NVA-1173 NetApp AIPod mit NVIDIA DGX Systemen

NetApp Solutions

NetApp
December 19, 2024

Inhalt

NVA-1173 NetApp AIPod mit NVIDIA DGX Systemen

NVA-1173 NetApp AIPod mit NVIDIA DGX Systemen – Einführung

POWERED BY

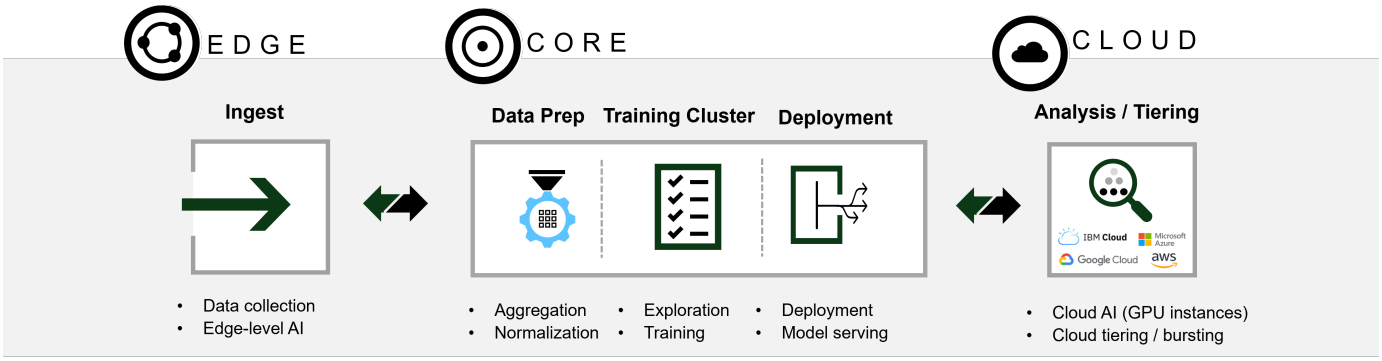
NVIDIA

NetApp Solution Engineering

Zusammenfassung

Das NetApp™ AIPod mit NVIDIA DGX™ Systemen und Cloud-vernetzten NetApp Storage-Systemen vereinfacht die Infrastrukturbereitstellung für Machine-Learning- (ML) und KI-Workloads (künstliche Intelligenz), indem Komplexität und Unsicherheiten bei der Systemaufsetzung beseitigt werden. Das Design basiert auf NVIDIA DGX BasePOD &#8482 und bietet außergewöhnliche Computing-Performance für Workloads der neuesten Generation. AIPod mit NVIDIA DGX Systemen fügt NetApp AFF Storage-Systeme hinzu. Damit können Kunden klein anfangen und unterbrechungsfrei wachsen. Gleichzeitig erhalten sie intelligente Datenmanagement-Funktionen, mit denen sich Daten zwischen Datenaufnahme, zentraler Datenplattform und Cloud frei verschieben lassen. NetApp AIPod ist Teil des größeren Portfolios an NetApp KI-Lösungen, wie in der Abbildung unten dargestellt.

NetApp AI Lösungsportfolio



Dieses Dokument beschreibt die wichtigsten Komponenten der AIPod Referenzarchitektur, Informationen zur

Systemkonnektivität und Konfiguration, die Ergebnisse der Validierungstests sowie Hinweise zur Dimensionierung der Lösung. Dieses Dokument richtet sich an Lösungstechniker von NetApp und Partnern sowie strategische Entscheidungsträger von Kunden, die an der Implementierung einer hochperformanten Infrastruktur für ML/DL- und Analyse-Workloads interessiert sind.

NVA-1173 NetApp AIPod mit NVIDIA DGX Systemen – Hardwarekomponenten

Der Abschnitt konzentriert sich auf die Hardwarekomponenten des NetApp AIPod mit NVIDIA DGX Systemen.

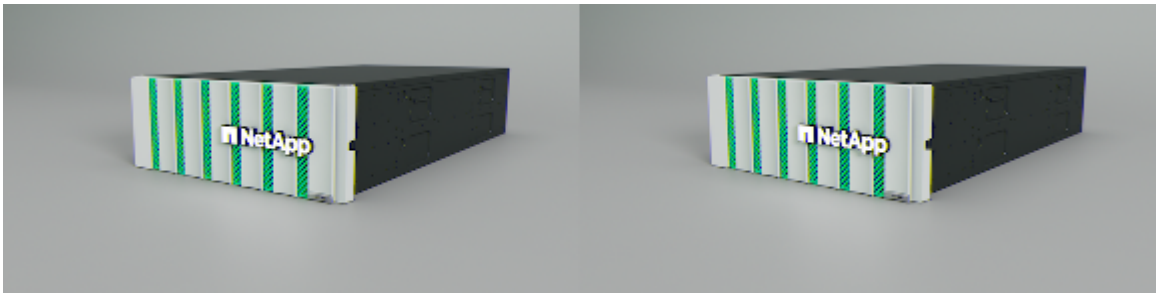
NetApp AFF Storage-Systeme

Mit den hochmodernen Storage-Systemen von NetApp AFF können IT-Abteilungen die Enterprise-Storage-Anforderungen mit branchenführender Performance, überlegener Flexibilität, Cloud-Integration und erstklassigem Datenmanagement erfüllen. AFF Systeme wurden speziell für Flash entwickelt. Sie helfen geschäftskritische Daten zu beschleunigen, zu managen und zu schützen.

AFF A90 Storage-Systeme

Die NetApp AFF A90 mit NetApp ONTAP Datenmanagement-Software bietet integrierte Datensicherung, optionale Funktionen zum Schutz vor Ransomware sowie die hohe Performance und Ausfallsicherheit, die zur Unterstützung der wichtigsten Geschäfts-Workloads erforderlich sind. Es verhindert Unterbrechungen geschäftskritischer Abläufe, minimiert Performance-Tuning und schützt Ihre Daten vor Ransomware-Angriffen. Die Lösung bietet: • Branchenführende Performance • kompromisslose Datensicherheit • vereinfachte, unterbrechungsfreie Upgrades

NetApp AFF A90 Storage-System



Erstklassige Performance

Die AFF A90 managt problemlos Workloads der nächsten Generation wie Deep Learning, KI und Hochgeschwindigkeitsanalysen sowie herkömmliche Unternehmensdatenbanken wie Oracle, SAP HANA, Microsoft SQL Server und virtualisierte Applikationen. Geschäftskritische Applikationen werden mit bis zu 2,4 Mio. IOPS pro HA-Paar und Latenz von bis zu 100µs Höchstgeschwindigkeit ausgeführt. Die Performance

steigt im Vergleich zu früheren NetApp Modellen um bis zu 50 %. Mit NFS over RDMA, pNFS und Session Trunking können Kunden mithilfe der vorhandenen Netzwerkinfrastruktur des Datacenters ein hohes Maß an Netzwerk-Performance für Applikationen der nächsten Generation erreichen. Einheitliche Multiprotokoll-Unterstützung für SAN-, NAS- und Objekt-Storage für Skalierung und Wachstum sowie maximale Flexibilität dank einheitlicher, einzelner ONTAP Datenmanagement-Software für Daten vor Ort oder in der Cloud Zudem kann der Systemzustand mit KI-basierten prädiktiven Analysen von Active IQ und Cloud Insights optimiert werden.

Kompromisslose Datensicherheit

AFF A90 Systeme enthalten die komplette Suite an integrierter und applikationskonsistenter Datensicherungssoftware von NetApp. Sie bietet integrierte Datensicherung und innovative Anti-Ransomware-Lösungen zur Vorbeugung und Wiederherstellung nach einem Angriff. Schädliche Dateien können nicht jemals auf Festplatte geschrieben werden. Anomalien im Storage werden einfach überwacht, um Erkenntnisse zu gewinnen.

Vereinfachte Unterbrechungsfreie Upgrades

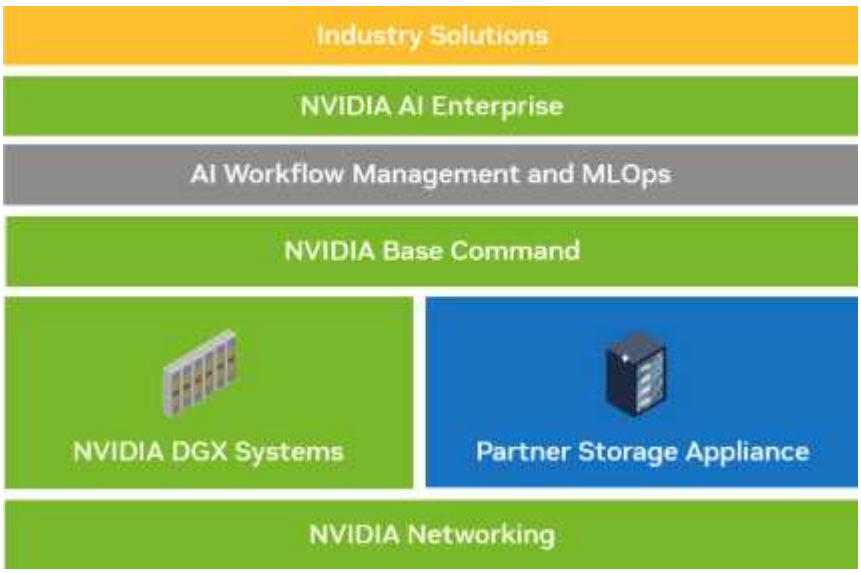
Das AFF A90 ist bei bestehenden A800 Kunden als unterbrechungsfreies Upgrade im Chassis erhältlich. Mit den RASM-Funktionen (Advanced Reliability, Verfügbarkeit, Wartungsfreundlichkeit und Management) von NetApp können Sie Störungen geschäftskritischer Betriebsabläufe ganz einfach aktualisieren. NetApp steigert darüber hinaus die betriebliche Effizienz und vereinfacht tägliche Aufgaben für IT-Teams, da die ONTAP Software automatisch Firmware-Updates für alle Systemkomponenten anwendet.

Für die größten Implementierungen bieten AFF A1K Systeme höchste Performance- und Kapazitätsoptionen, während andere NetApp Storage-Systeme wie AFF A70 und AFF C800 Optionen für kleinere Implementierungen zu geringeren Kosten bieten.

NVIDIA DGX BasePOD

NVIDIA DGX BasePOD ist eine integrierte Lösung, die aus NVIDIA Hardware- und Softwarekomponenten, MLOps-Lösungen und Storage von Drittanbietern besteht. Unter Nutzung von Best Practices im Scale-out-Systemdesign mit NVIDIA Produkten und validierten Partnerlösungen können Kunden eine effiziente und einfach zu managende Plattform für die KI-Entwicklung implementieren. Abbildung 1 zeigt die verschiedenen Komponenten von NVIDIA DGX BasePOD.

NVIDIA DGX BasePOD Solution



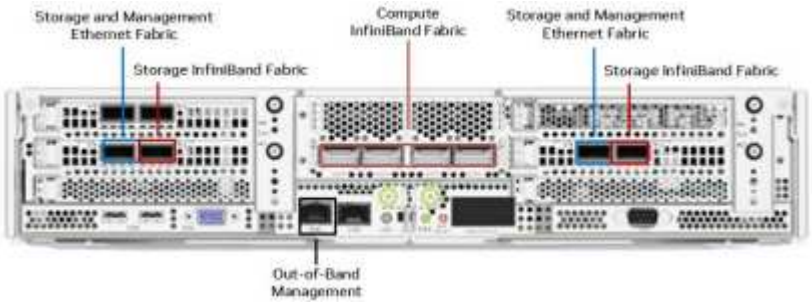
NVIDIA DGX H100-SYSTEME

Das NVIDIA DGX H100™-System ist das KI-Kraftpaket, das durch die bahnbrechende Performance der NVIDIA H100 Tensor Core GPU beschleunigt wird.

NVIDIA DGX H100-SYSTEM

Die wichtigsten Spezifikationen des DGX H100 Systems sind: • Acht NVIDIA H100 GPUs. • 80 GB GPU-Speicher pro GPU, insgesamt 640 GB. • Vier NVIDIA NVSwitch™-Chips. • Intel® Xeon® Platinum 8480 Dualcore-Prozessoren mit 56 Kernen und PCIe 5.0-Unterstützung. • 2 TB DDR5-Systemspeicher. • Vier OSFP-Ports für acht NVIDIA ConnectX- und#174;-7-Adapter (InfiniBand/Ethernet) und zwei NVIDIA ConnectX-7-Adapter (InfiniBand/Ethernet) mit zwei Ports. • Zwei M.2-NVMe-Laufwerke mit 1.92 TB für DGX OS, acht U.2-NVMe-Laufwerke mit 3.84 TB für Storage/Cache. • 10.2 kW maximale Leistung. Die hinteren Ports des DGX H100 CPU-Einschublade sind unten dargestellt. Vier der OSFP-Ports bedienen acht ConnectX-7-Adapter für die InfiniBand Computing-Fabric. Jedes Paar ConnectX-7-Adapter mit zwei Ports bietet parallele Pfade zu den Speicher- und Management Fabrics. Der Out-of-Band-Port wird für den BMC-Zugriff verwendet.

NVIDIA DGX H100-Rückwand



NVIDIA Networking

NVIDIA Quantum-2 QM9700-Switch

NVIDIA Quantum-2 QM9700 InfiniBand Switch

NVIDIA Quantum-2 QM9700 Switches mit InfiniBand-Konnektivität von 400 GB/s versorgen das Computing-Fabric in NVIDIA Quantum-2 InfiniBand BasePOD Konfigurationen. ConnectX-7 Single-Port-Adapter werden für die InfiniBand Computing-Fabric verwendet. Jedes NVIDIA DGX System verfügt über duale Verbindungen zu jedem QM9700 Switch und liefert damit mehrere Pfade mit hoher Bandbreite und niedriger Latenz zwischen den Systemen.

NVIDIA Spectrum-3 SN4600-Switch

NVIDIA Spectrum-3 SN4600-Switch

NVIDIA Spectrum ™-3 SN4600-Switches bieten insgesamt 128 Ports (64 pro Switch) für redundante Konnektivität zum bandinternen Management von DGX BasePOD. Der NVIDIA SN4600-Switch kann für Geschwindigkeiten zwischen 1 GbE und 200 GbE sorgen. Bei über Ethernet verbundenen Storage Appliances kommen darüber hinaus die NVIDIA SN4600-Switches zum Einsatz. Die Ports der NVIDIA DGX Dual-Port ConnectX-7-Adapter werden sowohl für die bandinterne Verwaltung als auch für die Storage-Konnektivität verwendet.

NVIDIA Spectrum SN2201-Switch

NVIDIA Spectrum SN2201-Switch

NVIDIA Spectrum SN2201-Switches bieten 48 Anschlüsse für die Out-of-Band-Verwaltung. Das Out-of-Band-Management bietet konsolidierte Managementkonnektivität für alle Komponenten in DGX BasePOD.

NVIDIA ConnectX-7-Adapter

NVIDIA ConnectX-7-Adapter

Der NVIDIA ConnectX-7-Adapter bietet einen Durchsatz von 25/50/100/200 G. NVIDIA DGX Systeme verwenden sowohl den Single- als auch den Dual-Port ConnectX-7-Adapter, um die Flexibilität von DGX BasePOD-Implementierungen mit 400 GB/s InfiniBand und Ethernet zu erhöhen.

NVA-1173 NetApp AIPod mit NVIDIA DGX Systemen – Softwarekomponenten

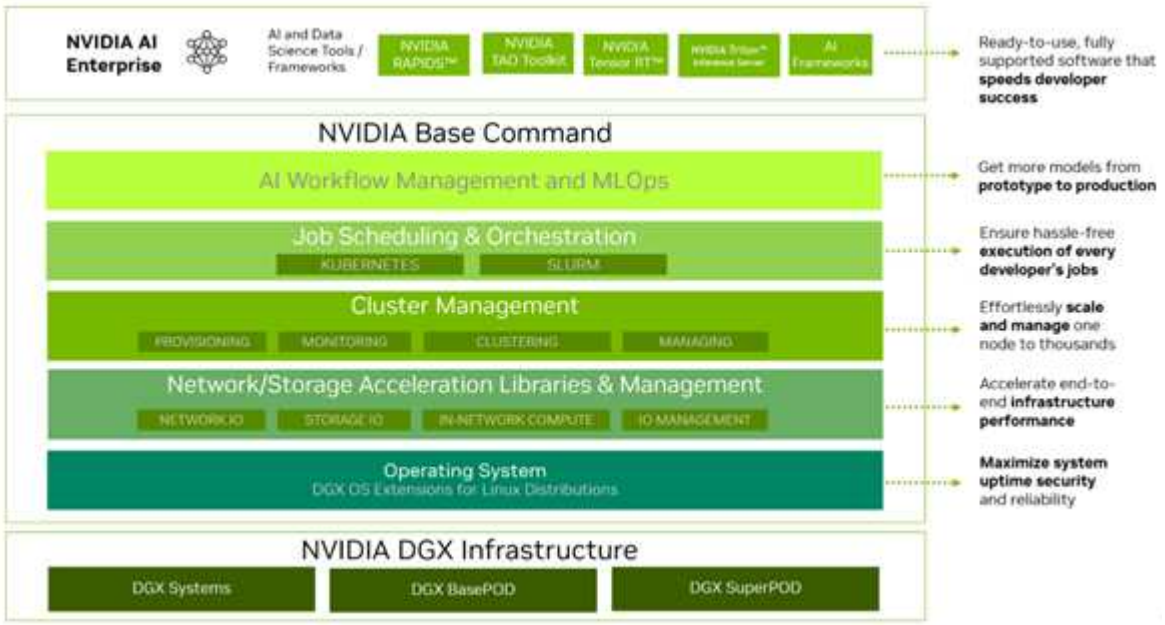
Der Schwerpunkt dieses Abschnitts liegt auf den Softwarekomponenten des NetApp AIPod mit NVIDIA DGX Systemen.

NVIDIA Software

NVIDIA Base-Befehl

NVIDIA Base Command™ unterstützt jeden DGX BasePOD, sodass Unternehmen das Beste aus der innovativen NVIDIA-Software ausschöpfen können. Unternehmen schöpfen das volle Potenzial ihrer Investitionen aus: Mit einer bewährten Plattform, die Orchestrierung und Cluster-Management der Enterprise-Klasse umfasst, Bibliotheken zur Beschleunigung von Computing-, Storage- und Netzwerkinfrastruktur sowie ein für KI-Workloads optimiertes Betriebssystem.

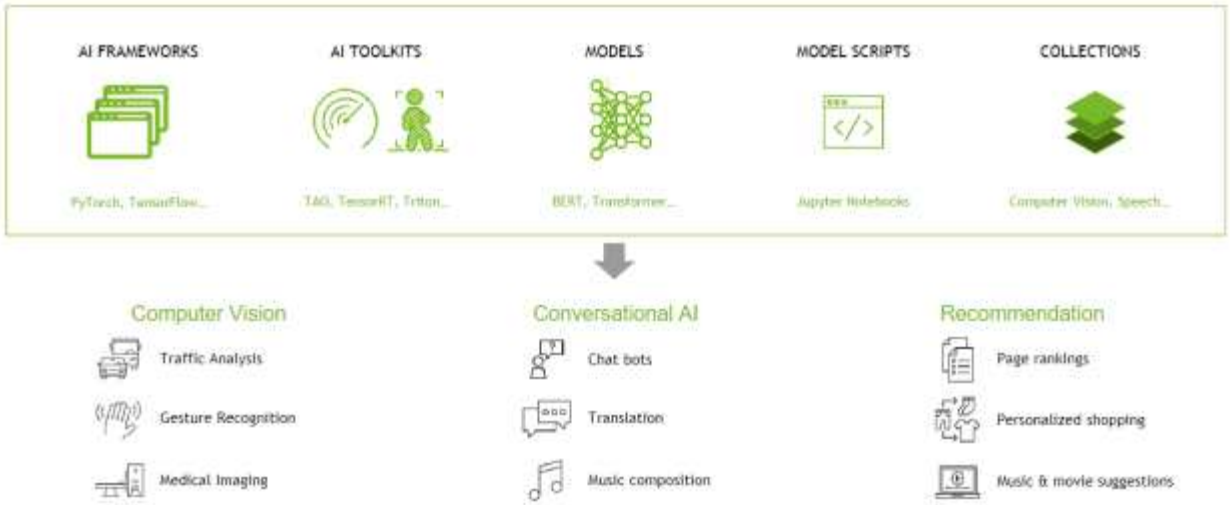
NVIDIA BaseCommand Solution



NVIDIA GPU CLOUD (NGC)

NVIDIA NGC™ bietet Software, die die Anforderungen von Data Scientists, Entwicklern und Forschern mit unterschiedlichen KI-Fachkenntnissen erfüllt. Software, die auf NGC gehostet wird, wird anhand einer aggregierten Reihe von gängigen Schwachstellen und Expositionen (CVEs), Crypto und privaten Schlüsseln untersucht. Die Lösung wurde getestet und zur Skalierung auf mehrere GPUs und in vielen Fällen auf Multi-Node-Systeme konzipiert, damit Benutzer ihre Investitionen in DGX-Systeme maximal ausschöpfen können.

NVIDIA GPU Cloud



NVIDIA AI Enterprise

NVIDIA AI Enterprise ist die End-to-End-Softwareplattform, die generative KI für jedes Unternehmen zugänglich macht und die schnellste und effizienteste Laufzeit für generative KI-Grundmodelle bietet, die für die Ausführung auf der NVIDIA DGX-Plattform optimiert sind. Mit Sicherheit, Stabilität und Verwaltbarkeit auf Produktionsniveau optimiert es die Entwicklung generativer KI-Lösungen. NVIDIA AI Enterprise ist in DGX BasePOD integriert, damit Entwickler Zugriff auf vortrainierte Modelle, optimierte Frameworks, Microservices, beschleunigte Bibliotheken und Enterprise-Support haben.

NetApp Software

NetApp ONTAP

ONTAP 9, die jüngste Generation der Storage-Managementsoftware von NetApp, ermöglicht Unternehmen eine Modernisierung der Infrastruktur und den Übergang zu einem Cloud-fähigen Datacenter. Dank der erstklassigen Datenmanagementfunktionen lassen sich mit ONTAP sämtliche Daten mit einem einzigen Toolset managen und schützen, ganz gleich, wo sich diese Daten befinden. Zudem können Sie die Daten problemlos dorthin verschieben, wo sie benötigt werden: Zwischen Edge, Core und Cloud. ONTAP 9 umfasst zahlreiche Funktionen, die das Datenmanagement vereinfachen, geschäftskritische Daten beschleunigen und schützen und Infrastrukturfunktionen der nächsten Generation über Hybrid-Cloud-Architekturen hinweg ermöglichen.

Beschleunigung und Sicherung von Daten

ONTAP bietet überdurchschnittliche Performance und Datensicherung, erweitert diese Funktionen auf folgende Weise:

- Performance und niedrige Latenz: ONTAP bietet höchstmöglichen Durchsatz bei geringstmöglicher Latenz, einschließlich Unterstützung für NVIDIA GPUDirect Storage (GDS) mit NFS over RDMA, Parallel NFS (pNFS) und NFS Session Trunking.
- Datensicherung ONTAP bietet integrierte Funktionen für die Datensicherung und die branchenweit stärkste Garantie für Ransomware-Schutz mit einem einheitlichen Management über alle Plattformen hinweg.
- NetApp Volume Encryption (NVE) ONTAP bietet native Verschlüsselung auf Volume-Ebene und unterstützt sowohl Onboard- als auch externes Verschlüsselungsmanagement.
- Storage-Mandantenfähigkeit und Multi-Faktor-Authentifizierung. ONTAP ermöglicht die gemeinsame Nutzung von Infrastrukturressourcen mit höchstmöglicher Sicherheit.

Vereinfachtes Datenmanagement

Für den Enterprise IT-Betrieb und die Data Scientists spielt Datenmanagement eine zentrale Rolle, damit für KI-Applikationen die entsprechenden Ressourcen zum Training von KI/ML-Datensätzen verwendet werden. Die folgenden zusätzlichen Informationen über NetApp Technologien sind bei dieser Validierung nicht im Umfang enthalten, können jedoch je nach Ihrer Implementierung relevant sein.

Die ONTAP Datenmanagement-Software umfasst die folgenden Funktionen, um den Betrieb zu optimieren und zu vereinfachen und damit Ihre Gesamtbetriebskosten zu senken:

- Snapshots und Klone ermöglichen Zusammenarbeit, parallele Experimente und erweiterte Daten-Governance für ML/DL-Workflows.
- SnapMirror ermöglicht die nahtlose Datenverschiebung in Hybrid-Cloud- und Multi-Site-Umgebungen, sodass Daten jederzeit und überall zur Verfügung stehen.
- Inline-Data-Compaction und erweiterte Deduplizierung: Data-Compaction reduziert den ungenutzten Speicherplatz in Storage-Blöcken, während Deduplizierung die effektive Kapazität deutlich steigert. Dies gilt für lokal gespeicherte Daten und für Daten-Tiering in die Cloud.
- Minimale, maximale und adaptive Quality of Service (AQoS): Durch granulare QoS-Einstellungen (Quality of Service) können Unternehmen ihre Performance-Level für kritische Applikationen auch in Umgebungen mit vielen unterschiedlichen Workloads garantieren.
- NetApp FlexGroups ermöglichen die Verteilung von Daten auf alle Nodes im Storage Cluster und sorgen für äußerst große Datensätze mit enormer Kapazität und höherer Performance.
- NetApp FabricPool: Bietet automatisches Tiering von „kalten“ Daten in Private- und Public-Cloud-Storage-Optionen, einschließlich Amazon Web Services (AWS), Azure und NetApp StorageGRID Storage-Lösung. Weitere Informationen zu FabricPool finden Sie unter "TR-4598: FabricPool Best Practices".
- NetApp FlexCache: Mit Remote-Caching-Funktionen für Volumes vereinfachen Sie die Dateiverteilung und senken die WAN-Latenz sowie die Kosten für die WAN-Bandbreite. FlexCache ermöglicht eine über mehrere Standorte verteilte Produktentwicklung sowie einen schnelleren Zugriff auf Unternehmensdatensätze von Remote-Standorten aus.

Zukunftssichere Infrastruktur

ONTAP bietet folgende Funktionen, um anspruchsvolle und sich ständig ändernde Geschäftsanforderungen zu erfüllen:

- Nahtlose Skalierung und unterbrechungsfreier Betrieb: Die ONTAP unterstützt das Online-Hinzufügen von Kapazität zu vorhandenen Controllern und das Scale-out von Clustern. Kunden können Upgrades auf die neuesten Technologien wie NVMe und 32 GB FC ohne teure Datenmigrationen oder Ausfälle durchführen.
- Cloud-Anbindung: ONTAP ist die Storage-Managementsoftware mit der umfassendsten Cloud-Integration und bietet Optionen für softwaredefinierten Storage (ONTAP Select) und Cloud-native Instanzen (Google Cloud NetApp Volumes) in allen Public Clouds.
- Integration in moderne Applikationen: ONTAP bietet Datenservices der Enterprise-Klasse für Plattformen und Applikationen der neuesten Generation, wie autonome Fahrzeuge, Smart Citys und Industrie 4.0, auf derselben Infrastruktur, die bereits vorhandene Unternehmensanwendungen unterstützt.

NetApp DataOps Toolkit

Das NetApp DataOps Toolkit ist ein Python-basiertes Tool zur Vereinfachung des Managements von Entwicklungs-/Trainings-Workspaces und Inferenzservern, die durch hochleistungsfähigen, horizontal skalierbaren NetApp Storage gesichert werden. Das DataOps Toolkit kann als eigenständiges Dienstprogramm ausgeführt werden. Noch effektiver ist es in Kubernetes-Umgebungen, in denen NetApp Trident Storage-

Vorgänge automatisiert. Die wichtigsten Funktionen:

- Schnelle Bereitstellung neuer JupyterLab Workspaces mit hoher Kapazität, die durch hochperformanten horizontal skalierbaren NetApp Storage unterstützt werden
- Schnelle Bereitstellung neuer NVIDIA Triton Inferenz Server Instanzen, die durch NetApp Storage der Enterprise-Klasse unterstützt werden
- Nahezu sofortiges Klonen von JupyterLab Workspaces mit hoher Kapazität für Experimentierfreudigkeit oder schnelle Iterationen
- Nahezu sofortige Snapshots von JupyterLab Workspaces mit hoher Kapazität für Backups und/oder Rückverfolgbarkeit/Baselining.
- Bereitstellung, Klonen und Snapshots hochperformanter Daten-Volumes nahezu sofort

NetApp Trident

Trident ist ein vollständig unterstützter Open-Source-Storage-Orchestrator für Container und Kubernetes-Distributionen, einschließlich Anthos. Trident kann mit dem gesamten NetApp Storage-Portfolio, einschließlich NetApp ONTAP, eingesetzt werden. Darüber hinaus werden NFS-, NVMe/TCP- und iSCSI-Verbindungen unterstützt. Trident beschleunigt den DevOps-Workflow, da Endbenutzer Storage über ihre NetApp Storage-Systeme bereitstellen und managen können, ohne dass ein Storage-Administrator eingreifen muss.

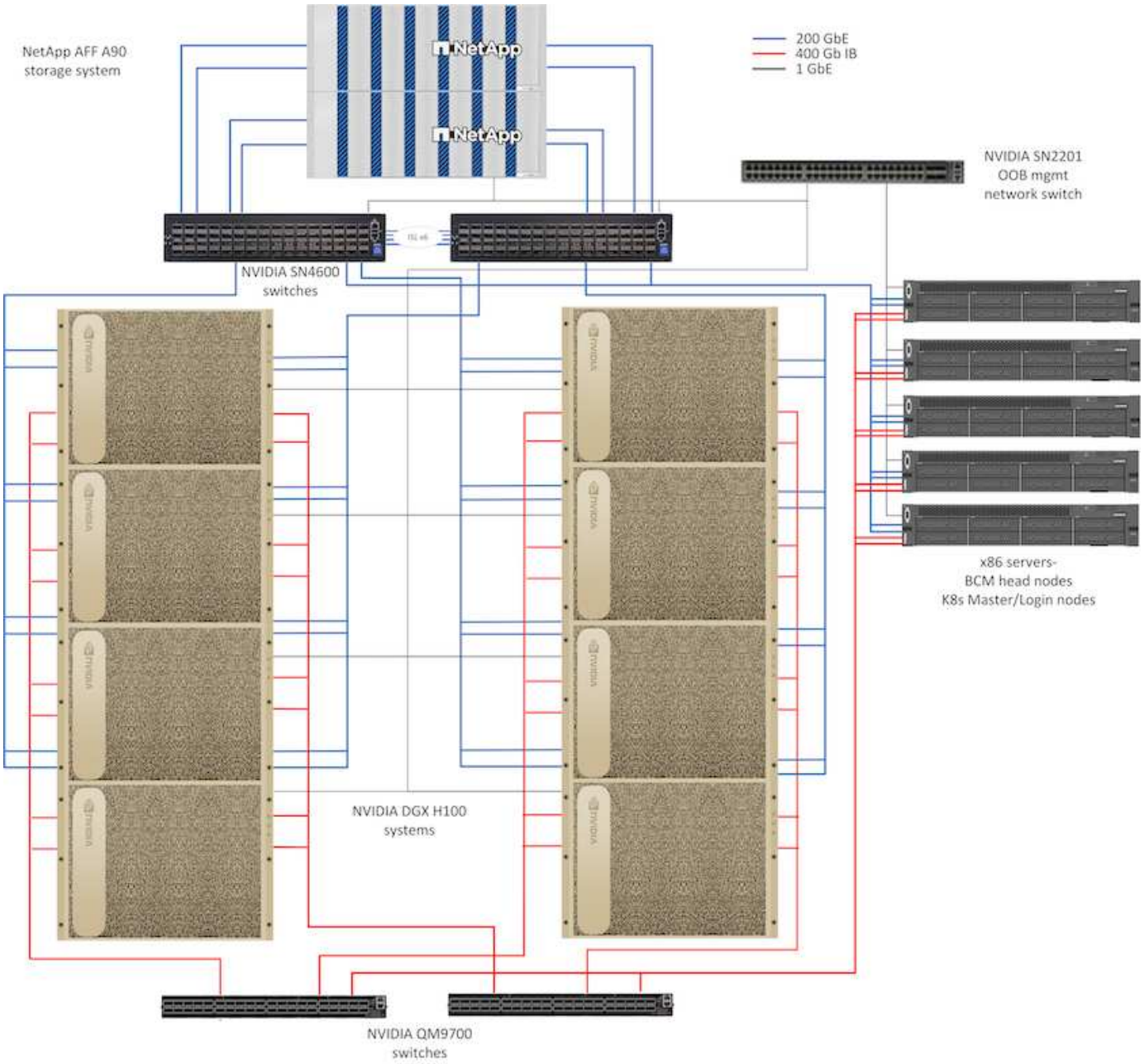
NVA-1173 NetApp AIPod mit NVIDIA DGX H100 Systemen – Lösungsarchitektur

Dieser Abschnitt konzentriert sich auf die Architektur des NetApp AIPod mit NVIDIA DGX-Systemen.

NetApp AIPod mit DGX Systemen

Diese Referenzarchitektur nutzt separate Fabrics für den Compute-Cluster-Interconnect- und Storage-Zugriff, wobei die InfiniBand (IB) mit 400 GB/s zwischen den Computing-Nodes verbunden ist. Die Abbildung unten zeigt die allgemeine Lösungstopologie von NetApp AIPod mit DGX H100 Systemen.

NetApp AIPOD-Lösungstopologie



Netzwerkdesign

In dieser Konfiguration verwendet die Computing-Cluster-Fabric ein Paar QM9700-400-GB/s-IB-Switches, die miteinander verbunden sind, um Hochverfügbarkeit zu gewährleisten. Jedes DGX H100-System ist über acht Verbindungen mit den Switches verbunden, wobei die Ports mit geraden Nummern mit einem Switch verbunden sind und die Ports mit ungeraden Nummern mit dem anderen Switch verbunden sind.

Für den Zugriff auf das Speichersystem, das bandinterne Management und den Client-Zugriff wird ein Paar SN4600 Ethernet-Switches verwendet. Die Switches sind mit Verbindungen zwischen Switches verbunden und mit mehreren VLANs konfiguriert, um die verschiedenen Datenverkehrstypen zu isolieren. Das grundlegende L3-Routing wird zwischen bestimmten VLANs aktiviert, um mehrere Pfade zwischen Client- und Speicherschnittstellen auf demselben Switch sowie zwischen Switches zu ermöglichen, um eine hohe Verfügbarkeit zu gewährleisten. Bei größeren Implementierungen kann das Ethernet-Netzwerk nach Bedarf durch zusätzliche Switch-Paare für Spine-Switches und zusätzliche Leaves zu einer Leaf-Spine-Konfiguration erweitert werden.

Neben dem Compute Interconnect und High-Speed-Ethernet-Netzwerken sind alle physischen Geräte zur Out-of-Band-Verwaltung auch mit einem oder mehreren SN2201 Ethernet-Switches verbunden. "Einzelheiten zur Implementierung"Weitere Informationen zur Netzwerkkonfiguration finden Sie auf der Seite.

Storage-Zugriffsübersicht für DGX H100 Systeme

Jedes DGX H100-System verfügt über zwei Dual-Port-ConnectX-7-Adapter für Management- und Storage-Datenverkehr. Bei dieser Lösung werden beide Ports auf jeder Karte mit demselben Switch verbunden. Ein Port von jeder Karte wird dann in einer LACP MLAG-Verbindung konfiguriert, wobei ein Port mit jedem Switch verbunden ist. VLANs für in-Band-Management, Client-Zugriff und Speicherzugriff auf Benutzerebene werden auf dieser Verbindung gehostet.

Der andere Port auf jeder Karte wird für die Verbindung zu den AFF A90 Storage-Systemen verwendet und kann je nach Workload-Anforderungen in mehreren Konfigurationen verwendet werden. Bei Konfigurationen, die NFS über RDMA verwenden, um NVIDIA Magnum IO GPUDirect Storage zu unterstützen, werden die Ports einzeln mit IP-Adressen in separaten VLANs verwendet. Für Implementierungen, die kein RDMA erfordern, können die Storage-Schnittstellen auch mit LACP Bonding konfiguriert werden, um Hochverfügbarkeit und zusätzliche Bandbreite zu gewährleisten. Mit oder ohne RDMA können Clients das Storage-System mit pNFS und Session-Trunking für NFS v4.1 mounten, um parallelen Zugriff auf alle Storage Nodes im Cluster zu ermöglichen. "Einzelheiten zur Implementierung"Weitere Informationen zur Client-Konfiguration finden Sie auf der Seite.

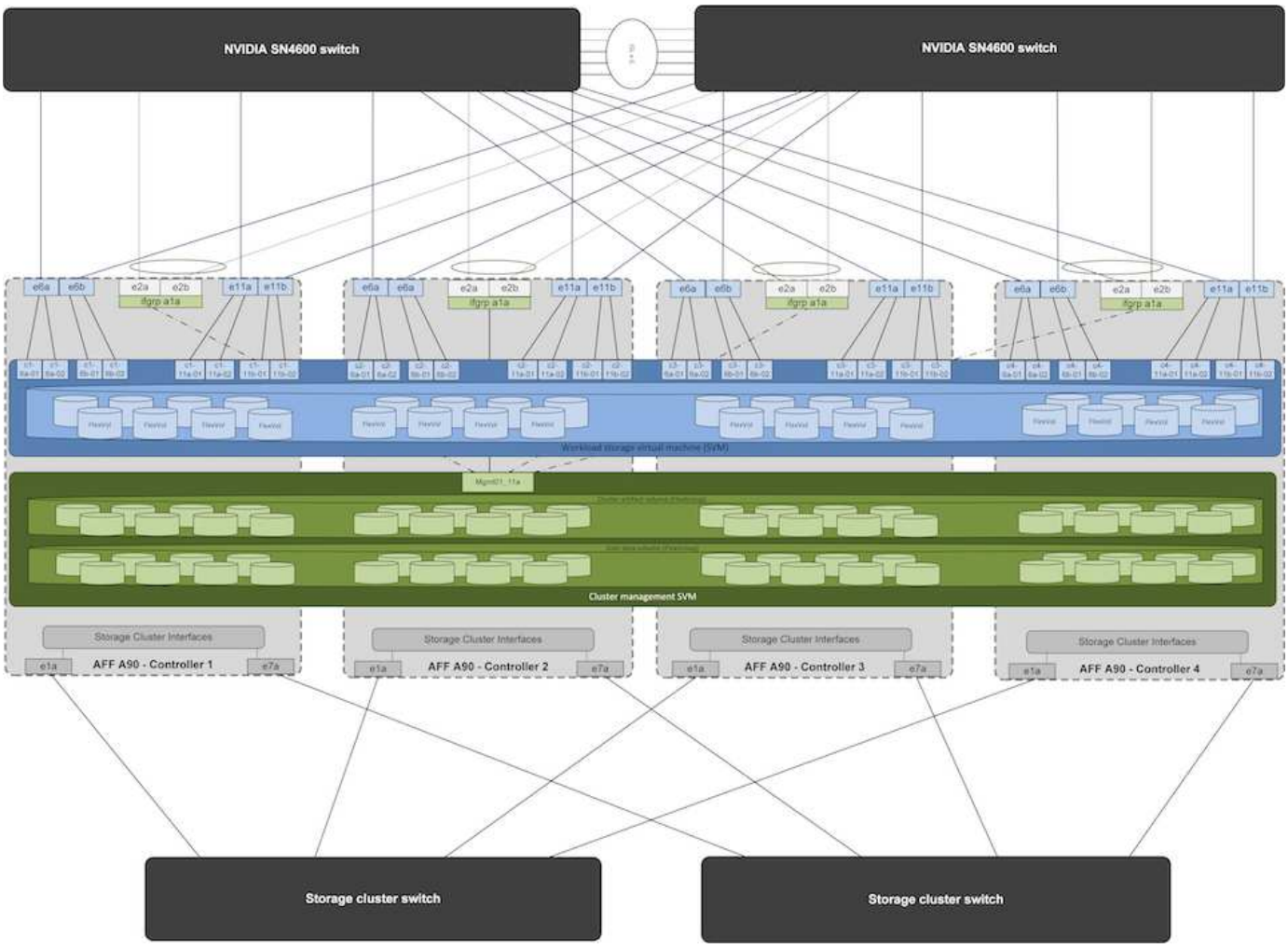
Weitere Informationen zur Konnektivität des DGX H100-Systems finden Sie in der "NVIDIA BasePOD-Dokumentation".

Design von Storage-Systemen

Jedes AFF A90 Storage-System ist über sechs 200-GbE-Ports von jedem Controller verbunden. Vier Ports von jedem Controller werden für den Workload-Datenzugriff aus den DGX-Systemen verwendet. Zwei Ports von jedem Controller werden als LACP-Schnittstellengruppe konfiguriert, um den Zugriff über die Server der Managementebene für Cluster-Managementartefakte und Benutzer-Home Directorys zu unterstützen. Der gesamte Zugriff auf die Daten aus dem Storage-System erfolgt über NFS, wobei eine Storage Virtual Machine (SVM) dediziert für den KI-Workload-Zugriff und eine separate SVM für das Cluster-Management vorgesehen ist.

Die Management SVM benötigt nur eine einzelne LIF, die sich auf den auf jedem Controller konfigurierten 2-Port-Schnittstellengruppen befindet. Andere FlexGroup Volumes werden auf der Management-SVM bereitgestellt, um Artefakte im Cluster-Management wie Cluster-Node-Images, Systemüberwachungsdaten und Home Directories der Endbenutzer zu beherbergen. Die nachfolgende Abbildung zeigt die logische Konfiguration des Storage-Systems.

NetApp A90 logische Konfiguration des Storage-Clusters



Server auf Managementebene

Diese Referenzarchitektur enthält außerdem fünf CPU-basierte Server für die Nutzung der Verwaltungsebene. Zwei dieser Systeme werden als Hauptknoten für NVIDIA Base Command Manager für die Cluster-Implementierung und -Verwaltung verwendet. Die anderen drei Systeme werden verwendet, um zusätzliche Cluster-Services wie Kubernetes-Master-Nodes oder Login-Nodes für Implementierungen bereitzustellen, die Slurm für die Jobplanung verwenden. Implementierungen mit Kubernetes können den NetApp Trident CSI-Treiber nutzen, um automatisierte Bereitstellungs- und Datenservices mit persistentem Storage für Management- und KI-Workloads auf dem AFF A900 Storage-System bereitzustellen.

Jeder Server ist physisch mit den IB-Switches und Ethernet-Switches verbunden, um Cluster-Implementierung und -Management zu ermöglichen. Er ist zur Speicherung von Clustermanagement-Artefakten wie oben beschrieben mit NFS-Mounts zum Storage-System konfiguriert.

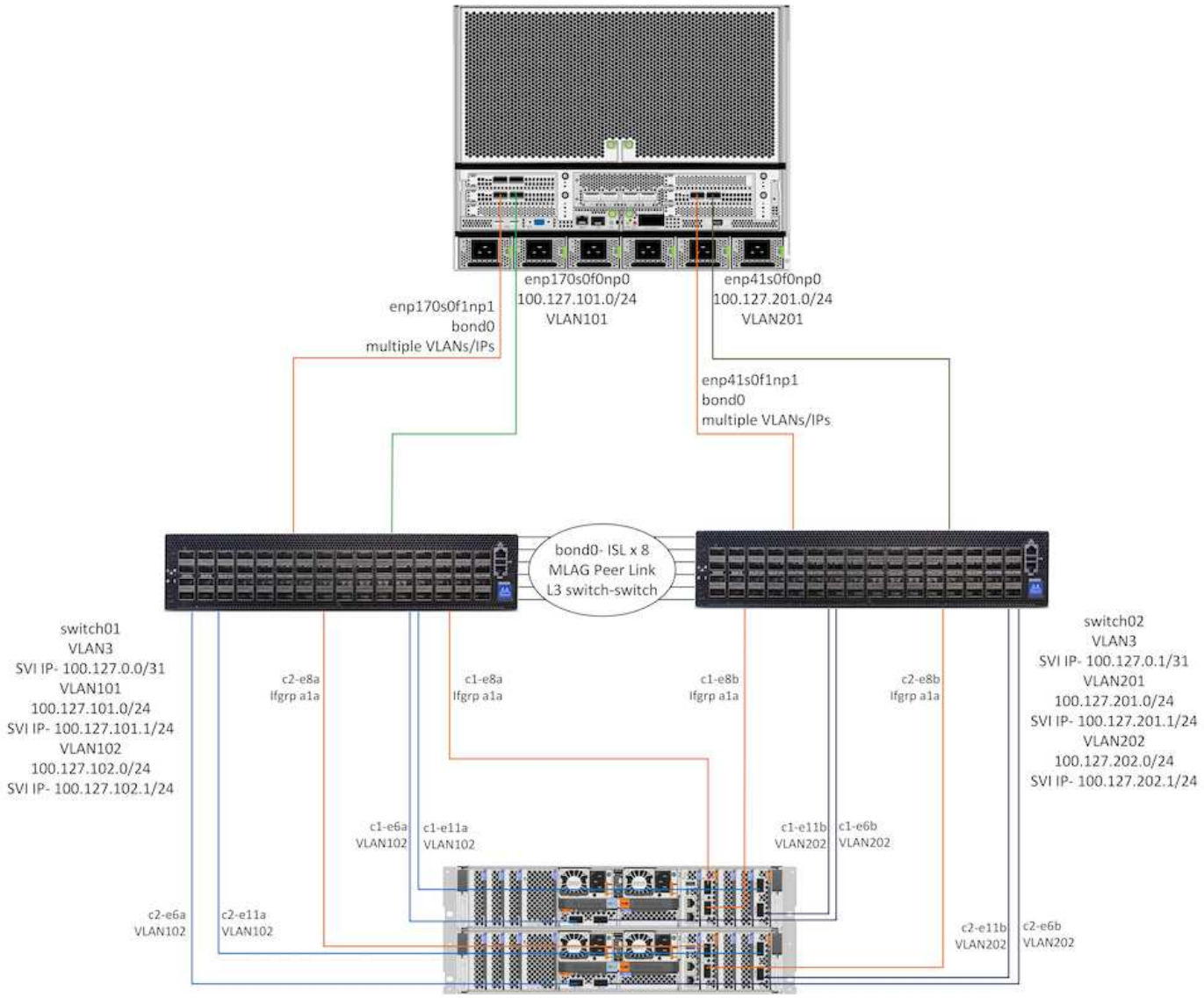
NVA-1173 NetApp AIPod mit NVIDIA DGX Systemen – Implementierungsdetails

In diesem Abschnitt werden die Einzelheiten zur Implementierung beschrieben, die bei der Validierung dieser Lösung verwendet werden. Die verwendeten IP-Adressen sind Beispiele und sollten entsprechend der Bereitstellungsumgebung geändert werden. Weitere Informationen zu bestimmten Befehlen, die bei der Implementierung dieser Konfiguration verwendet werden, finden Sie in der entsprechenden

Produktdokumentation.

Das folgende Diagramm zeigt detaillierte Netzwerk- und Konnektivitätsinformationen für 1 DGX H100-System und 1 HA-Paar AFF A90 Controller. Die Bereitstellungsleitlinien in den folgenden Abschnitten basieren auf den Details in diesem Diagramm.

NetApp AIPOD Netzwerkkonfiguration



Die folgende Tabelle zeigt Beispiel-Verkabelungszuweisungen für bis zu 16 DGX-Systeme und 2 AFF A90 HA-Paare.

Switch und Port	Gerät	Geräteanschluss
Switch 1-Ports 1-16	DGX-H100-01 bis -16	Enp170s0f0np0, Steckplatz 1, Anschluss 1
Switch 1-Ports 17-32	DGX-H100-01 bis -16	Enp170s0f1np1, Steckplatz 1, Anschluss 2
Switch 1-Ports 33-36	AFF-A90-01 bis -04	Port e6a
Switch 1-Ports 37-40	AFF-A90-01 bis -04	Port e11a

Switch und Port	Gerät	Geräteanschluss
Switch 1-Ports 41-44	AFF-A90-01 bis -04	Anschluss e2a
Switch 1-Ports 57-64	ISL zu switch2	Anschlüsse 57-64
Switch2-Ports 1-16	DGX-H100-01 bis -16	Enp41s0f0np0, Steckplatz 2, Port 1
Switch2-Ports 17-32	DGX-H100-01 bis -16	Enp41s0f1np1, Steckplatz 2, Anschluss 2
Switch2-Ports 33-36	AFF-A90-01 bis -04	Anschluss e6b
Switch2-Ports 37-40	AFF-A90-01 bis -04	Port e11b
Switch2-Ports 41-44	AFF-A90-01 bis -04	Anschluss e2b
Switch2-Ports 57-64	ISL zu switch1	Anschlüsse 57-64

Die folgende Tabelle zeigt die Softwareversionen für die verschiedenen Komponenten, die bei dieser Validierung verwendet werden.

Gerät	Softwareversion
NVIDIA SN4600-Switches	Cumulus Linux v5.9.1
NVIDIA DGX-System	DGX OS 6.2.1 (Ubuntu 22.04 LTS)
Mellanox OFED	24,01
NetApp AFF A90	NetApp ONTAP 9.14.1

Konfiguration des Storage-Netzwerks

In diesem Abschnitt werden wichtige Details zur Konfiguration des Ethernet Storage-Netzwerks erläutert. Informationen zur Konfiguration des InfiniBand-Compute-Netzwerks finden Sie im "NVIDIA BasePOD-Dokumentation". Weitere Informationen zur Switch-Konfiguration finden Sie im "NVIDIA Cumulus Linux Dokumentation".

Im Folgenden werden die grundlegenden Schritte zur Konfiguration der SN4600-Switches erläutert. Bei diesem Prozess wird vorausgesetzt, dass die Verkabelung und grundlegende Switch-Einrichtung (Management-IP-Adresse, Lizenzierung usw.) abgeschlossen sind.

1. Konfigurieren Sie die ISL-Verbindung zwischen den Switches, um Multi-Link-Aggregation (MLAG) und Failover-Datenverkehr zu aktivieren
   - Bei dieser Validierung wurden 8 Links verwendet, um mehr als genug Bandbreite für die zu testende Speicherkonfiguration bereitzustellen
   - Spezifische Anweisungen zur Aktivierung von MLAG finden Sie in der Dokumentation zu Cumulus Linux.
2. Konfigurieren Sie LACP MLAG für jedes Paar von Client-Ports und Speicherports an beiden Switches
   - Port swp17 auf jedem Switch für DGX-H100-01 (enp170s0f1np1 und enp41s0f1np1), Port swp18 für DGX-H100-02 usw. (bond1-16)
   - Port swp41 auf jedem Switch für AFF-A90-01 (e2a und e2b), Port swp42 für AFF-A90-02 usw. (bond17-20)

- nv-Set-Schnittstelle bondX Bond-Mitglied swpX
- nv-Set-Schnittstelle bondx Bond-Mlag-id X

3. Fügen Sie alle Ports und MLAG-Bindungen zur Standard-Bridge-Domäne hinzu
   - nv-Satz int swp1-16,33-40 Bridge-Domäne br_default
   - nv set int bond1-20 Bridge Domain br_default
4. Aktivieren Sie RoCE auf jedem Switch
   - nv-Einstellung roce-Modus verlustfrei
5. Konfigurieren von VLANs: 2 für Client-Ports, 2 für Speicherports, 1 für Verwaltung, 1 für L3-Switch zu Switch
   - Schalter 1-
     - VLAN 3 für L3-Switch-zu-Switch-Routing im Falle eines Client-NIC-Ausfalls
     - VLAN 101 für Storage-Port 1 auf jedem DGX-System (enp170s0f0np0, Steckplatz 1 Port 1)
     - VLAN 102 für Port e6a und e11a auf jedem AFF A90-Speicher-Controller
     - VLAN 301 für das Management mithilfe der MLAG-Schnittstellen zu jedem DGX-System und Storage Controller
   - Schalter 2-
     - VLAN 3 für L3-Switch-zu-Switch-Routing im Falle eines Client-NIC-Ausfalls
     - VLAN 201 für Storage Port 2 auf jedem DGX-System (enp41s0f0np0, Steckplatz 2 Port 1)
     - VLAN 202 für Port e6b und e11b auf jedem AFF A90-Speicher-Controller
     - VLAN 301 für das Management mithilfe der MLAG-Schnittstellen zu jedem DGX-System und Storage Controller
6. Weisen Sie jedem VLAN je nach Bedarf physische Ports zu, z. B. Client-Ports in Client-VLANs und Storage-Ports in Storage-VLANs
   - nv set int <swpX> Bridge Domain br_default Access <vlan id>
   - MLAG-Ports sollten als Trunk-Ports bleiben, um bei Bedarf mehrere VLANs über die verbundenen Schnittstellen zu ermöglichen.
7. Konfigurieren Sie virtuelle Switch-Schnittstellen (SVI) auf jedem VLAN, um als Gateway zu fungieren und L3-Routing zu aktivieren
   - Schalter 1-
     - nv setzt int vlan3 ip-Adresse 100.127.0.0/31
     - nv-Einstellung int vlan101 ip-Adresse 100.127.101.1/24
     - nv setzt int vlan102 ip-Adresse 100.127.102.1/24
   - Schalter 2-
     - nv setzt int vlan3 ip-Adresse 100.127.0.1/31
     - nv-Einstellung int vlan201 ip-Adresse 100.127.201.1/24
     - nv setzt int vlan202 ip-Adresse 100.127.202.1/24
8. Erstellen Sie statische Routen
   - Statische Routen werden automatisch für Subnetze auf demselben Switch erstellt
   - Für Switch-to-Switch-Routing im Falle eines Client-Link-Ausfalls sind zusätzliche statische Routen

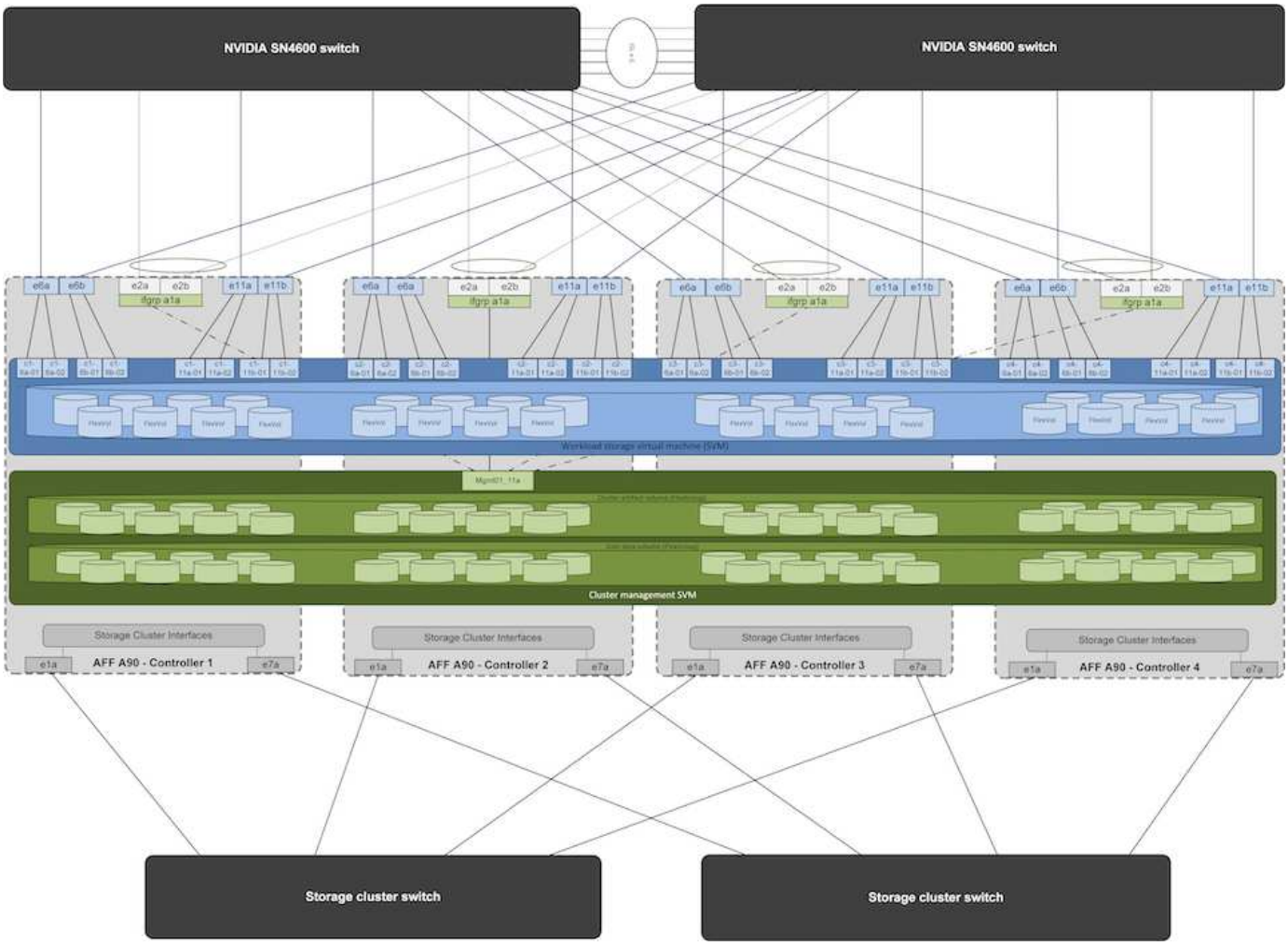
erforderlich

- Schalter 1-
  - nv vrf-Standardrouter statisch einstellen 100.127.128.0/17 über 100.127.0.1
- Schalter 2-
  - nv vrf-Standardrouter statisch einstellen 100.127.0.0/17 über 100.127.0.0

Konfiguration des Storage-Systems

In diesem Abschnitt werden die wichtigsten Details zur Konfiguration des A90-Speichersystems für diese Lösung beschrieben. Weitere Informationen zur Konfiguration von ONTAP Systemen finden Sie in der [ONTAP Dokumentation]. Das folgende Diagramm zeigt die logische Konfiguration des Storage-Systems.

NetApp A90 logische Konfiguration des Storage-Clusters



Im Folgenden werden die grundlegenden Schritte zur Konfiguration des Speichersystems beschrieben. Dabei wird vorausgesetzt, dass die grundlegende Installation des Storage-Clusters abgeschlossen ist.

1. Konfigurieren Sie auf jedem Controller 1 Aggregat mit allen verfügbaren Partitionen minus 1 Spare
   - aggr create -Node <node> -Aggregate <node>_data01 -diskcount <47>
2. Konfigurieren Sie ifrps auf jedem Controller
   - NET Port ifgrp create -Node <node> -ifgrp a1a -Mode Multimode_lacp -distr-Function Port

- NET Port ifgrp add-Port -Node <node> -ifgrp <ifgrp> -Ports <node>:e2a,<node>:e2b

3. Konfigurieren Sie den Management-vlan-Port auf ifgrp auf jedem Controller
   - NET Port vlan create -Node AFF-a90-01 -Port a1a -vlan-id 31
   - NET Port vlan create -Node AFF-a90-02 -Port a1a -vlan-id 31
   - NET Port vlan create -Node AFF-a90-03 -Port a1a -vlan-id 31
   - NET Port vlan create -Node AFF-a90-04 -Port a1a -vlan-id 31
4. Erstellen von Broadcast-Domänen
   - 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-a11a-04:e11a
   - Broadcast-Domain create -Broadcast-Domain vlan22 -mtu 9000 -Ports aaff-a90-01 04:e6b,AFF AFF-a90-01:e11b,AFF-a90-02:e6b,AFF-a90-02:e11b,AFF-a90-03:e6b,AFF-a90-03:e11b,AFF-a90-04:e6b
   - Broadcast-Domain create -Broadcast-Domain vlan31 -mtu 9000 -Ports AFF-a90-01:a1a-31,AFF-a90-02:a1a-31,AFF-a90-03:a1a-31,AFF-a90-04:a1a-31
5. Management-SVM erstellen *
6. Konfiguration der Management-SVM
   - Erstellung von LIF
     - NET int create -vserver basepod-mgmt -lif vlan31-01 -Home-Node AFF-a90-01 -Home-Port a1a-31 -Adresse 192.168.31.X -Netmask 255.255.255.0
   - FlexGroup Volumes erstellen –
     - vol. Erstellen -vserver Basepod-mgmt -Volume Home -size 10T -automatische Bereitstellung-als FlexGroup -Junction-path /Home
     - vol. Erstellen -vserver Basepod-mgmt -Volume cm -Größe 10T -automatische Bereitstellung-als FlexGroup -Verbindungspfad/cm
   - Erstellen der Exportrichtlinie
     - Regel für Export create -vserver basepod-mgmt -Policy default -Client-match 192.168.31.0/24 -rorule sys -rwrule sys -Superuser sys
7. Daten-SVM erstellen *
8. Daten-SVM konfigurieren
   - SVM für RDMA-Unterstützung konfigurieren
     - vserver nfs modify -vserver basepod-Data -rdma aktiviert
   - Erstellung der LIFs
     - 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 -Adresse 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 -Adresse 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. Konfigurieren Sie LIFs für RDMA-Zugriff
   - Für Implementierungen mit ONTAP 9.15.1 erfordert die RoCE-QoS-Konfiguration für physische Informationen Befehle auf betriebssystemebene, die in der ONTAP-CLI nicht verfügbar sind. Wenden Sie sich an den NetApp-Support, wenn Sie Hilfe bei der Konfiguration der Ports für den RoCE-Support benötigen. NFS über RDMA Funktionen ohne Probleme
   - Ab ONTAP 9.16.1 werden physische Schnittstellen automatisch mit den entsprechenden Einstellungen für eine End-to-End-RoCE-Unterstützung konfiguriert.
   - NET int modify -vserver basepod-Data -lif * -rdma-protocols roce
10. Konfigurieren Sie NFS-Parameter auf der Daten-SVM
    - nfs modify -vserver basepod-Data -v4.1 enabled -v4.1-pnfs aktiviert -v4.1-Trunking aktiviert -tcp-max -Transfer-size 262144
11. FlexGroup Volumes erstellen –
    - vol Create -vserver Basepod-Data -Volume Data -size 100T -Auto-Bereitstellung-als FlexGroup -Junction-Path /Data
12. Erstellen der Exportrichtlinie
    - Regel für Export-Policy create -vserver basepod-Data -Policy default -Client-match 100.127.101.0/24 -rorule sys -rwrule sys -Superuser sys
    - Regel für Export-Policy create -vserver basepod-Data -Policy default -Client-match 100.127.201.0/24 -rorule sys -rwrule sys -Superuser sys
13. Erstellen Sie Routen

- Route add -vserver basepod_Data -Destination 100.127.0.0/17 -Gateway 100.127.102.1 metrisch 20
- Route add -vserver basepod_Data -Destination 100.127.0.0/17 -Gateway 100.127.202.1 metrisch 30
- Route add -vserver basepod_Data -Destination 100.127.128.0/17 -Gateway 100.127.202.1 metrisch 20
- Route add -vserver basepod_Data -Destination 100.127.128.0/17 -Gateway 100.127.102.1 metrisch 30

DGX H100-Konfiguration für RoCE-Storage-Zugriff

In diesem Abschnitt werden die wichtigsten Details zur Konfiguration der DGX H100-Systeme beschrieben. Viele dieser Konfigurationselemente können in das OS-Image enthalten werden, das in den DGX-Systemen implementiert wurde, oder vom Base Command Manager beim Booten implementiert werden. Sie sind hier als Referenz aufgeführt. Weitere Informationen zur Konfiguration von Knoten und Software-Images in BCM finden Sie in der "BCM-Dokumentation".

1. Installieren Sie zusätzliche Pakete
   - ipmitool
   - python3-Pip
2. Installieren Sie Python-Pakete
   - Paramiko
   - matplotlib
3. Konfigurieren Sie dpkg nach der Paketinstallation neu
   - Dpkg --configure -a
4. Installieren Sie MOFED
5. Mst-Werte für Performance Tuning festlegen
   - Mstconfig -y -d <aa:00.0,29:00.0> set ADVANCED_PCI_SETTINGS=1 NUM_OF_VFS=0 MAX_ACC_OUT_READ=44
6. Setzen Sie die Adapter nach dem Ändern der Einstellungen zurück
   - Mlxfwreset -d <aa:00.0,29:00.0> -y zurücksetzen
7. Legen Sie MaxReadReq auf PCI-Geräten fest
   - Setpci -s <aa:00.0,29:00.0> 68.W=5957
8. Legen Sie die Größe des RX- und TX-Ringpuffers fest
   - Ethtool -G <enp170s0f0np0,enp41s0f0np0> rx 8192 tx 8192
9. Legen Sie PFC und DSCP unter Verwendung von mlnx_qos fest
   - Mlnx_qos -i <enp170s0f0np0,enp41s0f0np0> --pfc 0,0,0,1,0,0,0,0 --Trust=dscp --cable_len=3
10. Legen Sie ToS für RoCE-Traffic auf Netzwerk-Ports fest
    - Echo 106 > /sys/class/infiniband/<mlx5_7,mlx5_1>/tc/1/traffic_class
11. Konfigurieren Sie jede Speicher-NIC mit einer IP-Adresse im entsprechenden Subnetz
    - 100.127.101.0/24 für Speicher-NIC 1
    - 100.127.201.0/24 für Speicher-NIC 2
12. Bandinterne Netzwerk-Ports für LACP-Bonding konfigurieren (enp170s0f1np1,enp41s0f1np1)
13. Konfigurieren Sie statische Routen für primäre und sekundäre Pfade zu jedem Storage-Subnetz
    - Route addieren –net 100.127.0.0/17 gw 100.127.101.1 metrisch 20

- Route addieren –net 100.127.0.0/17 gw 100.127.201.1 metrisch 30
- Route addieren – netto 100.127.128.0/17 gw 100.127.201.1 metrisch 20
- Route addieren – netto 100.127.128.0/17 gw 100.127.101.1 metrisch 30

14. Mounten Sie /Home Volume
    - Mount -o vers=3,nconnect=16,rsize=262144,wsize=262144 192.168.31.X:/Home /Home
15. Mounten Sie /Data Volume
    - Beim Mounten des Daten-Volumes wurden die folgenden Mount-Optionen verwendet:
        - Vers=4.1 # ermöglicht pNFS für parallelen Zugriff auf mehrere Storage Nodes
        - Proto=rdma # setzt das Übertragungsprotokoll auf RDMA anstelle des Standard-TCP
        - max_connect=16 # ermöglicht das NFS-Session-Trunking zur aggregierten Storage-Port-Bandbreite
        - Write=Eager # verbessert die Schreib-Performance von gepufferten Schreibvorgängen
        - Rsize=262144,wsize=262144 # setzt die E/A-Übertragungsgröße auf 256 KB

NVA-1173 NetApp AIPod mit NVIDIA DGX Systemen – Leitfaden zur Lösungsvalidierung und Größenbestimmung

Der Abschnitt konzentriert sich auf die Lösungsvalidierung und die Anleitung zur Dimensionierung des NetApp AIPod mit NVIDIA DGX-Systemen.

Lösungsvalidierung

Die Storage-Konfiguration in dieser Lösung wurde mit Hilfe des Open Source-Tools FIO mit einer Reihe synthetischer Workloads validiert. Diese Tests schließen Lese- und Schreib-I/O-Muster ein, die darauf ausgelegt sind, den Storage-Workload zu simulieren, der von DGX-Systemen generiert wird, die Deep-Learning-Trainingsaufgaben durchführen. Die Storage-Konfiguration wurde mit einem Cluster aus 2-Socket-CPU-Servern validiert, auf denen die FIO-Workloads gleichzeitig ausgeführt wurden, um einen Cluster aus DGX-Systemen zu simulieren. Jeder Client wurde mit derselben oben beschriebenen Netzwerkkonfiguration konfiguriert, wobei folgende Details hinzugefügt wurden.

Für diese Validierung wurden die folgenden Mount-Optionen verwendet:

Vers=4.1	PNFS ermöglicht parallelen Zugriff auf mehrere Storage Nodes
Proto=rdma	Setzt das Übertragungsprotokoll auf RDMA anstelle des Standard-TCP
Port = 20049	Geben Sie den richtigen Port für den RDMA-NFS-Dienst an
max_Connect=16	Ermöglicht NFS Session Trunking zur Aggregation der Storage Port-Bandbreite
Write=Eager	Verbessert die Schreib-Performance von gepufferten Schreibvorgängen
Rsize=262144,wsize=262144	Legt die E/A-Übertragungsgröße auf 256 KB fest

Darüber hinaus wurden die Clients mit einem NFS max_Session_slots-Wert von 1024 konfiguriert. Als die Lösung mit NFS over RDMA getestet wurde, wurden die Storage-Netzwerk-Ports mit einem aktiv/Passiv-Bond konfiguriert. Für diese Validierung wurden die folgenden Bond-Parameter verwendet:

Mode=Active-Backup	Legt die Bindung auf den aktiven/passiven Modus fest
Primär = <interface name>	Die primären Schnittstellen aller Clients wurden über die Switches verteilt
mii-Monitor-interval=100	Gibt das Überwachungsintervall von 100 ms an
Failover-mac-Policy=aktiv	Gibt an, dass die MAC-Adresse der aktiven Verbindung die MAC der Verbindung ist. Dies ist für den ordnungsgemäßen Betrieb von RDMA über die gebundene Schnittstelle erforderlich.

Das Storage-System wurde mit zwei A900 HA-Paaren (4 Controllern) mit zwei NS224-Festplatten-Shelfs mit 24 1,9-TB-NVMe-Festplatten konfiguriert, die an jedes HA-Paar angeschlossen sind. Diese Beschreibung erfolgte unter Verwendung von zwei A900 HA-Paaren. Wie im Abschnitt zur Architektur erwähnt, wurde die Storage-Kapazität aller Controller mit einem FlexGroup Volume kombiniert, wobei die Daten aller Clients über alle Controller im Cluster verteilt wurden.

Leitfaden Zur Größenbemessung Für Storage-Systeme

NetApp hat die DGX BasePOD-Zertifizierung erfolgreich abgeschlossen. Die beiden getesteten A90 HA-Paare können problemlos ein Cluster mit sechzehn DGX H100-Systemen unterstützen. Für größere Implementierungen mit höheren Anforderungen an die Storage-Performance können dem NetApp ONTAP Cluster bis zu 12 HA-Paare (24 Nodes) in einem einzelnen Cluster zusätzliche AFF Systeme hinzugefügt werden. Mithilfe der in dieser Lösung beschriebenen FlexGroup Technologie kann ein 24-Node-Cluster in einem Single Namespace über 79 PB und einen Durchsatz von bis zu 552 Gbit/s bereitstellen. Andere NetApp Storage-Systeme wie die AFF A400, A250 und C800 bieten Optionen für niedrigere Performance und/oder höhere Kapazität für kleinere Implementierungen zu geringeren Kosten. Da ONTAP 9 Cluster mit gemischten Modellen unterstützt, können Kunden mit einem kleineren anfänglichen Platzbedarf beginnen und bei wachsenden Kapazitäts- und Performance-Anforderungen weitere oder größere Storage-Systeme zum Cluster hinzufügen. In der folgenden Tabelle ist eine ungefähre Schätzung der Anzahl der unterstützten A100- und H100-GPUs für jedes AFF-Modell aufgeführt.

Anleitung zur Dimensionierung des NetApp Storage-Systems

		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 mit NVIDIA DGX Systemen – Zusammenfassung und zusätzliche Informationen

Dieser Abschnitt enthält Referenzen zu weiteren Informationen zum NetApp AIPod mit NVIDIA DGX Systemen.

Schlussfolgerung

Die DGX-BasePOD-Architektur ist eine Deep-Learning-Plattform der nächsten Generation, für die gleichermaßen fortschrittliche Storage- und Datenmanagementfunktionen erforderlich sind. Durch die Kombination von DGX BasePOD mit NetApp AFF Systemen kann der NetApp AIPod mit der DGX Systemarchitektur in nahezu jeder Größenordnung implementiert werden. In Verbindung mit der erstklassigen Cloud-Integration und den softwaredefinierten Funktionen von NetApp ONTAP unterstützt AFF eine breite Palette an Daten-Pipelines, die Edge, Core und Cloud einschließen und für den Erfolg von DL-Projekten sorgen.

Weitere Informationen

Weitere Informationen zu den in diesem Dokument beschriebenen Daten finden Sie in den folgenden Dokumenten bzw. auf den folgenden Websites:

- NetApp ONTAP Datenmanagement-Software – ONTAP Informationsbibliothek

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

- NetApp AFF A90 Storage-Systeme –

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

- NetApp ONTAP RDMA-Informationen-

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

- NetApp DataOps Toolkit

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

- NetApp Trident

  "Überblick"

- NetApp GPUDirect Storage-Blog-

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

- NVIDIA DGX BasePOD

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

- NVIDIA DGX H100-SYSTEME

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

- NVIDIA Networking

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

- NVIDIA Magnum IO ™ GPUDirect ® Speicher

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

- NVIDIA Base-Befehl

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

- NVIDIA Base Command Manager

  "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/"

Danksagungen

Dieses Dokument ist die Arbeit der NetApp Solutions und ONTAP Engineering Teams - David Arnette, Olga Kornievskaia, Dustin Fischer, Srikanth Kaligotla, Mohit Kumar und Raghuram Sudhaakar. Zudem möchten sie sich bei NVIDIA und dem NVIDIA DGX BasePOD Engineering-Team für die fortgesetzte Unterstützung bedanken.

Copyright-Informationen

Markeninformationen