Tiered Adaptive Storage

An engineered system for large scale archives, tiered storage, and beyond

Harriet G. Coverston
Versity Software, Inc.
St. Paul, MN USA
harriet.coverston@versity.com

Scott Donoho, Craig Flaskerud, Nathan Schumann
Cray Inc.
St. Paul, MN USA
sdonoho@cray.com
cflaskerud@cray.com
nds@cray.com

Abstract— **A technical overview of Cray® Tiered Adaptive Storage (TAS) and its capabilities, with emphasis on the Cray TAS system architecture, scalability options and manageability. Also includes specific details about the Cray TAS architecture and configuration options, in addition, the Cray TAS software stack is covered in detail, including specific innovations that differentiate Cray TAS from other archive systems. Primary scalability factors of both the hardware and software layers of Cray TAS and the ease of Cray TAS integration with Lustre® via HSM capabilities available in Lustre 2.5 are characterized.**

Keywords-component; HSM, archive file system, tiered storage, Lustre

I. Introduction

With Cray Tiered Adaptive Storage (TAS), customers can transparently move data from a file system cache layer to archive media and back to a file system cache with un-compromised performance, efficiency, and reliability. At the center of Cray TAS is the Versity Storage Manager (VSM) software from Versity, Inc. VSM is a native port to Linux of the widely used open source Sun Microsystem® Storage Archive Manager (SAM) software, originally developed by LSC and Sun Microsystems for Solaris. The VSM software is comprised of a scalable, shared file system and a policy driven archive manager. The VSM file system provides a POSIX compliant cluster file system designed specifically for high performance tiered storage and dense archive environments. The VSM archiver is tightly-coupled with the VSM file system which enables efficient file classification and policy management capabilities.

II. Cray TAS an Engineered Archive System

Cray TAS is an engineered system comprised of server, storage and management software components selected for their performance, density, and economic efficiency with VSM at the core. These components are selected and optimized to provide a high-performing, flexible, highly available, and reliable system. A benefit of a Cray TAS solution is that it takes the burden of lengthy architecture planning, implementation, and testing out of the customers hands and provides a single support contact for all Cray TAS components and software. The following is a typical timeline of archive implementation from design to deployment:

- Production selection and evaluation
- Architecture Planning
- Networking
- Component compatibility
- Software integration
- Performance benchmarking
- High-availability implementation and testing
- Site planning (cabling, environmental, power)
- Testing
- Deployment
- Upgrades and patches

Cray's goal is to dramatically simplify complex data management architectures and shorten the time to implement a data management production solution while providing a single support contact.

The Cray TAS hardware stack consists of a flexible system of server and storage hardware components that scale from a single rack of fully factory-integrated components to multi-petabyte distributed systems that can leverage existing customer high-performance tape automation. Cray TAS supports up to four tiers of storage, including high-performance disk, dense archive disk, and several types of removable tape media. Remote tiers are supported for off site archiving as well.

A. Metadata Controller

A base Cray TAS system consists of a redundant pair of x86-based server nodes running Enterprise Linux as well as metadata and administrative storage array. The gateway configuration uses SAS for its private storage connections to the administrative storage. A Cray TAS gateway can be expanded by attaching fibre channel or InfiniBand® block storage arrays. These arrays serve as tier 0, the high-performance file system storage. For additional tiers the Cray

TAS gateway can connect with fibre channel attached LTO tape drives, Oracle StorageTek T10000 tape drives or the IBM® TS1100 family of tape drives.

B. Supported Tiers

Cray TAS with VSM supports four storage tiers that provide for transparent capacity increases and data protection. The tiers serve as a virtual expansion of the tier 0 file system capacity and provide for the capability to create one or more copies of data on any given tier. Tiers can be comprised of tape media as previously mentioned, disk media for faster data access, and remote storage media using the included network copy infrastructure.

Tape storage is a typical tier used for larger configurations because tape provides very high capacity at a low cost compared to disk storage. With Oracle's T10000D tape technology, each tape can store 8.5 TB natively, without compression, and can automatically compress data to further increase individual cartridge capacities. Current generation IBM TS1100 can store 4 TB natively per cartridge, with LTO at 2.5 TB, both supporting compression for additional capacity. Compare this to one of the largest hard drives on the market currently, the 6-TB Ultrastar He6 drive from HGST. While this drive does have a higher capacity than both the IBM and LTO tape formats, hard drives will likely require more power and cooling for daily operation and do not provide for portability like tape media can.

A disk system's place within the storage tiers is to provide much faster access to data than tape. Many archive architectures implement a disk-to-disk-to-tape archive configuration. In this configuration, the first disk system is comprised of very fast lower capacity drives for very fast access to data. The second disk system is used as a temporary archive space, where data may reside for days or weeks. This disk system provides fast access to recently used data that does not need to reside on the fastest tiers. Typically, data is copied to tape to ensure its protection. A disk system's place in a long-term data archive will constantly change with changes in disk technology. Higher capacity drives will continue to be developed. Object storage architectures with data coding can provide higher efficiencies and more flexibility than a standard RAID configuration. Additional power power savings can be realized with spin-down technology.

Storage tiers can also be located remotely that archive files on a local VSM client to a remote VSM server. File transfers between VSM sites is carried out efficiently by the VSM remote transfer deamon over TCP/IP. A common configuration for remote tiers is to have two VSM sites that provide host archive resources to each other in a high availability configuration. Each site maintians one or two archive copies of the other site. Another advantage of remote storage tiers allows archive resources to be consolodated in a central location and shared with remotes clients distributed over a large geographic area.

C. Data Movers

The data mover (DM) servers are VSM shared clients that mount the file system natively. VSM metadata access is provided over an internal TAS IP network, while data access is provided over a high-speed fibre channel or InfiniBand network. The data movers are stand-alone servers running CentOS Linux.

DMs are used for TAS external access to the VSM file systems, leaving the gateway available to manage file system metadata and archive data only. The DM can provide NFS, CIFS or the Lustre HSM Connector services.. DMs can also run customer applications and access the VSM file system via standard POSIX interfaces.

D. Management Server

The Cray TAS gateway also provides a Cray Integrated Management Services (CIMS) node that runs Bright Cluster Manager software for node provisioning and system health monitoring. A Cray TAS toolkit, which simplifies monitoring, administrative task, and reporting, is integrated into the Bright Cluster Manager health and metric framework. An Ethernet fabric for communication between TAS components provides a management network. This base Cray TAS gateway implements the VarsityVersity Storage Manger software stack which uses the SAS connected disk array for meta-data storage and administrative storage.

III. VSM Versity Storage Manager Architecture

The VSM software consists of two primary components; the VSM file system, which provides a caching layer and the VSM storage archive manager, which supports the storage tiers. The VSM file system has been developed with a superset of POSIX meta-data to enable the inclusion of many additional file meta-data attributes. Primarily, these include the location of up to 4 archive copies for each file, and control attributes that inform the storage archive manager about specific policies for the files.

A. Storage Archive Manager

There are four storage archive components: archive, release, stage, and recycler. The archive process copies files from the disk cache and makes multiple copies based on the archive policies. The release process manages the disk cache according to set thresholds. The stage process retrieves the file from the archive tier and restores it to the disk cache. The recycle process repacks the archive media.

The Versity storage archive manager VSM uses a file classification system called archive sets. A file can belong to a single archive set. The classification for each archive set is based on a rich set of meta-data policy attributes and includes size, user, group, minimum size, maximum size, access age, and name regular expressions.

New file system events are constantly evaluated by the storage archive manager, which is signaled when changes occur in the VSM file system. Archive file candidates are classified into archive sets based on administrative policy and queued for archival on archive media. The archive policies include parameters such as the activity state of the file (open or closed), number of copies of each file and additional polices attributes that implement and govern other components of the storage manager. The archive set is written to the media when the first of the three policy thresholds is reached: count of candidates, age of the archive set, or size of the candidates.

B. High Capacity File System

The VSM file system is a true 64-bit address space. Large files can be striped across storage devices (LUNs) or files can be allocated round robin on each LUN. The VSM file system supports over a thousand LUNs in one file system with a maximum size of 4.5 Petabytes per LUN. This means VSM can support up to 4.5 Exabytes in one file system and VSM does not limit the number of file systems on a server.

C. Metadata

File system metadata can be interleaved with files on the same physical disk devices or, for increased performance, can be on one or more physically separate disk devices. When metadata and data are separated, metadata is typically stored on very fast rotating disks or, as is the case with TAS, on Solid State Drives (SSDs) for maximum performance and low latency.

VSM provides an extended `ls` which outputs a detailed display of a VSM file. The POSIX attributes, VSM attributes, and archive copy information is displayed with the `sls -D` command.

D. Shared Clients

The shared client feature increases the scalability performance of the VSM archiver. This feature allows multiple file servers to mount VSM natively. The shared clients access data via a high-bandwidth fibre channel or InfiniBand network. Metadata is accessed through a dedicated internal IP network. The shared client feature enables a very scalable high-bandwidth file system. The end user can access the VSM file system via NFS/CIFS or directly from applications running on the shared clients.

E. Open Archive Format

As VSM identifies files to be archived, they are grouped into archive sets and written to the target tier (disk or tape archive). The format that the files are written in is based on the GNU TAR format. The main benefit to bundling archive data into an open format is archive data is always retrievable with standard UNIX utilities even outside of a Cray TAS/VSM environment. File system metadata or database look up operations are not required to retrieve individual files, which provides an excellent platform for disaster recovery. Another advantage of the TAR format is that it provides a performance benefit when writing to tape media. Tape is sequential access media and typically performs best when data is streamed to tape drives. By bundling smaller files into one larger TAR file, the overall archive performance improves as does system efficiency.

F. Removable Media Files

VSM supports a new file type, the removable media file. This file is a handle to a tape. You create a removable media file with the request command. For example, the following command creates a file that points to a LTO tape, labeled VSN 000083.

```
# request -m li -v 000083 /vsm1/rm/tapefile
```

You can see the removable media file in the namespace. This file provides device (drive/media) independence and enables VSM daemons to issue POSIX I/O. There is no tape logic in the archive and stage (copy) daemons. The removable media file enables VSM to deploy smaller, easy to maintain daemons. The request command also supports volume overflow. Multiple VSNs and a position for each volume is supported.

When a process opens a removable media file, the VSM file system suspends the process until the VSM Media Manager mounts and positions the media. Then the open returns and the process reads or writes to the tape with standard POSIX read/write system calls.

When the process issues the close, the file system suspends the process until the VSM Media Manager closes out the media. Now the tape is eligible to be rescheduled. If the process dies before issuing the close, the file system gets the close (this is standard UNIX) and properly terminates the tape. No polling is required and the tape drive is immediately available for rescheduling.

G. HSM Architecture

The VSM file system was first designed and developed as an archiving file system, resulting in a tightly coupled file system and archiver. Some HSMs added the archive feature to an existing file system, therefore requiring a daemon to mitigate I/O between the file system and archiver. Unlike other HSMs that use the XDSM protocol and have daemons in control, the VSM file system controls the HSM. The archive information is located directly in the VSM inode. Other HSMs have the archive information in a separate database requiring context switch callouts to get the archive information. More worrisome is that the database can get out of sync with the file system. VSM avoids metadata inconsistency issues by extending the file inode with the archive information. The enhanced inode also enables VSM to respond very fast to requests to restore (stage) files.

IV. CRAY TAS TOOLKIT FOR VSM

The Cray TAS archive solution provides a comprehensive collection of tools known as the Cray TAS toolkit for monitoring, reporting and administrative tasks.

The goal is the simplify the tasks for administering and understanding the Cray TAS data management system. These tools are divided into three categories: Monitoring, Administrative, and Reporting tools. The tools are designed to communicate through STDOUT, email and the `syslog` facility.

A. Administrative tools

The main administrative tool is the convenient script `tasha` that controls the active/passive MDC cluster. This script condenses a number of critical steps for failing over the active MDC to the standby MDC.

Management of the VSM file system dumps are maintained through the `tasdump` script that schedules dump times of each VSM file system. This ensures that the dumps are available in event of a catastrophic failure.

The `tasclean` script maintains a site defined number of dumps and VSM reports.

B. Monitoring tools

A number of the Cray TAS toolkit tools constantly monitor the Cray TAS and VSM environment. These tools target critical points in the system and alert the administrative staff via `syslog` or email to issues before they become serious.

C. Reporting tools

It is important that the archive administrative staff understands the performance of the Cray TAS environment. The `taschart` tool provides a daily snapshot of the performance at the device and server level as well as the file system usage.

The `tasvolumes` tool identifies problematic tape volumes in the automatic tape library. This script can be run as a periodic audit or on a daily basis.

V. LUSTRE HSM

With the community release of Lustre[1] server 2.5 which includes HSM capabilities, it is now possible to extend the functionality of Cray TAS data management platfrom to the Lustre file system. Similar functions native to Cray TAS and the Versity File System such astransparent file archival, and features such as releasing data associated with a file after it has been archived while still maintaining the file in the namespace and automatically restoring the data, are now all available directly from within the Lustre file system. Because these features are integrated directly within the Lustre file system, the Cray TAS can act as another strorage tier in the storage hierarchy. In a Cray TAS system, data is not considered protected until it is stored to long-term media with typically more than one copy of the data being created.
The Lustre server release included a reference implementation of a POSIX copy tool that allowed for data to be migrated from one Lustre file system to another Lustre file system. For Cray's implementation of TAS, it was important to provide flexibility with the backend storage, improved reliability, control, and monitoring over archive operations and finer granularity over performance scaling. Cray's Lustre HSM Connector (LHC) was designed with these requirements in mind. The LHC has three primary components, the Robinhood policy engine developed by the French Alternative Energies and Atomic Energy Commission (CEA), Lustre HSM Management Node (LMN) and one or more Lustre HSM Data Movers (LDM). With LHC the performance of a file copy can be scaled by distributing the data movement across multiple LDM nodes. Command and control of all LHC operations is performed by the LMN which is configured as a high-availability server pair for reliability.

A. Lustre HSM Management Node

The LMN is the interface between the Lustre file system and the LDMs. The role of the LMN is to receive requests from the Lustre file system coordinator then queue, sort, prioritize and schedule the requests across one or more LDM as required. Files are distributed across multiple LDMs based on a size threshold. The LMN sends updates to the Lustre file system during the copy operation and discards the request once a copy has completed. Two LMNs will typically be running within LHC to provide for fault tolerance. Both LMNs listen for events from the Lustre file system coordinator, but only one acts on the requests.

B. Lustre HSM Data Mover

The LDM is responsible for performing the data movement from the Lustre file system to Cray TAS. The LDM runs on a server that is a native client of both the Lustre file system being managed by LHC and the VSM file system used as the cache for the archive. The LDM does not require any specific knowledge of the file system because it is a simple data transfer mechanism that responds to requests from the LMN. Each LDM is independent from any one LMN and can accept requests from multiple LMNs that have registered with the LDM. This provides the capability to manage multiple Lustre file systems from the same LHC configuration. Each LDM can simultaneously copy multiple files on a single node.
Copy requests from the LMN are queued as they are received. When a copy slot is available the oldest request is removed from the queue and the file is copied. Partial reports are sent to the LMN regularly during the copy operation.

C. Lustre Attribute Storage

The layout of a file within the Lustre file system, stripe size and width, are stored along with the data in LHC. When a request to restore a file is received from the Lustre file system, the data is restored using the original stripe width and size along with any other attributes that were previously set. In order to maintain efficiency and consistency on the backend storage the attributes are packaged along with the data. Files are also organized using a similar scheme to the original copy tool in which the path is built from the Lustre file system file identifier (FID).

VI. Conclusions

Archive solutions are complex and time consuming to implement. Cray's goal with Cray TAS is to dramatically simplify deployment of a complete archive and data management solution. Cray TAS is a carefully engineered solution that provides a high-performance, reliable, and secure hardware and software stack.

References

[1] Lustre OpenSFS
http://lustre.opensfs.org