Snapshot Report

Using object storage as a target for backup, disaster recovery, archiving

By Deni Connor, founding analyst
Earl Follis, senior analyst
SSG-NOW
June 2015

In many organizations, backup data accounts for more than 50% of all stored data. Although this does not translate into 50% of storage equipment, this much stored data costs money and significant resources to maintain. Given fast-growing rates of data growth, and since corporations are famously reluctant to delete large amounts of data, this percentage is only going to grow in environments using traditional methods of data protection. This amount of data growth plus flat IT budgets is causing storage architects and backup admins to rethink their data protection strategies. Several issues are forcing this reassessment:

- *A shift from tape to nearline disk for recovery with high RTO requirements*. Tape can be a cost-effective choice for long-term data retention. But nearline disk has big advantages for recovery point objectives (RPO), recovery time objectives (RTO) and efficient backup windows.

- *Consolidating the data center*. IT admins prefer to consolidate equipment into fewer locations and racks. Many are choosing software platforms that offer full data management functionality instead of buying multiple specialized appliances.

- *The growth of public clouds*. Cloud technology represents a major shift in data protection thinking. The cloud acts as an extension to the data center for computing, storage, data management and disaster recovery. However, cloud

architecture is not the automatic answer. Slow bandwidth introduces latency into the data movement process. Cloud latency has a minimal impact on RPO but can strongly affect RTO and backup window times.

- *Reducing the complexity of private clouds*. Virtualizing Tier 1 applications has significant benefits, but the public cloud often cannot meet Tier 1 security and disaster recovery (DR) standards. Private clouds redress these issues with LAN speeds and corporate firewalls. However, private clouds built with industry-standard server and storage architectures can be highly complex and costly.

- *Increased regulatory compliance*. Tape may be a cost-effective solution for long-term compliant data retention. However, it is not ideal for active compliance and discovery as users search, extract, analyze and act on relevant data. In addition, tape schedules between on-premises equipment and vaults are very difficult to manage. And, whether data is stored on-premises or off-site, applying consistent retention policies is difficult.

- *Storage silos*. A deduplication domain is defined by the scope of the backup application: If there are multiple backup programs running in the same infrastructure, then there will be multiple deduplication silos. This increases storage requirements and locks in IT to specific deduplication and backup vendors.

The Changing Backup Landscape

These factors are changing the nature of the data protection marketplace in favor of easily accessible nearline disk, which is one key to faster RTOs and RPOs and shortened backup windows. Disk vs. tape performance is the largest single factor. Even with multi-threaded backup and high-performance tape drives, disk is still faster than the fastest tape.

Access also plays an important part. Tape is ideal for sequential access, an earmark of standard backup streams. However, disk does provide high-speed sequential and random access, which makes it the more flexible choice for backup and archiving writes and recovery.

Another shift in the backup marketplace is that in spite of huge data growth, IT budgets are shrinking. Admins are looking to do much more with quite a bit less. Earlier in backup and recovery development, the cost of backing up and storing large backup streams was prohibitive. Organizations invested in more and bigger backup storage systems just to store backups and archives. High cost drove the development of deduplication, a large subset (multi-billions of dollars worth) of the disk-based backup

market today. Deduplication granted fundamental storage efficiencies. At first, deduplication developers offered appliances to add deduplication backup streams. However, we are seeing significant changes in the market.

Although purpose-built deduplication appliances are still around, the backup vendor community has wasted no time adding deduplication to its applications – deduplication is quickly becoming a feature instead of a product. First generation backup servers benefitted from offloading dedupe operations to a separate appliance. This is not the case with today's fast processors, which manage far more load than older backup servers. And by adding deduplication to the backup software, IT does not need to buy separate dedupe appliances.

Disk-based backup also drove the development of snapshots and replication between NAS devices. Array-based replication was a strong addition to the data protection arsenal but the arrays with native replication were expensive and vendor-specific. This created hardware lock-in. These arrays are also limited in scalability and geographic distribution.

In response, backup software makers targeted disk for faster recovery and added replication features to their software. The software can be added to different types of arrays (or operate on higher levels in the stack) while replication processes and policies are universal.

How disk-based backup plus object storage can help

A third major factor driving disk-based backup is the growing popularity of object-based storage. Object storage creates a unified storage pool with a global namespace, easily accessible from any media agent with basic account credentials. The storage pool operates on scale-out hardware, running anywhere from a single location to many locations under a single namespace.

Although storage pools have existed for many years, this level of flexibility and scalability is much more difficult to do across data centers and multiple NAS heads. It is even more difficult if the enterprise is using deduplication appliances for its backup, instead of a backup application operating deduplication across the object store.

Object storage is a fundamental cloud computing and storage technology for public and private clouds. For security-conscious companies who want LAN speeds for their backup and recovery, the ideal solution is a private cloud that is massively scalable, very flexible

and highly available. This object-based private cloud scales out proxy resources, such as bandwidth, independently of scaling capacity with additional nodes. The architecture avoids common bottlenecks due to limited controller ports, array resources or appliances. True scalability across multiple locations and systems becomes possible, even simple for the enterprise.

IT can still backup long-term, lower value data to the public cloud or even to tape. But this private cloud architecture not only protects backed up and archived data, it enables swift and granular recovery, active archives and stored data protection that is second to none.

Eventual Consistency and Erasure Codes

Let's look at two additional data protection processes in this architecture. We'll take market leader SwiftStack as our example of a distributed object storage system that is rendered as a private cloud. Built-in availability allows multiple protection classes in the same cluster. Flexible protection features include eventually consistent replicas and erasure codes, which are conceptually similar to assigning different data protection schemes to different storage tiers.

'Eventually consistent' simply means that over time in an object store, all copies of unstructured data will adopt changes written to its most recent copy. This low-impact process takes minimal resources, is highly scalable and provides strong availability for unstructured data at a low cost. This process occurs in milliseconds, which achieves very fast backup without threat to restores.

SwiftStack replicas are replicated partitions (i.e., consistently sized directories represented by hash tables). Multiple replicas -- commonly three copies of each partition in a single-site cluster -- are placed in different cluster nodes as protection against node failure. The configuration also includes 'handoff locations' in the cluster where auditors and replicators made new replicas in the event that a replica drive should fail. SwiftStack recommends additional replicas for globally distributed clusters.

Erasure code (EC) technology enables data to be spread to multiple nodes, which delivers extreme data durability. The process breaks data into fragments that maintain their original identity with the addition of redundant data. These fragments are stored as uniquely as possible.

Should data fragments become corrupted, EC enables the storage system to rebuild data from the stored fragments. SwiftStack, for example, identifies a failure condition and

quickly checks neighboring segments of fragments for availability. The EC process rebuilds the failed data from the chunks and pushes them to their proper locations, including hand-off disks as necessary. SwiftStack also enables an additional EC ring to store EC data on different hardware as needed. This lets admins dedicate high CPU cycles for priority erasure-coded data and less expensive CPUs for higher density archived data.

Although EC can be CPU-intensive, it is more efficient with small files or chunks and is more storage-efficient than RAID. In fact, EC's data protection is similar to RAID 5 and 6, which respectively are distributed parity and dual parity. The number of parity fragments in an Erasure Codes policy is configurable, for example 8 data fragments and 4 parity fragments would tolerate up to 4 failures.

Integrating Simpana and SwiftStack, and the Importance of being Object

Now that the enterprise can deploy a scalable and flexible private cloud, they will want equally sophisticated backup software that can natively write to the new cloud. Commvault and SwiftStack have partnered to provide exactly that to provide significant savings to joint users of Simpana 10 and SwiftStack Object Storage. Before we turn to partnership details, let's lay the groundwork by looking at the founding companies and what they mean to object-based storage in the private cloud.

Let's set the stage with OpenStack Swift and SwiftStack. Swift is a widely adopted, open source object storage system that powers some of the world's largest clouds. Swift stores unstructured data of all types -- files, backups, images, videos, virtual machine (VM) snapshots and more -- in a highly available architecture with no single point of failure. Swift clusters are multiple nodes that run Swift's processes and consistency services.

This high availability architecture works very well as a stand-alone storage system. However, it comes into its own as a massively scalable, cloud computing infrastructure. Clusters are scalable from a simple three-node configuration to global distributed systems storing petabytes of data. Swift's native web functionality enables flexible application access with REST APIs or coding language libraries, including Python. All of this power is cost-effective: Open source Swift runs on standard Linux operating systems and industry standard server hardware with commodity hard drives.

Scale-out is another distinct advantage over file-based system scale-up architectures. The system scales linearly with no downtime and supports hot swapping. More importantly, object-based scale-out is simpler and less expensive to scale than NAS

alternatives. Scale-out NAS can provide petabytes of storage, but the scaling process impacts performance on controller and networking levels. Adding new systems while preserving performance is costly and increases the management load. In contrast, Swift's software-defined architecture runs on commodity servers and low-cost hard drives for cost savings.

The drawback is that for all the brilliance of its object storage architecture, adopting Swift for the enterprise requires significant integration work and configuration. Swift lacks native central management and global provisioning, and as an open source product it does not provide support. Swift is also a complex program where upstream errors can badly affect downstream processes.

Business must decide if they are going to go DIY with a complex deployment, or hire expensive specialty professional integrators who will not only install Swift but maintain it over months and years.

Fortunately there is a third option from SwiftStack, who builds-on the Swift engine, adds enterprise features, makes it a turnkey product and provides 24x7 professional support. SwiftStack essentially removes node management from the physical servers, replacing it with an out-of-band software controller that manages multiple Swift clusters. And unlike Swift, SwiftStack does not rely purely on code libraries or the REST API for application support: It also provides an optional gateway for NFS and CIFS.

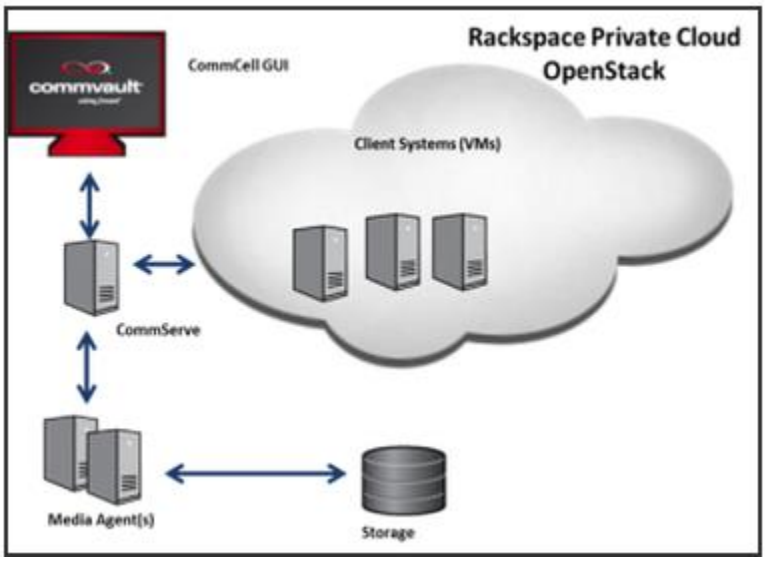
Only this set of advanced capabilities allows the enterprise to maintain central control over a widely distributed Swift storage system including optimal cluster tuning, capacity management, LDAP and Active Directory integration, and rolling upgrades.

Commvault on Object-Based Storage

Now we turn to the other partner: Commvault. Commvault makes no bones about Simpana being a full-featured data management platform. Simpana 10 supports block, file and object storage types; and writes to disk, tape and cloud tiers. In its capacity leveraging object storage, Simpana can transfer data to external object storage existing on-premises, remotely or in the cloud. Cloud data transfer occurs using object storage APIs, including OpenStack Swift.

Processes, such as deduplication, storage tiering and auxiliary copy creation, are built-in to Simpana. Its native policy management controls -- which Commvault first designed for public cloud usage -- automate inactive data movement to lower cost tiers. This reduces archive footprints up to 70%, which is a huge benefit to capital and operating costs.

For example, Rackspace Private Cloud is powered by OpenStack. Commvault partners with Rackspace on file-level backup and recovery of data inside individual instances. Commvault allows IT to configure and automate source-side deduped backups with instructions on data protection levels and retention within the Private Cloud.



1+1=3: Commvault Simpana and SwiftStack

Commvault is all about protecting data and so is SwiftStack. SwiftStack's close integration with Commvault lets customers use SwiftStack off-the-shelf for data management and protection on massively scaled clouds. By easily distributing data across different locations and systems, Commvault plus SwiftStack offers data protection in a class by itself.

When Commvault went looking for such a partner, Simpana already offered native support for Swift in the public cloud. Commvault wanted to develop a turnkey integration with a Swift partner to create integrated private clouds. They found their partner in SwiftStack, whose Swift-based private cloud architecture rivaled the scalability and flexibility of the Swift-based public clouds.

Now Commvault integrates with SwiftStack to power this enterprise-grade global private cloud. Simpana natively supports SwiftStack as an object storage backup target. The result is an integrated solution set that combines a powerful, global private cloud with a single data management platform for backup and recovery, replication, archiving and search/eDiscovery -- all in a single distributed object storage system.

Admins can painlessly introduce Simpana/SwiftStack into an existing infrastructure with a small three-node system if they wish; knowing that linear scaling to multi-petabyte levels with standard server hardware and standard hard drives is a simple matter. Commvault provides the backup and recovery, archiving and search capabilities including deduplication and auxiliary copies -- there is no need to buy separate appliances. RPO and RTO both benefit from object-based backup and recovery, and IT can centrally manage both Simpana software and the SwiftStack infrastructure.

Top Benefits of Using Commvault Simpana together with SwiftStack

- **Global clusters.** Automated replication of backups between data centers, with very simple restores from any media agent with account credential information. This automatic offsite protection can eliminate the need for most tape vaulting.

- **Storage policies.** Defined pools of storage within the global pool with different characteristics including geographical locations, protection levels, even media type.

- **Simplified retention policies.** Backing up data to object-based storage simplifies retention policies by replacing complex tape cycles and expensive array-based replication.

- **Faster backups.** Users can configure Simpana to take advantage of bandwidth across multiple nodes in the cluster, replacing the limited bandwidth available with tape or disk arrays.

- **Centralized views.** The global namespace enables a single integrated view into Simpana and SwiftStack, turning storage and data management into simple processes. The unified view also provides a single system of record for data.

Characteristics	Tape-based backup	Disk-based backup/Purpose-built backup appliances	Deduplication-based backup appliances/ software	Public cloud	Open Private cloud
Storage efficient	H	H	H	H	H
Resource scaling	L	L	L	H	H
System can be expanded a node at a time	L	L	L	H	H
Scale-out	L	L	M	H	H
Global management	L	M	L	H	H
Protection from media failure	L	L	L	M	H
Highly tolerant to drive failures	L	L	L	H	H
Geographic distribution	L	M	L	H	H
RESTful APIs	L	L	L	H	H
Near-line archive	L	H	H	H	H
Archive	H	L	M	H	H
Long-term retention	H	L	L	H	H
Search and discovery	M	L	L	H	H
Load-balancing	L	L	L	N/A	H
Fast restore	L	H	H	M	H
Eliminate vendor lock-in	L	L	L	H	H

Evolution

High Medium Low

Figure 2. The New Data Protection Continuum

- **Low-cost, Open scale-out architecture.** SwiftStack is built with standard server hardware and standard hard drives, independently scales bandwidth and capacity and includes standard HTTP load balancing across all nodes in a cluster. The private cloud protects data availability with no single point of failure; geographically distributed data is immediately available.

- **Highly secure multi-tenancy.** Commvault resellers and private cloud providers can offer the integrated package while preserving exceptional customer security.

Our Take

Tape is cheap … or not. Regulating tape schedules sucks up IT time and while media may be cheap, tape libraries are not. Disk-based solutions stepped into the backup performance gap and evolved to deliver a better infrastructure for backup, enabling more granular protection and faster restore. And, then the industry took another important step forward, combining disk backup with object-based storage. The result is that performance and reliability soar, while costs fall.

OpenStack Swift can be used in this architecture for a private storage cloud, slashing high costs with open source software and standard server hardware with standard hard drives. Eliminating storage vendor lock-in and expensive storage silos also minimizes expenses. SwiftStack provides enterprise features and off-the-shelf functionality, enabling enterprise-size businesses to create powerful private clouds with unlimited scalability in a familiar way.

There was one more important step to take after creating hugely scalable private clouds: Scale data protection to match. The enterprise struggled to store and protect massive volumes of unstructured data. They need the massive scalability of a cloud to do it, but did not want to sacrifice fast recovery and active archiving to slow bandwidth. They also need to watch costs even as data volumes grow. This is where Commvault steps in: Simpana's full-feature platform, natively integrated with SwiftStack.

This partnership is a standard-bearer of the continuing evolution of data protection and access. With SwiftStack's fully distributed architecture, the enterprise finally gets massive scalability, central control, and global data transparency from a private cloud.

And with Simpana's seamless integration, the business also gains powerful backup and recovery, active archiving, replication, and global search -- all running under system-wide master policies. Just a couple of years ago this kind of powerful, off-the-shelf protected cloud environment seemed impossible. Today it is more than possible -- it's here. ●