Disk-Based Technology for Multi-Petabyte Archives

Matthew O'Keefe, Alvarri, Inc., **Aloke Guha**, COPAN Systems, and **Peter Rigsbee**, Cray, Inc.

ABSTRACT: *Cray supercomputers create and process very large data sets, many of which are archived. The archives are commonly tape-based hierarchical storage management (HSM) systems. In this paper, we describe and discuss the usage requirements of current tape-based HSM systems and other data migration technologies. We then propose a potential alternative, a complementary strategy that uses disk-based MAID storage for deep archiving to speed access and improve data management scalability.*

KEYWORDS: *Network-Attached Storage (NAS), Storage Area Network (SAN), Hierarchical Storage Management (HSM), Massive Arrays of Idle Disks (MAID), HSM (hierarchical storage management).*

Introduction

Tape-based hierarchical storage management has been used for decades in data centers to store large, persistent datasets.[1,2] *Large* means approximately 1 to 10 Petabytes in 2007, scaling up to 1 Exabyte – 10^{18} bytes – in 2015. In the past, these datasets were too large to place on rotating magnetic disk storage due to the cost, space, and power requirements. These factors heavily outweighed the drawbacks of tape systems – lack of random access and slow speeds. In the past few years however, density increases (2x every 18-24 months) combined with price decreases have put disks at a par with tape.

Tape is more space-efficient and requires much less power than traditional disk-array-based storage. But new disk array technology known as Massive Arrays of Idle Disks (MAID) addresses the space and power differences with tape – it keeps most disks powered-down and uses aggressive error detection and prevention mechanisms to reduce drive failures and increase data reliability.

In this document, we outline requirements for large archival storage in HPC environments. We then review tape-oriented and file-oriented archive systems, as well as MAID storage. Finally, we address how MAID technology, combined with fast network attached storage, could address these large archive requirements in future systems.

2. Data Archiving and Migration Requirements for HPC

Large, frequently-accessed data archives found in some high performance computing data centers must meet an evolving set of requirements. These include:

Cost-effective, online archives that scale to 1 Exabyte and larger

Large digital data archives today are from 1 to 10 Petabytes. Growth rates are in the range of 25-100% per year, depending on site requirements. Therefore, by 2015 archive capacities that must be supported will be in the range of 1,000 Petabytes (1 Exabyte).

[1] J. Behnke, *et al.,* "EOSDIS Petabyte Archives: Tenth Anniversary," 22nd IEEE/13th NASA Goddard Conference on Mass Storage Systems and Technologies, Monterey, CA, April 2005, 81-93.

[2] R. Watson, "High Performance Storage System Scalability: Architecture, Implementation, and Experience," 22nd IEEE/13th NASA Goddard Conference on Mass Storage Systems and Technologies, Monterey, CA, April 2005, 145-159.

Data throughput to and from the archive should not reduce data usefulness

Tape archives like DMF have average transfer speeds in the 10s of MB/s range [(3)], assuming 10-20 relatively fast tape drives. Peak speeds for tape-based HSMs are in the range of several hundred MB/s [(4)]; but even large pools of fast tape drives cannot sustain this rate for workloads beyond simple large-file transfers. Today, MAID systems support transfer rates of over 1 Gigabyte/second per frame [(5)], where 20 frames would provide 10 Petabytes (assuming 750 Gigabyte drives) of archive storage, with raw transfer speeds over 20 Gigabytes/second. By 2015, disk-based archives should support speeds of from 1 to 10 Terabyte/second or faster over a range of archive access workloads.

Small File Accesses Should Be Reasonably Fast and Should Not Incur Excessive Overhead

Large supercomputing archives have been constructed with the assumption that small files are rare, but they are not rare.[6,7] As a consequence, performance of tape-based HSMs is typically very poor for small files. Small files should be supported in archives.

Whole Archive Migrations: reduce archive times from years to weeks

As tape capacities have increased, whole archive migration times have increased from months to years. In NASA's EOSDIS archive, for example, the latest 2.5 Petabyte migration took three years to complete.[8] Because of the slow speed of tape library systems, these long migrations are complex, error-prone with risk of data loss, and expensive in manpower and downtime.

Data Integrity: avoid data corruption end-to-end and point-in-time

Large archives imply large data transfers for creation and access. TCP/IP and storage technologies are vulnerable to silent data corruption when used in the development of large archives. Support for end-to-end data granule checksums can allow deep archives to detect and correct these errors.[9] Also, large archives based on disk and file system technologies need to avoid point-in-time data corruption errors caused by:

- user errors (for example, accidentally deleting a file or directory);
- operator errors (for example, accidentally initializing a disk, or a volume or partition misconfiguration);
- malicious user problems (for example, viruses); or
- system problems (including both hardware and software, for example, file system corruption or storage array failure resulting in unrecoverable disk block corruption).

Simple Maintenance: large archives should be maintained operationally by a small staff with standard industry skills

Operational aspects of large archives shall be simplified so that one full time equivalent employee with moderate levels of data center experience with the requisite technologies can maintain the archive indefinitely (including migrations to new, higher capacity media). In addition, the complexity to manage the archive must not increase as the scale of the archive increases.

Security: archive contents shall be protected against unauthorized access

Security mechanisms must be in place to prevent unauthorized access to archive contents. The archive space should be able to be partitioned, so that several domains of trust can exist within the same physical archive.

Indexing: large archives should support sophisticated metadata implementations with data classification and search capabilities.

When data is placed in the archive, metadata should be required to support data classification, indexing, and search. This can assist in both data placement and long-term data management in the archive.

[3] Ken Gacke, CR1 Silo Engineer Study, DIS Digital Archive, USGS National Center, EROS, Sioux Falls, SD, May, 2005.

[4] Watson.

[5] COPAN Systems current MAID implementation contains 896 disk drives and 8 Fibre Channel interfaces. Using 750 GB drives and factoring in 25% overhead for RAID and spares, each frame provides approximately 0.5 Petabytes of disk storage.

[6] M. Butler, "Storage Issues at NCSA," 19th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies, College Park, MD, April 2002, 93-108.

[7] J. Neil, "Characterizing Long Term Usage of a Mass Storage System At a Super Computer Site," 18th IEEE/10th NASA Symposium on Mass Storage Systems and Technologies, San Diego, CA, April 2001, 313-326.

[8] Behnke, *et al.*

[9] Ibid.

Open Architecture:

Historically tape-based HSMs have been closed, with a few exceptions. They are designed to work at the file system level on specific operating systems. Files are migrated from disk cache to a large tape backing store via a proprietary protocol specific to each HSM. An open, distributed architecture with published APIs, data formats, and an open source reference implementation could free large archive designers from dependence on a single company.

Unified Data View:

Ad hoc techniques are generally used to transfer files from other machines in the data center into a tape-based HSM, so that there is no unified view of the data (*e.g.,* its ownership, type, and associated management policies) across multiple systems. Data migration policies are implemented within the confines of the HSM, after a data copy from the original file location. There is no unified, open migration strategy across multiple, distributed systems, yet data centers continue to become more distributed over time.

The importance of each requirement is site-dependent, and certain requirements will be irrelevant to some sites.

3. Existing Archive Technologies

Data migration products manage the mapping of large, persistent datasets onto multiple tiers of secondary storage devices. Data accessed more frequently is kept on higher tiers (with low latency access, high bandwidth speeds, and small capacity) while data accessed less frequently is kept on lower tiers (with higher latency accesses, lower bandwidth speeds, and higher capacity).

Traditional tape-oriented hierarchical storage management (HSMs) used in high performance computing typically have two tiers of storage: disk and tape. Data is held primarily on tapes, but with complex mechanisms it can be migrated to disk cache for faster access. Therefore, tape storage systems view disk cache storage as a precious resource that must be conserved. This leads to *file-system-based data migration* where files are moved between tiers based on global statistics derived from the whole file system (most commonly, the amount of free space remaining on disk for the file system). In contrast, in *file-specific data migration*, files are moved based on per-file attributes (such as file type, name, pathname, and age). Data migration in commercial (non-HPC) products is typically file-specific, while filesystem-specific data migration is the norm in high performance computing. Hybrid approaches that use file-specific information to migrate sufficient files to meet a filesystem-based utilization goal are also possible.

File-System-Based Data Migration

There are several filesystem-based data migration products. They are tape-based HSMs that support large tape archives and a disk cache. As files in the disk cache age without further accesses, they are migrated from the disk cache to the tape-based backing store.

High Performance Storage System (HPSS)

Developed by the U.S. Department of Energy, the High Performance Storage System (HPSS) supports extremely large, petascale data archives.[10] HPSS is used in several dozen of the largest supercomputing centers in the world, as it is scalable in capacity and performance. The HPSS architecture is based on a *mover* abstraction that virtualizes sources (typically supercomputers or workstations) and sinks (typically file or tape servers or SAN-attached disks) of data. A *metadata service* tracks data as it is originally stored and then migrated within the HPSS storage tiers. In general, HPSS tries to create abstract interfaces that can be wrapped around commercial protocols and product implementations to reduce its dependence on such external technologies. There are significant costs in system complexity and development effort to support the abstractions in this architecture. However, this approach has allowed HPSS to use new technologies and evolve over a long period – a critical feature for large data archives.

HPSS is highly configurable and adaptable. It supports a wide variety of media, network interfaces, data transfer protocols, and operating systems. As a result, it is complex to configure and maintain, and it requires significant staffing (three to five or more mass storage administrators) per site.

SGI Data Migration Facility (DMF)

While DMF [(11)] is also a tape-based HSM like HPSS, it looks and feels like a traditional UNIX application and does not support a mover abstraction. DMF automatically detects a drop below the filesystem free-space threshold and migrates selected data from expensive online disk to cheaper secondary storage, such as tapes. DMF automatically recalls the file data from offline media when the user accesses the file with normal operating system commands. You can also manually force a file to be migrated or recalled.

[10] Watson

[11] DMF Administrator's Guide for SGI® InfiniteStorage, SGI, 2005.

Though simpler than HPSS, there are still 40+ commands in its CLI. DMF files also can be in more than 10 different states and, given the distributed nature of DMF storage tiers, there are complex state transitions that introduce intricate failure modes. This is common to all tape-based HSMs. A nearly 200-page manual is provided to cover maintenance and recovery from these complex failure cases. Relative to other HSMs on the market, DMF is relatively inexpensive.[12] Market data as of May 2005 suggests commercial HSM product pricing is in the range of $350,000 for 500TB, plus an additional 20-25% per year in support costs.[13]

Other File-System-Based Data Migration Systems

Other tape-based HSMs include Jasmine (14), Enstore (15), and Castor (16). All have architectures similar to HPSS and DMF: a disk cache in front of a large tape backing store with files migrated from disk-to-tape based on usage and file size, a metadata database to record and track file states, and a large user manual. Like DMF, migrations happen when file system reaches some limit on capacity usage (typically 60%-80%). Migrated files are moved back from tape to disk if accessed again. Commercial HSM products include Sun's SAM-FS and ADIC's Stornext. Both products are similar to SGI's DMF in functionality.

In all these tape-based archive systems, managing tape problems is a complex, error-prone, and labor-intensive process.[17,18,19] For example, in the Jasmine system (20), when a tape fails to mount in a tape drive, "both the tape and the tape drive are removed from the system and flagged as suspect. This allows a system administrator to examine both of them and make a decision. This prevents a bad drive from mangling multiple tapes, and it prevents a defective tape from being passed from drive to drive." Jasmine also tracks error counts per file, per volume, and per drive to spot impending tape or drive failures and then migrate data off marginal tapes.

Tape-based HSMs are being configured with more disk cache to improve performance and avoid some of the bottlenecks associated with tape, but large amounts of disk cache are required to get reasonable (>50%) hit rates. In a study comparing a MAID-based archive to a tape-based HSM with a disk cache providing a hit rate of 75% (this number is much higher than observed cache hit rates, and would require a very large disk cache), it was discovered that the throughput achievable with MAID was two to three orders of magnitude greater for general file access patterns.[21] Fundamentally, the disk cache-to-tape-backing store model adds performance overheads and management complexity. This limits the utility of disk caches in tape-based HSMs.

File-Specific Data Migration

In contrast to filesystem-based data migration products, file-specific data migration tools generally migrate files from more expensive disk storage to cheaper disk storage instead of to tape.[22,23] In addition, they use simple mechanisms such as symbolic links to connect one storage tier to another, instead of complex, kernel-dependent mechanisms like XDSM/DMAPI. A major advantage of the use of multiple disk tiers (without the use of tape) is that migrated files continue to be accessible with reasonable latency and bandwidth. (There are of course certain situations where files need maximum performance, such as in a transaction database, and migration for these is undesirable.) This eliminates much of the "cache management" complexity inherent in tape-based archives, where migrated files must be migrated back to disk before they can be used again.

BlueArc Data Migration

BlueArc Corporation (www.bluearc.com) has developed a fast NAS filer that migrates data from expensive FibreChannel storage to lower-cost SATA-based storage. Files are migrated from one tier to another based on a file's most recent access time, size, type, name, or directory location. Instead of migrating from a disk cache to a tape-based backing store, migrations are from one file system (located on a Fibre-Channel-based volume) to another (located on a SATA-based volume).

[12] Gacke.

[13] Ibid.

[14] B. Hess, *et al.*, "The Design and Evolution of Jefferson Lab's Jasmine Mass Storage System," 22nd IEEE/13th NASA Goddard Conference on Mass Storage Systems and Technologies, Monterey, CA, April 2005, 94-105.

[15] G. Oleynik, "Fermilab's Multi-Petabyte Scalable Mass Storage System," 22nd IEEE/13th NASA Goddard Conference on Mass Storage Systems and Technologies, Monterey, CA, April 2005, 73-80.

[16] O. Barring, *et al., "*Castor: Operational Issues and New Developments,*"* Computing in High Energy Physics, 2004.

[17] Behnke, *et al.*

[18] Hess, *et al.*

[19] Oleynik

[20] Hess, *et al.*

[21] Guha

[22] J. McCarron, "Acopia Adaptive Resource Switch Test Plan," Acopia Networks, 2005.

[23] J. Wendt, "Rein in NAS with File Virtualization," Storage, March 2007

Once the file contents are migrated to SATA storage, a symbolic link to the migrated file replaces the original file.

Migration templates are provided but system administrators can also configure migrations with precise rules. These rules are based on file attributes that include or exclude files according to sets of conditions related to file attributes (for example, migrate files > 100 megabytes that have not been accessed in the last 14 days). Data migration rules are fairly simple to configure. Once defined, they can be integrated into a data migration policy which applies preconditions based on file system usage to determine which rules to apply. A common policy configuration would be to have increasingly aggressive migration rules applied as more first-tier file system space is occupied. Policies can run on a regular schedule or be run at a single point in time. Both rules and policies can be configured using a straightforward GUI.

Note that migrations in the other direction, from SATA to FibreChannel, can be performed manually, if necessary.

Enigma Data Systems

Enigma Data Systems Inc. (www.enigmadata.com) has developed a data migration software package known as *SmartMove* that allows file migration between two file systems that may exist on separate servers or filers. The software typically runs on a third server and files are accessed through either NFS or CIFS. When migrated, a symbolic link is made from the original file to the migrated copy. Unlike most filesystem-specific data migration, this approach requires neither kernel modifications nor tight integration with the operating system.

SmartMove is written in Java. It is supported on Linux, Windows, and Solaris. It identifies files for migration based on file name, location, owner, size, or last time accessed or modified. Migration policies are developed around *containers* (top-level directories) and *projects* (subdirectories within containers) to which specific filters may be applied. Filters specify files that are migration candidates. A group of filters from which a group of file migration candidates is culled is called a *rule*. Filters can be configured in any combination to develop rules. Data migrations can be run manually or on a fixed schedule. Files may also be returned to their original location manually or via a regularly scheduled data migration back to the source. A variety of reports can be generated to provide statistics on file migrations. Predictive analysis for a particular rule set can be generated to determine if the rule and its filters can achieve the data migration goal (for example, migrating at least 500 MB of a certain project's data). Multiple storage destinations (the target file system for a migration) can be defined. All are managed by a single policy engine.

SmartMove also supports an archive function for projects and containers. Here, archived files are named, indexed, copied to the destination, and then deleted from the source file system. Files can be brought back from archive storage to the first tier file system.

Hybrid Data Migration

Hybrid data migration employs both file-specific and file-system specific techniques. Using hybrid data migration, the specific files to be moved are described, along with a target utilization metric for the file system.

Acopia Networks

Acopia Networks (www.acopia.com) has developed a file virtualization switch based on the NFS and CIFS protocols.[24] This switch is placed between clients and servers, and presents a single namespace that is constructed by coalescing the namespaces of the independent NFS mount points and CIFS shares. It is then possible to migrate files from one NFS file system or CIFS share to another based on usage goals for the source file system. The Acopia ARX switch can be configured to migrate specific files based on attributes. The ARX switch attains a high level of client-to-server transparency: files may still be accessed while they are migrating; and physical servers can be moved without impacting client configurations.

4. Copan MAID Storage: A SATA-based Replacement for Tape

A new kind of disk array, known as a Massive Array of Idle Disks (MAID) has been developed by COPAN Systems. It offers a promising alternative to tape archive and backup. The MAID concept, introduced by Grunwald and Collari in 2002, uses an array of disk drives that are powered down individually or in groups when not being accessed. [25] This lowers the non-disk overhead compared to traditional disk-based systems.

COPAN has applied this principle in a MAID storage array where 75% of the drives are powered down

[24] Acopia Adaptive Resource.

[25] Dennis Colarelli and Dirk Grunwald, "Massive Arrays of Idle Disks For Storage Archives," Proceedings: Supercomputing '02 Conference, 2002.

(yielding a power on duty cycle, D, of 25%), lowering power requirements and disk failure rates dramatically (by a factor of 1/D) and increasing data reliability. In addition, a technology known as disk aerobics is used to regularly exercise and test all drives to detect pending disk failures in advance through predictive monitoring. A drive that is expected to fail can be proactively retired and replaced by a spare to avoid a RAID rebuild due to an unexpected failure. In a sense, this predictive analysis and automated proactive replacement is an example of how the complex, error prone, labor intensive process of tape error management can be simplified and automated using disk technology.[26,27,28]

COPAN's MAID system uses a three-tier physical architecture that includes:

- a System Controller:
  - Fibre Channel interfaces for external attachment and routes external requests to storage shelfs;
  - runs application software to implement VTL (virtual tape) and Millenia Archive software interfaces;
  - runs monitoring and management software, the system controller is the centralized control unit for the COPAN MAID array;
- a Storage Shelf:
  - storage paths run from the system controller to each storage shelf;
  - the 25% drives-powered-on-ratio is applied per shelf, so that no more than 25% of the drives in a particular shelf are powered-on;
  - each shelf has 8 canisters;
- a Storage Canister:
  - implements basic physical packaging for drives;
  - there are 14 SATA drives per canister.

The drives are grouped into 3+1 RAID arrays. COPAN uses specialized RAID techniques that avoid striping writes and reads across all four drives in the RAID group. This reduces the number of drives and amount of power required for each read and write operation. There are 27 RAID groups per shelf, plus four spare disks, for a total of 27 x 4 = 108 disks assigned to RAID groups. There are 4 spares per shelf, which yields a total of 112 drives per shelf. COPAN's MAID system can host up to 8 shelves per frame for a total of 896 drives.

The RAID groups are arranged so that each disk in the group is in a different slide-out canister. RAID group 0 uses disks (0,0), (1,0), (2,0), (3,0) where the first number in each pair is the canister number, and the second number is the disk number.

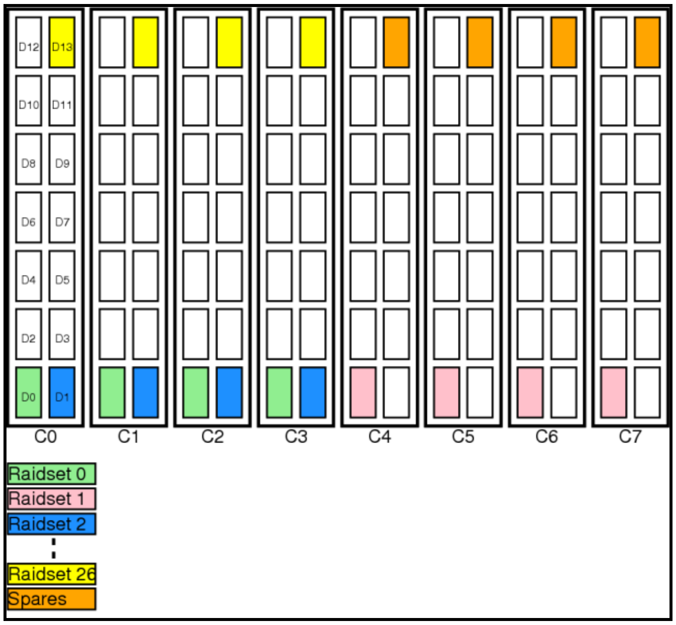

Figure 1: COPAN Array

This diagram represents the initial layout. It changes should a drive(s) fail and spares are pressed into service – failed drives are then replaced and become spares. So over time, the mapping of disks to RAID groups within the shelf is randomized. RAID groups and spares are organized per-shelf, so a RAID group can not span multiple shelves. Likewise, spares can only replace failed drives in the same shelf.

Basic MAID attributes — high density, low power requirements, high reliability and low failure rates — resemble tape in cost and physical characteristics. But MAID performs more like disk in terms of speed and access latency. If an access is made to a powered-down drive, it is powered up in about 14 seconds. All accesses after the disk is powered up are at disk drive speeds. Compare this to tape accesses which can range from 60 to 100s or more seconds.

The primary alternative to MAID is nearline storage arrays – standard always-on RAID arrays with SATA disk drives – with a Virtual Tape Library interface. These products have been successful in the market because they provide a mechanism for fast restore from backups, have faster throughput, and reduce data duplication. These products use much more power and space compared to tape and to MAID technologies.

[26] Behnke, *et al.*
[27] Hess, *et al.*
[28] Oleynik

5. Fast Network Attached Storage and MAID: Leveraging Strengths to Achieve End-to-End Data Management

The Data Archiving and Migration Requirements outlined in Section 2 render difficult the use of traditional tape-based HSM technology. The biggest problem is that tape error management is a labor-intensive process that requires significant system administrator resources. Small-file read/write activity is very hard on tape-based HSMs, because tape access is sequential and many transactions are required for each file state change in a complex system. Sophisticated resource schedulers are also required to manage scarce resources (tape drives) and to balance file access workloads, but these have been historically ineffective due to high tape access latency.

That said, there are many lessons to be learned from tape-based systems [(29)]. Some can be applied to resource scheduling of a limited number of powered-on MAID volumes. It is likely that tape-based HSMs will continue to be used for the foreseeable future for archive data that is rarely accessed. However, the toughest challenge for extremely large tape-based archives will be when they are migrated to new data formats and tape technologies.[30, 31]

File-specific data migration from one file system (expensive, fast) to another file system (cheaper, slower) is one way to accommodate large archive migration. Hybrid Data Migration (HDM) combines file-specific and filesystem-specific migration techniques. These techniques are used in file virtualization technology [(32)], such as switches from Acopia Networks. HDM works independently of operating systems, allows for fast small-file access, and satisfies most requirements discussed in Section 2, except those related to high-density, low-power, and cost-effective media. Tape technologies can meet high-density, low-power, and cost-effective media requirements, but tape can not meet all requirements.

It is impractical to map a destination file system directly to a tape drive. Since MAID systems are based on disk technology and support disk interfaces, MAID storage could potentially support file system accesses. A MAID-based filer could be integrated into a hybrid data migration solution to help meet all requirements outlined in Section 2.

There are two primary issues that must be addressed when implementing a distributed data migration strategy such as HDM across multiple, distributed file systems to a MAID-based filer. These are discussed here.

1. Implementing Data Migration

Files must be migrated from compute and file servers distributed across a network. One way to do this is with out-of-band software that migrates file data from distributed systems to a large MAID-based filer. Or, an in-band file-level switch can be used to create a global name space with file-specific and/or filesystem-specific data migration. Either of these two, software or switch, is the logical equivalent to migration software that moves data from disk cache to tape backing store in most HSM's. The difference is that this migration happens across a network of distributed systems rather than on a single server.

It is also important that a unified data model be supported. It should support data integrity checking from the application level on the source machine, across the network, to the destination machine. This model could be used to support storage resource management, lifecycle planning, and capacity management. It could also give a data center director a global view of all data on his or her storage resources, allowing more efficient and effective storage deployment planning and practices.

In addition to better performance and relative simplicity, MAID-based filers can use the sophisticated and widely-used NAS management features available in file systems today. These include snapshots, volume management, virus detection and removal, write-once / read-many file systems, NDMP backup, and replication. These techniques can be used to prevent or recover from user errors, prevent virus damage, and restrict access to sensitive archive content. Damage to disk and file systems can be prevented through metadata replication, RAID techniques, and partitioned file system structures.

2. Implementing a MAID-Based File System

A MAID-based filer requires a file system design that carefully manages metadata and data placement so that frequent, small-block metadata accesses are mapped to always-powered-on storage. Data accesses must be managed so that drive power cycling is kept to manageable levels while providing sufficient IO throughput to support data migration and on-line archive access.

[29] Behnke, *et al.*
[30] Ibid.
[31] Barring, *et al.*
[32] Wendt

MAID has powerful potential to replace tape as the backing store for an archive: it has sufficient density, low power requirements, and reasonable media costs.[33,34] And, since a file system can be mapped to a MAID storage device, it is possible to migrate file-specific data on low-cost media, and still reduce management costs and greatly increase throughput and performance compared to tape-based HSM.

6. Summary

As data archives continue to grow, large archive construction and migration becomes more and more difficult to accomplish on tape-based HSMs. A promising alternative is to use MAID-based filers and distributed data migration across multiple systems. This would both create and support extremely large archives with good performance (speed), scalability (capacity), and simplified management, all at reasonable cost.

About the Authors

Matthew O'Keefe is a founder and Vice-President of Engineering at Alvarri Inc., a start-up focusing on storage management software. Previously, Matthew founded Sistina Software, sold to Red Hat, Inc. in late 2003; he spent 10 years as a tenured Professor at the University of Minnesota, where he is currently a Research Associate Professor. He can be reached at okeefe@alvarri.com.

Aloke Guha is a cofounder and the CTO of Copan Systems, a storage system company delivering long-term persistent data solutions. His previous positions include CEO of Datavail (CreekPath/Opsware), Vice President and Chief Architect at StorageTek (Sun), and CTO of Network Systems (StorageTek). He is a senior member of IEEE, has authored over 25 patents (12 issued/allowed), and over 60 technical publications. He holds a B. Tech. (EE) from the Indian Institute of Technology (Kanpur), an MSEE and Ph.D. from the University of Minnesota, and is a graduate of Stanford University's School of Business' Executive Education Program.

Peter Rigsbee is a senior product marketing manager at Cray Inc., where he handles product management and marketing of Cray's storage strategies and products. Peter has over 20 years of experience in HPC at Cray Inc., SGI, and Cray Research, where he has held a variety of managerial and technical roles in both marketing and engineering. Peter also spent several years as a product marketing manager at StorageTek. Peter has a BS degree from the Massachusetts Institute of Technology (MIT) and an MBA from the University of St. Thomas (St. Paul, MN).

[33] W. C. Preston and G. Didio, "Disk at the Price of Tape," Glass House Technologies, 2004.
[34] Colarelli and Grunwald