Cray ClusterStor data services User Guide 2.0 (S-1239) Revision A

Part Number: S-1239
Published: December 2021
Edition: 1

Cray ClusterStor data services User Guide 2.0 (S-1239) Revision A

Abstract
This guide contains information about ClusterStor data services for non-administrator ClusterStor users.

Part Number: S-1239
Published: December 2021
Edition: 1

Table of contents

Notices

The information contained herein is subject to change without notice. The only warranties for Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

Confidential computer software. Valid license from Hewlett Packard Enterprise required for possession, use, or copying. Consistent with FAR 12.211 and 12.212, Commercial Computer Software, Computer Software Documentation, and Technical Data for Commercial Items are licensed to the U.S. Government under vendor's standard commercial license.

Links to third-party websites take you outside the Hewlett Packard Enterprise website. Hewlett Packard Enterprise has no control over and is not responsible for information outside the Hewlett Packard Enterprise website.

Acknowledgments

Kubernetes® is a registered trademark The Linux Foundation in the United States and/or other countries.

Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries.

The Lustre® trademark is jointly owned by OpenSFS and EOFS.

UNIX® is a registered trademark of The Open Group.

Support and other resources

Accessing Hewlett Packard Enterprise Support

- For live assistance, go to the Contact Hewlett Packard Enterprise Worldwide website:

  https://www.hpe.com/info/assistance

- To access documentation and support services, go to the Hewlett Packard Enterprise Support Center website:

  https://www.hpe.com/support/hpesc

Information to collect

- Technical support registration number (if applicable)
- Product name, model or version, and serial number
- Operating system name and version
- Firmware version
- Error messages
- Product-specific reports and logs
- Add-on products or components
- Third-party products or components

Accessing updates

- Some software products provide a mechanism for accessing software updates through the product interface. Review your product documentation to identify the recommended software update method.
- To download product updates:

  Hewlett Packard Enterprise Support Center

  https://www.hpe.com/support/hpesc

  Hewlett Packard Enterprise Support Center: Software downloads

  https://www.hpe.com/support/downloads

  My HPE Software Center

  https://www.hpe.com/software/hpesoftwarecenter
- To subscribe to eNewsletters and alerts:

  https://www.hpe.com/support/e-updates
- To view and update your entitlements, and to link your contracts and warranties with your profile, go to the Hewlett Packard Enterprise Support Center More Information on Access to Support Materials page:

  https://www.hpe.com/support/AccessToSupportMaterials

ⓘ **IMPORTANT:**

Access to some updates might require product entitlement when accessed through the Hewlett Packard Enterprise Support Center. You must have an HPE Passport set up with relevant entitlements.

Remote support

Remote support is available with supported devices as part of your warranty or contractual support agreement. It provides intelligent event diagnosis, and automatic, secure submission of hardware event notifications to Hewlett Packard Enterprise, which initiates a fast and accurate resolution based on the service level of your product. Hewlett Packard Enterprise strongly recommends that you register your device for remote support.

If your product includes additional remote support details, use search to locate that information.

HPE Get Connected

https://www.hpe.com/services/getconnected

HPE Pointnext Tech Care

https://www.hpe.com/services/techcare

HPE Complete Care

https://www.hpe.com/services/completecare

Warranty information

To view the warranty information for your product, see the links provided below:

HPE ProLiant and IA-32 Servers and Options

https://www.hpe.com/support/ProLiantServers-Warranties

HPE Enterprise and Cloudline Servers

https://www.hpe.com/support/EnterpriseServers-Warranties

HPE Storage Products

https://www.hpe.com/support/Storage-Warranties

HPE Networking Products

https://www.hpe.com/support/Networking-Warranties

Regulatory information

To view the regulatory information for your product, view the Safety and Compliance Information for Server, Storage, Power, Networking, and Rack Products, available at the Hewlett Packard Enterprise Support Center:

https://www.hpe.com/support/Safety-Compliance-EnterpriseProducts

Additional regulatory information

Hewlett Packard Enterprise is committed to providing our customers with information about the chemical substances in our products as needed to comply with legal requirements such as REACH (Regulation EC No 1907/2006 of the European Parliament and the Council). A chemical information report for this product can be found at:

https://www.hpe.com/info/reach

For Hewlett Packard Enterprise product environmental and safety information and compliance data, including RoHS and REACH, see:

https://www.hpe.com/info/ecodata

For Hewlett Packard Enterprise environmental information, including company programs, product recycling, and energy efficiency, see:

https://www.hpe.com/info/environment

Documentation feedback

Hewlett Packard Enterprise is committed to providing documentation that meets your needs. To help us improve the documentation, use the Feedback button and icons (located at the bottom of an opened document) on the Hewlett Packard Enterprise Support Center portal (**https://www.hpe.com/support/hpesc**) to send any errors, suggestions, or comments. All document information is captured by the process.

About the Cray ClusterStor data services User Guide

The Cray ClusterStor data services User Guide S-1239 provides useful information and procedures to ClusterStor users.

Table 1: Record of Revision

Publication Title	Date	Updates
Cray ClusterStor Data Services User Guide 2.0 (S-1239) Rev A	December 2021	Corrected first paragraph of "Write Data Movement Policies". Added "Acknowledgments" topic.
HPE Cray ClusterStor Data Services User Guide 2.0 (S-1239)	October 2021	Removed content about the deprecated `pfind` and `filestatus` tools.
Cray ClusterStor Data Services User Guide S-1239 (1.1)	May 2021	New `--json` option for `filestatus`
Cray ClusterStor Data Services User Guide S-1239 (1.0)	January 2021	Added `pfind` content, new section on command prompt conventions, new content on `filestatus` tool
Cray ClusterStor Data Services User Guide S-1239 (0.3)	August 2020	Initial release as a separate guide from Cray ClusterStor Data Services User and Administration Guide S-1237

Scope and Audience

This publication is for users of ClusterStor storage systems that include ClusterStor data services. This publication assumes that the reader is familiar Lustre commands, terminology, and architecture. This guide contains procedures and reference information to support users of ClusterStor data services systems running release 2.0. The reference material includes a thorough explanation of the aims, design principles, and architecture of ClusterStor data services. This guide includes procedures for using ClusterStor data services features.

Typographic Conventions

`Monospace`	Indicates program code, reserved words, library functions, command-line prompts, screen output, file and path names, and other software constructs.
`Monospaced Bold`	Indicates commands that must be entered on a command line or in response to an interactive prompt.
Oblique or *Italics*	Indicates user-supplied values in commands or syntax definitions.
Proportional Bold	Indicates a GUI Window, GUI element, cascading menu (Ctrl > Alt > Delete), or key strokes (press Enter).
`\` (backslash)	At the end of a command line, indicates a shell or command-line continuation character (lines joined by a backslash are parsed as a single line).

Trademarks

Command Example Conventions

Host names and accounts in command prompts

The host name in a command prompt indicates on what host or type of host to run the command. The prompt also indicates the account that must run the command.

- The `root` or super-user account always has the `#` character at the end of the prompt.
- Any non- `root` account is indicated with a `$` . A user account that is not `root` or `admin` is referred to as "user".

`install#`	Run the command as `root` on the system that hosts, serves, and executes the files that install ClusterStor data services on the three management nodes.
`ext-adm#`	Run the command as `root` on an external administration system for the ClusterStor data services cluster.
`csms$`	Run the command on a host that has (or will soon have) the CSMS CLI installed and initialized.
`VM-W#`	Run the command on a ClusterStor data services Kubernetes worker node virtual machine as `root`.
`CDS#`	Run the command on a ClusterStor data services node as `root`.
`CDS$`	Run the command on a ClusterStor data services node as a user.
`CDS1#`	Run the command on the primary ClusterStor data services node as `root`.
`CDS1$`	Run the command on the primary ClusterStor data services node as a user.
`MGMT0#`	Run the command on the primary ClusterStor management node as `root`.
`MGMT0$`	Run the command on the primary ClusterStor management node as `admin`.
`MGMT1#`	Run the command on the secondary ClusterStor management node as `root`.
`MGMT1$`	Run the command on the secondary ClusterStor management node as `admin`.
`OSS#`	Run the command on an OSS node as `root`.
`OSS$`	Run the command on an OSS node as a user.
`MGS#`	Run the command on an MGS node as `root`.
`MGS$`	Run the command on an MGS node as a user.
`MDS#`	Run the command on an MDS node as `root`.
`MDS$`	Run the command on an MDS node as a user.
`sw0#`	Run the command on the primary management switch with administrative privileges.
`sw1#`	Run the command on the secondary management switch with administrative privileges.
`client$`	Run the command on a Lustre client node as any user.
`client#`	Run the command on a Lustre client node as `root`.
`user@`*`hostname`*`$`	Run the command on the specified system as any user.

Lustre file system names

The name of the Lustre file system seen in command examples is `cls12345`, with a mount point of /lus on the Lustre clients. The following example demonstrates this convention:

```
client$ lfs df -h
UUID                        bytes        Used   Available  Use% Mounted on
cls12345-MDT0000_UUID         2.0T       59.1G        1.9T    4% /lus[MDT:0]
cls12345-MDT0001_UUID         2.0T       97.6M        1.9T    1% /lus[MDT:1]
cls12345-OST0000_UUID       112.0T        5.1T      105.8T    5% /lus[OST:0]
cls12345-OST0001_UUID       112.0T        5.1T      105.8T    5% /lus[OST:1]
cls12345-OST0002_UUID        15.3T      425.8G       14.8T    3% /lus[OST:2]
cls12345-OST0003_UUID        15.3T      423.4G       14.8T    3% /lus[OST:3]

filesystem_summary:       254.6T       11.0T      241.1T    5% /lus
```

Command Prompt inside a Kubernetes pod

If executing a shell inside a container of a Kubernetes pod where the pod name is *podName*, the prompt changes to indicate that it is inside the pod. Not all shells are available within every pod, the following is an example using a commonly available shell.

```
kubectl exec -it podName /bin/sh
pod#
```

Directory Path in Command Prompt

Example prompts do not include the directory path, because long paths can reduce the clarity of examples. Usually, the commands can run in any directory. When a command must run within a specific directory, examples use the `cd` command to change into the necessary directory.

For example, here are actual prompts as they appear on the system:

```
client:~ # cd /etc
client:/etc# cd /var/tmp
client:/var/tmp# ls file
```

And here are the same prompts as they appear in this publication:

```
client# cd /etc
client# cd /var/tmp
client# ls file
```

What ClusterStor data services Provides

ClusterStor E1000 introduces an additional flash storage tier into what has traditionally been a single-tier disk-based Lustre file system. ClusterStor data services provides flexible, fast, automated, and scalable movement of data between these tiers. As a result, users can fully use each tier with minimal effort.

Users of large and tiered Lustre storage systems have four vital needs:

- **Converged treatment of tiered data movement through the Lustre Hierarchical Storage Management (HSM) infrastructure** : The HSM functionality in Lustre cannot change data location or layouts within a single Lustre file system. It can only change them when moving a file to an external archive. Users of tiered storage systems, however, must be able to move data between tiers within the same Lustre file system. ClusterStor data services introduces a new `lfs migrate` option. This new option uses the Lustre HSM infrastructure to parallelize data movement on external dedicated nodes, yet is also able to move file data between different storage tiers.
- **Performant, scalable, efficient data movement**: Lustre users need a data movement solution optimized for extracting the full performance potential from the flash tier. ClusterStor data services parallelizes file data migration across multiple dedicated nodes.
- **Scalable and efficient external data movement request management:** Lustre itself only supports manual requests for data movement using `lfs` commands. It does not include, for example, an internal policy engine, forcing users to configure and integrate a third-party solution (software and hardware). Lustre does not provide a programmatic API for submitting, prioritizing, and managing data movement requests.
- **Fast and scalable search capabilities:** Data movement policy engines and user file searches on large Lustre namespaces both require search performance beyond what `find`, `lfs find`, and other search tools can deliver.

Lustre Tiering Limitations

Separate storage tiers can be accessed as separate OST pools (for example, separate OST pools for flash and disk drives) within Lustre without ClusterStor data services. ClusterStor users, however face two design issues when trying to use existing Lustre mechanisms for data management:

1. The `lfs hsm_archive` and `lfs hsm_restore` HSM commands initiate external copytools to move data between Lustre and a separate, external archive (such as a tape library). The `lfs mirror` and `lfs migrate` commands, however, do not work this way. Therefore, copytools are not initiated for tier management which requires data movement **within** a single Lustre file system.

2. These HSM commands rely on the MDS HSM Coordinator queue. The MDS HSM Coordinator is an inflexible first-in, first-out (FIFO) queue that does not have prioritization and other capabilities.

These limitations lead to several specific problems when users try to use these commands on tiered Lustre storage systems at scale:

- **Data movement with `lfs mirror` and `lfs migrate` is highly inefficient and does not scale:** Although these commands can move data between OSTs and OST pools, Lustre does not have internal mechanisms for queuing or distributing data transfers. Every data movement request which uses these commands sends all the data serially through the requesting node. This bottlenecks file I/O throughput and can lead to long wait times for large file movements.

- **Requests cannot be properly prioritized:** Since the Lustre HSM Coordinator Queue operates as a FIFO, requests cannot be prioritized or reprioritized. The result is a mismatch between the current data movement needs of the Lustre system and what data is moved at that moment.

- **The Lustre internal HSM coordinator queue consumes MDS resources:** The MDS actively maintains the list, tracks timeouts, reissues requests, cancels completed items, and reports status. All of these operations divert compute resources on the MDS away from its primary function: providing metadata to Lustre clients.

- **No internal policy engine:** Lustre does not have methods for automating requests based on policies. Users must configure and integrate an external policy engine.

- **Movement decisions require significant understanding of Lustre layouts** by the individual users issuing the data movement requests. This is because there are no standard rules or templates to follow.

- **Lustre does not have a data movement API:** Users have no way to request data movements in a programmatic fashion.

How ClusterStor data services Improves Upon Lustre

ClusterStor data services moves the entire HSM and tiering workflow out of Lustre and into a set of external, horizontally scalable microservices. These microservices are accessible through a REST API and the traditional `lfs` commands. ClusterStor data services expands the current HSM functionality within Lustre to efficiently manage hard disk and flash storage tiers. ClusterStor data services helps customers realize the full performance potential of the flash storage tier within hybrid ClusterStor systems.

What ClusterStor data services brings to existing ClusterStor and Lustre

Efficient and scalable data movement: large single file migrations from one tier to another are distributed over multiple dedicated transfer nodes, turning serial data transfer operations into faster parallel ones. Also, ClusterStor data services moves the data movement request queue from the MDS to a database managed outside of Lustre. As a result, requests can be prioritized and scheduled flexibly and the MDS can now reserve compute resources for fulfilling metadata requests. Also, performance is improved since ClusterStor data services can serve multiple data movement requests concurrently.

A programmatic way to move file data within Lustre through a REST API : ClusterStor data services allows requests for moving data within Lustre to originate outside of it. The result is better integration with automated data movement requesters, such as policy engines or job schedulers.

Integrated, full featured tiered storage management: ClusterStor data services incorporates data movement commands which have not traditionally been within the Lustre HSM mechanism, namely `lfs migrate`. ClusterStor data services also expands the `lfs hsm` command path to handle internal tiering management. This expansion solves a major problem with attempting to use Lustre to manage flash and hard disk tiers within a single file system. ClusterStor data services integrates the Lustre tiering mechanism with the Lustre internal data movement commands.

The ability to efficiently change the layouts of Lustre files by externalizing the `lfs migrate`: ClusterStor data services supports a new HSM version of `lfs migrate` and parallelizes the data movement.

Automatic space management of a flash tier in a hybrid disk-flash system: ClusterStor data services provides software tools and example policies for automating flash tier management. This automation lessens the administrative burden for users of hybrid (flash and disk) ClusterStor systems.

Efficient searches of file system metadata for reporting, analysis, or fileset operations: ClusterStor data services maintains an internal database of file system metadata. ClusterStor data services provides a database query tool (Query) for administrators to use from any Lustre file system client. The Query tool can efficiently and quickly provide lists of files matching search criteria in CSV or JSON form, as well as summary information.

Which Lustre HSM Requests Are Emitted to ClusterStor data services

In ClusterStor data services, the Emitter is able to forward all types of Lustre HSM requests, except for status queries through the `lfs hsm_action file` command. In this release, issuing this command has no effect.

The following table lists the types of Lustre HSM requests which are forwarded to ClusterStor data services by the Emitter. Refer to the Lustre documentation at www.lustre.org/documentation for more information on these `lfs` commands.

Table 2: Lustre HSM requests forwarded to ClusterStor data services

Lustre Request	Command	Notes
Move a file from the current OST or OST pool to another one.	`lfs migrate --hsm file`	
Cancel a file movement request.	`lfs hsm_cancel file`	Not implemented in current release

Manual Data Migration

ClusterStor data services can provide automatic, policy-driven migration of file data. Sometimes, however, file data must be manually migrated between OSTs for storage management or tiering purposes. To cover this use case, ClusterStor data services offers manual data migration.

Users can initiate a manual request to migrate file data by running an `lfs migrate --hsm` command on a Lustre client. The new `--hsm` flag directs Lustre to use ClusterStor data services (or any external coordinator) to move the data.

By design, ClusterStor data services operates transparently to Lustre and is invisible to ClusterStor users. The new HSM-based migration command operates asynchronously, like all other Lustre HSM commands.

See the following topics for commands to query the status of a manual migration request:

- Check File Movement Status with `lfs getstripe`
- Check File Movement Status Using File `ctime`

Request a File Migration from a Lustre Client

Lustre now provides an `--hsm` option for the `lfs migrate` command. This option instructs Lustre to forward the data movement task to external software instead of the performing it on the Lustre client. ClusterStor data services uses this option to ingest data movement requests from Lustre clients.

In general, any `lfs migrate` option will work with ClusterStor data services.

Prerequisites

- Log into a Lustre client node

Procedure

Submit a request to ClusterStor data services from a Lustre client to migrate a file residing in one tier to another tier.

In the following command example, the path *path/ to/ file* specifies a file currently on the flash pool, and `disk` is the destination pool.

```
client$ lfs migrate --hsm --pool disk /lus/path/to/file
```

Lustre will forward the data movement request to ClusterStor data services, which will store and perform the data movement request.

Check File Movement Status with `lfs getstripe`

Check the progress of a file migration request by retrieving the layout information from Lustre. If the OST pool name returned by Lustre for that file is different from that the original pool, then the movement request has been completed. If the OST pool location has not changed, then the movement request is either still pending, in progress, or was canceled.

Prerequisites

Issue a manual data movement request from one storage tier to another, following the instructions in Request a File Migration from a Lustre Client.

Procedure

1. Query Lustre for the layout information of the file for which a data movement request has been issued. The `lfs getstripe` command is issued from a Lustre client.

   ```
   client$ lfs getstripe -p example_file
   flash
   ```

   The `-p` or `--pool` option causes `lfs getstripe` to only return the name of the OST pool in which the file resides. If the file is not assigned to a specific pool, the previous command returns an empty line. In the previous example, `example_file` is in the flash tier, but a request has been submitted to move it to the disk tier.

2. Repeat the previous command as necessary until Lustre indicates that file has moved from the source OST pool to the destination OST pool.

   ```
   client$ lfs getstripe -p example_file
   disk
   ```

Check File Movement Status Using File `ctime`

Check the progress of a file movement request from one tier to another by querying the file metadata change time (`ctime`). This attribute updates on layout changes, and thus will update after migrating a file from one OST pool to another.

Prerequisites

Identify a file for data movement from one storage tier to another (for example, from the `flash` OST pool to the `disk` OST pool).

Procedure

1. Query the file `ctime` of a file for which a data movement request will be issued.

   ```
   client$ ls -lc /lus/test_dir/example_file
   -rw-rw-r-- 1 example_file 1318 Dec 12 04:00 /lus/test_dir/example_file
   ```

2. Issue a manual data movement request for the file. In the following example, the data movement requested is to move example_file from the disk tier to the flash tier.

   ```
   client$ lfs migrate --pool flash --hsm /lus/test_dir/example_file
   ```

3. Repeat the `ls -lc` command as necessary until the file `ctime` returned by Lustre indicates that the file migration has completed. The new `ctime` will be later than the one returned in Step 1.

   ```
   client$ ls -lc /lus/test_dir/example_file
   -rw-rw-r-- 1 example_file 1318 Dec 12 04:03 /lus/test_dir/example_file
   ```

Write Data Movement Policies

ClusterStor administrators can grant users permission to add their own policy files to the *LUSTRE_MOUNT_POINT*/.cray/cds/policy/ directory. Administrators can also move user-written policy files to this directory on behalf of a user. The ClusterStor data services Policy Engine will only process policy files which are in this directory.

Use this workflow to write policy files for automating data movement or deletion.

Prerequisites

- Obtain permission from the ClusterStor administrator to add files to the *LUSTRE_MOUNT_POINT*/.cray/cds/policy/ directory.
- Read the section "ClusterStor data services Policies" in Cray ClusterStor data services Administration Guide S-1237.

Procedure

1. Create a policy file per section "ClusterStor data services Policies" in Cray ClusterStor data services Administration Guide (S-1237).
2. Place the policy file in the .cray/cds/policy/ directory at the root of the mounted Lustre file system using one of the following methods:

   Choose from:

   - Obtain write permission on this directory and either move or copy the policy file into the policy directory.
   - Ask the ClusterStor administrator to move the policy file into the policy directory.