Dell™ FluidFS Solutions Brief

# Rendering Using Dell Fluid File System

FluidFS Systems Engineering
Dell Enterprise Storage

January 2015

# Acknowledgements

This white paper was written by Bryan Lusk and the FluidFS Systems Engineering Team

# Feedback

Please give us feedback on the quality and usefulness of this document by sending an email to:

FluidFS-System-Engineering@Dell.com

# Revision History

| Revision | Date | Description |
|---|---|---|
| A | January 2015 | Initial Release |

# Contents

# 1. Preface

Content development departments are under constant pressure to deliver increasingly complex content, on timetables that are always shrinking. This requires storage solutions that are high performance out of the box, while being able to seamlessly scale up and out as business needs grow. Organizations cannot take long downtimes for forklift upgrades when they need to scale up performance or capacity, while facing rigid deadlines. All the same, the data must be secure from loss, as well as highly available in case of unplanned hardware failures. Storage solutions in this industry must be high performance, highly scalable, and highly available.

As faster server and storage solutions are becoming available, these advances in computer power have allowed for much more complex images to be rendered, in the same timeframe (or better) as previous generations of lower quality CGI. This trend results in productions (industry-wide) that have much more detail, look more realistic, and are more visually stunning. In order for businesses to compete in terms of quality of rendered products, their IT infrastructure must be able to keep up with the high demands of the media market.

The Dell Fluid File System (FluidFS) is purpose-built specifically to meet these needs, and more. The FluidFS architecture is a fully virtualized, parallel active-active design. In environments, such as rendering, where performance is critical, and storage access is distributed across multiple hosts (such as render nodes), the Dell Fluid File System is designed for, and has proven to be able to deliver exceptional performance.

# 2. FluidFS – Built for Performance

The Dell Fluid File system uses an advanced clustered architecture that enables independent linear scaling of capacity and performance. When configured with a proper SAN and disk configuration, FluidFS can deliver exceptional performance for a variety of workloads.

FluidFS utilizes a sophisticated set of caching and data management policies that maximize performance while making very efficient use of physical resources. 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 flushed from cache based on a least recently used (LRU) algorithm. FluidFS maintains separate LRU's for data and metadata, ensuring metadata remains longer in cache, to deliver exceptional performance on metadata operations.

For rendering workloads, 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 ensure efficient use of the available cache for read hot spots.

Please refer to the whitepaper Dell Fluid File System Architecture for System Performance to read more on how FluidFS is purpose built for performance, scalability, and stability. This whitepaper, and more, can be found on the Dell FluidFS Tech Center website.

# 3. Rendering Workloads

A render farm is a collection of high performance computer systems/servers (render nodes), operating together on a common dataset, to render computer generated imagery (CGI). This computer generated imagery is most commonly used for television shows, commercials/advertisements, or films.

A rendering workload is a highly parallel workload, with many render nodes each rendering different frames. These frames are combined to form animated sequences. There are typically many of these sequences that are developed individually, grouped, and ultimately combined to create the end product/film/video. The render farm is controlled by a queue manager, which divides a job/shot/sequence into multiple pieces (some number of frames), and controls which render nodes will render each piece, and when. Each render node refers to a jobs "scene file", which is accessible by all render nodes on a NAS device, and renders its share of frames. Once each render node has finished its frames, it stores the rendered frames back to the NAS device to be reviewed/edited and built into a sequence or shot.

From the storage perspective, this translates into many hosts accessing a common set of files, performing a large number of read and metadata operations (for input), as well as write operations (for output), which is perfectly suited to Dell FluidFS. Most of the read operations typically are considered small (less than 1MB), with a random access pattern, as the render nodes read many small texture files, wireframe files, scene files, and other small files. The write operations are typically larger and more sequential than the read operations, since the write operations are the "output" of the render work. Additionally, in typical rendering workloads, metadata operations can account for a high percentage of the total IO.

An example rendering workload is as follows:

| Operation | Impact On Storage |
|---|---|
| 1. Parse a directory tree<br>Example directory structure:<br>/<studio>/shows/<show>/<sequence>//<shot> | Metadata operation |
| 2. Check for many files existence and their permissions. File may include texture files (pieces of image files), wireframe | Metadata operation |
| 3. Open the files | Metadata operation |
| 4. Read the header | Small read operation |
| 5. Seek to some point in each file | Metadata operation |
| 6. Read some of the data out of the file | Small read operation |
| 7. Close the file | Metadata operation |
| 8. (Local to the render node) perform render CPU work | None |
| 9. (Possibly) Use the NAS device to store temporary output (I.E. images that will be combined into a shot with motion) | Small write operation |
| 10. Repeat steps 1 through 9 until shot is complete, then write the final product | Large write operation |

After each image (frame) is rendered it is typically written to storage (either local to the render node, or to NAS). Then when the images (frames) are combined to form a shot, the video output is stored to NAS. For render farms with extremely high core counts, using separate storage solutions for read/metadata operations (input to render nodes) and write operations (output of render nodes) can result in significant performance benefits.

# 4. FluidFS Benchmarking

## 4.1. SpecSFS

Dell has conducted in house testing with the SpecSFS benchmark, and proven exceptional results (very low latencies and high throughput) using SSD-backed configurations.

SpecSFS is a common industry benchmark for NAS systems. The operations mix is outlined in the table below, and consists of 72% metadata operations. While it is not a perfect representation of a rendering workload, it is fairly close.

> Note: SpecSFS NFS operations per second is a composite of the operations below, and it should not be confused with IOPS. There is no easy or direct correlation between SpecSFS NFS Operations Per Second and SAN IOPS.

| NFS Version 3 Operation | SPECsfs2008 |
|---|---|
| LOOKUP | 24% |
| READ | 18% |
| WRITE | 10% |
| GETATTR | 26% |
| READLINK | 1% |
| READDIR | 1% |
| CREATE | 1% |
| REMOVE | 1% |
| FSSTAT | 1% |
| SETATTR | 4% |
| READDIRPLUS | 2% |
| ACCESS | 11% |

The SpecSFS results for FS8600, along with the tested configurations, are online at the locations below. When comparing the Dell FS8600 numbers below with the published numbers for other market leading scale-out NAS vendors, one can see that Dell FS8600 delivers better performance while using a fraction of the number of cluster nodes, CPU cores, RAM, drives, rack space, and most importantly at a fraction of the cost.

- 4 appliance FS8600 cluster with 2 Storage Centers - ~500,000 NFS Ops
  - Results and Details
  - Solution Diagram
- 2 appliance FS8600 cluster with 1 Storage Center - ~250,000 NFS Ops
  - Results and Details
  - Solution Diagram
- 1 appliance FS8600 cluster with 1 Storage Center - ~130,000 NFS Ops
  - Results and Details
  - Solution Diagram

# 5. FluidFS Configuration Guidance for Rendering

## 5.1. Hardware Configuration

For rendering workloads that demand the highest level of performance, Dell recommends using the Dell Compellent FS8600 FluidFS platform, along with an SSD-backed Storage Center SAN. These configurations are referenced in the previous section of this document, detailing the SpecSFS benchmarking that has been conducted.

All of the configurations included in the benchmarking activity utilize Solid State Disk (aka "flash"). With SSD media prices rapidly declining, and advanced tiering now a viable technology, Dell recommends flash for the rendering storage front-end (top tier), as well as dense HDD arrays with large capacity spinning disks for cold data/archive data (bottom tier).

However, depending on the storage demand and render farm CPU core count, acceptable performance can also be achieved with solutions backed with only spinning disk. Some studios that specialize in visual effects or commercials only render short shots, particles, or effects. These types of studios typically utilize workstations for traditional scene manipulation, offloading a significant portion of the storage demand to the workstations of the artists. If the core count/host count is low enough, these types of studios may not require an SSD-backed NAS system, and an FS8600 backed by a Storage Center with spinning disk can meet the storage performance demands.

For one Dell customer, Important Looking Pirates, a case study has been included in the links at the end of this document. Important Looking Pirates is a VFX rendering studio that is using FS8600 (4 appliance cluster), with a spinning disk SC8000 SAN that includes a 10K RPM tier (72 drives), and a 7.2K RPM tier (72 drives). Important Looking Pirates has roughly 50 clients accessing the FS8600 via NFS, and has a dataset comprised of many tiny files (bytes), and many large files (>2 GB). When a heavy simulation hits their render farm, they can easily have 50 nodes wanting to load 50 different 2GB files simultaneously. Their FS8600 handles their workload and delivers great performance. Important Looking Pirates uses the following applications (which utilize FS8600): Maya (3D effects), Houdini (3D effects and rendering), Vray (rendering), Nuke (composing), After Effects (composing), Photoshop (creating textures and 2D images), Rv & djv_view (playback and image review).

In addition to the disk configuration, additional FS8600 appliances can be added to achieve greater throughput or operations. Each FS8600 appliance can handle a maximum of 2.5Gbyte/second sequential throughput, or 130,000 NFS operations per second. FluidFS scales linearly, so as performance demands increase, a FluidFS cluster can be seamlessly scaled out to add more appliances, up to a maximum of 4 FS8600 appliances.

For sizing guidance on FS8600 (in terms of appliance type and count) and Compellent SAN (in terms of disk class, amount, size), please discuss the performance requirements with your Dell sales representative. The Dell Sales team has internal tools that help to provide sizing guidance.

## 5.2. Software Configuration

### 5.2.1 Protocol (NFS vs SMB)

In addition to the hardware configuration, the protocol that is used can impact performance as well. Testing has proven that the NFS v3 protocol delivers the best performance for rendering. The NFS v3 protocol has less overhead than SMB or NFS v4, and is geared more towards performance. However, FluidFS has proven that it can provide exceptional performance for Windows based/SMB render farms as well. NAS performance for a Windows based/SMB render farm is expected to scale linearly, in the same manner as it does for NFS. The default NFS mount options (which FluidFS negotiates with NFS Linux clients when no mount options are specified) are recommended to achieve the best performance.

**Note**: FluidFS v4 does not support NFS v4 delegations at this time.

### 5.2.2 Optimized inode distribution

FluidFS v4 includes a new feature that will distribute the ownership of newly written files across all FluidFS NAS volume domains. This feature is detailed in the FluidFS Migration Guide, however, it is also highly beneficial for rendering. A typical rendering workload results in many render nodes accessing a common set of data. Distributing the inodes across all FluidFS NAS volume domains results in much better performance, because it allows all NAS volume domains to share in the load of serving this common dataset. When the "optimize inode distribution" feature is enabled, all new writes will be distributed amongst all NAS volume domains. When they are subsequently read by the render farm, the load of the read operations is shared among all NAS volume domains. If this feature was disabled, and the data was written, then read, it is possible that only one NAS volume domain would be serving this common dataset, creating a bottleneck. Testing has shown that enabling this feature prior to migration results in up to a 2x improvement in rendering workloads (compared to results with the feature disabled).

### 5.2.3 Reduced Metadata Redundancy

By default, FluidFS always stores two copies of all filesystem metadata. It stores each copy on a different LUN. The purpose of this is to help protect and safeguard the filesystem in the very rare cases that one LUN might be corrupted, or there is some sort of cache loss. However, more modern SAN backends have proven that storing a redundant copy of filesystem metadata is not as critical as it was in the past. The metadata is inherently protected by RAID technology, not by storing redundant copies. FluidFS v4 includes a new feature that allows the administrator to configure FluidFS to store one copy of the file system metadata, instead of two. The end result is that for metadata operations, there is less load on the backing storage. This results in better performance with metadata operations, when only one copy of metadata is stored. Rendering workloads are very heavy on metadata operations, and therefore using this feature can result in performance benefits.

### 5.2.4 Fluid Data Reduction

FluidFS includes data de-duplication as well as compression technology, integrated into the file system, at no extra cost. This technology is called "Fluid Data Reduction", and is designed for data at rest. Data at rest is defined as files that haven't been accessed or modified in 5 days (or older, configurable). Deduplication and/or compression only applies to files that are 64Kbyte in size or larger. Many shops will develop content for a commercial, film, effect, etc and then never need it again, but still want to keep it stored. Utilizing Fluid Data Reduction can greatly cut down on IT costs for storing large amounts of archive data by reducing the amount of physical disk space that is needed. With Fluid Data Reduction, in conjunction with data progression, which propagates colder data to lower cost storage including dense arrays with 7K NL-SAS drives, rendering shops get the benefits of optimized performance for active rendering jobs, with low cost (and scalable) retention for cold data. This allows customers to defer on bulk data movement into archival systems, further driving down total storage solution costs.

> **Note**: Prior to enabling Fluid Data Reduction on an FS8600 running FluidFS v4, it is recommended to also enable the SCSI Unmap feature (if Storage Center OS 6.5 or later is installed as well).

# 6. Future Work

Dell is currently working closely with several rendering and FX customers on configuring FluidFS NAS for rendering workloads. As testing and configuration validation efforts continue, additional documentation will be provided to detail best practices around performance optimization, as well as key features including snapshot management, tiering, de-duplication, compression, backup, and disaster protection/recovery.

# 7. Conclusion

The rendering use case for NAS is one in which the phrase "time is money" rings most true. Animated film studios push the limits in terms of performance and protocol capability, and the Dell Fluid File System is well equipped to meet these high demands. In order to keep pace with the industry, content developers should not be limited by the IT solutions they rely on. FluidFS is an ideally suited solution to achieve the highest performance, reliability, and scalability.

# 8. Additional Resources

Media and Entertainment Specific:

FluidFS – Architected for Performance (whitepaper)

Information on how Dell Fluid File System is a great fit for Media and Entertainment

A case study on Important Looking Pirates – An acclaimed VFX rendering studio using Dell FS8600

Whitepaper from Axle Media on Dell FluidFS and axle Video's Gear appliance (MAM)

Whitepapers:

Dell Tech Center - IT Community where you can connect with Dell customers and employees to share knowledge, best practices, and information about Dell products and your installations.

General Product Information:

Documentation and Best Practices for Dell Compellent FS8600 and Dell Compellent Storage Center

Documentation and Best Practices for Dell EqualLogic FS7500 and FS76x0 as well as EqualLogic Arrays

Documentation for all other Dell products including PowerVault NX series, PowerEdge, PowerConnect, and Force10