



Data Archiving Best Practices Using a Perpetual Storage Tier



Contents

- Introduction 4**
- Rethinking the Storage Paradigm 4**
- What is the Paradigm Shift in Storage? 4**
 - The Traditional Storage Paradigm.....4
 - A New, Modern Storage Paradigm for
Managing and Storing Data.....5
- What is Modern Storage Lifecycle Management Software? 7**
- Dynamics of the Perpetual Storage Tier 8**
- Summary 9**
- Footnotes 9**

Copyright ©2020 Spectra Logic Corporation. All rights reserved worldwide. Spectra and Spectra Logic are registered trademarks of Spectra Logic. All other trademarks and registered trademarks are property of their respective owners. This document is provided for information purposes only, and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law. We specifically disclaim any liability with respect to this document, and no contractual obligations are formed either directly or indirectly by this document. All opinions in this white paper are those of Spectra Logic and are based on information from various industry reports, news reports and customer interviews.



Introduction

Many organizations have large amounts of data that are critical to their businesses and are growing rapidly. Often these organizations store all data, active and inactive, on an expensive Primary Tier of storage intended for active data. But upwards of 80 percent of data is typically inactive, and therefore, being stored on the wrong tier, costing millions of dollars a year. Lower-cost storage tiers are readily available, but existing solutions to migrate the data to these lower-cost tiers are complex and costly, inadequately protect migrated data, and don't address all of the migration workflows needed by organizations.

What is StorCycle?

StorCycle® by Spectra Logic is storage lifecycle management software that identifies inactive data and migrates it to a lower-cost Perpetual Tier of storage, which can include cloud, object storage disk, NAS, tape, and even replication sites, ensuring the data is stored safely while making it easily accessible to end users. Users can choose project-oriented directories to archive or let StorCycle automatically identify and migrate inactive data. By removing inactive data from the Primary Tier of storage, administrators can prevent their primary storage from filling up and reduce the need to purchase additional primary storage. A smaller primary storage tier will also reduce backup windows, reduce costs, and increase overall performance thus freeing capital that could also be used to move to a faster tier of primary storage that includes NVMe, Flash or other solid-state disk storage (SSD).

StorCycle can also be used to make additional copies of active data for disaster recovery (DR) purposes. Multiple copies of data can be made in the Perpetual Tier of storage and can be protected from ransomware and other types of malware by implementing an "air-gap strategy." Data in the Perpetual Tier is easily linked to its original source location and made accessible by users in a transparent or semi-transparent manner with the use of Symbolic links*, HTML links, and a web-based search.

StorCycle ensures that data is stored on the right tier throughout its lifecycle, reducing storage costs up to 70 percent and lessening the amount of primary storage needed, backup required, new storage purchases made, storage administration required, cloud egress costs paid, as well as power, cooling and floor space needed.

*In a future release

Rethinking the Storage Paradigm

For too long predictions of “storage doom” have lamented the growth of data. Given that data is now widely accepted as the single greatest asset of any given organization, it seems that data growth should be a point of anticipation, not dread. Often the naysayers shift the point of contention from data growth to the physical storage itself -- pointing to cost, physical footprint and management as the evils. Virtually every new storage technology is touted as the solution to these problems, yet no single technology has proven to be the answer. And counterintuitively, the combination of new technologies often leads to further complexities, costs and even isolation of information.

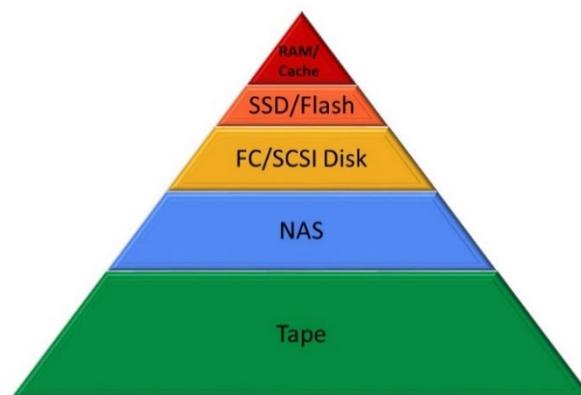
One of the few things that seems to have escaped the debate of long-term, large data storage is the *storage model* itself. This paper suggests a paradigm shift in how we approach the storage of information, including the management of information; access to information; sharing of information; protection of information; and the costs associated with all of the above.

What is the Paradigm Shift in Storage?

While it may seem like a daunting task, most organizations already have the majority of what they need to make the important paradigm shift – allowing IT administrators and content curators to welcome rather than dread the growth of data.

The Traditional Storage Paradigm

The storage pyramid is one of the most commonly applied storage models in our industry. It’s usually represented as three tiers, but could depict any number of tiers or combinations of storage technology entities – cache, RAM, SSD/flash, FC/SCSI disk, SAS/SATA disk, tape, optical, etc. It makes the important observation that the top of the pyramid is the most responsive, costliest, least dense, and smallest amount of storage in the ecosystem. All of those attributes flip as data proceeds down the pyramid. The lowest level of the pyramid is the least responsive, least costly, densest, and typically accounts for the largest amount of storage in the ecosystem.



Storage Paradigm



While that basic concept of the storage pyramid is as relevant today as it was 30 years ago, it's a model that doesn't address the newer challenges of modern storage.

With the introduction of the public cloud and object storage technologies, the hierarchical nature of the traditional paradigm becomes less effective. This is especially the case when different storage technologies are used in similar roles – SSDs and HDDs both used in the top tier; disk and tape both used in backup; and tape and cloud both used for disaster recovery (DR) or offsite storage. The roles of these technologies may be similar, but there are granular differences that enable them to meet the demands of individual data centers and significantly offset costs if those differences can be accounted for.

Likewise, today's storage model must consider the advent of new storage formats. As object storage enters the mainstream, there are many questions not answered by historical storage models. Does object storage apply to a single tier, or do we see Block, File and Object storage being used across multiple tiers and intermixed?

New use cases for data also call into question the usefulness of a hierarchical storage management or HSM. The increasing 'value' of data over the last decade means that data is held for longer periods of time – often forever. Object storage systems change the need to migrate data to newer forms of storage or even new formats. Administrators should be able to direct this between tiers or even within tiers.

New forms of data creation have also called for an updated storage paradigm. The storage pyramid was proposed when structured data dominated most data centers. The vast majority of data being stored today is unstructured data created by users, IoT, machines or sensors. A modern storage paradigm must address the need for some form of intelligence to move data in and out of various tiers and enable the data to be accessible to users and/or applications.

A modern storage approach should take into consideration all of these issues. By doing so, IT administrators, content creators and data curators will be able to create a better experience for external clients or internal data users to share research, monetize data, create a competitive edge in the market or satisfy whatever mandates their organization relies on to accomplish 'access to data.'

A New, Modern Storage Paradigm for Managing and Storing Data

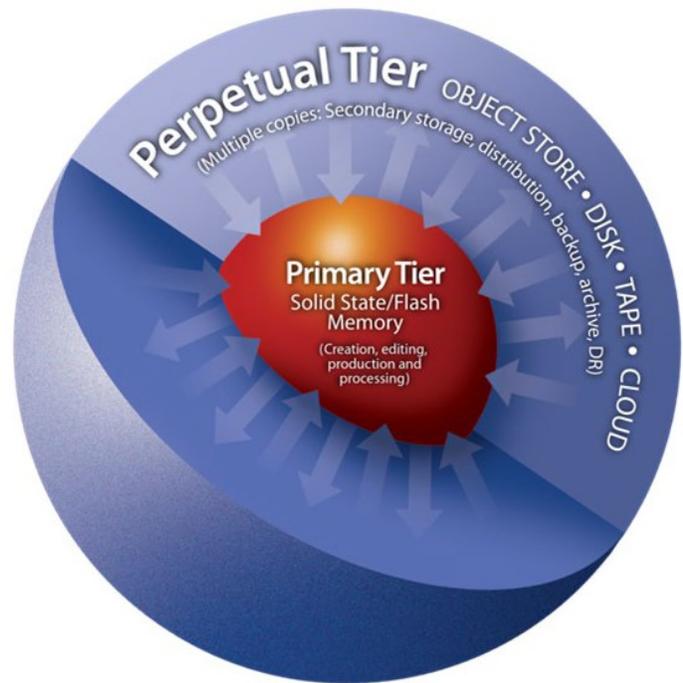
We live in a new world, and that is largely due to our ability to use data to enhance our world. It's interesting to look at data growth numbers for social media, but that's *not* at the heart of data growth for most organizations. Yet we are all finding new ways to use the data we collect or create to further organizational goals. In fact, recent studies show that 97.2% of organizations are investing in big data and artificial intelligence.¹

Data growth is not the problem, it's the opportunity. That is of course provided your organization is prepared to deal with this growth. Storage of information is no longer an afterthought. If it is, chances are that you're falling into a statistic of which you don't want to be a part of – as much as 80 percent of enterprise data may qualify as "inactive," yet it resides on primary storage.²

Traditional storage paradigms have focused too much on the storage "tier" without offering insight into the *data* that is being stored on any given tier. If inactive data is being stored on the wrong tier of storage, it could end up costing organizations millions of dollars a year as well as limiting their ability to

monetize and share this data properly. Ideally, the “knowledge” we need starts with knowledge of the data we already have. That’s the missing piece in traditional storage paradigms. Spectra solves that problem with the introduction of StorCycle.

StorCycle, Spectra’s groundbreaking storage lifecycle management software, enables a modern storage paradigm based on a two-tiered storage model. Rather than focusing exclusively on the storage medium, this model is based on the data or digital content that is actually being stored. We start by classifying data into two categories – “Active,” meaning it’s being edited, processed or changed in some way, and “Inactive” which quite simply refers to everything else. This results in a *Primary Tier* for the active data and a *Perpetual Tier* for inactive data.



Two-Tier Storage Model

The *Primary Tier* holds all active data and is most commonly composed of flash, NVMe and high-performance disk drives. By moving inactive data out of the Primary Tier and into the Perpetual Tier, organizations can significantly decrease the size of the Primary Tier. This allows administrators to better configure this tier using a combination of high-speed storage mediums in order to achieve the performance required for workflows associated with highly active data.

The *Perpetual Tier* is dedicated to inactive data and is designed to keep multiple copies of data on multiple storage mediums including NAS, object storage disk, cloud and tape. While the data is not considered “active” on the Perpetual Tier, there is quite a bit happening at this level. The Perpetual Tier is used for secondary storage; distribution; multiple copies (a responsive copy and DR copy); backup; archive; project archive; and traditional disaster recovery.

StorCycle allows organizations to configure the more economical Perpetual Tier to be as responsive as their workflows demand. Administrators can create “responsive copies” on low-cost NAS or another local format for data requiring quick but infrequent access. This copy can be used for data accessed directly by machine or application or may be used as a distribution copy. Simultaneously, StorCycle can create an identical copy on tape, for true air-gap protection from ransomware; a copy to cloud, for offsite DR; an additional copy to tape to avoid cloud restoration charges; or any combination thereof.

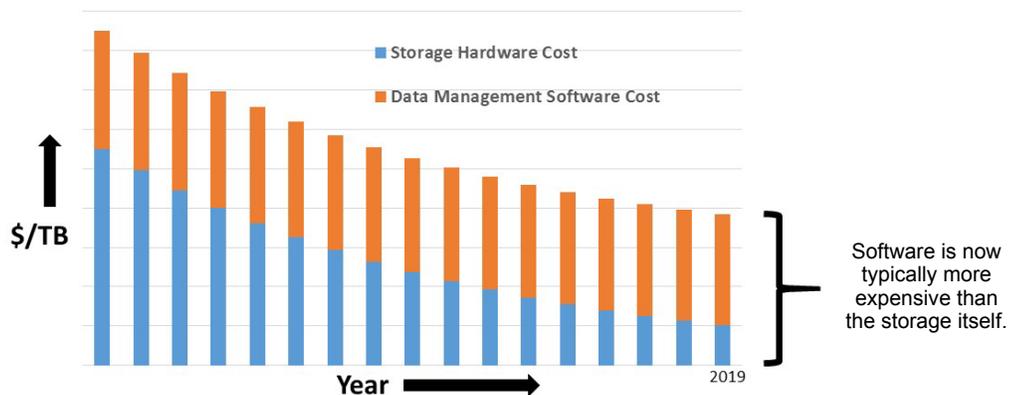
What is Modern Storage Lifecycle Management Software?

The basic concept of StorCycle is not new. Information Lifecycle Management (ILM) is a similar concept and has been accomplished with some success in very specific areas. There have been several approaches by both hardware manufacturers and software manufacturers to make staging of data a reality.

Providers of primary storage often offer excellent means for staging data between various targets within the Primary Tier of storage. While these solutions are very effective when working with a single hardware vendor, they often lose effectiveness when other vendor hardware is introduced. This encourages vendor lock-in and prevents a “best of breed” approach to implementing hardware. The real issue is that these solutions don’t offload primary storage to more affordable forms of storage for cold data, which includes low-cost NAS, tape and cloud. It’s not a design shortcoming, it simply wasn’t the design goal.

Hierarchical Storage Management (HSM) software is another approach, and one that is designed specifically to stage data on the appropriate level of storage hardware. There are a few effective, high-end HSM solutions on the market. These packages are often deployed in High Performance Computing (HPC) environments, which require scaling to hundreds of petabytes and often hundreds of tape drives. When required, such an HSM solution can be effective. It does, however, typically require dedicated support personnel, can take months to successfully deploy, and requires specialized file systems. Given that, prices for the overall implementation and supporting equipment can go into the millions for such HSM packages, these solutions are only justified in extreme data environments. Worldwide installations of such HSM packages probably number in the hundreds at most.

There are a few *new* software packages that fall into the category of “Storage Lifecycle Management Software.” Interestingly, they tend to fall victim to one of the two challenges found with HSM packages – either extremely complex or extremely expensive. There are even solutions that have capacity-based licensing that charges on the amount of data scanned regardless of whether the data is moved or not. As the cost per gigabyte of hardware storage has decreased significantly for NAS and Tape, the cost per gigabyte of the software required to move data to lower cost storage has increased exponentially.



Rising Cost Of Software Versus Falling Cost Of Hardware



If the total cost of implementing a data management solution negates any savings from moving data to a lower cost tier, or creates more management complications than it solves, it does not meet Spectra’s definition of *Modern Storage Management*.

Spectra’s design goal for StorCycle was to create a software package that could be used in virtually any situation, go from “box to production” in a matter of hours, decrease the cost of storage rather than be a cost multiplier, and assure that users retain access to data after it’s been migrated – even if it resides on tape or offsite cloud.

Furthermore, Modern Storage Management must offer complete data management. Automatically identifying inactive data and moving it to a lower cost storage is obviously a key component of StorCycle, but there are many times when administrators or users need to manually move data. StorCycle offers two migration modes. “Auto Migrate” enables policy-based movement of all files exceeding a given age, of a specified type, or matching other file identifiers. “Project Archive” allows logically grouped files associated with a given project or data collection to be manually migrated together as an archive. User-defined policies direct the data to NAS disk, object storage disk, tape, the cloud, a replicated site, or any combination of these.

And finally, Spectra’s design goal for successful Storage Lifecycle Management Software required the assurance of both data reliability and data protection.

In accomplishing these goals, Spectra has introduced a simple yet powerful software capable of increasing the value and effectiveness of existing storage infrastructure, allowing users to incorporate the best of new offerings in storage interfaces and technologies, and assuring the successful migration of all storage moving forward – even if at some point in the future your organization is not using StorCycle.

Dynamics of the Perpetual Storage Tier

The perpetual storage tier is a dynamic tier of storage. It can accommodate virtually any combination of storage platforms both old and new. Existing NAS, new NAS, object storage disk, tape, cloud and replication sites can all be part of this tier. This allows the Perpetual Tier to assist in data migration from older technology to new technology as newer storage platforms or interfaces are introduced.

An important aspect of the Perpetual Tier of storage is data integrity and safety. Because data stored in the Perpetual Tier is not part of the typical backup process used in the Primary Tier, keeping multiple copies of data in multiple locations is critical to reduce the likelihood that the data will be lost.

By creating multiple copies on various storage mediums, virtually any configuration can be achieved. A responsive copy can be kept on NAS for direct application/machine/user access. A DR copy can be made to tape, cloud or both. By using both cloud and tape, administrators could avoid expensive restoration charges from the cloud by restoring from a local tape copy when appropriate. Cloud copies can be used for sharing information with other groups and for use in cloud workflows such as transcoding and artificial intelligence (AI) tagging.



Most importantly, data in the Perpetual Tier is easily accessible by users in a transparent or semi-transparent manner with the use of Symbolic links, HTML links, and a web-based search. A properly implemented Perpetual Tier does not require additional headcount but rather reduces IT administration loads and manpower for management.

Summary:

If we inherently knew that all data resided on the appropriate level of storage, storage quotas could be a thing of the past, storage budgets could be more accurately forecasted and managed, and the benefits of any new storage medium could be easily implemented without an overhaul to the existing storage infrastructure. In this way, efficiencies could be maximized, data could be utilized to create further value and storage costs could actually go down instead of up... even when adding new technology.

Arguments over “end point” storage solutions – disk vs. tape, public cloud vs. private cloud, file vs. object – have consumed too much of the storage conversation and have deterred organizations from being able to focus on the real point behind storage – meeting the desired organizational goals that the information/data/content is used for.

Drawing from over 40 years of experience in the storage industry, Spectra is excited to introduce StorCycle storage lifecycle management software, and a new, two-tiered paradigm for storage that enables data to reside on the appropriate level of storage. This means more storage, lower costs, greater access, enhanced protection and fewer silos.

Footnotes:

1. <https://techjury.net/stats-about/big-data-statistics/>
2. <https://storageswiss.com/2019/04/12/inactive-data-how-much-do-you-really-have/>



About Spectra Logic Corporation

Spectra Logic develops data storage and data management solutions that solve the problem of long-term digital preservation for organizations dealing with exponential data growth. Dedicated solely to storage innovation for over 40 years, Spectra Logic's uncompromising product and customer focus is proven by the adoption of its solutions by leaders in multiple industries globally. Spectra enables affordable, multi-decade data storage and access by creating new methods of managing information in all forms of storage — including archive, backup, cold storage, private cloud and public cloud.

To learn more, visit www.SpectraLogic.com or contact our sales staff at sales@spectralogic.com