Comparing Tape and Cloud Storage for Long-Term Data Preservation

Matt Starr
Chief Technical Officer
Spectra Logic
Boulder, United States
matts@spectralogic.com

Matt Ninesling
Director, Hardware Engineering
Spectra Logic
Boulder, United States
mattn@spectralogic.com

Eric Polet
Program Manager, Emerging Markets
Spectra Logic
Boulder, United States
ericp@spectralogic.com

***Abstract*—With the retention life of data increasing, many organizations are trying to determine the best long-term storage strategy for the future. To do this, organizations must analyze the reliability, security and speed of data storage solutions, including both cloud and tape storage. This paper considers how a hybrid archive solution, consisting of both cloud and tape, may be deployed in the event of a cloud mandate, and how data growth and future costs can impact a storage solution selection.**

Keywords—data retention, storage strategy, long-term storage, data reliability, data security, speed of access, cloud, tape storage, hybrid archive, cloud mandate

I. Introduction

A quick web search could convince someone that cloud technology has taken over the computing and storage market. With Google, Amazon and Microsoft at the forefront of enterprise cloud storage, and hundreds of smaller, niche companies trying to get a piece of the cloud computing or storage market, it would seem that disk and tape storage has been abandoned and declared "dead."

Tape storage has existed for almost 50 years and was declared "dead" about 30 years ago when disk storage was the leading technology. As an archiving and backup tool, tape is still prominent in the market with a majority percentage of Fortune 100 companies still using a tape library. Tape technology continues to advance, with faster speeds, higher capacities, greater efficiencies and quicker retrieval times, all while outlasting the competition.

II. Perceptions of Leading Edge Technology

Prior to exploring the side-by-side comparisons of tape and cloud, the perceptions of both technologies must be addressed. While reliability, cost and other factors are important to consider in the adoption of any technology, the current industry opinion of that technology's inherent strengths and weaknesses is equally important to examine.

A. Cloud Technology Today

Today, cloud technology is seen as a leading-edge technology. The cloud boom started when Amazon released Amazon Web Services (AWS) in 2002 [1]. AWS included a suite of cloud-based infrastructure services. In 2006, when Amazon launched Elastic Compute Cloud (EC2) as a commercial web service that allows businesses to rent computers and create their own applications, Amazon directed its marketing to target technology startups and small businesses. Amazon touted that their services could get businesses up and running without spending tens of thousands of dollars to acquire on-premise servers, storage and networking equipment. By utilizing cloud computing and storage, "cloud-native" startup businesses started to emerge in the market [2]. Google followed suit and similarly targeted small business and startups. With Amazon and Google onboard, many enterprises started looking at Infrastructure as a Service (IaaS) models to compete with their smaller cloud-native competitors. When Microsoft finally came on the scene with the Platform as a Service (PaaS) and then extended their reach into IaaS, there was a huge spike in enterprises moving over to cloud-based platforms. With three large technology companies pushing cloud services, it's no surprise that cloud computing and digital storage developed the perception of being nimble, strategic, and appealing to stakeholders – all of which has led to numerous "Cloud First" mandates. A Cloud First mandate, which was first implemented by the U.S. Government in 2012, lays out the framework for how to move applications and storage to the cloud [3]. The government mandate has been adopted throughout some of the private sector as a plan to move enterprises away from physical computing and archiving to help cut costs and increase performance.

B. Tape Technology Perceptions

On the other hand, tape has been perceived as an older technology and does not generate the same excitement as cloud technology. If a business is only looking at the marketing of AWS, Google Cloud, and Microsoft Azure, it may think tape is costly and outdated, no matter how far from the truth that may be. In the past decade, tape technology has made strong and continuous technological advances and this trend shows no sign of slowing down. The technology improvements have lowered costs, increased capacity and improved the reliability and security of an already reliable and secure storage medium. According to a Tape Storage Council memo, in the last 10 years, LTO tapes have increased capacity by 1,400 percent, performance by 200 percent, and reliability by 9,900 percent. In 2017 alone, tape saw four major improvements, including a new generation of tape media. LTO-8 media doubled the native capacity of LTO-7, and improved throughput by 20 percent. Improvements have also been made to the tape libraries themselves, which are now centerpieces of many modern data centers [4]. Even with a "Cloud First" mandate, many organizations are finding that a combination of cloud and tape for archiving is extremely effective at ensuring reliability, security and cost savings. In addition, many cloud service providers themselves have discovered that tape storage provides an economical and reliable way to safeguard massive amounts of inactive data for long-term preservation [5].

III. Reliability

Reliability, stability and life expectancy are always top considerations for any organization when choosing a storage platform. Since the length of time data needs to be stored continues to expand, the need to store and retrieve data at any given point has become crucial to all businesses.

Tape is a reliable medium with a shelf life of 30 years in the right storage conditions [6]. The biggest concern for most businesses with tape technology is the hardware. Many organizations are worried that tape libraries, and their respective components like drives and robot arms, will wear out and eventually fail. What may not be realized is that, with annual maintenance, tape libraries have an almost infinite lifespan, ensuring that all of the data stored in that tape library can be retrieved at any point. LTO drives can read and write at least one generation back, so the data stored on a tape cartridge can continuously be migrated to newer generations [7]. While the physical media may only last 30 years, the data can be stored forever.

A lot of marketing for cloud storage guarantees a reliability and uptime between 99.999999999 percent and 99.99999999999 percent, between nine and eleven 9's [8]. While those are impressive numbers, they require replicating data to multiple cloud storage sites and placing multiple copies in each site. Storing data at only one cloud storage site is not a sound business decision with lots of recorded individual failures of data. Another part of the cloud guarantee is a Service Level Agreement (SLA). This agreement outlines what is considered to be an acceptable service provided by the cloud host. Both Amazon and Microsoft will provide a 10 percent monetary credit for most accounts if the reported cloud uptime is less than 99.9 percent [9]. This means that data might not be accessible for seven hours every month with no credit issued to the business for downtime. If those seven hours happen to fall in the middle of a large migration project, it could have a far-reaching impact on the timeliness and success of that migration.

IV. Security

In late 2010, Microsoft experienced what was called the first-ever cloud data and security breach [10]. This breach wasn't the work of malicious intent; it was due to an unspecified "configuration issue" that allowed private data to be downloaded by non-authorized users. Fast forward to August 2013, and three billion users on Yahoo were left vulnerable after hackers gained access to first and last names, email addresses, dates of birth, and questions and answers to security questions. Yahoo and its parent company Oath released a full press statement in October 2017 that outlined the severity of the security breach. The hacked information had been stored in Yahoo's vast network of servers, part of which was stored in cloud servers [11].

Unfortunately, data and security breaches have become all too common. According to the Identity Theft Resource Center (IRTC), three-quarters of U.S. businesses, including retailers, restaurants, and even hospitals and doctors' offices, have had cloud-related data breaches of some kind, and half of all those data breaches happened in 2017 alone. In fact, IRTC has found that many of these data breaches are due to the fact that most networks use a basic set up, and do not encrypt their data with any security software [12]. Cloud storage providers reassure customers that all data is securely encrypted, but with the constantly changing state of security software utilized by the cloud, updates can have unknown security vulnerabilities. It's these vulnerabilities that allow hackers to hit multiple accounts, or even multiple services, at the same time. Amazon and Microsoft are consistently updating software to keep their customer's data safe from vulnerabilities, and businesses are forced to increase their IT security budgets to hire and retain employees who are knowledgeable about cybersecurity to strengthen internal systems from external threats.

A. Ransomware

A particularly brutal external threat that has been gaining momentum is ransomware. Ransomware attacks are cyberattacks where malware is downloaded onto a computer or network. From there, some or all of the files on the network are encrypted using a private key. Organizations are told to pay an exorbitant amount of money to receive the private key to regain access to their data. While the malware itself is easily removed from the system without the encryption key, organizations have no way to get their data back. The first big ransomware attack, also known as a cryptolocker attack, happened in 2013 and took more than six months to completely resolve. Since 2013, the number of ransomware attacks has continued to grow, and this type of attack is now one of the top cybersecurity threats around [13]. Ransomware started out by using a network of infected devices known as a botnet, and the hackers behind it have now turned their sights to a much larger network of devices to infect: The Cloud.

Due to this rising threat, cloud providers like Microsoft and Amazon now offer an additional service to protect data in the cloud. For an extra charge, Microsoft will store copies in the cloud that are accessible for 30 days, and if a ransomware attack hits, data can be restored instantly. While this is a huge improvement over ransomware protection for cloud services, this service does not make cloud storage a bulletproof solution that will instantly save organizations from ransomware [14].

The only true way to protect from ransomware is by having a copy of data that is "air-gapped". With an Air Gap, a copy of the data is offline and not connected to any network at all, so is not vulnerable to a network attack. Because tape cartridges do not need power or connection to a network, tape is the only storage medium that keeps data completely safe from ransomware. When tape is the backup storage target and ransomware hits the network, the data that is being held hostage can be completely deleted to remove the malware, and data can be fully restored using the copy stored on tape. Tape provides two levels of protection from ransomware: (1) media sitting in a tape library can only be accessed when a robot inserts the tape into a drive; (2) media removed or ejected from a tape library can only be accessed with human intervention. The best part is that both levels of protection are available with no extra charge. And if the data on the media is encrypted, that data cannot be read, providing yet another layer of protection in the event of an attack [15].

V. Speed

Speed can be a major factor in choosing storage. Backing up and archiving massive amounts of data can take hours for any project or set of files to be transferred to the storage medium. Retrieving data is a unique process in and of itself. In the event that a project, movie, or a group of files needs to be recalled right away, or a worst-case scenario happens and disaster strikes, a long retrieval time may be unacceptable. Unfortunately, there is no storage medium that can recall data

instantaneously, so finding the fastest and most reliable retrieval method is critical to all data center environments.

A. Storing Data

Previously, tape had a reputation of being slow. This may have been true 40 years ago, but like all technologies, tape has advanced exponentially. Today tape can transfer data at an extremely fast rate, and like any on-premise device, speed is scalable. Organizations have full control over upgrades to the tape library and can add more tape drives and expansion frames to enhance performance. In fact, for full write workflows, tape even outpaces disk with speeds ranging from 400 MB/s uncompressed up to 900 MB/s per drive with compression [16]. A modest 24-drive installation can write over 75TB of data in an hour, 1.8PB in less than a day, and 13PB in a week with compressible data at a 2.5:1 compression ratio.

Cloud storage has the perception of being fast and agile, but to varying degrees, there is limited control and flexibility for how data can be uploaded. Uploading an initial dataset to the cloud depends on how much bandwidth an organization has, and how much bandwidth is available to the cloud provider [17]. Cloud providers, or backup software providers that write to the cloud, are able to quickly update data that is already stored in the cloud. Depending on the need, cloud bandwidth can be scalable to accommodate larger file transfers into the cloud, but the speed of the archive job is typically based on the organization's internet bandwidth. On average, if a business was transferring the same 60TB to the cloud instead of tape, with a 1Gbps bandwidth connection, this transfer would take a little over six days to archive. That is assuming the 1Gbps connection was completely dedicated to this transfer and not being used for anything else, and that the 1Gbps connection was completely stable the full six days.

Most of the time, a business is only able to use a quarter of their bandwidth for uploads to the cloud. With a dedicated 250Mbps connection, 60TB would take 25 days – almost a whole month to place that project into the cloud. A 1PB upload into the cloud on the same 250Mbps connection could take 418 days, or one year and two months, to complete.

If a business is only dealing with a small amount of data that needs to be archived, cloud is a feasible option, but in most scenarios, tape is going to be the fastest and most reliable option.

B. Retrieving Data

While storing data is fairly straightforward, retrieving data can be a bit more complicated. Typically, the speed of data recovery depends on how much data needs to be recovered. In the event of disaster recovery, tape is the fastest option. The process for retrieving data consists of a tape that is loaded back into a drive, and all the data is read from the tape and transferred back over the network to the designated source. If a set of specific files are being recalled though, tape could potentially take longer. The files that are needed could be written on multiple tapes, and since tape is read linearly, the file that is needed could be at the front of the tape or it could be at the end of the tape. If the tape is stored offsite, this will add a few hours to the retrieval time. There are modern applications that create objects and provide a type of on-premise cloud that helps the library identify where a specific file is located, which speeds restore times drastically.

The retrieval of archived cloud data has historically been more complicated than what has been portrayed. While there have been many improvements made, all cloud providers have a different protocol for retrieving data from the cloud. There are a lot of developers creating GUIs for cloud data retrieval, but these GUIs do not always work, and because they are third-party applications, cloud service providers cannot help troubleshoot issues. Even when the retrieval request works, it can take three to five hours to get a single file. If data needs to be retrieved for disaster recovery, it could take too long to get the data back.

The table below lays out Amazon's lowest cost tiers of storage and the SLA, both time and cost, to access data. Once a restore request has been made, it will take the time listed to download and access the data.

TABLE I. AMAZON'S LOWEST COST TIERS OF STORAGE AND SLAS

AWS Tier	Data Retrieval Costs			
	Level of retrieval	*Time to access*	*Cost per GB*	*Cost per TB*
Glacier	Expedited	1-5 minutes	$0.0300	$30.00
Glacier	Standard	3-5 hours	$0.0100	$10.00
Glacier	Bulk	5-12 hours	$0.0025	$2.50
Deep Archive	Standard	12 hours	$0.0200	$20.00
Deep Archive	Bulk	48 hours	$0.0025	$2.50

VI. DATA GROWTH

As technology grows, the amount of data it produces also grows, and this presents unique challenges for IT professionals. Data storage technology is improving year over year, and tape technology is continuously increasing its density every few years. The improvements are exciting, but they can lead organizations down a confusing path of how to navigate upgrades and how to cycle outdated storage technology and media.

In the next two years, many organizations are expected to quadruple the amount of data that needs to be stored. This expected increase caused tape technology experts to double down on efforts to maximize the density of the media. In late 2017, LTO-8 tape media was released with an astonishing 12TB per tape cartridge, doubling the capacity of LTO-7 media which was 6TB, native [18]. LTO-9 will again substantially increase the per-cartridge capacity. In the midst of these advancements, there is a confusion about when to buy new media and drives, and when to transfer data from old media onto new media. Another potential complication is when and how to upgrade the physical hardware. From tape drives to the whole library itself, upgrades can feel costly and complicated.

While it may seem counterintuitive, it's recommended to actually wait to upgrade tape media. Tapes have a shelf life of 30 years, and modern libraries are designed to have near-perfect conditions for tape media. Tape drives are also able to read and write back at least one generation, e.g., LTO-8 tape drives can read LTO-7 media. While this also means that organizations would need to keep older tape drives in the library, most tape libraries support multiple generations of tape drives. Many companies simply wait for the cost of a new generation of tape media to come down to a reasonable price

per gigabyte and then purchase new drives and media. By utilizing this method, a slow upgrade can occur over time instead of shutting down production on the tape library to write data to new media every three to five years.

Many organizations think that by utilizing cloud storage, they won't have to think about the ever-changing state of storage media because the cloud service provider will take care of everything for the business. This is only partially true. By utilizing the cloud, there is no need to worry about an on-premise storage device. Organizations don't have to build in a plan to transition to new generations as they come out. The drawback is that all cloud companies utilize physical storage devices like disk and/or tape to store data. This affects businesses because, after a certain tier of storage or a certain percentage of physical storage space is hit, most cloud providers have a new pricing structure that is more expensive than the previous one.

Moreover, most cloud storage costs are based on total capacity of files being stored, total number of files being stored, retrieval and access costs, transfer and transfer acceleration costs, and replication costs. All of these different categories also have costs that vary by region, so cloud costs for the eastern U.S. could differ from cloud costs for the western U.S. For example, Amazon S3 has six different pricing tiers for "standard" storage. Then, based on the storage method chosen, there are another five billing categories for data access, including whether or not an account moves from infrequent storage access into standard access. An account can transfer from infrequent access to standard access for two reasons: (1) the organization opts to transfer to standard access for faster access speeds; or (2) the business requests too many files in a month and is automatically transferred into standard access. The second scenario is a perfect example of how a hybrid cloud and tape environment can save organizations money. Data can be stored in the cloud at the deepest level, which is what Amazon Glacier does, and the same data can also be stored on tape. If a large portion of data needs to be retrieved, it can be retrieved from the tape copy at no additional cost, without the risk of being moved to a more expensive tier of cloud archive.

A. Redundancy

When calculating data growth, redundancy must always be factored into the equation. Multiple copies are not only important for disaster recovery, but also are critical for any type of data storage failure.

Storage failure can happen on any type of media and can affect a single file up to multiple terabytes worth of data. With tape storage, redundancy is easy. A set of files is written to two separate tapes. Ideally, one of these tapes is kept on premises and one is stored offsite in a secure location. If corruption occurs on a set of files, the secondary tape can be brought back onsite and used to restore the data.

Cloud storage can be as easy as putting two copies in the cloud, and hoping that they are stored in separate locations with both copies remaining uncorrupted. However, the uncertainty has led many users to store a single copy in the cloud, and store a copy on another storage media such as tape or disk. Putting multiple copies in the cloud can also impact pricing tiers, with twice the amount of data in the cloud doubling the price. Many experts who suggest that redundancy is necessary, and it almost always is, recommend that an on-premise storage option is best paired with cloud storage. If disaster strikes, impacting the data center, then copies can be retrieved from cloud storage. Once the data center is up and running, the data can be duplicated from the cloud and put back onto tape. Without a cloud mandate, organizations can still achieve best practices in data storage while saving money by leveraging tape: (a) making two copies to tape, with one ejected and taken offsite; or (b) making two copies to tape at two different customer-managed locations, by replicating the data to the second site.

VII. Expected Future Costs

The total cost of ownership (TCO) is perhaps the most important upfront factor when making the final decision for a storage option. Tape's TCO has been well studied through the years and is fairly predictable. In a 10-year span, the initial tape library purchases with media and drives produce the most upfront expenses, and after the initial purchase, the next concern is how to upgrade and migrate media and drives. Luckily, migrations can occur slowly, over time, and critical data can be migrated to new generations of tape media technology quickly. Many products can help to automate the tape migration process in the background so that users are unaware that the migration is occurring.

Cloud's TCO has not changed much over the last several years. The anticipated 50 percent drop in cloud prices (due to increased competition and more efficient processing and storing techniques used by large cloud providers) did happen, but with a twist: the cost to access the data is now twice that of the previous tier. Amazon Web Services recently announced a Deep Archive tier that allows organizations to store data for a staggering $0.001 per GB per month or $1.01 per TB per month compared to Amazon's Glacier tier that costs $0.0046 per GB per month. On the surface, this seems like a huge reduction in cost for using the cloud, but upon closer inspection, a standard retrieval from Deep Archive will cost twice as much and take three times longer to have data accessible than Glacier. So cloud storage costs have come down (in some cases quite dramatically) and this reduction will continue, but customers must be vigilant as to where cloud providers might be seeking to get their money back in other ways that are usually harder to plan for and budget for.

VIII. Parting Thoughts

While we recommend tape for many storage needs, cloud storage can provide some excellent usage models for different scenarios. A common storage idiom is: "Three copies, on two different media, and one offsite". With this in mind, a combination of tape and cloud storage could fulfill this basic storage principle. For these situations, it is recommended that data that is touched be stored in the on-premise storage and data that may never be touched be stored in the cloud – to be used as supplemental disaster recovery. For data that will only be retrieved in the worst possible circumstances, a deep layer of storage, Amazon Glacier, might be recommended. To control both on-premise storage and cloud storage, a single versatile device that acts like a hub should be considered.

Cloud costs can vary drastically, and it's important to keep in mind how much the cloud market structure is changing. It is very probable that a new cloud provider could appear, or an existing cloud provider could alter its structure, so a day may come when an organization will need or want to switch cloud providers. This is where an on-premise copy is imperative. Organizations that have on-premise storage provide themselves with the option to abandon cloud data, rather than

pay exorbitant egress fees. Furthermore, modern-day object storage solutions can enable a change in cloud vendor by uploading a local copy with the touch of a button, without having to pay any egress fees.

When trying to find the right storage strategy, the answer really comes down to the ideal storage solution that optimizes the strengths of each medium while reducing risks of storage failure. Tape and cloud together, behind an object storage device, is the most powerful combination. With the object storage system handling the workflow, tape being used as a primary back up, and cloud being used as a disaster recovery option, data can be protected elegantly, seamlessly and cost effectively from any unfortunate circumstances. If this hybrid combination is unable to be deployed, today's tape storage is the recommended platform for an economical long-term data storage and archival approach.

REFERENCES

[1] S. Carey, "The history of AWS: A timeline of defining moments from 2002 to now," Computer World, March 2019, Web. https://www.computerworld.com/article/3412382/the-history-of-aws--a-timeline-of-defining-moments-from-2002-to-now.html#slide14

[2] S. Hallet, "The evolution of cloud computing," Hanover Recruitment, May 2018, Web. https://hanrec.com/2018/05/31/the-evolution-of-cloud-computing/

[3] V. Kundra, "Federal cloud computing strategy," Obama White House Archives, pp. 2, February 2011, Web. https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/egov_docs/federal-cloud-computing-strategy.pdf

[4] "Tape takes aim on unprecedented data growth," Tape Storage Council Report, November 2019, Web. https://tapestorage.org/wp-content/uploads/TSC-2019-November.pdf

[5] E. Polet, "The marriage of tape and cloud: achieving cost-effective data preservation with the hybrid cloud approach," IT Pro Portal, December 2019, Web. https://www.itproportal.com/features/the-marriage-of-tape-and-cloud-achieving-cost-effective-data-preservation-with-the-hybrid-cloud-approach/

[6] The Commission on Preservation and Access, "Appendix 2: Estimation of magnetic tape life expectancies (LEs)," CLIR, National Media Library, Web. http://www.clir.org/pubs/reports/pub54/estimation_of_LEs/

[7] Ultrium LTO, "FAQ," Web. https://www.lto.org/about-the-lto-program/faq/.

[8] D. Friend, "What does 11 nines of durability really mean?" Wasabi, May 2019, Web. https://wasabi.com/blog/11-nines-durability/

[9] R. Al-Sayyed, W. Hijawi, A. Bashiti, I. Aljarah, N. Obeid and O. Adwan, "An investigation of Microsoft Azure and Amazon Web Services from Users' Perspectives," International Journal of Emerging Technologies in Learning (iJET), May 2019. 14. 217. 10.3991/ijet.v14i10.9902.

[10] K. Thomas, "Microsoft cloud data breach heralds things to come," PCWorld, December 2010, Web. https://www.pcworld.com/article/214775/microsoft_cloud_data_breach_sign_of_future.html

[11] N. Perlroth, "All 3 billion yahoo accounts were affected by 2013 attack," The New York Times, October 2017, Web. https://www.nytimes.com/2017/10/03/technology/yahoo-hack-3-billion-users.html

[12] "2017 annual data breach year-end review," Identity Theft Resource Center, Web. https://www.idtheftcenter.org/2017-data-breaches/

[13] J. De Groot, "A history of ransomware attacks: The biggest and worst ransomware attacks of all time," DataInsider, October 2019, Web. https://digitalguardian.com/blog/history-ransomware-attacks-biggest-and-worst-ransomware-attacks-all-time

[14] T. Warren, "Microsoft adds ransomware protection and file restore to OneDrive cloud storage," The Verge, April 2018, Web. https://www.theverge.com/2018/4/5/17201660/microsoft-onedrive-files-restore-feature-ransomware-protection

[15] F. Moore, "Data protection converges with cybersecurity: The tape air gap addresses cybercrime," August 2018, Web. https://edge.spectralogic.com/index.cfm?&fuseaction=home.displayFile&DocID=5044

[16] "Drive performance," IBM, Web. https://www.ibm.com/support/knowledgecenter/STQRQ9/com.ibm.storage.ts4500.doc/ts4500_ipg_3584_ircc4.html

[17] L. Ferreira da Silva, F. Brito e Abreu, Fernando, "Moving to the cloud: Estimating the internet connection bandwidth," January 2011.

[18] R. Gadomski, "How tape storage is changing the game for data centers," DataCenter Knowledge, July 2018, Web. https://www.datacenterknowledge.com/industry-perspectives/how-tape-storage-changing-game-data-centers