Dell HPC Lustre Storage

A Dell Technical White Paper

Quy Ta

Dell HPC Engineering Innovations Lab

September 2016 | Version 1.1

Contents

Figures

Tables

1. Introduction

In high performance computing (HPC), the efficient delivery of data to and from the compute nodes is critical and often complicated to execute. Researchers can generate and consume data in HPC systems at such speed that turns the storage components into a major bottleneck. Getting maximum performance for their applications require a scalable storage solution. Open source Lustre solutions can deliver on the performance and storage capacity needs. However, the managing and monitoring of such a complex storage system will 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 HPC Lustre Storage solution with Intel Omni-Path, referred to as Dell HPC Lustre Storage, 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. The solution uses the Intel® Enterprise Edition (EE) for Lustre® software v.3.0. It 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 deploying, 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 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 HPC Lustre Storage 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 and the Dell HPC Lustre Storage solution, followed by performance analysis, conclusions and Appendix.

2. The Lustre File System

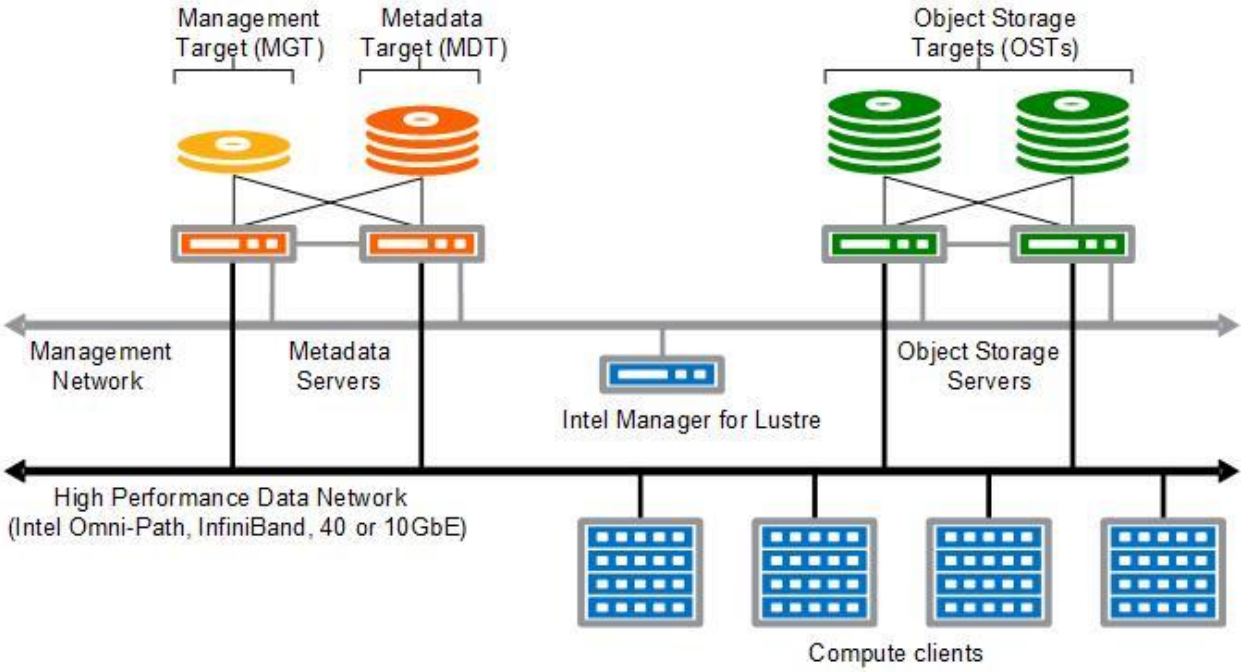
Lustre is a parallel file system, offering high performance through parallel access to data and distributed locking. A Lustre installation consists of three key elements: the metadata subsystem, the object storage subsystem (data) and the compute clients that access and operate on the data.

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. Dell HPC Lustre Storage also supports Lustre DNE (Distributed Namespace) configuration option.

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 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) ideally in

pairs. 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 HPC Lustre Storage utilizes Intel Omni-Path as the network to connect the clients, MDSs and OSSs, LNet enables Lustre to take advantage of the Performance Scaled Messaging (PSM) protocol, an performance optimized path for MPI traffic to provide faster I/O transport and lower latency compared to 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

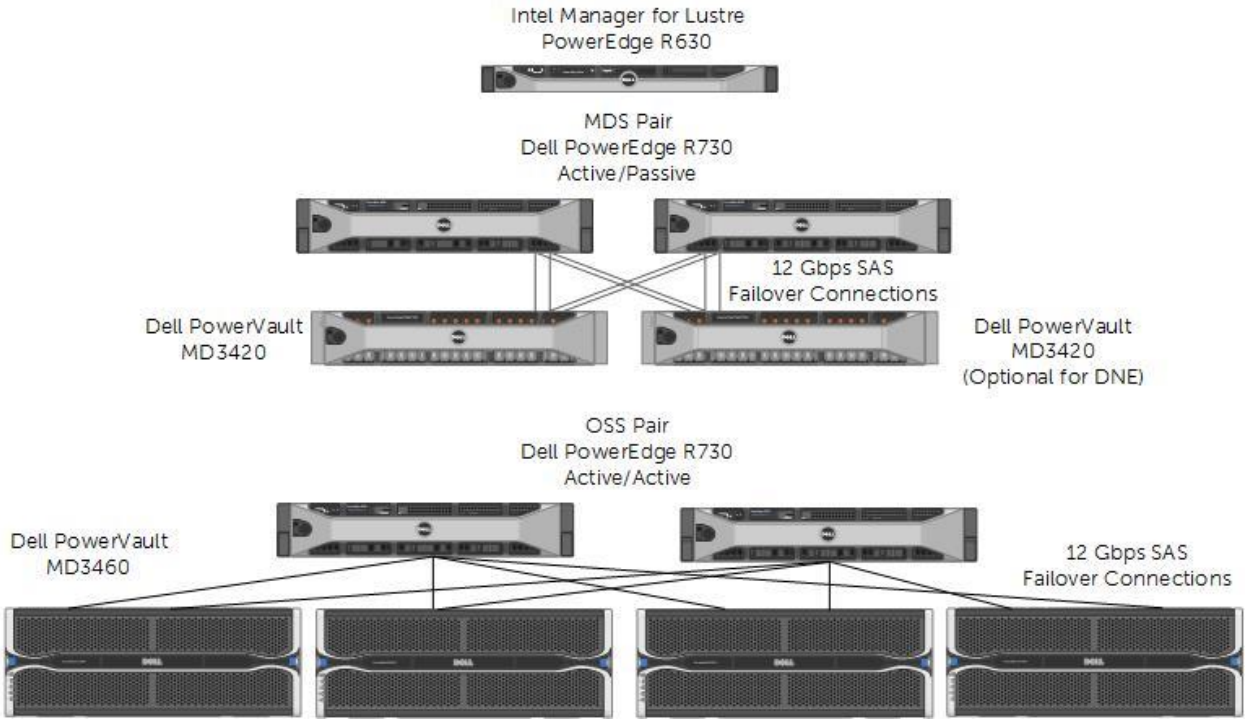
Typically, Lustre configurations and deployments are considered very complex and time-consuming tasks. Lustre installation and administration is normally done via a command line interface (CLI), requiring extensive knowledge of the file system operation, along with the auxiliary tools like LNet and the locking mechanisms. In addition, once the Lustre storage system 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 System 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 HPC Lustre Storage with Intel EE for Lustre Description

The Dell HPC Lustre Storage 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 HPC Lustre Storage 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)

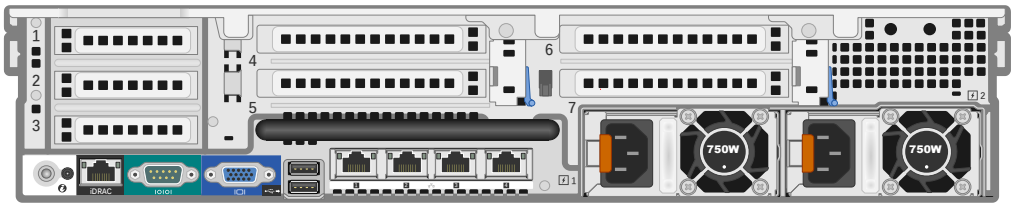
Figure 2: Dell HPC Lustre Storage Solution Components Overview



There are several new architectural changes in this release compared to previous release. The solution continues to use the Dell PowerEdge R630 server as the Intel Management Server, while the Object Storage Servers and Metadata Servers in the configuration will be based on the Dell PowerEdge R730. The PowerEdge R730 is a 2U dual socket server with 7 PCIe expansion slots available enabling future expansion. Also new to this solution is support of the new Intel Omni-Path Host Fabric Interface (HFI) adapter. The storage backend of the solution will utilize the PowerVault MD3460 fully populated with your choice of 4TB, 6TB or 8TB SAS hard disk drives. For the Lustre software, the solution supports version 3.0 of the Intel Enterprise Edition for Lustre software. More detailed overview of the components and solution architecture are in sections to follow.

To characterize the maximum capabilities of the Dell HPC Lustre Storage solution, this study will focus only on the performance of the Intel Omni-Path based configuration. This PowerEdge R730, shown in Figure 3, allows for server expansion, performance and serviceability of the solution server components in a 2U form factor.

Figure 3: Dell PowerEdge R730



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.

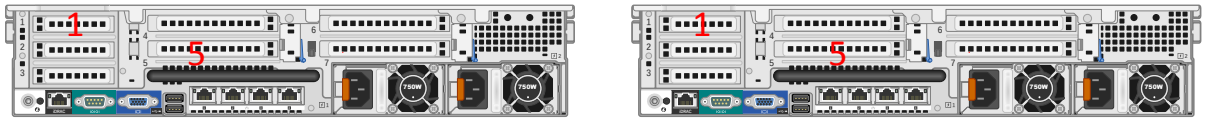
The management server is responsible for user interaction, as well as systems health management and basic monitoring data collected and provided via an interactive web GUI console, the Intel Manager for Lustre. All management and administration access to the Dell HPC Lustre Storage with Intel EE for 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 complexities from the installation process, minimizing time to deploy and configure the Lustre system. 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 730 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 – 300GB, 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 3TB 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 Intel Omni-Path HFI connection.

Figure 4: Metadata Server Pair

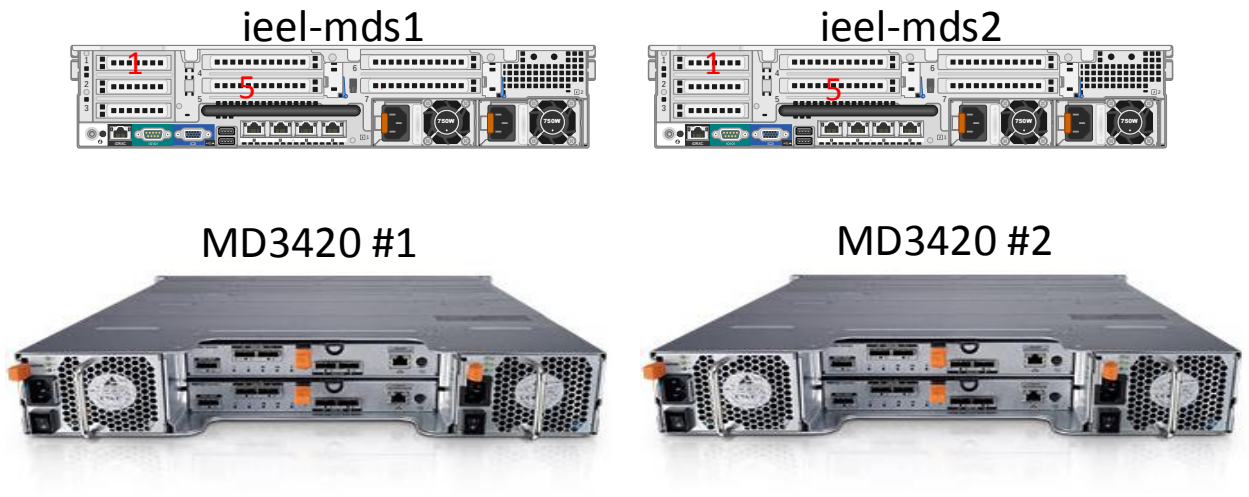


MD3420 #1

SERVER	SAS PCI SLOT	SAS PORT	MD3420 ARRAY	MD3420 CONTROLLER	MD3420 CONTROLLER PORT
ieel3-mds1	Slot 1	Port 0	MD3420 #1	Controller 0	Port 0
ieel3-mds1	Slot 5	Port 0	MD3420 #1	Controller 1	Port 0
ieel3-mds2	Slot 1	Port 0	MD3420 #1	Controller 0	Port 1
ieel3-mds2	Slot 5	Port 0	MD3420 #1	Controller 1	Port 1

With the Lustre DNE configuration option, shown in Figure 5, the two Dell PowerEdge R730 servers will be configured in an active/active, highly available cluster. Each server will be directly attached to two (2) Dell PowerVault MD3420 storage arrays. One MD3420 will house the Lustre MGT as well as one of the MDT while the second MD3420 will house the second MDT. The servers will be configured such that each server will be primary for one MDT while secondary for the other MDT. Both MD3420 will be fully populated with 24 – 300GB, 15K RPM, 2.5" SAS drives configured in a 22 disk RAID10 with 2 hot spares. The Lustre DNE option will effectively double the space available for file system metadata and the number of files that can be served.

Figure 5: Metadata Server Pair with Lustre DNE option

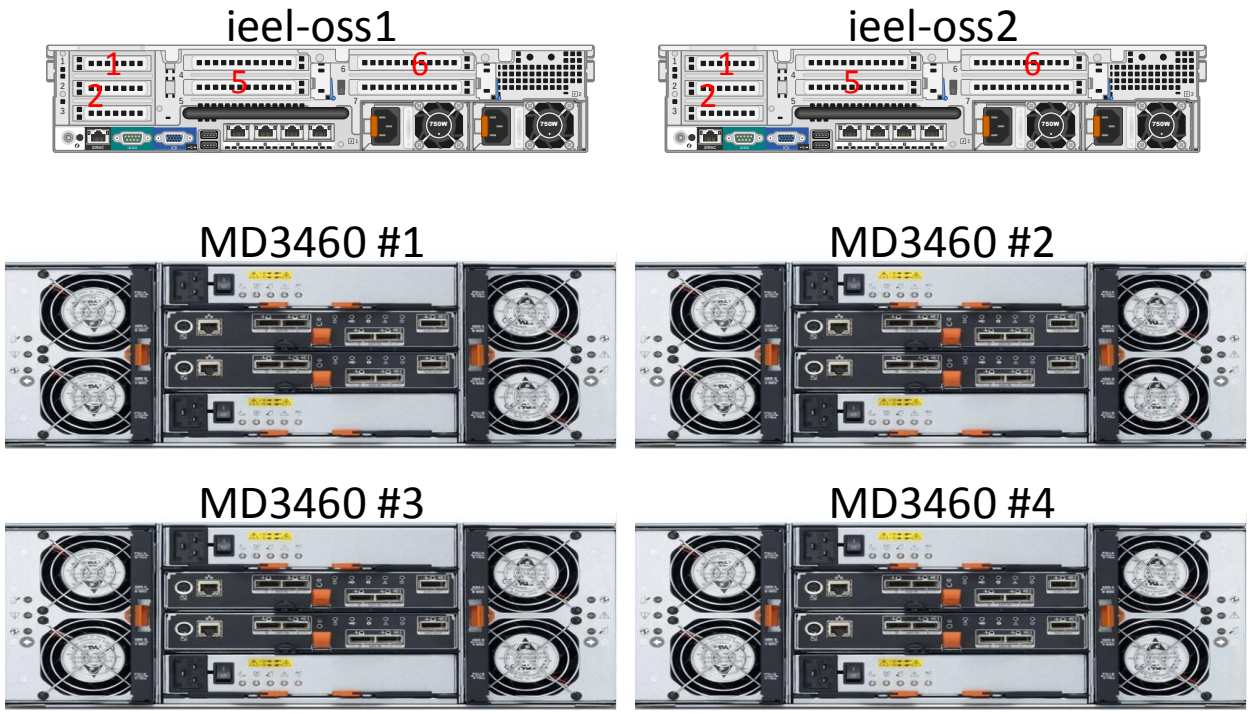


SERVER	SAS PCI SLOT	SAS PORT	MD3420 ARRAY	MD3420 CONTROLLER	MD3420 CONTROLLER PORT
ieel3-mds1	Slot 1	Port 0	MD3420 #1	Controller 0	Port 0
ieel3-mds1	Slot 5	Port 0	MD3420 #1	Controller 1	Port 0
ieel3-mds1	Slot 1	Port 1	MD3420 #2	Controller 0	Port 0
ieel3-mds1	Slot 5	Port 1	MD3460 #2	Controller 1	Port 0
ieel3-mds2	Slot 1	Port 0	MD3420 #1	Controller 0	Port 1
ieel3-mds2	Slot 5	Port 0	MD3420 #1	Controller 1	Port 1
ieel3-mds2	Slot 1	Port 1	MD3460 #2	Controller 0	Port 1
ieel3-mds2	Slot 5	Port 1	MD3460 #2	Controller 1	Port 1

3.3 Object Storage Servers

The Object Storage Servers, shown in Figure 6, are arranged in two-node high availability (HA) clusters providing active/active access to four Dell PowerVault MD3460 high-density storage arrays. Each PowerVault MD3460 array is fully populated with 60 – 4TB, 6TB or 8TB 3.5” NL SAS drives. The capacity of each PowerVault MD3460 array can be extended with one additional PowerVault MD3060e high-density expansion array. This standard configuration using 4TB drives offers each OSS pair with a raw storage capacity of 960TB.

Figure 6: Object Storage Server Pair



SERVER	SAS PCI SLOT	SAS PORT	MD3460 ARRAY	MD3460 CONTROLLER	MD3460 CONTROLLER PORT
ieel3-oss1	Slot 1	Port 0	MD3460 #1	Controller 0	Port 0
ieel3-oss1	Slot 1	Port 1	MD3460 #2	Controller 0	Port 0
ieel3-oss1	Slot 2	Port 0	MD3460 #2	Controller 1	Port 0
ieel3-oss1	Slot 2	Port 1	MD3460 #1	Controller 1	Port 0
ieel3-oss2	Slot 1	Port 0	MD3460 #1	Controller 0	Port 2
ieel3-oss2	Slot 1	Port 1	MD3460 #2	Controller 0	Port 2
ieel3-oss2	Slot 2	Port 0	MD3460 #2	Controller 1	Port 2
ieel3-oss2	Slot 2	Port 1	MD3460 #1	Controller 1	Port 2
ieel3-oss1	Slot 5	Port 0	MD3460 #3	Controller 0	Port 0
ieel3-oss1	Slot 5	Port 1	MD3460 #4	Controller 0	Port 0
ieel3-oss1	Slot 6	Port 0	MD3460 #4	Controller 1	Port 0
ieel3-oss1	Slot 6	Port 1	MD3460 #3	Controller 1	Port 0
ieel3-oss2	Slot 5	Port 0	MD3460 #3	Controller 0	Port 2
ieel3-oss2	Slot 5	Port 1	MD3460 #4	Controller 0	Port 2
ieel3-oss2	Slot 6	Port 0	MD3460 #4	Controller 1	Port 2
ieel3-oss2	Slot 6	Port 1	MD3460 #3	Controller 1	Port 2

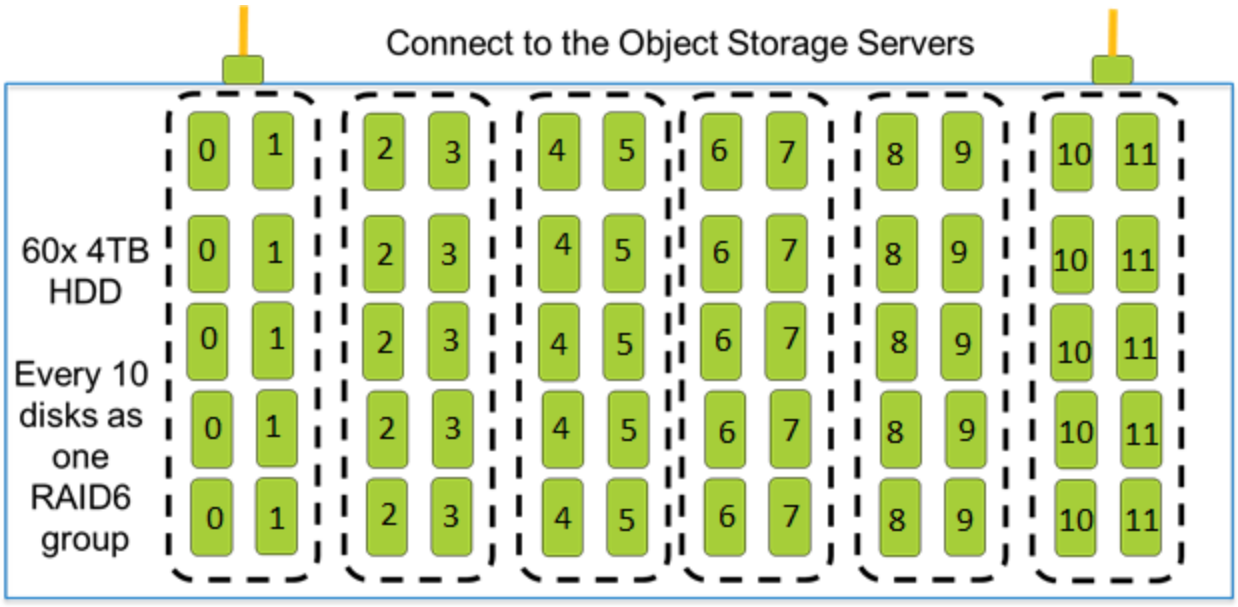
The Object Storage Servers are the building blocks of the solution. With four dual port 12Gb/s SAS controllers in each PowerEdge R730, the two servers are redundantly connected to each of the four PowerVault MD3460 high density storage arrays.

Figure 7 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 29TB of formatted object storage space when populated with 4TB HDD. With the Dell HPC Lustre Storage solution, each MD3460 provides 6 OSTs. The OSTs are exposed to clients with LNet via Intel Omni-Path 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, only with the enhancements of Lustre management.

Figure 7: RAID6 Layout on MD3460 arrays



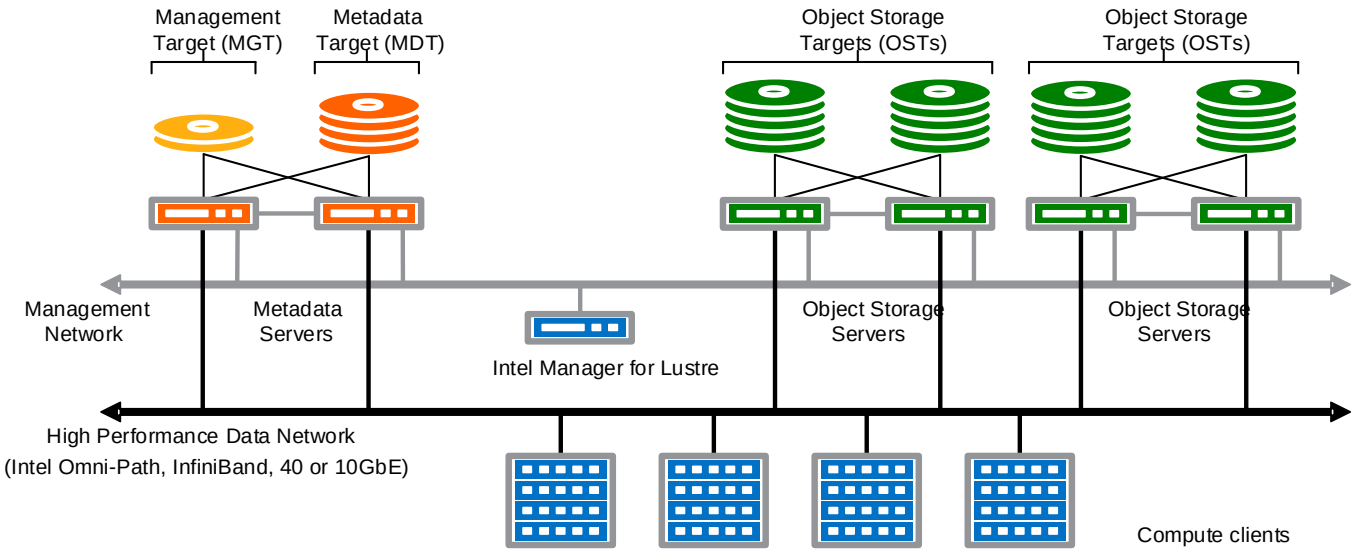
3.4 Scalability

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 R730 server provides performance and density. The solution provides 960TB of raw storage for each OSS pair and leverage Intel Omni-Path technology for very high-speed, low-latency storage transactions. The PowerEdge R730 takes advantage of the PCIe Gen3 x16 interface for the Omni-Path interconnect, helping achieve higher network throughput per OSS.

A Lustre client, version 2.7.15.3, for the RHEL7.2 kernel with HFI Omnipath support is available for access to the Dell HPC Lustre Storage solution. For details, see the Dell HPC Lustre Storage Solution Configuration Guide.

Figure 8: 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 8. This will increase both the total network throughput as well as 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. The default information 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 Intel Omni-Path fabric. This is the network used by the clients to access data. The Intel Manager for Lustre (IML) GUI interface provides an option to configure either a single or multiple Lustre Network Identifiers (NID) on MDS and OSS servers to participate in the Lustre Network. For instance, you could configure your Intel Omni-

Path HFI interface using IPoIB (i.e. ifcfg-ib0) as well as your 10GbE Ethernet interface (i.e. eth0) on your OSS servers to both participate in the Lustre Network.

In the Intel Omni-Path network, fast transfer speeds with low latency can be achieved. LNet leverages the Performance Scaled Messaging (PSM) protocol for rapid data and metadata transfer to and from MDTs and OSTs to the clients. The OSS and MDS servers take advantage of the Intel Omni-Path fabric with single port Intel HFI adapters.

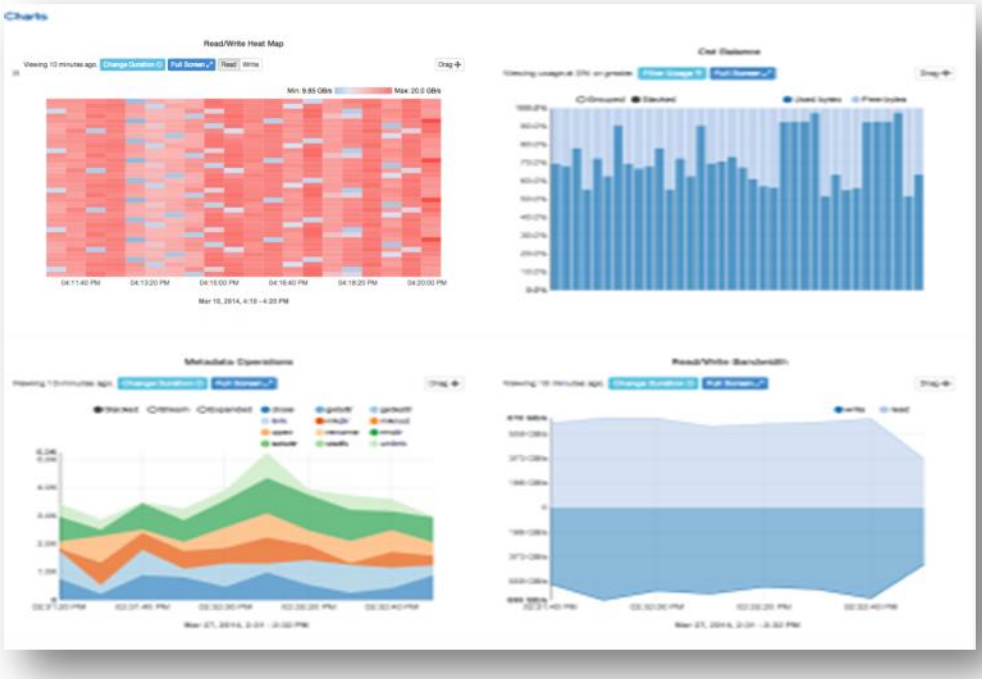
3.6 Managing the Dell Storage for HPC with Intel EE for Lustre

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 9 illustrates a few of the IML monitoring interfaces.

IML is a web-based management console to manage the solution (assuming all security requirements are satisfied). 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, and so forth.

Figure 9: Intel Manager for Lustre (IML) interface



4. Performance Evaluation and Analysis

The performance studies presented in this paper profile the capabilities of the Dell HPC Lustre Storage with Intel EE for Lustre software, in a 240-drive configuration. The configuration has 240 – 4TB disk drives (960TB 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 the Dell Zenith system, a Top500 class system based on Intel Scalable Systems Framework, 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. Intel Omni-Path was used for these studies since its high speed and low latency allows getting the maximum performance from the Dell HPC Lustre Storage solution, avoiding network bottlenecks.

We generally try to maintain a "standard and consistent" testing environment and methodology. There may be some areas where we purposely optimize server or storage configurations. We may have also take measures to limit caching effects. The goal is to better illustrate the impact to performance. This paper will detail the specifics of such configurations.

Table 1: Test Client Cluster Details

Component	Description
Compute Nodes:	Dell PowerEdge R630, 64 nodes
Node BIOS:	2.0.2
Processors:	Two Intel Xeon™ E5-2697 v4 @ 2.3GHz
Memory:	128GB DDR4 2400MHz
Interconnect:	Intel Omni-Path HFI
Lustre:	Lustre 2.7.15.3
OS:	Red Hat Enterprise Linux 7.2 (3.10.0-327.el7.x86_64)
Intel HFI firmware and driver:	10.1.1.06

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 HPC Lustre Storage with Intel EE for Lustre. We selected two benchmarks to accomplish our goal: IOzone and MDtest.

We used N-to-N load to test, where every thread of the benchmark (N clients) writes to a different file (N files) on the storage system. IOzone can be configured to use the N-to-N file-access method. For this study, we use IOzone for N-to-N access method workloads. See Appendix A for examples of the commands used to run these benchmarks.

Each set of tests was executed on a range of clients to test the scalability of the solution. The number of simultaneous physical clients involved in each test varied from a single client to 64 clients. The number of threads per node corresponds to the number of physical compute nodes, up to 64. The total

number of threads above 64 were simulated by increasing the number of threads per client across all clients. For instance, for 128 threads, each of the 64 clients ran two threads.

The test environment for the solution has a single *MDS* pair and a single *OSS* pair with a total of 960TB of raw disk space. The OSS pair contains two PowerEdge R730s, each with 256GB of memory, four 12Gbps SAS controllers and a single Intel Omni-Path HFI adapter. Consult the Dell HPC Lustre Storage Configuration Guide for details of cabling and expansion card locations. The MDS has identical configurations with 256GB of memory, an Intel Omni-Path HFI adapter and dual 12Gbps SAS controllers.

The Omni-Path fabric is comprised of a large port count Omni-Path fabric core switch for the client cluster, which the Lustre servers were also directly connected.

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

Table 2: Dell HPC Lustre Storage solution configuration

Configuration Size	960TB RAW
Lustre Server Version	2.7.15.3
Intel EE for Lustre Version	V3.0
OSS Nodes	2 x PowerEdge R730 Servers
OSS Memory	256GB DDR4 2133MT/s
OSS Processors	2 x Intel Xeon™ E5-2630V4 @ 2.20GHz 10 cores
OSS Server BIOS	2.0.2
OSS Storage Array	4 x PowerVault MD3460
Drives in OSS Storage Arrays	240 3.5" 4 TB 7.2K RPM NL SAS
OSS SAS Controllers	4 x SAS 12Gbps HBA LSI 9300-8e
MDS Nodes	2 x PowerEdge R730 Servers
MDS Memory	256GB DDR4 2133MT/s
MDS Processors	2 x Intel Xeon™ E5-2630V4 @ 2.20GHz 10 cores
MDS Server BIOS	2.0.2
MDS Storage Array	1 x PowerVault MD3420
Drives in MDS Storage Array	24 - 2.5" 300GB NL SAS
MDS SAS Controller	2 x SAS 12Gbps HBA LSI 9300-8e
Data network – Intel Omni-Path	
OSS, MDS Servers	Intel Omni-Path HFI adapter
Compute Nodes	Intel Omni-Path HFI adapter

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 was 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, a "*sync*" was performed on all the active servers (OSS and MDS) before each test and the kernel was 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 3.444. The throughput results presented in Figure 10 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 had an aggregate sample size of 2TB divided equally among the number of threads within that test. The block size for IOzone was set to 1MB to match the 1MB 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. The 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 10: Sequential Reads / Writes Dell HPC Lustre Storage with Intel Omni-Path

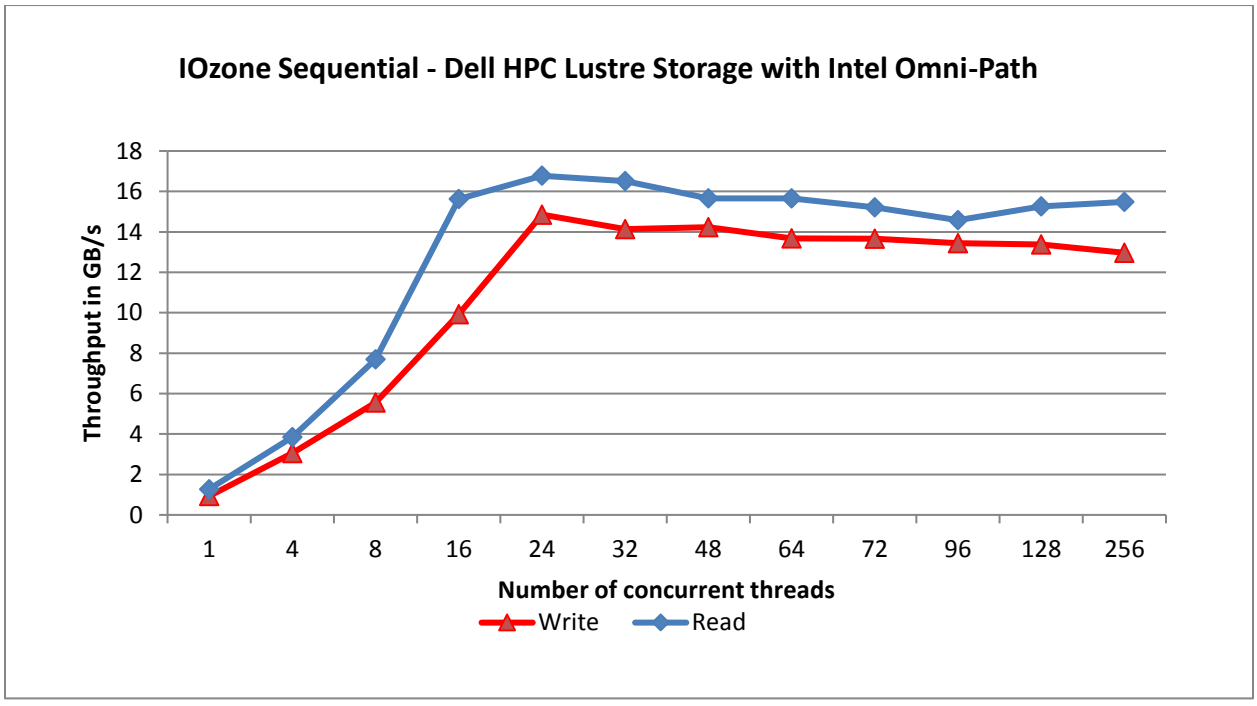


Figure 10 shows the sequential performance of the 960TB test configuration. With the test bed used, write performance peaks slightly less than 14.9GB/sec while read performance peaks near 16.8GB/sec.

We found that single client performance were consistent at 1GB/s to 1.3GB/s for writes and reads respectively. The write and read performance rise sharply as we increase the number of process threads up to 24 where we see level out to 256 with occasional dips. 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.

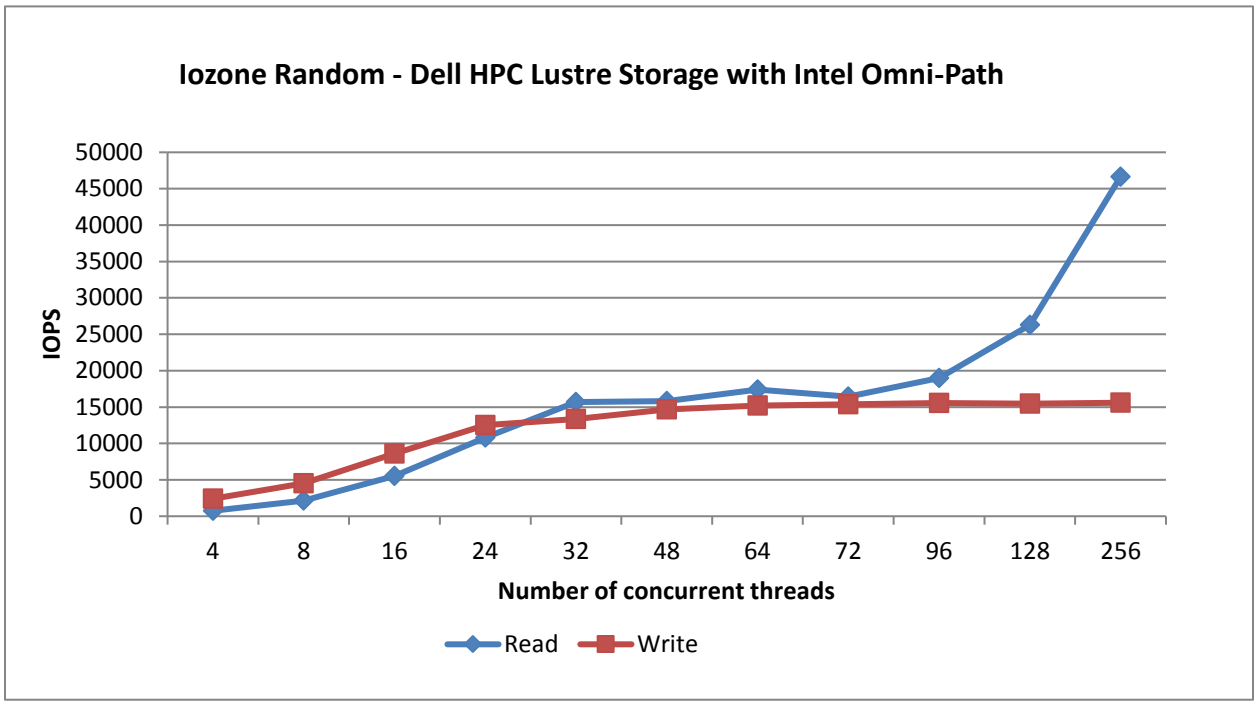
There are various OS, Intel Omni-Path IFS and Lustre level tuning parameters that can be used to optimize the Lustre storage servers for specific workloads. We cover the details of the tuning parameters that were configured on the test system below.

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 12 shows that the random writes peak at little shy of 15.6K IOPS and leveling after 48 threads, while random reads show a steady incline as threads increase. The IOPs of random reads increase rapidly from 4 to 32 threads and again at 72 to 256 threads. As the writes require a file lock per OST accessed, saturation is expected. Reads take advantage of Lustre's ability to grant overlapping read extent locks for part or all of a file.

Figure 12: N-to-N Random reads and writes



4.3 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 the Intel MPI distribution. 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. However, due to time constraints in testing, we only performed passes of files 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 were obtained with this Lustre topology.

Also during the preliminary metadata testing, we concluded 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 3 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 3: 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 13: File Metadata Operations

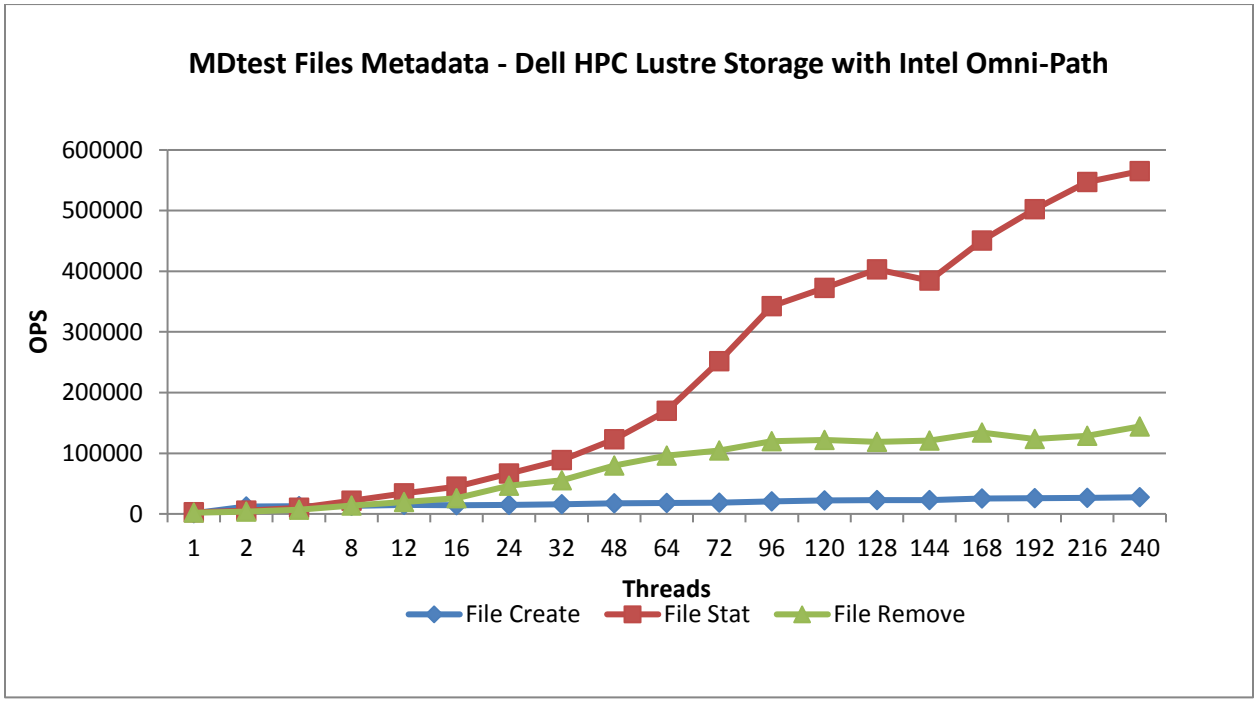


Figure 13 illustrates the file metadata results using MDtest.As shown in this graph, file create metadata operations start with less than 1100 OPS at 1 thread and scale to over 27K OPS with 240 concurrent threads. This may be due to the Lustre locks needed on the MDT, but also those on the OSTs, since 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 500K OPS with 240 concurrent threads. The increase in performance could be due to improvements made in Lustre version 2.7 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 has an advantage over the create operation when increasing in total threads, starting with over 1.9K OPS at 1 thread and scaling to over 144K OPS with increasing concurrent threads.

5. Performance Tuning Parameters

There are a number of performance tuning parameters that can be configured for an optimal system depending on intended workload patterns to be used. This section will detail the tuning parameters that we configured on the Lustre testbed system in the Dell HPC Engineering Innovations lab.

5.1 Lustre specific tunings

```
[root@node057 ~]# lctl set_param osc.*.max_pages_per_rpc=1024
```

The **max_pages_per_rpc** parameter is a tunable that sets the maximum number of pages that will undergo I/O in a single RPC to that OST.

[root@node057 ~]# lctl set_param osc.*.max_rpcs_in_flight=64
The **max_rpcs_in_flight** is a tunable that sets the maximum number of concurrent RPCs in flight to the OST. This parameter in majority of cases will help with small file IO patterns.

[root@node057 ~]# lctl set_param llite.*.max_read_ahead_mb=1024
The **max_read_ahead_mb** is a tunable that sets the maximum amount of data readahead on a file. These are read ahead in an RPC sized chunk.

[root@node057 ~]# lctl set_param osc.*.max_dirty_mb=1024
The **max_dirty_mb** is a tunable that sets how much dirty cache per OST can be written and queued. This generally benefits large memory IO workloads.

[root@node057 ~]# lctl set_param osc.*.checksums=0
This tunable will **disable checksums** and overhead associated.

[root@node057 ~]# lctl set_param llite.*.max_cached_mb=40960
The **max_cached_mb** tunable is the maximum amount of inactive dta that will be cached by the client.

[root@node057 ~]# lctl set_param llite.*.max_read_ahead_per_file_mb=1024
The **max_read_ahead_per_file_mb** is a tunable that sets maximum read ahead for each file.

[root@node057 ~]# lctl set_param debug=0
This tuning parameter will **disable debugging** completely and free the associated overhead.

[root@node057 ~]# lctl set_param subsystem_debug=0
The **subsystem_debug** parameter will control the debugging logs3 for subsystems.

5.2 Intel OPA specific tunings

```
[root@ieel3-oss1 ~]# cat /etc/modprobe.d/hfi1.conf
options hfi1 sge_copy_mode=2 wss_threshold=70 krcvqs=20
```

The **sge_copy_mode** tuning parameter introduces Adaptive Cache Memcpy feature that could improve throughput for high concurrency patterns such as outstanding requests in flight.

The **wss_threshold** parameter is the working set size threshold percentage that works with sge_copy_mode.

The **krcvqs** parameter is a tunable (kernel receive queues) that can improve parallel file system scalability.

5.3 Misc. tunings

NOTE: Setup of the Intel Omni-Path HFI interconnect will utilize IPoIB and the ifcfg-ib0 configuration file. Consult the Dell HPC Lustre Solution Configuration Guide for details.

Verify that the MTU setting in the IPoIB interface configuration file is set at 65520.

```
[root@ieel3-oss1 ~]# cat /etc/sysconfig/network-scripts/ifcfg-ib0
DEVICE=ib0
BOOTPROTO=static
…
ONBOOT=yes
NM_CONTROLLED=no
CONNECTED_MODE=yes
MTU=65520
[root@ieel3-oss1 ~]#
```

5. Conclusions

There is a well-known requirement for scalable, high-performance clustered file system solutions. The Dell HPC Lustre Storage 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 server platform, 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 14.9GB/s of write and 16.8GB/s of read throughput in a packaged component, is consistent with the needs of the high performance computing environments. The Dell HPC Lustre Storage solution is also capable of scaling in throughput as easily as it scales in capacity.

The continued use of generally available, industry-standard benchmark tools like IOzone 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 Solution with the requirements of many applications or group of applications.

The Dell HPC Lustre Storage 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.

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
```

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 HPC Lustre Solution 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 HPC Lustre Storage solution with Intel Omni-Path

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

Lustre Home Pages
http://www.intel.com/content/www/us/en/software/intel-solutions-for-lustre-software.html?cid=sem43700011015176072&intel_term=intel+lustre&gclid=COOMvL3w384CFQgOaQodbqsCeQ&gclsrc=aw.ds
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

Intel HPDD Wiki
https://wiki.hpdd.intel.com/display/PUB/HPDD+Wiki+Front+Page
LSI 12Gb/s SAS HBA
http://www.lsi.com/downloads/Public/Host%20Bus%20Adapters/LSI_PB_SAS9300_HBA_Family.pdf