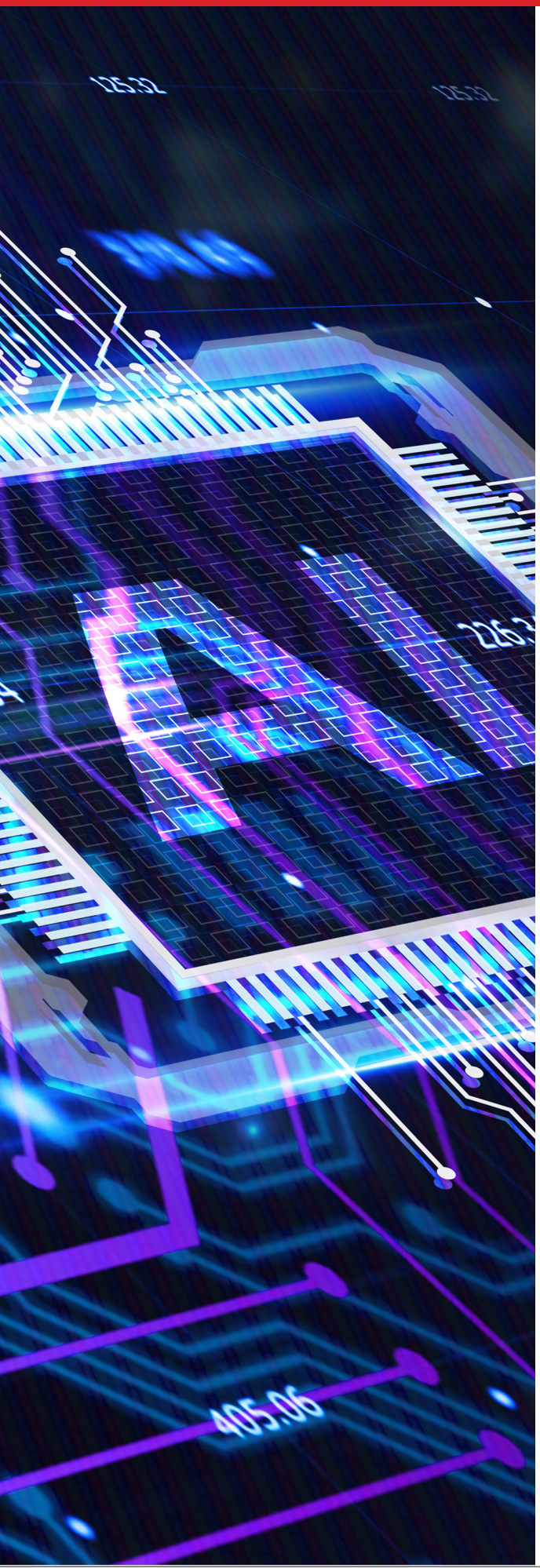


Voldoen aan de uitdagingen van AI-workloads met het Ai-portfolio van Dell

Een vergelijking van het AI-portfolio van Dell met vergelijkbare aanbiedingen van HPE

Er is geen twijfel dat kunstmatige intelligentie (AI) het bedrijfslandschap heeft getransformeerd en branches en organisaties van elke omvang in staat heeft gesteld om diepere inzichten te krijgen in hun data, bedrijfsprocessen te automatiseren, op maat gemaakte klant- en gebruikerservaringen te leveren en beter te concurreren in hun branche. Om de kracht van AI effectief te benutten, hebben organisaties een infrastructuurprovider nodig die een uitgebreid en geïntegreerd oplossingenportfolio kan bieden dat de hele AI-levenscyclus omvat.

Om klanten te helpen aan de groeiende eisen van AI te voldoen en de inherente complexiteit ervan te overwinnen, zijn er infrastructuurleveranciers zoals Dell Technologies en Hewlett Packard Enterprise (HPE). Deze leveranciers bieden AI-ready portfolio's en verschillende niveaus van AI-oplossingen die krachtige infrastructuuroplossingen on-premise en in de cloud combineren met strategische partnerschappen en een reeks support- en consultancyservices.

Dit rapport onderzoekt openbaar beschikbare informatie over de AI-portfolio's van Dell en HPE* met als doel specifieke architecturale, prestatie- en supportvoordelen te belichten waar klanten van kunnen profiteren door Dell Technologies te selecteren voor hun AI-behoeften. We vergelijken details van de servers die Dell heeft gebouwd om AI-implementaties te ondersteunen en verwijzen naar de resultaten van tests op branchestandaarden van ML Commons®. We verkennen ook aanvullende software- en serviceaanbiedingen die klanten ondersteunen in elke fase van hun AI-traject.

*Opmerking: PT heeft al het onderzoek op of vóór 5 december 2023 voltooid, dus dit artikel bevat geen aanbiedingen of wijzigingen van Dell of HPE-releases na die datum.

De uitdagingen van het implementeren van AI

Het implementeren van een AI-strategie brengt veel nieuwe uitdagingen met zich mee voor datacenters en het IT-personeel dat daar werkt, waaronder:

- De bestaande hiaten in vaardigheden van hun huidige personeel aanpakken door middel van interne AI-training of externe werving.
- Inzicht in de behoeften van AI voor datavoorbereiding, inclusief de kwaliteit, kwantiteit, locatie en huidige status van de data van het bedrijf.
- Het beoordelen van de specifieke zakelijke AI-doelen om beter te bepalen welke AI-modellen en -implementaties voordelen bieden.
- Het beoordelen van de computing-, netwerk- en storagebehoeften van de geplande AI-systemen en het bepalen van een acquisitieplan.

Dit zijn slechts een paar voorbeelden van de vele, vaak aanzienlijke obstakels waarmee een bedrijf wordt geconfronteerd wanneer het de voordelen van AI in hun datacenters wil benutten.

Met het AI-portfolio van Dell willen we klanten helpen deze uitdagingen aan te gaan door middel van professionele en adviserende services die klanten helpen bij het maken van een implementatieroadmaps en het voorbereiden van hun data op AI-modellen.[1] Het portfolio bevat ook trainingscursussen over machine learning-concepten (ML) en andere educatieve onderwerpen en biedt gevalideerde ontwerpen voor AI om te zorgen dat de implementatie een succes is.[2] Daarnaast werkt Dell samen met derden om extra AI-tools aan klanten te bieden, zoals een aangepaste Dell portal binnen de Hugging Face-community met speciale containers en scrips voor de implementatie van open source AI-modellen[3] en eenvoudige implementatie van het Meta Llama 2 Large Language Model (LLM).[4] Naast een grote selectie aan computing- en pc-aanbiedingen, van mobiele workstations tot servers die tot wel 8 hoogwaardige NVIDIA GPU's ondersteunen, biedt Dell ook de niet-gestructureerde datastorage-AI die vereist is bij een portfolio aan krachtige bestands- en objectstorage-arrays. Deze storageaanbiedingen, waaronder Dell PowerScale, ObjectScale, ECS en storage op de kaart, kunnen de niet-gestructureerde data verwerken die AI-workloads vaak gebruiken.[5] Dell werkt ook samen met Snowflake om Dell klanten een hybrid cloud-storageoplossing te bieden.[6] Volgens een analyse van Dell biedt Dell per augustus 2023 het 'breedste generatieve AI-portfolio' dat verder gaat dan alleen servers en storage, door resources te bieden tijdens het AI-implementatietraject.[7]

Prestaties van AI en versnelde computingopties: Dell versus HPE

AI-workloads kunnen CPU's, GPU's of beide gebruiken als computingresources, afhankelijk van de grootte of het type workload. Sommige CPU's bieden AI-specifieke versnellers, zoals Intel Advanced Matrix Extensions (Intel AMX) in de nieuwste Intel Xeon schaalbare processors.[8] GPU's zijn vaak beter voor grotere en/of veeleisendere workloads, maar de GPU-vormfactor kan de prestatieniveaus beïnvloeden. Sommige GPU's van het NVIDIA A100- en H100-model worden bijvoorbeeld geleverd in universele PCIe- of bedrijfseigen SXM-vormfactower, waarvan de laatste de krachtigere NVIDIA SXM-architectuur gebruiken.[9] Grote geheugencapaciteit en serverontwerpfuncties, zoals koelingsarchitectuur en energie-efficiëntie, hebben ook invloed op de prestaties. De meeste datacenters gebruiken nog steeds luchtkoeling, wat betekent dat high-performance computing-workloads (HPC) servers nodig hebben die zo effectief mogelijk met lucht kunnen koelen. Hieronder belichten we het aanbod van PowerEdge servers op het gebied van componenten, koelingsopties en meer, samen met hun gepubliceerde MLCommons® MLPerf®-scores.

Benchmarkprestaties van AI-modellen: Vergelijking van MLPerf-resultaten

MLPerf® is een benchmarksuite die AI-prestaties test voor zowel training als inferentie. Als een organisatie officiële MLPerf®-resultaten wil publiceren, moeten de resultaten voldoen aan de specifieke voorwaarden die zijn vastgesteld door de benchmarkontwikkelaar, MLCommons®.[10] Deze nalevingsrichtlijnen bieden standaarden die het makkelijker maken om prestaties te vergelijken. Voor inferentietests gebruikt MLPerf® de datasets Datacenter, Edge, Mobile en Tiny, en worden AI-scores en watt aan energie gepubliceerd die tijdens het testen zijn verbruikt. De benchmarksuite voor inferentie bevat tests voor veel gangbare AI-, ML- en DL-modellen; zie tabel 1.

Tabel 1: AI-, ML- en DL-modellen inbegrepen in MLPerf®-tests en typische gebruiksscenario's voor elk model. Bron: Principled Technologies.

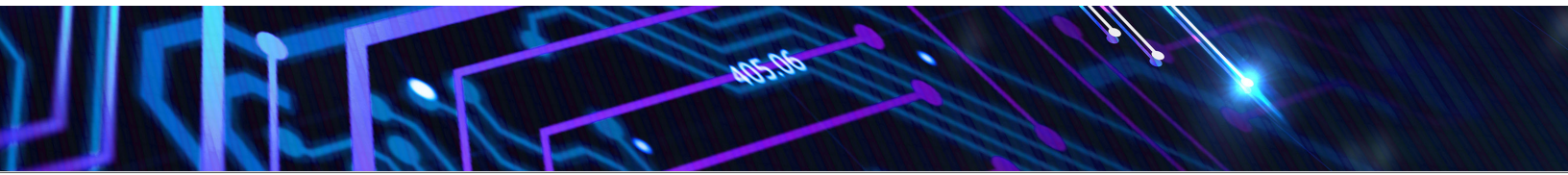
Gangbare AI-modellen	Typische gebruiksscenario's
ResNet	Een beeldclassificatiemodel waarmee computers verschillende beelden kunnen leren, onthouden en identificeren voor gebruiksscenario's zoals medische beeldvorming, contentmoderatie op sociale media en gezichtsherkenning
RetinaNet	Een type objectdetectie dat extra complexiteit kan verwerken vergeleken met ResNet. Het helpt computers om objecten in afbeeldingen of videoframes te identificeren en te lokaliseren en kan ze classificeren op belangrijkheid. Gebruikt voor zaken zoals autonoom rijden, auto-assist-technologie in voertuigen, bewaking, gezichtsherkenning
3D-UNet	Specifiek voor segmentatie van medische beelden
RNN-T	Spraakherkenning voor gebruiksscenario's zoals geautomatiseerde vertalingen
BERT	Natuurlijke taalverwerking voor gebruiksscenario's zoals tekstsamenvattingen, vertalingen en automatische voltooiing van taken
DLRM-v2-99.9	Aanbevelingsmodel voor gebruiksscenario's zoals gerichte advertenties en gepersonaliseerde productaanbevelingen
GPTJ-99 en 99.9	LLM voor natuurlijke taalverwerking dat uitblinkt bij het genereren van tekst voor gebruiksscenario's zoals chatbots en chatgebaseerde AI-tools

MLPerf

MLPerf®-resultaten bevatten verschillende parameters naast de AI-modellen zelf, wat ertoe kan leiden dat er veel data in één grafiek of tabel worden geparseerd. Hier volgt een beknopt overzicht van deze parameters:

- 99.0 en 99.9: Deze cijfers verwijzen naar de nauwkeurigheid waarmee het model is getraind. Hoe nauwkeuriger de uitvoer moet zijn, hoe complexer het model en hoe langer het kan duren om data te verwerken.
- Offline samples/sec: Modus waarbij de benchmark alle query's aan het begin van de test verzendt waarmee data worden simuleert die al in het systeem aanwezig zijn.
- Server queries/sec: Modus waarbij de benchmark query's verzendt tijdens de hele duur van de test, waarbij de analyse van een live datastroom wordt gesimuleerd.

Zie https://mlcommons.org/benchmarks/inference-datacenter/ voor meer informatie over de resultaten van MLCommons® en MLPerf®.



De resultaten in dit rapport zijn afkomstig van MLPerf® v3.1 Inference Datacenter-resultaten die in november 2023 op de MLCommons®-website zijn gepubliceerd.[11] Deze resultaten bevatten inzendingen van technologiefabrikanten en cloudserviceproviders en gaan over verschillende configuraties. In vergelijking met openbaar beschikbare inzendingen van HPE, leverden de servers van Dell betere resultaten op in bepaalde AI-modellen. (Opmerking: Verschillende GPU-configuraties op de servers kunnen het moeilijk maken om 1-op-1 vergelijkingen te maken.) Zie Tabel 2 voor details.

Tabel 2: Dell en HPE-servers die zijn meegenomen in de resultaten van MLCommons® MLPerf® 3,1, gepubliceerd op 29-11-2023. Bron: Principled Technologies.

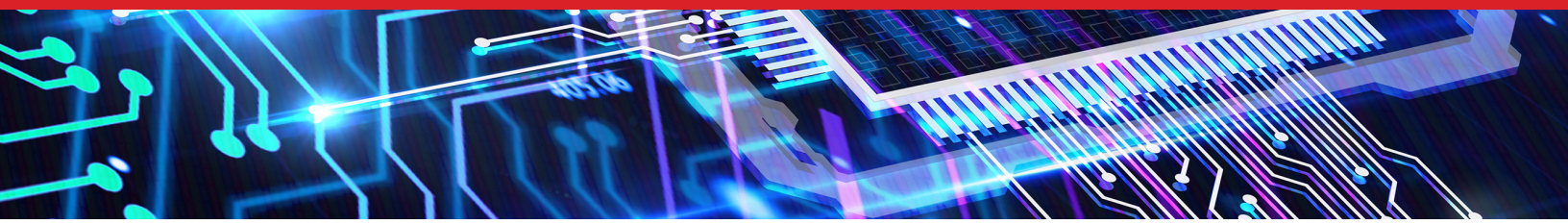
Indiener	Servermodel	aantal en model van GPU's	Omschrijving
Dell[12]	PowerEdge XE9680	8 x NVIDIA H100 SXM	Voor AI-training en -inferentie met grote workloads, zoals grote taalmodellen
	PowerEdge XE9640	4 x NVIDIA H100 SXM	Voor het trainen van grote AI-modellen in datacenters met een hoge dichtheid en vloeistofkoeling
	PowerEdge XE8640	4 x NVIDIA H100 SXM	Voor traditionele AI-training, HPC en data analyse-apps in een 4U-vormfactor voor luchtgekoelde datacenters
	PowerEdge R760xa	4 x NVIDIA H100 PCIe	Voor een breed scala aan workloads met hoge computingcapaciteit, waaronder AI-ML/DL-training en inferentie waarvoor geen toppresterende GPU's nodig zijn
HPE[13,14]	ProLiant XL675d Gen10 Plus	8 x NVIDIA A100 SXM	Voor high-performance computing en AI
	ProLiant DL380a Gen11	4 x NVIDIA H100 PCIe	2U-server voor gemiddelde AI-workloads

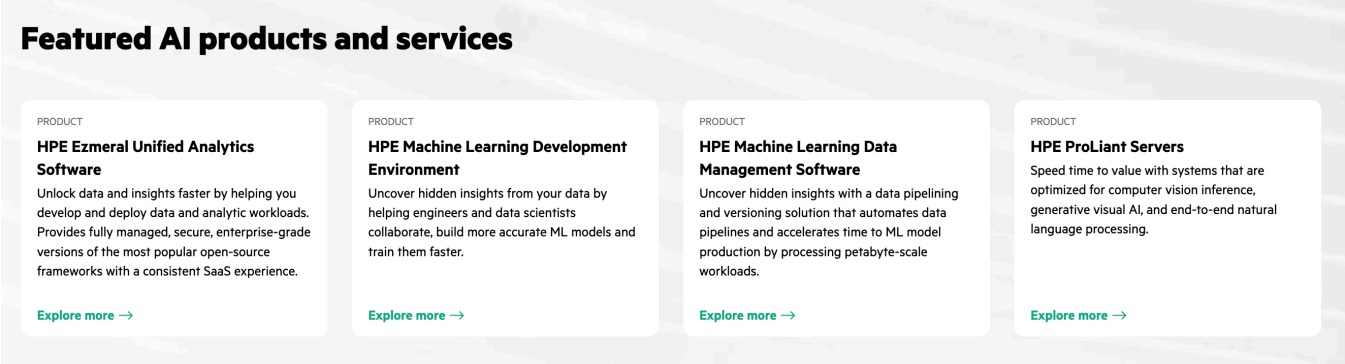
Directe vergelijking van Dell en HPE servers

Hoewel er voor een uitgebreide AI-strategie meer nodig is dan alleen hardware, is het garanderen van de krachtigste hardwareprestaties een van de meest essentiële factower voor het succes van AI-workloads. Naarmate nieuwere GPU's en andere technologieën beschikbaar komen, groeien ook de mogelijkheden voor AI-workloads. Toen de MLPerf® v3.1-resultaten voor het eerst werden gepubliceerd, was de beste NVIDIA-GPU die er beschikbaar was de H100 Tensor Core, waarmee Dell MLPerf®-resultaten voor verschillende van hun servers in zowel PCIe- als SXM5-vormfactower heeft gepubliceerd.[15] De gepubliceerde HPE-resultaten bevatten slechts één inzending met H100 en met alleen de PCIe-vormfactor. Uit ons onderzoek bleek dat slechts een paar van de beschikbare HPE-servers die geschikt zijn voor GPU de SXM5 H100-vormfactor ondersteunen voor de beste NVIDIA GPU-prestaties en dat geen van de HPE ProLiant-servers dat doen.[16] Zoals we hieronder laten zien, zorgen betere GPU's doorgaans voor betere prestaties van AI-workloads.

MLPerf-resultaten met acht GPU's

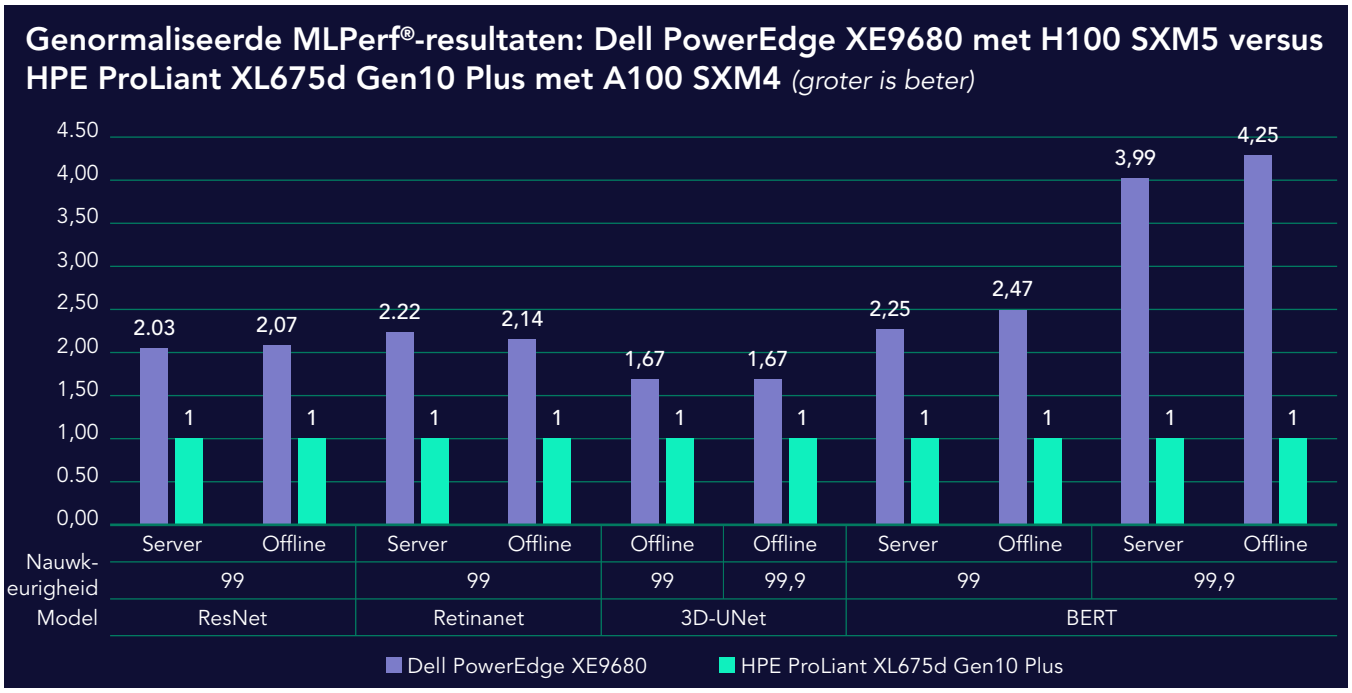
De Dell PowerEdge XE9680 biedt ondersteuning voor maximaal acht NVIDIA H100 SXM5-GPU's voor AI-versnelling en maximaal twee 4e generatie Intel® Xeon® schaalbare processors. De PowerEdge XE productlijn heeft een modulaire architectuur die SXM4 of SXM5 NVIDIA-GPU's of Open Compute Project Accelerator module (OAM) GPU-samenstellingen van AMD ondersteunt, waarmee de prestaties kunnen verbeteren in vergelijking met een standaard PCIe-GPU.[17] De PowerEdge XE9680 neemt slechts 6U aan rackruimte in en is een compacte acht-richtings NVIDIA H100 SXM5-server. De nieuwste HPE ProLiant Gen11-servers ondersteunen momenteel niet de H100 SXM-vormfactor,[18] hoewel sommige van de HPE Cray Supercomputing-servers dat wel doen.[19] Omdat HPE geen MLPerf®-resultaten heeft ingediend voor de Cray-servers en alleen hun ProLiant-servers op hun AI-portfoliopagina uitlicht, richten we ons voor dit artikel op ProLiant-servers. (Zie afbeelding 1.)





Afbeelding 1: Schermafbeelding van Uitgelichte AI-producten en -services op https://www.hpe.com/us/en/solutions/ai-artificial-intelligence.html met HPE ProLiant-servers per 05-12-2023.

In de MLPerf® v3.1-resultaten die voor het eerst werden gepubliceerd in november 2023 voor servers met acht GPU's presteerde de Dell PowerEdge XE9680 met NVIDIA SXM5 H100-GPU's tot 4,25x beter dan de HPE ProLiant XL675d Gen10 Plus met NVIDIA SXM4 A100-GPU's (zie afbeelding 2).



Afbeelding 2: Gepubliceerde MLPerf®-resultaten voor de Dell PowerEdge XE9680 en HPE ProLiant XL675d Gen10 Plus per 29-11-2023. Het Dell systeem gebruikt de NVIDIA H100-GPU, terwijl de GPU's in het HPE-systeem één generatie ouder zijn. Bron: Principled Technologies met gegevens van MLCommons®.[20,21]

Om makkelijker te kunnen vergelijken, hebben we de testresultaten in afbeelding 2 tot en met 5 genormaliseerd. Dit betekent dat we de waarde 1 toewijzen aan elk HPE ProLiant DL380a Gen 11-resultaat en dat we het bijbehorende resultaat van de Dell PowerEdge R760xa in relatie tot dat resultaat tonen. Zoals deze resultaten aantonen, kan zelfs één generatie verschil tussen GPU-modellen een aanzienlijk verschil maken in de prestaties die u kunt verwachten bij een groot aantal AI-workloads.

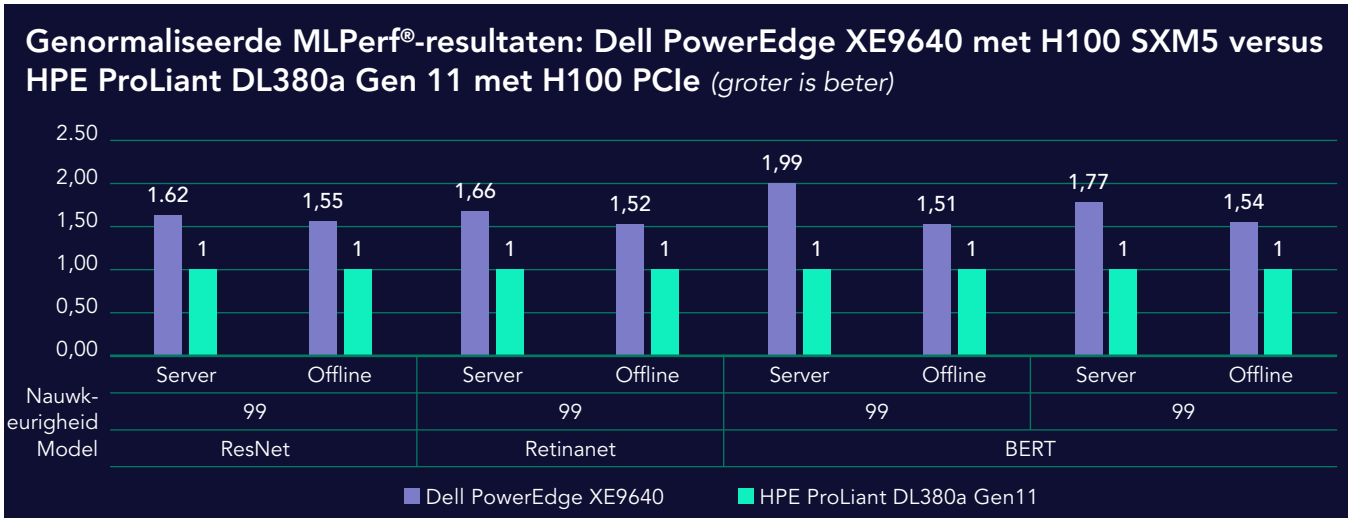
MLPerf-resultaten met vier GPU's

Wanneer het belangrijk is energie of ruimte te besparen in het datacenter, kan de 2U Dell PowerEdge XE9640 het antwoord zijn. Met tot wel vier NVIDIA H100 SXM-GPU's biedt de PowerEdgeXE9640 de helft van de GPU-rekenkracht van de XE9680 en neemt deze optie twee derde minder ruimte in.[22] De compacte Dell PowerEdge XE9640 bevat Dell Smart Cooling-technologie, die verschillende thermische technologieën biedt, waaronder directe vloeistofkoeling voor CPU's en GPU's.[23]

Het 2U-chassis van de PowerEdge XE9640 is geschikt voor verbeterde luchtstroommechanismen, waaronder grotere ventilatoren en koelplaten, om de andere essentiële componenten, zoals PCIe-kaarten en geheugen, te koelen.[24] De PowerEdge XE9640 is momenteel het enige aanbod van Dell of HPE dat wordt geleverd met vier-richtings HGX H100-GPU's bij 2U. Het HPE AI-portfolio biedt 1U en 2U Gen11 ProLiant-servers, maar deze zijn beperkt tot PCIe-vormfactor GPU's.[25]

De Dell PowerEdge XE9640 server ondersteunt ook Intel Max GPU serie 1550 OAM GPU's, die een energiezuinige GPU-optie met hoge dichtheid bieden, inclusief een PCIe-kaart en een OpenCompute Accelerator module (OAM).[26] Hoewel we niet konden vaststellen dat HPE per 12/05/2023 een server met deze GPU's aanbood, bieden ze wel de HPE ProLiant DL380 Gen11 en DL380a Gen11 met Intel Data Center GPU Max 1100 GPU's.[27] Dit betekent dat de Dell PowerEdge XE9640 misschien het enige huidige aanbod is met vier Intel Max 1550 OAM GPU's in een 2U-server. Voor bedrijven die zich bezighouden met ruimtebesparing en energie-efficiëntie in hun datacenter biedt een 2U-server met vier Intel Max 1550 GPU's een oplossing die krachtige computing en energie-efficiëntie combineert zonder te veel ruimte in het datacenter in beslag te nemen.

De Dell PowerEdge XE9640 met vier HGX H100 GPU's presteerde tot wel 1,99 keer beter dan de HPE ProLiant DL380a met vier PCIe H100-GPU's in de gepubliceerde MLPerf® 3.1-resultaten (zie Figuur 3).

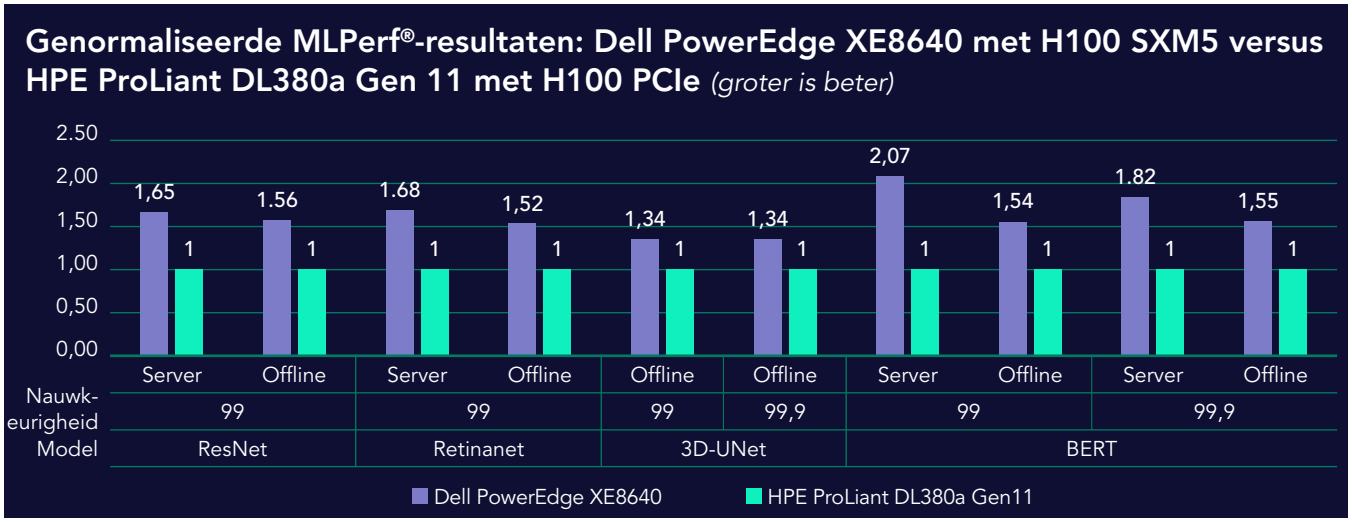


Afbeelding 3: Gepubliceerde MLPerf®-resultaten voor de Dell PowerEdge XE9680 en HPE ProLiant XL675d Gen10 Plus per 29-11-2023. Het Dell systeem gebruikt de NVIDIA H100-GPU, terwijl de GPU's in het HPE-systeem één generatie ouder zijn. Bron: Principled Technologies using data from MLCommons®.[28,29]

De PowerEdge XE8640 biedt een vier-richtings GPU-configuratie, met luchtkoeling voor processors en een Liquid Assisted Air Cooling Radiator voor de GPU's, waardoor er geen water-naar-rack-beschikbaarheid nodig is. Voor bedrijven die geen externe vloeistofkoeling gebruiken of kunnen gebruiken,[30] ondersteunt de 4U Dell PowerEdge XE8640 vier NVIDIA H100 SXM5-GPU's die dezelfde rekenkracht leveren als de PowerEdge XE9640, zonder dat directe vloeistofkoeling nodig is.[31]

De Dell PowerEdge XE8640 is uitgerust met de nieuwste 4e generatie Intel Xeon schaalbare processors en tot 4 TB geheugen[32] om de grote datasets en complexe berekeningen aan te kunnen die veel voorkomen bij AI en data-analyse. Nogmaals: HPE biedt de NVIDIA H100 SXM5-GPU's in HPE Cray-systemen, maar HPE ProLiant-servers met GPU-compatibiliteit ondersteunen dit niet.

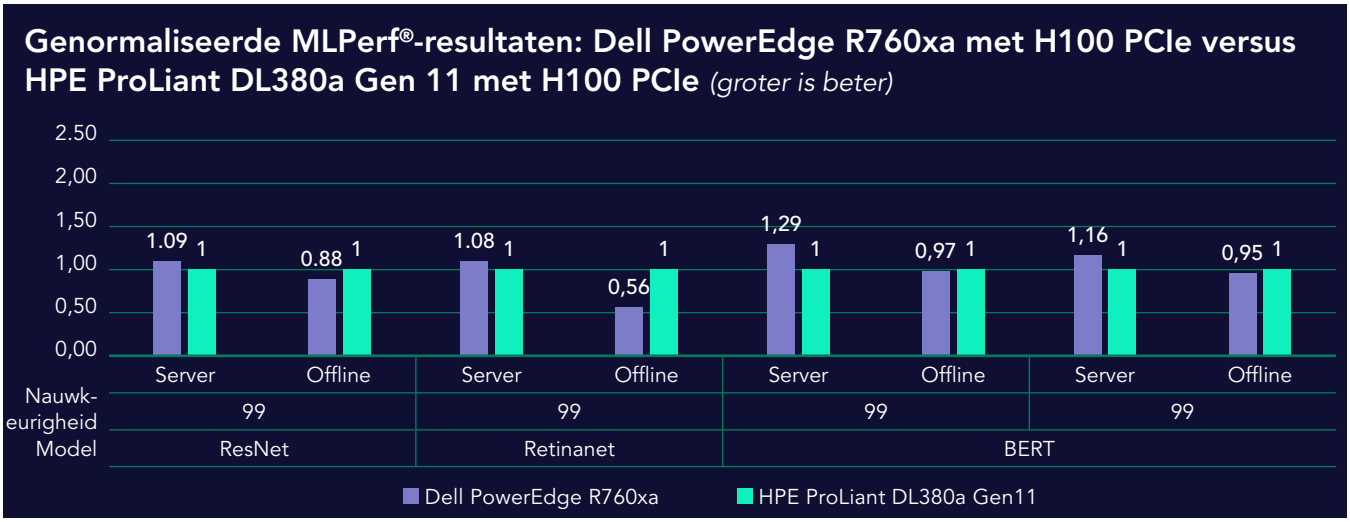
Vergeleken met MLPerf®-data die in november 2023 zijn gepubliceerd, behaalde de PowerEdge XE8640 server met vier NVIDIA H100 SXM5-GPU's de hoogste AI-uitvoer onder alle vier de GPU-inzendingen in negen verschillende categorieën. Zoals afbeelding 4 laat zien, scoorde deze tot 2,07 keer zo hoog in vergelijking met de HPE ProLiant DL380a-server.



Afbeelding 4: Gepubliceerde MLPerf®-resultaten voor de Dell PowerEdge XE8640 en HPE ProLiant DL380a Gen11 per 29-11-2023. Het Dell systeem gebruikt de NVIDIA H100 SXM-vormfactor, terwijl het HPE-systeem de minder krachtige PCIe-vormfactor gebruikt. Bron: Principled Technologies met gegevens van MLCommons®.[33,34]

Voor organisaties die kleiner willen starten en waar nodig willen kunnen groeien, bevat de 2U Dell PowerEdge R760xa server een reeks GPU's van NVIDIA, AMD en Intel, met ondersteuning voor maximaal vier PCIe Gen 5 GPU's met dubbele breedte of 12 PCIe GPU's met enkele breedte.[35] De server bevat 32 DIMM-slots, een bay met acht schijven voor 2,5 inch schijven en 12 PCIe-slots, die schaalbare storage bieden die kan groeien bij toenemende AI-datavereisten en ondersteuning voor maximaal 12 PCIe GPU's met enkele breedte of vier PCIe GPU's met dubbele breedte, zoals de NVIDIA H100 of L40S.[36] Deze schaalbaarheid betekent dat de server zich kan aanpassen aan veranderende AI-taken, van modeltraining voor machine learning tot geavanceerde dataverwerking.

Het luchtkoelsysteem van de PowerEdge R760xa ondersteunt computeromgevingen met hoge dichtheid en kan hogere TDP-versnellers (Thermal Design Power) tot 350 W bevatten.[37] Zo blijven de prestaties, goed, ook bij intensieve computingbelasting. In de MLPerf®-testresultaten voor ResNet, RetinaNet en BERT Server, gepubliceerd in november 2023, met de server-modus, presteerde de PowerEdge R760xa met vier NVIDIA H100 PCIe-GPU's beter dan de HPE ProLiant DL380a Gen 11, ook uitgerust met vier H100 PCIe-GPU's (zie Figuur 5).



Afbeelding 5: Gepubliceerde MLPerf®-resultaten voor de Dell PowerEdge R760xa en HPE ProLiant DL380a Gen11 per 29-11-2023. Beide systemen gebruiken de PCIe-vormfactor van de NVIDIA H100-GPU's. Bron: Principled Technologies using data from MLCommons®.[38,39]

Al met al tonen de MLPerf®-resultaten aan dat de prestaties sterk verschillen tussen servers en componenten. Het is dus belangrijk om de juiste opties te selecteren die uw workloads en de prestatievereisten daarvan ondersteunen. Dell PowerEdge servers voor AI-workloads bieden meerdere opties voor koeling en dichtheid om aan elke behoefte van een datacenter te voldoen. Bovendien scoren ze hoog op de MLPerf®-tests.

Uitgebreidere dekking van het AI-portfolio van Dell

Hoewel computingprestaties cruciaal zijn, zijn ze slechts één overweging bij het plannen van uw AI-workloads. U moet ook rekening houden met de rest van het AI-portfolio dat een leverancier aanbiedt wanneer u begint met een AI-implementatie. Hieronder bespreken we aanvullende categorieën die belangrijk zijn voor deze AI-portefeuilles, waaronder workstations voor klanten, cloudeigen producten, storage en nog veel meer. We bespreken ook gebieden waar het aanbod van Dell een voordeel kan bieden in vergelijking met HPE.

Workstations

Voor AI-ontwikkelaars en datawetenschappers bieden Dell Precision Data Science workstations NVIDIA RTX™-GPU's en Intel Xeon® CPU's, samen met een reeks tools voor datawetenschap.[40] Deze systemen gebruiken professionele computingopties met NVIDIA-GPU's die zijn gecertificeerd voor meer dan 100 professionele toepassingen[41] en Intel Xeon schaalbare processorversnellers zoals Intel DL Boost.[42] Precision workstations zijn verkrijgbaar in mobiele, tower- en rackformaten om te voldoen aan behoeften variërend van grotere, stationaire data-analyse tot wetenschappelijke veldmodellering onderweg.

Het aanbod van HPE-workstations is smaller en bestaat voornamelijk uit towers met één workstation die zijn uitgerust met NVIDIA L4S. HPE biedt geen optie voor mobiele workstations.[43] Hoewel hun aanbod aan workstationtowers geschikt is voor veel taken, biedt het niet dezelfde flexibiliteit en workloaddekking als het uitgebreidere assortiment van Dell. De verschillende opties voor grootte en draagbaarheid van Dell Precision workstations maken meer op maat gemaakte oplossingen mogelijk die voldoen aan de verschillende behoeften in omgevingen zoals laboratoria, kantower en werken in het veld.

Storage

Storage kan net zo belangrijk zijn als computing voor het uitvoeren van AI-workloads. Meer data verbetert de nauwkeurigheid van AI-modellen, maar het opslaan en beheren van enorme datasets is niet altijd mogelijk bij veel datacenters. Omdat modellen doorgaans worden getraind met behulp van niet-gestructureerde data moeten AI-ready storagesystemen bovendien veel verschillende datatypen met gemak kunnen verwerken.[44] Om capaciteit en schaalbaarheid te bieden voor AI-, ML- en DL-datasets, biedt Dell de PowerScale™ serie voor bestandstorage en Elastic Cloud Storage (ECS) of softwaregedefinieerde ObjectScale voor objectstorage.

Het PowerScale all-flash NAS portfolio biedt capaciteitsopties variërend van 3,84 TB tot 720 TB onbewerkte capaciteit per knooppunt, met geclusterde all-flash-capaciteiten die 186 PB onbewerkte capaciteit bereiken. De flexibiliteit en schaal van PowerScale werken voor een breed scala aan klanten en gebruiksscenario's voor AI.[45] Als de PowerScale F900 is geclusterd, kan deze tot 186 PB aan totale raw-storage bereiken.[46] Alle drie de all-flash PowerScale modellen (F200, F600 en F900) bevatten inline datacompressie en deduplicatie om de storage-efficiëntie te verbeteren.[47] Elk PowerScale storagemodel gebruikt het Dell OneFS™ bestandssysteem, dat beleid toepast om storage in lagen te plaatsen om prioriteit te geven aan de belangrijkste data op de snelste lagen voor workloadoptimalisatie.[48] Dell biedt ook OneFS-software in de AWS-marketplace met Dell APEX File Storage for AWS. Klanten kunnen OneFS gebruiken met hun AWS-compute-instanties voor een consistente gebruikerservaring met dezelfde functies die beschikbaar zijn in lokale OneFS-arrays.[49] Hoewel HPE public cloud-integratie biedt voor hybride storageoplossingen, vonden we geen cloudeigen optie zoals Dell APEX File Storage for AWS in het aanbod.

Objectstorageopties van Dell bevatten Zakelijke Dell objectstorage (ECS) dat "speciaal is gemaakt om niet-gestructureerde data op public cloud-schaal op te slaan."[50] Samen met ingebouwde compatibiliteit met Amazon S3-objectstorage voor hybrid cloud-functionaliteit leveren ECS storageknooppunten capaciteiten tot 14 PB per rack.[51] HPE biedt ook niet-gestructureerde storage met bestands- en objectstorageopties, maar het objectstorageaanbod verloopt via een partnerschap met Scality. Klanten kunnen HPE-oplossingen voor Scality kopen bij HPE.[52]

Professionele services

Dell biedt een breed scala aan professionele services, waaronder consulting, datavoorbereiding, implementatie, support en beheerde services om AI-implementaties te ondersteunen. Voor organisaties die op zoek zijn naar gevalideerde architecturen en oplossingen, biedt Dell Gevalideerde AI-ontwerpen, die zich richten op specifieke toepassingen om het giswerk uit het ontwerpen en implementeren van AI-resources te halen. Deze door Dell gevalideerde AI-oplossingen bevatten hardware- en softwarebundels, conversationele AI-modellen, machinelearning-activiteiten en meer. Door vooraf geconfigureerde, speciaal gebouwde oplossingen te combineren met AI-gerelateerde services, biedt Dell een uitgebreide AI-oplossing voor het hele spectrum aan AI-behoeften. Deze aanbiedingen kunnen een snellere en eenvoudigere weg naar AI-succes betekenen in vergelijking met het bouwen van ad-hocoplossingen.

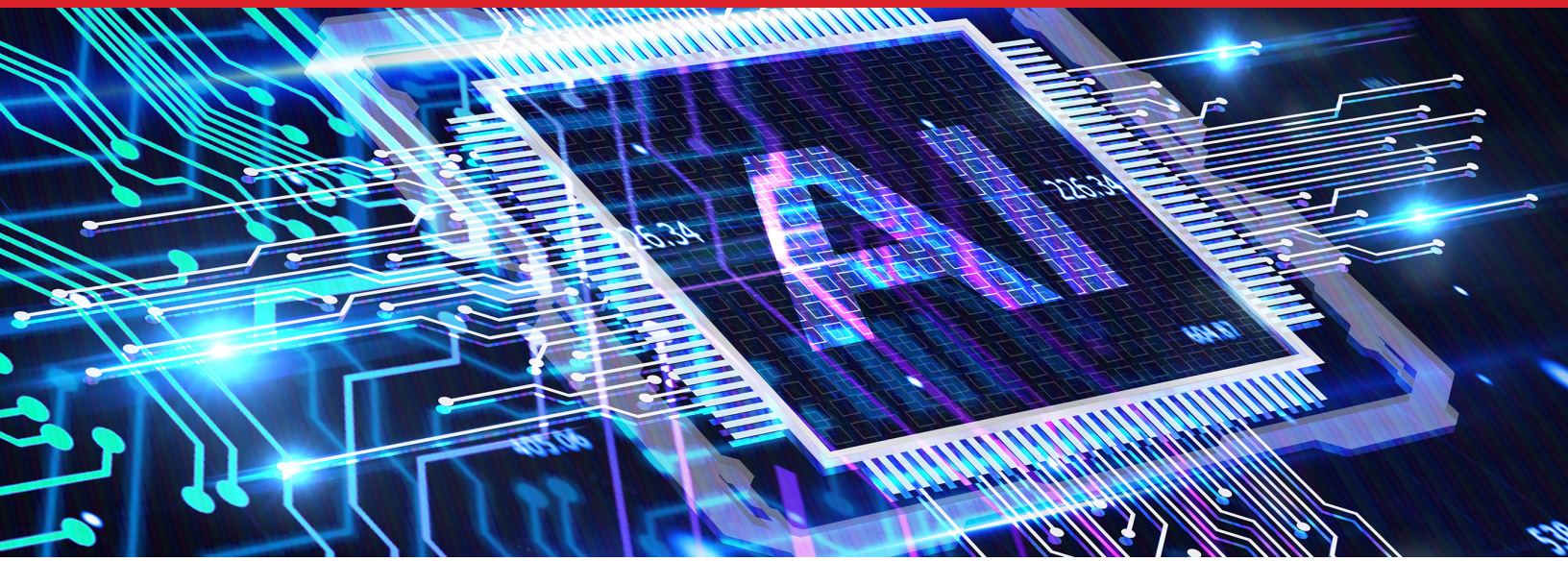
Dell services kunnen ook uw AI-traject begeleiden, van advies tot implementatie. Dell ProConsult Advisory Services helpt klanten in kaart te brengen waar gebruikers kunnen profiteren van het gebruiken van GenAI-processen en een roadmap te maken met de vereiste oplossingen en IT-vaardigheden. Dell services kunnen data voorbereiden op integratie van grote taalmodellen en IT-teams trainen op AI-kennis. Voor de volledige ingebruikname van GenAI controleren Dell teams uw specifieke toepassingen en bepalen, implementeren en configureren ze het beste AI-model dat aansluit bij uw behoeften. HPE biedt ook professionele services om bedrijven te ondersteunen bij hun AI-traject.[53,54]

Overwegingen voor beheer

Servers vereisen doorlopend beheer dat tijd van beheerders in beslag neemt. Firmware, software en drivers moeten periodiek worden bijgewerkt, IT-personeel moet de prestaties en temperaturen optimaliseren en onderhouden, en meer. In eerdere tests van Principled Technologies (PT) hebben we de beheermogelijkheden van Dell servers met Integrated Dell Remote Access Controller 9 (iDRAC9) beoordeeld.[55] Beheerders kunnen vertrouwen op geautomatiseerde online updates met iDRAC9 OpenManage™ Enterprise (OME) met configureerbare planning om hun servers up-to-date te houden en profielen te gebruiken om snel en eenvoudig nieuwe servers te integreren naarmate de workloads toenemen. Met iDRAC en OME hebben Dell klanten toegang tot meer functies voor extern beheer, kunnen ze servers eenvoudiger implementeren en firmware eenvoudiger bijwerken dan met HPE OneView en HPE iLO. Dell PowerEdge servers bevatten beheer en services van Dell die organisaties kunnen helpen door "te zorgen dat het minder tijd en moeite kost om taken uit te voeren zoals het bewaken van de systeemstatus of het bijwerken van firmware", waardoor IT-cycli vrijkomen voor innovatie en andere taken.[56]

Tabel 3: Samenvatting van de vergelijking tussen Dell en HPE-beheertools uit een PT-rapport van november 2022.[57]
Bron: Principled Technologies.

	Wat is het verschil met Dell beheertools	Hoeveel beter
Meer functies voor beheer op afstand iDRAC versus iLO	Meer HTML5-console- en BIOS-configuratiefuncties voor meer externe functionaliteit in iDRAC	2,5x de HTML5-consolefuncties en 13x de BIOS-functies
Eenvoudigere serverimplementatie OME versus OneView	Een-naar-veel profielimplementatie met OME	52% minder tijd om een server te implementeren dan met OneView
Eenvoudigere firmware-updates OME versus OneView	Geautomatiseerde online updates met OME	Werk meerdere servers bij door verbinding te maken met Dell.com, waardoor u de tijd bespaart die nodig is om servers bij te werken door handmatig bundels te uploaden met OneView
Makkelijkere waarschuwingen OME versus OneView	Stel waarschuwingsbeleid in OME in en voer geautomatiseerde acties uit op basis van waarschuwingen	Door dit proces te automatiseren bespaart u tijd en vermindert u de kans op fouten in plaats van acties handmatig uit te voeren telkens als u een waarschuwing krijgt in OneView
Eenvoudiger te gebruiken beveiligingsfuncties (systeemvergrendeling en dynamische USB) iDRAC versus iLO	Minder stappen, minder tijd, niet opnieuw opstarten met iDRAC	¼ van de stappen, 91% minder tijd voor systeemvergrendeling
Krachtigere analyses CloudIQ voor PowerEdge versus InfoSight	Aanpasbare rapporten, meer statusstatistieken voor betere beheercontrole met CloudIQ voor PowerEdge	Meer dan 15 x meer statistieken om uit te kiezen in vergelijking met InfoSight



Conclusie

De kracht van AI benutten om bedrijfsactiviteiten te stroomlijnen en te verbeteren kan een uitdagende taak zijn, met aanzienlijke zakelijke implicaties. Nu technologie sneller dan ooit vooruitgaat, is samenwerken met de juiste AI-leverancier cruciaal. Door te kiezen voor een bedrijf als Dell, dat niet alleen een uitgebreid AI-portfolio biedt, maar ook plannings-, voorbereidings-, implementatie- en beheerservices kan bieden, kunnen klanten deze uitdagingen het hoofd bieden. MLPerf®-benchmarktests tonen aan dat het aanbod in het AI-portfolio van Dell consistente, krachtige prestaties biedt voor AI-workloads. Met krachtige en flexibele serveropties, samen met meerdere storagekeuzes, gevalideerde oplossingen en professionele services die speciaal zijn afgestemd op AI, kan Dell bedrijven helpen om AI en de voordelen daarvan te omarmen.

---

1. Dell, "Increasing Your Data Value with Dell Generative AI Solutions", geraadpleegd op 19 december 2023, https://www.dell.com/en-us/blog/increasing-your-data-value-with-dell-generative-ai-solutions/.
2. Dell, "Dell AI solutions", geraadpleegd op 12 december 2023, https://www.dell.com/en-us/dt/solutions/artificial-intelligence/index.htm#accordion0&tab0=0.
3. Dell, "Dell Technologies and Hugging Face to Simplify Generative AI with On-Premises IT", geraadpleegd op 12 december 2023, https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2023~11~20231114-dell-technologies-and-hugging-face-to-simplify-generative-ai-with-on-premises-it.htm#/filter-on/Country:en-us.
4. Dell, "Dell and Meta Collaborate to Drive Generative AI Innovation", geraadpleegd op 12 december 2023, https://www.dell.com/en-us/blog/dell-and-meta-collaborate-to-drive-generative-ai-innovation/.
5. Dell, "Dell AI-Ready Data Platform", geraadpleegd op 12 december 2023, https://www.dell.com/nl-nl/dt/solutions/artificial-intelligence/storage-for-ai.htm?hve=explore+unstructured+storage#tab0=0.
6. Dell, "Snowflake and Dell Partnership Gains Momentum", geraadpleegd op 19 december 2023, https://www.dell.com/en-us/blog/snowflake-and-dell-partnership-gains-momentum/.
7. Robert McNeal, "Dell, VMware and NVIDIA Bring AI to Your Data", geraadpleegd op 17 januari 2024, https://www.dell.com/en-us/blog/dell-vmware-and-nvidia-bring-ai-to-your-data/. Volgens de bovenstaande link: "Based on Dell analysis, August 2023. Dell Technologies offers solutions engineered to support AI workloads from Workstations PCs (mobile and fixed) to Servers for High-performance Computing, Data Storage, Cloud Native Software-Defined Infrastructure, Networking Switches, Data Protection, HCI and Services."
8. Intel, "Accelerate Artificial Intelligence (AI) Workloads with Intel Advanced Matrix Extensions (Intel AMX)", geraadpleegd op 12 december 2023, https://www.intel.com/content/www/us/en/content-details/785250/accelerate-artificial-intelligence-ai-workloads-with-intel-advanced-matrix-extensions-intel-amx.html.

9. Vipera, "NVIDIA's H100 and A100 GPU Cards: Exploring the Intricacies of SXM and PCI-E Connections", geraadpleegd op 12 december 2023, https://www.viperatech.com/unraveling-the-mysteries-sxm-vs-pci-e-connections-in-nvidias-high-end-h100-and-a100-gpus/.
10. GitHub, "MLPerf® Results Messaging Guidelines", geraadpleegd op 16 januari 2024, https://github.com/mlcommons/policies/blob/master/MLPerf_Results_Messaging_Guidelines.adoc.
11. MLCommons®, "MLPerf® Inference: Datacenter Benchmark Suite Results", geraadpleegd op 12 december 2023, https://mlcommons.org/en/inference-datacenter-31/.
12. Dell, "PowerEdge AI servers", geraadpleegd op 12 december 2023, https://www.dell.com/nl-nl/dt/servers/specialty-servers/poweredge-xe-servers.htm?hve=explore+poweredge+xe#tab0=0.
13. HPE, "HPE ProLiant XL675d Gen10 Plus Configure-to-order Server", geraadpleegd op 12 december 2023, https://www.hpe.com/us/en/product-catalog/compute/proliant-servers/pip.1013142988.html.
14. HPE, "HPE ProLiant DL380a Gen11", geraadpleegd op 12 december 2023, https://www.hpe.com/us/en/product-catalog/compute/proliant-servers/pip.proliant-dl380-server.1014696168.html.
15. MLCommons®, "MLPerf® Inference: Datacenter Benchmark Suite Results v 3.1", geraadpleegd op 12 december 2023, https://mlcommons.org/benchmarks/inference-datacenter/.
16. HPE, "NVIDIA Accelerators for HPE ProLiant Servers", geraadpleegd op 12 december 2023, https://www.hpe.com/psnow/doc/c04123180.html?jumpid=in_pdp-psnow-qs.
17. Dell, "PowerEdge XE9680 Specification Sheet", geraadpleegd op 19 januari 2024, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe9680-spec-sheet.pdf.
18. HPE, "HPE & NVIDIA financial services solution sets new records in performance", geraadpleegd op 12 december 2023, https://community.hpe.com/t5/alliances/hpe-amp-nvidia-financial-services-solution-sets-new-records-in/ba-p/7197388.
19. HPE, "QuickSpecs: HPE Cray Supercomputing XD670", geraadpleegd op 12 december 2023, https://www.hpe.com/psnow/doc/a50004292enw.
20. Geverifieerde MLPerf®-score van v3.1 Inference Closed. Opgehaald van https://mlcommons.org/benchmarks/inference-datacenter/ op 5 december 2023, vermelding 3.1-0069. De naam en het logo van MLPerf® zijn geregistreerde en ongedeponeerde handelsmerken van MLCommons® Association in de Verenigde Staten en andere landen. Alle rechten voorbehouden. Ongeautoriseerd gebruik is strikt verboden. Zie www.mlcommons.org voor meer informatie.
21. Geverifieerde MLPerf®-score van v3.1 Inference Closed. Opgehaald van https://mlcommons.org/benchmarks/inference-datacenter/ op 5 december 2023, vermelding 3.1-0085. De naam en het logo van MLPerf® zijn geregistreerde en ongedeponeerde handelsmerken van MLCommons® Association in de Verenigde Staten en andere landen. Alle rechten voorbehouden. Ongeautoriseerd gebruik is strikt verboden. Zie www.mlcommons.org voor meer informatie.
22. Dell, "PowerEdge XE9640 rackserver", geraadpleegd op 12 december 2023, https://www.dell.com/nl-nl/shop/ipovw/poweredge-xe9640.
23. Accelsius, "Enabling the AI Revolution with Liquid Cooling", geraadpleegd op 12 december 2023, https://www.accelsius.com/blog/enabling-the-ai-revolution-with-liquid-cooling.
24. Dell, "Dell PowerEdge XE9640 Technical Guide", geraadpleegd op 12 december 2023, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe9640-technical-guide.pdf.
25. HPE, "HPE ProLiant DL380a Gen11", geraadpleegd op 12 december 2023, https://www.hpe.com/psnow/doc/PSN1014696168WWEN.pdf?jumpid=in_pdp-psnow-dds.
26. Intel, "Intel® Data Center GPU Max Series Technical Overview", geraadpleegd op 12 december 2023, https://www.intel.com/content/www/us/en/developer/articles/technical/intel-data-center-gpu-max-series-overview.html#gs.08874l.
27. HPE, "Intel Data Center GPU Max 1100 48GB Accelerator for HPE Data sheet", geraadpleegd op 12 december 2023, https://www.hpe.com/psnow/doc/PSN1014779728WWEN.
28. Geverifieerde MLPerf®-score van v3.1 Inference Closed. Opgehaald van https://mlcommons.org/benchmarks/inference-datacenter/ op 5 december 2023, vermelding 3.1-0066. De naam en het logo van MLPerf® zijn geregistreerde en ongedeponeerde handelsmerken van MLCommons® Association in de Verenigde Staten en andere landen. Alle rechten voorbehouden. Ongeautoriseerd gebruik is strikt verboden. Zie www.mlcommons.org voor meer informatie.
29. Geverifieerde MLPerf®-score van v3.1 Inference Closed. Opgehaald van https://mlcommons.org/benchmarks/inference-datacenter/ op 5 december 2023, vermelding 3.1-0084. De naam en het logo van MLPerf® zijn geregistreerde en ongedeponeerde handelsmerken van MLCommons® Association in de Verenigde Staten en andere landen. Alle rechten voorbehouden. Ongeautoriseerd gebruik is strikt verboden. Zie www.mlcommons.org voor meer informatie.

30. Dell, "AI and HPC —With Air or Liquid Cooling", geraadpleegd op 12 december 2023, https://www.delltechnologies.com/asset/en-us/products/servers/briefs-summaries/poweredge-xe9640-and-xe8640-infographic.pdf.
31. Dell, "PowerEdge XE8640: Drive AI, HPC modeling and simulation workloads with superior performance", geraadpleegd op 12 december 2023, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe8640-spec-sheet.pdf.
32. Dell, "PowerEdge XE8640 rackserver", geraadpleegd op 12 december 2023, https://www.dell.com/nl-nl/shop/ipovw/poweredge-xe8640.
33. Geverifieerde MLPerf®-score van v3.1 Inference Closed. Opgehaald van https://mlcommons.org/benchmarks/inference-datacenter/ op 5 december 2023, vermelding 3.1-0067. De naam en het logo van MLPerf® zijn geregistreerde en ongedeponeerde handelsmerken van MLCommons® Association in de Verenigde Staten en andere landen. Alle rechten voorbehouden. Ongeautoriseerd gebruik is strikt verboden. Zie www.mlcommons.org voor meer informatie.
34. Geverifieerde MLPerf®-score van v3.1 Inference Closed. Opgehaald van https://mlcommons.org/benchmarks/inference-datacenter/ op 5 december 2023, vermelding 3.1-0084. De naam en het logo van MLPerf® zijn geregistreerde en ongedeponeerde handelsmerken van MLCommons® Association in de Verenigde Staten en andere landen. Alle rechten voorbehouden. Ongeautoriseerd gebruik is strikt verboden. Zie www.mlcommons.org voor meer informatie.
35. Dell, "PowerEdge R760xa rackserver", geraadpleegd op 12 december 2023, https://www.dell.com/nl-nl/shop/dell-poweredge-servers/poweredge-r760xa-rack-server/spd/poweredge-r760xa/pe_r760xa_16902_vi_vp#features_section.
36. SANStorageWorks, "Dell EMC PowerEdge R760xa: Powerful and scalable for GPU workloads", geraadpleegd op 12 december 2023, https://www.sanstorageworks.com/PowerEdge-R760xa.asp.
37. Dell, "Dell PowerEdge Servers and NVIDIA GPUs", geraadpleegd op 12 december 2023, https://infohub.delltechnologies.com/l/design-guide-generative-ai-in-the-enterprise-inferencing/dell-poweredge-servers-and-nvidia-gpus-1/.
38. Geverifieerde MLPerf®-score van v3.1 Inference Closed. Opgehaald van https://mlcommons.org/benchmarks/inference-datacenter/ op 5 december 2023, vermelding 3.1-0064. De naam en het logo van MLPerf® zijn geregistreerde en ongedeponeerde handelsmerken van MLCommons® Association in de Verenigde Staten en andere landen. Alle rechten voorbehouden. Ongeautoriseerd gebruik is strikt verboden. Zie www.mlcommons.org voor meer informatie.
39. Geverifieerde MLPerf®-score van v3.1 Inference Closed. Opgehaald van https://mlcommons.org/benchmarks/inference-datacenter/ op 5 december 2023, vermelding 3.1-0084. De naam en het logo van MLPerf® zijn geregistreerde en ongedeponeerde handelsmerken van MLCommons® Association in de Verenigde Staten en andere landen. Alle rechten voorbehouden. Ongeautoriseerd gebruik is strikt verboden. Zie www.mlcommons.org voor meer informatie.
40. Dell, "Workstations klaar voor AI", geraadpleegd op 12 december 2023, https://www.dell.com/nl-nl/dt/ai-technologies/index.htm?hve=explore+dell+precision+for+ai#pdf-overlay=//www.delltechnologies.com/asset/nl-nl/products/workstations/briefs-summaries/ai-industry-brochure.pdf.
41. NVIDIA, "NVIDIA RTX in Professional Workstations", geraadpleegd op 12 december 2023, https://www.nvidia.com/en-us/design-visualization/desktop-graphics/.
42. Intel, "Intel® Deep Learning Boost (Intel® DL Boost)", geraadpleegd op 12 december 2023, https://www.intel.com/content/www/us/en/artificial-intelligence/deep-learning-boost.html.
43. HPE, "HPE ProLiant ML350 Gen11", geraadpleegd op 12 december 2023, https://buy.hpe.com/us/en/compute/tower-servers/proliant-ml300-servers/proliant-ml350-server/hpe-proliant-ml350-gen11/p/1014696172.
44. ComputerWeekly.com, "Storage requirements for AI, ML and analytics in 2022", geraadpleegd op 12 december 2023, https://www.computerweekly.com/feature/Storage-requirements-for-AI-ML-and-analytics-in-2022.
45. Dell, "PowerScale", geraadpleegd op 12 december 2023, https://www.dell.com/nl-nl/shop/powerscale-family/sf/powerscale.
46. Dell, "PowerScale vergelijken", geraadpleegd op 12 december 2023, https://www.dell.com/nl-nl/shop/powerscale-family/sf/powerscale#compare-module.
47. Dell, "Dell PowerScale all-flash", geraadpleegd op 12 december 2023, https://www.delltechnologies.com/asset/nl-nl/products/storage/technical-support/h15963-ss-powerscale-all-flash-nodes.pdf.
48. Dell, "Dell PowerScale OneFS Software Features", geraadpleegd op 12 december 2023, https://www.delltechnologies.com/asset/nl-nl/products/storage/technical-support/h18275-onefs-software-features-data-sheet.pdf.
49. Dell, "Dell APEX File Storage voor AWS", geraadpleegd op 12 december 2023, https://www.delltechnologies.com/asset/nl-nl/products/storage/briefs-summaries/h19575-so-apex-file-storage-for-aws.pdf.

50. Dell, "Dell ECS Enterprise objectstorage", geraadpleegd op 12 december 2023, https://www.dell.com/nl-nl/dt/storage/ecs/index.htm?hve=explore+ecs#tab0=0&tab1=0.
51. Dell, "Dell ECS Enterprise objectstorage", geraadpleegd op 12 december 2023, https://www.dell.com/en-us/dt/storage/ecs/index.htm#tab0=0&tab1=0&accordion0.
52. HPE, "Storage Solutions for Scality", geraadpleegd op 12 december 2023, https://www.hpe.com/us/en/storage/file-object/scality.html.
53. HPE, "Make AI work for you", geraadpleegd op 16 januari 2024, https://www.hpe.com/us/en/solutions/ai-artificial-intelligence.html.
54. HPE, "HPE AI Services – Generative AI Implementation", geraadpleegd op 16 januari 2024, https://www.hpe.com/us/en/services/generative-ai-implementation-service.html.
55. Principled Technologies, "Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE", geraadpleegd op 12 december 2023, https://www.principledtechnologies.com/Dell/Management-tools-vs-HPE-1122.pdf.
56. Principled Technologies, "Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE", geraadpleegd op 12 december 2023, https://www.principledtechnologies.com/Dell/Management-tools-vs-HPE-1122.pdf.
57. Principled Technologies, "Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE."

▶ Bekijk de oorspronkelijke, Engelstalige versie van dit rapport via https://facts.pt/zPmSx4c

Dit project is uitgevoerd in opdracht van Dell Technologies.

Principled Technologies®

Facts matter.®