NetApp Verified Architecture

NetApp AFF A90 Storage System Reference Design for NVIDIA Cloud Partners using NVIDIA HGX Servers

Design Guide

David Arnette & Stuart Oliver, NetApp
2025 | NVA-1182-DESIGN

In collaboration with

Abstract

The NetApp reference design (RD) for NVIDIA Cloud Partners (NCPs) combines the world-class computing performance of NVIDIA HGX™ Servers with NetApp cloud-connected storage systems to enable data-driven workflows for machine learning (ML), artificial intelligence (AI) and high-performance computing (HPC) in large-scale multi-tenant environments. This document describes the storage system components and configuration for NCP deployments of up to 41k GPUs using HGX H200, B200 and B300 servers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Executive Summary

The collaboration between NetApp and NVIDIA represents a transformative opportunity for hosting service providers to deliver scalable, high-performance AI infrastructure while unlocking new revenue streams. By integrating NetApp ONTAP's advanced data management capabilities with NVIDIA AI-optimized compute and networking platforms, service providers can offer enterprise-grade AI workload hosting, retrieval-augmented generation (RAG), and agentic AI solutions—all while maintaining consistent governance, multitenancy, and security across hybrid and multi-cloud environments.

NetApp and NVIDIA Cloud Partner Program Summary

In today's rapidly evolving digital landscape, service providers face mounting business challenges to deliver differentiated AI services for their customers due to the cost and complexity of modern infrastructure.

These high-level challenges are:

Data Silos Across Hybrid Environments

Fragmented data across on-premises systems, private clouds, and public clouds creates silos that hinder seamless data access and movement—critical for AI workflows. These silos limit the ability to train models efficiently or perform real-time inferencing.

Inefficiencies in AI Workflows

AI workloads typically need access to a performant infrastructure capable of handling massive datasets with low latency. Inefficient GPU utilization, slow data retrieval processes, and suboptimal storage architectures often lead to bottlenecks in training and inferencing workflows.

Data Security Concerns

Service providers must safeguard sensitive customer data across diverse hosting environments while adhering to strict compliance regulations such as ISO27001, SOC 1 Type II & SOC 2 Type II, HIPAA/HITECH, GLBA, PCI DSS v3.2.1, NIST SP 800-53, SOC 2 + HITRUST, ITAR, Privacy Shield and GDPR. This is particularly challenging when deploying AI applications that rely on large-scale data aggregation.

Multitenancy Challenges

Multitenancy is essential for service providers to maximize infrastructure utilization by serving multiple customers across shared resources. However, ensuring isolation between tenants while maintaining predictable performance can be difficult without advanced QoS mechanisms.

Quality of Service (QoS) Requirements

QoS is critical for managing diverse workloads on shared infrastructure. Service providers must prioritize latency-sensitive tasks while dynamically allocating resources to less time-critical AI workloads during off-peak hours. Without effective QoS policies, noisy neighbor issues can degrade performance across tenants.

NetApp provides service providers with a suite of solutions designed to optimize every stage of the AI lifecycle while addressing multitenancy and QoS requirements. NetApp ONTAP's certification for NCPs for NVIDIA HGX systems enables hosting providers to monetize value-added services tailored to enterprise AI workloads.

These services include:

Disaster Recovery as a Service (DRaaS)

With SnapMirror® technology, hosting providers can offer DRaaS by replicating customer datasets across hybrid cloud environments in real time. This ensures business continuity for enterprises while creating subscription-based revenue opportunities for service providers.

AI-Powered Analytics

NetApp ONTAP integrates seamlessly with NVIDIA AI Enterprise software to support analytics-as-a-service offerings. Hosting providers can leverage this capability to deliver predictive analytics, anomaly detection, or business intelligence solutions—targeting industries like finance, healthcare, and retail.

Retrieval-Augmented Generation (RAG)

NetApp's integration with NVIDIA NeMo™ Retriever microservices enables hosting providers to offer RAG solutions that combine generative AI models with secure data retrieval from customer datasets stored on ONTAP systems. These services can be monetized as premium packages for enterprises seeking intelligent copilots or enhanced productivity tools.

Agentic AI Hosting

By combining ONTAP's robust data management capabilities with NVIDIA Blackwell GPUs and the NVIDIA Dynamo inference library, hosting providers can deliver agentic AI solutions that support reasoning model inference workloads. These advanced services position providers as leaders in next-generation AI infrastructure.

Hybrid Cloud Data Management

NetApp ONTAP allows service providers to extend customer on-premises environments into public clouds seamlessly. This capability supports tiered storage, automated data pipelines, and compliance management—enabling upselling opportunities for hybrid cloud solutions tailored to enterprise needs.

Enterprise AI Workload Hosting

Hosting enterprise AI workloads requires infrastructure that integrates seamlessly with existing data environments while supporting the demanding requirements of modern AI applications. NetApp's collaboration with NVIDIA enables service providers to deliver scalable, secure, and high-performance enterprise AI workload hosting through the following capabilities:

SnapMirror for Seamless Data Integration

SnapMirror® technology allows hosting providers to replicate customer datasets from on-premises systems to cloud environments in real time. This ensures continuity between legacy systems and AI-specific workloads without disrupting existing workflows or creating silos.

Unified Data Management Across Hybrid Clouds

NetApp ONTAP provides a centralized platform for managing data across hybrid and multi-cloud environments. Hosting providers can leverage this functionality to deliver tailored solutions that extend enterprise infrastructure into the cloud while maintaining control and security over sensitive data.

Optimized GPU Utilization

Enterprise AI workloads often require high-performance computing resources for NVIDIA HGX systems for training large models or performing real time inferencing tasks. NetApp integrates with NVIDIA HGX systems to optimize GPU utilization through Magnum IO™ GPUDirect® Storage—reducing latency and accelerating processing times for resource-intensive operations.

Consistent Governance Across Environments

NetApp ONTAP ensures consistent governance by providing encryption, access controls, audit trails, and metadata cataloging across hybrid clouds. These features are critical for enterprise clients operating in regulated industries like healthcare or finance, where compliance is paramount.

Scalable Multitenancy Support

NetApp ONTAP supports secure multitenancy by isolating workloads across multiple customers while maximizing infrastructure utilization. Hosting providers can host diverse enterprise AI workloads on shared resources without compromising performance or security—a key differentiator in competitive markets.

NetApp AFF A90 with NVIDIA HGX Servers

NetApp has been supporting NVIDIA HGX systems with validated reference designs for every generation since the first purpose-built DL servers were released in 2016. ONTAP systems have a proven history of performance, reliability, and data management for data-driven AI environments. Using AFF A90 systems with NVIDIA HGX H200, B200, and B300 systems enables service providers to deliver high-performance infrastructure with integrated data management for the following use cases:

- ML at massive scale using traditional analytics tools.
- AI model training for Large Language Models (LLMs), computer vision/image classification, fraud detection, and countless other use cases.
- HPC such as seismic analysis, computational fluid dynamics, and large-scale visualization.

NetApp AFF A90 Storage Systems with NetApp ONTAP

NetApp AFF A90 hardware

The NetApp AFF A90 storage system is a 4RU chassis containing two controllers that operate as high-availability partners (HA Pair) for each other, with up to 48 internal NVMe solid state disks (SSDs). SSDs are available in sizes up to 15.3 TB, and up to six expansion shelves can be added for additional capacity. Each controller has two Intel CPUs and 1024 GB RAM, as well as a 64 GB NVRAM module. There are eight available hot-swappable PCIe Gen5 I/O slots available on each controller that can support up to 400 GbE.

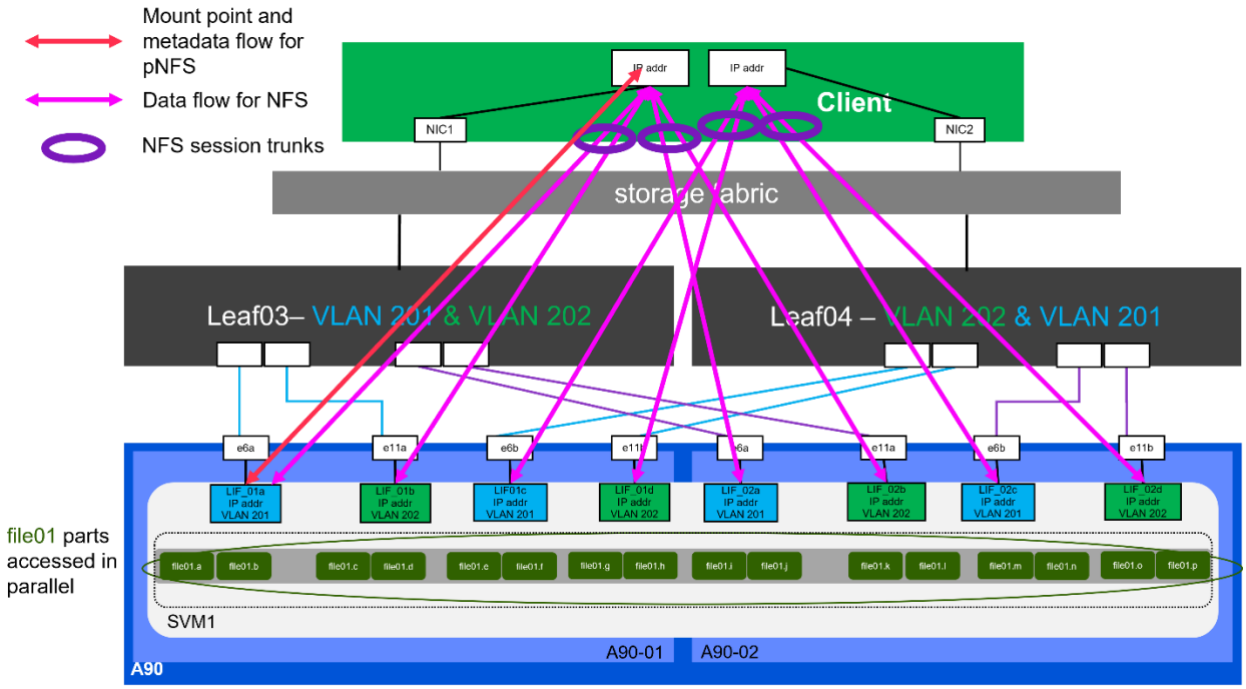
Figure 1 - NetApp AFF A90 storage system

To enable high performance and scalability, the storage controllers form a storage cluster that enables the performance and capacity of up to 40 cluster nodes to be combined into a single namespace called a

FlexGroup, with data distributed across the disks of every node in the cluster. Configurations with more than 40 controllers require multiple clusters and namespaces. The storage cluster supports NFS v4.1 with Parallel NFS (pNFS) that enables clients to establish connections directly to every controller in the cluster. Additionally, session trunking combines the performance from multiple physical interfaces into a single session, enabling even single-threaded workloads to access more network bandwidth than is possible with traditional ethernet bonding. Combining all of these features with RDMA enables the AFF A90 storage system to deliver low-latency and high throughput that scales linearly for workloads leveraging NVIDIA Magnum IO GPUDirect Storage.

With the new Granular Data Distribution feature released in ONTAP 9.16.1, individual files are segmented and distributed across the FlexGroup to enable the highest levels of performance for single-file workloads. Figure 2 below shows how pNFS and NFS session trunking work together with FlexGroups and GDD to enable parallel access to large files leveraging every network interface and disk in the storage system.

Figure 2 - pNFS, session trunking, FlexGroups, and GDD



NetApp ONTAP data management software

The NetApp AFF A90, powered by NetApp ONTAP data management software, provides built-in data protection, anti-ransomware capabilities, and the high-performance, scalability and resiliency required to support the most critical business workloads. It eliminates disruptions to mission-critical operations, minimizes performance tuning, and safeguards your data from ransomware attacks. NetApp AFF A90 systems with ONTAP deliver-

- **Intelligence—**Accelerate digital transformation with an AI-ready ecosystem built on data-driven intelligence, future-proof infrastructure, and deep integrations with NVIDIA and the MLOps ecosystem. Using ONTAP's snapshot and FlexClone capabilities, teams can instantly create space-efficient copies of datasets for parallel development and testing. FlexCache and Snapmirror replication technologies enable streamlined, space-efficient, and automated data pipelines from data sources across the enterprise. And multi-protocol access to data using NAS and object protocols enables new workflows optimized for ingest and data engineering tasks. Data and training checkpoints can be tiered to lower-cost storage to avoid filling primary storage.

Customers can seamlessly manage, protect, and mobilize data, at the lowest cost, across hybrid clouds with a single storage OS and the industry's richest data services suite.

- **Security—**NetApp ONTAP storage delivers enterprise-grade security through multiple layers of protection. At the infrastructure level, the solution implements robust access control mechanisms, including role-based access control (RBAC), multi-factor authentication, and detailed audit logging capabilities. The platform's comprehensive encryption framework protects data both at rest and in transit, utilizing industry-standard protocols and algorithms to safeguard intellectual property and maintain compliance with regulatory requirements. Integrated security monitoring tools provide real-time visibility into potential security threats, while automated response mechanisms help mitigate risks before they can impact operations. NetApp ONTAP is the only hardened enterprise storage that's validated to store top-secret data.
- **Multitenancy—**NetApp ONTAP delivers the widest range of features to enable secure multi-tenant usage of storage resources. Storage Virtual Machines provide tenant-based administrative delegation with RBAC controls. Comprehensive QoS controls guarantee performance for critical workloads while enabling maximum utilization, and security features such as tenant-managed keys for volume-level encryption guarantee data security on shared storage media.
- **Reliability—**NetApp eliminates disruptions to mission-critical operations through advanced reliability, availability, serviceability, and security capabilities, delivering the highest uptime available. For more information, see the ONTAP RASS whitepaper. In addition, system health can be optimized with AI-based predictive analytics delivered by Active IQ and Data Infrastructure Insights.

NetApp Architecture for NCP Deployments

The storage performance target for training or inference can vary depending on the type of model and dataset. The guidelines in Table 1 provide standard throughput for the various GPU system sizes and HPS sizing. The final HPS requirements for throughput and capacity will be specified for each NCP opportunity. The configurations described in the following sections meet the write requirements and significantly exceed the read requirements shown below.

Table 1- Guidance for Standard HPS aggregate storage performance

Description	Number of GPUs						
	1,024	2,304	4,096	8,192	16,384	29,952	41,472
Read throughput (GB/s)	160	360	640	1,280	2,560	4,680	6,480
Write throughput (GB/s)	80	180	320	640	1,280	2,340	3,240
Storage Configuration							
Number of appliances 1*	5	11	19	38	75	137	189
Number of appliances 2 (if required)							
Number of namespaces	1	1	1	2	4	7	10
Number of 400G storage ports**	20	44	76	152	300	548	756
Number of in-band management connections	10	22	38	76	150	274	378
Number of out-of-band management connections	12	24	40	80	158	288	398

Number of rack units	22	46	78	156	308	562	776
Power (KW)	9.125	20.07	34.675	69.35	136.875	250.025	344.925
Cooling (BTU/hr)	3.375	7.425	12.825	25.65	50.625	92.475	127.575

* Each AFF A90 appliance includes 2x controllers.

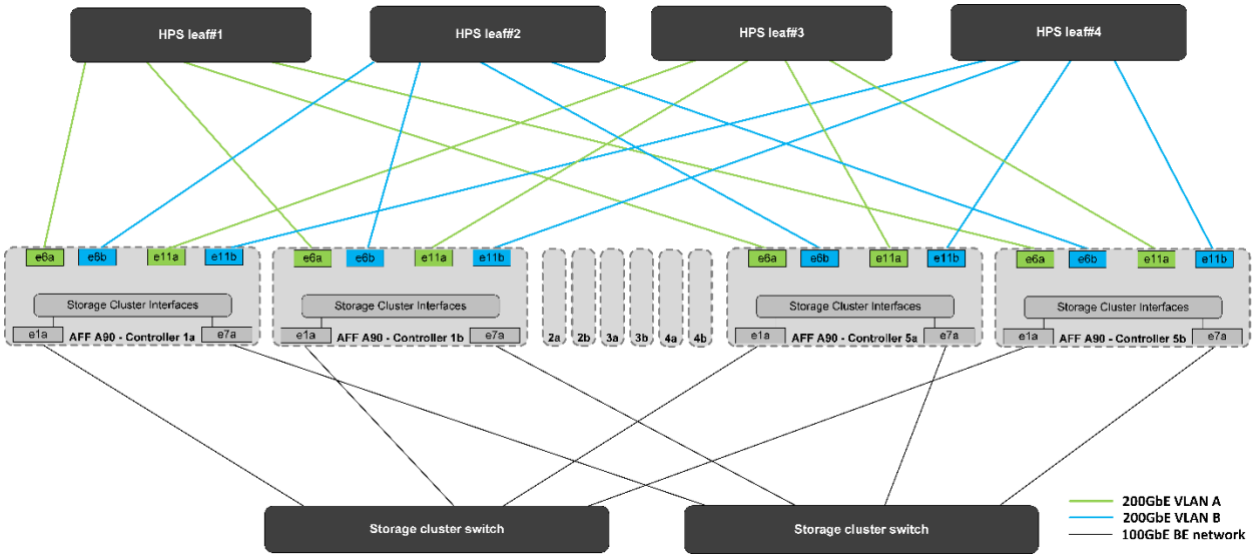
**400G switch ports are split into 2x200G ports using copper or optical splitter cables.

AFF A90 storage system connectivity

NetApp AFF A90 systems integrate seamlessly into the NVIDIA NCP Reference Design and connect to the NVIDIA Spectrum SN5600 leaf switches provided for HPS. Each AFF A90 controller has two NVIDIA ConnectX®-7 dual-port adapters for connecting to the HPS fabric, for a total of four 200 Gb connections per controller. For connectivity to the AFF A90 storage system, each 800 Gb port on the SN5600 switch is broken into four 200 Gb ports using the appropriate copper or optical breakout cables. For cable lengths over 5m, NetApp recommends MMA1Z00-NS400 QSFP112 flat-top SR4 transceivers in the ConnectX-7 cards in the AFF A90, connecting to MMA4Z00-NS 2x400 Gb/s twin-port OSFP multimode transceivers in the SN5600 using two split MPO-12/APC to MPO-12/APC cables. For distances shorter than 5m, NVIDIA offers a number of passive and active copper cables.

The AFF A90 uses a 100 Gb network for back-end cluster connectivity, with two ports per controller connected to redundant back-end network switches. The AFF A90 solution for NCPs uses NFS v4.1 pNFS to enable data transfer directly between every storage node and the clients. This eliminates data transfer across the back-end network, so the storage cluster network is only used for cluster synchronization and NVRAM mirroring between the controllers in each HA pair for high availability. Figure 3 below shows the specific ports used for front and back-end network connectivity in a system sized to support 1,024 GPUs.

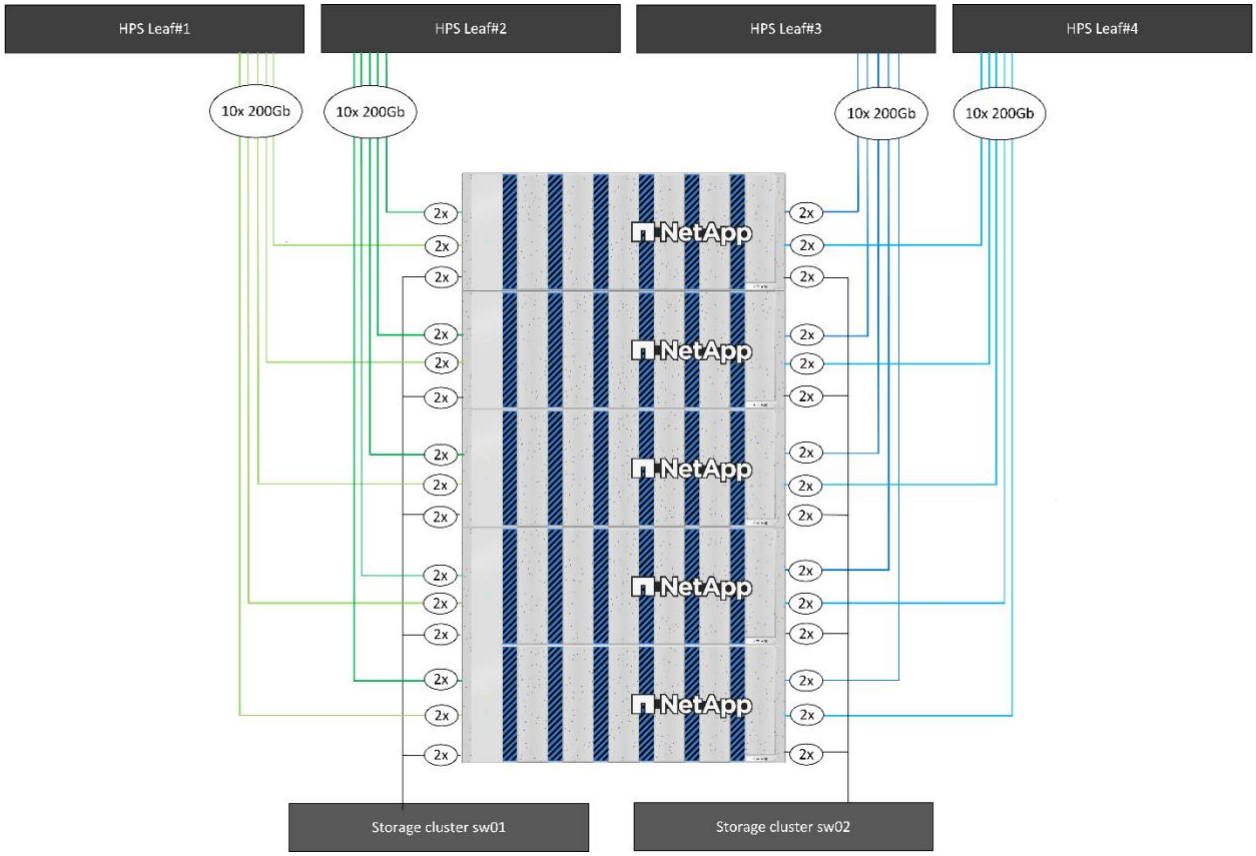
Figure 3 - Storage cluster connectivity



NCP Deployments with 1024 GPUs

Deployments with 1,024 GPUs require a total of five AFF A90 storage systems to meet the standard performance requirements. Figure 3 above shows the storage system network connectivity for this configuration, and Figure 4 below shows the HPS network architecture for a system of this size.

Figure 4 - NCP Deployments with 1,024 GPUs

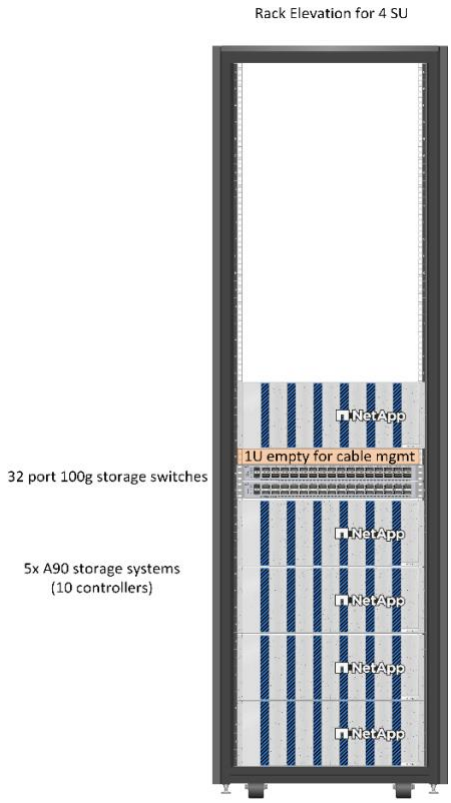


Configuration notes-

- The storage cluster is connected to four leaf switches for maximum redundancy and load distribution.
- Storage system HPS ports are 200 Gb ethernet. 400 Gb switch ports are split into two 200 Gb ports using optical or copper splitter cables.
- This configuration consumes 20 total 400 Gb switch ports, five on each HPS leaf switch.

Figure 5 below shows the rack elevation for a deployment with 1,024 GPUs.

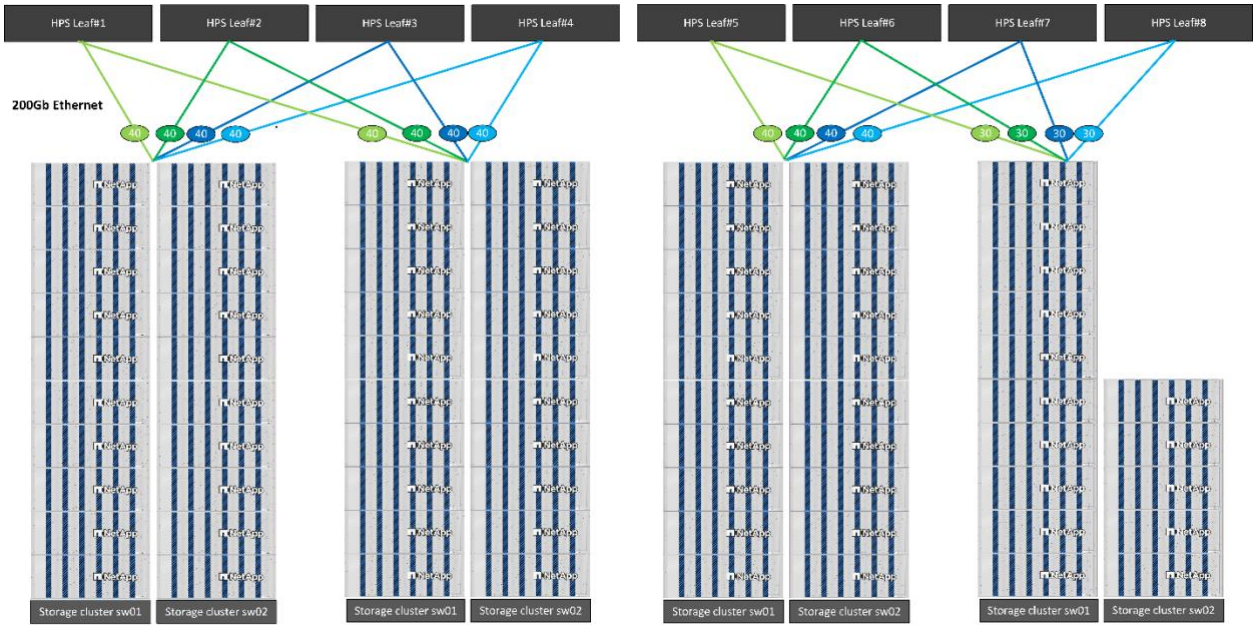
Figure 5 - Rack Elevation for 1,024 GPUs



NCP Deployments with 16,384 GPUs

Deployments with 16,384 GPUs require a total of 75 AFF A90 storage systems to meet the standard performance requirements. Figure 3 above shows the storage system network connectivity used for all deployments, and Figure 6 below shows the HPS network architecture for deployments using 16k GPUs.

Figure 6 - NCP Deployments with 16,384 GPUs



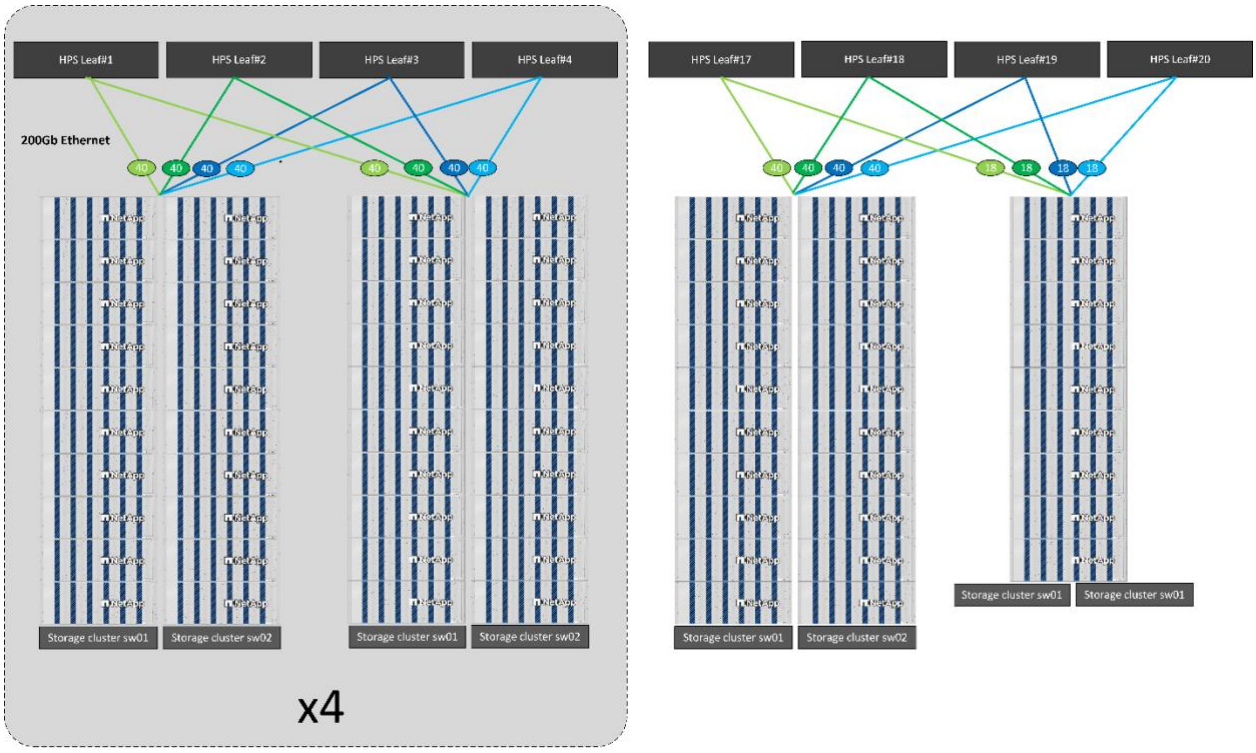
Configuration notes-

- Storage systems are configured into clusters of 20 AFF A90 systems with 40 controllers.
- Each storage cluster is connected to four leaf switches for maximum redundancy and load distribution.
- Storage system HPS ports are 200 Gb ethernet. 400 Gb switch ports are split into two 200 Gb ports using optical or copper splitter cables.
- This configuration consumes 300 total 400 Gb switch ports. Typical clusters with 40 nodes consume 40 400 Gb switch ports on each leaf the cluster is connected to.

NCP Deployments with 41,472 GPUs

Deployments with 41,472 GPUs require a total of 189 AFF A90 storage systems to meet the standard performance requirements. Figure 3 above shows the storage system network connectivity used for all deployments, and Figure 7 below shows the HPS network architecture for deployments using 41k GPUs.

Figure 7 - NCP Deployments with 41,472 GPUs



Configuration notes-

- Storage systems are configured into clusters of 20 AFF A90 systems with 40 controllers.
- Each storage cluster is connected to four leaf switches for maximum redundancy and load distribution.
- Storage system HPS ports are 200 Gb ethernet. 400 Gb switch ports are split into two 200 Gb ports using optical or copper splitter cables.
- This configuration consumes 756 total 400 Gb switch ports. Typical clusters with 40 nodes consume 40 400 Gb switch ports on each leaf the cluster is connected to.

Storage Partner Performance Validation

This storage solution was validated in multiple stages by NetApp and NVIDIA to ensure that performance and scalability meet the requirements for NCPs. The configuration was validated using a combination of synthetic workloads and real-world ML/DL workloads to verify both maximum performance and application interoperability. Functional tests were performed to demonstrate QoS isolation of workloads and the resilience and performance of the system under various failure conditions.

Summary

Service providers face mounting pressure to deliver high-performance AI-powered services while navigating the complexities of hybrid cloud environments. NetApp and NVIDIA offer a comprehensive suite of certified solutions designed to help service providers address these challenges head-on by:

- Leveraging NVIDIA NCP blueprints, SnapMirror®, FlexCache®, QoS mechanisms, and multitenancy features for unified data management across hybrid setups.
- Adopting NetApp and NVIDIA NCP certified solutions to accelerate deployment timelines and reduce time to revenue.
- Partnering with NetApp to align infrastructure strategies with emerging trends in responsible AI practices and cloud integrations.

By integrating NetApp's solutions into their operations, service providers can unlock new revenue streams through advanced services such as generative AI applications, predictive analytics, and edge computing—all while reducing costs and improving operational efficiency.

The future belongs to those who embrace innovation today—and the integrated NetApp and NVIDIA NCP partnership enables service providers to build thriving businesses in the age of artificial intelligence.

Appendix

NetApp ONTAP Solution for Data Lake

NetApp AFF A90 systems with ONTAP can also be deployed for use as a data lake and offers concurrent access to data with NFS and S3 (and SMB). For connectivity to the in-band network in the NCP RD, each A90 controller can be provisioned with additional NVIDIA dual-port 100 and 200 Gb cards, enabling connectivity to the SN4700 leaf switches in the in-band network fabric. Note that the sizing information above is based on HPS usage only, and additional controllers and/or clusters should be included to support data lake use cases. NetApp's industry-leading data management capabilities enable automated and seamless movement of data within and across storage clusters.

Where to find additional information

To learn more about the information that is described in this document, review the following documents and/or websites:

- NVA-11XX NetApp AFF A90 Storage System Reference Design for NVIDIA Cloud Partners using NVIDIA GB200, and GB300 NVL72 systems
- NVA-1175 NVIDIA DGX SuperPOD™ with NetApp AFF A90 Storage Systems Design Guide
- NVA-1175 NVIDIA DGX SuperPOD with NetApp AFF A90 Storage Systems Deployment Guide
- NVIDIA DGX™ B200 SuperPOD Reference Architecture
- NVIDIA DGX H200 SuperPOD Reference Architecture

- NVIDIA Spectrum SN5600 Ethernet switches
- NetApp Documentation
- NetApp AI Solutions Documentation
- NetApp ONTAP software
- NetApp Install and Maintain AFF Storage Systems
- NFS over RDMA
- What is pNFS (older doc with great pNFS info)

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.