Intel® Manager for Lustre*

User Guide

Software version 3.1.1
http://www.intel.com

Disclaimer and Legal Information

Contents

1 Introducing Intel® Manager for Lustre* Software

Enterprises and institutions of all sizes use high performance computing to solve today's most intense computing challenges. Just as compute clusters exploit parallel processors and development tools, storage solutions must be parallel to deliver the sustained performance at the large scales that today's applications require. The Lustre* file system is the ideal distributed, parallel file system for high performance computing.

Accordingly, as storage solutions continue to grow in complexity, powerful, yet easy-to-use software tools to install, configure, monitor, manage, and optimize Lustre-based solutions are essential. Intel® Manager for Lustre* software is purpose-built to simplify the deployment and management of Lustre-based solutions. Intel® Manager for Lustre* software reduces management complexity and costs, enabling storage superusers to exploit the performance and scalability of Lustre storage, and accelerate critical applications and work flows.

Intel® Manager for Lustre* software greatly simplifies the creation and management of Lustre file systems, using either the graphical user interface (GUI) or a command line interface (CLI). The GUI dashboard lets you monitor one or more distributed Lustre file systems. Real-time storage-monitoring lets you track Lustre file system usage, performance metrics, events, and errors at the Lustre level. Plug-ins provided by storage solution providers enable monitoring of hardware-level performance data, disk errors and faults, and other hardware-related information.

Intel® EE for Lustre* software, when integrated with Linux, aggregates a range of storage hardware into a single Lustre file system that is well-proven for delivering fast IO to applications across high-speed network fabrics such as InfiniBand* and Ethernet.

An existing Lustre file system that has been set up outside of Intel® Manager for Lustre* software can be monitored, but not managed by the manager. In this case, Lustre commands can be used to manage metadata or object storage servers in the Lustre file system.

1.1 Related Documentation

The following documents are pertinent to Intel® Enterprise Edition for Lustre* software. This list may not be current. Contact your Intel® support representative for the most current information.

- *Intel® Enterprise Edition for Lustre* Software Release Notes Version 3.1.n.n*
- *Intel® Enterprise Edition for Lustre* Software Installation Guide*
- *Lustre* Installation and Configuration using Intel® EE for Lustre* Software and OpenZFS*
- *Configuring LNet Routers for File Systems based on Intel* EE for Lustre* Software*

- *Configuring Snapshots for File Systems based on Intel® EE for Lustre* Software*
- *Hierarchical Storage Management Configuration Guide*
- *Installing Hadoop and the Hadoop Adapter for Intel® EE for Lustre* and the Job Scheduler Integration*
- *Creating an HBase Cluster and Integrating Hive on an Intel® EE for Lustre® File System*
- *Upgrading a Lustre file system to Intel® Enterprise Edition for Lustre* Software (Lustre only)*
- *Creating a Scalable File Service for Windows Networks using Intel® EE for Lustre* Software*
- *Intel® EE for Lustre* Hierarchical Storage Management Framework White Paper*
- *Architecting a High-Performance Storage System White Paper*

For more information beyond the documents listed above, see:
Intel® Solutions for Lustre* software - http://www.intel.com/content/www/us/en/software/intel-solutions-for-lustre-software.html

1.2 Overview of Intel® Enterprise Edition for Lustre* software

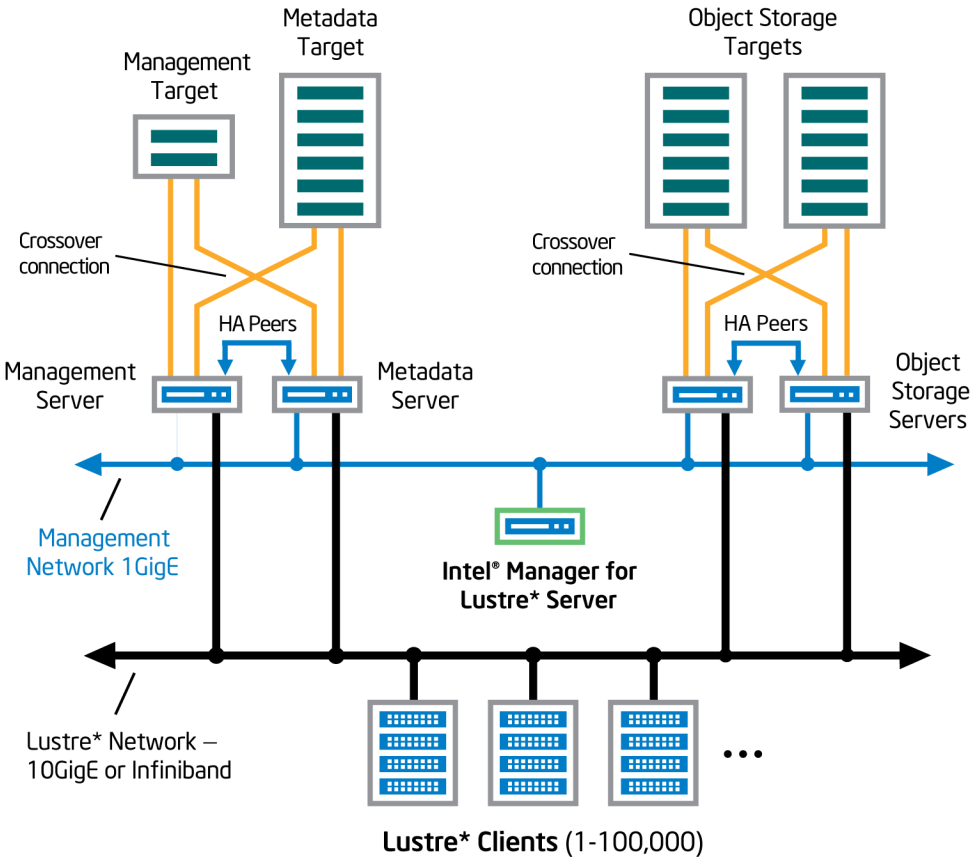
Intel® Enterprise Edition for Lustre* software is a global single-namespace file system architecture that allows parallel access by many clients to all the data in the file system, across many servers and storage devices. Designed to take advantage of the reliability features of enterprise-class storage hardware, Intel® EE for Lustre* software supports availability features such as redundant servers with storage failover. Metadata and data are stored on separate servers to allow each system to be optimized for the different workloads. The components of an Intel® EE for Lustre* software, file storage system include the following:

- Intel® Manager for Lustre* server: The server that hosts the Intel® Manager for Lustre* software and GUI, and is the server from which Lustre file systems are created, monitored, and managed. This server is connected to storage servers via the administrative LAN. *This is distinct from the management server*, which provides access to the management target.
- Management server(s) (MGS): Provide access to the management target. Paired, redundant management servers provide server failover (high availability) in the event of a server failure.
- Management target (MGT): The MGT stores configuration information for all the

Lustre file systems in a cluster and provides this information to other Lustre components. Each Lustre object storage target (OST) contacts the MGT to provide information, and Lustre clients contact the MGT to retrieve information. The MGT can be no larger than 10 gigabytes.

- Storage servers: Storage servers provide access to the management target, metadata target and the storage targets. Paired, redundant storage servers provide server failover (high availability) in the event of a server failure.
- Metadata target (MDT): The MDT stores metadata (such as file names, directories, permissions, and file layout) for attached storage, and makes this available to clients. Typically, each file system has one MDT, however Intel® EE for Lustre* software supports multiple MDTs.
- Object storage targets (OSTs) - User file data is stored in one or more objects that are located on separate OSTs in the file system. The number of objects per file is configurable by the user and can be tuned to optimize performance for a given workload.
- Lustre clients - Lustre clients are computational, visualization, or desktop nodes that are running Lustre client software, allowing them to mount the Lustre file system.

The servers on which the MGT, MDT, or OSTs are located can all be configured as high-availability (HA) servers, so that if a server for a target fails, a standby server can continue to make the target available.



1.3 Key Features

Following are key features provided by Intel® Enterprise Edition for Lustre* software and Intel® Manger for Lustre* software.

GUI-based creation and management of Lustre* file systems

The Intel® Manager for Lustre* software provides a powerful, yet easy-to-use GUI that enables rapid creation of Lustre file systems. The GUI supports easy configuration for high availability and expansion, and enables performance monitoring and management of multiple Lustre file systems. See Creating a new Lustre* file system.

Graphical charts display real-time performance metrics

Fully-configurable color charts display a variety of real-time performance metrics for single or multiple file systems, down to individual servers and targets, and reveal metrics such as read/write heat maps, OST balance, file system capacity, metadata operations, read/write operations, job statistics, and various resource usage parameters, among others. See View charts on the Dashboard.

Auto-configured high-availability clustering for server pairs

Pacemaker and Corosync are configured automatically when the system design follows configuration guidance. This removes the need for manually installing HA configuration

files on storage servers, and simplifies high-availability configuration. See High-availability file system support.

PDU configuration and server outlet assignments support automatic failover

The PDU window lets you configure and manager power distribution units. At this window you can add a detected PDU and assign specific PDU outlets to specific servers. When you associate PDU failover outlets with servers using this tool, STONITH is automatically configured.

IPMI and BMC Configuration

An alternative to PDU configuration, support for Intelligent Platform Management Interface and baseboard management controllers support server monitoring, high-availability configuration, and failover.

Support for Intel® Xeon Phi™ Coprocessor Clients

Intel® EE for Lustre* client software can be installed and configured to run on Intel® Xeon Phi™ Coprocessor clients. This means that the Intel® Xeon Phi™ Coprocessor clients can directly mount Lustre.

Hierarchical Storage Management

Intel® EE for Lustre* software includes support for hierarchical storage management. HSM provides a way to free up file system storage capacity by archiving the less-frequently accessed files into secondary, archival storage. You can configure the HSM framework directly from the Intel® Manager for Lustre* GUI.

Distributed Name Space

Distributed Namespace (DNE) allows the Lustre metadata to be distributed across multiple servers. DNE1 has been incorporated into Intel® EE for Lustre* software, and this featured is supported at the Intel® Manager for Lustre* GUI.

Robinhood Policy Engine

The Robinhood policy engine has been incorporated into Lustre and is included with Intel® EE for Lustre*. Intel® Manager for Lustre* software performs the provisioning of the Robinhood agent server, which is performed via the manager GUI. Robinhood can be used with the HSM capabilities described above to automate HSM archiving and report generation.

Apache Hadoop* adapter software

Intel® EE for Lustre* software is supported by the Apache Hadoop* adapter software, however the adapter software is a separate download. This Hadoop adapter for Lustre is compatible with the Apache Hadoop software, versions 2.3 and 2.5 as of this writing. Hadoop software allows users who run MapReduce jobs to bypass storing data in HDFS, and store the MapReduce output directly to Lustre instead. This allows the analytical

processes direct access to scientific output instead of transferring data from the compute cluster storage system to another file system. Optimizations have also been made to the shuffle step in MapReduce to take advantage of Lustre's high-speed network access to data. Many workloads will see an overall reduction in end-to-end processing time by using the Hadoop adapter with the Intel® EE for Lustre* software file system. For more information, see Installing Hadoop, the Hadoop Adapter for Intel® EE for Lustre* Software, and the Job Scheduler Integration.

Automated Provisioning of Custom Lustre Service Nodes

This feature allows users to create custom profiles for new Lustre client types and, based on a given profile, deploy and install custom code to provide new services. HSM copytool (above) is deployed in this way. Other services might include Samba file services, etc.

Simplified ISO-less installation and automated deployment mechanism streamlines overall installation

The installation strategy removes the need to manually install the software on each server. Intel® Manager for Lustre* software is quickly installed on the manager server, and from there, required packages are automatically deployed to all storage servers. Storage servers and the manager server can run the same standard operating system as the rest of your estate. Additional software built for CentOS or Red Hat will also work on servers managed by Intel® Manager for Lustre* software.

Note: The *manager server* is that server where the Intel® Manager for Lustre* software dashboard is installed.

Support for OpenZFS in Management Mode

Intel® EE for Lustre* software supports ZFS as a back-end file system replacement for ldiskfs. Intel® Manager for Lustre* software is able to configure and manage high-availability Lustre storage solutions, and Intel® EE for Lustre* software can discover and manage ZFS file systems. See Creating and Managing ZFS-based Lustre file systems.

Intel® EE for Lustre* Software ZFS Snapshots

The OpenZFS file system provides integrated support for snapshots, a data protection feature that enables an operator to checkpoint a file system volume. In Intel® EE for Lustre* software, as of version 3.0.0.0, Intel® has developed a mechanism in Lustre* that leverages ZFS to take a coordinated snapshot of an entire Lustre* file system, if all of the storage targets in the file system are formatted using ZFS.

HPC Job Scheduler integration with MapReduce

Intel® EE for Lustre* software works with the HPC job scheduler integration with MapReduce; however the job scheduler integration is a separate download. The HPC job scheduler integration supports Apache Hadoop. This adapter for job schedulers allows you to integrate common resource schedulers into your cluster. You have the choice of

installing the SLURM (Simple Linux Utility for Resource Management) job scheduler integration or the PBS (portable batch system) job scheduler integration. An integration guide is available: Installing Hadoop, the Hadoop Adapter for Intel® EE for Lustre* Software, and the Job Scheduler Integration.

Hadoop commonly uses Yarn to manage MapReduce jobs. However, virtually all HPC systems use a job scheduler, for example, SLURM, but having two job schedulers, e.g., SLURM and Yarn, in a single system can cause problems. The HPC Job Scheduler integration with MapReduce replaces YARN with an interface to the main resource manager for the system. This allows MapReduce applications to be run as normal HPC jobs.

Apache Hive compatibility

Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. Intel® has tested the Hadoop adapter for Lustre provided with Intel® EE for Lustre* software for compatibility with Apache Hive version 2.5.

Apache Hbase compatibility

HBase is a non-relational, distributed database modeled after Google's BigTable and written in Java*. Hbase runs on top of HDFS (Hadoop Distributed File System). Intel® has tested the Hadoop adapter for Lustre provided with Intel® EE for Lustre* software for compatibility with Apache Hbase version 2.5.

Lustre 2.7.x

This release of Intel® EE for Lustre* software is based on the Intel® Foundation Edition for Lustre* 2.7 release tree, representing a major update to the underlying Lustre* version for the Intel® Enterprise Edition for Lustre* software (as of version 3.0.0.0).

Online Lustre File System Consistency Checks (LFSCK)

LFSCK is an administrative tool that was first introduced in Lustre* software release 2.3 for checking and repairing attributes specific to a mounted Lustre* file system. LFSCK is similar in concept to an offline FSCK repair tool for a local file system, but LFSCK is implemented to run as part of the Lustre* file system while the file system is mounted and in use. LFSCK allows consistency checking and repair by the Lustre software without downtime, and can be run on the largest Lustre* file systems with negligible disruption to normal operations.

Distributed Namespace

Distributed Namespace (DNE) allows the Lustre metadata to be distributed across multiple metadata servers. Intel® EE for Lustre* software supports DNE1 (as of release 2.3.0.0), which supports the use of multiple MDTs. This enables the size of the Lustre namespace and metadata throughput to be scaled with the number of OSSs. This

featured is supported at the Intel® Manager for Lustre* GUI.

DNE II Striped Directories Support (Preview)

Striped directories support (Distributed Name Space, phase 2) is available in Intel® EE for Lustre* software, as of version 3.0, as a technology preview. Striped directories allow operators to shard directory entries across multiple metadata storage targets, providing both namespace and metadata performance scalability.

Single Client Metadata Concurrency

Also referred to as “multi-slot last_rcvd”, this update to the metadata communications interface between client and server allows multiple metadata RPCs to be in flight in parallel, per-client for both read and write transactions. Prior to this release, any client RPCs that modified file system metadata (for example, creates or unlinks), were sent serially to the server. With this update, this restriction is removed.

Differentiated Storage Services

Differentiated Storage Services (DSS) allows I/O data to be classified, sometimes referred to as “hinting". These hints pass seamlessly through Intel® EE Edition for Lustre* software, at which point data can be tiered and intelligently cached by the storage system. This enables a more efficient use of cache space, and decreases the likelihood of critical data being evicted when the cache fills. Intel® is working directly with storage and cache vendors to enable DSS hinting in Lustre appliances, and to provide optimized performance to Intel® EE Edition for Lustre* software deployments with a mix of SSD and traditional storage.

Support for Intel® Omni-Path Architecture

Intel® Omni-Path fabric support is available for Intel® EE for Lustre* software systems running RHEL 7.3. (Intel® OPA driver support requires RHEL 7.1 or newer, and so is not available for RHEL 6.x based systems.)

LNet Configuration

This feature assists in configuring LNet for a given server’s network interface by setting the LNet network ID for that port. This feature requires a single LNet. You can configure multiple LNets (i.e., with the use of routers), however in this release, additional LNets cannot be configured from the GUI.

Dynamic LNet Configuration

Dynamic LNet configuration (DLC) is a powerful extension of the LNet software to simplify system administration tasks for Lustre networking. DLC allows an operator to make changes to LNet (for example, network interfaces can be added and removed, or parameters changed,) without requiring that the kernel modules be removed and reloaded. Parameters can be altered while LNet is still running, meaning that tuning and optimization can be conducted while Lustre* is still running on the target node. Dynamic

LNet configuration also applies to LNet routers, so that routes can be added, removed and updated without affecting other Lustre network traffic.

Kerberos Network Authentication and Encryption

Kerberos provides a means for authentication and authorization of participants on a computer network, as well as providing secure communications through authentication. This functionality has been applied to Intel® EE for Lustre* software for the purposes of establishing trust between Lustre* servers and clients, and optionally, supporting encrypted network communications.

1.4 Management mode versus Monitor-only mode

What is Management Mode?

The Intel® Manager for Lustre* software lets you create and manage new HA Lustre file systems from its GUI. For each HA file system, the GUI and dashboard let you create, monitor, and manage all servers and their respective targets. The software lets you define failover servers to support HA. RAID-based fault tolerance for storage devices is implemented independent of Intel® Manager for Lustre* software.

To provide robust HA support, Intel® Manager for Lustre* software automatically configures Corosync and Pacemaker, and takes advantage of IPMI or PDUs to support server failover.

Note: Managed HA support requires that your entire storage system configuration and all interfaces be compliant with a pre-defined configuration. See the High Availability Configuration Specification in the *Intel® Enterprise Edition of Lustre, Partner Installation Guide* for detailed information.

Note: Management mode is supported in Intel® Enterprise Edition for Lustre* software, versions 1.0 and later. No claims of support are made for any versions of Lustre outside of that shipped with Intel® EE for Lustre* software.

What is Monitor-only Mode?

Monitor-only mode allows you to “discover” an existing Lustre file system using Intel® Manager for Lustre* software. You can then monitor the file system at the Intel® Manager for Lustre* dashboard. All of the charts presented on the manager dashboard to monitor performance and statistics, are available in monitor-only mode.

Monitor-only mode can be used to establish monitoring for file systems that don’t fully conform to the High Availability Configuration Specification. In this situation, the Corosync and Pacemaker configuration modules provided with Intel® Manager for Lustre* software are not automatically deployed. This means that Intel® Manager for Lustre*

software cannot configure the file system for server failover.

Note: RAID-based fault tolerance for storage devices are implemented independent of Intel® Manager for Lustre* software.

1.5 Overview of the graphical user interface

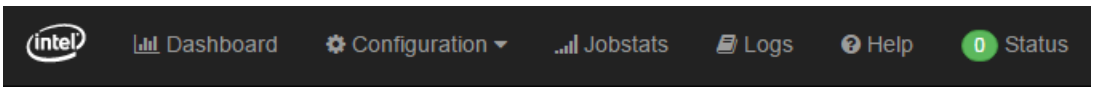
This section provides an overview of the Intel® Manager for Lustre* software GUI. For a *complete description* of the GUI, see Graphical User Interface.

The Intel® Manager for Lustre* software GUI presents a set of intuitive windows that let you set up, configure, monitor, and manage Lustre* file systems. The menu bar provides access to these capabilities. Click the following links for overview information:

- Menu bar
- Dashboard window
- Summary of charts
- Configuration menu
- Jobs Stats
- Logs window
- Help
- Status Indicator and window
- Alert bar

Menu bar

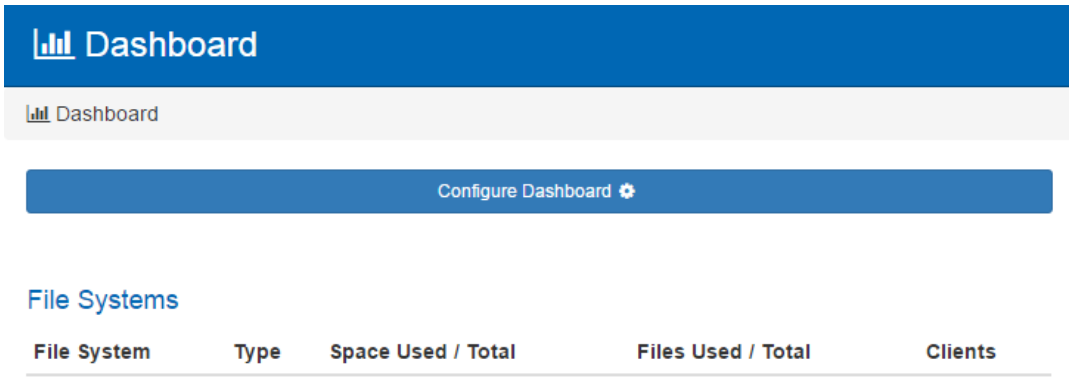
Following is the top menu bar. From here you can access the entire GUI, view the collective Status of all file systems and devices, and also access Help.



Dashboard window

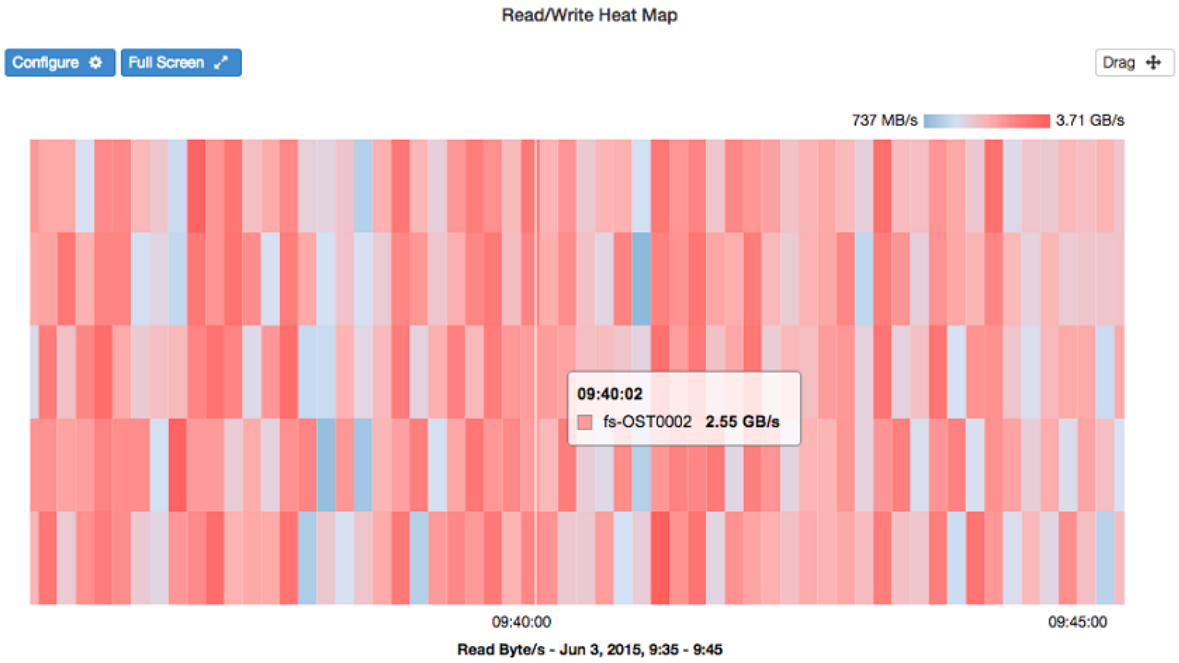
The Dashboard displays a set of charts that provide usage and performance data at several levels in the file systems being monitored. At the top level, this window displays an aggregate view of all file systems. You can select to view and monitor individual file systems and servers at the Dashboard. To view a single file system, click **Configure Dashboard** and under File System, select the desired file system.

The following is a partial view of the Dashboard.



Summary of charts

The Dashboard window presents several charts that display rich visual information about the current and historical performance of each Lustre file system. Following is an example of the Read/Write Heat Map, which is a color-coded map revealing the level of read/write activity per OST, over time.



The following twelve charts are presented. For more information, see View charts on the Dashboard.

- Read/Write Heat Map chart
- OST Balance chart
- Metadata Operations chart

- Read/Write Bandwidth chart
- Metadata Servers chart
- Object Storage Servers chart
- CPU Usage chart
- Memory Usage chart
- Space Usage chart
- File Usage chart
- Object Usage chart

Configuration menu

The Configuration drop-down menu provides access to the following several windows, where you can create, configure, and manage file systems:

- **Servers** - This window lets you add servers to the storage system and configure LNet for each server, provides server status information, and lets you start, stop, and remove servers. From here you can also automatically configure Corosync for managed HA servers.
- **Power Control** - This window lets you configure power control for each server. Here, you can add baseboard management controllers to configure IPMI to support server failover and also assign PDU outlets.
- **File Systems** - This window lists your current file systems and provides current configuration information. This window also provides access to step-by-step procedures to create and configure a file system and add system components. From this window, you can start, stop, or remove an entire file system, and you can start, stop, or remove management, metadata, or object storage targets.
- **HSM** - Hierarchical Storage Management. This window displays HSM information for one or all Lustre file systems for which HSM has been configured. After configuration, the HSM Copytool chart displays a moving time-line of waiting copytool requests, current copytool operations, and the number of idle copytool workers.
- **Storage** - This window lets you configure and view a custom storage system appliance provided by a storage solution provider. The features on this window are specific to the appliance provided by the storage solution provider.
- **Users** - This window lets you configure accounts for superusers and users.
- **Volumes** - This window provides features to configure primary and failover servers in file systems with servers configured for high availability. Each Lustre target corresponds to a single volume. If servers in the volume have been physically

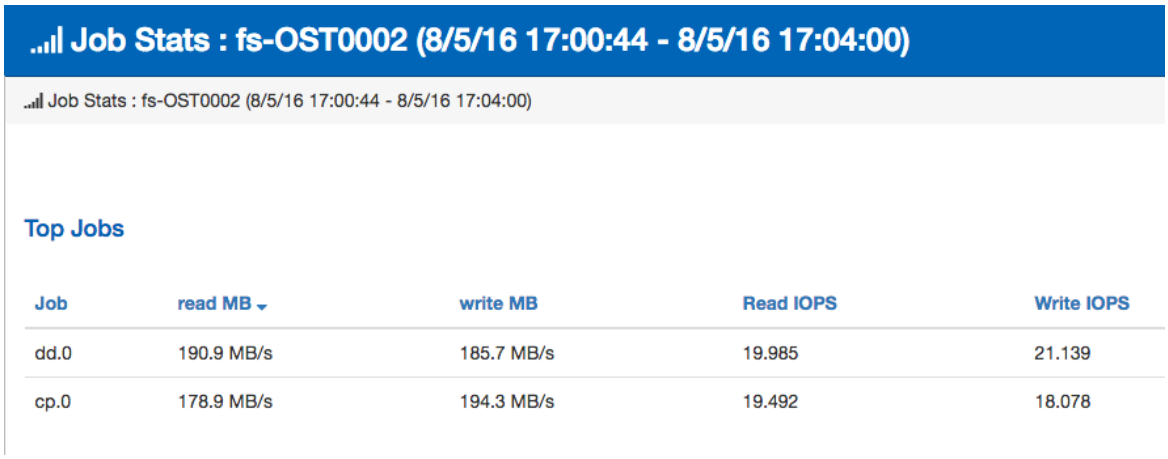
connected and then configured for high availability (using this Volumes window and the Power Control window), then primary and failover servers can be designated for a Lustre target. Only volumes that are not already in use as Lustre targets on local file systems are shown. A volume may be accessible on one or more servers via different device nodes, and it may be accessible via multiple device nodes on the same host.

- **MGTs** - This window provides features to create and configure a management target.

Job stats

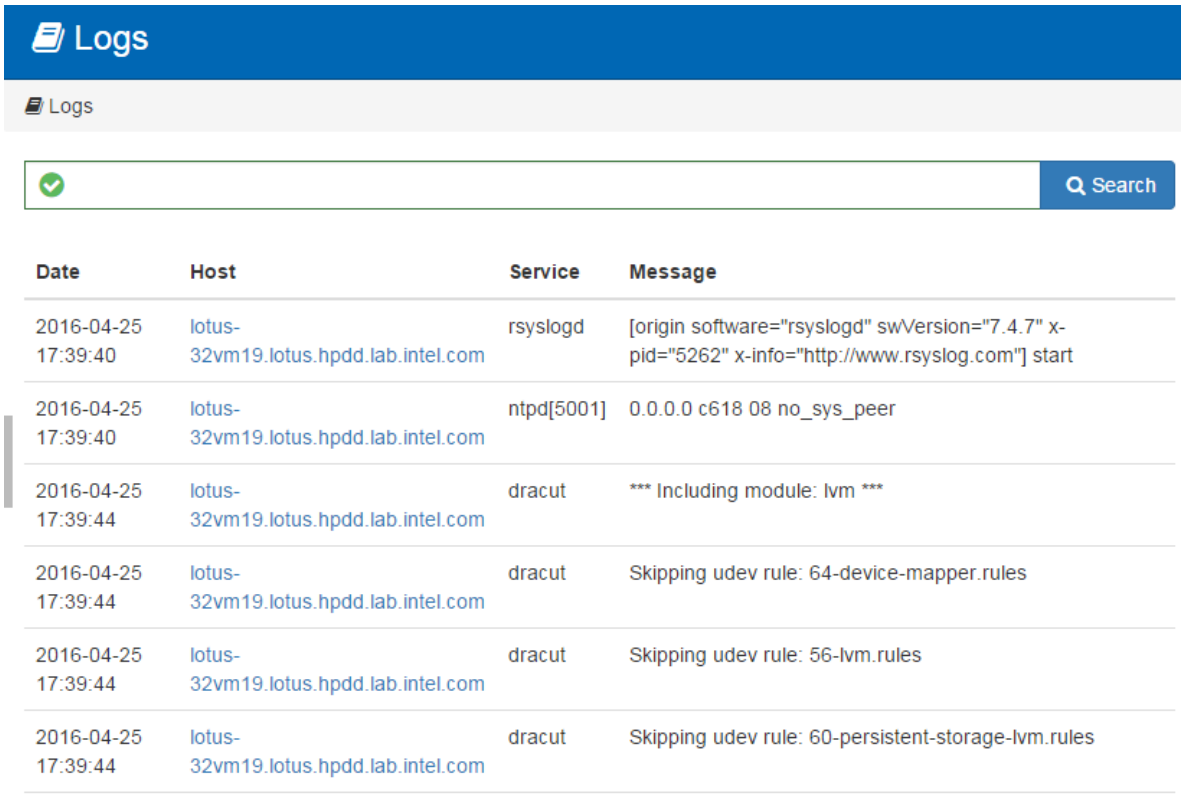
Clicking the **Jobstats** button on the top menu bar lists the top ten jobs currently in process. The listed jobs can be sorted by column and average duration can be selected. Column sorts and duration will be persistent when navigating away and back to the page.

Note: Job stats need to be enabled before then can be viewed. See View Job stats.



Logs window

The Logs window displays log information and lets you filter events by date range, host, service, and messages from Lustre or all sources. The logs window also features querying with auto-complete and linkable host names.



Help

Help is context-sensitive; Clicking Help at the menu bar opens online Help to the related topic. Internet access is not required.

Status Indicator and window

The Status indicator provides information about the functioning and health of each file system. *Alerts* are messages that indicate that the file system may be, or is, operating in a degraded mode.

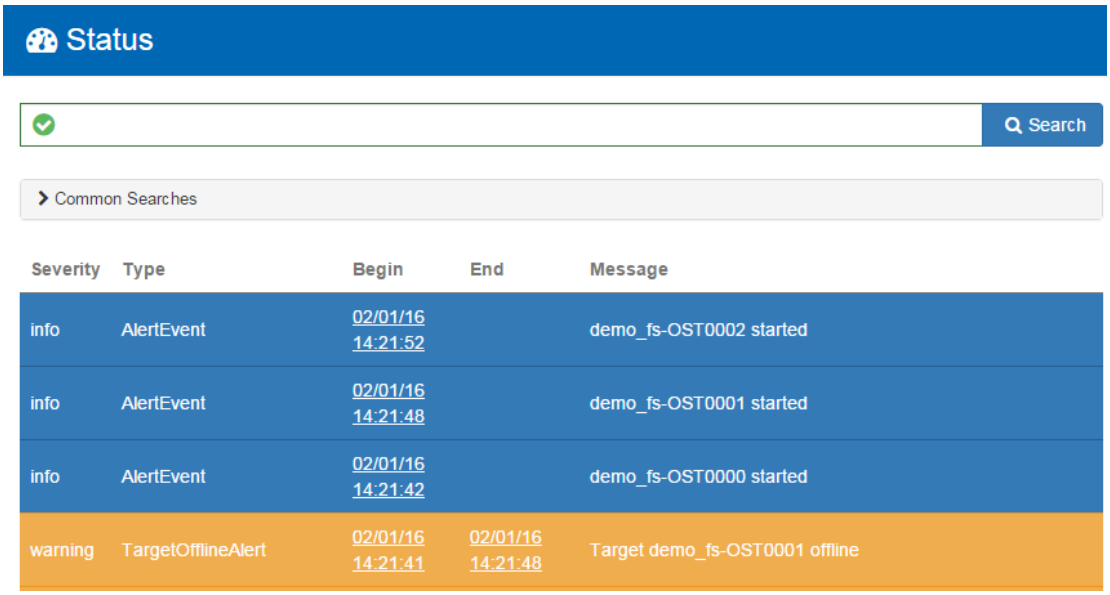
- A green light 

indicates that all is normal. Note that a green light does not indicate anything about file system performance.
- A yellow light 

indicates that one or more *warning alerts* have been received. The file system may be operating in a degraded mode, for example a target has failed over, so performance may be degraded.
- A red light 

indicates that one or more *errors alerts* have been received. This file system may be down or is severely degraded.

The Status window displays information alerts, commands that are executing, and events. For more information, see Status window.



Alert Bar

This red bar briefly appears if there are any active error or warning alerts on your system. Clicking **Details** opens the Status window and reveals the current, active alerts.

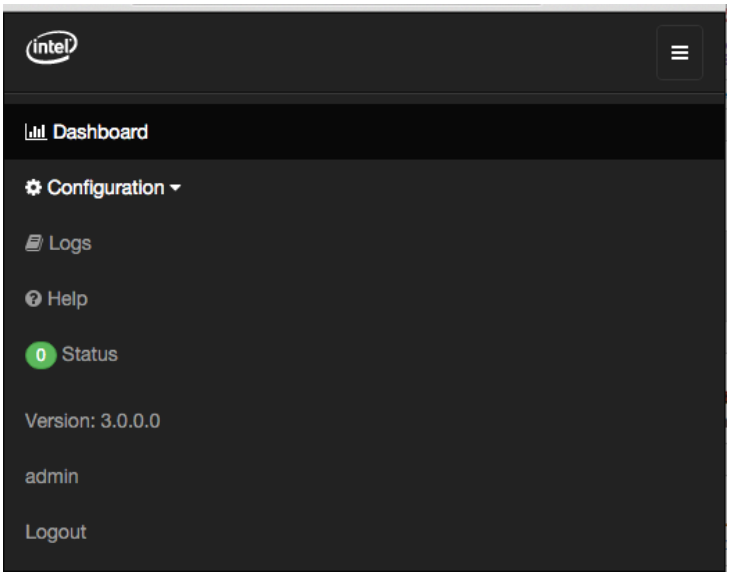
1.6 Access the Dashboard from a smart phone or tablet

You can access the Intel® Manager for Lustre GUI from your smart phone or tablet. To access the GUI from your smart phone or tablet, your device needs to be running the latest version of Chrome or Firefox browser:

1. Point your device's browser to the manager server running the Intel® Manager for Lustre software.

The window is responsive to fit within the display area.

2. To view the menu bar, click ☰. The menu bar is now displayed vertically along the left side of the window.



3. To hide the menu bar, click again.

2 Getting started

A high-availability Lustre file system managed by Intel® Manager for Lustre* software requires that your entire storage system configuration and all interfaces comply with a pre-defined configuration. For detailed information, see the section *High Availability Configuration Specification* in the *Intel® Enterprise Edition for Lustre* Software Installation Guide.* Also see the guide *Lustre* Installation and Configuration using Intel® EE for Lustre* Software and OpenZFS.*

Note: All references herein to the "manager" refer to the Intel® Manager for Lustre* software.

The Intel® Manager for Lustre* software can be used to:

- Create, monitor and manage high-availability Lustre* file systems, including systems running Open ZFS as the back-end.
- Monitor existing Lustre* file systems that have not been configured from the manager GUI.

See the following information to get started:

- For procedures for installing the Intel® Enterprise Edition for Lustre* software, including Intel® Manager for Lustre* software, and for completing initial configuration steps, see the documentation provided by your storage solution provider.
- To set up superuser and user accounts on Intel® Manager for Lustre* software see:

Creating user accounts.

- Also see: Setting up email notifications of alerts.
- To create a new Lustre file system using Intel® Manager for Lustre* software, see: Creating a new Lustre* file system.
- To detect and monitor an existing Lustre file system using Intel® Manager for Lustre* software, see: Detect and monitor existing Lustre* file systems.

WARNING: For Lustre* file systems created and managed by Intel® Manager for Lustre* software, the only supported command line interface is the CLI provided by Intel® Manager for Lustre* software. Modifying such a Lustre file system manually from a UNIX shell will interfere with the ability of the Intel® Manager for Lustre* software to manage and monitor the file system.

2.1 Creating user accounts

Note: Before creating user accounts, see the documentation provided by your storage solution provider for the initial setup procedure to be completed. The *first superuser* is created as part of *that* initial setup procedure.

To create user accounts:

1. At the menu bar, click the **Configuration** drop-down menu and click **Users**.
2. Click **+ Create user**.
3. At the *Create user* dialogue window, select the new user's role:
   a. File system user - A file system user has access to the full GUI, except for the Configuration drop-down menu, which is not displayed. A user cannot create or manage a file system, but can monitor all file systems using the Dashboard, and the Alerts and Logs windows. Users log in by clicking **Login** in the upper-right corner of the screen, and log out by clicking **Logout**.
   b. Superuser - A superuser has full access to the application, including the Configuration drop-down menu and all sub-menus. A superuser can create, monitor, manage, and remove file system and their components. A superuser create, modify (change passwords), and delete users. A superuser cannot delete their own account, but a superuser can create or delete another superuser.
4. Fill out the remainder of the *Create user* dialogue window and click **Create**.
5. To set up email notifications of alerts for a user, see Setting up email notifications of alerts.

More about roles

A superuser must be logged in to perform any actions that modify the system, such as starting a file system or adding a server.

After logging in, a user can modify their own account by clicking **Account** near the upper-right corner of the screen. A user can set these options:

- *Details* - Username, email address, and first and last name can be changed.
- *Password* - Password can be changed and confirmed.
- *Email Notifications* - The types of events for which this account will receive emailed notifications can be selected from a checklist. If no notifications are selected, email notifications will be sent for all alerts except “Host contact alerts”. See Setting up Email Notifications.

Note: Unauthenticated users can access the static HTML content present on the Intel® Manager for Lustre* GUI, but the display will not be populated with current system information unless the user is authenticated. See the documentation provided by your storage solution provider for how to configure Intel® Manager for Lustre* software to require all users to log in to see any data.

2.2 Setting up email notifications of alerts

This feature lets a superuser selectively turn on and turn off email notifications of specific classes of alerts for individual users. Users can also configure this capability. The alert email has specific information as to which component is affected.

Note: A mail handler needs to be established to forward alert emails before this feature will work. See *Enabling Email Notifications* in the *Intel® EE for Lustre* Software Installation Guide*.

To set up email notifications:

1. As the user, click **Account** in the upper right corner. Then click **Email Notifications**.
2. At the menu bar, click the **Configuration** drop-down menu and click **Users**. For the desired user, click **Edit**. Then click **Email Notifications**.
3. At the Email Notifications window, select the alert types for which you want to turn on notifications. Alert classes are listed here:
   - Host contact alert - Host lost contact with a server.
   - LNet offline alert - LNet is offline for a server.
   - LNet NIDs changed alert - See *Handling Network Address Changes*.
   - LNet NIDs changed on server <server name> - See *Handling Network Address Changes*.
   - Target offline alert - A target has gone offline.

- Target failover alert - A target is currently running on its secondary server.
- Target recovery alert - A target is in recovery.
- Storage resource offline - A monitored storage controller is offline or otherwise out of contact with chroma manager, monitoring data are not being received.
- Storage resource alert - A storage plug-in has raised an alert. This alert does not reveal the exact message generated by the storage plug-in.

4. With your selections made, click Save Changes. Clicking Reset Form returns the selections to their last saved state.

3 Creating a new Lustre* file system

This chapter describes how to create a new Lustre* file system, to be managed from the Intel® Manager for Lustre*, and how to mount file system clients.

Note: All references herein to the *manager* refer to the Intel® Manager for Lustre* software.

Note: The procedure for creating a Lustre file system *based on ZFS zpools* is different. For that information see, Creating and Managing ZFS-based Lustre file systems.

1. Important information about reconfiguring your file system
2. High-availability file system support (overview)
3. Prerequisites for creating an HA file system
4. Add one or more HA servers
5. Configure primary and failover servers
6. Add power distribution units (alternate to BMC configuration)
7. Assign PDU outlets to servers
8. Assign BMCs to servers (alternate to power distribution units and outlets)
9. Create the new Lustre file system
10. View the file system
11. Mount the Lustre file system

3.1 IMPORTANT PREREQUISITES to creating an HA Lustre file system

A high-availability Lustre file system managed by Intel® Manager for Lustre* software requires that your entire storage system configuration and all interfaces comply with a pre-defined configuration. Intel® Manager for Lustre* software performs LNet configuration assuming that each server is connected to a high-performance data

network, which is the Lustre network LNet. For detailed information, see the section *High Availability Configuration Specification* in the *Intel® Enterprise Edition for Lustre* Software Installation Guide.* Also see the guide *Lustre* Installation and Configuration using Intel® EE for Lustre* Software and OpenZFS.*

Caution: When initially setting up your storage system, take care in selecting block device names because these cannot be changed after the file system has been created using Intel® Manager for Lustre* software. **You should NOT make configuration changes to file system servers or their respective volumes/targets outside of Intel® Manager for Lustre* software**. Doing so will defeat the ability of Intel® Manager for Lustre* software to monitor or manage the file system, and will make the file system unavailable to clients. Re-labeling these device names during multipath configuration will break the HA configuration established by Intel® Manager for Lustre* software.

3.2 IMPORTANT INFORMATION about reconfiguring your file system

Caution: When initially setting up your storage system, take care in selecting block device names because these cannot be changed after the file system has been created using Intel® Manager for Lustre* software. **You should NOT make configuration changes to file system servers or their respective volumes/targets outside of Intel® Manager for Lustre* software**. Doing so will defeat the ability of Intel® Manager for Lustre* software to monitor or manage the file system, and will make the file system unavailable to clients. Re-labeling these device names during multipath configuration will break the HA configuration established by Intel® Manager for Lustre* software.

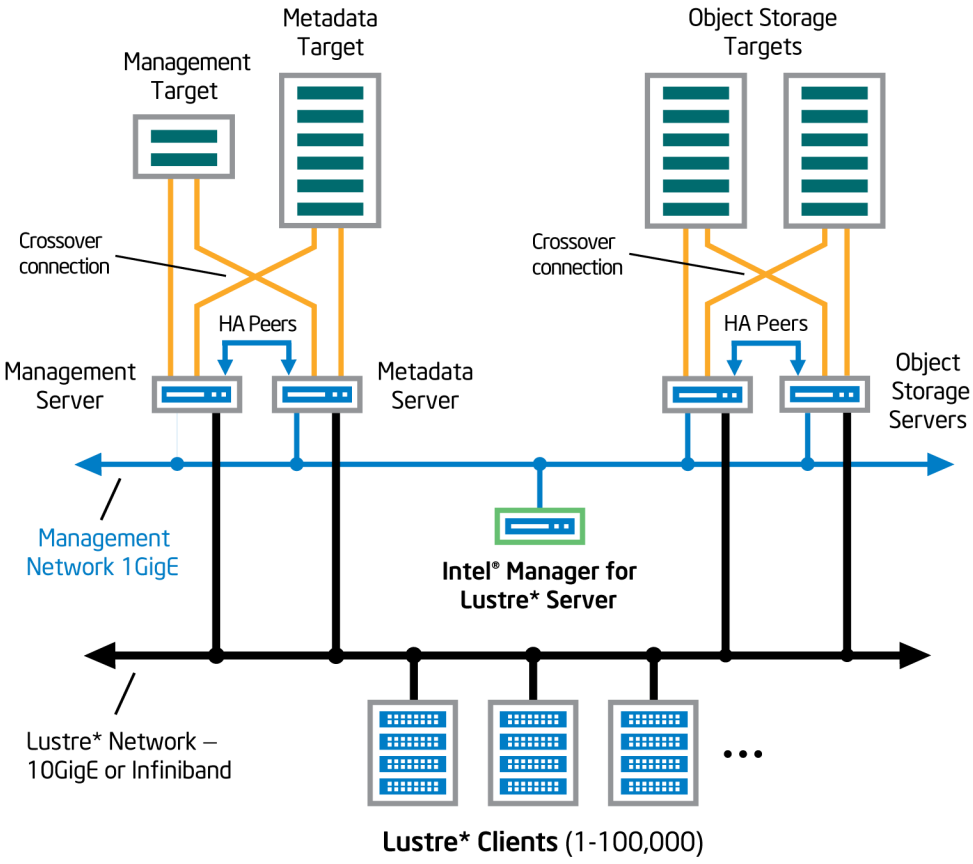
Caution: A known issue can result in a server being made unavailable. This can happen if the server has been added to a Lustre file system, (using Intel® Manager for Lustre* software) and then the user decides to Force Remove the server from the file system. The Force Remove command should only be performed if the Remove command has been unsuccessful. Force Remove will remove the server from the Intel® Manager for Lustre* configuration, but not remove Intel® Manager for Lustre* software from the server. All targets that depend on the server will also be removed without any attempt to unconfigure them. To completely remove the Intel® Manager for Lustre* software from the server (allowing it to be added to another Lustre file system), first contact technical support.

3.3 High-availability file system support

Intel® Manager for Lustre* software includes several capabilities for configuring and managing highly-available Lustre* file systems.

Generally, high availability (HA) means that the file system is able to tolerate server hardware failure without loss of service. The key components of this solution are the software components Corosync and Pacemaker. Corosync is responsible for maintaining intra-cluster control and heartbeat communications, and Pacemaker is responsible for managing HA resources (e.g., Lustre* targets).

To support automatic server failover, each HA server must have a dedicated crossover connection to the other server that will be its HA peer. During file system creation, each HA server is designated as a primary server for one or more targets, and as a failover, peer server for its peer server's targets. This crossover connection is configured as a redundant Corosync communications interface in order to reduce the likelihood of false failover events. Intel® Manager for Lustre* software uses a managed server profile to automatically configure Corosync and Pacemaker. The managed server profile is used to configure primary and failover servers. See the following figure.



Physically, HA peer servers must be cabled to provide equal access to the pool of storage targets allocated to those peers. For example: server 1 and server 2 are cabled as HA peers. Targets A and B have been configured with server 1 as their primary server and server 2 as their failover server. Targets C and D have been configured with server 2 as their primary server and server 1 as their failover server. If server 1 becomes unavailable, server 2 must have access to the block storage devices underlying targets A and B in order to mount them and make them available to Lustre clients. The end result is that server 1 is powered off and server 2 is now exporting targets A, B, C, and D to Lustre clients.

To support HA failover, each HA server must be able to automatically power-off its peer server if a failover is required. The process of powering off a faulty server is known as "node fencing" (also called "server fencing"), and ensures that a shared storage device is not mounted by more than one server at a time. Lustre includes protection against multiple simultaneous device mounts, but automatically powering off the faulty server ensures that failover works properly. Intel® Manager for Lustre* software supports the use of remotely-operable Power Distribution Units (PDUs) for this purpose. Alternative to the configuration of PDUs, Intel® Manager for Lustre* software also supports the Intelligent Management Platform Interface (IPMI) and associating baseboard management controllers (BMCs) with servers, to support server monitoring and control.

Note: See the *Intel® Manager for Lustre* Partner Installation Guide* for physical design and configuration guidelines required to support high availability.

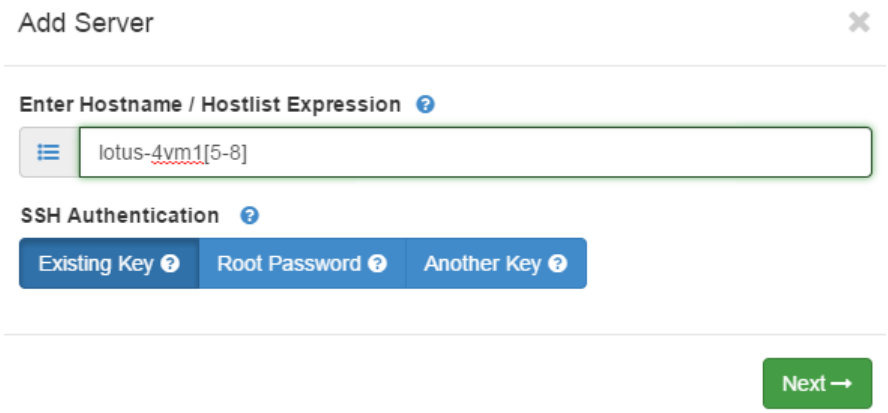
3.4 Add one or more HA servers

This procedure adds one or more servers. They may be storage servers, HSM agent nodes, Robinhood policy engine servers, or they may perform another function dictated by a custom server profile. Note that at least two storage servers are required for HA file systems.

Note: All authentication credentials are sent to the manager server via SSL and are not saved to any disk.

To add a server to be used for the file system:

1. At the menu bar, click the **Configuration** drop-down menu and click **Servers** to display the *Server Configuration* window.
2. Click **+ Add Servers.**



3. In the *Hostname / Hostlist Expression* field, enter the name of the server(s) to be added. You can enter a range of names, i.e., a "hostlist expression". For example, your can enter server[00-19] to generate a list of up to twenty servers (in this case). **Note**: These are all the server names that your expression expands to and may include servers that don't exist or are not connected to the network.
4. Select an authentication method:
   - Click **Existing Key** to use an existing SSH private key present on this server. There must be a corresponding SSH public key on each server you are adding.
   - Click **Root Password** and enter a root password for the server you are adding. This is standard password-based authentication. It is not uncommon to use the same root password for multiple servers.
   - Click **Another Key** and enter a private key that corresponds to a public key present on the server you are adding. If the key is encrypted, enter the passphrase needed to decrypt the private key.
5. Click **Next**. The software will attempt to verify the presence and readiness of all servers with names matching your hostname entry. Each server is represented by a square. A green square means that the server passed all readiness tests required for validation and this process can proceed for that server. A red square means that the server failed one or more readiness tests. Click on a red square to learn which tests the server failed. You can hover the pointer over the failed validation test to learn more.
6. For a server that failed validation, log into that server and work to address the failed validation. When the issue has been resolved, the GUI will update the failed validation test in real time, from a red x to green check mark. You can add the server when all failed validations are resolved.

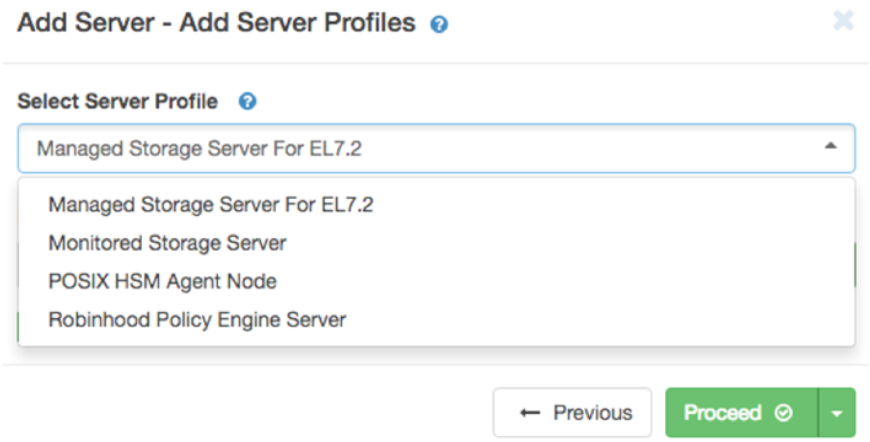
   Note: Many server names may be generated from your host list expression, and some of those servers may not exist. A red square is created for each server that doesn't exist.

7. Assuming that all servers pass the validation tests and all boxes are green, click **Proceed** to download agent software to each server. If one or more servers failed to pass validation tests, the green **Proceed** button changes to a yellow **Override** button. Clicking **Override** displays this warning: *You are about to add one or more severs that failed validation. Adding severs that failed validation is unsupported. Click **Proceed** to continue.*

   Caution: Although you can attempt to add a server that has failed validation, all of the capabilities exercised by the tests are needed for the management software and server to operate normally. The server will likely fail to operate normally if it has failed validation by this software. Adding a server that failed validation is not supported.
8. After clicking **Proceed**, agent software is deployed to each server and a **Commands** window opens to show progress. At completion, the message *Succeeded* is displayed. Click **Close** to close this **Commands** window.
9. If you decided to override servers that failed validation tests (this is not supported), expand any failed commands in the **Commands** window. Click on any failed jobs and examine the stack trace to learn the cause of the failure. Correct the cause of the failure and close the command and server windows. If the server exists in the server table, click **Actions** and select **Deploy Agent**. Otherwise open the **Add server** dialog and enter the failed server. In either case you should now see a green square for that server and be able to add it without issue.
10. With the **Commands** window now closed, the servers you added are listed as **Unconfigured** on the *Server Configuration* window. The next task is to add a server profile to each server; this step will configure the server for its purpose. For a given server, under the **Actions** drop-down menu, click **Setup server**.
11. At the *Add Server - Add Server Profiles* window, select the desired profile from the drop-down menu. Note that one profile type is selected for all servers you are adding in this process. The common profiles are listed next, but your software may have more server profiles.

    - **Managed Storage Server for EL7.2**: As above, this allows the manager GUI to configure Corosync and Pacemaker, configure NTP, etc., so that the manager software can monitor and manage the server. Managed storage servers must be physically configured for high-availability/server failover. This profile is for servers running RHEL 7.2. (In our example below, none of the servers being configuring are running RHEL 7.2, so the warning **Incompatible**, is displayed.)
    - **Monitored Storage Server**: This is for servers that are not correctly configured for HA/failover (as far as this software is concerned). A *monitored storage server* is monitored only; the manager GUI performs no such server configuration or management. Note that ZFS file systems use this profile. However the Dashboard will still display charts showing file system operations.

- **POSIX HSM Agent Node**: An HSM Agent node is used in hierarchical storage management to run an instance of Copytool. Copytool transfers certain files between the Lustre file system and the archive and deletes from the Lustre file system those files that have been archived. See Configuring and using Hierarchical Storage Management
- **Robinhood Policy Engine Server**: This server hosts the Robinhood policy engine, which enables automation of hierarchical storage management activities. See Configuring and using Hierarchical Storage Management.



12. Select the desired profile and click **Proceed**. The manager does an audit of the storage resources on each server. The manager then provisions the server by loading appropriate Lustre modules and making sure the Lustre networking layer is functioning. When all checks are completed, *LNet State* indicates *LNet Up* and each server is fully qualified as a Lustre server. Under the *Status* column, a green check mark is displayed for each new server. If server provisioning does not succeed, the *Status* will indicate a exclamation mark (!) and the *LNet State* may indicate *Unconfigured*. To learn the cause of the problem, click the exclamation mark for the failed server to see *Alerts*. For more information, click **Status** at the top menu bar. The *Status* window also lets you view related logs.

    Note: A certain profile may not be compatible with a server as the server is configured. If the profile you select is not compatible with the server(s) you specified, a warning is displayed: *Incompatible*. Each incompatible server is represented by a red box. To learn why a server is incompatible, click on a red box. A pop-up window reveals the problem. You can resolve the problem and the red box will change to green, indicating profile compatibility with the server.

    Caution: For servers with incompatible profiles, you have the option of clicking **Override**, however, this is not encouraged or supported. Each server's configuration

must be compatible with the selected profile, or the server will likely not function as required for the selected profile. The four available default server profiles are described above. For more information about the POSIX HSM Agent Node and Robinhood Policy Engine Server profiles, see Configuring and using Hierarchical Storage Management herein.

13. Click **Close**. This process is complete. For HA file systems, proceed to Configure primary and failover servers.

3.5 Configure primary and failover servers

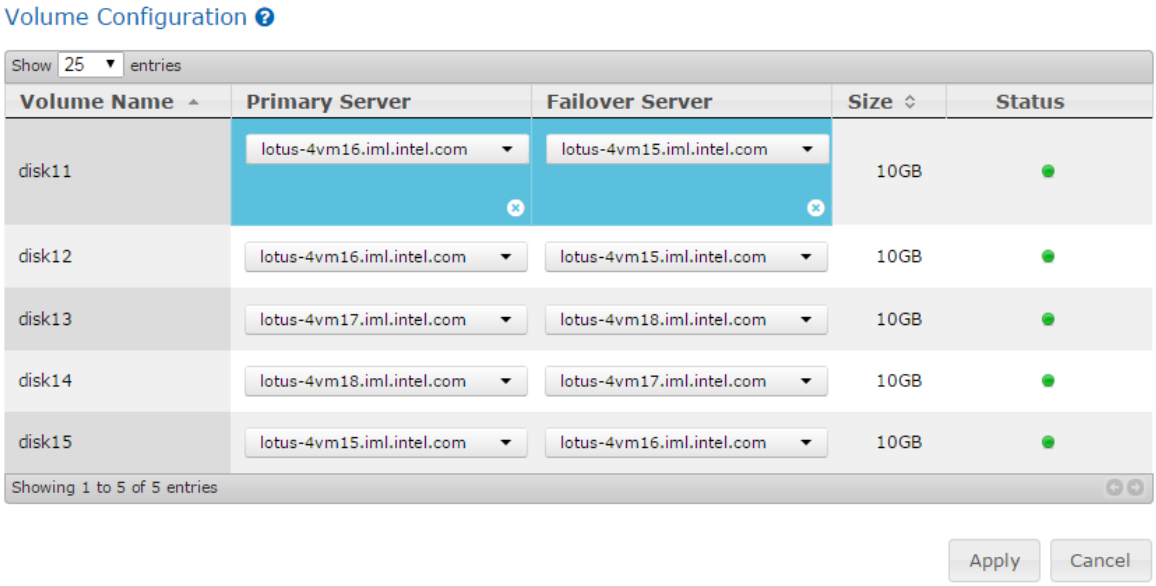
This section establishes the primary and failover storage servers for each volume, to support-high availability.

Caution: Configuring primary and failover servers on the *Volumes* window does not by itself enable high-availability. Automated failover also requires power management support by PDUs and outlet assignments or by assigning BMCs to servers. See Add power distribution units or Assign BMCs to servers. It is important to remember these server/volume configurations for when configuring power distribution units (PDUs) and outlet-server assignments.

Note: This section is for configuring managed storage servers, as previously set up in Add storage servers. This section does not apply to servers that are monitor-only.

To view the volumes that were discovered and make adjustments to volume configurations, complete these steps:

1. At the menu bar, click the **Configuration** drop-down menu and click **Volumes** to display the *Volume Configuration* window. A list of available volumes is displayed (if a volume does not contain unused block devices, it will not appear on this list).
2. The *Volume Configuration* window displays the current primary and failover servers for each volume. If required, you can change these primary and failover server assignments. To do so, for a given volume select the volume's *Primary Server* from the drop-down list. Then select the *Failover Server* from the drop-down list. Changes you make to volume/server configuration appear in blue, indicating that you have selected to change this setting, but have not applied it yet. To undo a change in-process, click the **x**.



3. Repeat step 2 for each volume that has a primary and failover server.
4. Click **Apply**. Then click **Confirm.**

Changes you select to make on this *Volumes Configuration* window will be updated and displayed after clicking **Apply** and **Confirm**. Other users viewing this file system's *Volume Configuration* window will see these updated changes after you apply and confirm them. To cancel all changes you have selected (but not yet applied), click **Cancel**.

Note: There is currently no lock-out of one user's changes versus changes made by another user. The most-recently applied setting is the one in-force.

Next, proceed to Add power distribution units or Assign BMCs to servers. It is important to remember these server/volume configurations for when configuring power distribution units (PDUs) and outlet-server assignments.

3.6 Add power distribution units

This section configures power distribution units (PDUs) and assigns PDU outlets to servers to support high availability (HA).

Note: A server cannot be associated with both a BMC and PDU outlets. Use PDUs or IPMI/BMCs to support failover.

Note: This section is for configuring managed storage servers, as previously set up in Add storage servers. You should configure PDUs based on primary and failover server configuration, per volume. This section (Add power distribution units) does not apply to servers that are monitor-only.

Issues Regarding Power Loss to the BMC or PDU

Regarding failover, if the method of power control is not functioning (e.g., loss of power to the fencing device, misconfiguration, etc.), then HA will be unable to fail the targets from the failed server to its failover server. This is because in order to complete failover, the failover server needs to be able to guarantee that the failed server can no longer access targets running on it. The only way to be sure of this is to power-down the failed server. Thus, the failover server needs to be able to communicate with the fencing device of the failed server for failover to occur successfully.

With IPMI, the power for each HA server and its fencing device is coupled together. This means there are more scenarios where both may lose power at once (chassis power failure, motherboard failure, etc.). In such a case, if a server suffers chassis power failure such that the BMC is no longer able to operate, HA will be unable to fail the targets over. To remedy this situation, restore power to the chassis of the failed server to restore the functionality of your file system. If HA coverage for the scenarios just described is important to you, we strongly recommend using smart PDUs, rather than IPMI as your fencing device.

For a PDU, power loss to the PDU will mean that HA will be unable to fail the targets over. As in the above situation, the remedy is to restore power to the PDU to restore the functionality of your file system. We recommend redundant PDUs if availability is critical. This approach is a necessary limit of HA to protect the integrity of the targets being failed over.

At the PDUs window you can add PDUs and then assign specific PDU outlets to specific servers. You should have at least two PDUs to support failover.

To add PDUs:

1. At the menu bar, click the **Configuration** drop-down menu and click **Power Control**.
2. With no power distribution units recognized, this window will read: *No power distribution units are configured*. Click **+ Add PDU**.
3. At the *Add PDU* dialogue window, select the PDU device type from the drop-down list.
4. Enter a name for this PDU. (If not entered, this field will default to the IP address.)
5. Enter the IP address for the PDU. If not known, enter a DNS-resolvable hostname. The address is stored as an IPv4 address. **Note**: This address is always stored as an IPv4 address, so if the mapping from hostname to IPv4 address later changes in DNS, it will need to be updated here as well.
6. Enter the port number. This port number must be unique to this PDU.
7. Enter your *Management user name*.
8. Enter your *Management password*. Then click **Save**.

Proceed to Assign PDU outlets to servers.

3.7 Assign PDU outlets to servers

Note: Be sure that electrical connection between a given power distribution unit (PDU) outlet and a specific server is performed by a qualified technician before PDUs are configured and assigned by this software.

Note: This section is for configuring managed storage servers, as previously set up in Add storage servers. This section does not apply to servers that are monitor-only.

Note: A server cannot be associated with both baseboard management controller (BMC) outlets and power distribution unit (PDU) outlets. Use PDUs or IPMI/BMCs to support failover.

Before assigning PDU outlets to servers, make note of the primary and failover server configurations for each volume on the *Volumes* window. Be sure to assign failover outlets from different PDUs than the primary outlets. When you associate PDU failover outlets with servers using this tool, STONITH is automatically configured.

To assign PDU outlets to servers:

1. At the menu bar, click the **Configuration** drop-down menu and click **Power Control**. The PDUs you already added should be displayed. If no PDUs are present, see Add power distribution units.
2. The left column shows all servers used in all file systems that you're currently managing. Each column to the right of the *Server* column shows outlet assignments for one PDU. If you have four PDUs configured, then there are four PDU columns. Each row represents an outlet-to-server assignment. To assign PDU outlets to servers:
   a. Pick a server row for which you want to assign outlets.
   b. Mouse over to the PDU column and click within the drop-down box to expose the outlets available from this PDU. Now select the desired outlet. (You can also use the tab key to move to the desired server/PDU. Then begin to enter the outlet name. This field auto-fills. Tab or press **Enter** to confirm this selection.)
   c. Move to the next server and assign outlets in the same way. Note that as an outlet is assigned to a server, it becomes unavailable for reassignment.
   d. To remove an outlet from a server, click the **X** next to the outlet name. It now becomes available to reassign.

3.8 Assign BMCs to servers

This section uses the Intelligent Management Platform Interface (IPMI) and associates baseboard management controllers (BMCs) with servers to support high availability.

Note: This section is for configuring managed storage servers, as previously set up in Add storage servers. This section does not apply to servers that are monitor-only.

Note: A server cannot be associated with both a BMC and PDU outlets. Use PDUs or BMCs to support failover.

Issues Regarding Power Loss to the BMC or PDU

Regarding failover, if the method of power control is not functioning (e.g., loss of power to the fencing device, misconfiguration, etc.), HA will be unable to fail the targets from the failed server to its failover server. This is because in order to complete failover, the failover server must be able to guarantee that the failed server can no longer access targets running on it. The only way to be sure this is true is to remove power from the failed server. Thus, the failover server must be able to communicate with the fencing device of the failed server for failover to occur successfully.

With IPMI, the power for each HA server and its fencing device is coupled together. Accordingly, there are more scenarios where both may lose power at once (chassis power failure, motherboard failure, etc.). If a server suffers chassis power failure such that the BMC is not operational, HA will be unable to fail the targets over. The remedy in this situation is to restore power to the chassis of the failed server to restore the functionality of your file system. *If HA coverage for the scenarios just described is important to you, we strongly recommend using smart PDUs, rather than IPMI as your fencing device.*

Power loss to a PDU will mean that HA will be unable to fail the targets over. As in the above situation, the remedy is to restore power to the PDU to restore the functionality of your file system. *We recommend redundant PDUs if availability is critical.*

This approach is a necessary limit of HA to protect the integrity of the targets being failed over.

To associate BMCs with servers:

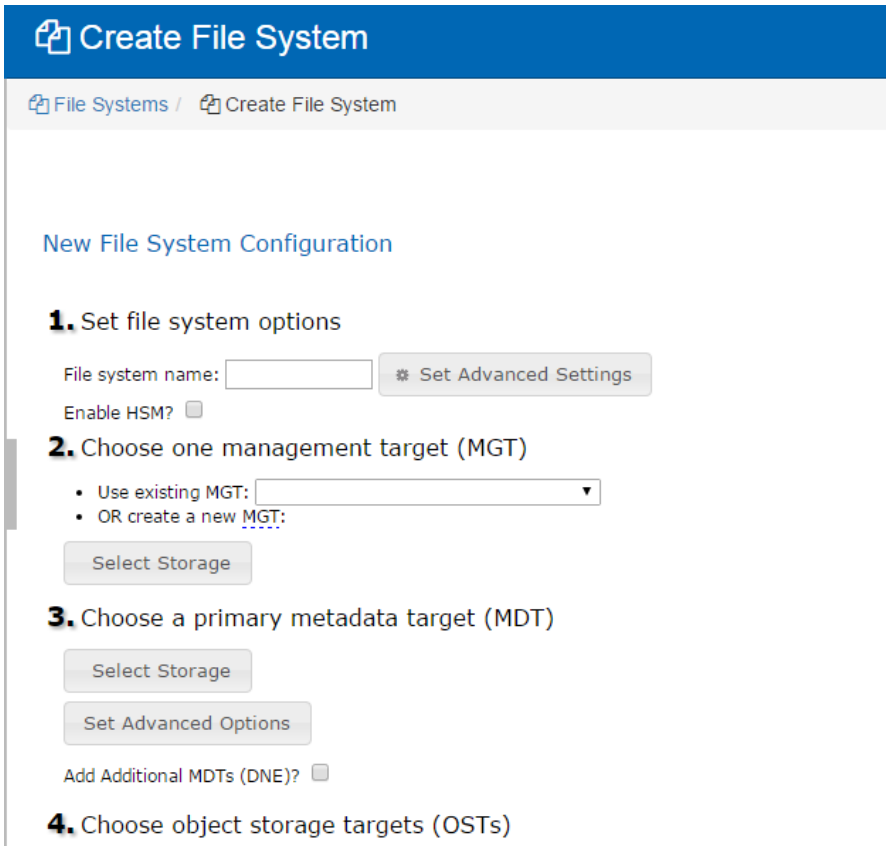
1. At the menu bar, click the **Configuration** drop-down menu and click **Power Control**.
2. Click **+ Configure IPMI**.
3. At the *Configure IPMI* dialogue window, enter your *Management username* and *Management password*. Click **Save**.
4. Each row is one server. For the desired server, under IPMI, click **+ Add BMC**.
5. In the *New BMC* window, enter an IP address or hostname for this BMC. **Note**: This address is always stored as an IPv4 address, so if the mapping from hostname to IPv4 address later changes in DNS, it will need to be updated here as well.
6. Click **Save**.

3.9 Create the new Lustre file system

This section is the last procedure to create the Lustre file system, after performing the previous configuration tasks outlined in this chapter. In this section, you will select servers.

To create the file system:

1. At the menu bar, click the **Configuration** drop-down menu and click **File Systems** to display the *File Systems* window.
2. Click **Create File System** to display *New File System Configuration*.



3. In the *File system name* field, enter the name of the new file system. The name can be no more than eight characters long and should conform to standard Linux naming conventions.
4. If this file system is to utilize Hierarchical Storage Management, click the check-box **Enable HSM**.
5. At step 2, Choose a management target (MGT). Intel® Manager for Lustre* software does not support an MGT larger than 10 gigabytes. *If a management target has been created previously*, the following options will be available. Use one of these options to select the MGT:

- If the MGT is to be installed on an existing server in the file system, you can select the target to be used from the *Use existing MGT* drop-down list.
- If a new MGT is to be created, click **Select Storage** to display a list of available servers and then click the server to be used.

Note: The MGT and MDT can be located on the same server. However, they cannot be located on the same volume/target on a server.

6. At step 3, choose a primary metadata target (MDT) by clicking **Select Storage**. Then at the drop-down menu, select the target to be used.
7. Notice the check-box labeled **Add Additional MDTs (DNE)**. After selecting the primary MDT, you can also add additional MDTs. DNE stands for Distributed Namespace. DNE allows the Lustre namespace to be divided across multiple metadata servers. This enables the size of the namespace and metadata throughput to be scaled with the number of servers. The primary metadata target in a Lustre file system is MDT 0, and added MDTs are consecutively indexed as MDT 1, MDT 2, and so on.

   To add an additional MDT, click the check-box**.** Then at the drop-down menu, select the additional MDT target or targets to be used. At the end of this process, after creating the file system, you will enter a command to configure this MDT.

   Note: You can also add additional MDTs after the file system has been created; see Add additional Metadata Targets. Any added MDT you create will be unavailable for use as an OST.
8. At step 4, choose the object storage targets (OSTs) for the file system by checking the boxes next to the targets to be included in the system.
9. Click **Create File System** now to create the file system.
10. To follow the process as the file system is created, click on **Status** on the top menu bar and select **Commands**. After the file system creation has completed successfully, perform the remaining steps if applicable.
11. If you selected to add additional MDT(s), then log into a client node and mount the Lustre file system. Then at the command line, for each added MDT beyond the primary MDT, enter the following command:

    ```
    lfs mkdir -i n <lustre_mount_point>/<parent_folder_to_contain_
    this_MDT>
    ```

    where the -i indicates that the following value, n is the MDT index. The first added MDT will be index 1.
12. Users can now create subdirectories supported by this MDT with the following command, as an example:

```
mkdir <lustre_mount_point>/<parent_folder_to_contain_
this_MDT>/<subdirectory_name>
```

Note: Intel® Manager for Lustre* software will automatically assign OST indices in a distributed fashion across servers to permit striping.

Note: If you plan to enable HSM for this file system, see the chapter Configuring and using Hierarchical Storage Management to setup HSM.

3.10 View the new file system

To view the file system configuration:

1. At the menu bar, click the **Configuration** drop-down menu and click **File Systems**.
2. At the *File Systems* window, select the name of the file system from the table displayed.
3. To view the dashboard metrics for the file system, at the menu bar, click **Dashboard** window and select **File Systems.** Select the file system in the fields displayed at the top of the window.

Note: For a new file system, some of the dashboard charts may appear blank until the file system has been running long enough to collect performance data.

3.11 Mount the Lustre file system

A compute client must mount the Lustre* file system to be able to access data in the file system. Before attempting to mount the file system on your Lustre clients, make sure the Intel® Enterprise Edition for Lustre* client software has been installed on each client. For instructions, see the documentation provided by your storage solution provider.

A client can mount the entire file system, or mount a file system sub-directory.

- Mount the entire file system
- Mount a file system sub-directory

3.11.1 Mount the entire file system

To obtain the command to use to mount an entire file system:

1. At the Dashboard menu bar, click the **Configuration** drop-down menu and click **File Systems**.
2. Each Lustre file system created using Intel® Manager for Lustre* is listed. Select the file system to be mounted. A window opens showing information for that file system.
3. On the file system window, click **View Client Mount Information**. The mount

command to be used to mount the file system is displayed. Following is an example only:
mount -t lustre 10.214.13.245@tcp0:/test /mnt/test

4. On the client server, enter the actual command.

3.11.2 Mount a file system sub-directory

To mount a file system, at the client computer, enter the following command at the command line. This syntax is generic and there are other options not described here.

```
mount -t lustre <mgsnid>[:<mgsnid>]:/<fsname>/<subdir path> <mount point>
```

4 Monitoring Lustre* file systems

You can easily monitor one or more file systems at the Dashboard, and Status, and Logs windows. The Dashboard window displays a set of charts that provide usage and performance data at several levels in the file systems being monitored, while the Status and Logs windows keep you informed of file system activity relevant to current and past file system health and performance.

- View charts on the Dashboard
- Check file systems status
- View alerts and events status messages
- View commands and status messages on the Status window
- View Logs
- View and change file system parameters
- View a server's parameters

4.1 View charts on the Dashboard

The Dashboard displays a set of graphical charts that provide real-time usage and performance data at several levels in the file systems being monitored. All Dashboards charts are available for both monitored-only and managed/monitored file systems.

At the top, the Dashboard lists the file system(s) being managed or monitored-only. The following information is provided for each file system:

- **File System**: The name assigned to this file system during its creation on the *Configuration* window.
- **Type**: **Monitored** or **Managed**. Managed file systems are configured and managed for high availability (HA). Managed file systems are both monitored and managed,

whereas monitored file systems are monitored-only and do not support failover via Intel® Manager for Lustre* software.

- **Space Used** / **Total**: This indicates the amount of file system capacity consumed, versus the total file system capacity.
- **Files Used** / **Total**: This indicates the total number of inodes consumed by file creation versus the total number of inodes established for this file system.
- **Clients**: Indicates the number of clients accessing the file system at this moment.

Persistent Chart Configuration

You can configure certain data display parameters for each chart, and your chart configuration will persist until you reload/refresh the Dashboard page, using the browser.

Following is an example of the OST Read/Write Heat Map chart.

See:

- View charts for one or all file systems (including all OSTs, MDTs, and servers)
- View charts for one or all servers
- View charts for an OST or MDT

4.1.1 View charts for one or all file systems

When you first login, the Dashboard displays the following six charts for all file systems combined. Click on the links here to learn more.

- Read/Write Heat Map chart
- OST Balance chart
- Metadata Operations chart
- Read/Write Bandwidth chart
- Metadata Servers chart
- Object Storage Servers chart

To these view these six charts for a single file system:

1. If it is not displayed, click **Dashboard** to access the Dashboard window. The default view is for all six charts to be displayed.
2. Click **Configure Dashboard.**
3. Under **File System**, selected the file system you wish to view.
4. Click **Update**.

4.1.2 View charts for one or all servers

When you first login, the Dashboard displays six charts for all file systems combined.

View charts for all servers combined

Viewing charts for all servers is the same thing as viewing charts for all file systems. To do this:

1. On the Dashboard, click **Configure Dashboard**.
2. Leave All Servers selected in the **Server** drop-down menu.
3. Click **Update**.

View charts for an individual server

4. On the Dashboard, click **Configure Dashboard**.
5. Select **Server**.
6. Under Server, select the server of interest and click **Update**.

The following charts are displayed for an individual server. Click on the links here to learn about these charts.

- Read/Write Bandwidth
- CPU Usage
- Memory Usage

4.1.3 View charts for an OST or MDT

To view charts for a specific OST or MDT:

1. On the Dashboard, click **Configure Dashboard**.
2. Select **Server**.
3. At the Server drop-down menu, select the sever hosting the desired target.
4. At the Target drop-down menu, select the desired target. Then click **Update**.

The following charts are displayed for OSTs. Click on the links here to learn about these charts.

- Read/Write Bandwidth
- Space Usage
- Object Usage

The following charts are displayed for MDTs:

- Metadata Operations
- Space Usage
- File Usage

4.2 Check file systems status

The file systems Status light 

provides a quick glance of the status and health of the *all* file systems managed by Intel® Manager for Lustre* software. This indicator is located along the top banner of the manager GUI. The indicator reflects the worst-case condition. For example, and Error message for any file system will always display a red Status light. Click **Status** to open the Status window and learn more about status.

- A green Status light 

indicates that all is normal. No errors or warnings have been received. The file system is operating normally.
- A yellow Status light 

indicates that one or more warning alerts have been received. The file system may be operating in a degraded mode, for example a target has failed over, so performance may be degraded.
- A red Status light 

indicates that one or more errors alerts have been received. This file system may be down or is severely degraded. One or more file system components may be currently unavailable, for example, both the primary and secondary servers for a target are not running.

Click **Status** to open the Status window. See View status messages on the Status window.

4.3 View job stats

Job statistics are available from two locations:

- Clicking the **Jobstats** button on the top menu bar lists the top ten jobs currently in process. The listed jobs can be sorted by column and average duration can be selected. Column sorts and duration will be persistent when navigating away and back to the page.
- Clicking on an OST cell on the *Read/Write Heat Map* chart. The job statistics feature displays the read and write throughput for the top ten jobs for that OST and selected time interval. The window also shows the top Read IOPS and Write IOPS for that OST and time interval. This feature supports the creation of plug-ins to display user account, command line, job size, and job start/finish times. Information is updated every ten seconds.

To view job statistics

1. Before viewing job statistics, you will need to run a command to enable this feature. Run this command for each file system. The following command is an example, to run on the management server (MGS):

   ```
   lctl conf_param <test1>.sys.jobid_var=procname_uid
   ```

   where `<test1>` is the file system name (refer to *Using jobstats with other job schedulers* for more information.
2. The variable testfs.mdt.job_cleanup_interval sets the period after which collected statistics are cleared out. If this interval is to short, statistics may get cleared while you're viewing job statistics. Set this interval to a value greater than your collection/ viewing period. As an example, you could set this interval to 70 minutes (4200 seconds) using the following command:

   ```
   lctl conf_param testfs.mdt.job_cleanup_interval=4200
   ```
3. View the *Read/Write Heat Map* chart on the dashboard window.
4. Each row on the *Read/Write Heat Map* corresponds to an OST, with consecutive columns from left-to-right, corresponding to consecutive time intervals. Mouse over a cell to find an OST and time interval of interest, and click on the desired cell.

   The **Jobs Stats** window opens. The top banner reveals the OST and time interval. Each job executing during that interval is displayed as a row, with its average data

throughput revealed for that interval. Only the top five read and write jobs are displayed. The window displays the Read Bytes, Write Bytes, Read OPS, and Write IOPS for the top five jobs, listed by Job ID.

5. To change the duration of the job statistics sampling period, return to the *Read/ Write Heat Map* chart. Click **Change Duration** and set the time period for the heat map. If you set the time period to one day (as an example), the 24-hour period will be divided into 20 equal, consecutive cells, starting 24 hours previous and ending now. Each Read/Write Heat Map cell now covers 1.2 hours. Clicking on a cell now will reveal a job statistics window that averages 1.2 hours of read/write operations.
6. To send this Job Stats window to another person, select and copy the URL from browser URL field. Then paste the URL into an email message body and send.

Note: The **Job Stats** window is static, specific to that time period and OST. To view another time period or OST, return to the **Read/Write Heat Map** chart and select the desired cell.

Using job stats with other job schedulers

The job stats code extracts the job identifier from an environment variable set by the scheduler when the job is started. Intel® EE for Lustre* software sets a `jobstats` environment variable to work with SLURM, however you can set the variable to work with other job schedulers. To enable job stats to work with a desired scheduler, specify the jobid_var to name the environment variable set by the scheduler. For example, SLURM sets the SLURM_JOB_ID environment variable with the unique job ID on each client. To permanently enable jobstats on the testfs file system, run this command on the MGS:

```
$ lctl conf_param testfs.sys.jobid_var=<environment variable>
```

- where <environment variable> is one of the following:

Job Scheduler	**environment variable**
Simple Linux Utility for Resource Management (SLURM)	SLURM_JOB_ID
Sun Grid Engine (SGE)	JOB_ID
Load Sharing Facility (LSF)	LSB_JOBID
Loadleveler	LOADL_JOBID
Portable Batch Scheduler (PBS)/MAUI	PBS_JOBID

Cray Application Level Placement Scheduler (ALPS)	ALPS_APP_ID

To disable job stats, specify `jobid_var` as `disable`:

```
$ lctl conf_param testfs.sys.jobid_var=disable
```

To track job stats per process name and user ID (for debugging, or if no job scheduler is in use), specify `jobid_var` as `procname_uid`:

```
$ lctl conf_param testfs.sys.jobid_var=procname_uid
```

4.4 View and manage file system parameters

After you have created a file system, you can view its configuration and manage the file system at the File System Details window.

4.5 View a server's detail window

To view all parameters available for a server, at the menu bar, click the **Configuration** drop-down menu and click **Servers**. Select the server to view the Server Details window.

4.6 View commands and status messages on the Status window

The Intel® Manager for Lustre* software provides status messages about the health of each managed file system.

View all status messages

Click **Status** to view all status messages. All messages are displayed most-recent first. Note that Warning and error messages are displayed as *alerts.* The Status window displays messages in five categories:

- *Command Running:* These messages are gray in color and inform you of commands that are currently in progress, running. These are commands that you have entered at the manager GUI.
- *Command Successful*: These messages are green in color and identify commands that have completed successfully. You can click **Details** and then click the command link to learn about underlying commands and their syntax.
- *Info messages* - These messages are displayed in blue. Events are normal transitions that occur during the creation or management of the file system, often in response to

a command entered at the GUI. A single command may cause several events to occur. An event message informs you of an event occurring at a single point in time.

- *Warning alerts*: Warnings are displayed in orange. A warning usually indicates that the file system is operating in a degraded mode, for example a target has failed over so that high availability is no longer true for that target. A warning message marks a status change that has a specific **Begin** and **End** time. A warning is active at the beginning of the status change and inactive at the end of the status change.
- *Errors alerts*: Errors are displayed in red. An error message indicates that the file system is down or severely degraded. One or more file system components are currently unavailable, for example both primary and secondary servers for a target are not running. An error often has a remedial action you can take by clicking the button.

For more information see Status window.

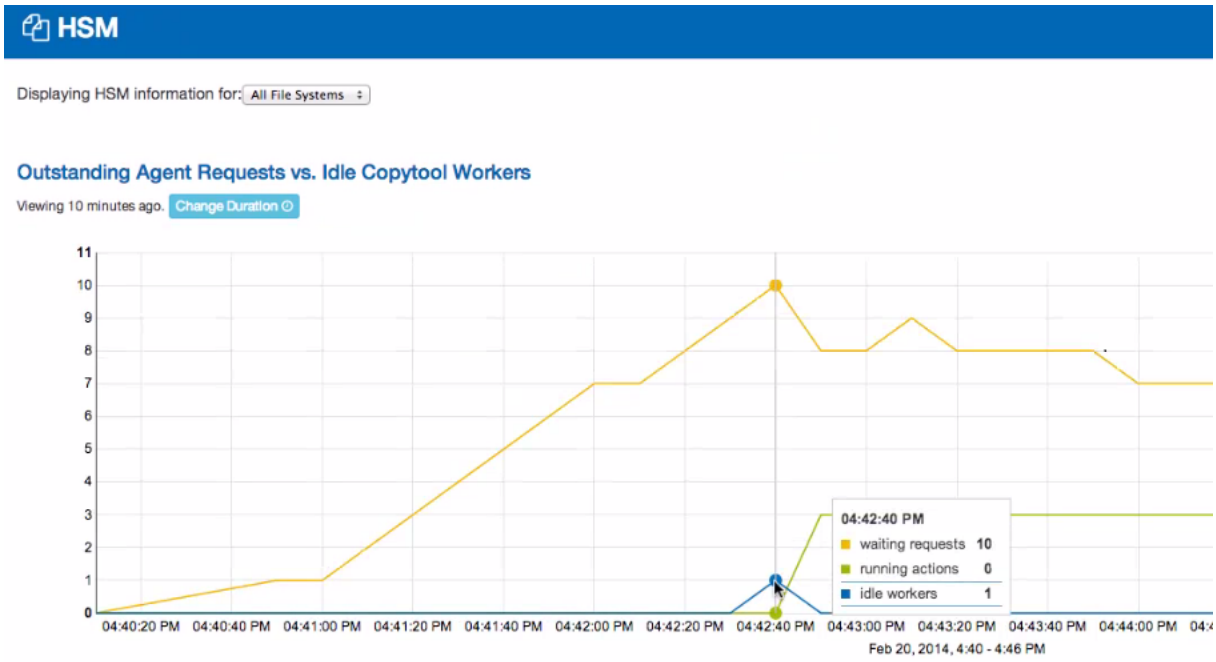
4.7 View Logs

Click **Logs** on the menu bar to view all system logs.

The Logs window displays log information and allows filtering of events by date range, host, service, and messages from Lustre or all sources.
The logs window also features querying with auto-complete and linkable host names.

4.8 View HSM Copytool activities

To view current copytool activities, click **Configuration** and select **HSM**. To learn about HSM capabilities supported in Intel® Enterprise Edition for Lustre* software, see Configuring and using Hierarchical Storage Management.
After HSM has setup for a file system, this HSM Copytool chart displays a moving time-line of waiting copytool requests, current copytool operations, and the number of idle copytool workers.



- Select to display copytool operations for all file systems (default) or one you select.
- Mouse over the graph to learn the specific values at a given point in time.
- Click **Actions** > **Disable** to pause HSM for this file system. New requests will be scheduled and HSM activities will resume after the HSM coordinator is enabled. To enable again, click **Actions** > **Enable**.
- Click **Actions** > **Shutdown** to stop the HSM coordinator for this file system. No new requests will be scheduled.
- Use **Change Duration** to change the time period for the range of data displayed on the HSM Copytool chart. chart. The chart begins at a start time set and ends now. You can set this to select **Minutes**, **Hours**, **Days** or **Weeks**, up to four weeks back in time and ending now. The most recent data displayed on the right. The number of data points will vary, based primarily on the duration.

5 Managing and Maintaining HA Lustre file systems

Warning: After you have created a Lustre file system using Intel® Manager for Lustre* software, you should not make any configuration changes outside of Intel® Manager for Lustre* software, to file system servers, their respective targets, or network connectivity. Doing so will likely defeat the ability of Intel® Manager for Lustre* software to monitor or manage the file system, and will make all or portions of the file system unavailable to clients.

Before performing any upgrades or maintenance on a primary HA server, all file system

targets attached to that server must be manually failed over to the secondary server, using the Intel® for Manager Lustre* software. DO NOT independently shut the server down.

In addition to the links below, see Advanced topics.

- Increase a file system's storage capacity
- Add an object storage target to a managed file system
- Start, stop, or remove a file system
- Start or stop an MGT, MDT, or OST
- Remove an OST from the file system
- Perform a single target failover from primary to secondary server
- Perform a single target failback from secondary to primary server
- Failover all targets from a primary to a secondary server
- Handling network address changes (updating NIDs)
- Reboot, power-off, or remove a server
- Reconfiguring Corosync and Pacemaker for a server
- Reconfiguring NIDs for a server
- Decommission a server for an MGT, MDT, or OST

5.1 Increase a file system's storage capacity

Perform the following procedures to increase a file system's storage capacity. This section applies to *managed*, high-availability file systems. For instructions on increasing the capacity a *monitored* file system, see Detect and monitor existing Lustre file systems.

Add a storage server

Adding another storage server may not be necessary if storage servers already present can accept additional OSTs. Remember that in HA systems, each OST is served by a primary and a failover server.

Perform the following procedures to add a storage server.

- Configure a storage server
- Configure primary and failover servers (required for HA)
- Add power distribution units (required for HA)
- Assign PDU outlets to servers (required for HA)

Add an Object Storage Target

See Add an Object Storage Target for instructions to add targets/volumes. Each target must already be connected to its server (or two servers in HA configurations).

5.2 Add an object storage target to a managed file system

To add another object storage target:

1. At the menu bar, click the **Configuration** drop-down menu and click **File Systems**.
2. In the file system table displayed, *click on the file system name* to display the file system window.
3. Under *Object Storage Targets*, click **+ Create new OST**.
4. Each available target device is displayed, with its Capacity, Type, HA-status, and server pair, if configured. Select the OST or OSTs to be added and click **OK**. The new OSTs will be displayed in the table of OSTs for the file system.

Note: Intel® Manager for Lustre* software will automatically assign OST indices in a distributed fashion across servers to permit striping.

5.3 Start, stop, or remove a file system

To start a file system:

1. At the menu bar, click the **Configuration** drop-down menu and click **File Systems**.
2. In the table entry for the file system, on the far right, click the **Actions** drop-down menu and click **Start**. The metadata and object store targets are started, enabling the file system to be mounted by clients.

To stop a file system:

1. At the menu bar, click the **Configuration** drop-down menu and click **File Systems**.
2. In the table entry for the file system, on the far right, click the **Actions** drop-down menu and click **Stop** to stop the metadata and object store targets. This action makes the file system unavailable to clients. Click **Confirm** to complete this action.

To remove a file system from the manager:

1. At the menu bar, click the **Configuration** drop-down menu and click **File Systems**.
2. To remove a file system, in the table entry for the file system, click the **Actions** drop-down menu and click **Remove**. File system contents will remain intact until volumes are re-used in another file system. Click **Confirm** to complete this action.

5.4 Start or stop an MGT, MDT, or OST

To start or stop a target:

1. At the menu bar, click the Configuration drop-down menu and click File Systems.
2. In the File System column, click the name of the file system in which the target is located. The file system window is displayed.
3. Under Management Target, Metadata Target or Object Storage Targets, locate the target name in the first column.
4. At the far right, click the Actions drop-down menu and click Start or Stop for that target. Note that Stop is only available if the server is running; Start is only available if the server is stopped. Click Confirm to complete this action.

Notes:

- When an MGT is stopped, clients are unable to make new connections to file systems using the MGT, but the MDT and OSTs stay up if they are currently running.
- When an MDT is stopped, the file system becomes inoperable until the MDT is started again.
- When an OST is stopped, clients are unable to access the files stored on this OST. Other OSTs on other servers are not affected.

5.5 Remove an OST from a file system

To remove an OST from a file system:

Caution: Upon removing an OST from a file system, the OST is no longer visible in the manager GUI. **When an OST is removed, files stored on the OST are no longer accessible**. To preserve data, manually create a copy of the data elsewhere before removing the OST.

1. At the menu bar, click the **Configuration** drop-down menu and click **File Systems**.
2. In the *File System* column, click the name of the file system in which the target is located. The file system window is displayed.
3. Under *Object Storage Targets*, locate the target name in the first column.
4. At the far right, click the **Actions** drop-down menu and click **Remove**. Click **Confirm** to complete this action.

5.6 Perform a single target failover from primary to secondary server

To force a manual failover of a storage target from its primary server to its secondary server, perform these steps:

1. At the menu bar, click the **Configuration** drop-down menu and click **File Systems**.

2. Select the file system to be modified.
3. In the entry for the target to be failed over, click the **Actions** drop-down menu and select **Failover**.

Note: The Failover button will be displayed only for targets that are configured for failover.

To initiate failover of a target from its primary server to its secondary server using the command line interface (CLI), enter:

```
$ chroma target-failover <target name, e.g. lustre-OST0000>
```

5.7 Perform a single target failback from secondary to primary server

To force a manual failback of a target to its primary server, perform these steps:

1. At the menu bar, click the **Configuration** drop-down menu and click **File Systems**.
2. Select the file system to be modified.
3. In the entry for the target to be failed back, click the **Actions** drop-down menu and select **Failback**.

Note: The Failback button will be displayed only for targets that are configured for failover and have failed over from a primary server.

To initiate failback of a target using the CLI, enter:

```
$ chroma target-failback <target name, e.g. lustre-OST0000>
```

5.8 Failover all targets from a primary to a secondary server

Warning: After you have created a Lustre* file system using Intel® Manager for Lustre* software, you should not normally make any configuration changes outside of Intel® Manager for Lustre* software, to file system servers, their respective targets, or network connectivity. Doing so will diminish or defeat the ability of Intel® Manager for Lustre* software to monitor or manage the file system, and will make all or portions of the file system unavailable to clients.

The following process is recommended only if the affected server has become unresponsive to commands and must be powered down or power cycled in order to perform recovery. This requires that you first failover all connected targets from the affected server to the secondary server.

Under normal circumstances, where the server is otherwise responsive, use the method described in the previous section (Perform a single target failback from secondary to primary server) to failover the targets before removing power from the host.

To manually failover all targets from a primary to secondary server, perform these steps:

1. At the menu bar, click the **Configuration** drop-down menu and click **Servers**.
2. For the primary server on which you want to perform maintenance, click the **Actions** drop-down and select **Power-Off**. Note that this action is visible only if PDUs have been added and outlets assigned to servers. This action will switch power off for this server. Any targets running on the primary server will be failed-over to the secondary server. Non-HA-capable targets (targets not supported by a secondary server) will be unavailable until power for the server is switched on again.

5.9 Handling network address changes

File system targets use a network address or network ID (NID) to refer to the server with which they are associated. A storage server NID may change if the network connecting the storage servers and clients is modified. If a server NID changes, the server NID record in the manager must be updated.

Note*:* The manager software detects and displays an alert "*NIDs changed on server %s*" for each server on which the NID has changed.

Caution: This procedure stops the file system and erases the configuration logs so that they will be regenerated when the servers are restarted.

To prompt the manager software to detect new NIDs and update file system targets:

1. At the menu bar, click the **Configuration** drop-down menu and click **Servers**.
2. Click **Re-write target configuration**.
3. A new column appears on the far right: *Select Server*. All servers are selected by default. Select the servers for which you want to rewrite NIDs. Then click **Re-write Target Configuration**.
4. The manager queries the network interfaces on the storage servers. Each target is updated with the current NID for the server with which it is associated. To check that the manager has detected the correct NID for a server, click the *Hostname* of the server to display a detailed view of the server. Scroll down to the NID Configuration section to view the network interface, IP Address, driver, and network for each NID.

You can also directly edit the NID configuration for a server, but to do this, the server *cannot belong to an existing Lustre file system*. See NID Configuration.

WARNING: For Lustre* file systems created and managed by Intel® Manager for Lustre* software, the only supported command line interface is the CLI provided by Intel® Manager for Lustre* software. Modifying such a Lustre file system manually from a UNIX shell will interfere with the ability of Intel® Manager for Lustre* software to manage and monitor the file system.

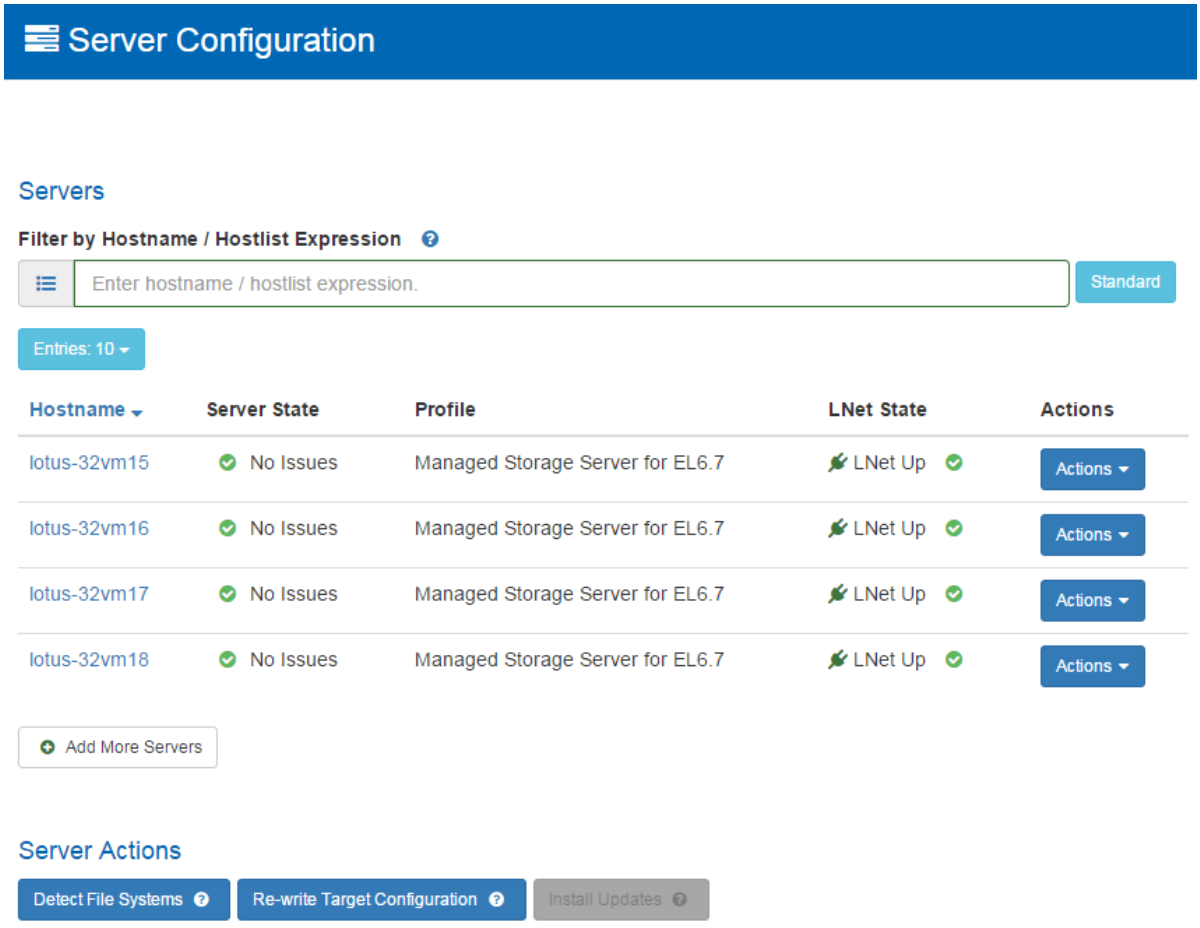
Lustre commands can, however, be used to manage metadata or object storage servers in

an existing Lustre storage system that has been set up *outside* the manager and is being monitored, *but not managed,* by Intel® Manager for Lustre* software.

5.10 Reboot, power-off, or remove a server

The **Actions** menu on the Server Configuration window let you perform the following commands for any single server. Some commands shown here may not be listed, depending on the state of the server.

- **Reboot**
- **Shutdown**
- **Power off**
- **Power cycle**
- **Remove**
- **Force Remove**



To perform any of these commands:

1. On the Dashboard, click **Configuration** > **Servers**.
2. For the desired server, at the **Actions** drop-down menu at the right, click the desired command.
3. After clicking on a command, a *Commands* window pops up to reveal the jobs that are run to perform this command, and the command Status shows pending. When each job completes, the command *Status* then shows *Succeeded*.

5.11 Reconfiguring Corosync and Pacemaker for a server

Pacemaker and Corosync configuration is required for each server if you are creating or expanding a high-availability file system. Intel® Manager for Lustre* software automatically configures Corosync and Pacemaker for each managed HA server that you add, so that manual configuration of Pacemaker and Corosync should not normally be required. See Add one or more HA servers.

An administrator may need to reset or configure Pacemaker or Corosync when performing maintenance on a server, altering the server's configuration, or troubleshooting problems with those services. See Pacemaker configuration and Corosync configuration for more information.

5.12 Reconfiguring NIDs for a server

Intel® Manager for Lustre* software automatically configures NIDs for each managed server that you add, so that manual NID configuration should not normally be required. However, an administrator may need to reconfigure NIDs for a server when performing maintenance on a server, altering the server's configuration, or troubleshooting problems network interfaces. See NID Configuration.

5.13 Decommission a server for an MGT, MDT, or OST

Caution: When a server for an MGT, MDT or OST, is removed (decommissioned), any file systems or targets that rely on this server will also be removed.

To remove (decommission) a server for an OST:

1. At the menu bar, click the **Configuration** drop-down menu and click **File Systems**.
2. In the *File System* column, click the name of the file system in which the target is located. The file system window is displayed.

3. Under *Object Storage Targets*, locate the target you want to decommission. Click the corresponding *Primary server* or *Failover server* that you want to decommission.
4. In the dialogue window that opens, click the **Actions** drop-down menu for that server and click **Remove** to remove the server from the file system. Click **Confirm** to perform this action.
5. Click the **Actions** drop-down menu and click **Stop LNet** to shut down the LNet networking layer and stop any targets running on this server.
6. Click the **Actions** drop-down menu and click **Unload LNet** to stop LNet, if it is running, and unload the LNet kernel module to ensure that it will be reloaded before any targets are started again. (Clicking **Start LNet** will reload the LNet kernel module and start the LNet networking layer again.)

Note: To remove the record for the server from the manager without attempting to contact the server, click the **Actions** drop-down menu and click **Force Remove**. Any targets that depend on this server will also be removed without any attempt to unconfigure them. **This action should only be used if the server is permanently unavailable**.

Warning: The **Force Remove** command will remove the server from the Intel® Manager for Lustre* configuration, but not remove Intel® Manager for Lustre* software from the server. All targets that depend on the server will also be removed without any attempt to unconfigure them. To completely remove the Intel® Manager for Lustre* software from the server (allowing it to be added to another Lustre file system), first contact technical support.

Note: Each server is also separately listed at **Configuration** > **Servers**, however the server configuration regarding which file system, target, and HA status is not shown on the *Servers* window.

5.14 Removing an unwanted server profile

You may want to remove a custom server profile that is no longer needed. Enter the following commands at the command line interface. Dollar sign prompts have been removed to allow copying and pasting.

```
chroma-config stop
chroma server_profile list  # Pick up the profile name you want to remove
chroma-config profile delete <profile name>
chroma server profile list  # Verify profile is now removed
chroma-config start
```

6 Configuring and using Hierarchical Storage Management

Hierarchical Storage Management (HSM) can help provide a cost-effective storage platform that balances performance and capacity. With HSM, storage systems are organized into tiers. The high-performance, primary tier is on the shortest path to the systems where applications are running and where the most data is generated and consumed. As the high-performance tier fills, data that is not being as actively accessed is migrated to lower-cost, higher-capacity storage archive for long-term retention. Data migration is generally managed automatically and transparently to users.

Intel® Enterprise Edition for Lustre*software provides a framework for incorporating HSM into a Lustre file system. When a new file is created, a replica is made on the associated HSM archive tier, so that initially, two copies of the file exist. As changes are made to the file, these are replicated onto the archive copy as well. As the available capacity is consumed on the high-performance tier, the least-frequently-used files are deleted from that tier and each file is replaced with stub file that points to the archive copy. Applications are not aware of the locality of a file. Applications do not need to be re-written to work with data stored on an HSM system. If a system call is made to open a file that has been deleted from the high-performance tier, the HSM software automatically dispatches a request to retrieve the file from the archive and restore it to the high-performance tier.

The HSM framework included with Intel® EE for Lustre* software includes the following components:

- An Agent: A Lustre client that runs an instance of Copytool to transfer certain files between the Lustre file system and the archive, and deletes from the Lustre file system those files that have been archived. There can be one instance of Copytool per agent.
- A POSIX Copytool: This is a reference implementation included in Lustre 2.5 and later. The copytool actually performs the data transfer between the file system and the archive.
- The HSM Coordinator: The HSM Coordinator gathers all archive requests and dispatches them to Agents. The HSM Coordinator thread coordinates HSM activities. (Some documents refer to this as the MDT Coordinator.)
- The Robinhood policy engine: Robinhood enables full automation of HSM activities. Robinhood lets an create file archiving policies based on the file class, as defined by file size, path, owner, age, extended attributes (xattrs), least-recently used, and other criteria. Multiple rules can be combined with Boolean operators. After copying files to archive, automatic file system purging can be set to occur based on the amount the

percentage of consumed file system capacity, file classes, etc. Robinhood can also be used to generate reports, and create packages.

Note: Robinhood is *not necessary to provide basic HSM capabilities* and this HSM framework as installed does not define Robinhood policies. Note that the Robinhood policy engine server requires a supported RDBMS.

Configure basic HSM capabilities for a Lustre file system

Perform these tasks to configure basic HSM capabilities for a Lustre file system.

- Add an HSM Agent node
- Configure power control (optional): HSM Agent nodes are NOT configured for high-availability with Pacemaker and Corosync. The HSM Coordinator schedules HSM tasks with multiple copytools, and if a copytool goes offline, the HSM Coordinator will assign HSM activities to the remaining copytool(s). However, you *can* configure power control, so that an HSM Agent can be power-controlled from the Intel® Manager for Lustre* GUI. You can either configure power distribution units (PDUs), or baseboard management controllers (BMCs) to control power to HSM Agent nodes.
  - To configure PDUs, see Add power distribution units and Assign PDU outlets to servers.
  - To configure IMPI/BMCs, see Assign BMCs to servers.
- Add a copytool to an HSM Agent node
- For an overview of manual HSM tasks, see Using HSM.

Add a Robinhood policy engine server

Robinhood can automate HSM activities. The section linked here discusses adding a Robinhood policy engine server, but does not discuss configuring Robinhood for HSM automation. For more information, see the related guide: *Hierarchical Storage Management Configuration*.

- Add a Robinhood policy engine server

6.1 Add an HSM Agent node

If you plan to enable Hierarchical Storage Management (HSM), perform the following procedures to create an HSM Agent node.

To add a copytool instance to an *existing* HSM Agent node, see Add a Copytool to an HSM Agent node.

Add an HSM Agent node:

1. Perform the steps under Add one or more servers. In that procedure, when selecting the server profile, select **POSIX HSM Agent Node**.
2. When you have added the server(s), perform the procedure in Add a Copytool to an HSM Agent node.
3. After the copytool has been added to the HSM Agent node, see Using HSM.

6.2 Add a Copytool to an HSM Agent node

1. At the menu bar, click the **Configuration** drop-down menu and click **HSM**.
2. At the bottom of the window, click **+ Add Copytool**.
3. At the Add Copytool form, set the following fields:
   a. **File system**: Specify file system for which this copytool will perform HSM actions.
   b. **Worker**: This is the POSIX HSM Agent node that you configured in Add an HSM Agent node. Each copytool instance has its own Agent node, so there may be several. Note that copytool is multi-threaded, so it is able to support multiple simultaneous HSM operations.
   c. **Path to the HSM agent binary**: The file system path to the copytool binary on the worker. For the POSIX copytool provided with Intel® EE for Lustre* software, the path is /usr/sbin/lhsmtool_posix). This was installed on the agent when you configured the HSM Agent node. If another copytool has been installed, it likely resides at another location.
   d. **HSM agent arguments (optional)**: This is a vendor-specific list of copytool arguments. Consult your HSM vendor documentation for the applicable arguments.

      Note: Do not provide any flags that will cause the copytool process to be run in the background (e.g. --daemon); this interferes with the ability of Intel® Manager for Lustre* software to control and monitor the copytool process.
   e. **File system mount point**: The file system mount point on the worker node. Copytool instances require client access to their associated file system.
   f. **Archive number**: The storage back-end number. Change this number only if your site policies require multiple storage back-ends. If there is only one archive available for the file system, set the archive number to "1" (the default). For more information, consult the "Lustre Operations Manual", Section 22.3.1: Archive ID, multiple back-ends.
4. To commit this configuration, click **Save**.

See Start the Copytool.

6.3 Start the Copytool

When a copytool is added to an Intel®EE for Lustre file system configuration, it is not automatically activated. Instead, the copytool will initially be set to Unconfigured. The configuration exists inside the Intel® Manager for Lustre* database but it has not been applied directly to the target HSM Agent.

To configure and launch the copytool on an HSM Agent:

1. Click the **Configuration** menu select **HSM**.
2. Locate the copytool instance in the Copytools table.
3. For the desired copytool, click the **Actions** drop down menu and select **Start**. The copytool status will change from Unconfigured to Idle and the graph will register that a new idle copytool instance has been added and is running on the file system.

As soon as copytool services are requested, the copytool worker will respond. See HSM window for more information.

6.4 Using HSM

After configuring the Copytool Agent node and adding Copytool to that agent, you can use HSM to manage file archiving, free-up file system storage, and improve overall file system performance.

1. To use HSM, log into a regular Lustre client node as the system superuser. The node is a compute client node not managed by Intel(R) Manager for Lustre software.
2. Issue lfs commands to initiate HSM actions (archive, restore, release, remove).
   For example: `root@client1234 #: lfs hsm_archive /mnt/lustre/path/to/big_file`
3. After issuing this archive command, the superuser can monitor progress on the operation at the Intel® Manager for Lustre* GUI. To monitor progress, click **Configuration and click HSM** to open the HSM window and observe copytool archive progress.
4. After the archive operation has completed, you can release command to remove the file from the Lustre file system and free up that space.
   For example: `root@client1234 #: lfs hsm_release /mnt/lustre/path/to/big_file`

After this command completes, the file's data exists in the HSM archive, but the file has been moved off of Lustre main storage. You may notice that the available space in the lustre file system has increased (if the file is large enough and the file system small enough - otherwise the change won't register in the graphs).

If you want the file to be copied back to the file system, issue an lfs restore command (below). Or simply wait for the next read attempt of that file by a client, and an implicit restore will return the file back to the file system.

Following are lfs hsm commands:

- `lfs hsm_archive /mnt/lustre/<path>/<filename>` - Copies the file to the archive.
- `lfs hsm_release /mnt/lustre/<path>/<filename>` - Removes the file from the Lustre file system; does not affect the archived file.
- `lfs hsm_restore /mnt/lustre/<path>/<filename>` - Restores the archived file back to the Lustre file system. This is an asynchronous, non-blocking restore. A client's request to access an archived file will also restore the file back the Lustre file system if is has been released; this will be a synchronous and blocking restore.
- `lfs hsm_cancel /mnt/lustre/<path>/<filename>` - Cancels an lfs_hsm command that is underway.

Displaying information about a current lfs_hsm request

To view the progress of HSM copytool activities, click **Configuration** and click **HSM** to open the HSM window and observe copytool progress. See Monitor HSM Copytool activities for more information.

The command `lctl get_param mdt.*.hsm. also requests` returns information about the currently executing HSM request.

6.5 Add a Robinhood Policy Engine server

The Robinhood policy engine can be used to automate HSM activities. Each instance of Robinhood and its RDBMS supports a single file system. A single server can support multiple instances of Robinhood. The following procedure adds a Robinhood server, however configuring policies are not discussed herein. See the implementation guide Hierarchical Storage Management Configuration Guide for more information.

To add a Robinhood policy engine server, perform the steps under Add one or more servers. In that procedure, when selecting the server profile, select **Robinhood Policy Engine Server**. For an overview, see Configuring and using Hierarchical Storage Management.

Creating Policies

Robinhood lets an superuser create file-archiving policies based on the file class, as defined by file size, path, owner, age, extended attributes (xattrs), least-recently used,

and other criteria. Multiple rules can be combined with Boolean operators. After copying files to archive, automatic file system purging can be set to occur based on the percentage of consumed file system capacity, file classes, etc. Robinhood can also be used to generate reports and create packages. See the implementation guide *Hierarchical Storage Management Configuration Guide* for more information.

7 Detecting and monitoring existing Lustre file systems

A Lustre file system that was created without using Intel® Manager for Lustre* software can be monitored, *but not managed*, from the manager GUI.

Before an existing Lustre file system can be monitored at the manager GUI, the servers must be added and then the file system detected by the manager.

- Detect file system
- Add OSTs and OSSs to a monitored file system

7.1 Detect file system

To make the Lustre file system appear on the Dashboard and Configuration > File System windows in the manager GUI, complete these steps:

1. At the menu bar, click **Configuration** > **Servers** and click **Detect File Systems**. A dialogue window listing hosts is displayed.
2. Select ALL of the servers on which the targets for the file system to be detected are running. Do this for the MGS and all OSSs for this file system, including those OSSs that were already present in this file system. Do NOT add servers that you don't want to add to this file system.
3. Click **Run**. A *Command* detail dialogue window appears showing progress. *Status* shows *Successful* when the process is complete.

Note: Due to a known issue, the software may report that the file system you added was not detected. However, you can confirm the creation of the file system.

You can add more servers and add more targets to an existing monitor-only file system. To do this, proceed to Add servers to be monitored only. Then, *you must detect the entire file system again*, using the steps above.

Note: To be detected, the file system must be running.

Note: To view the *Command Detail* after detection completes, click on **Notifications** on the left side of the screen and select **Commands** at the bottom of the notifications list. To the right of the Detecting file systems command, click **Open**.

The Lustre file system is now ready to be monitored at the manager GUI.

7.2 Add OSTs and OSSs to a monitored file system

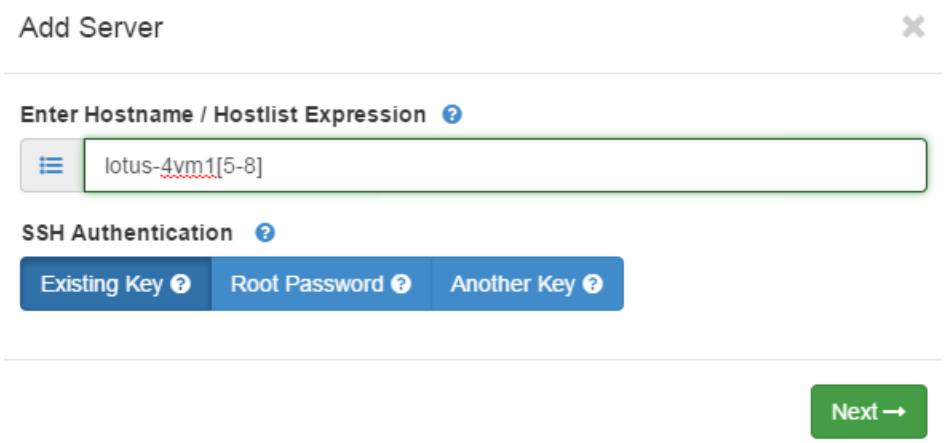
This procedure applies to an existing file system that is monitored only.

To begin, first add the new OSS and OST(s) to your Lustre file system via the command line. See the *Lustre Operations Manual*, for detailed instructions for adding and OSS and OST(s) to an existing file system. Then, to add one or more servers hosting the new OST(s) perform the following steps.

1. At the menu bar, click the **Configuration** drop-down menu and click **Servers** to display the *Servers Configuration* window.

   Note: All authentication credentials are sent to the manager server via SSL and are not saved to any disk.

2. Click **+ Add Servers.**



3. In the *Hostname / Hostlist Expression* field, enter the name of the server(s) to be added. You can enter a range of names, a "host list expression". For example, your can enter server[00-019] to generate a list of up to twenty servers (in this case). **Note**: These are all the server names that your expression expands to and may include names for servers that don't exist or are not connected to the network.
4. Select an authentication method:
   - Click **Existing Key** to use an existing SSH private key present on this server. There must be a corresponding SSH public key on each server you are adding.
   - Click **Root Password** and enter a root password for the server you are adding. This is standard password-based authentication. It is not uncommon to use the same root password for multiple servers.
   - Click **Another Key** and enter a private key that corresponds to a public key present on the server you are adding. If the key is encrypted, enter the passphrase needed to decrypt the private key.
5. Click **Next**. The software will attempt to verify the presence and readiness of all

servers with names matching your hostname entry. Each server is represented by a square. A green square means that the server passed all readiness tests required for validation and this process can proceed for that server. A red box means that the server failed one or more readiness tests. Click on a red box to learn which tests the server failed. You can hover the pointer over the failed validation test to learn more.

6. For a server that failed validation, log into that server and work to address the failed validation. When the issue has been resolved, the GUI will update the failed validation test in real time, from a red x to green check mark. You can add the server when all failed validations are resolved.

   Note: Many server names may be generated from your host list expression, and some of those servers may not exist. A red box is created for each server that doesn't exist.

7. Assuming that all servers pass the validation tests and all boxes are green, click **Proceed** to download agent software to each server. If one or more servers failed to pass validation tests, the green **Proceed** button changes to a yellow **Override** button. Clicking **Override** displays this warning: *You are about to add one or more severs that failed validation. Adding severs that failed validation is unsupported. Click **Proceed** to continue.*

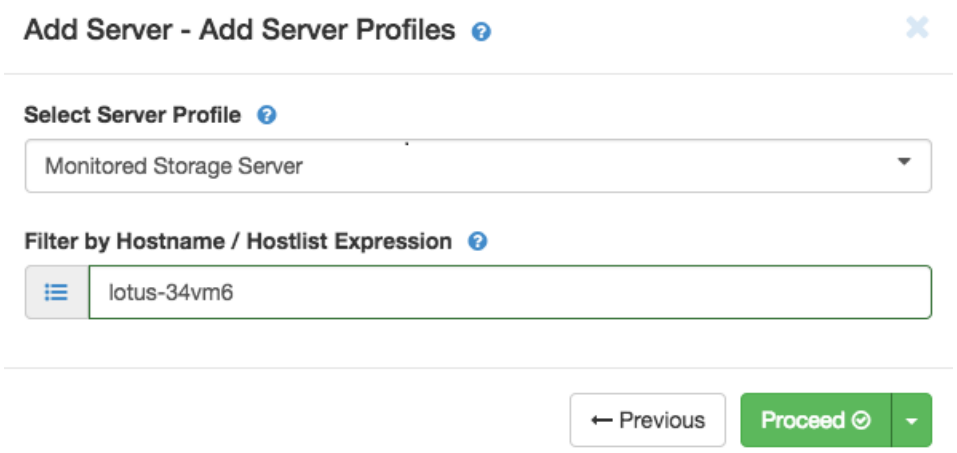
   Caution: Although you can attempt to add a server that has failed validation, all of the capabilities exercised by the tests are needed for the management software and server to operate normally. The server will likely fail to operate normally. Adding a server that failed validation is not supported.

8. After clicking **Proceed**, agent software is deployed to each server and a *Commands* window opens to show progress. Click **Close** to close this *Commands* window.

9. If you decided to override servers that failed validation tests (not supported), expand any failed commands in the Commands window. Click on any failed jobs and examine the stack trace to learn the cause of the failure. Correct the cause of the failure and close the command and server windows. If the server exists in the server table, click **Actions** and select **Deploy Agent**. Otherwise open the Add server dialog and enter the failed server. In either case you should now see a green square for that server and be able to add it without issue.

10. The next task is to add a server profile to each server. Here you select the desired profile from the drop-down menu. Note that *one profile type* is selected for all servers you are adding in this process.

    Select **Monitored storage server**: This is for servers that are not configured for HA/ failover (as far as this software is concerned). A *monitored storage server* is monitored only; the manager GUI performs no such server HA configuration or management. However the Dashboard will still display charts showing file system operations. In the image below, the Hostname is an example only.



11. Click **Proceed**. The manager does an audit of the storage resources on each server. The manager then provisions the server by loading appropriate Lustre modules and making sure the Lustre networking layer is functioning. When all checks are completed, *LNet State* indicates *LNet Up* and each server is fully qualified as a Lustre server. Under the *Status* column, a green check mark is displayed for each new server. If server provisioning does not succeed, the *Status* will indicate a exclamation mark (!) and the *LNet State* may indicate *Unconfigured*. To learn the cause of the problem, click the exclamation mark for the failed server to see *Alerts*. For more information, click **Status** at the top menu bar. The *Status* window also lets you view related logs.
12. You can proceed to add more servers. Otherwise, click **Close**.
13. When all the servers for a monitor-only file system have been added and configured using the manager GUI, at the menu bar, click **Configuration** > **Servers** and click **Detect File Systems**. A windows shows all detected hosts.
14. Select ALL of the servers on which the targets for the file system to be detected are running. Do this for all OSSs for this file system, including those that were already present in this file system. Do NOT add servers that you don't want to add to this file system.
15. Click **Run**. A Command detail dialogue window appears showing progress. Status shows “Successful” when the process is complete.

Note: Due to a known issue, the software may report that the file system you added was not detected. However, if you go to **Configuration** > **File Systems** and view the updated file system, the new OSS(S) and target(s) should be listed.

8 Creating and Managing ZFS-based Lustre file systems

Intel® Manager for Lustre* software is able to create and manage Lustre file systems that are based on OpenZFS object storage device (OSD) volumes. The software installs the necessary packages, formats Lustre targets from ZFS pools, and creates the high-

availability software framework for managing availability for Lustre + ZFS servers. The following topics are covered:

- Create a ZFS-based Lustre file system
- Importing and exporting ZFS pools in a shared-storage high-availability cluster
- Removing a ZFS-based Lustre file system
- Destroy an individual zpool
- Destroy all of the ZFS pools in a shared-storage high-availability cluster

8.1 Create a ZFS-based Lustre file system

The procedures in this section assume that you have first assembled and configured the physical hardware: servers, storage devices, network interfaces, etc., and installed Intel® Manager for Lustre* software as instructed in the *Intel® Enterprise Edition for Lustre* Software Installation Guide.*

To create and manage an OpenZFS-based Lustre file system that is highly-available and managed by Intel® Manager for Lustre* software, perform these steps:

1. Add all of the physical servers that will comprise your ZFS-based Lustre file system. To do this, perform the steps in Add one or more HA servers. Add each server as a *Managed Storage Server*.

Note: Steps 2 and 3 below are performed automatically by Intel® Manager for Lustre* software and do not need to be performed. They are included here for topic coverage only. Continue with Step 4.

2. Having added the servers, you now need to ensure that the server hostids are set before creating any zpools. Each server requires a unique hostid to be configured. Setting the hostid on each server will allow ZFS to apply an attribute on each zpool indicating which host currently has the pool imported. This provides some protection in ZFS against multiple simultaneous imports of zpools when the storage is connected to more than one server. To set the hostid, run the `genhostid` command on each host and reboot. In the document *Lustre* Installation and Configuration using Intel® EE for Lustre* Software and OpenZFS*, see the section *Protecting File System Volumes from Concurrent Access* for more information.
3. As a further protection against "double-importing" ZFS pools, and to prevent conflicts with the Pacemaker resource management software, disable the ZFS Systemd target by entering the following command:

```
systemctl disable zfs.target
```

This will stop the operating system from attempting to auto-import ZFS storage pools during system boot. Disabling the ZFS target affect all ZFS storage pools,

including any that are not being used for Lustre.

4. ZFS pool configuration is executed directly on each of the Lustre server HA pairs using the command line tools supplied with the ZFS software. The storage devices for each ZFS pool must be connected to, and accessible from, each Lustre server in the HA pair. For each ZFS pool, it is easiest to nominate a primary server and use that host to create the zpool. Log into each primary host and use this zpool command to create each storage pool. The general syntax for creating a zpool is as follows:

```
zpool create [-f] -O canmount=off \
-O mountpoint=none \
[ -o <option> ] \
-o cachefile=none \
-o failmode=panic \
<zpool name> <zpool specification>
```

ZFS is highly configurable and there are a wide range of optional parameters that can be applied to a ZFS pool, to improve performance or to modify functionality. The most important of these are as follows:

- cachefile=none: ZFS uses the cachefile to identify pools it can automatically import, but this does not make consideration for shared storage cluster environments. To prevent a ZFS pool from being added to the default cachefile and automatically imported at system boot, it is essential that all ZFS pools that are held on shared storage have the cachefile set to the special value “none”. This, in conjunction with setting the hostid provides a lightweight impediment against double-importing the pool.
- failmode=panic: zpools should have the "failmode" parameter set to "panic". This will cause a host that has lost connectivity to the underlying storage devices, or has experienced too many device failures, to shut down immediately. The node is thus prevented from issuing undesired writes to a zpool after catastrophic failures, including host disconnection from the zpool. When ZFS pools are operating within an HA cluster framework, panicking a node will also trigger a failover event in the cluster, allowing the zpools to be imported to a standby node, and services to continue operation.
- ashift: used to override ZFS auto-detection of the physical vdev storage sector size. Advanced format drives are known to cause issues by incorrectly reporting the native sector size. Setting the ashift property forces ZFS to use a specific sector size, and can yield improved performance. The value of ashift is an exponent to power of 2. Typically used to set 4KB sector size, (ashift=12).

In addition to zpool properties, there are also properties that can be applied to the ZFS datasets in a pool. ZFS dataset properties are set in the zpool command using a capital letter -O, rather than the lower case letter -o flag used for the cachefile and

ashift settings. The -O flag is used to set properties to the root file system data set in the pool, rather than for properties of the pool itself. Two ZFS dataset properties of particular importance are highlighted here:

- canmount=off: prevent the dataset from being mounted using the standard zfs mount command. Essential for Lustre OSDs, which must be mounted and unmounted using the mount.lustre (or mount -t lustre) command.
- recordsize: this sets a suggested block size for the file system. The block size cannot exceed this value. The default value, 128KB, is fine for MGT and MDT OSDs. For OSTs, where large block IO is the dominant workload, set the recordsize=1M, which is the maximum currently allowed in ZFS on Linux version 0.6.5. Future versions of ZFS are expected to be able to further increase this value.
- mountpoint=none: in monitored mode, the management software requires that ZfsDatasets have the following property values to prevent undesired automatic mounting.

For more details, refer to the section ZFS OSDs, in the guide titled Lustre* Installation and Configuration using Intel® EE for Lustre* Software and OpenZFS, as well as the system man pages for zfs(8) and zpool(8).

Following are three *examples* of typical pool configurations, one each for MGT, MDT and OST respectively:

- ZFS pool for MGT (typically a simple mirror):

```
zpool create -O canmount=off \
  -O mountpoint=none \
  -o cachefile=none \
  -o failmode=panic \
  mgspool mirror <dev A> <dev B>
```

- ZFS pool for MDT (typically a stripe of mirrors to maximize performance):

```
zpool create -O canmount=off \
  -O mountpoint=none \
  -o cachefile=none \
  -o failmode=panic \
  <fsname>-mdt<n>pool \
  mirror <dev A> <dev B> [ mirror <dev C> <dev D> ] ...
```

Note: The naming convention for the pool name is intended to reflect the Lustre file system name name (`<fsname>`) and the MDT index number (`<n>` usually 0, unless DNE is used). The number of mirrors used for the MDT depends on the requirements of the installation and can be scaled up accordingly.

- ZFS pool for OST (typically a RAIDZ2 pool, balancing reliability against optimal

capacity):

```
zpool create -O canmount=off \
  -O recordsize=1M \
  -O mountpoint=none \
  -o failmode=panic \
  -o cachefile=none \
  [ -o ashift=12 ] \
  <fsname>-ost<n>pool \
  raidz2 <dev A> <dev B> \
    <dev C> <dev D>  <dev E> <dev F> ...
```

Note: See the document *Lustre* Installation and Configuration using Intel® EE for Lustre* Software and OpenZFS* for descriptions of the `ashift` and `recordsize` properties. RAIDZ2 is the preferred vdev configuration for OSTs, and we recommend an arrangement of at least 11 disks (9+2) per RAIDZ2 vdev for best performance. The pool naming convention is based on the Lustre file system name and OST index number, starting at 0 (zero).

The remainder of this procedure is performed at the Intel® Manager for Lustre* software GUI.

5. For high-availability, configure your servers as primary and fail-over servers for each zpool. Perform the steps in Configure primary and fail-over servers.
6. If you are using power distribution units (PDUs) for power control, then for each server, perform the steps in Add power distribution units. Then perform the steps in Assign PDU outlets to servers for each server.
7. If you are using Baseboard Management Controllers (BMCs) for power control, then perform the steps in Assign BMCs to servers for each server.
8. Perform the steps in Create the new Lustre file system, using the zpools you created in step 4 above as object storage targets (volumes), rather than direct block devices. Each ZFS pool that you created will appear as a target, with the *Type* identified as a **ZfsPool.**

8.2 Importing and exporting ZFS pools in a shared-storage high-availability cluster

ZFS file system datasets are always members of a pool, and the pool must be imported to a host before any action can be taken to alter its content. In a shared-storage, high-availability cluster, some pools may be imported on different hosts. This is a common and recommended practice that balances distribution of Lustre OSDs equitably across the HA cluster nodes.

One must always manage ZFS pools and their contents from the host where the pool is imported, but it is easy to overlook and miss ZFS pools that are exported. Use the

combined output of `zpool list` and `zpool import` to gain a complete list of all storage pools available to a host, and import any pools that require administration.

1. List the currently imported ZFS pools:

```
zpool list
```

2. If any of the pools that are expected to be on this host are missing, it may be that the pools have been exported, or are imported on a different host. Use the `zpool import` command to identify these pools. For example:

```
[root@rh7z-oss1 ~]# zpool import
   pool: demo-ost0pool
     id: 12622396723776112603
  state: ONLINE
 action: The pool can be imported using its name or numeric identifier.
 config:

 demo-ost0pool  ONLINE
   raidz2-0  ONLINE
     sda     ONLINE
     sdb     ONLINE
     sdc     ONLINE
     sdd     ONLINE
     sde     ONLINE
     sdf     ONLINE

   pool: demo-ost1pool
     id: 617459513944251623
  state: ONLINE
 status: The pool was last accessed by another system.
 action: The pool can be imported using its name or numeric identifier
     and the '-f' flag.
   see: http://zfsonlinux.org/msg/ZFS-8000-EY
 config:

 demo-ost1pool                               ONLINE
   raidz2-0                                  ONLINE
     scsi-0QEMU_QEMU_HARDDISK_IEELOST0001    ONLINE
     scsi-0QEMU_QEMU_HARDDISK_IEELOST0007    ONLINE
     scsi-0QEMU_QEMU_HARDDISK_IEELOST0011    ONLINE
     scsi-0QEMU_QEMU_HARDDISK_IEELOST0010    ONLINE
     scsi-0QEMU_QEMU_HARDDISK_IEELOST0009    ONLINE
     scsi-0QEMU_QEMU_HARDDISK_IEELOST0008    ONLINE
```

In the example, there are two exported pools. Note that the second pool in the list has a `status` attribute that is not displayed in the first pool, and that the `action` attribute indicates that the pool can only be imported by force (indicated by the `'-f'` flag). This means that the pool has been imported on a different host. Do not import pools that are in this condition unless it can be absolutely determined that the other host is offline or does not have this pool running (this is commonly the case when a

server crashes or loses power in an unclean shutdown).

3. Use the zpool import command, along with the pool name, to import a pool onto the host:

   ```
   zpool import [-f] <pool name>
   ```

   For example:

   ```
   [root@rh7z-oss1 ~]# zpool import demo-ost0pool
   ```

A simple approach to managing all of the pools of an HA cluster as a batch, is to export all of the pools from each host connected to common shared storage, and then import all of the pools into a single host, as follows:

1. Export all of the pools on a host:

   ```
   zpool export –a
   ```

   Repeat this on every host in the HA framework

2. On one host, import all of the zpools:

   ```
   zpool import -a
   ```

8.3 Removing a ZFS-based Lustre file system

When removing a Lustre file system, the Intel® Manager for Lustre* software does not destroy the data content held on Lustre OSDs. This allows the operator an opportunity to recover data from the OSDs if the file system was removed from IML accidentally. As a consequence, after the file system is removed from the manager, the OSD volumes may need to be cleaned up as a separate operation before they are ready to be re-used.

LDISKFS volumes do not require any special treatment; they can simply be reformatted, either by hand or by re-adding the volumes into a new Lustre file system via Intel® Manager for Lustre* software. The manager software will detect the pre-existing Lustre data, and ask the user to confirm that the volume is to be re-used.

ZFS OSDs require some additional work. ZFS OSDs are file system datasets inside zpools. After a file system is removed using Intel® Manager for Lustre* software, any ZFS OSDs from that file system will still be present in the ZFS storage pools (zpools), and cannot be re-used directly. To reuse the storage in a zpool, the existing OSD datasets must first be removed, using the `zfs destroy` command. This is described in these sections:

- Destroy an individual zpool
- Destroy all of the ZFS pools in a shared-storage high-availability cluster

8.4 Destroy an individual zpool

Perform the following steps.

1. Run the `zpool list` command to display a list of zpools currently imported on the host.
2. If a pool is missing from the output, it may be imported on a different host or it may be in the exported state. Use the `zpool import` command to identify ZFS pools that have been created, but are not imported on the current host:

   ```
   zpool import
   ```

   If a zpool has a status field that says the pool was last accessed by another system, this pool may still be active on a different host connected to the same shared storage. The output will include fields similar to the following:

   ```
   ...
   status: The pool was last accessed by another system.
   action: The pool can be imported using its name or numeric identifier
   and the '-f' flag.
   ...
   ```

   Don't force an import of a pool in this state unless you are certain that the other host is offline. If possible, either export the zpool from the other host so that it can be imported, or log into the other host to perform the required work.
3. If necessary, use the `zpool import` command, along with the pool name, to import a pool onto the host:

   ```
   zpool import <pool name>
   ```

4. Run the command: `zpool destroy [-f] <zpool name>`

   For example:

   ```
   zpool destroy demo-ost0pool
   ```

5. Confirm that the pool has been removed with `zpool list`.

8.5 Destroy all of the ZFS pools in a shared-storage high-availability cluster

Perform the following steps.

1. Log into each server in the cluster framework that is connected to the same shared storage and export all of the ZFS pools:

   ```
   zpool export –a
   ```

2. On one host, import all of the zpools:

```
zpool import -a
```

3. List the zpools:

```
zpool list
```

4. For each pool, run the zpool destroy command:

```
zpool destroy [-f] <pool name>
```

The '-f' flag will force the destruction of the pool, unmounting any active datasets that may be present.

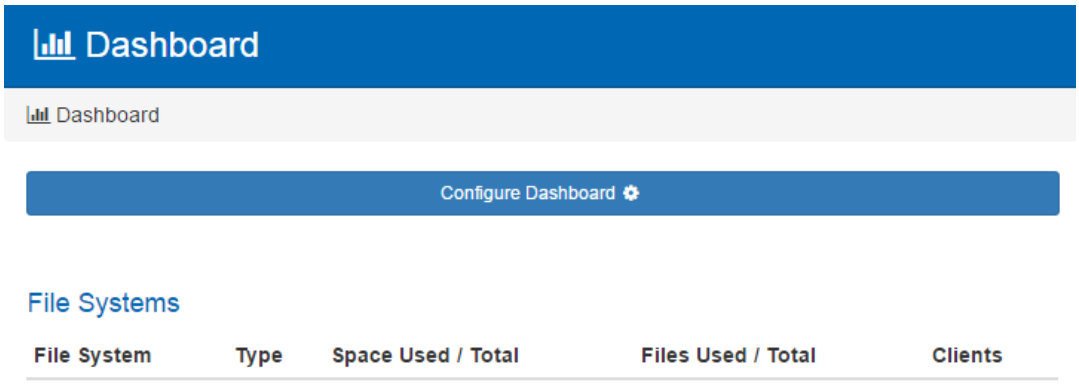
9 Graphical User Interface

This section details the Intel® Manager for Lustre* graphical user interface. Click the desired topic.

- Dashboard window
- Dashboard charts
- Configuration menu
- Job stats
- Logs window
- Status window
- Alert bar
- Resources tree view
- Breadcrumb navigation

9.1 Dashboard window

The Dashboard window is shown next.



The Dashboard displays a set of charts that provide usage and performance data at several levels for each file system. At the top level, this window displays an aggregate view of all file systems you're currently monitoring. You can select to monitor individual file systems and servers by clicking **Configure Dashboard**; See Configuring the Dashboard.

To view charts for OSTs and MST(s), select the specific file system. Then select the desired target(s).

At the top, the Dashboard lists the file system(s) being managed or monitored-only. The following information is provided for each file system:

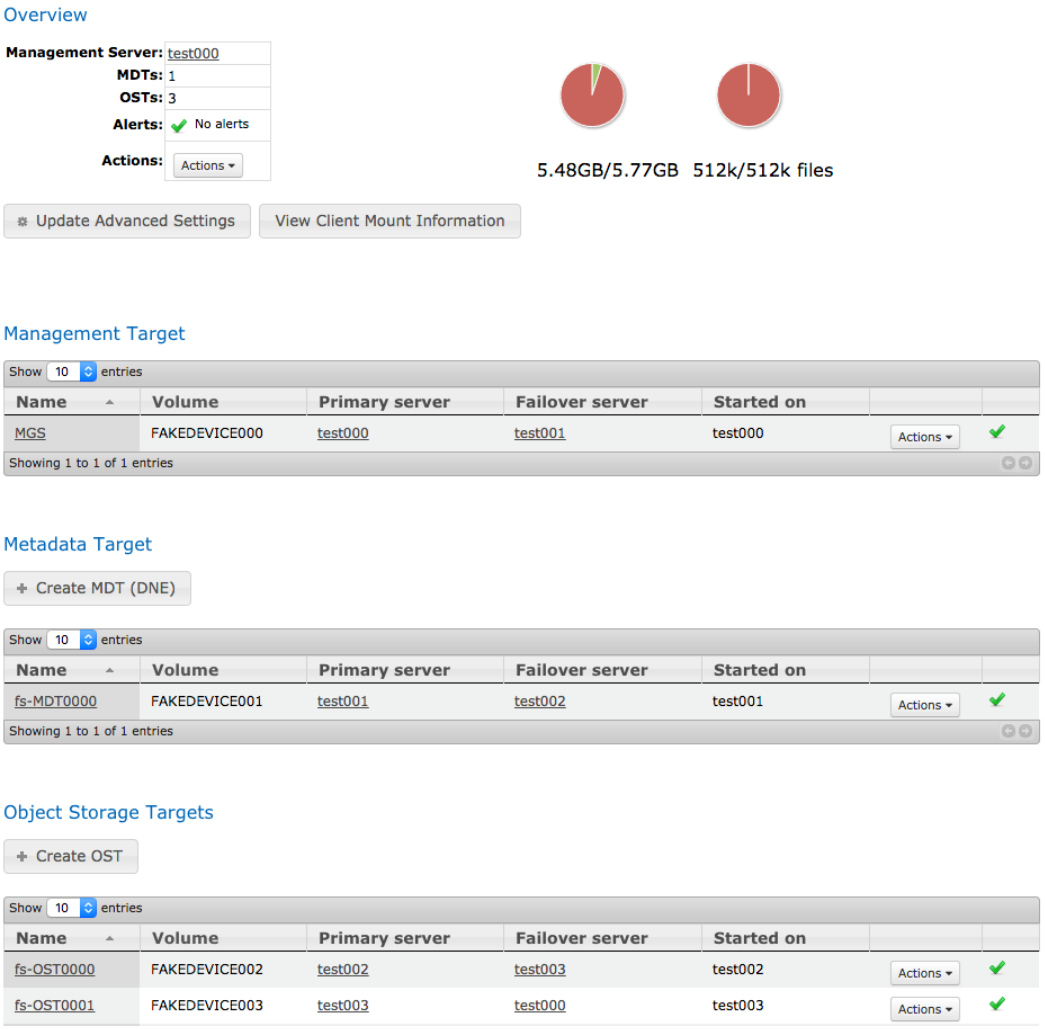
- *File System* name: The name assigned to this file system during its creation on the *Configuration* window.
- *Type*: Monitored or Managed. "Managed" file systems are configured and managed for high availability (HA). Managed file systems are both monitored and managed, whereas "monitored" file systems are monitored-only and do not support failover via Intel® Manager for Lustre* software.
- *Space Used / Total*: This indicates the amount of file system capacity consumed, versus the total file system capacity.
- *Files Used / Total*: This indicates the total number of inodes consumed by file creation versus the total number of inodes established for this file system.
- *Clients*: Indicates the number of clients accessing the file system at this moment.

Data used to produce the charts is saved for long-term use. Data is averaged and compressed over time so that the most recent data is stored and viewed at maximum resolution while aging data is stored and viewed at progressively lower resolutions over time.

9.1.1 File System Details window

After you have created a file system, you can view its configuration and manage the file system at the *File System Details* window.

To access the File System Details window, at the Dashboard, click the name of the file system of interest.



This window identifies the:

- *Management Target* (MGT)
- *Metadata Target* (MDT). There may be more than one MDT.
- *Object Storage Targets*
- Alert status
- Overall file system capacity and free space

This window also identifies the volume(s), primary server(s), and failover server(s) for the

MGS, MDT(s) and all OST(s). From this window you can Update Advanced Settings and View Client Mount Information.

9.1.2 Configuring the Dashboard

By default, the Dashboard displays information and charts for all file systems. Click **Configure Dashboard** to open a window to let you do the following:

- To view a file system's charts: Click **File System** (default). You can view information and charts for all file systems, or select a specific file system from the drop-down menu.
- To view a server's charts: Click **Select Server**. You can view information and charts for all servers (on all file systems), or select a specific server from the drop-down menu.
- To view charts for one or all targets: Click **File System**. Select the desired file system and then select **All Targets** or an individual target.

Click **Update** to apply your choices and **Cancel** to close.

9.2 Dashboard charts

Several Dashboard charts provide quick, detailed, visual representation of the performance of your Lustre file system(s). You can configure certain data display parameters for each chart, and your chart configuration will persist until you reload/ refresh the Dashboard page, using the browser.

Charts are presented as:

- File system charts
- Server charts
- MDT charts
- OST charts

File system charts

The Dashboard window displays the following six charts for one or more file systems:

- Read/Write Heat Map
- OST Balance
- Metadata Operations
- Read/Write Bandwidth

- Metadata Servers
- Object Storage Servers

Server charts

The Dashboard displays the following three charts for an individual server (MDS or OSS). To access, click **Configure Dashboard**. Then select **Servers** and select the desired server.

- Read/Write Bandwidth
- CPU Usage
- Memory Usage

MDT charts

The Dashboard window displays the following three charts for the selected MDT. To access, click **Configure Dashboard**. Then select the specific file system. Lastly, select the desired MDT.

- Metadata Operations
- Space Usage
- File Usage

OST charts

The OST Dashboard window displays the following three charts for the selected OST. To access, click **Configure Dashboard**. Then select the specific file system. Lastly, select the desired OST.

- Read/Write Bandwidth
- Space Usage
- Object Usage

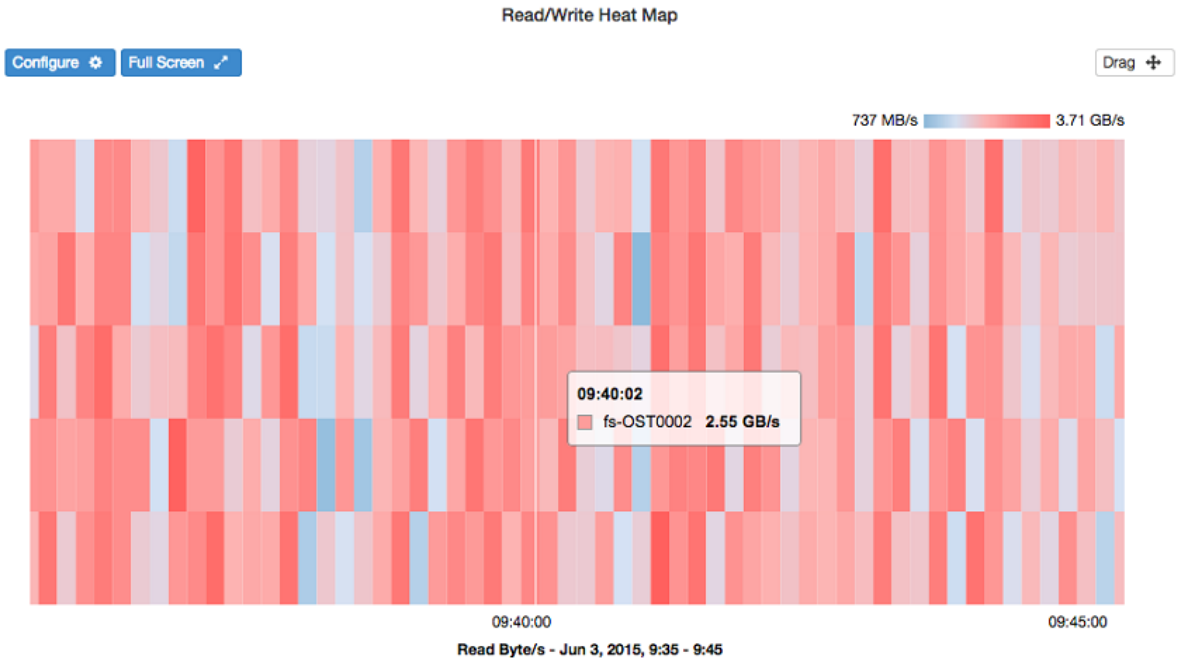
9.2.1 Read/Write Heat Map chart

The Read/Write Heat Map chart shows the level of read or write activity for each OST in all file systems. Each row is a single OST, and each column is a consecutive time sample. The chart updates from right to left, so the most recent sample for any OST is in the right-most column. This chart is displayed when all File Systems are selected on the Dashboard window (default). You can also view this chart for a single file system.

You can the monitor the level of read or write activity for a given OST over time by looking across the chart. Activity is displayed in shades, from light-blue to red. Displayed data transfer rates are not fixed: Light-blue represents the lowest percent of maximum for the

preceding twenty samples, while darkest-red represents the highest percent of maximum and the most read or write activity.

Note: Because of the way that activity information is averaged, the heat map may show slightly different information following a refresh of the display. This is normal.



Features

- Mouse over any cell on the heat map to learn which OST this is, its file system, its read or write activity, and the actual starting date and time of that measurement period.
- Click on a specific heat map cell to open the Job Stats window (job statistics) for that OST and read/write measurement. See View job statistics.
- To better view larger numbers of OSTs, for example, more than forty OSTs, click **Full Screen** to expand the map.

View this chart for a specific file system

This chart is displayed by default for all file systems. To view this chart for a single file system:

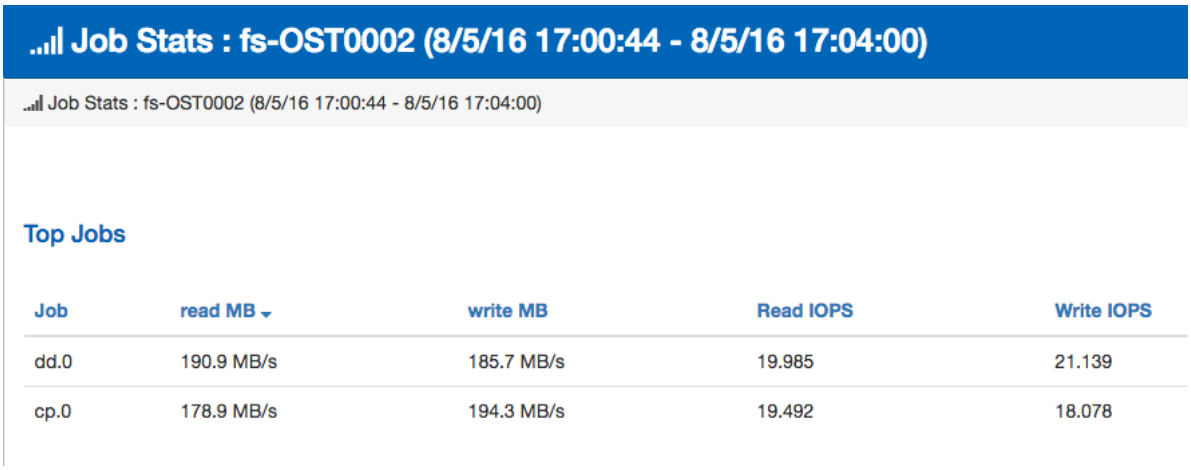
1. On the Dashboard, click **Configure Dashboard**.
2. Select **File System** (default).
3. At the File System drop-down menu, select the file system. Then click **Update.**

Configure the Heat Map chart

1. Click **Configure** to open the configuration window.
2. Click **Set Duration** to set the total time duration to Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4) for the entire map. This is a sliding duration. Based on your selection, the heat map is divided into columns of equal duration. Note that for long durations, the map will be divided over several days, with measurements taken at different times of the day. The value given in each cell is the average for that measurement period. After clicking **Update** to apply changes, the duration of measurements begins immediately.
3. Click **Set Range** to set the **Start** and **End** times and dates over which measurements will be displayed. This view of a heat map is a static snapshot, starting and ending as configured.
4. Click **Select data** to view to select **read bytes**, **write bytes**, **read IOPS**, or **write IOPS**.
5. Click **Update** to close this window and apply changes. Click **Cancel** to close.

Jobs Stats

Job statistics information is accessible from the Read/Write Heat Map chart. Simply click on an OST cell on the chart, and for that OST and time interval, a window opens that shows metrics for the top ten jobs for that OST. Current metrics include average, min, and max for read and write bandwidth and read and write IOPS per the time interval. Because this information is specific to a time period, it is static.

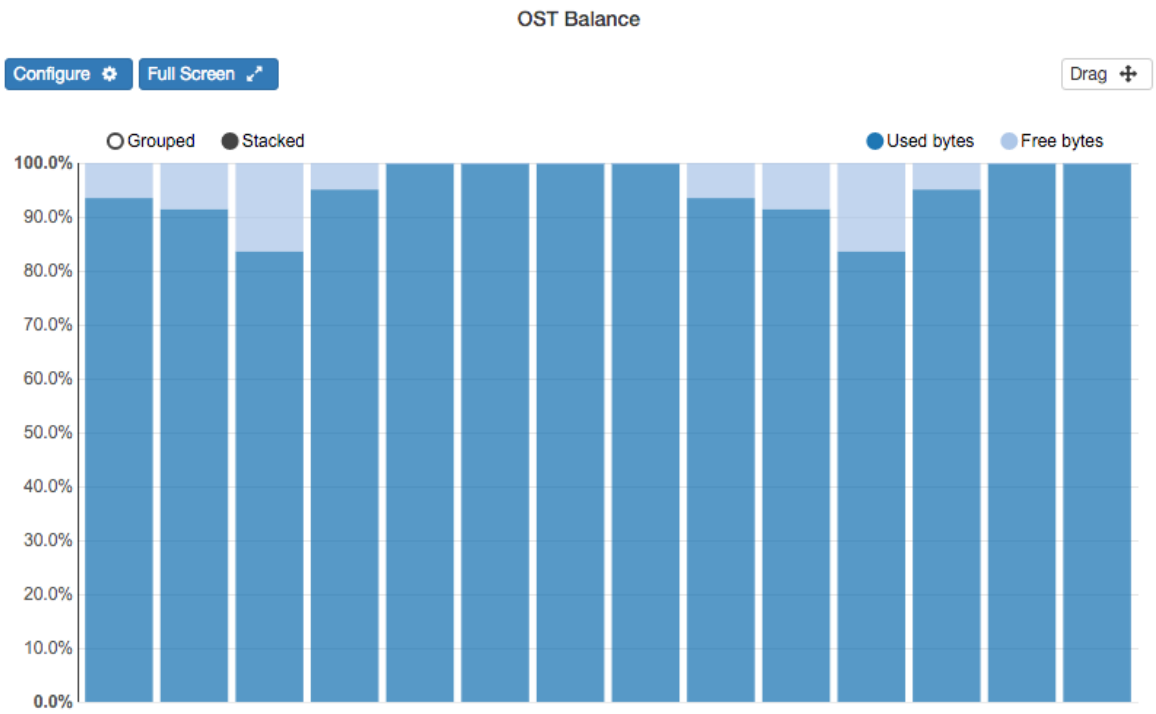


The Jobs Stats window is available for any dashboard window that has a heat map: These are the File Systems dashboard windows and the Servers dashboard window. This feature also supports the creation of plug-ins to display user account, command line, job size, and job start/finish times.

For statistics regarding the top ten jobs for all active file systems, click **Job Stats** at the top menu bar. This view updates in real time, showing a top-like interface of current jobs. Durations and sort-order are customizable.

9.2.2 OST Balance chart

This chart shows the percentage of storage capacity currently consumed for each OST. This chart is displayed when File Systems are selected on the Dashboard window (default). You can also view this chart for a single file system.



Features

- Click **Full Screen** to fill the browser window with this chart. Click Exit Full Screen to return to the normal view.
- Click **Stacked** to arrange the display so that the used and unused capacities are stacked for each OST.
- Click **Grouped** to arrange the display so that the used and unused capacities are shown separately for each OST.

View this chart for a specific file system

This chart is displayed by default for all file systems. To view this chart for a single file

system:

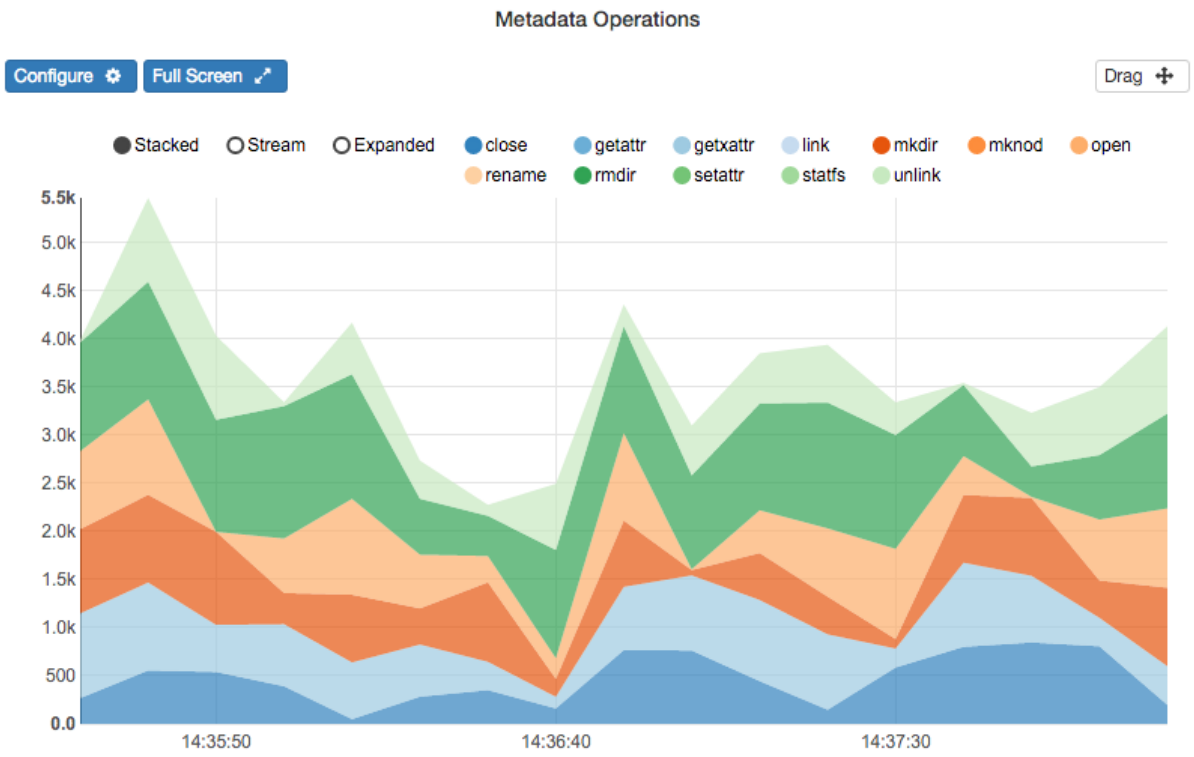
1. On the Dashboard, click **Configure Dashboard**.
2. Select **File System** (default).
3. At the File System drop-down menu, select the file system. Then click **Update.**

Configure the OST Balance chart

1. Click **Configure**:
2. This control lets you filter and display only those OSTs for which their usage (consumed capacity) is equal to or greater than the threshold you set. The default usage is set to zero percent, so that all OSTs are displayed. Set the desired threshold.
3. Click **Update**.

9.2.3 Metadata Operations chart

This chart is shown for file systems and for specific MDTs. The chart shows the number of metadata I/O operations over time, based on command type. These are system calls or commands performed on all file systems. You can also view this chart for a single file system or MDT.



Features

- Click **Full Screen** to fill the browser window with this chart. Click **Exit Full Screen** to return to the normal view.
- Mouse over any point on the chart to learn the values for each system call or command type executing at that time. Values shown vary, based on the chart type: For Stacked and Stream display, values are absolute. For Expanded display, values are relative percentages.
- Click on any area in the chart to display only information for that specific system call or command type. The vertical scale will adjust to better display that information.
- Click the command icons (e.g. **close**, **getattr**, etc.) to display or not display those command types on the chart.
- Click **Stacked** to show all displayed command types stacked, with the command types stacked alphabetically.
- Click **Stream** to display a "stream-graph" of the relative volume of each type of metadata operation. The display of each command-type (or layer) out from the horizontal center-line is ordered, from the least-varying volume to most-varying volume, per command type, over time.
- Click **Expanded** to show the percentage of each command type versus 100%.

View this chart for a specific file system

This chart is displayed by default for all file systems. To view this chart for a single file system:

1. On the Dashboard, click **Configure Dashboard**.
2. Select **File System** (default).
3. At the File System drop-down menu, select the file system. Then click **Update.**

View this chart for a specific MDT

To view this chart for a single MDT:

1. On the Dashboard, click **Configure Dashboard**.
2. Select **Server**.
3. At the **Server** drop-down menu, select the server hosting the desired target.
4. At the **Target** drop-down menu, select the desired MDT. Then click **Update**.

Configure the Metadata Operations chart

1. Click **Configure**.
2. Click **Set Duration** and enter at time period over which samples will be taken. Enter Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the chart will be divided over several days, with samples periods starting at different times of the day. The value given is an average for that sample period.
3. Click **Set Range** to set the **Start** and **End** times and dates over which measurements will be displayed. This view of the chart is a static snapshot, starting and ending as configured.
4. Click **Update** to apply and close this window.

9.2.4 Read/Write Bandwidth chart

The Read/Write Bandwidth chart shows read and write activity on all file systems, all servers one file system, or a specific server, or over time.

Depending on the view selected, the chart notation and display adjusts to occupy the full vertical range of the chart. This chart shows zero read or write operations across the center-line and values greater than zero expanding from the center-line. Read operations are shown above the center line; write operations are shown below the center line. This chart is displayed when File Systems are selected for display (default), or servers, or targets are selected.

Features

- Mouse over any point on the chart to learn the date/time of this measurement and the read and write values at that time.
- Click **Change Duration** to set the total time duration to Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day.
- Click **Full Screen** to fill the browser window with this chart. Click **Exit Full Screen** to return to the normal view.
- Click **Read** or **Write** to view only read or write information on the chart.

View this chart for a specific file system

This chart is displayed by default for all file systems. To view this chart for a single file system:

1. On the Dashboard, click **Configure Dashboard**.
2. Select **File System** (default).
3. At the File System drop-down menu, select the file system. Then click **Update.**

View this chart for a specific OST

To view this chart for a single OST:

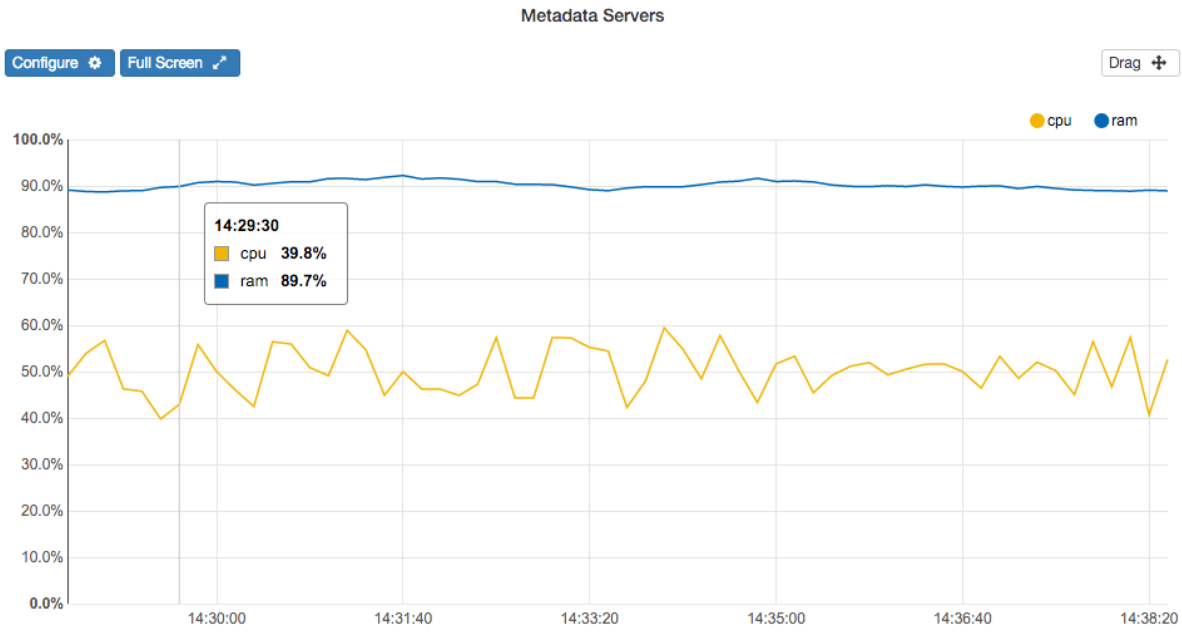
1. On the Dashboard, click **Configure Dashboard**.
2. Select **Server**.
3. At the **Server** drop-down menu, select the server hosting the desired target.
4. At the **Target** drop-down menu, select the desired OST. Then click **Update**.

Configure the Read/Write Bandwidth chart

1. Click **Configure**.
2. Click **Set Duration** and enter at time period over which samples will be taken. Enter Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day. The value given is an average for that sample period.
3. Click **Set Range** to set the **Start** and **End** times and dates over which measurements will be displayed. This view of the chart is a static snapshot, starting and ending as configured.
4. Click **Update** to apply and close this window.

9.2.5 Metadata Servers chart

This chart shows the percentage of CPU and RAM resources consumed on all metadata server(s) in all file systems, over time. This chart is displayed when all File Systems are selected on the Dashboard window (default). You can also view this chart for a single file system.



Features

- Mouse over any point on the chart to learn the date/time of this measurement and the values at that time.
- Click **Change Duration** to set the total time duration to Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day. The value given is an average for that sample period.
- Click **Full Screen** to fill the browser window with this chart. Click **Exit Full Screen** to return to the normal view.
- Click **CPU** or **RAM** to select/deselect to view only that information on the chart.

View this chart for a specific file system

This chart is displayed by default for all file systems. To view this chart for a single file system:

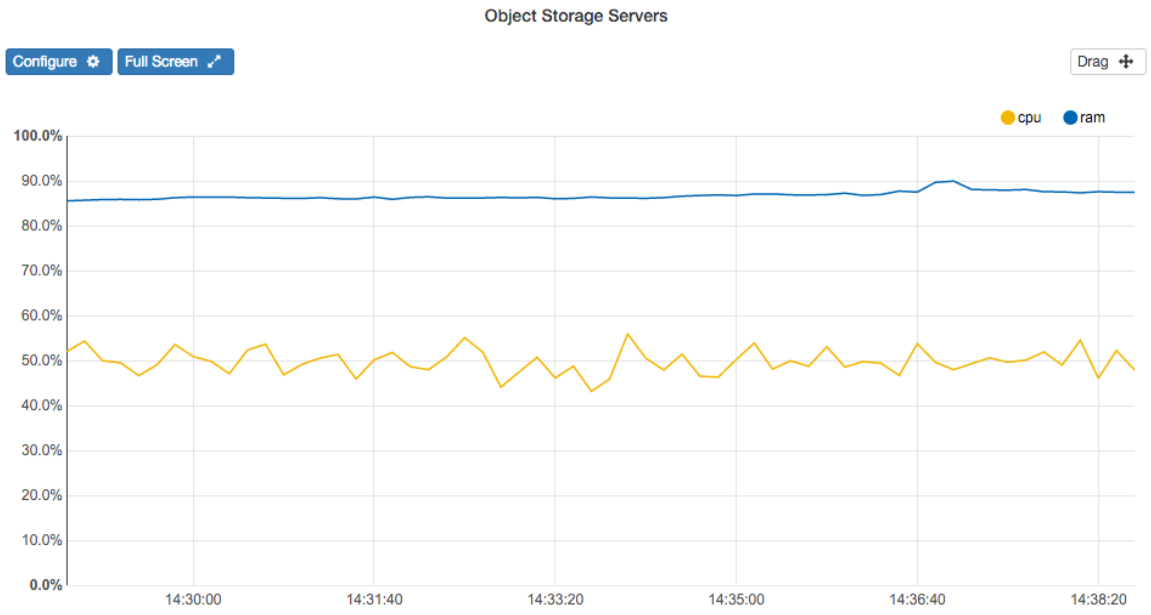
1. On the Dashboard, click **Configure Dashboard**.
2. Select **File System** (default).
3. At the **File System** drop-down menu, select the file system. Then click **Update.**

Configure the Metadata Servers chart

1. Click **Configure**.
2. Click **Set Duration** and enter at time period over which samples will be taken. Enter Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day. The value given is an average for that sample period.
3. Click **Set Range** to set the **Start** and **End** times and dates over which measurements will be displayed. This view of the chart is a static snapshot, starting and ending as configured.
4. Click **Update** to apply and close this window.

9.2.6 Object Storage Servers chart

The Object Storage Servers chart shows the percentages of CPU and RAM resources used on object storage servers (in all file systems) over time. This chart is displayed when File Systems are selected on the Dashboard window (default). This chart can also be displayed for a single file system.



Features

- Click **Change Duration** to set the total time duration to Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day. The value given is an average for that sample period.

- Click **Full Screen** to fill the browser window with this chart. Click **Exit Full Screen** to return to the normal view.
- Click **CPU** or **RAM** to select/deselect to view only that information on the chart.

View this chart for a specific file system

This chart is displayed by default for all file systems. To view this chart for a single file system:

1. On the Dashboard, click **Configure Dashboard**.
2. Select **File System** (default).
3. At the **File System** drop-down menu, select the file system. Then click **Update**.

Configure the Object Storage Servers chart

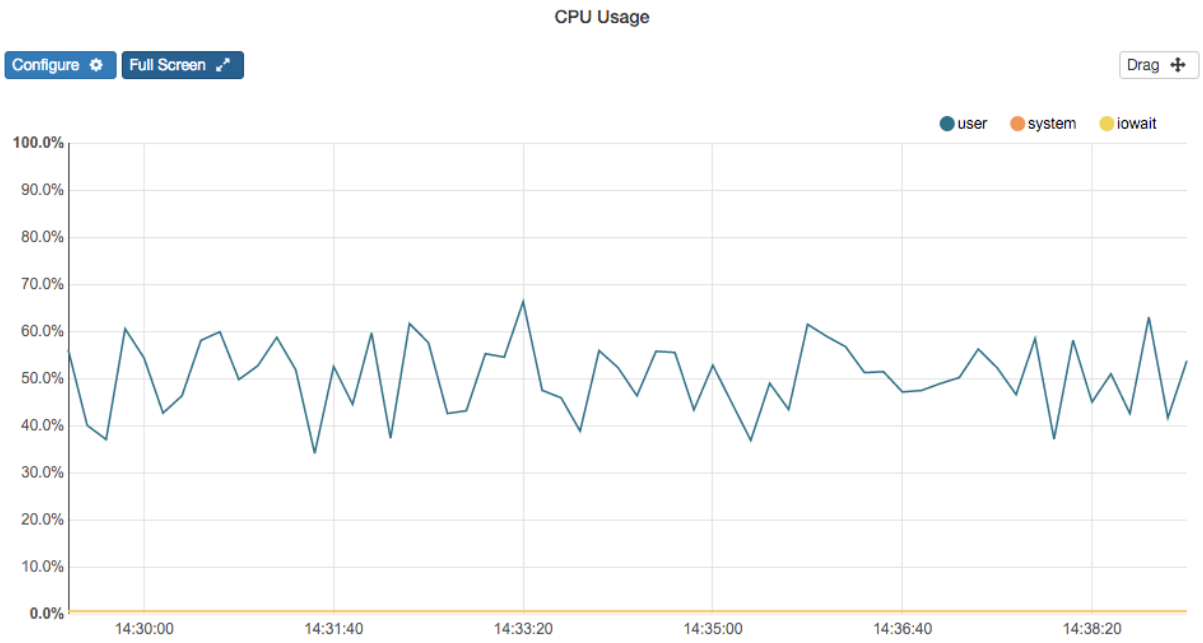
1. Click **Configure**.
2. Click **Set Duration** and enter at time period over which samples will be taken. Enter Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day. The value given is an average for that sample period.
3. Click **Set Range** to set the **Start** and **End** times and dates over which measurements will be displayed. This view of the chart is a static snapshot, starting and ending as configured.
4. Click **Update** to apply and close this window.

9.2.7 CPU Usage chart

This chart is visible for an individual server. The chart shows the percentages of CPU activity attributed separately to:

- user-level processes
- system-level processes
- processes in an IO Wait state

Data is displayed for the specific metadata server or object storage server selected, over time.



- Mouse over any point on the chart to learn the date/time of this measurement and the values at that time.
- Click **Full Screen** to fill the browser window with this chart. Click **Exit Full Screen** to return to the normal view.
- Click **user**, **system**, or **iowait** to select/deselect to view only that information on the chart.

View this chart

1. On the Dashboard, click **Configure Dashboard**.
2. Select **Server**.
3. Under *Server*, select the server of interest and click **Update**.

Configure the CPU Usage chart

1. Click **Configure**.
2. Click **Set Duration** and enter at time period over which samples will be taken. Enter Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day. The value given is an average for that sample period.
3. Click **Set Range** to set the **Start** and **End** times and dates over which measurements will be displayed. This view of the chart is a static snapshot, starting and ending as

configured.

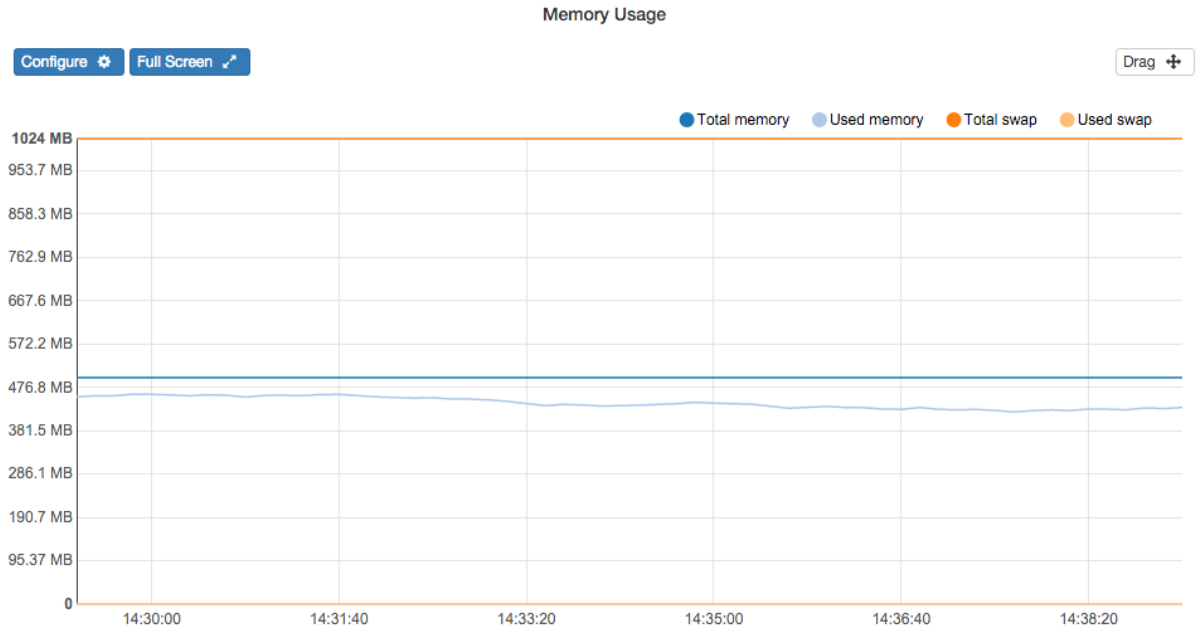
4. Click **Update** to apply and close this window.

9.2.8 Memory Usage chart

For an individual metadata server or object storage server selected, the Memory Usage chart shows:

- the total amount of RAM memory present
- the amount of RAM currently used
- the total swap space currently available
- the amount of swap space being used.

Data is displayed for the server selected, over time.



Features

- Mouse over any point on the chart to learn the date/time of this measurement and the values at that time.
- Click **Change Duration** to set the total time duration to Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day.
- Click **Full Screen** to fill the browser window with this chart. Click **Exit Full Screen** to

return to the normal view.

- Click any of the display icons: **Total memory**, **Used memory**, **Total swap**, **Used swap** to display only your selected parameters.

View this chart

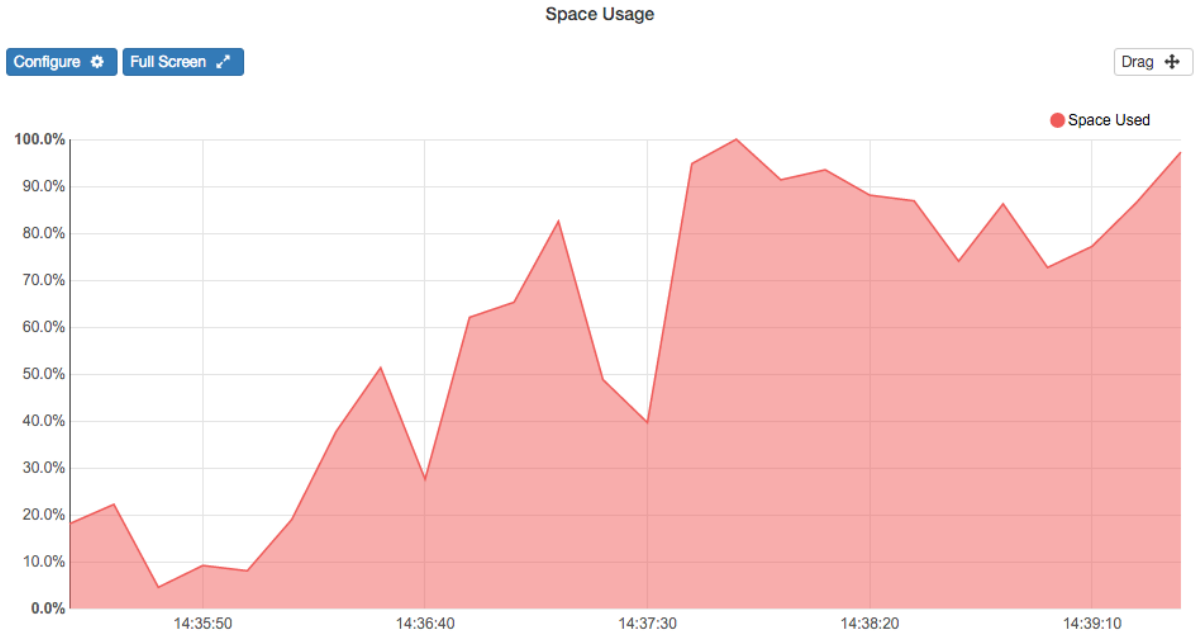
1. On the Dashboard, click **Configure Dashboard**.
2. Select **Server**.
3. Under *Server*, select the server of interest and click **Update**.

Configure the Memory Usage chart

1. Click **Configure**.
2. Click **Set Duration** and enter at time period over which samples will be taken. Enter Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day. The value given is an average for that sample period.
3. Click **Set Range** to set the **Start** and **End** times and dates over which measurements will be displayed. This view of the chart is a static snapshot, starting and ending as configured.
4. Click **Update** to apply and close this window.

9.2.9 Space Usage chart

This chart is displayed for a selected MDT or OST and shows percentage of file system space consumed on a target over time.



Features

- Mouse over any point on the chart to learn the date/time of this measurement and the values at that time.
- Click **Change Duration** to set the total time duration to Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day.
- Click **Full Screen** to fill the browser window with this chart. Click **Exit Full Screen** to return to the normal view.

View this chart

1. On the Dashboard, click **Configure Dashboard**.
2. Select **Server**.
3. At the **Server** drop-down menu, select the sever hosting the desired target.
4. At the **Target** drop-down menu, select the desired target. Then click **Update**.

Configure the Space Usage chart

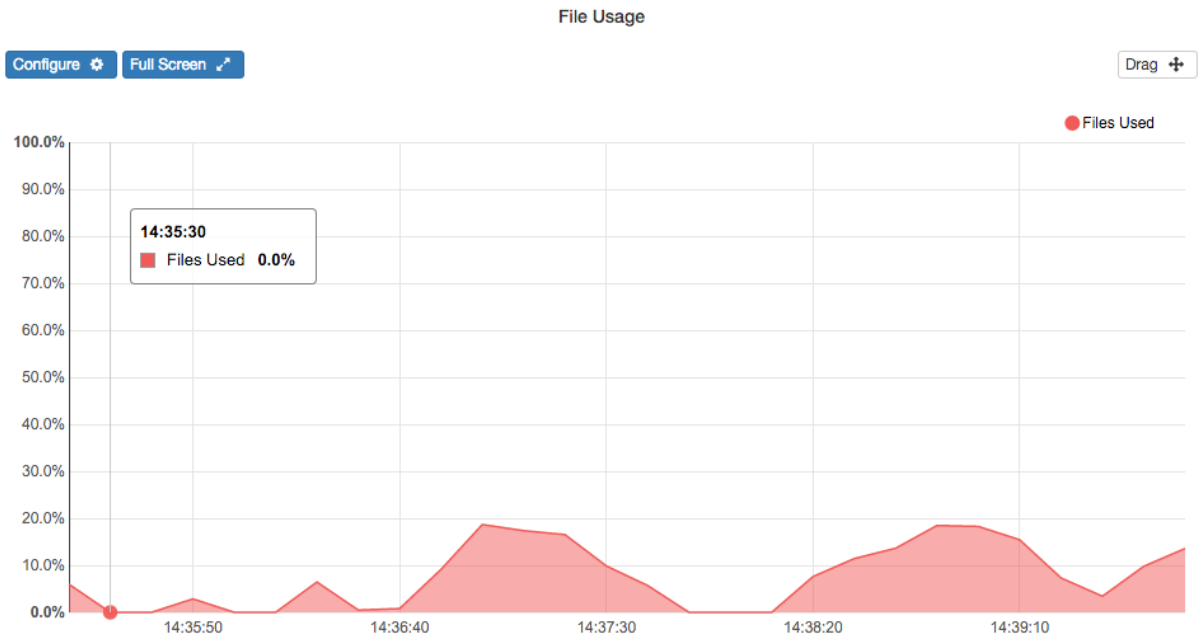
1. Click **Configure**.
2. Click **Set Duration** and enter at time period over which samples will be taken. Enter Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long

durations, the map will be divided over several days, with samples periods starting at different times of the day. The value given is an average for that sample period.

3. Click **Set Range** to set the **Start** and **End** times and dates over which measurements will be displayed. This view of the chart is a static snapshot, starting and ending as configured.
4. Click **Update** to apply and close this window.

9.2.10 File Usage chart

This chart is displayed for a selected MDT and shows the percentage of available files (inodes) used over time. Data is displayed for the specific metadata target selected.



Features

- Mouse over any point on the chart to learn the date/time of this measurement and the values at that time.
- Click **Change Duration** to set the total time duration to Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day.
- Click **Full Screen** to fill the browser window with this chart. Click **Exit Full Screen** to return to the normal view.

View this chart

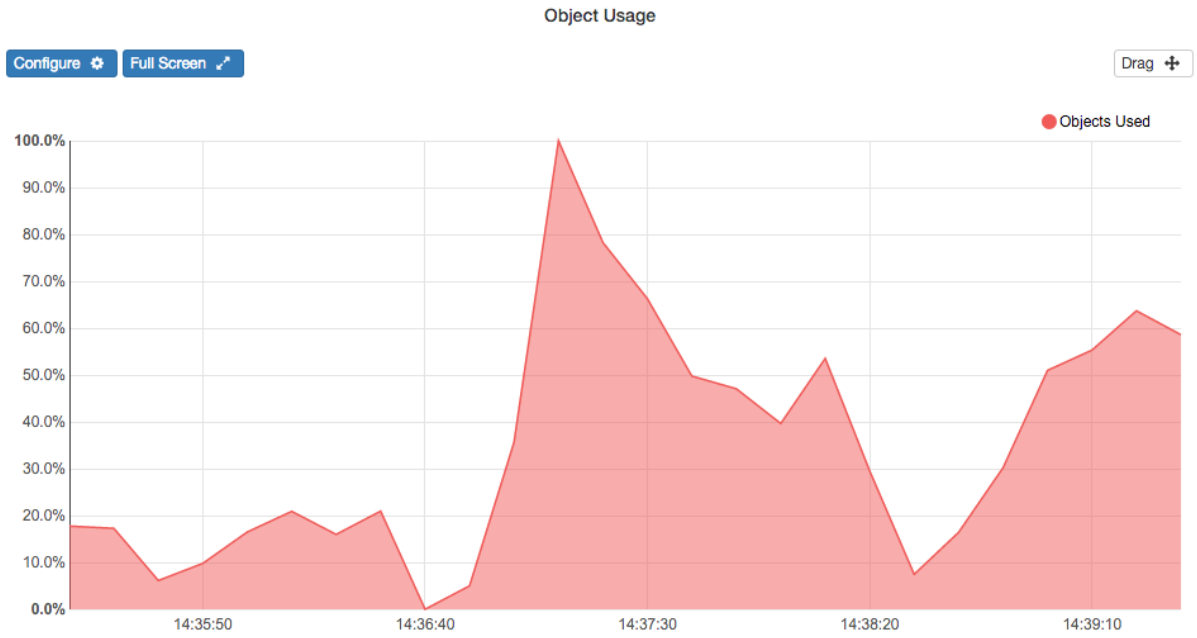
1. On the Dashboard, click **Configure Dashboard**.
2. Select **Server**.
3. At the **Server** drop-down menu, select the server hosting the desired target.
4. At the **Target** drop-down menu, select the desired MDT. Then click **Update**.

Configure the File Usage chart

1. Click **Configure**.
2. Click **Set Duration** and enter at time period over which samples will be taken. Enter Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day. The value given is an average for that sample period.
3. Click **Set Range** to set the **Start** and **End** times and dates over which measurements will be displayed. This view of the chart is a static snapshot, starting and ending as configured.
4. Click **Update** to apply and close this window.

9.2.11 Object Usage chart

This chart is displayed for a selected OST and shows the percentage of metadata objects used over time. Data is displayed for the object storage target selected.



View this chart

1. On the Dashboard, click **Configure Dashboard**.
2. Select **Server**.
3. At the Server drop-down menu, select the server hosting the desired OST.

Configure the Object Usage chart

1. Click **Configure**.
2. Click **Set Duration** and enter at time period over which samples will be taken. Enter Minutes (1-60), Hours (1-24), Days (1- 31), or Weeks (1-4). Note that for long durations, the map will be divided over several days, with samples periods starting at different times of the day. The value given is an average for that sample period.
3. Click **Set Range** to set the **Start** and **End** times and dates over which measurements will be displayed. This view of the chart is a static snapshot, starting and ending as configured.
4. Click **Update** to apply and close this window.

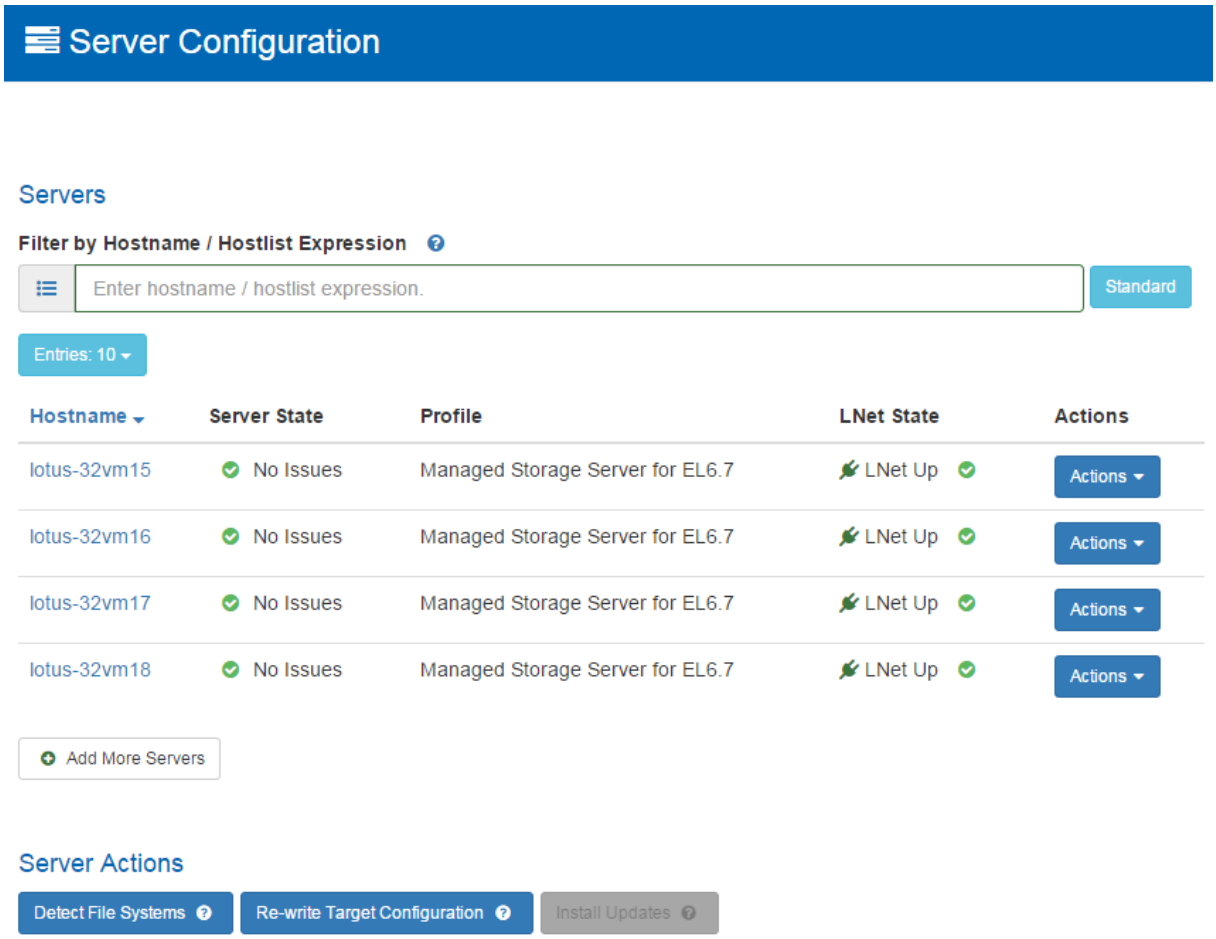
9.3 Configuration menu

The Configuration menu provides access to the following windows, to let you create and manage file systems:

- The Server window lets you configure a new server for a new file system or add a server to an existing file system.
- At the Power Control window, you can configure power distribution units and outlets, and assign servers to PDU outlets to support high availability/failover.
- At the File Systems window, you can create a new file system or manage a file system.
- The HSM window configure and monitor hierarchical storage management (HSM) activities. You can also add a copytool to a worker agent and assign that tool instance to a file system.
- The Storage window lists detected storage module plug-ins (provided by third parties), which may provide configuration, status, and/or failover control of RAID based storage devices, depending entirely on the plug-in.
- At the Users window, add and configure superusers and users. Superusers are administrators.
- Add volumes and configure those volumes for high availability at the Volumes window.
- The MGTs window lets you configure the management target.

9.3.1 Server Configuration window

The Server Configuration window is shown next. This is an example only configuration only.



This window supports the range of server configuration tasks. For instructions on how to add servers, see Add one or more HA servers.

Under **Server Configuration**, you can:

- Add an object storage server. Click **+ Add Server** or **+ Add More Servers**.
- View existing servers for all file systems.
- View **Server State**: This indicator tells you the alert status for that server. A green check mark indicates that all is well with that server. A red exclamation mark indicates an active alert has been generated for this server; you can mouse over the exclamation mark to learn the cause of the alert. See View all status messages for more information.
- View the **Profile** associated with each server. When you add a new server, you select the server profile for that server. The profile defines the role of that server. There are generally four server profiles available, however your installation may list more. The four common server profiles are:
  - *Managed storage server*

  - *Monitored storage server*
  - *POSIX HSM Agent Node*
  - *Robinhood Policy Engine server*
- Determine **LNet state** for a given server. Possible LNet states are: *LNet up*, *LNet down*, and *LNet unloaded*.
- Click on the server name (hostname) to open a Server Detail window to learn more about that server and access configuration options.
- Under **Actions**, specific to each server, you can perform the following commands. These commands are used primarily to decommission servers. See Decommissioning a server for an MGT, MDT, or OST.
  - **Reboot**: Initiate a reboot on this server. If this server is configured as the primary server of an HA pair, the file system will failover to the secondary server until this server is back online. The file system will then fail back to the primary server. If this is not configured as an HA server, then any file systems targets that rely on this server will be unavailable until rebooting is complete.
  - **Shutdown**: Initiate an orderly shutdown on this server. If this server is configured as the primary server of an HA pair, the file system will failover to the secondary server. If this is not configured as an HA server, then any file systems targets that rely on this server will be unavailable until this server is rebooted.
  - **Remove**: Remove this server. If this server is configured as the primary server of an HA pair, then the file system will failover to the secondary server.
  - **Warning**: If this is not configured as an HA server, then *any file systems or targets that rely on this server will also be removed*.
  - **Power Off**: Switch power off for this server. Any HA-capable targets running on the server will be failed-over to a peer. Non-HA-capable targets will be unavailable until power for the server is switched on again. This action is visible only if PDUs have been added and outlets assigned to servers.
  - **Power On**: Switch power on for this server. This action is visible only if PDUs have been added and outlets assigned to servers, and after the server has been powered-off at PDU.
  - **Power Cycle**: Switch power off and then back on again for this server Any HA-capable targets running on the server will be failed over to a peer. Non-HA-capable targets will be unavailable until the server has finished booting. This action is visible only if PDUs have been added and outlets assigned to servers.
  - **Remove**: Remove this server. If this server is configured as the primary server of an HA pair, then the file system will failover to the secondary server. If this not

configured as an HA server, then any file systems or targets that rely on this server will also be removed.

- **Force Remove**: This action removes the record for the storage server in the manager database, without attempting to contact the storage server. All targets that depend on this server will also be removed without any attempt to unconfigure them. **Warning**: **You should only perform this action if the server is permanently unavailable**.

Under **Server Actions**, you can perform the commands listed next. Note that these commands are *bulk action commands*. This means that when you click one of the following commands, you can then select which server(s) to perform this command on. You can enter a host name or host name expression in the file to generate a list of existing servers. You can chose **Select All**, **Select None**, or **Invert Selection**. At the far right, under *Select Server*, you can also select or deselect a server. After selecting the desired server(s), you can proceed to perform the command and it will be run on all selected servers.

- **Detect File Systems**: Detect an existing file system to be monitored at the manager GUI.
- **Re-write Target Configuration**: Update each target with the current NID for the server with which it is associated. This is necessary after making changes to server/target configurations and is done after rescanning NIDs. Also see Handling network address changes (updating NIDs).
- **Install Updates**: When an updated release of Intel® Manager for Lustre* software is installed at the *manager server*, a notification is displayed at the manager GUI that updated software is also available for installation on a managed server or servers. This button become enabled. After clicking the **Install Updates** button, a list of servers (default: all) to be included in this update operation is displayed in the Update dialog. Clicking the **Run** button in this dialog will cause the updated packages to be installed on the managed servers.

Server Detail window

Each *Server Detail* window contains the full extent of information for that server. To open a *Server Detail* window, click **Configuration** > **Servers**, and then click on the server of interest.

This window is divided into five sections:

- Server Detail
- Pacemaker configuration
- Corosync configuration
- LNet detail

- NID configuration

Server Detail

This section lists:

- **Address**: This is the IP address or the node name.
- **State**: The type of server, HA managed or unmanaged.
- **FQDN**: Fully qualified domain name
- **Node name**: The name previously assigned to this node.
- **Profile**: Indicates the profile assigned to this server during the Add Server process, including the OS.
- **Boot time**: Date of last boot
- **State changed**: Date of last State change; see State above.
- **Alerts**: Any alerts received pertinent to this server.

Click the **Actions** menu to access the following commands that are available for this server:

- **Reboot**: Initiate a reboot on this server. If this server is configured as the primary server of an HA pair, the file system will failover to the secondary server until this server is back online. The file system will then fail back to the primary server. If this is not configured as an HA server, then any file systems targets that rely on this server will be unavailable until rebooting is complete.
- **Shutdown**: Initiate an orderly shutdown on this server. If this server is configured as the primary server of an HA pair, the file system will failover to the secondary server. If this is not configured as an HA server, then any file systems targets that rely on this server will be unavailable until this server is rebooted. Remove: Remove this server. If this server is configured as the primary server of an HA pair, then the file system will failover to the secondary server. Warning: If this is not configured as an HA server, then any file systems or targets that rely on this server will also be removed.
- **Power Off:** This will switch power off for this server. If this is a primary server to any targets, those targets will be failed-over to the secondary server. Non-HA-capable targets (targets not supported by a secondary server) will be unavailable until power for the server is switched on again. This action is visible only if PDUs have been added and outlets assigned to servers.
- **Power Cycle:** Switch power off and then back on again for this server Any HA-capable targets running on the server will be failed over to a peer. Non-HA-capable targets will

be unavailable until the server has finished booting. This action is visible only if PDUs have been added and outlets assigned to servers.

- **Remove**: Remove this server. If this server is configured as the primary server of an HA pair, then the file system will failover to the secondary server. If this not configured as an HA server, then any file systems or targets that rely on this server will also be removed.
- **Force Remove**: This action removes the record for the storage server in the manager database, without attempting to contact the storage server. All targets that depend on this server will also be removed without any attempt to unconfigure them. Warning: You should only perform this action if the server is permanently unavailable.

Pacemaker configuration

Pacemaker configuration and enabling is performed automatically by Intel® Manager for Lustre* software. However, an administrator may need to reset or configure Pacemaker when performing maintenance on a server, altering the server's configuration, or troubleshooting problems with Pacemaker.

Click the **Actions** menu to access the following commands:

- **Stop Pacemaker:** This command stop Pacemakers. If this is a primary server, then failover to the secondary server occurs. The file system remains available but not in a high-availability state.
- **Unconfigure Pacemaker:** This command stops and unconfigures Pacemaker. If this is a primary server, then failover to the secondary server occurs. The file system remains available but not in a high-availability state.
- **Configure Pacemaker**: Visible if Pacemaker is unconfigured. This command configures Pacemaker, but does not start it. To start Pacemaker and restore this server to HA capability, click **Start Pacemaker.**
- **Start Pacemaker**: Visible if Pacemaker is stopped or unconfigured. Start Pacemaker to restore this server to HA capability. If failover has occurred from this server to the backup server, then after starting Pacemaker, manually failback the affected target(s) to this primary server. To do this, open the Status window, locate any warnings for target(s) running on the secondary server (and served by this primary server) and under Actions, click **Failback**.

Corosync configuration

Corosync configuration and enabling is performed automatically by Intel® Manager for Lustre* software. However, an administrator may need to reset or configure Corosync when performing maintenance on a server, altering the server's configuration, or

troubleshooting problems with Corosync.

Click the **Actions** menu to access the following commands:

- **Stop Corosync**: This command stops Corosync and also stops Pacemaker. If this is a primary server, then failover to the secondary server occurs. The file system remains available, but not in a high-availability state. Corosync must be restarted before Pacemaker can be started again.
- **Unconfigure Corosync**: This command stops and unconfigures Corosync and also stops Pacemaker. If this is a primary server, then failover to the secondary server occurs. The file system remains available, but not in a high-availability state. Corosync must be restarted before Pacemaker can be started again.
- **Configure Corosync**: Visible if Corosync is unconfigured. This command will configure Corosync, but not start it. To configure and start Corosync, click **Start Corosync**. After Corosync is started, you need to start Pacemaker.
- **Start Corosync**: Visible if Corosync is stopped or unconfigured. After Corosync is started, you also need to start Pacemaker. If failover occurred from this server to the backup server, then after Corosync and Pacemaker are running, you need to manually failback the affected target(s) to this primary server. See *Start Pacemaker*, above.

Clicking **Configure** to change the mcast port number.

LNet detail

LNet operations for a given server may need to be reset during maintenance. Doing so will take this server and any volumes it hosts offline, and depending on the server, will degrade or stop the file system.

Click the **Actions** menu to access the following commands:

- **Stop LNet**: Shut down the LNet networking layer and stop any targets running on this server.
- **Unload LNet**: If LNet is running, stop LNet and unload the LNet kernel module to ensure that it will be reloaded before any targets are started again.
- **Load LNet**: Load the LNet kernel module for this server.
- **Start LNet**: Start LNet.

NID configuration

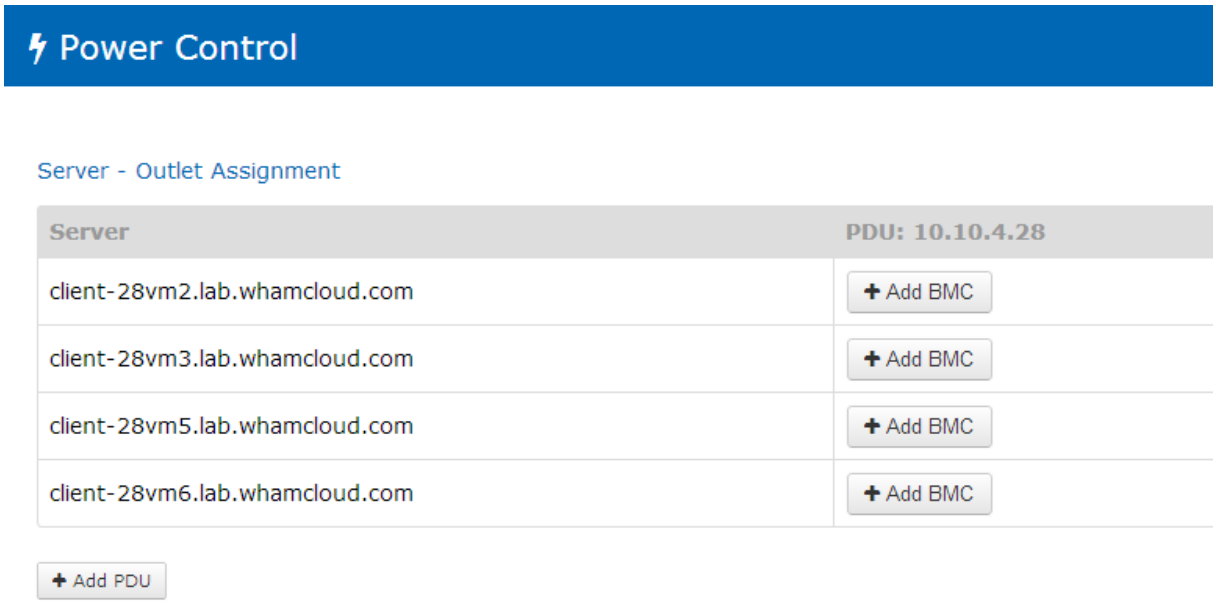
An administrator may need to reconfigure NIDs for a server when performing maintenance on a server, altering the server's configuration, or troubleshooting problems network interfaces. For each interface, you can set the network driver and assign the Lustre network. To be able to edit NID configuration, the file system first needs to taken

offline. Perform these steps:

1. At the menu bar, click **Configuration** > **File Systems**.
2. For this listed file system, at the right under **Actions**, select **Stop**.
3. Return to the Server Detail window for the server in question. Click **Configuration** > **Servers**. Click on the desired server.
4. Under NID Configuration, click **Configure**.
5. The IP address is not editable. At the Network Driver drop-down menu, the available driver types are dependent on the network interface. Select the appropriate driver.
6. If you are ready to place the file system online again, click **Configuration** > **File Systems**. Then for this file system, under **Actions**, select **Start**.

9.3.2 Power Control window

The *Power Control* window accessed from the *Configuration* menu is shown next.

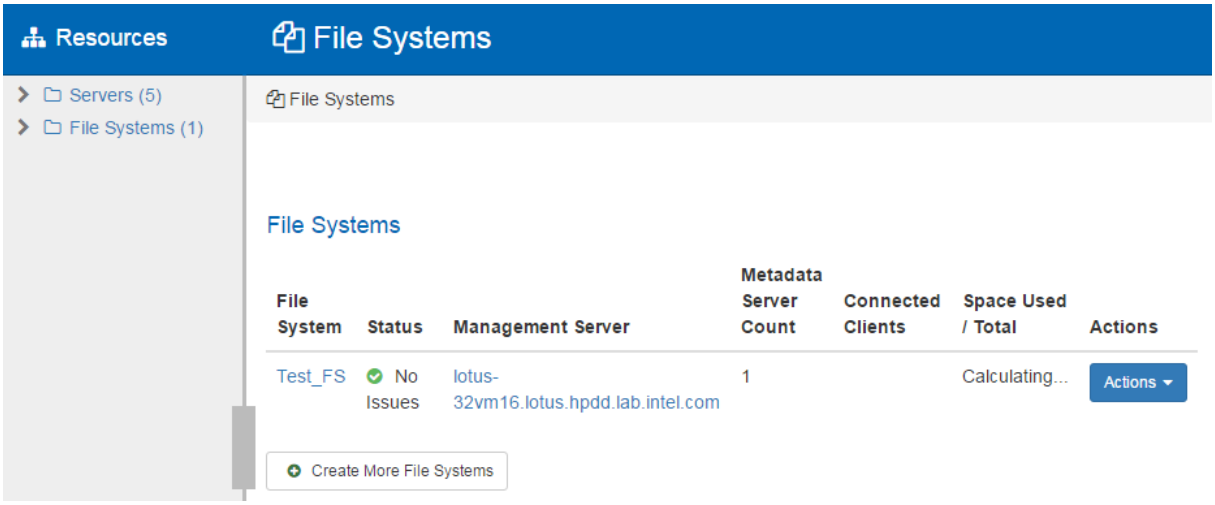


The *Power Control* window lets you configure and manager power distribution units. At this window you can add a detected PDU and then assign specific PDU outlets to specific servers. Once configured, this feature lets you check the status of PDUs and individual outlets. Based on server power requirements and your failover configuration, you may want to assign more than one outlet to a server. For improved failover performance, assign the failover outlet from a different PDU than the primary outlet. When you associate PDU failover outlets with servers using this tool, STONITH is automatically configured. Note that primary and secondary servers for each target must first be configured on the *Volumes* window.

See Add power distribution units.

9.3.3 File Systems window

The *File Systems* window accessed from the *Configuration* menu is shown next.



The *File Systems* window lets you configure, view and manage multiple file systems.

Click **Create File System** (or **Create More File Systems**) to begin the process of creating a new file system. See Create a new Lustre* file system.

Under Current File Systems, for each file system you can:

- view the file system name
- view the management server (MGS)
- view the metadata server (MDS)
- view the number of connected clients
- view total file system capacity (Size)
- view available free space
- check file system status. A green check mark ✔ indicates that the file system is operating normally. No warnings or error messages have been received.

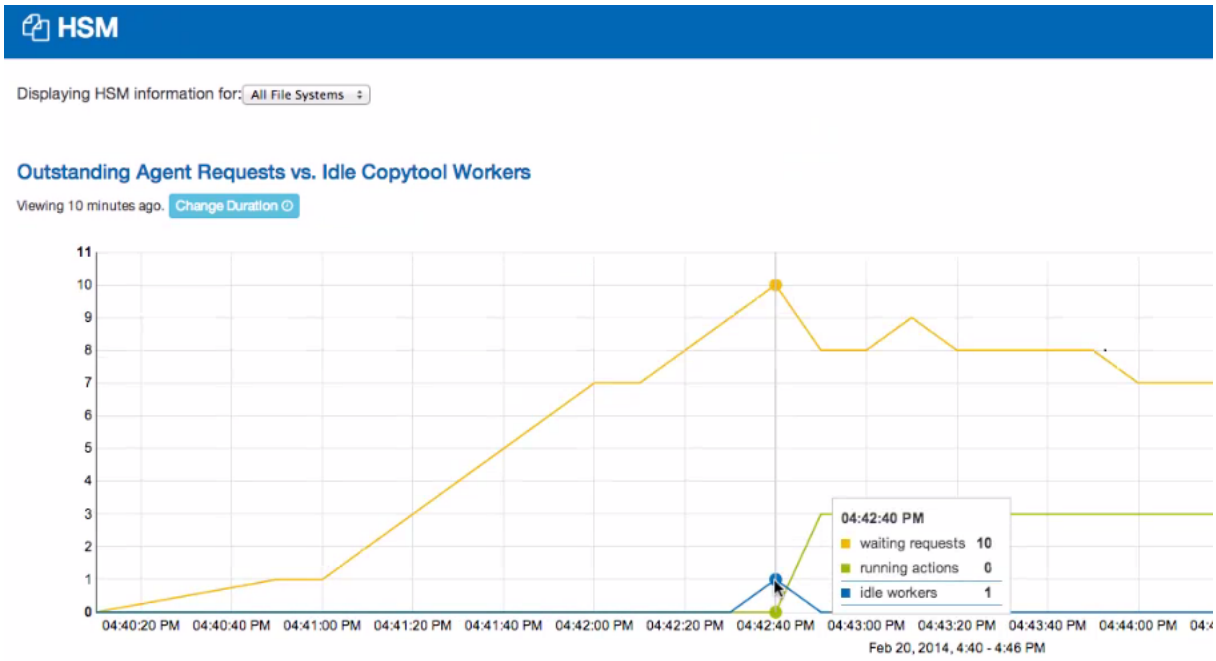
Under **Actions**, you can:

- **Remove** the file system: This file system is removed and will not be available to clients. However this file system's contents will remain intact until its volumes are reused in another file system.
- **Stop** the file system: This stops the metadata and object storage targets, thus making the file system unavailable to clients. If the file system has been stopped, to restart the file system, click **Start**.

To view the full display of file system parameters, click on the file system name in the left column. See View All File System Parameters.

9.3.4 HSM window

After Hierarchical Storage Management (HSM) has been configured for a file system, this HSM Copytool chart displays a moving time-line of waiting copytool requests, current copytool operations, and the number of idle copytool workers. For information about setting up HSM for a file system, see Configuring and using Hierarchical Storage Management.



On this window, you can:

- Select to display copytool operations for all file systems (default), or one you select.
- Mouse over the graph to learn the specific values at a given point in time.
- Use Change Duration to change the time period for the range of data displayed on the HSM Copytool chart. chart. The chart begins at a start time set and ends now. You can set this to select Minutes, Hours, Days or Weeks, up to four weeks back in time and ending now. The most recent data displayed on the right. The number of data points will vary, based primarily on the duration.
- Click **Actions** > **Disable** to pause the HSM coordinator for this file system (pause HSM activities). New requests will be scheduled and HSM activities will resume after the HSM coordinator is enabled. To enable again, click **Actions** > **Enable**.

- Click **Actions** > **Shutdown** to stop the HSM coordinator for this file system. No new requests will be scheduled.

If a copytool has been added but never configured or started, then click **Actions** to show the following menu:

- **Start** - Configure and Start this copytool to begin processing HSM requests.
- **Remove** - Deconfigure and remove this copytool from the manager database. It will no longer appear on this HSM window. This is best way to remove a copytool.
- **Configure** - Configure this copytool on the worker. Do not start the copytool. Status will show as Configured.
- **Force Remove** - Remove this copytool from the manager database without deconfiguring this copytool on the worker node. It will no longer appear on this HSM window. This is NOT the best way to remove a copytool, because a later attempt to add this copytool back will fail unless it is manually reconfigured. Only consider using Force Remove if Remove has failed.

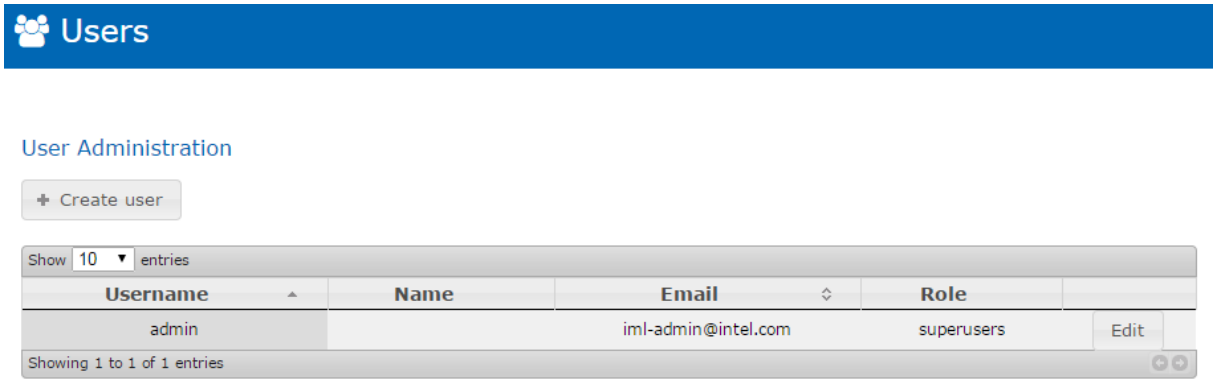
To learn about HSM capabilities supported in Intel® Enterprise Edition for Lustre* software, see Configuring and using Hierarchical Storage Management.

9.3.5 Storage window

The *Storage* window lists detected storage module plug-ins (provided by third parties), which may provide configuration, status, and/or failover control of RAID based storage devices, depending entirely on the plug-in. If no plug-ins are detected, none are listed. The layout and information displayed on this window is dependent on the storage plug-in(s).

9.3.6 Users window

The *Users* window accessed from the *Configuration* menu is shown next.



The *Users* window lets you create and manage the following accounts types:

- ***File system user*** - A file system user has access to the full GUI, except for the Configuration drop-down menu, which is not displayed. A file system user cannot create or manage a file system, but can monitor all file systems using the Dashboard, and the Alerts and Logs windows. Users log in by clicking **Login** in the upper-right corner of the screen, and log out by clicking **Logout**.
- ***Superuser*** - A superuser has full access to the application, including the Configuration drop-down menu and all sub-menus. A superuser can create, monitor, manage, and remove file system and their components. A superuser create, modify (change passwords), and delete users. A superuser cannot delete their own account, but a superuser can create or delete another superuser.

See Creating User Accounts for more information.

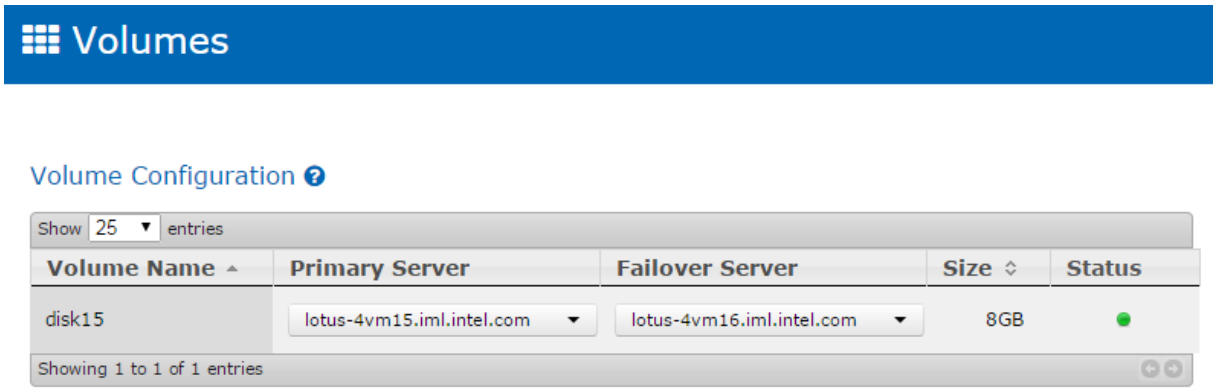
After logging in, a user can modify their own account by clicking **Account** near the upper-right corner of the screen. A user can set these options:

- ***Details*** - Username, email address, and first and last name can be changed.
- ***Password*** - Password can be changed and confirmed.
- ***Email Notifications*** - The types of events for which this account will receive emailed notifications can be selected from a checklist. If no notifications are selected, email notifications will be sent for all alerts except “Host contact alerts”. See Setting up Email Notifications.

See Creating User Accounts for more information.

9.3.7 Volumes window

The *Volumes* window accessed from the *Configuration* menu is shown next.



The *Volumes* window is for adding volumes to a file system. Volumes (also called LUNs or block devices) are the underlying units of storage used to create Lustre* file systems. Each Lustre target corresponds to a single volume. Only volumes that are not already in use as

Lustre targets or local file systems are shown. If servers in the volume have been configured for high availability, primary and secondary servers can be designated for a Lustre volume. A volume may be accessible on one or more servers via different device nodes, and it may be accessible via multiple device nodes on the same host.

On the *Volume* window, you can do the following:

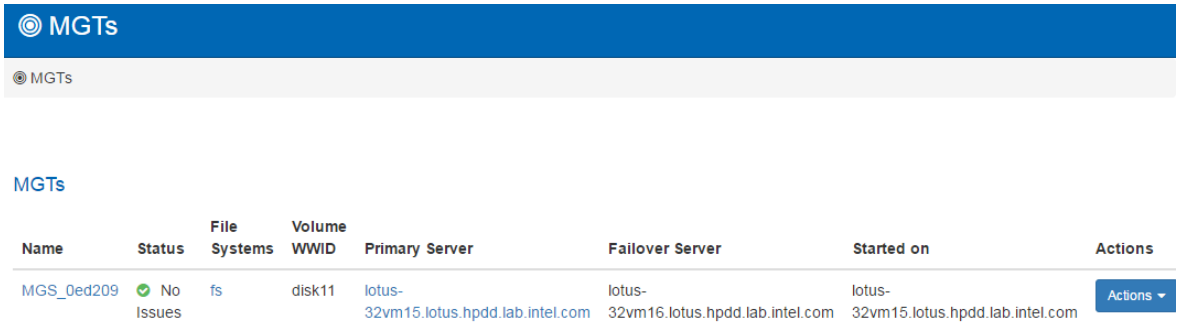
- Set or change the Primary Server and Failover Server for each volume. Each change you select to make will be displayed in orange, indicating that you have selected to change this setting, but have not applied it yet. Changes you make on this Volumes Configuration window will be updated and displayed after clicking **Apply** and **Confirm**. After confirming the change, the orange setting turns white. Other users viewing this file system's Volume Configuration window will see these updated changes after you apply and confirm them. If you select to change a setting (it becomes orange), you can click **X** to cancel that selection (it turns white and returns to the original setting). To cancel all changes you have selected (but not yet applied), click **Cancel**.

  Note: There is currently no lock-out of one user's changes versus changes made by another user. The most-recently applied setting is the one in-force and displayed.

- View the status of all volumes in all file systems.
- View each volume's name, primary server, failover server, volume size, and volume status.
  - A green Status light for the volume indicates that the volume has a primary and failover server.
  - A yellow Status light means that there is no failover server.
  - A red Status light indicates that this volume is not available.

9.3.8 MGTs window

The *MGT* window accessed from the *Configuration* menu is shown next.



At the MGT window, you can do the following:

- View your existing management target (if configured). Here you can determine the Capacity, Type, and high availability (HA) Status of the MGT. If this is an HA target, then the primary and secondary servers are identified. A green check mark ✔ indicates this target and server are functioning normally.
- Select storage for a new MGT and then create a new MGT. This task is not common; MGTs are created when you click **Create File System** at the *Configuration* > *File Systems* window.

Under MGT Configuration for an existing MGT, you can perform these actions under **Actions**:

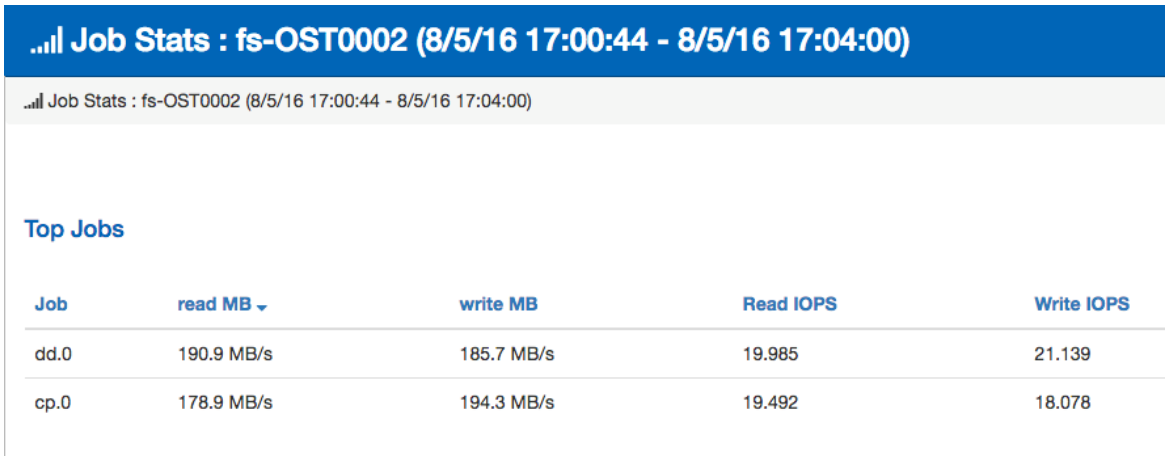
- **Stop**: Stop the MGT. When an MGT is stopped, clients are unable to make new connections to the file systems using this MGT. However, the MDT and OST(s) stay up if they were started before this MGT was stopped, and can be stopped and restarted while this MGT is stopped.
- **Failover**: Clicking Failover will forcibly migrate the target to its failover server. Clients attempting to access data on the target while the migration is in process may experience delays until the migration completes. If this action is not displayed, then the MGT has already failed-over and this button will display as Failback. Otherwise, a secondary server has not been configured.
- **Failback**: Migrate the target back to its primary server. Clients attempting to access data on the target while the migration is in process may experience delays until the migration completes. This action is displayed only after a target has failed-over.

9.4 Job Stats window

The Job Stats window is accessible at the top menu bar. Click **Job Stats**.

Clicking **Job Stats** opens the Job Stats window and reveals the top five jobs currently in process. The listed jobs can be sorted by column and average duration can be selected. Column sorts and duration selections are persistent if you leave and later return to this window.

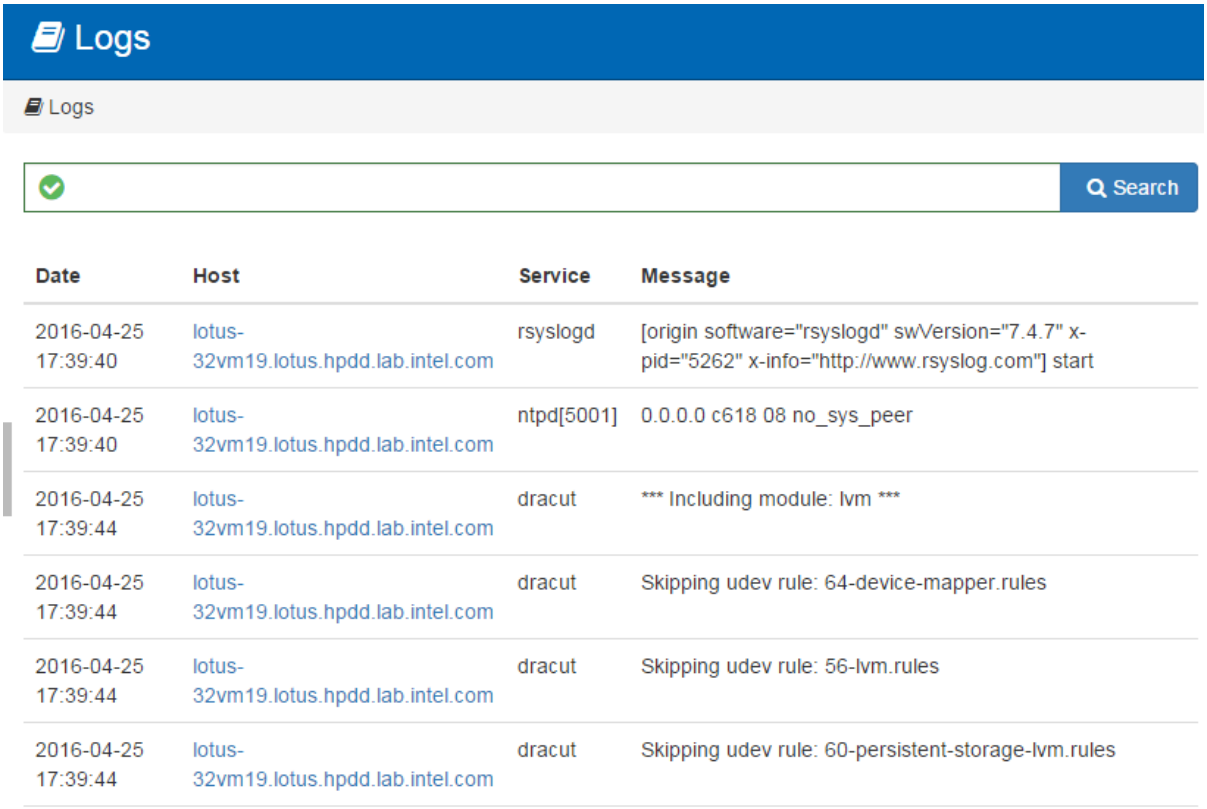
Note: Job stats need to be enabled before then can be viewed. See View Job stats.



On the Read/Write Heat Map (on the Dashboard), you can also click a heat map cell and go to the Job Stats screen for that OST. Doing so will present a static view of job stats for the selected OST. Because it is static, *Duration* is not selectable.

9.5 Logs window

The *Logs* window is shown next.



The *Logs* window displays log information and allows filtering of events by date range, host, service, and messages from Lustre or all sources. The Logs window also features auto-complete search functions and linkable host names.

The logs window also features querying with auto-complete and linkable host names.

For example, if a failover event takes place, the following occurs:

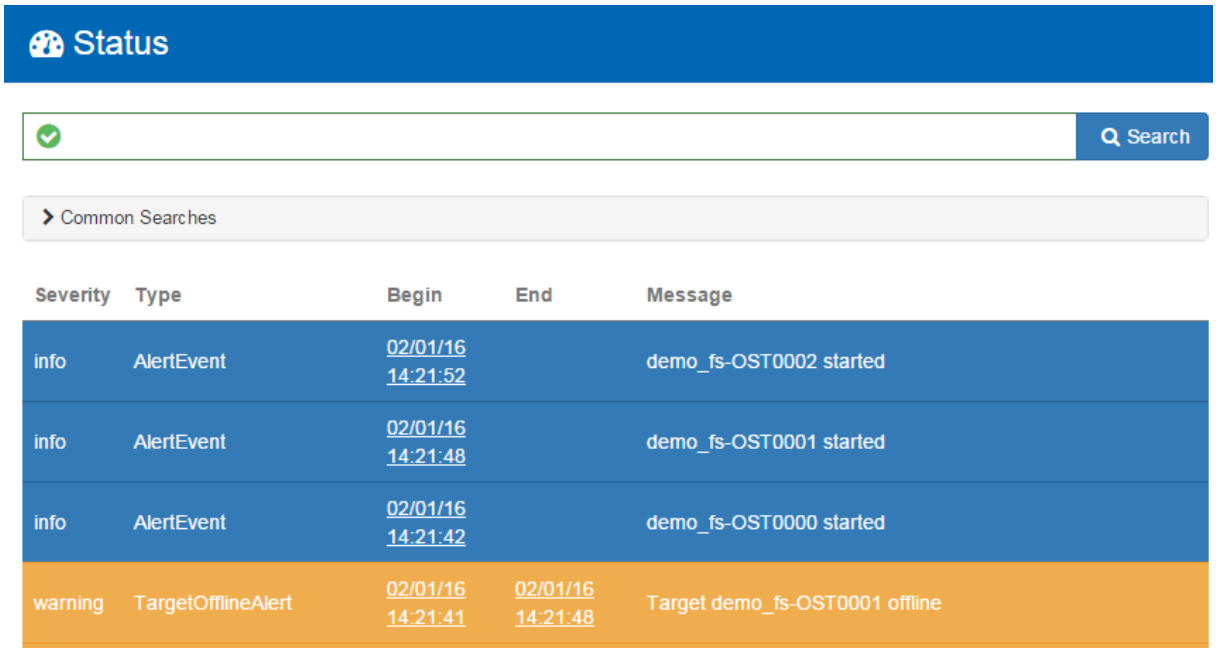
- The red Alert bar will appear briefly to notify you of active warning alerts related to the failover.
- An alert is displayed with a message on the Status window that a server has failed over. Other related alerts will be displayed.
- The alert icon appears on the Configuration > File System window for the file system. The server on which the target is now running is shown in the *Started On* column for that target.
- An email alert is sent to the superuser. See the documentation provided by your storage solution provider for how to configure your mail server to enable and set up email alerts.

Each of the above items generates a log message which is generated and displayed on the Logs window.

9.6 Status window

The *Status* window provides messages about the functioning and health of each managed file system.

The Status window is shown next.



The *Status* window shows current active and past alerts.

View all status messages

Click **Status** to view all status messages. All messages are displayed most-recent first. Note that *warning* and *error* messages are displayed as *alerts.* The Status window displays messages in five categories:

- *Command Running:* These messages are gray in color and inform you of commands that are currently in progress, running. These are commands that you have entered at the manager GUI.
- *Command Successful*: These messages are green in color and identify commands that have completed successfully. You can click **Details** and then click the command link to learn about underlying commands and their syntax.
- *Info messages* - These messages are displayed in blue. Events are normal transitions that occur during the creation or management of the file system, often in response to a command entered at the GUI. A single command may cause several events to occur. An event message informs you of an event occurring at a single point in time.
- *Warning alerts*: Warnings are displayed in orange. A warning usually indicates that the file system is operating in a degraded mode, for example a target has failed over so that high availability is no longer true for that target. A warning message marks a status change that has a specific **Begin** and **End** time. A warning is active at the beginning of the status change and inactive at the end of the status change. For

example, a warning message may inform you that an OST has gone offline, and that message is active until the OST becomes operational again Not all warnings necessarily signify a degraded state; for example a target recovery to a failover server signifies that the failover occurred successfully.

- *Errors alerts*: Errors are displayed in red. An error message indicates that the file system is down or severely degraded. One or more file system components are currently unavailable, for example both primary and secondary servers for a target are not running. An error often has a remedial action you can take by clicking the button.

Common Searches

At the Status window, under the **Common Searches** drop-down menu, you can select from the searches listed next.

Note that you can modify any of the searches below. First select the search type. Then edit the string that is displayed in the *Search* field, and click **Search** or press the Enter key.

- **Search active alerts**: Display active alerts (warnings and errors) that currently reflect the state of the file system. This search will list only active warnings and errors that have not been resolved.
- **Search alerts**: Display all alerts (warnings and errors) that have been generated since the file system was created. This includes active and inactive alerts (alerts that have been resolved).
- **Search commands**: Display all commands that have successfully executed and those that are currently in process.
- **Search events**: Display all information messages (events) that have occurred since the file system was created.

Alert bar

Note that the red Alert bar briefly appears on the GUI if there are any active error or warning alerts on your system. Clicking **Details** opens the Status window and reveals the current, active alerts.

Using the Search field

The Status window also incorporates an auto-complete search function. Simply begin enter text into the search field to use this.

You can run searches using the following rules:

1. Keywords can be filtered using the equals sign (**=**) or "in" keywords. Examples:
   - severity = ERROR
   - severity in [WARNING, ERROR]
2. Filters can be joined using the "and" keyword. Example:
   - severity = ERROR and active = true

The following table lists field names, associated types, and information about that field.

Field name	Type	Field description
active	boolean	True or False depending if record is active
record_type	string	Identifying type of the record
severity	string	String indicating the severity one of ['INFO', 'DEBUG', 'CRITICAL', 'WARNING', 'ERROR']

Here is an example query:

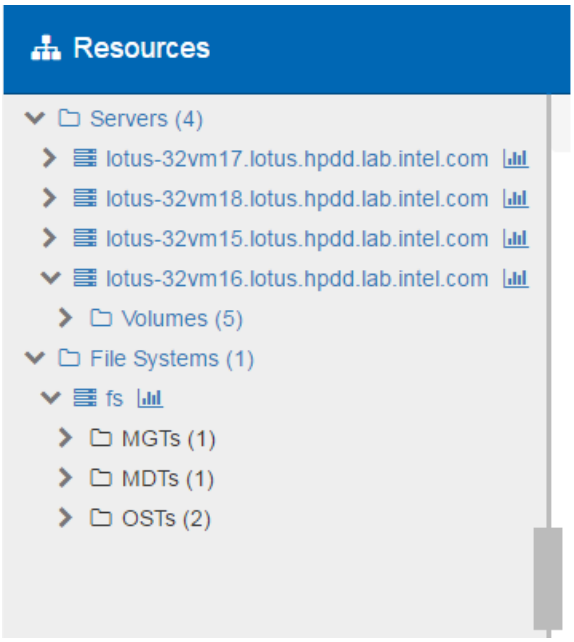
active = true

record_type = CorosyncNoPeersAlert

severity in [ERROR, WARNING]

9.7 Resources tree view

The following image is a partial display of the Resources tree view.

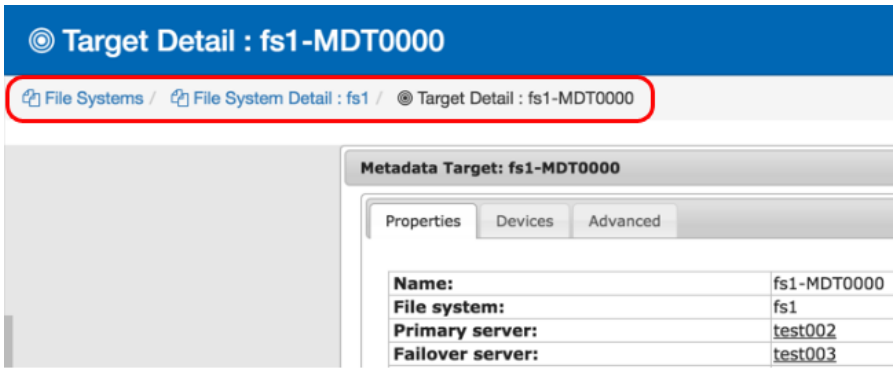


The Resources tree view is a tree listing of resources in the selected file system. It lists items in real time and lets you descend the file system hierarchy to the desired resource. You can click a resource (file systems, servers, volumes, and targets) to view that resource in the tree, and click its metrics link to view that resource's metrics. This pane displays pages when many records are available. You can size this pane by dragging its edge drawer.

9.8 Breadcrumb navigation

Breadcrumb navigation lets you see where in the hierarchy of the GUI you currently are.

Navigating between pages is now tracked by breadcrumbs. The breadcrumb path is shown at the upper-left of the page.



Breadcrumbs display a path shows you how you arrived at the current page. Breadcrumb navigation is reset when selecting a link from the Menu, or when selecting an item from the Resources tree view pane.

If you create a cycle, the breadcrumbs will automatically slice up to the current location, preventing an unnecessary build-up of items. Another key feature of this feature is that it always has a starting point of reference. If you drill down to a target and then refresh the page, the target will now be the only item in the breadcrumb list, because the page will start at that location.

If you click the Back button and the browser indicates that it is going to a previous page not in the breadcrumbs list, the new page will act as the starting breadcrumb location. This prevents a "reverse build-up" of breadcrumbs.

9.9 Alert bar

Alert Bar

This red bar briefly appears if there are any active error or warning alerts on your system. Click **Details** to open the Status window and reveal the current, active alerts.

10 Advanced topics

The following procedures are provided in this chapter:

- File system advanced settings

- Configure a new management target
- Detect and monitor existing Lustre* file systems

10.1 File system advanced settings

The following advanced settings are configurable for each file system.

Caution: Use care when changing these parameters as they can significantly impact functionality or performance. For help with these settings, contact your storage solution provider.

To access these settings:

1. At the menu bar, click the **Configuration** drop-down menu and click **File Systems**.
2. Under *Current File Systems*, select the file system in question.
3. At the File System parameters screen, click **Update Advanced Settings**.

Tunable Settings

- max_cached_mb - The maximum amount of inactive data cached by the client. Entered in megabytes. The default is 75% of RAM present on the OSS.
- max_read_ahead_mb - File read-ahead is triggered when two or more sequential reads by an application fail to be satisfied by the Linux buffer cache. The initial size of the read-ahead is 1 MB. Additional read-aheads grow linearly, and increment until the read-ahead cache on the client is full at 40 MB. This tunable setting controls the maximum amount of data read-ahead permitted on a file. Files are read ahead in RPC-sized chunks (1 MB or the size of read() call, if larger) after the second sequential read on a file descriptor. Random reads are done at the size of the read() call only (no read-ahead). Reads to non-contiguous regions of the file reset the read-ahead algorithm, and read-ahead is not triggered again until there are sequential reads again. To disable read-ahead, set this tunable to 0. The default value is 40 MB.
- max_read_ahead_whole_mb - This setting controls the maximum size of a file that is read in its entirety when the read-ahead algorithm regardless of the size of the read(). The default value is 2 MB.
- statahead_max - Many system commands will traverse a directory sequentially. To make these commands run efficiently, the directory stat-ahead and AGL (asynchronous glimpse lock) are enabled to improve the performance of traversing. This tunable sets the maximum number of files that can be pre-fetched by the stat-ahead thread. The default value is 32 bytes. Set to this 0 to disable.

Timeout Settings

These setting are pre-set to default values. Most of these settings are automatically adaptive so that a superuser should not need to change them. These settings are the

same timeout settings discussed in the Lustre Operations Manual.

- at_early_margin - Time in seconds of an advance queued request timeout at which the server sends a request to the client to extend the timeout time. The default value is 5.
- at_extra - Incremental time in seconds that a server requests the client to add to the timeout time when the server determines that a queued request is about to timeout. The default value is 30.
- at_history - Time period in seconds within which adaptive timeouts remember the slowest event that occurred. The default value is 600.
- at_max - Adaptive timeout upper limit in seconds. The default value is 600. Set to 0 to disable adaptive timeouts.
- at_min - Adaptive timeout lower limit or minimum processing time reported by a server, in seconds. Default value is 0.
- Idlm_timeout - Lustre distributed lock manager timeout: Time in seconds that a server will wait for a client to reply to an initial AST (local cancellation request). The default value is 20 seconds for an OST and 6 seconds for a MDT.
- timeout - Time in seconds that a client waits for a server to complete an RPC. The default value is 100.

10.2 Configure a new Management Target

The MGT is normally configured while creating the file system and doesn't need to be created separately on *MGT* window.

However, to configure the management target perform these steps:

1. At the menu bar, click the **Configuration** drop-down menu and click **MGTs** to display the *MGT Configuration* window.
2. Under *New MGT*, click **Select storage** and select the server for the MGT.

   Note: The MGT and metadata target (MDT) can be located on the same server. However, they cannot be located on the same volume on a server.
3. Click **+ Create new MGT** to create the new MGT.

10.3 Add additional Metadata Targets

You can add additional MDTs when creating the file system and later, after the file system has been created.

DNE stands for Distributed Namespace. DNE allows the Lustre namespace to be divided

across multiple metadata targets. This enables the size of the namespace and metadata throughput to be scaled with the size of the file system and the number of servers. The primary metadata target in a Lustre file system is MDT0. Added MDTs are indexed as MDT1, MDT2, and so on.

To add additional MDT(s):

1. At the top menu bar, click **Configuration** > **File Systems**.
2. Under **Current File Systems**, select the file system you wish to modify.
3. Under **Metadata Target**, click **+ Create MDT (DNE).**
4. At the **Create MDT** pop-up window, select the volume you wish to use as this new MDT. Click **Create**. After a moment, the new MDT will be listed on the file system window, under Metadata Target. You can create additional MDTs; simply repeat steps 3 and 4. When you have created the desired MDT(s), perform step 5.
5. Log into a client node and mount the Lustre file system. Then at the command line, for each added MDT beyond the primary MDT, enter the following command:

   lfs mkdir -i *n* <lustre_mount_point>/<parent_folder_to_contain this_MDT>

   where the -i indicates that the following value, *n* is the MDT index. The first added MDT will be index 1.

The new MDT is installed. Users can now create subdirectories supported by each added MDT with the following command, as an example:

mkdir <lustre_mount_point>/<parent_folder_to_contain this_MDT>/<subdirectory_name>

Note: Any added MDT you create will be unavailable for use as an OST.

11 Using the Intel® Manager for Lustre* command line interface

Intel® Manager for Lustre* software includes a command line interface (CLI) which can be used instead of the GUI to communicate with the Representational State Transfer (REST)-based API underlying the software GUI. The CLI is intended to be used in shell scripts by superusers and power users.

WARNING: For Lustre* file systems created and managed by Intel® Manager for Lustre* software, the only supported command line interface is the CLI provided by Intel® Manager for Lustre* software. Modifying such a Lustre file system manually from a UNIX shell will interfere with the ability of the Intel® Manager for Lustre* software to manage and monitor the file system.

This chapter provides the following procedures and information:

11.1 Accessing the command line interface

To access the Intel® Manager for Lustre* CLI:

1. Use SSH to log into the manager server as the UNIX superuser. Log in using your superuser account.
2. Enter CLI commands on the UNIX command line.

WARNING: To manage Lustre file systems from the command line, you must use the Intel® Manager for Lustre* command line interface. Modifying a file system manually from a shell on a storage server will interfere with the ability of Intel® Manager for Lustre* to manage and monitor the file system.

11.2 Creating a configuration file with login information

Although a superuser can enter a login name and password on the command line each time the Intel® Manager for Lustre* CLI is used, accessing login information in a configuration file is more convenient and more secure.

To set up an optional configuration file, complete these steps:

1. Create a configuration file $HOME/.chroma on the server hosting Intel® Manager for Lustre* software.
2. Edit the file to include content as shown below:

```
[chroma]
username = <user name of file system administrator>
password = <password>
```

Note: To minimize security risks, modify the permissions of the .chroma file to allow only the file owner to read from and write to it, using:

```
$ chmod 0600 ~/.chroma
```

11.3 Getting help for CLI commands

To access documentation for the CLI commands, use the chroma –h command shown next:

```
# chroma --help
usage: chroma [--api_url API_URL] [--username USERNAME]
 [--password PASSWORD]
 [--output {human,json,xls,yaml,csv,tsv,html,xlsx,ods}]
 [--nowait] [--help]

 {volume,fs,target,tgt,vol,cfg,oss,mgt,ost,nid,server,
 mgs,srv,filesystem,mds,configuration,mdt}
...
```

CLI

positional arguments:

```
 {volume,fs,target,tgt,vol,cfg,oss,mgt,ost,nid,server,mgs,srv
,
 filesystem,mds,configuration,mdt}
 configuration (cfg)
 dump, load
 filesystem (fs) list, show, add, remove, start, stop,
 detect, mountspec
 nid update, relearn
 server (srv, mgs, mds, oss)
 show, list, add, remove
 target (tgt, mgt, mdt, ost)
 list, show, add, remove, start, stop
 volume (vol) list, show
```

optional arguments:

```
 --api_url API_URL Entry URL for Chroma API
 --username USERNAME Chroma username
 --password PASSWORD Chroma password
```

```
--output, -o {human,json,xls,yaml,csv,tsv,html,xlsx,ods}
Output format
--nowait, -n Don't wait for jobs to complete
--help, -h Show this help message and exit
```

To view the command options available specific to a file system, enter:

```
# chroma filesystem --help
usage: chroma filesystem [-h]

{detect,show,list,stop,remove,start,add,
context,mountspec}
...
```

positional arguments:

```
{detect,show,list,stop,remove,start,add,context,mountspec}
list list all file systems
show show a filesystem
add add a filesystem
remove remove a filesystem
start start a filesystem
stop stop a filesystem
detect detect all file systems
mountspec mountspec for filesystem
context filesystem_name action (e.g. ost-list,
vol-list, etc.)
```

optional arguments:

```
-h, --help show this help message and exit
```

To show help for the server argument, enter:

```
# chroma server-show --help
usage: chroma server show [-h] server
```

positional arguments:

```
server
```

optional arguments:

```
-h, --help show this help message and exit
```

11.4 CLI command examples

This section includes examples of common operations executed using the CLI.

Note: Operations that modify the file system configuration can only be executed by a file system superuser. For a convenient way to access login information in a configuration file, see Creating a configuration file containing login information. If a configuration file containing the superuser's login information does not exist, include the `--username` and `--password` parameters in the CLI command.

To add the file system jovian to Intel® Manager for Lustre* , enter:

```
# chroma fs-add jovian --mgt autonoe:/dev/mapper/LustreVG-mgs
--mdt autonoe:/dev/mapper/
LustreVG-mdt --ost thyone:/dev/mapper/LustreVG-ost0 --ost
thyone:/dev/mapper/LustreVG-ost1 --ost thyone:/dev/mapper/
LustreVG-ost2 --ost thyone:/dev/mapper/LustreVG-ost3
```

To add a new server to be monitored and managed:

```
# chroma server-add thyone.jovian.private --server_profile
base_managed
Setting up host thyone.jovian.private: Finished
```

To add a new server to be monitored only:

```
# chroma server-add thyone.jovian.private --server_profile
base_monitored
Setting up host thyone.jovian.private: Finished
```

To list known servers:

```
# chroma server-list
| id | fqdn | state | nids | last_contact |
```

```
| 4 | autonoe.jovian.private | lnet_up | 10.141.255.2@tcp0 |
20:10:46 |
| 5 | thyone.jovian.private | lnet_up | 10.141.255.3@tcp0 |
20:10:46 |
```

To list known OSTs:

```
# chroma ost-list
| id | name | state | primary_path |
| 3 | jovian-OST0002 | mounted | thyone.jovian.private:/dev/
mapper/LustreVG-ost0 |
| 4 | jovian-OST0001 | mounted | thyone.jovian.private:/dev/
mapper/LustreVG-ost1 |
| 5 | jovian-OST0000 | mounted | thyone.jovian.private:/dev/
mapper/LustreVG-ost2 |
| 6 | jovian-OST0003 | mounted | thyone.jovian.private:/dev/
mapper/LustreVG-ost3 |
```

To list targets on a given server, limiting to primary targets:

```
# chroma server autonoe target-list --primary
| id | name | state | primary_path |
| 1 | MGS | mounted | autonoe.jovian.private:/dev/mapper/
LustreVG-mgs |
| 2 | jovian-MDT0000 | mounted | autonoe.jovian.private:/dev/
mapper/LustreVG-mdt |
```

To obtain client mount information:

```
# chroma filesystem-mountspec jovian
10.141.255.2@tcp0:/jovian
```

To detect existing (non-managed) Lustre file systems on servers that have been added to the Command Center, enter:

```
# chroma filesystem-detect
```

12 Errors and troubleshooting

The following topics are discussed in this chapter:

12.1 Unexpected file system events

This section discusses several unwanted file system events and how Intel® Manager for Lustre* software responds to them.

A server's connection to a storage target is lost

Immediate file system consequences:	Lustre clients will block if they have requested a file from an unavailable OST. The block will continue until connection to the OST is restored and the OST is again fully online. For OSTs that are still connected to their servers, client access continues unaffected.
Manager software / Peer server response:	No automatic failover. No alerts.
Suggested remedies:	Repair the connection to the target. In the meantime, the superuser may manually fail the target over to the peer server.

A server's connection to LNet is lost

Immediate file system consequences:	Lustre clients will block waiting for the connection to be re-established. Those portions of the file system that are presented by the affected server are unavailable until then.
Manager software / Peer server response:	No automatic failover. No alerts.
Suggested remedies:	Repair the server's connection to LNet. In the meantime, the superuser may manually fail the target over to the peer server.

Manager software connection to a server (via the management network, ring0) is lost

Immediate file system consequences:	No direct file system impact; the file system remains operational. However, Intel® Manger for Lustre* software can no longer manage or monitor the server.
Manager	Alerts to administer regarding loss of network connection to server.

software / Peer server response:	
Suggested remedies:	Re-establish the management network connection to the server.

A Lustre server loses connectivity with the power control device for its peer server (IPMI or PDU)

Immediate file system consequences:	None. The file system continues to operate normally. In the event of a peer server failure, the server that has lost connectivity to power control will be unable to power off the failed server and assume responsibility for its resources.
Manager software / Peer response:	No response to the loss of connectivity if the file system is operating normally. In the event of a server failure, automatic failover of Lustre targets from the failed server may be disabled.
Suggested remedies:	Repair the network link to power control (IPMI or PDU).

The Intel® Manager for Lustre* loses connection with a server's power control device (IPMI or PDU)

Immediate file system consequences:	The software's ability to shut down the server is lost.
Manager software/peer server response:	Alerts to administer regarding loss of connection to power control device.
Suggested remedies:	Restore the connection between the Manager software server and affected server's power control device.

A crossover cable between servers is disconnected or the network is down

Immediate file system consequences:	This is the loss of the ring1 network link, but the ring0 link (the management network) provides complete redundancy. The file system is not affected.
Manager software / Peer server response:	No automatic failover. No alerts.
Suggested remedies:	Replace/reconnect the cross-over cable, restore the network.

A primary server's OS kernel crashes

Immediate file system consequences:	Each server is used as both a primary and secondary server. Temporarily delayed access to served storage as failover occurs.
Manager software / Peer server response:	Peer server performs STONITH, failover occurs.
Suggested remedies:	None needed by Admin. Successful STONITH causes the server to be rebooted.

LBUG, a Lustre crash on a server

Immediate file system consequences:	This will also crash Linux on the affected server. Temporarily delayed access to served storage as failover occurs.
Manager software / Peer server response:	Peer server performs STONITH, failover occurs.
Suggested remedies:	No Admin action needed.

The primary server spontaneously reboots

Immediate file system consequences:	Temporarily delayed access to served storage as failover occurs.
Manager software / Peer server response:	Peer server performs STONITH, failover occurs.
Suggested remedies:	No Admin action needed.

The management network (ring0) and a peer crossover network (ring1) are both down

Immediate file system consequences:	The file system is not directly affected and client operations may continue. Affected peer servers may attempt STONITH.
Manager software / Peer server response:	Peer server performs STONITH and failover occurs. However, each affected server may attempt STONITH on its peer.
Suggested	This condition is unlikely and unstable. The superuser needs to restore

remedies: network connections for the management network and the cross-over link between affected servers.

12.2 Running Intel® Manager for Lustre* diagnostics

If Intel® Manager for Lustre* software is not operating normally and you require support from Intel® customer support, you may be asked to run chroma-diagnostics on any servers that are suspected of having problems, and/or on the server hosting the Intel® Manager for Lustre* dashboard. The results of running the diagnostics should be attached to the ticket you are filing describing the problem. These diagnostics are described next.

Run diagnostics

1. Log into the server in question. Admin login is required in order to collect all desired data.
2. Enter the following command at the prompt:

```
#chroma-diagnostics
```

This command generates a compressed tar.lzma file that you can email to Intel® customer support. Following are sample displayed results of running this command. (The resulting tar.lzma file will have a different file name.)

```
Collecting diagnostic files

Detected devices
Devices monitored
Listed installed packages
Listed cibadmin --query
Listed: pcs config show
Listed: crm_mon -1r
Finger printed Intel Manager for Lustre installation
Listed running processes
listed PCI devices
listed file system disk space.
listed cat /proc/cpuinfo
listed cat /proc/meminfo
listed cat /proc/mounts
listed cat /proc/partitions
Listed hosts
Copied 1 log files.
Compressing diagnostics into LZMA (archive)

Diagnostic collection is completed.
Size: 16K  /var/log/diagnostics_20150623T160338_lotus-4vm15.iml.intel.com.ta

The diagnostic report tar.lzma file can be sent to Intel(R) Manager for Lust
```

You can also decompress the file and examine the results. To unpack and extract the files, use this command:

```
tar --lzma -xvpf <file_name>.tar.lzma
```

Help for chroma-diagnostics

Generally, if requested you should run this command without options, as this will generate the needed data. Enter `chroma-diagnostics -h` to see help for this command, as follows:

```
# chroma-diagnostics -h
usage: chroma-diagnostics [-h] [--verbose] [--days-back DAYS_BACK]
Run this to save a tar-file collection of logs and diagnostic output.
The tar-file created is compressed with lzma.
Sample output: /var/log/diagnostics_<date>_<fqdn>.tar.lzma
optional arguments:
 -h, --help   show this help message and exit
 --verbose, -v   More output for troubleshooting.
 --days-back DAYS_BACK, -d DAYS_BACK
            Number of days back to collect logs. default is 1. 0 would mean to
```

13 Glossary

chroma-agent. An executable provided with the Intel® Manager for Lustre* software that can be installed as a service on Lustre* servers to enable monitoring of Lustre file systems not created by the Intel® Manager for Lustre* software.

Lustre clients. Lustre clients are computational, visualization, or desktop nodes that are running Lustre client software, allowing them to mount the Lustre file system.

Management target (MGT). The MGT stores configuration information for all the Lustre file systems in a cluster and provides this information to other Lustre components. Each Lustre object storage target (OST) contacts the MGT to provide information, and Lustre clients contact the MGT to retrieve information.

Metadata target (MDT). Each Lustre file system has one MDT. The MDT stores metadata (such as file names, directories, permissions, and file layout) for attached storage and makes them available to clients.

Object storage target (OST). User file data is stored in one or more objects that are located on separate OSTs in the Lustre file system. The number of objects per file is configurable by the user and can be tuned to optimize performance for a given workload.

Storage server. A server on which an MGT, MDT, or OST is located.

Target. See metadata target, management target, object storage target.

Volumes. (also called LUNs or block devices) are the underlying units of storage used to create Lustre file systems. Each Lustre target corresponds to a single volume. If servers in

the volume have been configured for high availability, primary and failover servers can be designated for a Lustre target. Only volumes that are not already in use as Lustre targets or local file systems are shown. A volume may be accessible on one or more servers via different device nodes, and it may be accessible via multiple device nodes on the same host.

14 Getting Help

If you need help with the Intel® Manager for Lustre* software, contact your storage solution provider.

Index

- A -

- B -

- C -

- D -

- E -

- F -

- G -

- I -

- L -

- M -

- N -

- O -

- P -

- S -

- U -

- V -