Configuring Snapshots for File Systems based on Intel® EE for Lustre* Software

Partner Guide

High Performance Data Division

Software Version: 3.0.0.0 or later

World Wide Web: http://www.intel.com

Disclaimer and legal information

Copyright 2016 Intel® Corporation. All Rights Reserved.

The source code contained or described herein and all documents related to the source code ("Material") are owned by Intel® Corporation or its suppliers or licensors. Title to the Material remains with Intel® Corporation or its suppliers and licensors. The Material contains trade secrets and proprietary and confidential information of Intel® or its suppliers and licensors. The Material is protected by worldwide copyright and trade secret laws and treaty provisions. No part of the Material may be used, copied, reproduced, modified, published, uploaded, posted, transmitted, distributed, or disclosed in any way without Intel's prior express written permission.

No license under any patent, copyright, trade secret or other intellectual property right is granted to or conferred upon you by disclosure or delivery of the Materials, either expressly, by implication, inducement, estoppel or otherwise. Any license under such intellectual property rights must be express and approved by Intel® in writing.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL® ASSUMES NO LIABILITY WHATSOEVER AND INTEL® DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL® PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel® Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL® AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL® OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL® PRODUCT OR ANY OF ITS PARTS.

Intel® may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel® reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Before using any third party software referenced herein, please refer to the third party software provider's website for more information, including without limitation, information regarding the mitigation of potential security vulnerabilities in the third party software.

Contact your local Intel® sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel® literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm.

Intel® and the Intel® logo are trademarks of Intel® Corporation in the U.S. and/or other countries.

* Other names and brands may be claimed as the property of others.

"This product includes software developed by the OpenSSL Project for use in the OpenSSL Toolkit. (http://www.openssl.org/)

Contents

About this Document ii
Document Purpose ii
Intended Audience ii
Conventions Used ii
Related Documentation ii
Introduction 1
An Overview of Snapshots 1
Requirements 1
Space Requirements 2
Configuring for Snapshots 2
Creating a Snapshot 3
Mounting Snapshots 4
Snapshot Features 4
List Snapshots 4
Modify Snapshot Attributes 5
Global Write Barriers 5
Snapshot Logs 7
Lustre Configuration Logs 7

About this Document

Document Purpose

This document describes how to configure, create, and use file system snapshots for a Lustre file system created with Intel® Enterprise Edition for Lustre* Software. This document introduces snapshot functionality, its related utilities, and explains their use.

Intended Audience

This guide is intended for the systems integrator with a strong technical background in Linux system administration and Lustre file system deployment and management, and who has a requirement to support access to Lustre from Windows clients. The guide tries to make no assumptions about the reader's experience HSM, but advanced concepts may require some knowledge of Windows network services, such as Active Directory or Windows NT4.

Readers should have:

- experience administering file systems and storage infrastructure, and familiarity with storage concepts such as RAID, SAN, and LVM
- system management experience sufficient to install and configure a storage platform compatible with the requirements as defined in this guide
- proficiency in setting up, administering, and maintaining computer networks, including Ethernet, TCP/IP and InfiniBand where necessary. Some knowledge of Lustre networking (LNET) is required.
- some experience with installing and managing Lustre file systems

This document is not intended for end-users or application developers.

Conventions Used

Conventions used in this document include:

- # preceding a command indicates the command is to be entered as root
- $ indicates a command is to be entered as a user
- *<variable_name>* indicates the placeholder text that appears between the angle brackets is to be replaced with an appropriate value

Related Documentation

- *Intel® Enterprise Edition for Lustre* Software Release Notes*
- *Intel® Manager for Lustre* Software User Guide*
- *Configuring LNet Routers for File Systems based on Intel* EE for Lustre* Software*

- *Intel® Enterprise Edition for Lustre* Software Installation Guide*
- *Installing Intel® EE for Lustre* Software on Intel® Xeon Phi™ Coprocessors*
- *Hierarchical Storage Management Configuration Guide*
- *Installing Hadoop, 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*
- *Creating a Monitored Lustre* Storage Solution over a ZFS File System*
- *Creating a High-Availability Lustre* Storage Solution over a ZFS 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*

Introduction

A Lustre* file system is a network-based, distributed parallel storage platform consisting of metadata servers (MDS) and object storage servers (OSS). Metadata servers manage the file system namespace (directory structure, file information – inodes), while object storage servers contain the file contents. A management server (MGS) provides directory services, storing the configuration information for Lustre file systems and presenting that information to the population of Lustre servers and clients.

Intel® Manager for Lustre* software is an additional service supplied with Intel® Enterprise Edition for Lustre that enables administrators to create, manage and monitor Lustre file systems using a powerful GUI-based dashboard. With this software, operators can easily configure and manage servers, and monitor file system health and performance.

Snapshots have been employed in data management workflows for several years, having become a common feature in enterprise and even workgroup storage systems. Snapshots are most commonly found in NAS solutions, but also available in some parallel file systems. Now, Intel® EE for Lustre* Software provides the capability to create a consistent snapshot of all storage targets in a 100% ZFS-based Lustre file system, and mount a snapshot as a separate file system on Lustre clients.

An Overview of Snapshots

Snapshots provide fast recovery of files from a previously created checkpoint without recourse to an offline backup. Snapshots provide a means to version-control storage, and can be used to recover lost files or previous versions of files. Recovery of lost files from a snapshot is usually considerably faster than from any offline backup or remote replica. However, note that snapshots do not improve storage reliability and are just as exposed to hardware failure as any other storage volume.

Requirements

All Lustre server targets must be ZFS file systems running Intel® EE for Lustre software, version 3.0.0.0 or later. In addition, the MGS must be able to communicate with ssh or another remote access protocol, without password authentication, to all other servers.

The feature is enabled by default and cannot be disabled. Management of snapshots is done through the `lctl` command on the MGS.

Space Requirements

Lustre snapshot is based on Copy-On-Write; the snapshot and file system may share a single copy of the data until a file is changed on the file system. If you snapshot a file system that is at 50% capacity, then if all the existing data is updated after making the snapshot, the system will be totally full. To guarantee space is available on the file system, it is best to make a snapshot of a file system before it is at 50% capacity. Also, removing old snapshots will free file system space. The file system administrator needs to make their own snapshot create/backup/remove policy according to their system's actual size and usage.

Configuring for Snapshots

No changes are required to the Lustre file system installation or configuration to enable snapshots. All that needs to be added to a new or existing Lustre file system to enable snapshots is a *snapshot configuration file*, which contains system configuration information. The system configuration information, such as each target's hostname, pool name, local filesystem name, role, index, and so on, must be contained in the snapshot configuration file at "/etc/lsnapshot/${fsname}.conf" on the MGS. The snapshot configuration file must be present on the MGS. If the system administrator wants to allow access or querying of snapshots from a different node, then a copy of this configuration file needs to be on that node. Each target (MGS/MDT/OST) must be listed on a separate line of the file in the following format:

```
host pool_dir pool local_filesystem role(s) index
```

where:

- *host* is the hostname of the server
- *pool_dir* is the ZFS file system's pool directory
- *pool* is the ZFS pool name
- *local_filesystem* is the name of the target's local ZFS file system
- *role* is the role that the target plays in the Lustre file system and can be "MGS", "MDT", "OST", or combined "MGS, MDT", and
- *index* is the index of the target as specified when creating the target with `mkfs`.

The following is a sample snapshot configuration file for a file system with a combined MDT and MGT, multiple MDTs, and four OSTs:

```
eagle-50 /dev lustre-mds1 mds1 MGS,MDT 0
eagle-51 /dev lustre-mds1 mds2 MDT 1
eagle-52 /dev lustre-ost1 ost1 OST 0
eagle-53 /dev lustre-ost2 ost2 OST 1
```

```
eagle-54 /dev lustre-ost3 ost3 OST 2
eagle-55 /dev lustre-ost4 ost4 OST 3
```

The configuration file is created and edited manually.

Once the snapshot configuration file is in place and reflects the current file system setup, you are ready to create a file system snapshot.

Creating a Snapshot

To create a snapshot of an existing Lustre file system, run the following `lctl` command on the MGS:

```
lctl snapshot_create [-b | --barrier [on | off]] [-c | --comment
comment] <-F | --fsname fsname> [-h | --help] <-n | --name ssname>
[-r | --rsh remote_shell][-t | --timeout timeout]
```

Options:

`-b`: set write barrier before creating snapshot, the default value is 'on'.

`-c`: describe what the snapshot is for, and so on.

`-F`: the filesystem name.

`-h`: for help information.

`-n`: the snapshot's name.

`-r`: the remote shell used for communication with remote target, the default value is 'ssh'.

`-t`: the life cycle (seconds) for write barrier, the default value is 60 seconds.

To delete a snapshot, run the following `lctl` command on the MGS:

```
lctl snapshot_destroy [-f | --force] <-F | --fsname fsname> [-h | --
help] <-n | --name ssname> [-r | --rsh remote_shell]
```

Options:

`-f`: destroy the snapshot by force.

`-F`: the filesystem name.

`-h`: for help information.

`-n`: the snapshot's name.

`-r`: the remote shell used for communication with remote target, the default value is 'ssh'.

Mounting Snapshots

Snapshots are treated as separate file systems and can be mounted on Lustre clients. The snapshot file system must be mounted as a read-only file system with the "-o ro" option. Mounting as read-only allows many normal Lustre file system functions to be skipped or ignored, such as starting quota threads, skip orphan clean up, etc. If the `mount` command does not include the read-only option, the mount will fail.

Before a snapshot can be mounted on the client, the snapshot must first be mounted on the servers using the `lctl` utility. To mount a snapshot, run the following `lctl` command on the MGS:

```
lctl snapshot_mount <-F | --fsname fsname> [-h | --help] <-n | --
name ssname> [-r | --rsh remote_shell]
```

Options:

-F: the filesystem name.

-h: for help information.

-n: the snapshot's name.

-r: the remote shell used for communication with remote target, the default value is 'ssh'.

To unmount a snapshot from the servers, first unmount the snapshot file system from all clients, using the `umount` command on each client. After all clients have unmounted the snapshot file system, run the following `lctl` command on a node where the snapshot is mounted:

```
lctl snapshot_umount [-F | --fsname fsname] [-h | --help] <-n | --
name ssname> [-r | --rsh remote_shell]
```

Options:

-F: the filesystem name.

-h: for help information.

-n: the snapshot's name.

-r: the remote shell used for communication with remote target, the default value is 'ssh'.

Snapshot Features

List Snapshots

To list the snapshots for a given file system, use the following `lctl` command on the MGS:

```
lctl snapshot_list [-d | --detail] <-F | --fsname fsname> [-h | --
help] [-n | --name ssname] [-r | --rsh remote_shell]
```

Options:

`-d`: list every piece for the specified snapshot.

`-F`: the filesystem name.

`-h`: for help information.

`-n`: the snapshot's name. If the snapshot name is not supplied, all snapshots for this file system will be displayed.

`-r`: the remote shell used for communication with remote target, the default value is 'ssh'.

Modify Snapshot Attributes

Currently, Lustre snapshot has five user visible attributes; snapshot name, snapshot comment, create time, modification time, and snapshot file system name. Among them, the former two attributes can be modified. Renaming follows the general ZFS snapshot name rules, such as the maximum length is 256 bytes, cannot conflict with the reserved names, and so on.

To modify a snapshot's attributes, use the following `lctl` command on the MGS:

```
lctl snapshot_modify [-c | --comment comment] <-F | --fsname fsname>
[-h | --help] <-n | --name ssname> [-N | --new new_ssname] [-r | --
rsh remote_shell]
```

Options:

`-c`: update the snapshot's comment.

`-F`: the filesystem name.

`-h`: for help information.

`-n`: the snapshot's name.

`-N`: rename the snapshot's name as the new_ssname.

`-r`: the remote shell used for communication with remote target, the default value is 'ssh'.

Global Write Barriers

Snapshots are non-atomic across multiple MDTs and OSTs, which means that if there is activity on the file system while a snapshot is being taken, there may be user-visible namespace inconsistencies with files created or destroyed in the interval between the MDT and OST snapshots. In order to create a consistent snapshot of the file system, we are able to set a global write barrier, or "freeze" the system. Once set, all metadata modifications will be

blocked until the write barrier is actively removed ("thawed") or expired. The user can set a timeout parameter on a global barrier or the barrier can be explicitly removed. The default timeout period is 60 seconds.

The snapshot create command will call the write barrier internally when requested using the '-b' option to `lctl snapshot_create`. So, explicit use of the barrier is not required when using snapshots but included here as an option to quiet the file system before a snapshot is created.

To impose a global write barrier, run the following `lctl` command on the MGS:

```
# lctl barrier_freeze
freeze write barrier on MDTs
usage: barrier_freeze <fsname> [timeout (in seconds)]
where timeout default is 60
```

To remove a global write barrier, run the following `lctl` command on the MGS:

```
# lctl barrier_thaw
```

Thaw write barrier on MDTs.

Usage: `barrier_thaw <fsname>`

To see how much time is left on a global write barrier, run the following `lctl` command on the MGS:

```
# lctl barrier_stat
```

Query write barrier status on MDTs.

Usage: `barrier_stat <fsname>`

In query the write barrier status, the possible status and related meanings are as follows:

- `init`: barrier has never been set on the system.
- `freezing_p1`: In the first stage of setting the write barrier.
- `freezing_p2`: In the second stage of setting the write barrier.
- `frozen`: the write barrier has been set successfully.
- `thawing`: In thawing the write barrier.
- `thawed`: The write barrier has been thawed.
- `failed`: Failed to set write barrier.
- `expired`: The write barrier is expired.

- `rescan`: In scanning the MDTs status, see the command `barrier_rescan`.
- `unknown`: Other cases.

If the barrier is in 'freezing_p1', 'freezing_p2' or 'frozen' status, then the remaining lifetime will be returned also.

To rescan a global write barrier to check which MDTs are active, run the following `lctl` command on the MGS:

```
# lctl barrier_rescan
```

Rescan the system to filter out inactive MDT(s) for barrier.

Usage: `barrier_rescan <fsname> [timeout (in seconds)]`

`where timeout default is 60`

Snapshot Logs

A log of all snapshot activity can be found in file `/var/log/lsnapshot.log`. This file contains information on when a snapshot was created, an attribute was changed, when it was mounted, and other snapshot information.

The following is a sample `/var/log/lsnapshot.log` file:

```
Mon Mar 21 19:43:06 2016
(15826:jt_snapshot_create:1138:scratch:ssh): Create snapshot lss_0_0
successfully with comment <(null)>, barrier <enable>, timeout <60>
Mon Mar 21 19:43:11 2016(16060:jt_snapshot_create:1138:scratch:ssh):
Create snapshot lss_0_1 successfully with comment <(null)>, barrier
<disable>, timeout <-1>
Mon Mar 21 19:44:38 2016 (17161:jt_snapshot_mount:2013:scratch:ssh):
The snapshot lss_1a_0 is mounted
Mon Mar 21 19:44:46 2016
(17662:jt_snapshot_umount:2167:scratch:ssh): the snapshot lss_1a_0
have been umounted
Mon Mar 21 19:47:12 2016
(20897:jt_snapshot_destroy:1312:scratch:ssh): Destroy snapshot
lss_2_0 successfully with force <disable>
```

Lustre Configuration Logs

A snapshot is independent from the original file system that it is derived from and is treated as a new file system with a new file system name. The file system name is part of the configuration log names and exists in configuration log entries. Two commands exist to manipulate configuration logs: `lctl fork_lcfg` and `lctl erase_lcfg`.

The snapshot commands will use configuration log functionality internally when needed. So, use of the barrier is not required to use snapshots but included here as an option. The following configuration log commands are independent of snapshots and can be used independent of snapshot use.

To fork a configuration log, run the following `lctl` command on the MGS:

```
# lctl fork_lcfg
```

Fork the configuration for the specified Lustre system.

Usage: `fork_lcfg <fsname> <newname>`

To erase a configuration log, run the following `lctl` command on the MGS:

```
# lctl erase_lcfg
```

Erase the configuration for the specified Lustre system

Usage: `erase_lcfg <fsname>`