SEAGATE

Inside Lustre HSM

Technology Paper

An introduction to the newly HSM-enabled Lustre 2.5.x parallel file system

Torben Kling Petersen, PhD

Introduction

Hierarchical Storage Management (HSM) has been the enterprise choice of multi-tier storage management deployments for many years. It first appeared in the mid-1970s as an IBM product for mainframes. Since then, a number of different implementations have been developed, and many solutions (both commercial and open source) exist today. The most notable are IBM's TSM (Tivoli Storage Manager), SAM-QFS (Storage Archive Manager – Quick File System) from either Oracle or Versity, and SGI's DMF (Data Migration Facility)[1].

The original idea of HSM was to allow users to automatically move data from near-line, expensive data storage to back end, and often tape-based, cheap archive systems in an automated and transparent fashion. While HSM is sometimes referred to as tiered storage or active archive, these processes differ significantly from tape in that most of the data, if not all of it, is online and fully available to users. In addition, the tiering and data movement is usually between NAND-based flash tiers, through SAS-based HDD arrays, and then to large SATA-based storage volumes[2].

HSM functionality is now available in Lustre® 2.5, closing one of the main requirement requests often voiced from the commercial technical computing community, which has traditionally relied on proprietary, full-feature parallel file systems such as IBM's GPFS. Lustre is now one of the most successful open source projects to date, with more than 70 active developers and representation from close to 20 companies and organizations[3]. One of these organizations, the French Atomic Energy Commission (CEA), leads the development of HSM for Lustre[4].

Note that even though Lustre HSM delivers the same functionality as other peers on the market, HSM is not to be considered a Lustre backup facility; it offers far more than this relatively simple feature.

Another important distinction of Lustre HSM is that Lustre by itself is *not* a full HSM solution; Lustre is HSM enabled. This means that Lustre adds several components of what makes up a complete HSM solution but lacks the downstream tiers. These are normally handled by either another file system or a full HSM solution in its own right.

[1] http://www-03.ibm.com/software/products/en/tivostormana

http://www.oracle.com/us/products/servers-storage/storage/storage-software/storage-archive-manager/overview/index.html

http://www.versity.com

https://www.sgi.com/products/storage/idm/dmf.html

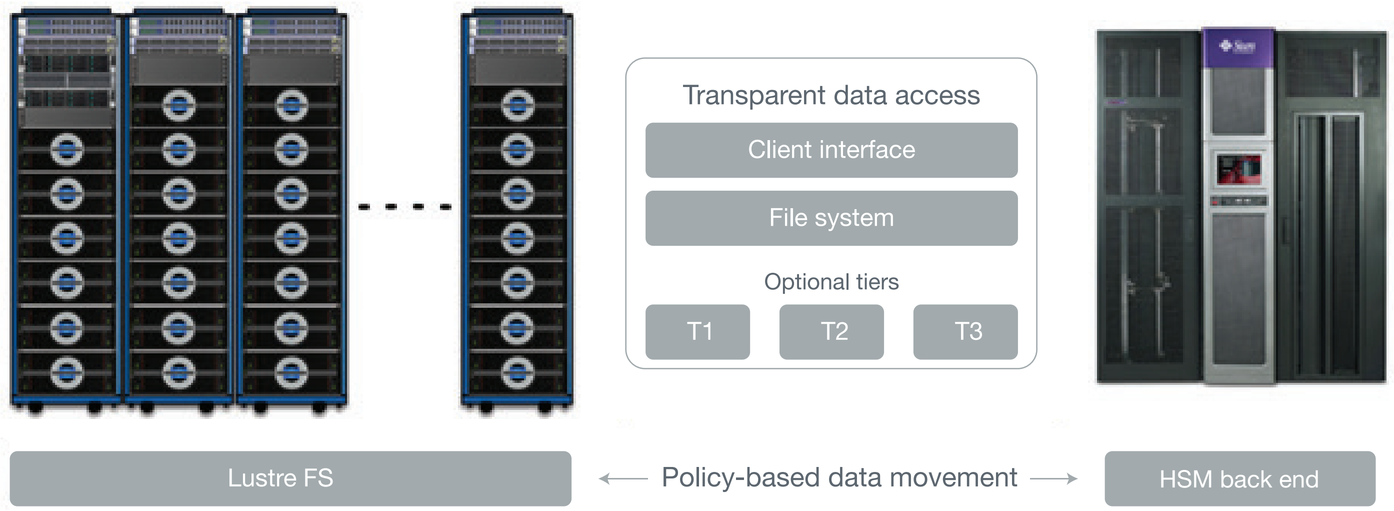
[2] http://www.activearchive.com

[3] This is a summary of characteristics for the largest supercomputer site.

For more information see http://top500.org

[4] http://www-hpc.cea.fr/en/red/equipements.htm

HSM basics

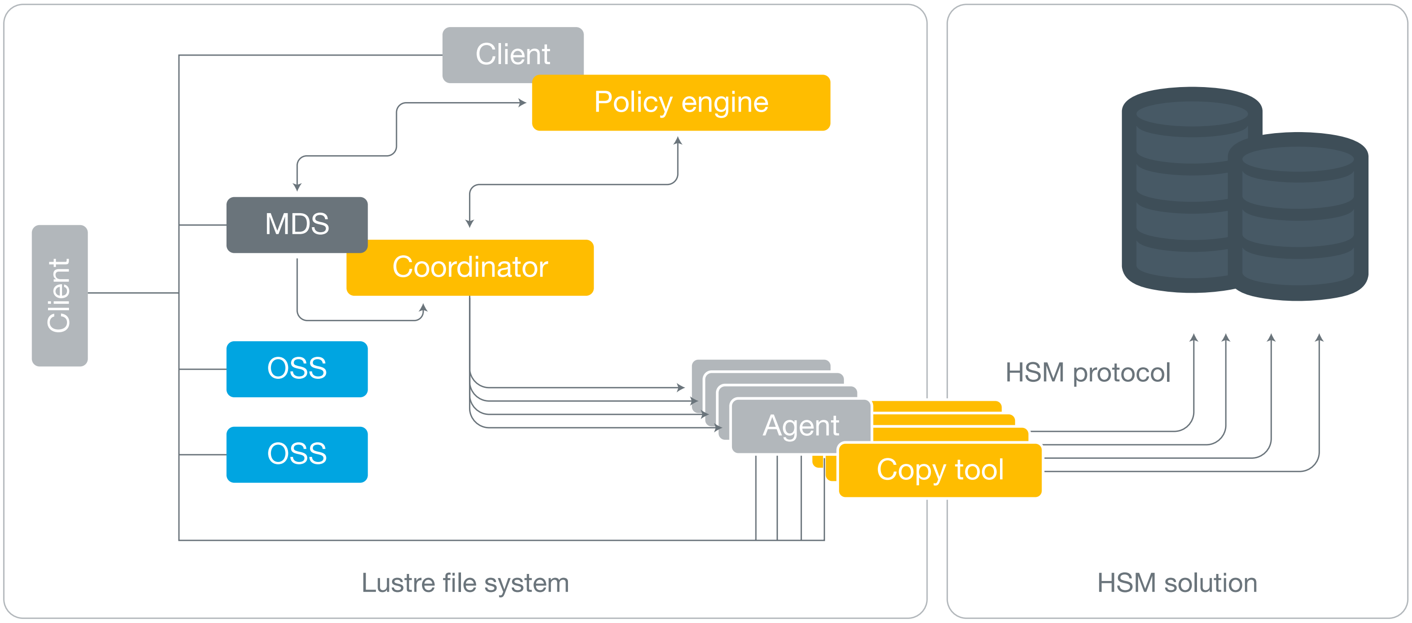


The goal of HSM is to free up space in the parallel file system's primary tier by automatically migrating rarely accessed data to a storage tier, which is usually significantly larger and less expensive.

The true benefit of HSM is that the metadata for the file (such as icons in folders, files and folders in ls - l, etc.) is NOT moved to the HSM back end. Instead, a "dirty bit" is added to the metadata of an archived file, informing the file system that the actual content of a file (in Lustre stored as one or more objects on the OSTs) had been moved. This means that when listing all files in a directory, even the files that have been moved are still visible. So from a user point of view, all the files are still available regardless of where the actual data is stored. Accessing data that have been archived requires copying the file(s) back to the file system, but again, this is what an HSM solution is designed to do.

The second important HSM concept involves the method and mechanisms used to automatically move data within storage subsystems. To put it simply: the file system keeps track of all files created, written, read, updated and deleted; this information is then used by a policy engine to either move data off or back to the file system. These policies can range from simple, such as "any file that has not been touched in 14 days," to complex, such as "any file located in the directory /mnt/lustre/large_files AND which is larger than 40 GB NOT ending in .tmp." The policy engine can have many different policies, which trigger one or more actions. Moving a file back to the file system is somewhat easier in that it doesn't require a policy, but when a user tries to read a file that has been moved, the file is copied back to the file system, after which the user gets access to the data.

Lustre HSM overview



Schematic design of a typical Lustre HSM setup

In the example above, we're only running a single policy engine, a single coordinator (i.e., single MDS) and four agents each running the CopyTool. While the example above is fairly basic, Lustre HSM is capable of managing multiple backends, each served by one or more agents. Based on the policy script, a migration of data can be done to multiple systems at the same time, thereby guaranteeing multiple copies if so desired.

While HSM is not a backup tool, a policy triggering an "Archive" action as soon as a new file is written but not followed immediately by a "release" will in essence generate a backup copy of said file. An even more advanced policy that triggers an additional "Archive" on an already archived file but using a new copy with a slightly new name would be able to work as a versioning tool.

Lustre HSM components

Agents – Agents are Lustre file system clients running CopyTool, which is a user space daemon that transfers data between Lustre and an HSM solution. A Lustre file system only supports a single CopyTool process, per ARCHIVE (i.e., the file or directories being archived) and per client node. Bundled with Lustre tools, the POSIX CopyTool can work with any HSM-capable external storage that exports a POSIX API.

Currently, the following open source CopyTools are available:

- POSIX CopyTool – Used with a system supporting a POSIX interface, such as Tivoli Storage Manager (TSM) or SAM/QFS.
- HPSS CopyTool – CEA development, which will be freely available to all HPSS sites. This tool requires HPSS 7.3.2 or higher.
- Other tools, such as a DMF CopyTool from SGI and an OpenArchive CopyTool from GrauData, are being actively developed.

Coordinator – Helper application designed to act as an interface between the policy engine, the metadata server(s) and the CopyTool(s).

Policy Engine – The policy engine used in Lustre is called "RobinHood." RobinHood[5] is a Lustre-aware multi-purpose tool that is external to the Lustre servers and can:

- Apply migration/purge policies on any POSIX file system
- Back up to HSM world
- Provide auditing and reporting
- Offer a multi-threaded architecture, developed for use in HPC
- Process Lustre changelogs to avoid an expensive file system scan
- Perform list/purge operations on files per OST
- Understand OST artifacts like OST usage

[5] http://sourceforge.net/projects/robinhood/

Policy examples

As mentioned above, all HSM functionality is triggered by policies managed by RobinHood. While the concept of policies might sound complicated, it is based on a very logical syntax and is easy to understand, even for those who do not have Lustre expertise.

In this example to the right,

- A check is run every 15 minutes if the OST usage exceeds 90%, then:
    - Files not modified in the last hour will be migrated
    - Files created more than 24 hours ago and not accessed within the last 12 hours will be released

Lustre HSM nomenclature

The steps to accomplish these tasks in Lustre are basically the same regardless of system, and are currently referred to as:

- **Archive** (“copyout”) – Copies data to external HSM system. Note that the data is still present on the Lustre file system, but a copy has been made on the HSM side.
- **Release** – Deletes objects that have been copied (N.B. on OSTs). The MDT retains metadata information for the file.
- **Restore** (“copyin”) – Copies data back when requested by a user. This is triggered by specific command (pre-stage) or a cache-miss.

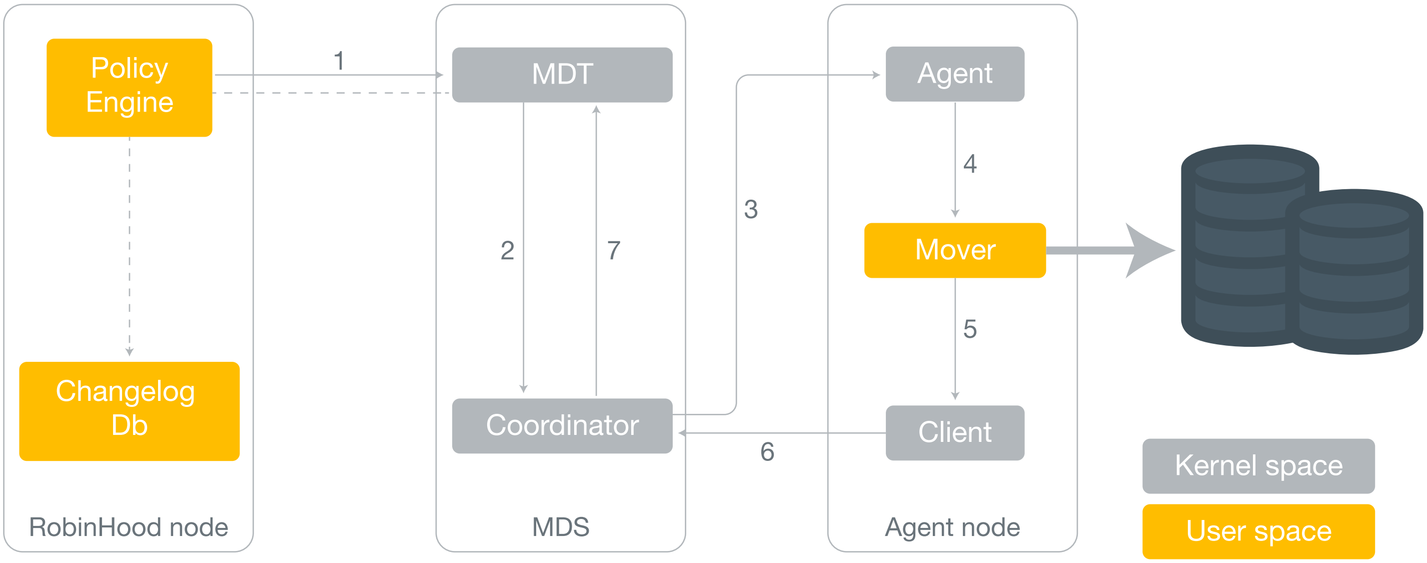
```
migration _ policies {
    policy default {
        condition { last _ mod > 1h }
    }
}
hsm _ remove _ policy {
    hsm _ remove = enabled;
    deferred _ remove _ delay = 24h;
}
purge _ policies {
    policy default {
        condition { last _ access > 12h
}
    }
}
purge _ trigger {
    check _ interval = 15min;
    trigger _ on = OST _ usage;
    high _ threshold _ pct = 90%;
}
```

Inside Lustre HSM

Schematic data flows within a Lustre HSM solution

To illustrate the inner workings of hierarchical storage management, the segments below are designed to show the basic steps of a Lustre Archive/Restore process.

Archiving (aka "Copy Out")



This example illustrates the steps HSM performs to move data out of the file system:

1. Policy engine "triggers" an event to migrate data
2. The MDT issues a CopyOut to the coordinator (including FID, extent, etc.)
3. Coordinator issues an HSMCopyOut to the agent
4. The agent launches a CopyOut command to the mover (CopyTool)
5. The CopyTool keeps the client updated on progress and completion
6. Upon completion, the agent node updates the coordinator that all data has been archived
7. Coordinator updates the metadata that data has been copied out by adding what's generally known as a "dirty bit" to the metadata

As previously noted, the process outlined above does not erase the data from the Lustre file system. To do so, the policy that triggered the data archiving process above (or a totally separate policy based on different criteria, such as "when fs > 80% full –> release archived data in chronological order"), a specific process needs to run which "releases" the data from the OSTs.

Summary

As Lustre has been transformed from a national lab supercomputer science project to an enterprise-quality file system, customer requirements have changed accordingly. One of the features about which the most questions have been asked is automated data movement within multi-tier storage systems. With Lustre 2.5 and the HSM functionality enabled, this goal has now been reached. Although this functionality is new to Lustre, it is important to remember that the French Atomic Energy Commission (CEA), which leads the development of HSM for Lustre, has been running HSM in production for several years within the Tera 1xx systems. Seagate plans to deliver a fully tested and supported version of Lustre 2.5.x with HSM-ready capability in our ClusterStor-engineered solution for HPC and Big Data by early 2015. The exact specification cannot be described at this time, and possible back-end implementations (outside the obvious HPSS and POSIX CopyTools) will have to be defined at a later stage.

It is interesting to see storage companies such as Seagate, EMC and NetApp participating in the development efforts through OpenSFS[6] and European Open File Systems (EOFS)[7], in addition to compute vendors such as Cray, Fujitsu and SGI augmenting the national labs and other high-profile end users. The continued support (both financially and through in-house development) is not only critical to the future of Lustre, but also forms the basis of the proven success of an open source file system.

This is especially true for developments involving enterprise features, which in addition to the actual code, need to undergo extensive scale testing to prove reliability and resiliency. And scale testing (which, both from the point of view of actual scale and capacity, also includes long-term stability testing) is an expensive proposition. While some early adapters are willing to work with code that may not be fully vetted, the commitment of vendors such as Seagate to support extremely large installations will benefit the entire community. And as the list of high-profile users keeps growing, the efforts in developing more enterprise features increases proportionally.

About Seagate Cloud Systems and Solutions

Seagate® is a world leader in storage solutions. Our new Cloud Systems and Solutions strategy brings innovation and an open approach to Intelligent Information Infrastructure™ to help all organizations manage their next-generation workloads—with scale, performance, and cost aligned to business needs. Our portfolio includes integrated high-performance computing solutions; do-it-yourself components and engineered solutions; custom, modularized systems for original equipment manufacturers (OEMs); and the EVault® line of cloud backup and restore, disaster recovery, and rapid archive serviceses.

Next Step

Find out more about the Seagate® ClusterStor™ line of HPC storage systems by calling 1.800.SEAGATE or visiting www.seagate.com/hpc

[6] http://opensfs.org/participants/

[7] European Open File Systems – http://www.eofs.org

seagate.com

AMERICAS Seagate Technology LLC 10200 South De Anza Boulevard, Cupertino, California 95014, United States, 408-658-1000
ASIA/PACIFIC Seagate Singapore International Headquarters Pte. Ltd. 7000 Ang Mo Kio Avenue 5, Singapore 569877, 65-6485-3888
EUROPE, MIDDLE EAST AND AFRICA Seagate Technology SAS 16–18, rue du Dôme, 92100 Boulogne-Billancourt, France, 33 1-4186 10 00