

Stärka företag med AI: En era av alternativ



Innehållsförteckning

Möjlighet att förändra branscher med AI 1

AI inom olika branscher 4

Vad IT-beslutsfattare måste tänka på..... 5

Komma igång: AI i detalj..... 5

Avgörande val..... 6

 Prestanda 6

 Datasäkerhet 6

Skalning av lösningar..... 7

 Balansera kostnader och innovation 7

 Enkelhet och flexibilitet..... 7

 Säkerställa förklarbarhet 7

Scenarier från verkligheten..... 8

Detaljhandel 8

Hälso- och sjukvård 9

Våra lösningar 10

AI är för alla: DELL och AMD demokratiserar AI 10

Hugging Face-samarbete 11

AMD EPYC™-processorer 11

AMD Instinct™ MI300X-acceleratorer 11

AMD ROCm™ 6-mjukvaruplattform med öppen källa 12

Produktblad över Dell PowerEdge™-servrarna 12

Sammanfattning 13

Möjlighet att förändra branscher med AI

Idag är möjligheten att omvandla ditt företag för framtidens innovation som störst tack vare AI. Data som samlats in från Accenture Vision Technology 2023 har visat att 98 % av de globala cheferna är överens om att AI-grundmodeller kommer att spela en viktig roll i organisationens strategier under de närmaste tre till fem åren.¹

AI har blivit otroligt användbart för företag inom områden som detaljhandel, hälso- och sjukvård och finansiella tjänster på grund av dess förmåga att öka effektiviteten i uppgifter, driva innovation och förbättra beslutsprocesser. Men trots fördelarna finns fortfarande ett upplevt hinder vid integrering av AI på grund av några vanliga missuppfattningar.



Du behöver ett team av AI-utvecklare för att komma igång:

Även om kompetens inom datavetenskap fortfarande är värdefullt för att utveckla avancerade AI-lösningar och förstå de underliggande principerna, är det inte längre en nödvändig förutsättning. Det har skett en spridning av användarvänliga AI-verktyg, plattformar som Hugging Face och uppgiftsspecifika modeller som undanröjer mycket av komplexiteten med att utveckla AI-lösningar.

Du måste spendera tiotals miljoner på hårdvara för att få resultat:

Denna missuppfattning undergräver allvarligt den mångfald av AI-resurser som finns tillgänglig idag. Även om dessa allmänt kända resurser ofta är kraftfulla och väl understödda, är de kanske inte alltid det mest lämpliga eller kostnadseffektiva valet för varje företag.

Du måste arbeta oavbrutet för att skaffa acceleratorer:

Även om acceleratorer utmärker sig vid tunga AI-arbetsbelastningar, kanske företag inte behöver så mycket datorkraft för sina AI-program. Att vänta alltför lång tid på att få tillgång till marknadsledande acceleratorer är helt enkelt inte realistiskt. I många fall kan AI-optimerade processorer faktiskt leverera den prestanda och effektivitet som krävs för att producera AI-assisterade analyser och beslut i realtid, vilket gör dem till en mycket mer kostnadseffektiv och anpassningsbar lösning.

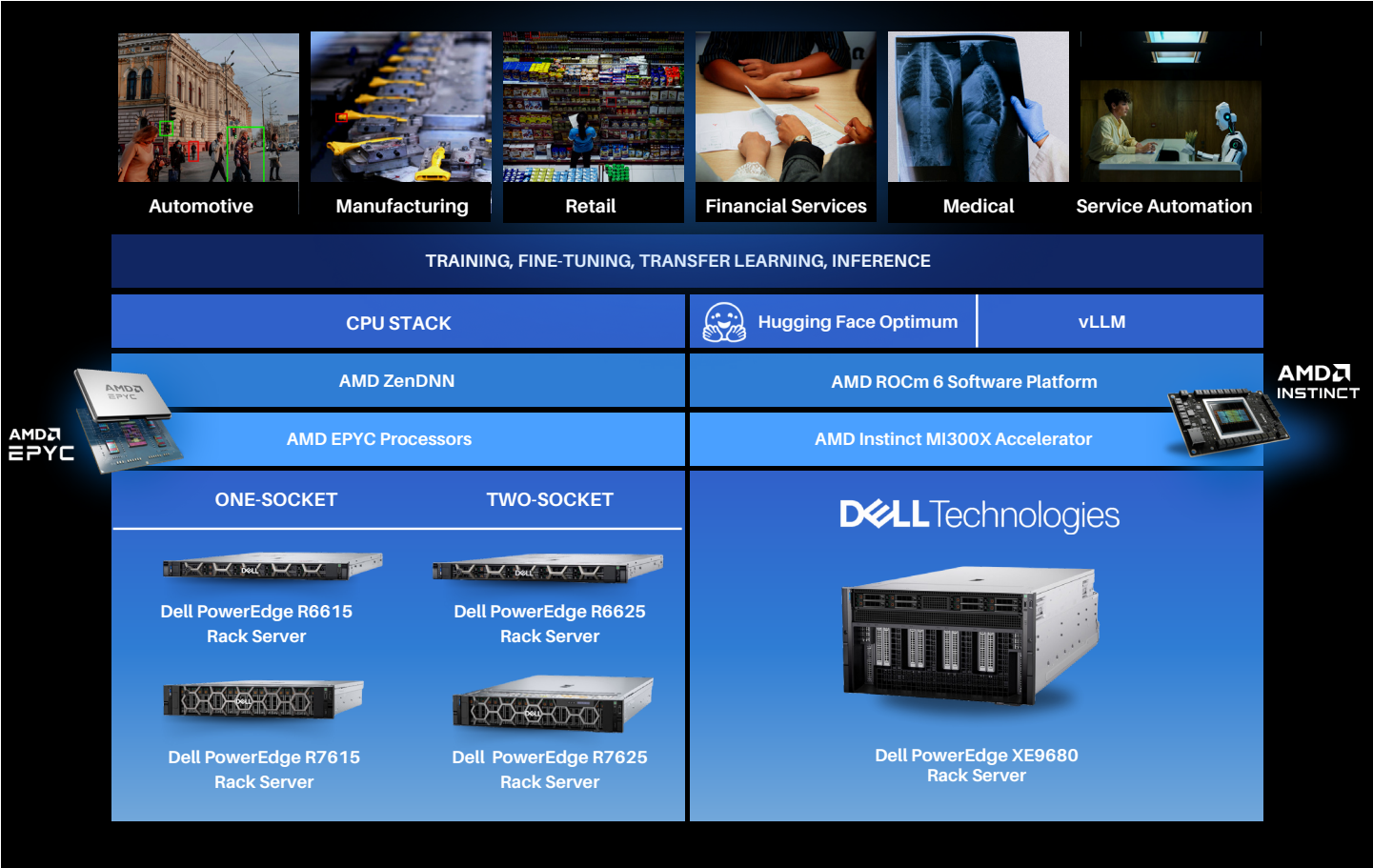
¹ Accenture, 30 mars 2023, "Accenture Technology Vision 2023: Generative AI to Usher in a Bold New Future for Business, Merging Physical and Digital Worlds", <https://newsroom.accenture.com/news/2023/accenture-technology-vision-2023-generative-ai-to-usher-in-a-bold-new-future-for-business-merging-physical-and-digital-worlds>



Lyckligtvis utvecklas AI-landskapet. **Dell** och **AMD** arbetar tillsammans för att slå hål på dessa myter genom att göra AI-teknik och AI-verktyg tillgängliga för fler användare med heltäckande infrastruktur utformad för att stödja dagens AI-krav.

Du kan komma igång med en redan optimerad modell, en tillförlitlig mjukvarustack och ett mångsidigt hårdvarusystem, som alla är öppet tillgängliga genom Dells och AMD:s samarbete. Att ha tillgång till allt knappare acceleratorer, en betydande grupp skickliga AI-tekniker eller resurser för att distribuera massiva molnkluster är inte längre ett krav för att dra nytta av AI.

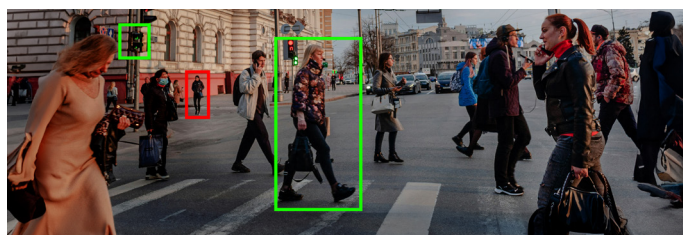
Dell och AMD:s samarbete är ett enhetligt ekosystem av hårdvara och mjukvara, utformat så att utvecklare kan skapa heltäckande AI-lösningar där överföringsinläring, finjustering och inferenser inkorporeras enkelt och effektivt. Med stöd från **Hugging Face** har vi nu en växande portfölj av modeller som körs på Dell PowerEdge-servrar med AMD EPYC™-processorer eller AMD Instinct™ MI300X-acceleratorer, så att utvecklare kan finjustera, tillämpa överföringsinläring och distribuera för inferens. Investeringarna i AMD ROCm™ och AMD ZenDNN™ samt samarbeten med PyTorch-, Tensorflow- och ONNX Runtime-ramverk är de grundläggande aktiverarna för utvecklare av tillämpad AI som upplever demokratiseringen av AI. Diagrammet nedan beskriver de komponenter som utgör Dells och AMD:s enhetliga AI-ekosystem.



AI inom olika branscher

Med diversifieringen av resurser och tonvikt på öppen källkodsinnovation migrerar AI till många olika branscher, som kundservice, finans och bank, hälso- och sjukvård samt detaljhandel för att nämna några. Inom dessa branscher gör AI det kollektivt möjligt för organisationer att frigöra potentialen hos sina egna äganderättsskyddade data och omarbete sina AI-arbetsflöden genom att ta itu med följande nyckelfunktioner: dataanalys, automatisering, personlig anpassning och prediktiv analys. AMD ROCm- och ZenDNN-bibliotek accelererar dessutom dessa AI-arbetsflöden för att leverera resultat i nära nog realtid.

Nedan kan du ta en närmare titt på hur AI influerar olika branscher.



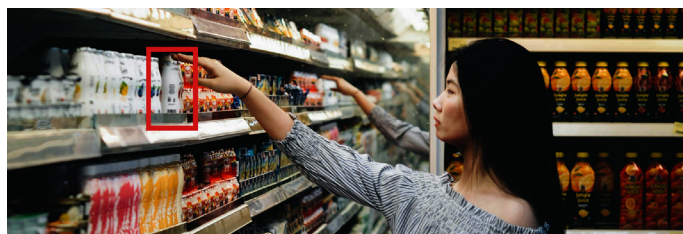
Fordon

AI används för objekt-detektering, körfältsspårning och beslutsfattande i autonoma fordon. AI kan också förutsäga när en fordonskomponent troligtvis slutar fungera, vilket möjliggör proaktivt underhåll och minskade driftavbrott.



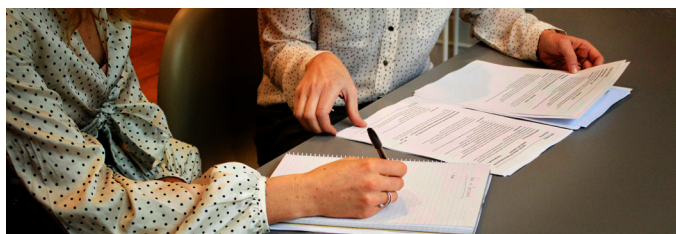
Tillverkning och industri

AI kan användas inom tillverkning och industri för förutsägbart underhåll, kvalitetskontroll, processoptimering och hantering av leveranskedjan, vilket ger förbättrad effektivitet och minskade driftavbrott.



Detaljhandel

AI kan analysera kundernas beteende för att ge personliga produktrekommendationer, förbättrat kundengagemang och ökad försäljning. Den kan också optimera lagernivåerna genom att förutsäga efterfrågan och minimera överskottslager eller att lager tar slut.



Finansiella tjänster

AI kan användas inom finans och bank för bedrägeridetektering, riskbedömning, kundservice och investeringsanalys, vilket leder till förbättrad säkerhet och bättre informerat beslutsfattande.



Medicin

AI kan användas inom hälso- och sjukvård för en mängd olika tillämpningar, inklusive medicinsk bildanalys, sjukdomsdiagnos, individuellt anpassad behandlingsplanering och framtagning av nya läkemedel, vilket leder till förbättrad patienthälsa och minskade kostnader.



Serviceautomatisering

AI-drivna chatbottar kan hantera kundförfrågningar och ge support, vilket minskar behovet av mänsklig inblandning. AI kan också automatisera repetitiva uppgifter som datainmatning eller dokumentbearbetning, vilket förbättrar effektiviteten och minskar antalet fel.

Vad IT-beslutsfattare måste tänka på

KOMMA IGÅNG: AI I DETALJ

Innan vi går igenom dessa användningsfall tar vi en närmare titt på AI-livscykeln. Livscykeln för AI (artificiell intelligens) avser de stadier som är involverade i att utveckla, distribuera och underhålla ett AI-system. Även om specifika metoder och terminologi kan variera, inkluderar en typisk AI-livscykel alltid modellträning och inferens.

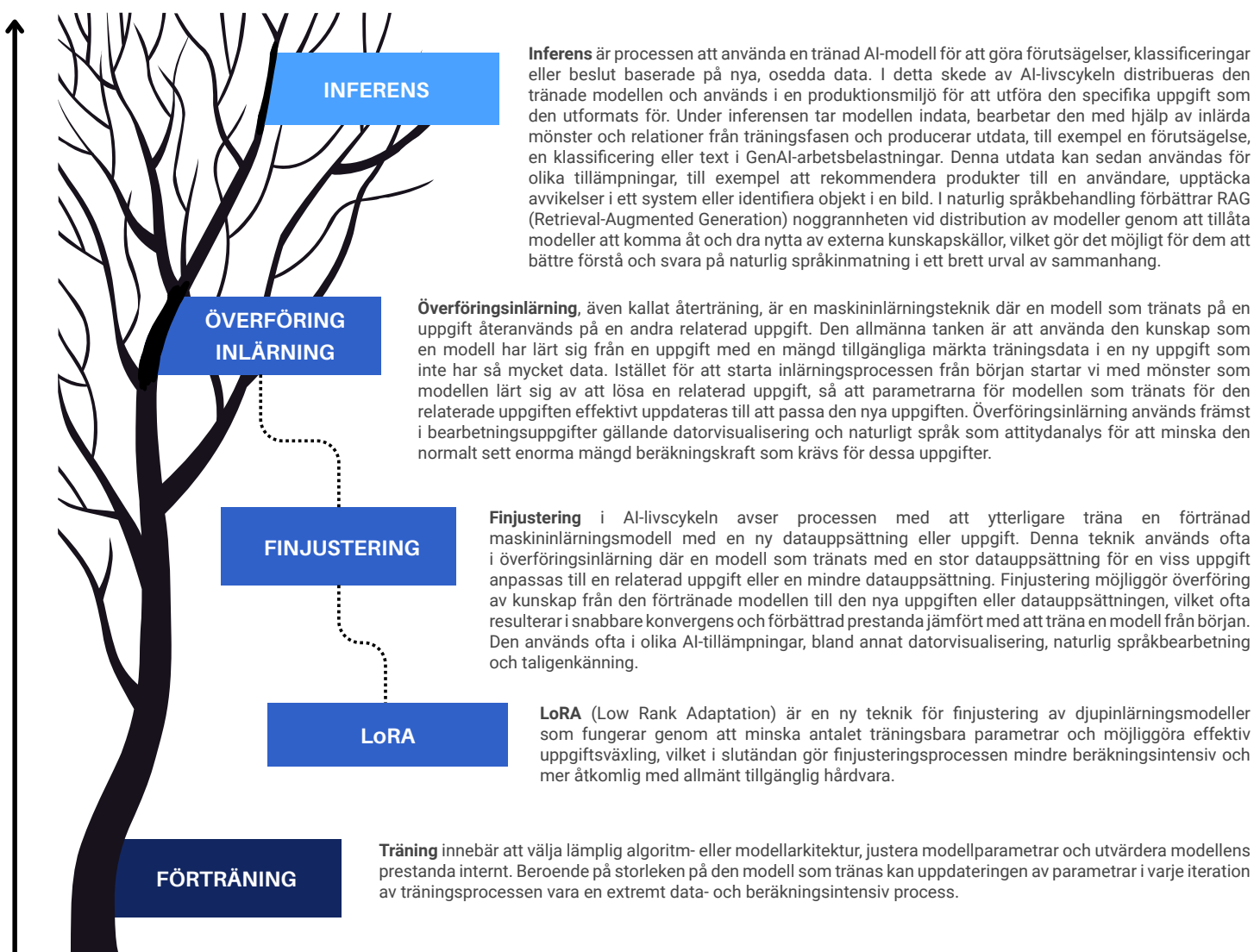


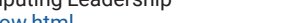
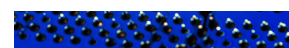
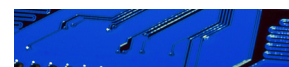
Bild 1: AI-livscykeln



AVGÖRANDE VAL

| Prestanda

I många av dessa verkliga tillämpningar är beslutsfattande i realtid eller nära nog realtid avgörande för framgång. Till exempel måste bedrägliga aktiviteter i finansiella transaktioner eller försäkringsanspråk snabbt identifieras för att förhindra ekonomiska förluster och skydda affärstillgångar. I ett tillverkningsscenario måste defekter i monteringslinjen eller fabriksinställningar övervakas dynamiskt för kvalitetssäkring. Processorn som hanterar din inferensarbetsbelastning måste optimeras för snabb och effektiv bearbetning av inkommande dataströmmar. Dell PowerEdge-serverar och AMD EPYC-processorer är en mångsidig kombination som är väl lämpad för hantering av kantinferensarbetsbelastningar samt uppgifter som omfattar högpresterande databehandling, molnanvändning och Big Data-analys.



| Datasäkerhet

Datasäkerhet är avgörande för AI-systems framgång, särskilt de som utnyttjar generativ AI, och är en viktig fråga för teknikledare som siktar på att införliva AI i sin verksamhet. AI-system förlitar sig vanligtvis på enorma mängder data, vilka kan inkludera känslig och konfidentiell information som personuppgifter, finansiella data eller äganderättsskyddad information. Att skydda dessa data är avgörande för att förhindra obehörig åtkomst eller datastöld samt för att säkerställa AI-modellernas och förutsägelsernas noggrannhet, tillförlitlighet och överensstämmelse.

Konfidentiell datoranvändning är en teknik som underlättar databehandling i en säker enklav och skyddar den från obehörig åtkomst eller manipulering av obehöriga parter, inklusive molnleverantören och andra användare.² Kryptering och andra säkerhetsåtgärder används för att isolera data under bearbetning. AMD Infinity Guard, en samling avancerade säkerhetsfunktioner som är integrerade i AMD EPYC-processorer, stöder konfidentiell datoranvändning genom att använda SEV (Secure Encrypted Virtualization), som krypterar virtuella maskiner (VM) med hjälp av en nyckel som bara är känd för processorn. Dessa tjänster syftar till att tillhandahålla hårdvarubaserade betrodda körningsmiljöer med hjälp av SEV-SNP (AMD SEV-Secure Nested Paging), vilket förbättrar gästskyddet för att skydda mot externa hot.

Federerad inlärning är en annan metod för att upprätthålla datasäkerhet. Den tränar en central modell över decentraliserade enheter eller servrar.³ Istället för att överföra alla data till en central plats tränar varje enhet modellen lokalt, och endast modelluppdateringar delas. Detta tillvägagångssätt bevarar integriteten och möjliggör samarbetande inlärning utan att dela rådata. Den federerade AI-plattformen från Dell Technologies gör det möjligt att köra beräkningsprocesser och AI- och ML-algoritmer på datauppsättningar i nätverkskanten när de samlas in, och endast dela matematiska modeller, metadata och frågeresultat över nätverket till andra kantenheter, datacenter eller molnet. Detta utbyte förbättrar resultaten genom att möjliggöra extraktion av användbara insikter i nära nog realtid från stora, distribuerade datauppsättningar utan att avslöja data och någon immateriell egendom.

² Advanced Micro Devices, Inc., 30 augusti 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, "Data Center Solutions, Confidential Computing", lösningsöversikt, <https://www.AMD.com/content/dam/AMD/en/documents/EPYC-business-docs/solution-briefs/confidential-computing-solution-brief.pdf>

³ Analytics Vidhya, december 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, "A federated learning platform for real-time artificial intelligence", lösningsöversikt, <https://www.Delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/dt-sb-analytics-anywhere.pdf>

SKALNING AV LÖSNINGAR

| Balansera kostnader och innovation

Att hitta rätt balans mellan kostnad och innovation säkerställer att AI-lösningar inte bara är ekonomiskt genomförbara utan att de även är effektiva, vilket ger verkligt värde för både företag och användare. En viktig komponent för att hitta denna balans ligger i att identifiera en hårdvara som både löser dina användningsfall och integreras enkelt i befintlig infrastruktur. Olika branschers ökade efterfrågan på acceleratorer bidrar, utöver begränsningar i produktionskapacitet, logistikutmaningar och halvledarbrist, även till brist på acceleratorer på den moderna AI-hårdvarumarknaden.

Processorer är emellertid redan en standardkomponent i de flesta datacenter, vilket gör integreringen enklare och mer kostnadseffektiv jämfört med om helt ny acceleratorhårdvara läggs till. AI-optimerade processorer kan dra nytta av befintlig mjukvara och befintliga verktyg, vilket minskar behovet av omfattande byte av verktyg eller återträning. Processorer erbjuder också större flexibilitet och effektivitet för ett brett spektrum av uppgifter utöver AI, vilket möjliggör en mer mångsidig användning av resurser inom datacentret. Genom att uppdatera datacentret med Dell PowerEdge-servrar som kör AMD EPYC-processorer stöds fullgörandet av dina befintliga arbetsbelastningar, samtidigt som det är redo för avancemang mot mer innovation och effektivitet med AI.

| Enkelhet och flexibilitet

Enkelhet och flexibilitet i ditt AI-system är avgörande för att bygga AI-lösningar som är effektiva, anpassningsbara och skalbara på lång sikt. Att ha tillgång till en rad olika ramverk och optimeringar som kompletterar din hårdvara förbättrar prestandan utan att du behöver ägna extra tid och ansträngning åt integration över plattformar. Dessa egenskaper är särskilt viktiga för att hantera blandade AI-arbetsbelastningar, vilket innebär en kombination av olika typer av AI-uppgifter som träning, inferens och databehandling.

AMD och Dell Technologies hanterar blandade AI-arbetsbelastningar genom en kombination av hårdvaru- och mjukvarulösningar. AMD EPYC-processorer ger högpresterande datorkraft, med funktioner som SMT (simultaneous multithreading) och ett stort antal kärnor, vilket möjliggör effektiv parallell bearbetning för AI-arbetsbelastningar. De här processorerna är optimerade för AI-uppgifter och erbjuder hög prestanda för både tränings- och inferensarbetsbelastningar. Dell PowerEdge-servrar, utrustade med AMD EPYC-processorer, erbjuder en skalbar och flexibel plattform för distribution av AI-arbetsbelastningar. Dessutom erbjuder Dell OpenManage-mjukvarusviten hanteringsverktyg för att optimera resursfördelning och prestandaövervakning för blandade AI-arbetsbelastningar.

AMD erbjuder också UIF (Unified Inference Frontend), som utnyttjar de prestandaförbättrade versionerna av var och en av dagens mjukvarustackar och bygger på AMD ZenDNN-biblioteket för AMD EPYC-processorer, AMD ROCm-stacken med öppen källa för AMD Instinct-acceleratorer samt en mjukvarustack för AMD Adaptive SoCs. AMD ROCm är dessutom utformad för att fungera med en mängd olika AMD-processorer och AMD-acceleratorer, som omfattar både professionella och konsumentklassade produkter.

| Säkerställa förklarbarhet

Förklarande AI har en avgörande roll för att säkerställa transparens, tillförlitlighet och effektivitet i tillämpningar med artificiell intelligens. Förklarande AI ger insikter i hur AI-modeller fattar beslut, och ger information om de underliggande faktorerna och resonemangsprocesserna. Denna transparens är avgörande för att få intressenternas förtroende, särskilt inom känsliga områden som hälso- och sjukvård, finans och straffrätt, där beslut direkt påverkar människors liv.

Human-in-the-loop-AI-system drar nytta av mänsklig intelligens för att förbättra AI-prestandan och minska algoritmisk snedvridning. Genom att integrera mänsklig tillsyn kan dessa system hantera komplexa och tvetydiga situationer mer effektivt, vilket säkerställer att AI-lösningar stämmer överens med etiska och sociala normer. Dessutom möjliggör mänsklig inblandning kontinuerlig förfining och anpassning av AI-modeller baserat på verklig feedback, vilket främjar iterativ förbättring och långsiktig pålitlighet. Dessa metoder är avgörande för att bygga ansvarsfulla och inkluderande AI-system som tjänar samhällets bästa.

Scenarier från verkligheten

Scalers AI samarbetade med Dell och AMD för att visa upp kapaciteten hos Dell PowerEdge-serverar utrustade med AMD-processorer. Läs om hur dessa tekniker utnyttjas för träning, överföringsinläring och inferens i scenarier inom detaljhandel samt hälso- och sjukvård.

DETALJHANDEL

Scalers AI byggde Retail Inventory Management-referenslösningen, ett system som är utformat för att övervaka och hantera lagernivåer på butikshyllor genom implementering av en AI-modell för objekt-detektering. Denna referenslösning utnyttjar SSD_MobileNet_V2-modellen för att identifiera och känna igen produkter på butikshyllor, vilket i slutändan möjliggör automatisk inventering och exakt övervakning av lagernivåer. Modellen genomgick överföringsinläring med hjälp av SKU110K-bilddatauppsättningen, bestående av 23 000 bilder från Roboflow. Genom att utnyttja datorvisualiserings- och maskininlärningsalgoritmer kan systemet upptäcka när artiklar håller på att ta slut eller är slut, och butikspersonalen informeras om att hyllorna eller lagret behöver fyllas på.

Den här lösningen använder Dell PowerEdge R7615-servern med AMD EPYC 9354P 32-kärnig processor.

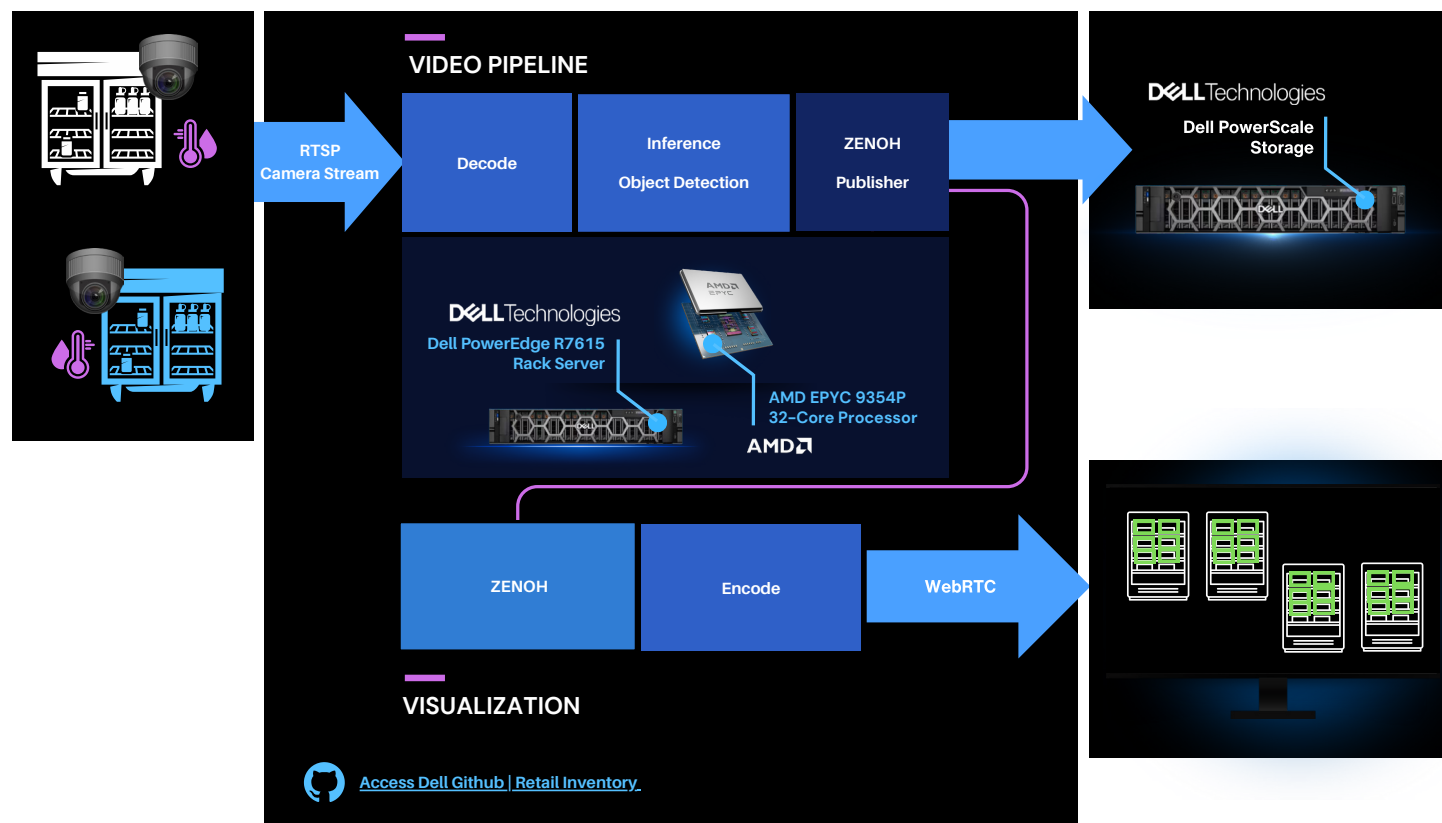


Bild 2: Arkitekturdiagram över Retail Inventory Management-referenslösningen

HÄLSO- OCH SJUKVÅRD

AI-driven medicinsk bildtagning är oerhört värdefull för sin förmåga att förbättra vården genom att förbättra diagnostisk noggrannhet och effektivitet, och ge vårdpersonal exakta insikter i förhållanden som kan vara svåra att upptäcka med blotta ögat. Genom att automatisera analysen av medicinska bilder minskar AI den tid som krävs för diagnos, vilket möjliggör snabbare behandlingsbeslut och i slutändan förbättrad patienthälsa.

Scalers AI utnyttjade kapaciteten hos Dell PowerEdge R7625-servern utrustad med AMD EPYC 9554 64-kärniga processorer för att skapa en AI-driven medicinsk bildtagningslösning för detektering av lunginflammation. Med avancerade algoritmer och maskininlärningstekniker för att analysera medicinska bilder, såsom röntgenbilder eller datortomografibilder, hjälper lösningen till att öka hastigheten och noggrannheten vid diagnos av lunginflammation hos patienter. I slutändan introducerar detta ett ytterligare lager av datorstödd granskning, vilket skapar potential att hjälpa vårdpersonal att hantera stora volymer bilddata mer effektivt.

Denna referenslösning använder ResNet50-modellen för att analysera röntgenbilder av bröstkorgen som erhållits från NIH Clinical Center-datauppsättningen. Dess primära mål är att upptäcka förekomst eller frånvaro av lunginflammation, och den utför i huvudsak en binär klassificering. Modellen har tränats med hjälp av DICOM-röntgendatauppsättningen från NIH Clinical Center-datauppsättningen, vilket innefattar överföringsinlärning med ResNet50-arkitekturen.

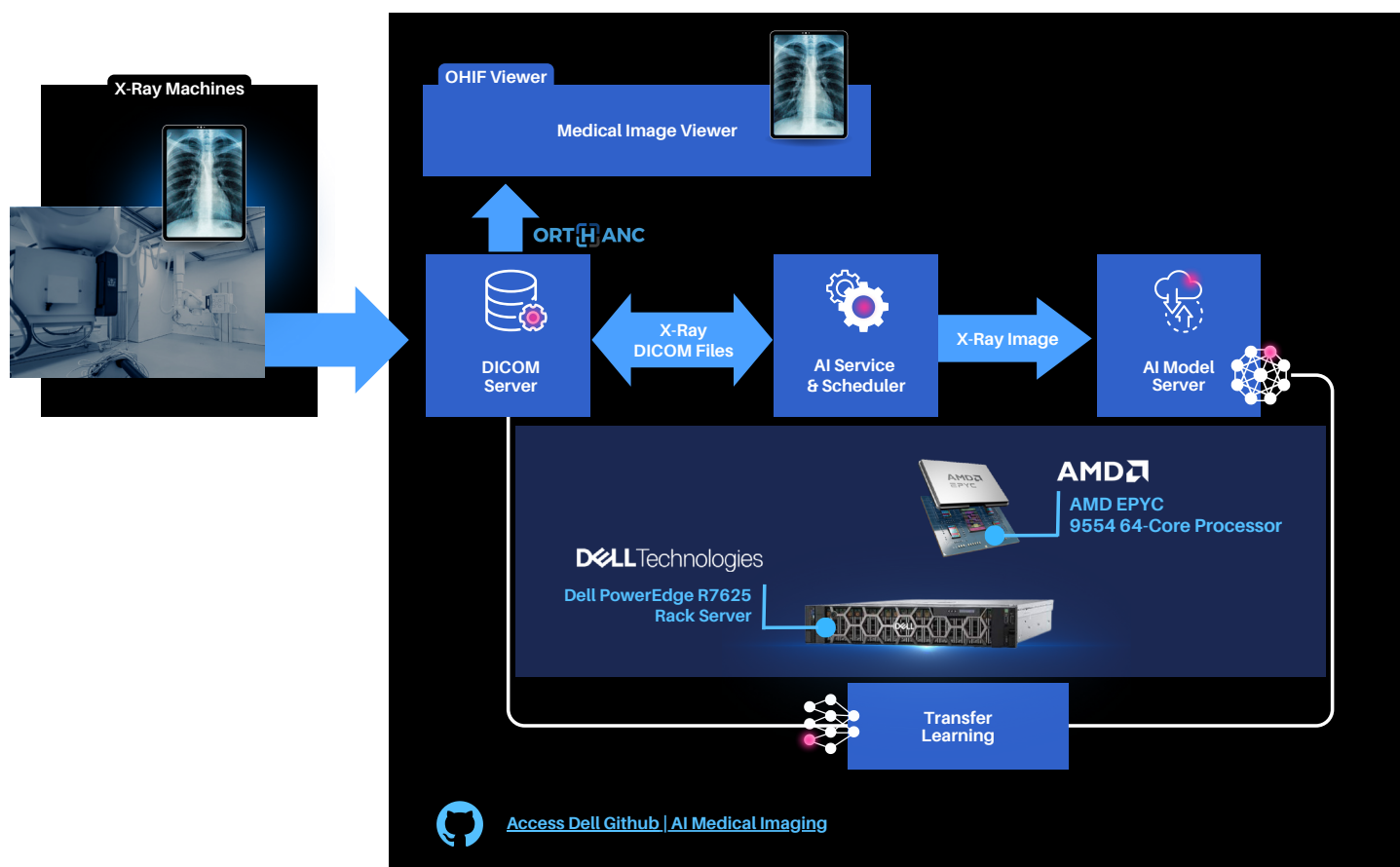
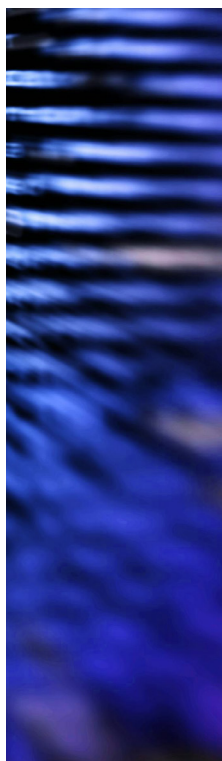
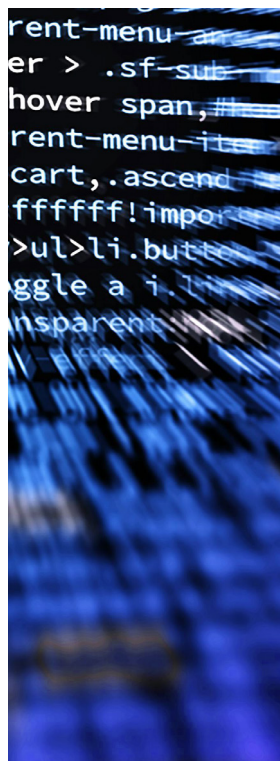


Bild 3: Arkitekturdiagram av den medicinska AI-bildtagningslösningen

Våra lösningar

AI ÄR FÖR ALLA: DELL OCH AMD DEMOKRATISERAR AI

Detta samarbete lägger grunden för demokratiseringen av AI, vilket är nödvändigt för att främja innovation och stödja inkludering i AI-ekosystemet. Dell och AMD uppnår detta resultat genom att ge individer och organisationer möjlighet att utnyttja AI och lösa unika utmaningar inom sina respektive områden med en tillgänglig uppsättning kraftfulla servrar utrustade med toppmodern AMD-processor- och acceleratorteknik. Dell PowerEdge-servrar med AMD Instinct MI300X-acceleratorerna kan hantera stora AI-arbetsbelastningar som att träna och finjustera stora språkmodeller (LLM), medan Dell PowerEdge-servrar utrustade med AMD EPYC-processorer utmärker sig när det gäller hantering av kantinferensarbetsbelastningar. Utöver den underliggande hårdvaruplattformen erbjuder AMD även mjukvarubiblioteket ZenDNN för optimering av djupinlärningsinferens på AMD-processorer, samt mjukvarubiblioteket AMD ROCm för att förbättra tränings-, finjusterings- och inferensfunktioner på AMD Instinct-acceleratorer. Alla dessa alternativ är smidigt sammankopplade i AMD:s UIF (Unified Inferencing Model), genom vilken användare kan konstruera heltäckande AI-lösningar, med flexibilitet i valet av mjukvaruramverk, mjukvaruoptimering och hårdvaruplattform.



HUGGING FACE-SAMARBETE

Företag som är angelägna om att införa AI kan börja med att utnyttja befintliga modeller eller AI-arbetsflöden som är skräddarsydda för deras specifika behov direkt från Hugging Face, en plattform med öppen källa avsedd för datavetenskap och maskininlärning. AMD har inlett ett samarbete med Hugging Face, med det gemensamma målet att leverera förstklassig omvandlingsprestanda genom att lägga till AMD-specifika mjukvaruoptimeringar till mjukvarubibliotek och ramverk som redan integreras sömlöst med AMD-plattformar. Hugging Face samarbetar aktivt med teknikteamet på AMD för att optimera nyckelmodeller för toppprestanda, införliva AMD ROCm i Transformers-biblioteket och förbättra Optimum-AMD, ett bibliotek särskilt utformat för AMD-plattformar, för att hjälpa Hugging Face-användare att använda dem med minimala kodändringar.

Dell Technologies har också nyligen gått samman med Hugging Face för att förenkla processen för företag att utveckla, finjustera och tillämpa sina egna modeller för generativ AI (Gen AI) med öppen källa med hjälp av Hugging Face-communityn, allt på branschledande produkter och tjänster för Dell-infrastruktur. På Hugging Face-plattformen håller en ny Dell-portal på att utvecklas, som kommer att innehålla anpassade, dedikerade behållare och skript för att hjälpa användare att på ett säkert och enkelt sätt distribuera modeller med öppen källa som finns tillgängliga på Hugging Face med Dells servrar och datalagringsystem. Företag kan nu dra full nytta av Hugging Face-resurser för att direkt distribuera modeller på Dell PowerEdge-servrar med AMD-processorer och konstruera heltäckande AI-lösningar med hjälp av egna äganderättsskyddade data.

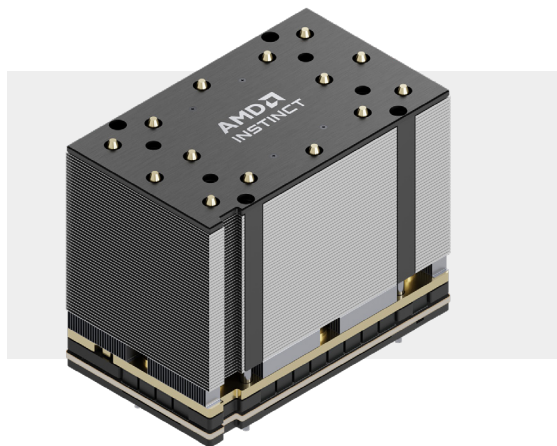


AMD EPYC-PROCESSORER

AMD tillhandahåller de tekniska framsteg som krävs för moderna molnbaserade datacenter genom sina AMD EPYC-processorer. Dessa processorer är ett systemchipp (system on chip, SoC) som är utformat från grunden för att effektivt möta kraven från nuvarande och framtida datacenter. Processorerna i AMD EPYC 9000-serien utrustar datacentret med upp till 128 kärnor, 256 trådar, 12 minneskanaler som stöder upp till 6 TB minne per sockel och 128 PCIe Gen5-banor. Detta kopplas till branschens banbrytande hårdvaruinbäddade x86-serversäkerhetslösning. Genom att integrera viktiga beräknings-, minnes-, I/O- och säkerhetsresurser i SoC ger AMD EPYC-processorererna toppprestanda och bidrar till en lägre total ägandekostnad (TCO).

AMD INSTINCT MI300X-ACCELERATORER

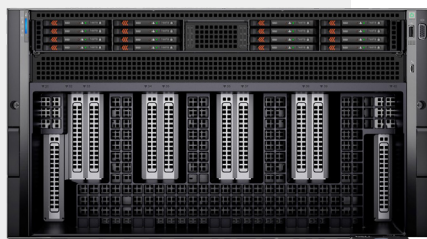
AMD Instinct MI300X-acceleratorn, som bygger på den senaste AMD CDNA 3-arkitekturen, erbjuder branschledande effektivitet och prestanda för de mest intensiva AI- och HPC-tillämpningarna. Den är utrustad med 304 högpresterande beräkningsenheter och har AI-specifika funktioner som stöd för nya datatyper och foto- och videoavkodning, samt ett enastående 192 GB HBM3-minne på en enda accelerator.



AMD ROCm 6-MJUKVARUPLATTFORM MED ÖPPEN KÄLLA

AMD ROCm 6-mjukvaruplattformen med öppen källa är optimerad för att maximera högpresterande datorsystem (HPC) och AI-arbetsbelastningsprestanda från AMD Instinct MI300X-acceleratorer. Den utökar också stödet för AMD Instinct MI300X-acceleratorer, vilket säkerställer kompatibilitet med branschramverk. AMD ROCm-plattformen omfattar en mängd olika drivrutiner, utvecklingsverktyg och API:er som underlättar acceleratorprogrammering från kärnnivå till slutanvändarprogram och kan skräddarsys för att passa dina specifika krav. AMD ROCm är särskilt lämpad för tillämpningar inom högpresterande datorsystem (HPC), artificiell intelligens (AI) och vetenskapliga beräkningar. Dessutom erbjuder AMD ROCm-plattformen stöd för beräkning med flera acceleratorer, inklusive RDMA (Remote Direct Memory Access) för server-nod-kommunikation.

AMD
ROCm



PRODUKTBLAD ÖVER DELL POWEREDGE-SERVERARNA

Dells investering i AMD skapar ett avgörande val på marknaden för att demokratisera AI, vilket framgår av deras fyra serverplattformar med EPYC och deras flaggskepp Dell PowerEdge XE9680-rackserver med AMD Instinct MI300X-accelerator. Den senaste generationen Dell PowerEdge-serverar som drivs av AMD EPYC-processorer förbättrar både verksamhetsagilitet och marknadsintroduktion med möjlighet att stödja transformativa arbetsbelastningar, som databaser och analyser, virtualisering, mjukvarudefinierad lagring, VDI (Virtual Desktop Infrastructure), containerisering, högpresterande datorsystem (HPC), AI och maskininlärning (ML). Deras rackserverar med en sockel (en processor) ger en kostnadseffektiv balans mellan prestanda och lagringskapacitet, utformad för att växa

sömlöst med ditt företag, medan deras rackserverar med två socklar (dubbla processorer) har utrymme för mer krävande arbetsbelastningar med ett brett utbud av funktioner.

Dells rackserver PowerEdge XE9680 är en robust databearbetningsmaskin som är anpassad specifikt för AI-uppgifter. Den har stöd för åtta acceleratorer, vilket gör den perfekt för träning och inferens med maskininlärning (ML)/djupinlärning (DL), särskilt för att träna stora språkmodeller (LLM). Utrustad med åtta MI300X-acceleratorer, var och en med 192 GB med 5,3 TB/s HBM3-minne (High Bandwidth Memory), vilket ger en total HBM3-kapacitet på 1,5 TB per server och över 21 petaflops av FP16-prestanda, är rackservern Dell PowerEdge XE9680 med AMD Instinct MI300X-accelerator redo att ytterligare utöka Gen AI-tillgängligheten till företag. Detta gör det möjligt för dem att träna större modeller, minimera datacentrets täckningsområde, minska den totala ägandekostnaden och få en konkurrensfördel.

Sammanfattning

Den snabba innovationstakten som drivs av AI revolutionerar arbetsbelastningar i datacenter snabbare än någon annan teknisk omvandling. För att stödja dessa tekniska framsteg arbetar Dell och AMD mot ett mer inkluderande, innovativt och etiskt utvecklat AI-ekosystem som uppmuntrar utvecklare från alla branscher att samarbeta om resurser med öppen källa och driva dagens Gen AI-innovation. Oavsett om din AI-lösning uppfyller dina prestandakrav på AMD EPYC-processorer eller på servrar som drivs med AMD Instinct-acceleratorer, tillhandahåller vi flexibiliteten att köra din AI-arbetsbelastning på våra hårdvaruplattformar, så att du kan dra nytta av det bästa som Dell och AMD har att erbjuda.

REFERENSER

AMD-bilder: AMD.com, AMD Partner Resource Library, <https://www.amd.com/en/partner/resources/resource-library.html>

Dell-bilder: [Dell.com](https://www.dell.com)