerisatie, high performance computing (HPC), AI en machine learning (ML) te ondersteunen. Hun rackservers met één socket (één CPU) bieden een kostenefficiënte balans tussen prestaties en storagecapaciteit, ontworpen om naadloos mee te groeien met uw bedrijf, terwijl hun rackservers met twee sockets (dubbele CPU) geschikt zijn voor veeleisendere workloads met een breed scala aan functies.

De Dell PowerEdge XE9680 rackserver is een robuuste machine op het gebied van dataverwerking die speciaal is ontworpen voor AI-taken. Deze ondersteunt acht versnellers, ideaal voor machine learning- (ML)/deep learning-training (DL) en inferentieworkloads, met name voor het trainen van grote taalmodellen (LLM's). De Dell PowerEdge XE9680-rackserver met AMD Instinct MI300X-versnellers is uitgerust met acht MI300X-versnellers, elk met 192 GB aan 5,3 TB/s High Bandwidth Memory (HBM3), wat leidt tot een totale HBM3-capaciteit van 1,5 TB per server en meer dan 21 petaflops aan FP16-prestaties. Hiermee wordt Gen AI nog toegankelijker voor ondernemingen. Hierdoor kunnen ze grotere modellen trainen, de voetafdruk van datacenters minimaliseren, de TCO verminderen en een concurrentievoordeel behalen.

Samenvatting

Het snelle tempo van innovatie dat door AI wordt aangestuurd, zorgt sneller dan enige andere technologische transformatie voor een revolutionaire versnelling van workloads in datacenters. Om deze technologische vooruitgang te ondersteunen, werken Dell en AMD aan een meer inclusief, innovatief en ethisch ontwikkeld AI-ecosysteem dat ontwikkelaars uit alle sectoren stimuleert om samen te werken aan open source-bronnen en de huidige generatie AI-innovatie te stimuleren. Of uw AI-oplossing nu voldoet aan uw prestatievereisten op AMD EPYC-processors of op servers met AMD Instinct Accelerators, wij bieden de flexibiliteit om uw AI-workload uit te voeren op al onze hardwareplatforms, zodat u kunt profiteren van het beste wat Dell en AMD te bieden hebben.

VERWIJZINGEN

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

Dell afbeeldingen: **Dell.com**