Riverbed Whitewater/Amazon Glacier ROI for Backup and Archiving

November, 2013

Saqib Jang

Abstract

This white paper demonstrates how to increase profitability by reducing the operating costs of backup and archiving using Riverbed Whitewater and Amazon Glacier cloud storage, instead of traditional tape backup methods. A convenient checklist summarizes operational cost categories used to calculate this ROI. These include ongoing monthly costs for media, labor, and offsite storage of backup data, and monthly costs for cloud storage of backup and archival data.

This Margalla Communications White Paper was commissioned by Riverbed Technology.

COPYRIGHT NOTICE

Executive Summary

The Riverbed Whitewater appliance enables enterprise IT to optimally leverage cloud storage services such as Amazon Glacier to store backup and archival data sets. Whitewater is an enterprise premise-based cloud storage appliance that uses the cloud as a backup target. By using data deduplication with cloud storage replication, the Whitewater appliance allows IT to take full advantage of the attractive pricing of cloud storage services. In addition, the ability of Riverbed Whitewater to cache data locally allows immediate access to recent data while avoiding unnecessary costs for transferring data from cloud storage.

Amazon Glacier is designed for low-cost storage of large datasets of infrequently accessed data, such as long-term archives, yearly full backups, and other data sets which typically, once written, are rarely read. This service directly addresses today's needs for a low-cost cloud storage solution; enabling companies to more cost effectively meet compliance and regulatory requirements. To deliver this extremely low baseline cost, Amazon Glacier has restricted data retrieval times to 3-5 hours, and provides recovery costs based on size of datasets as well as aggregate transactions for accessing data from Glacier cloud storage.

Whitewater appliances allow businesses to reap the low-cost benefits of Amazon's Glacier service, especially for very large backup and archiving data sets, while seamlessly addressing the challenges associated with restoring data from Glacier, such as 3-5 hour delays, expensive restoral fees, and the relatively cumbersome process for accessing data. Whitewater caches the most recent and frequently accessed data within the appliance, and there will be no delay for restores unless the data is not in the cache. This means that customers can rapidly restore their cached data from Whitewater instead of waiting hours to get it back from Glacier.

Whitewater also simplifies the process of accessing data from Glacier while minimizing Glacier restoral fees. In cases where the data for a restore request is not in its disk cache, Whitewater utilizes locally stored metadata for data objects in Glacier to automatically submit a Glacier restore request and alerts the user when the data has been received and is ready in the appliance. Whitewater also supports built-in capabilities that enable customers to take advantage of the complimentary retrieval amount Glacier service offers users each month, meaning that the additional data segments can be recovered at minimal cost.

Analysis of the return on investment (ROI) for a typical deployment shows that the use of Whitewater and Glacier delivers more than 70% operational cost savings when compared to on-premise tape infrastructure over a 3-year period. The results of this analysis can be extrapolated to customer scenarios having different characteristics to calculate corresponding savings that Riverbed/Glacier can deliver.

Introduction

Storage requirements for backup and long-term archiving for data protection and disaster recovery are growing at an unmanageable rate. Enterprises of all types need to back up active data regularly to avoid losses, stay compliant, and preserve data integrity. However, legacy data protection approaches such as tape backup are expensive and have become increasingly difficult to manage due to the large number of tapes frequently required for a restore procedure. Plus, new regulatory requirements are putting increasing pressure on existing legacy solutions that were not designed to meet the new realities of today's world.

In addition, enterprises frequently face a challenge getting their backup data offsite on a regular basis. The traditional solution has been to periodically send backup tapes to a records warehouse run by companies such as Iron Mountain. That creates a lot of work for backup administrators, as they have to create and manage additional jobs to duplicate backup data to tapes for offsite storage, then box up the tapes and deal with the courier. The result is that most companies send tapes offsite only once a week to once a month, which leaves their data highly vulnerable.

Amazon Glacier charges significantly less for data storage than do other cloud services and traditional online backup services for data storage. As a result, there is significant interest among IT professionals for utilizing Amazon Glacier to store backup and archive data. Amazon Glacier offers enterprises very cost-effective offsite backup and archiving storage capabilities, but using Glacier for backup storage may be less effective than it first seems, particularly as the size of backup data sets grows.

While some backup applications have added the ability to duplicate backup data to Amazon Glacier, most do not deduplicate the data before sending it to the cloud. Since enterprises will be paying Amazon for each and every gigabyte of data that is stored each month, deduplication would definitely be an imperative. In addition, it takes time to send and retrieve files to Glacier, time that is determined by WAN link speeds and as well as Glacier's 3-5 hour service level agreement (SLA) for data retrieval. Retrieving the data also incurs additional fees.

The Riverbed Whitewater appliance enables enterprise IT to optimally leverage cloud storage services such as Amazon Glacier to store these backup and archival data sets. Whitewater is an enterprise premise-based storage appliance that uses cloud storage as a backup target. By using data deduplication with cloud storage replication, the Whitewater appliance allows IT to take full advantage of the attractive pricing of cloud storage services designed for the long-term, in particular Amazon Glacier. In addition, the ability of Riverbed Whitewater to cache data locally allows immediate access to recent data while avoiding unnecessary costs for data transfer from Amazon Glacier.

Riverbed Whitewater Overview

Riverbed Whitewater appliances act as gateways between backup applications and the cloud provider of choice, including Amazon Glacier. Whitewater appears to backup applications as network-attached storage (NAS), accepting data via Common Internet File System (CIFS) and/or network file system (NFS) protocols. Once data is sent to the Whitewater appliance, it deduplicates, compresses, and encrypts it

before storing it in its local disk cache and transferring it to cloud storage, while retaining the metadata for data objects stored in the cloud.

Riverbed Whitewater deduplicates backup data inline, storing it to the local cache and sending the deduplicated data blocks to the cloud storage provider as quickly as it can over the wide-area networking (WAN) link. The Whitewater appliance compares incoming data against data seen before to see if the data block is a duplicate. If a backup is performed that contains data blocks that had been deduplicated by Whitewater once, Whitewater only creates metadata pointers to already-deduplicated data blocks in its disk cache. Such blocks will not be stored again in the Whitewater cache since the Whitewater appliance does not see it as new data.

As most data restore requests are for recent data, the Whitewater appliance will reassemble data for such requests using data blocks from its cache, which most likely includes all the blocks from the most recent backup, and the response is immediate. When restoring older backup data, the Whitewater appliance will transparently initiate a retrieval job for blocks stored on cloud storage, which, for Amazon Glacier, will typically complete in 3-5 hours. Regardless of the amount of data that needs to be retrieved from cloud storage, the deduplication, compression, and WAN optimization capabilities provided by Whitewater will maximize the speed of the data transfer.

A primary Whitewater appliance can replicate its cache with another peer Whitewater appliance, providing customers with a warm standby Whitewater appliance from which to recover data instantly in the event of a disaster at the primary site. The benefits of this approach include reduced recovery time object (RTO), reduced cloud recovery costs, and improved business continuity.

Whitewater appliances also enable backup administrators to actively "pin" specific data on the Whitewater cache, preventing that data from being evicted from the Whitewater cache later, to make space for new data. This guarantees immediate access to critical data, such as SQL backups, specific email archives, or key business data.

AWS Glacier Overview

The latest offering in the Amazon Web Services (AWS) portfolio is a new class or tier of storage service called Glacier, which (as its name implies) has reduced performance and enables support for very large datasets. In other words, Glacier enables very cost-effective support of inactive or seldom-accessed data in exchange for a longer RTO. Specifically, in exchange for an RTO in the order of 3-5 hours, Glacier offers baseline storage costs as low as 1 cent per Gigabyte (GB) or 12 cents per year per GB, in addition to optional fees for access to stored data.

Glacier is designed to help reduce the costs companies typically over-pay for archiving data. It eliminates the requirement that a company must purchase an expensive archiving solution (which does not include the ongoing cost for operational expenses such as power, facilities, staffing, and maintenance). By providing an elastic low-cost storage solution, businesses do not have to guess what their archiving capacity requirements and corresponding costs will be. This eliminates concerns about over-provisioning or under-provisioning these environments while correctly estimating the respective budgets tied to these projects. With Amazon Glacier, you pay only for what you use.

Glacier is designed for low-cost storage of large datasets of infrequently accessed data, such as long-term archives, yearly full backups, and other data sets which typically, once written, are rarely read. This service directly addresses today's needs for a low-cost cloud storage solution; enabling companies to more cost effectively meet compliance and regulatory requirements. To deliver this extremely low baseline cost, Amazon Glacier has restricted data retrieval times to several hours, and provides recovery costs based on size of datasets as well as aggregate transactions for accessing data from Glacier cloud storage.

Whitewater Value Proposition for Glacier

Whitewater appliances allow businesses to reap the low-cost benefits of Amazon's Glacier service, especially for backing-up and archiving very large data sets, while seamlessly addressing the challenges associated with restoring data from Glacier, such as 3-5 hours delays, expensive restoral fees, and the relatively cumbersome process for accessing data.

Whitewater caches the most recent and frequently accessed data within the appliance, and there will be no delay unless the data is not in the cache. Riverbed provides cache size guidelines based on backup dataset size for selecting the appropriate Whitewater model so as to maximize cache hits (the largest Whitewater physical device, the 3030, supports up to 96 TB cache size, and can therefore logically store up to 30 times 96 or 2.88 PB data locally). This means that customers can rapidly restore their cached data from Whitewater instead of waiting hours to get it back from Glacier.
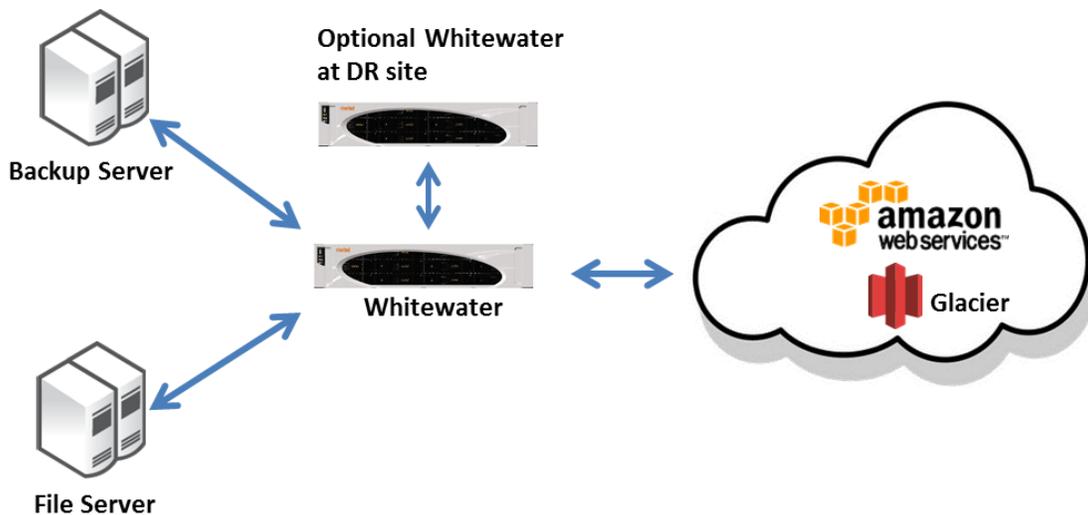

Figure 1. Riverbed Whitewater and Amazon Glacier enable the creation of a low-cost backup and archival storage tier (Source: Riverbed)

Whitewater also simplifies the process of accessing data from Glacier while minimizing restoral fees (based on restoral transactions as well as size of data transfer). In case that the data for a restore request is not in its disk cache, Whitewater utilizes locally stored metadata for data objects in Glacier to automatically submit a Glacier restore request and alerts the user when the data has been received and

is ready in the appliance. Customers can also specifically request files known to be in Glacier and not in the appliance's cache.

Whitewater also supports built-in capabilities that enable customers to take advantage of the complimentary 5% retrieval allowance Glacier service offers users each month, meaning that the additional data segments can be recovered at minimal cost. This means that by default, Whitewater can perform flow control of data retrieved from Amazon Glacier at a potentially slower rate than what the WAN connection allows. In the event of restores of large data sets (or restores which are needed faster than what is permitted by the complementary rate), the flow control capability can be increased or disabled on Whitewater to allow faster data retrieval. The tradeoff results in higher retrieval costs.

Riverbed Whitewater/AWS Glacier ROI Advantage

This section provides an estimate of the operational cost savings of using Riverbed Whitewater and Amazon Glacier over traditional on-premise tape backup methods for a typical data backup and archiving scenario. Using such an estimate, it becomes possible to analyze the ROI of moving to Whitewater/Glacier-based cloud backup and archiving and the profitability of reducing operational costs by investing in a cloud implementation for an actual real world scenario. The model also delineates additional soft benefits that, although difficult to quantify in financial terms, have very real value to a corporation that need to be considered in the cloud backup/archiving deployment model.

Operational Cost Overview: Tape backup systems

A range of enterprises continue to utilize tape for data backup and long-term archiving—and continue to invest in on-premise tape-based backup solutions. With tape-based backup and recovery, IT is responsible for the following tasks:

- Backing up to a tape-based system, then storing the tapes onsite or transporting them to offsite storage. In a remote office, non-technical staff may have to perform this work.
- Recovering data from tapes for restoration to end users or in response to discovery requests.

When preparing the business case showing the ROI for moving to cloud-based backup/archiving, the operational costs of a tape-based backup and recovery system need to be documented.

These cost categories include:

- Labor, whether performed by in-house staff or by contracted services
- Offsite backup tape pickup, storage, and maintenance

The first operational cost category is for labor costs (whether performed by in-house staff or by contracted services) for the range of functions comprising tape-based backup/recovery including:

- Ongoing daily tasks such as running backups, tape management, reviewing logs, and tracking down causes for backup errors.
- Hours spent per month recovering or restoring files for users (both end users and requests for discovery).

- Tasks performed on a periodic basis; for example, performing functional tests to validate that backups run correctly and that data can be restored within an acceptable amount of the time.

The second category of operational costs for tape-based backup and archiving is for offsite backup tape pickup, storage, and maintenance service costs. Such costs are based on the frequency of pickup, storage fees, and charges for maintaining an offsite backup tape collection, including an estimate for the number of times a recovery tape may need to be requested during one year.

Operational Cost Overview: Whitewater/Glacier for backup/archiving

Riverbed/Glacier-based data backup/recovery and archiving services are uniquely suited to address data protection and long-term preservation and offer the following benefits:

- Backup and archive data moves to an offsite location simplifying data management, while Whitewater continues to enable disk-speed access to backup data.
- The cloud option leverages Amazon's infrastructure and expertise in data durability and availability.
- Freed from attending to cumbersome, labor-intensive management of tape backup systems, IT personnel can align themselves better with business goals.

Following are the cost efficiencies of a Whitewater/Glacier-based data protection and preservation service:

First, there are no on-going costs for implementing cloud-based data protection and archiving, aside from IT staff's time spent deciding which types of files to back up, how often to back them up, and how long to retain them. Any costs to manage the life cycle for backup and archive data are borne by Amazon, rather than Glacier customers. It is also worth noting that there are additional possible benefits in end-user uptime by utilizing Whitewater replication capabilities as well as the high-availability capabilities of the Glacier service.

Second, the cloud option does not require any maintenance and media costs for storage infrastructure. Storage fees for backup and archive data (per backup schedule and data retention rules) are included in Glacier monthly fees. Importantly, Glacier passes on its massive economies of scale to customers in the form of very low, competitive rates.

Third, the cloud option allows significant savings for access to backup and archive data. Data on backup tapes is often unrecoverable, and independent studies suggest that a majority of recoveries fail because of errors in the backup process. Unlike the laborious task of tracking errors in tape-based backups for implementing restores, the cloud option allows immediate access to recent data from the Whitewater disk cache. In addition, Glacier provides average annual data durability of 99.999999999% or at an "11 nines" level, so that customers can rest assured that their data will be completely safe while stored within Glacier.

Fourth, with automatic, immediate backup and archival to an offsite, highly secure location, there is no additional cost to transport tapes offsite. Companies may decide to increase Internet connectivity

bandwidth with Whitewater, but typically these costs are low and leveraged by other services at the company, in addition to cloud backup.

ROI Case Study: Whitewater/Glacier vs. Tape Backup

This section focuses on comparing the 3-Year ROI for a Whitewater/Glacier cloud-based solution against the traditional tape option for a typical data backup and archival scenario. The base-level assumptions for the scenario can be extrapolated to estimate the ROI for a specific customer deployment.

Assumptions

Following are base-level assumptions that apply to both the cloud and tape backup scenarios:

- **Source dataset: 75TB**
- **Annual dataset growth rate: 25%**
- **Initial first pass deduplication rate: 3**

The ROI model assumes the following parameters regarding the tape backup process:

- **Cost per LTO 5 tape: $30**
- **Capacity per LTO 5 tape: 3TB (1.5TB base capacity times a compression factor of 2)**
- **Tape Vaulting Model**
  - **Cost: $400/TB/Year**
  - **Tape backups are transferred offsite every week**
  - **In addition, a monthly full backup archive is created for long-term retention offsite**
  - **Weekly backup tapes are recycled after 90 days**
- **Tape Administration: 10 hours/week or $22,500 annual cost**
- **Backup policy that keeps monthly full backup sets as an archive. All other backup data sets are kept for 90 days. After 90 days, this data is deleted and in the case of tape, the tapes are reused.**

These are the assumptions behind cloud-based backup and archiving using Whitewater/Glacier:

- **No duplicate blocks are ever written to Whitewater cache** which means that only 1% incremental data that is new is added each month to the Whitewater cache storing data processed earlier. The result is an effective deduplication ratio for Whitewater of greater than 20 during the 3-year period of interest.
- **Backup policy that keeps monthly full backup sets as an archive. All other backup data sets are kept for 90 days.**
- **Glacier storage costs $.01 per GB per month.**

ROI Comparison

		Whitewater/ Glacier Operational Costs	Tape Cartridge Costs	Tape Vaulting Costs	Tape Administration Costs	Tape Operational Costs
Year 1 Costs		$4,798	$22,410	$37,499	$22,500	$82,409
Year 2 Costs		$8,274	$16,020	$47,079	$22,500	$85,599
Year 3 Costs		$12,638	$19,950	$59,106	$22,500	$101,556
Totals		$25,711	$58,380	$143,684	$67,500	$269,564
Total Operational Costs		$51,422				$202,064
Savings with Glacier	$150,642					
Percent Saved	74.6%					

Table 1. Cost Savings for Whitewater/Glacier vs. Tape for backup and archiving

A summary of the ROI model is shown above in Table 1. For the Glacier option, the operational costs include the costs of storing backup and archival data in Glacier over the full 3-year period. For the tape backup case, the costs include the costs of tape cartridge hardware, labor costs for managing the tape infrastructure as well as the on-going costs for offsite tape vaulting. The analysis show that for a typical customer deployment over a 3-year period Whitewater and Glacier will deliver more than a 70% cost savings versus on-premises tape infrastructure. The results of this analysis can be extrapolated to customer scenarios with different base-level assumptions, such as initial source data size, growth data and deduplication ratio, to calculate corresponding potential savings that Riverbed/Glacier may deliver.

Conclusion

Riverbed Whitewater appliances allow businesses to reap the low-cost benefits of Amazon's Glacier service, especially for backing-up and archiving very large data sets, while seamlessly addressing the challenges associated with restoring data from Glacier, such as 3-5 hours delays, expensive restoral fees, and the relatively cumbersome process for accessing data. With the ability to pin data and also replicate to another appliance, the solution can provide immediate access to data while leveraging the cost advantages offered by Glacier.

A ROI analysis of operational costs for a typical customer deployment shows that over a 3-year period Whitewater and Glacier will deliver more than a 70% cost savings over on-premises tape infrastructure. Organizations looking for solutions to streamline their data protection process, eliminate the administrative workload of tape based systems while reducing the costs of storing backup data sets should consider the Whitewater/Glacier solution.