White Paper

Perché ha senso sviluppare e implementare la tecnologia IA sulle workstation

Sponsored by: Dell Technologies

Peter Rutten Dave McCarthy
July 2023

IL PARERE DI IDC

L'IA rappresenta oggi un'importante opportunità di differenziazione in tutti i settori e l'hardware richiesto per eseguirla è in rapida evoluzione. Il settore tecnologico spesso è molto focalizzato sulla crescita esponenziale delle dimensioni dei modelli di IA più avanzati. Le discussioni vertono su decine di miliardi di parametri, riduzione dei livelli di precisione, espansione della memoria, esigenze di addestramento e inferenza dell'IA simili a quelle dell'High Performance Computing (HPC), nonché su rack di server accelerati. In realtà, questa straordinaria scalabilità di elaborazione dell'IA rappresenta l'eccezione, soprattutto nell'azienda.

Oggi molte aziende si occupano di iniziative di IA, inclusa l'IA generativa, che non richiedono un supercomputer. Tuttavia, buona parte dello sviluppo dell'IA e, in misura sempre maggiore, del deployment dell'IA, soprattutto in corrispondenza dell'edge, viene eseguita su potenti workstation. Le workstation offrono numerosi vantaggi per lo sviluppo e il deployment dell'IA. Evitano agli scienziati e agli sviluppatori di intelligenza artificiale di negoziare l'ora del server, forniscono accelerazione delle GPU anche se le GPU basate su server non sono ancora facilmente accessibili nel data center, sono estremamente convenienti rispetto ai server e richiedono per un'istanza cloud un unico pagamento di entità ridotta, anziché costi continuativi che si accumulano rapidamente nel tempo, con la certezza che i dati sensibili siano archiviati on-premise in una posizione sicura. In questo modo, inoltre, evitano loro l'ansia dovuta all'accumulo dei costi anche solo in fase di sperimentazione di modelli di IA.

IDC rileva una crescita più rapida dell'edge come scenario di deployment dell'IA rispetto agli ambienti on-premise e cloud. Anche in questo caso, le workstation svolgono un ruolo sempre più vitale come piattaforme di inferenza dell'IA, spesso senza neppure richiedere l'uso di GPU, ma eseguendo l'inferenza su CPU ottimizzate per il software. I casi d'uso di inferenza dell'IA in corrispondenza dell'edge sulle workstation sono in rapido aumento e includono AIOps, risposta in caso di emergenza, radiologia, esplorazione di petrolio e gas, gestione del territorio, telemedicina, gestione del traffico, monitoraggio degli impianti di produzione e droni.

Questo white paper analizza il ruolo sempre più centrale delle workstation nello sviluppo e nel deployment dell'IA e accenna brevemente al portafoglio di workstation Dell per l'IA.

July 2023, IDC n. US51017223

PANORAMICA DELLA SITUAZIONE

Esplosione dell'IA e impatto sull'infrastruttura

Il numero di progetti di IA intrapresi dalle organizzazioni a livello globale è in rapido aumento. In tutti i settori, molte attività vengono già eseguite da software parzialmente o interamente basato su un modello di IA. IDC monitora l'IA su vari livelli: una metrica utile da considerare è l'importo di spesa previsto dalle aziende e dai provider di servizi cloud sui server per lo sviluppo e l'esecuzione dell'IA . Entro il 2026, la spesa prevista raggiungerà i 34,6 miliardi di dollari, equivalente a circa il 22% della spesa totale destinata ai server a livello globale.

Ma i server non sono gli unici elementi in gioco. Molte attività di preparazione, sviluppo, prototipazione e, in misura sempre maggiore, *deployment* avvengono sulle workstation. La sperimentazione di modelli di IA sta aumentando a dismisura man mano che organizzazioni di ogni dimensione scoprono che è possibile realizzare nuove opportunità di business integrando alcune funzionalità di IA nelle proprie applicazioni. Potenti workstation sono ideali a questo fine, grazie alla loro disponibilità immediata e alla prossimità ai dati.

Dal momento che gli algoritmi di IA vengono implementati da decenni, per quale motivo l'IA è diventata all'improvviso così predominante? Ciò è essenzialmente dovuto al fatto che, negli ultimi anni, sono state realizzate due condizioni tipiche per alimentare un tipo di algoritmo di IA particolarmente vincente come la rete neurale: l'elevata accessibilità a numerosi tipi di dati diversificati e a costi contenuti, tra cui dati strutturati e non strutturati, e il potenziamento dell'elaborazione lineare con un modello parallelo che consente di elaborare queste reti neurali in una frazione di tempo accettabile. Con queste due condizioni di base, i data scientist hanno realizzato enormi progressi nello sviluppo di reti neurali che apprendono automaticamente come eseguire attività sempre più determinanti. Se l'apprendimento automatico (ML) tradizionale rimane rilevante per i dati testuali e numerici, l'apprendimento approfondito (DL) è più efficace per video, audio, linguaggi e così via.

I modelli tradizionali di apprendimento automatico possono in genere essere sviluppati sulle CPU di una workstation, che includono al massimo alcune decine di core, ma le reti neurali richiedono coprocessori per parallelizzare l'elaborazione tra migliaia di core. Ciò è dovuto innanzitutto al fatto che, nell'apprendimento automatico, l'estrazione e la classificazione delle funzioni sono processi manuali, mentre nell'apprendimento approfondito sono automatizzati e prevedono l'addestramento del modello in modo costante e ripetitivo con data set di grandi dimensioni. A oggi, le GPU sono i coprocessori più comuni, ma iniziano a farsi strada anche nuovi processori specifici per l'IA sviluppati da start-up. Questo tipo di accelerazione con un coprocessore dedicato per l'elaborazione parallela ha rivoluzionato i mercati dei server e delle workstation, determinando l'ascesa di un fenomeno che IDC chiama elaborazione massicciamente parallela.

Nel 2022, i server accelerati rappresentavano un mercato globale di 21,8 miliardi di dollari, che dovrebbe raggiungere i 43,4 miliardi di dollari entro il 2026, con il 57% del totale costituito da server accelerati per l'esecuzione di IA. Al contempo, il numero di GPU dedicate vendute per l'uso nelle workstation ha raggiunto i 6,4 milioni nel 2022. Secondo le stime di IDC, il mercato delle workstation utilizzate per scopi scientifici o di progettazione software, sempre più basate sullo sviluppo di IA, si avvicinerà a 2 miliardi di dollari entro il 2026.

Fasi di sviluppo dell'intelligenza artificiale

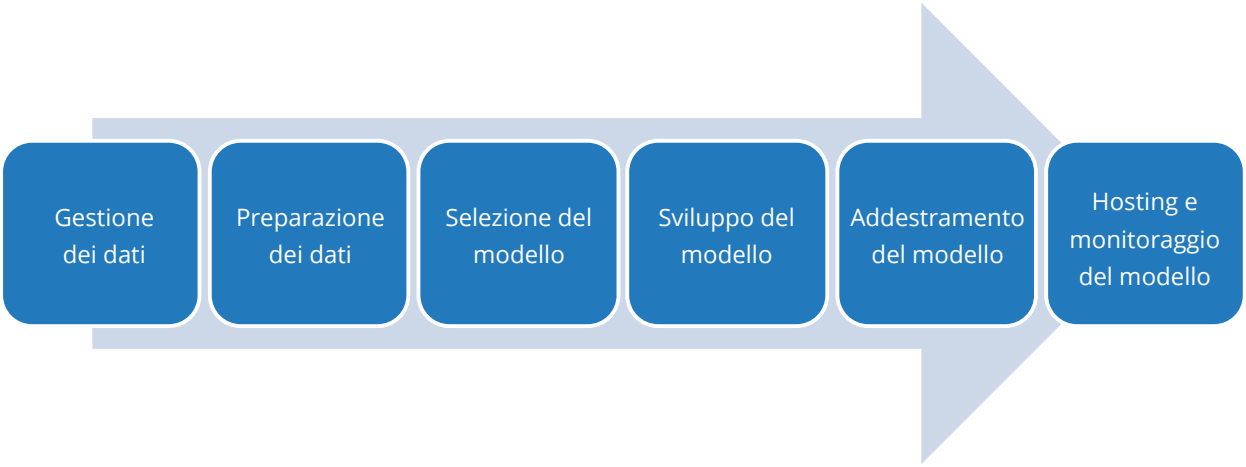
Come citato in precedenza, le reti neurali sono divenute percorribili per via dell'espansione dei tipi e dei volumi di dati, nonché di nuovi approcci all'elaborazione. La prima parte di questa equazione, ovvero i volumi e i tipi di dati, non è irrilevante. Secondo alcune stime, addirittura l'80% delle attività in un'iniziativa di IA con apprendimento approfondito riguarda la gestione e la preparazione dei dati. I dati devono essere acquisiti, gestiti e preparati prima della progettazione e dell'addestramento del modello. Queste sono le fasi di sviluppo dell'IA secondo IDC (figura 1):

- **Gestione dei dati:** identificazione e gestione dei dati pertinenti per il modello di IA dagli enormi volumi di dati nel data center, nell'edge e nel cloud che un'organizzazione acquisisce, genera e/o introduce (questi dati possono essere di qualsiasi tipo, guidati da eventi o in streaming, e molti possono richiedere una forma di governance).
- **Preparazione dei dati:** archiviazione dei dati (a livello di file, blocco oppure oggetto) in un data warehouse o un data lake, pulitura, verifica del fatto che siano completi e di alta qualità e successiva trasformazione al fine di renderli fruibili per il modello di IA, ad esempio con Spark o strumenti come Pandas.
- **Selezione del modello:** scelta del modello ideale per l'esecuzione dell'attività di IA per cui è stato programmato in termini di tasso di errore e/o prestazioni.
- **Sviluppo del modello:** progettazione del modello di IA mediante framework quali XGBoost, LightGBM, GLM, Keras, TensorFlow, PyTorch, Caffe, RuleFit, FTRL, Snap ML, scikit-learn o H2O.
- **Addestramento del modello:** addestramento del modello sull'infrastruttura di elaborazione con una combinazione sufficiente di processore e coprocessori per la parallelizzazione, includendo inoltre la possibilità di spiegare, convalidare e documentare le decisioni di un modello al fine di assicurare equità, responsabilità e trasparenza. Rientra in questo ambito anche la prototipazione, ovvero l'esecuzione di test sul modello addestrato tramite inferenza.
- **Hosting e monitoraggio del modello:** deployment del modello in un ambiente di produzione per eseguire l'attività per cui è stato progettato (processo in genere denominato "inferenza dell'IA") e monitoraggio delle prestazioni.

Le workstation possono ricoprire un ruolo importante in ognuna di queste sei fasi in combinazione con data center, cloud o infrastruttura edge.

FIGURA 1

Fasi di sviluppo dell'intelligenza artificiale



Fonte: IDC, 2023

SVILUPPO DI MODELLI DI IA SU WORKSTATION

Workstation e personal computer a confronto

È ormai chiaro che i personal computer (PC) non siano sufficientemente potenti per lo sviluppo dell'intelligenza artificiale. Data scientist e sviluppatori di IA sono in genere coinvolti in progetti importanti a livello strategico per le loro organizzazioni ed è della massima importanza evitare ostacoli alla produttività. Le workstation tendono ad avere prestazioni più prevedibili rispetto ai PC, in quanto vengono in genere sviluppate con componenti a prestazioni superiori e ottimizzate per il software che eseguono.

Questi componenti includono:

- **Processori di qualità superiore:** un esempio è rappresentato dai processori scalabili Intel Xeon.
- **GPU potenti:** un esempio è rappresentato dalle GPU professionali RTX NVIDIA, ad esempio NVIDIA RTX 6000 Ada.
- **Più storage:** alcune workstation offrono fino a 60 TB di storage e le velocità di I/O tendono a essere sensibilmente maggiori di quelle dei PC.
- **Più memoria:** alcune workstation includono ora fino a 6 TB di memoria.
- **Raffreddamento:** i componenti a prestazioni elevate generano una grande quantità di calore e i data scientist necessitano di una workstation con un livello adeguato di raffreddamento per evitare il surriscaldamento e mantenere prestazioni ottimali.
- **Scheda di rete:** per i data scientist che utilizzano data set di grandi dimensioni archiviati in server remoti, è essenziale una scheda di rete ad alta velocità per trasferire i dati in modo rapido ed efficiente.

- **Display:** un display di alta qualità è importante per svolgere attività di visualizzazione dei dati. I data scientist devono optare per un monitor con livelli elevati di risoluzione e precisione del colore e con uno schermo di grandi dimensioni.
- **Memoria ECC (Error Correction Code):** ECC rileva e corregge i tipi più comuni di danneggiamento dei dati interni, evitando schermate blu durante un'esecuzione prolungata di addestramento dell'IA in seguito a un errore irreversibile (bit non valido) o transitorio (bit invertito che genera valori errati); assicura inoltre precisione nei risultati, un requisito critico di importanza vitale, come il settore sanitario.
- **Silicio specializzato:** un esempio è rappresentato dalle VPU (Vision Processing Unit) Intel Movidius, coprocessori di elaborazione parallela per applicazioni di visione artificiale e IA nell'edge utilizzate in ambiti quali retail, sicurezza e automazione industriale. Nelle workstation vengono inoltre utilizzati dispositivi FPGA, ad esempio per le applicazioni finanziarie.
- **Software di ottimizzazione:** alcuni esempi includono OneAPI, il modello di programmazione basato su standard Intel per semplificare lo sviluppo e il deployment di carichi di lavoro incentrati sui dati tra CPU, GPU, FPGA e altri acceleratori, o CUDA, la piattaforma di elaborazione parallela e l'API NVIDIA per l'esecuzione di carichi di lavoro generici su GPU.

CPU e GPU per l'IA a confronto

Le workstation possono essere utilizzate in varie fasi dello sviluppo dell'IA e sono in genere in grado di eseguire una serie di funzionalità. Nonostante l'enfasi sulle GPU per l'elaborazione parallela, le CPU ricoprono un ruolo critico in fase di sviluppo di un modello di IA su una workstation. Esattamente come le GPU, le CPU possono essere utilizzate anche per la manipolazione dei dati e, naturalmente, per lo sviluppo di modelli di apprendimento automatico tradizionali. Le CPU vengono utilizzate anche per l'esplorazione dei dati, il processo di utilizzo di rappresentazioni visive di un data set per comprendere le caratteristiche dei dati.

Nell'addestramento dell'apprendimento approfondito è stato in qualche modo ridotto il ruolo delle CPU host, sostituite dalle GPU durante l'effettivo processo di addestramento. Nonostante ciò, le CPU continuano a rappresentare il livello di elaborazione per software critici, tra cui CUDA e il sistema operativo, e per l'orchestrazione di processi tra le GPU o con altri componenti elettronici. Inoltre, le CPU stanno iniziando ad assumere un nuovo ruolo sempre più preponderante come engine di inferenza dell'IA nei casi in cui una workstation venga utilizzata per l'esecuzione di un modello di IA in produzione. Secondo le previsioni di IDC, entro il 2024 la spesa sull'infrastruttura per l'inferenza dell'IA supererà la spesa sull'infrastruttura per l'addestramento dell'IA e una parte significativa (39%) di tale interferenza avverrà sulle CPU host.

Workstation e server a confronto: una relazione simbiotica

Per la maggior parte delle organizzazioni, pragmatismo è la regola generale nel caso in cui una workstation, un server on-premise, un'istanza cloud o una combinazione di questi tre elementi venga implementata per lo sviluppo dell'IA. Esiste una relazione simbiotica tra workstation, server e istanze cloud per le varie fasi di sviluppo di un progetto di IA.

Il vantaggio di una workstation rispetto a un server di data center è rappresentato dal fatto che i data scientist possono lavorare ovunque vogliano, un aspetto importante alla luce della recente pandemia, ma anche in circostanze normali. Possono inoltre sperimentare liberamente i loro modelli di IA, con tutte le iterazioni necessarie, in quanto la potenza delle moderne workstation, con GPU altrettanto potenti, consente spesso processi iterativi più interattivi, con feedback e risultati immediati, senza dover richiedere l'accesso ai server o incorrere in altre restrizioni del data center. Le workstation offrono loro la flessibilità

necessaria per avvicinare l'elaborazione ai dati, e non il contrario, risparmiando larghezza di banda, riducendo la congestione di rete e incrementando la portata. Non solo: le workstation possono essere configurate per esigenze diverse, tra cui attività tradizionali di apprendimento automatico e operazioni che richiedono un uso maggiore di apprendimento approfondito.

Inoltre, nonostante una crescita significativa nel mercato, i server accelerati non sono ancora ampiamente diffusi nei data center aziendali. Alla stesura di questo white paper, è stato accelerato in media il 4% dei server nei data center aziendali, ovvero molte organizzazioni non hanno i mezzi per sviluppare o eseguire l'intelligenza artificiale su GPU on-premise immediatamente disponibili. Anche per questo motivo, le workstation accelerate rappresentano un'utile alternativa per lo sviluppo dell'IA.

Le workstation altamente accelerate sono oggi sufficientemente potenti da eseguire l'addestramento dell'apprendimento approfondito a condizione che il modello di IA non raggiunga dimensioni eccessive, eliminando l'esigenza di addestramento sui server. I modelli addestrati su workstation con GPU possono essere implementati sulle workstation o sui server senza GPU, sfruttando le funzionalità di inferenza nelle CPU. Tecnologie software, quali Intel DL Boost e oneAPI, possono alimentare l'inferenza dell'IA sulla CPU, consentendo ai server non accelerati già implementati nei data center di supportare applicazioni di intelligenza artificiale.

Workstation e cloud a confronto

Il cloud computing ha rivoluzionato l'approccio mentale delle organizzazioni all'infrastruttura, ai dati e alle applicazioni. Con la promessa di scalabilità quasi illimitata, il cloud consente agli sviluppatori di eseguire il provisioning delle risorse on-demand, accelerando potenzialmente il ritmo dell'innovazione con minori vincoli. Apparentemente, il cloud rappresenta il paradigma perfetto per lo sviluppo dell'IA,

ma non è sempre così. Di fatto, come è emerso da una ricerca di IDC, le organizzazioni tendono sempre più a riportare alcuni carichi di lavoro dal cloud pubblico nell'infrastruttura on-premise. Ciò è dovuto a una serie di fattori.

- **Disponibilità del cloud:** chiunque si sia affidato ai servizi cloud ha sperimentato un'interruzione dell'alimentazione, per via di problemi con il provider di cloud stesso o a causa di un'interruzione nella connettività di rete in un punto qualsiasi tra il data center iperscalabile e l'utente finale. In questi casi, gli utenti sono in balia del fornitore di servizi per individuare una soluzione al problema, mentre rischiano un blocco totale della produttività.
- **Sicurezza e conformità:** in molti settori, le policy di governance aziendale determinano la posizione in cui i dati possono essere comunicati e archiviati, limitando l'utilizzo di servizi cloud. Anche normative pubbliche, come il GDPR in Europa e il California Consumer Privacy Act, applicano regole sulla sovranità dei dati.
- **Costi:** spesso le organizzazioni sottostimano quanto rapidamente possano aumentare le spese dei servizi cloud, soprattutto per i carichi di lavoro che richiedono funzionalità di elaborazione a prestazioni elevate e grandi quantità di storage. L'economia del cloud si basa sulla misurazione del consumo di tutti i tipi di risorse, incluso il ripristino dei dati nell'infrastruttura on-site.
- **Pressione associata all'apprendimento mediante sperimentazione:** la maggior parte delle iniziative di IA ha inizio con una porzione significativa di sperimentazione, in cui i modelli non funzionanti sono parte integrante del processo di sviluppo; in questo processo, i data scientist e gli sviluppatori di IA subiscono un certo livello di frustrazione nel momento in cui i costi di fatturazione del cloud si accumulano senza che riescano ancora a visualizzare risultati concreti.

Le workstation possono gestire queste limitazioni, sfruttando al contempo le tecnologie native per il cloud, come le architetture basate su microservizi e l'automazione basata su API. In questo modo è possibile usufruire degli stessi vantaggi derivanti dal confronto tra workstation e server di data center.

- **Lavoro agile:** rimuovendo la dipendenza dal cloud pubblico, sono ora possibili scenari disconnessi. Molti ambienti con sicurezza elevata sono isolati tramite air gap dalle reti pubbliche: le workstation di IA possono gestire in modo specifico questa esigenza. Inoltre, le risorse locali riducono la domanda di una connettività di rete a costi elevati.
- **Collocazione dei dati:** la proliferazione dei dispositivi IoT e di altre apparecchiature connesse contribuisce a una crescita esponenziale dei dati nell'edge. In molti casi, ha senso eseguire la colocation delle risorse di elaborazione con una workstation dedicata, il che permette anche di risolvere molti requisiti di conformità limitando lo spostamento dei dati.
- **Sperimentazione libera:** l'addestramento e l'ottimizzazione dei modelli di IA sono elementi di un processo iterativo che spesso include l'apprendimento mediante sperimentazione. Gli sviluppatori hanno bisogno di sperimentare liberamente senza compromessi, per via di possibili spese di servizi aggiuntivi. Le workstation offrono inoltre maggiore flessibilità per l'uso di strumenti personalizzati.

Riguardo a quest'ultimo punto, confrontare il prezzo di una workstation con un deployment nel cloud è relativamente semplice, in quanto gran parte dei provider di servizi cloud fornisce stime immediate sui costi di qualsiasi configurazione richiesta da un utente finale. Ad esempio, il costo di una normale macchina virtuale con una scheda NVIDIA T4 e un'istanza di storage SSD da 375 GiB operativa otto ore al giorno, cinque giorni a settimana, è di $ 140 con uno dei principali fornitori di cloud. Raddoppiando il numero di macchine virtuali, schede T4 e istanze di storage SSD, il costo raggiunge i $ 365 al mese. Rimanendo con due macchine virtuali, ma con quattro schede T4 e quattro istanze di storage da 375 GiB con addestramento full-time in esecuzione nell'ambiente, il costo è di $ 2.700 al mese. Possiamo dunque affermare che i costi del cloud per lo sviluppo dell'IA possono aumentare vertiginosamente a migliaia di dollari all'anno, decisamente maggiori rispetto all'ammortamento annuo di una workstation di fascia alta.

PROTOTIPAZIONE DELL'IA SULLE WORKSTATION

Rispetto ai server on-premise e al cloud, le workstation offrono un netto vantaggio in termini di prototipazione dei modelli di IA. I server nel data center possono essere impiegati al loro massimo utilizzo oppure essere eccessivamente mission critical per la prototipazione e i test di IA e, come già spiegato, le istanze cloud possono determinare rapidamente uno sforamento del budget se sovrautilizzate come ambiente di test. Le workstation evitano ai data scientist o agli sviluppatori di IA l'onere di dover negoziare l'accesso al server o la preoccupazione legata all'accumulo dei costi del cloud in fase di prototipazione. I loro costi ridotti una tantum offrono massima libertà di prototipazione ovunque e in qualsiasi momento, senza spese aggiuntive.

DEPLOYMENT DI MODELLI DI IA SU WORKSTATION

Sebbene il deployment di modelli di IA su una workstation sia una strategia comune da alcuni anni, IDC rileva un incremento di casi d'uso di *deployment* di un modello di IA su una workstation, generalmente nell'edge: in altre parole, significa inserire il modello di IA in produzione sulla workstation mediante inferenza. L'edge è in rapida ascesa come posizione per il deployment di IA per i server. È più che triplicato dal 2020 al 2024 in termini di spesa hardware annua e le workstation non sono da meno, nel momento in cui gli utenti finali ne individuano i vantaggi in corrispondenza dell'edge.

IDC definisce l'edge come un paradigma di elaborazione distribuita che include il deployment dell'infrastruttura e delle applicazioni all'esterno di data center cloud e on-premise centralizzati, il più vicino possibile alla posizione in cui i dati vengono generati e consumati, tra cui uffici remoti e filiali, nonché sedi specifiche del settore, quali fabbriche, magazzini, ospedali e punti di vendita al dettaglio.

I carichi di lavoro che richiedono un uso elevato di dati ed elaborazione vengono sempre più implementati on-premise o nell'edge. Questo avviene per mitigare le limitazioni intrinseche nei cloud pubblici, ad esempio la quantità di tempo necessaria per eseguire il caricamento di data set di grandi dimensioni e i costi variabili associati all'addestramento dell'IA, soprattutto nei casi che richiedono una significativa quantità di sperimentazione di Data Science.

Dalla ricerca di IDC è emerso che l'edge è uno scenario di deployment in rapida crescita per l'IA, con organizzazioni che hanno investito 2,9 miliardi di dollari in elaborazione di IA nell'edge nel 2023, cifra che dovrebbe arrivare a 6,9 miliardi di dollari nel 2026 (consultare *Worldwide AI Hardware Forecast, 2022-2026: Strong Market Growth for AI Compute and Storage,* IDC n. US49671722, settembre 2022). Inoltre, l'edge è sempre più richiesto come opzione di deployment per carichi di lavoro HPC, ad esempio in ambito tecnico e di progettazione, con alcune aziende che al momento investono circa 1 miliardo di dollari in questi carichi di lavoro nell'edge, fino ad arrivare a una previsione di spesa pari a 2,4 miliardi di dollari entro il 2027 (vedere *Worldwide High-Performance Computing Server Forecast, 2023-2027: Enterprise Will Overtake HPC Labs,* IDC n. US50525123, aprile 2023). Queste sono aree in cui è comprensibile implementare una workstation di IA.

In fase di deployment di un modello di IA su una workstation nell'edge, non sono sempre necessarie GPU di fascia alta, come nel caso della fase di sviluppo. GPU più leggere possono eseguire l'inferenza dell'IA e, in alcuni casi, non sono neppure necessarie. In questi casi, le CPU possono svolgere in modo appropriato l'attività di inferenza, soprattutto se utilizzate con ottimizzazioni come Intel DL Boost, un set di funzioni impostate con istruzioni sui microprocessori Intel per accelerare i carichi di lavoro di IA, inclusa l'inferenza. Con Intel DL Boost, Intel afferma di aver rilevato una portata di inferenza INT8 in tempo reale di 1,45 volte superiore con un processore scalabile Intel Xeon di quarta generazione che supporta Intel DL Boost rispetto alla generazione precedente (BERT-Large SQuAD). In questo modo, la workstation risulta anche più idonea per il deployment nell'edge, in cui aspetti quali potenza, mobilità e gestione termica richiedono voltaggi inferiori. Intel Movidius Myriad (M2) è la scelta ideale in tal senso, grazie a un ingombro energetico ridotto a 12 W.

Casi d'uso per il deployment dell'IA sulle workstation

Alcune situazioni si prestano naturalmente al deployment dell'IA su workstation implementate in locale. Tra le caratteristiche comuni rientrano volumi di grandi dimensioni di dati non strutturati e temporali generati da macchine, come i flussi video e le immagini. In alcuni casi, inoltre, gli esperti in materia devono potenziare i modelli di IA con interpretazione umana.

Ecco alcuni esempi:

- **AIOps:** con la crescita dei sistemi IT in termini di scalabilità e complessità, è sempre più determinante passare da un incident management reattivo a un monitoraggio proattivo, tanto più quando l'infrastruttura e le applicazioni sono distribuite nell'edge, in cui il personale tecnico è ridotto, se non addirittura assente. Modellando una baseline di prestazioni normali, è possibile identificare anomalie e automatizzare le procedure di correzione.

- **Risposta in caso di emergenza:** durante un'emergenza, gli operatori di primo intervento devono valutare rapidamente una situazione, monitorare le apparecchiature critiche e implementare risorse per aiutare chi è più in difficoltà. Il tutto deve spesso avvenire in un ambiente senza connettività di rete che richiede una workstation locale in grado di aggregare i feed di dati, eseguire inferenza sui modelli di IA e automatizzare le comunicazioni per il personale chiave.
- **Radiologia:** i progressi nella tecnologia di imaging hanno determinato un aumento di volume dei dati generati da una singola scansione, che devono essere analizzati on-site in modo tempestivo. I modelli di IA addestrati da milioni di esempi precedenti possono identificare schemi in modo più accurato rispetto all'occhio umano, incrementando i livelli di precisione.
- **Esplorazione di petrolio e gas:** le aziende del settore gas-petrolifero upstream usano una combinazione di dati telemetrici, sismici e di imaging per individuare riserve di risorse naturali, scegliere le zone di perforazione e ottimizzare le prestazioni delle apparecchiature nel processo di produzione. Questa condizione richiede spesso l'analisi di informazioni in aree in cui sono disponibili solo dispendiose comunicazioni satellitari.
- **Ricerca contro il cancro e sviluppo di farmaci:** i ricercatori negli ospedali e nelle accademie utilizzano IA ed elaborazione del linguaggio naturale per aiutare gli oncologi a determinare le cure individuali contro il cancro più efficaci per i loro pazienti. Combinano inoltre apprendimento automatico con visione artificiale per permettere ai radiologi di comprendere meglio l'avanzamento dei tumori dei pazienti e utilizzano algoritmi per capire come si sviluppano e quali cure applicare per contrastarli.
- **Elaborazione dei claim di indennizzo assicurativo:** l'elaborazione manuale dei claim di indennizzo richiede molto lavoro ed è soggetta a errore umano. Un livello di IA in grado di valutare la validità dei claim riduce i costi, consentendo ai periti assicurativi di concentrarsi sui casi che richiedono maggiori indagini. Tutto questo determina un aumento della portata complessiva dell'operazione senza compromettere la precisione.
- **Telemedicina:** l'IA migliora i tassi di recupero dei pazienti grazie alla creazione di piani di trattamento individuali basati sui segni vitali acquisiti in tempo reale da dispositivi indossabili. Queste informazioni vengono combinate con le cartelle cliniche dei pazienti e una knowledge base di casi simili. Tutto questo è particolarmente importante nelle aree rurali che si affidano maggiormente alla telemedicina.
- **Sicurezza nel retail (antifurto):** vengono utilizzate analisi in tempo reale applicate ai flussi video per prevedere comportamenti umani che possono sfociare in attività criminali. A tale scopo, è necessario analizzare insieme più feed video per tracciare i movimenti di una persona all'interno di un negozio. Considerata la tempestività insita nell'identificazione di un evento materiale, si tratta di un processo che idealmente deve essere eseguito in locale.
- **Gestione del traffico:** gli enti pubblici responsabili delle operazioni di trasporto adottano in misura sempre maggiore l'intelligenza artificiale per il coordinamento dei semafori e dei segnali digitali al fine di migliorare il flusso dei veicoli e la sicurezza dei cittadini. Tutto questo richiede una combinazione di input, tra cui videocamere e telemetria, acquisiti da sensori stradali per ottimizzare i modelli di traffico.
- **Monitoraggio degli impianti di produzione:** per un responsabile di stabilimento, è della massima importanza assicurare il tempo di utilizzo di processi critici e rispettare i piani di produzione. Ciò si traduce nella manutenzione predittiva delle apparecchiature chiave, nel rilevamento automatizzato dei difetti e nell'ottimizzazione della supply chain all'interno e all'esterno del sito. Si tratta di un'area in cui l'intelligenza artificiale può aiutare gli operatori umani a migliorare le prestazioni, pur rispettando gli standard di sicurezza.

- **Droni:** l'analisi automatizzata delle immagini acquisite dai droni permette di monitorare un'ampia gamma di condizioni a un livello precedentemente impossibile. Questo determina un impatto significativo sull'ispezione degli impianti di fornitura di gas ed energia elettrica, sulle indagini assicurative, sulle iniziative di ricerca e soccorso, sull'agricoltura di precisione e sulla manutenzione degli allevamenti ittici e delle riserve naturali.
- **Ambienti di lavoro giornalieri:** gli ambienti di lavoro giornalieri risultano enormemente migliorati grazie all'uso di strumenti di produttività basati su IA, come Microsoft Copilot.
- **Energie rinnovabili:** i siti di energie rinnovabili, come le centrali eoliche, le dighe idroelettriche e i parchi fotovoltaici, richiedono monitoraggio, manutenzione e raccolta in tempo reale di dati che devono essere generati e analizzati in loco.

WORKSTATION DELL PER L'IA

Dell offre un'ampia gamma di workstation per vari livelli di sviluppo e/o implementazione di intelligenza artificiale, tutti riuniti sotto il nome di Data Science Workstation (DSW). Questa sezione fornisce un accenno alle specifiche e quindi affronta vari ruoli e applicazioni in ambito di intelligenza artificiale, ad esempio i data scientist e i vantaggi derivanti dalla tecnologia Dell DSW. Queste workstation di Data Science predisposte per l'IA sono state progettate nello specifico per i data scientist. Le workstation Precision Data Science di nuova generazione utilizzano la funzionalità di IA per perfezionare i dispositivi per le prestazioni ottimizzate delle applicazioni più utilizzate dai data scientist, consentendo loro di completare in anticipo le attività più importanti. Inoltre, le workstation Dell Precision sono testate e certificate da ISV indipendenti per verificare che supportino le applicazioni a prestazioni elevate necessarie ai clienti Dell per svolgere le loro attività quotidiane.

Elementi di differenziazione delle workstation Dell

Le workstation Dell Precision con GPU NVIDIA RTX sono progettate per fornire livelli affidabili di scalabilità e prestazioni per le analisi e le iniziative di IA di un'organizzazione. Dell Technologies offre soluzioni hardware complete, ottimizzate per l'esecuzione delle applicazioni software di IA più recenti del settore.

- **Configurazione hardware affidabile:** le workstation Dell Precision offrono una gamma di efficaci configurazioni hardware, che includono processori multi-core, RAM a elevata capacità e varie opzioni di GPU. Questi componenti forniscono le risorse di elaborazione necessarie per attività di IA, offrendo livelli efficienti di addestramento e inferenza.
- **Scalabilità e personalizzazione:** le workstation Dell Precision sono scalabili e personalizzabili, consentendo agli utenti di adattare la configurazione hardware in base ai loro specifici requisiti di IA. Grazie a questo grado di flessibilità, le workstation possono essere ottimizzate per le singole esigenze dei carichi di lavoro di IA.
- **Certificazione e ottimizzazione:** Dell collabora con NVIDIA alla certificazione delle workstation Precision per compatibilità e prestazioni con GPU NVIDIA RTX, tra cui le schede NVIDIA RTX 6000 Ada Generation. Questa certificazione assicura livelli eccellenti di integrazione e prestazioni ottimizzate in caso di utilizzo di workstation Dell Precision con GPU NVIDIA RTX per attività di IA.
- **Potente funzionalità di elaborazione:** le workstation Dell Precision dotate di processori Intel forniscono la potenza di elaborazione necessaria per le attività di IA. Con processori multi-core ed elevate velocità di clock, queste workstation offrono le prestazioni richieste per l'addestramento e l'inferenza nei flussi di lavoro di IA.

- **Supporto per software e strumenti:** le workstation Dell Precision includono software e strumenti precaricati che supportano lo sviluppo e il deployment dell'IA. Sono inclusi stack software ottimizzati, framework di IA e librerie che utilizzano le GPU NVIDIA RTX, rendendo più semplice per gli utenti iniziare a lavorare con progetti di IA.

Inoltre, le tecnologie menzionate nelle sezioni seguenti rappresentano altre aree importanti in cui si differenziano le workstation Dell.

Reliable Memory Technology

Dell fornisce una tecnologia basata su ECC denominata Reliable Memory Technology Pro (RMT Pro), progettata per massimizzare il tempo di utilizzo. Opera congiuntamente con la memoria ECC per rilevare e correggere errori di memoria in tempo reale. Secondo Dell, RMT Pro elimina praticamente questi errori evitando che la memoria danneggiata venga rivisitata anche se il modulo DIMM continua a essere in uso. Dopo un riavvio del sistema, RMT Pro isola l'area della memoria danneggiata rendendola non visibile al sistema operativo. Di conseguenza, i data scientist e gli sviluppatori di IA evitano di subire arresti anomali continui a causa del fatto che la memoria danneggiata continua a essere indirizzabile: un importante incremento della produttività.

Dell Optimizer for Precision

Dell include inoltre Dell Optimizer for Precision nella maggior parte delle workstation, che regola automaticamente le impostazioni di sistema in modo da consentire l'esecuzione di varie applicazioni commerciali popolari alla massima velocità possibile. Tutto questo migliora la produttività di data scientist e sviluppatori. Lo strumento crea inoltre report delle prestazioni in tempo reale per l'IT sull'utilizzo di processori, storage, memoria e scheda grafica. DOP non è ancora disponibile su Linux ed è quindi essenzialmente utile per il deployment dell'IA, in quanto lo sviluppo continua a essere eseguito tendenzialmente con software open source basato su Linux. Dell Optimizer for Precision fornisce inoltre ExpressSign-in, Express Charge (per dispositivi mobili), Intelligent Audio e strumenti di reporting e analisi per ottimizzare la workstation.

SFIDE E OPPORTUNITÀ

Per le aziende

IDC rileva una biforcazione nel mercato dell'intelligenza artificiale. Da un lato, le aziende implementano strategie di dati per rimanere competitive, tra cui l'integrazione dell'IA su vasta scala. Ad esempio, si confrontano con colleghi che hanno raggiunto importanti traguardi utilizzando offerte di infrastrutture IA effettivamente registrate nei primi 100 supercomputer. D'altro canto, le aziende osservano la realtà quotidiana di piccole iniziative di intelligenza artificiale attuate su server disponibili nel data center o nel cloud, spesso con budget insufficiente e hardware meno performante.

Per molte aziende, il primo scenario non è pertinente e il secondo è fin troppo reale. Per loro, la sfida consiste nel fornire ai data scientist e/o agli sviluppatori di IA gli strumenti più adatti a eseguire l'addestramento dell'IA in modo tempestivo, senza investire eccessivamente su istanze cloud o server di data center accelerati tramite GPU. Per IDC, queste aziende sono attrezzate in modo adeguato a fornire a data scientist e sviluppatori potenti workstation accelerate da GPU.

Per Dell

Nel mercato si è diffusa la credenza che lo sviluppo e il deployment dell'IA richiedano hardware del server costosi e accelerati, spesso addirittura in cluster. Può essere vero per i più grandi algoritmi di intelligenza artificiale, con miliardi di parametri, ma la maggior parte delle aziende non sviluppa algoritmi di questa entità. Con le loro iniziative di IA realizzano qualcosa di utile, incisivo e gestibile e molte aziende non comprendono che i modelli di IA su scala così comune possono essere sviluppati e implementati su workstation. La sfida di Dell è quella di abbattere i preconcetti ed educare il mercato in merito alle possibilità con il suo portafoglio di workstation.

Al contempo, Dell deve verificare che le workstation siano all'altezza delle aspettative e non si tramutino in colli di bottiglia tecnologici nel tempo. Questo significa attuare una rapida innovazione costante per non deludere gli utenti finali che utilizzano le workstation in modo appropriato (ovvero, che non tentano di eseguire un algoritmo con miliardi di parametri). Significa anche che, per i clienti in procinto di eseguire una scalabilità molto rapida o i cui algoritmi iniziano ad assumere dimensioni significative, il passaggio dalla workstation alla linea di server di IA Dell è semplice. In ciò risiede anche l'opportunità per Dell di fornire la soluzione più adatta a ogni cliente, indipendentemente dal volume dell'iniziativa di IA a cui stanno lavorando.

CONCLUSIONI

Secondo IDC, oggi le workstation non sono valutate appieno come validi strumenti per lo sviluppo e il deployment dell'intelligenza artificiale in molti casi d'uso. Forniscono ai data scientist e agli sviluppatori di IA un'affidabile piattaforma accelerata tramite GPU che determina un CAPEX inferiore rispetto ai server e un OPEX nettamente inferiore rispetto alle istanze cloud, offrendo al contempo maggiore libertà di sperimentare i modelli di IA. La maggior parte delle aziende, che sviluppano iniziative di IA senza algoritmi con miliardi di parametri, deve considerare la possibilità di potenziare i team di IA con workstation per processi di sviluppo dell'intelligenza artificiale senza vincoli e deployment semplificato basato sull'edge.

About IDC

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

Global Headquarters

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