White Paper

Varför är det en god ide att utvecka och implementera AI-teknik på arbetsstationer?

Sponsored by: Dell Technologies

Peter Rutten Dave McCarthy
July 2023

IDC:S ÅSIKT

AI har tagit fart som en viktig, differentierande funktion i alla branscher, och den hårdvara som krävs för att köra AI utvecklas snabbt. Inom teknikbranschen fokuserar man ofta på den exponentiella tillväxt sosm de mest avancerade AI-modellerna genomgår. Diskussionerna handlar om tiotals miljarder parametrar, minskad precision, expanderande minne, beräkningar med hög prestanda (HPC)-liknande behov av AI-inlärning och inferens, samt rack med snabbare servrar. I verkligheten är AI-datoranvändning i en sådan extraordinär skala ovanlig, i synnerhet inom företag.

Idag arbetar många företag hårt med AI-initiativ, inklusive generativ AI som inte kräver en superdator. Faktum är att en stod del av AI-utvecklingen – och i allt högre grad AI-distribution, särskilt vid kanten – sker på kraftfulla workstation-datorer. Workstation-datorer har många fördelar när det gäller utveckling och distribution av AI. De ser till att AI-forskaren och AI-utvecklare slipper förhandla om servertid, de ger GPU-acceleration även då serverbaserade GPU:er fortfarande inte är lättillgängliga i datacentret, de är extremt prisvärda i jämförelse med servrar och de representerar en mindre engångskostnad, snarare än en snabbt ackumulerande faktura som i fallet med en molninstans. Dessutom ger de användarna den trygghet det innebär att veta att känsliga data lagras på plats. Workstation-datorerna gör det också möjligt för forskarna och utvecklarna att experimentera med olika AI-modeller utan att behöva bekymra sig över kostnadsökningar.

IDC menar att AI-distribution i kanten kommer växa snabbare än AI-distribution på plats eller i molnet. Även här spelar workstation-datorer en allt viktigare roll som AI-inferensplattformar, eftersom de inte ens kräver GPU:er, utan utför inferens på mjukvaruoptimerade processorer. Användningsfallen för AI-inferens vid kanten på workstation-datorer växer snabbt och omfattar AIOps, katastrofinsatser, radiologi, olje- och gasprospektering, markförvaltning, telehälsa, trafikledning, övervakning av tillverkningsanläggningar samt drönare.

Det här informationsdokumentet tittar på den allt större roll som workstation-datorer spelar inom utveckling och distribution av AI, och behandlar kortfattat Dells portfölj av workstation-datorer för AI.

July 2023, IDC #US51017223

LÄGESÖVERSIKT

AI-explosionen och dess inverkan på infrastrukturen

Antalet AI-projekt som organisationer världen över är engagerade i växer snabbt. Redan idag utförs, inom alla branscher, en mängd uppgifter av mjukvara som helt eller delvis drivs av en AI-modell. IDC spårar AI på många nivåer, och ett mått som är användbart här är det belopp som företag och molntjänstleverantörer förväntas spendera på servrar för att kunna utveckla och driva AI. År 2026 kommer förväntas detta belopp vara 34,6 miljarder dollar, vilket motsvarar nära 22 % av de totala serverutgifterna världen över.

Men servrar utgör inte hela bilden. En hel del AI-relaterade förberedelser, utveckling, prototypframställning och, i allt högre grad, *distribution* sker på workstation-datorer. I och med att organisationer, små som stora, har upptäckt att nya affärsmöjligheter kan realiseras genom att förse programmen med AI-funktioner, har experimenterandet med AI-modeller skjutit i höjden. De robusta workstation-datorernas omedelbara tillgänglighet och närhet till data gör dem idealiska för detta syfte.

Hur blev AI plötsligt så utbredd, med tanke på att AI-algoritmer har använts i årtionden? Det beror främst på att två avgörande villkor för att driva en särskilt framgångsrik typ av AI-algoritm, det neurala nätverket, har realiserats under de senaste åren: tillgången till stora, billiga och olika typer av data, såsom ostrukturerade och semistrukturerade data, samt förstärkningen av linjär beräkning med en parallell modell för att bearbeta dessa neurala nätverk inom en acceptabel tidsram. Med dessa två grundläggande villkor uppfyllda har dataforskare gjort enorma framsteg med att utveckla neurala nätverk som automatiskt lär sig hur man utför allt mer imponerande uppgifter. Medan traditionell maskininlärning (ML) fortfarande är relevant för textbaserade eller numeriska data, är djupinlärning (DL) mer effektivt för video, ljud, språk och så vidare.

Traditionella maskininlärningsmodeller kan vanligtvis utvecklas på en workstation-dators processorer, som har upp till flera dussin kärnor. Men neurala nätverk kräver samprocessorer för att kunna parallellisera sin bearbetning över tusentals kärnor. Den främsta anledningen till detta är att inom ML är funktionsextraktion och klassificering en manuell process. I DL, däremot, är denna process automatiserad, vilket kräver att modellen tränas genom konstant upprepning med hjälp av stora datamängder. I dagsläget är den vanligaste samprocessorn GPU, men nya AI-specifika processorer utvecklade av nystartade företag kommer också finnas tillgängliga. Denna typ av acceleration, där en separat samprocessor för parallell bearbetning används, har revolutionerat marknaderna för servrar och workstation-datorer, och gett upphov till vad IDC kallar massivt parallell beräkning.

År 2022 utgjorde snabba servrar en världsomspännande marknad på 21,8 miljarder US-dollar. År 2026 förväntas den har växt till 43,4 miljarder US-dollar, varav 57 % representerar snabba servrar avsedda att köra AI. Samtidigt ökade antalet separata GPU:er som såldes för användning i workstation-datorer till 6,4 miljoner år 2022. IDC uppskattar att marknaden för workstation-datorer som används för vetenskapliga eller programvarutekniska ändamål, som i allt högre grad drivs av AI-utveckling, kommer att ha ökat till nära 2 miljarder US-dollar år 2026.

AI-utvecklingens olika stadier

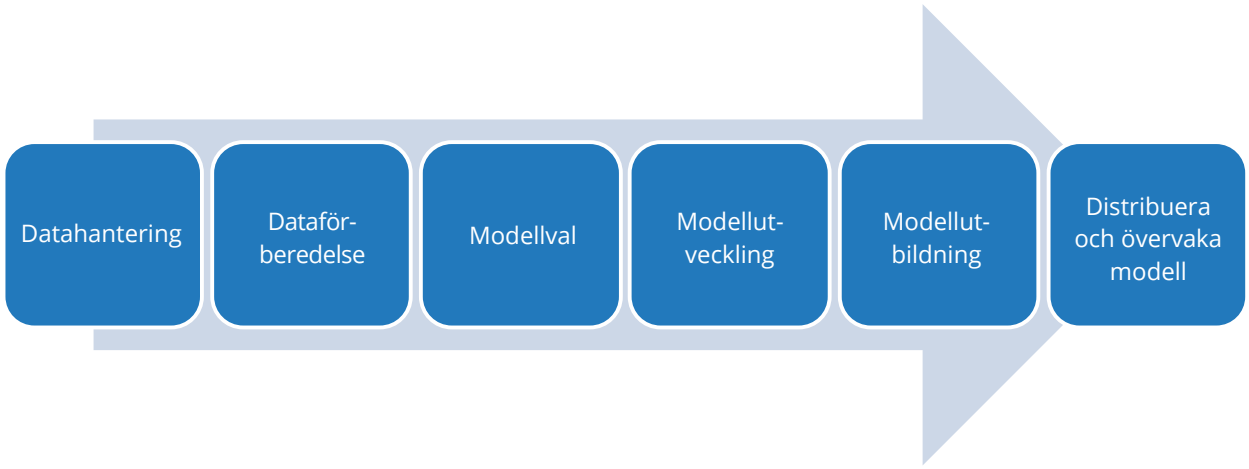
Som tidigare nämnts blev neurala nätverk genomförbara på grund av expanderande datatyper och datavolymer, samt nya metoder för beräkning. Den första delen av denna ekvation, datavolymer och datatyper, är inte försumbar: enligt vissa bedömningar utgörs hela 80 % av ett AI-baserat djupinlärningsinitiativ av datahantering och förberedelser. Data måste tas in, hanteras och förberedas innan modelldesign och utbildning kan påbörjas. Enligt IDC är följande utvecklingsstadier för AI (se bild 1):

- **Datahantering:** Identifiera och hantera relevanta data för AI-modellen från de enorma datavolymer i datacentret, kanten och molnet som en organisation tar emot, genererar och/eller skaffar (dessa data kan vara av vilken typ som helst, händelsedriven eller strömmad, och stora delar kan kräva någon typ av styrning.)
- **Dataförberedelse:** Lagring av data (fil, block eller objekt) i ett datalager eller en datasjö, rengöra dem, se till att de är kompletta och av hög kvalitet, samt därefter omvandla dessa data till en form som gör dem användbara för AI-modellen – till exempel med Spark eller verktyg som Pandas
- **Modellval:** Att bestämma vilken modell som bäst utför AI-uppgiften den är programmerad för med avseende på felfrekvens och/eller prestanda
- **Modellutveckling:** Utforma AI-modellen med ramverk såsom XGBoost, LightGBM, GLM, Keras, TensorFlow, PyTorch, Caffe, RuleFit, FTRL, Snap ML, scikit-learn eller H2O
- **Modellutbildning:** Utbilda modellen i beräkningsinfrastruktur med adekvata processor- och/eller samprocessorkärnor för parallellisering (allt oftare inbegriper detta förmågan att förklara, validera och dokumentera besluten i en modell för att säkerställa rättvisa, ansvarsskyldighet och transparens). (Detta inkluderar prototyper – testa den utbildade modellen genom att köra inferens på den.)
- **Distribuera och övervaka modellen:** Distribuera modellen i en produktionsmiljö så att den kan utföra uppgiften den är utformad för, vanligtvis kallad ”AI-inferens”, samt övervaka resultatet

Workstation-datorer kan spela en viktig roll i alla dessa sex steg, i kombination med datacenter, moln eller kantinfrastruktur.

BILD 1

AI-utvecklingens olika stadier



Källa: IDC, 2023

UTVECKLA AI-MODELLER PÅ WORKSTATION-DATORER

Workstation-datorer kontra persondatorer

Det är allmänt känt att persondatorer (PC) inte är tillräckligt kraftfulla för AI-utveckling. Dataforskare och AI-utvecklare är vanligtvis involverade i projekt som är strategiskt viktiga för deras organisationer, och därför är obehindrad produktivitet av yttersta vikt. Workstation-datorer tenderar att prestera mer förutsägbart än vad persondatorer gör, eftersom de vanligtvis är byggda av komponenter med högre prestanda och optimerade för mjukvaran som körs på dem.

Exempel på dessa nyckelkomponenter:

- **Högkvalitativa processorer:** Ett exempel är Intel Xeon skalbara processorer.
- **Kraftfulla GPU:er:** Ett exempel är NVIDIAs RTX professionella GPU:er såsom NVIDIA RTX 6000 Ada.
- **Mer lagring:** Vissa workstation-datorer kan leverera så mycket som 60 TB, och I/O-hastigheterna tenderar att vara betydligt högre än för persondatorer.
- **Mer minne:** Workstation-datorer med minne på hela 6 TB finns nu tillgängliga.
- **Nedkylning:** Högpresterande komponenter genererar mycket värme och dataforskare behöver en workstation-dator som har tillräckligt med kylning för att förhindra överhettning och bibehålla optimal prestanda.
- **NIC (Nätverksgränssnittskort):** Dataforskare som arbetar med stora datauppsättningar som lagras på fjärrservrar behöver ett höghastighets-NIC för att snabbt och effektivt kunna överföra data.
- **Skärm:** En högkvalitativ skärm är viktig för datavisualiseringsuppgifter, och dataforskare bör leta efter en stor skärm med hög upplösning och exakt färgprecision.

- **ECC-minne (Error Correcting Code):** ECC detekterar och korrigerar de vanligaste typerna av intern datakorruption, vilket förhindrar blå skärmar orsakade av antingen ett hårt fel (felaktig bit) eller ett mjukt fel (omvänd bit, som orsakar felaktiga värden). ECC säkerställer också noggrannhet i resultaten, vilket är helt avgörande inom områden såsom hälso- och sjukvård.
- **Specialiserat kisel:** Ett exempel är Intel Movidius visuell bearbetningsenheter (VPU:er). Dessa samprocessor för parallell bearbetning för datorseende- och kant-AI-program används i miljöer såsom detaljhandel, säkerhet och industriell automation. FPGA:er används också i workstation-datorer, till exempel för ekonomiprogram.
- **Optimeringsmjukvara:** Exempel inkluderar OneAPI, som är Intels standardbaserade programmeringsmodell avsedd att förenkla utveckling och distribution av datacentrerade arbetsbelastningar över CPU:er, GPU:er, FPGA:er och andra acceleratorer, eller CUDA, som är NVIDIA:s programprogrammeringsgränssnitt för att köra allmänna arbetsbelastningar på GPU:er.

CPU:er kontra GPU:er för AI

Workstation-datorer kan användas i olika stadier av AI-utvecklingen, och de är vanligtvis utrustade med en mängd olika funktioner. Trots betoningen på GPU:er för parallell bearbetning spelar processorerna en avgörande roll när man utvecklar en AI-modell på en workstation-dator. Precis som GPU:er kan också CPU:er användas för datamanipulation och, naturligtvis, för att utveckla traditionella ML-modeller. CPU: er används också för datautforskning, dvs. att använda visuella representationer av en datamängd för att förstå dessa datas egenskaper.

Inom DL-utbildning spelar värd-CPU:erna en något mindre viktig roll, eftersom GPU:erna tar över under själva utbildningsprocessen. Men även då fortsätter CPU:erna att fungera som bearbetningslager för kritisk mjukvara såsom OS eller CUDA, samt för orkestreringsprocesser bland GPU:er eller med annat kisel. CPU:erna har också i allt större utsträckning tagit på sig en ny roll som AI-inferensmotorer i de fall en workstation-dator används för att köra en AI-modell under produktion. IDC förväntar sig att 2024 års utgifter för AI-inferensinfrastruktur kommer överstiga utgifterna för AI-utbildningsinfrastruktur, och att en betydande andel (39 %) av denna inferens kommer att ske på värd-CPU:erna.

Workstation-datorer kontra servrar: ett symbiotiskt förhållande

De flesta organisationer ser pragmatism som tumregeln för när en workstation-dator, en lokal server, en molninstans eller en kombination av dessa tre, ska distribueras för AI-utveckling. Ett symbiotiskt förhållande råder mellan workstation-datorer, servrar och molninstanser, under de olika utvecklingsstadierna av ett AI-projekt.

Fördelen med workstation-datorer kontra datacenterservrar är att dataforskare kan arbeta var de vill – en viktig faktor under den pågående pandemin, men också under normala omständigheter. De kan också experimentera fritt på sina AI-modeller, och iterera så ofta de anser nödvändigt, eftersom moderna workstation-datorer, med deras kraftfulla GPU:er, tillåter att den iterativa processen blir mer interaktiv samt ger omedelbar feedback och resultat, utan att man behöver begära åtkomst till servrar eller stöter på andra datacenterbegränsningar. Workstation-datorerna ger också dataforskarna flexibiliteten att flytta datorn närmare data, istället för tvärtom, vilket sparar bandbredd, minskar nätverksträngsel och ökar genomströmningen. Dessutom kan workstation-datorer konfigureras för olika behov: till exempel traditionella ML-uppgifter eller mer DL-intensivt arbete.

Även om det nu sker en betydande tillväxt på marknaden för snabba servrar är snabba servrar fortfarande inte allmänt tillgängliga i företagsdatacenter. När detta informationsdokument skrevs var i genomsnitt 4 % av servrarna i företagsdatacenter snabba servrar, vilket innebär att många organisationer inte har möjlighet att utveckla eller köra AI på lättillgängliga GPU:er på plats. Också detta gör snabba workstation-datorer till ett användbart alternativ för AI-utveckling.

Snabba workstation-datorer är idag tillräckligt kraftfulla för att utföra DL-utbildning så länge AI-modellen inte är extremt stor, vilket eliminerar behovet av utbildning på servrar. Och modeller som utbildas på workstation-datorer med GPU:er kan distribueras antingen på workstation-datorer eller på servrar utan GPU:er, och därmed dra nytta av CPU:ernas inferenskapacitet. Mjukvarutekniker såsom Intels DL Boost och oneAPI kan driva AI-inferens på CPU:n, vilket gör det möjligt för icke-snabba servrar som redan är utplacerade i datacenter att stödja AI-program.

Workstation-datorer kontra molnet

Molnanvändning har revolutionerat hur organisationer tänker kring infrastruktur, data och program. Med löftet om nästan gränslös skalbarhet gör molnet det möjligt för utvecklare att provisionera resurser på begäran, vilket potentiellt kan öka innovationstakten med färre begränsningar. Vid första anblicken verkar molnet vara det perfekta paradigmet för AI-utveckling.

Detta är dock inte alltid fallet. Faktum är att IDC-forskning har visat att organisationer i allt större utsträckning drar tillbaka vissa arbetsbelastningar från det offentliga molnet till infrastruktur på plats. Detta beror av flera faktorer:

- **Molntillgänglighet:** Alla som har förlitat sig på molntjänster har upplevt avbrott, antingen orsakat av ett problem hos molnleverantören, eller på grund av att nätverksanslutningen har försvunnit någonstans mellan hyperskaladatacentret och slutanvändaren. I dessa situationer är användarna helt utlämnade till tjänsteleverantören för att lösa problemet och under tiden står produktiviteten stilla.
- **Säkerhet och efterlevnad:** Inom många branscher är det policyer för bolagsstyrning som dikterar var data kan kommuniceras och lagras, vilket begränsar användningen av molntjänster. Statliga regelverk såsom Dataskyddsförordningen (GDPR) i Europa och California Consumer Privacy Act innehåller också regler för datasuveränitet.
- **Kostnad:** Det är vanligt att organisationer underskattar hur snabbt molntjänstavgifterna kan stiga, i synnerhet för arbetsbelastningar som kräver beräkningar med hög prestanda och stort lagringsutrymme. Molnekonomi baseras på mätning av alla typer av resursförbrukning, inklusive data som tas ut och skickas tillbaka till infrastrukturen på plats.
- **Press att leverera:** De flesta AI-initiativ börjar med experiment och att modeller misslyckas är en del av utvecklingsprocessen. Detta kan dock leda till att AI-forskare och AI-utvecklare hamnar under en press då molnfakturorna stiger, utan att de har kunnat visa några körbara resultat ännu.

Workstation-datorer kan hantera dessa begränsningar samtidigt som de fortfarande utnyttjar molnbaserade tekniker som t.ex. mikrotjänstbaserade arkitekturer och API-driven automation. Detta möjliggör vissa av de fördelar man ser när man jämför workstation-datorer med datacenterservrar:

- **Arbeta var som helst:** Genom att ta bort beroendet av det offentliga molnet möjliggörs frånkopplade scenarier. Många högsäkerhetsmiljöer saknar offentliga nätverk, och AI-workstation-datorer kan på ett helt unikt sätt möta detta behov. Lokala resurser minskar också behovet av dyra nätverksanslutningar.

- **Datalokalitet:** Spridningen av IoT-enheter och annan ansluten utrustning bidrar till en exponentiell tillväxt av data på kantplatser. I många situationer är det vettigt att samlokalisera datorresurserna med en dedikerad workstation-dator. Genom att på så vis begränsa förflyttning av data kan man uppfylla många efterlevnadskrav.
- **Experimentera fritt:** Utbildningen och optimeringen av AI-modeller är en iterativ process, som ofta inkluderar ett inslag av att testa sig fram. Utvecklare behöver känna att de är fria att utföra experiment utan att behöva kompromissa på grund av det kan leda till ytterligare serviceavgifter. Workstation-datorer ger också mer flexibilitet för anpassade verktyg.

När det gäller den senare punkten är det relativt enkelt att jämföra priset på en workstation-dator med en molninstallation eftersom de flesta molntjänstleverantörer kan ge en omedelbar kostnadsuppskattning för alla konfigurationer som en slutanvändare vill distribuera. Till exempel är kostnaden för en vanlig virtuell maskin (VM) med en NVIDIA T4 och en instans av 375GiB SSD-lagring som används åtta timmar om dagen, fem dagar i veckan, 140 US-dollar en stor molnleverantör. Om man dubblar antalet VM, T4:or och SSD-enheter stiger kostnaden till 365 US-dollar per månad. Om man väljer en (1) VM, fyra T4:or och ett lagringsutrymme på 4 x 375 GiB, och kör en heltidsutbildning på miljön, går kostnaden upp till 2 700 US-dollar per månad. Molnkostnaderna för AI-utveckling kan med andra ord lätt stiga till tiotusentals dollar per år, betydligt mer än den årliga avskrivningen av en avancerad workstation-dator.

PROTOTYPFRAMSTÄLLNING AV AI PÅ WORKSTATION-DATORER

Jämfört med både servrar på plats och med molnet ger workstation-datorer en tydlig fördel när det gäller prototypframställning av AI-modeller. Det kan hända att servrarna i datacentret redan är i full användning eller att de är för uppdragskritiska för AI-prototypframställning och AI-testning. Dessutom, vilket har diskuterats tidigare, kan molninstanser snabbt leda till kostnadsöverskridanden när de används som en testmiljö. Workstation-datorer ser till att AI-forskaren eller AI-utvecklaren slipper förhandla om serveråtkomst och/eller slipper oroa sig över att molnfakturorna stiger under prototypframställningsstadiet. Workstation-datorernas låga engångskostnader möjliggör prototypframställning var som helst, när som helst. Utan några extra kostnader.

DISTRIBUERA AI-MODELLER PÅ WORKSTATION-DATORER

Att utveckla AI-modeller på en workstation-dator har varit en vanlig strategi under flera års tid, men nu ser IDC ett ökande antal användningsfall för att *distribuera* en AI-modell på en workstation-dator, vanligtvis vid kanten – med andra ord, att låta workstation-datorn köra inferens på AI-modellen. Kanten som en AI-distributionsplats för servrar växer snabbt – mellan 2020 och 2024 har de årliga hårdvaruutgifterna mer än tredubblats – och workstation-datorer ligger inte långt efter, när slutanvändarna nu börjar upptäcka fördelarna med att använda dem vid kanten.

IDC definierar kanten som ett distribuerat beräkningsparadigm som inkluderar distribution av infrastruktur och program som ligger utanför centraliserade moln och datacenter på plats, så nära den plats där data genereras och konsumeras som krävs. Här ingår distanskontor och filialer, samt branschspecifika platser såsom fabriker, lager, sjukhus och butiker.

Data- och datorintensiva arbetsbelastningar distribueras i allt högre grad antingen på plats eller på kantplatser. Detta görs för att mildra begränsningarna med offentliga moln, t.ex. den tid det tar att ladda upp stora datamängder och de rörliga kostnaderna för att genomföra AI-utbildning, i synnerhet i situationer som kräver en betydande mängd datavetenskapliga experiment.

IDC-forskning visar att kanten är ett snabbväxande distributionsscenario för AI. Under 2023 investerade organisationer 2,9 miljarder US-dollar i AI-beräkningar vid kanten, och denna siffra spås ha stigit till 6,9 miljarder US-dollar 2026 (se *Worldwide AI Hardware Forecast, 2022-2026: Strong Market Growth for AI Compute and Storage,* IDC #US49671722, september 2022). Dessutom blir kanten ett allt vanligare alternativ för distribution för HPC-arbetsbelastningar såsom konstruktion och teknik. I dagsläget investerar företag nästan 1 miljard US-dollar i sådana arbetsbelastningar vid kanten, och denna siffra spås ha stigit till 2,4 miljarder US-dollar 2027 (se *Worldwide High-Performance Computing Server Forecast, 2023-2027: Enterprise Will Overtake HPC Labs,* IDC #US50525123, april 2023). Inom dessa är det vettigt att distribuera en AI-workstation-dator.

När man distribuerar en AI-modell på en workstation-dator vid kanten, behövs inte alltid avancerade GPU:er, som i fallet med AI-utveckling. Lättare GPU:er kan utföra AI-inferens, och i ganska många fall behövs inga GPU:er överhuvudtaget. I dessa fall kan CPU:er utföra inferensuppgiften på ett adekvat sätt, i synnerhet när de används med optimeringar såsom Intel DL Boost, en uppsättning instruktionsfunktioner på Intels mikroprocessorer utformade för att påskynda AI-arbetsbelastningar, inklusive AI-inferens. Intel hävdar att Intel DL Boost ger 1,45 gånger högre INT8-inferensgenomströmning i realtid med 4:e generationens Intel Xeon skalbar processor som stöder Intel DL Boost, jämfört med tidigare generationer (BERT-Large SQuAD). Detta gör också en workstation-dator bättre lämpad för distribution vid kanten, där överväganden såsom strömförsörjning, mobilitet och värmehantering kräver färre watt. Här passar Intel Movidius Myriad (M2) bra, tack vare dess lilla energiavtryck på 12 W.

Användningsfall för distribution av AI på workstation-datorer

Det finns flera situationer som lämpar sig för att distribuera AI på lokalt distribuerade workstation-datorer. Vanligtvis handlar det om stora volymer maskingenererade tidsseriedata och ostrukturerade data som t.ex. videoströmmar och bilder. Det finns också tillfällen då ämnesexperter måste utöka AI-modellerna med mänsklig tolkning.

Exempel:

- **AIOps:** I takt med att IT-systemen växer i omfattning och komplexitet ökar behovet av att gå från reaktiv incidenthantering till proaktiv övervakning. Detta gäller särskilt då infrastruktur och program distribueras till kantplatser där det finns lite eller ingen teknisk personal. Genom att modellera en baslinje för normal prestanda kan man identifiera avvikelser och automatisera åtgärdssteg.
- **Katastrofinsatser:** I en nödsituation måste räddningspersonal snabbt kunna utvärdera en situation, spåra kritisk utrustning och sätta in resurser för att hjälpa de som behöver det mest. Detta måste ofta ske i en miljö utan nätverksanslutning, vilket kräver en lokal workstation-dator som kan aggregera dataflöden, inferera mot AI-modeller och automatisera kommunikation till nyckelpersonal.
- **Radiologi:** Framsteg inom bildteknik har lett till en ökning av storleken på data som genereras i skanning, vilket kräver att dessa data förblir på plats för att den ska kunna analyseras inom en rimlig tid. AI-modeller som har utbildats med hjälp av miljontals tidigare exempel kan identifiera mönster mer exakt än vad det mänskliga ögat kan, vilket ökar noggrannheten.
- **Prospektering av olja och gas:** Uppströmsbolag inom olja och gas använder en kombination av telemetridata, seismiska data och bilddata för att lokalisera naturresursreserver, välja borrplatser och optimera utrustningens prestanda under produktionsprocessen. Detta kräver ofta att man analyserar data i områden där endast dyr satellitkommunikation finns tillgänglig.

- **Cancerforskning och läkemedelsutveckling:** Forskare på forskningssjukhus och universitetssjukhus använder AI och naturlig språkbehandling för att hjälpa onkologer att fastställa den mest effektiva, skräddarsydda cancerbehandlingen för sina patienter. De kombinerar också maskininlärning med datorseende för att ge radiologer en bättre förståelse för hur patienternas tumörer förändras. Och de använder algoritmer för att bättre förstå hur cancer utvecklas och vilka behandlingar som fungerar bäst för att bekämpa dem.
- **Bedömningar av försäkringsskador:** Manuell hantering av fordringar är arbetskrävande och mänskliga misstag är vanliga. AI som kan utvärdera anspråkets giltighet minskar kostnaderna genom att låta personalen fokusera på de fall som kräver mer utredning. Det ökar den totala genomströmningen utan att göra avkall på noggrannheten.
- **Telehälsa:** AI förbättrar patienternas återhämtning genom att skräddarsy individuella behandlingsplaner baserat på vitala parametrar som kan avläsas i realtid från bärbara enheter. Denna information kombineras med historiska patientjournaler och en kunskapsdatabas om liknande fall. Detta är särskilt viktigt i landsbygdsområden som är mer beroende av hälso- och sjukvård på distans.
- **Säkerhet inom detaljhandeln (stöldskydd):** Realtidsanalyser som tillämpas på videoströmmar används för att förutsäga mänskligt beteende som kan leda till kriminell aktivitet. Detta kräver vanligtvis att flera videoflöden sammanfogas för att spåra en individs rörelser i en butik. Med tanke på att en väsentlig händelse av denna typ måste identifieras snabbt, är detta en process som bäst körs lokalt.
- **Trafikledning:** Statliga enheter som ansvarar för transportverksamhet använder allt oftare AI för att tillhandahålla samordning av trafikljus och digital skyltning, i syfte att förbättra flödet av fordon och hålla medborgarna säkra. Detta kräver en kombination av indata från bland annat videokameror och telemetri från vägsensorer för att optimera trafikmönster.
- **Övervakning av tillverkningsanläggningar:** För en fabrikschef är det av största vikt att säkerställa drifttiden för kritiska processer och uppfylla produktionsscheman. Detta översätts till förutsägbart underhåll av viktig utrustning, automatisk detektering av defekter och optimering både i och utanför anläggningens leverantörskedja. Det här är ett område där AI kan hjälpa operatörer öka prestandan samtidigt som de upprätthåller viktiga säkerhetsstandarder.
- **Drönare:** Automatisk analys av bilder tagna av drönare gör det möjligt att övervaka ett brett spektrum av förhållanden i en skala som tidigare inte varit möjlig. Det här har en betydande inverkan på inspektionen av gas- och elnätsinfrastruktur, försäkringsundersökningar, sök- och räddningsinsatser, precisionsjordbruk och underhåll av fiske och naturreservat.
- **Dagliga kontorsmiljöer:** Kontorsmiljöer idag har förbättrats avsevärt tack vare AI-baserade produktivitetsverktyg såsom Microsoft Copilot.
- **Förnybar energi:** Platser för förnybar energi, såsom vindkraftverk, vattenkraftsdammar och solkraftverk kräver övervakning, underhåll och datainsamling i realtid som måste genereras och analyseras lokalt.

DELL WORKSTATION-DATORER FÖR AI

Dell erbjuder ett brett utbud av workstation-datorer för olika nivåer av AI-utveckling och/eller implementering, allt under paraplyet Data Science Workstation (DSW). I det här avsnittet ges en kortfattad beskrivning av specifikationerna och dataforskares olika AI-roller/AI-program, samt fördelarna med Dell DSW-teknik. Dessa AI-förberedda datavetenskapsworkstation-datorer har utformats särskilt för dataforskare. De senaste Precision Data Science workstation-datorerna använder AI-funktioner för att optimera enheterna för de program som dataforskare använder mest. Det innebär att de kan slutföra sina viktigaste uppgifter snabbare. Dessutom testas och certifieras Dell Precision workstation-datorer av oberoende ISV:er för att säkerställa att de stöder de högpresterande program som Dells kunder behöver för att kunna utföra sina dagliga uppgifter.

Hur Dells workstation-datorer utmärker sig

Dell Precision workstation-datorer som drivs av NVIDIA RTX GPU:er är utformade för att ge den kraftfulla skalbarhet och prestanda som krävs för en organisations analyser och AI-initiativ. Dell Technologies levererar omfattande hårdvarulösningar optimerade för att köra branschens senaste AI-programvara:

- **Robust hårdvarukonfiguration:** Dell Precision workstation-datorer erbjuder en rad kraftfulla hårdvarukonfigurationer, inklusive flerkärniga processorer, högkapacitets-RAM och flera GPU-alternativ. Dessa komponenter tillhandahåller de nödvändiga beräkningsresurser som krävs för AI-uppgifter, genom att möjliggöra effektiv inlärning och inferens.
- **Skalbarhet och anpassningsbarhet:** Dell Precision workstation-datorer är skalbara och anpassningsbara, vilket innebär att användare kan skräddarsy hårdvarukonfigurationen efter sina specifika AI-krav. Denna flexibilitet säkerställer att workstation-datorn kan optimeras för specifika AI-arbetsbelastningsbehov.
- **Certifiering och optimering:** Dell samarbetar med NVIDIA för att certifiera Precision workstation-datorer för kompatibilitet och prestanda med NVIDIA RTX GPU, inklusive NVIDIA RTX 6000 Ada Generation-kort. Denna certifiering säkerställer sömlös integration och optimerad prestanda när du använder Dell Precision workstation-datorer med NVIDIA RTX GPU:er för AI-uppgifter.
- **Kraftfull bearbetningsförmåga:** Dell Precision workstation-datorer utrustade med Intel-processorer ger den beräkningskraft som behövs för AI-uppgifter. Med flerkärniga processorer och höga klockhastigheter levererar dessa workstation-datorer den prestanda som krävs för inlärning och inferens i AI-arbetsflöden.
- **Mjukvaru- och verktygssupport:** Dell Precision workstation-datorer levereras förinstallerade med mjukvara och verktyg som stöder AI-utveckling och AI-distribution. Detta inkluderar optimerade mjukvarustackar, AI-ramverk och bibliotek som drar fördel av NVIDIA RTX GPU:er, vilket gör det lättare för användare att komma igång med AI-projekt.

Teknikerna som diskuteras i avsnitten nedan är andra viktiga områdena där Dell workstation-datorer sticker ut.

Reliable memory technology

Dell tillhandahåller en teknik utöver ECC som kallas Reliable Memory Technology Pro (RMT Pro), som är utformad för att maximera drifttiden. Den arbetar tillsammans med ECC-minne för att detektera och korrigera minnesfel i realtid. Enligt Dell eliminerar RTM Pro praktiskt taget minnesfel genom att förhindra att defekta minnesområden återbesöks trots att DIMM-kortet är i full användning. Efter en omstart av systemet kommer RTM Pro att isolera det defekta minnesområdet och dölja det från operativsystemet. Som ett resultat kommer AI-dataforskare och AI-utvecklare kunna undvika fortsatta krascher, eftersom det defekta minnet kan fortsatt åtgärdas – en stor produktivitetsökning.

Dell Optimizer for Precision

Dell inkluderar även Dell Optimizer for Precision i de flesta av sina workstation-datorer, som automatiskt justerar systeminställningarna för att workstation-datorn ska köra olika populära företagsprogram i snabbast möjliga hastighet. Detta förbättrar en dataforskares eller datautvecklares produktivitet. Verktyget skapar också IT-prestandarapporter i realtid för processor, lagring, minne och grafikanvändning. DOP körs inte på Linux ännu och är därför mest användbart för att distribuera AI, eftersom AI-utveckling tenderar att göras med Linux-baserad mjukvara med öppen källkod. Dell Optimizer for Precision tillhandahåller även ExpressSign-in, Express Charge (på mobiler), Intelligent Audio, samt rapporterings- och analysverktyg för att finjustera workstation-datorn.

UTMANINGAR/MÖJLIGHETER

För företag

IDC ser en splittring på marknaden för AI. Å ena sidan använder företag datastrategier för att förbli konkurrenskraftiga, inklusive att komma igång med AI i stor skala. Till exempel ser de hur kollegor har utfört extraordinärt arbete med hjälp av företagsinriktad AI-infrastruktur som är registrerad bland de 100 bästa superdatorerna. Å andra sidan ser företag dagligen de små AI-initiativ som testas på tillgängliga servrar i datacentret eller i molnet, ofta med otillräcklig budget och underpresterande hårdvara.

För många företag är det första scenariot inte relevant och det andra alltför verkligt. För dem är utmaningen att ge sina AI-dataforskare och/eller AI-utvecklare rätt verktyg för att snabbt kunna utföra AI-utbildning, utan att spendera enorma summor pengar på molninstanser eller GPU-accelererade datacenterservrar. IDC anser att dessa företag är väl betjänta av att förse sina forskare och utvecklare med kraftfulla GPU-accelererade workstation-datorer.

För Dell

Det finns ett vanligt missförstånd på marknaden om att AI-utveckling och AI-distribution kräver dyr, snabb serverhårdvara, ofta även i ett kluster. Det kan vara sant för de största AI-algoritmerna, som har miljarder parametrar, men de flesta företag utvecklar inte så massiva algoritmer. De gör något användbart, effektfullt och hanterbart med sitt AI-initiativ, och många företag inser inte att sådana vanliga AI-modeller kan utvecklas – och distribueras – på workstation-datorer. Dells utmaning är att bryta igenom förutfattade meningar och upplysa marknaden om möjligheterna med deras workstation-datorer.

Samtidigt måste Dell se till att deras workstation-datorer levererar och inte blir tekniska flaskhalsar med tiden. Det innebär en snabb pågående innovation för att slutanvändare som använder workstation-datorerna på rätt sätt (dvs. som inte försöker köra en algoritm med flera miljarder parametrar) inte ska bli besvikna. Det innebär också att det för kunder som plötsligt börjar skala väldigt snabbt eller vars algoritmer faktiskt blir väldigt stora, sker en sömlös övergång från workstation-datorn till Dells AI-serversortiment. Däri ligger naturligtvis också möjligheten för Dell: att erbjuda rätt lösning för varje kund, oavsett vilken storlek på AI-initiativ de arbetar med.

SAMMANFATTNING

IDC anser att workstation-datorernas betydelse för AI-utvecklingen och distributionen av många användningsfall. Dessa datorer ger AI-forskare och utvecklare en kraftfull GPU-accelererad plattform som representerar lägre kapitalkostnader än servrar, dramatiskt lägre OPEX än molninstanser och mycket större frihet att experimentera med AI-modeller. Företag som utvecklar AI-initiativ som inte kräver en algoritm med flera miljarder parametrar (dvs. de allra flesta) bör överväga att ge sina AI-team workstation-datorer för obegränsad AI-utveckling och enkel kantbaserad distribution.

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