# Dell Storage for HPC with Intel Enterprise Edition for Lustre

## A Dell Technical White Paper

**Quy Ta**

Dell HPC Engineering

November 2014 | Version 1.0

THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

© 2014 Dell Inc. All rights reserved. Reproduction of this material in any manner whatsoever without the express written permission of Dell Inc. is strictly forbidden. For more information, contact Dell.

*Dell*, the *DELL* logo, and the *DELL* badge, *PowerConnect*, and *PowerVault* are trademarks of Dell Inc. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. Dell Inc. disclaims any proprietary interest in trademarks and trade names other than its own.

# Contents

Figures ..... iv
Tables ..... v
1. Introduction ..... 1
2. The Lustre File System ..... 1
3. Dell Storage for HPC with Intel EE for Lustre Description ..... 3
3.1 Management Server ..... 4
3.2 Metadata Servers ..... 4
3.3 Object Storage Servers ..... 5
3.4 Scalability ..... 7
3.5 Networking ..... 8
3.5.1 Management Network ..... 8
3.5.2 Data Network ..... 8
3.6 Managing the Dell Storage for HPC with Intel EE for Lustre ..... 9
4. Performance Evaluation and Analysis ..... 10
4.1 N-to-N Sequential Reads / Writes ..... 13
4.2 Random Reads and Writes ..... 13
4.3 IOR N-to-1 Reads and Writes ..... 14
4.4 Metadata Testing ..... 16
5. Conclusions ..... 19
Appendix A: Benchmark Command Reference ..... 20
References ..... 21

## Figures

Figure 1: Lustre based storage solution components ..... 2
Figure 2: Dell Storage for HPC with Intel EE for Lustre Components Overview ..... 3
Figure 3: Dell PowerEdge R630 ..... 4
Figure 4: Metadata Server Pair ..... 5
Figure 5: Object Storage Server Pair ..... 6
Figure 6: RAID6 Layout on MD3460 or MD3060e arrays ..... 7
Figure 7: OSS Scalability ..... 8
Figure 8: Intel Manager for Lustre (IML) interface ..... 10
Figure 9: Sequential Reads / Writes Dell Storage for HPC with Intel EE for Lustre ..... 13
Figure 10: N-to-N Random reads and writes ..... 14
Figure 11: N-to-1 IOR Read / Write ..... 16
Figure 12: File Metadata Operations ..... 18
Figure 13: Directory Metadata Operations ..... 19

# Tables

Table 1: Test Client Cluster Details ..... 10
Table 2: Dell Storage for HPC with Intel EE for Lustre Configuration ..... 12
Table 3: IOR Shared File Size ..... 15
Table 4: Parameters used on MDtest ..... 17

# 1. Introduction

In High-Performance computing, the efficient delivery of data to and from the compute nodes is critical and usually implicates some complications. Researchers can generate and consume data in HPC systems at such speed that renders the storage components a bottleneck. Managing and monitoring such complex storage systems add to the burden on storage administrators and researchers. The data requirements around performance and capacity keep increasing rapidly. Increasing the throughput and scalability of storage systems supporting the high performance computing system can require a great deal of planning and configuration.

The Dell Storage for HPC with Intel Enterprise Edition for Lustre software, referred to as Dell Storage for HPC with Intel EE for Lustre for the rest of this document, is designed for academic and industry users who need to deploy a fully-supported, easy-to-use, high-throughput, scale-out, and cost-effective parallel file system storage solution. Dell Storage for HPC with Intel EE for Lustre is a scale-out storage solution appliance capable of providing a high performance and high availability storage system. Utilizing an intelligent, extensive and intuitive management interface, the Intel Manager for Lustre (IML), the solution greatly simplifies managing and monitoring all of the hardware and storage system components. It is easy to scale in capacity, performance or both, thereby providing a convenient path to grow in the future. The storage solution uses Lustre®, the leading HPC open source parallel file system[1].

The storage solution utilizes the 13th generation of enterprise Dell PowerEdge™ servers and the latest generation of high density PowerVault™ storage products. With full hardware and software support from Dell and Intel, the Dell Storage for HPC with Intel EE for Lustre solution delivers a superior combination of performance, reliability, density, ease of use and cost-effectiveness.

The following sections of this paper will describe the Lustre File System, the Dell Storage for HPC with Intel EE for Lustre solution followed by performance analysis and conclusions. Appendix A: Benchmark Command Reference

# 2. The Lustre File System

Lustre is a parallel file system, offering high performance through parallel access to data and distributed locking. A Lustre installation consists of three key elements: the metadata subsystem, the object storage subsystem (data) and the compute clients that access and operate on the data.

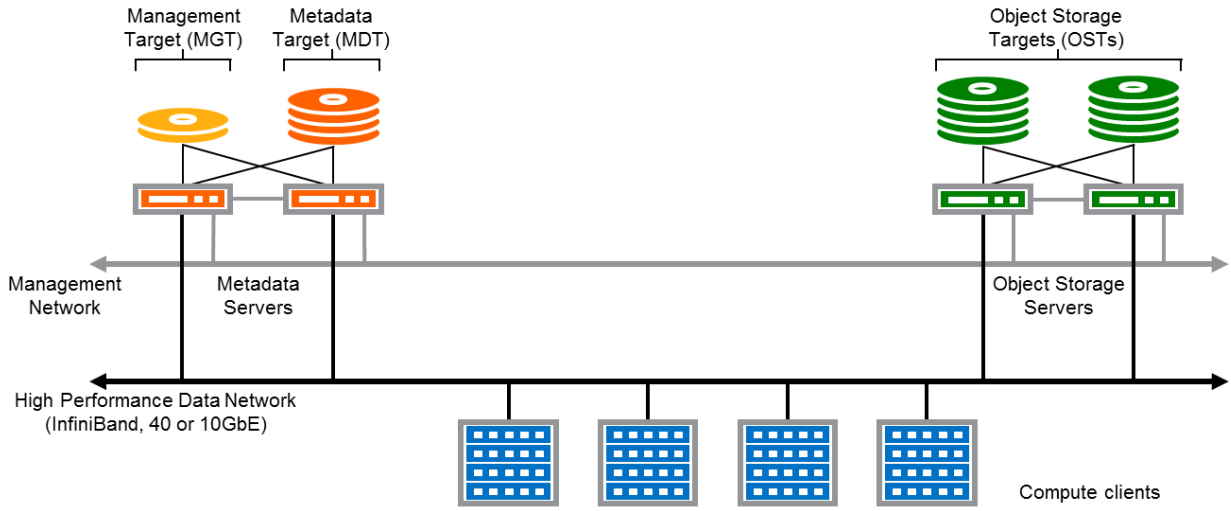
The metadata subsystem is comprised of the Metadata Target (MDT), the Management Target (MGT) and a Metadata Server (MDS). The MDT stores all metadata for the file system including file names, permissions, time stamps, and the location of data objects within the object storage system. The MGT stores management data such as configuration information and registry. The MDS is a dedicated server that manages the MDT.

The object storage subsystem is comprised of one or more Object Storage Targets (OST) and one or more Object Storage Servers (OSS). The OSTs provides storage for file object data, while each OSS

[1] A New Generation of Lustre Software Expands HPC into the Commercial Enterprise

manages one or more OSTs. Typically, there are several active OSSs at any time. Lustre is able to deliver increased throughput by increasing the number of active OSSs (and associated OSTs). Each additional OSS increases the existing networking throughput, while each additional OST increases the storage capacity. Figure 1 shows the relationship of the MDS, MDT, MGS, OSS and OST components of a typical Lustre configuration. Clients in the figure are the HPC cluster's compute nodes.

**Figure 1: Lustre based storage solution components**




A parallel file system, such as Lustre, delivers performance and scalability by distributing data ("striping" data) across multiple Object Storage Targets (OSTs), allowing multiple compute nodes to efficiently access the data simultaneously. A key design consideration of Lustre is the separation of metadata access from IO data access in order to improve the overall system performance.

The Lustre client software is installed on the compute nodes to allow access to data stored on the Lustre file system. To the clients, the file system appears as a single namespace that can be mounted for access. This single mount point provides a simple starting point for application data access, and allows access via native client operating system tools for easier administration.

Lustre includes a sophisticated and enhanced storage network protocol, Lustre Network, referred to as LNet. LNet is capable of leveraging certain types of network features. For example, when the Dell Storage for HPC with Intel EE for Lustre utilizes InfiniBand as the network to connect the clients, MDSs and OSSs, LNet enables Lustre to take advantage of the RDMA capabilities of the InfiniBand fabric to provide faster I/O transport and lower latency than experienced with typical networking protocols.

To summarize, the elements of the Lustre file system are as follows:

- Metadata Target (MDT) – Stores the location of "stripes" of data, file names, time stamps, etc.
- Management Target (MGT) – Stores management data such as configuration and registry
- Metadata Storage Server (MDS) – Manages the MDT, providing Lustre clients access to files.
- Object Storage Target (OST) – Stores the data stripes or extents of the files on a file system.
- Object Storage Server (OSS) – Manages the OSTs, providing Lustre clients access to the data.

- Lustre Clients – Access the MDS to determine where files are located, then access the OSSs to read and write data

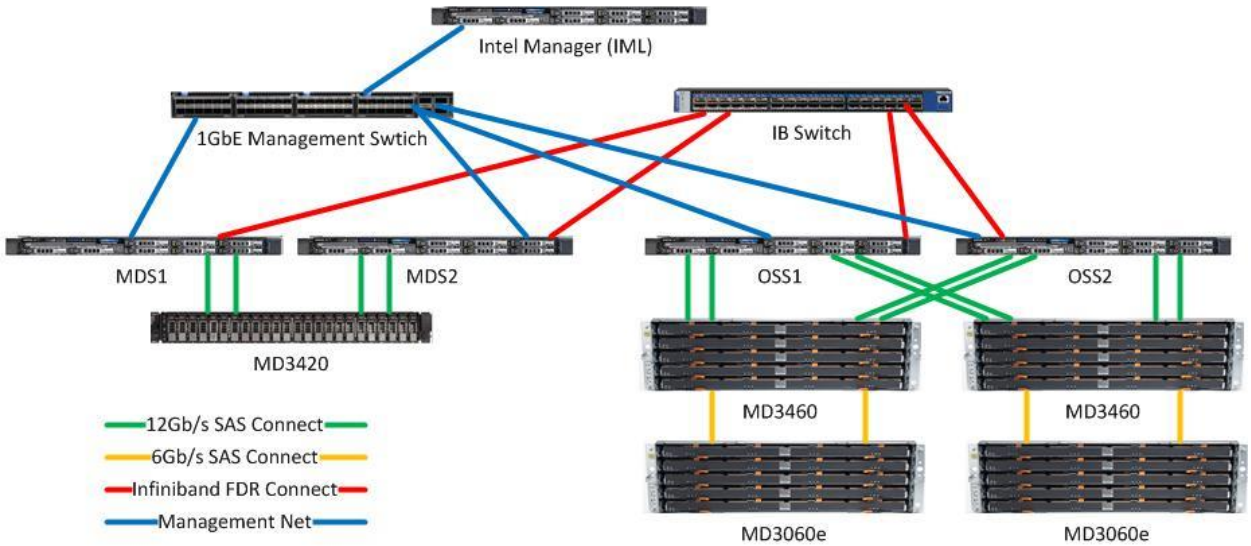
Typically, Lustre deployments and configurations are considered very complex and time consuming tasks. Lustre installation and administration is normally done via a command line interface (CLI), requiring extensive knowledge of the file system operation, along with the auxiliary tools like LNet and the locking mechanisms. In addition, once the Lustre storage sytem is in place, maintaining the system and performance optimizations can be a daunting undertaking. Such requirements, and the steep learning curve associated with them, may prevent Systems Administrators unfamiliar with Lustre from performing an installation, possibly preventing their organization from experiencing the benefits of a parallel file system. Even for experienced Lustre System Administrators, maintaining a Lustre file system can take a big portion of their time.

# 3. Dell Storage for HPC with Intel EE for Lustre Description

The Dell Storage for HPC with Intel EE for Lustre solution provides a storage solution consisting of a management server, Lustre Metadata Servers, Lustre Object Storage Servers and the associated backend storage. The solution provides storage using a single namespace that is easily accessed by the cluster's compute nodes and managed through the Intel Manager for Lustre (IML), an extensive Web-based interface. Figure 2 shows an overview of a Dell Storage for HPC with Intel EE for Lustre system in a 240-drive configuration, including some basic information about the different components described in detail later in this section. Note the three major components:

- Management Server (IML)
- Metadata Server pair (MDS)
- Object Storage Server pair (OSS)

**Figure 2: Dell Storage for HPC with Intel EE for Lustre Components Overview**



The Dell Storage for HPC with Intel EE for Lustre solution utilizes the Dell PowerEdge R630 server platform as the Management Server, Object Storage Servers and Metadata Servers in the configuration. The solution supports Mellanox ConnectX-3 InfiniBand FDR (56 Gb/s) adapters, which takes advantages of the PCIe 3.0 supported by Dell's 13th generation servers. Alternatively, there is also support for 10 Gb/s Ethernet to connect to clients. To characterize the maximum capabilities of the Dell Storage for HPC with Intel EE for Lustre solution, this study focuses only on the performance of the InfiniBand based configuration at FDR speeds. This PowerEdge R630, shown in Figure 3, allows for server density, performance and serviceability of the solution server components in a 1U form factor.

**Figure 3: Dell PowerEdge R630**


## 3.1 Management Server

The Intel Manager Server is a single server connected to the Metadata servers and Object Storage servers via an internal 1GbE network.

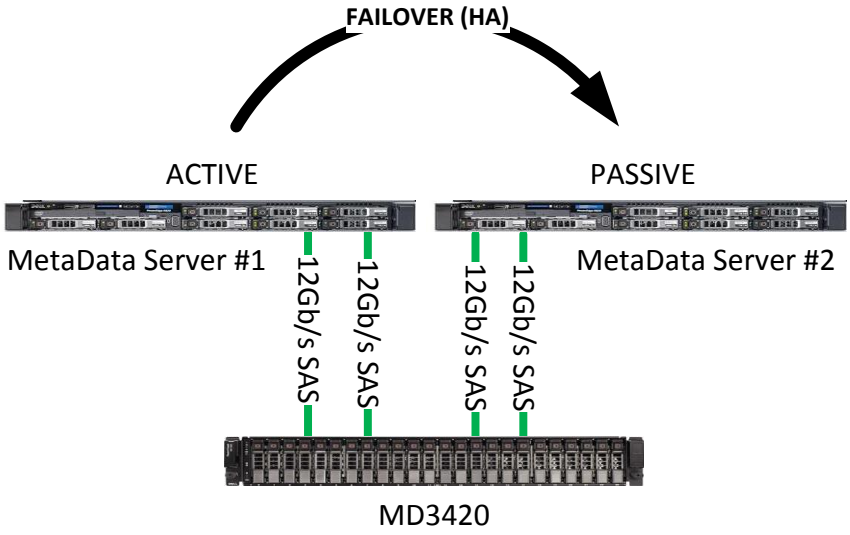
The management server is responsible for user interaction, as well as systems health management and basic monitoring data collected and provided via an interactive web GUI console, the Intel Manager for Lustre. All management and administration access to the Dell Storage for HPC with Intel EE for Lustre will be conducted through this server. While the management server is responsible for collecting Lustre file system related data and provides management for the solution, it does not play an active operational role in the Lustre file system or the data path itself.

The Intel Manager for Lustre (IML) GUI removes the complexities of installation, minimizing Lustre deployment and configuration time. It also automates the monitoring of the health and performance of the different components. The decrease in time and effort for deployment and configuration speeds up the general preparations for production use. The automation of the monitoring provides a better service to end users without increasing the burden for system administrators. In addition, the solution provides tools that help troubleshoot problems related to performance of the file system. Finally, the monitoring tools ability of keeping historical information, allow for better planning for expansion, maintenance and upgrades of the storage appliance.

## 3.2 Metadata Servers

The Metadata Server pair, shown in Figure 4, is comprised of two Dell PowerEdge R630 servers configured as an active/passive highly available cluster. Each server is directly attached to a single Dell PowerVault MD3420 storage array housing the Lustre MDT and MGT. The Dell PowerVault MD3420 is fully populated with 24 - 300 GB, 15K RPM, 2.5" near-line SAS drives configured in a 22 disks RAID10 with 2 hot spares. In this Metadata Target (MDT), the solution provides about 3TiB of space for file system metadata. The MDS is responsible for handling file and directory requests and routing tasks to the appropriate Object Storage Targets for fulfillment. With a single MDT of this size, the maximum number of files that can be served will be in excess of 1.6 billion. In this solution, storage requests are handled across LNet by a single 56 Gb/s FDR InfiniBand connection.

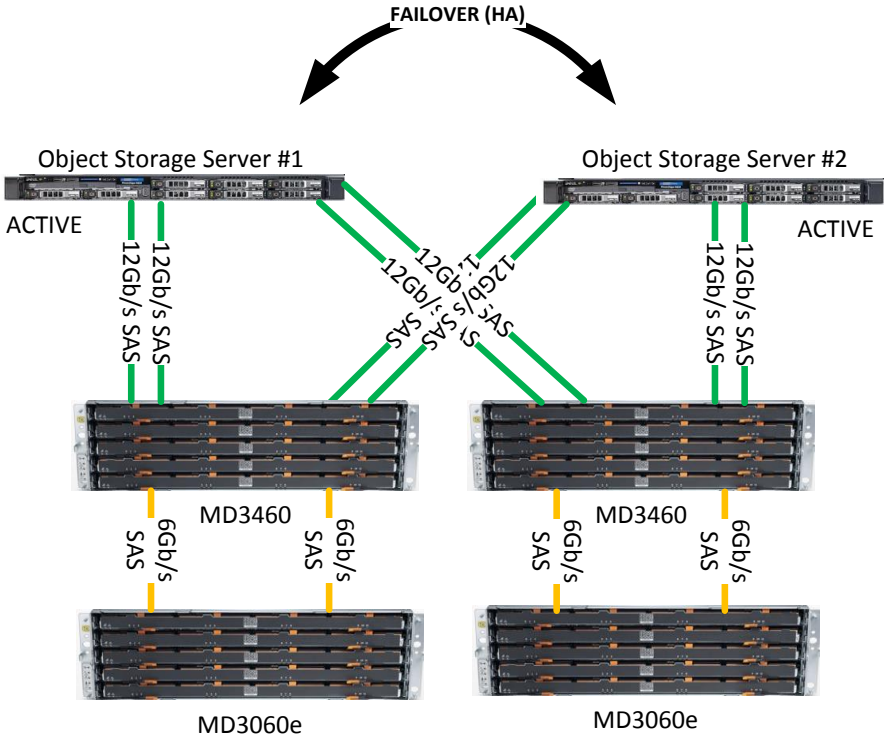
**Figure 4: Metadata Server Pair**




## 3.3 Object Storage Servers

The Object Storage Servers, shown in Figure 5, are arranged in two-node high availability (HA) clusters providing active/active access to two Dell PowerVault MD3460 high-density storage arrays each with MD3060e expansions. Each PowerVault MD3460 array is fully populated with 60 - 4TB 3.5" NL SAS drives. The capacity of each PowerVault MD3460 array is extended with one additional PowerVault MD3060e high density expansion array. This configuration offers each OSS pair with a raw storage capacity of 960TB.

**Figure 5: Object Storage Server Pair**

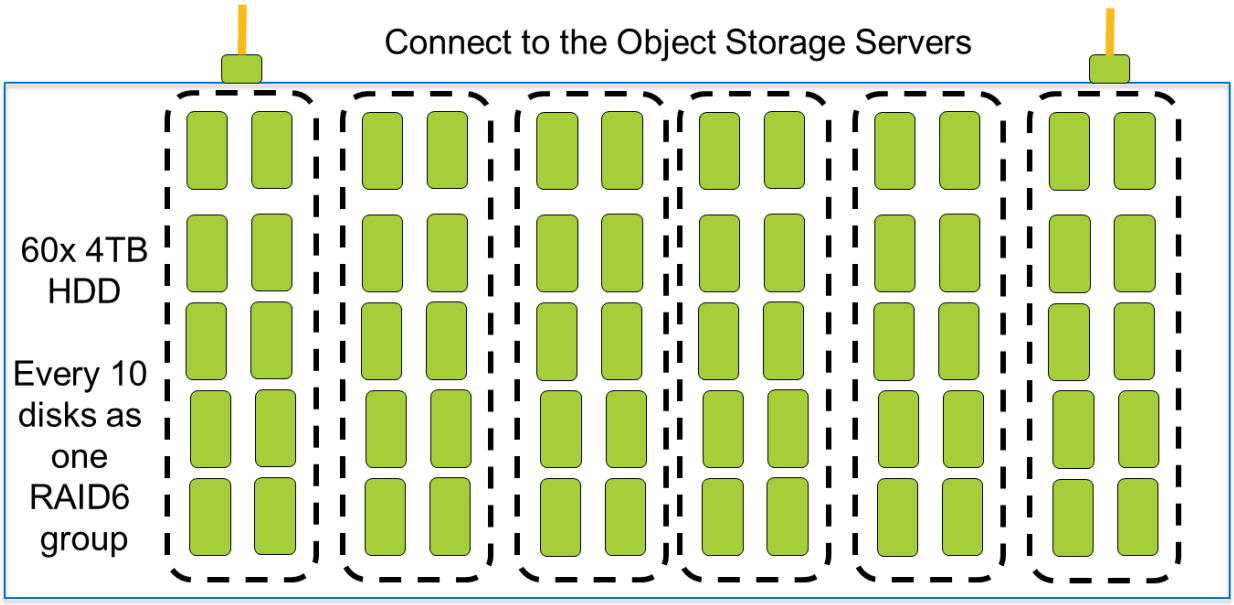



The Object Storage Servers are the building blocks of the solution. With two dual port 12Gb/s SAS controllers in each PowerEdge R630, the two servers are redundantly connected to each of two PowerVault MD3460 high density storage arrays.

Figure 6 illustrates how each storage array is divided into six RAID 6 virtual disks, with eight data and two parity disks per virtual disk, using two disks per tray of the array. That yields six Object Storage Targets per enclosure. By using RAID 6, the solution provides higher reliability at a marginal cost on write performance (due to the extra set of parity data required by each RAID 6). Each OST provides about 29TiB of formatted object storage space. With the Dell Storage for HPC with Intel EE for Lustre solution, a single OSS pair has 24 OSTs by adding the PowerVault MD3060e expansion arrays to the MD3460 arrays. The OSTs are exposed to clients with LNet via 56 Gb/s Infiniband FDR or 10Gb/s Ethernet connections.

When viewed from any compute node equipped with the Lustre client, the entire namespace can be viewed and managed like any other file system, but with the enhancements of Lustre management.

**Figure 6: RAID6 Layout on MD3460 or MD3060e arrays**



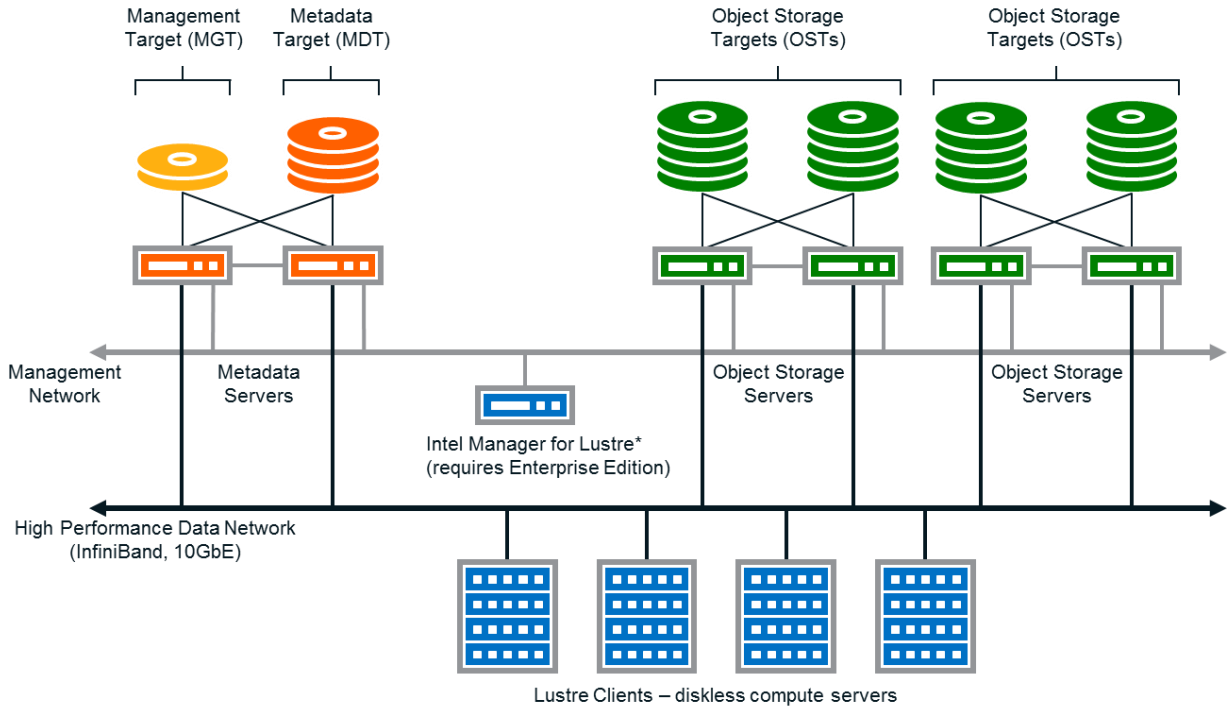

## 3.4 Scalability

Providing the Object Storage Servers in active/active cluster configurations yields greater throughput and product reliability. This configuration provides high availability, decreasing maintenance requirements and consequently reducing potential downtime.

The PowerEdge R630 server provides performance and density. The solution provides 960TB of raw storage for each OSS pair. The solution also leverage FDR InfiniBand interconnect for very high-speed, low-latency storage transactions or 10Gb/s Ethernet can be used for high speed, lower cost and to allow the use of existing 10GbE infrastructures. The PowerEdge R630 takes advantage of the PCIe Gen3 interface for FDR InfiniBand, helping achieve higher network throughput per OSS.

An RPM based Lustre client, version 2.5.23, for the RHEL6.5 kernel with Mellanox OFED v.2.2-1 is available for access to the Dell Storage for HPC with Intel EE for Lustre (for details see **Error! Reference source not found.** Storage for HPC with Intel EE Lustre Configuration Guide).

**Figure 7: OSS Scalability**




Scaling the Dell Storage for HPC with Intel EE for Lustre can be achieved by adding additional OSS pairs with storage backend, demonstrated in Figure 7. Thus will increase both the total network throughput and increase the storage capacity at once. This allows for an increase in the volume of storage available while maintaining a consistent maximum network throughput.

## 3.5 Networking

### *3.5.1 Management Network*

The private management network provides a communication infrastructure for Lustre and Lustre HA functionality as well as storage configuration, monitoring and maintenance. This network creates the segmentation required to facilitate day-to-day operations and to limit the scope of troubleshooting and maintenance. The management server uses this network to interact with the different solution components to query and collect systems health information, as well as to perform any management changes initiated by administrators. Both OSS and MDS servers interact with the management server to provide health information, performance data, and to interact during management operations. The PowerVault MD3420 and MD3460 controllers are accessed via the out of band (ethernet ports) to monitor storage array health, and to perform management actions on the storage backend.

This level of integration allows even an inexperienced operator to efficiently and effortlessly monitor and administer the solution. Information provided is summarized for quick inspection, but users can zoom in to a level of detail of messages from server or storage components.

### *3.5.2 Data Network*

The Lustre file system is served via a Lustre Network (LNet) implementation on either InfiniBand FDR or 10GbE or both. This is the network used by the clients to access data. The Intel Manager for Lustre

(IML) GUI interface provides an option to configure multiple Lustre Network Identifiers (NID) on MDS and OSS servers to participate in the Lustre Network. For instance, you could configure your Infiniband interface (i.e. ib0) as well as your 10GbE Ethernet interface (i.e. eth0) on your OSS servers to both participate in the Lustre Network.

In the InfiniBand network, fast transfer speeds with low latency can be achieved. LNet leverages the use of RDMA for rapid data and metadata transfer to and from MDTs and OSTs to the clients. The OSS and MDS servers take advantage of the FDR InfiniBand fabric with single port Mellanox ConnectX-3 56 Gb adapters. The FDR InfiniBand HBAs can be integrated to existing QDR or DDR networks if needed. With the 10GbE network, Lustre can still benefit from fast transfer speeds and take advantage of the lower cost and pervasiveness of Ethernet technology, leveraging any existing 10GbE infrastructure.

## 3.6 Managing the Dell Storage for HPC with Intel EE for Lustre

The Intel Manager for Lustre (IML) takes the complexity out of administering a Lustre file system by providing a centralized web GUI for management purposes. For example, IML can be used as a tool to standardize the following actions: initiating failover of the file system from one node to another (for either OSS or MDS), formatting the file system, issue mount and unmounts of the targets and monitoring performance of the Lustre file system and the status of its components. Figure 8 illustrates a few of the IML monitoring interfaces.

IML is a web based management console to manage the solution (assuming all security requirements are met). It provides a view of the hardware, software and file system components, while allowing monitoring and management.

Using IML, many tasks that once required complex CLI instructions, can now be completed easily with a few mouse clicks. IML can be used to shut down a file system, initiate a failover from one MDS to another, monitor, etc.

**Figure 8: Intel Manager for Lustre (IML) interface**


# 4. Performance Evaluation and Analysis

The performance studies presented in this paper profile the capabilities of the Dell Storage for HPC with Intel EE for Lustre 240-drive configuration. The configuration has 240 – 4TB disk drives (960 TB raw space). The goal is to quantify the capabilities of the solution, points of peak performance and the most appropriate methods for scaling. The client test bed used to provide I/O workload to test the solution is a Dell HPC compute cluster based on the PowerEdge M610 blade servers, with configuration as described in Table 1.

A number of performance studies were executed, stressing the configuration with different types of workloads to determine the limitations of performance and define the sustainability of that performance. InfiniBand was used for these studies since its high speed and low latency allows getting the maximum performance from Dell Storage for HPC with Intel EE for Lustre, avoiding network bottlenecks.

**Table 1: Test Client Cluster Details**

| Component | Description |
|---|---|
| **Compute Nodes:** | Dell PowerEdge M610, 64 nodes |
| **Node BIOS:** | 6.3.0 |
| **Processors:** | Two Intel Xeon™ X5650 @ 2.67GHz six core processors |
| **Memory:** | 24GiB DDR3 1333MHz |
| **Interconnect:** | InfiniBand - Mellanox Technologies M3601Q (QDR) |

| | |
|---|---|
| **Lustre:** | Lustre 2.5.23 + Mellanox OFED Client |
| **OS:** | Red Hat Enterprise Linux 6.5 (2.6.32-431.el6.x86_64) |
| **IB SOFTWARE:** | Mellanox OFED 2.2-1 |

Performance analysis was focused on three key performance markers:

- Throughput, data sequentially transferred in GB/s.
- I/O Operations per second (IOPS).
- Metadata Operations per second (OP/s).

The goal is a broad but accurate review of the capabilities of the Dell Storage for HPC with Intel EE for Lustre. We selected three benchmarks to accomplish our goal: IOzone, IOR and MDtest.

There are two types of file access methods used with the 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. For this study, we use IOzone for N-to-N access method workloads. The second file access method is N-to-1, where every thread writes to the same file (N clients, 1 file). For this study, we use IOR for N-to-1 access method workloads. IOR can use MPI-IO, HDF5, or POSIX to run N-to-1 file-access tests. For purpose of our analysis, we used POSIX. N-to-1 testing determines how the file system handles the overhead introduced with multiple concurrent requests when multiple threads write or read to the same file. The overhead encountered comes from threads dealing with Lustre's file locking and serialized writes. See Appendix A for examples of the commands used to run these benchmarks.

Each set of tests was run on a range of clients to test the scalability of the solution. The number of simultaneous physical clients involved in each test was varied from a single client to 64 clients. The number of threads corresponds to the number of physical servers up to 64. Total numbers of threads above 64 were achieved by increasing the number of threads per client across all clients. For instance, for 128 threads, each of the 64 clients runs 2 threads.

The test environment for the solution has a single *MDS* and a single *OSS Pair* with a total of 960TB of raw disk space. The OSS pair contains two PowerEdge R630s, each with 256GB of memory, two 12Gbps SAS controllers and a single Mellanox ConnectX-3 FDR HCA. Consult the Dell Storage for HPC with Intel EE for Lustre Configuration Guide for details of cabling and expansion card locations. The MDS has identical configurations with 256GB of memory, a Mellanox ConnectX-3 FDR HCA and dual 12Gbps SAS controllers.

The InfiniBand fabric is comprised of a 32-port Mellanox M3601Q QDR InfiniBand switch for the client cluster and a 36-port Mellanox SX6025 FDR InfiniBand switch for the Dell Storage for HPC with Intel EE for Lustre servers. Three ports from the M3601Q switch were also connected to the SX6025 switch.

Table 2 shows the details about the characteristics for the different software and hardware components.

**Table 2: Dell Storage for HPC with Intel EE for Lustre Configuration**

| **Configuration Size** | 960TB RAW |
|---|---|
| Lustre Server Version | 2.5.23 |
| Intel EE for Lustre Version | v2.1 |
| | |
| **OSS Nodes** | 2 x PowerEdge R630 Servers |
| OSS Memory | 256GiB DDR4 2133MT/s |
| OSS Processors | 2 x Intel Xeon™ E5-2660V3 @ 2.60GHz 10 cores |
| OSS Server BIOS | 0.3.28 |
| OSS Storage Array | 2 x PowerVault MD3460, 2 x PowerVault MD3060e |
| Drives in OSS Storage Arrays | 240 3.5" 4 TB 7.2K RPM NL SAS |
| OSS SAS Controllers | 2 x SAS 12Gbps HBA LSI 9300-8e |
| | |
| **MDS Nodes** | 2 x PowerEdge R630 Servers |
| MDS Memory | 256GiB DDR4 2133MT/s |
| MDS Processors | 2 x Intel Xeon™ E5-2660V3 @ 2.60GHz 10 cores |
| MDS Server BIOS | 0.3.28 |
| MDS Storage Array | 1 x PowerVault MD3420 |
| Drives in MDS Storage Array | 24 - 2.5" 300GB NL SAS |
| MDS SAS Controller | 1 x SAS 12Gbps HBA LSI 9300-8e |
| **Data network - InfiniBand** | |
| OSS, MDS Servers | Mellanox FDR HCA MT27500 |
| Compute Nodes | Mellanox QDR HCA MT26428 |
| Client QDR IB Switch | Mellanox M3601Q |
| HSS5.5 FDR IB Switch | Mellanox 36 Ports SX6036 |
| IB Switch Connectivity | Clients: QDR Cables; Servers: FDR Cables<br>3 uplinks from the QDR switch to the FDR switch |

To prevent inflated results due to caching effects, tests were performed with a *cold cache* established with the following technique. Before each test started, a remount of the Lustre File System under test was executed. A *sync* was performed and the kernel is instructed to drop caches on all the clients with the following commands:

- `sync`
- `echo 3 > /proc/sys/vm/drop_caches`

In addition, to simulate a cold cache on the server, before each test started, on all the active servers (OSS and MDS) a "*sync*" was performed and the kernel is instructed to drop caches with the same commands used on the client.

In measuring the performance of the Dell Storage for HPC with Intel EE for Lustre solution, all tests were performed with similar initial conditions. The file system was configured to be fully functional and the targets tested were emptied of files and directories prior to each test.

## 4.1 N-to-N Sequential Reads / Writes

The sequential testing was done with the IOzone testing tool version 3.429. The throughput results presented in Figure 9 are converted to MB/s. The file size selected for this testing was such that the aggregate sample size from all threads was consistently 2TB. That is, sequential reads and writes have an aggregate sample size of 2TB divided equally among the number of threads within that test. The block size for IOzone was set to 1 MiB to match the 1 MiB Lustre request size.

Each file written was large enough to minimize cache effects from OSS and clients. In addition, the other techniques to prevent cache effects helped to avoid them as well. Files written were distributed evenly across the OSTs (Round Robin). This was to prevent uneven I/O loads on any single SAS connection or OST, in the same way that a user would expect to balance a workload.

**Figure 9: Sequential Reads / Writes Dell Storage for HPC with Intel EE for Lustre**

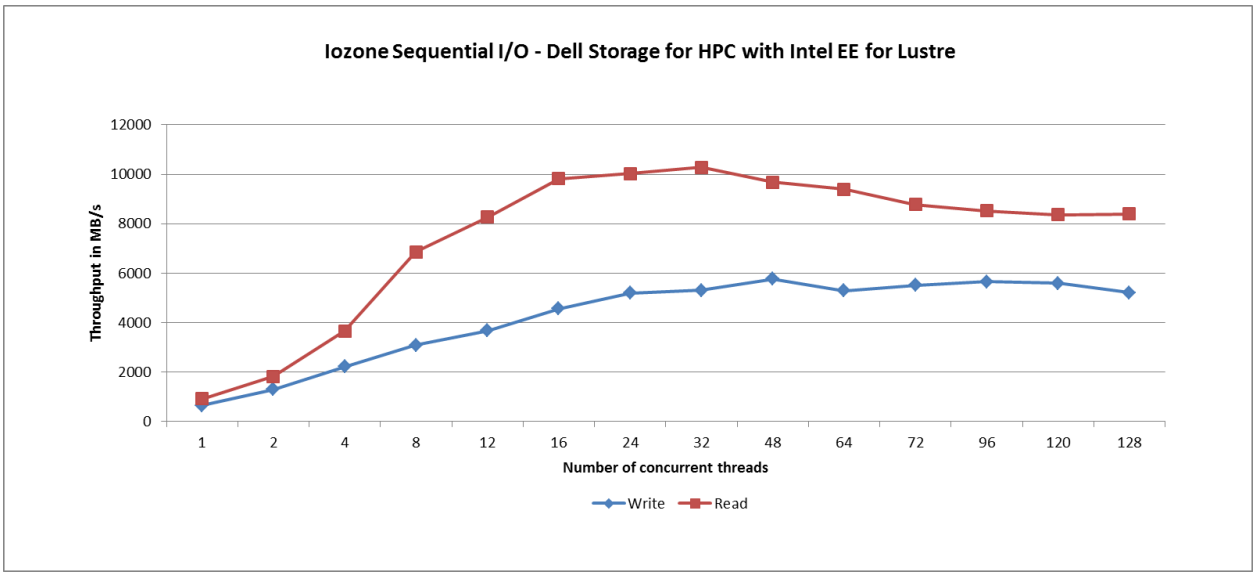



Figure 9 shows the sequential performance of the 960TB test configuration. With the test bed used, write performance peaks near 6GB/sec while read performance peaks near 10GB/sec. Single client performance has reads at 926MB/sec with writes at 645MB/sec. The write and read performance rises steadily as we increase the number of process threads up to 32 for reads and 48 for writes. This is partially a result of increasing the number of OSTs utilized, as the number of threads is increased (up to the 24 OSTs in our system).

To maintain the higher throughput for an even greater number of files, increasing the number of OSTs is likely to help. A review of the storage array performance using the tools provided by the Dell PowerVault Modular Disk Storage Manager, Performance Monitor was performed to independently confirm the throughput values produced by the benchmarking tools.

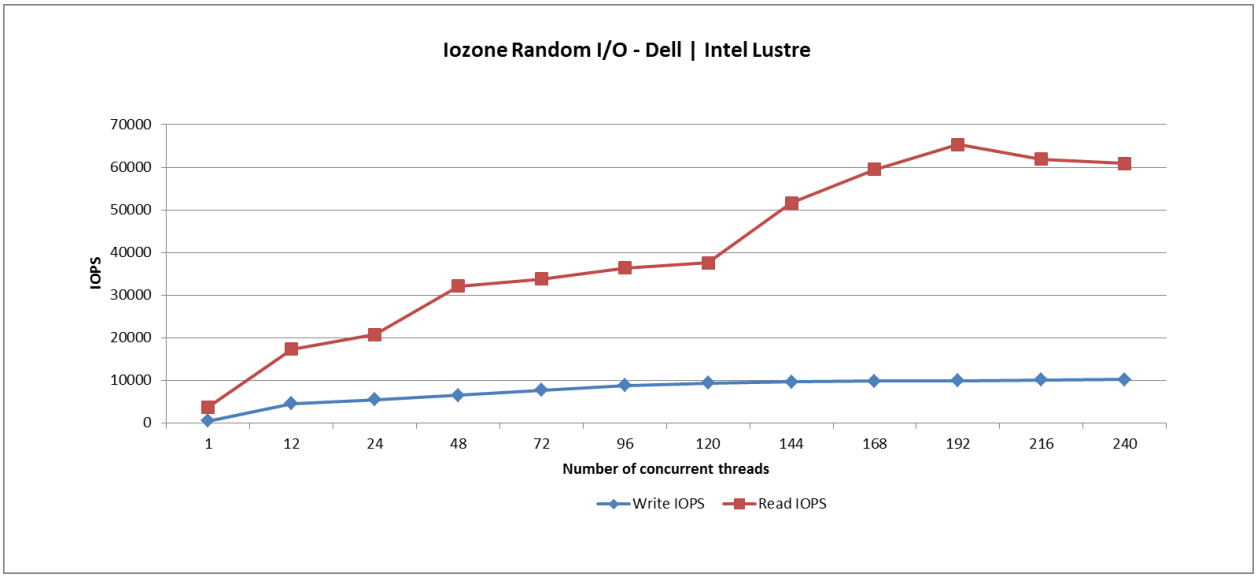
## 4.2 Random Reads and Writes

The IOzone benchmark was used to gather random reads and writes metrics. The file size selected for this testing was such that the aggregate size from all threads was consistently 1TB. That is, random reads and writes have an aggregate size of 1TB divided equally among the number of threads within that test. The IOzone host file is arranged to distribute the workload evenly across the compute nodes. The storage is addressed as a single volume with a stripe count of 1 and stripe size of 4MB. A 4KB

request size is used because it aligns with Lustre's 4KB file system block size and is representative of small block accesses for a random workload. Performance is measured in I/O Operations per second (IOPS)

Figure 10 shows that the random writes peak at little over 10K IOPS with 240 threads, while random reads peak at 65K IOPS with 192 threads. The IOPs of random reads increase rapidly from 120 to 192 threads and then slight decline before leveling. As the writes require a file lock per OST accessed, saturation is not unexpected. Reads take advantage of Lustre's ability to grant overlapping read extent locks for part or all of a file.

**Figure 10: N-to-N Random reads and writes**




## 4.3 IOR N-to-1 Reads and Writes

Performance analysis of the Dell Storage for HPC with Intel EE for Lustre solution with reads and writes to a single file was done with the IOR benchmarking tool. IOR accommodates MPI communications for parallel operations and has support for manipulating Lustre striping. IOR allows several different IO interfaces for working with the files. For purpose of our tests, we used the POSIX interface to exclude the advanced features and associated overhead of the other available IO interfaces. This gives us an opportunity to review the file system and hardware performance independent of those additional enhancements.

IOR benchmark version 3.0.1 was used in this study. The MPI stack used was Intel MPI version 5.0 Update 1.

The configuration for the write test included a directory set with striping characteristics designed to stripe across 24 OSTs with a stripe size of 4MB. Therefore, all threads write to a file that is striped across all 24 OSTs. In this test, the request size for Lustre was set to 1MB, however, a transfer size of 4MB was used to match the stripe size used on the target file.

In order to reduce the cache effects from server and client memory, it was decided to use a file size that was twice the combined memory size of the OSSs and the clients' memory, according to the following formula and rounding to whole values where necessary:

> **File Size** = 2 * (2 OSSs*256 GiB memory per OSS + Number of physical clients * 24 GiB memory per client).

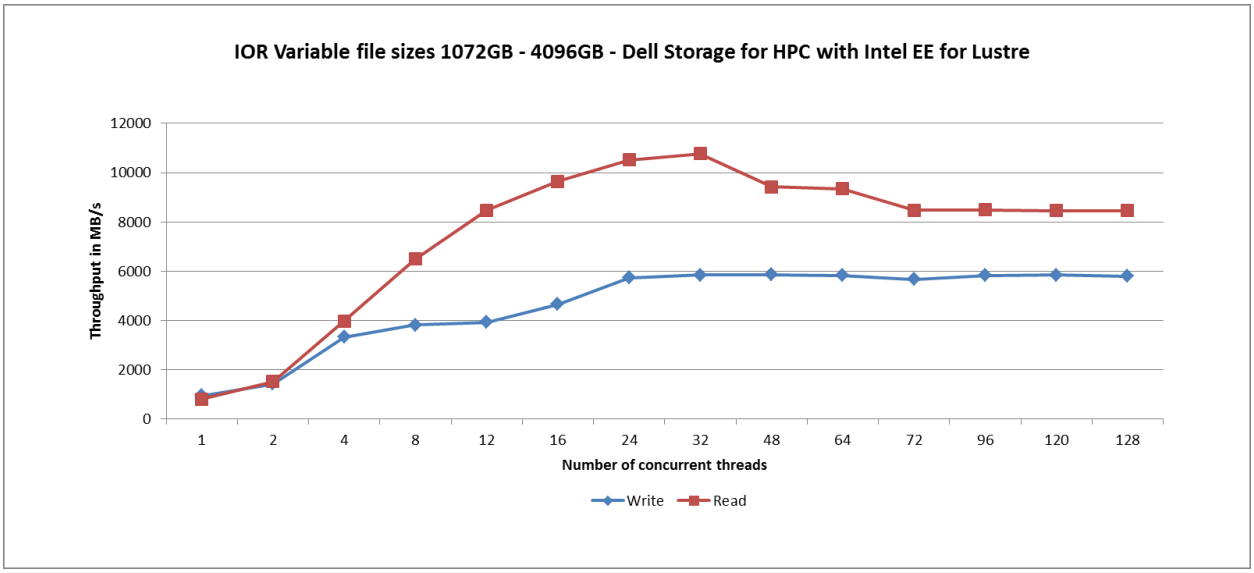
Table 3 shows the size of the data manipulated by each set of clients, the number of threads, and the size of the total shared file.

**Table 3: IOR Shared File Size**

| Number of Threads | Number of Physical Clients | Data Written per Thread (GB) | Shared File Size (GB) |
|---|---|---|---|
| 1 | 1 | 1072 | 1072 |
| 2 | 2 | 560 | 1120 |
| 4 | 4 | 304 | 1216 |
| 8 | 8 | 176 | 1408 |
| 12 | 12 | 133 | 1600 |
| 16 | 16 | 112 | 1792 |
| 24 | 24 | 91 | 2176 |
| 32 | 32 | 80 | 2560 |
| 48 | 48 | 69 | 3328 |
| 64 | 64 | 64 | 4096 |
| 72 | 64 | 57 | 4096 |
| 96 | 64 | 43 | 4096 |
| 120 | 64 | 34 | 4096 |
| 128 | 64 | 32 | 4096 |
| 144 | 64 | 28 | 4096 |
| 168 | 64 | 24 | 4096 |
| 192 | 64 | 21 | 4096 |
| 216 | 64 | 19 | 4096 |
| 240 | 64 | 17 | 4096 |
| 256 | 64 | 16 | 4096 |

Figure 11 shows the IOR results, where reads have the advantage and peak at 10.7GB/s at 32 threads with trends very similar to the sequential N-to-N performance. Write performance increases steadily as threads to OST ratio increases reaching a plateau at 24 threads, the same number of OSTs available in the test system. Peak write is at 6.2GB/s. Single client IOR performance has reads at 806MB/sec and writes at 948MB/sec.

**Figure 11: N-to-1 IOR Read / Write**




## 4.4 Metadata Testing

Metadata testing measures the time to complete certain file or directory operations that return attributes. *MDtest* is an MPI-coordinated benchmark that performs Create, Stat, and Remove operations on files or directories. This study used MDtest version 1.9.3. The MPI stack used for this study was Intel MPI version 5.0 Update 1. The metric reported by *MDtest* is the rate of completion in terms of operations per second (OP/sec). MDtest can be configured to compare metadata performance for directories and files. For this test, we perform a pass with file operations, then a separate pass with directory operations.

On a Lustre file system, OSTs are queried for object identifiers in order to allocate or locate extents associated with the metadata operations. This interaction requires the indirect involvement of OSTs in most metadata operations. In lab experiments with the test bed, it was found that using OST count of 1 was more efficient for most metadata operations, especially for higher thread counts. The experiments consisted of conducting tests with up to 64 clients for the following Lustre topologies (results not presented here), to select the most efficient configuration for the metadata performance tests for this version of the solution:

- 1 OST, 1MB Stripes
- 1 OST, 4MB Stripes
- 24 OSTs, 1MB Stripes
- 24 OSTs, 4MB Stripes

The most efficient configuration was found to be with 1 OST with 1MB stripes, therefore the results presented in this section are with such Lustre topology.

Also during the preliminary metadata testing, it was found that the number of files per directory significantly affects the results, even while keeping constant the total number of files created. For

example when testing with 64 threads creating 3125 files per directory in 5 directories per thread OR creating 625 files per directory in 25 directories per thread, both result in the creation of 1 million files, but the measured performance in IOPS is not the same. This is due to overhead of seeks performed on the OSTs when changing directories. In order to present coherent results, the number of files per directory was fixed at 3125 while we varied the number of directories per thread to yield total number of files to be no less than 1 million and steadily increasing as thread counts increase. Table 4 represents the values used in each test. For both the file operation tests as well as the directory operations tests, we performed the tests with eight iterations each taking the average value for recorded results.

**Table 4: Parameters used on MDtest**

| Number of Threads (N) | Number of Files per Directory | Number of Directories per thread | Total number of Files |
|---|---|---|---|
| 1 | 3125 | 320 | 1000000 |
| 2 | 3125 | 160 | 1000000 |
| 4 | 3125 | 80 | 1000000 |
| 8 | 3125 | 40 | 1000000 |
| 12 | 3125 | 27 | 1012500 |
| 16 | 3125 | 20 | 1000000 |
| 24 | 3125 | 14 | 1050000 |
| 32 | 3125 | 10 | 1000000 |
| 48 | 3125 | 7 | 1050000 |
| 64 | 3125 | 5 | 1000000 |
| 72 | 3125 | 5 | 1125000 |
| 96 | 3125 | 4 | 1200000 |
| 120 | 3125 | 3 | 1125000 |
| 128 | 3125 | 3 | 1200000 |
| 144 | 3125 | 3 | 1350000 |
| 168 | 3125 | 2 | 1050000 |
| 192 | 3125 | 2 | 1200000 |
| 216 | 3125 | 2 | 1350000 |
| 240 | 3125 | 2 | 1500000 |

**Figure 12: File Metadata Operations**

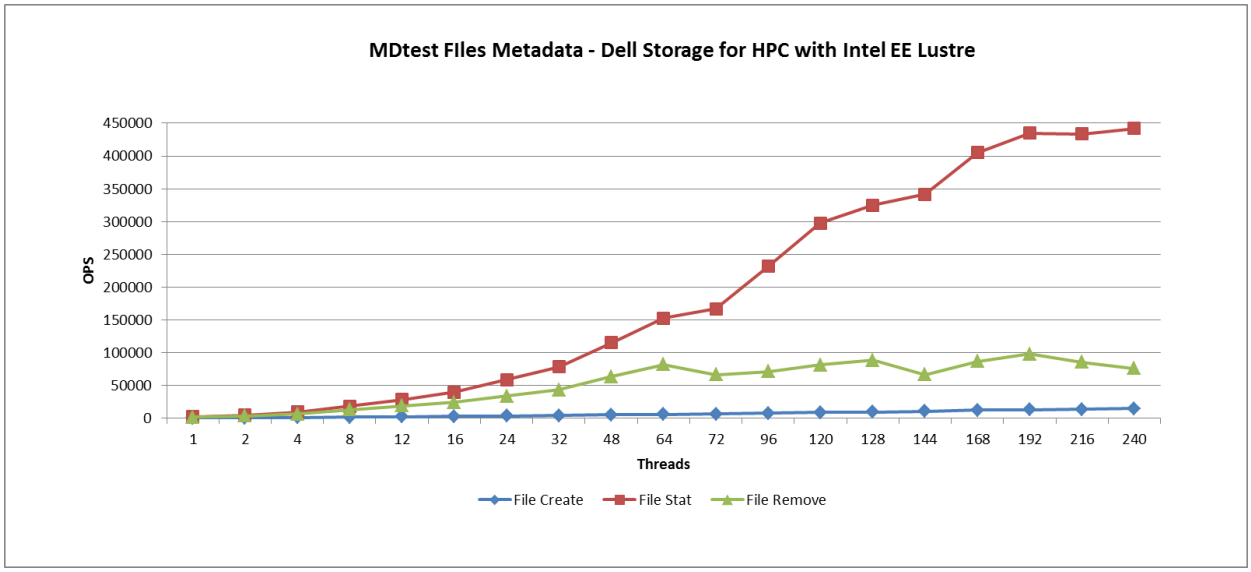


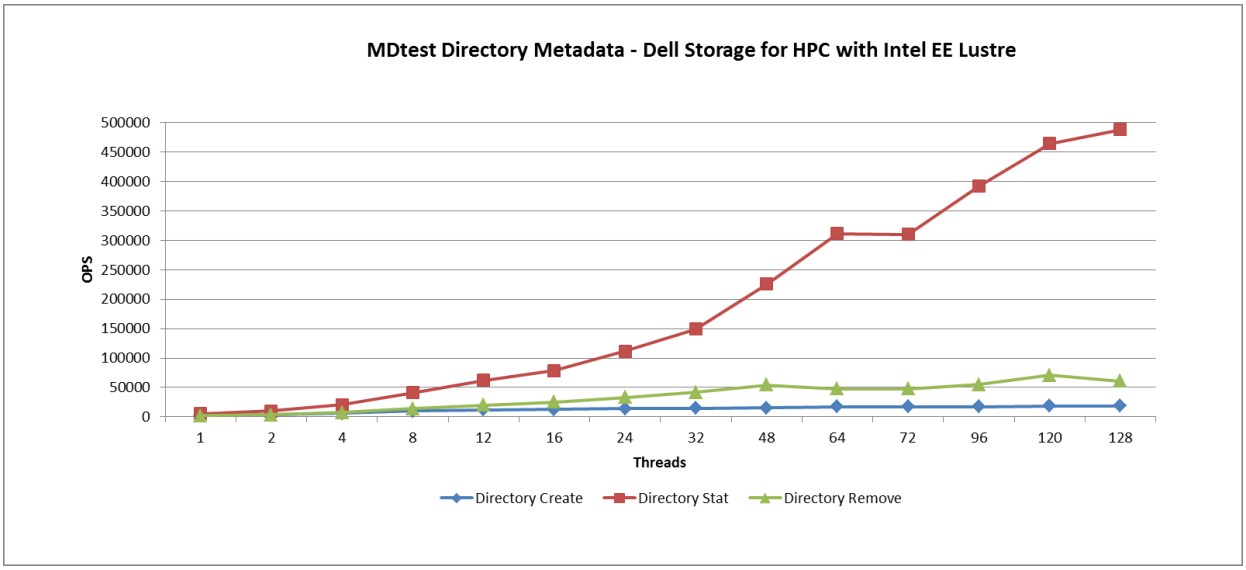

Figure 12 illustrates the file metadata results using MDtest. From this graph, file create metadata operations start with less than 504 OPS at 1 thread and scale to almost 15K OPS with 240 concurrent threads. This may be due to the Lustre locks needed on the MDT, but also those on the OSTs, since using a stripe count of 24 had a significant decrease in performance. At 240 threads, we had 2 directories (-b 2) and 2.2 million files were created.

File stat metadata operations is overwhelmingly the lightest metadata operation of the three observed. A single thread test yield over 2K OPS and scale to more than 400K OPS with 240 concurrent threads. The increase in performance could be due to improvements made in Lustre version 2.5 metadata operations. Also of note is the use of 15K RPM drives in the MDT volumes.

Removal of files is also limited by accesses to OSTs, similar to the create operation. However, the remove operation have advantage over create operations when increasing in total threads, starting with over 1.8K OPS at 1 thread and scaling to almost 100K OPS with 192 concurrent threads.

Figure 13 illustrates the directory metadata results using MDtest. From this graph, directory create metadata operations is again the most expensive operation for most cases, starting with 1.8K OPS at 1 thread reaching a maximum of almost 19K at 128 threads. Directory operations are also affected by the number of top directories used (-b), but to a lesser degree than file operations. Remove operations is almost as expensive as directory createsas it starts with almost 1.9K OPS at 1 thread and reaches a maximum of over 70K OPS at 120 threads. Directory Stats operation is again overwhelmingly the lightest of the three observed. A single thread test yield over 5K OPS at 1 thread and grows to over 450K at 128 threads.

**Figure 13: Directory Metadata Operations**




# 5. Conclusions

There is a well-known requirement for scalable, high-performance clustered file system solutions. The Dell Storage for HPC with Intel EE for Lustre addresses this need with a well-designed solution that is easy to manage and fully-supported. The solution includes the added benefit of the Dell PowerEdge™ 13th generation servers platform and PowerVault™ storage products and Lustre® technology, the leading open source solution for a parallel file system. The Intel Manager for Lustre (IML) unifies the management of the Lustre File System and solution components into a single control and monitoring panel for ease of use.

The scale of the raw storage, 960TB per Object Storage Server Pair and up to 10GB/s of read and 6GB/s of write throughput in a packaged component, is consistent with the needs of the high performance computing environments. The Dell Storage for HPC with Intel EE for Lustre is also capable of scaling in throughput as easily as it scales in capacity.

The performance studies demonstrate a high throughput for both reads and writes for N-to-N as well as N-to-1 file type access. Results from MDtest show an elevated capacity for the metadata file operations. With the PCI-e 3.0 interface, IB FDR HCAs contributes to excel in high bandwidth applications.

The continued use of generally available, industry-standard benchmark tools like IOzone, IOR and MDtest provide an easy way to match current and expected growth with the performance outlined. The profiles reported from each of these tools provide sufficient information to align the configuration of the Dell Storage for HPC with Intel EE for Lustre with the requirements of many applications or group of applications.

The Dell Storage for HPC with Intel EE for Lustre solution delivers all the benefits of a scale-out parallel file system-based storage for your high performance computing needs.

# Appendix A: Benchmark Command Reference

This section describes the commands used to benchmark the Dell Storage for HPC with Intel EE Lustre solution.

**IOzone**

IOzone Sequential Writes -

```
iozone -i 0 -c -e -w -r 1024K -I -s $Size -t $Thread -+n -+m /root/list.$Thread
```

IOzone Sequential Reads -

```
iozone -i 1 -c -e -w -r 1024K -I -s $Size -t $Thread -+n -+m /root/list.$Thread
```

IOzone IOPS Random Reads / Writes -

```
iozone -i 2 -w -c -O -I -r 4K -s $Size -t $Thread -+n -+m /root/list.$Thread
```

| IOzone Command Line | 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 |

The O_Direct command line parameter ("-I") allows us to bypass the cache *on the compute nodes* where the IOzone threads are running.

**IOR**

IOR Writes -

```
mpirun -np $Threads -rr --machinefile /root/list.$Threads /cm/share/IOR -a POSIX -v -i $rep -d 3
-e -k -o /mnt/boulder/perf/mytestfile -w -s 1 -t 4m -b $SizePerThread
```

IOR Reads -

```
mpirun -np $Threads -rr --machinefile /root/list.$Threads /cm/share/IOR -a POSIX -v -i $rep -d 3
-e -k -o /mnt/boulder/perf/mytestfile -r -s 1 -t 4m -b $SizePerThread
```

| IOR Command Line Arguments | Description |
|---|---|
| -a S | api -- API for I/O [POSIX\|MPIIO\|HDF5\|NCMPI] |
| -v | verbose -- output information (repeating flag increases level) |
| -I N | repetitions -- number of repetitions of test |
| -d N | interTestDelay -- delay between reps in seconds |
| -e | fsync -- perform fsync upon POSIX write close |
| -k | keepFile -- don't remove the test file(s) on program exit |
| -o S | testFile -- full name for test |
| -w | writeFile -- write file |
| -r | readFile -- read existing file |
| -s N | segmentCount -- number of segments |
| -t N | transferSize -- size of transfer in bytes (e.g.: 8, 4k, 2m, 1g) |
| -b N | blockSize -- contiguous bytes to write per task (e.g.: 8, 4k, 2m, 1g) |

**MDtest - Metadata**

Files Operations -

```
mpirun -np $Threads -rr --hostfile /share/mdt_clients/mdtlist.$Threads /share/mdtest/mdtest.intel
-v -d /mnt/lustre/perf_test24-1M -i $Reps -b $Dirs -z 1 -L -I $Files -y -u -t -F
```

Directories Operations -

```
mpirun -np $Threads -rr --hostfile /share/mdt_clients/mdtlist.$Threads /share/mdtest/mdtest.intel
-v -d /mnt/lustre/perf_test24-1M -i $Reps -b $Dirs -z 1 -L -I $Files -y -u -t -D
```

| MDtest Command Line Arguments | Description |
|---|---|
| -d | the directory in which the tests will run |
| -v | verbosity (each instance of option increments by one) |
| -i | number of iterations the test will run |
| -b | branching factor of hierarchical directory structure |
| -z | depth of hierarchical directory structure |
| -L | files only at leaf level of tree |
| -I | number of items per directory in tree |
| -y | sync file after writing |
| -u | unique working directory for each task |
| -t | time unique working directory overhead |
| -F | perform test on files only (no directories) |
| -D | perform test on directories only (no files) |

# References

Dell Storage for HPC with Intel EE for Lustre Solution Brief
http://salesedge.dell.com/doc?id=0901bc82808e334f&ll=d&pm=160376162

Dell Storage for HPC with Intel EE for Lustre Configuration Guide
* Contact your Dell Sales Rep for this document

Dell PowerVault MD3420
http://www.dell.com/support/home/us/en/04/product-support/product/powervault-md3420/research

Dell PowerVault MD3460 & MD3060e
http://www.dell.com/support/home/us/en/04/product-support/product/powervault-md3460/research

Lustre Home Pages
http://www.whamcloud.com/lustre/
http://wiki.lustre.org/index.php/Main_Page

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

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

Intel Home Page
http://www.intel.com

Dell Storage for HPC with Intel Enterprise Edition for Lustre sofware

Intel HPDD Wiki
https://wiki.hpdd.intel.com/display/PUB/HPDD+Wiki+Front+Page

Mellanox Technologies Home Page
http://www.mellanox.com

LSI 12Gb/s SAS HBA
http://www.lsi.com/downloads/Public/Host%20Bus%20Adapters/LSI_PB_SAS9300_HBA_Family.pdf