WHITE PAPER
Intel® High Performance Data Division

Lustre* — The High Performance File System

Getting the Most from Your Data
Intel® High Performance Data Division

Abstract

Around the world and across all industries, high-performance computing (HPC) is being used to solve today's most important and demanding problems. More than ever, storage solutions that deliver high sustained throughput are vital for powering HPC and "Big Data" workloads. As storage challenges grow rapidly and unlocking the value within storage becomes even more important, today's high performance storage solutions require the optimal solutions designed for today's demanding needs. **Lustre***[1] is a performance distributed file system. From its roots in HPC, Lustre is expanding to meet the needs of data storage in the government, public, and private sectors. The Lustre file system provides world class data storage performance and scalability. It is open source, and capable of working on a broad set of commodity based hardware. The Lustre file system can lead you into the future with exabyte storage capabilities. When used with commodity-based hardware, it allows users to readily access large datasets, improve data management and storage utilization, while minimizing costs and data loss risks. The inherent performance characteristics of Lustre allow it to collect, sort and store data under one global namespace. As a result, the Lustre file system provides an affordable, high performance computing solution for data. Storage costs will be further reduced with the Hierarchical Storage Management (HSM) capabilities as an upcoming feature of Lustre.

Lustre has the performance, scalability and flexibility to meet your needs at an affordable price point. To further complement in-house personnel, Intel offers Intel® Enterprise Edition for Lustre* Software (IEEL), a commercial support capability, to speed up implementation and problem resolution. The improved manageability provided by Intel® Manager for Lustre* and the Intel® Hadoop Adaptor for Lustre*, also in Intel® Enterprise Edition for Lustre* Software, allows Hadoop users seamless access to data on Lustre file systems.

Lustre Overview

Introduction

Lustre is a well-respected file system from the world of high performance computing (HPC) and is an exciting answer for Big Data computation. From the largest U.S. Government laboratories with the world's fastest supercomputers, Lustre became the file system of choice to answer the need for performance. For applications running on HPC and commodity hardware computing systems, the file system is an essential part of the input/output (I/O) software stack and has a significant impact on performance. The Lustre file system is one of the leading parallel file system options for managing and storing large amounts of data for Linux.* The Lustre community is developing a number of capabilities needed for current petascale (1015 Bytes) users and for future exascale (1018 Bytes) users.

The Lustre 2.4 release improves performance with multiple metadata server (MDS) capability and initial work on an HSM capability. In Lustre 2.4, one can

Table of Contents

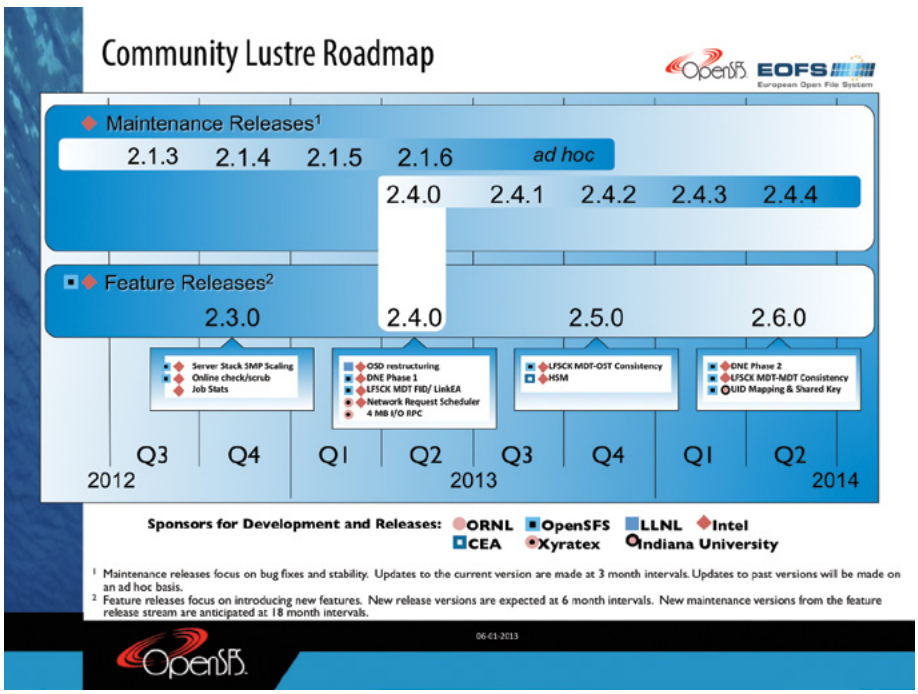
now select ZFS as a back-end filesystem option that Lustre uses for data and metadata storage.

The Lustre 2.5 release planned features complete the HSM capability as well as other features including: asynchronous remote updates, file rename and migration, and Lustre file system check. Additionally, the 2.5 release plans to provide copy tools initially for High Performance Storage System (HPSS) and POSIX archives, a client extended attribute cache and work on a Lustre client to be included in Linux.

History and Maturity

Lustre came from Department of Energy laboratories and is a premier high performance scalable network-attached storage (NAS) file system with the ability to manage many large files. Dr. Peter Braam started Lustre as a research project in 1999. He founded his own company, Cluster File Systems, and released Lustre 1.0 in 2003. In 2007, Sun Microsystems* acquired Cluster File Systems. In 2010, Oracle Corporation* acquired Sun and began to manage and release Lustre, but in late 2010 they announced they would cease Lustre 2.x development and place Lustre 1.8 into maintenance-only support. It is currently developed and maintained by the Lustre community under the guidance of the **Open Scalable File Systems (OpenSFS)**.[2] Important users and developers of Lustre also include organizations such as Lawrence Livermore National Laboratory (LLNL)*, Oak Ridge National Laboratory (ORNL)*, Commissariat à l'Énergie Atomique (CEA)*, Indiana University* and the National Center for Supercomputing Applications (NCSA.)* For years these supercomputer labs have used Lustre to manage and store data sets as large as those industries need today. As Lustre matured into an easy-to-manage file system with rich storage management capabilities, it has been embraced by markets looking for a single unified global

Figure 1
Current Lustre Roadmap from OpenSFS
(https://wiki.hpdd.intel.com/display/PUB/Community+Lustre+Roadmap)



solution to ever increasing collections of difficult to manage distributed data.

With the involvement of OpenSFS, new releases of Lustre are being developed as detailed in Figure 1, Current Lustre Roadmap from OpenSFS. OpenSFS is a vendor neutral, non-profit organization supported by its members, including Intel Corp.

Appendix A on page 9 contains details of Lustre releases in the recent past and in the near future.

A Strong Ecosystem

Lustre has a strong ecosystem across the computer industry. Intel, in particular, has a leadership position in the effort to expand Lustre into new markets outside of HPC and is providing strong support to the Lustre open source community through OpenSFS and the **Lustre User Group (LUG.)**[3] The counterpart to OpenSFS in Europe is the **European Open File System group (EOFS)**.[4] The OpenSFS and EOFS groups support vendor-neutral development and promotion of Lustre. In addition to these organizations and corporations supporting Lustre, key storage vendors like Hitachi Data Systems*, NetApp*, and EMC* use Lustre as the software foundation for their scalable solutions.

Lustre's Key Attributes

Below highlights some key capabilities:

- **Performance** Object storage target (OST) striping allows performance to increase with increased capacity, in one case achieving 1.2 TB/sec sustained write and over 2.0 TB/sec sustained read performance.[5]
- **Scalability** Testing[6] shows that Lustre can handle over 26,000 clients and over 500 PB of data.[7]
- **Manageability** New manageability features with the release of the Intel Manager for Lustre Tool.
- **Additional details of Intel® Manager for Lustre*** are presented in the following section.
- **Multi-Vendor and Open Source Lustre** requires no vendor lock-ins, one license agreement, and offers an inclusive and supportive user group for development.
- **Flexibility** Lustre is a flexible file system, able to handle different types of data and tasks on a variety of block storage hardware devices from an array of vendors.
- **Reliability** Lustre enables persistent state recovery and resiliency, includes a tool for disaster recovery, and can resynchronize the storage cluster without lengthy file system checking.
- **Availability** Lustre provides availability with redundant access to data. Software packages enable a complete Lustre failover solution.
- **Serviceability** Lustre allows for failing nodes to be taken offline and repaired while maintaining access to all the data in the file system.
- **Support** Intel® Enterprise Edition for Lustre* offers un-paralleled service level agreements for your 24/7/365 business needs.

Performance

Lustre's data striping from concurrently running nodes is a key performance enhancing factor. Good parallel performance is obtained when multiple threads write to individual files, or when accessing a single file striped mode access is used. Performance often reaches over 90 percent of the available I/O bandwidth. The sustained performance of the Lustre file systems at National Center for Supercomputing Association (NCSA) has exceeded 1 (TB) per second sustained performance, and on the Fujitsu K system[8] over 2 (TB) of sustained performance.[9,10] Striping makes it possible to increase the bandwidth available when accessing the file since the bandwidth from several up to thousands of object storage servers (OSS) can be combined. Striping also can allow the cache from OSSs to be aggregated and it provides more available disk space for storing the file since the striped file can be larger than an object storage target.

Scalability

Lustre has been shown to support over 26,000 clients and hundreds of Object Storage Servers (OST) with their associated disk storage, allowing it to scale up to petabytes (PB). Currently available versions of Lustre support maximum file sizes of 32 PB and a maximum file system size of 512 PB. For example, **NCSA**[11] uses Lustre to provide more than 25 PB of usable online storage while providing sustained aggregate performance of over 1 TB/second sustained performance. Lawrence Livermore National Laboratory (LLNL)[12] has 55 PB of Lustre storage and up to 1.4 TB/sec sustained performance.

The Lustre file system achieves great scalability and performance by separating metadata operations from data operations. This results in recoverability from failure conditions by providing the advantages of both journaling and distributed file systems. As long as the metadata device is appropriately sized when creating a Lustre file system, there is no need to take the file system offline to resize the metadata device. In the Lustre 2.4 release and beyond, the metadata service will have its own cluster, which may consist of dozens of nodes.

The following are some optimal scalability and performance results that are from the Intel Corp and can be found at:

http://www.whamcloud.com/lustre/support/new-to-lustre/

- File I/O as a percent of raw bandwidth: >90%
- Achieved single object storage server I/O: >6 GB/sec
- Achieved single client I/O: >3 GB/sec
- Achieved aggregate sustained I/O: (1.2 TB/sec write and 2 TB/sec read)
- Metadata transaction rate: 60,000 ops/sec
- Maximum file size / maximum file system size: 32 PB / > 512 PB
- Maximum number of clients: > 50,000[13, 14]

Manageability

The perception that you need a team of computer scientists to manage Lustre is outdated. Using the graphical or command line interface, the new Intel® Manager for Lustre* – included within the Intel® Enterprise Edition for Lustre* software product, is designed to allow users to experience the performance and scalability benefits of Lustre software faster and easier than ever before.

Intel® Manager for Lustre* brings together information about your Lustre configuration and provides a unified, consistent view of what's going on inside the storage system, while also simplifying the installation, configuration, monitoring and management of Lustre. Its key features include browser-based administration, real-time system monitoring, advanced troubleshooting tools and open interfaces; these features lower management complexity and costs and help users and enterprises of all sizes exploit the performance and scalability of Lustre. Additional details of Intel® Manager for Lustre* are presented in the section below.

Multi-Vendor and Open Source

Lustre is a vendor neutral scalable file system. While other comparable file systems require a customer to commit with a specific vendor and system, Lustre does not require commitment to one vendor. Lustre can work on different kinds of storage and appliances and on a variety of commodity hardware. This allows customers to select hardware and vendors based on their own needs and priorities rather than those of the file system.

Lustre's vendor neutrality results from its open source status. As mentioned earlier, the Lustre community provides maintenance and development under the guidance of OpenSFS. While other file systems require multiple agreements, Lustre has only a single **GNU GPL license**[15] to download and use, allowing the customer to use the software with minimal restrictions. While additional agreements may be necessary for vendor support, the customer can acquire them based solely on specific needs and priorities. Another benefit of the open source nature of Lustre is that groups are constantly working to update and upgrade it with new features and abilities to meet the collective needs of customers. It allows for easier rolling upgrades and no changes are needed in the agreement or necessary hardware when upgrading from one version to the next. This is not the case with other file systems that may require hardware upgrades to use the new software.

Flexibility

Lustre offers remarkable flexibility, and can manage data storage with files of varying sizes. While specialized for large files, recent improvements such as the ability to use multiple metadata servers and HSM capabilities allow Lustre to work with loads that include large and small files. In addition, Lustre has the ability to work with, and improve, other data management and analysis tools such as Hadoop*, which is discussed below. These abilities make Lustre ideal for handling a variety of data loads and situations. Lustre has demonstrated its ability to handle large data and simulations in the HPC world at national laboratories and universities. Recent studies have looked at Lustre's ability to manage incoming data for high definition (HD) video capture applicable to security cameras used in surveillance discussed in greater detail below.

Another example of the flexibility of Lustre is in the use of storage appliances. Many enterprises have a strong preference for integrated, pre-configured and validated storage solutions. Intel® Enterprise Edition for Lustre* Software, coupled with hardware storage platforms from an array of vendors, creates a simple to manage and efficient storage appliance solution. Aeon*, Bull*, Dell*, HDS*, NetApp* and many others have Lustre solutions available with support from Intel Corp. However, these are not the only vendors offering Lustre appliances and many more are in development. Over 35 partners currently offer Intel backed Lustre solutions and support.

Reliability

The Object Storage Servers (OSSs) will take over the task of modifying objects in storage so that client systems do not have to, ensuring scalability for large-scale clusters as well as improved security and reliability. In contrast, shared block-based file systems must allow all clients in the file system to have direct access to the underlying storage. This allows clients to directly modify blocks of storage without any safeguards, meaning that misbehaving or defective clients create a risk to the file system by potentially corrupting blocks of data in storage. Lustre technology utilizes the underlying Linux journaling file systems to enable persistent state recovery, as well as resiliency and recoverability from failed OSTs.

The Lustre API includes a file system checking tool ***(lfsck)*** for disaster recovery, but journaling and failure recovery protocols typically resynchronize the cluster without the need for lengthy file system checks. These file servers can further augment reliability mechanisms, using techniques such as RAID, to protect the content on the individual servers.

Availability

Lustre provides built-in availability via failover mechanisms at the file system level with the OSSs having redundant access to the OSTs. OSSs are typically deployed in an active/active configuration that provides redundancy in case of a failure. Metadata servers are set up for failover in an active/passive pair configuration. Lustre can be used with several high-availability (HA) software packages including **Red Hat Cluster Manager*** or **Pacemaker*** to enable a complete Lustre failover solution. It is assumed that two Lustre server nodes share a number of Lustre targets. Each node provides a number of targets and, in case of a failure; the active or non-failed

node takes over the targets of the failed nodes and makes them available to the Lustre clients. Lustre failover requires power control and management capability to verify that a failed node is shut down before I/O is directed to the failover node. IML can simplify the failover configuration; in fact, IML can install and configure Pacemaker for MDS and OSS.

Serviceability

Lustre's serviceability is closely tied to its availability. The ability to take failing nodes offline and repair them while maintaining access to all the data in the Lustre file system is a key consideration. Failing disks can be repaired while allowing continued access to the other data in the file system; only the data on the failed disk is inaccessible unless it was previously replicated. These capabilities allow users to continue their work with a minimum of outage.

Support

The Intel® Enterprise Edition for Lustre* Software (IEEL) product offering provides you with the option of utilizing Intel's expertise to help with your Lustre implementation and problem resolution. This addresses one of the perceived shortcomings of being on your own when it comes to Lustre support. Intel® Enterprise Edition for Lustre* Software provides a superset of management tools, certification testing and expanded support options for Lustre. With Intel® Enterprise Edition for Lustre* Software, Intel commits to worldwide enterprise grade Service Level Agreements (SLAs) including 24x7 coverage for critical Severity One problems.

Lustre Architecture Overview

Lustre runs on most of the large Linux* clusters in the world and is used by many HPC customers. Lustre is supported on servers running Red Hat Enterprise Linux 6.3* or CentOS 6.3*; it provides client support for SUSE 11* and Linux 2.6* kernels with the underlying block allocation based on the *ext4* file system.

Lustre 2.4.0 was released in May of 2013. Table 1 shows the supported server and client operating systems in the Lustre 2.4.0 support matrix.

Server Support	Client Support
RHEL 6.4	RHEL 6.4
CentOS 6.4	CentOS 6.4
	SLES 11 SP2 (SUSE)
	FC/18 (Fedora)

Table 1 - Lustre 2.4.0 Test Matrix

Lustre uses object-based storage in which each file or directory can be thought of as an object with attributes. Attributes can be assigned a value such as file type, file location, number of data stripes, ownership and permissions. An OSS allows you to specify for each file where to store the blocks allocated to the file via the MDS and OSTs. Extending the storage attribute, you can specify how many targets to stripe onto and what level of redundancy is needed. It also allows creating information containers and adding attributes to data for information management. Lustre creates a separation of functionality between computing and storage for efficient use of resources and flexibility.

Several factors allow Lustre to provide a high performance, expandable, shared-file global namespace with manageable storage. First, Lustre's separation of the data and metadata is critical in high capacity file systems. The file system efficiently handles how and where the data is stored and who has access. By processing a file's metadata separate from the data, the file system provides a direct pipe to large data files while processing other requests uninterrupted. Second, by allowing the file system to handle backend storage, the user does not need to know exactly where data resides; he only needs to access data in a familiar fashion. Lustre builds upon RAID architecture created to increase I/O performance by placing the I/O requests into a stripe of data chunks to "Just a Bunch of Disks" (JBOD). This provided a parallel I/O path to the

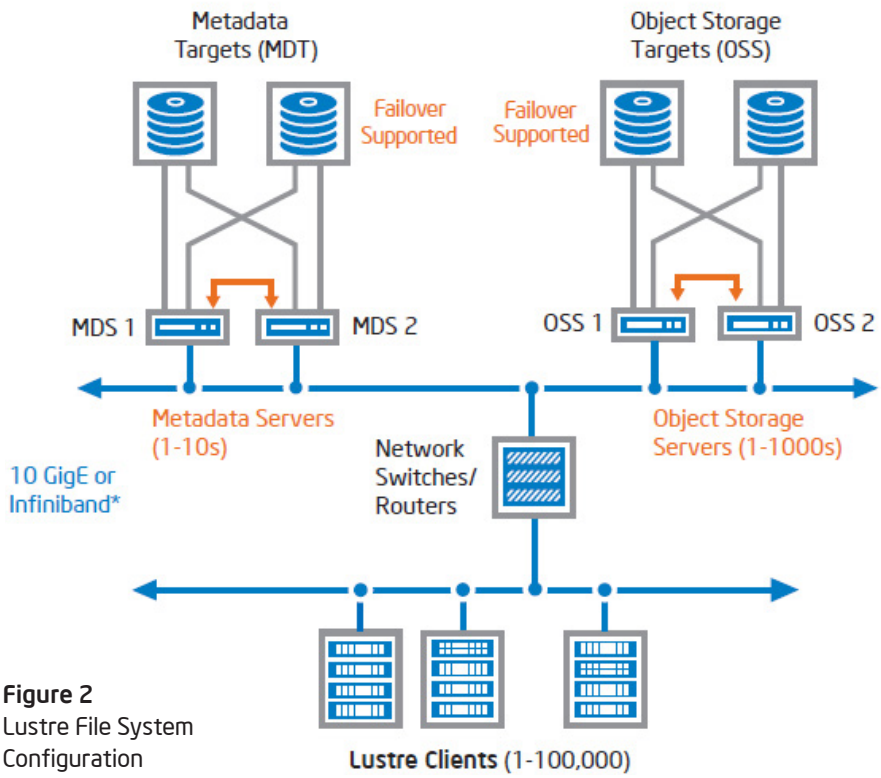


Figure 2
Lustre File System Configuration

disks for increased I/O bandwidth and performance over previous non-parallel file systems. Lustre builds on this by creating multiple object storage targets (OST) from individual RAID logical units. The Lustre file system stripes the data across the OSTs similar to a RAID stripe across disks to make increases in storage bandwidth possible. In addition, these OSTs are grouped onto OSSs to provide optimal front-end network bandwidth to back-end storage bandwidth. By increasing your storage capacity through the addition of OSS/OST units, you have the benefit of increasing the performance of your storage bandwidth at no cost to your management of the global namespace, since adding capacity adds bandwidth without increasing management needs.

A Lustre storage cluster consists of a Lustre MDS and Lustre OSSs, each of which has associated disk storage as shown in Figure 2 - Lustre File System Configuration. Client systems access these servers through supported network connections including TCP/IP (Gigabit Ethernet) and InfiniBand. Lustre file operations bypass the MDS completely and utilize the parallel data paths to all OSSs in the cluster. Like other UNIX and Linux file systems, Lustre files are represented by inodes, but in the Lustre file system, these inodes contain references to the objects storing the file data.

By default, the Lustre stores file metadata in a 4KB inode, resulting in each file requiring an additional 4KB on top of the actual data stored in the file. This should change in the Lustre 2.4 release to allow a file system to be formatted with inodes of 256 bytes in size, which can hold the striping of an inode over a small number of OSS nodes or a form of the new proposed wide striping layout, as well as other Lustre extended attributes. This change will allow the metadata for billions of files to be stored on an MDS with a large storage volume.

Intel® Manager for Lustre Description

As a key feature in Intel® Enterprise Edition for Lustre* software (See Figure 3 below), Intel® Manager for Lustre* software is a simple but powerful management solution that was developed specifically for the Lustre file system. It brings together information about your Lustre configuration and provides a unified, consistent view about what's going on inside the storage system and also simplifies the installation, configuration, monitoring and overall management of Lustre.

Around the world and across all industries, high-performance computing is being used to solve today's most important and demanding problems. More than ever, storage solutions that deliver high sustained throughput are vital for powering HPC and Big Data workloads. As storage challenges grow rapidly and unlocking the value within storage becomes even more important, today's high performance storage solutions have become too large and complex to be managed using software tools that were not designed for today's demanding needs.

Designed to make Lustre-based storage solutions easier to deploy and manage, Intel® Manager for Lustre* software maximizes return on storage investments. Designed to unleash the powerful benefits of distributed, parallel storage–from scalability to absolute performance–installing, configuring and managing Lustre has never been simpler.

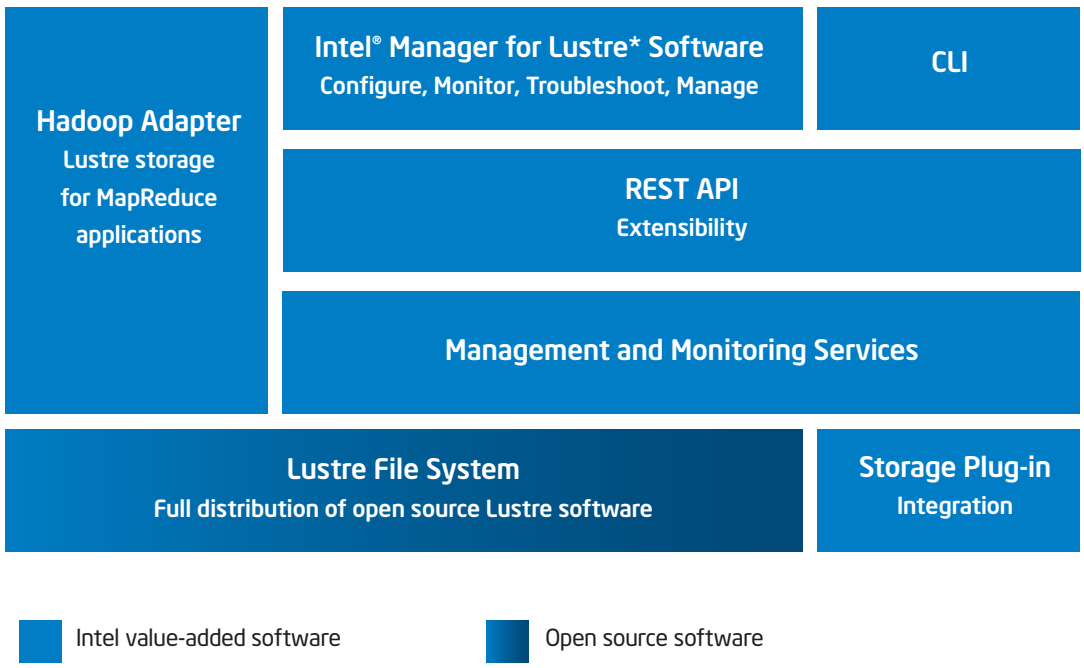


Figure 3
Intel® Enterprise Edition for Lustre* Software Stack

Key Features

- Browser-based administration provides simple but powerful graphical and scriptable command line interfaces to easily define and manage common administrative tasks.
- Real-time system monitoring allows you to monitor storage health and key performance indicators in real time, to give a view of high level system performance or individual components and generate historical and real-time charts and reports.
- Advanced troubleshooting tools give a consolidated view of cluster-wide storage log files, allow intelligent log-scanning for efficient problem isolation and analysis, and provide you with configurable event notifications.
- Its open interfaces provide a Representational State Transfer (REST) compliant API for easy integration with other systems and applications.

Intel® Manager for Lustre* Benefits

- Low total cost of ownership since the Intel® Manager for Lustre* software allows enterprises to scale their storage deployments horizontally, yet they can manage them with the same efficiency and precision as before. The benefits in this area include streamlined management, a shorter learning curve and lower operational expenses.
- High return on investment as the Intel® Manager for Lustre* software gives enterprises access to detailed system statistics, allowing them to analyze performance indicators in real-time. This more precise information enables a more efficient use of IT resources, better business and IT intelligence to lower business risk.
- For system administrators, the Intel® Manager for Lustre* software allows for faster installation and configuration and demystifies the process of configuring and managing Lustre deployments, since it provides a certified technology stack of software, hardware and networking as well as point-and-click tools for provisioning new Lustre nodes. It provides centralized management since Intel® Manager for Lustre* consolidates all Lustre information in a central, browser- accessible location, which enables administrators to have immediate access to real-time status information and proactively identify and correct faults before they become full-blown crises.
- For IT service managers and CIOs, Intel® Manager for Lustre* saves time and resources through browser-based management tools that reduce management complexity and shorten the learning curve, saving time and money on human resource acquisition, training, and support. It mitigates business risk by providing real-time performance and status indicators, giving IT managers much improved visibility into the health and viability of their Lustre deployment, which can help reduce and mitigate against the risk of technical failure and downtime. Intel® Manager for Lustre* also potentially improves business decisions by providing detailed metrics on Lustre file system usage, enabling IT executives to base investment decisions on accurate data and assisting in improving the decision-making and predictive ability of IT departments.

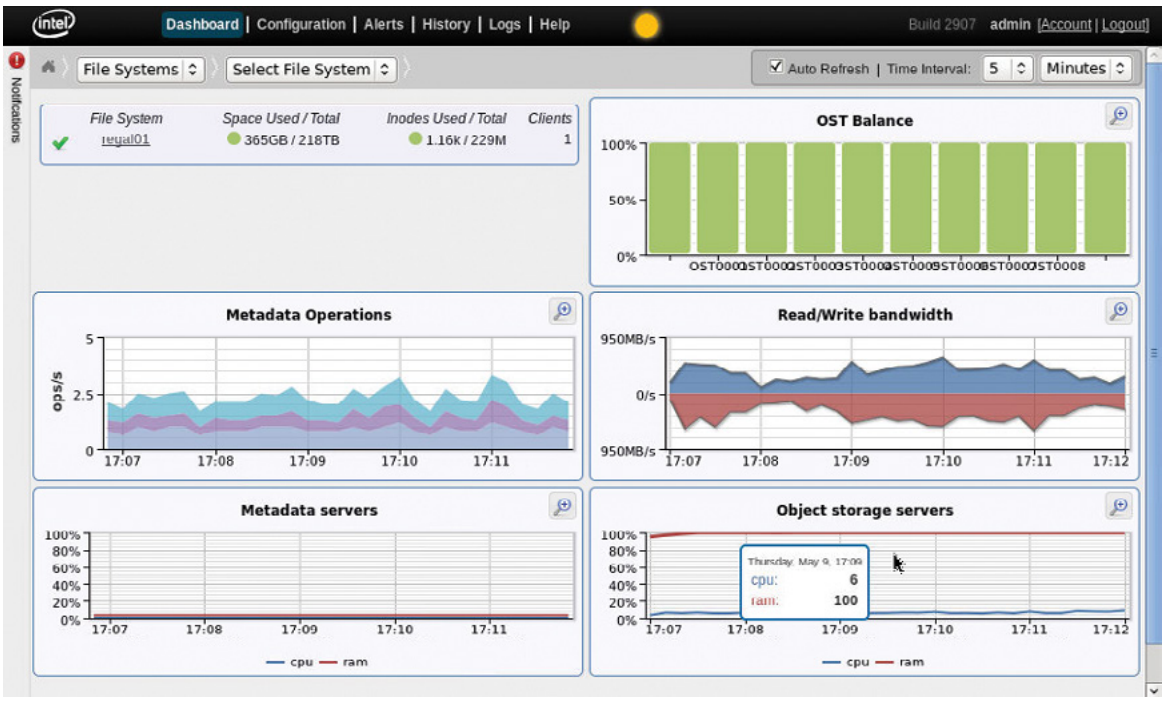


Figure 4
Intel® Manager for Lustre* GUI Showing File System

Purpose-built to solve today's most challenging storage issues, Intel® Manager for Lustre* software lowers management complexity and costs, helping users and enterprises of all sizes exploit the performance and scalability of Lustre, the most powerful storage software, to accelerate your critical applications and workflows. Created and supported by the Lustre experts at Intel, and backed by the brightest set of partners and storage integrators anywhere, Intel® Manager for Lustre* makes Lustre faster, smarter and more productive than it would be otherwise.

Use Cases

As a high performance file system, Lustre has been integral to HPC for many years with up to 400 of the top 500 supercomputers in the world using Linux and cluster architecture ideal for Lustre. Recently, Lustre's performance qualities have drawn the attention of other areas of information technology as the file system has been tested in combination with other offerings.

As data collections approach the exabyte levels and become commonplace, file systems must be able to work with software to analyze the data. Lustre demonstrated this ability by working with and improving the performance of **Hadoop***,[16, 17] which reduces data into key values for analysis. The two interact by using Hadoop's built-in LocalFileSystem class and adjusting Lustre so Hadoop can find it in its configuration files. Combining Lustre with Hadoop results in improvements in performance over the Hadoop Distributed File System (HDFS.) Lustre's throughput for reads and writes was double that of HDFS. In addition, overall runtime of Hadoop using Lustre was 15 percent faster than HDFS alone.[18] More testing is being conducted so that Lustre can further improve the performance of Hadoop, and Intel is working to incorporate upgrades coming with Lustre 2.4 into improving other tools in the Hadoop stack, such as Hive and HBase.

High Definition (HD) Video Capture

In addition to aiding software, Lustre is also demonstrating itself as a practical tool in industries outside the HPC environment. One specific real-world application is Lustre's use in HD video capture. In an internal study by Intel® in early 2013 on video capture,[19, 20] Lustre was able to take in data from over 500 HD cameras at 30 frames per second without dropping many frames. After tuning to improve performance, Lustre was able to handle data from 1,000 HD cameras at 30 frames per second while dropping less than 1 percent of frames per second.[21] This means that over 99 percent of all camera frames were successfully stored for future study. Analysis showed that this was because of Lustre's rich feature set that allowed for performance, scalability, load balancing, and management of petabyte-sized storage. Lustre was not only able to store the data from a large number of HD cameras, but testing suggested that Lustre is an ideal platform for analyzing this data, although testing was too limited for definitive findings. Areas of interest for these findings include both Federal, state and local government entities, like the Department of Homeland Security or police departments, and commercial entities, such as casinos and large malls, which are reliant on networks of cameras to provide surveillance over large areas.

Additional Use Cases include

- **Use of Lustre in a WAN environment**[22, 23]
- **Large Data**[24, 25]
- **Lustre Performance over the InfiniBand WAN**[26, 27]

Conclusion

From its roots in HPC, Lustre is expanding to meet the needs of data storage in the government, public, and private sectors. The Lustre file system provides world class data storage performance and scalability, it is open source, and it can work on commodity hardware. As a result, the Lustre file system offers a HPC solution for data storage at a low cost for the user.

Lustre is moving forward in the Lustre 2.4 release with capabilities to improve performance, such as multiple metadata servers, interfacing to an HSM system and using ZFS as an optional underlying file system to address the missing features and capabilities of *ldiskfs*, based on *ext4*. These capabilities will only increase with the Lustre 2.5 release.

There are a number of advantages to Lustre including:

- Excellent parallel performance can be obtained when multiple processes or threads write individual files or in striped mode, when accessing a single file concurrently
- Easy manageability since there are open source tools available as well as the Intel® Manager for Lustre* product
- Since it does not have server and client licensing costs as well as support costs, Lustre is less expensive than other file system options
- Appliances are now available in a variety of configuration options that have had thorough testing

Appendix A - Details of Recent and Future Lustre Releases

Lustre 2.3

Release 2.3.0 was made available on October 22, 2012 from the Intel High Performance Data Division (known at that time as Whamcloud Inc, prior to its acquisition by Intel Corp in 2012) and contained the following highlighted capabilities:

- Object Index (OI) scrub - *lfsck* rebuilds the OI file after MDT file-level backup/restore
- Improved SMP performance of shared directory
- Wide community involvement with code contributions from Bull, CEA, Cray, DDN, EMC, Fujitsu, LLNL, ORNL, TACC, Ultrascale, UVT, Intel High Performance Data Division and Xyratex

Details of the Lustre 2.3.0 release are available at:

Lustre 2.4

https://wiki.hpdd.intel.com/display/PUB/Lustre+2.4

Lustre 2.4 was released in spring 2013. This release contained the following features:

- Adding support for SLES11-SP2 clients
- Adding support for Fedora 18 clients

New Features for Lustre 2.4:

- Network request scheduler
- 4MB remote procedure call (RPC)
- File identifier (FID) on OST infrastructure
- Update Wireshark support for LNET and Lustre
- Client performance improvements
- Hierarchical storage management (HSM) layout lock
- FID-in-Dirent, linkEA (LFSCK Phase 1.5)
- Remote directories (DNE Phase I)
- Server stack
- Object storage device (OSD) - LLOG
- Layered object device / object storage proxy (LOD/OSP)
- OSD-MDT
- Quota enforcement
- Management server (MGS) over OSD
- Changelog
- ZFS OSD Utilities
- LNET networks hashing
- Ability to disable pinging
- FID on OST
- HSM: Manage dirty flag for HSM
- HSM: Data version per file
- HSM: Flags
- HSM: POSIX copy tool
- HSM: Layout swapping MDT part

Further details of the planned Lustre 2.4 release are available at:

http://wiki.whamcloud.com/display/PUB/Lustre+2.4+Scope+Statement

Of particular interest in the 2.4 release are the capabilities of having multiple metadata servers, initial work on interfacing to an HSM system and using ZFS as an optional underlying file system.

Lustre 2.5

Lustre Release 2.5 includes the following:

- Distributed namespace (DNE): Asynchronous remote updates, rename, migration
- Lustre file system check: MDT/OST consistency check or repair
- Data migration (parts included in 2.4 and 2.5): Uses layout lock and object data version from HSM; move file data between OSTs safely; files can be open and in use
- HSM: Copy tools initially for HPSS and POSIX archives; uses CEA Robin Hood for policy engine; infrastructure usable for migration, replication, etc.
- Client extended attribute cache: Fetch xattrs from MDS with other attributes; avoid RPC round-trip for each xattr access; avoid RPC round-trip for xattrs that do not exist; important for Samba/CIFS exporting performance and for SELinux and similar labeling systems
- Client update: Desire to include Lustre client in upstream Linux kernel for easier customer installations and to reduce or eliminate lag for new kernels; need to clean up ten years of legacy code; changes included incrementally in versions 2.3, 2.4 and 2.5
- File replication (parts to be included in versions 2.5, 2.6 and 2.7): Mirror files across multiple OSTs; phase 1 - out-of-band replication; phase 2 - synchronous replication; phase 3 - asynchronous replication

The presentation at **http://www.opensfs.org/wp-content/uploads/2012/12/Dilger_Lustre-2.5-Beyond-LUG-2013.pdf** provides details of features in the Lustre 2.5 release and in some of the later releases.

[1] http://wiki.lustre.org/index.php/Main_Page

[2] http://www.opensfs.org/

[3] http://wiki.lustre.org/index.php/

[4] http://www.eofs.eu/

[5] The IEEE 1541-2002 Standard defines the use of prefixes for binary multiples of units of measurement related to computing and includes GiB, TiB, PiB, etc. However, in this document the more common GB, TB, etc. are used for simplicity.

[6] http://www.whamcloud.com/lustre/support/new-to-lustre/ . These figures are based on figures published by WhamCloud in early 2013.

[7] Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. This data is for open source Lustre and is not exclusive to Intel® Enterprise Edition for Lustre*.

[8] http://www.opensfs.org/wp-content/uploads/2013/04/LUG2013-FJ-20130410-final-1.pdf page 18, results in presentation at Lustre User Group 2013, San Diego by Fujitsu on April 18, 2013

[9] Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase. This data is for open source Lustre and is not exclusive to Intel® Enterprise Edition for Lustre*.

[10] http://insidehpc.com/2013/03/05/terabyte-per-second-file-systems-come-to-big-fast-data/ , Findings are from NCSA and Cray work and were announced October 7, 2012.

[11] http://www.ncsa.illinois.edu/. Results and figures are from testing and operations conducted by NCSA and Cray in October 2012.

[12] http://zfsonlinux.org/docs/LUG12_ZFS_Lustre_for_Sequoia.pdf, results in presentation by LLNL at Lustre User Group 2012, April 23, 2012

[13] Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase. This data is for open source Lustre and is not exclusive to Intel® Enterprise Edition for Lustre*.

[14] All figures and findings are from tests and operations conducted by WhamCloud in early 2013.

[15] http://www.gnu.org/licenses/gpl.html

[16] Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. This data is for open source Lustre and is not exclusive to Intel® Enterprise Edition for Lustre*.

[17] http://hadoop.apache.org/ , Testing conducted by Omkar Kulkarni for Intel and a Lustre User Group Presentation: Hadoop MapReduce over Lustre*, April 16, 2013.

[18] Omkar Kulkarni, Lustre User Group Presentation: Hadoop MapReduce over Lustre*, April 16, 2013.

[19] Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. This data is for open source Lustre and is not exclusive to Intel® Enterprise Edition for Lustre*.

[20] Results have been simulated and are provided for informational purposes only. Results were derived using simulations run on an architecture simulator or model. Any difference in system hardware or software design or configuration may affect actual performance. This data is for open source Lustre and is not exclusive to Intel® Enterprise Edition for Lustre*.

[21] Internal study by Intel® to measure Lustre*'s effectiveness in handling, managing, and storing HD video data. Testing was conducted from February to April 2013.

[22] Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase. This data is for open source Lustre and is not exclusive to Intel® Enterprise Edition for Lustre*.

[23] http://wiki.lustre.org/images/3/3a/JamesHoffman.pdf , James Hoffman, Naval Research Lab and David McMillen, System Fabric Works. Testing conducted for the Large Data JCTD, April 17 2009.

[24] Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase. This data is for open source Lustre and is not exclusive to Intel® Enterprise Edition for Lustre*.

[25] http://wiki.lustre.org/images/b/b8/LUG2010_Filizetti_NRL.pdf, Jeremy Filizetti for NRL, April 16, 2010.

[26] Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase. This data is for open source Lustre and is not exclusive to Intel® Enterprise Edition for Lustre*.

[27] http://wiki.lustre.org/images/6/60/LUG2010_Filizetti_SMSi.pdf, Jeremy Filizetti for SMSi, April 16, 2010.

Notices and Disclaimers

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Some results have been simulated and are provided for informational purposes only. Results were derived using simulations run on an architecture simulator or model. Any difference in system hardware or software design or configuration may affect actual performance.

Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase.

Information and data in this document is regarding open source Lustre* and is not exclusive to Intel® Enterprise Edition for Lustre*.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel.

Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel.

Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804