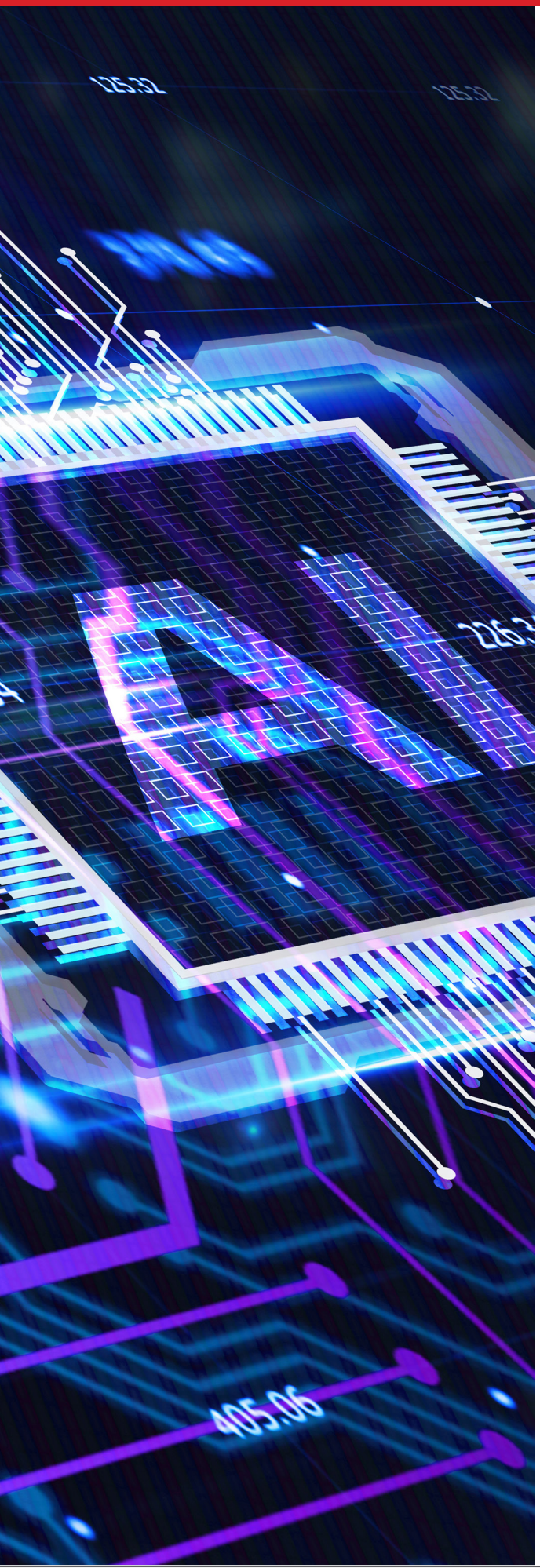


# Hantera utmaningarna med AI-arbetsbelastningar med hjälp av Dells AI-portfölj

## En jämförelse av Dells AI-portfölj jämfört med liknande erbjudanden från HPE

Det råder ingen tvekan om att artificiell intelligens (AI) har förändrat affärslandskapet och gjort det möjligt för branscher och företag i alla storlekar att få djupare insikter från sina data, automatisera affärsprocesser, leverera skräddarsydda kund- och användarupplevelser och vara starkare konkurrenter i sin bransch. Men om företag ska kunna utnyttja kraften i AI på ett effektivt sätt behöver de en infrastrukturleverantör som kan erbjuda en omfattande och integrerad lösningsportfölj som omfattar hela AI-livscykeln.

Då kan infrastrukturleverantörer som Dell Technologies och Hewlett Packard Enterprise (HPE) hjälpa kunder att hantera de växande kraven från AI och navigera i det komplexa landskapet. Dessa leverantörer har AI-förberedda portföljer och erbjuder olika nivåer av AI-lösningar som förenar högpresterande infrastrukturlösningar på plats och i molnet med strategiska partnerskap och ett urval av support- och rådgivningstjänster.

Den här rapporten undersöker offentligt tillgänglig information om Dells och HPE:s AI-portföljer* med målet att lyfta fram specifika arkitektur-, prestanda- och supportfördelar som kunder kan dra nytta av genom att välja Dell Technologies för sina AI-behov. Vi jämför information om de servrar som Dell har byggt för att stödja AI-distributioner och hänvisar till testresultat för branschen från ML Commons®. Vi utforskar även ytterligare mjukvaru- och tjänsteutbud som stöder kunderna i varje steg av deras AI-resa.

*Obs! PT slutförde all forskning på eller före den 5 december 2023, så det här dokumentet återspeglar inte erbjudanden eller ändringar, varken från Dell eller HPE, efter det datumet.

## Utmaningarna med att börja använda AI

Det finns många utmaningar för datacenter och IT-personal när de ska börja använda en AI-strategi:

- Hantera befintliga kunskapsluckor hos den nuvarande personalen genom antingen intern AI-utbildning eller extern anställning.
- Förstå alla förberedelser och databehov för AI, inklusive kvalitet, kvantitet, plats och aktuellt tillstånd för företagets data.
- Bedöma företagets specifika AI-mål för att bättre avgöra vilka AI-modeller och implementeringar som ger fördelar.
- Bedöma beräknings-, nätverks- och lagringsbehoven för de planerade AI-systemen och fastställa en förvärvsplan.

Det här är bara några exempel på de många, ofta betydande, hinder som ett företag står inför när det gäller att dra nytta av fördelarna med att implementera AI i sina datacenter.

Dells AI-portfölj syftar till att hjälpa kunder att hantera dessa utmaningar genom professionella och rådgivande tjänster som hjälper kunderna att bygga upp handlingsplaner för implementering och förbereda sina data för AI-modeller.[1] Portföljen innehåller även utbildningar som omfattar koncept för maskininlärning (ML) och andra utbildningsämnen och erbjuder validerade utformningar för AI för att säkerställa att implementeringen lyckas.[2] Dessutom samarbetar Dell med tredje part för att ge kunderna ytterligare AI-verktyg, till exempel en anpassad Dell-portal inom Hugging Face-communityn med dedikerade behållare och skript för driftsättning av AI-modeller med öppen källkod[3] och enkel distribution av den stora språkmodellen (LLM) Meta Llama 2.[4] Tillsammans med ett stort urval av beräknings- och datorerbjudanden, från mobila workstation-datorer till servrar som stöder upp till åtta avancerade NVIDIA-grafikprocessorer, erbjuder Dell även den ostrukturerade datalagring som AI kräver tillsammans med en portfölj av högpresterande fil- och objektlagringsdisksystem. Dessa lagringserbjudanden, inklusive Dell PowerScale, ObjectScale, ECS och inbyggd lagring, kan hantera de ostrukturerade data som AI-arbetsbelastningar ofta använder.[5] Dell har också samarbetat med Snowflake för att tillhandahålla en hybridmolnlagringslösning för Dell-kunder.[6] Enligt Dells analys från augusti 2023 erbjuder de den ”bredaste portföljen med generativ AI” som sträcker sig bortom bara servrar och lagring genom att tillhandahålla resurser under hela AI-implementeringsresan.[7]

## AI-prestanda och accelererade beräkningsalternativ: Dell jämfört med HPE

AI-arbetsbelastningar kan använda processorer, GPU:er eller båda som beräkningsresurser beroende på storlek eller typ av arbetsbelastning. Vissa processorer tillhandahåller AI-specifika acceleratorer, till exempel Intel Advanced Matrix Extensions (Intel AMX) i de senaste Intel Xeon skalbara processorerna.[8] GPU:er är ofta bättre för större och/eller mer krävande arbetsbelastningar, men GPU-formfaktorn kan påverka prestandanivåerna. Till exempel finns vissa NVIDIA A100- och H100-modellgrafikprocessorer i antingen universella PCIe- eller tillverkarspecifika SXM-formfaktorer där den senare använder den högpresterande NVIDIA SXM-arkitekturen.[9] Stor minneskapacitet och serverdesignfunktioner som nedkylningsarkitektur och energieffektivitet påverkar också prestanda. De flesta datacenter använder fortfarande luftkylning. Det innebär att högprestandaberäkning (HPC) behöver servrar som är byggda för att kyla med luft så effektivt som möjligt. Nedan lyfter vi fram PowerEdge-servererbjudanden för bland annat komponenter och kylningsalternativ tillsammans med deras publicerade MLCommons® MLPerf-®poäng.

## Prestandatest av AI-modell: jämförelse av MLPerf-resultat

MLPerf® är en prestandasvit som testar AI-prestanda för både träning och inferens. Om ett företag ska kunna publicera officiella MLPerf®-resultat måste resultaten överensstämma med de specifika villkor som fastställts av prestandautvecklaren MLCommons®.[10] Dessa riktlinjer för överensstämmelse innehåller standarder som gör det enklare att jämföra prestanda. För inferenstestning använder MLPerf® datacenter, Edge, Mobile och små datauppsättningar, och rapporterar AI-poäng och watt som förbrukas under testning. Sviten för inferensprestanda innehåller tester för många vanliga AI-, ML- och DL-modeller, se tabell 1.

Tabell 1: AI-, ML- och DL-modeller ingår i MLPerf®-testning och typiska användningsfall för varje. Källa: Principled Technologies.

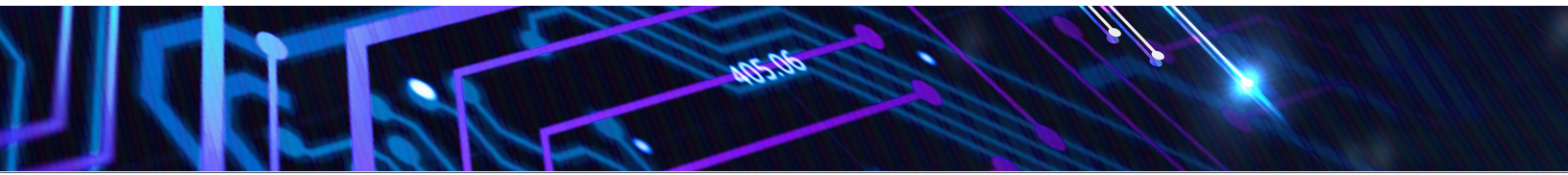
| Vanliga AI-modeller | Vanliga användarfall |
|---|---|
| ResNet | En bildklassificeringsmodell som hjälper datorer att lära sig, komma ihåg och identifiera olika bilder för användningsfall som medicinsk bildbehandling, moderering av innehåll på sociala medier och ansiktsigenkänning |
| RetinaNet | En typ av objektdetektering som kan hantera ytterligare komplexitet jämfört med ResNet. Det hjälper datorer att identifiera och hitta objekt i bilder eller videobildrutor och kan klassificera dem efter betydelse. Används för bland annat autonom körning, automatisk assistansteknik för fordon, övervakning och ansiktsigenkänning |
| 3D-UNet | Specifikt för medicinsk bildsegmentering |
| RNN-T | Taligenkänning för användningsfall som automatisk språköversättning |
| BERT | Naturlig språkbehandling för användningsfall som textsammanfattning, språköversättning och automatiskt slutförande av uppgifter |
| DLRM-v2-99.9 | Rekommendationsmodell för användningsfall som riktade annonser och anpassade produktrekommendationer |
| GPTJ-99 och 99.9 | LLM för naturlig språkbearbetning som utmärker sig vid textgenerering för användningsfall som chattbotar och chattbaserade AI-verktyg |

### MLPerf

MLPerf®-resultaten omfattar flera parametrar utöver själva AI-modellerna. Det innebär att mycket data kan analyseras i ett enda diagram eller en tabell. Här är en snabbreferens till dessa parametrar:

- 99,0 och 99,9: dessa siffror avser den noggrannhet som modellen har tränats med. Ju mer exakt utdata du behöver, desto mer komplex är modellen och desto längre tid kan det ta att bearbeta data.
- Offlineprover/sek: läge där prestandatestet skickar alla frågor i början av testet för att simulera data som redan finns i systemet.
- Serverfrågor/sek: läge där prestandatestet skickar frågor under hela testperioden för att simulera analys av en livestream av data.

Mer information om MLCommons®- och MLPerf®-resultat finns på https://mlcommons.org/benchmarks/inference-datacenter/.



Resultaten i denna rapport kommer från MLPerf® v3.1 Inference Datacenter-resultat som publicerades på webbplatsen för MLCommons® från november 2023.[11] Dessa resultat inkluderar bidrag från tekniktillverkare och molntjänstleverantörer och omfattar en rad konfigurationer. Jämfört med offentligt tillgängliga bidrag från HPE gav Dell-servrar bättre resultat i vissa AI-modeller. (Obs! Olika GPU-konfigurationer på servrarna kan göra det svårt att jämföra mellan olika enheter.) Mer information finns i tabell 2.

Tabell 2: Dell- och HPE-servrar som ingår i MLCommons® MLPerf® 3.1-resultaten publicerade den 29 november 2023. Källa: Principled Technologies.

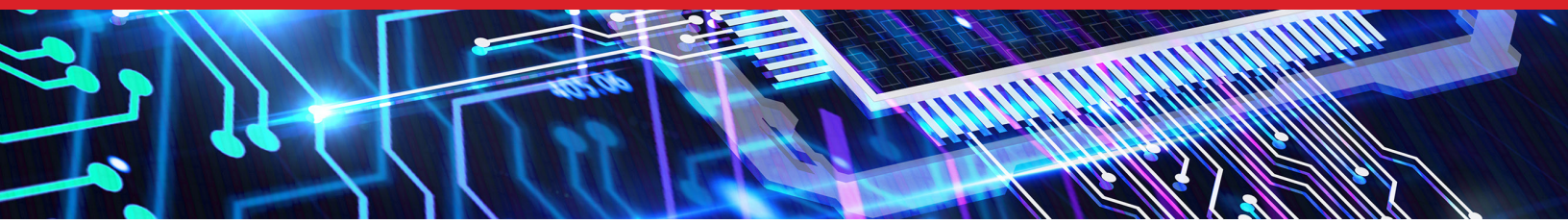
| Inlämnare | Servermodell | Antal och modell för GPU:er | Beskrivning |
|---|---|---|---|
| Dell[12] | PowerEdge XE9680 | 8x NVIDIA H100 SXM | För AI-utbildning och inferens med stora arbetsbelastningar som stora språkmodeller |
| | PowerEdge XE9640 | 4x NVIDIA H100 SXM | För träning av stora AI-modeller i datacenter med hög densitet och vätskekylda datacenter |
| | PowerEdge XE8640 | 4x NVIDIA H100 SXM | För drift av traditionella appar för AI-träning, HPC och dataanalys i en 4U-formfaktor för luftkylda datacenter |
| | PowerEdge R760xa | 4x NVIDIA H100 PCIe | För ett stort utbud av arbetsbelastningar med hög beräkning, inklusive AI-ML/DL-utbildning och inferens som inte kräver högpresterande GPU:er |
| HPE[13,14] | ProLiant XL675d Gen10 Plus | 8x NVIDIA A100 SXM | För högpresterande beräkning och AI |
| | ProLiant DL380a Gen11 | 4x NVIDIA H100 PCIe | 2U-server för måttliga AI-arbetsbelastningar |

## Direkt jämförelse mellan Dell- och HPE-servrar

Även om en omfattande AI-strategi består av mer än hårdvara är en av de viktigaste faktorerna för att lyckas med AI-arbetsbelastningar att säkerställa den bästa hårdvaruprestandan. I takt med att nyare GPU:er och annan teknik blir tillgängliga utvecklas även AI-arbetsbelastningsfunktionerna. När MLPerf® v3.1-resultaten först publicerades var den bästa tillgängliga NVIDIA GPU:n H100 Tensor Core med vilken Dell publicerade MLPerf®-resultat i flera av sina servrar i både PCIe- och SXM5-formfaktorer.[15] De publicerade HPE-resultaten omfattade endast ett H100-bidrag och med endast PCIe-formfaktorn. Vår forskning visade att få av de tillgängliga GPU-kompatibla HPE-servrarna stödde SXM5 H100-formfaktorn för bästa NVIDIA GPU-prestanda och ingen av HPE ProLiant-servrarna gör det.[16] Som vi visar nedan innebär bättre GPU:er vanligtvis en förbättring av AI-arbetsbelastningen.

## Åtta GPU MLPerf-resultat

Dell PowerEdge XE9680 har stöd för upp till åtta NVIDIA H100 SXM5-grafikprocessorer för AI-acceleration och upp till två 4:e generationens skalbara Intel® Xeon®-processorer. PowerEdge XE-produktserien har en modulär arkitektur med stöd för GPU-enheter med SXM.4 eller SXM5 NVIDIA GPU:er eller OAM (Open Compute Project Accelerator Module) från AMD, något som kan öka prestandan jämfört med en vanlig PCIe GPU.[17] PowerEdge XE9680 kräver bara 6 enheter rackutrymme och är en kompakt 8-vägs NVIDIA H100 SXM5-server. De senaste HPE ProLiant Gen11-servrarna har för närvarande inte stöd för H100 SXM-formfaktorn[18] även om vissa av HPE Cray Supercomputing-servrarna har det.[19] Eftersom HPE inte skickade in några MLPerf®-resultat med Cray-servrarna och endast belyser sina ProLiant-servrar på sidan för AI-portföljen, fokuserar vi på ProLiant-servrar i det här dokumentet. (Se bild 1)



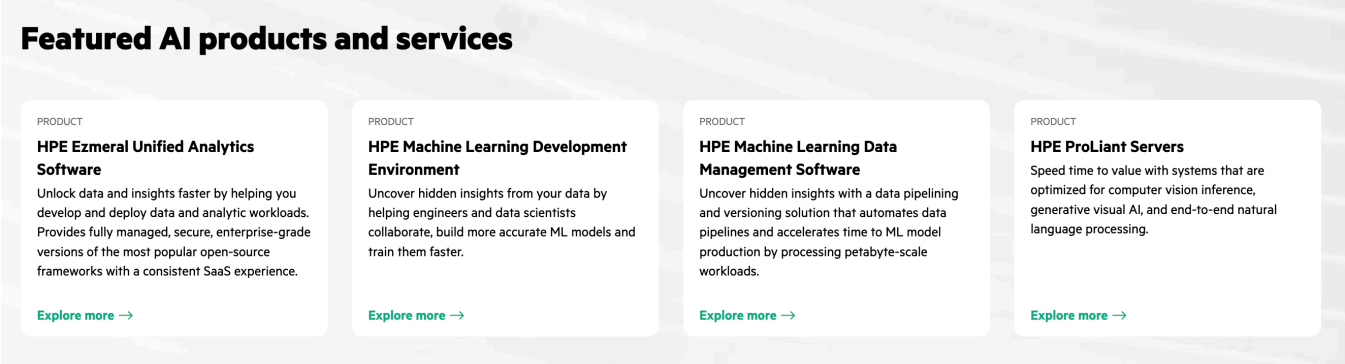



Bild 1: Skärmbild av utvalda AI-produkter och -tjänster på https://www.hpe.com/us/en/solutions/ai-artificial-intelligence.html som visar HPE ProLiant-servrar från och med den 5 december 2023.

I MLPerf® v3.1-resultaten som först publicerades i november 2023 för åtta GPU-servrar presterade Dell PowerEdge XE9680 med NVIDIA SXM5 H100 GPU:er bättre än HPE ProLiant XL675d Gen10 Plus med NVIDIA SXM.4 A100 GPU:er med upp till 4,25 gånger (se bild 2).

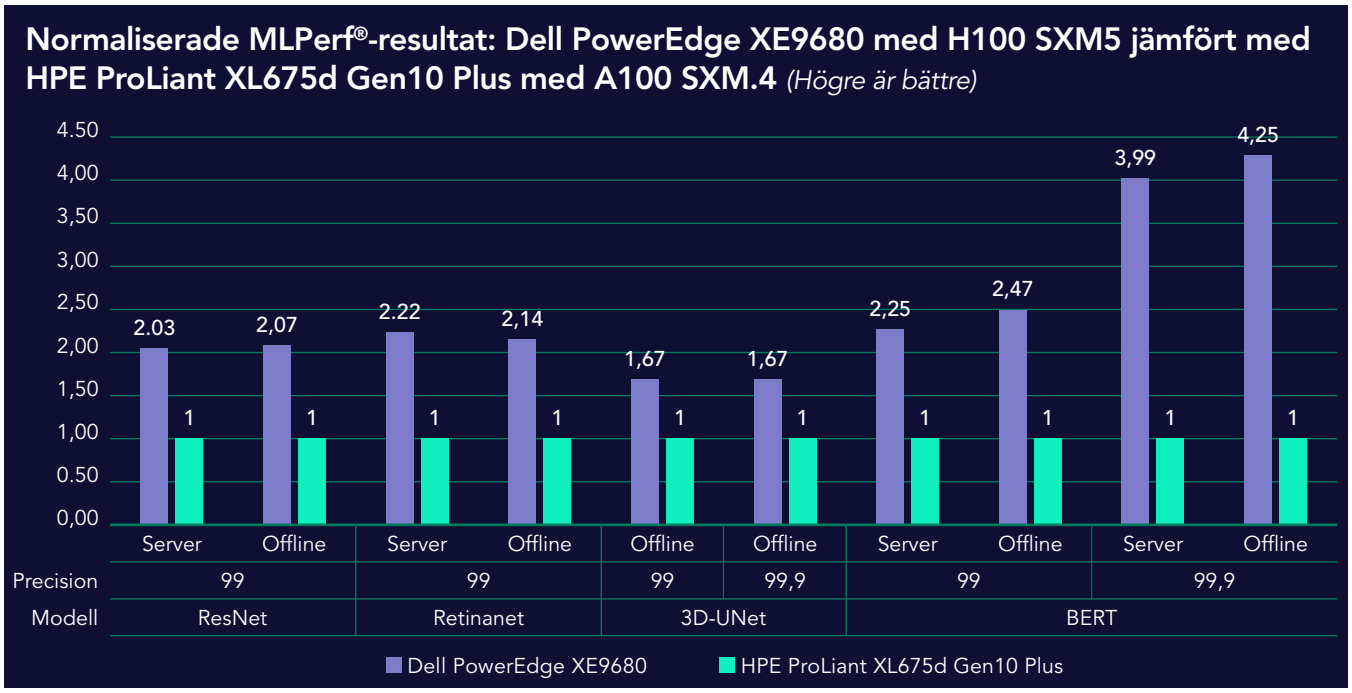



Bild 2: Publicerade MLPerf®-resultat för Dell PowerEdge XE9680 och HPE ProLiant XL675d Gen10 Plus från och med den 29 november 2023. Dells system använder NVIDIA H100 GPU, medan grafikprocessorerna i HPE-systemet är en generation äldre. Källa: Principled Technologies med data från MLCommons®.[20,21]

För att underlätta jämförelsen har vi normaliserat testresultaten i bilderna 2 till 5. Det innebär att vi tilldelar värdet 1 till varje HPE ProLiant DL380a Gen 11-resultat och visar motsvarande Dell PowerEdge R760xa-resultat i förhållande till det. Som dessa resultat visar kan även en generations skillnad mellan GPU-modeller göra en betydande skillnad i den prestanda du kan förvänta dig att se över en mängd olika AI-arbetsbelastningar.

## MLPerf-resultat med fyra GPU:er

När det är viktigt att spara energi i datacentret eller utrymme kan 2U Dell PowerEdge XE9640 ge svaret. Med upp till fyra NVIDIA H100 SXM GPU:er erbjuder PowerEdgeXE9640 hälften av GPU-beräkningskraften hos XE9680, på två tredjedelar mindre utrymme.[22] Den tätt packade Dell PowerEdge XE9640 har Dell Smart Cooling-teknik med en rad olika termiska tekniker, inklusive direkt vätskekylning för processorer och GPU:er.[23]

2U-chassit på PowerEdge XE9640 rymmer förbättrade luftflödesmekanismer, större fläktar och kylflänsar för att hjälpa till att kyla andra viktiga komponenter som PCIe-kort och minne.[24] PowerEdge XE9640 är för närvarande det enda erbjudandet från antingen Dell eller HPE som levereras med fyrvägs HGX H100-GPU:er på 2U. HPE AI-portföljen erbjuder 1U och 2U Gen11 ProLiant-servrar, men de är begränsade till GPU:er med PCIe-formfaktor.[25]

Dell PowerEdge XE9640-servern har även stöd för Intel Max GPU Series 1550 OAM GPU:er, som ger ett GPU-alternativ med låg effekt och hög densitet som inkluderar ett PCIe-kort och en OAM (OpenCompute Accelerator Module).[26] Även om vi inte kunde fastställa att HPE från och med 5 december 2023 erbjöd en server med dessa GPU:er erbjuder de HPE ProLiant DL380 Gen11 och DL380a Gen11 med Intel data Center GPU Max 1100 GPU:er.[27] Det innebär att Dell PowerEdge XE9640 kan vara det enda aktuella erbjudandet med fyra Intel Max 1550 OAM GPU:er i en 2U-server. För företag som är intresserade av datacenterutrymme och energieffektivitet erbjuder en 2U-server med fyra Intel Max 1550 GPU:er en lösning som kombinerar högpresterande beräkning och energieffektivitet utan att offra datacenterutrymme.

Dell PowerEdge XE9640 med fyra HGX H100 GPU:er överträffade HPE ProLiant DL380a med fyra PCIe H100 GPU:er med upp till 1,99 gånger i de publicerade MLPerf® 3.1-resultaten (se bild 3).

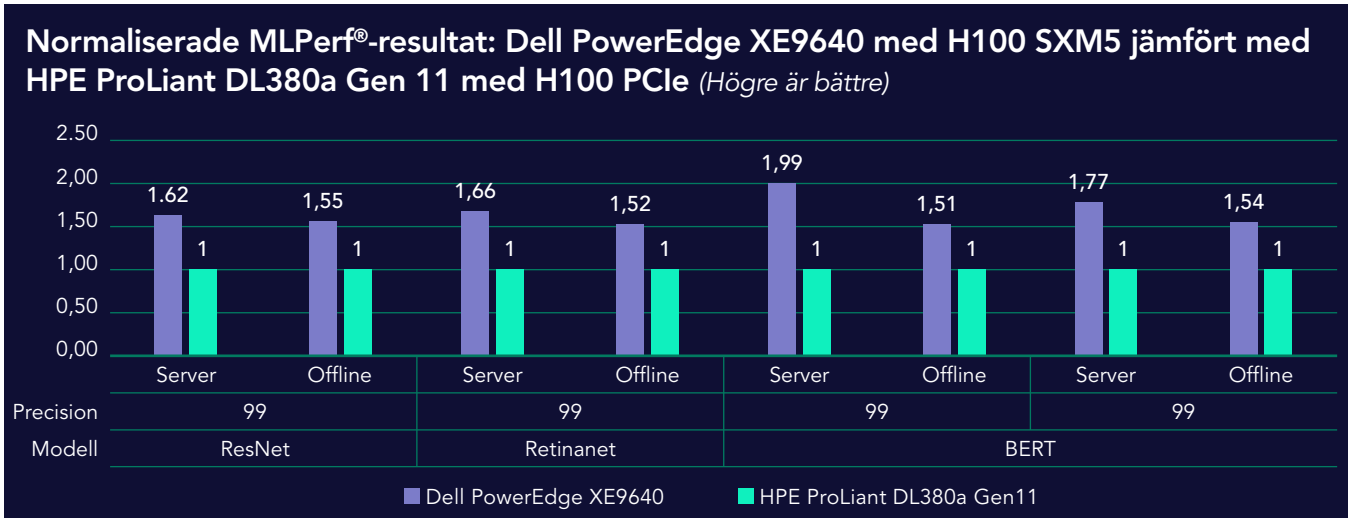



Bild 3: Publicerade MLPerf®-resultat för Dell PowerEdge XE9680 och HPE ProLiant XL675d Gen10 Plus från och med den 29 november 2023. Dells system använder NVIDIA H100 GPU, medan grafikprocessorerna i HPE-systemet är en generation äldre. Källa: Principled Technologies med data från MLCommons®.[28,29]

PowerEdge XE8640 erbjuder en fyrvägs GPU-konfiguration med luftkylning för processorer och en kylare med vätskeassisterad luftkylning för GPU:er som inte kräver vatten från anläggningen. För de som inte använder eller inte kan använda extern kylning[30] har 4U Dell PowerEdge XE8640 stöd för fyra NVIDIA H100 SXM5-grafikprocessorer som ger samma beräkningskraft som PowerEdge XE9640 utan behov av direkt vätskekylning.[31]

Dell PowerEdge XE8640 har den senaste 4:e generationens skalbara Intel Xeon-processorer och upp till 4 TB minne[32] för att hantera stora datauppsättningar och komplexa beräkningar som är vanliga inom AI och dataanalys. Återigen erbjuder HPE NVIDIA H100 SXM5 GPU:er i HPE Cray-system, men HPE ProLiant GPU-aktiverade servrar stöder inte det.

Jämfört med MLPerf®-data som publicerades i november 2023 uppnådde PowerEdge XE8640-servern med fyra NVIDIA H100 SXM5-GPU:er den högsta AI-genomströmningen bland alla fyra GPU-bidrag i nio olika kategorier. Som bild 4 visar fick den upp till 2,07 gånger så hög poäng jämfört med HPE ProLiant DL380a-servern.

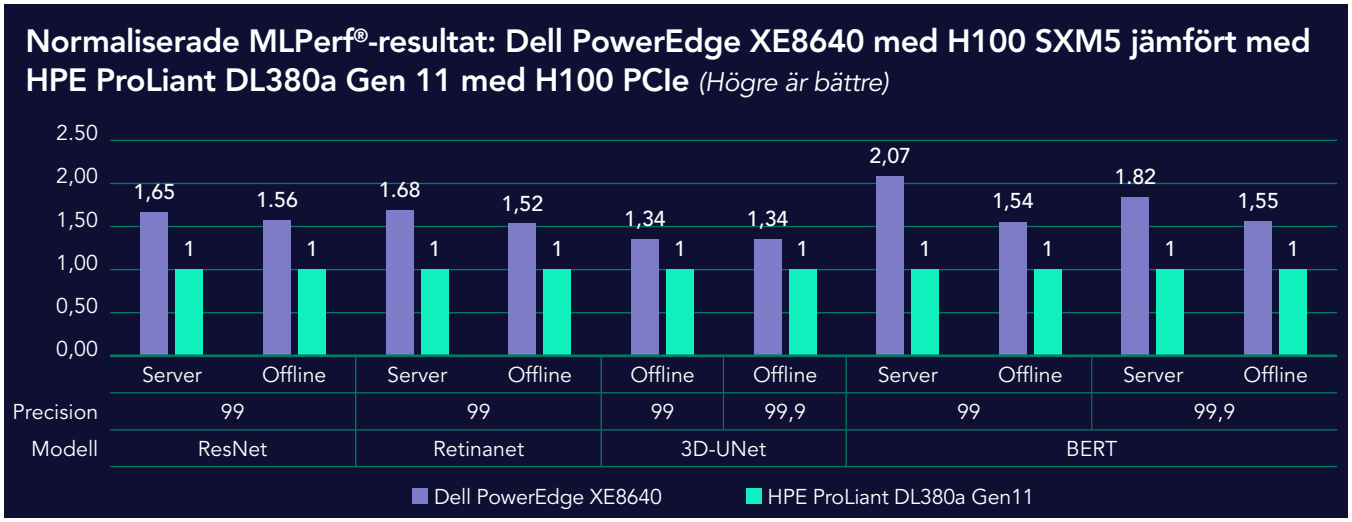



Bild 4: Publicerade MLPerf®-resultat för Dell PowerEdge XE8640 och HPE ProLiant DL380a Gen11 från och med den 29 november 2023. Dell-systemet använder NVIDIA H100 SXM-formfaktorn, medan HPE-systemet använder den mindre kraftfulla PCIe-formfaktorn. Källa: Principled Technologies med data från MLCommons®.[33,34]

För företag som kanske vill börja i en mindre skala och utöka efter behov har 2U Dell PowerEdge R760xa-servern stöd för en rad GPU:er från NVIDIA, AMD och Intel, med stöd för upp till fyra PCIe Gen 5 GPU:er med dubbel bredd eller 12 PCIe GPU:er med enkel bredd.[35] Den har 32 DIMM-kortplatser, ett fack för åtta enheter för 2,5-tumsdiskar och 12 PCIe-kortplatser. Det ger en skalbar lagring som kan växa med ökande AI-datakrav och stöd för upp till 12 PCIe-grafikprocessorer med enkel bredd eller fyra PCIe-GPU med dubbel bredd som NVIDIA H100 eller L40S.[36] Den här skalbarheten innebär att servern kan anpassa sig till nya AI-uppgifter, från träning av maskininlärningsmodeller till avancerad databehandling.

Luftkylningssystemet i PowerEdge R760xa har stöd för beräkningsmiljöer med hög densitet och kan rymma acceleratorer med högre termisk designeffekt (TDP) upp till 350 W,[37] en kapacitet som kan hjälpa IT-avdelningen att bibehålla prestanda under intensiva beräkningsbelastningar. I testresultaten för MLPerf® ResNet, RetinaNet och BERT Server som publicerades i november 2023 med ”server”-läget överträffade PowerEdge R760xa med fyra NVIDIA H100 PCIe-GPU:er HPE ProLiant DL380a Gen 11 även utrustad med fyra H100 PCIe-GPU:er (se bild 5).

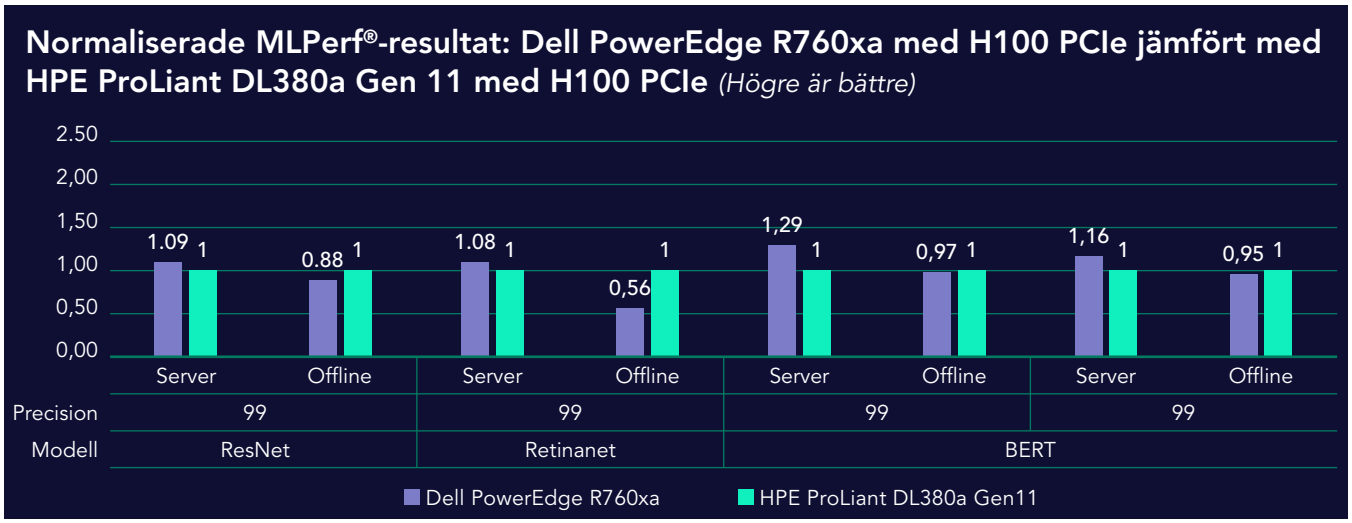



Bild 5: Publicerade MLPerf®-resultat för Dell PowerEdge R760xa och HPE ProLiant DL380a Gen11 från och med den 29 november 2023. Båda systemen använder PCIe-formfaktorn i NVIDIA H100-grafikprocessorerna. Källa: Principled Technologies med data från MLCommons®.[38,39]

Sammantaget visar MLPerf®-resultaten att prestandan varierar mycket mellan servrar och komponenter. Därför är det viktigt att välja rätt alternativ för att stödja dina arbetsbelastningar och deras prestandakrav. Dell PowerEdge-servrar för AI-arbetsbelastningar erbjuder flera alternativ för kylning och densitet för att passa alla datacenterbehov som ett företag kan ha samtidigt som de ger stark MLPerf®-prestanda.

## Mer detaljerad täckning av Dells AI-portfölj

Även om beräkningsprestanda är viktigt är det bara en av alla saker att tänka på när du planerar dina AI-arbetsbelastningar. Du måste också överväga resten av AI-portföljen som en leverantör erbjuder när du påbörjar en AI-implementering. Nedan tar vi upp ytterligare kategorier som är avgörande för dessa AI-portföljer, bland annat klient-workstation, molnbaserade produkter, lagring med mera. Vi lyfter också fram områden där erbjudanden från Dell kan ge en fördel jämfört med HPE.

### Workstations

För AI-utvecklare och dataforskare erbjuder Dell Precision data Science workstations NVIDIA RTX™ GPU:er och Intel Xeon®-processorer, tillsammans med en svit av datavetenskapliga verktyg.[40] Dessa system använder professionella beräkningsalternativ med NVIDIA GPU:er som är certifierade för över 100 professionella program[41] och Intel Xeon skalbara processoracceleratorer som Intel DL Boost.[42] Precision-workstations finns i mobila format, torn- och rackformat för att tillgodose behov som sträcker sig från större, stationära dataanalyser till vetenskapliga fältmodeller på språng.

HPE erbjuder workstation-lösningar som är smalare, främst med enkla workstation-torn utrustade med NVIDIA L4s, men HPE erbjuder inget mobilt workstation-alternativ.[43] Även om de är tillräckliga för många uppgifter ger deras tornerbjudanden för workstations inte samma flexibilitet och arbetsbelastningstäckning som det mer omfattande utbudet som Dell erbjuder. De olika alternativen för storlek och portabilitet i Dell Precision workstations möjliggör mer skräddarsydda lösningar som tillgodoser olika behov i miljöer som laboratorier, kontor och fältverksamhet.

## Lagring

Lagring kan vara lika viktigt som beräkning för att köra AI-arbetsbelastningar. Mer data förbättrar AI-modellens noggrannhet, men lagring och hantering av enorma datauppsättningar kan vara en utmaning för många datacenters kapacitet. Eftersom modeller vanligtvis tränas med ostrukturerade data måste dessutom AI-förberedda lagringssystem kunna hantera många olika datatyper enkelt.[44] För att tillhandahålla kapacitet och skalning för AI-, ML- och DL-datauppsättningar erbjuder Dell PowerScale™-serien för fillagring och Elastic Cloud Storage (ECS) eller mjukvarudefinierad ObjectScale för objektlagring.

PowerScale All-Flash NAS-portföljen erbjuder kapacitetsalternativ från 3,84 TB upp till 720 TB rå kapacitet per nod, med klustrad All-Flash-kapacitet som når 186 PB rå kapacitet. Flexibiliteten och skalan hos PowerScale kan stödja en mängd olika kunder och AI-användningsfall.[45] När de är grupperade kan PowerScale F900 nå upp till 186 PB total rå lagringskapacitet.[46] Alla de tre All-Flash-baserade PowerScale-modellerna F200, F600 och F900 omfattar inlinedatakomprimering och deduplicering för att förbättra lagringseffektiviteten.[47] Varje PowerScale-lagringsmodell använder Dell OneFS™-filsystemet som använder policyer för nivåindelning av lagring för att prioritera de viktigaste data på de snabbaste nivåerna för optimering av arbetsbelastningar.[48] Dell erbjuder även OneFS-mjukvara på AWS-marknaden med Dell APEX File Storage for AWS. Kunderna kan utnyttja OneFS med sina AWS-beräkningsinstanser för en konsekvent användarupplevelse med samma funktioner som finns i OneFS-disksystem på plats.[49] HPE erbjuder offentlig molnintegrering för hybridlagringslösningar, men vi hittade inte något molnbaserat alternativ som Dell APEX File Storage for AWS bland deras erbjudanden.

Objektlagringsalternativ från Dell inkluderar Dell Enterprise Object Storage (ECS) som är ”specialbyggd för att lagra ostrukturerade data i offentlig molnskala”.[50] Tillsammans med inbyggd kompatibilitet med Amazon S3-objektlagring för hybridmolnfunktioner levererar ECS-lagringsnoder kapacitet på upp till 14 PB per rack.[51] HPE erbjuder även ostrukturerad lagring med alternativ för fil- och objektlagring, även om dess objektlagringserbjudande sker via ett samarbete med Scality. Kunderna kan köpa HPE Solutions for Scality från HPE.[52]

## Professionella tjänster

Dell erbjuder ett brett utbud av professionella tjänster, som rådgivning, dataförberedelse, distribution, support och hanterade tjänster för att stödja AI-distributioner. Företag som letar efter validerade arkitekturer och lösningar kan titta närmare på Dells validerade utformningar för AI som riktar in sig på specifika användningsfall för att få ut det mesta av utformningen och distributionen av AI-resurser. Dessa Dell-validerade AI-lösningar omfattar hårdvaru- och mjukvarupaket, konversationsbaserade AI-modeller, maskininlärningsåtgärder med mera. Genom att kombinera förkonfigurerade, specialbyggda lösningar med AI-relaterade tjänster erbjuder Dell en omfattande AI-lösning för alla typer av AI-behov. Dessa erbjudanden kan ge en snabbare och enklare väg till AI-framgång jämfört med att bygga ad hoc-lösningar.

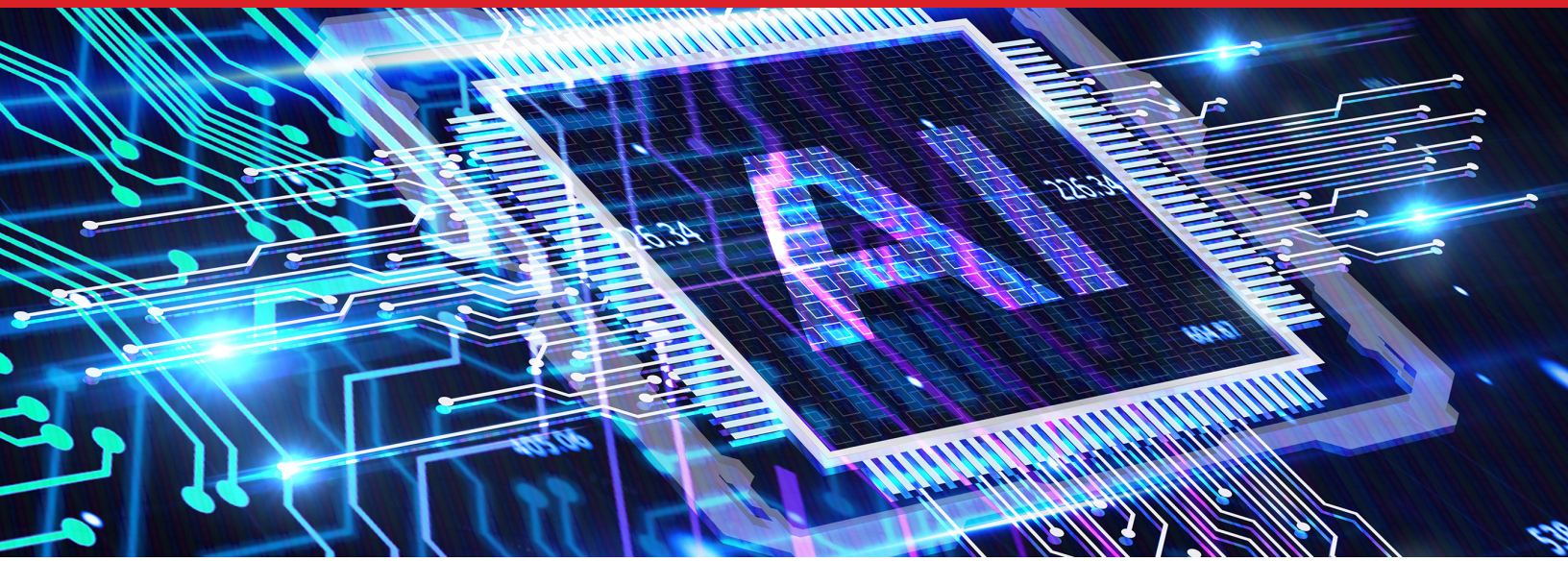
Dells tjänster kan också vägleda din AI-resa från rådgivning till implementering. Dell ProConsult Advisory Services hjälper kunderna att identifiera var användare kan dra nytta av att implementera GenAI-processer och skapa en handlingsplan med nödvändiga lösningar och IT-färdigheter. Dells tjänster kan förbereda data för integrering av stora språkmodeller och utbilda IT-team inom AI-kunskap. För fullständig GenAI-implementering granskar Dell Teams dina specifika användningsfall och fastställer, driftsätter och konfigurerar den bästa AI-modellen för att passa dina behov. HPE erbjuder också professionella tjänster för att stödja företag i deras AI-arbete.[53,54]

## Överväganden vid hantering

Servrar kräver löpande hantering som tar upp administratörernas tid. Fast mjukvara, mjukvara och drivrutiner behöver regelbundna uppdateringar, IT-personal måste optimera och bibehålla prestanda och temperaturer med mera. I tidigare Principled Technologies-tester (PT) utvärderade vi hanteringsfunktionerna hos Dell-servrar med Integrated Dell Remote Access Controller 9 (iDRAC9).[55] Administratörer kan lita på automatiserade onlineuppdateringar med iDRAC9 OpenManage™ Enterprise (OME) med konfigurerbar schemaläggning för att hålla servrarna uppdaterade och använda profiler för att snabbt och enkelt introducera nya servrar i takt med att arbetsbelastningarna växer. Med iDRAC och OME kan Dell-kunder få åtkomst till fler funktioner för fjärrhantering, driftsätta servrar enklare och uppdatera fast mjukvara enklare än de kunde med hjälp av HPE OneView och HPE ILO. Med Dell PowerEdge-servrar följer Dells hantering och tjänster som kan hjälpa organisationer genom att ”minska tid och ansträngning för uppgifter som övervakning av systemhälsa eller uppdatering av fast mjukvara”. Det frigör IT-cykler för innovation och andra uppgifter.[56]

Tabell 3: Sammanfattning av en jämförelse mellan hanteringsverktyg från Dell och HPE från en PT-rapport från november 2022.[57]
Källa: Principled Technologies.

| | Det som är annorlunda med Dells hanteringsverktyg | Så mycket bättre |
|---|---|---|
| **Fler fjärrhanteringsfunktioner**<br>iDRAC jämfört med ILO | Fler konfigurationsfunktioner för HTML5-konsol och BIOS för fler fjärrfunktioner i iDRAC | 2,5 gånger fler HTML5-konsolfunktioner och 13 gånger fler BIOS-funktioner |
| **Enklare serverdistribution**<br>OME jämfört med OneView | Profildistribution en-till-många med OME | 52 % kortare tid för att driftsätta en server än med OneView |
| **Enklare uppdateringar av fast mjukvara**<br>OME jämfört med OneView | Automatiserade onlineuppdateringar med OME | Uppdatera flera servrar genom att ansluta till Dell.com, vilket sparar in på den tid det tar att uppdatera servrar genom att manuellt ladda upp paket med OneView |
| **Enklare varningar**<br>OME jämfört med OneView | Konfigurera varningspolicyer i OME och utför automatiserade åtgärder baserat på varningar | Automatisering av den här processen sparar tid och minskar risken för fel jämfört med att utföra åtgärder manuellt varje gång du får en varning i OneView |
| **Enklare att använda säkerhetsfunktioner (systemlåsning och dynamisk USB)**<br>iDRAC jämfört med ILO | Färre steg, kortare tid, inga omstarter med iDRAC | ¼ av stegen, 91 % kortare tid för systemlåsning |
| **Mer robusta analyser**<br>CloudIQ för PowerEdge jämfört med InfoSight | Anpassningsbara rapporter, fler hälsomätvärden för bättre administratörskontroll med CloudIQ för PowerEdge | Över 15 gånger fler mätvärden att välja mellan jämfört med InfoSight |



## Sammanfattning

Det kan vara en utmanande uppgift att utnyttja kraften i AI för att effektivisera och förbättra affärsverksamheten, med betydande konsekvenser för verksamheten. När tekniken utvecklas snabbare än någonsin är det viktigt att samarbeta med rätt leverantör för AI. Genom att välja ett företag som Dell som inte bara erbjuder en omfattande AI-portfölj utan även erbjuder tjänster inom planering, förberedelser, implementering och hantering, kan kunderna ta sig an dessa utmaningar. Prestandatestning av MLPerf® visar att erbjudanden i Dells AI-portfölj erbjuder konsekvent, stark prestanda för AI-arbetsbelastningar. Dell kan hjälpa företag att använda AI och alla dess fördelar genom att erbjuda högpresterande och flexibla serveralternativ tillsammans med flera lagringsalternativ, validerade lösningar och professionella tjänster som är särskilt anpassade för AI.

---


1. Dell, "Increasing Your Data Value with Dell Generative AI Solutions", hämtat den 19 december 2023, https://www.dell.com/en-us/blog/increasing-your-data-value-with-dell-generative-ai-solutions/.
2. Dell, "Dell AI solutions", hämtat den 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", hämtat den 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", hämtat den 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", hämtat den 12 december 2023, https://www.dell.com/en-us/dt/solutions/artificial-intelligence/storage-for-ai.htm?hve=explore+unstructured+storage#tab0=0.
6. Dell, "Snowflake and Dell Partnership Gains Momentum", hämtat den 12 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", hämtat den 17 januari 2024, https://www.dell.com/en-us/blog/dell-vmware-and-nvidia-bring-ai-to-your-data/. Enligt länken ovan: "Baserat på Dells analys, augusti 2023. Dell Technologies erbjuder lösningar som är utformade för att stödja AI-arbetsbelastningar från workstation-datorer (mobila och stationära) till servrar för beräkning med hög prestanda, datalagring, molnbaserad och mjukvarudefinierad infrastruktur, nätverksswitchar, dataskydd, HCI och tjänster."
8. Intel, "Accelerate Artificial Intelligence (AI) Workloads with Intel Advanced Matrix Extensions (Intel AMX)", hämtat den 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", hämtat den 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", hämtat den 16 januari 2024, https://github.com/mlcommons/policies/blob/master/MLPerf_Results_Messaging_Guidelines.adoc.
11. MLCommons®, "MLPerf® Inference: Datacenter Benchmark Suite Results", hämtat den 12 december 2023, https://mlcommons.org/en/inference-datacenter-31/.
12. Dell, "PowerEdge XE Servers", hämtat den 12 december 2023, https://www.dell.com/en-us/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", hämtat den 12 december 2023, https://www.hpe.com/us/en/product-catalog/compute/proliant-servers/pip.1013142988.html.
14. HPE, "HPE ProLiant DL380a Gen11", hämtat den 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", hämtat den 12 december 2023, https://mlcommons.org/benchmarks/inference-datacenter/.
16. HPE, "NVIDIA Accelerators for HPE ProLiant Servers", hämtat den 12 december 2023, https://www.hpe.com/psnow/doc/c04123180.html?jumpid=in_pdp-psnow-qs.
17. Dell, "PowerEdge XE9680 Specification Sheet", hämtat den 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", hämtat den 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", hämtat den 12 december 2023, https://www.hpe.com/psnow/doc/a50004292enw.
20. Verifierad® MLPerf-poäng för v3.1 Inference Closed. Hämtad från https://mlcommons.org/benchmarks/inference-datacenter/ den 5 december 2023, post 3.1-0069. Namnet och logotypen MLPerf® är varumärken som tillhör MLCommons® Association i USA och andra länder. Med ensamrätt. Obehörig användning är strängt förbjuden. Besök www.mlcommons.org för mer information.
21. Verifierad® MLPerf-poäng för v3.1 Inference Closed. Hämtad från https://mlcommons.org/benchmarks/inference-datacenter/ 5 december 2023, post 3.1-0085. Namnet och logotypen MLPerf® är varumärken som tillhör MLCommons® Association i USA och andra länder. Med ensamrätt. Obehörig användning är strängt förbjuden. Besök www.mlcommons.org för mer information.
22. Dell, "PowerEdge XE9640 Rack Server", läst den 12 december 2023, https://www.dell.com/en-us/shop/ipovw/poweredge-xe9640.
23. Accelsius, "Enabling the AI Revolution with Liquid Cooling", hämtat den 12 december 2023, https://www.accelsius.com/blog/enabling-the-ai-revolution-with-liquid-cooling.
24. Dell, "Dell PowerEdge XE9640 Technical Guide", hämtat den 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", hämtat den 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", hämtat den 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", hämtat den 12 december 2023, https://www.hpe.com/psnow/doc/PSN1014779728WWEN.
28. Verifierad® MLPerf-poäng för v3.1 Inference Closed. Hämtat från https://mlcommons.org/benchmarks/inference-datacenter/ den 5 december 2023, post 3.1-0066. Namnet och logotypen MLPerf® är varumärken som tillhör MLCommons® Association i USA och andra länder. Med ensamrätt. Obehörig användning är strängt förbjuden. Besök www.mlcommons.org för mer information.
29. Verifierad® MLPerf-poäng för v3.1 Inference Closed. Hämtad från https://mlcommons.org/benchmarks/inference-datacenter/ 5 december 2023, post 3.1-0084. Namnet och logotypen MLPerf® är varumärken som tillhör MLCommons® Association i USA och andra länder. Med ensamrätt. Obehörig användning är strängt förbjuden. Besök www.mlcommons.org för mer information.
30. Dell, "AI and HPC —With Air or Liquid Cooling", hämtat den 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", hämtat den 12 december 2023, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe8640-spec-sheet.pdf.
32. Dell, "PowerEdge XE8640 Rack Server", hämtat den 12 december 2023, https://www.dell.com/en-us/shop/ipovw/poweredge-xe8640.
33. Verifierad® MLPerf-poäng för v3.1 Inference Closed. Hämtat från https://mlcommons.org/benchmarks/inference-datacenter/ 5 december 2023, post 3.1-0067. Namnet och logotypen MLPerf® är varumärken som tillhör MLCommons® Association i USA och andra länder. Med ensamrätt. Obehörig användning är strängt förbjuden. Besök www.mlcommons.org för mer information.
34. Verifierad® MLPerf-poäng för v3.1 Inference Closed. Hämtad från https://mlcommons.org/benchmarks/inference-datacenter/ 5 december 2023, post 3.1-0084. Namnet och logotypen MLPerf® är varumärken som tillhör MLCommons® Association i USA och andra länder. Med ensamrätt. Obehörig användning är strängt förbjuden. Besök www.mlcommons.org för mer information.
35. Dell, "PowerEdge R760xa Rack Server", hämtat den 12 december 2023, https://www.dell.com/en-us/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", hämtat den 12 december 2023, https://www.sanstorageworks.com/PowerEdge-R760xa.asp.
37. Dell, "Dell PowerEdge Servers and NVIDIA GPUs", hämtat den 12 december 2023, https://infohub.delltechnologies.com/l/design-guide-generative-ai-in-the-enterprise-inferencing/dell-poweredge-servers-and-nvidia-gpus-1/.
38. Verifierad® MLPerf-poäng för v3.1 Inference Closed. Hämtat från https://mlcommons.org/benchmarks/inference-datacenter/ 5 december 2023, post 3.1-0064. Namnet och logotypen MLPerf® är varumärken som tillhör MLCommons® Association i USA och andra länder. Med ensamrätt. Obehörig användning är strängt förbjuden. Besök www.mlcommons.org för mer information.
39. Verifierad® MLPerf-poäng för v3.1 Inference Closed. Hämtad från https://mlcommons.org/benchmarks/inference-datacenter/ 5 december 2023, post 3.1-0084. Namnet och logotypen MLPerf® är varumärken som tillhör MLCommons® Association i USA och andra länder. Med ensamrätt. Obehörig användning är strängt förbjuden. Besök www.mlcommons.org för mer information.
40. Dell, "Workstations for AI", hämtat den 12 december 2023, https://www.dell.com/en-us/dt/ai-technologies/index.htm?hve=explore+dell+precision+for+ai#pdf-overlay=//www.delltechnologies.com/asset/en-us/products/workstations/briefs-summaries/ai-industry-brochure.pdf.
41. NVIDIA, "NVIDIA RTX in Professional Workstations", hämtat den 12 december 2023, https://www.nvidia.com/en-us/design-visualization/desktop-graphics/.
42. Intel, "Intel® Deep Learning Boost (Intel® DL Boost)", hämtat den 12 december 2023, https://www.intel.com/content/www/us/en/artificial-intelligence/deep-learning-boost.html.
43. HPE, "HPE ProLiant ML350 Gen11", hämtat den 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", hämtat den 12 december 2023, https://www.computerweekly.com/feature/Storage-requirements-for-AI-ML-and-analytics-in-2022.
45. Dell, "PowerScale AI-Ready Data Platform", hämtat den 12 december 2023, https://www.dell.com/en-us/shop/powerscale-family/sf/powerscale.
46. Dell, "Compare PowerScale", hämtat den 12 december 2023, https://www.dell.com/en-us/shop/powerscale-family/sf/powerscale#compare-module.
47. Dell, "Dell PowerScale All-Flash", hämtat den 12 december 2023, https://www.delltechnologies.com/asset/en-us/products/storage/technical-support/h15963-ss-powerscale-all-flash-nodes.pdf.
48. Dell, "Dell PowerScale OneFS Software Features, hämtat den 12 december 2023, https://www.delltechnologies.com/asset/en-us/products/storage/technical-support/h18275-onefs-software-features-data-sheet.pdf.
49. Dell, "Dell APEX File Storage for AWS", hämtat den 12 december 2023, https://www.delltechnologies.com/asset/en-us/products/storage/briefs-summaries/h19575-so-apex-file-storage-for-aws.pdf.
50. Dell, "Dell ECS Enterprise Object Storage", hämtat den 12 december 2023, https://www.dell.com/en-us/dt/storage/ecs/index.htm?hve=explore+ecs#tab0=0&tab1=0.

51. Dell, "Dell ECS Enterprise Object Storage", hämtat den 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", hämtat den 12 december 2023, https://www.hpe.com/us/en/storage/file-object/scality.html.
53. HPE, "Make AI work for you", hämtat den 16 januari 2024, https://www.hpe.com/us/en/solutions/ai-artificial-intelligence.html.
54. HPE, "HPE AI Services – Generative AI Implementation", hämtat den 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", hämtat den 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", hämtat den 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."



▶ Se den ursprungliga, engelska versionen av den här rapporten på https://facts.pt/zPmSx4c

Det här projektet har beställts av Dell Technologies.

Principled Technologies®

Facts matter.®