Overview: Replacing Tape with Cloud in Backup Workflows

A Comparison and Three Reference Architectures

June 2017

Notices

Contents

Introduction 1

Tape Characteristics 1

- Backup 2
- Archive 2
- Cost Factor Analysis 3
- Portability 4

Reliability 4

- Performance 5

Cloud Characteristics 5

Cost Factor Analysis 5

- Durability and Availability 6
- Complementary Services 7
- Performance 8

Reference Architectures 9

- Backup to Cloud via Backup Software Integration 9
- Backup to Cloud through a Hybrid Cloud Virtual Tape Library 10
- Backup to Cloud via an On-Premises NFS Interface 12

Conclusion 14

Contributors 14

Further Reading 14

Abstract

IT organizations have looked for decades for options to replace tape-based backup and archival processes. Most recently, cloud services have become an option for reducing the cost and complexity of tape and tape automation. Surveys show that IT professionals now list backup as one of the top use cases for cloud platforms. [1, 2] This paper compares and contrasts the cost and performance factors of traditional magnetic tape with those of cloud storage for backup workflows. It also provides three typical reference architectures for simple, minimally disruptive cloud backup projects, and highlights why AWS is a compelling platform for backup storage.

Introduction

Tape media, automation systems, and backup software have been an important part of traditional data protection strategies. However, many organizations have grown weary of the cost and effort associated with buying, handling, and storing tape media, as well as the budget impacts of replacing tape libraries and associated software licensing. The backup function remains critical for protecting applications and their data, but there must be an alternative to the complexity of, and talent spent on, tape management and processes. AWS offers a variety of options to improve or replace traditional tape backups, on-premises tape libraries, and offsite physical archiving services with solutions that leverage the durability, massive scalability, pay-as-you-go consumption, short retrieval times, and automated management capabilities of the cloud.

Tape Characteristics

For over 60 years, the IT industry has relied on magnetic tape to store data, move it offsite, and preserve it for long periods of time. This longevity is due to the increases in density that have consistently driven down the per-gigabyte cost of tape storage, the low cost of power and cooling for tape systems compared to hard disk drive media ("disk"), as well as the familiar constructs around tape management that have been in use for years.

In a typical tape-based backup architecture, primary data used for production applications is stored on either networked storage or local disk storage. On a set schedule, the data is sent via application servers or storage array snapshots to a backup server that collects and writes the data to magnetic tape. These tapes are stored in large, on-site tape libraries that are manually managed by tape operator staff or automatically managed by some form of robotics. Some organizations replicate their mission-critical application data over a wide area network (WAN) to other disk arrays, to smaller tape libraries offsite, or hire a third-party to manually move the tapes into an off-site storage facility.

Although backups written to magnetic tape offer easy storage and simple replication, they have inherent drawbacks, such as proprietary formats and sequential read behavior that lengthen backup windows and recovery processes, making it difficult for IT teams to meet Recovery Point Objectives (RPO) and Recovery Time Objectives (RTO). Additionally, it is impractical to test magnetic tape backups frequently, and tapes can suffer from cartridge and media failures during backup and recovery, particularly if they are not kept in pristine, dust-free environments. Indeed, tape media units and revenue have been in steady decline, with vendors such as TDK, Imation, and Tandberg Data consolidating or exiting the market due to rising pressures from disk and cloud-based alternatives.

These flaws typically render tape-based backup solutions inadequate in a recovery event. If you currently use a tape-based backup solution, moving to the cloud not only removes the complexities and risks of physical media handling but may also reduce the total costs involved with backing up to physical tape storage media. It has the additional benefit of supporting application recovery in the cloud.

Backup

This paper defines "backups" as copies of data and application state stored in a secondary location to protect against system downtime and malicious or accidental data loss. Backups typically have a useful lifespan measured in days or weeks and are often managed on-site, traditionally in a tape library. Most backup software products maintain a backup catalog with metadata about the location and time of the backup jobs. The data management role such products play is usually very important to IT administrators. Disk-to-disk-to-tape (D2D2T) solutions may offer shorter backup windows and faster restores in exchange for increased disk capacity consumption. Data reduction technologies such as deduplication and compression may help address this increased consumption of more costly disk.

Backup copies may be retained for years—or even decades—to comply with regulations or internal mandates. The durability, reliability, and associated system costs of physical tapes is often a major concern for IT organizations when tape backups are kept as a part of a long-term data retention policy.

Archive

Archival applications may also leverage tape as a common storage medium, but unlike backup they often retain the only copy of a digital asset, stored for extended periods for regulatory compliance, preservation, audit, re-use, analysis, or commercial license value. Archive processes may be driven by industry or regulatory compliance requirements or by organizational best practices. This often results in an approach that essentially "locks away" data in service of the lowest $/GB, and the data is only available for use when specifically requested. RTO is often measured in days, and third-party vendors may offer specialized services and solutions for specific types of data and applications (such as email) that cater to compliance or records management professionals.

This paper is focused primarily on backups, although data management applications may be able to leverage cloud technologies to seamlessly convert backups into archives and simplify the archive process. It should also be noted that while third-party solutions (such as Index Engines) exist to assist you in converting and transferring existing tape archives into the cloud, they are not discussed in this paper.

Cost Factor Analysis

The most compelling advantage of tape is the nominal low media cost, which typically is the lowest cost and highest density media for data storage. The second major cost advantage of tape is low power consumption cost since power is only consumed when tapes are written, read, or transported by robotics systems. But considering only these factors in a Total Cost of Ownership (TCO) analysis provides an incomplete picture. Additional cost factors include:

- Hardware costs for tape libraries, robotics, and expansion arrays, which can cost millions of dollars at large scale, and are amortized/depreciated over 5- to 10-year life spans.
- Administrative and media costs associated with bulk media migration, when data is moved between generations of tape media, to protect against long-term "bit-rot." Media refreshes and migrations are commonly performed at 6- to 8-year intervals.
- Costs for tape drives and network fabrics to provide appropriate throughput to meet backup window requirements. New drives are acquired in support of new tape formats or to boost performance.
- Support contract costs, typically represented as 20% of capital spend on hardware purchases.
- Labor costs of tape backup administrators, including the time and effort to retrieve tapes from offline sources.

- Contracts for offsite vaulting for services (such as Iron Mountain).
- Hardware and administration costs, which are often multiplied due to the common practice of retaining multiple on-site and off-site data copies to protect against media or mechanical failure.
- Total costs may also increase due to factors of data overhead and media underutilization. Compression may mitigate this factor, though this is an unpredictable benefit that may not be realized at all since most modern backup application workflows deduplicate, compress, and/or encrypt data. Media file (e.g., video, audio, or image) archival workflows often already include some compression in the file format or codec, cancelling out native compression.

In summary, the nominal low-cost of tape media is often offset in detailed TCO analyses by mitigating OpEx and non-media CapEx factors.

Portability

Tape cartridges are designed to be physically simple to handle, transport, and store, and adhere to a very narrow range of industry standards. High-capacity hard disk drive (HDD) storage media is not designed to handle the shock and vibration of travel, while enterprise solid-state drives (SSD) are more resilient than HDDs. SSD prices per GB are not currently on par with enterprise HDDs, and IDC does not forecast parity in the foreseeable future.[3]

Reliability

While tape-based storage is engineered for long-term preservation, it presents a variety of reliability risks in the real world. As a result, you can create multiple tape copies (both on-site and off-site), which further erodes the nominal cost benefits of tape. Cartridges, library robotics, and tape drives are mechanical in nature. Failures can result in data loss or extended availability loss. Data integrity on tape can unpredictably erode over time, forcing the need for periodic bulk media migration programs. The act of mounting and reading a tape is destructive in nature, as oxide coatings erode and strain is applied to tape cartridges, as are specific behaviors like “shoeshine” when retrieving random data from a linear device. Tape systems designed for frequent access have unique design elements and are only available at a significant price premium, which erodes the presumed cost advantage.

Performance

Well-provisioned tape systems can deliver high read and write throughput. However, performance can be unpredictable, especially when under load. Retrieving large datasets can range from hours to days, depending on the volume of data, tape drive specifications, the number of tape drives multiplexed together, and the speed of robotic tape retrieval. Although warehouse-based offsite archival provides the safety of potential geographic separation from the data center, recovering data tapes may take days, unless you pay for "rush" service to accelerate retrieval.

Disk-based systems are commonly deployed to hold recently protected data to mitigate the high-retrieval latencies associated with tape. Both file-system and virtual tape library (VTL) solutions are popular because they supplement existing tape management software and hardware with the speed of magnetic disk. These designs provide performance improvements for data stored on the disks, but not for data that has passed through the staging disks and migrated completely onto physical tapes.

Cloud Characteristics

Cloud providers leverage aggregated customer demand to invest in infrastructures that deliver economies of scale in cost and reliability, including the development of highly redundant and efficient data centers, Availability Zones (in the case of AWS), and regional redundancies. These economies of scale also enable the development of complementary value-added services that can leverage stored data, such as search, analytics, and cloud-based disaster recovery for applications. As scale increases, cost savings can be passed on to customers, who consume more, driving up revenues and then investment in a virtuous cycle. Long-term storage platforms in the cloud offer low-cost storage tiers that tie in with other cloud storage offerings and (as of this writing) cost as little as $0.004 per GB/month, with data retrieval times ranging from milliseconds to hours.

Cost Factor Analysis

Public cloud storage delivers a substantially different consumption and cost model, wherein all costs are "pay as you go" variable monthly charges based on usage. When considering the broader costs of on-premises tape solutions against the costs and agility benefits of cloud-based storage, note that cloud storage today is able to deliver a TCO that is on par with tape.

- Unlike tape-based backup and archival, cloud storage has no up-front capital investment. Consume any amount of capacity and scale as fast as data can be stored.
- No costs associated with media and system obsolescence and data migration.
- No administrative costs associated with scaling, repairing, and configuring tape systems.
- No administrative costs associated with retrieving offline media for restores, or populating new media into libraries.
- No additional support contract costs.
- No additional durability and availability cost (it's all just a part of the service).
- Benefit from general pricing reductions over time as the cloud provider passes along efficiency savings (the "commodity curve").

Durability and Availability

Cloud platforms are designed to offer very high reliability against component and facilities failures, with storage solutions such as Amazon Simple Storage Service (S3) and Amazon Glacier delivering a durability service-level agreement (SLA) as high as 99.999999999% (11 nines). The likelihood of data loss due to infrastructure failure is extremely low, and generally lower that what is easily achievable with on-premises storage solutions. The underlying storage architecture and media choice is invisible to you, in contrast with traditional solutions where tape formats and backup software applications may change and require you to maintain backward compatibility with aging formats and peripherals.

In addition, cloud platforms are intended to be highly available. Industry-leading cloud storage services are based on multi-site regions designed to withstand failures of media, systems, networks, and even complete data centers. Data may also be replicated between geographic regions of the world, providing further reliability and global availability benefits. Assembling a comparable solution using on-premises tape or disk systems with the durability and availability levels of the cloud is likely cost-prohibitive, particularly in today's commonly CapEx-constrained IT budgets.

Complementary Services

Cloud storage platforms can envelop data with value-added services, including management tools that help you automatically sort and move data to achieve price-performance goals. These management tools can be integrated with, and driven by, security, audit trails, access controls, and behaviors within other cloud-based applications. Write once, read many (WORM) policies may be applied to assist with compliance requirements in industries such as healthcare or financial services.

Here are some examples of cloud-based data lifecycle management tools applied during a tape replacement project:

- New backups with low RTO requirements can be written to a high-performance object storage tier for 30 days, moved to a lower cost tier after 90 days, and then moved to a lowest-cost archival tier, where they can be locked for compliance and deleted after 7 years.
- Newly imported backups within a specific timestamp range can be moved directly into the archive tier and optionally locked.
- Newly imported backups can be tagged with metadata (such as job IDs, dates, physical locations) so that policies may be applied by tag. This offers an additional range of indexing and management within your growing cloud tape archive.

In some cases, backup datasets contain valuable historical data that line-of-business users are interested in analyzing for compliance or for actionable business value. However, in tape-based systems, accessing data for bulk analysis can be difficult or impossible. In cloud storage, however, there are adjacent compute services available (on a pay-per-use basis) to readily extract value from data if/when the need emerges. With this in mind, if you are moving data from tape to cloud, you may seek to avoid proprietary backup formats in favor of preserving native data and file formats. Thus, moving to the cloud can "turn dumb bits into smart bits", potentially extracting big-data value for internal stakeholders.

Performance

Cloud storage is available in a variety of price-performance options, from milliseconds to hours, which can be tuned to meet the needs of your workflows. For performance-sensitive workflows, such as bulk data retrieval for a restore operation, RTO may be affected more by WAN connectivity (Internet connection, for example) than by native storage performance. Therefore, when shifting from tape to cloud you should carefully consider the requirements for network connectivity as well as on-premises caching of more recent data.

The lowest-cost cloud storage tiers are often the lowest performing for reads, and designed for long-term, seldom-accessed archival use cases. Archival platforms, such as Amazon Glacier object storage, may offer a standard retrieval time that delivers data in hours with potential "expedited" retrieval that delivers data in minutes. Sometimes these are based on "best-effort" SLAs, so pre-paid, guaranteed performance options may also be offered. Retrieving a petabyte in a day, for use cases such as media or geospatial applications, may be done through "bulk" retrieval options or even off-line solutions.

One challenge when replacing tape is moving backup data into the cloud. Vendors may offer a range of on-premises network transport methods (such as AWS Direct Connect or AWS Storage Gateway), physical transport methods (such as AWS Snowball and AWS Snowmobile devices), or integrations with existing backup software vendors (so that backups may be written directly to the cloud).

You should evaluate your project needs for scale, compliance, RPO, and RTO when planning your tape-to-cloud projects considering these questions:

1. Do I need my historical backups in the cloud, or can I simply start backing up new data to the cloud and gradually reduce on-premises tape dependency as data reaches end-of-life?
2. If I have a large initial dataset to move to the cloud, what is the right approach to getting it there?
3. When I move data from tape to cloud, are there processing options that should be performed first such as indexing, transcoding, or repackaging?
4. Is my move to cloud natively supported by my backup vendor? Are new feature licenses or major version updates required first? Does my backup

vendor support direct restore of data into cloud virtual machines (VMs)? How do my plans for my backup software vendor (and associated backup catalog) impact my decision?

5. Can my backup software use cloud-native interfaces (such as the Amazon S3 protocol) to write data directly to the cloud? Do my backup windows support this?
6. Should I have my backup software use a traditional storage protocol such as a network file system (NFS)?
7. Is there value for my organization to maintain existing tape-like workflows with a cloud-backed VTL?
8. Do I have an appropriate amount of tape drive capacity to make large scale changes without disrupting ongoing backup activity?
9. Do I want to access my data natively in the cloud? What are the implications for file format and cataloging systems?
10. Can I simplify widely distributed (e.g., ROBO) backup workflows by leveraging the global reach of the cloud?
11. Which cloud features, such as analytics, lifecycle management, or compliance capabilities, are must-have requirements?
12. Given the scale and complexity of the job, should I explore the use of external integrators or specialists to accomplish the migration?

It may be a formidable job to read and re-write an entire tape archive into the cloud. You may want to explore using a third-party partner with expertise in tape handling, software, equipment, and processes to source, read, standardize, and migrate those tapes.

Reference Architectures

Consider the following example architectures for replacing tape backup with backup to the AWS Cloud.

Backup to Cloud via Backup Software Integration

Amazon S3 and Amazon Glacier are natively API-based and available over the Internet. This allows backup software vendors to directly integrate their applications with AWS storage solutions as shown in the following figure.

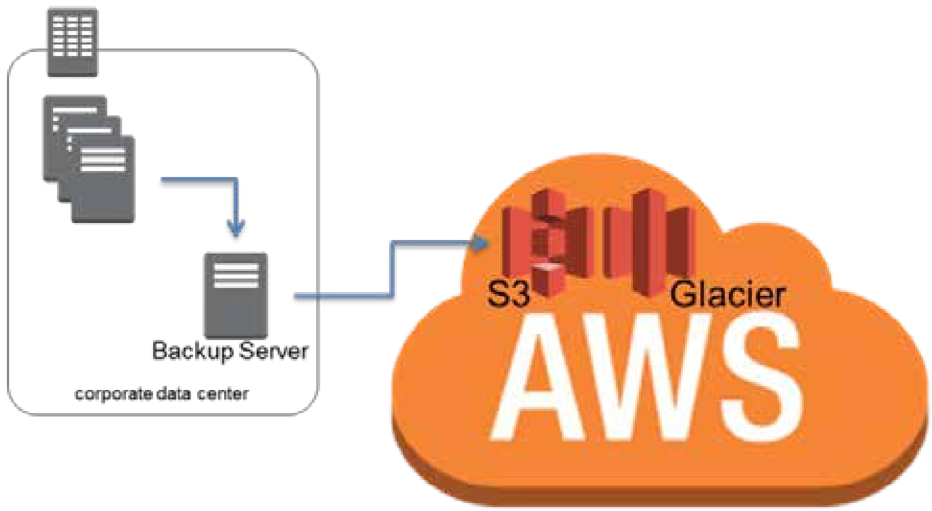

Figure 1: Backup connector to Amazon S3 or Amazon Glacier

In this scenario, existing on-premises backup and archive software that has traditionally been used to put backups on tape interfaces directly with AWS. The backup software simply writes to AWS storage tiers as the backup target, and maintains complete visibility over the data and the backup catalog. It means that you can cut over from tape to cloud without disrupting operations and still maintain continuity with backup jobs performed to tape and stored in tape libraries or offsite archives. Over time, those backup jobs, tapes, and tape hardware may quietly expire. While using backup software that interfaces directly with AWS is a simple and efficient way to begin using the cloud, we recommend confirming feature integration specifics and software versions with backup software vendors. It is also a good idea to confirm that WAN performance will meet backup window demands.

Backup to Cloud through a Hybrid Cloud Virtual Tape Library

Since tape backup has existed inside IT organizations for decades, many organizations have developed tried-and-true processes around the tape construct. A second non-disruptive method to replace tape media with cloud storage for backups is a virtual tape library that uses those same processes and also bridges from your on-premises environment into AWS storage. This design uses existing on-premises tape backup software to write to an on-premises virtual tape library with local disk for low-latency access to recently backed-up data, and cloud storage tiers instead of tape media for durable retention. Since it appears to back up products, such as a tape library with robotics, tape slots, and tapes, the software and processes operate the way they always have (except for the manual work of touching, labeling, moving, storing, and retrieving physical tapes). Backups are actually performed to local disk, then pushed up to the cloud.

The AWS Storage Gateway service can be configured to act as an industry-standard iSCSI-based VTL that connects your on-premises environment and your production applications to Amazon S3 and Amazon Glacier. This AWS Storage Gateway configuration appears to your existing backup application as a virtual media changer with virtual tape drives and tapes. Existing backup applications and workflows write to a collection of virtual tapes stored in the on-premises virtual tape library on the AWS Storage Gateway. The virtual tapes in the library are asynchronously backed up to Amazon S3. When you no longer require quick or frequent access to data contained on a virtual tape, you can have your backup application archive it from the virtual tape library into Amazon Glacier to further reduce storage costs. AWS Storage Gateway is compatible with a variety of leading backup applications, including Veritas NetBackup and Backup Exec, EMC Networker, Arcserve, Veeam, Dell Netvault, HPE Data Protector, Microsoft Data Protection Manager, and other products that directly communicate with industry-standard, iSCSI-compatible VTLs.

The AWS Storage Gateway configured as a VTL, or “Tape Gateway,” eliminates large upfront tape automation capital expenses, multi-year maintenance contract commitments, and ongoing media costs. You pay only for the amount of data written to tape, and the service scales as your needs grow. The need to transport storage media to offsite facilities and handle tape media manually goes away, and backups and archives benefit from the design and durability of the AWS Cloud storage platform.

The following diagram shows a typical AWS Storage Gateway topology when deployed as a VTL.

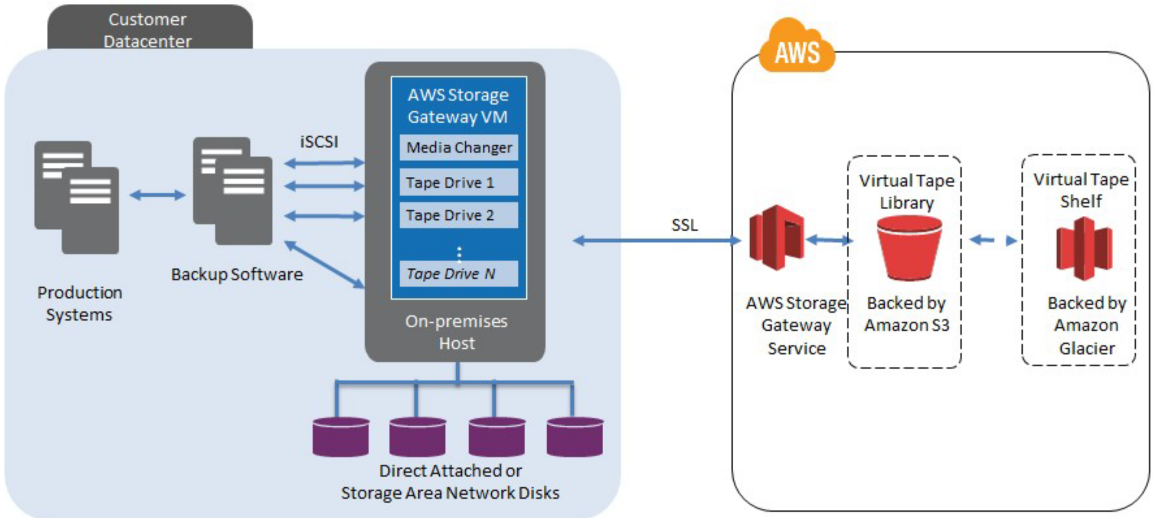

Figure 2: AWS Storage Gateway deployed as industry-standard iSCSI VTL

This topology enables low-latency, on-premises access to data that is also backed up in Amazon Cloud storage. If the data is in the local cache, recoveries happen at local disk and network speed, and recovery times are sped up by the amount of data to be restored that is in the local VTL. This removes the complexity of dealing with off-site tape storage or VTL site-to-site replication. Your backup software can encrypt your backups, or you may use AWS Key Management Service with the AWS Storage Gateway.

You can download an AWS Storage Gateway appliance from AWS as a VM and install it on a local hypervisor. The VTL configuration can be deployed on VMware ESXi or Microsoft Hyper-V. If you are deploying the gateway on-premises, you download and deploy the gateway VM and then activate the gateway. Note that the gateway may also be deployed on an Amazon Elastic Compute Cloud (EC2) instance in the cloud, which is useful for companies that want to maintain consistent organizational processes around the tape backup construct, even after moving application workloads into Amazon EC2.

Backup to Cloud via an On-Premises NFS Interface

AWS Storage Gateway may also be configured as a “file gateway”, which can act as an on-premises NFS backup storage target and push backup files to the AWS Cloud for durable retention. The file gateway deployed in a privately hosted VMware environment can be used as a backup target, acting as a performance-optimized connection between NFS-compatible backup systems in a private data center and Amazon S3 buckets hosted in a given AWS Region. The file gateway uses locally attached storage to provide a read/write cache to reduce latency for backup servers (acting as NFS clients) in the same local area network (LAN) as the file gateway.

The gateway service then transports these backup files to a supported AWS Region for storage as Amazon S3 objects. This makes it a hybrid cloud option for backup products that can use NFS. Data can be stored on low-cost, highly durable cloud storage and tiered to progressively lower cost storage as the likelihood of restoration diminishes. Figure 3 shows an example architecture that assumes backups are retained for one year. After 30 days the likelihood of restoration is infrequent, and after 60 days it becomes extremely rare.

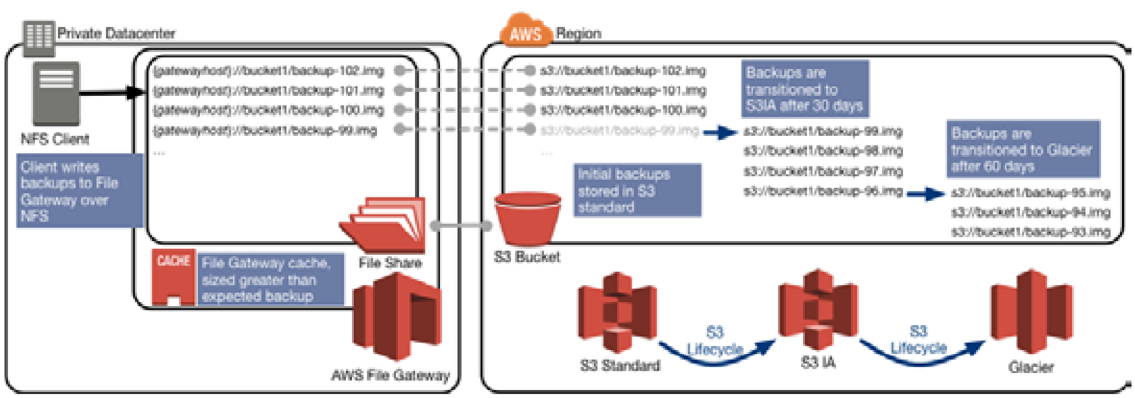

Figure 3: File gateway storing files as objects in Amazon S3 Standard and transitioning to Amazon S3 Standard – IA and Amazon Glacier

In this solution, Amazon S3 Standard is the initial target for the first 30 days. The backup software or scripts write backups to the NFS share. Note that larger files offer better cost optimization in the end-to-end solution, including lower storage costs in Amazon S3 Standard – Infrequent Access (IA) and Amazon Glacier and lower lifecycle transition costs because fewer transitions are required.

After another 30 days, the backups are transitioned to Amazon Glacier. Here they are held until a full year has passed since they were first created, at which point they are deleted.

When sizing cache for the file gateway in this type of solution, it is important to understand the backup process itself. Cache size should be large enough to contain a complete full backup, allowing restores directly from the cache and much more quickly than over a WAN link.

If the backup solution uses software that consolidates backup files by reading existing backups before writing ongoing backups, it's important to factor that into the sizing of cache also. This is because reading from the local cache during these types of operations reduces cost and increases overall performance of ongoing backup operations. For more information on using the AWS Storage Gateway as an NFS-based target for backing up to the cloud, please read the whitepaper File Gateway for Hybrid Architectures: Overview and Best Practices.[4]

Conclusion

Tape media costs are undoubtedly low on a pure $/GB basis. However, you may want to re-evaluate the additional costs of your tape-based backup solution, including labor, management, and the costs of initially purchasing, maintaining, refreshing, and licensing tape media and tape automation systems. Replacing tape backup with cloud backup is a high-priority and simple task that can unlock further benefits by moving your organization's data into the cloud.

Contributors

The following individuals and organizations contributed to this document:

- Peter Baltatzidis, product manager, AWS Storage Gateway
- Peter Levett, solutions architect, Amazon Web Services
- Dylan Locsin, product marketing manager, Amazon Web Services
- Mike Davis, business development, Amazon Web Services

Further Reading

For additional help, please consult the following sources:

- Simple Project and implementation guide detailing the steps for replacing tape backup with cloud storage[5]

- Whitepaper: File Gateway for Hybrid Architectures Overview & Best Practices[6]
- Veritas blog on using Backup Exec with the VTL[7]
- Veritas blog on using NetBackup with the S3 Connector or the VTL[8]

Notes

[1] IDC, Public Cloud Storage Spending Becomes Strategic and Business-Driven — European Storage Manager Survey, 2015, #EMEA40869715

[2] IDC, Storage User Demand Study, 2014 — Fall Edition: Growing Demand for Private Storage Cloud, #259067

[3] "Looking at HDD and SSD Markets Through the Lens of the 'Laws of Economics,'" Nov 2016, IDC, Jeff Janukowicz, John Rydning

[4] https://d0.awsstatic.com/whitepapers/Storage/aws-storage-gateway-file-gateway-for-hybrid-architectures.pdf

[5] https://aws.amazon.com/getting-started/projects/replace-tape-with-cloud/

[6] https://d0.awsstatic.com/whitepapers/Storage/aws-storage-gateway-file-gateway-for-hybrid-architectures.pdf

[7] https://vox.veritas.com/t5/Backup-Exec/Backup-Exec-and-AWS-Storage-Gateway/ba-p/828669

[8] https://vox.veritas.com/t5/Netting-Out-NetBackup/Use-Amazon-Web-Services-to-Store-Your-Backups/ba-p/826330