Dell Fluid File System v3.0

A Dell Technology White Paper

Dell Product Group

June 2013

June 2013, version 3.0

Contents

1 Abstract 4
2 Dell Fluid File System Architecture 5
2.1 Layered Approach to FluidFS Logical Architecture 5
2.1.1 Network Load-Balancing Layer 5
2.1.2 File Protocol Access 6
2.1.3 NAS Volumes 6
2.1.4 NAS Pool 6
2.1.5 Block Device LUNs 7
2.2 The Physical Building Blocks 7
2.2.1 NAS Appliance 7
2.2.2 Back-end SAN Storage 8
2.3 Summary 8
3 Data Integrity 10
3.1 Cache Mirroring and Metadata 10
3.2 High availability 10
3.3 Summary 10
4 Data Protection 11
4.1 Snapshots 11
4.2 Backups 11
4.3 Replication 12
4.4 Summary 12
5 Data Reduction 13
5.1 Data Deduplication 13
5.2 Compression 13
5.3 Summary 14
6 Fluid File System Solutions 15
6.1 Dell Compellent 15
6.1.1 FS8600 15
6.2 Dell EqualLogic 16
6.2.1 FS76x0 16
7 Summary 17
7.1 Additional Reading 17

1 Abstract

Traditional approaches to handling file data growth have proven to be costly, hard to manage, and difficult to scale effectively and efficiently. Dell™ Fluid File System (FluidFS) is designed to go beyond the limitations of traditional file systems with a flexible architecture that enables organizations to scale out and scale up non-disruptively. Hence, it addresses organizational challenges by allowing them to gain control of their data, reduce complexity, and meet growing data demands over time.

The FluidFS architecture is an open-standards based network attached storage (NAS) file system that supports industry standard protocols including NFS v4 and CIFS/SMB v2.1. It provides innovative features proving high availability, performance, efficient data management, data integrity, and data protection. As a core component of the Dell Fluid Data architecture, FluidFS brings differentiated value to the various Dell storage offerings including the Dell Compellent FS8600 and FS8610 as well as the Dell EqualLogic FS7600 and FS7610. Continue reading to learn about additional features and enhancements that make it unique from other NAS products available in the market today.

2 Dell Fluid File System Architecture

The relentless growth of unstructured file 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 disaster recovery are complex, consuming administrative time and resources.
- Meeting data growth by deploying more and more storage systems increases both the administrative burden and capital expenditure at a time when businesses need to run lean.
- Traditional file systems have scalability limitations that make them unwieldy for organizations with rapidly expanding file data.

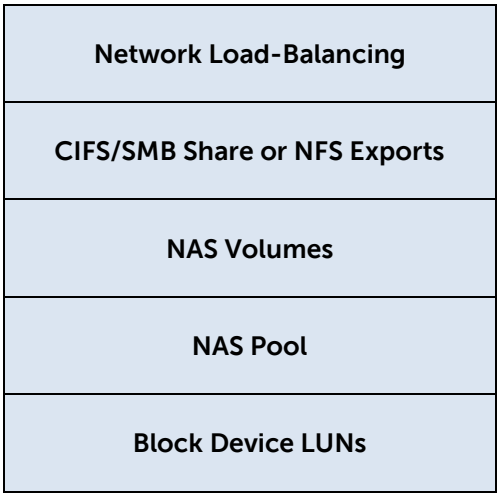
FluidFS is an enterprise-class, fully distributed file system that provides customers with the tools necessary to manage file data in an efficient and simple manner. The underlying software architecture leverages a symmetric clustering model with distributed metadata, native load balancing, advanced caching capabilities and a rich set of enterprise-class features. FluidFS removes the scalability limitations such as limited volume size associated with traditional file systems and supports high capacity, performance-intensive workloads via scaling up (adding capacity to the system) and by scaling out (adding nodes, or performance, to the system).

FluidFS operates across a symmetric cluster of purpose-built NAS controllers (housed in pairs within a 2U appliance), which interface over a fabric to shared back-end SAN storage, the EqualLogic PS Series and Compellent Storage Center in particular.

The following sections provide descriptions of the logical and physical layers of FluidFS.

2.1 Layered Approach to FluidFS Logical Architecture

The FluidFS architecture approaches layering with performance in mind. The layered architecture presents a traditional file system to network clients while performing a range of special functions at the back end. The specific goal of this design is to *utilize all available resources* at the network, server and disk levels to support the fastest possible response times.



The *logical* layers that comprise a FluidFS system and influence system performance are described below.

2.1.1 Network Load-Balancing Layer

FluidFS presents multiple network ports from multiple controllers to the client network. To simplify access and administration, FluidFS supports Virtual IP addresses and physical distribution of client sessions is

managed by a built-in load balancing capability. The load is balanced between the interfaces within a controller using one of two industry standard algorithms — Adaptive Load Balancing (ALB) or Link Aggregation Protocol (LACP). ALB is a MAC address-based balancing mechanism that does not require any configuration on the network switch. LACP, which is also known as 802.3ad, is another available option and is supported by most major manufacturers' managed switches; some configuration is required on the switch.

For client access within the same layer 2 network, FluidFS balances client connections across all available network interfaces within the cluster based on client MAC addresses. When load balancing across a routed network boundary, FluidFS utilizes DNS Round Robin in addition to the MAC-based address lookup. Each NAS controller can access and serve all data stored in the FluidFS system. Figure 1 shows load balancing in a Fluid File System cluster on a single layer 2 network.

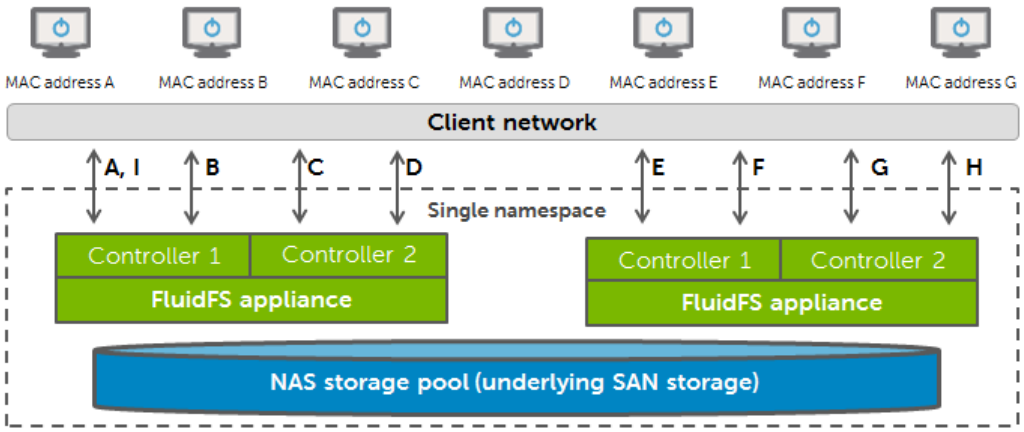


Figure 1 - Load Balancing in a FluidFS Cluster

2.1.2 File Protocol Access

FluidFS supports the CIFS/SMB and NFS protocols. All shares and exports configured by the system administrator are available from all NAS controllers at all times in a manner that is completely transparent to the client.

2.1.3 NAS Volumes

NAS volumes are virtual entities that provide policy-based management of snapshots, replication, quotas, deduplication, backup and security style as appropriate for different workloads. NAS volumes can be created on the fly and can shrink or grow, non-disruptively, in capacity up to the physical limits of the back-end storage.

2.1.4 NAS Pool

At its core, FluidFS creates a single distributed file system which spans across all SAN block device LUNs that are presented to the system. This layer, represented as the NAS Pool, provides file system services to the NAS volumes defined by the administrator in an optimized manner. All NAS controllers actively participate in the FluidFS distributed file system and can write or read any data housed within the file system.

2.1.5 Block Device LUNs

FluidFS integrates with best-in-class Dell SAN solutions EqualLogic and Compellent. LUN resources are automatically provisioned to FluidFS by defining an arbitrary amount of desired NAS capacity. Figure 2 depicts

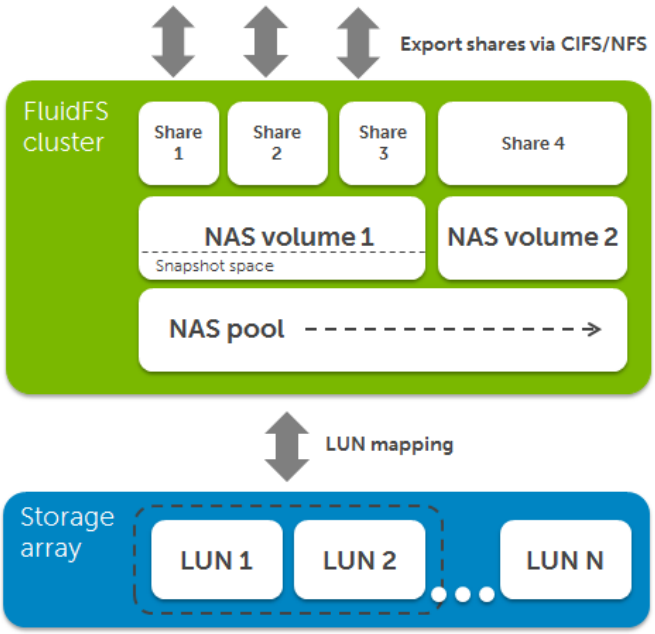


Figure 2 – Logical Constructs of a FluidFS Cluster

2.2 The Physical Building Blocks

FluidFS interacts with two main hardware elements:

- NAS gateway appliance(s)
- Back-end SAN storage fabric and array

The minimum hardware configuration of a FluidFS system is a single NAS appliance and a single SAN "unit" consisting of RAID controllers and disks.

This configuration can scale capacity by adding additional disks to the SAN and scale performance by adding additional NAS appliances or SAN storage controllers. The NAS system can scale capacity and performance independently and online, without disrupting system availability.

2.2.1 NAS Appliance

Two active-active, hot-swappable NAS controllers are housed side by side within a 2U FluidFS NAS appliance. A 40Gb/s mid-plane between the two controllers enables cache mirroring and write-back operations. Internal hard drives and backup batteries protect metadata and ensure data integrity. The appliance is connected to a block-based storage system – a Compellent or EqualLogic SAN – via 8Gb Fibre Channel or 10Gb iSCSI. The cluster can scale out to multiple NAS appliances according to client workload characteristics and business needs. To achieve data distribution and maintain high availability, each controller pair in a cluster has access to all other controller pairs in the cluster

through a dedicated and redundant interconnect network. Each FluidFS NAS controller can also read from and write to any file in any file system. Figure 3 represents the logical architecture of a FluidFS cluster with 2 NAS appliances.

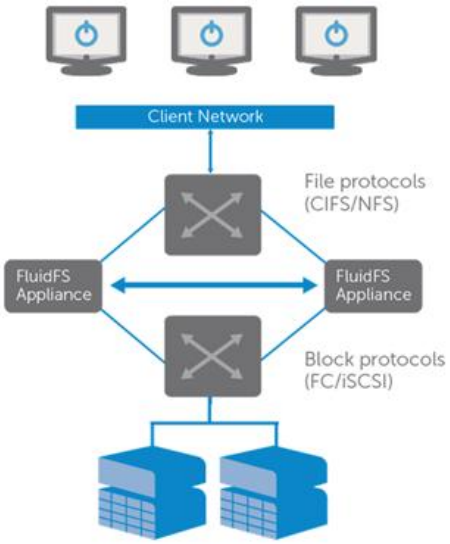


Figure 3 – A Dual NAS Appliance Cluster

The number of NAS appliances clustered together in a FluidFS system dictates the aggregate compute, cache and client network bandwidth available on the system. However, proper system sizing for performance must also consider the performance of the SAN back-end.

2.2.2 Back-end SAN Storage

FluidFS is integrated with Dell's leading block-level SAN storage systems, either Compellent or EqualLogic. Whether purchased as a complete system (including the SAN block device) or as an add-on to existing Compellent or EqualLogic components, FluidFS can be configured to meet organizational and needs.

One hundred percent of the back-end storage system can be assigned to provide block-level LUNs for the NAS controllers. Or the SAN resources can be shared with other applications that utilize direct block access.

Key configuration options on the back-end storage system include:

- Number of SAN controllers
- Number and type of disks
- RAID level and number of tiers

The specific configuration will dictate the overall performance of the SAN. For more information on FluidFS performance, read the Dell FluidFS Architecture for System Performance White Paper.

2.3 Summary

Every file accessed by a client on the network passes through the logical layers described above. Starting at the Network layer, the client is load-balanced to a specific NAS controller and continues

through the protocol layer via a CIFS/SMB share or NFS export to the NAS volume layer, the distributed file system layer and culminating with a write or read operation by the block device.

The configuration and number of NAS controllers combined with back-end storage system components defines the system characteristics in terms of capacity and performance. The FluidFS scale-up and scale-out capabilities allow the system's initial configuration to evolve over time to support the changing needs of the environment.

3 Data Integrity

FluidFS provides a series of mechanisms to provide a high level of integrity and system resiliency for data at rest and in transit. This section details those mechanisms.

3.1 Cache Mirroring and Metadata

During normal operation, the write cache, which includes the data and the metadata, is mirrored between the controller pairs in the FluidFS cluster. Additionally, important metadata is journaled to back-end storage capacity by a journaling process that runs continuously. This ensures file system consistency, in the event of a failure.

When a controller fails, cache mirroring is not possible, so the surviving controller immediately journals its data and metadata to the storage presented by the back-end storage subsystem. This ensures that the file system remains consistent and that all data is protected in case of additional failures.

In the event of a power outage, the write cache is journaled to a temporary staging location within the controller. This ensures that all I/O in flight remains consistent and prevents data loss. In addition, this ensures data consistency is not compromised, irrespective of the duration of the outage, as seen by traditional battery backed systems, which will lose data if power is lost before the write completes.

These mechanisms provide resiliency and redundancy for a multitude of failure scenarios to ensure that data is always intact and accessible.

3.2 High availability

In a FluidFS cluster, any single controller 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 ensure data store consistency.

FluidFS monitors the health of the NAS appliance, including temperature and power conditions, to ensure cluster reliability and maximize data availability in cases of hardware or software failures. If failures occur, hardware components in the storage subsystem are redundant and hot- swappable.

Each controller receives its power from the power grid and a dedicated backup power supply (BPS), which is regularly monitored to ensure that the BPS maintains a minimum level of power for normal operation. The BPS has sufficient battery power to allow the controllers to execute their shutdown procedures and use the cache as NVRAM. The BPS also provides enough time to write all the data from the cache to disk in the event of a loss of power.

3.3 Summary

FluidFS data integrity features provide both redundancy and high availability for several different scenarios. By working together you can rest assure that your data will always be available.

4 Data Protection

FluidFS enables data protection within a single system, across systems, and to external NAS repositories. This section will discuss some of the features and benefits of using FluidFS to store and protect data long-term.

4.1 Snapshots

Snapshots provide the first level of data protection by providing the ability to recover data instantly and are an integral part of FluidFS. Each NAS volume has its own snapshot policy to allow flexibility and to optimize space management. The administrator has the ability to schedule, point-in-time, automatic snapshots as frequently as every five minutes and can initiate manual snapshots at any time.

FluidFS incorporates redirect-on-write snapshots, instead of the copy-on-write solutions typical of other file systems. Redirect-on-write requires only one I/O operation, thereby preventing performance degradation. Figure 4 shows the redirect-on-write mechanism and the different stages of the snapshot process.

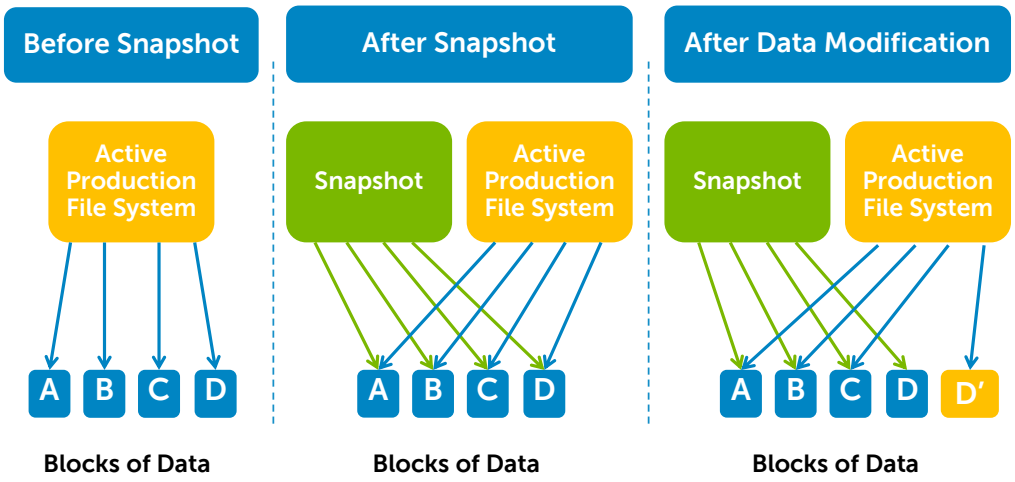


Figure 4 - Redirect-on-Write Snapshot Process

Snapshots are available to users as a read-only copy of the file system, which allows them to restore their documents in a simple manner without administrator intervention. Administrators also can easily restore very large data sets (terabyte scale) as a whole to a particular point in time. This eliminates long file copies and the need for free space for the recovery process.

4.2 Backups

The Dell Fluid File System supports standard backup software using Network Data Management Protocol (NDMP) version 4 with no changes required to existing backup workflows. Dell has partnered with industry leaders to provide comprehensive backup solutions that integrate with Fluid File System. Currently supported backup software includes:

- Dell Quest NetVault Backup 9
- CommVault® Simpana® 9.x
- Symantec™ BackupExec™ 7.x
- Symantec™ NetBackup™ 2010R3 and 2012
- Tivoli Storage Manager 6.3

Please refer to the product specification sheets for the latest information on backup software supported with FluidFS products.

4.3 Replication

FluidFS allows fast and reliable snapshot-based replication of any number of volumes to a partner. After the initial synchronization, only incremental changes are replicated, which improves network bandwidth utilization. This replication is native to FluidFS and does not require any additional hardware. The data is always consistent on the partner site and available as read-only.

In addition to data, NAS configurations (volumes, exports, etc.) are replicated. This reduces administrative burden and enables continuous access to data in the event of a disaster or site failure to assure business continuity.

Replication is bi-directional, meaning that the same system can host both source and destination volumes. In addition, the direction can be reversed without requiring a full resynchronization. FluidFS also supports "one-to-many" and "many-to-one" replication between NAS systems using unique volumes. Figure 5 shows the supported replication options.

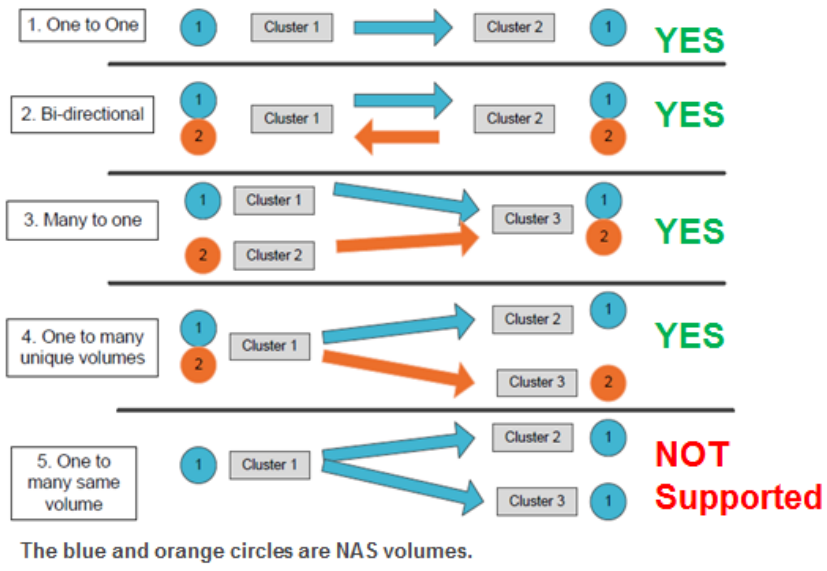


Figure 5 - FluidFS Replication

4.4 Summary

The use of FluidFS data protection features enables users to prevent data loss and to ensure continued operations for most storage environments, with limited impact on users and applications.

5 Data Reduction

FluidFS v3 provides policy based data reduction capability which can be used to reduce data foot print and thus maximize return on investment in Dell Storage. Data reduction polices are set on a per volume basis. FluidFS data reduction is designed to align with typical life cycle management policies of unstructured data. Data reduction is available as part of FluidFS v3 and does not require separate hardware or licensing. This section will discuss some of the features and benefits of using data reduction and how it can benefit your data.

5.1 Data Deduplication

FluidFS deduplication utilizes a post-write process. Files are written to the FluidFS appliance in their native form and then deduplicated once they meet a pre-specified set of reduction criteria. The administrator can set deduplication policies for each NAS volume based on the last time a file was accessed or modified, and can also schedule deduplication activity for a certain time window. For instance, a data set that will be rarely or never read or modified again, such as backup, can be set to be immediately reduced, and to do so after hours. Hot data that will typically experience many writes, such as home shares, can be scheduled to deduplication after a longer period of inactivity. When a file is modified, FluidFS will only rehydrate the portion of the file with write activity, minimizing I/O.

Fluid Data Reduction uses a variable-block sliding-window strategy. It crawls each file, looking for matching chunks of data sized between 64KB and 192KB with a general goal of 128KB. The reduction engine then eliminates redundancies, creates an object block map and stores that information back on the NAS volume.

Under normal circumstances, FluidFS will not optimize data that has not been accessed or modified in fewer than 30 days. However, there will be times where data will rarely be accessed after it is written to the file system. This is the case with backup/archive data or large image stores as one might find in a PACS system. For these instances, a setting called “archive mode” allows the NAS administrator to set a policy that enable data deduplication immediately, without waiting for the minimum active data policy to run.

5.2 Compression

Compression is the second sub feature of the core data reduction capabilities of FluidFS. Compression of files in FluidFS is accomplished using the LZPS algorithm. This was developed by Dell and uses less CPU overhead than traditional compression algorithms. This cuts down on the overhead associated with reading/writing small files.

Any NAS volume can be configured for only deduplication or deduplication plus compression. The NAS administrator can schedule when to run data reduction and on which NAS volumes. As more NAS appliances or controller pairs are added to a NAS cluster and as the new controllers own more data, not only does the NAS cluster scale in file performance, but also scales in the speed at which data can be reduced.

Once the data reduction job starts, all files meeting the data reduction criteria are processed for data reduction. Since only the files that are not actively used are reduced, FluidFS does not incur any performance impact for active data. For read access to data that has been reduced, the blocks requested by the client are rehydrated or changed to native form on the fly by FluidFS and passed to the client. Due to rehydration, read performance can be impacted. Reads do not alter how the blocks are stored on the disk- reduced blocks remain in reduced state. Any blocks within reduced data that are modified are stored in native form. When such files with modified blocks meet the data reduction policies, the modified blocks will be again reduced and stored in reduced form.

5.3 Summary

As unstructured data continues to grow, data reduction will become increasingly important. By taking advantage of FluidFS data reduction features it will be possible to store large amounts of data efficiently all while following typical data lifecycle management policies.

6 Fluid File System Solutions

FluidFS is being implemented as a core component in a number of Dell storage solutions that serve the needs of different customer scenarios and data workloads. Each solution has unique feature, function and value propositions, and each is differentiated by the type of controllers being used in a FluidFS cluster and the architecture of the underlying block-based back-end storage subsystem. This approach to design, coupled with the distributed and clustered architecture results in solutions that can easily leverage future technology enhancements.

Current solution that incorporate FluidFS are:

- Dell Compellent FS8600 – Unified Fibre Channel and iSCSI storage platforms that offer high-performance, scale out capability and efficient use of storage resources through features such as thin provisioning and Automated Tiered Storage technology that moves data to the optimal storage tier and/or RAID level.
- Dell EqualLogic FS76x0 – Unified iSCSI storage platforms that offer high-performance scale out capability. The scale out architecture native to both FluidFS and EqualLogic storage systems provides linear performance scalability in line with capacity growth.

6.1 Dell Compellent

Dell Compellent provides an extremely agile platform for constantly evolving block and file storage requirements. Performance and capacity scale non-disruptively to accommodate growing storage needs without forcing a platform rip-and-replace.

6.1.1 FS8600

The Compellent FS8600 scale-out architecture supports a single namespace across as many as four appliances and capacity that expands up to

2PB of manageable space with two Storage Center arrays. The inherent resilience of the FS8600 provides robust data protection without adding complexity.

For additional information refer to the product datasheet located on DellStorage.com.

6.2 Dell EqualLogic

Dell EqualLogic NAS platforms are high-performance solutions that enable organizations to easily configure and manage iSCSI, SMB, and NFS storage from a single interface. As storage needs change, block and file capacity can be modified without disrupting applications and storage systems. A single file system can be expanded up to the capacity of the EqualLogic back end. NAS service can be configured and added to EqualLogic arrays that have been deployed quickly and efficiently. EqualLogic NAS products include 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.

6.2.1 FS76x0

The FS7600 and FS7610 are the second-generation NAS appliances in the EqualLogic product line. They are based on FluidFS v2 and work with all EqualLogic PS arrays to provide highly available scale-out NAS and unified storage solutions. These solutions deliver the core benefits of the EqualLogic platform, including peer scaling, ease-of-use and all inclusive software licensing and they enable both performance and capacity to scale without SAN or application downtime.

The FS7600 has 1GbE connectivity to the SAN and client network and the FS7610 has 10GbE connectivity to the SAN and client network. Both appliances work with existing PS Series arrays to help provide comprehensive unified storage and NAS functionality for midsize and smaller deployments. With the PS storage back end, block and file storage capacity can expand up to 509TB, and performance can be increased by scaling out across two FS76x0 appliances. Most notably, with FluidFS, a NAS Volume or share can scale to the capacity of the backend storage allocated to file storage, all within a single namespace.

For additional information refer to the product datasheet located on DellStorage.com.

7 Summary

Dell's Fluid File System adds non-disruptive, scale out and scale up NAS capabilities to Dell's primary storage products: Compellent and EqualLogic. Additionally, it removes the scalability limitations associated with traditional monolithic NAS architectures. It also provides innovative features proving high availability, performance, efficient data management, data integrity, and data protection. With the added data reduction features of FluidFS v3.0 handling file data growth is now within the grasp of organizations of all sizes.

7.1 Additional Reading

For more detailed information on FluidFS performance, read the Dell FluidFS Architecture for System Performance White Paper.

For more detailed information on each of the Dell Storage Products mentioned in this document, please review the product datasheets located on DellStorage.com.