The Big Data Challenge: Intelligent Tiered Storage at Scale

Michael Feldman

White paper
November 2013

EXECUTIVE SUMMARY

The explosion of data and the consequent allure of deriving value from it is highlighting the value of tiered storage technologies. Such systems are designed to support applications whose data naturally conforms to different access patterns and capacity profiles. By utilizing the range of storage technologies – solid state, hard disk, and tape — tiered storage can deliver cost-effective solutions, along with the flexibility, performance, and capacity demanded by extreme-scale data environments.

The inherent complexity of tiered storage has led to the use of Hierarchical Storage Management (HSM), software that provides transparent data access across multiple storage subsystems. Its principle goals are to virtualize the storage components from the application and user viewpoint, as well as deliver automated data management. HSM also encompasses a set of tools for system administrators to implement site-specific storage and data policies.

Tiered storage/HSM systems are especially suited to data-intensive applications in which storage capacity and data access patterns are especially demanding. This encompasses both technical and business applications (in Intersect360 Research parlance, High Performance Technical Computing and High Performance Business Computing), and includes analytics associated with scientific high performance computing, oil and gas exploration, financial services, healthcare, security/surveillance, and media/entertainment. The common denominator among these applications is their demand for performance and large, continuously growing data repositories.

There are a number of tiered storage/HSM solutions on the market, the latest being Cray's Tiered Adaptive Storage (TAS). TAS offers an integrated, end-to-end storage system that can be configured according to the customer's application needs. It supports up to four storage tiers of solid state drives, disk or tape, and is designed to manage data with rapid growth rates and indefinite lifetimes. Based on open standards, TAS is built to serve large-scale enterprise and HPC data sets with a need for unified data access. As such, it offers one of the most advanced and full-featured tiered storage solutions in the industry.

TABLE OF CONTENTS

TIERED STORAGE FOR BIG DATA

One of the most important enabling technologies for big data applications is that of tiered storage. The tiered model enables users to build storage systems from heterogeneous technologies, which allows applications to be optimized for price, performance, capacity, and functionality. With data capacities for a widening array of applications reaching into petabytes, and soon, exabytes, the demand for robust tiered storage systems is almost certain to expand.

At the most fundamental level, tiered storage aggregates two or more storage types, based on tape, hard disk drives, and solid state disks (SSDs). For example, there's a spectrum of hard drive types — from SATA to SAS. Each offers trade-offs in price, performance, capacity, and reliability. Analogous differentiation is available in tapes and SSDs.

Traditionally, users have kept data tiers in separate silos: tape systems for the archival data, disk systems for online data, and SSDs or fast disks for metadata and content that requires low-latency access. Archives, in particular, were often considered a "write once, read never" (WORN) model, but with an increasing focus on data mining, even offline stores are now fair game for applications. Besides the inflexibility of a siloed setup, it also increases complexity, requiring users to procure and maintain multiple distinct hardware and software platforms. Interoperability between the storage systems is often left to customized solutions or ad hoc solutions from third-party providers. The cost and complexity of planning, designing, and building these systems should not be underestimated.

Tiered storage provides an environment where the storage components can be unified under a single namespace. The overarching philosophy is to store infrequently accessed data on cheaper, higher capacity media, while data that needs to be accessed continuously resides on more expensive but faster media. Data with access patterns that fall in between those extremes is stored on intermediate tiers.

Fortunately for most applications, the majority of the data is accessed infrequently (an average of 60 to 90 days, depending on the application and usage model) while the working dataset is usually much smaller. As data sizes grow — think petascale — and the data naturally aligns to different usage profiles, the cost advantages of tiered storage environments becomes more compelling.

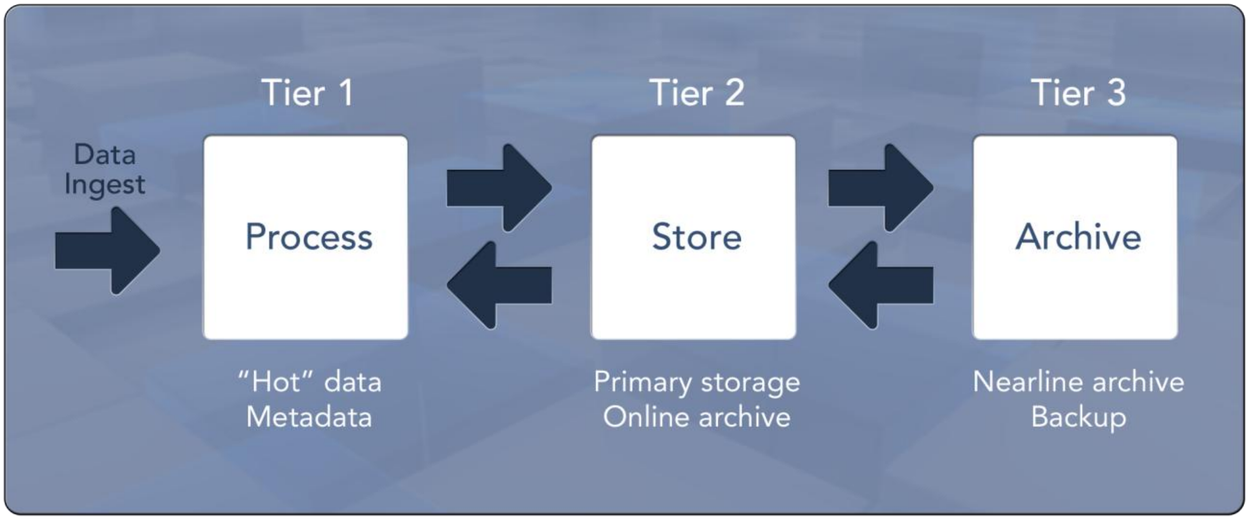
While a tiered storage system can be built entirely from various disk products providing different cost/capacity/performance profiles, for archival data, in particular, tape still offers the best price/performance. Despite advances in capacity, cost, and power efficiency for hard disk components, 30% of users in our latest (2012) HPC User Site Census survey reported that they use tape for archival storage[1].

At the other end of the cost and performance spectrum are SSDs, which offer a "Tier 0" level of storage for data that is expected to be accessed continuously and requires the highest transfer rates possible. With some vendors offering SSD arrays aimed at competing with spinning disk arrays on a cost basis, the use of solid state storage should become more attractive for many applications, including hierarchical storage environments.

The general data workflow of a three-tiered storage system is provided in Figure 1.

Figure 1: Workflow in a three-tiered storage system

Source: Intersect360 Research



UNIFYING THE TIERS: HIERARCHICAL STORAGE MANAGEMENT

It's not enough just to hook together a system of connected storage tiers. Software must do the work of migrating data across tiers to deal with data access patterns dynamically and to enforce data policy decisions. Accomplishing this requires Hierarchical Storage Management (HSM), a set of data management tools that automate the movement and replication of data across tiers. There are number of HSM systems in the marketplace, including the High Performance Storage System (HPSS), the Data Migration Facility (DMF),

[1] "HPC User Site Census: Storage," Intersect360 Research, 2013

and the Oracle's SAM (Storage Archive Manager), which came through Sun's Storage Archive Manager/Quick File System (SAM-QFS) during the Sun acquisition, among others.

The value of these tools is that they provide automated tiering of storage, which reduces many of the manual tasks that fall to system administrators when they need to maintain primary and archive storage together. That not only translates to cost savings, but makes the process of storage migration more efficient and dynamic, as well as less error-prone.

The challenge for HSM is to manage heterogeneous storage systems, often from multiple hardware and software vendors. Making all the pieces work seamlessly requires an intimate understanding of how the hardware, software drivers, network interfaces, and operating system interact. In many ways, HSM is the analogue to heterogeneous computing, where different processing elements are integrated together to maximize compute throughput.

In the same way, HSM must virtualize non-uniform storage components so as to present a global storage pool to users and applications. In other words, the tiers must be abstracted; users and applications must be able to access their files transparently, without regard to their physical locations. This allows the customers and their software to remain independent of the underlying hardware.

Arguably HSM has been the weakest link in tiered storage systems, since these tools must encompass the complexity of the hardware and all the software layers mentioned above. This has led customers to cobble together their own systems for their own limited subset of components, despite the inherent disadvantages in such an approach.

A general-purpose tiered storage/HSM system is much preferred if it can encompass the basic requirements, including:

- **Scalability:** The system must scale to a level that accommodates the information flow.

- **Performance:** The data must be retrievable in a timeframe that makes it useful for the applications.

- **Virtualization:** The storage pool should be presented as a single namespace to the users and applications.

- **Facilities efficiency:** The hardware must fit within the facilities budget for floor tiles covered and watts consumed.

- **Data management:** There must be tools for migrating files efficiently between the storage tiers.

- **Data integrity:** The user needs to be confident that the stored data is valid and will still be good data years down the road.

- **Data accessibility:** Data lives forever, unlike the hardware that stores it, so the system should accommodate both planned and unplanned upgrades of storage components without disruption.

To provide this continuous accessibility, the HSM should incorporate transparent data replication and migration to new storage, such that downtime due to component failures and system upgrades can be minimized or even eliminated entirely. This is especially desirable when there is a large archival component to the dataset that demands frequent capacity upgrades to handle a continuous influx of new data.

The Achilles heel of some HSM systems, especially in large-scale archives, is their architectural design. For example, HSMs that are tied to DMAPI (Data Management API) can exhibit namespace consistency problems. By nature, DMAPI provides an interface between a file system and database, which can become out of sync at scale. If the database needs to go offline to sync up or to provide a consistency check, that takes the entire archive offline.

On a more practical note, many HSM systems tie customers to proprietary technologies, often on hardware sold by the originating vendor. That type of vendor lock-in is tolerable for certain types of applications, but for storage systems that must house data with indefinite lifetimes, such an arrangement can be particularly problematic.

APPLICATION LANDSCAPE

There is a growing set of uses for tiered storage/HSM environments. The expanding application set associated with big data workloads has the potential to drive more customers to consider such systems. Essentially any application that must deal with large volumes of persistent file-based data is a candidate.

A report[2] compiled by researchers from Indiana University, Oxford University, and Microsoft Research describes the data deluge inundating scientific and commercial enterprises. For example, the authors point out that the Large Hadron Collider generates 15 petabytes of

[2] http://grids.ucs.indiana.edu/ptliupages/publications/Where does all the data come from v7.pdf

data per year, while the next-generation radio telescope, known as the Square Kilometer Array, will collect more than three times that in just an hour. Not all of that data will be stored, but it shows that volumes of data with which these large science projects must contend.

The principle categories, for both scientific and business applications, include:

- **Scientific Research.** Traditional simulation and modeling for research (i.e., bioscience and medicine, astronomy, molecular dynamics, earth science, climatology/meteorology, fluid dynamics, food science, particle physics, and nuclear energy) generates or uses large datasets. As the models increase in fidelity, data capacity increases accordingly. Access speed associated with file metadata and the working data set is a driving factor for many of these applications. Archival requirements vary, but areas like climatology, earth science, and astronomy demand particularly extreme levels of capacity.

- **Oil and Gas Exploration**. Energy companies and their partners collect high-definition seismic imagery in order to discover new oil and gas reservoirs or to uncover untapped resources in existing ones. Growing sets of 3D seismic data and time-lapse seismic (4D) require scalable storage environments, which must be flexible enough to feed performance-demanding seismic analytics applications as well as to hold that data for long periods of time for possible future analysis. The favorable economics of the oil and gas industry are driving the acquisition of larger datasets as companies expand their exploration activities.

- **Financial Services.** Banks and other financial institutions rely on big data analytics for applications such as asset risk management, compliance monitoring, and investment research. Although data velocity tends to be more critical than capacity, datasets are growing, and for a number of reasons (legal compliance, risk reduction, and new models), much of the historical data is accumulated indefinitely rather than thrown away.

- **Healthcare.** Medical and patient data represents some of the fastest-growing and most heterogeneous data sets. Medical imaging alone is producing data at the rate of more than an exabyte per year, a level that is expected to be matched in the future with human genome sequencing. Due to the personal nature of the data, most of it will demand storage timeframes of (at least) decades.

- **Security/Surveillance.** Collections of video, audio, and personal records (phone, email, and social media data) make up the majority of security data. Video surveillance, one of the fastest-growing applications, is estimated to generate hundreds of petabytes per year. A lot of real-time security data has a relatively short life, requiring a lot of storage performance at the top tiers closest to the analytics, but very little at the archive end.

- **Media and Entertainment.** Video and audio files from live streams and recorded content are the heart of intellectual property for companies in this sector, which includes web-based content aggregation and distribution. Archival content for news, sports, and the increasing influx of social media content from individual users creates some of the most demanding capacity challenges of any commercial sector. Content distribution and analytics are the principle applications that shape storage requirements. The ability to scale the storage capacity continuously and provide an environment where data must be kept indefinitely are critical requirements for this application set.

In applications where data is clearly accessed according to different performance expectations, tiered storage is a natural fit. For example, in the case of web-based video distribution, one could envision a three-level tiered storage model: At the highest access points, website-embedded images depicting scenes from a video could be stored on SAS disks, while the video trailers could be housed on SATA drives. The complete video files could reside on a tape-based archive, which are transferred to disk as needed. The most popular videos would tend to reside on the disk tiers or perhaps even be migrated to an additional SSD cache tier.

AN INTEGRATED SOLUTION FOR HPC AND BIG DATA

To meet the needs of these users, Cray has introduced Tiered Adaptive Storage (TAS), an end-to-end, integrated tiered storage solution designed for the most demanding HPC and big data environments. TAS aims to integrate primary and archive storage under a single virtualized storage environment. The solution is based on Linux and encompasses all of the associated hardware and software components.

Physically, TAS consists of a management network, gateway servers (file system and management, as well as metadata storage), file system storage, and data movers. Storage networking comes in the form of InfiniBand or Fibre Channel for disk and Fibre Channel for tape. The tape, disk and SSD components are sourced from a variety of partners (including Spectra Logic and NetApp), with Cray acting as the reseller, system integrator, and single point of support.

Cray is doing all of the heavy lifting in terms of configuration, integration, and testing. The company selects and configures device drivers for all components, and validates them against the hardware. In essence, Cray has encapsulated a significant amount of expertise as part of their product build and deployment.

Up to four physical tiers are supported, with the choice of media up to the customer. The typical arrangement is as follows:

- **Tier 0** – Performance-optimized for high I/O and throughout (disk or SSD)
- **Tier 1** – Primary storage where live data lives most of time (disk or SSD)
- **Tier 2** – Capacity-optimized nearline storage (disk or tape)
- **Tier 3** – Extreme capacity- and cost-optimized for deep archives (usually tape)

From the user and application point of view, data access is transparent. TAS is front-ended by the user's local file system — NFS, CIFS, or (soon) Lustre. Underneath the covers is the HSM storage management engine, which acts as an intermediary between the native file system and the actual storage tiers. As such, it automates the data migration from tier to tier based on policies established for the user's environment. This automatic data management is designed to offer customers a flexible, cost-effective, and predictable way to optimize use of storage tiers.

In TAS, the HSM storage engine is provided by Versity, which Cray has partnered with to provide this key building block. Known as the Versity Storage Manager (VSM), it virtualizes all storage in a tiered environment. VSM is based on SAM-QFS (Storage and Archive Manager-Quick File System), an open source HSM platform originally developed by Sun Microsystems. Specifically, VSM offers a Linux version (the original client OS supported by SAM-QFS was Solaris) that also includes advanced data management features for enterprise and HPC users. Versity uses the open TAR data format to ensure long-term access to stored content.

The Versity manager provides an array of tools with which the user can define and execute storage policies. The general idea is that a file is migrated to a specific tier when its attributes match the criteria set up by the system administrator. For example, policies can be set up by file location, size, owner, and age. Multiple copies of a file can be defined to provide data security and failover in case of system problems. VSM also performs file validation to ensure data integrity.

TAS was designed for environments where the data needs to live forever and is dependent on continuous growth. Through its active migration/replication strategies, the system enables the storage infrastructure to be upgraded without downtime. This is a critical asset in many environments, especially for the growing number of applications that need 24/7 data accessibility.

With TAS, Cray is moving the tiered storage/HSM solution space forward in a number of ways, not least of which is simply providing a solution that can be delivered and maintained as a single system. With the upcoming integration of Lustre support on the front end, TAS will also provide a critical feature for HPC users with a need for managed tiered storage attached to supercomputers. Finally, by enabling data to be preserved indefinitely, the solution provides a level of protection and longevity that is rarely found with other commercial offerings.

CONCLUSIONS

Tiered storage under HSM presents the most general-purpose solution to data deluge problems. By balancing hardware performance and capacity requirements with the way data is actually used, this model promises the most cost-effective solution for a wide array of applications across HPC, oil and gas exploration, financial services, healthcare, security, and media/entertainment.

The challenge is complexity, particularly as associated with specific requirements for particular applications. That has relegated most installations to ad hoc solutions, leading to higher costs and uncertain futures. It also requires that system administrators and users learn how to deal with non-standard interfaces and protocols.

Cray's TAS solution addresses this complexity through end-to-end integration and support for open standards and interfaces (Linux, SAM-QFS, TAR). The company does so without tying the customer to particular hardware platforms. TAS also delivers the core requirements of a tiered storage system, namely scalability, performance, virtualization, facility efficiency, data management, data integrity, and data access. To that it adds single-point support, as well as Cray's reputation as a reliable provider of extreme-scale systems.