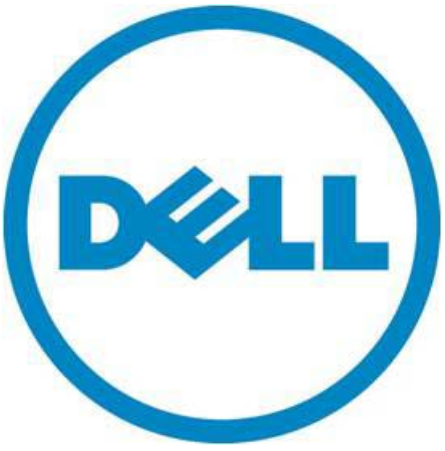
DELL | Terascala
HPC Storage Solution (DT-HSS2)

A Dell Technical White Paper

Dell – Li Ou, Scott Collier

Terascala – Rick Friedman

Dell HPC Solutions Engineering



November 2010

Contents

Tables

Figures

Introduction

Cluster computing is the most popular platform for high-performance computing in use today. In all clusters, compute nodes and networking and storage I/O main subsystems must be designed in such a way that clusters run efficiently.

For compute nodes, the primary challenges are to ensure the availability of scalable tools, that servers deployed as compute nodes can be easily installed and maintained, and that the nodes themselves deliver high performance per watt for an optimal cost of ownership.

Once an efficient processing solution is available, networking is the next key element to a successful cluster. As processing nodes become more efficient, internode communication performance becomes critical. A cluster interconnect must deliver high bandwidth and low-latency communication while being easy to manage on a large scale. Dell offers InfiniBand and 10Gb Ethernet technologies to help address these challenges.

With a powerful computing platform and high-speed internode communication available, the final challenge to creating an efficient and balanced compute cluster is the storage and I/O subsystem. With multiple compute nodes simultaneously accessing data, today's single access storage solutions cannot deliver data fast enough to keep the cluster running at peak efficiency. One alternative has been to use a parallel file system approach that delivers high throughput, parallel access, and scalable capacity in one system. However, historically, such solutions have been complex to deploy and maintain or proprietary and too expensive.

The Dell | Terascala High-Performance Computing (HPC) Storage Solution (DT-HSS) is a unique new storage solution that provides high-throughput Lustre storage as an appliance. The DT-HSS is targeted toward usage as a scratch file system. The DT-HSS solution consists of Metadata Servers (MDS), Object Storage Servers (OSS), and one or more pre-configured storage arrays. With performance of up to 1.2 GB/sec per Object Server, the latest generation offering (referred to as DT-HSS2 in this paper) delivers the performance necessary to get maximum utilization from your high-performance computing infrastructure. The available redundant metadata infrastructure ensures high availability of all metadata through an active–passive MDS architecture and use of RAID controllers in the storage arrays. The Dell | Terascala HPC Storage Solution (DT-HSS2) is a refresh of the previous generation offering (DT-HSS1). Standard DT-HSS2 configurations are designed to scale from 48TB installations up to 336TB, are delivered as a fully-configured storage solution that requires minimal deployment, and are available with full hardware and software support. Leveraging the Dell™ PowerVault™ MD3200, MD3220, and MD1200 storage arrays, the DT-HSS2 delivers a great combination of performance, reliability and cost effectiveness.

This paper describes the Dell | Terascala HPC Storage Solution which delivers all the benefits of a parallel file system based storage solution in a simple to use, cost effective appliance.

Lustre Overview

Lustre is an open source, high-performance parallel file system for applications that require very high throughput, scalability, and capacity. It is used in some of the largest supercomputing sites in the world, delivering hundreds of GB/sec of throughput and supporting multiple petabytes of data in production environments for the last ten years.

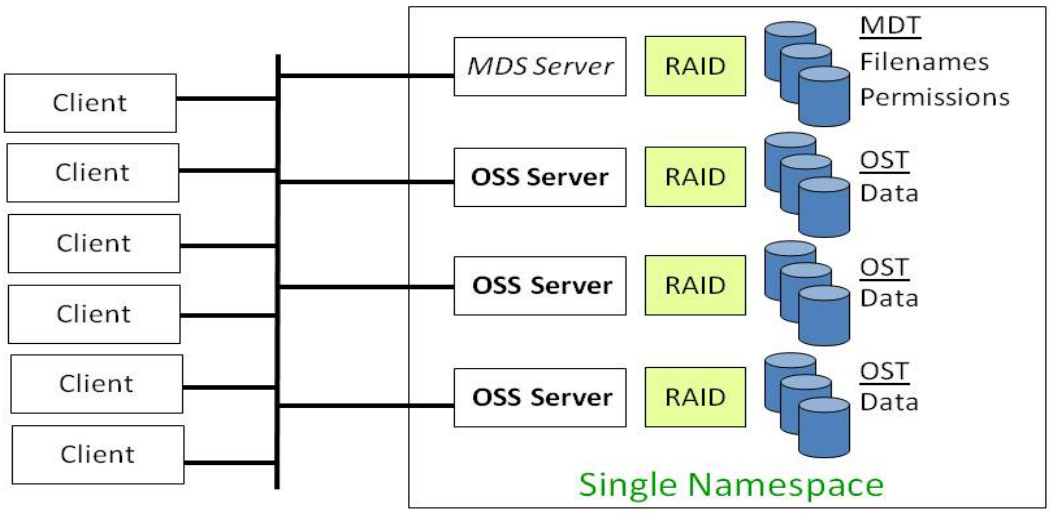
A parallel file system like Lustre delivers its performance and scalability by distributing (or striping) data across multiple access points, allowing multiple compute engines to access data simultaneously. A

Lustre installation consists of three key systems: the metadata system, the object system, and the compute clients.

The metadata system consists of the Metadata Target (MDT) and the Metadata Server (MDS). The MDT stores all metadata for the file system including file names, permissions, time stamps, and where the data objects are stored within the object system. The MDS is the dedicated server that manages the MDT. There is only one active MDS running at any time.

The object system consists of the Object Storage Target (OST) and the Object Storage Server (OSS). The OST provides storage for file object data, while the OSS manages one or more OSTs. There are typically multiple active OSSs at any time. Lustre is able to deliver increased throughput with the addition of OSSs, since each additional OSS provides additional networking and processing throughput and capacity. See Figure 1 for more information.

Figure 1. Lustre Overview



The Lustre client software is installed on the compute nodes to allow access to data stored within the Lustre file system. To the clients, the file system appears as a single namespace, a single entity, making application data access simple.

To summarize the functionality of the different elements of the Lustre parallel file system:

- Metadata Target (MDT) — Tracks the location of "chunks" of data
- Object Storage Target (OST) — Stores the "chunks," which are really just blocks on a disk)

- Lustre Client — Accesses the MDS to determine where a file is located (i.e., on which OSTs), and accesses the OSSs to read and write data

Typically, Lustre deployments and configurations are considered complex and time consuming. Open source Lustre is generally installed and administered via a command line interface, which may hinder a Systems Administrator who is not familiar with Lustre and therefore won't reap the benefits of such a powerful file system. The Dell | Terascala HPC Storage Solution removes these complexities and minimizes both Lustre deployment time and configuration so the file system can be tested and production ready as soon as possible.

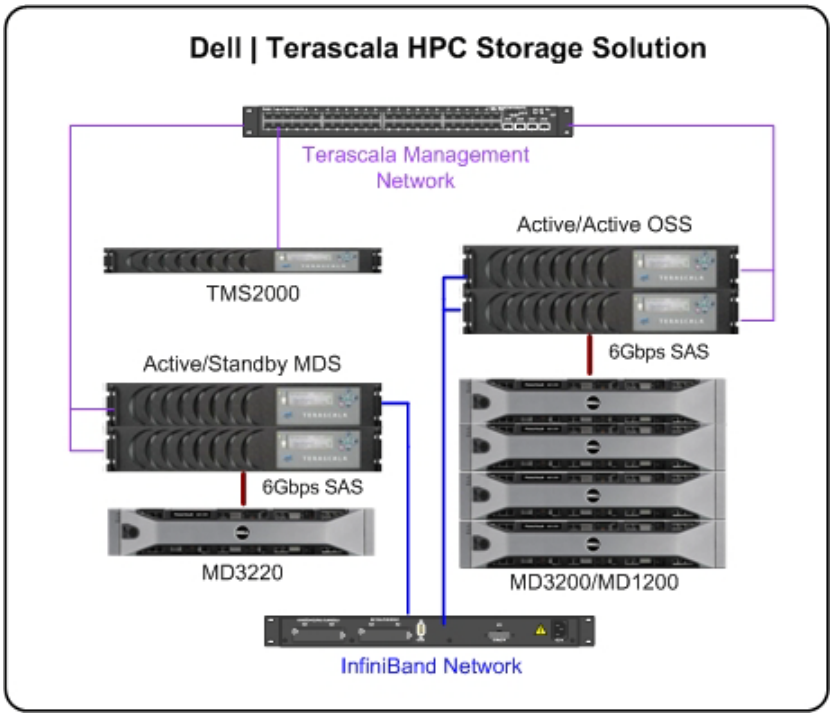
Dell | Terascala HPC Storage Solution Description

The key hardware systems in the Dell | Terascala HPC Storage Solution include the MDS (Metadata Servers), the OSS (Object Storage Servers), and the TMS2000 management appliance.

There are two ways in which the DT-HSS2 can be configured. The first is a non-redundant configuration that contains just one Terascala MDS server attached to a PowerVault MD3220 and one Terascala OSS server attached to a PowerVault MD3200. The second way the DT-HSS2 can be configured is in a redundant configuration where two Terascala MDS servers are attached to a PowerVault MD3220 and two Terascala OSS servers are attached to a PowerVault MD3200.

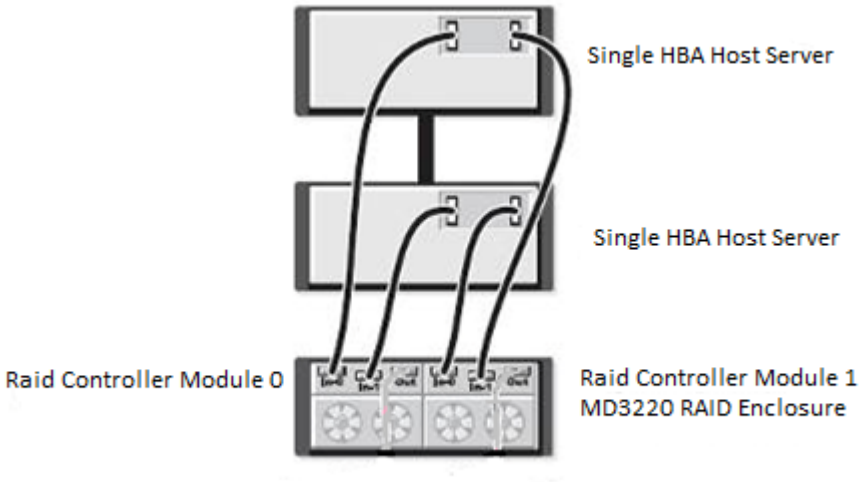
The MDS and OSS nodes are connected to the compute nodes via QDR InfiniBand. This allows file system traffic to traverse a high-speed, low-latency network to improve performance. A sample configuration of a 96TB DT-HSS2 is shown in Figure 2.

Figure 2. Sample DT-HSS2 96TB Entry Level Configuration



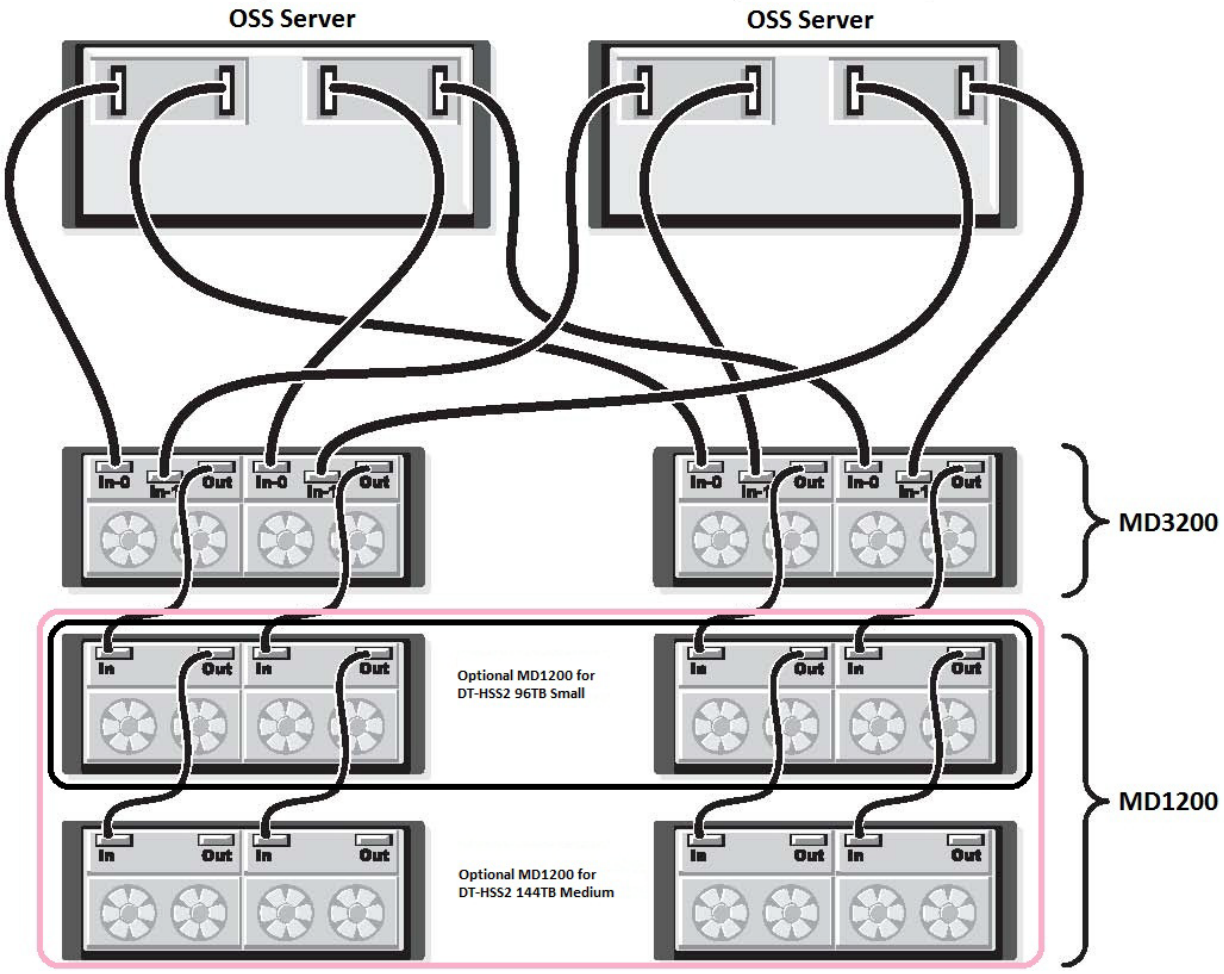
The MDS portion consists of two Terascala storage servers connected in an active/passive configuration to a MD3220 storage array as shown in Figure 3. The active/passive configuration provides high availability and reliability to metadata. It uses advanced monitoring and management to enable rapid, complete failover in case of MDS failure while insuring no spurious writes. The MDS can deliver up to 15,000 file creates/sec and store over 5 TB of metadata information in a RAID 10 volume on the MD3220 array.

Figure 3. Example MDS Cable Configuration



The DT-HSS2 has a building block approach that consists of stacking object pairs together. For redundant configurations, multiple OSS object pairs can be stacked together to increase both performance and capacity. The redundant OSS object pairs consist of two Terascala storage servers and two PowerVault MD3200 storage arrays cross connected in an active/active configuration as shown in Figure 4. The active/active configuration allows both servers to see data from both storage arrays. With this configuration, all of the object data can be accessed through redundant paths.

Figure 4. Example OSS Cable Configuration



The DT-HSS2 has three networks. The primary data network is the InfiniBand fabric on which the Lustre file system traffic traverses. This network is also, typically, the primary message-passing network used by the compute nodes. The DT-HSS2 is configured with QDR InfiniBand HCAs. This allows integration into existing DDR or QDR InfiniBand fabrics, utilizing QDR to DDR crossover cables as required. The second network is an Ethernet network that the Terascala Management Console uses to collect data from DT-HSS2 components and to present that data via the GUI. The third network is a private Ethernet network that provides a heartbeat between the MDS nodes and the OSS nodes used for failover traffic (for redundancy).

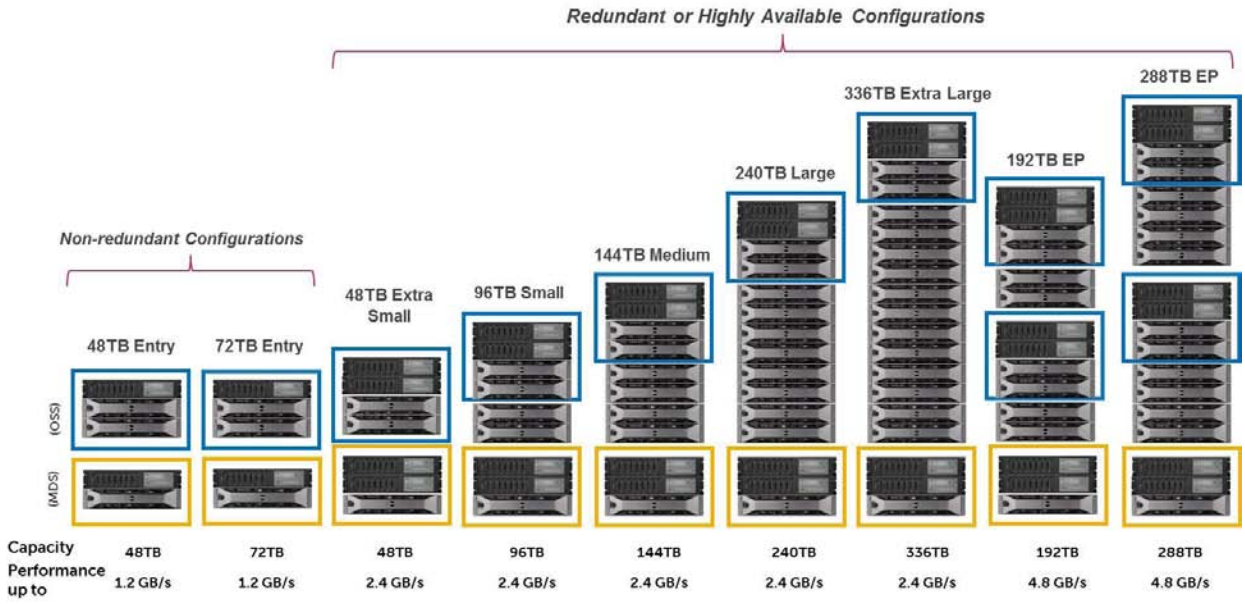
The complete Dell | Terascala HPC Storage Solution is managed through the TMS2000 management appliance. This appliance maintains management paths to all elements of the solution to capture system status and file system information and configuration. Users interact with the TMS2000 through the Terascala Management Console.

There are two ways the storage capacity can be expanded. The first is by adding PowerVault MD1200 storage arrays attached to the existing MD3200 arrays. The second is by adding additional OSS object pairs. When expanding capacity using the second method, both capacity and performance increase.

In addition to scaling capacity, the DT-HSS2 can also scale performance. Figure 5 shows a variety of configuration options. First are the DT-HSS2 non-redundant configurations. These configurations range from 48TB to 72TB of capacity, with large sequential write performance of approximately 1.2GB/s based on a single MDS and OSS server. Next are the DT-HSS2 redundant configurations. The redundant standard configurations scale from 48TB to 336TB of capacity, with large sequential write performance of approximately 2.4GB/s. The performance is increased due to an additional OSS within the configuration. Finally, the DT-HSS2 solutions consist of the Enhanced Performance (EP) configurations. These configurations scale from 192TB to 288TB of capacity. Again, by adding one object pair, the performance doubles to approximately 4.8GB/s for large sequential writes. Custom solutions are available to scale beyond the configurations covered in this paper.

The Dell | Terascala HPC Storage Solution provides a plug and play Lustre file system appliance by providing pre-configured storage arrays and integrating Lustre administration into the Terascala Management Console. Both Systems Administrators and cluster users alike can benefit from such a solution, because they can focus on the work at hand instead of file system and storage administration. For more details on DT-HSS2 configurations, see Appendix B.

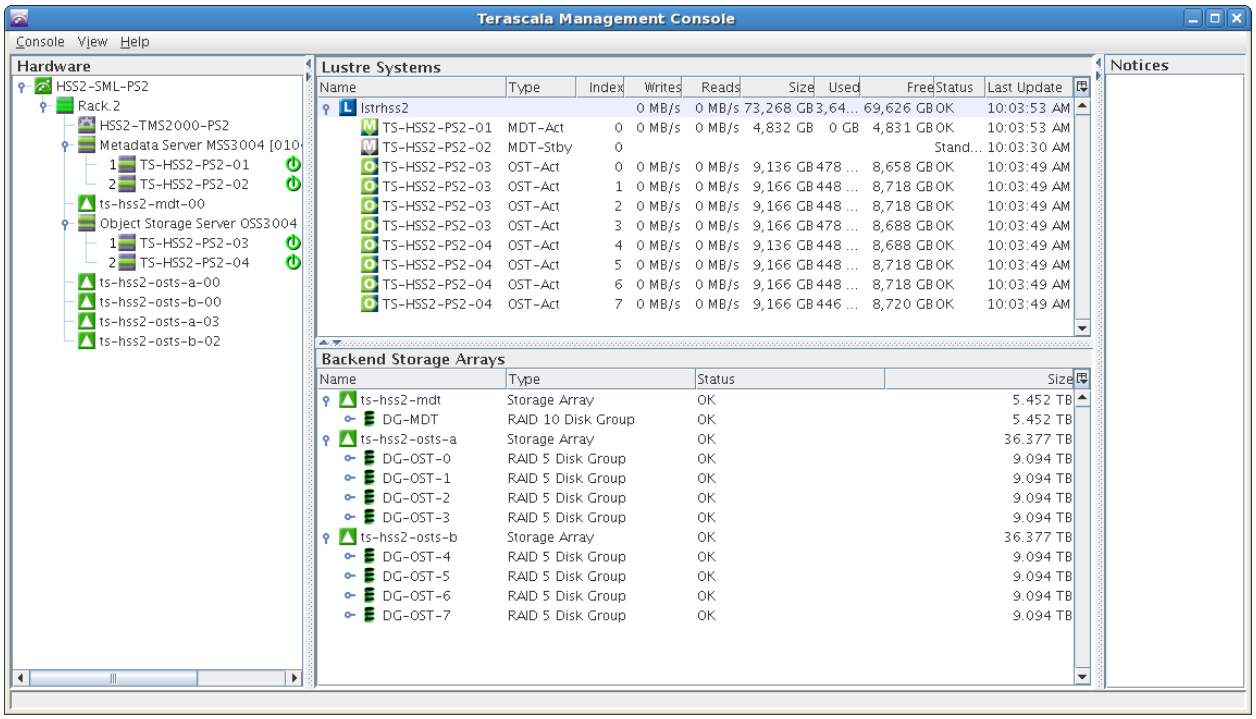
Figure 5. DT-HSS2 Expansion Options



Managing the Dell | Terascala HPC Storage Solution

The Terascala Management Console (TMC) takes the complexity out of administering a Lustre file system by providing a centralized GUI for management purposes. The TMC can be used as a tool to standardize the following actions: to mount and unmount the file system, to initiate failover of the file system from one node to another, and to monitor the performance of the file system and the status of its components. Figure 6 illustrates the TMC main interface.

Figure 6. Terascala Management Console Summary



The TMC is a Java-based application that can be run from any computer and that remotely manages the complete solution (assuming all security requirements are met). It provides a complete view of both the hardware and file system, while allowing complete management of the solution.

Figure 6 shows the initial window view of a DT-HSS2 system. In the left pane of the window are all the key hardware elements of the system. Each element can be selected to get additional information. In the center pane is a view of the system from a Lustre perspective, showing the status of the MDS and various OSS nodes. In the right pane is a message window that highlights any conditions or status changes. The bottom pane displays a view of the PowerVault storage arrays.

Using the TMC, many tasks that required complex CLI instructions, can now be completed easily with a few mouse clicks. The following figures show how to shut down a file system (see Figure 7), initiate a failover (see Figure 8) and monitor the MD3200 array (see Figure 9).

Figure 7. Unmount the File System in TMC

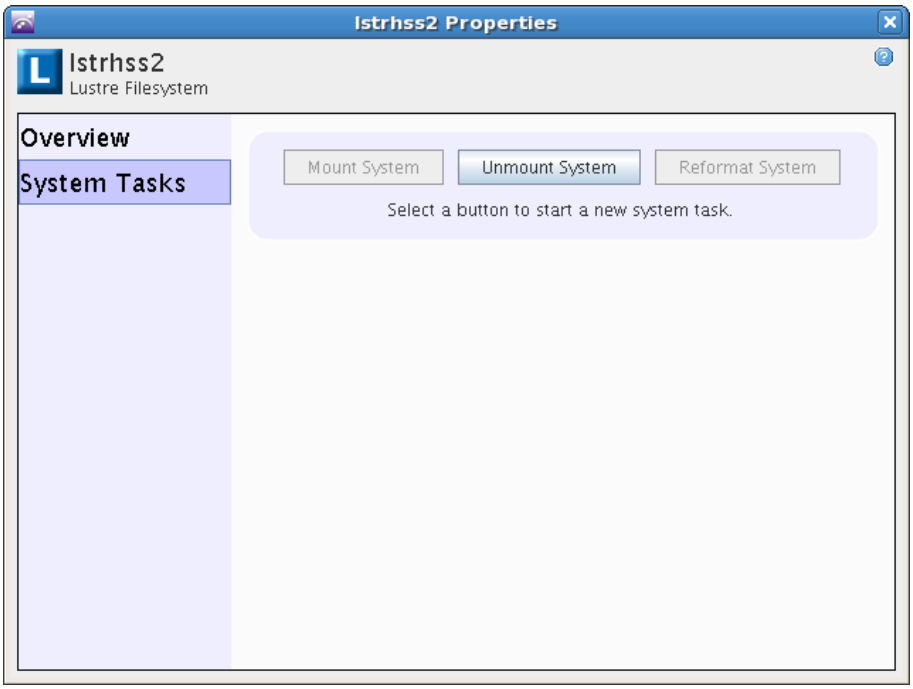


Figure 8. Initiate a Failover in TMC

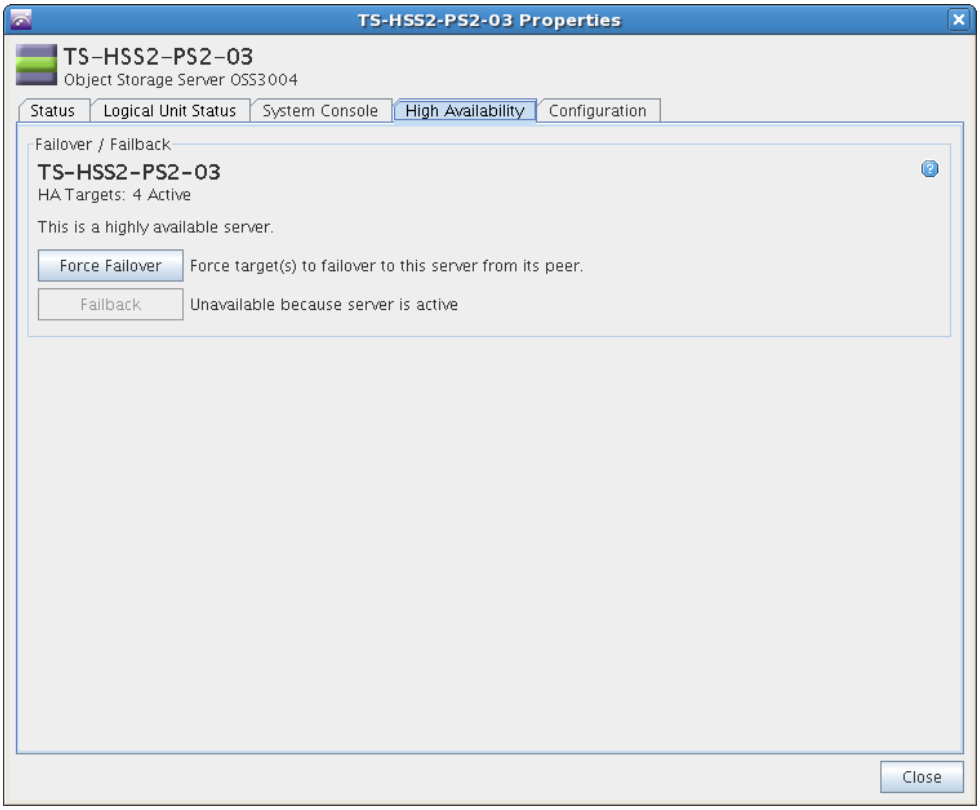
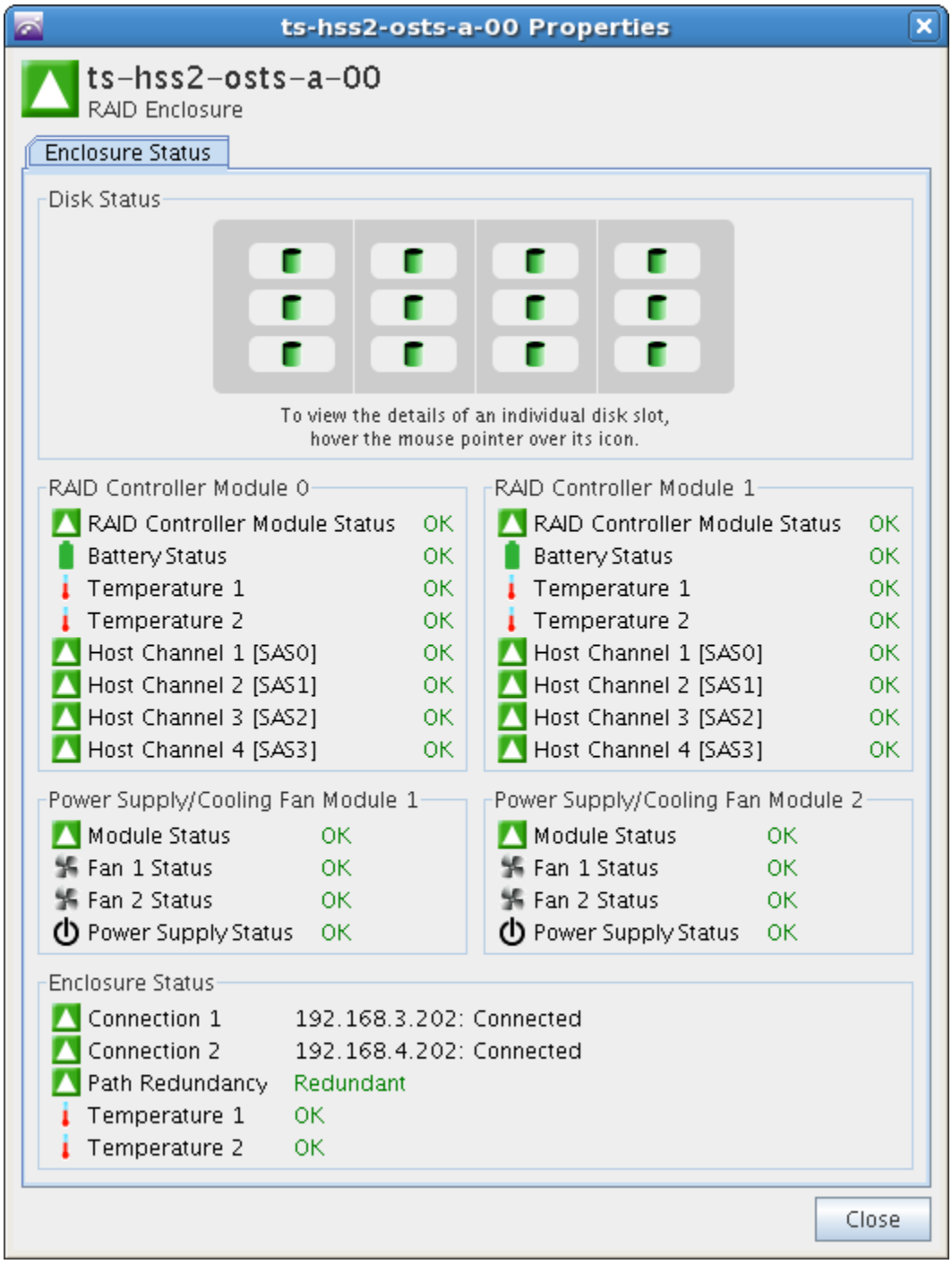


Figure 9. Monitor the MD3200 in TMC



PowerVault MD3200 / MD3220 and MD1200 Overview

DT-HSS2 utilizes three kinds of latest generation 6Gb/s SAS PowerVault direct-attached storage arrays. The first is the PowerVault MD3200 for the OSTs, the second is the PowerVault MD1200 that provides expansion storage for the PowerVault MD3200s, and, finally, the PowerVault MD3220 used for the MDT.

The Dell | Terascala HPC Storage Solution leverages the PowerVault MD3200 storage array as the primary data repository for the OSTs. The MD3200 is a modular disk storage unit that houses up to twelve 3.5-inch disk drives in a single 2U rack enclosure. The MD3200 direct-attached storage array is expandable by adding up to seven additional MD1200 expansion enclosures for a total of 96 drives of capacity. In DT-HSS2 offerings, both MD3200 and MD1200 arrays are configured with twelve 2TB near line SAS drives per shelf to provide the maximum number of spindles to deliver maximum performance. The MD3200 series array also introduces a new premium feature called *High Performance Tier*. This option is included by default on MD3200 arrays as part of the DT-HSS2 solution, and it significantly improves performance for large sequential reads and writes.

The MDS is attached to an MD3220 that serves as the MDT. The MD3220 can house up to twenty-four 2.5-inch drives in a single 2U rack enclosure. The MD3220 provides an efficient direct-attached storage solution for the MDT, because it has more spindles to benefit metadata performance.

Both direct-attached storage arrays support removing and replacing redundant components without disruption to operation. In the rare event of component failure, disk drives, storage controllers, power supplies, and cooling fan modules are all hot-pluggable (i.e., those components can be removed and replaced for easy repair while the system remains up and running). These disk storage array enclosures contain two integrated power supply and fan modules; if one is removed, the other can handle the demand, virtually eliminating downtime to the storage array. The storage arrays are architected to avoid single points of failure. Both the MD3200 and the MD3220 direct-attached arrays are loaded with data-protection features to keep applications up and running. A key feature of both the MD3200 and the MD3220 are the two active-active RAID controllers that provide the logic to govern everything that occurs within both of the disk storage arrays, by performing necessary calculations, controlling I/O operations, handling communication with management applications, and storing firmware.

The performance, reliability, and expandability of both the MD3200 and MD3220 provide a highly effective base for the Dell | Terascala HPC Storage Solution.

Integrating Dell | Terascala HPC Storage Solution into a High Performance Cluster

Dell, Platform Computing, and Terascala collaborated to create a Terascala Lustre Kit that allows seamless integration and deployment of Lustre clients onto compute nodes in a cluster. To ensure consistency of the software and configuration on the cluster, use Platform Cluster Manager (PCM) 2.0.1 to deploy the Terascala Lustre Kit and manage Lustre software. As new patches are released, revision control can be performed by using the kit deployment method as well.

The following integration steps were performed using PCM 2.0.1 Dell Edition as the cluster management software. PCM provides the cluster middleware solution that eases deployment and administration of the cluster. Platform Computing and Dell have partnered to test and validate PCM on various Dell solutions to ensure software and hardware compatibility. Some of the key features of PCM are the inclusion of common HPC tools (compilers, MPI stacks, etc.), Web-based management, bare-metal cluster provisioning, simplified node management, and job submission/management tools.

Terascala Lustre Kit

PCM software uses the following terminology to describe software provisioning concepts:

1. Installer Node – Runs services such as DNS, DHCP, HTTP, TFTP, etc.
2. Components – RPMS
3. Kits – Collection of components
4. Repositories – Collection of kits
5. Node Groups – Allow association of software to a particular set of compute nodes

The following is an example of how to integrate a Dell | Terascala HPC Storage Solution into an existing PCM cluster.

1. Access a list of existing repositories on the head node:

```
# kusu-repoman -l
Repo name:      rhel-5.5-x86_64
Repository:     /depot/repos/1000
Installers:     172.20.0.1;192.168.254.83
Ostype:         rhel-5-x86_64
Kits:           base-2.0-1-x86_64, dell-vendor-5.5-1-x86_64,
                java-jre-1.5.0-16-x86_64, nagios-2.12-7-x86_64,
                PCM_GUI-2.0-1-x86_64, platform-hpc-2.0-3-x86_64,
                platform-isf-ac-1.0-1-x86_64, platform-lsf-7.0.6-1-x86_64,
                platform-mpi-7.1-1-x86_64, platform-ofed-1.5.1-1-x86_64,
                platform-rtm-2.0.1-1-x86_64, rhel-5.5-x86_64,
                mellanox-1.5.1-1-x86_64
```

   The key here is the “Repo name:” line.

2. Add the Terascala kit to the cluster:

```
# kusu-kit-install -r "rhel-5.5-x86_64" kit-terascala-lustre-1-
1.x86_64.iso
```

3. Confirm the kit has been added:

```
# kusu-kitops -l
```

4. Associate the kit to the compute node group on which the Lustre client should be installed:
   a. Launch **ngedit** at the console: `# ngedit`
   b. On the **Node Group Editor** screen, select the compute node group to add the Lustre client.
   c. Select the **Edit** button on the bottom.
   d. Accept all default settings until you reach the **Components** screen.
   e. Use the down arrow, and select the **Terascala Lustre** kit.
   f. Expand and select the **Terascala Lustre** kit component.
   g. Accept the default settings, and on the **Summary of Changes** screen, accept the changes and push the packages out to the compute nodes.

On the front-end node, there is now a **/root/terascala** directory that contains sample directory setup scripts. There is also a **/home/apps-lustre** directory that contains Lustre client configuration parameters. This directory contains a file that the Lustre file system startup script uses to optimize clients for Lustre operations.

The Terascala kit utilizes the IPoIB network so clients can access the Lustre file system over an InfiniBand network. Also, the Terascala kit installs patchless Lustre clients. The Lustre client is version 1.8.2 for the Terascala 1.1 kit.

Enter the following to verify that the Lustre file system is mounted and accessible from all clients:

```
# pdsh -a service lustre start
# pdsh -a mount | grep lustre | dshbak -c
```

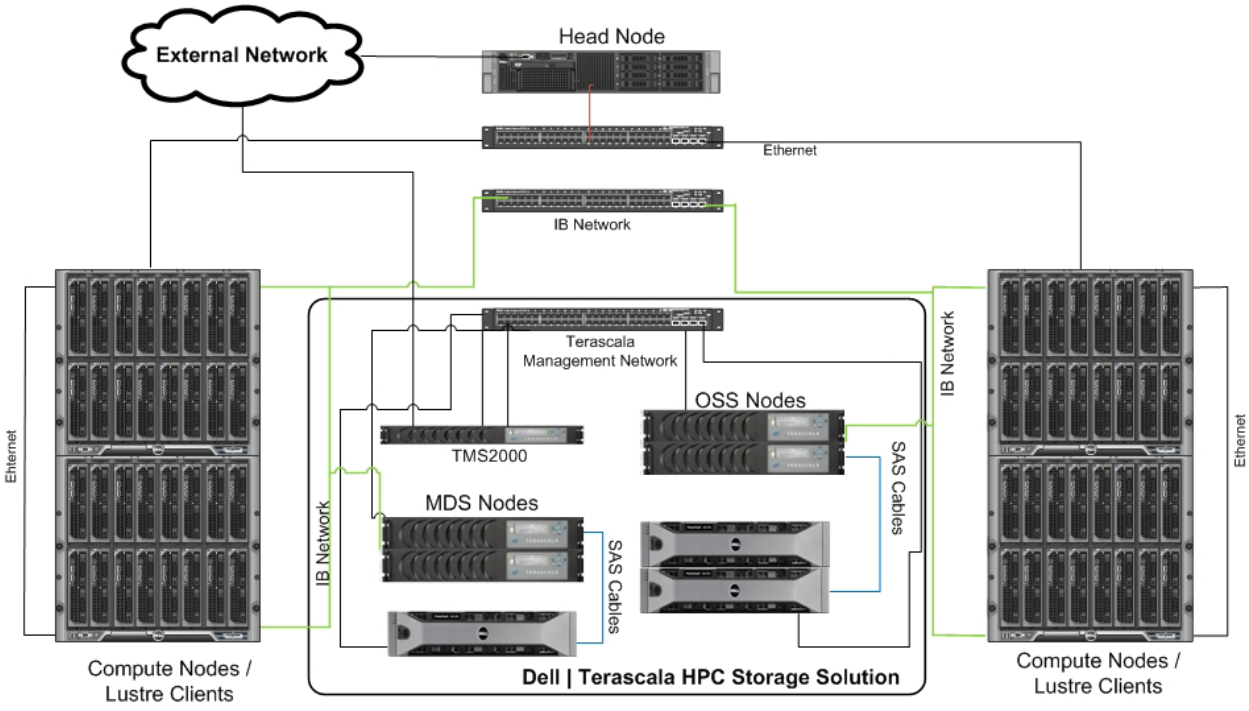
Performance Studies

The performance studies described in this paper characterize the types of applications that benefit from using DT-HSS2 as a storage appliance. Using a Dell HPC compute test bed, a number of performance studies were performed. The goal was to stress a DT-HSS2 configuration with different types of workloads to both determine the maximum performance and define how well the solution can sustain that performance.

The study tested different aspects of the solution, specifically: Throughput, I/O Operations Per Second (IOPS), and Metadata Operations of the system. Throughput is the amount of data that can be carried from one point to another point within a particular time frame. IOPS results can be influenced by multiple factors including number of disks, average seek time of a disk, latency, and rotational speed of the disks. IOzone and IOR benchmarks were used to measure throughput and IOPS. Metadata is data about data. The benchmark *mdtest* was used to examine the speed with which DT-HSS2 creates files and directories, stats files, and directories and, finally, deletes files and directories.

There are two types of file access methods used in these benchmarks. The first file access method is N-to-N, where every thread of the benchmark (N clients) writes to a different file (N files) on the storage system. IOzone and IOR can both be configured to use the N-to-N file-access method. The second file access method is N-to-1, which means that every thread writes to the same file (N clients, 1 file). IOR can use MPI-IO, HDF5, or POSIX to run N-to-1 file-access tests. N-to-1 testing determines how the file system handles the overhead introduced with multiple concurrent requests when multiple clients (threads) write to the same file. The overhead encountered comes from threads dealing with single-file locking and serialized writes. See Appendix A for examples of the commands used to run these benchmarks. Figure 10 shows a diagram of the cluster configuration used for this study.

Figure 10. DT-HSS2 Cluster Diagram



In this study, the storage used is a DT-HSS2 96TB small configuration:

- Two MDSs connected to a single MD3220 that has twenty-four 500GB, 7,200 RPM Nearline SAS drives
- Two OSSs connected in a redundant configuration to two MD3200s
- Each OSS MD3200 has two OSTs (virtual disks) for a total of four OSTs
  - Each OSS has an MD1200 expansion enclosure with two OSTs (virtual disks) for a total of eight OSTs

Sixty-four PowerEdge M610 blade servers were used as compute nodes running Platform PCM 2.0.1, which includes Red Hat Enterprise Linux® 5.5 and Platform OFED kit (v.1.5.1). The compute nodes also ran a patchless Lustre client version 1.8.4. The compute nodes are connected to a QDR InfiniBand switch via QDR InfiniBand cables to which the MDS and OSS servers are connected. For a more detailed description of the cluster components, see Table 1.

Table 1 - Cluster Setup

Compute Nodes (PowerEdge M610)	
Processor	Two Intel Xeon™ X5650 2.67 GHz quad core processors
Memory	24 GB (Six 4 GB DDR3-Registered DIMMs)
OS	Red Hat Enterprise Linux 5 U5
Lustre	Lustre 1.8.4 Patch-less Client
Kernel	2.6.18-194.el5
BIOS	2.0.14

DT-HSS2 Configuration	
Configuration	96TB Small
OSS Nodes	Terascala Servers
OSS Storage Array	2 x MD3200 / 2 x MD1200
Drives in OSS Storage Arrays	48 3.5" 2TB 7200RPM Near Line SAS
MDS Nodes	Terascala Servers
MDS Storage Array	1 x MD3220
Drives in MDS Storage Array	24 2.5" 500GB 7200RPM Near Line SAS
InfiniBand Network	
Terascala Servers	Mellanox ConnectX-2 QDR HCA
Compute Nodes	Mellanox ConnectX-2 QDR InfiniBand HCA
External QDR IB Switch	Mellanox 36 Port IS5030
IB Switch Connectivity	QDR Cables

IOzone Benchmark

IOzone is an industry-standard benchmark that can generate many types of loads on a disk subsystem. In this study, IOzone is used to test the sequential read and write throughput of the Dell | Terascala HPC Storage Solution. IOzone can be run in one of two ways: In standalone mode where it launches locally and can spawn X number of threads and report the throughput, or launched in cluster mode. In cluster mode, IOzone spawns a master process on the head node and launches IOzone threads on the compute nodes that each write to an individual file. When the IOzone tests are finished, the IOzone master process collects the results and reports the aggregate bandwidth. The study used IOzone 3.283, which is available at www.iozone.org.

For this study, the caches on the compute nodes, the OSS server, and the MD3200 were saturated for all tests. For example, for large sequential writes, a file size of 48GB was used for each IOzone thread to ensure that any cache effects in the system were minimized and the tests would generate realistic results.

Figure 11. IOzone N-to-N Sequential Read / Write Performance

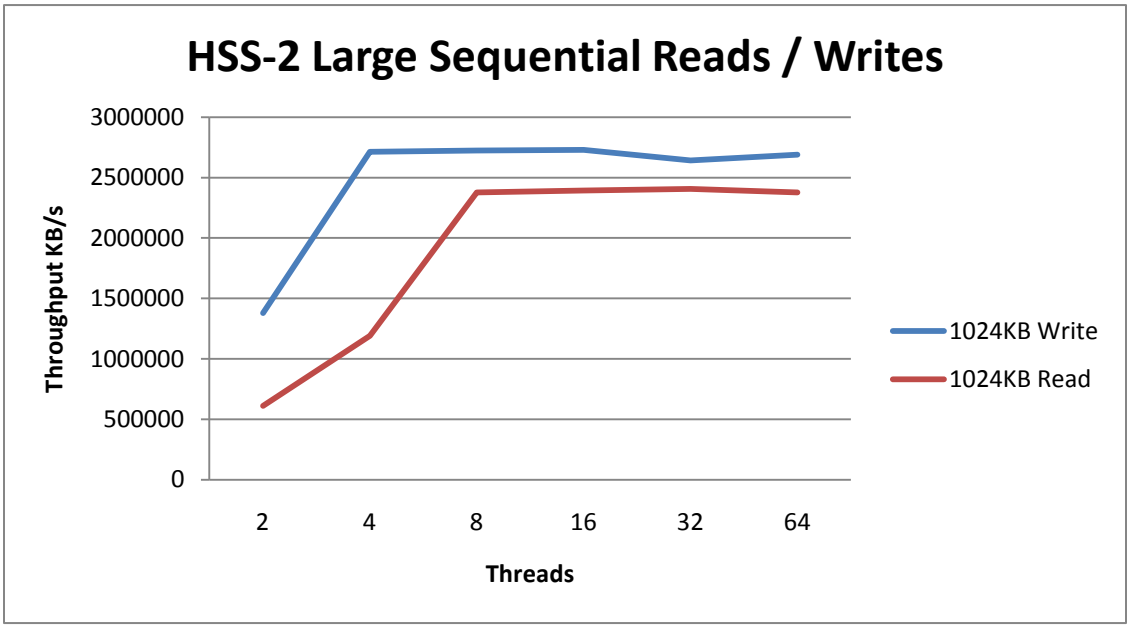


Figure 11 shows that the DT-HSS2 small configuration can sustain sequential write bandwidth of 2,600 MB/s starting at four nodes. The quick climb of the graph to maximum performance can be attributed to Lustre file system traffic traveling over the InfiniBand network. The performance is maintained even with a 64 concurrent requests, which satisfies the I/O requirements of most parallel applications. The request size is set to 1024KB to align with the Lustre 1MB RPC packet size. This test was run on a DT-HSS2 small configuration that has one active-active pair of base object storage. When additional base objects are added, the performance increases accordingly. Figure 11 also shows that sequential reads can saturate the link at approximately four threads at around 2400MB/s with *cold cache*, where both the client and server cache are purged between each run.

Figure 12. IOzone IOPS – Random Reads / Writes

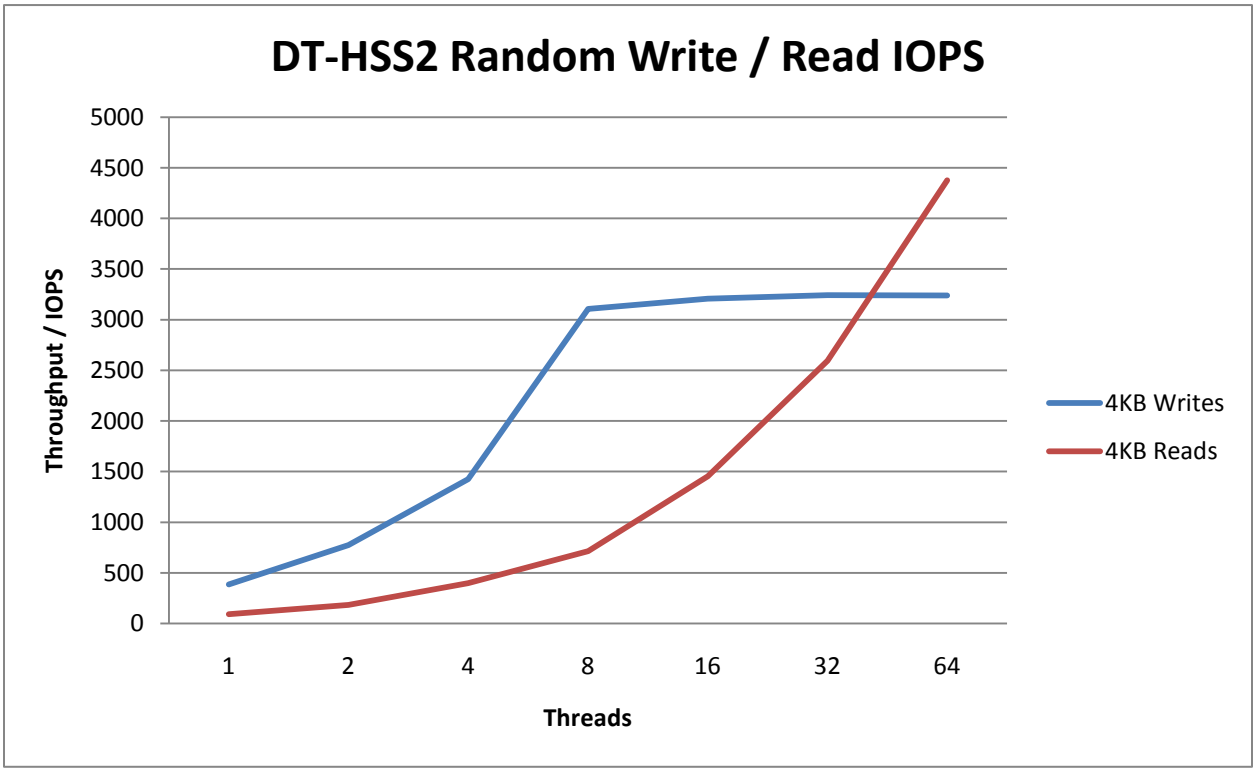


Figure 12 shows that random writes are saturated at sixteen threads at around 3,200 IOPS using a 4KB request size. A 4KB request size was used because it aligns to the Lustre file system block size. It is possible to increase IOPS by adding more MD1200s to the solution, because adding more spindles increases the performance of the I/O pattern.

IOR N-to1 Testing

IOR is an industry standard tool used to benchmark parallel file systems using POSIX, MPI-IO or HDF5 interfaces. In this study, IOR was run using the POSIX API to test the raw throughput of the file system and avoid the overhead of the HDF5 or MPI-IO, which involve using collectives and file views. IOR is launched with *mpirun,* and creates a thread on each compute node where each thread can either write to a separate file (N-to-N) or all threads can write to a single file (N-to-1). For this study, N-to-1 was used to determine the level of overhead encountered based on the file locking and serialization involved in coordinating multiple threads writing to a single file.

IOR benchmark version 2.10.3 was used in this study, and is available at http://sourceforge.net/projects/ior-sio/. The MPI stack used for this study was Open MPI version 1.4.1, and was provided as part of the PCM OFED kit.

Figure 13. IOR – Parallel IO Using Sequential Reads and Writes

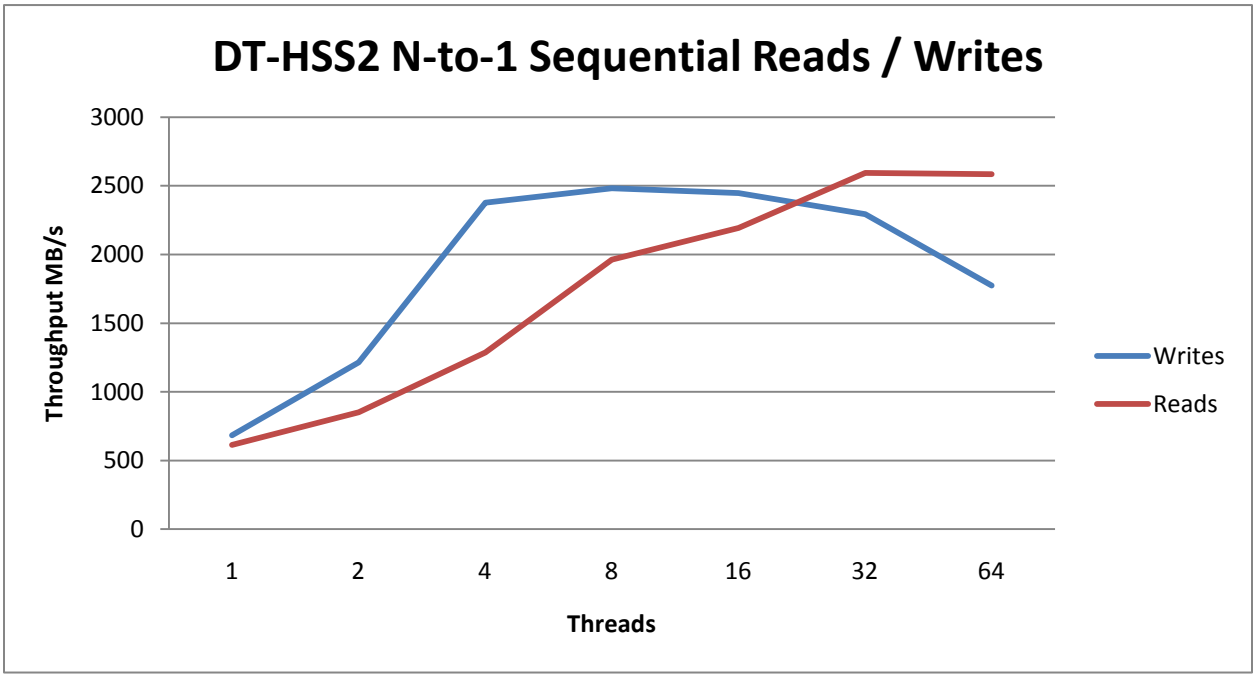


Figure 13 shows that the link can be saturated at around 2,500 MB/s with sequential N-to-1 writes, and the link is able to sustain that throughput up to approximately 32 nodes. Although IOR performs N-to-1 tests, despite some overhead because of locking and operation serialization on a single file, the write throughput is still 90% of the N-to-N IOzone tests. Figure 13 illustrates that DT-HSS2 performance satisfies the most critical parallel I/O requirements. Lustre is optimized for better performance with writes. The Lustre driver is able to profile write requests, and when appropriate it will collect those smaller requests into larger requests and then issue a single write that reduces locks and improves performance.

Metadata Testing

Metadata is data about data. *Mdtest* is an MPI-coordinated benchmark that performs `create`, `stat`, and `delete` operations on files and directories and then provides timing results. This study used mdtest version 1.8.3, which is available at http://sourceforge.net/projects/mdtest/. The proper way to launch *mdtest* is with *mpirun,* which creates a thread on each compute node. Mdtest can be configured to compare metadata performance between directories and files. Version 1.8.3 allows all three options to be separated, which enables purging of the cache between each run to provide the most accurate results. Millions of files and directories were created to ensure that the cache on the storage controller was saturated.

On a Lustre file system, the directory metadata operations are all performed on the MDT; however, the file metadata operations require the involvement of both MDTs and OSTs. The wider the stripe width of a file, the more operations are required to complete a metadata request. As a result, this study selected three scenarios to represent metadata performance. The first is directory metadata operations within the MDT. The second is file metadata operations demonstrating the performance of files that are not striped and hosted on a single OST. The third is file metadata operations that demonstrate the performance of files that are striped across eight OSTs. The performance results (Figure 14, Figure 15, and Figure 16) indicate that metadata operations of create/stat/remove of

directories have less overhead than those of files which require two distinct operations: one on the MDT and one on the OSTs. Also, file operation overhead increases as the striping width increases. For example, a stat operation of a file striped across multiple OSTs requires that the clients contact each OST to calculate the file size. A file striped to a single OST performs better than a file striped to eight OSTs. Considering the metadata overheard ofstriped files, Lustre also implements optimizations to reduce the latency from operations requiring actions from both OSS and MDS. For example, the MDS has a cache of pre-allocated blocks on each OST for file creation, causing the file create on one OST to have similar performance results as a file create that is striped across multiple OSTs.

One exception is that a file remove is faster than a directory remove, because to remove a directory, the client must first contact the MDS to open the directory and ensure it is empty prior to removal. This involves multiple RPCs between the client and the MDS, which is less efficient than file removal.

Figure 14. Metadata File / Directory Create

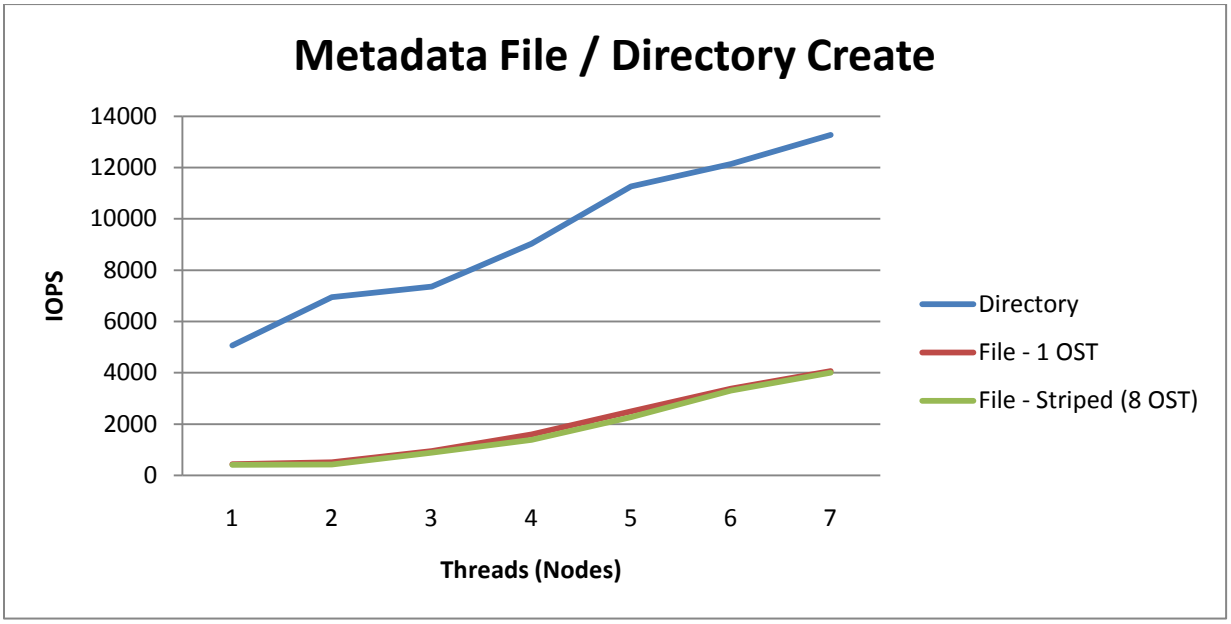


Figure 15. Metadata File / Directory Stat

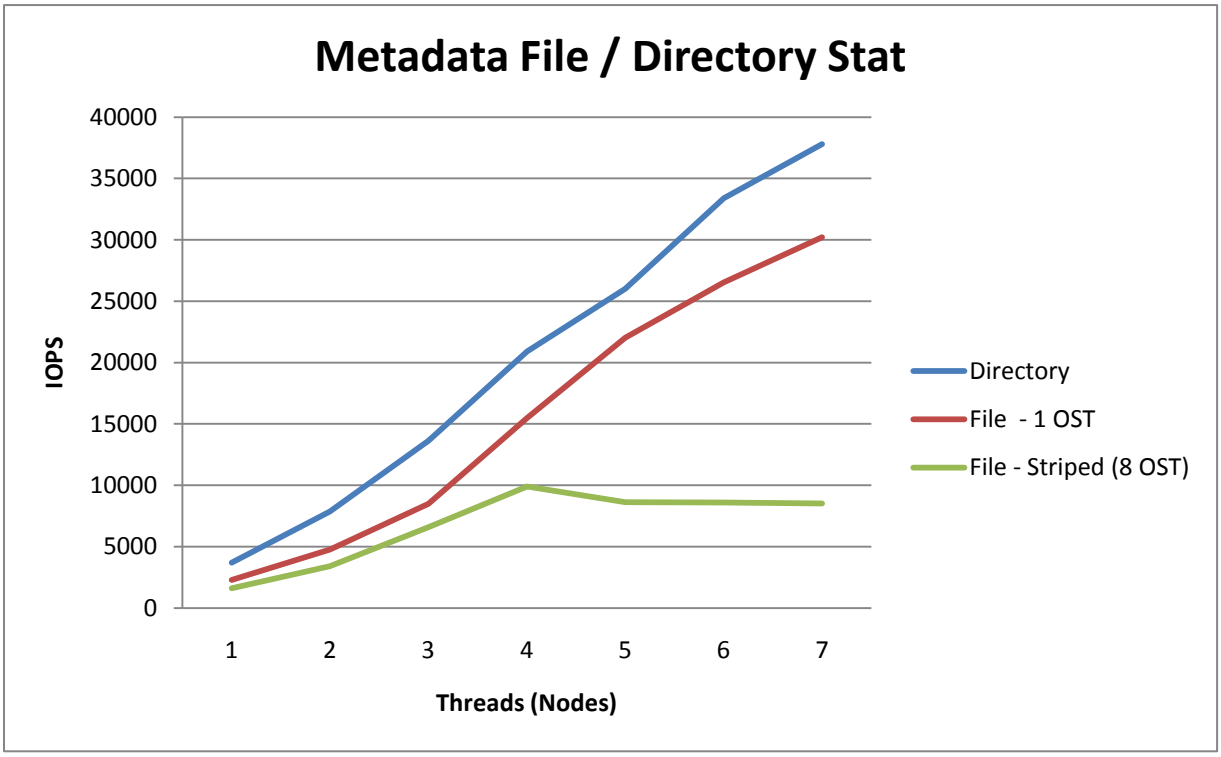
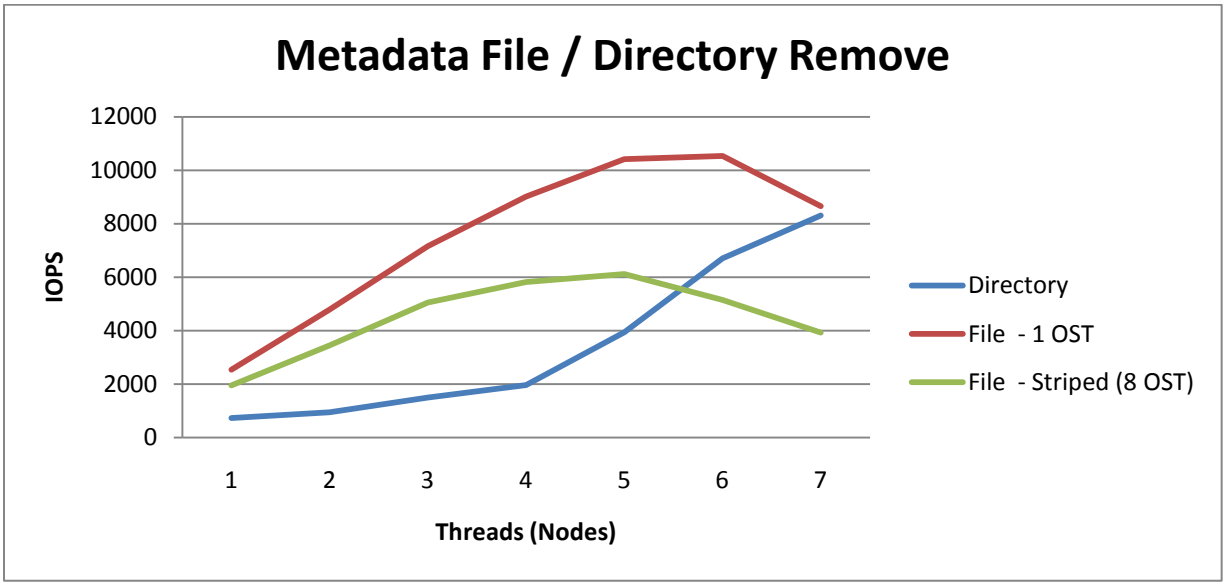


Figure 16. Metadata File / Directory Remove



Conclusion

Some applications require a scalable high-performance and high-capacity file system as a scratch space to ensure that user jobs run within a reasonable time frame. Apart from providing high performance

and ability to scale, the file system solution must be easy to deploy, manage, and integrate into the compute cluster. The Dell | Terascala HPC Storage Solution (DT-HSS2) addresses these issues and provides a scratch space appliance for HPC deployments.

By providing 48TB to 336TB of storage within a single 42U rack and 1.2 GB/s to 4.8 GB/s throughput in packaged configurations, the DT-HSS2 meets the capacity and performance requirements for HPC environments. It is also architected using open and standard building blocks to deliver the price and performance levels required in HPC environments.

The DT-HSS2 is a plug-and-play appliance that is managed via the TMC. This console provides a centralized administration interface that can be used to perform all necessary actions on the Lustre file system. By using the TMC, there is no need to manually issue Lustre commands or become a Lustre file system expert — that is all addressed by the console.

The performance studies in the paper clearly show that the DT-HSS2 provides very high throughput and IOPS for both N-to-N and N-to-1 file access types. Industry-standard tools like IOzone, IOR, and *mdtest* provide a way to capture performance results from the DT-HSS2 and also to characterize the types of applications that will run well with this type of storage backend.

The DT-HSS2 provides a scalable, easy-to-administer storage array that removes the complexity associated with deploying and administering Lustre. By providing a fully-supported and pre-configured hardware and software Lustre file system appliance, the DT-HSS2 is quick to deploy and easy to maintain. From the data presented in this paper, it is clear that the DT-HSS2 solution delivers excellent performance, helping users to maximize the application performance and server utilization.

References

Dell | Terascala HPC Storage Solution Brief
http://i.dell.com/sites/content/business/solutions/hpcc/en/Documents/Dell-terascala-hpcstorage-solution-brief.pdf

Dell PowerVault MD3200 / MD3220
http://www.dell.com/us/en/enterprise/storage/powervault-md3200/pd.aspx?refid=powervault-md3200&s=biz&cs=555

Dell PowerVault MD1200
http://configure.us.dell.com/dellstore/config.aspx?c=us&cs=555&l=en&oc=MLB1218&s=biz

Dell HPC Solutions Home Page
http://www.dell.com/hpc

Dell HPC Wiki
http://www.HPCatDell.com

Terascala Home Page
http://www.terascala.com

Platform Computing Home Page
http://www.platform.com

Lustre Home Page
http://wiki.lustre.org/index.php/Main_Page

Appendix A: Benchmark Command Reference

This section describes the commands used to benchmark the Dell | Terascala HPC Storage Solution 2.

IOzone

IOzone Sequential Writes –

```
/root/iozone/iozone -i 0 -c -e -w -r 1024k -s 45g -t $THREAD -+n -+m ./hosts
```

IOzone Sequential Reads -

```
/root/iozone/iozone -i 1 -c -e -w -r 1024k -s 45g -t $THREAD -+n -+m ./hosts
```

IOzone IOPS Random Reads / Writes –

```
/root/iozone/iozone -i 2 -w -O -r 4k -s $SIZE -t $THREAD -I -+n -+m ./hosts
```

Description of command line arguments:

IOzone Command Line Arguments	Description
-i 0	Write test
-i 1	Read test
-i 2	Random IOPS test
-+n	No retest
-c	Includes close in the timing calculations
-e	Includes flush in the timing calculations
-r	Records size
-s	File size
+m	Location of clients to run IOzone on when in clustered mode
-I	Use O_Direct
-w	Does not unlink (delete) temporary file
-+n	No retests selected
-O	Return results in OPS

By using *-c* and *-e* in the test, IOzone provides a more realistic view of what a typical application does.

The O_Direct command line parameter allows us to bypass the cache on the compute node on which we are running the IOzone thread.

IOR

IOR Writes –

```
mpirun -np $i --hostfile hosts $IOR -a POSIX -i 3 -d 32 -e -k -w -o $IORFILE
-s 1 -b $SIZE -t 1m
```

IOR Reads –

```
mpirun -np $i --hostfile hosts $IOR -a POSIX -i 3 -d 32 -e -k -r -o $IORFILE
-s 1 -b $SIZE -t 1m
```

Description of command line arguments:

IOR Command Line Arguments	Description
-np	Number of processes
--hostfile	File with names of compute nodes
"$IOR"	Location of IOR executable
-a	Defines which API to use
-i	Number of repetitions of test
-d	Delay between repetitions in seconds
-k	Keeps the file on program exit
-r	Reads the existing file
-e	fsync -- perform fsync upon POSIX write close
-o	Name of test file
-s	Number of segments
-b	Contiguous bytes to write per test
-t	Size of transfer in bytes
$SIZE	Size of the file that IOR will write / read
-r	Writes a file

mdtest - Metadata

Create Files -

```
mpirun -np $THREAD --nolocal --hostfile $HOSTFILE $MDTEST -d $FILEDIR -i 6 -b
$DIRS -z 1 -L -I $FTP -y -u -t -F -C
```

Stat files -

```
mpirun -np $THREAD --nolocal --hostfile $HOSTFILE $MDTEST -d $FILEDIR -i 6 -b
$DIRS -z 1 -L -I $FTP -y -u -t -F -R -T
```

Remove files -

```
mpirun -np $THREAD --nolocal --hostfile $HOSTFILE $MDTEST -d $FILEDIR -i 6 -b
$DIRS -z 1 -L -I $FTP -y -u -t -F -r
```

Create Directories –

```
mpirun -np $THREAD --nolocal --hostfile $HOSTFILE $MDTEST -d $FILEDIR -i 6 -b
$DIRS -z 1 -L -I $FTP -y -u -t -D -C
```

Stat Directories –

```
mpirun -np $THREAD --nolocal --hostfile $HOSTFILE $MDTEST -d $FILEDIR -i 6 -b
$DIRS -z 1 -L -I $FTP -y -u -t -D -R -T
```

Remove Directories –

```
mpirun -np $THREAD --nolocal --hostfile $HOSTFILE $MDTEST -d $FILEDIR -i 6 -b
$DIRS -z 1 -L -I $FTP -y -u -t -D -r
```

Description of command line arguments:

mdtest Command Line Arguments	Description
-d $FILEDIR	Directory where the tests will run
-F	Perform test on files only (no directories)
-i	Number of iterations each test runs
-y	Ensures the file gets synced after each operation
-D	Only test directories
-u	Creates a unique working directory for each task
-t	Time the unique working directory overhead
$HOSTFILE	Location of hostfile
$MDTEST	Location of mdtest
-np $THREAD	How many threads to create
-b $DIRS	How many directories to create
-I $FPT	Number of items per directory in tree
-C	Only create files/dirs
-L	Files only at leaf level of tree
-z	Depth of hierarchical directory structure
-R	Randomly stat files
-r	only remove files or directories left behind by previous runs
-T	only stat files/dirs

Appendix B: Dell | Terascala HPC Storage Solution (DT-HSS2) Details

The DT-HSS2 is designed to be delivered in a set number of combinations and expansion units. The first two configurations are the DT-HSS2 non-redundant configurations:

48TB Entry Level (Non-Redundant)	**72TB Entry Level (Non-Redundant)**	**48TB Extra Small (Redundant)**
• 48 TB (using 24 2TB NL SAS Drives) of raw space for user data and 6TB (using 12 500GB NL SAS drives) of raw space for metadata • Up to 1.2 GB/s performance • Non redundant OSS and MDS servers. • One Object Storage Server connected to a dual-controller PowerVault MD3200 with PowerVault MD1200 as expansion storage • One Metadata Storage Server connected to a single-controller PowerVault MD3220 • Terascala Management software running on a dedicated management server • Installation service and three years of Dell ProSupport • Upgradable to a fully-redundant configuration	• 72TB (using 36 2TB NL SAS Drives) of raw space for user data and 6TB (using 12 500GB NL SAS drives) of raw space for metadata • Up to 1.2 GB/s performance • Non redundant OSS and MDS servers • One Object Storage Server connected to a dual-controller PowerVault MD3200 with two MD1200s as expansion storage • One Metadata Storage Server connected to a single-controller PowerVault MD3220 • Terascala Management software running on a dedicated management server • Installation service and three years of Dell ProSupport • Upgradable to a fully-redundant configuration	• 48TB (using 24 2TB NL SAS Drives) of raw space for user data and 12TB (using 24 500GB NL SAS drives) of raw space for metadata • Up to 2.4 GB/s performance • Redundant OSS and MDS servers • Two active/active Object Storage Servers connected to two dual-controller PowerVault MD3200s • Two active/passive Metadata Storage Servers connected to a dual-controller PowerVault MD3220 • Terascala Management software running on a dedicated management server • Installation service and three years of Dell ProSupport
96TB Small (Redundant)	**144TB Medium (Redundant)**	**240TB Large (Redundant)**
• 96TB (using 48 2TB NL SAS Drives) of raw space for user data and 12TB (using 24 500GB NL SAS drives) of raw space for metadata • Up to 2.4 GB/s performance • Redundant OSS and MDS servers • Two active/active Object Storage Servers connected to two dual-controller PowerVault MD3200s, each with one MD1200 as expansion storage	• 144TB (using 72 2TB NL SAS Drives) of raw space for user data and 12TB (using 24 500GB NL SAS drives) of raw space for metadata • Up to 2.4 GB/s performance • Redundant OSS and MDS servers • Two active/active Object Storage Servers connected to two dual-controller PowerVault MD3200s, each with two MD1200 arrays as expansion storage	• 240TB (using 120 2TB NL SAS Drives) of raw space for user data and 12TB (using 24 500GB NL SAS drives) of raw space for metadata • Up to 2.4 GB/s performance • Redundant OSS and MDS servers • Two active/active Object Storage Servers connected to two dual-controller PowerVault MD3200s, each with four MD1200 arrays as expansion storage

• Two active/passive Metadata Storage Servers connected to a dual-controller PowerVault MD3220 • Terascala Management software running on a dedicated management server • Installation service and three years of Dell ProSupport	• Two active/passive Metadata Storage Servers connected to a dual-controller PowerVault MD3220 • Terascala Management software running on a dedicated management server • Installation service and three years of Dell ProSupport	• Two active/passive Metadata Storage Servers connected to a dual-controller PowerVault MD3220 • Terascala Management software running on a dedicated management server • Installation service and three years of Dell ProSupport
336TB Extra Large (Redundant)	**192TB Enhanced Performance (EP) (Redundant)**	**288TB Enhanced Performance (EP) (Redundant)**
• 336TB (using 168 2TB NL SAS Drives) of raw space for user data and 12TB (using 24 500GB NL SAS drives) of raw space for metadata • Up to 2.4 GB/s performance • Redundant OSS and MDS servers • Two active/active Object Storage Servers connected to two dual-controller PowerVault MD3200s, each with six MD1200 arrays as expansion storage • Two active/passive Metadata Storage Servers connected to a dual-controller PowerVault MD3220 • Terascala Management software running on a dedicated management server. • Installation service and three years of Dell ProSupport	• 192TB (using 96 2TB NL SAS Drives) of raw space for user data and 12TB (using 24 500GB NL SAS drives) of raw space for metadata • Up to 4.8 GB/s performance • Redundant OSS and MDS servers • Four active/active Object Storage Servers connected to four dual-controller PowerVault MD3200s, each with one MD1200 array as expansion storage • Two active/passive Metadata Storage Servers connected to a dual-controller PowerVault MD3220 • Terascala Management software running on a dedicated management server. • Installation service and three years of Dell ProSupport	• 288TB (using 144 2TB NL SAS Drives) of raw space for user data and 12TB (using 24 500GB NL SAS drives) of raw space for metadata • Up to 4.8 GB/s performance • Redundant OSS and MDS servers • Four active/active Object Storage Servers connected to four dual-controller PowerVault MD3200s, each with two MD1200 arrays as expansion storage • Two active/passive Metadata Storage Servers connected to a dual-controller PowerVault MD3220 • Terascala Management software running on a dedicated management server. • Installation service and three years of Dell ProSupport