ANALISI DI MERCATO

Sponsorizzato da: Dell

Gli investimenti nell'infrastruttura di elaborazione di IA stanno aumentando a un ritmo accelerato. La buona notizia è che, per quanto riguarda l'infrastruttura di elaborazione di IA, oltre il 50% dei sistemi non si evolverà nel 2024 e potrà essere eseguito su reti Ethernet e server standard.

GenAI si evolve a velocità record di pari passo con le aziende che iniziano il percorso verso l'intelligenza artificiale generativa

Febbraio 2024

Redatto da: Brandon Hoff, Research Director, Enabling Technologies: Networking and Comm, e Vijay Bhagavath, Research Vice President, Cloud and Datacenter Networks

Introduzione

La ricerca di IDC fornisce le previsioni e i fattori sottostanti che, a nostro avviso, influiranno sugli investimenti IT nel 2024 e oltre. I leader tecnologici e le rispettive controparti nelle linee di business (LOB) possono basarsi su questo documento per orientare le attività di pianificazione strategica.

I team operativi acquisiscono dati, creano data lake e sfruttano il cloud per archiviarli. Ora, con ChatGPT sempre più popolare e l'IA generativa (GenAI) nel suo "momento iPhone", i team operativi sanno come utilizzare i propri data set. Tutti conoscono i vantaggi dell'IA generativa, pertanto anche gli investitori, i dirigenti e il mercato esercitano pressioni affinché i team operativi implementino una strategia GenAI efficace. Per migliorare le operazioni aziendali e la produttività dei dipendenti si possono sfruttare diverse tecnologie e un'ampia gamma di opzioni, da GenAI a ML, fino ai gemelli digitali e altro ancora. L'implementazione corretta della tecnologia adeguata rappresenta un KPI essenziale per i team operativi e per il business in generale.

IN BREVE

CONSIDERAZIONI PRINCIPALI

Inizia a pianificare l'infrastruttura di GenAI:

- Integra l'intelligenza artificiale più velocemente della concorrenza, spostando subito on-premise almeno una copia dei dati, un'iniziativa volta a recuperare i dati dal cloud.
- Investi nella comprensione del valore degli algoritmi alla base dell'IA generativa, dell'IA, della ML e dei gemelli digitali nel business e assegna le priorità in base al valore per il business.
- Tre passaggi: implementazione di rete Ethernet e server standard per cui è possibile valutare l'adozione di GenAI. Scale-out di GenAI per carichi di lavoro di livello enterprise in base alle esigenze, sfruttando l'Ethernet standard. Ribilanciamento dei carichi di lavoro tra l'infrastruttura on-premise e off-premise per ottimizzare CAPEX e OpEx nei prossimi tre-cinque anni.

Comprendere l'IA generativa oggi

Con l'esplosione della domanda di soluzioni IA, i provider di servizi cloud e le aziende stanno costruendo la loro infrastruttura a un ritmo accelerato. I fornitori di servizi cloud utilizzano la stragrande maggioranza degli acceleratori per l'IA per costruire la propria infrastruttura di IA, tuttavia, questi acceleratori sono costosi e fanno lievitare i prezzi dei servizi di IA che offrono. Gli acceleratori per l'IA sono costituiti dalle tecnologie GPU, TPU, FPGA, ASSP e ASIC. Questi fornitori di servizi cloud stanno costruendo fabbriche di IA per carichi di lavoro enormi che abbracciano le esigenze di un'ampia gamma di aziende e sono implementate principalmente negli ambienti IT più grandi del mondo, nove aziende in totale.

L'infrastruttura di GenAI per i carichi di lavoro di livello enterprise, invece, può essere creata con sistemi standard senza necessità di accelerare l'innovazione. IDC prevede che oltre il 50% dei sistemi di GenAI non si evolverà nel 2024; pertanto, chiunque può iniziare a implementare la propria infrastruttura di GenAI con server e reti standard. Per chi le necessiti, sono anche disponibili le GPU. Esistono diverse opzioni per l'implementazione dell'infrastruttura di IA, nonché diversi tipi di soluzioni di GenAI, IA, ML e gemelli digitali che apportano molteplici vantaggi alle aziende. Gli stack software GenAI sono generalmente supportati sui server standard, pertanto l'utilizzo di GenAI su questi ultimi risulta vantaggiosa. Le aziende che investono nell'infrastruttura standard on-premise accelerano più di altre l'attuazione delle rispettive iniziative di GenAI. È essenziale che i team IT inizino a valutare i vari algoritmi di GenAI, IA, ML e gemelli digitali per identificare quelli che esercitano il maggiore impatto sul loro business.

Vantaggi

GenAI e altri modelli fondamentali stanno cambiando le regole del gioco, portando a un nuovo livello le tecnologie assistive e offrendo potenti funzionalità agli utenti non tecnici. GenAI ha il potenziale di aumentare l'efficienza e la produttività, aprire nuove opportunità di crescita, ridurre i costi e fornire un vantaggio competitivo alle aziende che la sfruttano.

Per creare la propria infrastruttura di GenAI è opportuno iniziare con l'integrazione di questa tecnologia rivoluzionaria nelle operazioni aziendali e con lo sviluppo on-site di competenze nello stack tecnologico GenAI. Dando priorità agli investimenti tecnologici per creare un'infrastruttura di GenAI iniziale on-premise basata su rete Ethernet e server aziendali standard, le aziende che sfruttano questa tecnologia trasformativa otterranno un vantaggio in termini di time to market.

Considerazioni

Agisci ora per sfruttare GenAI

Dati i risultati impressionanti offerti da ChatGPT e da altri modelli, la tecnologia GenAI desta molto interesse. È vero che GenAI offre valore, ma il valore varia in base alla fonte dei dati proprietari e agli algoritmi implementati. I consigli di amministrazione, gli investitori e i dirigenti dovranno porre domande e cercare di comprendere in che modo GenAI può aiutarli nelle loro attività.

Per i business che hanno acquisito grandi quantità di dati proprietari non strutturati, l'IA generativa promette di creare, a partire dai dati proprietari esistenti, contenuti originali che dovrebbero contribuire a ridefinire l'organizzazione in termini di innovazione continua. È opportuno adottare un approccio graduale per capire i vantaggi che GenAI può portare all'azienda e in che modo procedere.

Creazione dell'infrastruttura di GenAI iniziale su Ethernet

Dal punto di vista dei carichi di lavoro di livello enterprise, i sistemi standard offrono le prestazioni necessarie per iniziare il percorso di adozione di GenAI. Inoltre, basare l'infrastruttura di GenAI su rete Ethernet e server standard significa sfruttare i sistemi operativi, gli strumenti di gestione e gli strumenti di gestione della rete aziendali. Una volta compresi i requisiti di elaborazione per gli LLM vantaggiosi per il business, è possibile migliorare le prestazioni di elaborazione selezionando gli acceleratori per l'IA o la GenAI adeguati. Il punto è che l'infrastruttura di IA deve essere ben architettata. Una fabric ben architettata può supportare da decine a migliaia di nodi di elaborazione per l'IA.

Per quanto possano esistere diverse opzioni di rete per i carichi di lavoro GenAI, l'opzione onnipresente, aperta e multi-vendor preferita è la rete Ethernet. I deployment iniziali di GenAI possono essere supportati attraverso la rete Ethernet standard oggi disponibile per i cluster GenAI.

Creazione di un'infrastruttura di GenAI con Ultra Ethernet

Per ogni azienda che inizia gradualmente a sviluppare l'IA generativa può essere opportuno creare la propria infrastruttura di GenAI scale-out. La creazione di un'infrastruttura di GenAI scale-out richiede due aggiunte chiave: reti e acceleratori di IA nel data center. In un'infrastruttura di GenAI scale-out tipica, vengono implementate otto GPU del data center su ciascun server e per ogni GPU è implementata una scheda di rete o DPU ad alta velocità per offrire una rete a prestazioni elevate.

"Si prevede che il mercato degli switch Ethernet per data center GenAI nel segmento enterprise crescerà a un tasso composto annuo del 158,2%, da 41,9 milioni di dollari nel 2023 a 1,0 miliardi di dollari nel 2027", Vijay Bhagavath, IDC.

Un requisito chiave per un'infrastruttura di IA scale-out è una rete a prestazioni elevate. Per gli LLM GenAI, il collo di bottiglia dell'elaborazione è il tempo che i dati trascorrono nella rete. Per alcuni carichi di lavoro, il tempo nella rete può rappresentare fino al 60% del tempo di elaborazione di un LLM: l'infrastruttura di elaborazione viene lasciata inattiva mentre i dati si spostano tra i cluster di elaborazione. Per la rete di IA, è oggi disponibile una rete migliorata, erogata tramite Ultra Ethernet Consortium, che promette un'interconnessione performante come le reti di supercomputing, scalabile nel data center cloud e conveniente e onnipresente come l'Ethernet. La rete di IA è essenziale per sostenere la crescita delle esigenze di rete di GenAI e HPC su larga scala. La buona notizia è che la rete Ultra Ethernet Consortium è supportata dalla maggior parte dei fornitori di switch Ethernet.

Per prestazioni ottimizzate, sono necessari tre componenti tecnologici chiave: fibre ottiche, PHY e SerDes ad alta velocità. Queste tre tecnologie vengono utilizzate nell'Ethernet e in altre tecnologie di rete, quindi, in sostanza, l'utilizzo di una tecnologia di rete specifica non conferisce alcun vantaggio in termini di prestazioni. Per ottenere massime prestazioni dall'Ethernet, la InfiniBand Trade Association ha lanciato l'iniziativa RDMA over Converged Ethernet (RoCE) e ha definito il protocollo RoCE. RoCE è supportato sugli switch per data center standard e sono disponibili sul mercato ulteriori miglioramenti per aumentare le prestazioni, come switch Ethernet su radice elevata, switch cut-through, bilanciamento del carico e larghezza di banda superiore con link fino a 800 GbE (4 da 200 GbE).

I test iniziali degli LLM GenAI possono fornire informazioni preliminari sui vantaggi che GenAI può apportare al business e supportare la creazione di una strategia di implementazione degli LLM GenAI, nonché offrire dettagli sui tipi di infrastruttura necessari. In sostanza, lo stack software determina i requisiti al livello di semiconduttori che consentono il passo successivo nell'evoluzione della GenAI aziendale. Comprendere lo stack software aiuterà a implementare un'infrastruttura hardware ottimizzata.

Ribilanciamento dell'infrastruttura on-premise e off-premise grazie alla stabilizzazione dei costi dei semiconduttori

Con l'aumento dell'offerta di GPU per i data center, sempre più fornitori offriranno GPU per i data center, saranno disponibili più acceleratori per l'IA e maggiore potenza di elaborazione GenAI sarà disponibile per le distribuzioni locali. Parallelamente, i colli di bottiglia associati ai fornitori di servizi cloud scompariranno e si prevede che i costi si stabilizzeranno. Quando ciò si verificherà, tra tre-cinque anni, il ribilanciamento dei carichi di lavoro GenAI tra l'infrastruttura on-premise e l'infrastruttura cloud ottimizzerà CAPEX e OpEx.

Conclusioni

GenAI è la tecnologia rivoluzionaria per l'IA. Per proseguire il percorso verso l'integrazione di questa tecnologia rivoluzionaria nelle operazioni aziendali, i business dovranno disporre di una strategia/un piano per l'IA generativa da avviare ora per i carichi di lavoro di livello enterprise.

"Il mercato ha continuato a sottovalutare l'espansione di GenAI e IDC prevede una crescita rilevante dei semiconduttori e dell'infrastruttura GenAI", Brandon Hoff, IDC.

La domanda elevata fa lievitare i prezzi dei componenti e il costo dei provider di servizi cloud. Allo stesso tempo, IDC prevede che oltre il 50% dei sistemi GenAI non si evolverà nel 2024; pertanto, chiunque può iniziare a implementare la propria infrastruttura GenAI con server e rete standard. Per chi le necessiti, sono anche disponibili le GPU. Esistono diverse opzioni per l'implementazione dell'infrastruttura di IA, nonché diversi tipi di soluzioni di GenAI, IA, ML e gemelli digitali che apportano molteplici vantaggi alle aziende.

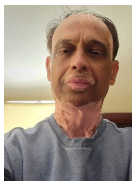
Secondo la previsione di IDC, per ridurre i costi OpEx, le aziende recupereranno i dati dal cloud per l'elaborazione GenAI. Le aziende avvieranno lo sviluppo di GenAI e i relativi test su hardware di rete Ethernet ed elaborazione standard e investiranno per apprendere quali sono gli LLM funzionali per la loro azienda e il valore che possono estrarre dai rispettivi dati proprietari.

Costruire l'infrastruttura per i test LLM di GenAI su reti Ethernet aziendali e server pronti all'uso consentirà all'azienda di sfruttare appieno il valore di GenAI.

Informazioni sugli analisti

***Brandon Hoff,** Research Director, Enabling Technologies: Networking and Comm*

Brandon Hoff si occupa di infrastruttura di rete e comunicazione all'interno del team Enabling Technologies di IDC. Il Sig. Hoff analizza le tendenze tecnologiche, i carichi di lavoro, i prodotti, i fornitori, la supply chain e le strategie di adozione degli utenti finali nell'IT aziendale e nei data center dei fornitori di servizi web, cloud e di telecomunicazione.



***Vijay Bhagavath,** Research Vice President, Cloud and Datacenter Networks*

Vijay Bhagavath promuove la leadership di pensiero fruibile e approfondimenti pragmatici sulle tecnologie e sui mercati delle reti per data center e cloud. Vijay ha una profonda conoscenza del mercato globale delle soluzioni di rete, delle tecnologie, delle roadmap dei prodotti, della differenziazione competitiva e delle strategie di deployment, che gli consente di fornire indicazioni e commenti approfonditi a vendor, provider di cloud, nonché ad acquirenti e operatori IT aziendali.

MESSAGGIO DELLO SPONSOR

IDC Custom Solutions

Il contenuto di questo documento è un adattamento della ricerca di IDC pubblicata su www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
Tel. 508.872.8200
Fax 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com