WHITE PAPER

Addressing the Problem of Inactive Data Filling Up Expensive Active Disk Silos

Creating an Active Archive Strategy to Address Both Archive and Backup In the Midst of Data Explosion

By Floyd Christofferson, SGI
April, 2011

According to virtually every study, analysis, pundit or perspective, the growth of file-based enterprise data is skyrocketing. Gartner research from March 2011 forecasts a compound annual growth rate in raw terabytes of external storage arrays at 55% over the next 5 years. This translates in a growth from the 2010 number of 11.8 million terabytes sold to a projected volume of 107.5 million terabytes in 2015.

This growth rate does not include drives inside laptops or desktop computers. It doesn't include drives that might be used in a myriad of other devices and technologies. This growth directly represents the increase in storage infrastructure for business, new file-oriented applications used in both enterprise and technical computing, disk-based backup and archive deployments, and expanded server and desktop virtualization projects.

Gartner is but one to make this prediction. IDC, Frost and Sullivan and almost any IT manager will tell you the same thing. All data is exploding exponentially and in the process causing primary storage and backup infrastructures to grow massively and expensively to keep pace.

There is a companion problem that goes along with the issue of data and infrastructure growth. The problem is that even though more and more files are filling up ever-larger disk silos, the utilization of those files does not necessarily increase at the same rate. In other words, people may be creating more and more files, but they are still using them only a few at a time. My own hard drive has tens of thousands of presentations, documents, photos, emails, and other files, the large majority of which I have not touched in months or years. And yet, I want them available at all times for when I might need them.

Translate this to an enterprise and the problem becomes astounding. Because the problem moves from the realm of personal preference: (I want my files available all the time) to business necessity: (my business needs to have access to its data at all times).

Researching the Problem

In a 2008 study done by the University of California, Santa Cruz funded by the National Science Foundation, an active storage pool of 22TB used by 1500 employees in business and technical workflows was analyzed for utilization of network file system workloads. In other words, they studied usage patterns for the type of workloads used in virtually every enterprise in the world.

What did they find? Files live an order of magnitude longer than in previous studies. Files are rarely reopened; 95 percent are reopened fewer than five times. Over 60% of file reopens occur within a minute of the first open. Over 76% of files are never opened by more than one client, and of those that are opened by others, 90% of sharing is read only. Finally, most files are not re-opened once they are closed.

The net of this is that in that environment of 22TB, most of the files sitting in those arrays are not going to be reopened or changed. And yet, just like the files on my laptop that I haven't touched in a couple of years, business users have a very difficult time determining which files to delete or remove from active storage. And so datacenter disk infrastructures keep growing at an astronomical rate. This problem is compounded by the cost of that growth. Not just in the acquisition of new disk arrays, but in the cost of backing up those arrays, the cost of adding data center space, and the cost in electricity and cooling of the data centers for disk drives that are spinning continually, but are seldom being used.

Translated into real numbers, a leading network attached disk-based storage array uses about $91 per TB of power when operational, or 32 KwH per cabinet. For a 2 petabyte system, this translates into $190,000 per year in operational power costs alone, at typical U.S. utility rates, not including data center space and cooling costs. Yes, the data is available for users to access at all times. But at what cost?

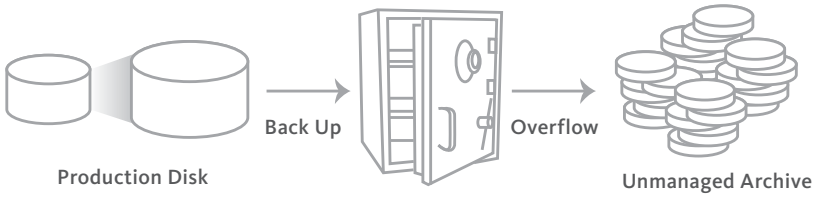
Inactive Data Sitting on Active Disk

The solution is as simple as it has often been beguiling. Create an active archive where the data is available in an 'online' state for easy access, where the data is protected for extremely long-term retention, and where the operational cost is extremely low. If any of those three elements is missing, archives strategies tend to fall flat.

The problem is that most archive solutions only address part of the issue. When backup (protection of active data) gets confused with archive (retention of inactive data), data paralysis occurs. Backup and restore times become impossible to manage because they involve inactive as well as active data. Seldom-accessed data becomes hard to find. Operating costs skyrocket when additional production disks are needed just to keep up with the relentless growth of data.

What's worse, the excessive growth of data stored on production disks can be a contributing factor to data becoming segmented into incompatible silos. This makes collaboration between different areas either impossible, or at best a manual process prone to error and wasted effort.

Users deal with files but are forced to work within file systems. The job of a proactive data management strategy—an active archive strategy—is to let the users focus on their work, and not waste time, infrastructure, or energy on just setting up to do their work. That is what the tool kit that comprises active archiving solutions enables. Namely, a way for IT managers to keep data accessible, affordable, and protected without requiring users to add work to their day understanding how it is done or where it is, or what steps need to be taken to get there.



Key Concepts: Distinguishing Between Backup and Archive

At the heart of resolving the layers of problems described herein is sorting out an integrated approach to archive and backup. This makes data accessibility the priority to users while still allowing for cost containment and data protection priorities for IT managers.

The problem is that backup and archive are often confused. Not necessarily in concept, but in day-to-day practice.

The process often goes something like this:
As data continues to grow, primary or production disk arrays fill up and must be expanded. As noted above, this is typically a mixture of active data with older, seldom-used data. Backup is needed to provide protection for primary disk, and as the overall data volume grows, the backup windows grow as well.

Whether it is due to the inability to even get backups done within the available time windows or because the backup environment simply fills up, IT managers are often left with no choice but to take excess backup data and put it on a shelf as an 'archive.' It is either that or grow their environment. The problem is that this 'archive' is not a true archive at all, and is often unmanaged. Data with a high time value is mixed in with low-value data.

Also, when archive data is taken from backup data it becomes haphazard and incomplete. The data, which may have significant time value, is offline to users and is often irretrievable without significant cost, effort and time.

Worse, this approach means that the backup environment must continually grow to keep pace with the expansion of the production disk environment. That adds costs without actually solving anything.

This is a common problem across industries that are otherwise extremely careful with their data production and protection. Indeed, it is because of the high volume of data and the difficulty in managing the distinction between high-value and low-value data that many IT managers are left with few options but to keep everything, which merely compounds the problem.

The solution to this problem is to create a clear distinction between backup and archive, and to decouple the needs of data protection from those of data retention.

Backup strategies should be for short-term production data. They are to protect what is done in the short term in case of catastrophic failures.

Archive or data retention strategies on the other hand are long-term by nature. Disaster recovery protection is still needed for this data, but does not need to be done within the tight time window needed by backup.

Backup	Archive
Copies Data	Moves Data
Supports Operation and Recovery	Supports Business and Compliance
Supports Availability	Supports Operational Efficiencies
Short Term In Nature	Long Term In Nature
Data Typically Overwritten	Data Typically Secured, Not Overwritten
No Historic Revelance	For Historic Information
Not Easily Searched	Easily Searched

Building an Active Archive Strategy

An active archive is one in which all data is always available in an 'online' state all the time. An 'online' state does not mean that it is taking up expensive primary disk capacity. In the context of an active archive, 'online' means that the data is available in an environment that is immediately and easily accessible to users, that is not drawing power or taking up space, and one in which the data is protected for very long retention.

In fact, when properly applied an active archive strategy significantly reduces the overall storage and data management costs while at the same time increasing efficiencies and the ability of users to access all data. Using an active archive strategy, costs are contained because the production disk does not need to grow very often. Inactive data that still has retention value is moved into an archive tier storage that, although 'online' and visible to the user, is typically in a powered-down state using MAID technology that completely removes power from the array. These archives, while still available to users, can be managed with very different disaster recovery techniques, and at a fraction of the operational costs of conventional disk-based file stores.

Thus, data lives where it is most efficient. Online, active data is contained only on the primary disk arrays. Seldom used data can be moved to an archive tier, either programmatically or manually, where it remains available, but in a low-cost protected state. Since data growth is not managed by increasing production disk, backup of those arrays also does not mushroom. Backup is reserved only for active data, keeping costs down while reducing backup and recovery times.

Implementing an Active Archive

There are numerous tools that can be employed to implement an active archive strategy. These will vary by industry, by use case and workflow. Not all are needed in every circumstance. In fact, what is most important in devising a strategy is to approach data growth proactively from this whole-system perspective rather than reactively by throwing more disk arrays at it to solve short-term problems. As we have seen, short-term solutions compound the overall problems, always leading to higher costs and risk.

Some tools to consider:

- **Digital Asset Management Solutions:**
  A key problem in determining whether data is active or not is in the strategy used to classify it. This problem is compounded when production data is distributed across multiple silos.

  Leading digital asset management solutions such as LiveArc™ from SGI enable content to be indexed automatically in multiple ways as it is created and modified. Users can search for data, and administrators can easily set policies to determine which data should remain on production disk and which can migrate to second or third tier storage.

  Another key benefit of a digital asset management platform like LiveArc is the ability to bridge multiple namespaces, or data silos, to provide a global view across all storage, data and metadata types needs. In this way, IT managers have complete control over their environment and can implement back-end changes without impacting users. Users don't need to know or care where the data actually is in the hierarchy of storage infrastructure because it is always visible to them within the management interface on their desktop.

- **Hierarchical Storage Management (Tier Virtualization):**
  Another key practice that can aid in developing an active archive is to virtualize tiers of storage through the use of a hierarchical storage management solution, such as SGI DMF (Data Migration Facility). DMF enables multiple tiers of disk and tape to appear to the users as one large aggregated volume even though the data is actually distributed across multiple storage types.

  For example, production disk is typically higher performing (and thus higher cost) disk, according to the needs of the industry. But only a fraction of the data is active, so expensive disk arrays are used to house inactive data.

  With an HSM solution like DMF, the expensive high performance disk is linked with 'nearline' or cheaper capacity disk. This in turn can sit in front of a MAID solution to power disks down or a tape library.

The beauty of this system is that all the data appears to the user to be online in the expensive production disk at all times. But in reality, even though the file appears to be right where the user put it in the file system, it is actually migrated to lower cost disk which results in dramatic overall savings, without the need for users to wonder where their content is located.

With a solution like DMF, the rules for when or if the bulk data migrates can be established by policies, such as file type, time since it was last accessed, etc. In addition, since DMF can manage multiple copies of the same file, backup becomes optimized to a significantly smaller amount of data.

- **Low power mass storage using MAID (Massive Array of Idle Disks):** A MAID solution is another significant tool in creating an active archive by selectively powering down whole sections of the disk array until the data is needed. This dramatically reduces the power and cooling requirements of the data center, much like tape libraries do, but with the added advantage of much higher performance and proactive data protection. SGI® COPAN™ 400 MAID and VTL solutions are best of breed in this category, providing extremely high density and low power options. See additional SGI White Papers on COPAN MAID technology solutions.

- **SGI's ArcFiniti™:**
  A file-based archive solution blends the best of these technologies into a fully integrated network accessible archiver. Go to http://www.sgi.com/arcfiniti/ for more information about this innovative new platform that makes the dramatic cost savings of active archiving accessible and easy to every data environment.

Protecting the Data That Is Your Business

An active archive strategy requires planning and tools, but when done properly it can dramatically reduce the overall costs of managing a growing pool of data. More importantly, by de-coupling production disk from other tiers of storage, single points of failure are virtually eliminated. Individual components can be upgraded or changed without impacting overall utilization for users, scalability then becomes an asset in this scenario, and not a headache.

Corporate Headquarters
46600 Landing Parkway
Fremont, CA 94538
tel 510.933.8300
fax 408.321.0293
www.sgi.com

Global Sales and Support
North America +1 800.800.7441
Latin America +55 11.5185.2860
Europe +44 118.912.7500
Asia Pacific +61 2.9448.1463