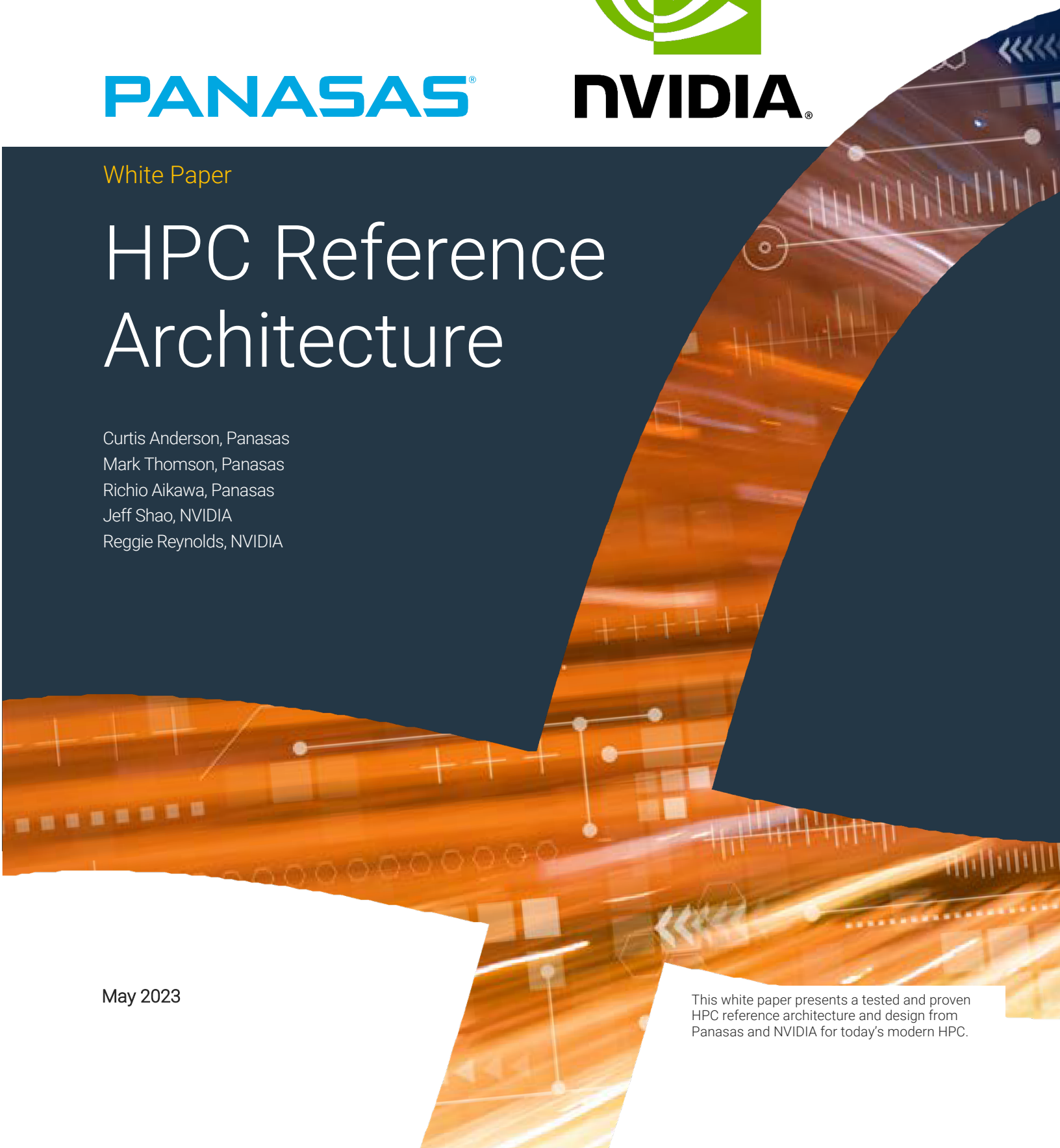
PANASAS®

NVIDIA®

White Paper

HPC Reference Architecture

Curtis Anderson, Panasas
Mark Thomson, Panasas
Richio Aikawa, Panasas
Jeff Shao, NVIDIA
Reggie Reynolds, NVIDIA



May 2023

This white paper presents a tested and proven HPC reference architecture and design from Panasas and NVIDIA for today's modern HPC.

Revisions

Date	Description
May 2023	Initial release

Table of Contents

Introduction

The High-Performance Computing (HPC) market continues to grow as it expands further into enterprise data centers, edge locations, and public clouds. The five-year 2021–2026 on-premises HPC server forecast from Hyperion Research predicts 6.9% growth, with storage historically the highest growth HPC component at an 8.6% compounded annual growth rate (CAGR) [1].

Coupled with the convergence of traditional modeling-simulation HPC, machine learning (ML), and high-performance data analysis (HPDA), today's modern HPC deployments require systems that are adaptable and performant under this mixture of workloads. In fact, this mixed workload market of HPC-enabled AI is expected to exceed 22% growth over the five-year period from 2021 to 2026, according to Hyperion Research [1].

The increase in the use of HPC continues to grow beyond the research labs in academia and government, as industries of all types are requiring more computational and storage resources and faster networks. Whether it's driving innovations in genomics and bioinformatics, accelerating modeling and simulation in commercial product development and manufacturing, improving energy exploration, or creating the latest advancements in gaming visual EFX, selecting an appropriate and simple to manage HPC infrastructure is critical.

However, HPC environments by their very nature tend to be large and are usually quite complex. Thus, familiarity with HPC components is important, and designing scalable and efficient systems to meet these new requirements can be challenging. Extremely beneficial are reliable blueprints from tested and proven reference architectures clarifying how to effectively design, install, use, and maintain today's modern HPC systems.

The Panasas® + NVIDIA® HPC Reference Architecture, referred to in the document as the "HPC Reference Architecture", was developed to address and mitigate the challenges that arise when piecing together components to build HPC systems of any scale, anywhere. This document includes a prescriptive full-rack reference design based on an integrated, tested, and supported configuration. For more details on installation and usage, refer to the latest Panasas ActiveStor® Installation and Service Guide and Panasas ActiveStor User Guide. Instructions on how to download these documents are provided in the Support section below.

Additional information not supported in the HPC Reference Architecture or reference design is also included in the document for technology, product, and feature set completeness of Panasas and NVIDIA solutions. As this information is supplemental, annotations to this effect are made in the document.

HPC Reference Architecture

The Panasas + NVIDIA HPC Reference Architecture is designed to address the needs of modern HPC from smaller workgroups to the largest supercomputing clusters, and for primary, remote, and edge data centers. Using NVIDIA's fast, reliable, and efficient networking infrastructure, CPU- and GPU-based client compute nodes can access the Panasas ActiveStor storage solution using the Panasas DirectFlow® Client driver over NVIDIA Spectrum™ Ethernet and Quantum InfiniBand® (IB)-based fabrics or NFS and SMB/CIFS protocols. Panasas PanFS®, the storage operating system and parallel file system that runs on the ActiveStor solution, and the other major components of the HPC Reference Architecture are detailed later in the document.

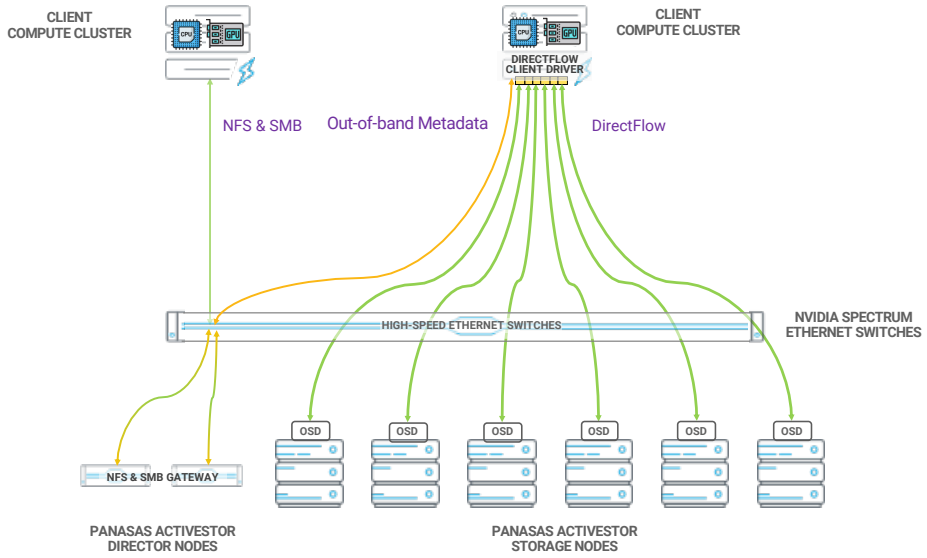


Figure 1. Panasas + NVIDIA HPC Reference Architecture – Ethernet-based Compute Cluster.

The HPC Reference Architecture shown in Figure 1 is for an Ethernet-based compute cluster. The basic components of this architecture include Panasas ActiveStor Director and Storage nodes, Panasas DirectFlow Client drivers on Linux compute servers, NVIDIA Spectrum™ Ethernet switches, and NVIDIA interconnects. These components are described in more detail later in the document.

The HPC Reference Architecture is expanded for an NVIDIA Quantum InfiniBand-based compute cluster (see Figure 2). Additional basic components include NVIDIA Quantum switches and a Panasas ActiveStor ASR-400.

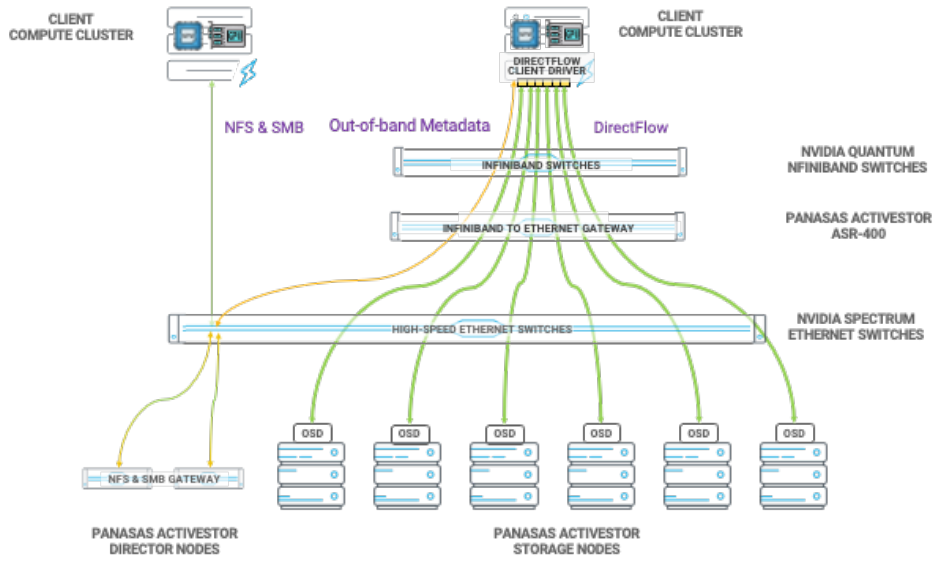


Figure 2. Panasas + NVIDIA HPC Reference Architecture – IB-based Compute Cluster.

Full Rack Reference Design

Based on the HPC Reference Architecture, a full-rack reference design for an NVIDIA Spectrum Ethernet-based compute cluster and a reference design for an NVIDIA Quantum InfiniBand-based compute cluster is provided.

Ethernet Reference Design

The components used in a full-rack reference design using Panasas ActiveStor Directors and Panasas ActiveStor Ultra storage enclosures with an NVIDIA Spectrum Ethernet-based compute cluster are described in Table 1.

Table 1. Full-Rack Reference Design (Ethernet-based Compute Cluster) Components.

Quantity	Component
8	Panasas ActiveStor Ultra ASU-100 enclosure (shelf) • 4 Storage nodes per enclosure (32 total Storage nodes)
2	Panasas ActiveStor Director ASD-200 enclosure • 4 Director nodes per enclosure (8 total Director nodes)
2	NVIDIA Spectrum SN3420 or SN2410 25GbE/100GbE Open Ethernet Switch • 48x 25GbE and 12x 100GbE QSFP28 ports or 48x 25GbE and 8x 100GbE QSFP28 ports, respectively, for high-speed networking • The SN3420 is being added to the reference design; the SN2410 is included
1	NVIDIA Spectrum SN2201 or AS4610-54T Ethernet Switch • 48x 10/100/1000Base-T RJ45 and 4x 100GbE QSFP ports or 48x 10/100/1000Base-T RJ45 and 4x 10GbE SFP+ ports, respectively, for management • The SN2201 is being added to the reference design; the AS4610-54T is included

A full rack of eight Panasas ActiveStor Ultra storage enclosures (see Figure 3) provides between 890 TB and 3.32 PB of raw data storage and 61.4 TB of metadata storage capacity.

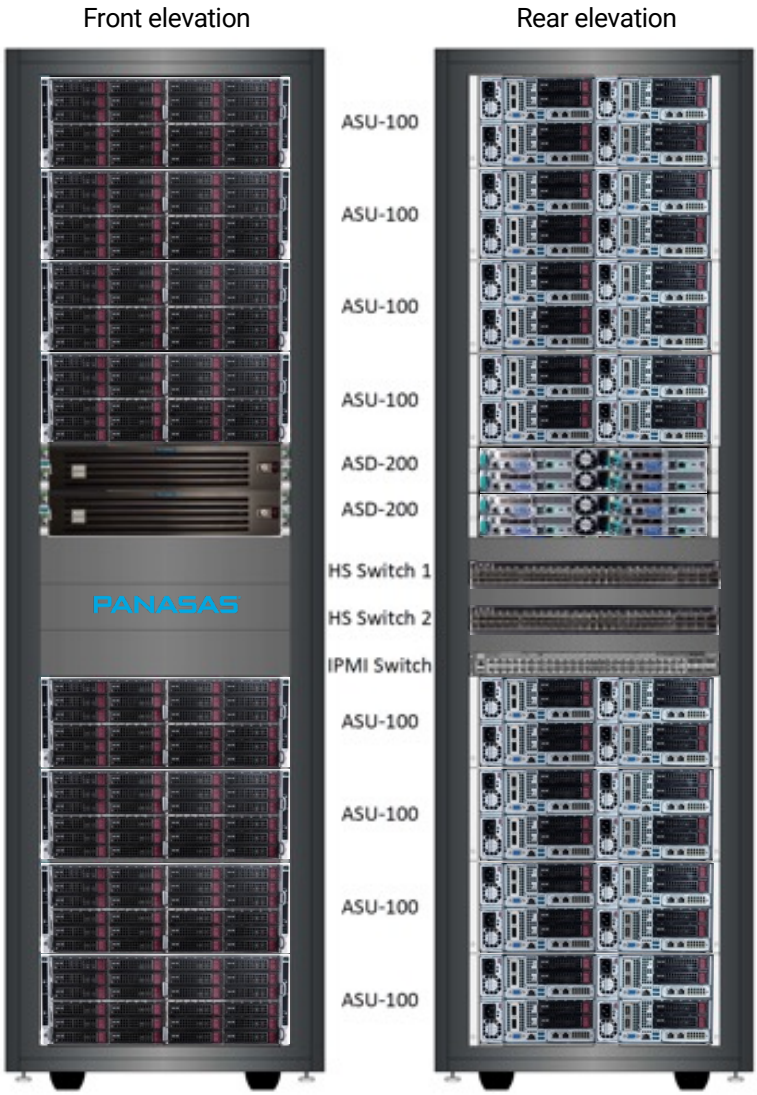


Figure 3. Panasas + NVIDIA HPC ActiveStor Ultra Full Rack Reference Design Front and Rear Elevations.

Reference Design Network Topology

The full-rack reference design network topology (see Figure 4) shows a no-single-point-of-failure networking topology for the Panasas + NVIDIA HPC ActiveStor Ultra full-rack reference design installation.

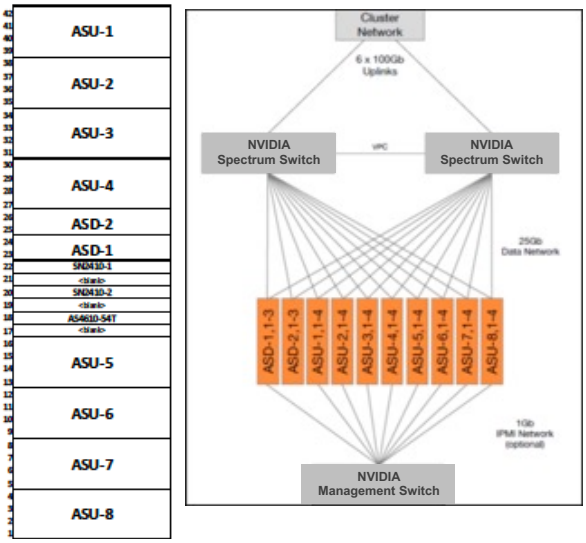


Figure 4. Panasas + NVIDIA HPC ActiveStor Ultra Full Rack Reference Design Network Topology.

InfiniBand Reference Design

In addition to the full-rack reference design shown in Figure 3, the reference design with an IB-based compute cluster includes NVIDIA Quantum InfiniBand switch(es) and a Panasas ActiveStor ASR-400 as described in Table 2.

Table 2. Full rack reference design (IB-based compute cluster) components.

Quantity	Component
8	Panasas ActiveStor Ultra ASU-100 enclosure (shelf) • 4 Storage nodes per enclosure (32 total Storage nodes)
2	Panasas ActiveStor Director (ASD-200 enclosure • 4 Director nodes per enclosure (8 total Director nodes)
2	NVIDIA Spectrum SN3420 or SN2410 25GbE/100GbE Open Ethernet Switch • 48x 25GbE and 12x 100GbE QSFP28 ports or 48x 25GbE and 8x 100GbE QSFP28 ports, respectively, for high-speed networking • The SN3420 is being added to the reference design; the SN2410 is included
1	NVIDIA Spectrum SN2201 or AS4610-54T Ethernet Switch • 48x 10/100/1000Base-T RJ45 and 4x 100GbE QSFP ports or 48x 10/100/1000Base-T RJ45 and 4x 10GbE SFP+ ports, respectively, for management • The SN2201 is being added to the reference design; the AS4610-54T is included
1 or more	NVIDIA Quantum-2 QMS9700/QM9790 or Quantum 8700 InfiniBand Switch • 64 ports of 400Gb/s InfiniBand per port or 40 ports of 200Gb/s InfiniBand per port, respectively • The quantity of Quantum/Quantum-2 InfiniBand Switches required is dependent on the compute cluster size
1	Panasas ActiveStor ASR-400 enclosure • 2 ASR-400 nodes per enclosure

If fiber optic cabling is used, the qualified optical transceivers are:

- NVIDIA LinkX® 25GbE SR SRP28 MMF Optical Transceiver model MMA2P00-AS on both sides of the fiber optic cable for the ASU-100 and on the switch side for the ASD-200
- Chelsio SM25G-SR 25G Short Reach SFP28 Optical Module on the ASD-200 side.

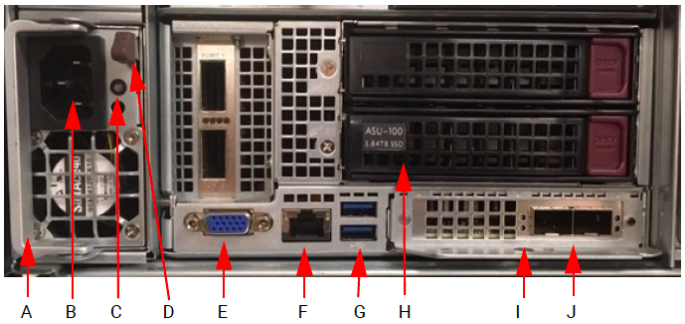
If copper cabling is used, NVIDIA LinkX Ethernet DAC cables are qualified.

Physical Connectivity

The Panasas ActiveStor Ultra Storage nodes defined in the Reference Design support 25 Gigabit Ethernet (GbE) networks via two network ports in the rear of each node (see Figure 5). Details on their physical connectivity are described below.

The default configuration upon initial installation is link aggregation across two ports – a 2 x 25GbE configuration using two 25GbE SFP28 cables, with one attached to each port. The Panasas ActiveStor Ultra Storage nodes support Link Aggregation Control Protocol (LACP) by default; static Link Aggregation Group (LAG), single link, and failover modes are also available.

Panasas ActiveStor Ultra Storage nodes also contain a single 1GbE port that may be used as a general administrative network port or for troubleshooting.



Label	Description
A	Power supply handle (folded closed)
B	Power supply outlet
C	Power supply status LED
D	Power supply latch
E	VGA monitor port
F	1GbE RJ45 admin port (1000Base-T)
G	2 USB ports
H	SATA SSD
I	Network adapter LEDs
J	Dual 25GbE network ports

Figure 5. Panasas ActiveStor Ultra Node and Power Supply Rear View.

Network Configuration Options

As described in the Physical Connectivity section. there are four network configuration options for the HPC Reference Architecture:

- Dynamic LACP
- Static LAG
- Single link
- Failover network.

The default network configuration for the Panasas ActiveStor Ultra Storage nodes is LACP across the dual 25GbE ports.

Generally, protocols other than LACP and static LAG operate in active/passive mode.

Active/Active Link Aggregation Mode

When load balancing is required to optimize performance, Panasas ActiveStor Ultra Storage nodes can be configured to use either dynamic LACP or static LAG. LACP is preferred, as it is significantly more robust than static LAG.

In static LAG mode, the physical ports are bonded with the IEEE 802.3ad static LAG link-layer protocol. This provides both load balancing and fault tolerance if a port loses its physical carrier status. Static LAG may fail to detect when the port stops functioning properly, but its carrier state will remain active.

In LACP mode, the physical ports are bonded with the IEEE 802.3ad LACP link-layer protocol. This provides load balancing, better fault tolerance, and protection against misconfiguration than static LAG.

Single Link Mode

While single link mode is supported on Panasas ActiveStor Ultra Storage nodes, it is not optimal since it is a single point of failure and suffers reduced bandwidth. Thus, single link mode should be used with caution, as loss of the one, single link will make the node inaccessible.

Network Failover Mode

Network failover is used on Panasas ActiveStor Ultra Storage nodes when active/passive redundancy is required.

Storage Configuration Options

Panasas provides two mechanisms to manage namespace and capacity: bladesets and volumes.

Bladesets

The bladeset is a physical mechanism that groups storage nodes into a uniform storage pool. It is a collection of three or more storage enclosures grouped together to store data. You can grow a bladeset by adding more hardware, and you can move data within a bladeset. (Refer to the Panasas ActiveStor User Guide for details on how to add to a bladeset and move data within one.)

Volumes

A volume is a logical mechanism, a sub-tree of the overall system directory structure. A read-only top-level root volume ("/"), under which all other volumes are mounted, and a /home volume are created during setup. All other volumes are created by the user on a particular bladeset, with up to 3,600 per realm.

A volume does not have a fixed space allocation, but instead has an optional quota which can set a maximum size for the volume. Effectively, the volume is a flexible container for files and directories, and the bladeset is a fixed size container for multiple volumes.

When planning volume configuration, keep the following points in mind:

- Volumes can be used to manage capacity.
- Volumes can be created to complement backup strategies.
- Performance can be enhanced by assigning volumes to different directors.

RAID and Erasure Coding

Instead of relying on hardware RAID controllers that protect data at a drive level and computing RAID on the disks themselves, the Panasas PanFS architecture uses per-file distributed RAID in software using erasure codes, i.e., per-file erasure coding rather than hardware RAID. Files in the same bladeset, volume, and even directory can have different RAID/erasure code levels. A file can be seen as a single virtual object that is sliced into multiple component objects.

PanFS uses objects to store POSIX files differently than how S3 objects are typically used to store files. Instead of storing each file in an object, PanFS stripes a large POSIX file across a set of component objects and adds additional component objects into that stripe that store the P and Q data protection values of N+2 erasure coding. Using multiple objects per POSIX file enables the file striping that is one of the sources of a parallel file system's performance [2].

Users have three RAID levels available: RAID 6, RAID 5, and RAID 10.

- RAID 6 is the default RAID for all volumes, as it provides a balance between performance, capacity overhead, and RAID protection.
- RAID 5 volumes are optimized for performance and capacity overhead.
- RAID 10 volumes combine striping and mirroring to provide high performance for applications that perform small random writes, while at the same time providing resiliency against a single disk failure.

You can mix the RAID levels and any volume layout together in the same bladeset, with each volume evaluated independently for availability status.

Compute System

The primary component of HPC is the compute cluster, which is composed of usually dozens, if not hundreds or even tens of thousands, of interconnected servers that work in parallel as a single unit. In a cluster, each server is referred to as a *node*. Nodes are typically all identical, but some clusters consist of nodes that incorporate accelerators, including NVIDIA GPUs, for accelerated computing. Computationally intensive workloads consisting of individual parallelized tasks are distributed among the nodes in the cluster.

Client compute cluster hardware requirements are determined in part by the applications the cluster will run. The applications help determine the type of processor and/or accelerator, the memory size, the amount and type of local storage, and possibly operating system requirements. Additionally, IT organizations often seek solutions that help lower power consumption and operational costs while enhancing scalability of their HPC infrastructure amidst budget limitations, energy efficiency initiatives, and growing demands for increased processing and capacity.

Today's HPC is also evolving with the end of Moore's law, a changing HPC ecosystem, and the growing use and importance of AI/ML in both applications and systems [3]. Beyond the current state-of-the-art combination of HPC and AI/ML, new workflows are being created that integrate HPC simulations with AI/ML driven predictions and observations. The result is an HPC shift towards distributed computing and data workflows for optimized time to completion [3].

CPUs and GPUs

The use of CPUs by themselves for HPC processing is no longer enough to meet the computational requirements of large and complex problems., Work is underway to create a workload division framework between CPUs and GPUs based on computational and data transfer rate capabilities [4].

Computer architects, programmers, and researchers are already shifting from CPU-only *homogenous computing* to a CPU plus GPU paradigm, or *heterogeneous computing*, where the best of both can be combined for more computational gains [5]. Modern GPUs, such as NVIDIA's H100, accelerate both vector and matrix operations spanning a range of numerical format and by themselves, NVIDIA GPUs accelerate over 2,500 HPC applications over a spectrum of fields from material science, quantum chemistry, seismic processing, weather, and climate [6].

The next steps involve novel techniques to update conventional CPU-only or GPU-only optimizations. Such updates will continue to realize the promise of heterogeneous CPU plus GPU or FPGA computing for exascale performance [5].

File System and Storage

With growing cluster sizes, higher performance servers, and higher speed networks, getting enormous quantities of data in and out of storage can be a challenge. For example, innovations in genomics, cryo-EM, drug discovery, and other life science disciplines depends on the ability to rapidly store, analyze, and retrieve unprecedented volumes of data, but legacy storage systems just aren't fast, scalable, or adaptable enough. In manufacturing, powerful simulation tools require storage solutions that can reliably deliver huge volumes of test and simulation data with high performance and availability to support today's streamlined design workflows. Other industries including energy, media and entertainment, financial services, and academic and government research also require high-performance storage for the acceleration of workloads to workflows, and from product development to discovery.

With such large quantities of data and the need to keep high-performance compute nodes busy, only a storage system with a true parallel file system architecture will meet the performance requirements. As today's IT HPC departments have neither the time or personnel for the HPC storage "science project" approach of the past, the need for a turnkey approach with an optimized file system, optimized hardware, unrivaled reliability, easy manageability, and enterprise-class support exists.

Panasas PanFS Parallel File System

The Panasas PanFS storage operating system is the integrated software that manages and directs Panasas systems (see Figure 6). PanFS dynamically distributes file activity across Panasas ActiveStor storage nodes and clusters ActiveStor Director nodes to coordinate file activity, balance system performance, and manage failovers. This distributed cluster-based approach eliminates performance bottlenecks and eases the administrative burden for the Panasas storage network.

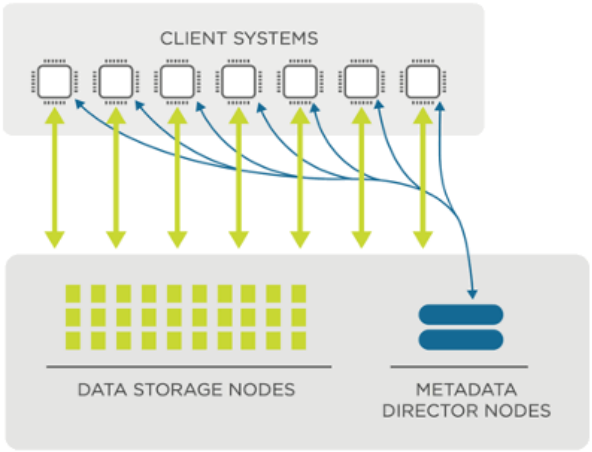


Figure 6. Panasas PanFS Architecture.

PanFS is an object-based parallel file system. Files are divided into components that are stored in different objects and accessed in parallel for high performance and redundancy. Data objects are containers for data and attributes that describe how the data fits into the overall distributed file system.

PanFS orchestrates multiple storage services into a single entity that serves your data to your compute cluster. Through sophisticated software, multiple storage servers that each contain HDDs and/or SSDs work together to support hundreds of gigabytes per second (GB/s) of data being read and written by your HPC applications. PanFS manages this orchestration without manual intervention, automatically recovering from any failures and continuously balancing both the load across those storage servers and scrubbing the stored data for the highest levels of data protection.

Each file stored by PanFS is individually striped across many ActiveStor storage nodes, allowing each file to be read and written in parallel, increasing the performance of accessing each file. Each file has its own *map* that defines the striping pattern for that file. This mapping allows different files to have different striping patterns and supports a choice of different performance and redundancy characteristics. PanFS uses erasure coding as part of that striping to ensure the highest level of data integrity and reliability. PanFS is also a *direct file system* that allows the client systems to talk over the network directly to all the storage nodes.

Director Nodes

Panasas ActiveStor architecture scales data and metadata independently and is purpose-built for adaptability and flexibility to handle a wide range of use cases. Panasas ActiveStor Directors function as the *control plane* of the system, managing metadata services instead of storing user data. ActiveStor Director nodes control distributed filesystem operations such as file-level and object-level metadata consistency, client cache coherency, recoverability from interruptions to client I/O, storage node allocation operations, and secure multiuser access to files.

In addition, ActiveStor Director nodes control many other aspects of the overall storage system including namespace management, system health, failure recovery actions, and gateway functionality. ActiveStor Directors also facilitate scalability and virtualize data objects across all available storage nodes, enabling the system to be viewed as a single, easily managed global namespace. The pool of ActiveStor Director nodes can be scaled independently of the pool of ActiveStor storage nodes to scale metadata performance.

NFS and SMB/CIFS Gateway

One of the roles of ActiveStor Director nodes in PanFS is to act as gateways that translate Network File System (NFS) and Server Message Block/Common Internet File System (SMB/CIFS) operations into DirectFlow operations, allowing clients such as laptops and workstations to access the same namespace and data as the HPC compute cluster. PanFS provides high performance NFS and SMB/CIFS access, but

as a result of its parallel and direct nature, the DirectFlow protocol will always be the highest performance path to PanFS storage.

Storage Nodes

Panasas ActiveStor storage nodes are the core of the *data plane*. In scale-out storage systems like PanFS, there simply is no maximum performance or maximum capacity and PanFS has been architected to provide linear scale-out; performance and capacity scale linearly as storage nodes are added. For more capacity or more storage performance, just add more storage nodes.

Storage nodes in PanFS are highly sophisticated Object Storage Devices (OSDs). Instead of storing each file in an object (see Figure 7), PanFS stripes a large POSIX file across a set of component objects and adds additional component objects into that stripe that store the P and Q data protection values of N+2 erasure coding. Using multiple objects per POSIX file enables the file striping that is one of the sources of a parallel file system's performance.

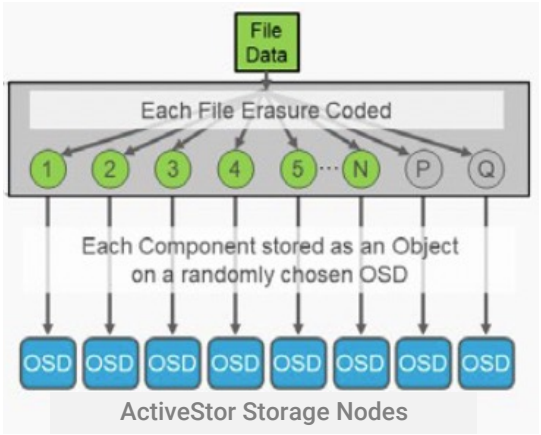


Figure 7. Panasas PanFS Per File Object Erasure Coding.

While large POSIX files are stored using erasure coding across multiple component objects, small POSIX files use triple replication across three component objects. This approach delivers both better performance than can be achieved by using erasure coding on such small files and better space efficiency.

ActiveStor storage nodes feature a multi-tier intelligent data placement architecture, called Dynamic Data Acceleration, that matches the right type of storage media to each type of data to deliver the highest performance at the lowest cost. For an ActiveStor Ultra Storage node:

- An NVDIMM-based intent-log protects inflight data and metadata operations.
- Metadata is stored in a database on low-latency NVMe SSDs.
- Small files are stored on high-IOPS flash SSDs.
- Large files are stored on low-cost, high-capacity, high-bandwidth HDDs.
- Unmodified data and metadata are cached in system DRAM.

Data Reconstruction

If a ActiveStor storage node were to fail, PanFS would reconstruct only those component objects that were on the failed node, not the entire raw capacity of the node like a RAID array would. PanFS reads the component objects for each affected file from all the other ActiveStor storage nodes and use each file's erasure code to reconstruct the component objects that were on the failed node.

PanFS also continuously scrubs the data integrity of the system in the background by slowly reading through all files in the system, validating that the erasure codes for each file match the data in that file.

DirectFlow Client Driver

Similar to massively parallel supercomputer architectures, Panasas DirectFlow Client software supports fully parallel reads and writes from the DirectFlow clients to OSDs without encountering bottlenecks (see Figure 8). The DirectFlow Client software module is a loadable kernel module that adds support for the PanFS filesystem protocols to Linux compute nodes, i.e. a file system implementation installed on the compute servers and used by application programs like any other file system. DirectFlow Client works with the ActiveStor director and storage nodes to deliver fully POSIX-compliant file system behavior from a single namespace, across all servers in the compute cluster, directly to the storage nodes. The DirectFlow Client also uses a map made by PanFS to each file to know which storage nodes to access both directly and in parallel.

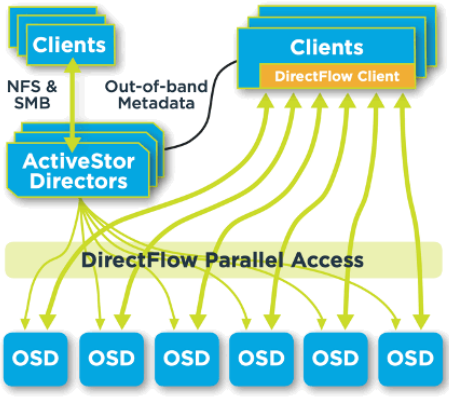


Figure 8. Panasas PanFS DirectFlow Linux Client Driver.

Panasas ActiveStor Product Line

Panasas provides a product line of turnkey appliance systems called ActiveStor that are powered by the PanFS operating system. ActiveStor includes a portfolio of storage systems: ActiveStor Flash, ActiveStor Ultra Edge, ActiveStor Ultra XL, ActiveStor Ultra, ActiveStor Director, and ActiveStor ASR-400. (While the HPC Reference Architecture supports the entire ActiveStor storage portfolio, the full rack reference designs are based on the ActiveStor Ultra storage system.)

ActiveStor Flash

ActiveStor Flash (ASF-100) is a high-performance all-flash data storage appliance built on a balanced hardware platform chosen for its high-density form factor, the AMD EPYC processor, and all-NVMe SSD support. It delivers outstanding scratch storage performance, enhanced support for AI/ML training projects, and higher rack density. ActiveStor Flash complements the ActiveStor Ultra and ActiveStor Ultra XL storage appliances as a higher performance tier.

ActiveStor Ultra Edge

ActiveStor Ultra Edge (ASU-100E) is a small form factor, high-performance storage appliance with flexible configuration options and endless scalability. ActiveStor Ultra Edge with PanFS delivers industry-leading mixed HPC and AI/ML workload performance for today's modern HPC applications where you need it, such as at primary and remote data centers, colocation sites, and edge HPC locations. With ActiveStor Ultra Edge, performance, reliability, simplicity, and flexibility are delivered in a compact and low-cost entry point.

ActiveStor Ultra XL

ActiveStor Ultra XL (ASU-100XL) is a large-capacity data storage appliance that can deliver 5.76 PB in a single 42U rack. It is the industry's best price/TB parallel file system solution and is an ideal choice for high-performance applications using large files and cooler dataset workloads, such as seismic resource exploration, manufacturing, and media and entertainment.

ActiveStor Ultra

ActiveStor Ultra (ASU-100) is a hybrid, mixed-workload storage appliance. It is a scale-out, parallel file system storage appliance that is built on industry-standard hardware chosen for its carefully balanced architecture, with emphasis on mixed media storage. Each ActiveStor Ultra Storage node is powered by the PanFS parallel filesystem to deliver the extreme performance, enterprise grade reliability, unlimited scalability, and ease of management required to process datasets of the size and complexity associated with high-performance computing and AI/ML in manufacturing, life sciences, energy, financial services, media and entertainment, and academic and government research.

PanFS uses a scale-out architecture that grows storage capacity, DRAM caching, and network bandwidth incrementally and linearly as you add more ActiveStor Ultra storage enclosures. It delivers data from storage nodes in parallel to the application, multiplying the bandwidth an application can achieve to a single file, not just aggregate bandwidth. Data flows directly from the storage nodes to the application without any hops through intermediate servers or even extra network links.

The ActiveStor Ultra enclosure is a 4U, 19" rackmount, four-node chassis. Each of the four ASU-100 nodes contains six HDDs (see Figure 9).

Figure 9. Panasas ASU-100 Enclosure Front – Four Nodes.

The rear of the ASU-100 enclosure has one SSD, dual 25GbE ports, and a 1GbE (1000Base-T) port for IPMI (see Figure 4). The drive options and capacities for the ActiveStor Ultra node and enclosure (4 nodes/enclosure) are detailed in Table 3.

Table 3. ActiveStor Ultra Drive Configuration Options and Capacities.

Configuration Options	HDDs	SATA SSDs	NVMe SSD
# Drives	6	1	1
Capacity/drive	4 TB, 12 TB, or 16TB	3.84 TB or 7.68 TB	3.84 TB
Capacity/node	24 TB – 96 TB	3.84 TB or 7.68 TB	3.84 TB
Capacity/enclosure	96 TB – 384 TB	15.36 TB – 30.72 TB	15.36 TB

ActiveStor Director

The Panasas ActiveStor Director (ASD-200) is powered by PanFS and controls many aspects of the overall storage solution including filesystem semantics, namespace management, distribution, and consistency of user data on storage nodes, system health, failure recovery, and gateway functionality.

The ASD-200 is a 2U enclosure containing up to four ActiveStor Director nodes. It is backward and forward compatible with the entire portfolio of ActiveStor storage nodes. The ASD-200 takes full advantage of the latest technologies, such as dual 25GbE networking ports for high-bandwidth data movement and non-volatile dual-inline memory modules (NVDIMM) technology for ultra-low latency persistent storage.

ActiveStor ASR-400

The Panasas ActiveStor ASR-400 provides scalable, high-performance InfiniBand connectivity to Panasas ActiveStor Ethernet-based storage systems. With the addition of ASR-400 nodes, the Panasas system gains scalable, high-bandwidth, fault-tolerant networking paths to compute clusters that use InfiniBand.

An ASR-400 node is a server that has both Ethernet and InfiniBand network interfaces and is configured to route IP traffic between the two network interfaces. The ASR-400 enclosure can be configured with up to four nodes, providing cost-effective high performance and low-latency InfiniBand fabric connectivity to the Panasas ActiveStor system. (The ASR-400 is an optional component of the HPC Reference Architecture.)

System Limits and Requirements

General limits and requirements of the PanFS filesystem and the ActiveStor Ultra storage system are provided in Table 4.

Table 4. Panasas PanFS and ActiveStor Ultra System Limits and Requirements.

Limit	Description
Minimum Ultra system	• 3 Panasas ASU-100 enclosures (12 OSDs) with 1 Panasas ASD-200 director enclosure
Largest Ultra system	• No enforced limit • Panasas has tested up to 1300 OSDs
Maximum number of files or directories	• No enforced limit
Maximum file size	• CIFS: 32 TB • NFS: 32 TB (FreeBSD buffer cache limitation) • 32-bit Linux DirectFlow: 16 TB • 64-bit Linux DirectFlow: 8192 PB
Maximum number of volumes	• Per director: no limit • Per bladeset: no limit • Per realm: 3600
Maximum number of Realm Managers (RMs) in repset	• 5
Maximum bladeset size	• No enforced limit
Maximum number of snapshots	• 32 (enforced) per volume
Maximum time between snapshots	• 10 minutes (enforced) • 30 minutes (recommended) You can schedule many volumes within the same bladeset to have a snapshot taken at the same time; these will be batched and optimized. The system does snapshot work in the background and, in some cases, will reject a snapshot if activity associated with a previous snapshot is still in progress. You can also attempt snapshots more frequently using the manual "Take a Snapshot Now" button.
Maximum number of entries per directory	• 1,000,000 (enforced) • 250,000 (recommended)
Maximum number of clients	• 30,000
Maximum CIFS clients per Director	• No limit
Maximum path length	• 4096 bytes

High-Performance Network and Interconnect

Once the compute cluster is defined, the cluster must be connected into a network. The *interconnect* is in fact the central nervous system running through the cluster, and choosing the right technology to interconnect all the hardware components for best performance delivery is key. To accelerate and maximize the compute and storage performance, NVIDIA offers an end-to-end high-performance network providing the highest throughput and lowest latency (see Figure 10). The NVIDIA networking solutions consist of network switches, smart network interface cards (Smart NICs), data processing unit (DPU) accelerator cards, and cables and transceiver modules, supporting both InfiniBand and Ethernet protocols up to 400G speed. The NVIDIA networking solutions accelerate all high-performance computing and AI/ML workloads in data centers and clouds. 72% of the TOP500 Supercomputers, including several of the top ten HPC and AI Supercomputers on the June 2022 TOP500 list, have an NVIDIA Quantum InfiniBand or Spectrum Ethernet backbone.

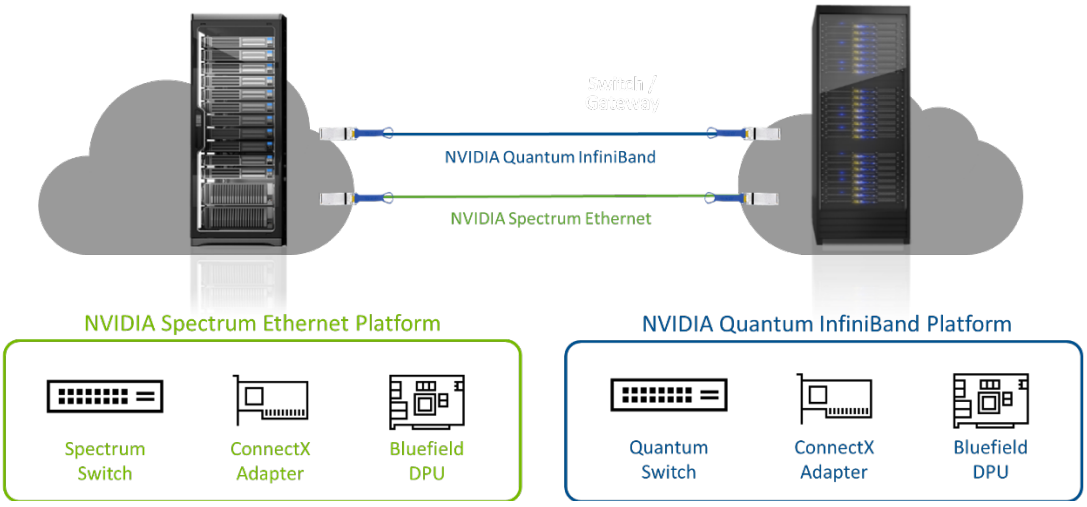


Figure 10. End-to-end NVIDIA Networking Solutions.

NVIDIA Spectrum Ethernet Switches

The NVIDIA Spectrum Ethernet Switch product family is a broad portfolio of top-of-rack and aggregation switches, deployed in layer-2 and layer-3 cloud designs, in overlay-based virtualized networks, or as part of high-performance, mission-critical Ethernet storage fabrics. In any deployment, you'll get:

- Industry-leading performance and predictability
- High-performance Remote Direct-Memory Access over Converged Ethernet (RoCE)
- Cloud scalability and open networking
- Comprehensive visibility with advanced telemetry
- Advanced virtualization and security
- Time to deployment with automated configuration and management.

NVIDIA Spectrum SN3420 and SN2410 Ethernet Switch

The NVIDIA Spectrum SN3420 and SN2410 Ethernet switches (see Figures 11 and 12) are top-of-rack (ToR) switches for high-performance storage. They meet and exceed the growing demands of predictable, consistent high performance, density, and power consumption for today's database, storage, and data center environments.

Figure 11. NVIDIA Spectrum SN3420 Ethernet Switch.

Figure 12. NVIDIA Spectrum SN2410 Ethernet Switch.

NVIDIA Spectrum SN3420/2410 Ethernet Switch Highlights

- Ideal spine and top-of-rack (ToR) solutions allow for maximum flexibility.
- Port speeds span from 10Gb/s to 100Gb/s per port.
- High port density enables full rack connectivity to any server at any speed.
- Uplink ports allow a variety of blocking ratios that suit any application requirement.

Powered by an NVIDIA Spectrum switching ASIC, the SN3420 Ethernet switch carries a whopping switching capacity of up to 2.4Tb/s with a landmark 3.58Bpps processing capacity in a compact 1RU form factor. (While RoCE is supported by NVIDIA Spectrum Ethernet Switches, it is not an element of the HPC Reference Architecture. The Spectrum SN3420 Ethernet Switch is being added to the reference design; the Spectrum SN2410 Ethernet Switch is included.)

NVIDIA 1G Spectrum SN2201 and AS4610-54T Ethernet Switch

NVIDIA offers 1G Ethernet switches for network management, the Spectrum SN2201 and AS4610-54T. The NVIDIA 1G Ethernet switches provide 48x 10/100/1000Base-T RJ45 ports with 10G or 100G uplinks. The NVIDIA 1G Ethernet switches support all standard compliances and are fully interoperable with third-party systems. (The Spectrum SN2201 Ethernet Switch is being added to the reference design; the AS4610-54T Ethernet Switch is included.)

NVIDIA Quantum InfiniBand Switches

Essential for HPC, AI, and big data, NVIDIA Quantum InfiniBand Switches are the clear leader that introduces in-network computing capabilities into the network specifically designed to achieve maximum performance. NVIDIA provides complete end-to-end solutions supporting InfiniBand networking technologies.

NVIDIA Quantum/Quantum-2 InfiniBand Switches

Faster servers, high-performance storage, and increasingly complex computational applications drive data bandwidth requirements to new heights. NVIDIA Quantum InfiniBand Switches deliver a complete switch system and fabric management portfolio that enables managers to build highly cost-effective and scalable switch fabrics ranging from small clusters up to thousands of nodes, reducing operational costs and infrastructure complexity. Quantum InfiniBand Switches bring a high-speed, extremely low-latency and scalable solution that incorporates state-of-the-art technologies such as Remote Direct Memory Access (RDMA), adaptive routing, and NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™. Static routing, adaptive routing, and advanced congestion management optimize computing efficiencies, making NVIDIA Quantum InfiniBand Switches ideal for top-of-rack leaf connectivity or for small to extremely large clusters.

NVIDIA Quantum-2 InfiniBand Switches

NVIDIA Quantum-2 is the industry-leading switch platform in power and density, with 400Gb/s InfiniBand throughput that provides AI developers and scientific researchers with the highest networking performance available to take on the world's most challenging problems. The NVIDIA Quantum-2-based QM9700 and QM9790 switch systems deliver an unprecedented 64 ports of 400Gb/s InfiniBand per port in a 1U standard chassis design. NVIDIA Quantum-2 InfiniBand Switches deliver extremely high networking performance by delivering up to 16Tb/s of non-blocking bandwidth.

Figure 13. NVIDIA Quantum QM8700 InfiniBand Switch.

NVIDIA Quantum QM8700 InfiniBand Switches (see Figure 13) provide extremely high networking performance by delivering up to 16Tb/s of non-blocking bandwidth with extremely low latency. NVIDIA QM8700 switches provide up to 40 ports of 200Gb/s full bi-directional bandwidth per port. And novel NVIDIA HDR100 technology supports up to 80 ports of 100Gb/s, enabling Quantum switches to provide double-density radix for 100Gb/s data speeds, reducing the cost of network design and network topologies.

NVIDIA ConnectX® InfiniBand Adapters

InfiniBand connectivity in an ActiveStor ASR-400 node is provided by a pair of dual-port NVIDIA smart InfiniBand adapter cards. NVIDIA provides a family of ConnectX InfiniBand Adapters, including the ConnectX-6 and ConnectX-5 adapters (see Figures 14 and 15). ConnectX adapters offer low latency, a high message rate, plus an embedded PCIe switch and Fabric offloads.

Each adapter in the ASR-400 provides one port of 100Gb/s InfiniBand and one port of 100GbE Ethernet connectivity. (The ConnectX InfiniBand Adapter is an optional component of the reference design. The ConnectX-6 is being added to the reference design; the ConnectX-5 is included.)

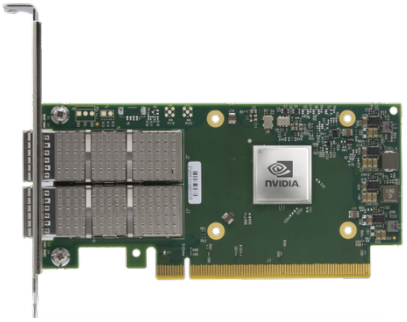


Figure 14. NVIDIA ConnectX-6 IB Adapter.

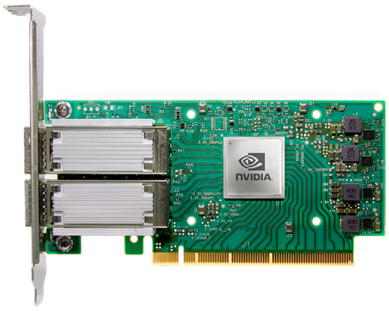


Figure 15. NVIDIA ConnectX-5 IB Adapter.

Cables and Transceivers

The NVIDIA LinkX® product family of cables and transceivers provide the industry's most complete line of 10, 25, 40, 50, 100, 200, and 400GbE in Ethernet and 100, 200, and 400Gb/s NDR in InfiniBand products for high-performance computing, AI/ML, and other data center/cloud applications.

LinkX cables and transceivers are used to link top-of-rack switches downwards to network adapters in GPUs and CPU servers and storage, and/or upwards in switch-to-switch applications throughout the network infrastructure.

Management Network

Panasas uses a 1GbE management network for IPMI to the ASU-100 and ASD-200 nodes in the HPC Reference Architecture. While IPMI is optional, it makes remote administration much easier. You will need 1GbE links for the NVIDIA SN2201 or AS4610-54T Ethernet switch connectivity.

Software and System Management

Panasas PanFS Software Suite

The Panasas PanFS software suite (see Figure 16) is a family of software built around the PanFS filesystem.

Figure 16. PanFS Software Suite.

In addition to PanFS itself, the suite includes software for data mobility and protection, data visibility and analytics, PanFS and realm management, and security features (see Table 5).

Table 5. Panasas PanFS Software Suite Components.

Components	Description
PanFS OS	POSIX-compliant parallel filesystem with Panasas DirectFlow Client driver and self-healing, self-managing data engine featuring client-based per-file object erasure coding, auto-tuning, capacity balancing, and auto-recovery from catastrophic failure using quadruple-redundant directory copies
PanMove*	Data movement and data protection software family for data center, cloud, edge and S3-object store
PanView*	Discovery, analytics, and visualization software family for space consumption and asset availability insight
PanActive Manager	System-level monitoring and management of Panasas PanFS realms through single simple to use GUI
PanCLI	Panasas PanFS command-line interface
Realm Manager	Panasas PanFS and director clustering
Security	Hardware-based encryption-at-rest, SELinux file labels, and granular ACLs

* Collaboration with Atempo®

The PanMove solutions bring a balance of simplicity and functionality to solve data mobility challenges in the data center, at the edge, and in the cloud. PanMove provides both a scalable and robust client-side parallel data replication solution with state-of-the-art parallelized rsync features and a seamless copy, move, and sync functionality for data movement between PanFS-based ActiveStor storage appliances, between the ActiveStor appliances and the cloud including support for AWS, Azure, and Google Cloud, and between Panasas appliances and on-prem and cloud-based S3 object storage.

With PanView, Panasas addresses the critical challenge of understanding the space consumption and asset availability of storage systems, helping organizations reduce data duplication and storage costs and organize their data so that they can find the right assets at the right time. PanView provides clear visibility into PanFS-based ActiveStor storage appliances, with extensive analytical information and insights to make informed data placement and management choices. Comprehensive visualization reports and displays bring data into focus enabling intelligent data management choices.

System Management

Unquestionably, HPC systems are more than just compute clusters, networks, and storage solutions. Effective HPC system architectures must also provide a seamless system management software environment for control and management of the HPC system.

Panasas PanActive Manager

The Panasas ActiveStor Ultra storage system is a single entity you manage from one graphical user interface (GUI) or command-line interface (CLI), no matter how many ActiveStor enclosures you integrate into it (see Figure 17). With a single point of management, an intuitive interface, and many automated enterprise management functions, Panasas PanActive Manager makes it easy to manage hundreds of terabytes to hundreds of petabytes of storage. One IT administrator is all it takes.

Figure 17. Panasas PanActive Manager.

Even in the largest Panasas deployments, all data resides within a single namespace, with a single management GUI and CLI, delivered with high reliability and availability. It is possible to quickly add more ActiveStor enclosures, and each added unit will immediately contribute more capacity and performance.

The Panasas ActiveStor solution automatically rebalances capacity across the realm as new ActiveStor Storage enclosures are added, or if nodes become unbalanced,

automatically reconstructs the full levels of erasure-coded data protection for all files in the event of a failure, and continuously scans all files in the background to scrub out any latent issues. Consolidating many different unstructured data workloads into a single scale-out ActiveStor solution not only reduces storage complexity in your data center, it also reduces operational expenses.

Support

Outstanding support is an essential ingredient in minimizing downtime in technical computing environments. For the Panasas + NVIDIA HPC Reference Design, Panasas working with NVIDIA provides worldwide enterprise-class support. Refer to the Panasas website at https://www.panasas.com/support/support-plans/ for details on support plans. In addition, customized offerings to meet specific support requirements are available.

All Panasas support services include access to the MyPanasas online self-service portal, where you can access support resources anytime, anywhere. MyPanasas also offers access to custom software downloads and a robust knowledge base that can help quickly resolve technical issues.

Panasas support services satisfy most companies' requirements. However, for companies with sensitive data that demands a high level of confidentiality and security, as well as the ability to meet classified data center requirements, Panasas offers SecureDisk and SecureBlade services. The Panasas Support Services team brings many years of operational experience in highly secure environments and understands how to optimize secure installations.

Summary

Navigating the open waters of high-performance computing is a difficult task, made more so by the deluge of new technologies emerging every day. For many manufacturing, life sciences, energy, financial services, media and entertainment organizations, and academic and government research centers, the effective use of HPC environments provides substantial improvements, from processing and analyzing data, to discoveries and product development, and even to developing and enhancing finance and business strategies.

Design, implementation, and support of HPC infrastructures are often complicated and confusing and can benefit from a turnkey approach. To that end, the Panasas + NVIDIA HPC Reference Architecture provides reference designs for both NVIDIA Spectrum Ethernet- and NVIDIA Quantum InfiniBand-based compute clusters, a full-rack configuration of Panasas ActiveStor Ultra storage systems, networking, interconnects, software, and management components that are integrated, tested, and supported.

The reference designs are both prescriptive and descriptive, as well as instructive in providing future paths. Included are component details and configuration options to guide HPC solution architects and system administrators in their planning of high-performance systems that can accelerate and support the converged application workloads of today's modern HPC users, including concurrent and integrated workflow of modeling and simulation with HPDA and ML. The reference designs are tested to support workloads with both CPUs and accelerators such as GPUs, with both high-speed NVIDIA Spectrum Ethernet and NVIDIA Quantum InfiniBand networks, and with high-performance Panasas ActiveStor Ultra storage systems with an architecture and engine designed to meet the rigors of many different users running many different applications at the same time.

To learn more about Panasas and NVIDIA technologies and products, visit us at www.panasas.com and www.nvidia.com, respectively.

Take the key next step and engage with Panasas and NVIDIA.

References

1. Joseph, E., et al, “Hyperion Research HPC Market Update,” Nov 2022

2. Anderson C., “PanFS 9: Architectural Overview.” [Online}. Available: http://www.panasas.com/resources/whitepapers/panfs-9-architectural-white-paper/

3. Ang J.A., Mountain, D.J., “New Horizons for High-Performance Computing,” Computer, 15 Nov 2022, DOI: 10.1109/MC.2022.3200859.

4. Chandrashekahr, B.N., Sanjay, H.A., “Performance Framework for HPC Applications on Homogeneous Computing Platform,” I.J. Image, Graphics and Signa Processing, 2019, 8, 28-39, DOI: 10.5815/ijigsp.2019.08.03.

5. Mittal S., Vetter J.S., “A Survey of CPU-GPU Heterogeneous Computing Techniques,” ACM Computing Surveys, Vol. X, No. Y, Article 1, Feb 2015, DOI: 10.1145/2788396.

6. “High-Performance Computing for the Age of Ai and Cloud Computing.” [Online]. Available: http://resources.nvidia.com/en-us-hpc-ebooks/hpc-for-the-age-of-ai?xs=409135

05/2023

PANASAS®

2680 N. First St., Suite 150
San Jose, CA 95134

PH +1.888.PANASAS
F +1.408.215.6801
www.panasas.com