Inferenza AI Edge con NetApp e Lenovo

NetApp artificial intelligence solutions

NetApp
February 12, 2026

Sommario

Inferenza AI Edge con NetApp e Lenovo

TR-4886: Inferenza AI all'Edge - NetApp con Lenovo ThinkSystem - Progettazione della soluzione

Sathish Thyagarajan, NetApp Miroslav Hodak, Lenovo

Questo documento descrive un'architettura di elaborazione e storage per implementare l'inferenza dell'intelligenza artificiale (IA) basata su GPU sui controller di storage NetApp e sui server Lenovo ThinkSystem in un ambiente edge che soddisfa gli scenari applicativi emergenti.

Riepilogo

Diversi scenari applicativi emergenti, come i sistemi avanzati di assistenza alla guida (ADAS), l'Industria 4.0, le città intelligenti e l'Internet delle cose (IoT), richiedono l'elaborazione di flussi di dati continui con una latenza prossima allo zero. Questo documento descrive un'architettura di elaborazione e storage per implementare l'inferenza dell'intelligenza artificiale (AI) basata su GPU sui controller di storage NetApp e sui server Lenovo ThinkSystem in un ambiente edge che soddisfa questi requisiti. Questo documento fornisce anche dati sulle prestazioni per il benchmark standard del settore MLPerf Inference, valutando varie attività di inferenza su server edge dotati di GPU NVIDIA T4. Esaminiamo le prestazioni di scenari di inferenza offline, a flusso singolo e multi-flusso e dimostriamo che l'architettura con un sistema di archiviazione di rete condiviso ed economico è altamente performante e fornisce un punto centrale per la gestione di dati e modelli per più server edge.

Introduzione

Le aziende generano volumi sempre più ingenti di dati ai margini della rete. Per ottenere il massimo valore dai sensori intelligenti e dai dati IoT, le organizzazioni sono alla ricerca di una soluzione di streaming di eventi in tempo reale che consenta l'edge computing. Per questo motivo, i lavori più impegnativi dal punto di vista computazionale vengono sempre più svolti ai margini, al di fuori dei data center. L'inferenza dell'intelligenza artificiale è uno dei motori di questa tendenza. I server edge forniscono una potenza di calcolo sufficiente per questi carichi di lavoro, soprattutto quando si utilizzano acceleratori, ma lo spazio di archiviazione limitato rappresenta spesso un problema, soprattutto negli ambienti multiserver. In questo documento mostriamo come implementare un sistema di storage condiviso nell'ambiente edge e come ciò possa apportare vantaggi ai carichi di lavoro di inferenza dell'IA senza imporre una penalizzazione delle prestazioni.

Questo documento descrive un'architettura di riferimento per l'inferenza dell'IA all'edge. Combina più server edge Lenovo ThinkSystem con un sistema di storage NetApp per creare una soluzione facile da implementare e gestire. È concepito come una guida di base per implementazioni pratiche in varie situazioni, come ad esempio in fabbriche con più telecamere e sensori industriali, sistemi POS (Point-of-Sale) nelle transazioni al dettaglio o sistemi Full Self-Driving (FSD) che identificano anomalie visive nei veicoli autonomi.

Questo documento riguarda i test e la convalida di una configurazione di elaborazione e storage composta da Lenovo ThinkSystem SE350 Edge Server e un sistema di storage NetApp AFF ed EF-Series entry-level. Le architetture di riferimento forniscono una soluzione efficiente e conveniente per le distribuzioni di intelligenza artificiale, offrendo al contempo servizi dati completi, protezione dati integrata, scalabilità senza interruzioni e archiviazione dati connessa al cloud con il software di gestione dati NetApp ONTAP e NetApp SANtricity .

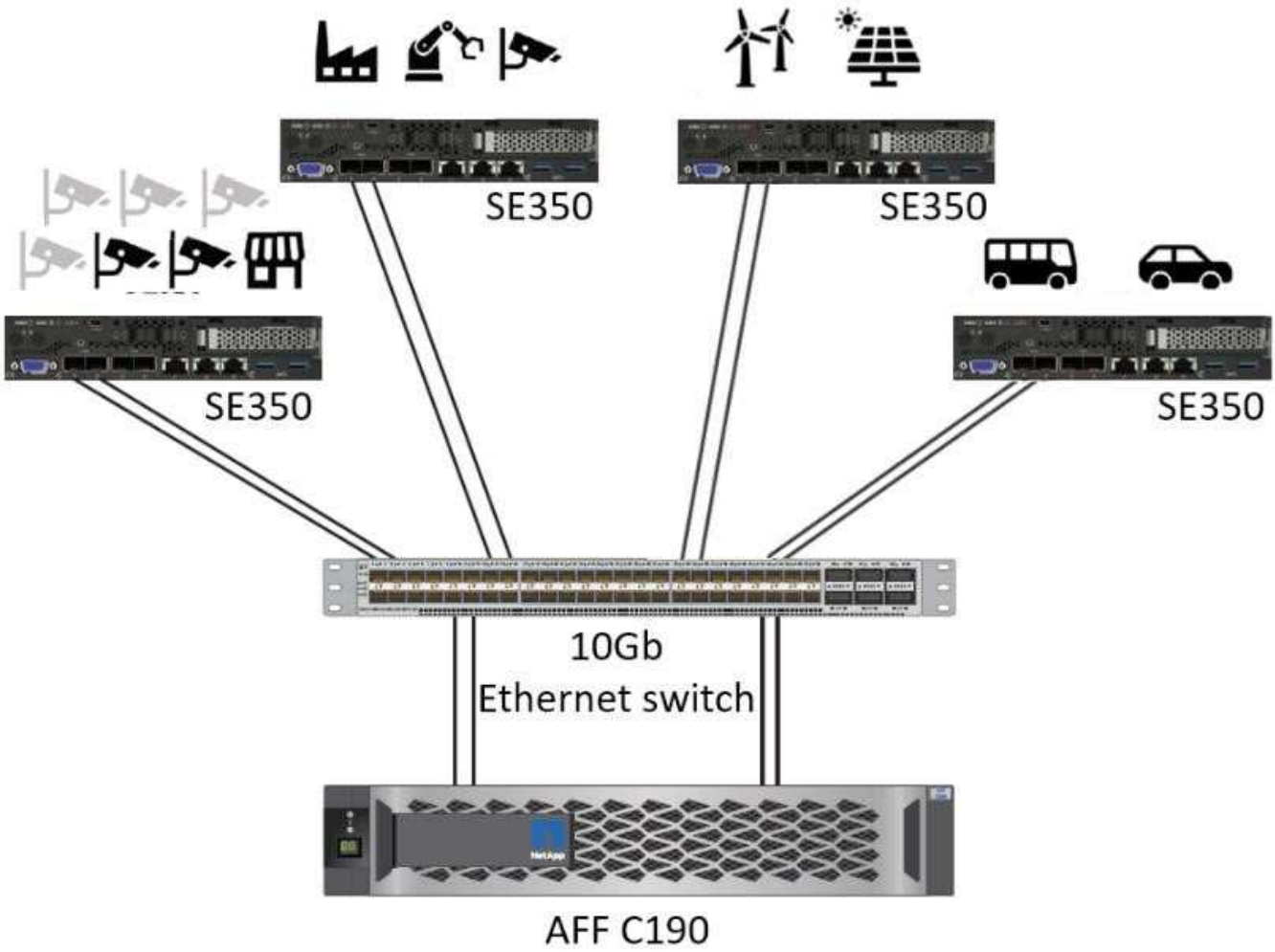
Pubblico di destinazione

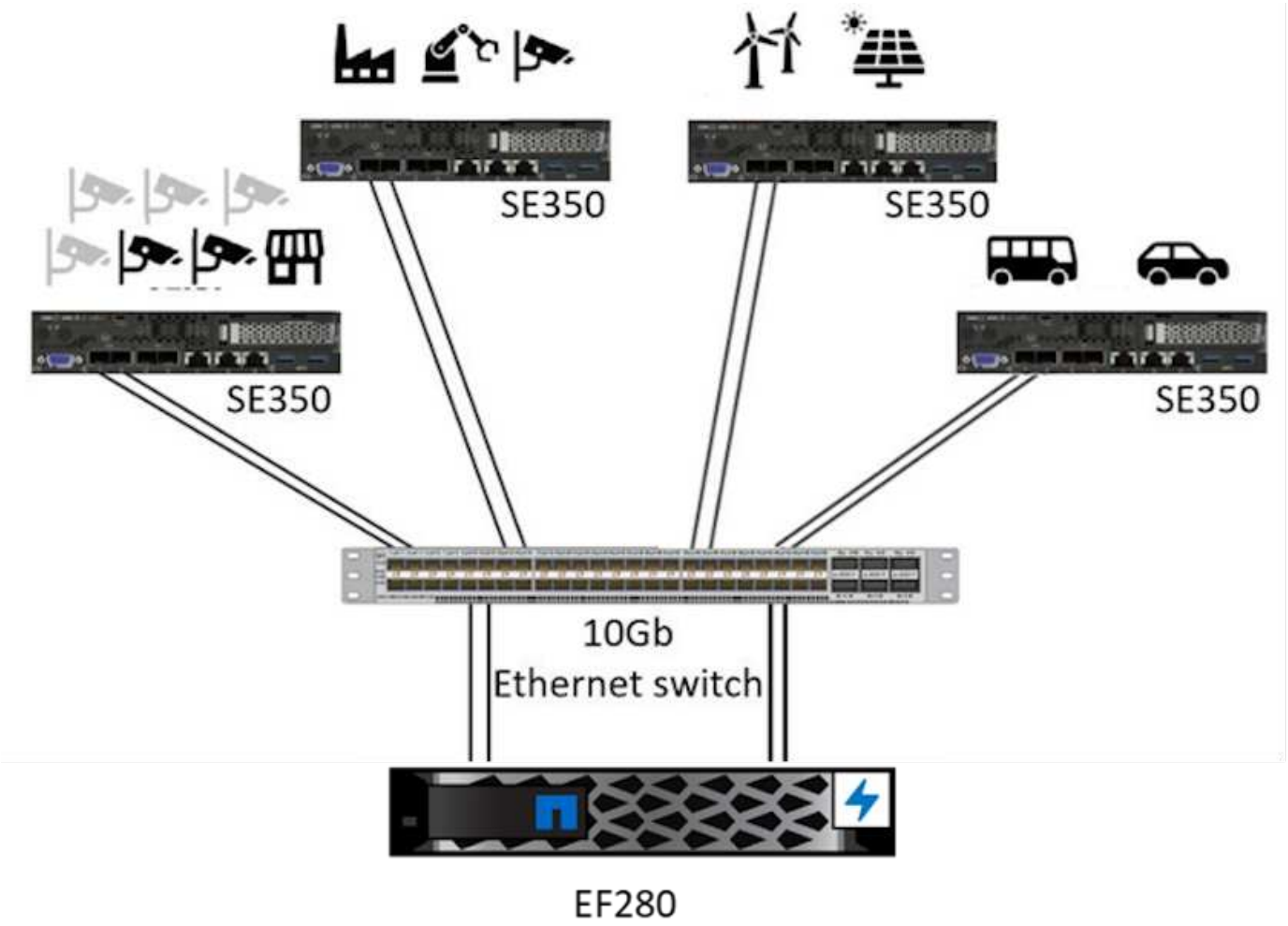
Il presente documento è destinato ai seguenti destinatari:

- Leader aziendali e architetti aziendali che desiderano rendere produttiva l'intelligenza artificiale ai margini della rete.
- Data scientist, data engineer, ricercatori di intelligenza artificiale/apprendimento automatico (ML) e sviluppatori di sistemi di intelligenza artificiale.
- Architetti aziendali che progettano soluzioni per lo sviluppo di modelli e applicazioni AI/ML.
- Data scientist e ingegneri dell'intelligenza artificiale cercano modi efficienti per implementare modelli di deep learning (DL) e ML.
- I gestori dei dispositivi edge e gli amministratori dei server edge sono responsabili dell'implementazione e della gestione dei modelli di inferenza edge.

Architettura della soluzione

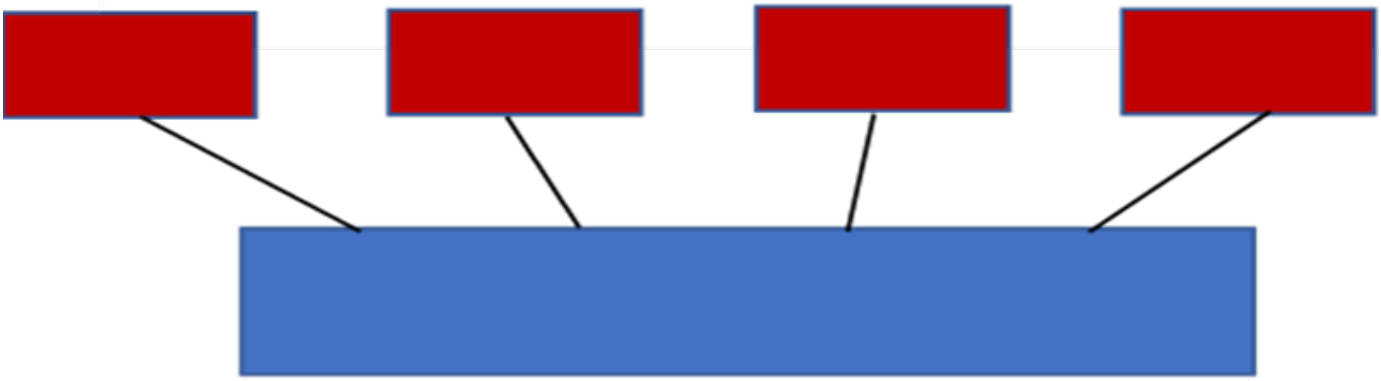
Questa soluzione di server Lenovo ThinkSystem e storage NetApp ONTAP o NetApp SANtricity è progettata per gestire l'inferenza AI su grandi set di dati utilizzando la potenza di elaborazione delle GPU insieme alle CPU tradizionali. Questa convalida dimostra elevate prestazioni e una gestione ottimale dei dati con un'architettura che utilizza uno o più server edge Lenovo SR350 interconnessi con un singolo sistema di storage NetApp AFF , come mostrato nelle due figure seguenti.





La panoramica dell'architettura logica nella figura seguente mostra i ruoli degli elementi di elaborazione e di archiviazione in questa architettura. Nello specifico, mostra quanto segue:

- Dispositivi di edge computing che eseguono inferenze sui dati ricevuti da telecamere, sensori e così via.
- Un elemento di archiviazione condiviso che svolge molteplici funzioni:
  - Fornisce una posizione centrale per i modelli di inferenza e altri dati necessari per eseguire l'inferenza. I server di elaborazione accedono direttamente allo storage e utilizzano modelli di inferenza in tutta la rete senza doverli copiare localmente.
  - Qui vengono pubblicati i modelli aggiornati.
  - Archivia i dati di input che i server periferici ricevono per un'analisi successiva. Ad esempio, se i dispositivi edge sono collegati alle telecamere, l'elemento di archiviazione conserva i video acquisiti dalle telecamere.



rosso	blu
Sistema di elaborazione Lenovo	Sistema di archiviazione NetApp AFF
Dispositivi edge che eseguono inferenze su input provenienti da telecamere, sensori e così via.	Archiviazione condivisa che contiene modelli di inferenza e dati provenienti da dispositivi edge per analisi successive.

Questa soluzione NetApp e Lenovo offre i seguenti vantaggi chiave:

- Elaborazione accelerata tramite GPU all'edge.
- Distribuzione di più server edge supportati e gestiti da uno storage condiviso.
- Protezione dati affidabile per soddisfare obiettivi di punto di ripristino (RPO) e obiettivi di tempo di ripristino (RTO) bassi senza perdita di dati.
- Gestione ottimizzata dei dati con copie e cloni NetApp Snapshot per semplificare i flussi di lavoro di sviluppo.

Come utilizzare questa architettura

Questo documento convalida la progettazione e le prestazioni dell'architettura proposta. Tuttavia, non abbiamo testato determinati componenti a livello di software, come la gestione di container, carichi di lavoro o modelli e la sincronizzazione dei dati con il cloud o il data center in locale, perché sono specifici di uno scenario di distribuzione. Qui le scelte sono molteplici.

A livello di gestione dei container, la gestione dei container Kubernetes è una buona scelta ed è ben supportata sia in una versione completamente upstream (Canonical) sia in una versione modificata adatta alle distribuzioni aziendali (Red Hat). IL"Piano di controllo AI NetApp" che utilizza NetApp Trident e il nuovo aggiunto "Kit di strumenti NetApp DataOps" fornisce tracciabilità integrata, funzioni di gestione dei dati, interfacce e strumenti che consentono a data scientist e data engineer di integrarsi con lo storage NetApp . Kubeflow, il toolkit ML per Kubernetes, fornisce funzionalità di intelligenza artificiale aggiuntive insieme al supporto per il controllo delle versioni dei modelli e KFServing su diverse piattaforme come TensorFlow Serving o NVIDIA Triton Inference Server. Un'altra opzione è la piattaforma NVIDIA EGX, che fornisce la gestione del carico di lavoro insieme all'accesso a un catalogo di contenitori di inferenza AI abilitati per GPU. Tuttavia, queste opzioni potrebbero richiedere notevoli sforzi e competenze per essere messe in produzione e potrebbero richiedere l'assistenza di un fornitore di software indipendente (ISV) o di un consulente di terze parti.

Aree di soluzione

Il vantaggio principale dell'inferenza AI e dell'edge computing è la capacità dei dispositivi di calcolare, elaborare e analizzare i dati con un elevato livello di qualità e senza latenza. Gli esempi di casi d'uso dell'edge

computing sono davvero troppi per essere descritti in questo documento, ma eccone alcuni tra i più importanti:

Automobili: veicoli autonomi

L'esempio classico dell'edge computing è rappresentato dai sistemi avanzati di assistenza alla guida (ADAS) nei veicoli autonomi (AV). L'intelligenza artificiale nelle auto senza conducente deve elaborare rapidamente una grande quantità di dati provenienti da telecamere e sensori per garantire la sicurezza dei conducenti. Impiegare troppo tempo per interpretare i dati tra un oggetto e un essere umano può significare vita o morte, quindi è fondamentale riuscire a elaborare tali dati il più vicino possibile al veicolo. In questo caso, uno o più server di edge computing gestiscono l'input proveniente da telecamere, RADAR, LiDAR e altri sensori, mentre l'archiviazione condivisa contiene modelli di inferenza e memorizza i dati di input provenienti dai sensori.

Assistenza sanitaria: monitoraggio dei pazienti

Uno degli impatti più significativi dell'intelligenza artificiale e dell'edge computing è la loro capacità di migliorare il monitoraggio continuo dei pazienti affetti da malattie croniche, sia nell'assistenza domiciliare che nelle unità di terapia intensiva (UTI). I dati provenienti da dispositivi edge che monitorano i livelli di insulina, la respirazione, l'attività neurologica, il ritmo cardiaco e le funzioni gastrointestinali richiedono un'analisi istantanea dei dati su cui è necessario intervenire immediatamente, perché il tempo a disposizione per salvare la vita di qualcuno è limitato.

Vendita al dettaglio: pagamento senza cassiere

L'edge computing può potenziare l'intelligenza artificiale e l'apprendimento automatico per aiutare i rivenditori a ridurre i tempi di pagamento e ad aumentare il traffico pedonale. I sistemi senza cassiere supportano vari componenti, come i seguenti:

- Autenticazione e accesso. Collegare l'acquirente fisico a un account convalidato e consentire l'accesso allo spazio di vendita.
- Monitoraggio dell'inventario. Utilizzo di sensori, tag RFID e sistemi di visione artificiale per aiutare gli acquirenti a confermare la selezione o la deselezione degli articoli.

  Qui, ciascuno dei server edge gestisce ogni cassa e il sistema di archiviazione condiviso funge da punto di sincronizzazione centrale.

Servizi finanziari: sicurezza delle persone ai chioschi e prevenzione delle frodi

Le organizzazioni bancarie utilizzano l'intelligenza artificiale e l'edge computing per innovare e creare esperienze bancarie personalizzate. I chioschi interattivi che utilizzano l'analisi dei dati in tempo reale e l'inferenza basata sull'intelligenza artificiale consentono ora agli sportelli bancomat non solo di aiutare i clienti a prelevare denaro, ma anche di monitorare proattivamente i chioschi attraverso le immagini catturate dalle telecamere per identificare rischi per la sicurezza umana o comportamenti fraudolenti. In questo scenario, i server di edge computing e i sistemi di storage condiviso sono collegati a chioschi interattivi e telecamere per aiutare le banche a raccogliere ed elaborare dati con modelli di inferenza basati sull'intelligenza artificiale.

Produzione: Industria 4.0

La quarta rivoluzione industriale (Industria 4.0) è iniziata, insieme a tendenze emergenti come la Smart Factory e la stampa 3D. Per prepararsi a un futuro basato sui dati, la comunicazione machine-to-machine (M2M) su larga scala e l'IoT vengono integrati per una maggiore automazione senza la necessità dell'intervento umano. La produzione è già altamente automatizzata e l'aggiunta di funzionalità di intelligenza artificiale è la naturale continuazione di questa tendenza a lungo termine. L'intelligenza artificiale consente di automatizzare operazioni che possono essere automatizzate con l'ausilio della visione artificiale e di altre funzionalità dell'intelligenza artificiale. È possibile automatizzare il controllo qualità o le attività che si basano sulla visione

umana o sul processo decisionale per eseguire analisi più rapide dei materiali sulle linee di assemblaggio negli stabilimenti produttivi, aiutando così gli stabilimenti di produzione a soddisfare gli standard ISO richiesti in materia di sicurezza e gestione della qualità. Qui, ogni server edge di elaborazione è connesso a una serie di sensori che monitorano il processo di produzione e i modelli di inferenza aggiornati vengono inviati all'archiviazione condivisa, secondo necessità.

Telecomunicazioni: rilevamento della ruggine, ispezione delle torri e ottimizzazione della rete

Il settore delle telecomunicazioni utilizza tecniche di visione artificiale e intelligenza artificiale per elaborare immagini che rilevano automaticamente la ruggine e identificano le torri cellulari che presentano corrosione e, pertanto, necessitano di ulteriori ispezioni. Negli ultimi anni è aumentato l'uso di immagini di droni e modelli di intelligenza artificiale per identificare le diverse aree di una torre e analizzarne ruggine, crepe superficiali e corrosione. La domanda di tecnologie di intelligenza artificiale (IA) che consentano di ispezionare in modo efficiente le infrastrutture di telecomunicazione e le torri cellulari, di valutarne regolarmente il degrado e di ripararle tempestivamente quando necessario è in continua crescita.

Inoltre, un altro caso d'uso emergente nel settore delle telecomunicazioni è l'uso di algoritmi di intelligenza artificiale e apprendimento automatico per prevedere modelli di traffico dati, rilevare dispositivi compatibili con il 5G e automatizzare e potenziare la gestione energetica MIMO (multiple-input and multiple-output). L'hardware MIMO viene utilizzato nelle torri radio per aumentare la capacità della rete; tuttavia, ciò comporta costi energetici aggiuntivi. I modelli di apprendimento automatico per la "modalità di sospensione MIMO" implementati nei siti cellulari possono prevedere l'uso efficiente delle radio e contribuire a ridurre i costi di consumo energetico per gli operatori di rete mobile (MNO). Le soluzioni di inferenza AI e di edge computing aiutano gli operatori di rete mobile a ridurre la quantità di dati trasmessi avanti e indietro ai data center, ad abbassare il costo totale di proprietà (TCO), a ottimizzare le operazioni di rete e a migliorare le prestazioni complessive per gli utenti finali.

Conclusione

L'automazione basata sull'intelligenza artificiale e l'edge computing rappresentano un approccio leader per aiutare le aziende a realizzare la trasformazione digitale e massimizzare l'efficienza operativa e la sicurezza. Grazie all'edge computing, i dati vengono elaborati molto più velocemente perché non devono viaggiare da e verso un data center. Di conseguenza, i costi associati all'invio e alla ricezione dei dati ai data center o al cloud risultano ridotti. Una latenza inferiore e una maggiore velocità possono rivelarsi vantaggiose quando le aziende devono prendere decisioni quasi in tempo reale utilizzando modelli di inferenza AI implementati all'edge.

I sistemi di storage NetApp garantiscono prestazioni uguali o migliori rispetto allo storage SSD locale e offrono i seguenti vantaggi a data scientist, data engineer, sviluppatori di intelligenza artificiale/apprendimento automatico e responsabili delle decisioni aziendali o IT:

- Condivisione semplice dei dati tra sistemi di intelligenza artificiale, analisi e altri sistemi aziendali critici. Questa condivisione dei dati riduce il sovraccarico dell'infrastruttura, migliora le prestazioni e semplifica la gestione dei dati in tutta l'azienda.
- Elaborazione e archiviazione scalabili in modo indipendente per ridurre al minimo i costi e migliorare l'utilizzo delle risorse.
- Flussi di lavoro di sviluppo e distribuzione semplificati mediante copie Snapshot e cloni integrati per spazi di lavoro utente istantanei e salvaspazio, controllo delle versioni integrato e distribuzione automatizzata.
- Protezione dei dati di livello aziendale per il disaster recovery e la continuità aziendale. La soluzione NetApp e Lenovo presentata in questo documento è un'architettura flessibile e scalabile, ideale per

implementazioni di inferenza AI di livello aziendale in ambito edge.

Ringraziamenti

- JJ Falkanger, Direttore senior, Soluzioni HPC e IA, Lenovo
- Dave Arnette, ingegnere tecnico di marketing, NetApp
- Joey Parnell, responsabile tecnico delle soluzioni di intelligenza artificiale della serie E, NetApp
- Cody Harryman, ingegnere QA, NetApp

Dove trovare ulteriori informazioni

Per saperne di più sulle informazioni descritte nel presente documento, fare riferimento ai seguenti documenti e/o siti web:

- Pagina del prodotto degli array NetApp AFF A-Series

  "https://www.netapp.com/data-storage/aff-a-series/"

- Software di gestione dati NetApp ONTAP: libreria di informazioni ONTAP 9

  http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286

- TR-4727: Introduzione alla serie EF NetApp

  https://www.netapp.com/pdf.html?item=/media/17179-tr4727pdf.pdf

- Scheda tecnica del software NetApp E-Series SANtricity

  https://www.netapp.com/pdf.html?item=/media/19775-ds-3171-66862.pdf

- NetApp Persistent Storage per container: NetApp Trident

  "https://netapp.io/persistent-storage-provisioner-for-kubernetes/"

- MLPerf
  - "https://mlcommons.org/en/"
  - "http://www.image-net.org/"
  - "https://mlcommons.org/en/news/mlperf-inference-v11/"
- Copia e sincronizzazione NetApp BlueXP

  "https://docs.netapp.com/us-en/occm/concept_cloud_sync.html#how-cloud-sync-works"

- Benchmark di TensorFlow

  "https://github.com/tensorflow/benchmarks"

- Server edge Lenovo ThinkSystem SE350

  "https://lenovopress.com/lp1168"

- Array di archiviazione flash unificato Lenovo ThinkSystem DM5100F

"https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array"

Informazioni sul copyright

Informazioni sul marchio commerciale