nexenta®

Global Leader in Software-Defined Storage.

NexentaStor & NexentaFusion
Software-Defined Storage for Enterprise Applications

Product Guide

July 2018

Table of Contents

Preface

Copyright, Trademarks, and Compliance

Copyright © 2018 Nexenta Systems ™, ALL RIGHTS RESERVED

1 Solution Overview

NexentaStor is the industry's leading software-defined storage (SDS) platform, delivering full-featured, enterprise-class unified block and file storage services. As of version 5, NexentaStor is used by thousands of customers in configurations ranging from tens of terabytes to multiple petabytes as a replacement for legacy storage area network (SAN) and network attached storage (NAS) hardware appliances.

Unlike hardware appliances from traditional storage vendors, NexentaStor is delivered as software that delivers consistent functionality whether it is deployed on bare metal industry-standard x86 servers, in virtual machine on private cloud, or in public cloud. NexentaStor 5 incorporates a state of the art management framework exposing an exhaustive set of developer friendly, self-documenting REST APIs that are the foundation for all user management interfaces (CLI and GUI) and ecosystem plugins. NexentaFusion ships with NexentaStor 5 and provides a single pane of glass management interface for multi-system deployments.

Being delivered as software only gives customers new levels of flexibility and performance, eliminates hardware lock-in, facilitates migration to public cloud, significantly reduces storage costs, and generally simplifies storage management and scaling.

1.1 Introduction to NexentaStor SDS

NexentaStor SDS provides enterprise-class file (NFS and SMB) and block (FC and iSCSI) services. The software is deployed as a full storage operating system on standard x86 bare metal servers or in virtual machines on private clouds and hyper-converged infrastructure. The same software is also available as a service on public cloud through the NexentaCloud portfolio.

Figure 1-1: One Software Stack for Bare-Metal, HCI, Private and Public Clouds
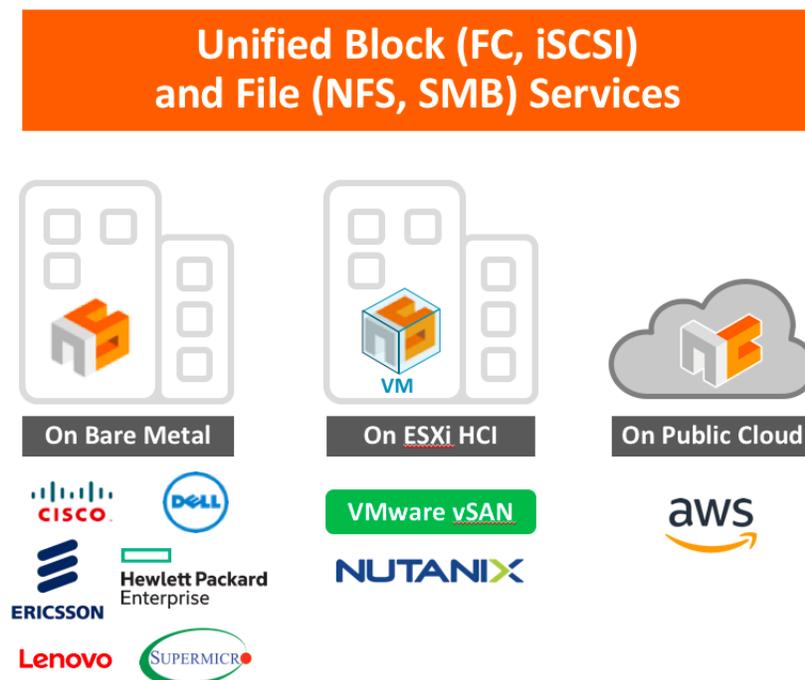
Having the ability to run the exact same storage software on this variety of environments allows customers to simplify and optimize their file services infrastructure. It is increasingly common to see customers run NexentaStor on a hyper-converged cluster in a remote office to provide file services to virtual machines and virtual desktop, configured to replicate data to a backup and archive bare metal configuration back in the main datacenter, or to a NexentaCloud instance in the public cloud.

NexentaStor builds on the open source OpenZFS file system to deliver high-performance, enterprise-class data services that effectively compete and win against legacy storage appliance vendors. Its mature NFS and SMB file services have been field proven in some of the most complex Enterprise environments.

Figure 1-2: NexentaStor 5 Feature Support

Protocols	File: NFSv3, NFSv4, SMB 1.0, SMB 2.1, SMB 3 Block: Fibre Channel, iSCSI
Configurations	Single node – x86 bare metal Dual-node High Availability clusters – x86 bare metal MetroHA Stretched clusters – x86 bare metal Virtual Storage Appliance – VMware ESXi, KVM NexentaCloud in Amazon Web Services
Data Management	All-Flash, Hybrid, or All-Disk pools RAID 10, N+1, N+2, N+3 OpenZFS end-to-end data integrity Unlimited snapshots & clones Unlimited file system size High performance inline data reduction Thin provisioning Storage Quality of Service Snapshot based Scheduled Replication Continuous Asynchronous Replication
Management	Self documenting REST API, CLI, SNMP NexentaFusion (HTML 5 based)
Client OS Support	VMware ESXi, Microsoft Windows, CentOS, RHEL, Ubuntu
Ecosystem Integration	Microsoft AD, LDAP, Kerberos SMB 3 ODX for Microsoft Hyper-V VMware VAAI Block VMware Virtual Volume (VVOL) 2.0 VMware vCenter Plugin VMware Site Recovery Manager OpenStack Cinder & Manila Docker Volume Plugin Kubernetes Persistent Volume

This mature set of features has allowed Nexenta to successfully sell into a wide variety of market verticals and use cases, from high-performance low-latency all-flash arrays supporting virtual environments and business-critical applications, all the way to large-scale cost-optimized backup and archive repositories.

The types of configurations and target use cases for NexentaStor are show below:

Figure 1-3: NexentaStor Use Cases
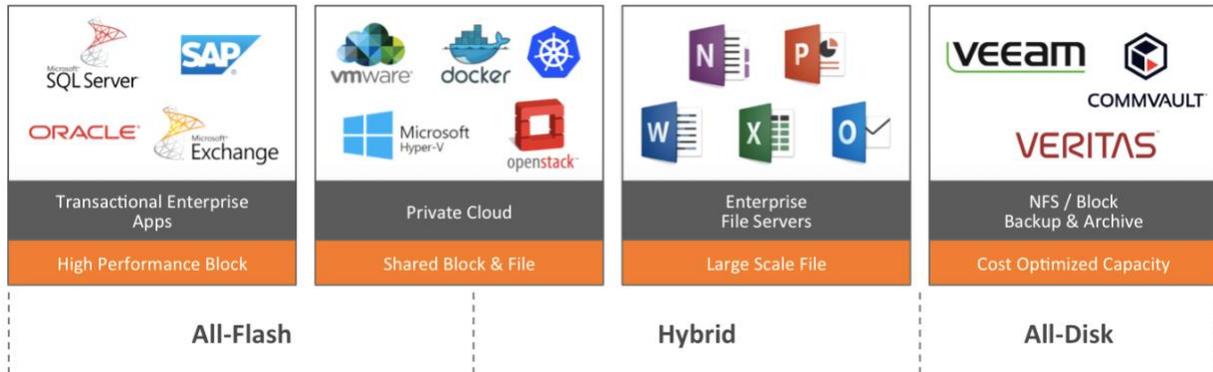
The ability to select the right hardware configuration for the right workload gives customers maximum flexibility. Having the same software stack running on high-performance all-flash, mixed hybrid and cost-optimized backup and archive setups simplifies data movement across tiers and management of the full storage infrastructure.

1.2 NexentaStor Software Architecture

NexentaStor 5 builds on the open source OpenZFS file system for its core data services and complements it with high performance block and file services, high-availability clustering, kernel based multi-site replication and the Nexenta Management Framework (NMF).

As shown in the figure below, NMF provides a tight cover on the underlying storage operating system and handles all provisioning and management operations. From a user perspective, NMF presents an exhaustive self-documenting REST API that provides a consistent foundation for the system's Command Line Interface (CLI), the NexentaFusion graphical user interface and all supported ecosystem plugins.

Figure 1-4: NexentaStor High Level Architecture
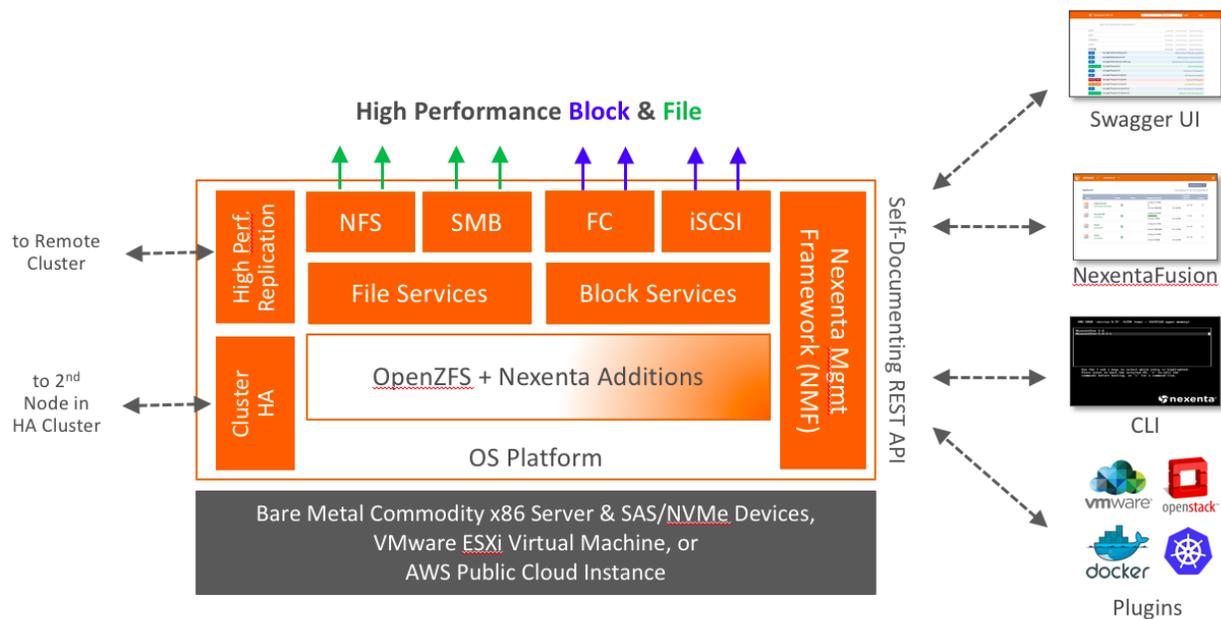
NexentaStor 5 benefits from all the scalability, performance, and reliability improvements in the core platform that were implemented over the last few years in NexentaStor 4. NexentaStor 5 also incorporates a number of new and enhanced device drivers, adding support for 25Gb/s and 40 Gb/s Ethernet interfaces, as well as 32 Gb/s Fibre Channel HBAs.

The Nexenta Management Framework was first introduced with NexentaStor 5.0 and fully replaced the legacy Nexenta Management Server (NMS), Nexenta Management Console (NMC), and Nexenta Management View (NMV) that were part of NexentaStor 4.x and all previous releases.

By design, NMF is a high-performance, multi-threaded, fault-tolerant management plane that provides a streamlined and simplified storage-centric management experience. Building the NexentaStor CLI as well as all management products on top of the NMF REST API ensures that all operations and all statistics available through the CLI and GUI are supported via the REST API.

The NexentaStor REST API is self-documenting and provides an interactive Swagger user interface for developers and technology partners looking to build storage automation and orchestration.

Figure 1-5: NexentaStor REST API Management Screen
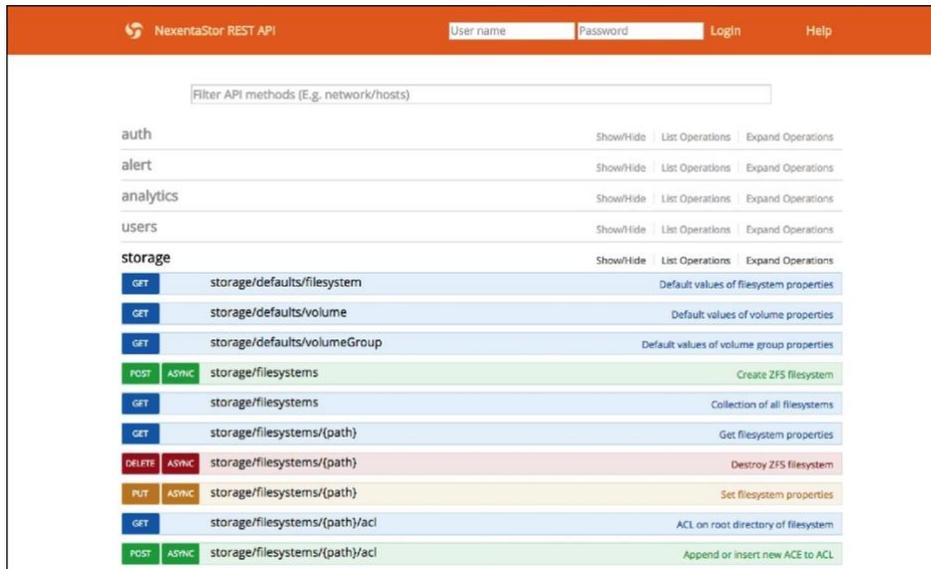
Day to day management operations are accomplished using the NexentaFusion management server, freely available with any NexentaStor 5 license. NexentaFusion complements any NexentaStor 5 deployment and provides a single pane of glass Graphical User Interface for multiple NexentaStor appliances. NexentaFusion is available as a VMware OVA and as a Docker container for simple deployment on any Linux host, physical or virtual. It incorporates intuitive workflows for all provisioning and fault management tasks and delivers advanced performance analytics dashboards for detailed monitoring and troubleshooting.

Figure 1-6: NexentaFusion Management Screen
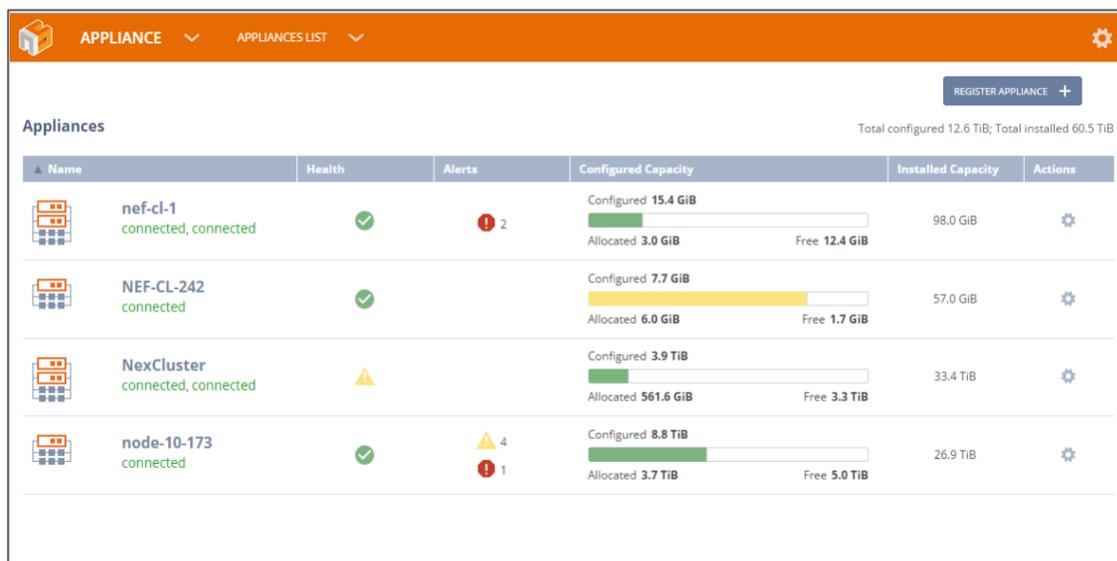
1.3 NexentaStor High Availability

NexentaStor high availability (HA) allows you to configure two NexentaStor nodes to provide redundant access to storage pools. Active-passive (two NexentaStor nodes, one pool) and active-active (two nodes, two pools or more) configurations are supported. If a node is unable to deliver storage services for some reason, the other node automatically takes over.

Technically, a NexentaStor HA configuration consists of an *HA cluster* where one or more *HA services* are running:

- An HA cluster is a pair of NexentaStor nodes that share access to backend storage devices, have network connectivity to each other, and on which the high availability feature is enabled.

- An HA service specifies the Virtual IP (VIP) address(es) or Fibre Channel ports for the HA cluster, the storage pool(s) to be protected, and which of the nodes initially provides clients access to the storage.

To facilitate HA configuration for file and iSCSI services, each node should have at least one static IP address assigned to it, and each VIP should have its own static IP address.

Figure 1-7: Example NexentaStor High Availability Configuration
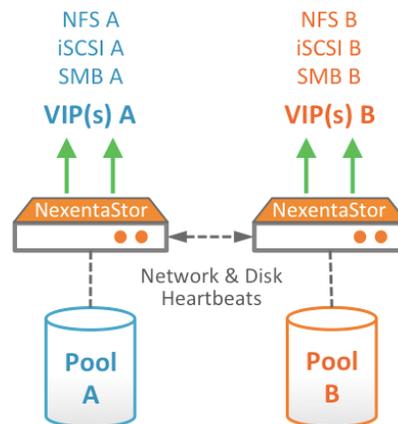
In this example, Node-A and Node-B comprise an HA cluster. There are two storage pools, Pool-A and Pool-B, that consist of backend storage devices that are accessible to both nodes in the HA cluster. There are two HA services in this configuration: one that provides access to Pool-A using VIP-A and another that provides access to Pool-B using VIP-B. In this example, Node-A serves as the backup for Pool-B, and Node-B serves as the backup for Pool-A.

Node-A and Node-B continuously exchange status information using a number of different network and storage heartbeat mechanisms. If all of the heartbeat mechanisms are disrupted for a given length of time, the main node is considered to have failed and the backup node automatically takes over its storage and service resources. In this example, failure of Node-A would lead to Node-B taking over Pool-A and VIP-A, and failure of Node-B would lead to Node-A taking over Pool-B and VIP-B. During an automated failover operation, the failed node is automatically fenced off from accessing the backend pool to protect the data from unintended write operations.

1.4 NexentaStor Reference Architecture and Appliances

NexentaStor deployments on bare metal hardware span from single node configurations on SATA or NVMe devices to dual node clusters of industry-standard x86 servers with SAS connected shared storage enclosures.

Figure 1-8: NexentaStor Hardware Appliances Overview
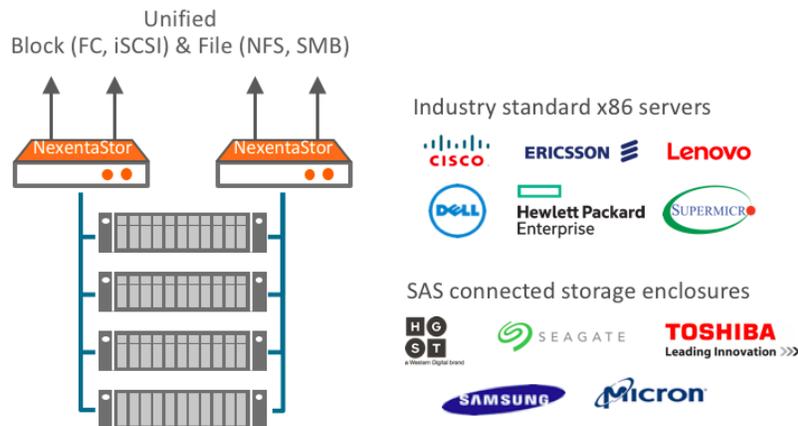
While the maturity, flexibility and cost-effectiveness of NexentaStor software-defined storage make it attractive for a broad set of Enterprise use cases, meeting the needs of Enterprise customers also requires specific attention to deployment and supportability. This is why Nexenta works closely with hardware partners and server OEMs to certify full stack Reference Architectures (RAs) that address all-flash, hybrid and all-disk storage requirements and scale from tens of terabytes to multi-petabyte appliances.

Figure 1-9: Reference Architecture Examples
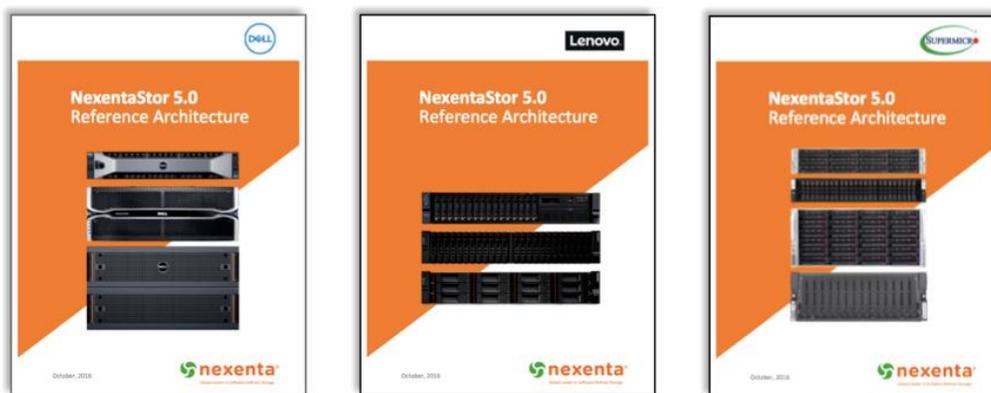
These end-to-end solutions are delivered through server partners such as Cisco, Dell EMC, Ericsson, HPE, Lenovo or Supermicro with many providing full stack support, taking first call on both software and hardware and providing a seamless support experience to NexentaStor users.

In addition to reference architectures, a select set of server OEMs offer NexentaStor based hardware appliances where the software is pre-loaded and pre-configured in manufacturing providing a turn-key hardware appliance experience to end customers.

Certifying a reference architecture entails the following:

- Working with the hardware partner to create Nexenta skus for the specific hardware configurations, explicitly specifying CPU, DRAM, HBA and NIC components and firmware, to deliver the right performance and capacity scaling

- Performing extensive hardware and software interoperability testing to ensure that the solution will operate and perform as expected in production environments

- Developing hardware specific chassis management plugins for NexentaStor and NexentaFusion to simplify day-to-day operations related to hardware management, device replacement, etc.

Figure 1-10: NexentaFusion Chassis Management for Supermicro Appliance

For more details on NexentaStor 5 reference architectures and appliances, please see the NexentaStor 5 Hardware Compatibility List.

1.5 NexentaStor Licensing

NexentaStor 5 is available as an Enterprise Edition or Community Edition:

- *NexentaStor Enterprise Edition* is sold as a perpetual software license based on raw capacity under management (measured in TB). Support and services are sold separately. Pricing is tiered on the amount of raw capacity required for a system, with a larger configuration getting a lower price per GB than a smaller one.

  The Enterprise Edition includes all core storage functionality such as snapshots, clones, inline data reduction, software RAID and snapshot based scheduled replication. It also includes the right to use NexentaFusion as the Graphical User Interface for the system running that license. Additional options are sold on a per-node basis and include features such as High-Availability Cluster, MetroHA, Continuous Replication or Fibre Channel support.

- *NexentaStor Community Edition* is a limited-functionality, limited-capacity, free version of the software that can be used for non-production deployments. Outside of the online Nexenta Community forums, no support services are available for NexentaStor Community Edition. A NexentaFusion server can manage at most one NexentaStor Community Edition appliance. For more details, see the online Nexenta Community forums.

Figure 1-11: NexentaStor Software Licenses

License Type	Time Limit	Capacity Limit	Separately Licensed Options
Enterprise Edition	None	Variable	- High Availability Cluster - HPR Continuous Replication - Fibre Channel Target - MetroHA
Trial	45 days	Variable	Trial period includes Enterprise Edition optional features
Community Edition	None	10TB Allocated	None

NexentaStor 5 implements a simple licensing engine that seamlessly integrates with the customer license portals on Nexenta.com. Activating a license on a NexentaStor 5 appliance (single node or HA cluster) is as simple as entering the activation token provided by the customer license portal and letting the system automatically register, download a license file and activate the license.

The following features can be added to an Enterprise Edition license:

- High-Availability Cluster – Configures a pair of physical servers to run as a highly available NexentaStor appliance, providing redundant access to shared storage pools. If one of the NexentaStor nodes should fail, all storage services are seamlessly failed over to the other node. See section 1.3 for more details.

- Continuous Replication – The Enterprise Edition license includes snapshot based scheduled replication services to a remote appliance, with a minimum Recovery Point Objective (RPO) of 15 minutes. The Continuous Replication option actually unlocks 2 advanced replication features: it allows snapshot based scheduled replication as often as every minute; and it unlocks the continuous asynchronous replication functionality that delivers close to zero Recovery Point Objective (RPO) over any distance, without affecting application performance. See section 4.6 for more details.

- Fibre Channel functionality – Allows you to configure FC targets on the appliance and serve LUNs over Fibre Channel.

- MetroHA – Enables NexentaStor to deliver zero RPO continuous availability disaster recovery between sites in the same metro area. See section 2.7 for more details.

Contact Nexenta Systems for information on purchasing a license for the software and individual features.

1.6 Benefits of NexentaStor Software-Defined Storage

With its performance, enterprise-grade reliability and feature set, NexentaStor is unique in the market in its ability to bring the flexibility, agility and cost saving benefits of Software-Defined Storage to traditional enterprise storage environments.

More specifically, NexentaStor delivers the following benefits:

- **Ultimate flexibility.** NexentaStor software can be deployed on a variety on hardware reference architectures. It can also be deployed as virtual machines on private cloud and hyper-converged infrastructure to complement these platforms with enterprise-class file services. And through the NexentaCloud portfolio, the exact same feature set and management interfaces are available in public cloud for seamless hybrid cloud storage management.

- **No compromise**. NexentaStor includes all the advanced storage functionality that can be expected of high-end, enterprise-class storage arrays supporting unified block and file services. The solution also supports highly differentiated features such as high-performance inline data reduction for all-flash configurations, triple parity RAID for large archive setups, continuous replication for near zero RPO disaster recovery scenarios, and unlimited file system sizes.

- **Great performance**. NexentaStor reference architectures make use of the latest hardware components and generally take advantage of larger amounts of DRAM in the NexentaStor nodes than is typically found in traditional storage appliances. The combination of high-performance software and higher-performance hardware components in the NexentaStor reference architectures translates into solutions that generally outperform the competition.

- **Total control**. Customers running NexentaStor solutions effectively are in full control of what to deploy for their storage needs and when to refresh their storage infrastructure. Rather than having a traditional storage appliance vendor dictate when a particular array needs to be replaced, NexentaStor customers get to choose when to refresh the standard x86 server components supporting their storage infrastructure. Having the option to run a NexentaStor appliance for five or even seven years on the same hardware can translate into significant cost savings over the life of a system.

- **Simplified management and scaling**. NexentaStor 5 implements a developer-friendly, self-documented REST API for ease of automation and orchestration. For more traditional operational environments, NexentaFusion delivers a modern and intuitive graphical user interface (GUI) for all provisioning and monitoring workflows.

- **Lower total cost of ownership**. By running on standard x86 servers, NexentaStor allows customers to acquire high-performance, enterprise-grade storage solutions at industry-standard x86 server prices, avoiding the typical high margin charged by traditional storage appliance vendors. Furthermore, NexentaStor's perpetual software licenses allow customers to simply repurpose licenses and save over the long run as hardware refreshes only incur hardware costs, not software costs.

1.7 Where to Find More Information

| Document	Purpose
Quick Start Guides	These documents provide basic information about essential functionality of NexentaStor and concise procedures to get you up and running as quickly as possible.
	The quick start guides cover the following topics:
	• NexentaStor and NexentaFusion installation
	• NexentaStor CLI Configuration
	• NexentaStor CLI Reference Guide
	• NexentaStor High Availability
	• NexentaStor High Performance Replication
	• NexentaStor REST API
	• NexentaStor Data-at-Rest Encryption
	• NexentaStor vCenter plugin
	• NexentaStor VVOL Administration
NexentaStor CLI manual pages	Enter `<command> -h` for high level information on sub-commands and options and `man <command>` to get detailed usage instructions
NexentaFusion documentation	NexentaFusion's online help provides step-by-step procedures for common configuration and monitoring tasks.
	Additional documentation:
	• NexentaFusion installation
	• NexentaFusion User Guide
Release notes	High-level overview of new functionality in a given release and detailed review of resolved and known issues.

The NexentaStor quick start guides and NexentaFusion documentation are available online in PDF and HTML format from nexenta.com at: https://nexenta.com/products/documentation

2 Deployment Scenarios & Use Cases

2.1 Use Cases

NexentaStor software is successfully used by thousands of customers over a wide range of use cases, from high performance transactional databases to private cloud and cost-optimized very large-scale backup and archive. The ability to use the same storage software, with the same rich feature set and management interfaces and APIs across all these workloads enables customers to massively simplify their storage infrastructure.

Figure 2-1: Typical NexentaStor Use Cases
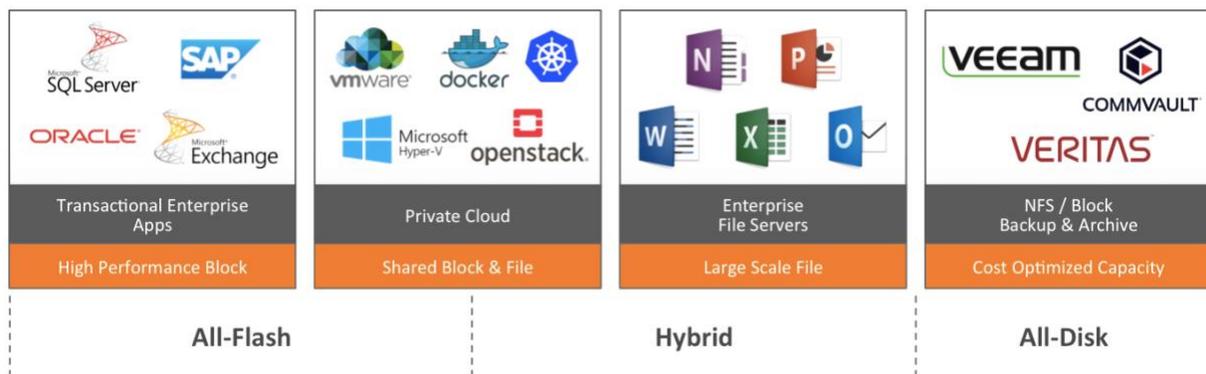
All-flash configurations are ideal for high performance applications, databases and cloud hosting use cases. NexentaStor on all-SSD configurations provides high performance, low latency storage services. All-flash NexentaStor reference architecture can hit very low $/GB price points that make all-flash storage cost-effective for a larger share of applications. Combining the low capex of software-defined on all-flash (built on industry standard x86 servers and SSDs) with the low opex benefits of all-flash storage in general (easier to run, more reliable, less power, less space) has allowed a number of Nexenta customers to fully standardize on all-flash systems, deploying multiple petabytes of all-flash NexentaStor for both block and file services.

Hybrid configurations combine a few SSDs for write logging and read caching with NL-SAS HDDs for long-term data storage. These configurations are popular for private cloud and the wide variety of applications requiring unstructured data storage, NAS services and full-featured NFS and SMB services. These configurations often scale from hundreds of terabytes to petabytes with single or dual parity raid to optimally balance cost and performance requirements.

All-Disk configurations are ideal for low cost backup and archive use cases. Depending on the backup software used, customers leverage NexentaStor Fibre Channel, iSCSI, NFS or SMB services. In this use cases, systems scale to multi-petabytes, with large 10TB or 12TB HDDs and triple parity raid. To further simplify data management, few very large file systems are generally configured, from hundreds of TB to single digit PB scale. NexentaStor built-in OpenZFS data integrity protection mechanisms ensure that faithful long-term data retention and the software's advanced remote replication functionality can be used for multi-site disaster recovery configurations.

2.2 Deployment Scenarios

A NexentaStor 5 appliance by itself can be managed via CLI or REST APIs. To manage the appliance via graphical user interface, it must be registered on a NexentaFusion server, a VMware vCenter plugin, or both.

Figure 2-2: NexentaStor 5 Deployment Scenarios
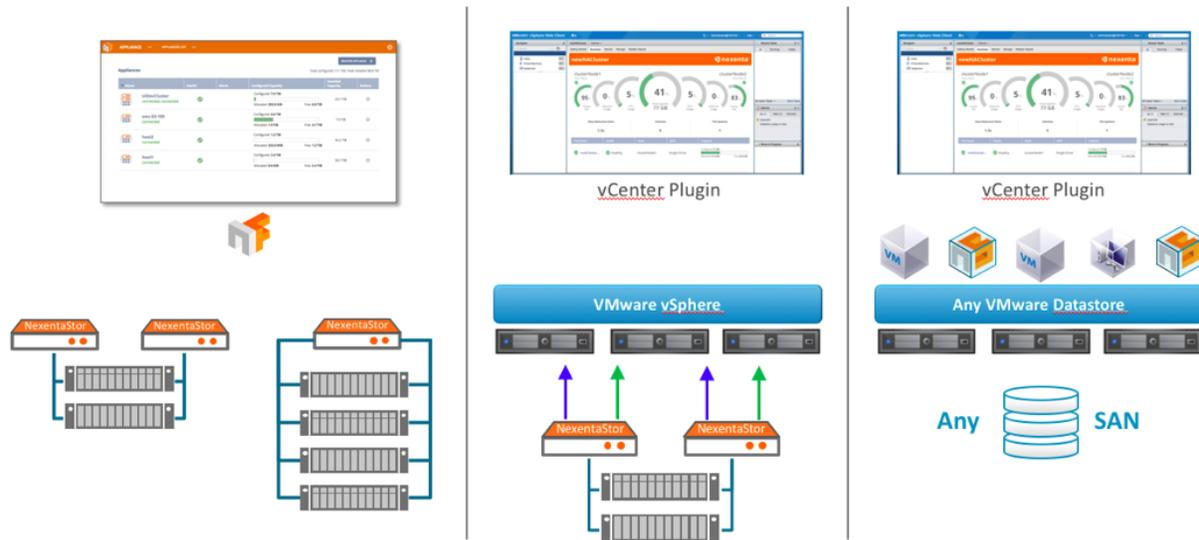
A typical customer deployment includes one or more NexentaStor appliances managed by a single NexentaFusion server, providing a single-pane-of-glass GUI. Note that the NexentaFusion server must be deployed on a separate Linux host or Linux virtual machine.

For VMware environments, NexentaStor 5 bare metal appliances can be used to provide high-performance FC, iSCSI, or NFS datastores to vSphere. In this scenario, the Nexenta vCenter Plugin can be used by VMware admins to streamline datastore provisioning tasks, manage capacity, configure snapshot schedules and replication services, etc. The vCenter Plugin can be used as an alternative, or as a complement to a NexentaFusion server.

Last but not least, NexentaStor 5 can also be deployed as virtual machines to provide file services and complement block only SANs. This is a popular solution for customers deploying transitioning to block-only All-Flash SANs looking for a cost effective and easy to manage NAS replacement solution. The same approach can also be used in cloud environments looking to provide virtual NAS appliances for different tenants, with each tenant having delegated management of its own virtual appliance.

2.3 File Services for Hyper-Converged Infrastructure

An increasingly popular deployment scenario for NexentaStor 5 is to provide full featured enterprise class services on ESXi based hyper-converged infrastructure such as clusters running VMware Virtual SAN or Nutanix. In this scenario, customers deploy NexentaStor 5 in virtual machines providing full featured NFS and SMB file services, with snapshots, inline data reduction and long-distance replication to any other NexentaStor instance (physical, virtual or in public cloud).

Figure 2-3: NexentaStor 5 on VMware Hyper-Converged Infrastructure
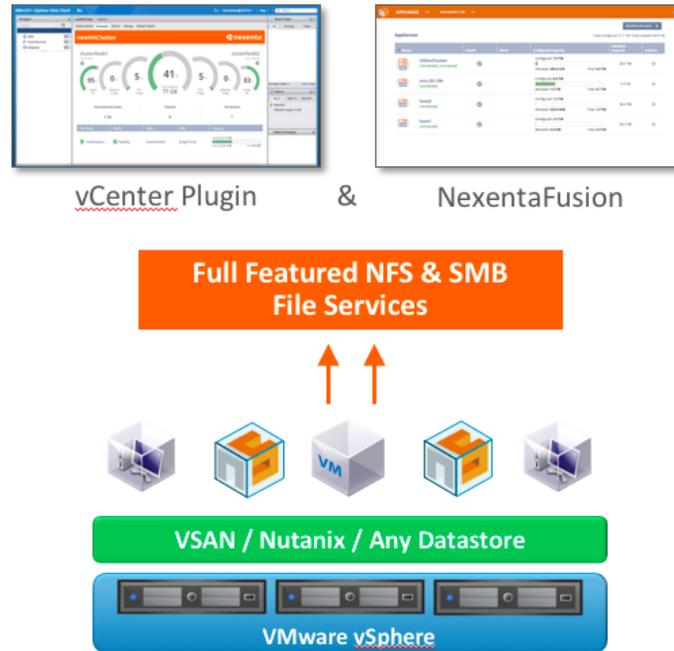
NexentaStor storage pools are built on vmdks directly supported by the underlying HCI cluster, benefiting from the durability provided by the HCI solution. For environments that require the highest level of file services availability, NexentaStor clustering services can be enabled between NexentaStor virtual machines to provide fast services failover capability in case of hardware or software faults. The solution leverages special integration with vCenter to automatically control the state of the NexentaStor virtual machines and protect file system data in the event of split brain scenarios.

2.4 VMware Integration

Ecosystem integration is an important benefit of the NexentaStor portfolio. NexentaStor 5 supports VAAI for block services, and comes with a new VMware vCenter Plugin that enables VMware administrators to perform advanced storage management and storage operation tasks directly from vCenter.

Figure 2-4: vCenter Plugin Appliance Dashboard
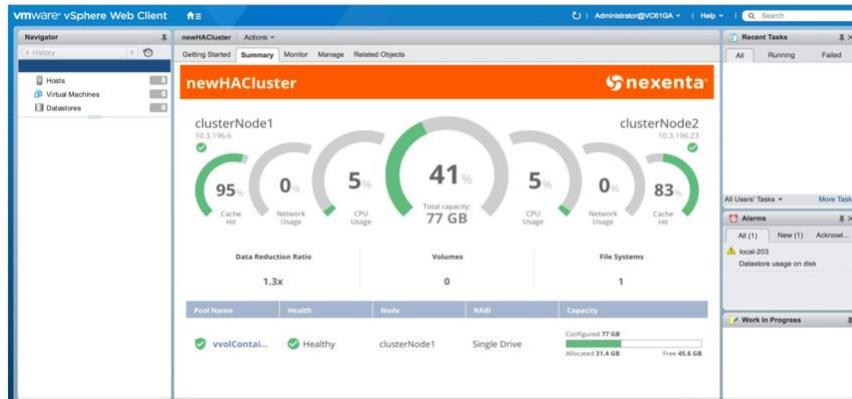
NexentaStor 5 also supports VMware Virtual Volumes (VVOL) over NFS protocol endpoints for scalability. NexentaStor 5.1 also adds support for VMware Site Recovery Manager with NexentaStor 5 High Performance Replication.

2.5 NexentaCloud in Amazon Web Services

NexentaCloud allows customers to extend their NexentaStor file services from on-premises to public cloud while keeping 100% consistency of features and functionality, managed and operated via tried and trusted NexentaFusion.

Figure 2-5: NexentaCloud in AWS Overview

www.nexenta.com
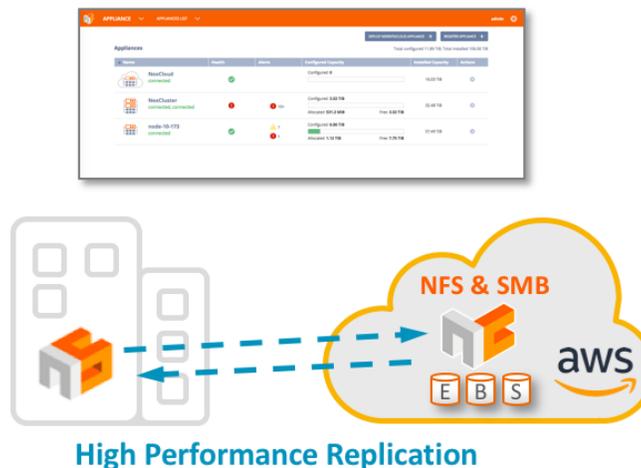
An instance of NexentaCloud in AWS uses Amazon Elastic Block Store (EBS) devices for raw storage pool capacity, runs the same software as NexentaStor and is managed using the same CLI, API or NexentaFusion. The combination of NexentaStor on premises and NexentaCloud in AWS provides an ultra-flexible solution for Test/Dev in public cloud or Disaster Recovery to public cloud using High-Performance Replication to replicate datasets between on-premises and AWS. As such, NexentaCloud delivers full featured file services with enterprise class AD / LDAP support for seamless hybrid cloud deployments.

NexentaFusion, deployed on-premises or in AWS, provides automated deployment of NexentaCloud instances from the main dashboard. NexentaCloud OneClick enables a NexentaFusion admin to provide AWS credentials, select a desired NexentaCloud performance type and capacity and let NexentaFusion automatically take care of provisioning the instance in AWS, attaching the right amount of Elastic Block Storage devices to it, and registering the appliance on the NexentaFusion dashboard. From then on, the NexentaCloud instance is managed just like any other NexentaStor appliance: create a pool and file systems, configure NFS and SMB services, set data protection and data replication services, etc.

Figure 2-6: NexentaFusion Powered NexentaCloud Deployment
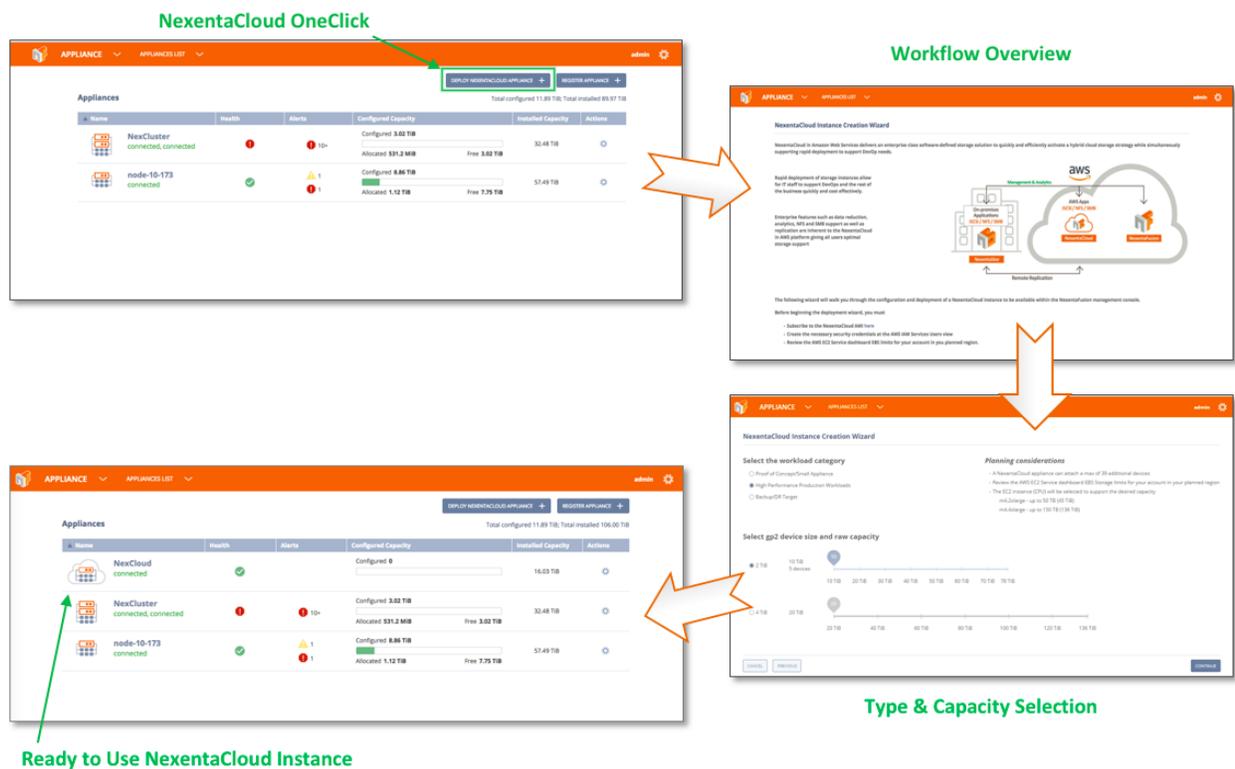
You can try NexentaCloud any time by accessing the NexentaFusion instance available on the AWS Marketplace here.

2.6 OpenStack Integration

Nexenta has been an early and active member of the OpenStack storage community, building its first set of Cinder drivers for the OpenStack Essex release, all the way back in April 2012. As the OpenStack platform has matured, so has NexentaStor integration with NexentaStor 5 supporting Cinder over iSCSI and NFS protocols, as well as Manila storage services.

Figure 2-7: OpenStack Cinder over iSCSI or NFS
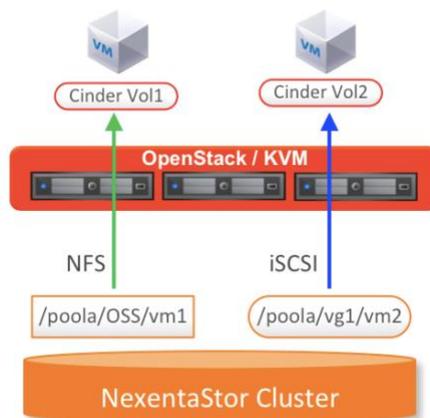
More detailed information on OpenStack support, latest Cinder and Manila drivers for a host of releases, as well as best practice guidelines are available on nexenta.github.io/openstack.

2.7 Stretch Clusters and MetroHA Configurations

For applications that require continuous operations, NexentaStor MetroHA delivers continuous availability, synchronous mirroring and zero RPO disaster recovery. The solution can be deployed between sites connected via a stretched Fibre Channel SAN on the same campus or in the same metro area, over distances up to 50 miles / 80 km.

Functionally, NexentaStor MetroHA stretches a NexentaStor HA cluster across 2 sites, with one NexentaStor head node in each site. SAS backend storage enclosures in both sites are connected using high performance ATTO Technology XstreamCORE FC 75xx Controllers to a shared stretched Fibre Channel Fabric. The NexentaStor software manages this Fibre Channel backend storage and synchronously mirrors data across both sites to ensure zero data loss in the event of a device, node or site failure.

The solution provides automatic failover for clear-cut failure scenarios and operator driven site failover for site wide disaster situations.

Figure 2-8: MetroHA with Four ATTO XstreamCore Controllers
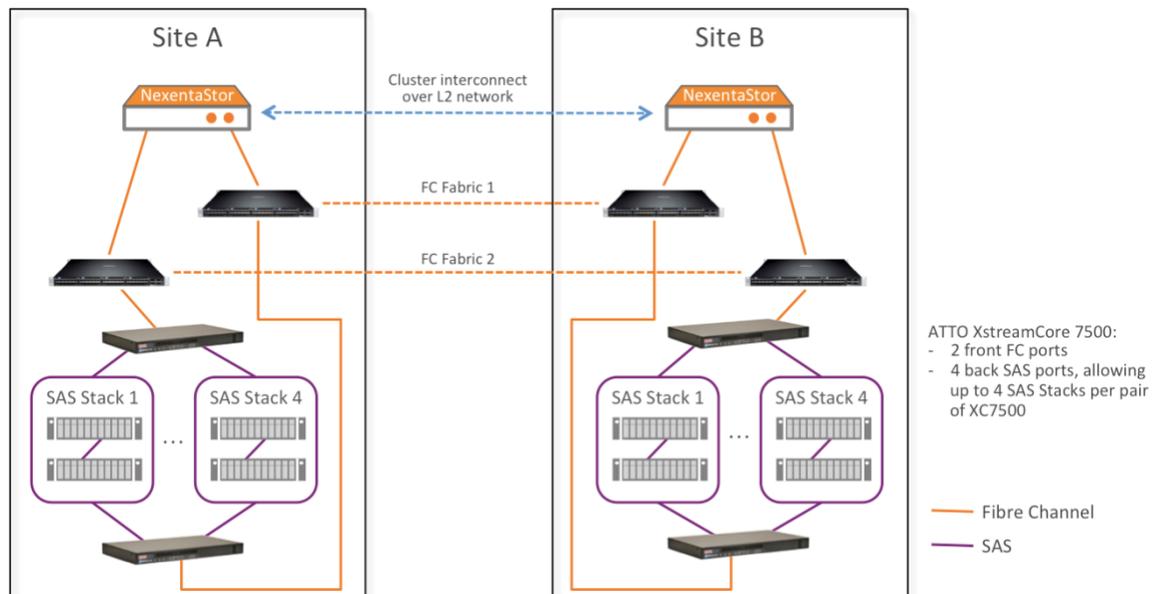
3 Unified Block and File Services

3.1 Storage Pooling Architecture

NexentaStor SDS delivers true unified block and file storage services. Physical capacity is organized in logical pools of storage that can support both file systems and block volumes concurrently. A single NexentaStor node with a single backend pool can thus concurrently deliver block and file services to client applications.

NexentaStor 5 builds on OpenZFS for its core data layer. In a typical NexentaStor HA cluster, the physical capacity will be organized in one or two storage pools. Each storage pool is a collection of virtual devices (vdevs), which are themselves made up of physical devices that are connected to the NexentaStor nodes. NexentaStor builds on OpenZFS to protect data against silent data corruption, providing strong data integrity, automated self-healing, and advanced software-based data protection.

Selecting the right pool configuration boils down to balancing the following 3 items:

- Resilience to device failure(s): NexentaStor vdevs can be configured as mirrors, or with single parity, dual parity or triple parity protection. A pool with mirror vdevs can sustain a single device failure per vdev. It is however exposed to data loss in the event of a double-failure in the same vdev. Larger physical devices translate into higher risks of hitting a double failure in a given vdev. This is one reason why Nexenta best practices use dual parity configurations or higher for pools of large HDDs.

- Usable capacity: a pool with mirror vdevs will use 50% of raw capacity in parity overhead while a pool with (4+2) dual parity vdevs will only spend 30% of raw capacity on parity overhead. Given the same raw capacity, the (4+2) pool will provide more usable capacity.

- Performance: ignoring the effects of ZIL, SLOG, ARC and L2ARC, all data stored in a pool is automatically stripped across the set of vdevs that make up the pool. More vdevs in a pool generally translate into higher pool performance, particularly for random IO workloads. Put differently, given the same number of physical devices, a pool with mirror vdevs will generally deliver more IOPS than a pool built with (4+2) vdevs.

Figure 3-1: NexentaStor Storage Pools
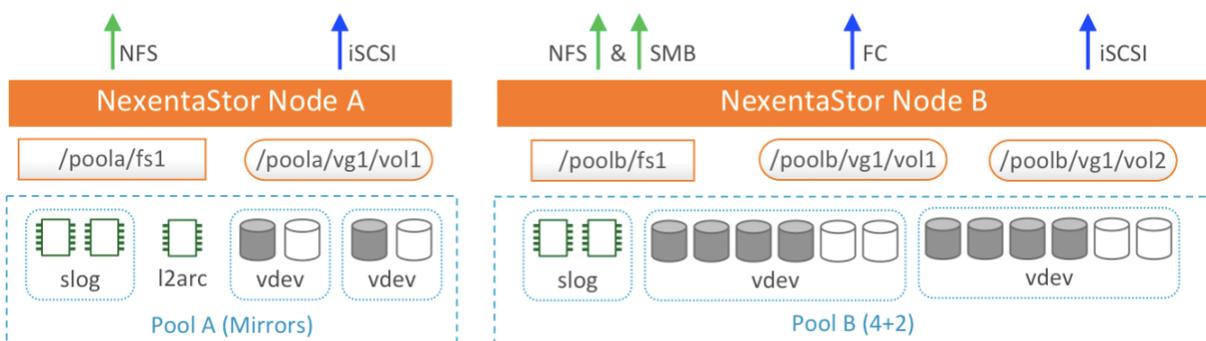
The example above shows a NexentaStor HA cluster with 2 hybrid pools. Pool A is configured with mirror vdevs and SSD based mirror write log (slog) and read cache (l2arc), while pool B is configured with (4+2) dual-parity vdevs and SSD based mirror write log.

File systems and block volumes are then carved out of each logical pool. All data stored in a pool is automatically striped across all vdevs in the pool for maximum performance. This allows any dataset (file system or volume), large or small, to get access to the full performance capability of the backend pool.

NexentaStor appliances typically run with a few large pools (potentially scaling up to petabytes) supporting large file systems. Being able to scale pools and file system sizes minimizes the number of logical entities that need to be managed and generally simplifies storage management.

To simplify system configuration, NexentaStor 5 supports a new *create-auto* function through both CLI and NexentaFusion that automatically creates pool data vdevs based on user provided criteria. For example, using this feature a user can specify automated creation of a "pool of RAID-Z2 (4+2) vdevs across 120 4TB HDDs across 3 enclosures encl_A, encl_B and encl_C with enclosure-level redundancy". The system takes care of laying out all the vdevs in the right way and automates the entire setup.

3.2 Performance Management

Nexenta is a long time contributor to OpenZFS across a range of areas covering data management, scalability, reliability, failure management and performance. NexentaStor 5 supports pools of SAS SSDs, SAS HDDs (generally NL-SAS) and NVMe SSDs. Combining modern x86 hardware, large amounts of DRAM and fast SSDs with OpenZFS allows NexentaStor hybrid appliances to generally out-perform legacy storage hardware appliances.

OpenZFS was specifically designed to deliver optimal performance and availability from commodity hardware. It is well known for its hybrid pool performance capabilities and its innovative use of ZFS Intent Log (ZIL), flash based secondary log devices (SLOG), DRAM based read cache (Adaptive Replacement Cache - ARC) and large size flash based level 2 reach cache (L2ARC).

While extensive information can be found online on the inner workings of OpenZFS, it is worthwhile to call out the following when it comes to performance and hybrid pools:

- NexentaStor makes optimal use of all available DRAM. DRAM is primarily used for read caching and the Adaptive Replacement Cache (ARC). ARC concurrently implements 2 caching algorithms to optimally allocate memory to cache most recently used and most frequently used data. Over the last decade of usage in a wide range of workloads, ARC has proven its effectiveness and its ability to out-perform legacy Least Recently Used (LRU) caching implementations.

- Less frequently accessed data is stored on SSD based L2ARC devices. NexentaStor refers to the L2ARC simply as *cache*. When creating a storage pool, you can designate devices to be used as the cache.

- The ZFS intent log (ZIL) accumulates write operations before full write transaction groups are issued to the backend pool devices. To maximize performance of small synchronous write operations, hybrid pools are configured with high performance, low latency SSD based secondary log devices (SLOG) that allow the NexentaStor node to pro-actively acknowledge write operations to client applications.

- In NexentaStor 5, all data is compressed before being stored on backend media. The default algorithm is LZ4, recommended because of its performance and compression efficiencies. ARC operates on compressed data, further optimizing its use of system DRAM. In NexentaStor HA clusters, SSD based level 2 ARC are also persistent, allowing cached data to be immediately available to a NexentaStor node following a pool failover.

One important parameter when tuning pools for performance involves configuring the pool and dataset record size. In NexentaStor 5, the default record size for file systems is 128KB. For volumes, it is 32KB based on the fact that FC and iSCSI LUNs are more generally used for applications and workloads that generate small(er), random IOs. For streaming type workloads such as large backup and archive or large video streaming systems, NexentaStor 5 supports larger record sizes, all the way to 1MB.

NexentaStor 5 supports both inline compression and inline deduplication. Inline compression is enabled by default on all storage pools with the high-performance LZ4 algorithm by default. Inline deduplication can be used but is generally recommended for lower capacity appliances. NexentaStor 5 continually monitors the size of the deduplication table and automatically turns off deduplication if it detects a risk of the deduplication table spilling out of fast DRAM onto slower pool storage, thereby protecting the system from unexpected performance degradations.

To further facilitate storage pool configuration, NexentaStor 5 includes profiles containing settings tuned especially for a given storage type: *All-Flash* and *Generic*. For example, if your NexentaStor appliance has all SSD storage devices, you can apply the *All-Flash* profile, which optimizes performance for this kind of configuration. All storage pools you subsequently create on the NexentaStor appliance will automatically have the profile settings applied to them.

For all-SSD configurations, NexentaStor 5 supports automated or scheduled TRIM / UNMAP on the backend SSDs. For some SSDs, this can help maintain performance, lower I/O latency, and increase the useful life of the SSDs in the pool. This feature can also be used for pools configured on top of thin LUNs presented by a backend storage array that supports UNMAP operations to reclaim unused capacity in the backend array.

3.3 Device Failure Handling

As mentioned above, besides performance, the other main objective of storage pool configuration is durability and protecting data and data services from device failures. NexentaStor supports raid 0 striped, mirror, single-parity, dual-parity and even triple parity pool configurations.

As of late 2017, NexentaStor 5 reference architectures support NL-SAS HDDs ranging from 2TB all the way to 12TB. Pools of 2TB HDDs are generally configured with mirror vdevs, while pools with disks larger than 4TB generally get configured with dual-parity (4+2), (6+2) or larger vdevs. An important consideration in the vdev configuration is whether the system should be resilient to loss of a full storage enclosure. For example, a physical setup with 3 separate storage enclosures can support a (4+2) pool with enclosure level redundancy (no more than 2 disks per enclosure for a given vdev).

NexentaStor supports the allocation of any number of hot spare devices to a pool. Both SSDs and HDDs can be configured as hot spares in a hybrid pool. In the event of a device failure in a pool, NexentaStor 5 will automatically activate the 'best' hot spare: a spare of the same type (SSD or HDD), of the same size and preferentially in the same storage enclosure as the failed device.

Nexenta led the work in OpenZFS on resilver performance, enabling device resilver (aka device rebuilds) operations to go as fast as the device can handle sequential IOs. This NexentaStor 5 enhancement is particularly critical for pools with large HDDs to reduce data exposure to failed devices.

3.4 Data at Rest Encryption

NexentaStor 5 supports data at rest encryption for pools built with TCG Enterprise Self-Encrypting Drives. The solution is transparent to all other features and does not have any impact on overall system performance. For data at rest encryption to be enabled on a particular pool, all devices in it must be SED compliant, including the SSDs used as SLOG or L2ARC devices in hybrid pools.

Figure 3-2: Data at Rest Encryption Overview
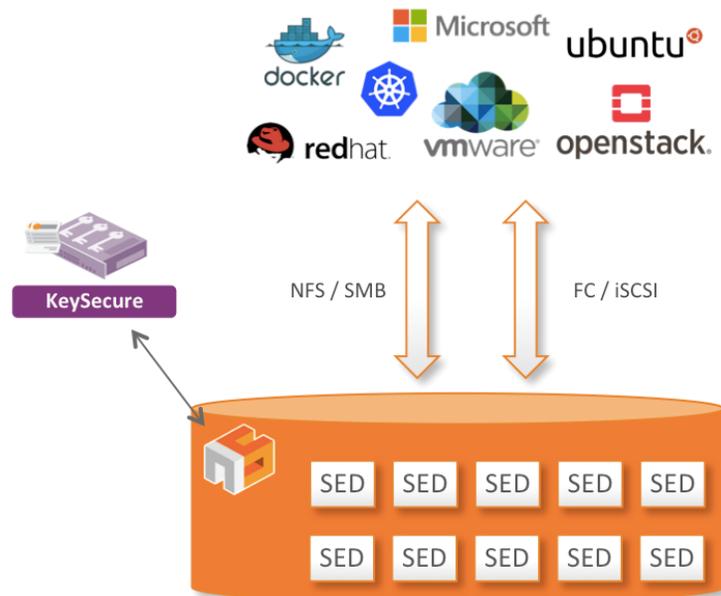
The solution relies on an external KMIP compliant key management infrastructure (e.g. KeySecure from Gemalto) to provide the keys required to access data on the pool devices.

3.5 File Systems and Shared File Services

A NexentaStor file system is a dataset configured in a storage pool to provide POSIX compliant storage of files and directories. NexentaStor benefits from OpenZFS' advanced data integrity, functionality, and scalability. As a result, NexentaStor supports file systems of virtually unlimited sizes (effectively only limited by the physical size of a pool), easily scaling to hundreds of terabytes and even petabytes.

By default, file systems share all the capacity in the underlying pool for maximum flexibility. Clients access the files and directories using standard SMB and NFS protocols. Microsoft Windows clients commonly use SMB to access file system data, while clients using UNIX and UNIX-like operating systems commonly use NFS.

A NexentaStor file system has a number of properties associated with it. These properties control such things as inline compression settings, default block size for files in the file system, users, groups, access control lists, quotas, reservations, ICAP virus scanning, and so on. Using the NexentaStor API and CLI, you can set these properties when you create the file system, or you can modify them as needed.

You can configure multiple file systems within a storage pool, and a file system can have additional file systems nested below it (up to 16 levels deep). By default, nested file systems inherit the properties of their parent. Nested file systems can also be combined with data protection services such as snapshot and replication schedules to maintain write consistency across the full hierarchy whenever snapshots are taken on the parent file system.

3.5.1 Shared File Services

To allow Windows clients access to data in a NexentaStor storage pool using the SMB protocol, you can create a file system and configure SMB-related settings for the file system, such as a workgroup to join, Active Directory domain, share names, etc. NexentaStor 5 provides SMB services through a Nexenta developed kernel based SMB server for better performance and scalability.

NexentaStor 5 ships with SMB 2.1 as the default SMB protocol version and adds support for SMB 3 to provide a high-performance, file-based storage backend for Microsoft Hyper-V environments, including Offloaded Data Transfer (ODX) to accelerate Hyper-V copy operations over SMB 3.

To allow clients running UNIX and UNIX-like operating systems to access storage pool data using the NFS protocol, you can configure NFS properties for a file system. NexentaStor 5 supports NFSv3 (default) and NFSv4 through a mature, high performance kernel based NFS server.

3.5.2 Access Control Lists

NexentaStor implements NFSv4-style ACLs on its file systems. A file system's ACL consists of one or more Access Control Entries (ACEs) that specify which users and groups are allowed to read, write, and modify files and directories in the file system. You can create and modify individual ACEs using the NexentaStor CLI and API.

NexentaStor ACLs are compatible with both NFS and SMB, so that the ACL you create for a file system applies to clients using either protocol. This provides the foundation for full-featured, concurrent SMB and NFS sharing of NexentaStor file systems using Microsoft ID Mapping.

3.5.3 Quotas, Reservations and Capacity Management

A critical implication of the NexentaStor pooling model is that all file systems potentially have access to all available capacity in the pool. This is a powerful feature that can greatly simplify capacity management in a large system: rather than managing capacity at each individual file system level, the storage administrator can simply monitor available capacity at the aggregate pool level.

An important OpenZFS best practice is to keep utilization of storage pools below 80%. As a copy on write / redirect on write file system, OpenZFS needs a certain amount of headroom to effectively manage available space in a storage pool. Pushing utilization above 85% will generally translate in degraded performance and higher latencies, particularly on all-disk and hybrid pools.

To control how much capacity a particular file system, a particular user or a particular group can consume, NexentaStor supports configuring quotas and reservations. A reservation pro-actively carves out capacity from the pool and keeps it aside for the said file system. A quota on the other hand defines the maximum amount of capacity that a particular user or group is allowed to consume.

3.5.4 Storage QoS For File Services

NexentaStor all-flash configurations serving out high performance, low latency, NFS datastores to VMware vSphere private clouds is a very popular use case. It is popular because users get the management simplicity of NFS VMware datastores, with the performance and ease of use of all-flash, and the cost savings of running software-defined storage on industry standard servers and SSDs. For hosting providers running large-scale multi-tenant virtualized infrastructure, all-flash storage also has the benefit of simplifying day-to-day operations and practically eliminating storage related customer escalations.

In some instances however, there is a need to control storage performance for some tenants; either to ensure fair sharing of the backend storage resources between tenants, or ensure that tenants get access to the resources they are entitled to, and not much more. For these environments, NexentaStor 5 supports Quality of Service for file services.

Figure 3-2: Example Use Case for NexentaStor Storage QoS
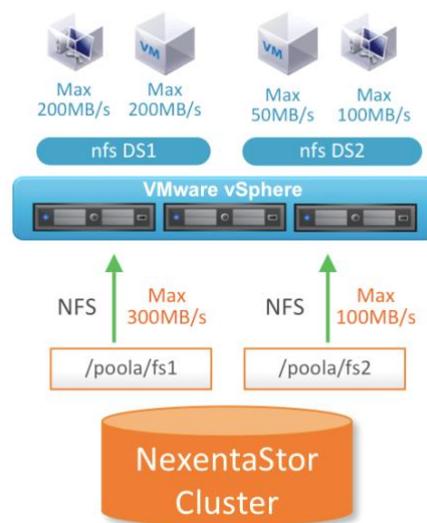
In the example above, NexentaStor QoS is used to deliver differentiated storage performance between a couple of tenants. Tenant A has access to NFS datastore DS1 while Tenant B has access to NFS datastore DS2. On NexentaStor, DS1 is configured to provide up to 300MB/s of bandwidth while DS2 is configured to provide no more than 100MB/s of bandwidth.

Configuring QoS on the NexentaStor NFS share allows the hosting provider to effectively cap the aggregate storage resources consumed by a particular tenant while leaving configuration of Virtual Machine level QoS (better enforced by the hypervisor) up to the tenant itself.

3.6 Volumes and Shared Block Services

Similar to how file systems are supported on storage pools, NexentaStor supports the creation of raw block volumes on storage pools, to be shared as either iSCSI or Fibre Channel LUNs. As with file systems, you can use the NexentaStor CLI to configure properties for volumes such as block size, compression mode, and whether the volume is read-only. All volumes configured on a NexentaStor pool are thin by default. Configuring a reservation on a volume will effectively make it thick.

3.6.1 Volume Groups

NexentaStor 5 requires that all volumes be contained within a predefined volume group. Volume groups simplify the process of configuring volumes by allowing you to set properties once at the volume group level and rely on property inheritance for the underlying volumes in the group.

For example, to change the compression property for the volumes in a volume group, you could simply change the volume group compression property rather than each volume's compression property. When you set the property for the volume group, all underlying volumes in the group automatically inherit it.

Volume groups also provide the critical function of consistency groups for data protection services, providing a natural control point for transaction level consistent snapshots across all underlying volumes. For example, NexentaStor can be configured to take recursive snapshots of all volumes in a volume group according to a schedule. The snapshots can be cloned or rolled back for the entire volume group at once.

3.6.2 iSCSI Block Services

NexentaStor volumes can be shared as iSCSI LUNs. To do this, you first create one or more iSCSI targets within NexentaStor, specifying their associated network portal(s). Nexenta recommends creating one iSCSI target per IP interface intended to support iSCSI traffic.

You then add the iSCSI targets you created to an *iSCSI target group*. An iSCSI target group is a group of one or more iSCSI targets that will provide access to a set of volumes. NexentaStor also allows you to configure *iSCSI host groups*, which are groups of one or more iSCSI initiators.

Finally, you specify one or more mappings between the volume and the iSCSI target group and iSCSI hostgroups and LUN identifier. The initiators in the iSCSI host group then have access to the volume using the LUN identifier on the specified iSCSI target group. The relationship between iSCSI targets, iSCSI target group, host initiator groups and mappings is shown below:

Figure 3-3: Typical NexentaStor iSCSI Configuration
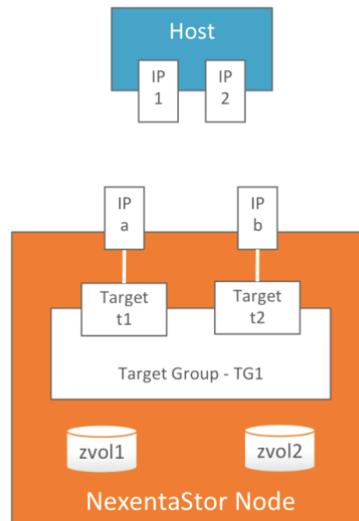
In a NexentaStor HA cluster, iSCSI targets should be created on Virtual IP addresses so that the iSCSI targets and iSCSI target groups that contain them can seamlessly fail over between nodes, with the backend storage pool containing the volumes they are sharing.

3.6.3 Fibre Channel Block Services

NexentaStor 5 supports implicit ALUA configurations when sharing FC LUNs. In order to share volumes as FC LUNs, your license must include the optional Fibre Channel Target feature. This allows you to create Fibre Channel target groups that contain FC targets from both nodes of a NexentaStor HA cluster.

Similar to how logical mappings are managed with iSCSI, creating Fibre Channel host groups and FC LUN mappings that combine a FC target group with a specific FC host group and volume provide granular control of LUN visibility across the Fibre Channel fabric.

4 NexentaStor Data Protection

4.1 Instant Space Optimized Snapshots and Clones

NexentaStor supports virtually unlimited numbers of high-performance, space-efficient, instant snapshots and clones. It is possible to take individual snapshots of file systems and volumes to get crash-consistent, point-in-time copies of their content.

Recursive snapshots can also be taken on nested file systems or volume groups to get transaction-consistent, point-in-time copies of all their underlying children file systems or volumes, respectively.

Figure 4-1: NexentaStor Hierarchical Data Consistency
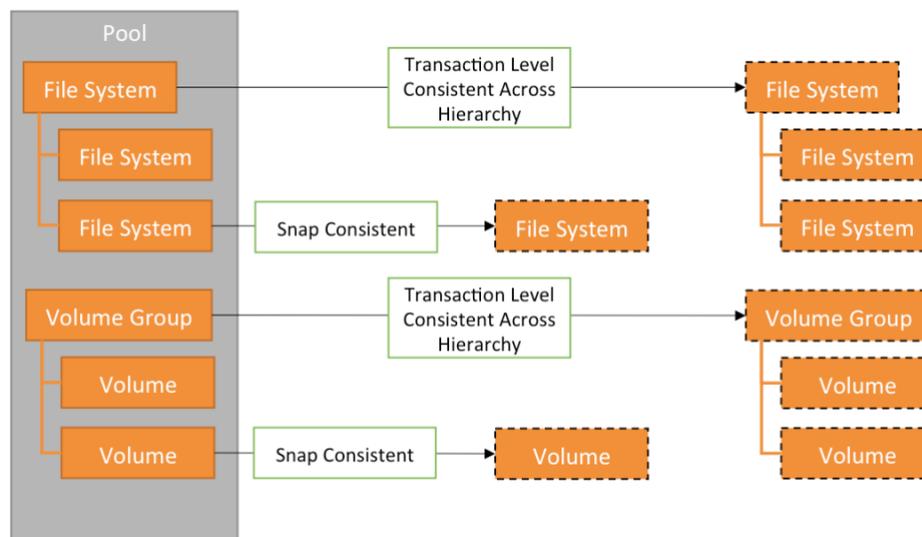
A snapshot is a read-only representation of a dataset (that is, a file system, volume group, or volume) at a specific point in time. Snapshots do not consist of the actual data, but pointers to where the data is located in the underlying pool. Because they do not contain data, snapshots are much smaller in size than the data they reference, allowing NexentaStor to collect and maintain a practically unlimited number of snapshots.

Snapshots can be cloned. When you clone a snapshot, it creates a new dataset that is initially populated with the contents of the snapshot. Unlike snapshots, clones are writable, so that you can add data to them as necessary.

A clone is linked to the original snapshot so that the original snapshot cannot be deleted while any clones of it exist. A clone can be promoted, which makes it independent of its original snapshot.

Using the NexentaStor snapshot rollback feature, you can restore a dataset to the state it existed in any of the snapshots retained on the appliance, assuring that all versions of the data captured by snapshots are protected.

4.2 Snapshot Services

Snapshots of a dataset can be created as needed using the CLI or REST API tools. NexentaStor 5 also supports advanced scheduled snapshot services (also called *snapping jobs*), which collect snapshots on a recurring basis and specify how many snapshots to maintain per schedule. Scheduled snapshot services are configured to take snapshots according to a predefined schedule, from every minute to every hour, day, week or even month.

Snapshot services provide great scale and flexibility. A given dataset can be configured with multiple snapshot schedules, each with its own retention policy specified in terms of number of snapshots to keep. For example, a recursive service could be configured on a volume group with 3 schedules: "every 2 hours, keep 12", "every day, keep 7" and "every Saturday, keep 52". This would result in a year's worth of data retention, with 52 weekly snapshots, daily snapshots for the past week and bi-hourly snapshots for the past 24 hours.

4.3 High Performance Replication Overview

NexentaStor 5 High Performance Replication (HPR) provides full featured, high performance remote replication for disaster recovery over any distance. The NexentaStor 5 replication engine has been moved to the operating system kernel for maximum performance and fine-grained control of Recovery Point Objectives. The result is a solution that supports two types of replication services:

- Scheduled Replication (SR) – this is snapshot based scheduled replication. It is included in Enterprise Edition licenses and allows Recovery Point Objectives (RPO) of 15 minutes and up. On systems with the additional continuous replication license option, SR can be configured for even tighter RPOs, as low as 1 minute.

- Continuous Replication (CR) – this is continuous asynchronous replication that can deliver close to zero RPO over any distance without affecting application performance on the primary site. This feature is only available on systems running Enterprise Edition with the additional continuous replication license. CR works by asynchronously replicating every write transaction group on the source dataset.

As shown in the example below, HPR services can be configured between file systems, volumes or volume groups and are managed at the dataset level, allowing bi-directional replication between sites, albeit between different source and destination datasets.

**Figure 4-2: High-Performance Replication Examples**
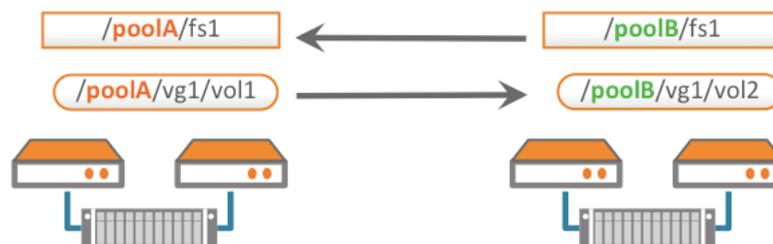
The HPR services can also be used between on-premises NexentaStor and in public cloud NexentaCloud instances. This provides a simple way to move datasets to the cloud for test/dev use cases, or application migration from on-premises to public cloud.

Figure 4-3: High-Performance Replication Between NexentaStor and NexentaCloud
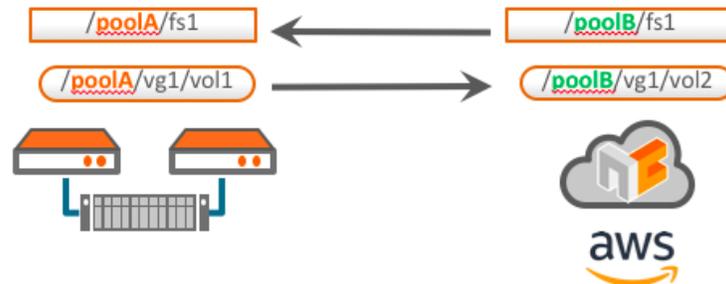
HPR services can either be configured on a particular file system or volume, or they can be set to run recursively on a parent file system and all its nested children file systems, or on a volume group and all the volumes it contains. This guarantees write transaction level consistency across the set of nested datasets and can be used to implement application level consistency groups.

Figure 4-4: Application Level Consistency Groups and HPR
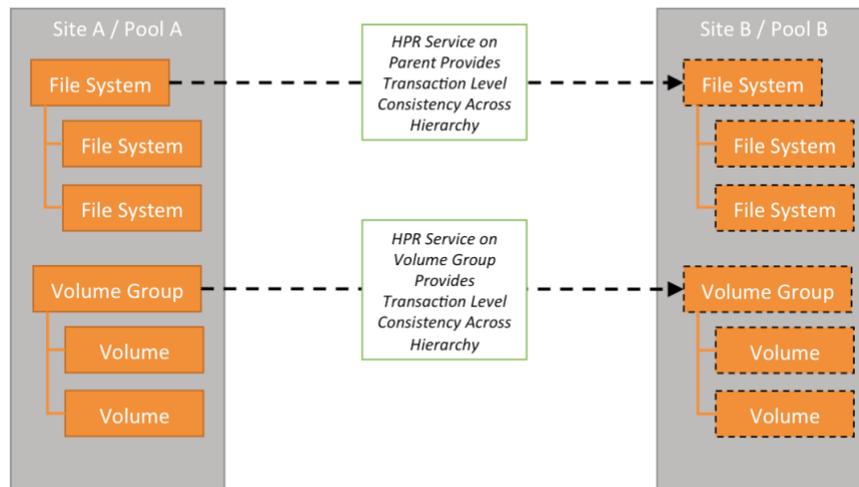
NexentaStor High Performance Replication services can be configured:

- To move datasets between 2 pools on the same NexentaStor appliance
- To replicate datasets between NexentaStor HA clusters
- To replicate datasets between a mix of NexentaStor HA clusters and single node appliance

4.4 Licensing Considerations for HPR

High Performance Replication is supported between systems running NexentaStor version 5 and above with Enterprise Edition licenses.

A system with the base Enterprise Edition license will support Scheduled Replication services delivering RPO as low as 15 minutes.

A system with the Enterprise Edition license and the additional continuous replication option provides support for Scheduled Replication with RPO as low 1 minute, as well as Continuous Replication services for near zero RPO over any distance.

4.5 Scheduled Replication Services

NexentaFusion and the NexentaStor VMware vCenter plugin provide simple configuration and management of HPR services. More advanced configuration options are available using the NexentaStor CLI and REST API.

Before configuring HPR services, you must ensure that all NexentaStor appliances share the same Replication password. This shared password is typically set at software installation time. It can also be reset at any time using the NexentaStor CLI. The shared Replication password provides a simple control mechanism to protect NexentaStor appliances from unauthorized data transfers.

A scheduled replication (SR) service is defined by the Primary (typically source) and Secondary (typically destination) datasets and a set of one or more schedules. When a schedule is created as part of a SR service, the user specifies how many snapshots for that schedule must be kept on the Primary and on the Secondary site. It is important to note that the retention policy is site-specific and does not change depending on the direction of the replication streams.

Figure 4-5: Scheduled Replication Service with Site Specific Retention Policies

By allowing multiple schedules to be configured as part of a SR service, it is possible, for example, to configure a service with an hourly schedule that keeps 12 snapshots on Primary and 24 on Secondary, and a daily schedule that keeps 7 snapshots on Primary and 30 on Secondary.

NexentaStor SR services only replicate the snapshots that are created as part of the service. Snapshots created on the source dataset (by local snapping jobs for example) are not replicated by HPR.

As part of a SR service, it is possible to create an on-demand snapshot and have it immediately replicated to the destination dataset. Schedule-specific retention policies do not apply to these on-demand snapshots, and they must be managed (i.e. eventually removed) by the end user.

While a SR service is active, the user may create a clone out of any of the snapshots on the destination dataset. Such clones can be used for disaster recovery testing, for example:

Figure 4-6: Scheduled Replication Service with non-disruptive DR Test
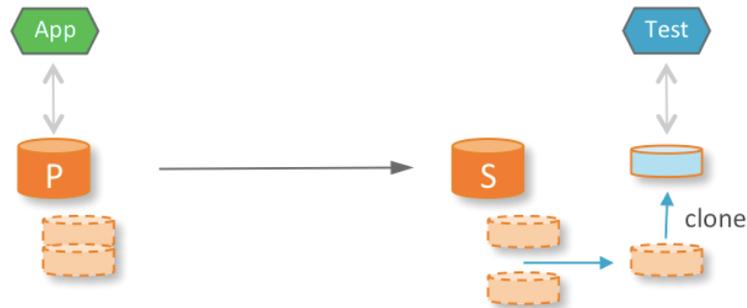
While a clone exists, its parent snapshot is protected and will be kept on the destination dataset, independent of the configured retention policy. Once the clone is deleted, its parent snapshot is automatically returned to the SR service and will be kept or deleted according to the retention policy of the schedule that originally created the snapshot.

It is also possible to share the destination file system of an active HPR service as 'read-only', allowing non-disruptive access to its data for backup and restore use cases.

HPR is a robust solution that can recover from wide variety of unexpected disruptions, including full site outages, as long as there are common snapshots between the Primary and Secondary sites. The NexentaStor hpr recover CLI command provides a simple solution to automatically restart a replication service that is interrupted by changes made to the source or destination datasets.

Another important reliability feature of HPR is the ability to automatically resume from interruptions of long running snapshot transfers (for example, during the initial synchronization operation): HPR is able to automatically restart from where it left off once network connectivity is restored without having to retransmit previously transferred data.

4.6 Continuous Replication Services

Continuous Replication services can be configured on file systems, volumes and volume groups. There are no schedules, retention policies, or visible snapshots to manage with CR services, making them seemingly much simpler to operate than Scheduled Replication services.

It is possible to concurrently configure both a CR and a SR service on the same source dataset, replicating to 2 different destination datasets, possibly on the same NexentaStor system. In such a configuration, the CR service provides near zero RPO protection for disaster recovery while the SR service can be used for more traditional backup and restore use cases.

4.7 Multi-Destination

NexentaStor HPR supports multi-destination configurations by allowing multiple SR services to be configured on the same primary dataset. HPR does not support cascaded replication, however, so the state of the various SR services that share a common primary dataset must be consistent with that limitation.

One and only one continuous replication (CR) service is allowed on a primary dataset. As a result, HPR supports having one CR and one or more SR services configured on that primary dataset, enabling configurations where, for example, dataset A is continuously replicated to site B, and replicated every day to site C via a separate SR service.

Figure 4-7: HPR Multi-Destination Configurations
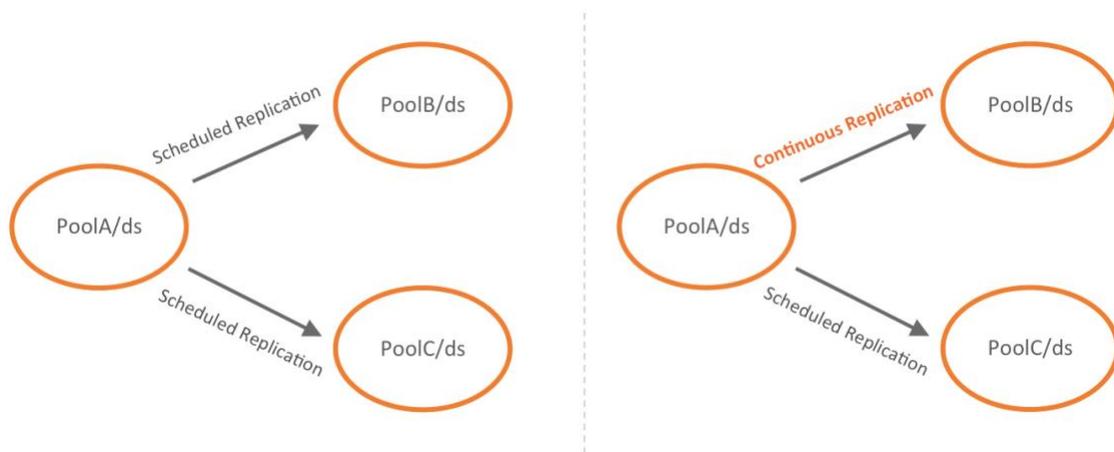
4.8 Moving Dataset To and From 3rd Party OpenZFS Systems

Independent of HPR, NexentaStor 5 also supports transfers of snapshots using standard OpenZFS send & receive over ssh to and from system running standard OpenZFS on Linux, FreeBSD or illumos.

Figure 4-8: Snapshot Transfer Using Standard OpenZFS send & receive
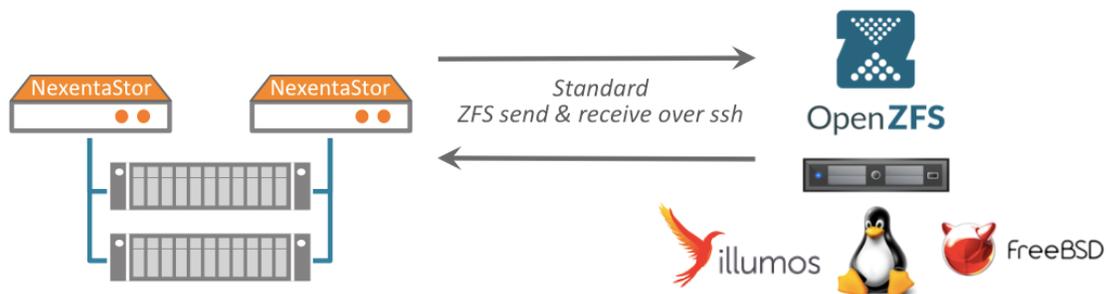
Contrary to HPR, this feature can only be used for on-demand snapshot transfer and does not support schedules or automatic retention policies.

Nexenta Systems, Inc.
Global Headquarters
2025 Gateway Place, Suite 160
San Jose, CA 95110   USA
Office: +1 (408) 791-3300
sales@nexenta.com