EF-Series ai con NVIDIA

NetApp Solutions

NetApp
September 10, 2024

This PDF was generated from https://docs.netapp.com/it-it/netapp-solutions/ai/ai-dgx-superpod.html on September 10, 2024. Always check docs.netapp.com for the latest.

Sommario

NetApp EF-Series ai con NVIDIA 1
EF-Series ai con sistemi NVIDIA DGX A100 e BeeGFS 1
NVIDIA DGX SuperPOD con NetApp - Guida alla progettazione 1
NVA-1156-DESIGN: NetApp EF-Series ai con sistemi NVIDIA DGX A100 e BeeGFS 5
NVA-1156-DEPLOY: NetApp EF-Series ai con sistemi NVIDIA DGX A100 e BeeGFS 5

NetApp EF-Series ai con NVIDIA

Panoramica delle soluzioni di infrastruttura convergente EF-Series ai di NetApp e NVIDIA.

EF-Series ai con sistemi NVIDIA DGX A100 e BeeGFS

- "Guida alla progettazione"
- "Guida all'implementazione"
- "Guida all'implementazione di BeeGFS"

NVIDIA DGX SuperPOD con NetApp - Guida alla progettazione

NVIDIA DGX SuperPOD con NetApp - Guida alla progettazione

Amine Bennani, David Arnette e Satish Thyagarajan, NetApp

Sintesi

Anche se l'ai migliora la vita dei consumatori e aiuta le organizzazioni di tutti i settori di tutto il mondo a innovare e far crescere i propri business, può rappresentare una rivoluzione per L'IT. Per supportare il business, i reparti IT stanno cercando di implementare soluzioni di high-performance computing (HPC) in grado di soddisfare le esigenze estreme dei workload ai. Con la gara per l'ai che entra nel vivo, la necessità di una soluzione facile da implementare, scalare e gestire diventa sempre più urgente.

NVIDIA DGX SuperPOD rende l'infrastruttura di supercalcolo facilmente accessibile per qualsiasi organizzazione e fornisce l'estrema potenza computazionale necessaria per risolvere anche i problemi ai più complessi. Per aiutare i clienti a eseguire un'implementazione su larga scala oggi, questa soluzione pronta all'uso di NVIDIA e NetApp elimina la complessità e le approssimazioni dalla progettazione dell'infrastruttura e offre una soluzione completa e validata che comprende calcolo, rete, storage e software Best-in-class.

Riepilogo del programma

NVIDIA DGX SuperPOD con i sistemi NVIDIA DGX H100 e NVIDIA base Command riunisce una combinazione ottimizzata per il design di calcolo ai, fabric di rete, storage, software e supporto. BeeGFS sull'architettura NetApp è stato precedentemente validato in un cluster di accettazione dedicato di NVIDIA. L'architettura più recente estende tale processo di validazione mantenendo un design comprovato, incorporando al tempo stesso il supporto per l'hardware più recente di NVIDIA.

Panoramica della soluzione

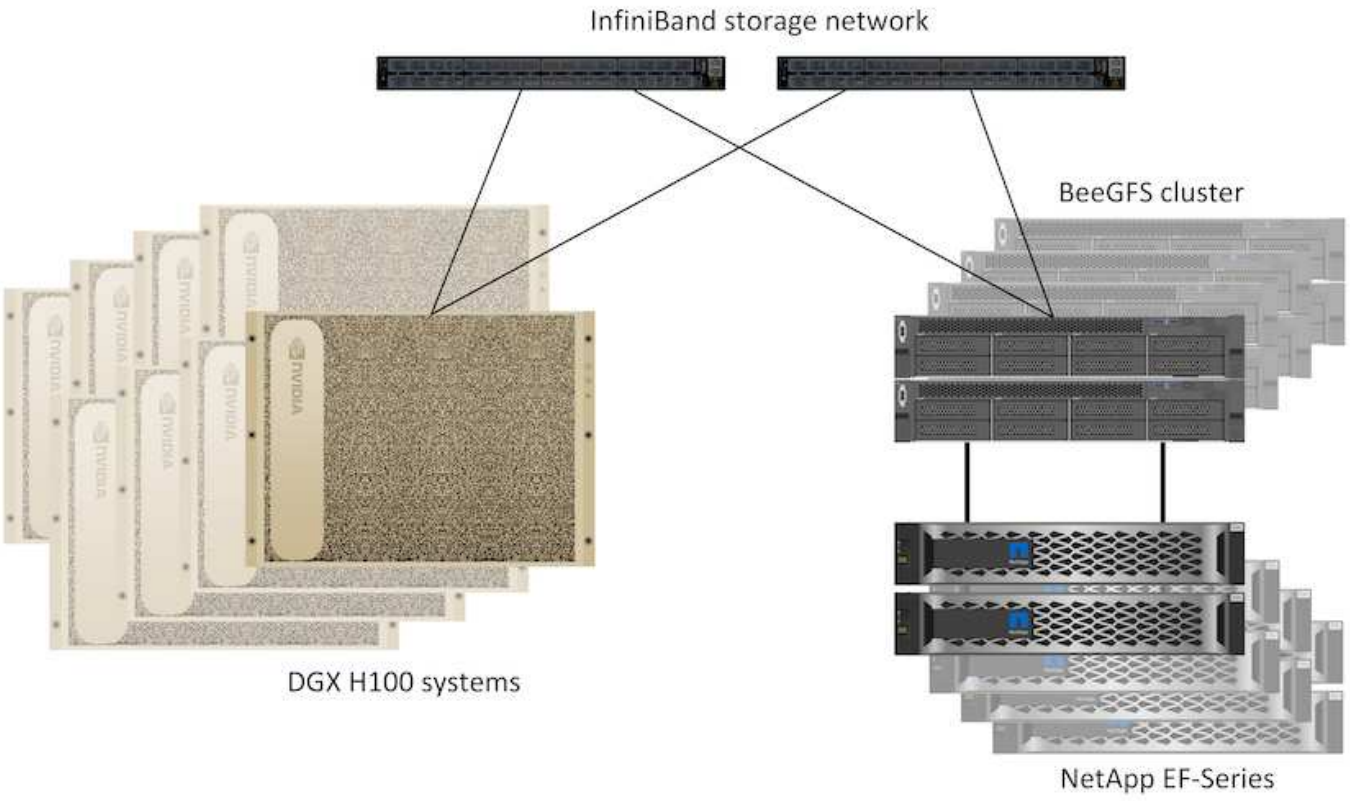
NVIDIA DGX SuperPOD è una piattaforma infrastrutturale per data center ai fornita come soluzione pronta all'uso per L'IT, che supporta i workload ai più complessi che devono affrontare le aziende moderne. Semplifica l'implementazione e la gestione e offre scalabilità a performance e capacità virtualmente illimitate. In altre parole, DGX SuperPOD ti permette di concentrarti sui dati utili invece che sull'infrastruttura.

Con gli array all-flash NetApp EF600 alla base di NVIDIA DGX SuperPOD, i clienti ottengono una soluzione ai agile che scala in modo semplice e perfetto. La flessibilità e la scalabilità della soluzione consentono all'IT di supportare e adattarsi ai carichi di lavoro in continua evoluzione, rappresentando una base solida per soddisfare i requisiti di storage attuali e futuri. I building block di storage modulari consentono un approccio granulare alla crescita e scalare senza problemi da terabyte a petabyte. Aumentando il numero di building block di storage, i clienti possono scalare in verticale performance e capacità del file system, consentendo alla soluzione di gestire con facilità i carichi di lavoro più estremi.

Tecnologia della soluzione

- I sistemi NVIDIA DGX SuperPOD con NVIDIA DGX H100 sfruttano i sistemi DGX H100 con storage condiviso convalidato esterno:
  - Ogni unità scalabile DGX SuperPOD (su) è composta da 32 sistemi DGX H100 ed è in grado di erogare 640 petabyte di performance ai con una precisione di FP8. In genere contiene almeno due building block NetApp BeeGFS, a seconda dei requisiti di prestazioni e capacità di una particolare installazione.

Una visione di alto livello della soluzione

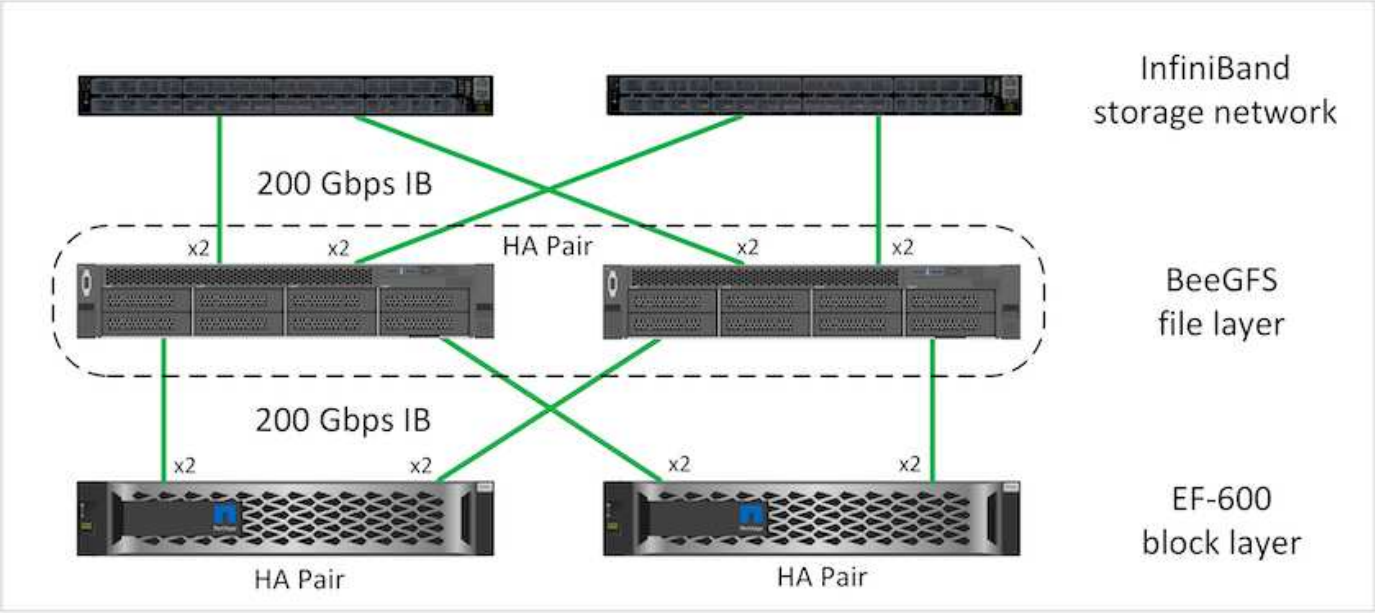


- I building block di NetApp BeeGFS sono costituiti da due array NetApp EF600 e due server x86:
  - Con gli array all-flash NetApp EF600 alla base di NVIDIA DGX SuperPOD, i clienti ottengono una base storage affidabile supportata da sei 9s di uptime.

- Il livello del file system tra i sistemi NetApp EF600 e NVIDIA DGX H100 è il file system parallelo BeeGFS. BeeGFS è stato creato dal Fraunhofer Center for High-Performance Computing in Germania per risolvere i punti critici dei file system paralleli legacy. Il risultato è un file system con una moderna architettura dello spazio utente, ora sviluppata e fornita da ThinkParQ e utilizzata da molti ambienti di supercomputing.
- Il supporto di NetApp per BeeGFS allinea l'eccellente organizzazione di supporto di NetApp con i requisiti dei clienti in termini di performance e uptime. I clienti possono accedere a risorse di supporto di livello superiore, all'accesso anticipato alle release di BeeGFS e a funzioni aziendali di BeeGFS selezionate, come l'applicazione delle quote e l'alta disponibilità (ha).

- La combinazione dei building block NVIDIA SuperPOD SUS e NetApp BeeGFS offre una soluzione ai agile in cui si scala in modo semplice e perfetto le risorse di calcolo o storage.

Building block BeeGFS di NetApp



Riepilogo dei casi d'utilizzo

Questa soluzione si applica ai seguenti casi di utilizzo:

- Intelligenza artificiale (ai) che include machine learning (ML), deep learning (DL), elaborazione del linguaggio naturale (NLP), comprensione del linguaggio naturale (NLU) e ai generativa (Genai).
- Formazione ai su media e grande scala
- Computer vision, voce, audio e modelli linguistici
- HPC include applicazioni accelerate tramite l'interfaccia di trasmissione messaggi (MPI) e altre tecniche di elaborazione distribuita
- Workload delle applicazioni caratterizzati da:
  - Lettura o scrittura su file di dimensioni superiori a 1 GB
  - Lettura o scrittura sullo stesso file da parte di più client (10, 100 e 1000)
- Set di dati da più terabyte o petabyte
- Ambienti che richiedono un unico spazio dei nomi di storage ottimizzabile per una combinazione di file di grandi e piccole dimensioni

Requisiti tecnologici

Questa sezione descrive i requisiti tecnologici della soluzione NVIDIA DGX SuperPOD con NetApp.

Requisiti hardware

La tabella 1 riportata di seguito elenca i componenti hardware necessari per implementare la soluzione per una singola su. Il dimensionamento della soluzione inizia con 32 sistemi NVIDIA DGX H100 e due o tre building block NetApp BeeGFS.
Un singolo building block BeeGFS di NetApp è composto da due array NetApp EF600 e due server x86. I clienti possono aggiungere building block aggiuntivi con l'aumento delle dimensioni dell'implementazione. Per ulteriori informazioni, consultare "Architettura di riferimento NVIDIA DGX H100 SuperPOD" e. "NVA-1164-DESIGN: BeeGFS su NetApp NVA Design".

Hardware	Quantità
NVIDIA DGX H100	32
Switch NVIDIA Quantum QM9700	8 foglia, 4 colonna vertebrale
Blocchi di costruzione BeeGFS NetApp	3

Requisiti software

La tabella 2 riportata di seguito elenca i componenti software necessari per implementare la soluzione. I componenti software utilizzati in una particolare implementazione della soluzione possono variare in base ai requisiti del cliente.

Software
Stack software NVIDIA DGX
NVIDIA base Command Manager
File system parallelo ThinkParQ BeeGFS

Verifica della soluzione

NVIDIA DGX SuperPOD con NetApp è stato validato in un cluster di accettazione dedicato presso NVIDIA utilizzando i building block NetApp BeeGFS. I criteri di accettazione si basano su una serie di test applicativi, prestazionali e di stress eseguiti da NVIDIA. Per ulteriori informazioni, consultare "NVIDIA DGX SuperPOD: Architettura di riferimento NetApp EF600 e BeeGFS".

Conclusione

NetApp e NVIDIA hanno una lunga storia di collaborazione per fornire un portfolio di soluzioni ai sul mercato. NVIDIA DGX SuperPOD con l'array all-flash NetApp EF600 è una soluzione comprovata e validata che i clienti possono implementare in tutta sicurezza. Questa architettura completamente integrata e pronta all'uso elimina i rischi legati all'implementazione e aiuta chiunque a vincere la corsa alla leadership dell'ai.

Dove trovare ulteriori informazioni

Per ulteriori informazioni sulle informazioni descritte in questo documento, consultare i seguenti documenti e/o siti Web:

- "Architettura di riferimento NVIDIA DGX SuperPOD"
- "Guida di riferimento per la progettazione di data center NVIDIA DGX SuperPOD"
- "NVIDIA DGX SuperPOD: NetApp EF600 e BeeGFS"

NVA-1156-DESIGN: NetApp EF-Series ai con sistemi NVIDIA DGX A100 e BeeGFS

Abdel Sadek, Tim Chau, Joe McCormick e David Arnette, NetApp

NVA-1156-DESIGN descrive un'architettura verificata di NetApp per i carichi di lavoro di machine learning (ML) e intelligenza artificiale (ai) che utilizzano i sistemi storage NetApp EF600 NVMe, il file system parallelo BeeGFS, i sistemi NVIDIA DGX A100 e gli switch IB NVIDIA Mellanox Quantum QM8700 a 200 Gbps. Questo design è dotato di InfiniBand (IB) a 200 Gbps per lo storage e il fabric di interconnessione del cluster di calcolo per offrire ai clienti un'architettura completamente basata su IB per carichi di lavoro dalle performance elevate. Questo documento include anche i risultati dei test di benchmark per l'architettura implementata.

"NVA-1156-DESIGN: NetApp EF-Series ai con sistemi NVIDIA DGX A100 e BeeGFS"

NVA-1156-DEPLOY: NetApp EF-Series ai con sistemi NVIDIA DGX A100 e BeeGFS

Abdel Sadek, Tim Chau, Joe McCormick e David Arnette, NetApp

Questo documento descrive un'architettura verificata di NetApp per i carichi di lavoro di apprendimento automatico (ML) e intelligenza artificiale (ai) che utilizzano i sistemi storage NetApp EF600 NVMe, il file system parallelo ThinkParQ BeeGFS, i sistemi NVIDIA DGX A100 e gli switch NVIDIA Mellanox Quantum QM8700 200Gbps InfiniBand (IB). Questo documento include anche istruzioni per l'esecuzione dei test di benchmark di convalida al termine dell'implementazione.

"NVA-1156-DEPLOY: NetApp EF-Series ai con sistemi NVIDIA DGX A100 e BeeGFS"

Informazioni sul copyright

Copyright © 2024 NetApp, Inc. Tutti i diritti riservati. Stampato negli Stati Uniti d'America. Nessuna porzione di questo documento soggetta a copyright può essere riprodotta in qualsiasi formato o mezzo (grafico, elettronico o meccanico, inclusi fotocopie, registrazione, nastri o storage in un sistema elettronico) senza previo consenso scritto da parte del detentore del copyright.

Il software derivato dal materiale sottoposto a copyright di NetApp è soggetto alla seguente licenza e dichiarazione di non responsabilità:

IL PRESENTE SOFTWARE VIENE FORNITO DA NETAPP "COSÌ COM'È" E SENZA QUALSIVOGLIA TIPO DI GARANZIA IMPLICITA O ESPRESSA FRA CUI, A TITOLO ESEMPLIFICATIVO E NON ESAUSTIVO, GARANZIE IMPLICITE DI COMMERCIABILITÀ E IDONEITÀ PER UNO SCOPO SPECIFICO, CHE VENGONO DECLINATE DAL PRESENTE DOCUMENTO. NETAPP NON VERRÀ CONSIDERATA RESPONSABILE IN ALCUN CASO PER QUALSIVOGLIA DANNO DIRETTO, INDIRETTO, ACCIDENTALE, SPECIALE, ESEMPLARE E CONSEQUENZIALE (COMPRESI, A TITOLO ESEMPLIFICATIVO E NON ESAUSTIVO, PROCUREMENT O SOSTITUZIONE DI MERCI O SERVIZI, IMPOSSIBILITÀ DI UTILIZZO O PERDITA DI DATI O PROFITTI OPPURE INTERRUZIONE DELL'ATTIVITÀ AZIENDALE) CAUSATO IN QUALSIVOGLIA MODO O IN RELAZIONE A QUALUNQUE TEORIA DI RESPONSABILITÀ, SIA ESSA CONTRATTUALE, RIGOROSA O DOVUTA A INSOLVENZA (COMPRESA LA NEGLIGENZA O ALTRO) INSORTA IN QUALSIASI MODO ATTRAVERSO L'UTILIZZO DEL PRESENTE SOFTWARE ANCHE IN PRESENZA DI UN PREAVVISO CIRCA L'EVENTUALITÀ DI QUESTO TIPO DI DANNI.

NetApp si riserva il diritto di modificare in qualsiasi momento qualunque prodotto descritto nel presente documento senza fornire alcun preavviso. NetApp non si assume alcuna responsabilità circa l'utilizzo dei prodotti o materiali descritti nel presente documento, con l'eccezione di quanto concordato espressamente e per iscritto da NetApp. L'utilizzo o l'acquisto del presente prodotto non comporta il rilascio di una licenza nell'ambito di un qualche diritto di brevetto, marchio commerciale o altro diritto di proprietà intellettuale di NetApp.

Il prodotto descritto in questa guida può essere protetto da uno o più brevetti degli Stati Uniti, esteri o in attesa di approvazione.

LEGENDA PER I DIRITTI SOTTOPOSTI A LIMITAZIONE: l'utilizzo, la duplicazione o la divulgazione da parte degli enti governativi sono soggetti alle limitazioni indicate nel sottoparagrafo (b)(3) della clausola Rights in Technical Data and Computer Software del DFARS 252.227-7013 (FEB 2014) e FAR 52.227-19 (DIC 2007).

I dati contenuti nel presente documento riguardano un articolo commerciale (secondo la definizione data in FAR 2.101) e sono di proprietà di NetApp, Inc. Tutti i dati tecnici e il software NetApp forniti secondo i termini del presente Contratto sono articoli aventi natura commerciale, sviluppati con finanziamenti esclusivamente privati. Il governo statunitense ha una licenza irrevocabile limitata, non esclusiva, non trasferibile, non cedibile, mondiale, per l'utilizzo dei Dati esclusivamente in connessione con e a supporto di un contratto governativo statunitense in base al quale i Dati sono distribuiti. Con la sola esclusione di quanto indicato nel presente documento, i Dati non possono essere utilizzati, divulgati, riprodotti, modificati, visualizzati o mostrati senza la previa approvazione scritta di NetApp, Inc. I diritti di licenza del governo degli Stati Uniti per il Dipartimento della Difesa sono limitati ai diritti identificati nella clausola DFARS 252.227-7015(b) (FEB 2014).

Informazioni sul marchio commerciale

NETAPP, il logo NETAPP e i marchi elencati alla pagina http://www.netapp.com/TM sono marchi di NetApp, Inc. Gli altri nomi di aziende e prodotti potrebbero essere marchi dei rispettivi proprietari.