Hierarchical Storage Management

2016-10-20 Zhang, Hongchao

Legal Information

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at http://www.intel.com/content/www/us/en/software/intel-solutions-for-lustre-software.html.

Intel technologies may require enabled hardware, specific software, or services activation. Check with your system manufacturer or retailer.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

The products described 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.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

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.

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

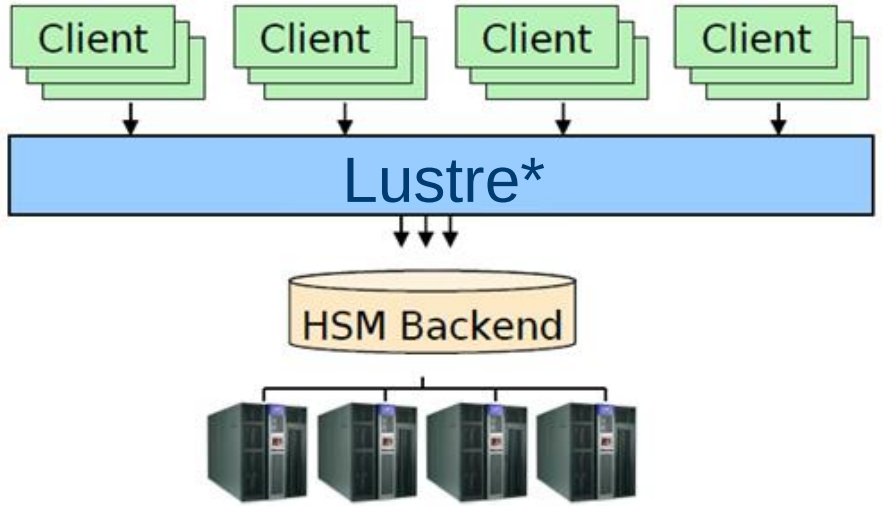
* Other names and brands may be claimed as the property of others.

Lustre* HSM Table of Contents

- Objectives
- Architecture
- Copytool
- PolicyEngine - Robinhood
- Call Path
- Usages

Objectives

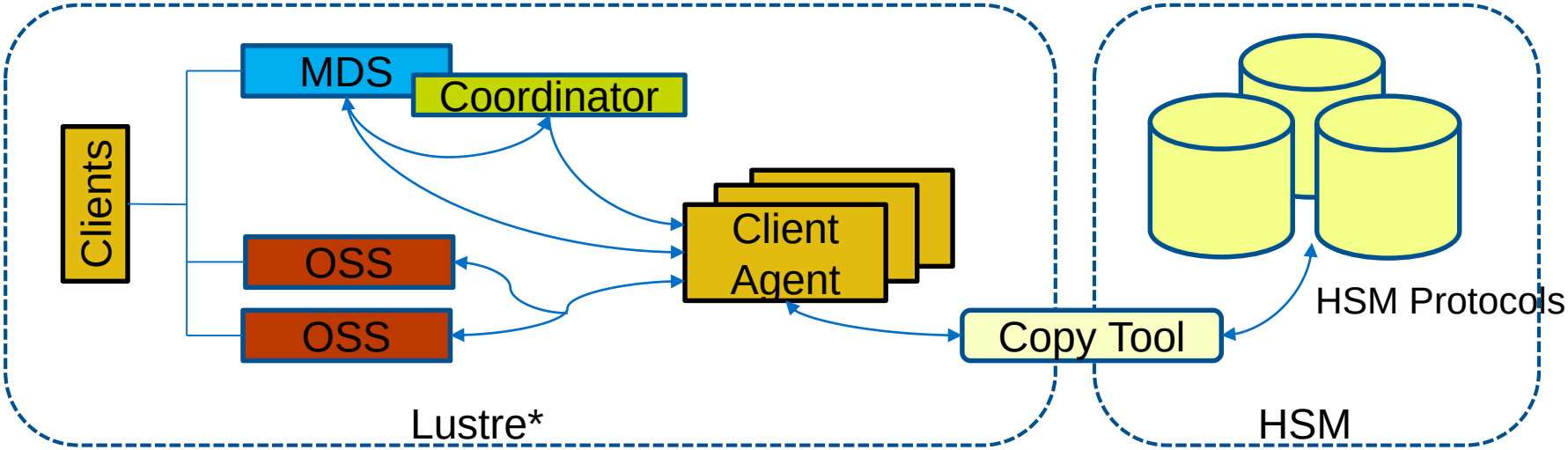
- Provide an archive solution with better scalability and performance
- Integrate with the Lustre filesystem seamlessly



Objectives – Take the best of each world

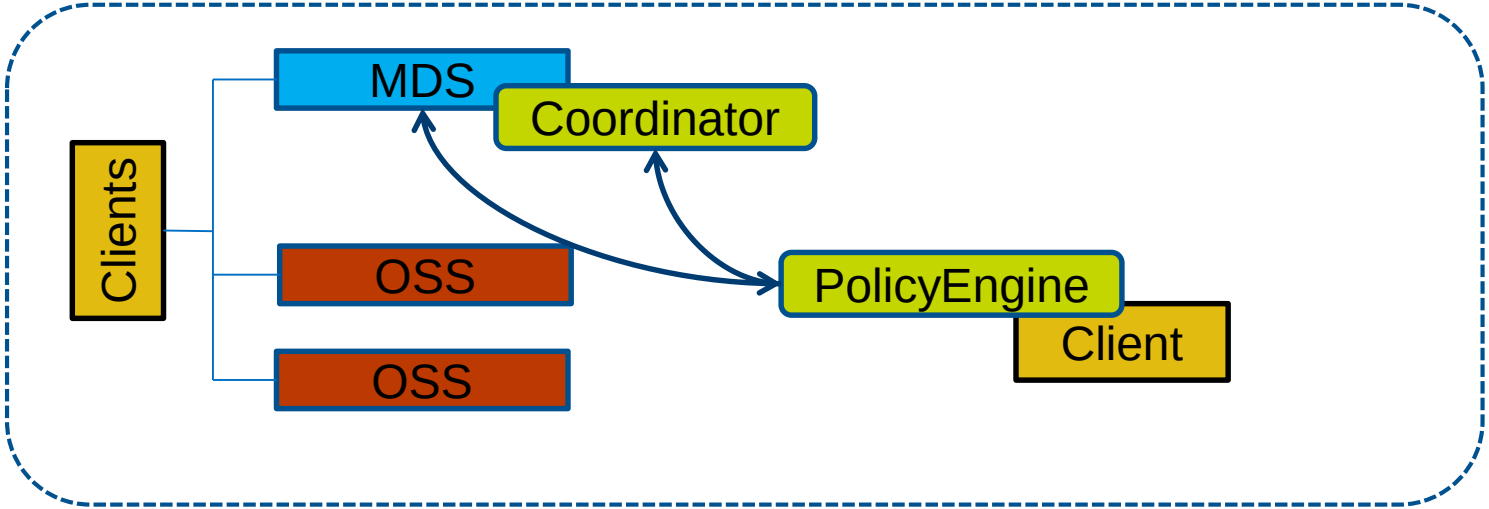
- Lustre
  - parallel cluster filesystem
  - high-performance disk cache in front of the HSM
  - high I/O performance, POSIX access
- HSM
  - Manage large number of disks and tapes
  - Huge storage capacity
  - Long-term data storage

Architecture



- The coordinator gathers archiving requests and dispatches them to agents
- Agent is a client which runs a copytool to transfer data between Lustre and HSM
- Copytool is a backend specific user-space daemon

Architecture – Policy Engine



- PolicyEngine is a user-space tool to communicate with MDT and Cooridinator
- Watch the filesystem changes by processing the changelog
- Trigger actions like pre-migration, purges and removal in backend

Copytool

- It is the interface between Lustre and the HSM (by KUC pipe).
- It reads and writes data between them (HSM specific)
- It is running on a standard Lustre client (called Agent)
- A example is available in Lustre (Posix Copytool)
  - #ls lustre/utils/lhsmtool_posix.c
  - #./lustre/utils/lhsmtool_posix --hsm_root $hsmroot --daemon /mnt/lustre

PolicyEngine - Robinhood

- PolicyEngine is the specification
- Robinhood is an implementation
  - Is a user-space daemon for monitoring and purging large filesystems
  - CEA opensource development: http://robinhood.sf.net
- Policies
  - File class definitions based on attributes (path, size, owner, age, xattr…)
  - Rules can be combined with boolean operators
  - LRU-based migration/purge policies
  - Entries can be white-listed

Call Path

- User or PolicyEngine initiates HSM request by “llapi_hsm_request()”
- Client sent *MDS_HSM_REQUEST* by “mdc_ioc_hsm_request()”
- MDT creates archive action and saves it to llog (LLOG_AGENT_ORIG_CTXT)
- CDT is waken up to process the HSM action stored in llog and send HSM request back to specified client where the Copytool runs (mdt_coordinator())
- Copytool receives the request by KUC
- Copytool could send progress report by “llapi_hsm_action_progress()”
- Copytool send completion to MDT by “llapi_hsm_action_end()”

Call Path – Data Version and Lock

- Data Version
  - Archive: compare the data_version before copy and after copy, will mark the Archiving as Failed.
  - Release: compare the current data_version with the data_version stored in the file (XATTR_NAME_HSM), will stop release if mismatch
- Lock
  - Release: open with exclusive OPEN lock (lease open)
  - Restore: hold a exclusive LAYOUT lock during restoring the file

Usages – Various HSM File State

- Exist: some copy could exist in a HSM storage
- Archived: A full copy was done for this file
- Dirty: The Lustre file has been modified since last copy
- Released: the LOV information and LOV objects are removed
- Lost: the file copy has been lost, it could not be restored.
- Non-release: the file can’t be released after archived.
- Non-archive: the file can’t be archived to HSM storage

Usages – How-to Start

- Format your devices as usual
- Start(Stop) HSM Coordinator
  - #lctl set_param mdt.lustre-MDT0000.hsm_control enabled
  - #lctl set_param mdt.lustre-MDT0000.hsm_control disabled
- On each Lustre* client which will act as Agent, starts a copytool
  - #lhsmtool_posix --daemon --hsm_root /tmp/hsm /mnt/lustre
- Your filesystem is ready to archive
  - #lfs hsm_archive /mnt/lustre/my_file
  - #lfs hsm_release /mnt/lustre/my_file

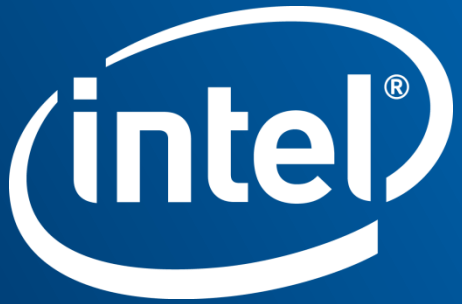
Usages – Archive, Release

- Archive

```
[root@zhanghc lustre-release]#
[root@zhanghc lustre-release]# ./lustre/utils/lfs hsm_state /mnt/lustre/myfile
/mnt/lustre/myfile: (0x00000000), archive_id:1
[root@zhanghc lustre-release]# ./lustre/utils/lfs hsm_archive /mnt/lustre/myfile
[root@zhanghc lustre-release]# ./lustre/utils/lfs hsm_state /mnt/lustre/myfile
/mnt/lustre/myfile: (0x00000009) exists archived, archive_id:1
[root@zhanghc lustre-release]#
```

- Release

```
[root@zhanghc lustre-release]#
[root@zhanghc lustre-release]# ./lustre/utils/lfs hsm_state /mnt/lustre/myfile
/mnt/lustre/myfile: (0x00000009) exists archived, archive_id:1
[root@zhanghc lustre-release]# ./lustre/utils/lfs hsm_release /mnt/lustre/myfile
[root@zhanghc lustre-release]# ./lustre/utils/lfs hsm_state /mnt/lustre/myfile
/mnt/lustre/myfile: (0x0000000d) released exists archived, archive_id:1
[root@zhanghc lustre-release]#
```


intel®