Technical Report

NetApp High-Performance Storage Solution for Lustre

Solution Design

Narjit Chadha, NetApp
October 2014 | TR-4345-DESIGN

Abstract

The NetApp[®] High-Performance Storage Solution (HPSS) for Lustre, based on the E-Series platform, is designed for scalable, reliable, and high-performance data requirements. It is designed for extreme I/O performance and massive file system capacity. The NetApp HPSS for Lustre reduces the time to start benchmarking workloads from a typical six to nine months with a build-it-yourself system down to one to three months with its preconfigured and prebenchmarked design. The HPSS for Lustre can be deployed on site within two business days and provides up to 2.88PB of raw disk space, accessible at up to 20GB/sec in a single 42U rack. The racks may be aggregated together for increased capacity and throughput.

The solution is managed with Terascala TeraOS integrated management software, which provides an overview of the entire system, giving a “red/orange/green light” status update and real-time charts for performance, utilization, and balance. Terascala has pioneered its TeraOS software to lower total cost of ownership by managing and optimizing data, performance, and reliability. This management layer provides system-level high availability to reduce costly degraded performance and downtime. Customers are also insulated from the details of Lustre and can focus on obtaining results.

The system layout and a description of the system components are contained within this technical report, along with actual standard benchmark results.

TABLE OF CONTENTS

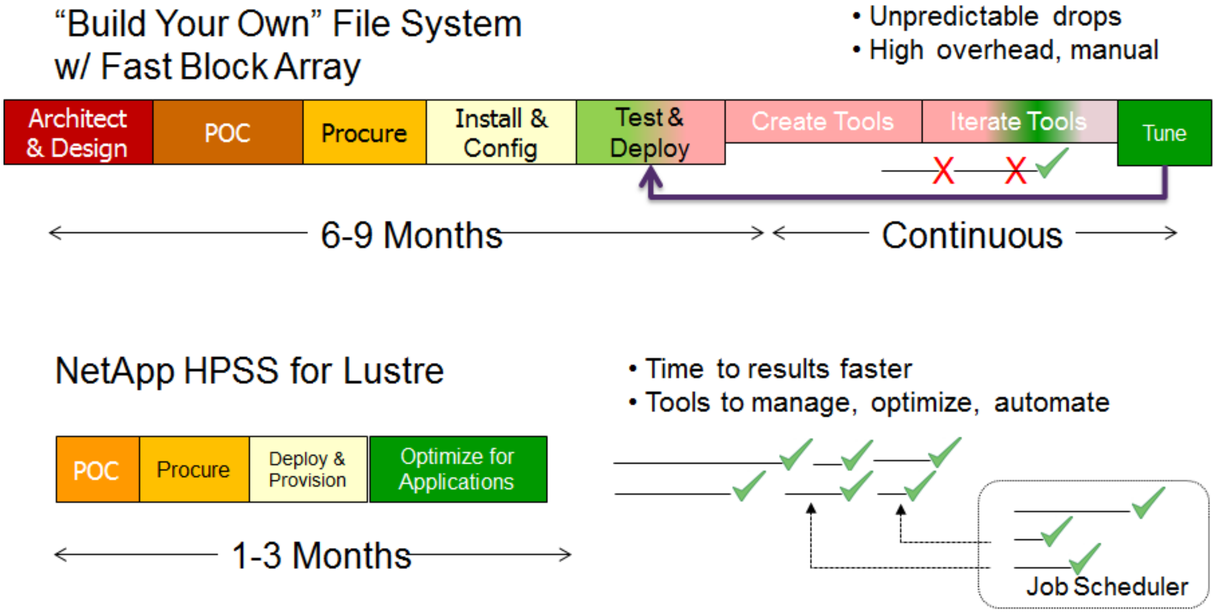
LIST OF TABLES

LIST OF FIGURES

1 Solution Overview

Organizations are faced with challenges involving the deployment of high-performance computing solutions. A typical time frame of six to nine months is required for architecting and designing, a proof of concept, procurement, installation and configuration, and testing and deployment of the chosen HPC solution. The NetApp High-Performance Storage Solution (HPSS) for Lustre is a prebuilt and field-deployed solution that is optimized for performance and reliability. The total time to be in a position to benchmark workloads is reduced typically to one to three months, and this includes the time for the proof of concept, procurement, and application optimization. The benefit is that end users can concentrate on their application work, rather than on infrastructure setup.

Figure 1) Time to complete cluster deployment.



When the environment is set up, the Lustre file system becomes simple to format, use, and monitor. The included Terascala TeraOS management software monitors the performance and health of the HPSS and the results are displayed in an intuitive dashboard.

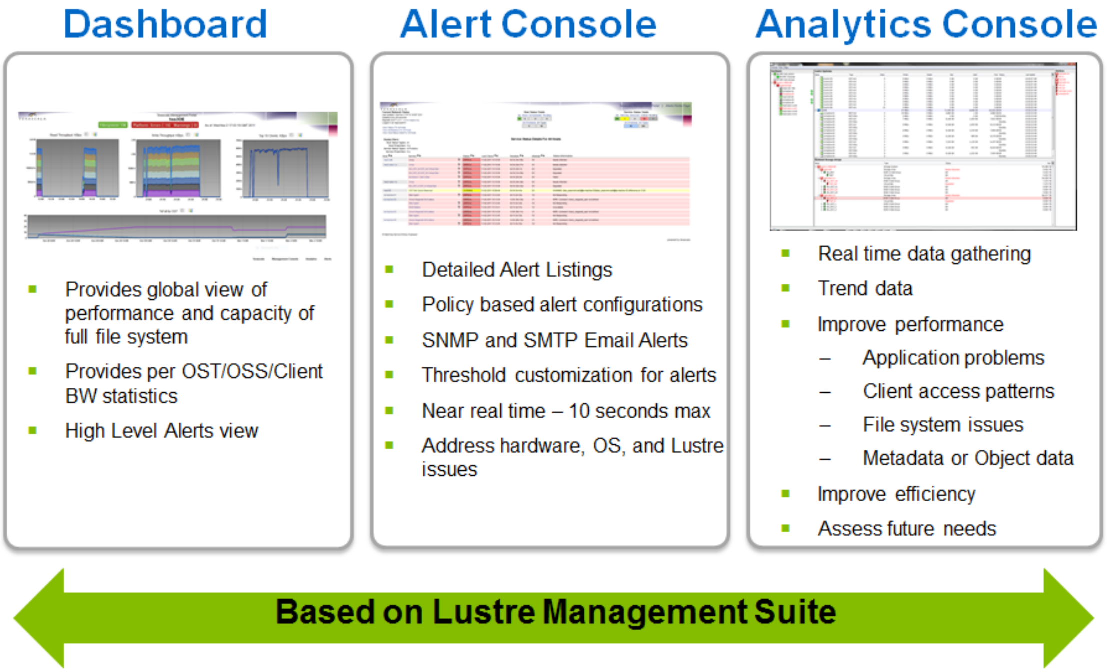
The TeraOS Dashboard homepage provides an overview of the entire system, giving a “red/orange/green light” status update and real-time charts for performance, utilization, and balance. From the dashboard, users can push down to one of the three other interfaces.

The TeraOS Management Console delivers a single view of the underlying storage platform. This console handles all the “maintenance” tasks associated with the file system, including disk failures and replacements, file system errors, and any other issues or parameter updates that need to be addressed.

The Alert Console provides a simple alert review and notification setup for the storage solution. Within the console, users can review and react to any system alerts that are generated. Additionally, administrators can tune alert notifications so that issues are properly reported. Examples include adjusting temperature thresholds to match “real” environment conditions and minimizing false reports, or adjusting parameters to minimize flagging so that repeated errors are not reported as separate events. Additionally, administrators can review the event history to determine patterns or trends.

The Operations Analytics Console provides the tools and views for analyzing and understanding system operations over time. With this console, an administrator can analyze system performance to look for overloaded individual OSS servers, poor data striping, client access patterns, or unbalanced application codes. This console enables users to detect patterns to improve performance of the file system, understand and justify the need for additional capacity or performance, and help users understand unproductive data access patterns.

Figure 2) HPSS for Lustre graphical interface.



Dashboard

- Provides global view of performance and capacity of full file system
- Provides per OST/OSS/Client BW statistics
- High Level Alerts view

Alert Console

- Detailed Alert Listings
- Policy based alert configurations
- SNMP and SMTP Email Alerts
- Threshold customization for alerts
- Near real time – 10 seconds max
- Address hardware, OS, and Lustre issues

Analytics Console

- Real time data gathering
- Trend data
- Improve performance
  - Application problems
  - Client access patterns
  - File system issues
  - Metadata or Object data
- Improve efficiency
- Assess future needs

Based on Lustre Management Suite

1.1 Target Audience

The target audience for the solution includes the following groups:

- HPC system administrators—Those who administer HPC clusters will enjoy the easy-to-use management interface that allows easy monitoring, tuning, and maintenance of the Lustre file system.
- Lab managers—Those who manage the lab can easily understand the HPSS for Lustre environment.
- HPC benchmark personnel—The easy-to-use interface and reliable hardware allow busy professionals to concentrate on obtaining results rather than on managing the cluster.

1.2 Solution Technology

The HPSS for Lustre contains all of the components of a traditional Lustre file system in a tightly integrated and easy-to-manage package. A dedicated Dell R320 server handles the TeraOS management interface and the sensor feedback across the entire system. The HPSS for Lustre includes:

1. Management Switch—A 1 Gigabit Ethernet switch for IPMI and file system monitoring traffic
2. Terascala Management Server— A Dell R320 server to manage and monitor the entire HPSS and attached Lustre clients
3. Lustre Metadata Servers (MDSs)—Two servers in an active-passive failover manner
4. NetApp E2724 Metadata Target—Metadata storage device

5. Lustre Object Storage Servers (OSSs)—Multiple, depending on the file system performance desired to push object data to and from storage
6. NetApp E5560+ Optional DE6600 Data Storage Devices—Multiple, depending on performance and capacity desired

Figure 3 shows the technical components of the solution from a front view for a half-loaded 42U rack.

Figure 3) HPSS for Lustre half-loaded rack, front view.

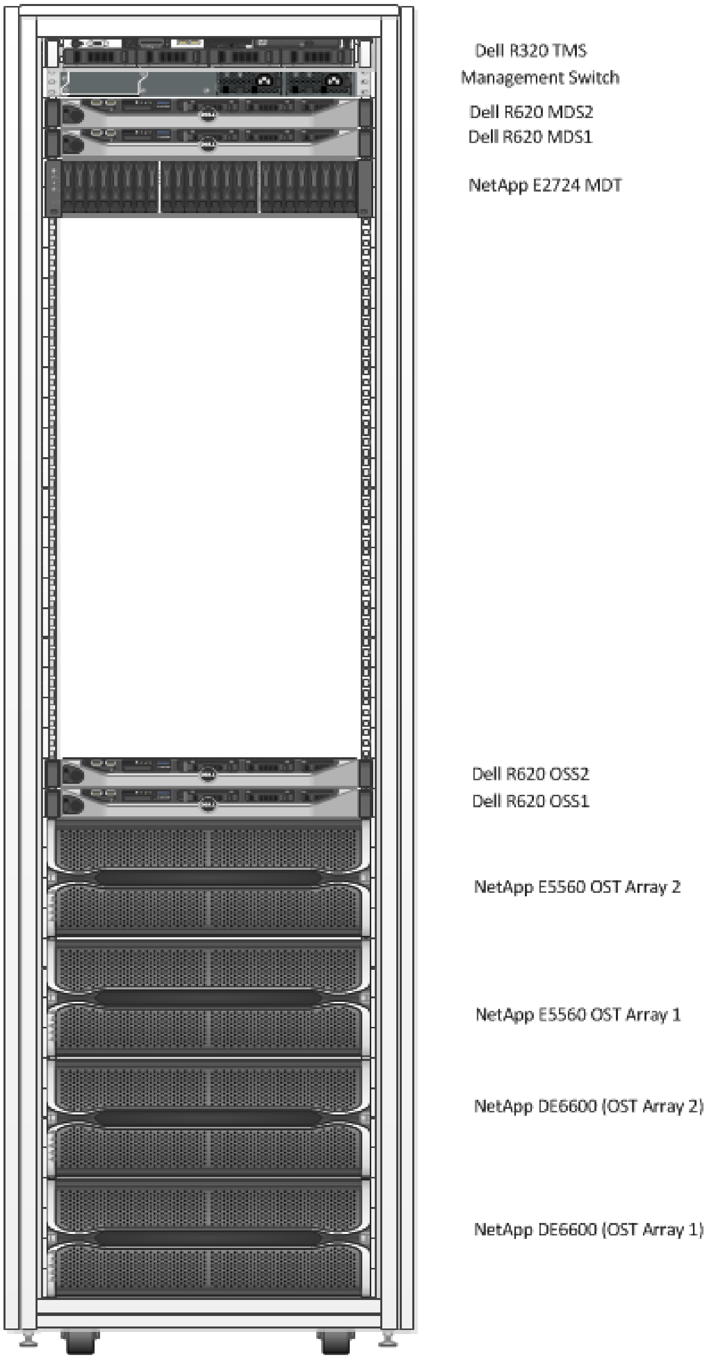
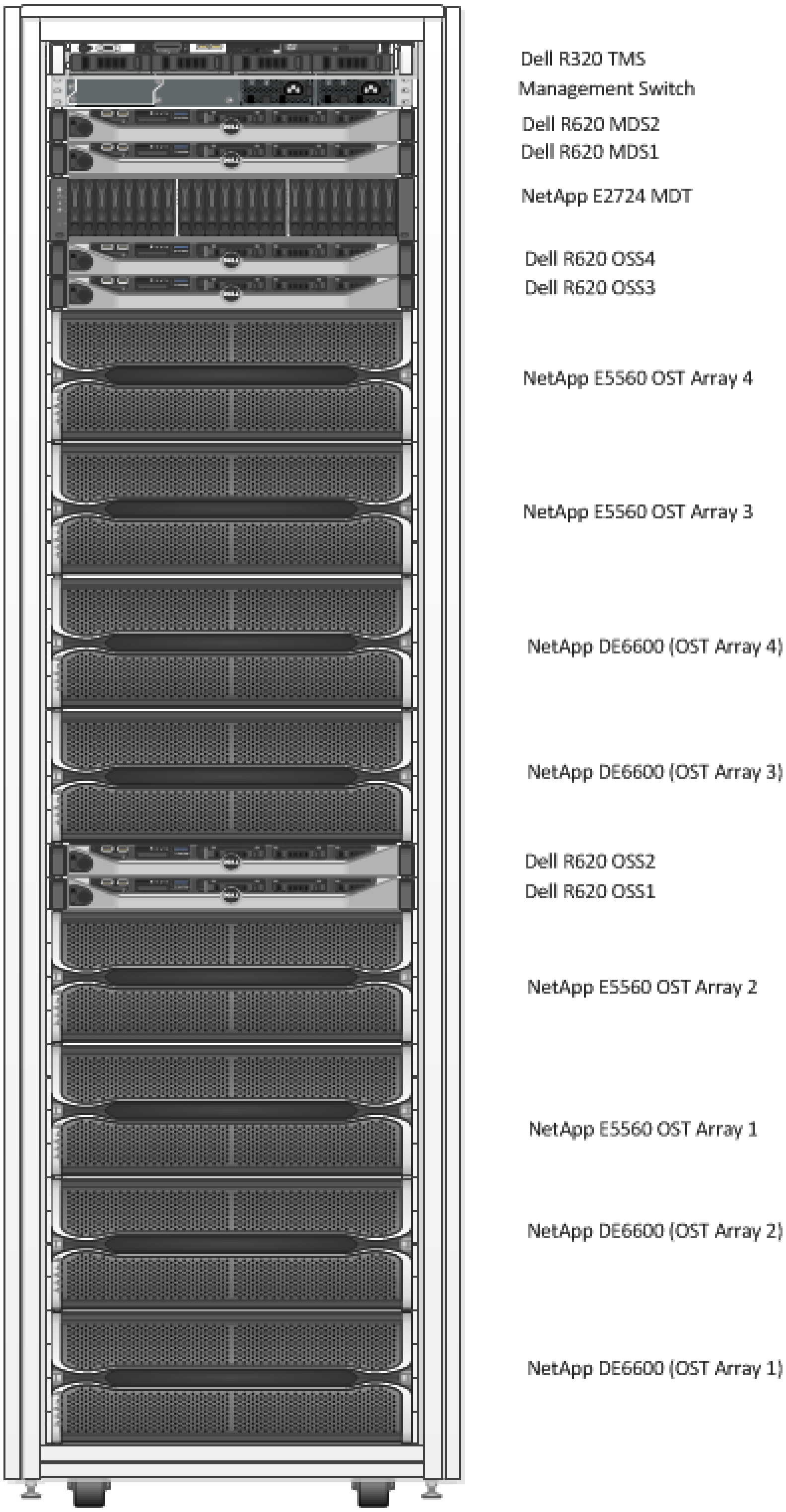


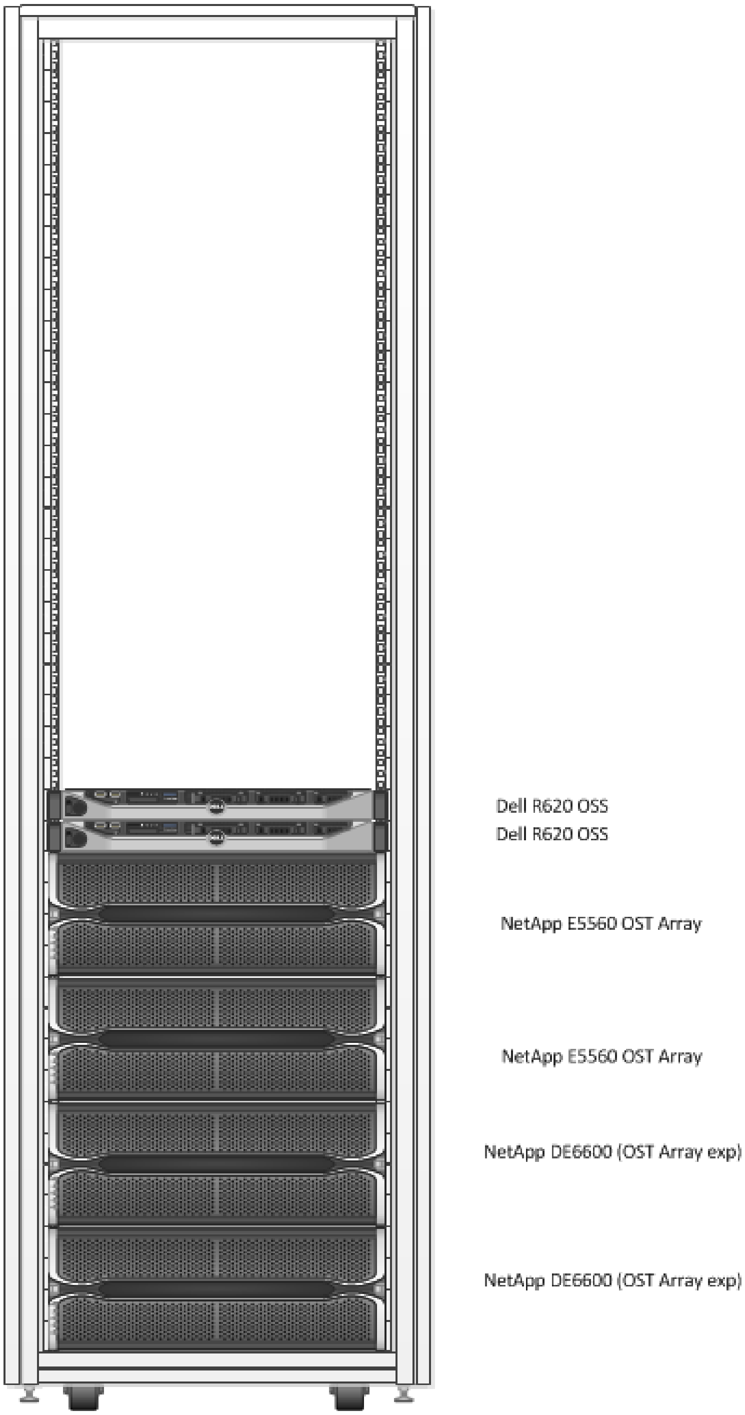
Figure 4 shows the technical components of the solution from a front view for a fully loaded 42U rack.

Figure 4) HPSS for Lustre 42U, front view.



The solution is highly scalable and can grow with the addition of two OSSs and two 120-drive E5560 arrays. The Scalable Storage Unit (SSU) is illustrated in Figure 5, below. At each level of growth, failover redundancy is maintained while allowing both capacity and bandwidth growth. Bandwidth is designed to grow at 5GB/s per 120 drives, or 10GB/s per SSU if a 56Gbps cluster networking system is used. Capacity growth varies based on the drive type chosen.

Figure 5) Scalable Storage Unit (SSU).



1.3 Use-Case Summary

This solution applies to the following use cases:

- The requirement to quickly and easily deploy and use Lustre
- The need to monitor Lustre visually, as opposed to using command line syntax

2 Value of the NetApp High-Performance Storage Solution for Lustre

The NetApp HPSS for Lustre provides high-capacity and high-performance E-Series storage platforms that enable the Lustre file system to support very large scalability and extremely high I/O throughput in the most demanding environments, such as modeling and simulation. The scalable and highly reliable design provides the ability to meet current and future requirements for performance and growth. The HPSS is designed to allow an organization to be in a position to benchmark within one to three months versus the more usual six to nine month timeframe. Deployment is handled onsite by a team of experts from Terascala and is completed within two business days.

Businesses, governments, universities, and research organizations will find that the HPSS for Lustre meets the challenge of supporting tens of thousands of Lustre clients accessing tens of petabytes of storage with scalable I/O throughput.

3 Technology Requirements

3.1 Hardware Requirements

The HPSS for Lustre is self-contained and includes all Lustre components such as metadata servers, metadata targets, object storage servers, and object storage targets, with the addition of the TMS and TeraOS for simple manageability. However, the client end is left entirely to the end user. This includes the data networking switch, and the client servers, which are completely agnostic to the solution. The only requirement is that the client is able to support Lustre, and Lustre performance is dictated by the actual client servers chosen.

Table 1 lists the hardware components required to implement the solution. The hardware components used in any particular implementation of the solution may vary based on end-customer requirements.

Table 1) Hardware requirements.

Hardware	Quantity
Client Servers	Unspecified
Data Network Switch	1 or greater (40/56GigE) depending on HPSS size
Network Cables	40 or 56 GigE based on network chosen

3.2 Software Requirements

Table 2 lists the software components required to implement the solution. The software components used in any particular implementation of the solution may vary based on customer requirements.

Table 2) Software requirements.

Software	Version or Other Information
Red Hat Enterprise Linux® or CentOS Linux on client	5.8+

Software	Version or Other Information
servers	
NetApp SANtricity® software (optional) on networked server (Windows® or Linux) routable to storage	11.10.xxxx.xxxx +

4 HPSS for Lustre Management Interface

HPSS for Lustre TeraOS software was designed as an all-encompassing management interface to HPSS for Lustre. It is designed for configuration of the underlying Lustre environment and for monitoring and reporting on the HPSS environment.

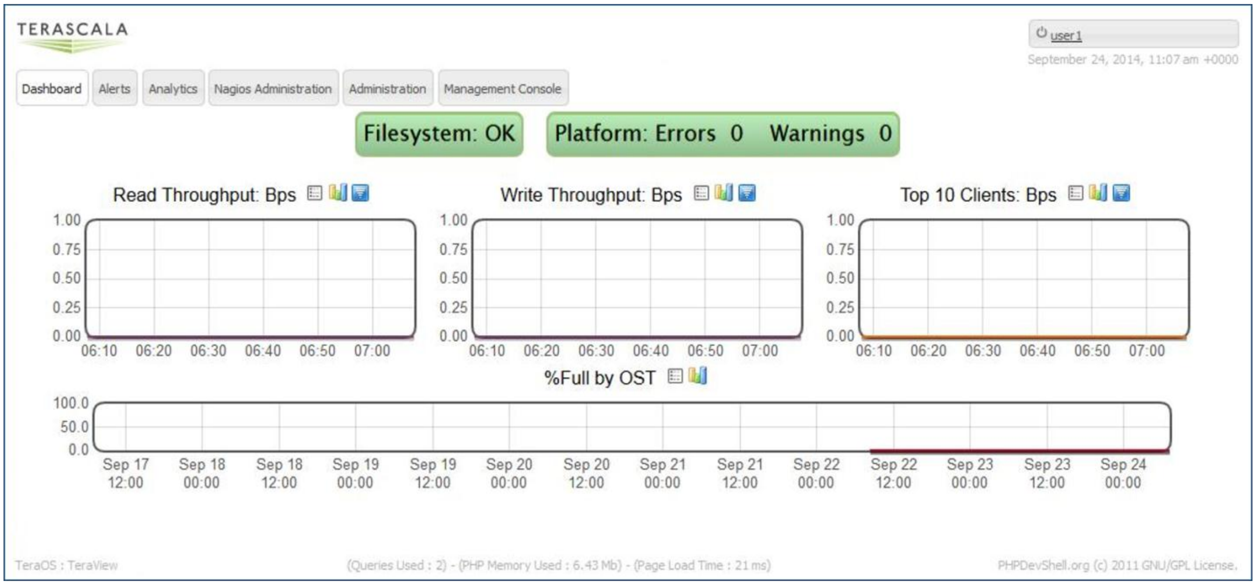
HPSS for Lustre software is tightly integrated with HPSS hardware and the Lustre file system environment to function in an optimal manner. Any hardware or software changes to the rack configuration must be precipitated through NetApp and/or Terascala to enable compliance.

The HPSS for Lustre Management Interface may be accessed via the Firefox or the Google Chrome browser.

Note: Microsoft® Internet Explorer is not supported at this time.

You must also know the IP address of the TMS and have a client machine that is routable to that address. When the client web browser is pointed at the TMS address, the Terascala Management Portal is displayed. The default user name and password combination is admin/admin.

Figure 6) HPSS for Lustre Management Portal.



The HPSS for Lustre Management Portal displays the current status of the file system, platform error count, throughput graphs to the underlying Lustre file system marked with time of day, and file system capacity utilization, and it provides links at the top of the window to other HPSS screens.

Terascala provides detailed documentation of HPSS for Lustre Management capabilities, including screens for monitoring, administration, and analytics. Alerts may be set up and sent automatically to e-mail accounts to obtain information remotely on the system health.

5 Capacity

Table 3, below, shows expected raw and usable Lustre storage capacities utilizing the NetApp HPSS for Lustre. The table moves up in 60-drive increments.

Table 3) Raw and usable capacities by drive count and drive size.

Drive Capacity (TB)	Drive Count	Raw Capacity (TB)	Usable Capacity (TB)
3	60	180	133.5
3	120	360	267
3	240	720	534
3	480	1440	1068
4	60	240	178
4	120	480	356
4	240	960	712
4	480	1920	1424
6*	60	360	267^
6*	120	720	534^
6*	240	1440	1068^
6*	480	2880	2136^

*Firmware version 08.20 or greater.

^Projected capacity availability.

6 Benchmark Results

Various benchmark tests were carried out to gauge the performance of a full 42U HPSS for Lustre, as shown in Figure 4, above. The results given in the sections below are from the use of 40GigE or 56GigE interconnects between the OSS and MDS units and 10GigE in use to the end Lustre clients. In all cases a Mellanox SX1024 10/40/56 GigE switch was employed for the benchmarking. FDR IB has not been tested to date, although results are expected to be higher with the use of this technology, given the added network bandwidth available and the lower CPU usage required in each server with the TCP/IP elimination.

6.1 Servers Used for Benchmarking

The HPSS for Lustre spec allows higher-performance servers. However, the following servers were employed in the HPSS tested:

Metadata Servers

Dell R620 1U server

2 x Intel® E5-2670 (2.6GHz) CPUs

128GB RAM

Lustre 2.1.5 (Terascala TeraOS v5.0.4)

Object Storage Servers (1,2)

Dell R620 1U server

2 x Intel® E5-2670 (2.6GHz) CPUs

64GB RAM

Lustre 2.1.5 (Terascala TeraOS v5.0.4)

Object Storage Servers (3,4)

Dell R620 1U server

2 x Intel® E5-2640 (2.4GHz) CPUs

192GB RAM

Lustre 2.1.5 (Terascala TeraOS v5.0.4)

Client Servers (8 clients)

Dell R620 1U server

1 x Intel® E5-2640 (2.4GHz) CPUs

64GB RAM

Lustre 2.5.0 x86_64 with CentOS 6.5 x86_64

6.2 Storage Used for Benchmarking

MDT

24-drive E2724

24 x 4TB NL SAS drives

1 12Gbps SAS link from each controller to each MDS

OST Devices

4 x E5560+DE6600 (120 4TB NL SAS drives each) for entire system

Each E5560 controller has 2 6Gbps SAS links, 1 to each OSS of the OSS pair

6.3 Results

6.3.1 Obdfilter-survey

Obdfilter-survey is an OSS-based utility that allows for the measurement of the raw disk performance of the file system, negating the Lustre network and clients. It is generally a measurement of the maximum performance that a Lustre file system is capable of.

By running numerous iterations of obdfilter-survey, we determined that the disk component itself was capable of 19.8GB/sec of performance throughput. This is a result of using the default RAID 6 8+2 LUN configuration per OST and keeping a generic 128kB segment size per OST.

```
ost 48 sz 1610612736K rsz  512K obj  192 thr 1536 write 17002.32 [  79.00, 542.43] rewrite 17490.54 [
251.42, 532.46] read 19822.10 [  10.50,6432.59]

ost 48 sz 1610612736K rsz  512K obj  384 thr  384 write 19851.85 [ 134.99, 649.98] rewrite 19208.80 [
121.00, 565.43] read 17650.17 [ 101.50,8114.50]
```

It has been observed that greater than 20GB/sec of performance is standard for large block I/O when using a segment size of 512kB, although there may be a performance impact if smaller file sizes are used. Such a change, if desired, should be made before the HPSS system is brought up.

6.3.2 IOzone

IOzone measures maximum throughput to the storage device from a single client node or multiple client nodes. For the purposes of this testing, cluster mode was employed and IOzone version 3_428 was utilized for the benchmarking. It became quickly obvious that 256 threads, the IOzone maximum allowed, was not sufficient to saturate the entire file system using the client servers available. A new strategy was then employed to utilize each of the four OSSs independently to obtain a gauge of what the entire file system (4 120-drive E5560 arrays) is capable of. For the testing, each OSS wrote to half of the LUNs on each of the two 120-drive E5560 arrays directly connected to it. The results were aggregated to obtain approximations of what the entire file system is capable of.

For the testing, 8 clients and 192 threads (32GB/thread) were used to write independently from each of the OSSs.

Table 4) 40GigE IOzone results (192 client threads).

OSS	Sequential Write (GB/s)	Sequential Rewrite (GB/s)	Random Write (GB/s)
1	4.4	4.4	2.7
2	4.4	4.5	2.6
3	4.3	4.5	2.5
4	4.3	4.5	2.5

Note: Servers 1,2 are different from 3,4 as described in Section 6.1, above, and utilize the faster CPUs with less memory

Table 5) 56GigE IOzone results (192 client threads).

OSS	Sequential Write (GB/s)	Sequential Rewrite (GB/s)
1	4.8	5.0
2	5.1	5.1
3	4.7	4.7
4	4.7	4.8

The numbers above show that approximately a 20GB/sec write performance is obtainable on the overall larger file system when using E5-2670 processors in the OSSs and a 56Gbps Ethernet interconnect between the OSSs and the MDSs. Using 40Gbps Ethernet, there is moderate network contention, and 17GB/s write performance appears obtainable by aggregating the results of the individual OSSs. Read performance was measured as slightly higher than write performance.

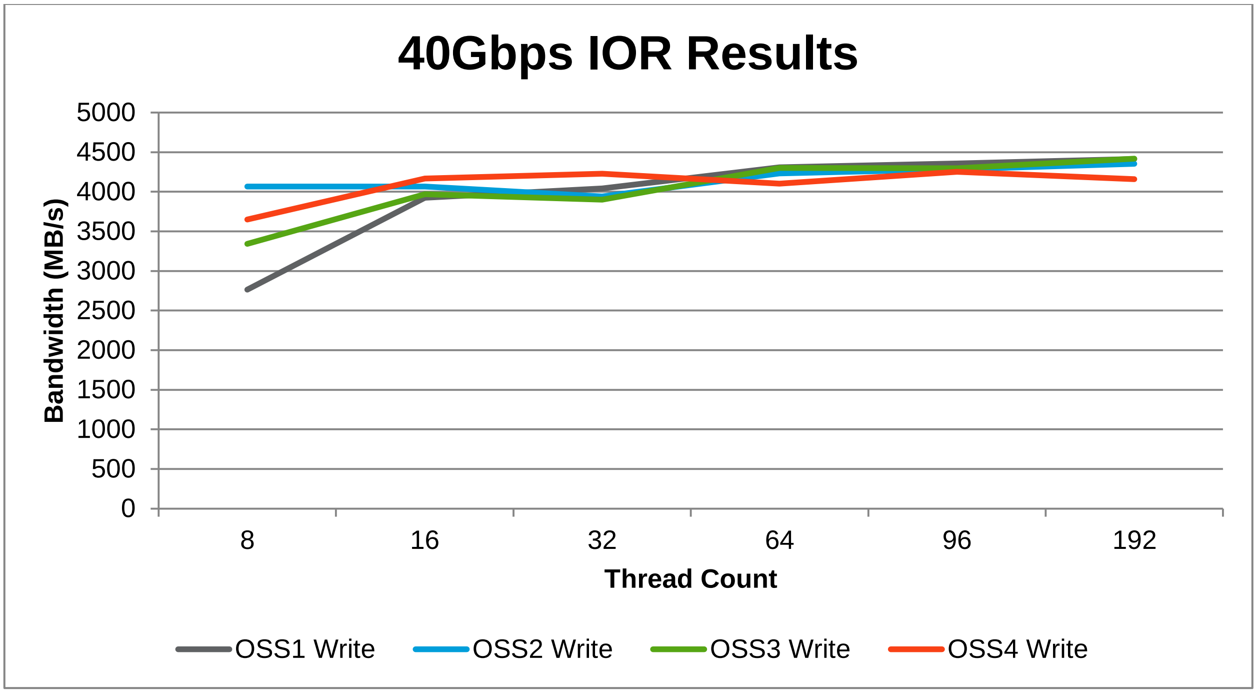
6.3.3 IOR

With IOR, clients can first write data and then read data written by another client, thus avoiding the problem of having to clear a client's cache. This method works well for more than two clients. After compiling with MVAPICH2, IOR was run.

The same strategy was employed as with IOzone. Independent OSSs were used for the runs.

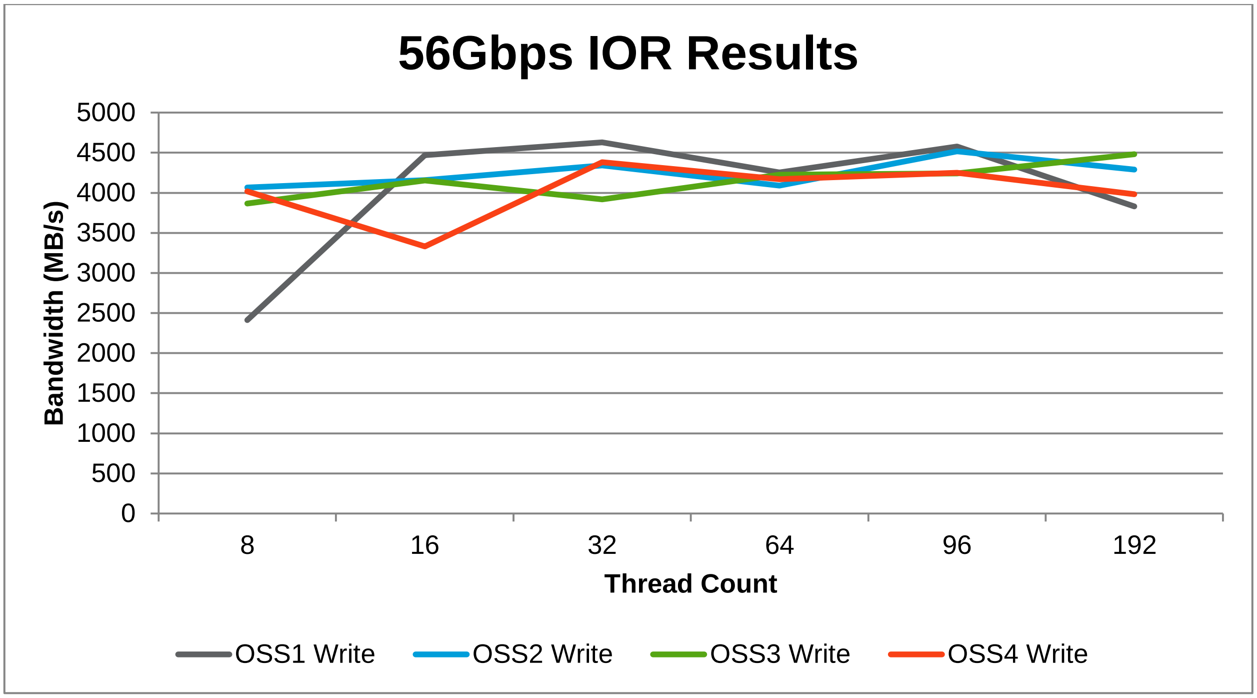
In Figure 7, below, the 40Gbps network results are illustrated (writes). Numbers are approximately the same across all OSSs and peak at about 4400MB/sec per OSS.

Figure 7) 40Gbps IOR results.



The 56Gbps network results are shown below in Figure 8. Little difference in performance between 40 and 56Gbps networks was observed with IOR, although the 56Gbps write results peak much closer to 5 GB/sec per OSS.

Figure 8) 56Gbps IOR results.



6.3.4 mdtest

mdtest is a program that measures the performance of various metadata operations. It uses MPI to coordinate the operations and to collect the results. mdtest requires compilation under MPI for multiple servers to push data out simultaneously to the E2724 MDT. mvapich2 was chosen as the mpi version for this testing because it is a generic and freely available mpi implementation. Other mpi versions such as pgi or Intel mpi may produce better results.

Testing was done with mdtest using the recommended E2724 segment size of 128KB and all 24 drives of the device. The 24 (900GB) drives in the E2724 were in a single RAID 10 group (12+12), per the HPSS architecture. Essentially, mdtest was run in a loop using 8 clients, all 4 OSSs, and all 48 OSTs, with a single MDS/E2724 MDT. The entire file system was employed for this test.

When 40Gbps was used as the network interconnect, it was observed that directory creates peaked around 49,000 operations per second, directory stats around 31,000 per second, and directory removals at 25,500 per second. For file operations, file creations were at 39,700 per second, file stats at 31,200 per second, and file removals at 25,800 per second.

Figures 9 to 11, below, show the 40Gbps results for 100, 1,000, and 10,000 directories.

Figure 9) 40Gbps 100-directory mdtest results.

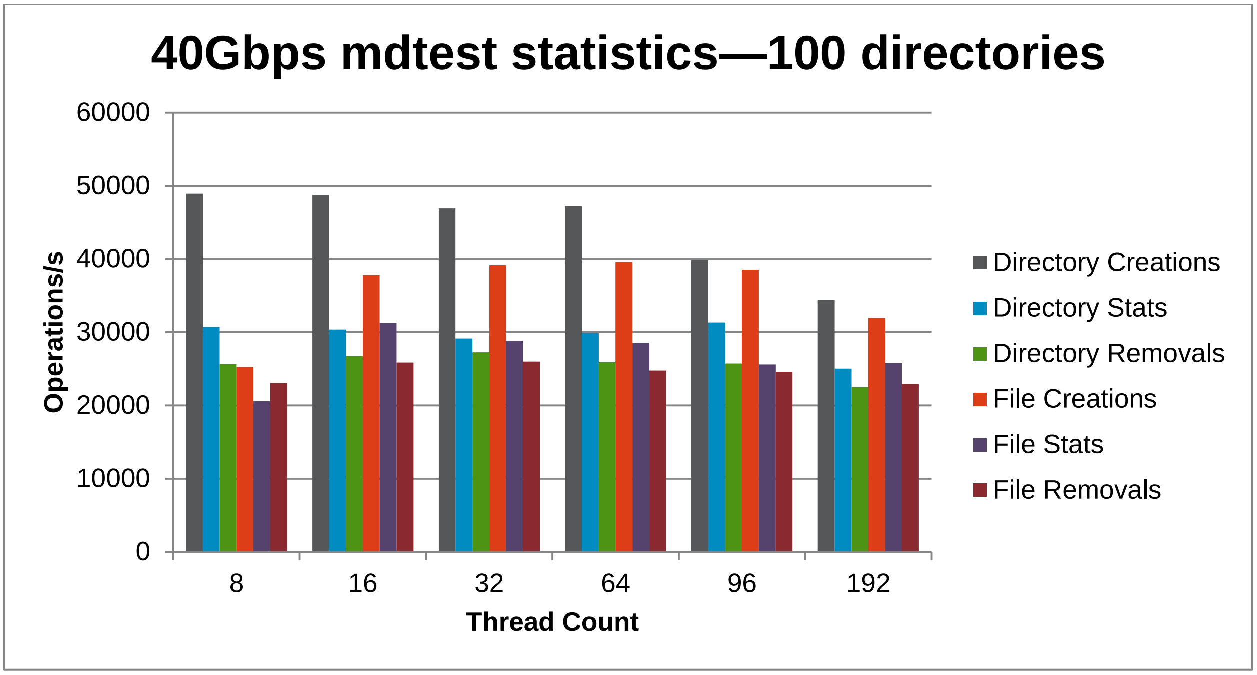


Figure 10) 40Gbps 1,000-directory mdtest results.

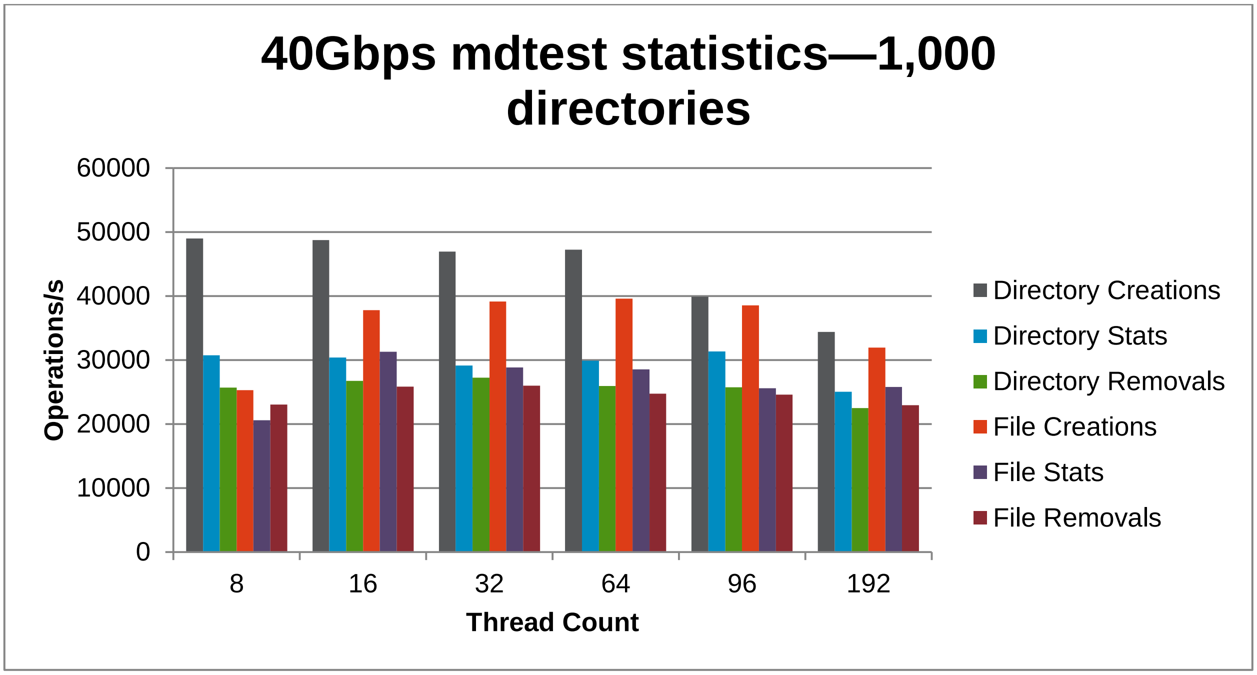
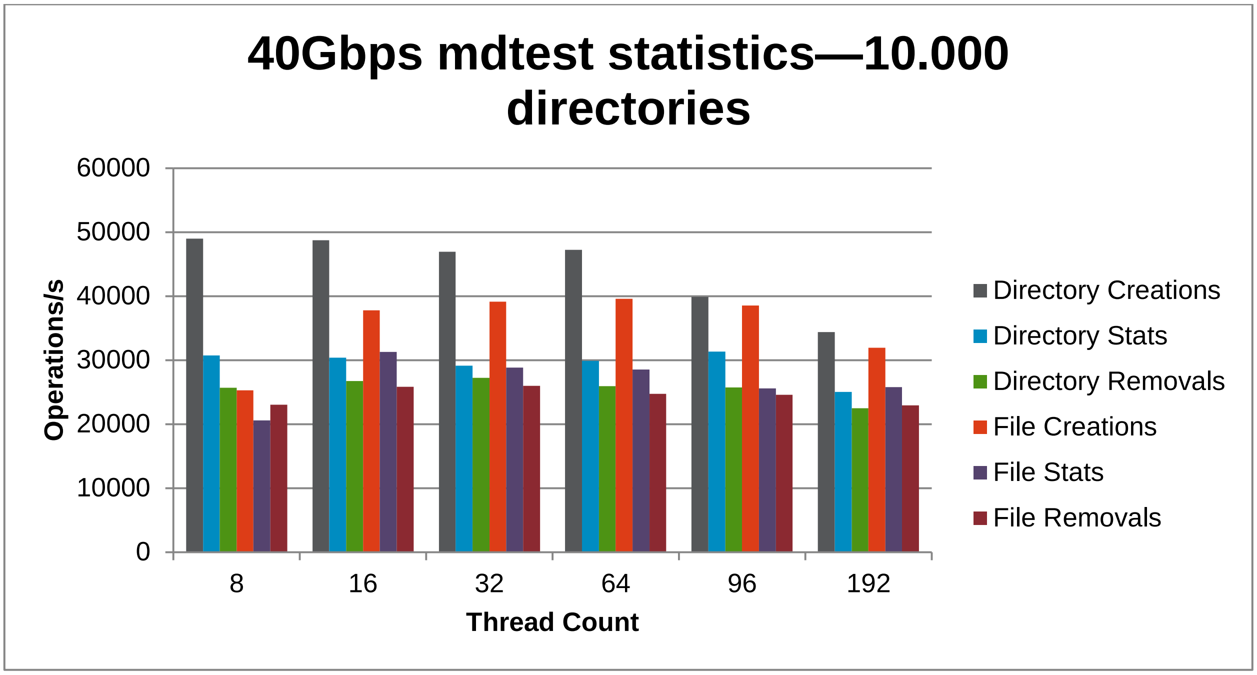


Figure 11) 40Gbps 10,000-directory mdtest results.



The use of 56Gbps Ethernet as the interconnect between the OSSs and MDSs did not appear to improve metadata results significantly. The numbers seen for directory creates using 56Gbps peaked at 47,400 per second, directory stats at 30,500 per second, and directory removals at 26,700 per second. File creates were observed at 39,000 per second, file stats at 30,600 per second, and file removals at 25,700 per second. The results of the 56Gbps mdtest runs are shown below in Figures 12 to 14.

Figure 12) 56Gbps 100-directory mdtest results.

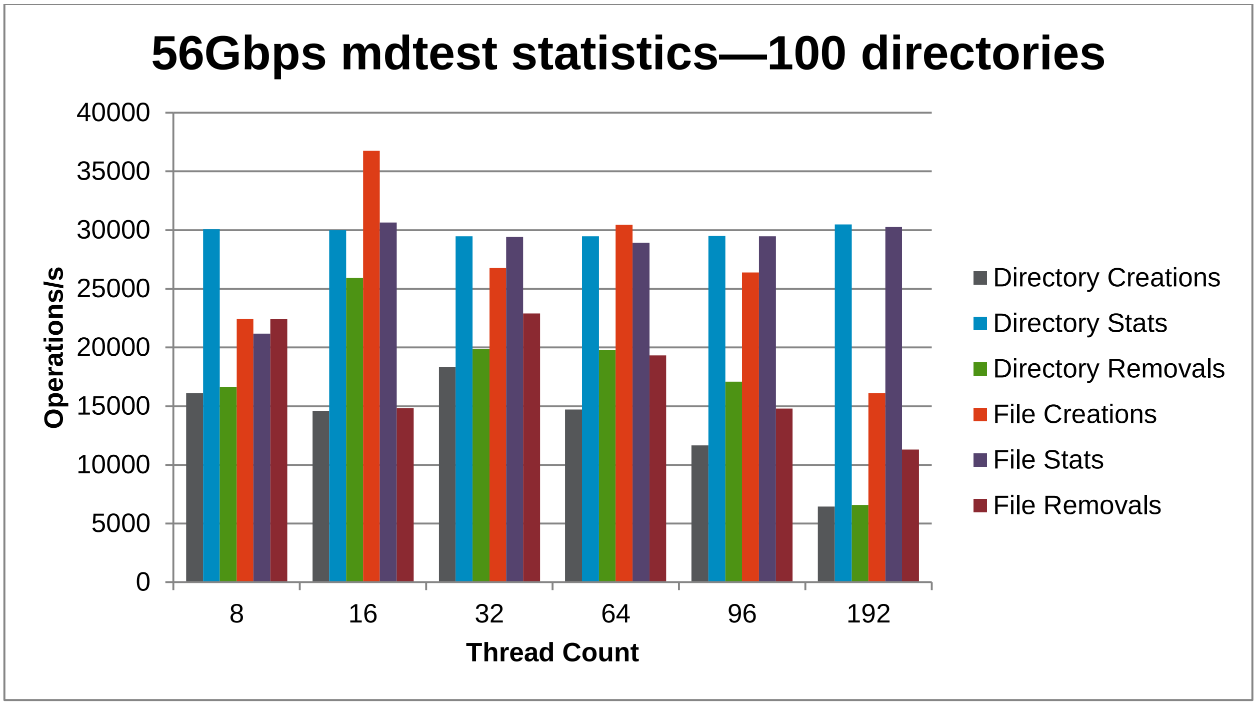


Figure 13) 56Gbps 1,000-directory mdtest results.

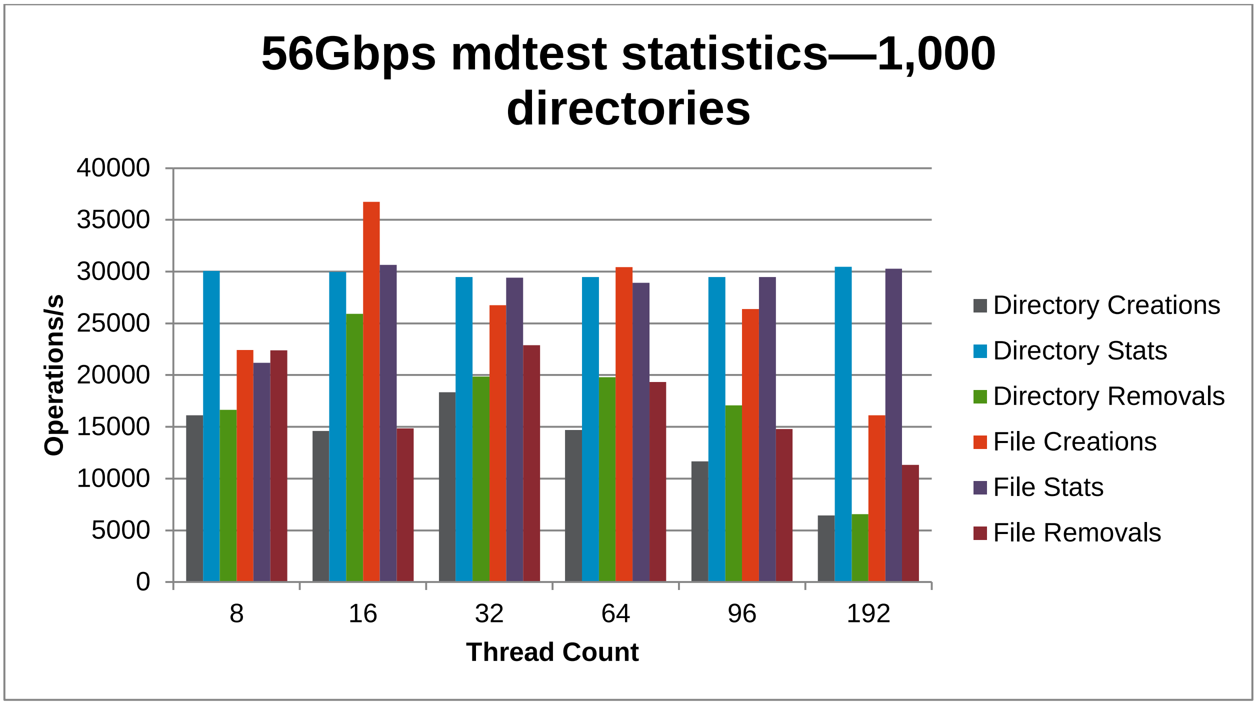
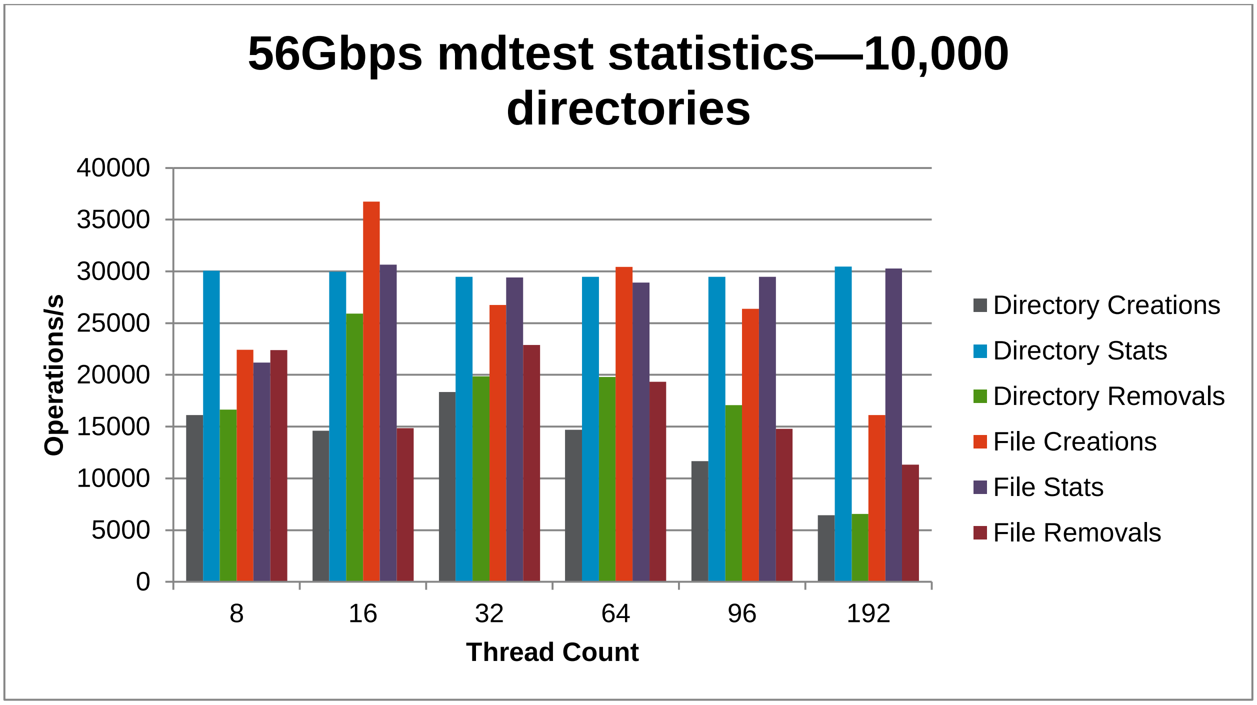


Figure 14) 56Gbps 10,000-directory mdtest results.



7 Best Practices

Replace failed components immediately.

Use a failback environment following hardware failures.

Do not directly modify storage LUNs.

8 Support

Support for the NetApp High-Performance Storage Solution for Lustre is handled through our partners at Terascala. Terascala developed the TeraOS software for the solution and is very knowledgeable about the system internals. It was determined that this is the most efficient way to assist customers. NetApp will assist at the back end with issue resolution if it is determined that problems are storage related.

9 Conclusion

To resolve the technical and time challenges of deploying a high-performance file system from scratch, NetApp designed the High-Performance Storage Solution (HPSS) for Lustre, which typically compresses the time for full system deployment from six to nine months down to one to three months from the time an order is received. Installing and bringing up the HPSS is done on site within two business days by a trained Terascala technician.

NetApp HPSS for Lustre, based on the E-Series platform, is designed for scalable, reliable, and high-performance computational requirements for extreme I/O performance and massive file system capacity. Businesses, governments, universities, and research organizations will find that HPSS for Lustre meets the challenge of supporting tens of thousands of Lustre clients accessing tens of petabytes of storage with scalable I/O throughput.

The integrated Terascala HPSS for Lustre software was designed as an all-encompassing management interface to HPSS. It is designed to configure the underlying Lustre environment and to monitor and report on the HPSS environment. This design provides a feature set above and beyond that of standard Lustre.

Benchmarking results reveal that approximately 20GB/sec of Lustre throughput and 1.424PB (1.92PB raw) of usable storage space are available in a single 42U rack with this solution using 56GigE as the interconnect and 4TB near-line SAS drives. 40GigE has also been vetted as a network interconnect for the solution. Following the 08.20 firmware release slated for December 2014 and with the use of 6TB near-line SAS drives, 2.88PB of raw disk space will be available in a 42U rack.

Version History

Version	Date	Document Version History
Version 1.0	October 2014	Initial document including results produced.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

NetApp provides no representations or warranties regarding the accuracy, reliability, or serviceability of any information or recommendations provided in this publication, or with respect to any results that may be obtained by the use of the information or observance of any recommendations provided herein. The information in this document is distributed AS IS, and the use of this information or the implementation of any recommendations or techniques herein is a customer's responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. This document and the information contained herein may be used solely in connection with the NetApp products discussed in this document.

www.netapp.com