REFERENCE ARCHITECTURE

High-performance Storage for NVIDIA Cloud Partners with Pure Storage

Validated reference design guide for NVIDIA Cloud Partners, featuring Pure Storage® FlashBlade//S500 and NVIDIA HGX™ H200, B200, and B300 servers

Contents

Executive overview

The Pure Storage FlashBlade//S500 is a certified, high-performance, unified storage platform tailored for NVIDIA Cloud Partners (NCPs) building infrastructure for AI workloads on NVIDIA HGX servers.

FlashBlade® delivers the performance, operational efficiency, scalability, and simplicity required to support the rapid evolution of GPU-accelerated workloads across training, inference, and data preprocessing pipelines. By integrating seamlessly into the NCP ecosystem, FlashBlade//S™ enables NCPs to build scalable AI-as-a-service environments with consistent performance and secure multi-tenancy.

This document presents a validated reference design guide for NCPs that combines Pure Storage FlashBlade//S500 and Pure Fusion™, a federated FlashBlade//S500 fleet management solution, with NVIDIA HGX, H200, B200, and B300 platforms. The solution is optimized to deliver predictable, low-latency data access across thousands of GPUs. It supports the high-throughput demands of modern AI pipelines and adheres to the NCP HGX storage design guide specifications for high-performance storage (HPS).

Designed for unstructured data at scale, FlashBlade//S500 helps NCPs accelerate time-to-value while reducing infrastructure complexity and power consumption. The rest of this document details how FlashBlade//S500 meets the NCP HPS requirements, with a focus on deployment, scalability, and multi-tenant support.

FlashBlade//S simplifies and accelerates GPU-provider AI services for NCPs

The FlashBlade//S500 offers a unique set of key features and benefits as an HPS solution tailored for NCPs. These include:

- **Ease of use**: For NCPs scaling multi-tenant AI services, the simplified, scale-out architecture of FlashBlade reduces operational overhead and accelerates deployment timelines.
- **Distributed everything**: Metadata, data, and control functions are distributed across the FlashBlade//S500 solution. This massively **parallel architecture** avoids bottlenecks and allows performance and capacity to scale with the addition of more blades.
- **Erasure coding**: Each FlashBlade//S500 supports wide-stripe erasure coding with N+2 and N+4 redundancy schemes. This **enhances resiliency** by fragmenting data and distributing parity information across the FlashBlade//S500 array, ensuring that in the event of failures, the original data remains accessible and can be reconstructed without significant performance impact.
- **Scalability**: Designed for managing extensive unstructured data, a single FlashBlade//S500 solution provides highly efficient scalability. **Hot-swappable** components facilitate vertical capacity increases, while horizontal scaling allows for seamless expansion. This ensures **continuous availability and optimal performance** for NCP configurations and their associated workloads during growth.
- **Reliability**: The FlashBlade architecture is designed to provide reliable unstructured data storage. The disaggregated architecture of a FlashBlade//S500 solution allows **compute and storage to scale independently**, optimizing infrastructure spend and minimizing footprint in dense GPU clusters. The FlashBlade//S500 is engineered for both exceptional reliability and operational simplicity. Its fully N+2 and N+4 redundant, nondisruptive architecture delivers six nines (**99.9999%**) of availability.

- **Efficiency, predictability, and sustainability**: The FlashBlade//S500 architecture is designed to maximize energy efficiency and density, operating at just 1.3 watts per terabyte. This contributes to a reduced data center footprint and lower power consumption, aligning with sustainability goals and supporting future scalability with increasingly capable components.
- **Scalable performance and availability**: The scale-out nature of a single FlashBlade//S500 for NCP solutions allows for **nondisruptive, granular capacity and performance** expansion. This ensures predictable performance scaling concurrent with growing AI workload demands and provides sustained high throughput and availability for long-term infrastructure planning.
- **High performance**: The FlashBlade//S500 for NCP delivers exceptional performance through a massively distributed transactional database that handles metadata control, large file chunks, and protocol processing with precision. It's optimized for both small and large files, using high-performance metadata handling and variable-block encoding for efficient random access.
- **Speed and scalability across workloads**: All data is intelligently distributed across the system, ensuring speed and scalability for multiple workloads. DirectFlash® technology from Pure Storage provides direct access to flash media and exposes the full concurrency of flash devices, enabling the architecture to maximize throughput and minimize latency at every layer.

About the FlashBlade//S500 hardware

FlashBlade//S500 is a scale-out storage appliance engineered for high-throughput, low-latency data services across AI workloads. All components in the FlashBlade//S500 solution work in combination to handle all data I/O to and from the FlashBlade//S500 array.

For this NCP design guide, naming conventions describing the FlashBlade//S500 architecture used in this document are as follows:

- **DirectFlash Module**: A physical all-flash storage device created by Pure Storage. In the case of NCPs, the DirectFlash Module is 37.5TB in size.
- **Blade**: A single physical piece of hardware that is used to house DirectFlash Modules.
- **Chassis**: A single physical piece of hardware used to contain multiple blades.
- **External fabric module (XFM)**: A switch *pair* used to connect multiple chassis to the NCP converged network fabric and aggregate all data traffic to blades.

About the FlashBlade//S500 operating system

Purity for FlashBlade (Purity//FB) is the operating environment that manages the FlashBlade hardware, networking, and storage components. The Purity//FB software is part of the FlashBlade solution. As a single install package, it manages all hardware and software components, enabling the FlashBlade//S500 to function as a highly scalable storage solution for NCP environments.

Key Pure Storage service provider features for NCP

Multi-tenancy

When delivering data for AI as a service using the NCP design guide, a Pure Storage FlashBlade//S500 storage unit provides secure separation of data and management for multiple tenants. Customers can leverage secure, logical separation within a single FlashBlade//S500 storage unit, eliminating the need for costly investments in multiple physical systems. This multi-tenancy capability lowers the operational and administrative costs associated with managing multiple FlashBlade//S500 storage units for multiple tenants and reduces the rack space required to house multiple storage units.

In addition to segregated data access for each tenant, those tenants can also self-administer their allocated resources with role-based access controls. All data is encrypted at rest, and each tenant is unable to access the resources of any other tenant.

The system is simple to scale and manage while providing a high degree of flexibility and security. Creation of data services for tenants takes minutes to configure, with resources available in seconds.

Multi-tenancy network setup

Pure Storage introduced two key concepts to its Purity operating environment for multi-tenancy:

- **Realms** for delegated administrative control
- **Servers** for isolated data access

Realms and servers can be used individually or in combination to achieve true secure, multi-tenant storage with the most agility and optionality for Purity users.

FlashBlade//S500 simplifies multi-tenancy networking using servers for isolated data access. Each server requires a data ingress IP and provides an isolated network space for data access management. Servers encapsulate network and identity information, allowing for unique network interfaces, file system exports, and authentication accounts (Active Directory/LDAP/NIS) independent of the parent FlashBlade//S500 authentication configuration.

Figure 1 illustrates the aforementioned multi-tenancy concepts of realms and servers. Server 1 is within Realm 1, and therefore a Realm 1 administrator only sees file systems within their realm. Server 1 and Server 2 are on two different subnets and virtual local-area networks (VLANs). File systems on Server 1 are exposed to clients via Pure Storage file system exports. A client accessing data via Server 2 VLAN connections cannot view and does not have access to the file systems on Server 1 due to network isolation and discrete identity domains. Similarly, a client accessing data via the Server 1 connection cannot view and does not have access to the file systems on Server 2.

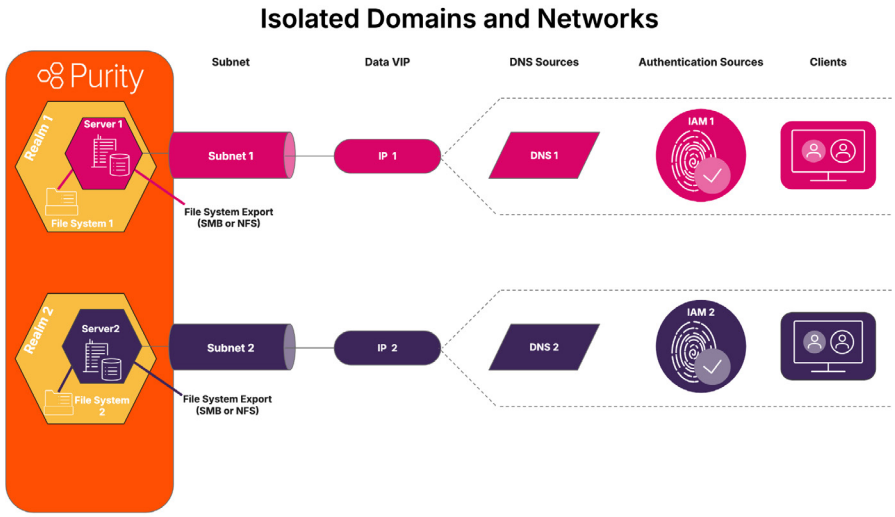

FIGURE 1 Multi-tenancy isolated domains and client networks

Quality of service

Pure Storage FlashBlade//S500 supports a quality of service (QoS) ceiling. With multi-tenancy, each tenant can leverage the QoS ceiling to effectively limit IOPS and/or bandwidth consumption for file systems that they have access to.

FlashBlade//S500 utilizes a policy-based approach for QoS, and a single QoS policy can have a one-to-many relationship. This means that for each tenant, a policy can be attached to a file system or multiple file systems that have identical performance requirements. This streamlined approach enhances scalability and simplifies management for customers with diverse performance needs.

Security

Pure Storage FlashBlade//S500 has clear control and data plane segregation that enhances security and enforces role-based access. Data plane operations are restricted to authenticated users through designated data protocols.

Administrative tasks within the control plane, such as file system/object store lifecycle management, network configuration, and access control policies, are exclusively accessible via administrative interfaces (CLI, GUI, REST API). This separation prevents unauthorized data manipulation by administrators and restricts system-level changes by data users.

Pure Fusion

Pure Fusion is a federated FlashBlade//S500 fleet management solution. It introduces the concept of the “Fleet View” to the FlashBlade//S500 array user interface, allowing an administrator to search for, modify, or create remote resources without having to manage multiple user interfaces or endpoints. When Fleet View is enabled, all storage resources from all FlashBlade//S500 systems that are joined to a fleet can be viewed in one place.

The standard filtering and sorting features in the user interface can then be used to discover remote FlashBlade//S500 systems, providing the ability to edit a specific FlashBlade//S500 system configuration as though you were locally logged into the FlashBlade//S500 management interface.

Each FlashBlade//S500 system in a fleet has its own identity and retains its local workflows and interfaces, but an administrator can manage any FlashBlade//S500 system from any other FlashBlade//S500 within the fleet. With Pure Fusion, you can create, modify, and join a fleet and provision storage on any FlashBlade//S500 in the fleet without the need to switch endpoints.

Pure Fusion is fully integrated into the Purity operating system, available free of charge with a nondisruptive Purity upgrade to Purity//FB 4.5.6 or later for FlashBlade//S500 systems. Review the Additional Information section for Pure Fusion technical requirements.

Pure Storage architecture for NCP deployments

NCP storage targets: scaling FlashBlade//S500

The storage performance target for training or inference can vary depending on the type of model and data set. 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.

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

TABLE 1 NVIDIA HGX standard performance guidelines for NCPs

FlashBlade//S500 in NCP environments

Pure Storage FlashBlade//S500 solutions for NCP deployments are based on a core building block of a **single 10-chassis FlashBlade//S500 per 1,000–1,100 GPUs**. This base building block ensures that performance consistency is maintained as the NCP cluster size grows. The recommended scaling model involves **adding a FlashBlade//S500 incrementally per 1,000–1,100 GPUs**.

Each FlashBlade//S500 for NCP is fully capable of performing in both single-tenant and multi-tenant configurations at 1,000–1,100 GPUs, maintaining high performance for 1,000–1,100 GPU read and write benchmarks in both scenarios.

With clientless native Network File System (NFS) over Remote Direct Memory Access (RDMA) support, FlashBlade//S500 provides simultaneous data access to thousands of data connections with minimal administrative overhead.

All data on a FlashBlade//S500 is accessible via a **single data network IP**.

FlashBlade//S500 array to NVIDIA NCP HPS network connectivity

The FlashBlade//S500 connects to the NCP converged network using 16 400GbE ports. Each XFM, shown in Figure 2, includes eight of these ports. Since XFMs are deployed in pairs, the total number of ports doubles to 16.

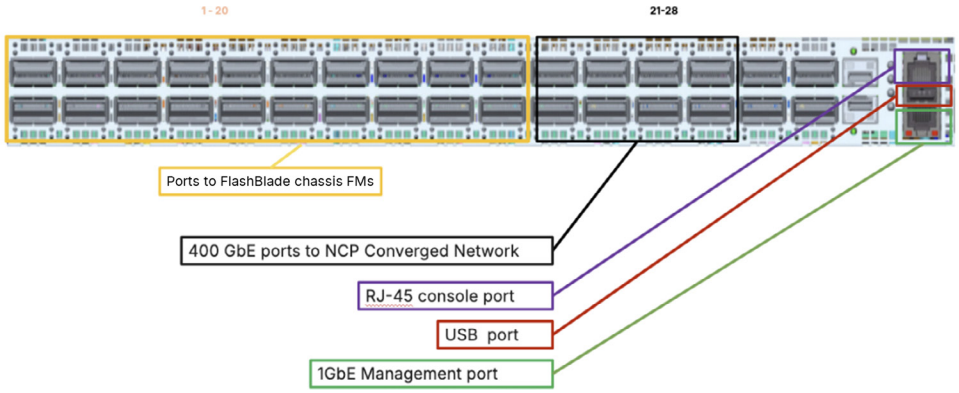

FIGURE 2 XFM port map

Each FlashBlade//S500 is connected to disparate pairs of NVIDIA Spectrum™ SN5600/5610 Ethernet switches as needed, based on workload distribution.

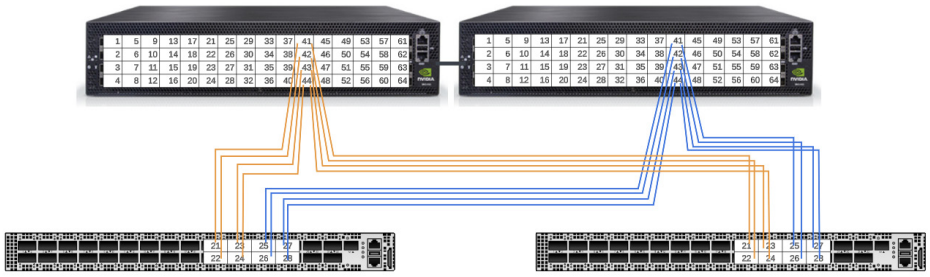

FIGURE 3 Aggregate connection uplinks from XFMs to the NCP HPS converged network NVIDIA Spectrum SN5600/5610 Ethernet switches

FlashBlade//S500 XFMs require an active-active port channel to be configured on the NVIDIA Spectrum SN5600/5610 Ethernet switches to ensure proper distribution of data traffic. Once connected to an NVIDIA NCP converged network via a pair of NVIDIA Spectrum SN5600/5610 Ethernet switches, changes to the number of chassis or blades within the FlashBlade//S500 do not affect the network connectivity and remain transparent to the NVIDIA converged fabric.

Aggregated data traffic connections to the FlashBlade//S500 are passed from the XFMs to the chassis and then to the blades within each chassis, **fully distributing all connections across all blades in the system**. All blades within the FlashBlade//S500 have access to all storage across the FlashBlade//S500 and can respond to any client request (see Figure 4).

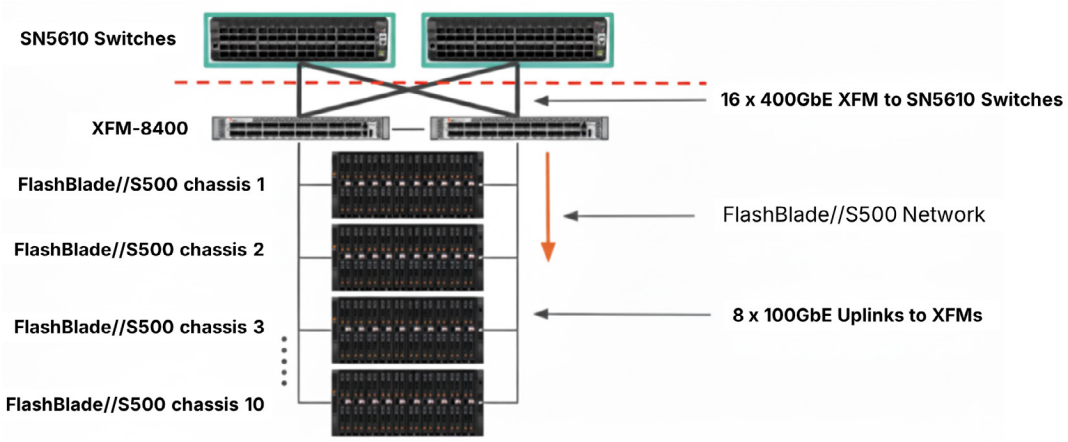

FIGURE 4 FlashBlade//S500 and NVIDIA Spectrum SN5600/5610 HPS network connectivity

Cabling information

Connections from any single FlashBlade//S500 to an NCP converged network use the same number of data and out-of-band (OOB) connections.

Each FlashBlade//S500 for NCP utilizes the number of data and OOB network connections shown in Table 2.

Network links per FlashBlade//S500 array	Count
OOB management ports per FlashBlade//S500 array	2
400GbE data ports per FlashBlade//S500 array	16

TABLE 2 FlashBlade//S500 cable counts for external communication

The following section provides guidance for 1,000–1,100 GPU building blocks for NCP with a FlashBlade//S500 (for FlashBlade//S500 sizing, see Table 3).

Number of GPUs	Number of FlashBlade//S500 arrays	XFMs	NVIDIA NCP converged network uplinks
1,000–1,100	1	1 pair	16

TABLE 3 FlashBlade//S500 sizing table

NCP FlashBlade//S500 deployments for 1,000–1,100 GPUs

Figure 5 and Table 4 provide deployment guidance for scaling FlashBlade//S500 to support 1,000–1,100 GPU building blocks in an NCP environment.

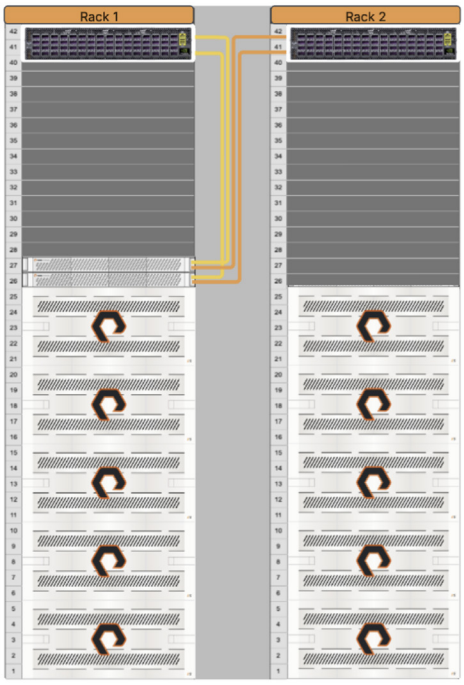

FIGURE 5 One FlashBlade//S500

FlashBlade//S500 component list for 1,000–1,100 GPUs	
Item	**Quantity**
XFM	1 pair (total of 2)
Chassis	10
Accessory Kit list for 4SU	
Item	**Quantity**
FB-XFM8400-MC-Accessory-Kit	10
FB-MPO24-Fiber-10M	10
FB-Higher-Performance-400G-DR4-Accessory-Kit	2
OOB connectivity	
Item	**Quantity**
OOB RJ45 Cat5/6 1Gb/E	2

TABLE 4 Bill of materials for two FlashBlade//S500 arrays

Pure Storage solution performance validation

This reference design guide was validated by NVIDIA using a Pure Storage FlashBlade//S500. To assess the overall cluster performance typical of AI workloads, an NVIDIA benchmarking test suite was used. The tests specifically evaluated the write demands of large-scale checkpointing, alongside the read requirements for token loading and checkpoint restoration.

In addition to this testing, Pure Storage has a long history of providing storage solutions for AI and high-performance computing workloads in customer environments, with highly variable workload types that require consistent linear performance and proven 99.9999% uptime.

Summary

This document presents a validated reference design guide for the Pure Storage platform, adhering to the NCP design guide specifications for NVIDIA HGX platforms. NCPs can successfully pair with the Pure Storage platform to deliver the high performance, scalability, and operational simplicity needed for AI infrastructure deployments.

The Pure Storage platform delivers a unified, multidimensional solution built on 15 years of relentless software innovation and flash technology. It empowers organizations to seamlessly execute every stage of the AI pipeline, from data curation and model training to serving and inference, with autonomously tuned HPS—all with Pure Storage efficiency and simplicity in a single, powerful platform. Pure Storage supports many AI customers across diverse stages of their innovation journeys, including some of the largest AI environments in existence.

Additional information

- FlashBlade//S
- Pure Storage platform for AI
- FlashBlade//S technical specifications
- Automating storage with Pure Fusion

Pure Fusion technical requirements

Table 5 shows the Pure Fusion technical requirements for networking, authentication, API access, and hardware support.

Pure Fusion technical requirements	
Description	**Requirement**
Networking	All arrays must be able to reach each other. Pure Fusion uses the lowest numbered virtual management port for array-to-array communication.
Authentication	Active Directory (AD)/LDAP is required. To join arrays to a fleet, the arrays must be in the same AD/LDAP domain. GUI/CLI users must be authenticated by AD/LDAP for remote operations. Local users cannot perform remote operations.
API access	API access requires an API token that is mapped to an AD/LDAP account. The API token must exist on the array initiating the operation.
Hardware support	A maximum of 32 Pure Storage FlashBlade//S500 systems are supported in the same fleet.

TABLE 5 Pure Fusion technical requirements

ID0784-01-en-11/25