NetApp AIPod con sistemi NVIDIA DGX

NetApp artificial intelligence solutions

NetApp
February 12, 2026

Sommario

NetApp AIPod con sistemi NVIDIA DGX

NVA-1173 NetApp AIPod con sistemi NVIDIA DGX - Introduzione

POWERED BY

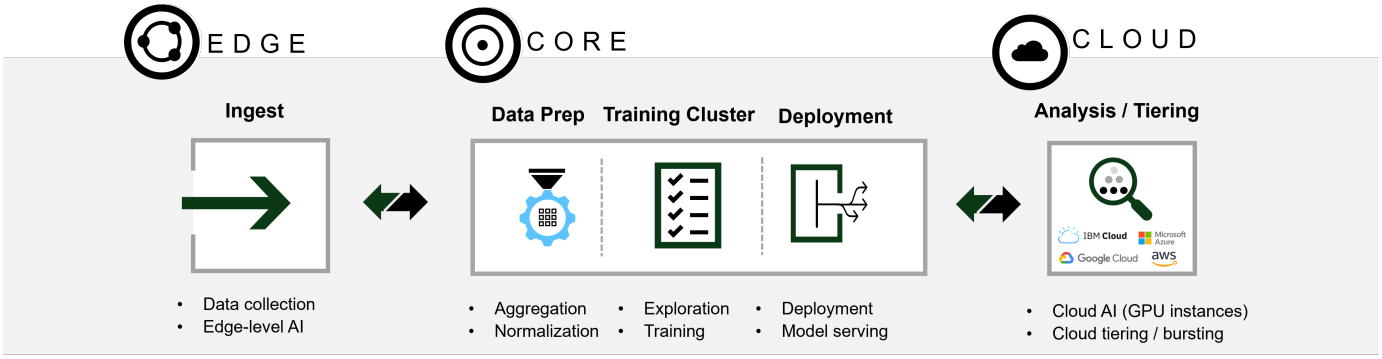
NVIDIA

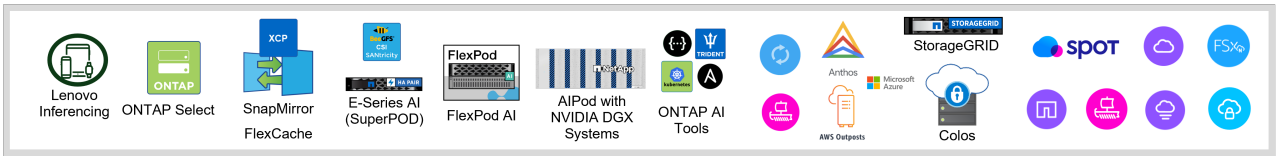
Ingegneria delle soluzioni NetApp

Sintesi

L' AIPod NetApp con sistemi NVIDIA DGX e sistemi di storage NetApp connessi al cloud semplifica le distribuzioni dell'infrastruttura per carichi di lavoro di apprendimento automatico (ML) e intelligenza artificiale (AI), eliminando la complessità di progettazione e le congetture. Basandosi sul design NVIDIA DGX BasePOD per offrire prestazioni di elaborazione eccezionali per carichi di lavoro di nuova generazione, AIPod con sistemi NVIDIA DGX aggiunge sistemi di storage NetApp AFF che consentono ai clienti di iniziare in piccolo e crescere senza interruzioni, gestendo in modo intelligente i dati dall'edge al core, al cloud e viceversa. NetApp AIPod fa parte del più ampio portafoglio di soluzioni AI NetApp , illustrato nella figura seguente.

Portafoglio di soluzioni AI NetApp





Questo documento descrive i componenti chiave dell'architettura di riferimento AIPod , le informazioni sulla connettività e sulla configurazione del sistema, i risultati dei test di convalida e le indicazioni per il dimensionamento della soluzione. Questo documento è destinato agli ingegneri delle soluzioni NetApp e dei partner e ai decisori strategici dei clienti interessati a implementare un'infrastruttura ad alte prestazioni per

carichi di lavoro di ML/DL e analisi.

NVA-1173 NetApp AIPod con sistemi NVIDIA DGX - Componenti hardware

Questa sezione si concentra sui componenti hardware per NetApp AIPod con sistemi NVIDIA DGX.

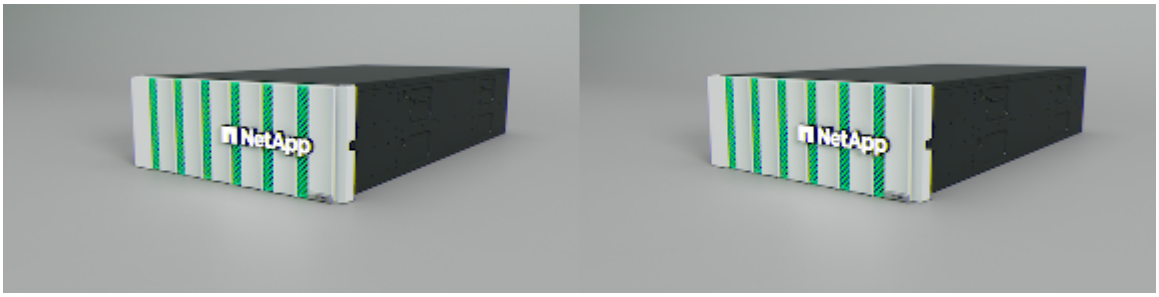
Sistemi di archiviazione NetApp AFF

I sistemi di storage all'avanguardia NetApp AFF consentono ai reparti IT di soddisfare i requisiti di storage aziendale con prestazioni leader del settore, flessibilità superiore, integrazione cloud e la migliore gestione dei dati della categoria. Progettati specificamente per flash, i sistemi AFF aiutano ad accelerare, gestire e proteggere i dati aziendali critici.

Sistemi di stoccaggio AFF A90

NetApp AFF A90 , basato sul software di gestione dati NetApp ONTAP , offre protezione dati integrata, funzionalità anti-ransomware opzionali e le elevate prestazioni e la resilienza necessarie per supportare i carichi di lavoro aziendali più critici. Elimina le interruzioni delle operazioni mission-critical, riduce al minimo l'ottimizzazione delle prestazioni e protegge i dati dagli attacchi ransomware. Offre: • Prestazioni leader del settore • Sicurezza dei dati senza compromessi • Aggiornamenti semplificati e senza interruzioni

_Sistema di archiviazione NetApp AFF A90



Prestazioni leader del settore

L' AFF A90 gestisce facilmente carichi di lavoro di nuova generazione come deep learning, intelligenza artificiale e analisi ad alta velocità, nonché database aziendali tradizionali come Oracle, SAP HANA, Microsoft SQL Server e applicazioni virtualizzate. Mantiene le applicazioni aziendali critiche in esecuzione alla massima velocità con un massimo di 2,4 milioni di IOPS per coppia HA e una latenza fino a 100 µs, e aumenta le prestazioni fino al 50% rispetto ai precedenti modelli NetApp . Grazie a NFS su RDMA, pNFS e Session Trunking, i clienti possono raggiungere l'elevato livello di prestazioni di rete richiesto per le applicazioni di nuova generazione utilizzando l'infrastruttura di rete del data center esistente. I clienti possono inoltre scalare e crescere con il supporto multiprotocollo unificato per SAN, NAS e storage di oggetti e garantire la massima

flessibilità con un software di gestione dati ONTAP unificato e singolo, per i dati in sede o nel cloud. Inoltre, lo stato di salute del sistema può essere ottimizzato con analisi predittive basate sull'intelligenza artificiale fornite da Active IQ e Cloud Insights.

Sicurezza dei dati senza compromessi

I sistemi AFF A90 contengono una suite completa di software di protezione dei dati NetApp integrato e coerente con le applicazioni. Fornisce protezione dei dati integrata e soluzioni anti-ransomware all'avanguardia per il recupero preventivo e post-attacco. È possibile bloccare la scrittura su disco dei file dannosi e monitorare facilmente le anomalie di archiviazione per ottenere informazioni.

Aggiornamenti semplificati e non distruttivi

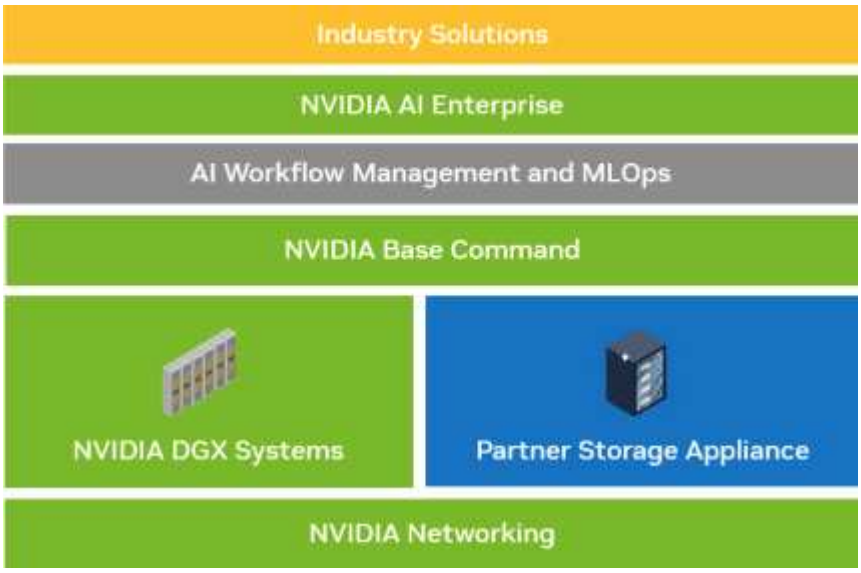
L' AFF A90 è disponibile come aggiornamento non invasivo del telaio per i clienti A800 esistenti. NetApp semplifica l'aggiornamento e l'eliminazione delle interruzioni nelle operazioni mission-critical grazie alle nostre funzionalità avanzate di affidabilità, disponibilità, manutenibilità e gestibilità (RASM). Inoltre, NetApp aumenta ulteriormente l'efficienza operativa e semplifica le attività quotidiane dei team IT, poiché il software ONTAP applica automaticamente gli aggiornamenti del firmware per tutti i componenti del sistema.

Per le distribuzioni più grandi, i sistemi AFF A1K offrono le prestazioni e le capacità più elevate, mentre altri sistemi di storage NetApp , come AFF A70 e AFF C800, offrono opzioni per distribuzioni più piccole a costi inferiori.

NVIDIA DGX BasePOD

NVIDIA DGX BasePOD è una soluzione integrata composta da componenti hardware e software NVIDIA , soluzioni MLOps e storage di terze parti. Sfruttando le migliori pratiche di progettazione di sistemi scalabili con prodotti NVIDIA e soluzioni partner convalidate, i clienti possono implementare una piattaforma efficiente e gestibile per lo sviluppo dell'intelligenza artificiale. La figura 1 evidenzia i vari componenti di NVIDIA DGX BasePOD.

Soluzione NVIDIA DGX BasePOD



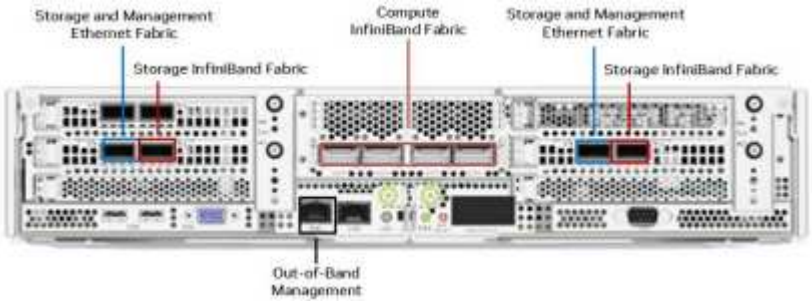
Sistemi NVIDIA DGX H100

Il sistema NVIDIA DGX H100 è un concentrato di intelligenza artificiale accelerato dalle prestazioni rivoluzionarie della GPU NVIDIA H100 Tensor Core.

Sistema NVIDIA DGX H100

Le specifiche principali del sistema DGX H100 sono: • Otto GPU NVIDIA H100. • 80 GB di memoria GPU per GPU, per un totale di 640 GB. • Quattro chip NVIDIA NVSwitch. • Doppi processori Intel Xeon Platinum 8480 a 56 core con supporto PCIe 5.0. • 2 TB di memoria di sistema DDR5. • Quattro porte OSFP che servono otto adattatori NVIDIA ConnectX-7 (InfiniBand/Ethernet) a porta singola e due adattatori NVIDIA ConnectX-7 (InfiniBand/Ethernet) a doppia porta. • Due unità M.2 NVMe da 1,92 TB per DGX OS, otto unità U.2 NVMe da 3,84 TB per archiviazione/cache. • Potenza massima 10,2 kW. Di seguito sono illustrate le porte posteriori del vassoio CPU DGX H100. Quattro delle porte OSFP servono otto adattatori ConnectX-7 per la struttura di elaborazione InfiniBand. Ogni coppia di adattatori ConnectX-7 a doppia porta fornisce percorsi paralleli verso le strutture di archiviazione e gestione. La porta fuori banda viene utilizzata per l'accesso BMC .

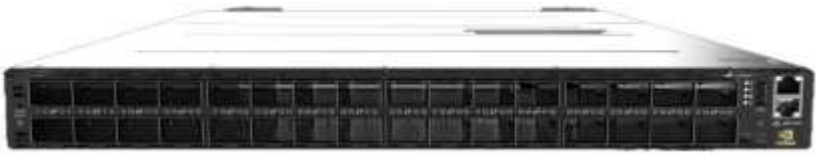
Pannello posteriore NVIDIA DGX H100



NVIDIA

Switch NVIDIA Quantum-2 QM9700

Switch NVIDIA Quantum-2 QM9700 InfiniBand



Gli switch NVIDIA Quantum-2 QM9700 con connettività InfiniBand da 400 Gb/s alimentano la struttura di elaborazione nelle configurazioni NVIDIA Quantum-2 InfiniBand BasePOD. Per la struttura di elaborazione InfiniBand vengono utilizzati gli adattatori a porta singola ConnectX-7. Ogni sistema NVIDIA DGX è dotato di doppie connessioni per ogni switch QM9700, fornendo più percorsi ad alta larghezza di banda e bassa latenza tra i sistemi.

Switch NVIDIA Spectrum-3 SN4600

Switch NVIDIA Spectrum-3 SN4600

Gli switch NVIDIA Spectrum-3 SN4600 offrono 128 porte totali (64 per switch) per fornire connettività ridondante per la gestione in-band del DGX BasePOD. Lo switch NVIDIA SN4600 può fornire velocità comprese tra 1 GbE e 200 GbE. Per gli apparecchi di archiviazione collegati tramite Ethernet vengono utilizzati anche gli switch NVIDIA SN4600. Le porte sugli adattatori ConnectX-7 a doppia porta NVIDIA DGX vengono utilizzate sia per la gestione in banda che per la connettività di archiviazione.

Switch NVIDIA Spectrum SN2201

Switch NVIDIA Spectrum SN2201

Gli switch NVIDIA Spectrum SN2201 offrono 48 porte per garantire la connettività per la gestione fuori banda. La gestione fuori banda fornisce una connettività di gestione consolidata per tutti i componenti in DGX BasePOD.

Adattatore NVIDIA ConnectX-7

Adattatore NVIDIA ConnectX-7

L'adattatore NVIDIA ConnectX-7 può fornire una velocità di trasmissione di 25/50/100/200/400 G. I sistemi NVIDIA DGX utilizzano sia gli adattatori ConnectX-7 a porta singola che doppia per garantire flessibilità nelle distribuzioni DGX BasePOD con InfiniBand ed Ethernet da 400 Gb/s.

NVA-1173 NetApp AIPod con sistemi NVIDIA DGX - Componenti software

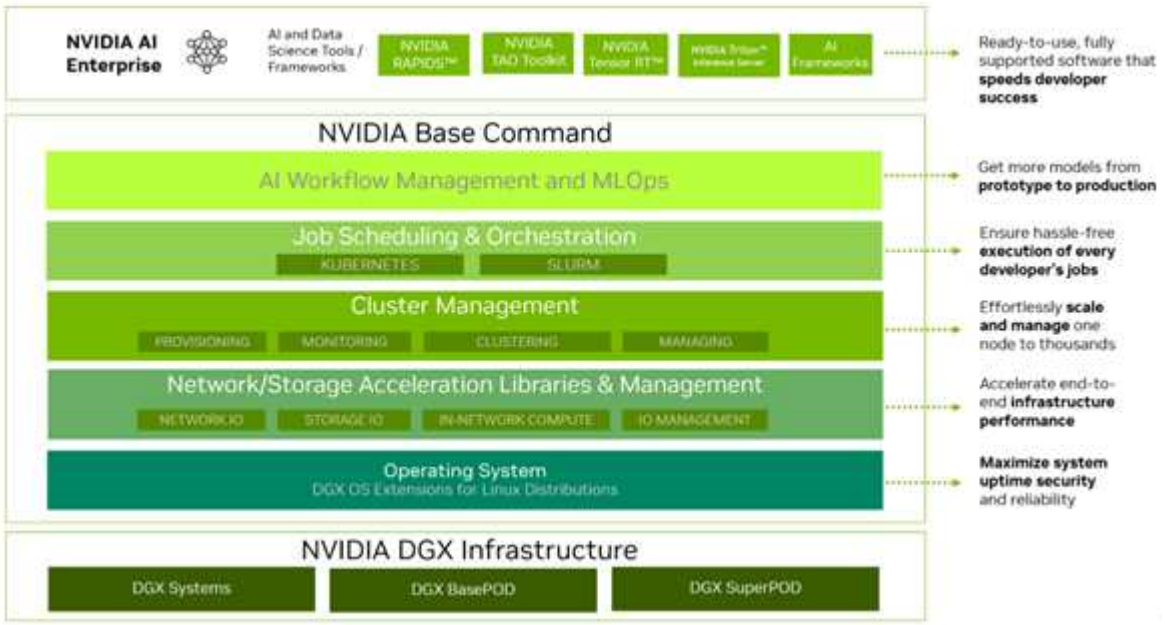
Questa sezione si concentra sui componenti software di NetApp AIPod con sistemi NVIDIA DGX.

Software NVIDIA

Comando di base NVIDIA

NVIDIA Base Command è alla base di ogni DGX BasePOD, consentendo alle organizzazioni di sfruttare il meglio dell'innovazione software NVIDIA . Le aziende possono sfruttare appieno il potenziale del loro investimento con una piattaforma collaudata che include orchestrazione e gestione dei cluster di livello aziendale, librerie che accelerano l'infrastruttura di elaborazione, storage e rete e un sistema operativo (SO) ottimizzato per i carichi di lavoro di intelligenza artificiale.

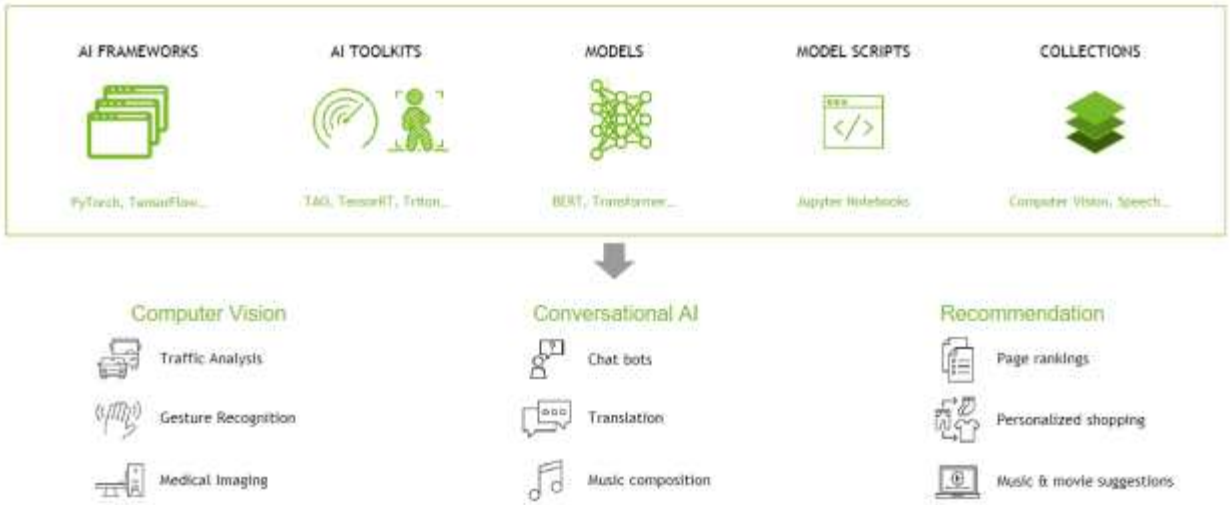
Soluzione NVIDIA BaseCommand



Cloud GPU NVIDIA (NGC)

NVIDIA NGC fornisce software per soddisfare le esigenze di data scientist, sviluppatori e ricercatori con diversi livelli di competenza in materia di intelligenza artificiale. Il software ospitato su NGC viene sottoposto a scansioni su un set aggregato di vulnerabilità ed esposizioni comuni (CVE), chiavi crittografiche e private. È testato e progettato per essere scalabile su più GPU e, in molti casi, su più nodi, garantendo agli utenti di massimizzare il loro investimento nei sistemi DGX.

Cloud GPU NVIDIA



NVIDIA AI Enterprise

NVIDIA AI Enterprise è la piattaforma software end-to-end che rende l'intelligenza artificiale generativa accessibile a tutte le aziende, offrendo il runtime più rapido ed efficiente per i modelli di base dell'intelligenza artificiale generativa ottimizzati per l'esecuzione sulla piattaforma NVIDIA DGX. Grazie a sicurezza, stabilità e gestibilità di livello produttivo, semplifica lo sviluppo di soluzioni di intelligenza artificiale generativa. NVIDIA AI Enterprise è incluso in DGX BasePOD per consentire agli sviluppatori aziendali di accedere a modelli pre-addestrati, framework ottimizzati, microservizi, librerie accelerate e supporto aziendale.

Software NetApp

NetApp ONTAP

ONTAP 9, l'ultima generazione di software di gestione dello storage di NetApp, consente alle aziende di modernizzare l'infrastruttura e passare a un data center pronto per il cloud. Sfruttando le funzionalità di gestione dei dati leader del settore, ONTAP consente la gestione e la protezione dei dati con un unico set di strumenti, indipendentemente da dove risiedano. È inoltre possibile spostare liberamente i dati ovunque siano necessari: edge, core o cloud. ONTAP 9 include numerose funzionalità che semplificano la gestione dei dati, accelerano e proteggono i dati critici e abilitano le funzionalità infrastrutturali di nuova generazione nelle architetture cloud ibride.

Accelerare e proteggere i dati

ONTAP garantisce livelli superiori di prestazioni e protezione dei dati ed estende queste capacità nei seguenti modi:

- Prestazioni e latenza più bassa. ONTAP offre la massima produttività possibile con la latenza più bassa possibile, incluso il supporto per NVIDIA GPUDirect Storage (GDS) tramite NFS su RDMA, NFS parallelo (pNFS) e trunking di sessione NFS.
- Protezione dei dati. ONTAP offre funzionalità integrate di protezione dei dati e la garanzia anti-ransomware più forte del settore, con gestione comune su tutte le piattaforme.
- Crittografia del volume NetApp (NVE). ONTAP offre la crittografia nativa a livello di volume con supporto sia per la gestione delle chiavi integrate che per quella esterna.
- Multitenancy di archiviazione e autenticazione a più fattori. ONTAP consente la condivisione delle risorse infrastrutturali con i massimi livelli di sicurezza.

Semplificare la gestione dei dati

La gestione dei dati è fondamentale per le operazioni IT aziendali e per gli scienziati dei dati, in modo che vengano utilizzate risorse appropriate per le applicazioni di intelligenza artificiale e per la formazione di set di dati di intelligenza artificiale/apprendimento automatico. Le seguenti informazioni aggiuntive sulle tecnologie NetApp esulano dall'ambito di questa convalida, ma potrebbero essere rilevanti a seconda della distribuzione.

Il software di gestione dati ONTAP include le seguenti funzionalità per semplificare e snellire le operazioni e ridurre i costi operativi totali:

- Gli snapshot e i cloni consentono la collaborazione, la sperimentazione parallela e una governance dei dati migliorata per i flussi di lavoro ML/DL.
- SnapMirror consente lo spostamento fluido dei dati in ambienti cloud ibridi e multi-sito, fornendo i dati dove e quando sono necessari.
- Compattazione dei dati in linea e deduplicazione estesa. La compattazione dei dati riduce lo spazio sprecato all'interno dei blocchi di archiviazione, mentre la deduplicazione aumenta significativamente la capacità effettiva. Ciò vale sia per i dati archiviati localmente sia per i dati archiviati a livelli nel cloud.
- Qualità del servizio minima, massima e adattiva (AQoS). I controlli granulari della qualità del servizio (QoS) aiutano a mantenere i livelli di prestazioni per le applicazioni critiche in ambienti altamente condivisi.
- NetApp FlexGroups consente la distribuzione dei dati su tutti i nodi del cluster di storage, garantendo un'enorme capacità e prestazioni più elevate per set di dati estremamente grandi.
- NetApp FabricPool. Fornisce la suddivisione automatica dei dati inattivi in opzioni di archiviazione cloud pubbliche e private, tra cui Amazon Web Services (AWS), Azure e la soluzione di archiviazione NetApp StorageGRID . Per ulteriori informazioni su FabricPool, vedere "TR-4598: Buone pratiche FabricPool" .
- NetApp FlexCache. Fornisce funzionalità di memorizzazione nella cache di volumi remoti che semplificano la distribuzione dei file, riducono la latenza WAN e abbassano i costi della larghezza di banda WAN. FlexCache consente lo sviluppo di prodotti distribuiti su più sedi, nonché l'accesso accelerato ai set di dati aziendali da postazioni remote.

Infrastruttura a prova di futuro

ONTAP aiuta a soddisfare le esigenze aziendali più esigenti e in continua evoluzione grazie alle seguenti funzionalità:

- Scalabilità fluida e operazioni senza interruzioni. ONTAP supporta l'aggiunta online di capacità ai controller esistenti e ai cluster scalabili. I clienti possono effettuare l'aggiornamento alle tecnologie più recenti, come NVMe e FC da 32 Gb, senza costose migrazioni di dati o interruzioni.
- Connessione cloud. ONTAP è il software di gestione dello storage più connesso al cloud, con opzioni per lo storage definito dal software (ONTAP Select) e istanze cloud-native (Google Cloud NetApp Volumes) in tutti i cloud pubblici.
- Integrazione con applicazioni emergenti. ONTAP offre servizi dati di livello aziendale per piattaforme e applicazioni di nuova generazione, come veicoli autonomi, città intelligenti e Industria 4.0, utilizzando la stessa infrastruttura che supporta le app aziendali esistenti.

Kit di strumenti NetApp DataOps

NetApp DataOps Toolkit è uno strumento basato su Python che semplifica la gestione degli spazi di lavoro di sviluppo/formazione e dei server di inferenza supportati da storage NetApp ad alte prestazioni e scalabile. DataOps Toolkit può funzionare come utility autonoma ed è ancora più efficace negli ambienti Kubernetes che sfruttano NetApp Trident per automatizzare le operazioni di archiviazione. Le principali funzionalità includono:

- Fornisci rapidamente nuovi spazi di lavoro JupyterLab ad alta capacità supportati da storage NetApp scalabile e ad alte prestazioni.
- Fornisci rapidamente nuove istanze di NVIDIA Triton Inference Server supportate da storage NetApp di classe enterprise.
- Clonazione quasi istantanea di spazi di lavoro JupyterLab ad alta capacità per consentire la sperimentazione o l'iterazione rapida.
- Snapshot quasi istantanei di spazi di lavoro JupyterLab ad alta capacità per backup e/o tracciabilità/baselining.
- Provisioning, clonazione e snapshot quasi istantanei di volumi di dati ad alta capacità e ad alte prestazioni.

NetApp Trident

Trident è un orchestratore di storage open source completamente supportato per container e distribuzioni Kubernetes, tra cui Anthos. Trident funziona con l'intero portfolio di storage NetApp , incluso NetApp ONTAP, e supporta anche connessioni NFS, NVMe/TCP e iSCSI. Trident accelera il flusso di lavoro DevOps consentendo agli utenti finali di effettuare il provisioning e gestire lo storage dai propri sistemi di storage NetApp senza richiedere l'intervento di un amministratore dello storage.

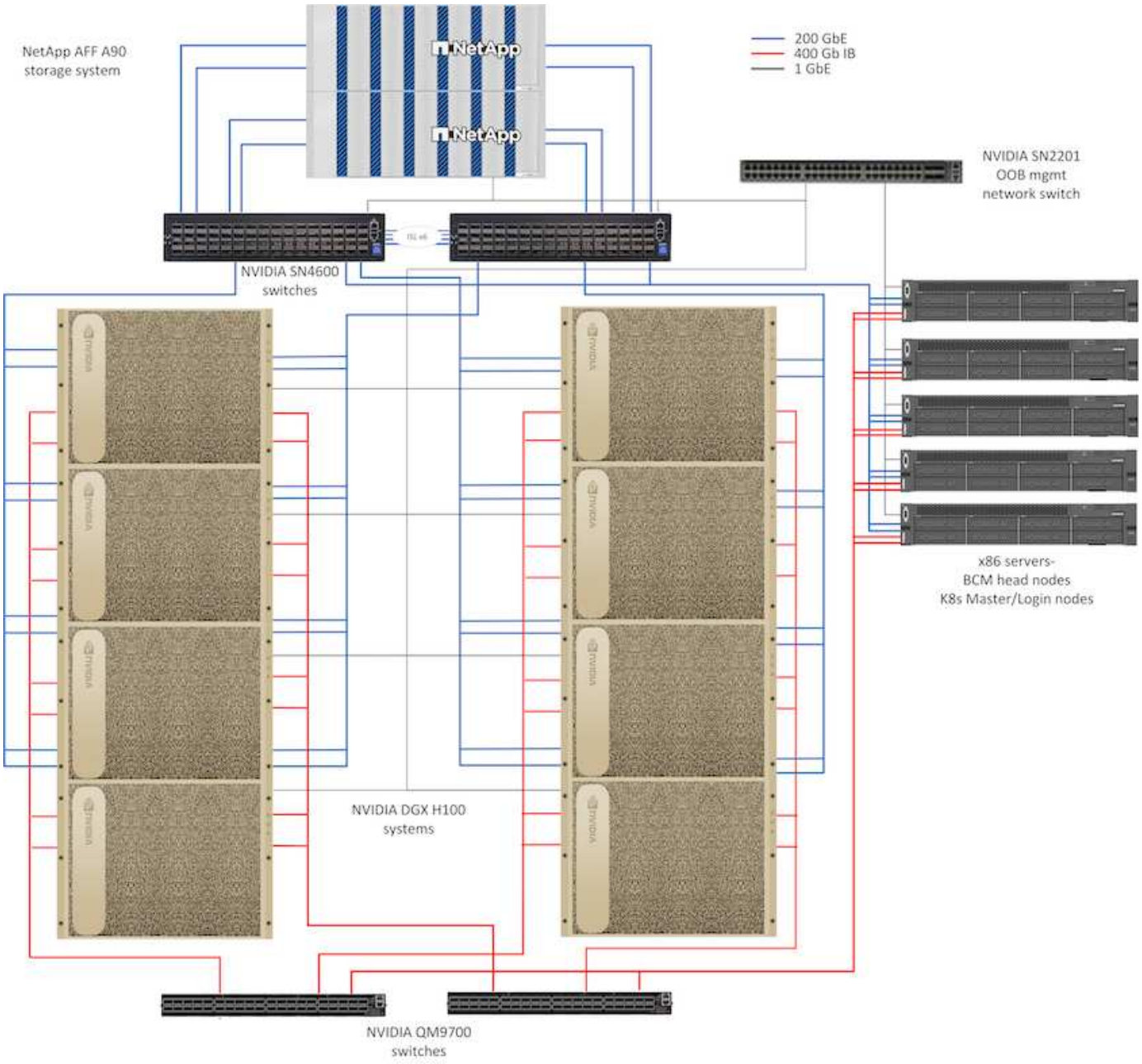
NVA-1173 NetApp AIPod con sistemi NVIDIA DGX H100 - Architettura della soluzione

Questa sezione si concentra sull'architettura per NetApp AIPod con sistemi NVIDIA DGX.

NetApp AIPod con sistemi DGX

Questa architettura di riferimento sfrutta strutture separate per l'interconnessione dei cluster di elaborazione e l'accesso allo storage, con connettività InfiniBand (IB) da 400 Gb/s tra i nodi di elaborazione. Il disegno seguente mostra la topologia complessiva della soluzione NetApp AIPod con sistemi DGX H100.

Topologia della soluzione NetApp AIpod



Progettazione di rete

In questa configurazione, il cluster di elaborazione utilizza una coppia di switch IB QM9700 da 400 Gb/s, collegati tra loro per garantire un'elevata disponibilità. Ogni sistema DGX H100 è collegato agli switch tramite otto connessioni, con le porte con numero pari collegate a uno switch e le porte con numero dispari collegate all'altro switch.

Per l'accesso al sistema di storage, la gestione in banda e l'accesso client, viene utilizzata una coppia di switch Ethernet SN4600. Gli switch sono collegati tramite collegamenti inter-switch e configurati con più VLAN per isolare i vari tipi di traffico. Il routing L3 di base è abilitato tra VLAN specifiche per abilitare percorsi multipli tra interfacce client e di archiviazione sullo stesso switch, nonché tra switch per un'elevata disponibilità. Per implementazioni più ampie, la rete Ethernet può essere estesa a una configurazione leaf-spine aggiungendo ulteriori coppie di switch per gli switch spine e ulteriori leaf, se necessario.

Oltre all'interconnessione di elaborazione e alle reti Ethernet ad alta velocità, tutti i dispositivi fisici sono collegati anche a uno o più switch Ethernet SN2201 per la gestione fuori banda. Si prega di vedere il"dettagli di

distribuzione" pagina per maggiori informazioni sulla configurazione di rete.

Panoramica dell'accesso all'archiviazione per i sistemi DGX H100

Ogni sistema DGX H100 è dotato di due adattatori ConnectX-7 a doppia porta per la gestione e il traffico di archiviazione e, per questa soluzione, entrambe le porte su ciascuna scheda sono collegate allo stesso switch. Una porta di ogni scheda viene quindi configurata in un bond LACP MLAG con una porta collegata a ogni switch e le VLAN per la gestione in banda, l'accesso client e l'accesso allo storage a livello utente sono ospitate su questo bond.

L'altra porta su ciascuna scheda è utilizzata per la connettività ai sistemi di archiviazione AFF A90 e può essere utilizzata in diverse configurazioni a seconda dei requisiti del carico di lavoro. Per le configurazioni che utilizzano NFS su RDMA per supportare NVIDIA Magnum IO GPUDirect Storage, le porte vengono utilizzate singolarmente con indirizzi IP in VLAN separate. Per le distribuzioni che non richiedono RDMA, le interfacce di archiviazione possono anche essere configurate con bonding LACP per garantire elevata disponibilità e larghezza di banda aggiuntiva. Con o senza RDMA, i client possono montare il sistema di archiviazione utilizzando NFS v4.1 pNFS e Session trunking per abilitare l'accesso parallelo a tutti i nodi di archiviazione nel cluster. Si prega di vedere il"dettagli di distribuzione" pagina per maggiori informazioni sulla configurazione del client.

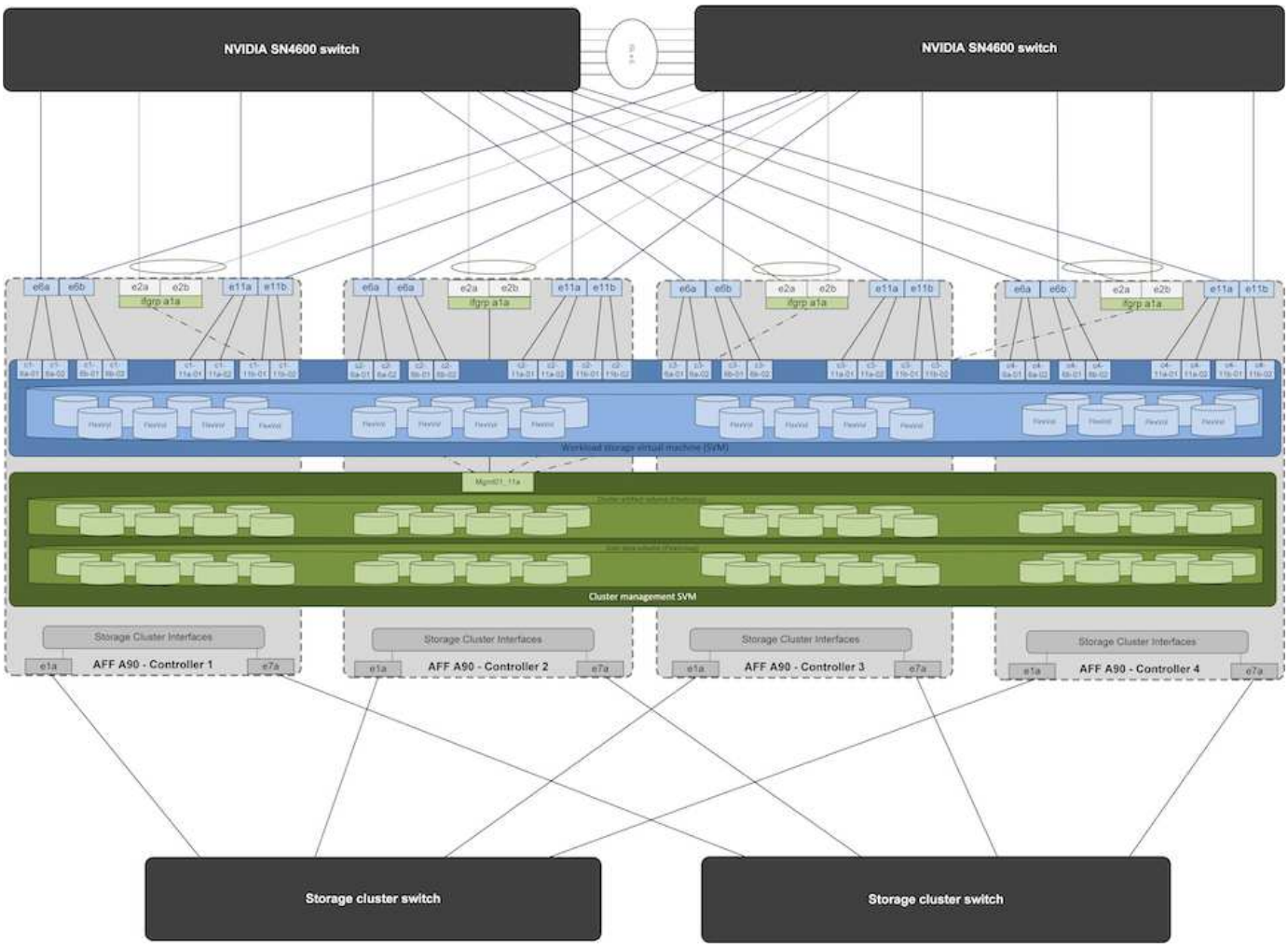
Per maggiori dettagli sulla connettività del sistema DGX H100 fare riferimento a"Documentazione NVIDIA BasePOD" .

Progettazione del sistema di archiviazione

Ogni sistema di storage AFF A90 è connesso tramite sei porte da 200 GbE da ciascun controller. Quattro porte di ciascun controller vengono utilizzate per l'accesso ai dati del carico di lavoro dai sistemi DGX e due porte di ciascun controller sono configurate come gruppo di interfacce LACP per supportare l'accesso dai server del piano di gestione per gli artefatti di gestione del cluster e le directory home degli utenti. L'accesso ai dati dal sistema di storage avviene tramite NFS, con una macchina virtuale di storage (SVM) dedicata all'accesso al carico di lavoro dell'intelligenza artificiale e una SVM separata dedicata agli utilizzi di gestione del cluster.

La SVM di gestione richiede solo un singolo LIF, ospitato sui gruppi di interfacce a 2 porte configurati su ciascun controller. Altri volumi FlexGroup vengono forniti sulla SVM di gestione per ospitare artefatti di gestione del cluster come immagini dei nodi del cluster, dati storici di monitoraggio del sistema e directory home degli utenti finali. Il disegno seguente mostra la configurazione logica del sistema di archiviazione.

Configurazione logica del cluster di storage NetApp A90



Server del piano di gestione

Questa architettura di riferimento include anche cinque server basati su CPU per l'utilizzo nel piano di gestione. Due di questi sistemi vengono utilizzati come nodi principali per NVIDIA Base Command Manager per la distribuzione e la gestione dei cluster. Gli altri tre sistemi vengono utilizzati per fornire servizi cluster aggiuntivi, come nodi master Kubernetes o nodi di accesso per distribuzioni che utilizzano Slurm per la pianificazione dei lavori. Le distribuzioni che utilizzano Kubernetes possono sfruttare il driver NetApp Trident CSI per fornire provisioning automatizzato e servizi dati con storage persistente sia per i carichi di lavoro di gestione che di intelligenza artificiale sul sistema di storage AFF A900 .

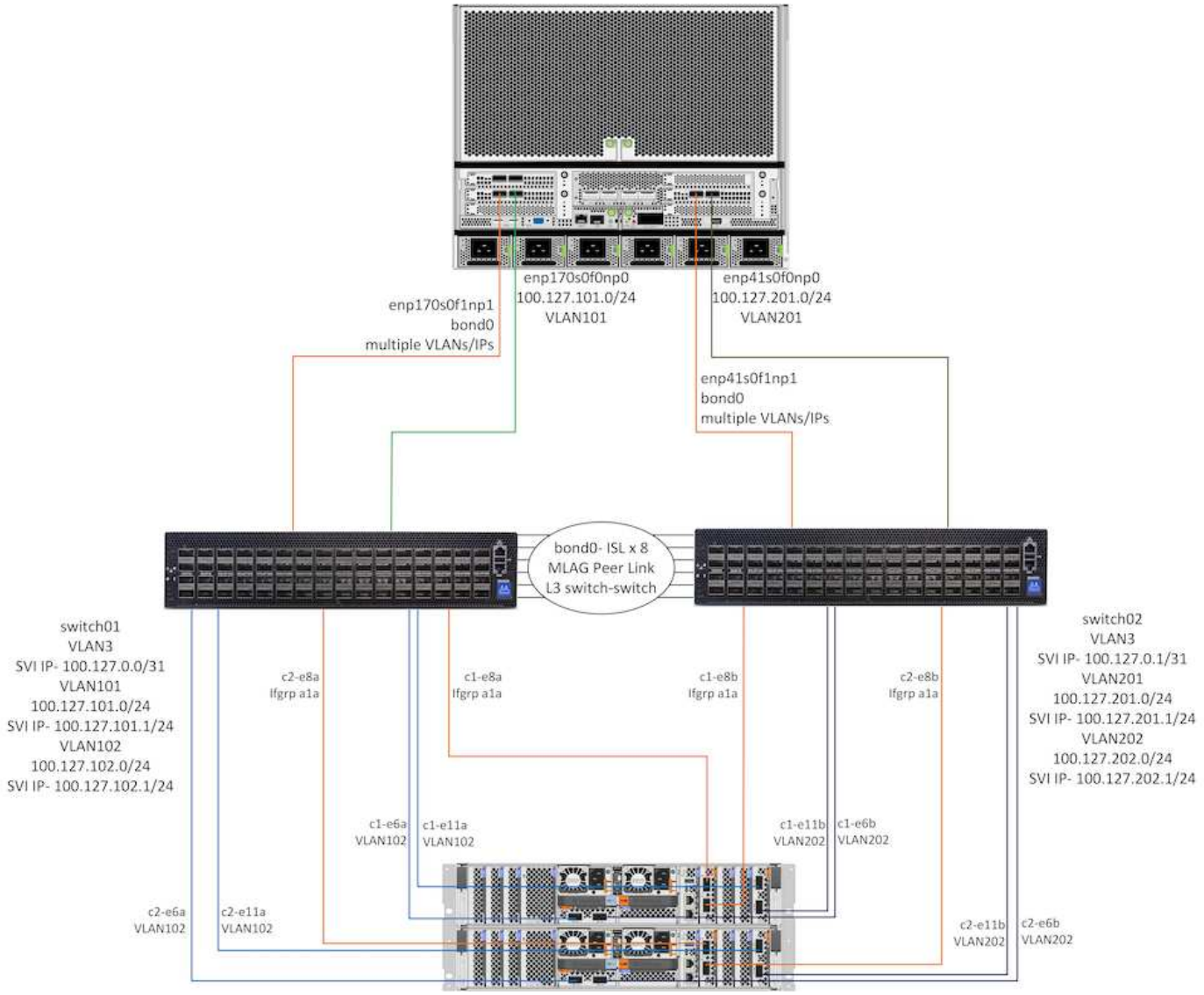
Ogni server è fisicamente connesso sia agli switch IB che agli switch Ethernet per consentire la distribuzione e la gestione del cluster, ed è configurato con montaggi NFS sul sistema di archiviazione tramite la SVM di gestione per l'archiviazione degli artefatti di gestione del cluster, come descritto in precedenza.

NVA-1173 NetApp AIPod con sistemi NVIDIA DGX - Dettagli di distribuzione

Questa sezione descrive i dettagli di distribuzione utilizzati durante la convalida di questa soluzione. Gli indirizzi IP utilizzati sono esempi e devono essere modificati in base all'ambiente di distribuzione. Per ulteriori informazioni sui comandi specifici utilizzati nell'implementazione di questa configurazione, fare riferimento alla documentazione del prodotto appropriata.

Il diagramma seguente mostra informazioni dettagliate sulla rete e sulla connettività per 1 sistema DGX H100 e 1 coppia HA di controller AFF A90 . Le istruzioni per l'implementazione nelle sezioni seguenti si basano sui dettagli riportati in questo diagramma.

Configurazione di rete NetApp AIpod



La tabella seguente mostra esempi di assegnazioni di cablaggio per un massimo di 16 sistemi DGX e 2 coppie AFF A90 HA.

Switch e porta	Dispositivo	Porta del dispositivo
porte switch1 1-16	DGX-H100-01 fino a -16	enp170s0f0np0, slot1 porta 1
porte switch1 17-32	DGX-H100-01 fino a -16	enp170s0f1np1, slot1 porta 2
porte switch1 33-36	AFF-A90-01 attraverso -04	porta e6a
porte switch1 37-40	AFF-A90-01 attraverso -04	porta e11a
porte switch1 41-44	AFF-A90-01 attraverso -04	porta e2a
porte switch1 57-64	ISL per switch2	porte 57-64

Switch e porta	Dispositivo	Porta del dispositivo
porte switch2 1-16	DGX-H100-01 fino a -16	enp41s0f0np0, slot 2 porta 1
porte switch2 17-32	DGX-H100-01 fino a -16	enp41s0f1np1, slot 2 porta 2
porte switch2 33-36	AFF-A90-01 attraverso -04	porta e6b
switch2 porte 37-40	AFF-A90-01 attraverso -04	porta e11b
porte switch2 41-44	AFF-A90-01 attraverso -04	porta e2b
porte switch2 57-64	ISL per passare1	porte 57-64

Nella tabella seguente sono riportate le versioni software dei vari componenti utilizzati in questa convalida.

Dispositivo	Versione del software
Switch NVIDIA SN4600	Cumulus Linux v5.9.1
Sistema NVIDIA DGX	DGX OS v6.2.1 (Ubuntu 22.04 LTS)
Mellanox OFED	24,01
NetApp AFF A90	NetApp ONTAP 9.14.1

Configurazione della rete di archiviazione

Questa sezione descrive i dettagli chiave per la configurazione della rete di archiviazione Ethernet. Per informazioni sulla configurazione della rete di elaborazione InfiniBand, consultare"Documentazione NVIDIA BasePOD" . Per maggiori dettagli sulla configurazione dello switch fare riferimento a"Documentazione NVIDIA Cumulus Linux" .

Di seguito sono descritti i passaggi di base utilizzati per configurare gli switch SN4600. Questo processo presuppone che il cablaggio e la configurazione di base dello switch (indirizzo IP di gestione, licenze, ecc.) siano stati completati.

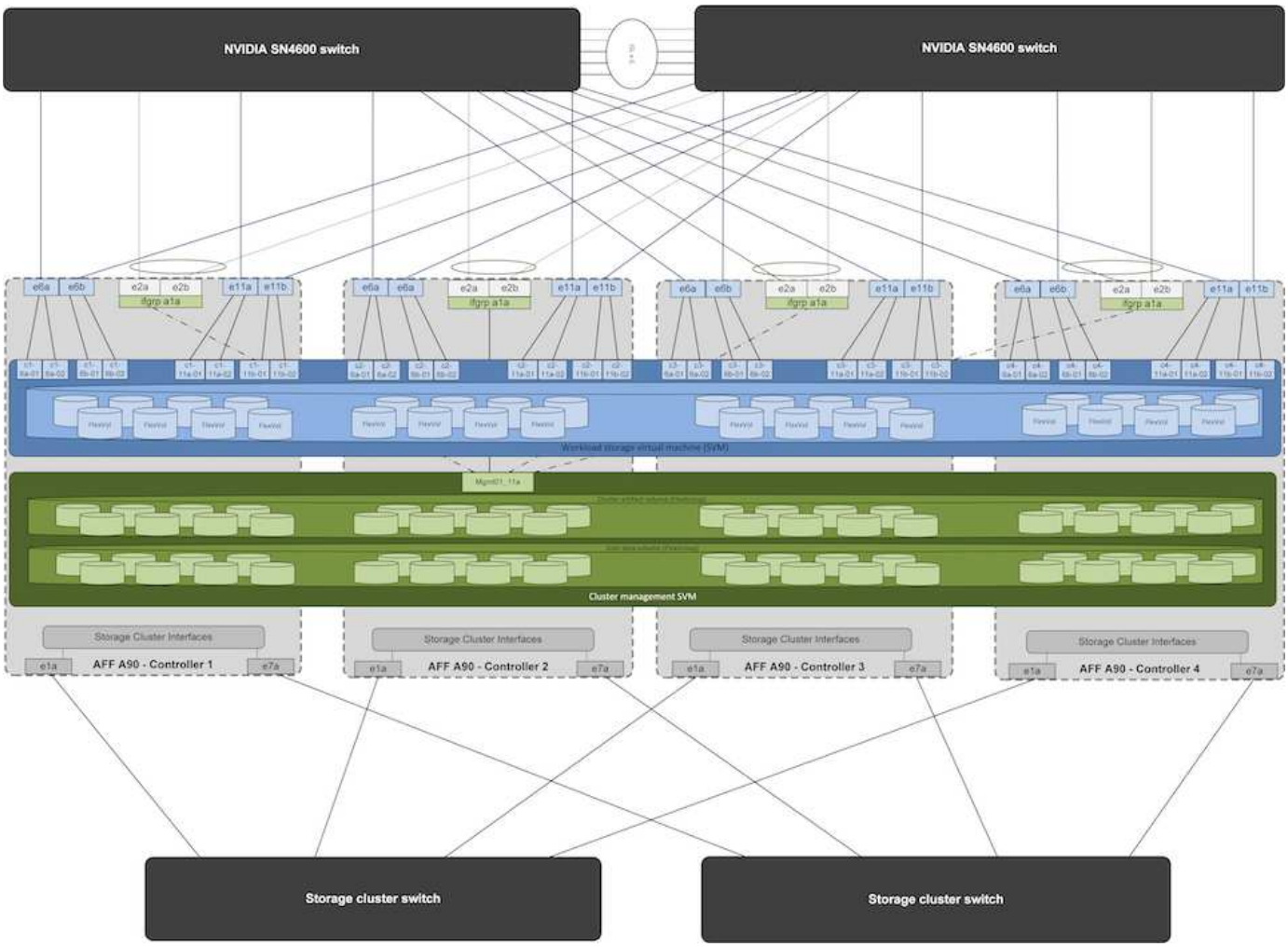
1. Configurare il legame ISL tra gli switch per abilitare l'aggregazione multi-link (MLAG) e il traffico di failover
    - Questa convalida ha utilizzato 8 collegamenti per fornire una larghezza di banda più che sufficiente per la configurazione di archiviazione in fase di test
    - Per istruzioni specifiche sull'abilitazione di MLAG, fare riferimento alla documentazione di Cumulus Linux.
2. Configurare LACP MLAG per ogni coppia di porte client e porte di archiviazione su entrambi gli switch
    - porta swp17 su ogni switch per DGX-H100-01 (enp170s0f1np1 e enp41s0f1np1), porta swp18 per DGX-H100-02, ecc. (bond1-16)
    - porta swp41 su ogni switch per AFF-A90-01 (e2a ed e2b), porta swp42 per AFF-A90-02, ecc. (bond17-20)
    - nv set interfaccia bondX membro del legame swpX
    - nv set interfaccia bondx bond mlag id X
3. Aggiungere tutte le porte e i legami MLAG al dominio bridge predefinito
    - nv set int swp1-16,33-40 dominio bridge br_default
    - nv set int bond1-20 dominio bridge br_default

4. Abilita RoCE su ogni switch
    - nv imposta la modalità roce senza perdita di dati
5. Configurare le VLAN: 2 per le porte client, 2 per le porte di archiviazione, 1 per la gestione, 1 per lo switch L3 da switch a switch
    - interruttore 1-
        - VLAN 3 per il routing da switch a switch L3 in caso di guasto della NIC del client
        - VLAN 101 per la porta di archiviazione 1 su ciascun sistema DGX (enp170s0f0np0, slot1 porta 1)
        - VLAN 102 per la porta e6a e e11a su ciascun controller di archiviazione AFF A90
        - VLAN 301 per la gestione tramite le interfacce MLAG per ciascun sistema DGX e controller di archiviazione
    - interruttore 2-
        - VLAN 3 per il routing da switch a switch L3 in caso di guasto della NIC del client
        - VLAN 201 per la porta di archiviazione 2 su ciascun sistema DGX (enp41s0f0np0, slot2 porta 1)
        - VLAN 202 per la porta e6b e e11b su ciascun controller di archiviazione AFF A90
        - VLAN 301 per la gestione tramite le interfacce MLAG per ciascun sistema DGX e controller di archiviazione
6. Assegnare porte fisiche a ciascuna VLAN in modo appropriato, ad esempio porte client nelle VLAN client e porte di archiviazione nelle VLAN di archiviazione
    - nv set int <swpX> dominio bridge br_default access <id vlan>
    - Le porte MLAG dovrebbero rimanere porte trunk per abilitare più VLAN sulle interfacce collegate, secondo necessità.
7. Configurare le interfacce virtuali dello switch (SVI) su ciascuna VLAN per fungere da gateway e abilitare il routing L3
    - interruttore 1-
        - nv set int vlan3 indirizzo IP 100.127.0.0/31
        - nv set int vlan101 indirizzo IP 100.127.101.1/24
        - nv set int vlan102 indirizzo IP 100.127.102.1/24
    - interruttore 2-
        - nv set int vlan3 indirizzo IP 100.127.0.1/31
        - nv set int vlan201 indirizzo IP 100.127.201.1/24
        - nv set int vlan202 indirizzo IP 100.127.202.1/24
8. Creare percorsi statici
    - Le rotte statiche vengono create automaticamente per le subnet sullo stesso switch
    - Sono necessari percorsi statici aggiuntivi per il routing da switch a switch in caso di guasto del collegamento client
        - interruttore 1-
            - nv imposta il router predefinito vrf statico 100.127.128.0/17 tramite 100.127.0.1
        - interruttore 2-
            - nv imposta il router predefinito vrf statico 100.127.0.0/17 tramite 100.127.0.0

Configurazione del sistema di archiviazione

In questa sezione vengono descritti i dettagli chiave per la configurazione del sistema di archiviazione A90 per questa soluzione. Per maggiori dettagli sulla configurazione dei sistemi ONTAP fare riferimento a"Documentazione ONTAP" . Il diagramma seguente mostra la configurazione logica del sistema di archiviazione.

Configurazione logica del cluster di storage NetApp A90



Di seguito sono descritti i passaggi di base utilizzati per configurare il sistema di archiviazione. Questo processo presuppone che l'installazione del cluster di archiviazione di base sia stata completata.

1. Configurare 1 aggregato su ciascun controller con tutte le partizioni disponibili meno 1 di riserva
   - aggr create -node <nodo> -aggregate <nodo>_data01 -diskcount <47>
2. Configurare ifgrps su ciascun controller
   - net port ifgrp create -node <nodo> -ifgrp a1a -mode multimode_lacp -distr-function port
   - net port ifgrp add-port -node <nodo> -ifgrp <ifgrp> -ports <nodo>:e2a,<nodo>:e2b
3. Configurare la porta VLAN mgmt su ifgrp su ciascun controller
   - porta di rete vlan crea -nodo aff-a90-01 -porta a1a -id-vlan 31
   - porta di rete vlan crea -nodo aff-a90-02 -porta a1a -id-vlan 31

- porta di rete vlan crea -nodo aff-a90-03 -porta a1a -id-vlan 31
- porta di rete vlan crea -nodo aff-a90-04 -porta a1a -vlan-id 31

4. Crea domini di trasmissione
   - broadcast-domain create -broadcast-domain vlan21 -mtu 9000 -ports aff-a90-01:e6a,aff-a90-01:e11a,aff-a90-02:e6a,aff-a90-02:e11a,aff-a90-03:e6a,aff-a90-03:e11a,aff-a90-04:e6a,aff-a90-04:e11a
   - broadcast-domain create -broadcast-domain vlan22 -mtu 9000 -ports aaff-a90-01:e6b,aff-a90-01:e11b,aff-a90-02:e6b,aff-a90-02:e11b,aff-a90-03:e6b,aff-a90-03:e11b,aff-a90-04:e6b,aff-a90-04:e11b
   - creazione dominio broadcast -dominio broadcast vlan31 -mtu 9000 -porte aff-a90-01:a1a-31,aff-a90-02:a1a-31,aff-a90-03:a1a-31,aff-a90-04:a1a-31
5. Crea SVM di gestione *
6. Configurare la gestione SVM
   - creare LIF
     - net int create -vserver basepod-mgmt -lif vlan31-01 -home-node aff-a90-01 -home-port a1a-31 -address 192.168.31.X -netmask 255.255.255.0
   - creare volumi FlexGroup
     - vol create -vserver basepod-mgmt -volume home -size 10T -auto-provision-as flexgroup -junction -path /home
     - vol create -vserver basepod-mgmt -volume cm -size 10T -auto-provision-as flexgroup -junction -path /cm
   - creare una politica di esportazione
     - regola export-policy create -vserver basepod-mgmt -policy default -client-match 192.168.31.0/24 -rorule sys -rwrule sys -superuser sys
7. Crea dati SVM *
8. Configurare i dati SVM
   - configurare SVM per il supporto RDMA
     - vserver nfs modify -vserver basepod-data -rdma abilitato
   - creare LIF
     - net int create -vserver basepod-data -lif c1-6a-lif1 -home-node aff-a90-01 -home-port e6a -address 100.127.102.101 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-6a-lif2 -home-node aff-a90-01 -home-port e6a -address 100.127.102.102 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-6b-lif1 -home-node aff-a90-01 -home-port e6b -address 100.127.202.101 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-6b-lif2 -home-node aff-a90-01 -home-port e6b -address 100.127.202.102 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-11a-lif1 -home-node aff-a90-01 -home-port e11a -address 100.127.102.103 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-11a-lif2 -home-node aff-a90-01 -home-port e11a -address 100.127.102.104 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-11b-lif1 -home-node aff-a90-01 -home-port e11b -address 100.127.202.103 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-11b-lif2 -home-node aff-a90-01 -home-port e11b

-address 100.127.202.104 -netmask 255.255.255.0

- net int create -vserver basepod-data -lif c2-6a-lif1 -home-node aff-a90-02 -home-port e6a -address 100.127.102.105 -netmask 255.255.255.0
- net int create -vserver basepod-data -lif c2-6a-lif2 -home-node aff-a90-02 -home-port e6a -address 100.127.102.106 -netmask 255.255.255.0
- net int create -vserver basepod-data -lif c2-6b-lif1 -home-node aff-a90-02 -home-port e6b -address 100.127.202.105 -netmask 255.255.255.0
- net int create -vserver basepod-data -lif c2-6b-lif2 -home-node aff-a90-02 -home-port e6b -address 100.127.202.106 -netmask 255.255.255.0
- net int create -vserver basepod-data -lif c2-11a-lif1 -home-node aff-a90-02 -home-port e11a -address 100.127.102.107 -netmask 255.255.255.0
- net int create -vserver basepod-data -lif c2-11a-lif2 -home-node aff-a90-02 -home-port e11a -address 100.127.102.108 -netmask 255.255.255.0
- net int create -vserver basepod-data -lif c2-11b-lif1 -home-node aff-a90-02 -home-port e11b -address 100.127.202.107 -netmask 255.255.255.0
- net int create -vserver basepod-data -lif c2-11b-lif2 -home-node aff-a90-02 -home-port e11b -address 100.127.202.108 -netmask 255.255.255.0

9. Configurare LIF per l'accesso RDMA
   - Per le distribuzioni con ONTAP 9.15.1, la configurazione RoCE QoS per le informazioni fisiche richiede comandi a livello di sistema operativo che non sono disponibili nella CLI ONTAP . Contattare l'assistenza NetApp per ricevere assistenza nella configurazione delle porte per il supporto RoCE. NFS su RDMA funziona senza problemi
   - A partire da ONTAP 9.16.1, le interfacce fisiche verranno automaticamente configurate con le impostazioni appropriate per il supporto RoCE end-to-end.
   - net int modifica -vserver basepod-data -lif * -rdma-protocols roce
10. Configurare i parametri NFS sulla SVM dei dati
    - modifica nfs -vserver basepod-data -v4.1 abilitato -v4.1-pnfs abilitato -v4.1-trunking abilitato -tcp-max -transfer-size 262144
11. Crea volumi FlexGroup
    - vol create -vserver basepod-data -volume data -size 100T -auto-provision-as flexgroup -junction-path /data
12. Creare una politica di esportazione
    - regola export-policy create -vserver basepod-data -policy default -client-match 100.127.101.0/24 -rorule sys -rwrule sys -superuser sys
    - regola export-policy create -vserver basepod-data -policy default -client-match 100.127.201.0/24 -rorule sys -rwrule sys -superuser sys
13. creare percorsi
    - percorso aggiungi -vserver basepod_data -destinazione 100.127.0.0/17 -gateway 100.127.102.1 metrica 20
    - percorso aggiungi -vserver basepod_data -destinazione 100.127.0.0/17 -gateway 100.127.202.1 metrica 30
    - percorso aggiungi -vserver basepod_data -destinazione 100.127.128.0/17 -gateway 100.127.202.1 metrica 20

- percorso aggiungi -vserver basepod_data -destinazione 100.127.128.0/17 -gateway 100.127.102.1 metrica 30

Configurazione DGX H100 per l'accesso allo storage RoCE

Questa sezione descrive i dettagli chiave per la configurazione dei sistemi DGX H100. Molti di questi elementi di configurazione possono essere inclusi nell'immagine del sistema operativo distribuita sui sistemi DGX o implementati da Base Command Manager al momento dell'avvio. Sono elencati qui per riferimento, per maggiori informazioni sulla configurazione dei nodi e delle immagini software in BCM consultare"Documentazione BCM" .

1. Installa pacchetti aggiuntivi
   - ipmitool
   - python3-pip
2. Installa i pacchetti Python
   - paramiko
   - matplotlib
3. Riconfigurare dpkg dopo l'installazione del pacchetto
   - dpkg --configure -a
4. Installa MOFED
5. Imposta i valori mst per l'ottimizzazione delle prestazioni
   - mstconfig -y -d <aa:00.0,29:00.0> imposta ADVANCED_PCI_SETTINGS=1 NUM_OF_VFS=0 MAX_ACC_OUT_READ=44
6. Reimpostare gli adattatori dopo aver modificato le impostazioni
   - mlxfwreset -d <aa:00.0,29:00.0> -y ripristina
7. Imposta MaxReadReq sui dispositivi PCI
   - setpci -s <aa:00.0,29:00.0> 68.W=5957
8. Imposta la dimensione del buffer ad anello RX e TX
   - ethtool -G <enp170s0f0np0,enp41s0f0np0> rx 8192 tx 8192
9. Imposta PFC e DSCP utilizzando mlnx_qos
   - mlnx_qos -i <enp170s0f0np0,enp41s0f0np0> --pfc 0,0,0,1,0,0,0,0 --trust=dscp --cable_len=3
10. Imposta ToS per il traffico RoCE sulle porte di rete
    - echo 106 > /sys/class/infiniband/<mlx5_7,mlx5_1>/tc/1/traffic_class
11. Configurare ogni scheda di rete di archiviazione con un indirizzo IP sulla subnet appropriata
    - 100.127.101.0/24 per la scheda di rete di archiviazione 1
    - 100.127.201.0/24 per la scheda di rete di archiviazione 2
12. Configurare le porte di rete in banda per il bonding LACP (enp170s0f1np1,enp41s0f1np1)
13. configurare percorsi statici per percorsi primari e secondari verso ciascuna subnet di archiviazione
    - percorso aggiungi –net 100.127.0.0/17 gw 100.127.101.1 metrica 20
    - percorso aggiungi –net 100.127.0.0/17 gw 100.127.201.1 metrica 30
    - percorso aggiungi –net 100.127.128.0/17 gw 100.127.201.1 metrica 20

- percorso aggiungi –net 100.127.128.0/17 gw 100.127.101.1 metrica 30

14. Monta /volume home
    - monta -o vers=3,nconnect=16,rsize=262144,wsize=262144 192.168.31.X:/home /home
15. Monta /volume dati
    - Le seguenti opzioni di montaggio sono state utilizzate durante il montaggio del volume di dati:
        - vers=4.1 # abilita pNFS per l'accesso parallelo a più nodi di archiviazione
        - proto=rdma # imposta il protocollo di trasferimento su RDMA invece del TCP predefinito
        - max_connect=16 # abilita il trunking della sessione NFS per aggregare la larghezza di banda della porta di archiviazione
        - write=eager # migliora le prestazioni di scrittura delle scritture bufferizzate
        - rsize=262144,wsize=262144 # imposta la dimensione del trasferimento I/O a 256k

NVA-1173 NetApp AIPod con sistemi NVIDIA DGX - Guida alla convalida e al dimensionamento della soluzione

Questa sezione si concentra sulla convalida della soluzione e sulle linee guida per il dimensionamento dei sistemi NetApp AIPod con NVIDIA DGX.

Validazione della soluzione

La configurazione di archiviazione in questa soluzione è stata convalidata utilizzando una serie di carichi di lavoro sintetici utilizzando lo strumento open source FIO. Questi test includono modelli di I/O di lettura e scrittura pensati per simulare il carico di lavoro di archiviazione generato dai sistemi DGX che eseguono attività di formazione di deep learning. La configurazione di archiviazione è stata convalidata utilizzando un cluster di server CPU a 2 socket che eseguono contemporaneamente i carichi di lavoro FIO per simulare un cluster di sistemi DGX. Ogni client è stato configurato con la stessa configurazione di rete descritta in precedenza, con l'aggiunta dei seguenti dettagli.

Per questa convalida sono state utilizzate le seguenti opzioni di montaggio:

versione=4.1	abilita pNFS per l'accesso parallelo a più nodi di archiviazione
proto=rdma	imposta il protocollo di trasferimento su RDMA invece del TCP predefinito
porta=20049	specificare la porta corretta per il servizio RDMA NFS
max_connect=16	consente il trunking della sessione NFS per aggregare la larghezza di banda della porta di archiviazione
scrivere=impaziente	migliora le prestazioni di scrittura delle scritture bufferizzate
rsize=262144,wsize=262144	imposta la dimensione del trasferimento I/O a 256k

Inoltre, i client sono stati configurati con un valore NFS max_session_slots pari a 1024. Poiché la soluzione è stata testata utilizzando NFS su RDMA, le porte delle reti di archiviazione sono state configurate con un collegamento attivo/passivo. Per questa convalida sono stati utilizzati i seguenti parametri di legame:

modalità=backup attivo	imposta il legame in modalità attiva/passiva
primary=<nome interfaccia>	le interfacce primarie per tutti i client sono state distribuite sugli switch

mii-monitor-intervallo=100	specifica un intervallo di monitoraggio di 100 ms
fail-over-mac-policy=attivo	specifica che l'indirizzo MAC del collegamento attivo è il MAC del legame. Ciò è necessario per il corretto funzionamento dell'RDMA sull'interfaccia collegata.

Il sistema di archiviazione è stato configurato come descritto con due coppie A900 HA (4 controller) con due ripiani per dischi NS224 da 24 unità disco NVMe da 1,9 TB collegate a ciascuna coppia HA. Come indicato nella sezione dedicata all'architettura, la capacità di archiviazione di tutti i controller è stata combinata utilizzando un volume FlexGroup e i dati di tutti i client sono stati distribuiti tra tutti i controller del cluster.

Guida al dimensionamento del sistema di archiviazione

NetApp ha completato con successo la certificazione DGX BasePOD e le due coppie A90 HA testate possono supportare facilmente un cluster di sedici sistemi DGX H100. Per distribuzioni più grandi con requisiti di prestazioni di storage più elevati, è possibile aggiungere sistemi AFF aggiuntivi al cluster NetApp ONTAP fino a 12 coppie HA (24 nodi) in un singolo cluster. Utilizzando la tecnologia FlexGroup descritta in questa soluzione, un cluster da 24 nodi può fornire oltre 79 PB e fino a 552 GBps di throughput in un singolo namespace. Altri sistemi di storage NetApp , come AFF A400, A250 e C800, offrono prestazioni inferiori e/o opzioni di capacità più elevate per implementazioni più piccole a costi inferiori. Poiché ONTAP 9 supporta cluster a modello misto, i clienti possono iniziare con un ingombro iniziale più piccolo e aggiungere al cluster sistemi di storage più grandi o più numerosi man mano che aumentano i requisiti di capacità e prestazioni. La tabella seguente mostra una stima approssimativa del numero di GPU A100 e H100 supportate su ciascun modello AFF .

Guida al dimensionamento del sistema di storage NetApp

		Throughput[2]	Raw capacity *(typical[3] / max)*	Connectivity	# NVIDIA A100 GPUs supported[4]	# NVIDIA H100 GPUs supported[5]
NetApp® AFF A1K	**1 HA pair**[1]	56 GB/s	368TB / 14.7PB	**200 GbE**	1-160	1-80
	12 HA pairs	672 GB/s	**4.4PB / 176.4PB**		**1920**	**960**
AFF A90	**1 HA pair**	46 GB/s	368TB / 6.6PB	**200 GbE**	1 – 128	1-64
	12 HA pairs	552 GB/s	**4.4PB / 79.2PB**		**1536**	**768**
AFF A70	**1 HA pair**	21 GB/s	368TB / 6.6PB	**200 GbE**	1-48	1-24
	12 HA pairs	252 GB/s	**4.4PB / 79.2PB**		**576**	**288**

NVA-1173 NetApp AIPod con sistemi NVIDIA DGX - Conclusione e informazioni aggiuntive

Questa sezione include riferimenti per informazioni aggiuntive sui sistemi NetApp AIPod con NVIDIA DGX.

Conclusione

L'architettura DGX BasePOD è una piattaforma di apprendimento profondo di nuova generazione che richiede capacità di archiviazione e gestione dei dati altrettanto avanzate. Combinando DGX BasePOD con i sistemi

NetApp AFF , l'architettura NetApp AIPod con i sistemi DGX può essere implementata praticamente su qualsiasi scala. In combinazione con l'integrazione cloud superiore e le funzionalità software-defined di NetApp ONTAP, AFF consente una gamma completa di pipeline di dati che abbracciano l'edge, il core e il cloud per progetti DL di successo.

Informazioni aggiuntive

Per saperne di più sulle informazioni descritte nel presente documento, consultare i seguenti documenti e/o siti web:

- Software di gestione dati NetApp ONTAP — Libreria di informazioni ONTAP

  "https://docs.netapp.com/us-en/ontap-family/"

- Sistemi di archiviazione NetApp AFF A90

  https://www.netapp.com/pdf.html?item=/media/7828-ds-3582-aff-a-series-ai-era.pdf

- Informazioni NetApp ONTAP RDMA-

  "https://docs.netapp.com/us-en/ontap/nfs-rdma/index.html"

- Kit di strumenti NetApp DataOps

  "https://github.com/NetApp/netapp-dataops-toolkit"

- NetApp Trident

  "Panoramica"

- Blog di NetApp GPUDirect Storage -

  "https://www.netapp.com/blog/ontap-reaches-171-gpudirect-storage/"

- NVIDIA DGX BasePOD

  "https://www.nvidia.com/en-us/data-center/dgx-basepod/"

- Sistemi NVIDIA DGX H100

  "https://www.nvidia.com/en-us/data-center/dgx-h100/"

- NVIDIA

  "https://www.nvidia.com/en-us/networking/"

- Archiviazione NVIDIA Magnum IO™ GPUDirect®

  "https://docs.nvidia.com/gpudirect-storage"

- Comando di base NVIDIA

  "https://www.nvidia.com/en-us/data-center/base-command/"

- Gestore dei comandi di base NVIDIA

"https://www.nvidia.com/en-us/data-center/base-command/manager"

- NVIDIA AI Enterprise

  "https://www.nvidia.com/en-us/data-center/products/ai-enterprise/"

Ringraziamenti

Questo documento è frutto del lavoro dei team NetApp Solutions e ONTAP Engineering: David Arnette, Olga Kornievskaia, Dustin Fischer, Srikanth Kaligotla, Mohit Kumar e Raghuram Sudhaakar. Gli autori desiderano inoltre ringraziare NVIDIA e il team di progettazione NVIDIA DGX BasePOD per il loro continuo supporto.

Informazioni sul copyright

Informazioni sul marchio commerciale