BOSTON
Servers I Storage I Solutions

PANASAS®

Panasas ActiveStor Solution: Architectural Overview

WHITE PAPER

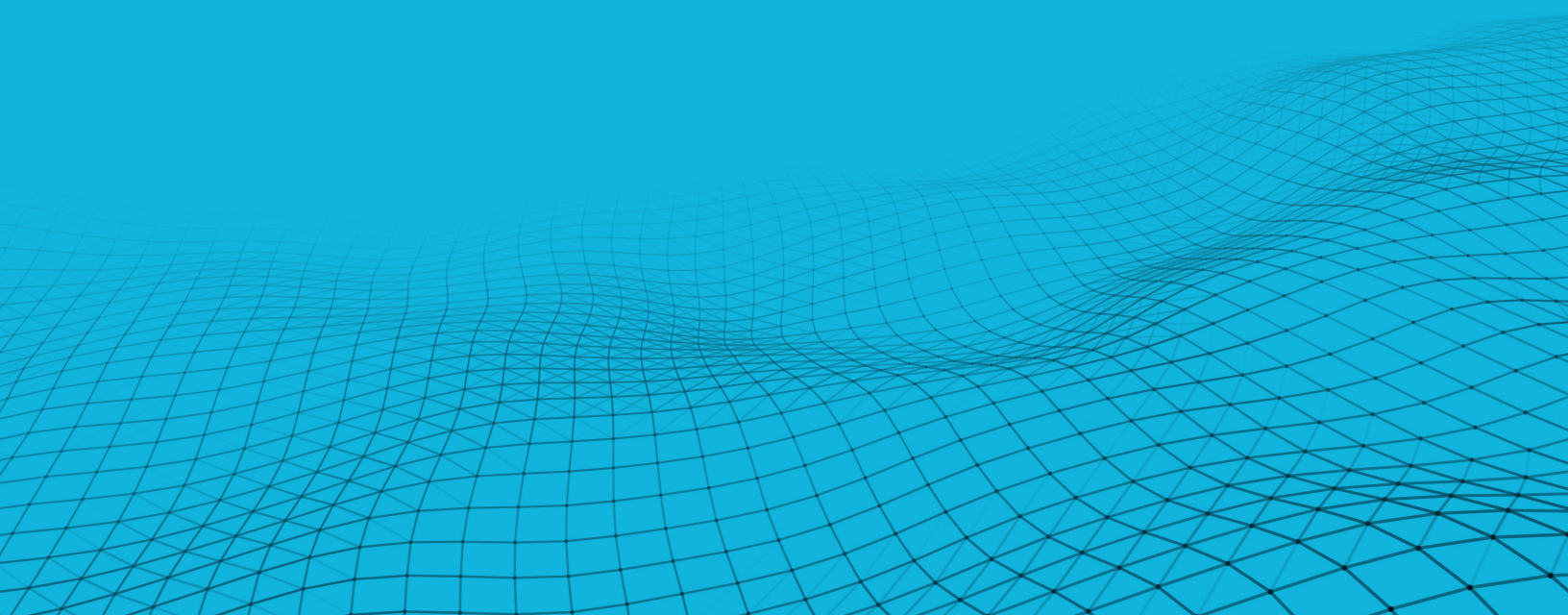


TABLE OF CONTENTS

Panasas ActiveStor 3

- High Performance 3
- Surprising Simplicity 3
- Extreme Flexibility 4
- Low Cost to Own and Operate 4
- Timely, High Quality Support 4

Panasas PanFS 6

- Director and Storage Nodes 6
- Scale-out Object Store Back End 6
- POSIX on Objects 6
- Volumes in the Namespace 7
- Storage Media 7

Panasas PanFS Data Protection 7

- Erasure-Coded Data Protection 7
- Reliability That Increases with Scale 8
- Continuous Data Scrubbing 8
- Extended File System Availability 8

Panasas PanFS Software Architecture 9

Panasas PanFS Director Node Software 10

- Scalable Metadata Services 11
- File Manager 11
- Storage Manager 11
- NFS and SMB Protocol Gateways 12
- Cluster Management Services 12
- Panasas PanActive Manager 12

Panasas PanFS Storage Node Software 12

- Object Storage Device File System 12
- Separation of Metadata and Data 12
- File-Level Versus Block-Level Metadata 13

Panasas PanFS Client System Software 14

- Panasas DirectFlow Parallel Data Protocol 14
- Network-Distributed Erasure Coding per File 14

Panasas ActiveStor Deployment Options 15

- Panasas ActiveStor Prime Family 15
- Panasas ActiveStor Ultra Family 15

Conclusion 16

Panasas ActiveStor

Enterprises that are working on some of the world's most innovative and complex high-performance computing (HPC) projects need to deliver insights and products fast. They need to improve their time to results and can't afford to wait as their storage systems try to keep up with intensive computational processing and time-sensitive data delivery.

We've architected the Panasas® ActiveStor® solution for simplicity and limitless performance. We've innovated our solution, based on 20 years of file system experience, to take advantage of the latest advances, including porting to Linux, offering a wider range of networking speed options, and leveraging the latest industry-standard hardware technologies.

With our fast and simple solution, you can forget about storage and focus on what really matters: building the next Dreamliner, winning a Formula One race, creating mind-bending visual effects, curing disease, or modeling climate change.

HIGH PERFORMANCE

Panasas ActiveStor Ultra high-performance storage uses a scale-out architecture that grows storage capacity, DRAM caching, and network bandwidth incrementally and linearly as you add more ActiveStor 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. And data flows directly from our storage nodes to the application without any hops through intermediate servers or even extra network links. ActiveStor Ultra stores metadata on ultra-low-latency NVMe media, small files onto cost-efficient SSDs, and large files onto high-bandwidth, low-cost HDDs. Placing each type of data onto the right type of device is one of the reasons we're so good at mixed workloads.

SURPRISING SIMPLICITY

The Panasas scale-out ActiveStor solution 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. It takes only a couple hours to install the ActiveStor solution. Just rack it up, plug it in, power it on, answer a few questions, and it's ready for anything you can throw at it.

Even in the largest Panasas deployments, all data resides within a single namespace, with a single management GUI and CLI, delivering data at very 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 will automatically rebalance capacity across the ActiveStor enclosures as you add them or if they become unbalanced; automatically reconstruct the full levels of erasure-coded data protection for all files in the event of any failures; and continuously scan all files in the background to scrub out any latent issues.

EXTREME FLEXIBILITY

The Panasas architecture scales data and metadata independently and is purpose-built for adaptability and flexibility to handle a wide range of use cases.

The Panasas ActiveStor solution automatically adapts to dynamically changing workloads and increasing demands. Its scale-out nature inherently spreads the workload, reducing the impact of hot spots, as well as simply growing capacity and performance.

ActiveStor product families offer flexible configuration options for each node to meet specific workflow needs, now and in the future. The solution supports mixing generations of ActiveStor products within a single namespace.

LOW COST TO OWN AND OPERATE

The number of people it takes to keep a storage system performing at its peak also affects its value. Part of our focus on HPC storage for the past two decades has been on reducing complexity, automating failure management, and supporting wide and widely mixed workloads without requiring tuning or retuning. It takes only part-time attention from a single person to manage the Panasas ActiveStor solution, no matter how large or high performance it is.

TIMELY, HIGH QUALITY SUPPORT

Unlike open-source solutions and even commercial alternatives from broad portfolio vendors, Panasas offers timely world-class L1-L4 support with just one phone call to resolve issues within minutes to hours rather than days and weeks.

- Single-vendor problem resolution, with centralized responsibility for addressing your issues
- Focused attention on your issues the moment you need it
- Support expertise from highly qualified technicians at your fingertips
- Rapid time to problem resolution

In HPC, high performance is a must. But both acquisition costs and ongoing management costs are also part of the value of a storage system. The ActiveStor product families are hybrid systems intelligently optimized for HPC workloads.

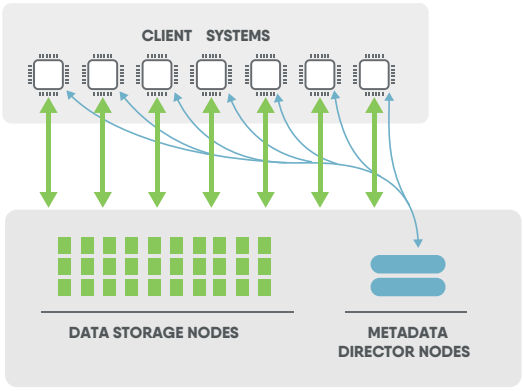
KEY PANASAS ARCHITECTURAL CONCEPTS	
INTELLIGENT DATA PLACEMENT	The Panasas architecture leverages three types of storage media—NVMe, flash and hard drives. File system metadata are stored on ultra-low latency NVMe media, small files are stored on flash, while large files are stored on hard drives. The streaming bandwidth of hard drives makes them ideal to store large sequential file data; and the high random IOPS rates of flash are a good match for storing small files maximizing overall price/performance.
SCALE-OUT PARALLEL FILE SYSTEM	You can easily and incrementally grow a single pool of Panasas storage with one global namespace, adding capacity and performance as needed while remaining online and accessible.
SEPARATION OF DATA AND METADATA	All the data and metadata in a Panasas storage system is stored in a highly scalable, massively parallel, object store. Panasas software coordinates the operation of the object store to provide local file system POSIX semantics via the Panasas DirectFlow® protocol plus support for standard NFS and SMB NAS protocols.
OBJECT SCALABILITY	All of the characteristics of operating on file system metadata are completely different than operating on bulk user data. Metadata items are typically very small and are usually modified independently of each other; it's an IOPs dominated workload. Data is typically large or very large compared to metadata and is most frequently read or written in sequential order within a file. It's a bandwidth-dominated workload. The Panasas solution separates the storage and management of metadata and data, with components dedicated to and optimized for each: director nodes and storage nodes.
PARALLEL AND DIRECT DATA FLOW	The Panasas solution stripes every file across different subsets of storage nodes in its architecture. Each client system directly, and in parallel, connects over the network to all the storage nodes that hold each file without having to transit through some other node in the storage system.
RELIABILITY THAT INCREASES WITH SCALE	Panasas uses software-based network-distributed erasure codes to separately protect individual files rather than using traditional RAID to protect whole raw storage drives. As each file is striped across storage nodes, erasure codes are generated and stored on other storage nodes. The time to rebuild a traditional RAID group is limited by the write bandwidth of the replacement drive, while the time to rebuild the missing part of all the affected files in a Panasas system onto other storage nodes (in parallel) is determined by the total write bandwidth of the storage cluster. This reduction in "exposure time" is crucial to increasing data reliability and explains how reliability increases with scale.

Panasas PanFS

The Panasas PanFS® parallel file system powers the Panasas ActiveStor high performance storage system. It is designed for extreme performance, reliability, and manageability. It combines the functions of a distributed and clustered file system, a scalable software-based erasure-coding data-protection engine, and intelligent management and failure recovery algorithms. The result is a high-performance storage system that serves up to hundreds of gigabytes of data per second from a single namespace and increases in reliability as it grows *(see figure 1).*

The Panasas DirectFlow parallel data protocol delivers the highest-performance access to the ActiveStor solution, while traditional Network File System (NFS) and Server Message Block (SMB) protocols can provide access to the same namespace via standards-defined protocols already built into every client system.

Figure 1. High scalability and efficiency



PanFS Parallel File System

YOU GET

Linear scalability of performance and capacity
Limitless performance as you scale-out
Reliability that improves with scale
Excellent mixed workload performance
Consistent user experience

YOU AVOID

Bottlenecks and hot spots
Tuning/re-tuning when workloads change
Noisy neighbor problems
Productivity impact from failures
Frustrated application users

DIRECTOR AND STORAGE NODES

The PanFS solution divides the labor of implementing a file system between director nodes that process metadata (information about files) and storage nodes that process data (files). Director nodes and storage nodes have a complex working relationship with each other. The unique value of the PanFS operating environment is that it reliably and transparently handles complexity at very high speed.

SCALE-OUT OBJECT STORE BACK END

All the data and metadata in the ActiveStor solution is stored in a highly scalable, massively parallel object store that the PanFS storage operating system manages. The PanFS platform coordinates the operation of the object store to provide all the same semantics and behavior as a traditional local file system.

Scalability occurs by growing the number of ActiveStor storage nodes in the object store, with each one adding to the capacity, performance, and network bandwidth of the solution.

You can configure the ratio of director to storage nodes and the number of each type of storage nodes to meet the demands of specific application workloads, making changes non-disruptively while the system is online.

POSIX ON OBJECTS

Object-based data layout is one of the key design principles in the Panasas architecture. Files in the PanFS environment are stored inside objects. Per-file erasure coding separately protects each file by breaking them up into a set of component objects instead of just one object per file.

The PanFS system maintains a map for each file that identifies the set of component objects, plus the striping parameters for that file, replicating the map across the component objects of each file. File system metadata (information about files) and directories are also stored inside objects and kept in the object store.

The DirectFlow protocol is the native high-performance method of access to the object store. Clients interact with the director nodes using the DirectFlow protocol to obtain access permissions and location information for the component objects for each file (e.g., the map). Clients then use the DirectFlow protocol to read and write component objects directly on multiple storage nodes in parallel.

During this process, the director nodes are manipulating the component objects in the object store to make it appear to the client systems that they're talking to a Portable Operating System Interface (POSIX)-compliant file system.

VOLUMES IN THE NAMESPACE

Storage administrators can easily create multiple volumes if they need different administrative regions of the PanFS global namespace.

Volumes are normal hierarchies of directories and files that share the common pool of storage capacity in the PanFS system and appear as top-level directories in the global PanFS namespace; but they separate the namespace into areas that can have different administrative controls. For example, you can define per-user capacity quotas that are unique to each volume and take snapshots per volume.

Multiple-user authentication schemes are available to control data access permissions for smooth integration into Linux, macOS, and Windows environments.

Each volume has a set of management processes associated with it that implements the behavior of the volume. The volume managers are distributed across the available director nodes, allowing overall metadata performance to scale out as needed.

STORAGE MEDIA

The Panasas architecture features intelligent data placement of data on three types of storage media: NVMe, flash-based solid-state drives (SSDs) and traditional hard disk drives (HDDs). PanFS stores file system metadata on NVMe, small files on SSDs, and large files on HDDs. The streaming bandwidth of HDDs makes them ideal to store large sequential file data; and the high random input/output operations per second (IOPS) of SSDs are a good match for storing small files and file system metadata, maximizing overall price/performance.

Panasas PanFS Data Protection

The PanFS system is now portable to industry- standard hardware platforms, making it cost- efficient in the hardware market, as well as up to date with the latest technologies.

The simplicity of managing the Panasas ActiveStor solution reduces the man-hours required to keep your storage humming; and you don't need to tune or retune the PanFS system to adapt to the changing environment. Typically, it requires only one administrator, no matter how large your Panasas ActiveStor solution is. And there will be time left over for other projects, too.

PANASAS PANFS DATA PROTECTION

There are multiple layers of advanced, scalable protection for your data in the PanFS architecture, resulting in a system that is both highly available and very tolerant of faults, while still providing high performance.

ERASURE-CODED DATA PROTECTION

As total storage capacity increases in a storage system, drive failures can become more frequent, simply because there are more drives in the system. Panasas addresses this challenge with advanced per-file distributed erasure-coding. An N+2 erasure code is applied as data in each file distributed across storage nodes in a cluster, correcting up to two simultaneous failures, whether it is an individual drive or a whole node.

Traditional hardware or software RAID architectures rebuild a whole physical drive onto another physical drive. Rebuilding a physical drive is limited by the bandwidth of writing to the replacement drive, and the entire replacement drive is written to, even if the drive that failed is not full.

The Panasas solution rebuilds only the files that are affected by the failure; and it rebuilds them onto free space distributed throughout the storage cluster, not onto a dedicated and otherwise idle "hot spare" drive. Because it uses the whole cluster to rebuild any affected files, it can recover from failures in a fraction of the time traditional RAID architectures require.

RELIABILITY THAT INCREASES WITH SCALE

Traditional storage products based on physical RAID groups, whether implemented in hardware or even in software, base their data protection strategies on using multiple groups of drives and an algorithm that can recover from one or possibly two drive failures within each group. Those groups are usually quite small, on the order of 12 to 24 drives, and usually fixed in size. Upon any failures, the RAID controller will pull an idle and unused drive from a hot spare pool and use an algorithm to fill that new drive with the same data the failed drive had.

There are several important and potentially performance-impairing consequences to this approach:

- "Rebuild" of a RAID group can only go as fast as the new drive is able to accept the rebuilt data.
- Write bandwidth is not growing at nearly the same rate as drive capacity, if at all.
- Probability of a drive failing in a RAID group that is recovering from an earlier failed drive is dependent upon how long it takes to rebuild the first drive. In other words, the longer it takes to rebuild, the higher the likelihood of a "dual failure" and data loss.
- You paid for a hot spare pool of drives that are not contributing to the performance or capacity you need.

Using the RAID architectural approach means overall data reliability decreases as the scale increases. The PanFS platform, on the other hand, is designed to deliver increasing data reliability as scale (and performance) increases. One of the most important factors behind this is parallel reconstruction. The PanFS system rebuilds files instead of drives. And because files are striped across differing subsets of all the storage nodes, you can rebuild all the files affected by a failure in parallel using the bandwidth of all the storage nodes at once.

All the director nodes in the system cooperate to simultaneously read the existing portions of each file from the storage nodes, reconstruct the missing portion, and write the reconstructed portion of each file to other storage nodes.

PanFS rebuild is not limited by the bandwidth of a single drive. Rebuild performance scales just as data access performance does; and as more director nodes are added to the system, reconstruction performance increases linearly.

The shorter rebuild time dramatically reduces the window of vulnerability where a file is not fully covered by all aspects of data protection in the PanFS solution.

CONTINUOUS DATA SCRUBBING

Each file is separately checked continuously in the background to ensure that all its data protection information is consistent. If the erasure code doesn't match the data, for example, the PanFS system will correct it via file-level reconstruction. Data scrubbing catches latent failures early so you can fix them before there is any risk of data loss.

EXTENDED FILE SYSTEM AVAILABILITY

In addition to the erasure-coded protection for file data, the PanFS solution provides another layer called extended file system availability (EFSA) for the namespace: the directory hierarchy and filenames.

In the extremely unlikely event of errors that the erasure coding cannot recover from, such as three or more storage nodes failing at the same time, the PanFS system can still tell which files have been affected and which have not and will fence off the affected files against inadvertent access.

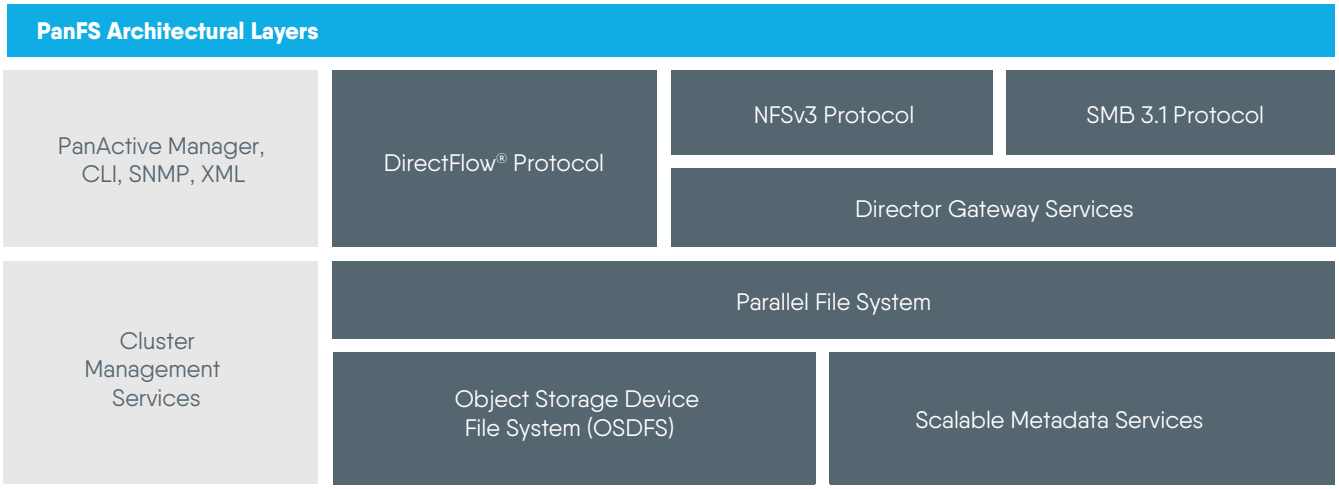
This additional layer of protection for the directory hierarchy and filenames can tell administrators the full path names of the affected files. This provides users with uninterrupted access to all files unaffected by the failure (typically the vast majority of user files), while the smaller number of affected files is logged for restoration from backup or some other source.

PanFS Software Architecture

The PanFS architecture *(see Figure 2)* is composed of the following functional blocks:

- **PanActive® Manager** – CLI, Simple Network Management Protocol (SNMP), and XML management interfaces
- **DirectFlow protocol** – Native high- performance, cache-coherent file access for Linux and macOS
- **NFS protocol** – Standard NFSv3 protocol for Linux, macOS, and others
- **SMB protocol** – Standard SMBv3.1 protocol for Windows and macOS
- **Director gateway services** – Support for access to files and data stored in the PanFS system via a protocol implementation layered above
- **Cluster management services** – Nodes that are currently healthy and in the "quorum" and those that aren't; recovery actions if nodes fail, configuration changes, etc.
- **Parallel file system** – Coordination of the actions of the object storage devices and metadata processing
- **Object Storage Device File System (OSDFS)** – Storage and retrieval of data from drives on behalf of client systems
- **Scalable metadata services** – Cache coherency and file metadata processing

Figure 2. Overcome the limitations of more traditional scale-out NAS solutions



Panasas PanFS Director Node Software

Director nodes make up the "control plane" of the Panasas architecture, managing metadata rather than storing user data. Each director node has a processor, DRAM, and several high-bandwidth Ethernet ports and runs a Panasas-developed software image that contains control and interface processes for many different aspects of the overall storage system.

Director nodes track the health and "quorum membership" of all director and storage nodes (whether they are alive and well and part of the Panasas cluster or not). They manage the namespace (filenames and the hierarchy of directories), the distribution and consistency of user data on storage nodes, and failure recovery actions such as data scrubbing, and rebuilds and host a GUI that treats the whole storage cluster as a single entity.

Director nodes also provide a "gateway" functionality, translating between the native Panasas DirectFlow protocol of the Panasas architecture and the standards-defined storage protocols NFS and SMB.

Director nodes do all of this without being in the data path. User data in files does not pass through a director node unless it is being translated to or from NFS or SMB.

SCALABLE METADATA SERVICES

PanFS metadata services implement all file system semantics. They control distributed file system 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.

POSIX mandates that each modification to the directory hierarchy or file metadata be atomic. Director nodes use a transaction log to ensure atomicity. Fault tolerance for the log is based on synchronously replicating each local transaction log to another director node, which is, in addition to its other duties, the designated "backup node" for this director node. The backup node relationships are assigned and reassigned automatically.

FILE MANAGER

The file manager is responsible for nearly everything having to do with the POSIX behavior of files and directories. It directly implements POSIX directories and filenames, using component objects to store them.

It grants and denies access to files based on the user's credentials. It retrieves the credentials for a user from identity managers, such as Lightweight Directory Access Protocol (LDAP) or Active Directory, and compares those credentials to the access control list (ACL) associated with the file being accessed.

It directly manipulates the component objects that make up files to emulate the behavior of POSIX files, creating empty component objects when a file is created, and deleting the component objects when the file is deleted, for example.

It manages cache coherency across the client systems by dynamically granting and revoking the right to access a byte-range region of a component object, separately controlling the right to read and write. It also controls whether each client can cache the regions it has been granted access to.

STORAGE MANAGER

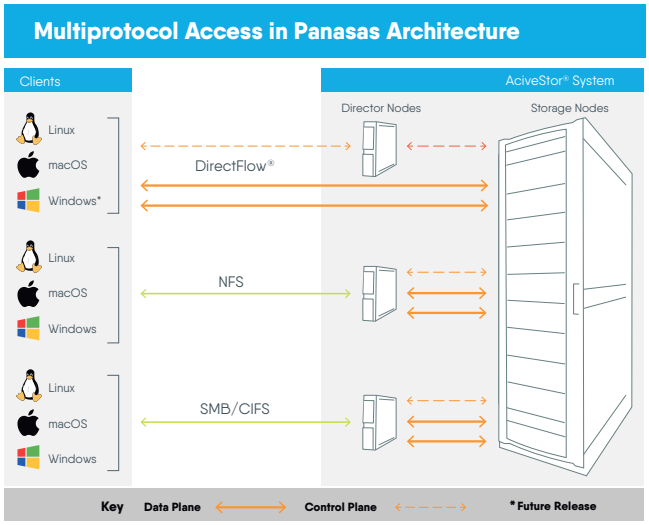
The storage manager is responsible for the management of component objects.

It decides which storage node should host each component object and maintains the "maps" that show which component objects are part of erasure code groups (i.e., files).

It will "reconstruct" all the missing component objects if a storage node fails. It will use all the maps to know which ones are missing and which ones to use as part of erasure code groups to regenerate the missing component objects. It will also decide which storage nodes will host the new component objects It is creating.

It ensures that all the storage nodes are roughly balanced in capacity and will move component objects around if it needs to in a process called "automatic capacity balancing."

Figure 3. Panasas PanFS operating environment provides scalable access for client systems via NFS or SMB protocols



NFS AND SMB PROTOCOL GATEWAYS

In addition to native DirectFlow access, the PanFS solution provides scalable access for client systems via the standard NFSv3 or SMB protocols. Running on director nodes as a "gateway" *(see Figure 3)*, this enables the PanFS system to integrate into heterogeneous IT environments consisting of a combination of Linux, macOS, and Windows clients.

Using the SMB protocol, you can easily manage files created by a Microsoft Windows, macOS, or Linux environment. A variety of options are available to manage user authentication, including Active Directory and LDAP. The PanFS solution provides administrators with the capability to map Windows security identifiers (SIDs) to Linux user IDs so storage quotas can apply simultaneously to a given user's Windows, macOS, and Linux accounts.

CLUSTER MANAGEMENT SERVICES

Every ActiveStor storage system is a cluster of storage and director nodes called a "realm." Each node in the realm runs a common PanFS cluster management service enhanced with additional services that provide hardware monitoring, configuration management for the node, and control of services running on the node.

You select an arbitrary subset of three or more director nodes in the realm to be part of the "repset"—director nodes that will host a replicated copy of the global configuration database.

The director nodes in the repset vote to elect one of their members as the "realm president." The president runs the master cluster management service for the realm. This process avoids so-called "split-brain" conditions commonly found in other scale-out storage architectures and is the reason it requires at least three director nodes in a realm for normal operation.

The realm president is responsible for modifications to the global system configuration for deciding which services are started on which nodes in the realm, detecting and responding to cluster node failures, and handling software upgrades.

PANASAS PANACTIVE MANAGER

Storage administrators interact with the PanFS system via Panasas PanActive Manager, an intuitive web-based GUI plus an industry-standard style CLI. PanActive Manager allows storage administrators to manage the cluster of ActiveStor enclosures as a single entity regardless of scale, providing a single point of storage management for the entire namespace.

PanActive Manager automates key workflows such as new storage discovery, load balancing to streamline performance, and enterprise data services such as reporting, snapshots, and user quota enforcement. Standards-based SNMP is also available for PanFS status and reporting.

Panasas PanFS Storage Node Software

Storage nodes hold all the user data and file system metadata of the Panasas system and make up the "data plane" of the architecture.

The storage nodes run a Panasas-developed software image called OSDFS. It interfaces to the rest of the Panasas architecture, including directly to the client systems, and stores the client's data and metadata on SSDs and HDDs.

The PanFS system stripes files across storage nodes, creating independent component objects for each stripe portion of each file and distributing them across differing subsets of the storage nodes.

OBJECT STORAGE DEVICE FILE SYSTEM

Panasas supports two different versions of OSDFS. Version 2 runs on the ActiveStor Prime family of products, and version 4 runs on the ActiveStor Ultra family of products. See the "Panasas ActiveStor Deployment Options" section for more information on the two families. Both versions of OSDFS fulfill

the same role in the Panasas architecture, but each is optimized for the hardware platform it runs on.

To deliver the highest performance, OSDFS includes advanced caching capabilities and intelligent object data placement on the available SSDs and HDDs.

In addition to a read cache to improve data retrieval, OSDFS caches newly written data in power-protected DRAM or nonvolatile dual in-line memory modules (NVDIMMs). Accumulating newly written data in larger sequential regions reduces data fragmentation, so later reads of the data will also be sequential.

OSDFS determines the best physical storage device within the storage node for each object based on the object's size. PanFS metadata and user files smaller than 60KB reside on the storage node's SSDs; and all user data larger than 60KB resides on the storage node's HDDs.

SEPARATION OF METADATA AND DATA

The Panasas solution extends the separation of metadata and data at the PanFS level in OSDFS version 4 for the same reasons.

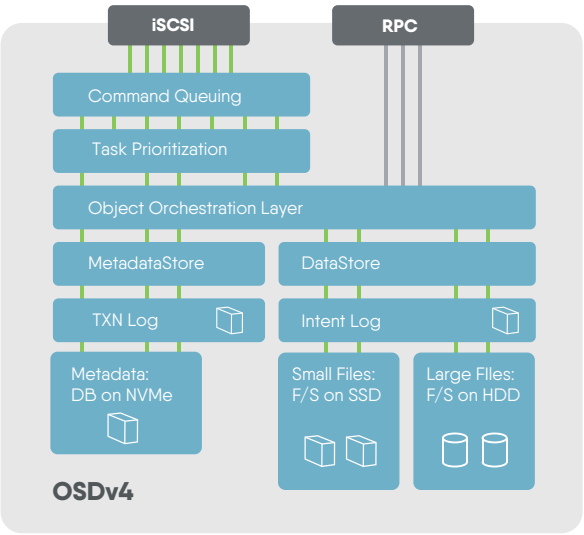
Essentially all the characteristics of operating on file system metadata are completely different than operating on bulk user data. Metadata items are typically very small and are usually modified independently of each other. It's an IOPs-dominated workload. Data is typically large or very large compared to metadata and is most frequently read or written in sequential order within a file. It's a bandwidth-dominated workload.

OSDFS version 4 internally has three different techniques to store information, each optimized for a different purpose. It stores metadata in a database optimized for small random reads/writes held on a nonvolatile memory express (NVMe) SSD. It stores the data for small files as regular files in a typical Linux local file system held on Serial ATA (SATA) SSDs, and it stores the bulk data for large files as regular files in a typical Linux local file system on SATA HDDs *(see Figure 4)*.

OSDFS version 2 is internally structured more like a typical local file system. It stores the metadata for an object as part of that object and separates all the metadata and the small files onto an SSD while storing large files on HDDs.

Figure 4. The right storage for each data type

3 TIER INTELLIGENT DATA PLACEMENT



Use the right kind of storage for each kind of data

- Transaction logos stored on MVDIMM
- Unmodified data/metadata cached in DRAM
- Metadata stored in database on low-latency NVMeSSD
  - DB may be used for Map/Reduce data analytics in the future
- Small files stored on cost-effective SATA HDD
- Large files stored on low-cost, high-B/W SATA HDD

Full data stability with fully async performance

- NVDIMM intent-log for both data & metadata operations
- Allows other drives to run full async for best performance
  - e.g. coalesce writes into contiguous runs for later read-back perf

FILE-LEVEL VERSUS BLOCK-LEVEL METADATA

The PanFS system maintains several different types of metadata about files, including the typical user-visible information such as the owner, size, and modification time. In addition, it maintains a per-file map of where the client system can find the set of component objects that make up each file.

Only the OSDFS software stack inside each storage node knows on which SSD or HDD each component object is held.

In traditional NAS systems, the management of “block-level metadata” often consumesa large portion of available processing power. By delegating low-level drive management and data placement optimization in a scale-out fashion to each storage node, PanFS director nodes have an erasure code algorithm is run over those pieces to generate data protection information for that region of the file. The protection information is treated as additional shards of the file and stored in additional component objects. All the component objects are written to their respective storage nodes independently and in parallel.

Panasas PanFS Client System Software

Standards-defined NAS protocols such as NFS and SMB are not capable of parallel transfers directly interacting with massively parallel scale-out object stores. Supporting NFS and SMB in the PanFS system serves a valuable purpose by connecting client systems that do not or cannot accept a custom protocol implementation. Installing the Panasas DirectFlow protocol on a client system (Linux and macOS) brings significant benefits.

PANASAS DIRECTFLOW PARALLEL DATA PROTOCOL

The Panasas DirectFlow parallel data access protocol avoids traditional performance

bottlenecks by allowing client systems to directly access the storage nodes hosting each file, without having to transit through some other node in the storage system, in parallel.

This results in higher performance than typical industry-standard protocols such as NFS and SMB while also adding support for cache coherency of file access across client systems that those protocols do not support.

The DirectFlow protocol implementation on a client system is delivered as a file system module that runs inside the kernel of the client's operating system. It implements a standard virtual file system (VFS) interface. Clients can access the PanFS solution as a standards-compliant POSIX file system, just like a local file system.

NETWORK-DISTRIBUTED ERASURE CODING PER FILE

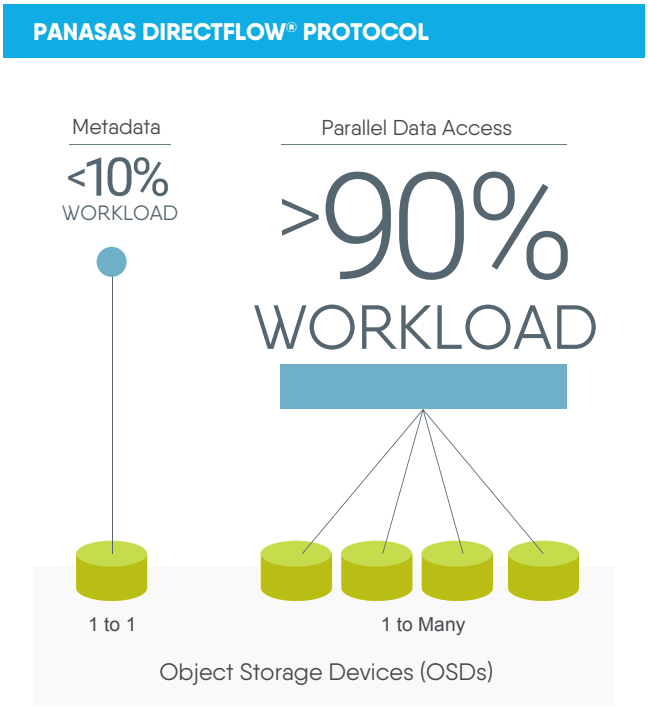
The PanFS solution generates an N+2 erasure-coded level of data protection for each file as it is distributed across the storage nodes in the cluster, correcting up to two simultaneous failures, whether

individual drives or whole nodes. DirectFlow clients "shard" files across storage nodes, creating component objects for each shard that makes up a file and distributing them across a subset of the storage nodes.

When file data is written, it is first divided into component objects via striping; then an erasure code algorithm is run over those pieces to generate data protection information for that region of the file. The protection information is treated as additional shards of the file and stored in additional component objects. All the component objects are written to their respective storage nodes independently and in parallel.

Similarly, clients read files by fetching the map of the component objects for a file for the director node, then reading those component objects directly from those storage nodes in parallel and combining them back into the desired file *(see Figure 5)*.

Figure 5. Native high-performance method of direct client access to object storage device data



Panasas ActiveStor Deployment Options

The Panasas PanFS distributed clustered file system runs on a choice of two different families of ActiveStor hardware platforms. The two families differ in the storage nodes they include but share the same director nodes. They are all forward and backward compatible with each other.

PANASAS ACTIVESTOR PRIME FAMILY

The Panasas ActiveStor Prime family includes ActiveStor director nodes and ActiveStor Prime storage nodes.

One to four ActiveStor director nodes fit into an industry-standard 2RU hardware enclosure. Each node has a powerful processor, significant amounts of DRAM, and a pair of high-speed redundant Ethernet ports to connect it to the client systems and the other director and storage nodes. It also includes NVDIMM technology to store power-protected data and metadata rather than using an uninterruptable power supply (UPS).

Eleven ActiveStor Prime storage nodes can fit into a 4RU hardware enclosure. Each node is a "hybrid" device with one flash-based SATA SSD to store metadata and small files, two SATA HDDs to store large files, a processor, and a pair of redundant Ethernet ports to connect to the client systems and the other director and storage nodes.

Every ActiveStor Prime enclosure includes a dedicated internal UPS that allows the software running on the storage nodes within that enclosure to treat their DRAM capacity as a power-protected cache of newly written data and metadata.

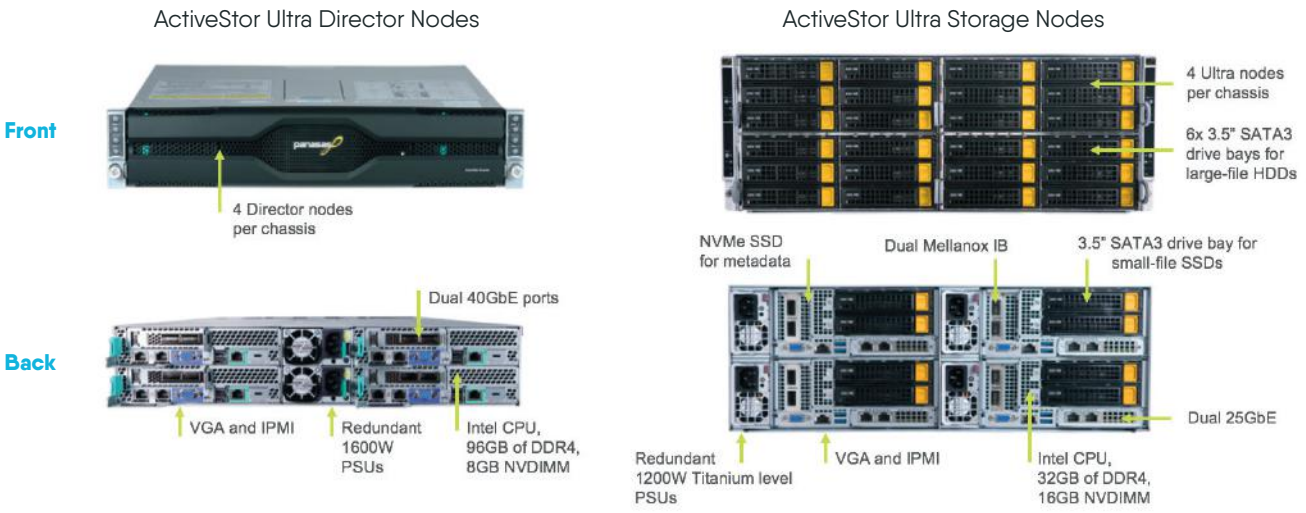
PANASAS ACTIVESTOR ULTRA FAMILY

Built on cost-effective industry standard hardware, The Panasas ActiveStor Ultra family includes ActiveStor director nodes and ActiveStor Ultra storage nodes *(see Figure 6)*.

One to four ActiveStor director nodes can fit into an industry-standard 2RU hardware enclosure. Each node has a powerful processor, significant amounts of DRAM, and a pair of high-speed redundant Ethernet ports to connect to the client systems and the other director and storage nodes. It also includes NVDIMM technology to store power-protected data and metadata rather than using a UPS.

Figure 6. ActiveStor Ultra hardware options

BUILT FOR PRICE/PERFORMANCE, ENGINEERED FOR MANAGEABILITY



ActiveStor Ultra storage nodes are higher capacity and higher bandwidth than ActiveStor Prime storage nodes. Four ActiveStor Ultra storage nodes fit into an industry standard 4RU hardware enclosure. Each is a "hybrid" device with a flash-based NVMe SSD to store metadata, zero to two flash-based SATA SSDs to store small files, six to eight SATA HDDs to store large files, a processor, and a pair of high-speed redundant Ethernet ports to connect to the client systems and the other director and storage nodes. It also includes NVDIMM technology to store power-protected data and metadata rather than using a UPS.

CONCLUSION

The Panasas ActiveStor architecture running the PanFS storage operating system breaks through the performance constraints of other parallel file systems. Our comprehensive and tightly integrated solution enables high-performance direct parallel access to petabytes of data while avoiding the stability problems inherent in legacy NAS systems as they grow.

FASTEST, MOST EFFICIENT PARALLEL FILE SYSTEM AT ANY PRICE POINT

The Panasas DirectFlow protocol can support your highest data-transfer requirements, accelerating time to results via clients doing direct parallel data I/O to multiple storage nodes.

CONSISTENTLY HIGH PERFORMANCE, REGARDLESS OF WORKLOAD

The I/O performance of the Panasas ActiveStor architecture remains consistent across large numbers of concurrently executing applications, even when they are reading and writing a mixture of large and small unstructured data sets.

BUILT FOR UNLIMITED LINEAR PERFORMANCE SCALABILITY

With the Panasas solution, you can easily and seamlessly scale the performance and capacity of your storage. Growing from 10 to 100, or even 1,000 ActiveStor enclosures will increase performance, capacity, and client access by the same factor with nearly perfect linear scaling.

UNPARALLELED RELIABILITY

Data reliability and availability increase as you expand the ActiveStor scale-out solution. The Panasas advanced per-file distributed erasure-coding software implementation detects and corrects problems when a drive or even a storage node fails.

SURPRISING SIMPLICITY

The Panasas ActiveStor solution is a fully integrated scale-out parallel file system with a single point of management regardless of scale. It includes enterprise data services you expect from a NAS system and is optimized for high-performance commercial market workloads.

By combining high performance with ease of use, Panasas reduces storage complexity in your data center by consolidating many different unstructured data workloads into a single scale-out ActiveStor solution.