



Multi-Petabyte Storage Environments: Sharing and Storing Strategies for Data- Intensive Organizations

2022



Contents

- Abstract.....3**
- Introduction4**
- Challenges in a Multi-Petabyte Environment4**
- HPC Application Example5**
- Weather Data5**
- High Energy Physics.....6**
- Hybrid Cloud for Archive7**
- StorCycle: A Modern Approach to Storage Lifecycle Management7**
- Programmatic Application Integration8**
- StorCycle can be Part of a Disaster Recovery Strategy9**
- StorCycle and Project Archive Examples9**
- Capacity Offload Utilization with Disk for Near Instant Access for Rarely Used Data9**
 - Genomics Testing Laboratory Upgrading Current Primary Tier Storage to an All-Flash Array + BlackPearl Capacity Storage using StorCycle..... 10
- What Would an Exabyte in an Archive Tape Library Look Like? 12**
 - A Large University Supporting Multiple Research Projects 12
 - Production or Post-Production Studio in Media & Entertainment Supporting Advertising Agencies..... 14
 - Government-Sponsored Research Organization 15
- Summary17**

Copyright ©2022 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.



Abstract

In various organizations, there is often a desire to share research data files with other research groups, whether internal or external. Due to the large volume of data often generated, it is typical to archive this data to a longer term, lower cost storage medium, such as cloud or tape. A solution must be provided that can both archive the data and restore the archived data files to a more accessible location when they need to be shared. StorCycle® Storage Lifecycle Management Software can be used to curate large collections of data as well as archive large data sets, or even completed project data sets, and then easily restore them when further use or examination is needed. The software simplifies the migration of multi-petabyte data sets to longer term, lower-cost storage platforms, such as cloud or tape.

Organizational data sets that could benefit from such a solution include:

- Research & scientific data
 - Engineering data
 - Genomics data
 - HPC
 - University IT
 - Email archiving (PST files)
 - Medical research – project and data archive
 - Video surveillance content
 - Computer simulation
 - Architectural drawings
 - Federal, state and local government data
 - Petroleum exploration and seismic data
 - Drug research and development
 - Legal retention hold
 - IT archival of laptops and data from departing employees
 - VM image archival
 - Digital camera video offloading
 - Chip development and lifecycle management
 - IoT data capture/storage/archival
 - Database archive (version, transaction log files)
 - Artificial Intelligence (AI) data kept for training of algorithm
-



Introduction

Managing and storing petabytes of data is more difficult and more expensive than storing and managing terabytes of data. If the management of data is done incorrectly, it has the potential to cost the organization thousands of dollars in lost productivity; or even worse, lead to data loss that could cost the organization millions. Data management and storage challenges that arise with terabytes of data can sometimes be ignored or addressed through legacy techniques, but can become overwhelming with petabytes of data. The key to managing data repositories of this magnitude is to understand what the challenges are and address them with a forward-thinking management approach.

Challenges in a Multi-Petabyte Environment

One major factor that has impacted the overall data landscape in data-intensive environments is the advancement of machine-generated data. Machine-generated data is information automatically generated by a computer process, application, or other mechanism without the active involvement of a person. In short, there are a growing number of machines that generate data today, including microscopes, satellites, IoT, automobiles, sensors, medical devices, etc. New technological advancements have resulted in an exponential growth in data worldwide, equating to an enormous challenge as to how to store, and even more importantly, utilize all the data being collected.

Prior to the expansion of machine-generated data, storing and using data was much different. Data capacities were smaller and, as a result, data could be managed and stored manually. But, as the data sets and capacities have ballooned, new methods of managing data had to be developed. Previously all data was stored and accessed from an organization's high-performance storage hardware and parallel file systems. As the data repositories continued to grow in size, a tiered storage structure was needed to help lower the overall cost of storing data long term. That's because data often cannot be deleted (either for governance requirements or future evaluation). And older data is not often accessed, so finding a low-cost, long-term storage solution was needed.

Since the introduction of machine-generated data, a new challenge emerged as data was then being offloaded to a lower-cost tier, which sometimes made the data unusable and took many steps to rehydrate the data and make it usable again.

As archived data sets grow and organizations look to use or share that data, the challenge of how to access and distribute that data becomes relevant. Without the ability to use and share archived data, organizations are unable to leverage or monetize their data. To find and share data, many factors need to be considered, including security, permissions, accessibility, portability, usability. Data users and IT professionals alike have been looking for a self-service solution where groups/users could search for particular data, then restore and access it themselves. This type of workflow would eliminate many of the tasks and responsibilities associated with the movement and restoration of data, leaving data users to focus on daily activities, rather than where their data is stored and how they will access it.

Following is a collection of use cases illustrating how organizations today are curating, managing, using and storing massive amounts of data.

HPC Application Example

The world of High-Performance Computing (HPC) relies on supercomputers which are used for a wide range of computationally intensive tasks in various fields, including quantum mechanics, weather forecasting, climate research, oil and gas exploration, molecular modeling, and physical simulations (such as simulations of the early moments of the universe, airplane and spacecraft aerodynamics, automobile design, the detonation of nuclear weapons, and nuclear fusion). As one might guess, High Performance Computing requires and creates enormous data sets. First, computational test scenarios manipulate a digital environment or situation. Data is then run through these test scenarios and the output of the tests or experiments is collected. A single experiment can create hundreds of terabytes up to multiple petabytes. Likewise, a single experiment may take months or even years to run and cost millions of dollars. This creates a challenge for any experimental computing facility. Experiments are often repeated and reexamined. All inputs and outputs—that is, all the digital data related to a project—must be preserved. HPC sites rival data repositories of any vertical market and are probably second in size only to Internet operations.




Dual TFinity ExaScale Libraries that are deployed at Rutherford Appleton Laboratory.

Although data sets are regularly shared among engineers, scientists and developers who work in the HPC world, their workflows are more “manual” than those of general IT. Once the output of an experiment or modeling session is done, it does not change, and so it can be moved to a data repository requiring lower access and, hopefully, much greater density and energy efficiency for storage. Government national labs are a large part of the HPC community. They create and store vast amounts of data for both civilian and military use. In the 2020s, each of these government labs will grow in capacity to one or more Exabytes. The tape-based Exascale archive is already deployed at most HPC sites. The “TOP 500 list” of supercomputers shows the size and capacities of the world’s top sites.

Weather Data

Weather prediction uses sophisticated modeling running on supercomputers, which are continuously being improved. Weather data consists of collecting data from thousands to tens of thousands of sensors. This also fits into the category of the “Internet of Things.” These sensors may include ground-based sensors, satellite-based sensors, sea-water temperature sensors, aircraft collected data and occasionally weather balloons. This data is stored forever, along with the resultant forecasts. Weather scientists can tell if their models are working by evaluating the gaps and differences between actual outcome and what their model predicted for that particular time, whether currently or in the past. Therefore, weather data archives are enormous. Most large national weather forecasters will exceed one Exabyte before 2025. There are in excess of 25 major worldwide organizations which include the



National Weather Service (US), Meteo France, The European Center for Medium Range Forecasts, the Korean Meteorological Association, etc. As sensors become more prevalent and higher in resolution, they will collect ever more data. As supercomputers grow in computational power, we will see greater growth in required archive storage.

High Energy Physics

When the limits of human understanding are being pushed by the desire to answer some of life's most complex questions, the amount of time and complexity of the research demands the most secure and scalable storage available. CERN uses some of the world's largest and most complex scientific instruments to study the fundamental particles of matter, quite literally discovering the "God Particle", the missing cornerstone in our knowledge of nature. They do this with the LHC (Large Hadron Collider), the world's largest and most powerful particle accelerator. The LHC is 27 km (16.8 miles) of superconducting magnets in the shape of a ring, that sends high-energy particle beams close to the speed of light until they collide, with the goal of figuring out how the universe was formed. This research means the CERN Data Centre is producing data at an astronomical one petabyte of data per day, pushing their CERN Advance Storage System (CASTOR) to a massive 330PB of data stored on tape. According to CERN, this is equivalent to 2000 years of 24/7 HD video recording.

Until 2022 the LHC is shut down while it gets a big performance upgrade. This upgrade will keep it shut down for two years, just long enough to upgrade their data archival software and tape system to handle the much higher data volume. Once the LHC resumes its next run, data creation is expected to double during the years 2021 to 2023. This means that they will be storing an additional 600PB or more after run three in 2023. This will put CERN very close to if not over 1 Exabyte of data on tape at the end of 2023. The LHC will then shut down again, upgrades to many of the sensors, magnets, and testing devices in the collider will be made, and when it resumes from 2026 to 2029, we can expect an increase of five times the current level of data stored. This means that during run four there will be over 1.5 Exabytes of additional data that needs to be archived to tape.

The new CERN Tape Archive (CTA) software will replace CASTOR and will store the existing 330 petabytes of data, as well as ALL new data created. It is being designed to be able to handle these massive amounts of data. The data that is produced at CERN is extremely valuable and must be preserved for future generations of physicists making tape the ideal archive storage technology to use. It is also shared worldwide. CERN has transferred 830PB of data and 1.1 billion files to other HEPIX research organizations all over the world. This allows other physicists to conduct research and it also means that the data is archived geographically with multiple copies so that it can safely be kept forever.

CTA has the ability to store an Exabyte in native capacity and CERN is rapidly approaching the need for an Exabyte of storage. They are counting on the industry to continue to innovate to store multiple Exabytes in the coming decade to be preserved for future generations and with the roadmap of LTO tape it looks to be a partnership that will stand the test of time.



Hybrid Cloud for Archive

One of the largest growth areas in our industry is that of hybrid cloud. “Hybrid Cloud,” is a term that’s been thrown about ever since the introduction of public cloud. In this discussion, hybrid cloud refers to the ability to keep data on premises as well as in the cloud. Some data may reside exclusively on premises, such as highly confidential information or amounts of information too large to store in the cloud indefinitely given monthly charges. Other data may reside exclusively in the cloud, such as transitive data not needed once a given calculation or compute service is completed. It may be desired to have copies of the same data stored both in the cloud and on premises simultaneously.

The ultimate definition of hybrid cloud allows for the management of data to occur from a “single pane of glass” or control mechanism which provides a universal view of the data – regardless of where that data physically resides.

Both public cloud and on-premises storage provides certain advantages and challenges. Hybrid cloud offers the possibility of getting the best of both worlds if it’s done right. Let’s start by comparing cloud storage and on-premises storage. Cost and performance will not bode as well for cloud storage of 20PB or even upwards of an Exabyte, as it does for on-premises storage. That isn’t necessarily a strike against cloud storage as much as it is a reason for a hybrid approach of combining on-premises storage and public cloud. Before discussing Hybrid Cloud, it’s important to understand the opportunities/challenges with “cloud-only” or “on-premises only” models.

StorCycle: A Modern Approach to Storage Lifecycle Management


Spectra’s StorCycle storage lifecycle management software is an easy, affordable and efficient solution built specifically to identify inactive data that consumes expensive primary storage capacity. It then migrates identified data to a more affordable protected tier of storage called the Perpetual Storage Tier, while leaving the data fully accessible to users. The Perpetual Storage Tier holds inactive data, which is tracked, protected and is available at a future time if needed, or even deleted when the data has reached the end of its lifecycle. This tier is also used for data distribution, backup, archive, disaster recovery and more.



IT professionals intuitively know they are not storing their data efficiently because they have never been given the correct tools to do so... until now. Spectra Logic’s StorCycle software is the answer to deploying a modern storage lifecycle management workflow.

The Perpetual Tier of storage is not limited to older data. The perpetual storage tier should also serve as an archive tier for large data sets and projects which may be moved immediately after creation or collection.

Universities, Government Agencies, Genomics, Research Labs – these organizations create enormous amounts of data on an ongoing basis that needs to be managed outside the confines of high-performance, Tier-1 storage. Users have no way to “group” various data sets or track them once moved from primary storage, so they sit on high-performance, high-cost storage indefinitely.



Likewise, IT departments have data sets which could be moved to lower cost storage tiers immediately after completion, such as year-end financials, corporate videos, marketing collateral, and email archives, just to name a few.

StorCycle can be used to migrate any files/folders/directories, or objects/buckets, associated with a given project. The migration job will create a manifest file which shows migrated data, where it was migrated from and where it was migrated to. Simply click on the archive job to display the manifest. The archive can also be tagged with searchable information related to the project. This allows for simple search and restore even years or decades into the future, by either graphical user interface or programmatically through API's

Protection of data is not lost to StorCycle, and when paired with the BlackPearl® platform, superior ransomware resiliency can be achieved through Spectra's Attack Hardened™ feature set. Using StorCycle's encryption and BlackPearl's triggered snapshots, data that is migrated with StorCycle can be protected from a ransomware attack, and achieve resiliency to bounce back in the event of an attack.

Programmatic Application Integration

When a customized workflow and specific use cases need to be addressed, organizations can utilize existing tools to connect their applications to the StorCycle management software. This in turn fully integrates StorCycle into existing workflows and allows for seamless integration and implementation.

[StorCycle includes an API](#) that programmatically interacts with the application to provide an easy way to integrate a storage lifecycle management strategy into any organization's workflow.

The exposed commands do the following:

- Authenticate
- Create and manage storage locations (sources and targets)
- Create and manage scan, migrate/archive, and restore operations

These commands are used for actions such as bulk creation of storage locations, scripting of migrations that first require actions of other systems, and user restores through a separate web portal. The StorCycle API is built using a standard RESTful interface and is detailed in a full set of documentation for each command. Moreover, code samples in Python that can be found on [StorCycle's GitHub platform](#). In fact, with [Open API Generator](#), a wrapper for the API can be created to interact in Python or another language of a customer's choosing.

As more organizations experience the pain of managing and storing massive amounts of data, new and modern approaches to managing data is being implemented. By integrating StorCycle into existing workflows and applications, organizations can now create a limitless, storage destination that is not just accessible, but also affordable. With easy integration and an exposed API, StorCycle is perfectly designed for all organizations facing data growth and retention requirements.



StorCycle Can be Part of a Disaster Recovery Strategy

Many organizations make multiple copies of data as an effective part of their strategy to ensure that data is available and can be restored in the event of a disaster. Disasters happen as a result of many different reasons, including extreme weather, natural disasters, human error, and cyberattack. StorCycle provides the option to make multiple copies of data on multiple storage mediums, including automatic tape libraries, offsite tape storage, disk storage, and even public cloud storage. When utilizing the cloud as a storage target, StorCycle can prioritize restorations from the cloud as a last resort as the charges and egress fees from the cloud can add up quickly. StorCycle can be part of a cost-effective disaster recovery strategy to ensure data is always recoverable.

StorCycle and Project Archive Examples

As we continue to look at use case examples, it is important to keep in mind that the Perpetual Tier of storage is not limited to older data. The above use cases focus on moving data into the Perpetual Tier based on age and date of last access. The Perpetual Tier should also serve as an archive tier for data sets that may be moved immediately after creation or collection. By allowing users to archive recent or even older “project-based” files or directories, historically critical data can be maintained and protected indefinitely.



Any of the above use cases could also choose to use StorCycle’s unique Project Archive feature. The Project Archive functionality is beneficial to organizations creating and saving enormous amounts of data on an ongoing basis such as universities, government agencies, genomics testing, research labs, media and entertainment, weather research and healthcare.

The following use cases are as varied as the organizations deploying project archives, but there are commonalities which allow any organization to gain insight into better ways of managing and archiving large data sets.

Capacity Offload Utilization with Disk for Near Instant Access for Rarely Used Data

In a modern organization, the need and desire for a flash-based storage system is in high demand and many organizations are struggling with the balance between fast performing primary storage and the challenge of storing all their data in an affordable manner. All-flash solutions increase the speed at which large amounts of data are captured and processed, creating better value for businesses. By delivering greater storage performance within a similar footprint, all-flash are quickly becoming the gold standard for data-intensive organizations.

The increased performance that comes with a flash-based system also comes with a higher price tag, meaning that the higher the performance and the higher the capacity, the greater the cost. Finding creative and effective ways to balance the amount of storage needed for active data with the storage needed for inactive data can be an effective way of finding a solution that meets the performance requirements, and fits into the required budget.

By leveraging Spectra’s StorCycle to identify inactive data and effectively migrate that data to a lower cost, accessible disk-based tier of storage, organizations can finally purchase a solution that increases performance, and provides a Perpetual Tier of storage designed to hold long-term data leaving it accessible to users and applications.

Genomics Testing Laboratory Upgrading Current Primary Tier Storage to an All-Flash Array + BlackPearl Capacity Storage using StorCycle

One data-intensive organization we’re working with was considering a performance upgrade of their primary storage tier. They had originally targeted a combination of flash storage and clustered NetApp HDD infrastructure. They have over 5PB of primary storage, most of it used for storing lab results—but they expect a 50% or greater annual growth rate. They originally decided to forego the upgrade to an all-flash solution due to cost.

They estimate 80 percent of their data currently stored on the Primary Tier is inactive and can, therefore, be moved to the a Perpetual Tier, offered by Spectra Logic. The majority of that migrated data (4.5 PB) will reside on low-cost BlackPearl NAS, roughly 12¢/GB or \$120/TB. Cost for the Perpetual Tier NAS storage will be roughly \$550,000.

Thus this customer has now created a 500TB flash-only Primary Tier platform for a fraction of the initial quotes, allowing the overall solution to fit into the budget they currently have. The total savings from this simple application of a Perpetual Tier using StorCycle is in the millions of dollars (which includes all the of cost of StorCycle and BlackPearl system and support for 5 years). But more importantly, they are able to significantly increase the performance of their data center and work with state-of-art technology, and users have seamless access to migrated data on the NAS tier of perpetual storage.

The expandability of the BlackPearl solution makes this a solution that can expand over time to truly be a forward-looking solution that not only solves today’s challenges but also the future challenges of managing and storing vast amounts of data.



Spectra BlackPearl NAS. Beginning at year 1 with 4.6 PB of capacity and growing over 5 years to over 20 PB in a single 42U rack

BlackPearl in conjunction with StorCycle includes a unique Spectra Attack Hardened™ feature, which is enabled via StorCycle. Any data written to the BlackPearl may be “locked” via inline snapshots from accidental deletion or wholesale destruction from hacking or ransomware.

StorCycle allowed this organization to rethink the most effective way to store their data. By removing inactive and “low transaction” data from the Primary Tier, the organization can now more affordably upgrade to a high-performance Primary Tier that uses SSD, NVMe flash or other cutting-edge high-performance technologies. StorCycle will also be used to make a low-cost, DR copy to a public service provider for offsite DR, such as Amazon Glacier or Wasabi (final selection has not been made at the time this was written).

A detailed cost breakdown of the hardware, software and support needed to store and expand over a five- year period is shown below, beginning with 4.5PB and growing to over 20PB of uncompressed raw storage in a single 42U rack. Note the price includes hardware, software, installation, and hardware and software maintenance for the five-year period. All disk drives are enterprise-grade SAS, with Self-Encryption using the BlackPearl KMS, to provide encryption at rest. Although the BlackPearl system includes very effective built-in data compression, no data compression has been assumed in the capacities and costs below:

	Year 1	Year 2	Year 3	Year 4	Year 5
Solution Cost	455,785.30	262,197.63	235,977.87	212,380.08	191,142.07
Support Cost expiring year 6 (co-term)	99,427.03	112,200.00	84,150.00	56,100.00	28,050.00
Total Cost per Year	\$ 555,212.33	\$ 374,397.63	\$ 320,127.87	\$ 268,480.08	\$ 219,192.07
TCO for Solution	\$ 555,212.33	\$ 929,609.96	\$ 1,249,737.83	\$ 1,518,217.91	\$ 1,737,409.98
Capacity Raw	4.76 PB	9.04 PB	13.32 PB	17.6 PB	21.88 PB
Capacity Raw in GB	4,760,000 GB	9,040,000 GB	13,320,000 GB	17,600,000 GB	21,880,000 GB
Total cost per GB	\$0.12 Per GB	\$0.10 Per GB	\$0.09 Per GB	\$0.09 Per GB	\$0.08 Per GB
Total cost per TB	\$116.64 Per TB	\$102.83 Per TB	\$93.82 Per TB	\$86.26 Per TB	\$79.41 Per TB
Capacity Formatted	4.16 PB	8 PB	11.52 PB	15.36 PB	19.2 PB
Capacity Formatted in GB	4,160,000 GB	8,000,000 GB	11,520,000 GB	15,360,000 GB	19,200,000 GB
Total cost per GB Formatted	\$0.13 Per GB	\$0.12 Per GB	\$0.11 Per GB	\$0.10 Per GB	\$0.09 Per GB
Total cost per TB Formatted	\$133.46 Per TB	\$116.20 Per TB	\$108.48 Per TB	\$98.84 Per TB	\$90.49 Per TB

Formatted = Double parity with 18 drive stipes and global spare drives

Note, in the above model, each “stripe” of disk drives has dual parity protection, and an ample pool of global spare disk drives have been included to automatically replace any failing disk drives over the system service life.

BlackPearl offers a unique spin-down Object Storage mode for disk storage. This feature substantially lowers the carbon-footprint of the solution, while extending the life of the disk drives for several more years. That option was not selected by this customer, as they wished to have instant access to data by any of their NFS and CIFS applications.

What Would an Exabyte in an Archive Tape Library Look Like?

Bringing the discussion closer to home, an Exabyte is 1 quintillion bytes or a 1 with 18 zeros behind it. 1,000,000,000,000,000,000 bytes.

In the “good old days” storage products were rated in terms of compressed capacity, but storage software either pre-compresses or encrypts most content or data. The Spectra® TFinity® Exascale Tape Library is the world’s largest capacity storage system. The TFinity Tape Library is capable of holding over 1EB of data based on compressed capacity using LTO-8 or over 2EB of data based on compressed capacity using TS Tape Technology. This was an industry first. With the announcement of LTO-9 in September 2020, the TFinity now holds an Exabyte of **uncompressed** data –making it the only tape library in the world capable of holding an Exabyte of uncompressed information – or as Spectra Logic calls it – ‘Exascale’ storage.

A 45-frame TFinity Tape library, configured with 48 LTO-9 tape drives, holds 55,990 LTO-9 tape cartridges. At 18 terabytes of native data per tape, this equals 1.008 EB of uncompressed data in a single tape library. Physically, this is a large library. It measures 109’ long and 3.6’ deep which means that it will use 391 square feet of real estate. Although the footprint is large, the data density is the best available on the market; a TFinity of this configuration offers 2,558 TB per square foot which is extremely compact for use in a modern data center.

Just as no data center “starts” with an Exabyte of data, the TFinity can be configured as a 3-frame unit and ‘transcaled’ up from that point. The examples below will showcase a variety of configurations, sizes and costs. Spectra Logic offers single frame as well as rack-mounted libraries that can all be ‘transcaled’ into future and larger Spectra libraries ensuring investment protection and fewer technology refreshes along the way.

 TFinity 3-frame library footprint




TFinity 45-frame library footprint

Max capacity in tapes/drives

A Large University Supporting Multiple Research Projects

One of the data centers Spectra Logic works with is part of a large university that supports the research efforts of over 50 different groups within the university. They offer various service level agreements (SLAs) based on the performance of the storage. Each research group is billed for the amount of storage they use based on the SLA they select.



The university has standardized on three storage performance levels: an SSD-based tier; high-speed, disk-based tier; and a NAS-based tier targeted for archive of projects. There are multiple challenges for both the university and the individual researchers that the university would like to overcome.

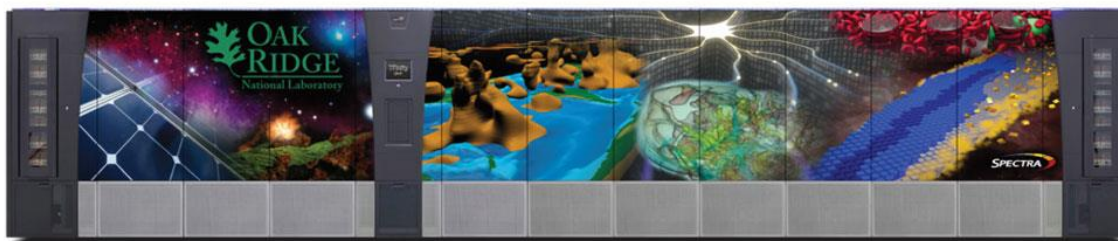
While NAS is the lowest cost repository in the current storage model, archiving large amounts of fixed content becomes extremely expensive over time. Researchers have asked for a lower cost solution for archiving. The university would like to introduce tape storage as part of the Perpetual Tier for archiving, but they have no way to introduce “rule-based” file movement across the storage infrastructure to a Perpetual Tier of tape.

The university actively encourages the researchers to move data off the high-speed Primary Tier with a bill-back system that offsets their costs. However, when the university runs out of high-speed storage, they don’t always have the funds to acquire more before the offsets come in, which can hamper research efforts.

The researchers would like to move their data to a low-cost storage tier, but they have a challenge to accomplish this. The data can be human-generated, application-generated, or machine-generated, and it has been accumulating for years. The researchers have access to the data, but they have no way to identify what is actively used, what needs to be archived, and what data is orphaned. And if the data is manually moved, how can it be accessed after the researcher leaves the university? Grants for research often require storage of the data for periods longer than the researchers who work for the lab or university involved.

StorCycle is a perfect solution for both parties. StorCycle offers a seamless view across all of the storage it manages. Both the university and the researchers can have access across all data on the Perpetual Tier.

Researchers can use the scanning capabilities of StorCycle to identify and target inactive data sets for archive. StorCycle’s Project Archive will assure this situation doesn’t reoccur. Project Archive allows users to identify any files or directories associated with a project, then archive them as a group. This can be done immediately after a large project is completed.



13-Frame TFinity ExaScale Tape Library with the ability to hold over 250PB of uncompressed data.

Archived data sets can be tagged, and grouped, with additional information to identify anything of importance to the project, be it grants associated with the project, researchers involved, project names, etc. This metadata can easily be searched at any point in the future. Likewise, StorCycle produces a manifest for each project archive which can be accessed as a digital file. The manifest shows exactly what was moved, when it was moved, where it originated and where it landed. The information can be digitally displayed by clicking on the finished project archive and stored with other files in the project. It does not require a query into the database and can be worked into existing workflows.

In a 250PB archive stored over five years, a 5-frame, 24-drive TFinity and a single BlackPearl X would be recommended. The unit will grow by 50PB per year with media purchases of 12.5PB quarterly. The TFinity will be expanded to a total of 13 frames over the course of the five-year period. The costs purchase, installation, and onsite maintenance for these two examples can be seen in the below table.

Capacity in PB	Year 1 Cost	Year 2 Incremental Cost	Year 3 Incremental Cost	Year 4 Incremental Cost	Year 5 Incremental Cost	Total Cost	\$/GB	TB/SqFt
250	\$ 978,058	\$ 403,844	\$ 363,153	\$ 335,425	\$ 325,289	\$ 2,405,769	\$ 0.01	\$ 2,197

The university will now be able to deploy an affordable tape storage specifically for archiving. The cost to the university for automated tape will be around 2.5 cents/GB (U.S. dollars) and additional copies on tape will be around 1.2 cents/GB. Researchers will be able to identify data and move it to a reliable secondary storage tier, and that migrated data will be protected and easily accessible into perpetuity. All of this relieves pressure on the most expensive Primary Tier of storage, which the university is responsible for purchasing and maintaining.

Production or Post-Production Studio in Media & Entertainment Supporting Advertising Agencies

Movies, television series, reality TV and commercials – They seem relatively straightforward when we watch them, but what goes on behind the scenes is tremendously complex. Many actions go into creating a simple commercial – film ingest, QC, logging, audio synch, creating a proxy or mezzanine, editing, rendering, adding special effects, dubbing – it’s all done on high-speed, high-cost disk storage.

In this manner, a single commercial is typically 50 to 100TB. That’s the output; all of the input (think 1200:1 ratio) is kept as well. On average 10 to 15 final versions of the commercial will be presented for review and selection by the ad agency. Here’s where the numbers really become staggering – A single post-production house will typically support multiple advertising agencies and a single

advertising agency will support multiple clients. A single client can easily run four to eight campaigns a year, each requiring multiple commercials, which easily creates 500TB to 1PB per client per year.

Unlike large movie or television studios, many post-production studios don’t use Media Asset or Production Asset Management software for archiving content. Those solutions are often cost prohibitive.



Spectra Stack plus BlackPearl platform delivering over 5PB of uncompressed capacity.

This particular post-production studio will use StorCycle’s Project Archive feature to solve this problem. All digital assets will be stored in directories associated with a given client or commercial. As soon as the commercial is completed, all associated directories and files will be archived off of their Primary Tier of storage to their Perpetual Tier of storage via StorCycle Project Archive. Additional metadata will be added for more granular searches in the future, such :commercial type/seasonal/geography/length/cost/end user client/ad agency client, etc.

Due to the volume of data created in each case, the content will all be archived to tape via StorCycle and the fully integrated Spectra BlackPearl platform. This allows them to manage the content on their newly created Perpetual Tier with the advantages of using object storage even when written to tape. Future migration is seamless. The life of a tape library is long -- 10 to 12 years vs. three to five years for disk. In this scenario, additional data copies can be made for vault storage or for sending back to the ad agency, which will cost under 2¢/GB.

Capacity in PB	Year 1 Cost	Year 2 Incremental Cost	Year 3 Incremental Cost	Year 4 Incremental Cost	Year 5 Incremental Cost	Total Cost	\$/GB	TB/SqFt
5	\$ 77,244	\$ 31,894	\$ 28,681	\$ 26,491	\$ 25,690	\$ 190,000	\$ 0.38	\$ 526

Instead of spending roughly 50¢/GB for high performance disk, the content can be archived to tape for roughly 3.8¢/GB (including the cost for BlackPearl). Given those numbers, this example takes the cost of storage from \$500,000 per PB to \$38,000 per PB. The solution solves three issues: They can now afford to keep all data associated with an ad agency or client indefinitely; they can seamlessly move that data to a Perpetual Tier of storage and easily bring it back when needed; and StorCycle’s Project Archive feature assures all content will be available and searchable as a single project.

Government-Sponsored Research Organization

Virtually any governmental agency, in any country across the globe, deals with large amounts of data.

This particular agency creates, collects and distributes scientific information used by both U.S. and international governmental offices; non-governmental agencies; other researchers; and even individual citizens.



Data gathered can be generated by application output, field sensors, machines, cameras, individuals, or other data creation methods. After data is gathered, it is further analyzed, categorized or simply stored for possible future use or reference. No data gathered is considered “disposable”. As technology, science and even the earth itself evolve, new exploration often draws on historical data – weather patterns, ocean currents, agricultural yields, and mineral exploration, to name a few.

After extensive search for a storage lifecycle management software application, this particular agency was drawn to StorCycle specifically for its ability to archive data based on the data's association with a given project. How else could multiple forms of data from multiple types of data generators be collected and archived if not through some form of "project identification"?


Even after this, they still had a challenge, because much of the data they collect is machine-generated, such as data from sensors that may detect physical phenomena such as light and sound and turn it into a data stream, and calculations from algorithms predicting the risk of earth movement based on other seismic data sets. The output may not be analyzed immediately, or the researcher may deem it unnecessary for a current project, but they wish to keep it for future reference. Most of the machine-generated data they receive initially requires high-speed disk as a landing zone. The researchers had no way to move it to lower cost storage and bring it back when needed. Therefore, it stayed on the Primary Tier of storage – at great expense – even if it was never accessed.

The ideal solution combines StorCycle, Spectra's BlackPearl platform and a Spectra TFinity Tape Library. By setting up an Archive Directory with StorCycle, even machine-generated data can be immediately archived as it comes in – using high-speed disk storage for ingest, but immediately moving it to a lower performance storage tier. It's then deleted from the primary storage. Failover can be achieved when clustering StorCycle instances through a virtual machine on a single server, or across multiple servers.



Capacity in PB	Year 1 Cost	Year 2 Incremental Cost	Year 3 Incremental Cost	Year 4 Incremental Cost	Year 5 Incremental Cost	Total Cost	\$/GB	TB/SqFt
1000	\$ 2,679,290	\$ 1,536,882	\$ 1,362,234	\$ 1,264,155	\$ 1,211,729	\$ 8,054,290	\$ 0.008	2,558.32
500	\$ 1,763,583	\$ 761,945	\$ 712,044	\$ 669,422	\$ 637,267	\$ 4,544,261	\$ 0.009	2,392.03

As the output storage target for StorCycle, a flash-based BlackPearl platform can not only ingest the archived data at great speeds, but it can also direct it to the tape library at great speeds – easily streaming more than 100 LTO tape drives simultaneously. It's as if the high-speed Primary Tier of storage is extended indefinitely at pennies per gigabyte. Due to the infinite retention period and vast amount of data being collected and stored, the capacity can grow quickly. For this example, capacities



of 500PB and 1EB will be used to show the high-end scale of the Spectra solutions and the capacities needed for such a workflow.



A fully expanded TFinity library can reach an amazing 45 frames with the ability to store over an exabyte in a single tape library.

Most importantly, the Archive Directory can be associated with the project which created the data so that the data captured can be seamlessly tracked as part of the larger project. Individual researchers can designate the storage layer for data or content, but the ability to query the StorCycle database means other individuals, not originally associated with the research, can find it throughout its lifecycle.

Other data sets within the agency are posted to the cloud for distribution or sharing. StorCycle's ability to migrate to multiple targets allows the agency to direct data to the same long-term, tape storage tier – which is not externally accessible – as well as direct a copy to the cloud for distribution or sharing. The cloud copy can expire in a year or two while the DR copy will remain on tape into perpetuity.

All of the above examples for Project Archive require a Perpetual Tier that is simply accessed, easily searchable by project name or other tagged metadata, and built for “forever” retention.

StorCycle is the only data management software package that offers Project Archive, making it an ideal solution for preserving large data sets generated by data-intensive organizations that produce and manage project data as part of their mission.

These are but a few of the ways in which StorCycle is making a simple, two-tiered storage paradigm a reality for organizations working with a few terabytes on up to those working with hundreds of petabytes and beyond.

Summary

Today's storage landscape is one of growth and expansion. As data continues to be generated at rates that have never been seen before, organizations need to find new ways to manage and retain the data that is so valuable to their day-to-day operations. StorCycle was designed to help organizations identify, migrate, preserve, and access data regardless of where it is stored or when it was created. With StorCycle, storage budgets can be more accurately forecasted and managed, and the benefits of any new storage medium can be easily implemented without an overhaul to the existing workflow (that might drastically change user experiences and create IT headaches). In this way, the IT infrastructure is



streamlined and optimized while data is protected to achieve long-lasting value even as data storage costs are reduced.

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 our 42 years of experience in the storage industry, Spectra’s 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.



About Spectra Logic Corporation

[Spectra Logic](#) develops a full range of Attack Hardened™ data management and data storage solutions for a multi-cloud world. Dedicated solely to data storage innovation for more than 40 years, Spectra Logic helps organizations modernize their IT infrastructures and protect and preserve their data with a broad portfolio of solutions that enable them to manage, migrate, store and preserve business data long-term, along with features to make them ransomware resilient, whether on-premises, in a single cloud, across multiple clouds, or in all locations at once.

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