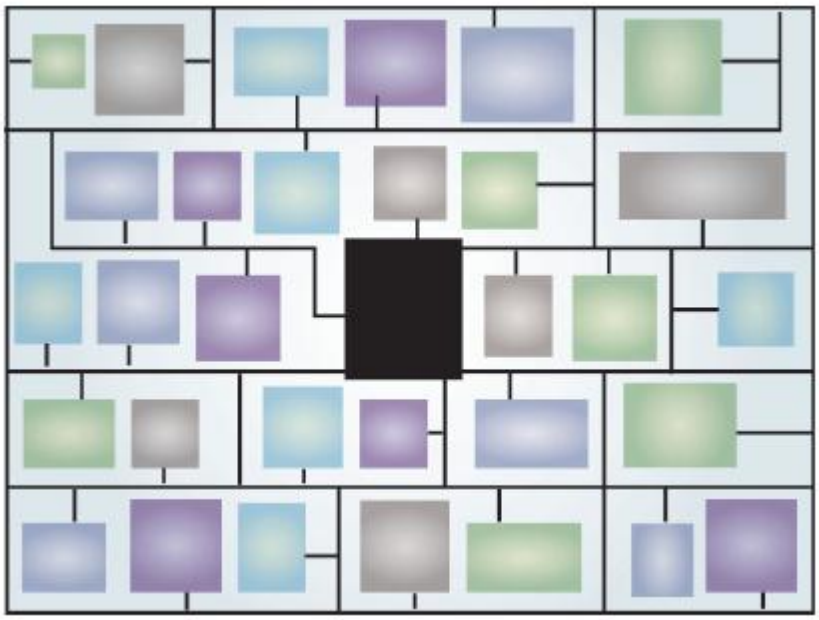
# DELL SCALABLE FILE SYSTEM

**A Dell Technology White Paper**
Version 3.5



June 2011

## Contents

## 1. Abstract

Traditional approaches to handling file data growth have proven costly, hard to manage, and difficult to scale effectively. Dell Scalable File System File System (DSFS) is designed to go beyond the limitations of traditional file systems with a flexible architecture that enables organizations to scale linearly—that is, to grow capacity without diminishing returns on performance. The DSFS architecture is enterprise class and standards based, supports multiple protocols including CFS and NFS, and incorporates innovative features for high availability, performance, efficient data management, and data protection.

## 2. File systems overview

The relentless growth of unstructured data is accelerating the need for network file storage systems. Organizations coping with data growth are confronted with several challenges. Data silos prevent easy access to vital business information. Data migration, backup, and DR are complex, consuming administrative time and resources. Meeting data growth by deploying more and more storage systems increases both the administrative burden and CAPEX, at a time when businesses need to run lean. And traditional file systems have scalability limitations that make them unwieldy for organizations with rapidly expanding file data.

A file system is a software layer running on the controller that manages how the data is stored, accessed and protected. Because the file system manages read and write access and maintains data integrity, its architecture and design directly impact storage system scalability, performance and reliability.

Today, a majority of organizations manage block and file data on separate, non-integrated storage systems. Both SAN (Storage Area Network, primarily Fibre Channel, and iSCSI) and NAS (Network Attached Storage, primarily CIFS and NFS) can be used to store unstructured data, but traditionally IT environments have selected NAS for unstructured data. NAS storage provides a consistent file system that enables the same storage space to be accessed by multiple heterogeneous clients.

Unified Storage combines SAN and NAS protocols in the same physical hardware chassis and typically has a single management interface. Organizations typically use Unified Storage to avoid the costs of deploying and managing separate block and file storage systems. However, as unstructured (file) data continues to grow, several challenges have emerged with this approach.

Traditional Unified Storage typically has rigid architectural boundaries that sometimes limit performance, scalability and size of file shares. Spikes in NAS storage demand can cause SAN performance bottlenecks. Read/write access to small files tends to be slow, a small subset of nodes are used for heavy file writers and multiple clients read large files concurrently. Unified Storage typically is priced like SAN, which costs much more in $/GB than traditional NAS. Forklift upgrades to the next-generation platform require tedious data migration and often weeks of system and network downtime.

File systems optimized for unified storage systems must address these challenges and overcome the associated limitations. Dell believes that a modern and forward-looking file system design should include the following features:

- No single point of failure and failover-on-write capability for data protection.
- Scale-*out* as well as scale-*up* architecture. Simply adding disk to handle data growth results in diminishing returns on investment. When scaling capacity, an organization must also scale the entire system that manages data storage and access. Otherwise, performance will lag, resulting in poorer and poorer access to data as capacity increases, and lower and lower utilization of the investment in new storage hardware.
- Unlimited file system/share size. Rather than imposing arbitrary limits, the architecture should enable file systems and shares to utilize all available storage capacity.
- The ability to handle multiple data access protocols, primarily CFS, NFS.

With these requirements in mind, Dell developed Dell Scalable File System (DSFS), a high performance, scalable, standards-based, enterprise-class solution to meet the challenges of rapidly expanding file data. This technology is being implemented in Dell's family of storage solutions to deliver the advantages of scale-out unified to its customers.

## 3. Dell Scalable File System

### DSFS Development History

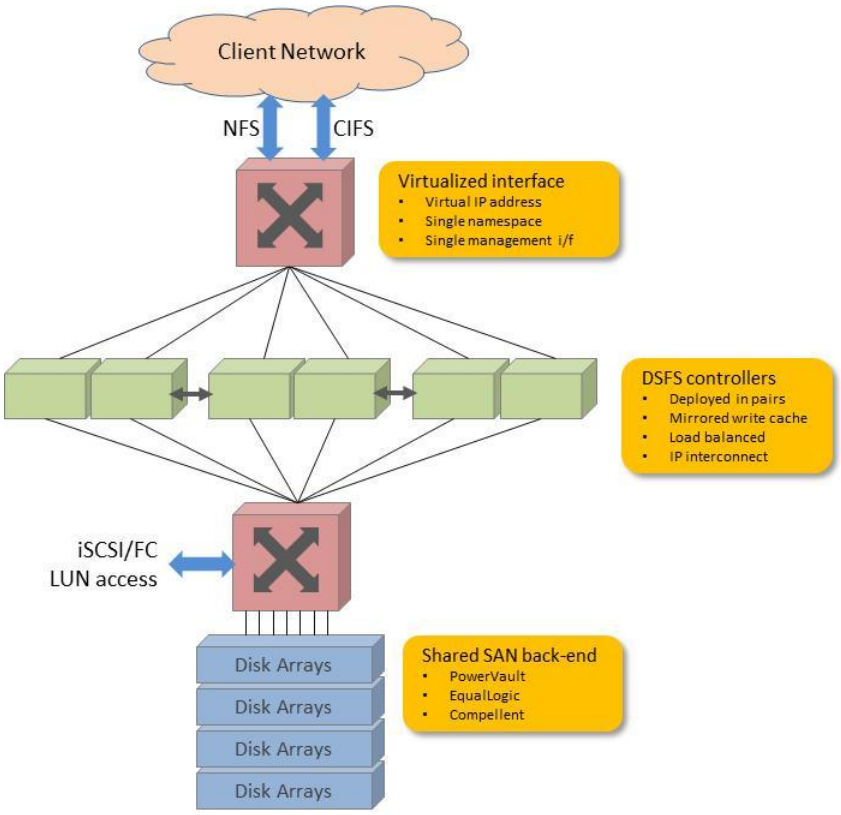
DSFS is the result of Dell's focus on offering superior technology that will enables our customers to meet their most critical enterprise IT challenges. In 2010, Dell acquired intellectual property from Exanet, Ltd., a firm whose assets included a hardware-independent, scalable NAS storage solution. Dell further developed the Exanet file system to support NDMP backup and Ocarina deduplication and compression technology and adapted it to work with its EqualLogic and PowerVault storage platforms.

### DSFS Solution Architecture

Dell Scalable File System (DSFS) represents an exceptional level of innovation in distributed file systems technology. It removes the scalability limitations associated with traditional file systems, supports both scale-out performance and scale-up capacity expansion, and provides a single global namespace for easy administration. DSFS is a high-performance scale-out file system that presents a single file-system namespace through a virtual IP address, regardless of cluster size. It offers an optimal combination of performance and scalability, making it an excellent choice for a wide range of use cases and deployment environments, including file-intensive user shares, highly available NAS and unified storage for SMB and public sector deployments, and virtual server environments with extensive NFS data and enterprise-level storage consolidation projects.

DSFS's active-active software architecture supports redundant controller pairs in a clustered configuration. DSFS systems typically consists of DSFS nodes (deployed in pairs) and the underlying storage arrays. DSFS nodes are based on x86 commodity hardware, offering organizations Dell's economies-of-scale as the leading vendor of server hardware for Enterprise customers.

DSFS stores file data on conventional storage arrays, from a single storage array to SAN configurations with multiple controllers, on multiple platforms. This enables organizations to balance cost/performance tradeoffs and optimize compatibility with existing infrastructure as they require. In a DSFS cluster, any single node can fail without affecting data availability or causing data loss, even if write operations were in-flight.




**DSFS System Architecture**

## 5. DSFS features and functions

### High Utilization of Storage Hardware

The design of Dell Scalable File System separates users' data and access from the underlying hardware configuration so that servers, CPUs, cache memory and disk drives are optimally utilized. As data gets written to the virtual server, it is distributed across internal file servers, and eventually to all disks connected to the storage cluster.

### Seamless File Sharing Among Heterogeneous Clients

With different client platforms accessing the same shared file system, DSFS provides fully interoperable multi-protocol file sharing for UNIX, Linux, and Windows clients using standard CIFS and NFS file access protocols and authentication methods (NIS, AD, LDAP). Multi-protocol support ensures that files can be shared via both CIFS and NFS protocols concurrently. DSFS also includes support for mapping users and permissions among users, and for authentication domains.

### On-Demand Virtual Storage Provisioning

DSFS volumes enable administrators to provision storage as needed, so that capacity can be allocated independently of physical storage configuration. The large pool of storage (i.e., the aggregate of all disks seen by the nodes) is split into smaller, virtual containers, each providing administrators with a full set of policies such as security style, quotas, snapshots, and alerting.

### Speedy Restoration of Large Volumes of Data

DSFS provides the ability to restore very large data sets that need to be recovered as a whole to a particular point in time. It helps administrators to recover large data sets (terabyte scale) easily, eliminating long file copies and the need for free space for the recovery process. It gives the ability for end-users to restore previous versions of files directly without contacting IT.

### Simple and Easy Management

Managing terabytes of NAS storage is simple with the administrative functions supported by DSFS. From installation and initial configuration to ongoing monitoring and storage operations, all functionality is provided via easy-to-use screens and wizards. The management interface can easily be used to set up policies, quotas, snapshots, and replication and engage NMDP backup.

A DSFS cluster is managed as a single NAS system regardless of how many nodes are in the cluster. After DSFS nodes are discovered and added to the cluster, there are no node-specific operations for the administrator. Volumes are virtual entities that span underlying LUNs and provide a context for setting policies related to access, quotas,

snapshots, replication, etc. When new LUNS are added, an administrator can choose to expand the NAS volume accordingly.

A DSFS cluster is accessed through a single VIP or network name, which means that as storage scales, customers don't need to worry about managing multiple mounts, balancing data across them, and redesigning applications to accommodate a fragmented namespace.

**Transparent capacity scaling**

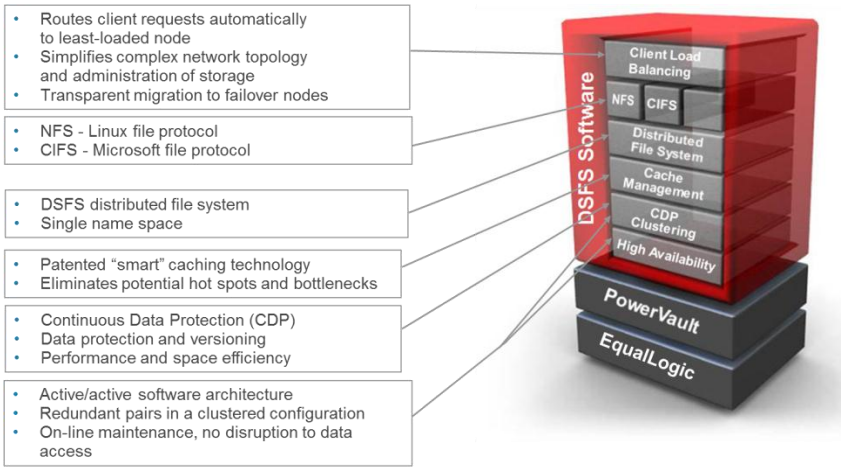
In a scale-on-demand model, organizations don't need to provision excess capacity in anticipation of future growth, which makes scale-out solutions ideal for customers that expect rapid growth over time or phased consolidation of applications. DSFS-based products support the transparent addition of capacity. Additional arrays can be added to a cluster, and those LUNs will be seamlessly mapped into the file system's virtual volume without downtime.

**Performance optimization and scaling**

In a scale-out implementation, customers may add nodes to transparently increase performance on-demand.

Because all nodes in a DSFS clustered system support active I/O, organizations benefit from high intrinsic performance without exotic protocols or the need to distribute application load across multiple filers. Load balancing sends client requests automatically to the node with the least-current workload. Network traffic is load balanced across the cluster. Load balancing is implemented natively within DSFS, and organizations can further benefit from network-based load balancing solutions.

Storage platforms using DSFS are also load- and capacity-balanced in the back end. For example, write traffic is load balanced across LUNs, and capacity is monitored to insure balancing across them.



**High availability**

In a DSFS cluster, any single node can fail without affecting data availability or causing data loss, even if write operations were in-flight. Cross-cluster reliability is achieved through a variety of mechanisms including a high speed cluster interconnect, write cache mirroring, failsafe journaling, and data integrity checks to insure data store consistency.

A DSFS Fault Management System (FMS) uses several mechanisms for insuring cluster reliability, maximizing data availability in cases of hardware or software failures. Other services monitor the health of the server platform, including temperature and power condition. Every DSFS implementation includes at least two active redundant components, making it invulnerable to any single point of failure, including network interfaces, PSU's, and the controllers themselves. In some recent implementations, including the EqualLogic FS7500 and PowerVault NX3500, cache mirroring may be supplemented by internal or external battery backup that supports cache dumping to disk in the event of node loss.

**Metadata**

File metadata, including name, owner, permissions, date created/modified and a soft link to its storage location, is contained in an i-node which is automatically stored on the backend array(s). It is managed by the file system through a separate caching scheme and is used to handle client requests and ensure data integrity.

Because all nodes are peer nodes in managing the file system, there is no metadata server. Checksums protect the metadata and directory structure and a background process constantly checks and fixes wrong checksums. Every metadata item is replicated on two separate logical units and all metadata updates are constantly journalled.

**File access**

Dell Scalable File system uses file system daemons to manage file access. The daemons provide file services, working together in one instance across all nodes to present a single file system and manage client requests. Each file system daemon manages the metadata for files that it created and file data may be handled by other FSDs.

**Optimization for large and small file sizes**

DSFS has no file size limit and is optimized for both large and small file sizes to ensure performance, reliability, and capacity efficiency associated with specialized workloads are optimized. For large files, data is distributed intelligently across all available storage at 1MB block granularity to improve performance and minimize fragmentation. Data striping provides performance and capacity balancing. DSFS aggregates small files for more efficient write operations. Files smaller than 4KB are stored in the i-node (assigned by file system, also contains metadata) so they can be read with a single disk I/O operation to improve read access times and overall file system performance.

Directories, metadata, and map files are distributed across all nodes and domains. Directory read-ahead improves directory listing and browsing efficiency. Read cache propagation is incorporated to improve data access to multiple clients that are reading large files concurrently.

Because DSFS is a cache-centric file system, writes to disk are done in such a manner that for a given file, there is minimum fragmentation on disk.

**File system operation in failover mode**

Any write to one node cache in a DSFS cluster is mirrored to the peer node before the operation is acknowledged. In the case of node failure, all dirty cache is duped to local, on-disk servers and the failed node is detached from the cluster. The cluster is put into Journal Mode, which triggers the mirror to be written to a journal file. DSFS client load balancing makes migration to failover nodes transparent.

## 6. Data protection and management

**Snapshots**

DSFS incorporates redirect-on-write snapshots instead of the copy-on-write solutions typical of other file systems. This approach requires only one I/O operation and delivers higher write performance. Each NAS volume has its own snapshot policy, wherein point-in-time "frozen" images are available to the end user as read-only views. DSFS creates a point-in-time backup of the data and applies volume-based policies that can be set for each volume.

Snapshots provide the first level of data protection by providing both end-users and administrators with the ability to recover data instantly from an unlimited number of recovery points.

**Asynchronous Replication**

DSFS allows fast and reliable snapshot-based replication of any number of volumes to a peer (local or remote) site. Only incremental changes are replicated and data is always consistent on the peer site, available as read-only. NAS configurations (volumes, exports, etc.) are replicated, enabling continuous access to data in the case of a disaster or site failure to assure business continuity.

**Quotas**

DSFS allows quotas to be set at the User and Group levels, and enabled or disabled without disruption. A Dell NAS volume is an extremely flexible mechanism for managing capacity allocation among multiple applications. In addition to User and Group quotas, quotas can be set for the volume itself as well as for the maximum amount of space that can be consumed by snapshots.

**Backup and DR**

DSFS supports standard backup software using NDMP. No changes are required to existing backup workflows. Backup streams are load balanced among available nodes.

DSFS allows immediate file and volume recovery using snapshots and replicas.

**Data protection**

A DSFS solution is configured to consist of two controllers in a cluster to ensure that there will be no single point of failure. Read-write operations are handled through mirrored non-volatile RAM (NVRAM). Cache data is mirrored between the paired NAS controllers to maintain complete data integrity. Data from the cache to permanent storage is transferred asynchronously, through optimized data-placement schemes. The controllers in a DSFS configuration connect to a RAID storage subsystem, designed to eliminate single points of failure.

Active components in the storage subsystem are redundant and hot-swappable. Each controller receives its power from a dedicated BPS (Backup Power Supply) and from the power grid. Each controller regularly monitors the BPS battery status, which requires the BPS to maintain a minimum level of power for normal operation. The BPS has sufficient battery power to allow the controllers to execute it's shutdown procedure. The BPS enables the controllers to use the cache as NVRAM. The BPS provides the clustered solution enough time to write all the data from the cache to the disk if the controller experiences a loss of power.

Continuous data protection is applied to metadata to help with versioning and prevent corruption in the storage system. Every metadata item and directory is replicated on two separate logical units. Every metadata item and the directory structure are protected with checksums. A background process constantly checks and fixes wrong checksums.

## 7. Deduplication

DSFS integrates Ocarina content-aware deduplication technology and this feature will be available in future platform updates. Deduplication is post-process, based on policies that can be set on a NAS volume basis and/or file attributes. Snapshots are fully integrated into the deduplication implementation, and deletion of snapshots is not required to accrue space savings. The policy-based deduplication process ensures that data that is not current is not optimized, minimizing performance impact. It also enables efficient and parallel enumeration of changed files and partial file optimization based on access patterns to a file. Post-processing data compression is also planned for future platform updates.

## 8. Product integration

DSFS is being implemented in a number of Dell storage solutions that serve the needs of small, midsize and large organizations. The first two product that incorporate this technology are:

- The Dell PowerVault NX3500 Unified Storage Platform, provides an easy-to-manage solution for both file and block-based applications in small to midsize deployments.
- The Dell EqualLogic FS7500 system offers high performance scale-out Unified Storage for midsized SANs. It is currently the only scale-out unified storage solution optimized for SMBs.

More information about these solutions is located in the appendix.

## 9. Appendix

**PowerVault NX3500**

The PowerVault NX3500 is the first in a series of products based upon Dell Scalable File System that delivers enterprise class file services to Microsoft Windows and Linux clients. It works with PowerVault MD32x0i and MD36x0i storage arrays, providing affordable unified storage with iSCSI, CIFS and NFS access to block and file data. The NX3500 lowers the barriers posed by traditional clustered file system implementations by reducing deployment complexity and offering clustered file systems benefits such as high availability, load balancing etc.

Organizations can use the PowerVault NX3500 to consolidate user data as well as other file and block applications into a single, easy-to-manage unified storage system with best-of-breed data management and scaling capabilities. The PowerVault NX3500's scale-up architecture delivers a flexible, load-balanced pool of high performance storage, making it easy to grow capacity up while avoiding the scalability constraints and challenges of managing separate block and file systems. With dual active-active file controllers and backup power supply, the PowerVault NX3500 gives you data protection and excellent performance with no single point of failure. More information about the NX3500 is available at Dell.com/NX3500.

### PowerVault NX3500 and Dell Scalable File System Technical Specifications

| Specification | Max Value (2-node system) |
|---|---|
| Max system size | 192 TB |
| Max file size | 4 TB |
| Max files | ~32 billion |
| Number of directories | ~34 billion |
| Max NAS volumes | 512 |
| Max snapshots per volume | 512 |
| Max snapshots per NX3500 system | 10,000 |
| Memory per NX3500 system with 2 quad-core CPUs | 24 GB per controller |
| Max LUNs | 16 |
| File name length | 255 bytes |
| Max NFS mounts | 1024 |
| Max CIFS shares | 1024 |
| Max Quota rules per NX3500 system (user quotas) | 65,536 |
| Max quota rules per volume | 256 |
| Max block level replication policies | 256 |
| Max directory depth | 1,024 |

**EqualLogic FS7500**

The FS7500 is a high performance solution that enables organizations to easily configure and manage iSCSI, CIFS, and NFS storage from a single interface. Its unique, DSFS-based architecture lets organizations scale both capacity and performance and pay as they grow. As storage needs grow and change, block and file capacity can be modified without disrupting existing applications and storage systems. A single file system can be expanded up to the capacity of the EqualLogic backend (up to 509TB usable storage). NAS service can be configured and added to EqualLogic arrays that have been deployed quickly and efficiently. The EqualLogic FS7500 includes a file-based snapshot capability (separate from iSCSI snapshots). Users can restore previous versions of files from a directory of these snapshots themselves, without contacting IT.

A dual active/active controller architecture and sizable onboard cache give the EqualLogic FS7500 outstanding performance. Each controller contains 24GB mirrored cache protected by a backup power supply. The EqualLogic FS7500 supports all new and existing EqualLogic arrays running a current version of the EqualLogic firmware. A dual active/active controller architecture and sizable onboard cache give the EqualLogic FS7500 outstanding performance. Each system provides 48GB of battery protected cache and traffic is automatically load balanced across all nodes.

The EqualLogic FS7500 supports all new and existing EqualLogic arrays. A single FS7500 system can support up to eight EqualLogic PS Arrays with the ability to add another FS7500 system into the same namespace to improve file performance. As with all Dell EqualLogic products, the FS7500's features, software licensing and future firmware enhancements are included in the base price.

| Feature | Dell™ EqualLogic™ FS7500 with Dell™ Scalable File System |
|---|---|
| Max system size | 509 TB |
| Max file size | 4 TB |
| Max files | ~64 billion |
| Number of directories | ~34 billion |
| Max NAS file systems | 256 per 2-controller FS7500 system, 512 per 4-controller FS7500 system |
| Max snapshots per NAS File system | 512 |
| Max snapshots | 10,000 per 2-controller system or 4-controller solution |
| Memory per FS7500 2-node system | 48 GB/24 GB per controller |
| File name length | 255 bytes |
| Max NFS mounts | 1024 per 2-controller FS7500 system, 2048 per 4-controller FS7500 solution |
| Max CIFS shares | 1024 per 2-controller FS7500 system, 2048 per 4-controller FS7500 solution |
| Max Quota rules per FS7500 system (user quotas) | 100,000 |
| Max quota rules per volume | 512 |
| Max directory depth | 512 |

**Glossary**

Understanding the terminology related to a Dell Scalable File System based PowerVault NAS system will help you successfully deploy, manage, and maintain your unified storage environment.

**Backup Power Supply (BPS)** – Provides back up battery power in the event of a power loss.

**Client access VIP** – Virtual IP addresses that clients use to access CIFS shares and NFS exports hosted by a PowerVault NAS system. The PowerVault NAS system supports multiple client access VIPs.

**Controller (NAS controller or nodes)** – Server appliance installed with the PowerVault NX3500 Dell Scalable File System software.

**Controller pair** – Two NAS controllers that are configured as pair in a PowerVault NAS clustered system. Cache data is mirrored between the paired NAS controllers.

**Dell Management Application (DMA)** – Data Management Application is also known as a Backup Application Server.

**Dell PowerVault Modular Disk Storage Manager (MDSM)** – The management software that ships with the PowerVault MD32x0i/MD36x0i array.

**Dell Scalable File System (DSFS)** – High-performance, scalable file system software installed on NAS controllers.

**Host Port Identifier** – Unique ID used to identify hosts in a network.

**Internal network A (peer connection)** – The PowerVault NX3500's internal network consists of two independent Gigabit Ethernet ports. The internal network is the infrastructure for PowerVault NX3500 clustering, including the heartbeat monitor, data transfer, and mirroring information between the controllers.

**Internal network B (internal management/IPMI)** – The PowerVault NX3500 internal management network (also known as internal network b) connects both controllers. All administrative related functions and controller reboots are performed on this network.

**LAN/client network (primary network)** – The network through which clients access NAS shares/exports. The PowerVault NAS system is connected to customer IT environment and its NAS clients using this network. NAS storage pool Virtual disks created on the PowerVault MD32x0i/MD36x0i storage arrays dedicated to the PowerVault NX3500 system.

**NAS volume (NAS container or virtual volume)** – A virtualized volume that consumes storage space in the NAS storage pool. Administrators can create CIFS shares and NFS exports on a NAS volume and share them with authorized users. A PowerVault NAS system supports multiple NAS volumes.

**NAS replication** – Replication between two PowerVault NAS systems or between two NAS volumes.

**NAS replication partners** – PowerVault NAS systems participating in a replication activity.

**PowerVault NX3500 Architecture** – The PowerVault NX3500 clustered NAS solution consists of a pair of controllers and the PowerVault Modular Disk (MD) iSCSI storage array. In addition, both controllers are protected by a backup power supply (BPS), which helps protect data during power failure.

**Network Data Management Protocol (NDMP)** – Network Data Management Protocol is used for backup and restore.

**Peer controller** – The peer NAS controller with which a specific NAS controller is paired in a PowerVault NAS system.

**Power module (battery unit)** – One of the battery units in a Backup Power Supply.

**PowerVault MD3xx0i** – Refers to the PowerVault MD3200i, MD3220i, MD3600i, MD3620i iSCSI storage solutions.

**PowerVault NAS Configuration Utility** – The setup wizard used to initially discover and configure a PowerVault NAS system. This utility is only used for initial setup.

**PowerVault NAS Manager** – The web-based user interface, which is part of the PowerVault NX3500 software, used to manage the PowerVault NAS system.

**PowerVault NAS system** – A fully configured, highly-available and scalable NAS appliance, providing NAS (CIFS and/or NFS) services, which is comprised of a pair of NAS controllers, a BPS, a PowerVault storage subsystem and the NAS Manager.

**Standby controller** – A server appliance that is not installed with the PowerVault NX3500 DSFS software. For example, a new or replacement controller from the Dell factory is considered as a standby controller.

**SAN network (iSCSI network)** – The network that carries the block level (iSCSI) traffic and to which the storage subsystem is connected. NOTE: It is recommended that this network be isolated from the LAN/client network.