Dell Fluid File System Architecture for System Performance

A Dell Technology White Paper

Dell Product Group

June 2013

June 2013, version 1.0

Contents

1 Why Performance Matters

Organizations coping with multi-dimensional data growth face a mission-critical challenge: how can they obtain the capacity and performance required to meet their business needs, and do so while being mindful of their IT budget? While sizing to support a specific *capacity* is a straightforward process (expressed as the number and size of disks required) achieving the *performance* required (that is, the response time for reading and writing files) can pose a daunting technological challenge.

Business scenarios that are particularly demanding for file systems include:

- **High Performance Computing (HPC),** where many terabytes of data must be processed in a timely manner.
- **Media, Digital Video Editing,** in which rendering farms requires high capacity and bandwidth performance.
- **Home Directories,** where hundreds of users access thousands of files in a random manner.
- **Web Server farms,** which may receive hundreds of thousands of simultaneous queries that translate into a high number of file OPS managed by the file system.

In all of these cases, performance is impacted by a number of parameters such as compute power and storage system optimization. Today, CPU is a relatively inexpensive commodity that can be scaled up by using multiple servers. The major potential bottleneck to achieving the performance goals of a networked file system is the efficiency of the storage system.

Dell Fluid File System (FluidFS) is designed to go beyond the limitations of traditional file systems to meet the needs of diverse workloads, including those listed above.

Dell FluidFS ensures performance through:

- An advanced clustered architecture that enables independent scaling of capacity and performance.
- A sophisticated set of caching and data management policies that maximize performance while making very efficient use of physical resources.

When sizing a FluidFS environment – including the back-end storage array – to meet the performance requirements of a particular workload, users must consider all aspects of the storage architecture, the ways in which they contribute to performance and the impact they have on one another. A working understanding of this architecture is critical to preventing potential performance bottlenecks at the file system layer, the network layer or within the backend storage array. Refer to the Dell Fluid File System v3 technology white paper for a detailed overview of the physical and logical FluidFS architecture.

2 Designed for Performance: FluidFS Features

The Fluid File System (FluidFS) is the core IP for products like the Compellent FS8600 and EqualLogic FS7600/FS7610 NAS appliances. FluidFS operates across a symmetric cluster of purpose-build NAS controllers (housed in pairs within a 2U appliance), which interface over a fabric to shared back-end SAN storage, namely the EqualLogic PS Series and Compellent Storage Center.

The FluidFS software architecture utilizes a rich set of enterprise-class mechanisms like distributed metadata, native load balancing and advanced caching capabilities to support high capacity, performance-intensive workloads via scaling up (adding capacity to the system) and by scaling out (adding nodes, or performance, to the system). This section details those mechanisms.

2.1 Parallel Architecture and Load Balancing

The FluidFS architecture is a fully parallel, virtualized and active-active design. Any NAS controller can serve any file request, and all physical resources are available for I/O, even on a single file system. This means that client I/O activity can be distributed across the cluster to take advantage of multiple physical uplinks and parallel file I/O processes. Cluster parallelism is virtualized through the use of single or multiple virtual IP addresses.

Load balancing across the cluster can leverage multiple native FluidFS mechanisms including ARP, ALB, and LACP. Additionally, customers may take advantage of DNS-level load balancing depending on the client network topology

This parallelism has been shown to deliver near-linear performance increases as more NAS controllers are added to the cluster. These controllers can be added transparently, while the system is online, to scale performance as workload requirements scale.

2.2 Dynamic Read, Write and Metadata Caching

One of the major bottlenecks of traditional NAS systems is the inability to efficiently manage file system metadata.

The FluidFS cache is organized as a pool of 4KB pages. This cache is used for data as well as metadata. Data is evicted from cache based on the least recently used (LRU) algorithm. FluidFS maintains separate LRUs for data and metadata, ensuring metadata is retained longer in cache. This allows FluidFS to deliver high metadata performance.

In addition, FluidFS adapts to read-intensive, write-intensive and mixed workloads by maintaining separate LRUs for read and write, as well as using dynamic cache, automatically adjusting the size of the shared read/write cache at all times. Each controller reads and caches the data that is accessed by the clients connected to it. All subsequent access to the same data is serviced from cache, reducing back-end disk operations thus improving response time.

2.3 Distributed Caching and Hot Spot Mitigation

For workloads such as content distribution or film rendering, concurrent demand for a small set of data can impose performance bottlenecks. The distributed FluidFS architecture is ideally suited to support these types of workloads.

With FluidFS, every controller stores recently-accessed files in its own read cache. Frequent access to the same files on a controller will lead to those file constantly being served from cache. This enables extremely fast responses to read requests of hot files throughout the file system

As additional requests for the same data get distributed across the cluster, multiple (or even all) of the FluidFS controllers will cache copies of the data, preventing I/O bottlenecks to that data. This caching occurs at a block range level to insure efficient use of the available cache for read hot spots.

2.4 Write-Back and Data Coalescing

Physical write operations to disk constitute one of the most "costly" system functions in terms of performance, as they involve the physical movement of data through networking fabrics, RAID controllers, and ultimately to disk spindles before the write can be acknowledged back to clients. This normally results in highly asymmetric performance where write throughput and IOPS is significantly slower than read performance.

To reduce latency associated with read operations, FluidFS employs a write-back caching model. In a write-back model, writes are acknowledged to clients as soon as they land reliably in the write cache of the NAS controller and in the write cache of the peer controller. FluidFS insures reliability through this peer-node replication as well as a battery backup mechanism for cache data.

The decision to "write-back" the data to disk is based on internal FluidFS policies. FluidFS normally coalesces the write-cache contents and performs a physical write operation only when the aggregated write items in cache constitute approximately 1MB. Write-back can greatly reduce the load on the back-end from random traffic (i.e., large amounts of small file items), effectively turning transactional and small file application workloads into streaming operations.

2.5 Utilizing All LUNs for Write/Read Operations

Traditional NAS systems link specific disk groups or LUNs to specific file systems. This not only limits practical file-system scale, but imposes a performance bottleneck associated with the number of spindles in that LUN.

The fully distributed FluidFS utilizes *all* LUNs available to the system. Write operations that are directed from clients to the NAS reach the file system, at which point FluidFS spreads the write operation across multiple LUNs when applicable. Utilizing disk resources in parallel yields higher performance.

2.6 Minimizing Small-File I/O

Traditional file systems create and maintain inodes for each data item (file or directory) and use this structure to store metadata for the item (such as permissions, location mapping and so on). This means traditional file systems often perform multiple I/O operations for every read and write, which imposes a high performance overhead for transaction-oriented and small-file workloads.

For files smaller than 4KB – including emails, XML records, image thumbnails, etc. – FluidFS utilizes the inode to store both the metadata *and* the data. This innovative approach divides in half the number of disk operations required for small files.

2.7 Quicker Access: Read-Ahead for File and Directory

FluidFS employs a sophisticated read-ahead algorithm to identify sequential read operations for large files and complex directories. This special handling enhances performance by pre-fetching the next byte-range in the file or directory listing from disk, mitigating round trip latencies associated with the application and networking stack. Similarly, directory read-ahead pre-fetches lists of sub-directories and files to speed up browsing and crawling operations.

2.8 SAN Performance Optimizations

Although SAN capabilities are outside the scope of FluidFS, performance optimizations in the back-end can directly contribute to NAS solution performance, and should be considered in system sizing. Variables influencing NAS solution performance include fabric performance (switch configuration, RAID controller capabilities) and back-end configuration choices (RAID selection, disk types and quantity).

One of the more advanced performance optimizations available in Dell SAN products is automated tiering, automatically maximizing performance for hot data and maximizing economics for cold data. Compellent's Data Progression capability, for example, supports tiering across media types and RAID levels, and even supports platter edge-placement for the hottest data. Placement policies in the Compellent Storage Center controllers ("Data Progression Profiles") are tunable to meet the needs of different workloads.

FluidFS NAS solutions such as the Compellent FS8600 leverage features like Data Progression to extract maximum performance from the back-end without the need to overprovision expensive SSD or 15K drive resources.

3 FluidFS Performance Benchmarking

Generally speaking, the design configurations for a FluidFS NAS system involve:

- Required capacity
- Workload scenarios and required performance
- Future growth plans for capacity and performance

While sizing to support a specific *capacity* is a straightforward process (expressed as the number and size of disks required), meeting *performance* requirements is a more challenging task. Performance requirements are based on the workload scenarios by the organization. Performance requirements are satisfied by configuring the elements listed below in accordance with the specific workload scenario:

- Number and type of NAS appliances
- Number and type of back-end storage controllers
- Number and type of disks

As explained above, sizing considerations include both the capacity and the workload required. For example, the configuration for achieving best performance for 100 terabytes of capacity may differ from one workload to another. Consider the following hypothetical examples:

Workload Scenario A may require one NAS appliance, one back-end array and 60 high-capacity disks to achieve required performance and a capacity of 100 terabytes.

Workload Scenario B may require four NAS appliances, three back-end arrays and 500 lower-capacity drives to support a 100 terabyte capacity and the performance required for this workload.

The following sections describe some of the common workload needs that can be effectively met with FluidFS.

3.1 Random Access and Transactional Workloads

A random access workload is generally defined as a large number of various types of small file operations executed in an unpredictable manner (from the perspective of the FluidFS NAS system). In this scenario, write and read operation are commonly comprised of small files and file metadata.

Scenarios commonly categorized as random file access include home directories, compilation, Multi-tenant web-serving, and CAD/CAM, as they share the following characteristics: random traffic across many simultaneous clients generating read and write operations on relatively small files.

The access pattern of a random workload generates a large number of file system map queries and random access operations across back-end storage. The metric used to measure performance of random workloads is the number of file operations a system can execute per second (file OPS). File OPS should

not be confused with SAN IOPS benchmarking, as the former measures the aggregate operations across diverse read, write and metadata transactions.

3.1.1 How FluidFS Supports Random Workloads

The FluidFS architecture supports random access workload performance requirements by utilizing a built-in mechanism as well as the inherent scale up and scale out capabilities of the system, accommodating a system configuration that supports high random performance requirements. These built-in mechanisms, discussed in Section 3 above, include:

- Optimized caching
- Read cache propagation
- Cache coalescing (review section 3.4 for additional details)
- Storing small files in inodes

3.1.2 Number and type of NAS appliances

In a random workload, the aggregate amount of cache (and CPU) plays an important role in the system's performance characteristics. The cache allows the system to absorb a large amount of small file operations into cache and, in turn, apply the cache-related built-in mechanism which enhances system performance.

The greater the amount of cache the system supports, the higher number of small file items can be fit into cache and coalesced into large write operations to the SAN. This reduces random writes activity to the back-end storage array.

Read cache reduces random read access from the SAN, as hot files reside in the read cache and need to be read only once from disk. FluidFS dynamic caching will automatically adjust read/write cache ratios, devoting up to 90% of cache to read or to write at a given point in time, according to actual load.

As the environment's file performance requirements grow over time, the administrator may choose to add additional cache, compute and client network bandwidth to the system by adding additional NAS appliances to a live system.

3.1.3 SAN storage system configuration

NAS performance in random access workloads is highly dependent on back-end SAN performance. SAN IOPS performance is a function of the drive type, number of drives and the level of RAID protection applied. In a random traffic pattern, the SAN is required to read and write small bits of data that are spread across the storage system, hence response time affects the overall system performance.

Additionally, more drives allocated to the system will lead to more parallelism in I/O requests. Multiple I/O requests will be served by multiple drives in parallel.

As the environment performance requirements grow over time, the administrator may choose to scale up the back-end storage performance by adding additional drives, or in some cases by adding an

additional SAN array. As FluidFS NAS controllers utilize all available resources in parallel, any resource that is added to the back-end storage array is translated to higher FluidFS NAS performance.

3.1.4 Note on the Use of Solid-State Drives (SSDs):

An EqualLogic or Compellent SAN can be configured with various types of disks (ex. SSD, SAS, near-line SAS). To support random workloads, it is often best to use SSDs (fully or partially) due to their high IOPS performance. Common configurations often include a hybrid mix of SSDs and less expensive SAS disks to support capacity requirements. In hybrid configurations, the SSDs will provide Tier 1 performance while older data is migrated to SAS/near-line (NL) SAS disks residing on lower tiers.

The systems used in SPECsfs testing below used only SSDs, representing the minimum number and type of drives needed to deliver the measured file OPS performance. Customers requiring higher capacity may deploy lower-cost SAS or NL-SAS disks which are leveraged as Tier 2 or Tier 3 layers in tiering policies. Performance levels on active data (Tier 1) will remain consistent while cold data will be migrated to lower-cost, high-performance drives. The benefit to customers is a lower marginal cost of scaling because the hottest data is always migrated to the fastest media.

3.1.5 FS8600 Performance Benchmarks — SPECsfs

The SPECsfs benchmark suite (http://www.spec.org) is a standardized method used to compare performance data across file storage vendors. SPECsfs defines a workload which is designed to emulate random traffic based on data captured on thousands of servers; the SPECsfs benchmark generates a workload of which 72% of the operations are metadata operations, 18% are reads and 10% are writes.

SPECsfs results can provide a general indication of performance in workload scenarios consisting of home directories, compilation and so on. The table to the right details the technical distribution of the workload for SPECsfs.

SPECsfs benchmark results summarize the system capabilities in terms of number of operations per second (file OPS), and the average time (latency) to complete an operation.

NFS Operation	IO Mix
LOOKUP	24%
READ	18%
WRITE	10%
GETATTR	26%
READLINK	1%
READDIR	1%
CREATE	1%
REMOVE	1%
FSSTAT	1%
SETATTR	4%
READDIRPLUS	2%
ACCESS	11%
COMMIT	NA

3.1.6 Linear Scaling: FluidFS SPECsfs Results

Benchmark testing using SPECsfs performed at Dell labs demonstrated the exceptional performance and scaling capabilities of FluidFS.

- System tested: Dell FluidFS FS8600 10GbE model with a Compellent SAN storage system using three configurations
- The result per single appliance demonstrates exceptional performance per NAS appliance
- Scaled up and out, the system demonstrates near linear scaling capabilities of FluidFS

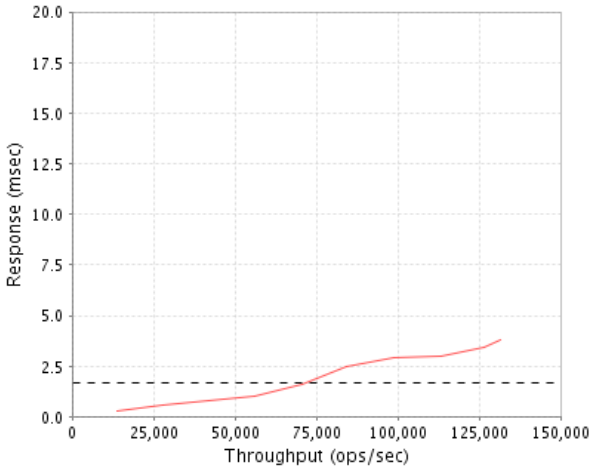
Read the FS8600 SPECsfs Solution Brief for a complete analysis of Dell's SPECsfs submissions.

3.1.7 SPECsfs configuration 1: Single-appliance FS8600

The single FS8600 NAS appliance configuration consisted of:

- 1x 10GbE FS8600 NAS appliance – (2 FluidFS controllers)
- 1x Compellent Storage Center (2 SAN controllers) housing 36 SSDs

The maximum SPECsfs performance achieved using this system was 131,684 operations per second with an overall response time of 1.68 milliseconds.



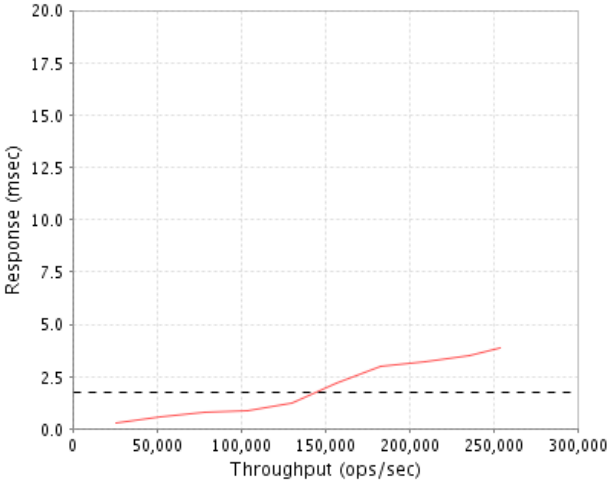
3.1.8 SPECsfs configuration 2: Two-appliance FS8600

Following the initial single-appliance test, the FluidFS system was scaled out by adding one more NAS appliance (to boost cache and CPU) and adding additional drives to the back-end storage array.

The second FluidFS system configuration used in the test consisted of:

- 2x 10GbE FS8600 NAS appliances – (4 clustered NAS controllers)
- 1x Compellent Storage Center (2 SAN controllers) housing 72 SSDs

The maximum SPECsfs performance achieved using this system is 254,412 operations per second with an overall response time of 1.71 milliseconds.



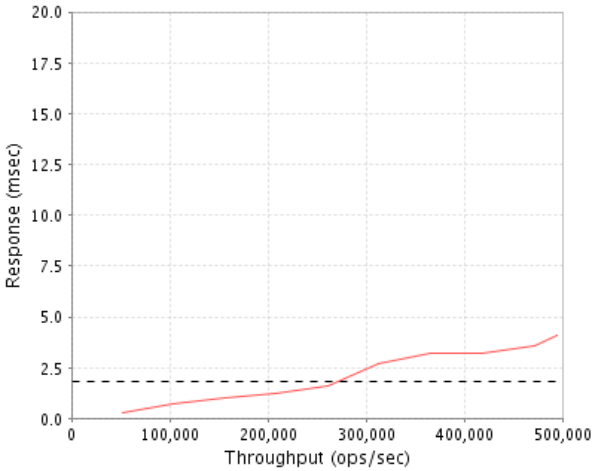
The second system demonstrated near linear scaling (2X) in the number operations per second over the first system.

3.1.9 SPECsfs configuration 3: Four-appliance FS8600

Finally, the FluidFS system was scaled out by adding two additional NAS appliances, adding a second Compellent Storage Center (2 additional SC8000 controllers) and doubling the number of drives.

The scaled up and scaled out FluidFS system configuration used in the test consisted of:

- 4x NAS appliances – FS8600 10GbE (8 clustered NAS controllers)
- 2x Compellent Storage Centers (4 SAN controllers) containing 144 SSDs



The maximum SPECsfs performance achieved using this system is 494,244 operations per second with an overall response time of 1.85 milliseconds.

The three configurations (1-, 2-, and 4-appliance clusters) demonstrated near linear scaling, doubling the number of operations per second while maintaining the same level of latency per any single operation.

3.2 FluidFS in Sequential Access Workloads

Sequential access workloads are generally defined as those that read and write large files in a consecutive manner. Common examples include backup, video streaming and seismic processing. Typically, a sequential workload involves relatively few metadata operations (as the number of files accessed is relatively small) and consists largely of file data access (read and write). The measurement unit used for this workload is MB/sec per single stream and MB/sec aggregate throughput of the entire system.

3.2.1 How FluidFS Supports Throughput Workloads

FluidFS scale out and scale up capabilities provide the flexibility to cater to the initial bandwidth workload requirement and grow the system as the environment evolves to require more performance. Read-ahead algorithms play a key role in supporting and accurately sizing bandwidth workloads.

3.2.2 Number of NAS appliances — Network Bandwidth

As accessing large files in a sequential manner requires high network bandwidth, the number and speed of client network NICs is critical to performance. In a bandwidth workload environment, it is recommended to use 10GbE appliance models, which can provide an aggregate 40 Gbps of client network bandwidth per NAS appliance (or 2x 10GbE LAN ports per NAS controller). Additionally, the number of NAS appliances will determine the available client network bandwidth to the NAS system.

As the environment performance requirements grow over time, the administrator may choose to add additional client network bandwidth capacity to the system by adding additional NAS appliance(s) to a live system.

Note that network infrastructure for both LAN and SAN fabric traffic must be appropriately selected to meet the needs of the workload.

3.2.3 SAN Storage System

In sequential workloads, most drives can support high bandwidth regardless of the drive type; thus the type of drive is less critical, and near-line (NL) SAS drives can often meet the requirements. The important factor is the *number* of disks allocated to the system which, in aggregate, can support the required bandwidth and the level of RAID protection applied. A common configuration might include a hybrid mix of SAS disks and less expensive NL-SAS to support the capacity. In this configuration, SAS disks will provide Tier 1 performance while older data is migrated to NL-SAS drives residing on Tier 2.

3.2.4 Throughput Benchmarking Results

Bandwidth tests conducted in Dell labs demonstrated an aggregated bandwidth of [1]2.5 GB/sec using a single FS8600 NAS appliance. The back-end Compellent Storage Center was configured with 15K SAS disks. Scaling up the back-end and adding an additional FS8600 NAS appliance (scaling out) doubled the total throughput of the system to 5 GB/sec.

In summary FluidFS can cater to environments that require high sustained bandwidth, scaling up and out to support higher bandwidth requirements in a linear manner.

[1] Testing based on FS8600 (FluidFS v2) with SC8000, IOzone load generation, and load designed to defeat read caching,

4 Summary: FluidFS Architecture for Performance

The FluidFS architecture – both logical and physical – is designed to enable excellent scalability and performance by leveraging advanced architectural concepts as well as SAN-level performance optimizations like Compellent Data Progression. Customers requiring high transactional performance and customers requiring high bandwidth for sequential workloads are both well served by the FluidFS architecture. The SAN-based architecture enables the deployment of FluidFS across both of these diverse workloads as well as for economically sensitive multi-petabyte archival applications without requiring a new learning curve for storage systems administrators.

Recent benchmarking of the Compellent FS8600 NAS appliance with SC8000 Storage Center SAN documents the high performance capabilities, demonstrating performance linearity with cluster size and efficient use of physical network, cache and drive resources to deliver that performance.

4.1 Additional Reading

For a more detailed review of the FluidFS architecture and capabilities, read the Fluid File System Technical White Paper**.**

For more detail on FluidFS SPECsfs performance, read the SPECsfs Solution Brief. Filings are available at www.spec.org.