White Paper

Waarom het ontwikkelen en installeren van AI-technologie op werkstations een logische keus is

Sponsored by: Dell Technologies

Peter Rutten Dave McCarthy
July 2023

IDC OPINION

AI heeft in alle sectoren een vlucht genomen als een belangrijke, onderscheidende voorziening en de hardware die nodig is om AI uit te voeren, ontwikkelt zich snel. De technologie-industrie is vaak erg gefocust op de exponentiële groei qua omvang die de meest geavanceerde AI-modellen doormaken. De discussies gaan over tientallen miljarden parameters, het verminderen van precisie, het uitbreiden van geheugen, HPC-achtige (High-Performance Computing) behoeften voor AI-training en -deducering en racks met snellere servers. In werkelijkheid is deze buitengewone schaal van AI-computing de uitzondering, vooral in het bedrijfsleven.

Tegenwoordig werken veel bedrijven hard aan AI-initiatieven, waaronder generatieve AI, waarvoor geen supercomputer nodig is. Veel AI-ontwikkeling, en in toenemende mate ook AI-implementatie, met name aan de edge, vindt namelijk plaats op krachtige workstations. Workstations bieden talloze voordelen voor de ontwikkeling en implementatie van AI. AI-wetenschappers en ontwikkelaars hoeven niet langer te onderhandelen over servertijd en kunnen beschikken over GPU-acceleratie, zelfs als servergebaseerde GPU's nog steeds niet eenvoudig beschikbaar zijn in het datacenter. Daarnaast zijn ze extreem betaalbaar ten opzichte van servers en vereisen ze een slechts kleinere, eenmalige uitgave in plaats van een snel oplopende rekening voor een cloud-instantie. En dan is er ook nog het comfort van de wetenschap dat gevoelige data veilig on-premise worden opgeslagen. Hierdoor bevrijden ze de wetenschappers en ontwikkelaars ook van de zorg dat ze hoge kosten maken terwijl ze goed beschouwd alleen maar experimenten met AI-modellen uitvoeren.

IDC ziet de edge als scenario voor AI-implementatie sneller groeien dan on-premise of de cloud. Ook hier spelen workstations een steeds belangrijkere rol als platforms voor AI-deducering, waarbij vaak niet eens GPU's nodig zijn, maar deducering in plaats daarvan wordt uitgevoerd op softwarematig geoptimaliseerde CPU's. De gebruiksscenario's voor AI-deducering op workstations aan de edge groeien snel en omvatten AIOps (asynchrone input/output per seconde), respons bij noodgevallen, radiologie, olie- en gasexploratie, landbeheer, telegezondheidszorg, verkeersleiding, bewaking van productiefaciliteiten en drones.

In deze whitepaper wordt gekeken naar de steeds grotere rol die workstations spelen bij de ontwikkeling en implementatie van AI en wordt kort de Dell portfolio van workstations voor AI besproken.

July 2023, IDC-nr. US51017223

OVERZICHT VAN DE SITUATIE

De AI-explosie en de impact op de infrastructuur

Wereldwijd zijn organisaties bezig met een steeds groeiend aantal AI-projecten. Nu al worden in alle bedrijfstakken veel taken uitgevoerd door software die geheel of gedeeltelijk wordt aangestuurd door een AI-model. IDC volgt AI op vele niveaus en één cijfer dat nuttig is om te bekijken, is het bedrag dat bedrijven en cloudserviceproviders naar verwachting zullen uitgeven aan servers voor het ontwikkelen en uitvoeren van AI. In 2026 zal dit 34,6 miljard dollar zijn, wat neerkomt op bijna 22% van de totale uitgaven aan servers wereldwijd.

Maar servers vormen niet het hele plaatje. De voorbereiding, ontwikkeling, prototyping en, in toenemende mate, *implementatie* van AI gebeurt op workstations. Nu organisaties, klein en groot, ontdekken dat er nieuwe zakelijke kansen kunnen worden gerealiseerd door hun applicaties te voorzien van een zekere mate van AI-functionaliteit, is het experimenteren met AI-modellen explosief gestegen. En robuuste workstations zijn ideaal voor dit doel, omdat ze direct beschikbaar zijn en zich dicht bij data bevinden.

Hoe komt het dat AI plotseling zo wijdverbreid is geworden? AI-algoritmen worden immers al tientallen jaren toegepast. Dat komt vooral omdat de afgelopen jaren twee essentiële voorwaarden voor het toepassen van een specifiek succesvol type AI-algoritme, het neurale netwerk, zijn gerealiseerd: de eenvoudige beschikbaarheid van enorme, goedkope en uiteenlopende soorten data, zoals niet-gestructureerde en semi-gestructureerde data, en de uitbreiding van lineaire rekencapaciteit met een parallel model waarmee deze neurale netwerken binnen een acceptabel tijdsbestek kunnen worden verwerkt. Nu aan deze twee basisvoorwaarden is voldaan, hebben datawetenschappers enorme vooruitgang kunnen boeken met het ontwikkelen van neurale netwerken die automatisch leren hoe ze steeds indrukwekkendere taken kunnen uitvoeren. Terwijl traditionele machine learning (ML) relevant blijft voor tekstuele of numerieke data, is deep learning (DL) effectiever voor video, audio, talen, enzovoort.

Traditionele modellen voor machine learning kunnen meestal worden ontwikkeld met de CPU's van een workstation, die hooguit enkele tientallen kernen hebben, maar neurale netwerken hebben coprocessoren nodig om taken verdeeld over duizenden kernen parallel te kunnen verwerken. De belangrijkste reden hiervoor is dat bij ML de extractie en classificatie van kenmerken een handmatig proces is, terwijl dit bij DL geautomatiseerd verloopt, waardoor het model met behulp van grote datasets moet worden getraind aan de hand van constante herhaling. Op dit moment is de GPU de meest gebruikte coprocessor, maar er komen ook nieuwe AI-specifieke processoren beschikbaar die door start-ups worden ontwikkeld. Dit type versnelling, waarbij een discrete coprocessor wordt gebruikt voor parallelle verwerking, heeft een revolutie teweeggebracht in de markt voor servers en workstations. Dit heeft geleid tot wat IDC massively parallel compute noemt.

In 2022 vormden versnelde servers een wereldwijde markt van 21,8 miljard dollar, die zal groeien tot 43,4 miljard dollar in 2026. 57% van dat totaal bestaat uit versnelde servers voor het uitvoeren van AI. Tegelijkertijd groeide het aantal discrete GPU's dat werd verkocht voor gebruik in workstations tot 6,4 miljoen in 2022. IDC schat dat de markt voor workstations die worden gebruikt voor wetenschappelijke doeleinden of software-engineering, waar AI-ontwikkeling een steeds belangrijkere rol speelt, zal toenemen tot bijna 2 miljard dollar in 2026.

Ontwikkelingsstadia voor AI

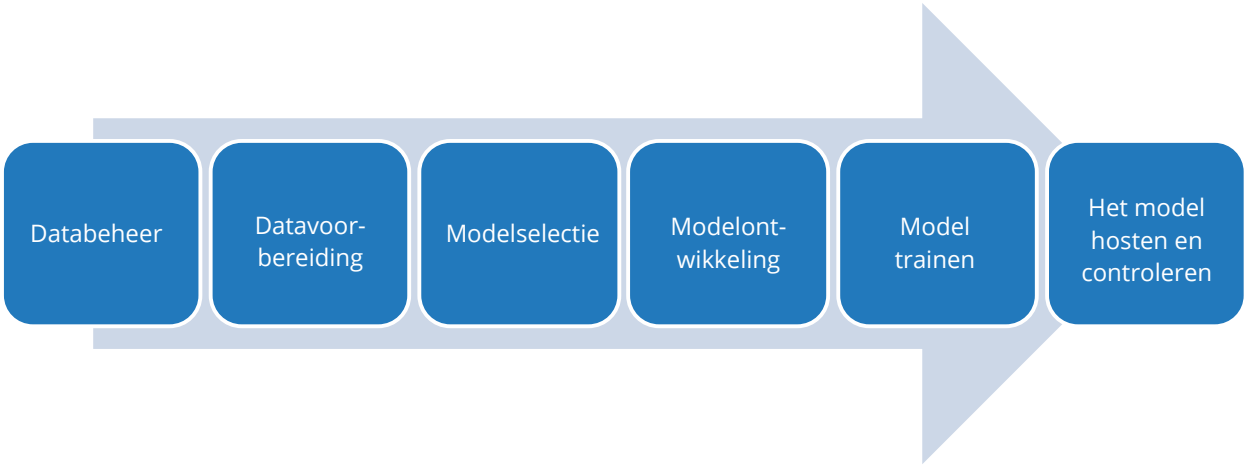
Zoals eerder vermeld, zijn neurale netwerken haalbaar geworden door de uitbreiding van datatypes en -volumes en nieuwe benaderingen voor het uitvoeren van berekeningen. Het eerste deel van deze vergelijking, datavolumes en -types, is bepaald geen onbelangrijk deel. Volgens sommigen gaat maar liefst 80% van de inspanningen in een AI-initiatief voor deep learning naar het beheren en voorbereiden van data. Data moeten worden opgenomen, beheerd en voorbereid voordat met het modelontwerp en de training kan worden gestart. Volgens IDC kan AI-ontwikkeling worden verdeeld in de volgende stadia (zie afbeelding 1):

- **Databeheer:** het identificeren en beheren van relevante data voor het AI-model uit de enorme hoeveelheden data in het datacenter, de edge en de cloud die door een organisatie worden opgenomen, gegenereerd en/of verworven. Deze data kunnen van elk type zijn, zowel gebeurtenisgestuurd of streaming, en voor veel ervan is mogelijk een vorm van governance nodig.
- **Datavoorbereiding:** het opslaan en opschonen van data (bestanden, blokken of objecten) in een datawarehouse of data lake, ervoor zorgen dat de data volledig en van hoge kwaliteit zijn en ze vervolgens transformeren in een vorm die ze bruikbaar maakt voor het AI-model, bijvoorbeeld met Spark of tools zoals Pandas.
- **Modelselectie:** bepalen welk model de AI-taak waarvoor het is geprogrammeerd het beste uitvoert wat betreft foutenpercentage en/of prestaties.
- **Modelontwikkeling:** het AI-model ontwerpen met frameworks zoals XGBoost, LightGBM, GLM, Keras, TensorFlow, PyTorch, Caffe, RuleFit, FTRL, Snap ML, scikit-learn of H2O.
- **Model trainen:** het trainen van het model op een compute-infrastructuur met voldoende processor- en/of coprocessorkernen voor parallelle verwerking van taken (steeds vaker ook met de mogelijkheid om de beslissingen van een model uit te leggen, te valideren en te documenteren om daarmee eerlijkheid, verantwoording en transparantie te garanderen). (Dit stadium omvat ook prototyping, het testen van het getrainde model door er deduceringen mee uit te voeren).
- **Het model hosten en controleren:** het model implementeren in een productieomgeving om de taak uit te voeren waarvoor het is ontworpen, meestal 'AI-deducering' genoemd, en de prestaties controleren.

Workstations kunnen in elk van deze zes stadia een belangrijke rol spelen, in combinatie met datacenter-, cloud- of edge-infrastructuur.

AFBEELDING 1

Ontwikkelingsstadia voor AI



Bron: IDC, 2023

AI-MODELLEN ONTWIKKELEN OP WORKSTATIONS

Workstations in vergelijking met personal computers

Het is algemeen bekend dat personal computers (pc's) niet krachtig genoeg zijn voor de ontwikkeling van AI. Datawetenschappers en AI-ontwikkelaars zijn meestal betrokken bij strategisch belangrijke projecten voor hun organisaties, en onbelemmerde productiviteit is van het grootste belang. Workstations presteren doorgaans voorspelbaarder dan pc's, omdat ze meestal gebouwd zijn met onderdelen die beter presteren en geoptimaliseerd zijn voor de software die erop wordt uitgevoerd.

Deze componenten zijn onder andere:

- **Hoogwaardige processoren:** bijvoorbeeld de Intel Xeon schaalbare processoren.
- **Krachtige GPU's:** een voorbeeld hiervan zijn de professionele RTX GPU's van NVIDIA, zoals de NVIDIA RTX 6000 Ada.
- **Meer storage:** sommige workstations kunnen wel 60 TB bieden en de I/O-snelheden zijn vaak aanzienlijk hoger dan die van pc's.
- **Meer geheugen:** er zijn nu workstations beschikbaar met wel 6 TB geheugen.
- **Koeling:** krachtige componenten genereren veel warmte en datawetenschappers hebben een workstation nodig met voldoende koeling, om oververhitting te voorkomen en optimale prestaties te behouden.
- **Netwerkinterfacekaart (NIC):** voor datawetenschappers die werken met grote datasets die zijn opgeslagen op externe servers, is een snelle netwerkinterfacekaart essentieel om data snel en efficiënt te kunnen overdragen.
- **Beeldscherm**: een beeldscherm van hoge kwaliteit is belangrijk voor het visualiseren van data; datawetenschappers hebben een monitor nodig met een hoge resolutie en kleurnauwkeurigheid, en een groot schermformaat.

- **ECC-geheugen:** ECC (Error-Correcting Code; foutcorrectiecode) detecteert en corrigeert de meest voorkomende soorten interne databeschadiging. Hierdoor kunnen blauwe schermen tijdens een lange AI-trainingsrun door ofwel een onherstelbare fout (slechte bit) of een herstelbare fout (omgekeerde bit, waardoor slechte waarden ontstaan) worden voorkomen. ECC zorgt ook voor nauwkeurige resultaten, een essentiële vereiste voor levenskritisch werk, zoals in de gezondheidszorg.
- **Gespecialiseerd silicium:** een voorbeeld hiervan zijn de Intel Movidius VPU's (Vision Processing Units), coprocessoren voor parallelle verwerking. Deze worden gebruikt voor *computer vision* en edge AI-toepassingen die worden gebruikt in omgevingen zoals de detailhandel, beveiliging en industriële automatisering. In workstations worden ook FPGA's (Field Programmable Gate Arrays) gebruikt, bijvoorbeeld voor financiële applicaties.
- **Optimaliseringssoftware:** voorbeelden zijn OneAPI, het op standaarden gebaseerde programmeermodel van Intel, om de ontwikkeling en implementatie van datacentrische workloads op CPU's, GPU's, FPGA's en andere versnellers te vereenvoudigen. Een ander voorbeeld is CUDA, het platform voor parallel computing en de API (Application Programming Interface) van NVIDIA, voor het uitvoeren van algemene workloads op GPU's.

CPU's versus GPU's voor AI

Workstations kunnen worden gebruikt in verschillende stadia van AI-ontwikkeling en ze zijn meestal uitgerust voor diverse toepassingen. Hoewel vooral naar GPU's wordt gekeken als het gaat om parallelle verwerking, spelen ook CPU's een cruciale rol bij het ontwikkelen van een AI-model op een workstation. Net als GPU's kunnen CPU's ook worden gebruikt voor het manipuleren van data, en natuurlijk ook voor het ontwikkelen van traditionele ML-modellen. CPU's worden ook gebruikt voor dataverkenning, het proces waarbij visuele weergaven van een dataset worden gebruikt om de kenmerken van de data in kaart te brengen.

Bij DL-training wordt de rol van de host-CPU's iets kleiner omdat de GPU's het overnemen tijdens het eigenlijke trainingsproces, maar zelfs dan blijven de CPU's dienen als verwerkingslaag voor kritische software zoals het besturingssysteem of CUDA en voor het orkestreren van processen tussen de GPU's of met andere processoren. Bovendien zijn CPU's steeds meer een nieuwe rol gaan spelen als engine voor AI-deducering in gevallen waarin een workstation wordt gebruikt om een AI-model in productie uit te voeren. IDC verwacht dat in 2024 de uitgaven aan infrastructuur voor AI-deducering hoger zullen zijn dan de uitgaven aan AI-infrastructuur voor AI-training en dat een aanzienlijk deel (39%) van deze deducering zal plaatsvinden op de CPU's van de host.

Workstations versus servers: een symbiotische relatie

Voor de meeste organisaties is pragmatisme de vuistregel wanneer een workstation, een server op locatie, een cloud-instantie of een combinatie van deze drie wordt gebruikt voor AI-ontwikkeling. Er is een symbiotische relatie tussen workstations, servers en cloud-instanties voor de verschillende ontwikkelingsstadia van een AI-project.

Het voordeel van workstations in vergelijking met servers in een datacenter is dat datawetenschappers overal kunnen werken. Dit was een belangrijke factor tijdens de pandemie en behoort inmiddels tot de normale werkomstandigheden. Het biedt hen de mogelijkheid om vrij te experimenteren met hun AI-modellen en zo vaak te itereren als ze nodig achten, omdat met de kracht van moderne workstations, met krachtige GPU's, het iteratieve proces interactiever kan zijn en directe feedback en resultaten oplevert, zonder dat ze toegang hoeven te vragen tot servers of tegen andere beperkingen van datacenters aanlopen. En workstations bieden hen de flexibiliteit om de computer dichter bij de data te brengen in plaats van andersom, wat bandbreedte bespaart, netwerkcongestie vermindert en de doorvoersnelheid verhoogt. Bovendien kunnen workstations worden geconfigureerd voor verschillende behoeften: traditionele ML-taken, bijvoorbeeld, of werk dat DL-intensiever is.

Ook al is er een aanzienlijke groei in de markt voor versnelde servers, zijn deze servers nog steeds niet algemeen beschikbaar in de datacenters van bedrijven. Op het moment dat deze whitepaper werd geschreven, bestond gemiddeld 4% van de servers in de datacenters van bedrijven uit versnelde servers, wat betekent dat veel organisaties niet de middelen hebben om AI te ontwikkelen of uit te voeren op direct beschikbare on-premise GPU's. Ook om deze reden zijn versnelde workstations een nuttig alternatief voor AI-ontwikkeling.

Sterk versnelde workstations zijn nu krachtig genoeg om DL-training uit te voeren zolang het AI-model niet te groot is, waardoor training op servers niet meer nodig is. En modellen die zijn getraind op workstations met GPU's kunnen worden ingezet op workstations of op servers zonder GPU's, waarbij gebruik wordt gemaakt van de deduceringsvoorzieningen in de CPU's. Softwaretechnologieën zoals DL Boost en oneAPI van Intel kunnen AI-deducering op de CPU ondersteunen, waardoor AI-toepassingen kunnen worden ondersteund door niet-versnelde servers die reeds in datacenters zijn geïmplementeerd.

Workstations in vergelijking met de cloud

Cloudcomputing heeft een revolutie teweeggebracht in de manier waarop organisaties denken over infrastructuur, data en applicaties. Vanwege de bijna onbeperkte schaalbaarheid stelt de cloud ontwikkelaars in staat om resources op aanvraag beschikbaar te stellen, waardoor er minder beperkingen zijn die de versnelling van het innovatietempo in de weg staan. Op het eerste gezicht lijkt werken in de cloud het perfecte scenario voor AI-ontwikkeling.

Dit is echter niet altijd het geval. Uit onderzoek van IDC blijkt zelfs dat organisaties steeds vaker bepaalde workloads van de public cloud naar de on-premise infrastructuur verplaatsen. Hierbij spelen verschillende factoren een rol:

- **Beschikbaarheid van de cloud:** iedereen die afhankelijk is van cloudservices, heeft wel eens een storing meegemaakt, of deze nu te wijten was aan problemen bij de cloudprovider zelf of aan een storing in de netwerkverbinding ergens tussen het hyperscale datacenter en de eindgebruiker. In deze situaties zijn gebruikers overgeleverd aan de serviceprovider om het probleem op te lossen, terwijl ondertussen de productiviteit stagneert.
- **Beveiliging en naleving:** in veel bedrijfstakken is de bedrijfsgovernance bepalend voor waar data mogen worden gedeeld en opgeslagen, waardoor het gebruik van cloudservices wordt beperkt. Overheidsvoorschriften zoals de Algemene Verordening Gegevensbescherming (AVG) in Nederland, de GDPR in Europa en de California Consumer Privacy Act in de VS leggen ook dwingend regels op over de datasoevereiniteit.
- **Kosten:** het komt vaak voor dat organisaties onderschatten hoe snel de kosten van cloudservices kunnen stijgen, vooral voor workloads die krachtige compute-voorzieningen en een grote storagecapaciteit vereisen. De economische aspecten van de cloud zijn gebaseerd op het meten van alle soorten resourceverbruik, inclusief de terugvoer van data naar de on-premise infrastructuur.
- **De druk van proefondervindelijk werken:** de meeste AI-initiatieven beginnen met een aanzienlijke reeks experimenten, waarbij modellen die mislukken deel uitmaken van het ontwikkelingsproces. In dit proces is er een psychologische belasting die AI-wetenschappers en -ontwikkelaars voelen wanneer de cloudkosten zich opstapelen terwijl zij nog geen uitvoerbare resultaten kunnen laten zien.

Met workstations kunnen deze beperkingen worden weggenomen en kan toch gebruik worden gemaakt van cloud-native technologieën zoals op microservices gebaseerde architecturen en API-gestuurde automatisering. Dit biedt enkele van dezelfde voordelen die we hebben gezien in de vergelijking tussen workstations met datacenterservers:

- **Overal werken:** door de afhankelijkheid van de public cloud weg te nemen, worden loskoppelingsscenario's mogelijk. Veel hoogbeveiligde omgevingen zijn afgesloten van openbare netwerken en AI-workstations kunnen op unieke wijze in deze behoefte voorzien. Bij lokale resources is er ook minder vraag naar dure netwerkconnectiviteit.
- **Locatie van de data:** de proliferatie van IoT-apparaten en andere verbonden apparatuur draagt bij aan een exponentiële groei van data op edge-locaties. In veel situaties is het zinvol als de computing-resources en een toegewezen workstation zich op dezelfde locatie bevinden. En hierdoor wordt ook tegelijk aan veel nalevingsvereisten voldaan, omdat de verplaatsing van data wordt beperkt.
- **Vrij experimenteren:** het trainen en optimaliseren van AI-modellen is een iteratief proces waarbij het vaak een kwestie van vallen en opstaan is. Ontwikkelaars moeten de vrijheid hebben om experimenten uit te voeren zonder compromissen te hoeven sluiten vanwege mogelijke extra servicekosten. Workstations bieden ook meer flexibiliteit voor aangepaste tools.

Wat dit laatste punt betreft: het is relatief eenvoudig om de prijs van een workstation te vergelijken met een cloudimplementatie, aangezien de meeste cloudserviceproviders direct kostenramingen geven voor elke configuratie die een eindgebruiker wil implementeren. De kosten van één gewone virtuele machine (VM) met één NVIDIA T4 en één instantie van 375 GiB SSD-storage die vijf dagen per week acht uur per dag wordt gebruikt, bedragen bijvoorbeeld bij een bepaalde grote cloudprovider USD 140. Verdubbel de VM's, T4's en SSD's en de kosten stijgen naar USD 365 per maand. Blijf bij twee VM's, maar verdubbel de T4's naar vier en de storage naar 4 x 375 GiB en voer een volledige trainingsrun uit in de omgeving, en de kosten stijgen naar USD 2700 per maand. De cloudkosten voor AI-ontwikkeling kunnen dus gemakkelijk oplopen tot tienduizenden dollars per jaar, aanzienlijk meer dan de jaarlijkse afschrijving van een high-end workstation.

AI-PROTOTYPING OP WORKSTATIONS

Vergeleken met on-premise servers en de cloud bieden workstations een duidelijk voordeel als het gaat om prototyping van AI-modellen. Servers in het datacenter zijn mogelijk volledig bezet of te missiekritisch voor het prototypen en testen van AI, en zoals eerder besproken kan het onbeperkt gebruiken van cloud-instanties als testomgeving al snel tot kostenoverschrijdingen leiden. Workstations zorgen ervoor dat de AI-wetenschapper of -ontwikkelaar niet hoeft te onderhandelen over servertoegang en zich niet voortdurend zorgen hoeft te maken over hoge kosten van cloudservices tijdens het prototypestadium. De lage eenmalige kosten van een workstation bieden volledige vrijheid om overal en op elk moment zonder extra kosten te kunnen prototypen.

AI-MODELLEN IMPLEMENTEREN OP WORKSTATIONS

Hoewel het ontwikkelen van AI-modellen op een workstation al jaren een veelgebruikte strategie is, ziet IDC steeds meer gebruiksscenario's voor het *implementeren* van een AI-model op een workstation, meestal aan de edge, met andere woorden: het AI-model in productie nemen op het workstation door het daar deduceringen te laten uitvoeren. De edge groeit snel als AI-implementatielocatie voor servers. Er is sprake van meer dan een verdrievoudiging van 2020 tot 2024 wat betreft de jaarlijkse hardware-uitgaven, en voor workstations is dit nauwelijks minder, naarmate eindgebruikers de voordelen ervan aan de edge ontdekken.

IDC definieert de edge als een scenario voor gedistribueerde computing waarbij infrastructuur en applicaties buiten gecentraliseerde datacenters, zowel in de cloud en on-premise, worden geïmplementeerd en zo dicht bij de plaats waar data worden gegenereerd en verbruikt als nodig. Hieronder vallen externe kantoren en filialen, maar ook branchespecifieke locaties zoals fabrieken, magazijnen, ziekenhuizen en winkels.

Data- en computing-intensieve workloads worden steeds vaker on-premise of aan de edge geïmplementeerd. Dit wordt gedaan om minder last te hebben van de beperkingen die inherent zijn aan public clouds, zoals de tijd die het kost om grote datasets te uploaden en de variabele kosten van het uitvoeren van AI-training, vooral in situaties die een aanzienlijke hoeveelheid datawetenschappelijke experimenten vereisen.

Uit onderzoek van IDC blijkt dat de edge een snel groeiend implementatiescenario voor AI is, waarbij organisaties in 2023 USD 2,9 miljard investeren in AI-computers aan de edge, met een groei naar USD 6,9 miljard in 2026 (zie *Worldwide AI Hardware Forecast, 2022-2026: Strong Market Growth for AI Compute and Storage,* IDC nr. US49671722, september 2022). Bovendien wint de edge steeds meer aan populariteit als voorkeurslocatie voor implementatie van HPC-workloads zoals engineering en technische workloads. Bedrijven investeren momenteel bijna USD 1 miljard in dergelijke workloads aan de edge, met een groei naar USD 2,4 miljard in 2027 (zie *Worldwide High-Performance Computing Server Forecast, 2023-2027: Enterprise Will Overtake HPC Labs,* IDC nr. US50525123, april 2023). Dit zijn gebieden waar het zinvol is om een AI-workstation te implementeren.

Wanneer een AI-model wordt geïmplementeerd op een workstation aan de edge, is er niet altijd behoefte aan high-end GPU's zoals het geval is bij AI-ontwikkeling. Lichtere GPU's kunnen ook de AI-deducties uitvoeren en in veel gevallen zijn GPU's helemaal niet nodig. In deze gevallen kunnen CPU's de deductietaak adequaat uitvoeren, vooral als ze worden gebruikt in combinatie met optimalisaties zoals DL Boost van Intel. DL Boost is een set instructiefuncties op Intel-microprocessoren die is ontworpen om AI-workloads, waaronder AI-deducties, te versnellen. Volgens Intel levert Intel DL Boost met de 4e generatie Intel Xeon schaalbare processor, die Intel DL Boost ondersteunt, een 1,45 keer hogere INT8 real-time deductiedoorvoer op dan de vorige generatie (BERT-Large SQuAD). Dit helpt ook om een workstation geschikter te maken voor gebruik aan de edge, waar zaken zoals vermogen, mobiliteit en temperatuurbeheer kleinere wattages vereisen. Movidius Myriad (M2) van Intel past goed in deze vermogenssituatie, omdat de coprocessor slechts 12 W vraagt.

Gebruiksscenario's voor het implementeren van AI op workstations

Er zijn verschillende situaties die zich van nature goed lenen voor het implementeren van AI op lokaal geïmplementeerde workstations. Gemeenschappelijke kenmerken zijn grote hoeveelheden machinaal gegenereerde tijdreeksdata en niet-gestructureerde data zoals videostreams en beelden. Er zijn ook gevallen waarin de interpretatie van materiedeskundigen noodzakelijk is als aanvulling op AI-modellen.

Voorbeelden hiervan zijn:

- **AIOps:** naarmate IT-systemen groter en complexer worden, groeit de behoefte om van reactief beheer van incidenten over te stappen op proactieve bewaking. Dit geldt vooral als infrastructuur en applicaties worden gedistribueerd naar locaties aan de edge van de organisatie waar weinig tot geen technisch personeel aanwezig is. Door een uitgangssituatie met normale prestaties te modelleren, is het mogelijk om afwijkingen te identificeren en herstelstappen te automatiseren.

- **Respons bij noodgevallen:** in een noodsituatie moeten eerstehulpverleners snel een situatie beoordelen, kritische apparatuur opsporen en middelen inzetten om hen te helpen die als eerste hulp behoeven. Dit moet vaak gebeuren in een omgeving zonder netwerkverbinding, waardoor een lokaal workstation nodig is dat datafeeds kan verzamelen, informatie kan deduceren uit AI-modellen en de communicatie met belangrijke medewerkers kan automatiseren.
- **Radiologie:** de vooruitgang in beeldvormingstechnologie heeft geleid tot een toename in de omvang van data die door een enkele scan worden gegenereerd. Daarom is het belangrijk dat deze data ter plaatse zijn om tijdig te kunnen worden geanalyseerd. AI-modellen die zijn getraind op basis van miljoenen bestaande voorbeelden kunnen patronen nauwkeuriger identificeren dan het menselijk oog, waardoor de nauwkeurigheid toeneemt.
- **Opsporing van gas- en olievelden:** upstream olie- en gasbedrijven gebruiken een combinatie van telemetrische, seismische en beelddata om natuurlijke bronnen te lokaliseren, boorlocaties te selecteren en de prestaties van apparatuur in het productieproces te optimaliseren. Dit vereist vaak analyse van informatie in gebieden waar alleen kostbare satellietcommunicatie beschikbaar is.
- **Kankeronderzoek en de ontwikkeling van geneesmiddelen:** onderzoekers in onderzoeksziekenhuizen en universiteiten gebruiken AI en natuurlijke taalverwerking om oncologen te helpen bij het bepalen van de meest effectieve, geïndividualiseerde kankerbehandeling voor hun patiënten. Ze combineren ook machine learning met computer vision om radiologen een beter inzicht te geven in hoe de tumoren van patiënten zich ontwikkelen. En ze gebruiken algoritmes om beter te begrijpen hoe kankersoorten zich ontwikkelen en welke behandelingen het beste werken om ze te bestrijden.
- **Beoordeling van verzekeringsclaims:** handmatige verwerking van claims is arbeidsintensief en vatbaar voor menselijke fouten. Als de geldigheid van claims door AI kan worden beoordeeld, kunnen de kosten worden verlaagd doordat verzekeringsagenten zich kunnen richten op zaken die meer onderzoek vereisen. Dit verhoogt de algehele doorvoersnelheid van de activiteit zonder dat dit ten koste gaat van de nauwkeurigheid.
- **Telegezondheid:** AI verbetert het herstelpercentage van patiënten door individuele behandelplannen op maat te maken op basis van real-time informatie van vitale functies uit draagbare medische hulpmiddelen. Deze informatie wordt gecombineerd met de medische geschiedenis van patiënten en een kennisbank met vergelijkbare gevallen. Dit is vooral belangrijk in landelijke gebieden die meer afhankelijk zijn van telegezondheidszorg.
- **Winkelbeveiliging (anti-diefstal):** er wordt gebruik gemaakt van real-time analyse van videostreams om menselijk gedrag dat kan leiden tot criminele activiteiten te voorspellen en herkennen. Hiervoor moeten doorgaans meerdere videofeeds worden samengevoegd om de bewegingen van een persoon in een winkel te volgen. Omdat het zaak is om een daadwerkelijke gebeurtenis snel te identificeren, is dit een proces dat het beste lokaal kan worden uitgevoerd.
- **Verkeersleiding:** overheidsinstanties die verantwoordelijk zijn voor transportactiviteiten maken steeds vaker gebruik van AI voor de coördinatie van verkeerslichten en digitale bewegwijzering om de doorstroming van voertuigen te verbeteren en de veiligheid van burgers te garanderen. Hiervoor is een combinatie van verschillende soorten invoer nodig, waaronder videocamera's en telemetrie van wegsensoren, om verkeerspatronen te optimaliseren.
- **Bewaking van productielocaties:** voor een fabrieksmanager is het van het grootste belang om de uptime van kritische processen te kunnen garanderen en zonder vertraging de productieschema's uit te voeren. Dit vertaalt zich in voorspellend onderhoud van belangrijke apparatuur, automatische detectie van defecten en optimalisatie zowel binnen als buiten de leveringsketen voor de locatie. Dit is een gebied waar AI menselijke operators kan helpen om de prestaties te verhogen en tegelijkertijd de veiligheidsnormen te handhaven.

- **Drones:** geautomatiseerde analyse van beelden die zijn vastgelegd door drones maakt het mogelijk om een breed scala aan situaties te monitoren op een schaal die voorheen niet mogelijk was. Dit heeft een grote invloed op de inspectie van de infrastructuur van gas- en elektriciteitsnetwerken, verzekeringsonderzoeken, zoek- en reddingsacties, precisielandbouw en het onderhoud van visgronden en natuurreservaten.
- **Dagelijkse kantooromgevingen:** dagelijkse kantooromgevingen worden steeds meer verbeterd met op AI gebaseerde productiviteitstools zoals Microsoft Copilot.
- **Hernieuwbare energie:** locaties voor het opwekken van hernieuwbare energie, zoals windmolenparken, waterkrachtcentrales en zonneparken, vereisen real-time bewaking, onderhoud en dataverzameling die lokaal moeten worden gegenereerd en geanalyseerd.

DELL WORKSTATIONS VOOR AI

Dell biedt een breed scala aan workstations voor verschillende niveaus van AI-ontwikkeling en/of -implementatie, allemaal onder de noemer Data Science Workstation (DSW). In deze paragraaf worden kort de specificaties beschreven en worden vervolgens talrijke AI-persona's/applicaties besproken, zoals datawetenschappers en de voordelen van de DSW-technologie van Dell. Deze AI-ready Data Science Workstations zijn speciaal ontworpen voor datawetenschappers. De nieuwste Precision Data Science Workstations maken gebruik van AI-functionaliteit om de apparaten nauwkeurig af te stellen voor optimale prestaties van de applicaties die datawetenschappers het meest gebruiken. Hierdoor kunnen ze hun belangrijkste werk sneller afronden. Bovendien worden Dell Precision workstations getest en gecertificeerd door onafhankelijke ISV's om te garanderen dat ze de krachtige applicaties ondersteunen die de klanten van Dell nodig hebben om hun dagelijkse taken uit te voeren.

Waarin de workstations van Dell zich onderscheiden

Dell Precision workstations met NVIDIA RTX GPU's zijn ontworpen om sterke schaalbaarheid en prestaties te bieden voor de analytics- en AI-initiatieven van een organisatie. Dell Technologies levert uitgebreide hardwareoplossingen die zijn geoptimaliseerd voor de nieuwste AI-software:

- **Robuuste hardwareconfiguratie:** Dell Precision workstations bieden een reeks krachtige hardwareconfiguraties, waaronder multi-core processoren, RAM met hoge capaciteit en meerdere GPU-opties. Deze componenten leveren de benodigde rekenkracht voor AI-taken, waardoor efficiënte training en deductie mogelijk zijn.
- **Schaalbaarheid en aanpasbaarheid:** Dell Precision workstations zijn schaalbaar en aanpasbaar, zodat gebruikers de hardwareconfiguratie kunnen aanpassen aan hun specifieke AI-vereisten. Deze flexibiliteit zorgt ervoor dat het workstation kan worden geoptimaliseerd voor de specifieke behoeften van AI-workloads.
- **Certificering en optimalisering:** Dell werkt samen met NVIDIA om Precision workstations te certificeren voor compatibiliteit en uitstekende prestaties met NVIDIA RTX GPU's, waaronder NVIDIA RTX 6000 Ada Generation-kaarten. Deze certificering garandeert naadloze integratie en geoptimaliseerde prestaties bij het gebruik van Dell Precision workstations met NVIDIA RTX GPU's voor AI-taken.
- **Krachtige verwerkingscapaciteit:** Dell Precision workstations die zijn uitgerust met Intel-processoren leveren de rekenkracht die nodig is voor AI-taken. Met multi-core processoren en hoge klokfrequenties leveren deze workstations de prestaties die nodig zijn voor training en deducering in AI-workflows.

- **Support voor software en tools:** op Dell Precision workstations zijn vooraf software en tools geladen die de ontwikkeling en implementatie van AI ondersteunen. Dit omvat geoptimaliseerde softwarestacks, AI-frameworks en bibliotheken die gebruikmaken van NVIDIA RTX GPU's, waardoor gebruikers gemakkelijker aan de slag kunnen met AI-projecten.

Daarnaast zijn er andere belangrijke gebieden waarop de Dell workstations zich onderscheiden. De technologieën die hiervan deel uitmaken, worden in de volgende paragrafen besproken.

Reliable Memory Technology

Dell biedt een technologie, Reliable Memory Technology Pro (RMT Pro) geheten, die bovenop ECC werkt en is ontworpen om uptime te helpen maximaliseren. Deze technologie werkt samen met ECC-geheugen om geheugenfouten in realtime op te sporen en te corrigeren. Volgens Dell elimineert RMT Pro geheugenfouten vrijwel volledig door te voorkomen dat slechte geheugensectoren opnieuw wordt gebruikt, zelfs als de DIMM volledig in gebruik blijft. Na een herstart van het systeem isoleert RMT Pro het defecte geheugengebied en verbergt het voor het besturingssysteem. Omdat slechte geheugengebieden niet langer kunnen worden geadresseerd, hebben AI-datawetenschappers en -ontwikkelaars ook geen last meer van het probleem van aanhoudende crashes. Dit geeft de productiviteit een belangrijke boost.

Dell Optimizer for Precision

Dell voorziet de meeste van zijn workstations ook van Dell Optimizer for Precision (DOP), dat automatisch systeeminstellingen aanpast zodat het workstation verschillende veelgebruikte zakelijke applicaties op de hoogst mogelijke snelheid uitvoert. Dit verbetert de productiviteit van een datawetenschapper of -ontwikkelaar. De tool maakt ook real-time prestatierapporten voor IT over gebruik van processoren, storage, geheugen en grafische kaarten. DOP draait nog niet op Linux en is daarom vooral nuttig voor het implementeren van AI, omdat het ontwikkelen van AI meestal gebeurt met op Linux gebaseerde open source software. Dell Optimizer for Precision biedt ook ExpressSign-in, Express Charge (op mobiele apparaten), Intelligent Audio en rapportage- en analysetools om het workstation nauwkeurig in te richten.

UITDAGINGEN EN KANSEN

Voor bedrijven

IDC ziet een tweedeling in de markt voor AI. Aan de ene kant zetten bedrijven datastrategieën in om concurrerend te blijven, waaronder de grootschalige toepassing van AI. Bij wijze van voorbeeld krijgen ze collega's te zien die buitengewoon werk hebben verricht met behulp van AI-infrastructuurproducten voor bedrijven die daadwerkelijk in de top 100 van supercomputers worden vermeld. Aan de andere kant bestaat de dagelijkse realiteit van bedrijven uit kleine AI-initiatieven die worden uitgeprobeerd op beschikbare servers in het datacenter of in de cloud, vaak met onvoldoende budget en ondermaats presterende hardware.

Voor veel bedrijven is het eerste scenario niet relevant en het tweede maar al te realistisch. Zij staan voor de uitdaging om hun AI-datawetenschappers en/of -ontwikkelaars de juiste tools te geven om tijdig AI-trainingen te kunnen uitvoeren zonder enorme bedragen uit te geven aan cloud-instanties of GPU-versnelde datacenterservers. IDC is van mening dat deze bedrijven er goed aan doen om hun wetenschappers en ontwikkelaars te voorzien van krachtige GPU-versnelde workstations.

Voor Dell

Er heerst een misverstand in de markt dat AI-ontwikkeling en -implementatie dure, versnelde serverhardware vereist, vaak zelfs in een cluster. Dit is misschien waar voor hele grote AI-algoritmen, met miljarden parameters, maar de meeste bedrijven ontwikkelen niet zulke enorme algoritmen. Ze doen iets met hun AI-initiatief dat nuttig, invloedrijk en beheersbaar is en veel bedrijven realiseren zich niet dat dergelijke AI-modellen op algemene schaal kunnen worden ontwikkeld, en geïmplementeerd, op workstations. De uitdaging voor Dell is om door het vooroordeel heen te breken en de markt te informeren over de mogelijkheden die de workstations van het bedrijf bieden.

Tegelijkertijd moet Dell ervoor zorgen dat zijn workstations de verwachtingen waarmaken en na verloop van tijd geen technologische bottlenecks worden. Dit betekent snelle voortdurende innovatie om eindgebruikers die de workstations op de juiste manier gebruiken (met andere woorden, die niet proberen een algoritme met miljarden parameters uit te voeren) nooit teleur te stellen. Het betekent ook dat er voor klanten die plotseling zeer snel beginnen op te schalen of wiens algoritmen inderdaad heel groot worden, een naadloze overgang is van het workstation naar de AI-serverlijn van Dell. Daar ligt natuurlijk ook de kans voor Dell: om de juiste oplossing te hebben voor elke klant, ongeacht de omvang van hun AI-initiatief.

CONCLUSIE

IDC is van mening dat workstations momenteel voor veel gebruiksscenario's worden ondergewaardeerd als de werkpaarden van de ontwikkeling en implementatie van AI. Ze bieden AI-wetenschappers en -ontwikkelaars een krachtig GPU-versneld platform met een lagere capex (kapitaaluitgave) dan servers, een veel lagere opex (operationele kosten) dan cloud-instanties en een veel grotere vrijheid om te experimenteren met AI-modellen. Bedrijven die AI-initiatieven ontwikkelen waarvoor geen algoritmes met miljarden parameters nodig zijn (zoals in de meeste gevallen), zouden moeten overwegen om hun AI-teams uit te rusten met workstations voor onbeperkte AI-ontwikkeling en voor eenvoudige implementatie aan de edge.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com