Enterprise Backup and Recovery

On-Premises to AWS

Curd Zechmeister, Alex Tomic, Radhika Ravirala, Jeff Nunn

December 2014

Contents

Abstract

Many enterprise businesses struggle to build and deploy a comprehensive, cloud-based backup and recovery strategy for on-premises systems or systems that run in hosted environments. This whitepaper provides best practices for designing and building hybrid backup architectures that leverage Amazon Web Services (AWS) storage and compute services. The whitepaper is especially designed for those who manage backup and recovery, disaster recovery, and storage processes for enterprise businesses. We include reference architectures to guide you in the design process and to help you get started with using the AWS cloud as the backup and recovery environment for your on-premises systems.

Introduction

For most enterprise businesses, on-premises backup and recovery solutions are costly endeavors in terms of resources, and the return on investment can be minimal. Today, many businesses are turning to the cloud for backup and recovery solutions instead of building and maintaining complex, costly on-premises environments.

Amazon Web Services (AWS) offers a broad range of highly secure, scalable, and cost-effective storage options for backup and recovery. You can use AWS services to augment your existing on-premises backup and recovery environment, or you can use AWS services to build solutions that are based solely in the cloud. In this whitepaper, we discuss both options and help you choose the AWS services that are right for your business.

Traditional Backup and Recovery Methods

First, let's look at typical scenarios that we encounter frequently when we work with customers who want to simplify their backup and recovery environment and leverage best practices for cloud-based solutions. The most common scenarios range from traditional backup technologies that use magnetic tape-based systems to solutions that are completely disk-based and either implement virtual tape on disk or are based on data snapshots written to disk.

Tape Systems

In a typical tape-based backup architecture, data is persisted on either network attached storage or local disk storage. On a set schedule, the data is sent to a backup server that collects and writes the data to magnetic tape, which is stored in large, on-site tape libraries that are manually managed by tape operators or automatically managed by some form of robotics. Many organizations also replicate their mission-critical application data over WAN to smaller tape libraries offsite.

Although backups written to magnetic tape offer easy storage and simple replication, they have inherent drawbacks, such as sequential reads, lack of frequent testability, long backup windows, and cartridge failure during backup and recovery.

These flaws render tape-based backup solutions inadequate in a disaster event. If you currently use a tape-based backup solution for your on-premises environment, then moving to the cloud not only will remove the complexities of physical media handling, but also will greatly reduce the cost involved with physically storing magnetic media.

Disk-Based Backups

In the backup-to-disk (B2D) approach, data is stored in the form of snapshots of the primary storage device or is first written in tape format to disks and later moved to tape for long-term storage. The benefits of using a disk-based backup solution over magnetic tape are speed, reliability, and flexibility.

Despite the advantages that B2D solutions offer, using disks as replacements for tape can be costly. With disk-based backup solutions, you frequently must make a backup in the form of a snapshot of the volumes that contain your backup data. Thus, you are actually increasing the amount of disk that is needed to store your data, which potentially could increase the total cost of the solution.

Virtual Tape Library (VTL)

A virtual tape library (VTL) essentially is a disk-based file storage that emulates a traditional tape medium. Virtual tape library technology is by far the most frequently deployed backup technology for on-premises environments today. With VTLs, businesses can backup data from local or remote sites to a back-up server where the data is persisted on disk and managed as virtual tape cartridges. Frequently, VTL solutions allow for optimization and size reduction of the backup data. Once data is sent to the VTL appliance across the network, or in some cases even prior to moving the data, the data to be stored is run through deduplication and compression techniques. These techniques can greatly reduce the disk space that is needed to store the backup or reduce the backup data size if it is transported across a WAN to a data center hub.

Amazon Services for Backup and Recovery

AWS offers a number of services that are great for building backup and recovery architectures for both hybrid and fully hosted solutions. Storage services on AWS include block storage devices in the form of Amazon Elastic Block Store (Amazon EBS), Amazon Simple Storage Service (Amazon S3), and Amazon Glacier. For tight integration with on-premises solutions, AWS also offers a storage gateway service called AWS Storage Gateway available as an on-premises appliance or on the Amazon Elastic Compute Cloud (Amazon EC2), as well as a rich set of third-party solutions available in the AWS Marketplace. For effective movement of data between an on-premises

environment and the cloud, AWS provides a variety of options that enable you to connect your data centers to the AWS cloud with AWS VPN CloudHub and AWS Direct Connect. The following sections summarize the benefits of these services for backup and recovery.

Amazon Elastic Compute Cloud (Amazon EC2)

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web scale computing easier for developers and system administrators. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use. Amazon EC2 provides developers and system administrators the tools to build failure-resilient applications and isolate themselves from common failure scenarios. For more information, see Amazon EC2.

Amazon Simple Storage Service (Amazon S3)

Amazon Simple Storage Service (Amazon S3) provides highly secure, scalable object storage.

You can use the Amazon S3 console to store and retrieve any amount of data, at any time, from anywhere on the web. You can use Amazon S3 alone or together with Amazon EC2, Amazon Elastic Block Storage (Amazon EBS), Amazon Glacier, and third-party storage repositories and gateways to provide cost-effective object storage for a wide variety of use cases. Amazon EBS allows you to create storage volumes and attach them to Amazon EC2 instances. Amazon S3 is particularly well suited for backup data due to its high degree of design durability (it is designed to provide 99.999999999% durability) and low cost.

Amazon S3 stores data as objects within resources called *buckets*. AWS Storage Gateway and many third-party backup solutions can manage Amazon S3 objects on your behalf. You can store as many objects as you want within a bucket, and you can write, read, and delete objects in your bucket. Single objects can be up to 5 TB in size. For more information, see Amazon S3.

Amazon Glacier

Amazon Glacier is an extremely low-cost, cloud archive storage service that provides secure and durable storage for data archiving and online backup. To keep costs low, Amazon Glacier is optimized for data that is frequently accessed and for which retrieval times of several hours are acceptable. With Amazon Glacier, you can reliably store large or small amounts of data for as little as $0.01 per gigabyte per month, a significant savings compared to on-premises solutions. Amazon Glacier is well suited for storage of backup data with long or indefinite retention requirements and for long-term data archiving.

By enabling you to both scale on-demand and pay only for the capacity you use, both Amazon S3 and Amazon Glazier remove the need for backup storage capacity planning. For more information, see Amazon Glacier.

Amazon Elastic Block Store (Amazon EBS)

Amazon EBS provides persistent block level storage volumes for use with Amazon EC2 instances in the AWS cloud. Each Amazon EBS volume is automatically replicated within its Availability Zone to protect you from component failure, offering high availability and durability. Amazon EBS volumes offer the consistent and low-latency performance that is needed to store and retrieve backups. With Amazon EBS, you can scale your usage up or down within minutes, all while paying a low price only for what you use. Storage gateways frequently use Amazon EBS volumes to persist backup data in the AWS cloud. For more information, see Amazon EBS.

AWS Storage Gateway

AWS Storage Gateway connects an on-premises software appliance with cloud-based storage to provide seamless and highly secure integration between your on-premises IT environment and the AWS storage infrastructure. For more information, see AWS Storage Gateway.

AWS Direct Connect

AWS Direct Connect makes it easy to establish a dedicated network connection from your on-premises environment to AWS. Using AWS Direct Connect, you can establish private connectivity between AWS and your data center, office, or co-location environment, which in many cases can reduce your network costs, increase bandwidth throughput, and provide a more consistent network experience than Internet-based connections. For more information, see AWS Direct Connect.

AWS Import/Export

AWS Import/Export accelerates moving large amounts of data into and out of the AWS cloud by using portable storage devices for transport. AWS Import/Export transfers your data directly onto and off of storage devices using the AWS high-speed internal network and bypassing the Internet. For significant data sets, AWS Import/Export is often faster than Internet transfer and more cost effective than upgrading your connectivity. For more information, see AWS Import/Export.

General Hybrid Backup and Recovery Approaches

To back up your on-premises data to the AWS cloud, you can choose between two common approaches:

- Write backup data directly to Amazon S3 by making API calls to the AWS platform, and by placing or retrieving backup data through secure HTTP PUT and GET requests directly across the Internet. Here, the endpoint itself makes a direct connection with Amazon S3 to write data and retrieve data.
- Write backup data to an intermediate device known as a storage gateway, and then let the storage gateway device move the data to the AWS cloud. For more information about storage gateways, see Role of Storage Gateways in this whitepaper. Solutions that facilitate gateway technology are typically hybrid architecture solutions with a portion of the solution residing on-premises as well as within the AWS ecosystem.

Role of Storage Gateways

The approach you take to writing backup data to the cloud and the way you retrieve backup data from the cloud can have both performance as well as cost implications. For example, using storage gateway technology requires on-premises hardware, while writing data directly to Amazon S3 might require more Internet bandwidth. How can storage gateways help to create a better on-premises backup and recovery experience for your users?

Storage gateways link on-premises environments to the cloud. Gateways are hardware or software appliances that act as an intermediary between a physical location within your organization and the remote cloud storage in the AWS cloud. You can deploy a single storage gateway with your IT infrastructure or deploy multiple gateways at multiple locations.

To reduce the need for high bandwidth network connections to and from your locations, it's best to compress and deduplicate your data in your on-premises environment before you move it across a wide area network (WAN). Backup products and storage gateway appliances typically perform data compression and, in many cases, data deduplication as built-in features. Compression is the encoding of data that results in the consumption of fewer bits as compared to uncompressed data. Data compression is typically expressed in the ratio between compressed and uncompressed data. Deduplication is a way to further compress data by removing duplicate copies of repeat data. Applications and appliances capable of data deduplication compare patterns in the data and replace duplicate occurrences of data with reference pointers that consume significantly less storage space. Storage gateways are particularly good at these tasks.

You can use AWS Storage Gateway to perform all of the tasks we just discussed. Further, you can take advantage of a broad range of storage gateway technologies (including both hardware and software appliances) that partners within the Amazon Partner Network (APN) provide in the AWS Marketplace. For example, NetApp and CTERA both have gateway technologies that provide a way to connect your on-premises environment to the AWS cloud.

For the remainder of this whitepaper, we focus on AWS Storage Gateway to illustrate the use of gateway technology for designing hybrid storage architectures.

What are the benefits of using AWS Storage Gateway?

- **Highly Secure** — AWS Storage Gateway encrypts data when it's uploaded and downloaded through SSL, and encrypts it at rest in Amazon S3 using AES 256.
- **Durably Backed by Amazon S3** — AWS Storage Gateway stores your data as EBS snapshots in Amazon S3. Amazon S3 is designed to sustain the concurrent loss of data in two facilities, redundantly storing your data on multiple devices across multiple data centers in a region.
- **Compatible** — AWS Storage Gateway presents volumes in industry standard formats (such as iSCSI) or as virtual tape libraries (VTLs). The iSCSI interface means there's no need to re-architect your on-premises applications.
- **Cost-Effective** — You pay only for what you use.
- **Integrates with Amazon EC2** — AWS Storage Gateway allows you to easily mirror data from your on-premises applications to applications running on Amazon EC2.
- **Network Efficient** — AWS Storage Gateway uploads only changed data, and compresses all data on upload and download.

AWS Storage Gateway Configuration Options

AWS Storage Gateway supports three configurations:

- **Gateway-Cached Volumes** — You can store your primary data in Amazon S3, and retain your frequently accessed data locally. Gateway-cached volumes provide substantial cost savings on primary storage, minimize the need to scale your storage on-premises, and retain low-latency access to your frequently accessed data.
- **Gateway-Stored Volumes** — In the event that you need low-latency access to your entire data set, you can configure your on-premises data gateway to store your primary data locally, and asynchronously back up point-in-time snapshots of this data to Amazon S3.
- **Gateway-Virtual Tape Library (Gateway-VTL)** — With gateway-VTL, you can have a limitless collection of virtual tapes. Each virtual tape can be stored in a virtual tape library backed by Amazon S3 or a virtual tape shelf backed by Amazon Glacier.

AWS Storage Gateway Deployment

The following diagram shows a typical deployment of the AWS Storage Gateway when used in gateway-stored volume mode.

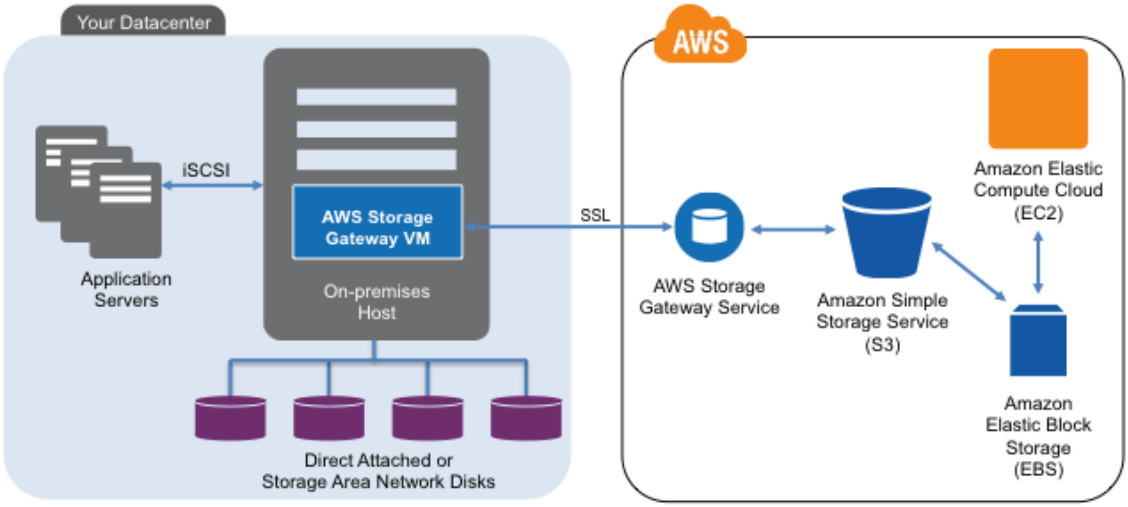

Figure 1: AWS Storage Gateway Deployment in Gateway-Volume Mode

The diagram shows AWS Storage Gateway deployed in an on-premises environment and running as a virtual appliance on a local host system. In this scenario, you can decide how much direct attached storage you need for AWS Storage Gateway, based on how you want to use the gateway. Communication with AWS Storage Gateway is through a secure socket layer (SSL) connection. Data is stored in Amazon S3 and can be recovered to the system at your data center. Data can also be recovered to Amazon EBS devices, which then can be attached to Amazon EC2 instances to, for example, implement a disaster recovery strategy.

AWS Storage Gateway and other storage gateway technologies can also function as virtual tape libraries. With gateway-VTL, you can have a limitless collection of virtual tapes. Each virtual tape can be stored in a virtual tape library that is backed by Amazon S3, or a virtual tape shelf that is backed by Amazon Glacier. The virtual tape library exposes an industry standard iSCSI interface, which gives your backup application on-line access to the virtual tapes. When you no longer require immediate or frequent access to data contained on a virtual tape, you can use your backup application to move it from its virtual tape library to your virtual tape shelf in order to further reduce your storage costs. One key advantage is that you can continue to use backup and recovery software applications that you already have deployed within your IT infrastructure, such as Veeam, Backup Exec, TSM, and Robocopy, which can directly communicate with industry-standard, iSCSI-compatible virtual tape libraries.

The following diagram shows a typical gateway topology when deployed in gateway-VTL mode.

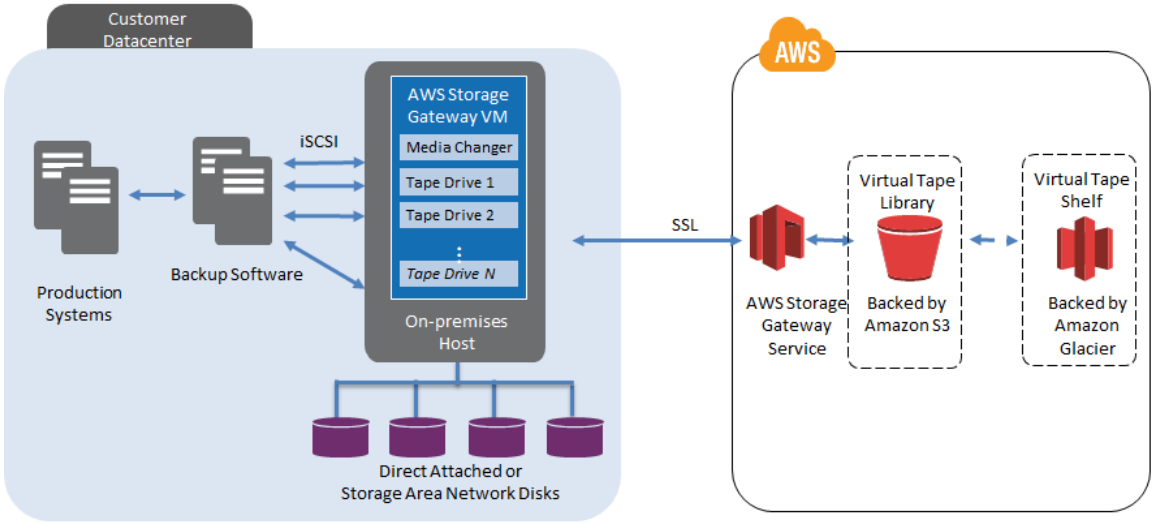

Figure 2: AWS Storage Gateway Deployed as Industry-Standard iSCSI VTL

This topology enables you to continue to store your backups on-premises for low-latency access while using Amazon S3 for your off-site backups. Thus, you can remove the complexity of dealing with off-site tape storage or VTL site-to-site replication.

You can download AWS Storage Gateway from AWS as a virtual machine (VM) and install it under the hypervisor of your choice: VMware ESXi or Microsoft Hyper-V. If you are deploying the gateway on-premises, you download and deploy the gateway virtual machine, and then activate the gateway. If you are deploying the gateway on an Amazon EC2 instance, you launch an Amazon Machine Image (AMI) that contains the gateway VM image, and then activate the gateway.

As a best practice, you should consider the amount of storage that you want to locally attach to a storage gateway in your on-premises environment. The following table summarizes key considerations for sizing local storage for the AWS Storage Gateway.

Gateway Storage Mode	Consideration
Gateway-Stored Volumes	In gateway-stored volume mode, snapshots of a complete volume are stored in the AWS cloud. Therefore, you need to attach the actual amount of desired disk space to the storage volume. For example, if you plan to serve 50 TB of iSCSI volumes from the storage gateway to systems that are on-premises, then you need to attach 50 TB of local storage to the gateway. In this mode, the primary copy of your data is stored locally on AWS Storage Gateway, and backups of the gateway volumes are stored on AWS.
Gateway-Cached Volumes	Gateway-cached volume mode greatly reduces the amount of local storage that needs to be attached to the storage gateway. In cached volume mode, the primary copy of your data is stored on the AWS cloud,

Gateway Storage Mode	Consideration
	and only cache data is stored on the locally attached drives. The amount of storage required locally on the storage gateway is driven by the amount of data that should be cached locally.

Backup and Recovery Architecture Best Practices

The following three architecture blueprints describe a selection of common backup and recovery problems. From an architecture perspective, a complete backup and recovery solution typically contains a backup software component installed on the endpoint that can be combined with an on-premises gateway solution.

Architecture Blueprint 1: On-Premises Backup with Third-Party Gateways

AWS Storage Gateway and similar products from APN partners provide storage gateway systems that integrate on-premises backup infrastructure with Amazon S3 and Amazon Glacier services in AWS. Conceptually, these gateways act as local caches for unlimited storage in AWS. In the context of backup and recovery, the primary goal of storage gateway deployments is to replace tape and disk storage systems that are traditionally used for backup systems. In turn, that reduces the complexity of the backup infrastructure and removes constraints in capacity.

Design Pattern

This architecture blueprint reviews a single deployment of a storage gateway system in a primary data center. Enterprise backup servers use the storage gateway as the primary storage for backup archives. The storage gateway provides a local disk cache for improved backup performance and asynchronously uploads all backup archives to Amazon S3. Depending on business need, backup data can also be periodically moved from Amazon S3 to Amazon Glacier for longer-term storage.

As discussed earlier, storage gateways come in the form of physical or virtual appliances. To integrate with the backup infrastructure, a storage gateway such as AWS Storage Gateway exposes interfaces that are compatible with backup software, such as an iSCSI target, CIFS/SMB or NFS share, or VTL emulation. The gateway interfaces with Amazon S3 or Amazon Glacier APIs over SSL-protected HTTPS connections.

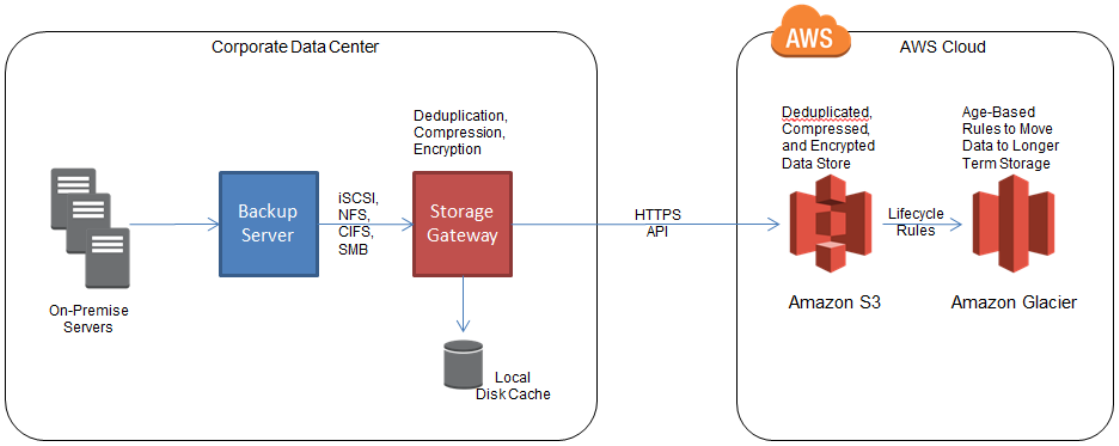

Figure 3: Storage Gateway Integration with AWS

Performance Considerations

You should consider two primary performance factors when evaluating a storage gateway solution:

- Throughput and volume of data transfer between the backup server and the storage gateway
- Ratio of data transfer volume to Internet bandwidth between the storage gateway and Amazon S3

While the RTO is affected by both of those factors, the RPO is primarily affected by the ratio of data volume to the available bandwidth capacity to reach Amazon S3.

Backup Server to Storage Gateway

The following table lists features that commonly affect the throughput between a backup server and a storage gateway.

Feature	Description
Local Disk Size	Larger local disk storage provides a larger cache, reducing the latency when storing or fetching data between a backup server and a storage gateway
Local Disk Performance	Faster disks on a storage gateway enable faster delivery of backup data from a backup server to a storage gateway
CPU and Memory	More powerful compute resources allow for faster software execution of features such as deduplication, compression, encryption, and WAN optimization

Feature	Description
LAN Speed	Faster LAN equipment and Ethernet interfaces provide higher throughput from a backup server to a storage gateway

Throughput between the backup server and the storage gateway determines the length of the backup window. It is primarily influenced by the hardware included with the storage gateway, for example, the size and the speed of the disk subsystem as well as the line speed of the LAN connection. In the case of virtual appliances, the throughput is dictated by resources assigned to the VM. The efficiency of the backup server software and the storage gateway software also plays a significant role.

Storage Gateway to AWS

The following table shows a list of factors that affect the performance between a storage gateway and Amazon S3.

Feature	Description
Available Bandwidth	A storage gateway utilizes HTTPS API calls over the public Internet or direct connect lines to store data in Amazon S3
Compression	Compression reduces the amount of data that needs to be stored in local storage and in Amazon S3, as well as reducing the volume of data transfer between the storage gateway and Amazon S3
Deduplication	Deduplication can provide a much higher level of data compression, but might require additional compute and memory resources on the storage gateway
WAN Optimization	WAN optimization uses additional techniques to reduce bandwidth use between the storage gateway and Amazon S3

Performance between the storage gateway and Amazon S3 is the primary factor for meeting the desired recovery point objective (RPO). This performance is influenced not only by the Internet connection speed and quality, but also by the use and efficiency of software features that the storage gateway system might include. The software features include deduplication, compression, and WAN optimization. Depending on the type of content that is being backed up, these techniques could reduce the amount of data transfer to Amazon S3 and Amazon Glacier storage by over 90%.

Example Scenario

You must weigh the combination of available Internet bandwidth, compression, deduplication, and WAN optimization efficiency to evaluate the viability of the overall backup system design. For example, let's consider this backup system design: after

deduplication and compression, the estimate is 50 GB of daily data transfer, while the Internet bandwidth is a single T-1 line and the required RPO is 24 hours. This solution will not be viable and would require a bandwidth upgrade to a minimum of 6 Mbps to ensure that daily backups are moved to Amazon S3 within a 24-hour window.

Security Considerations

The following table summarizes the most critical security features common to many storage gateways. Some APN partners might offer additional features.

Feature	Description
In-Transit Encryption	All data should be transferred to and from Amazon S3 through SSL-protected transmission
At-Rest Encryption	All data stored in the storage gateway, Amazon S3, and Amazon Glacier should be encrypted using strong encryption
Key Storage	The storage gateway might allow you to store encryption keys in a discrete, physical location
Key Rotation	The storage gateway should provide an option for periodic updates of encryption keys

Storage gateways interact with AWS over the public Internet and store data in Amazon S3 and Amazon Glacier. Most gateways support encryption of data at rest; typically, data is encrypted using the AES-128 or AES-256 algorithm prior to persisting the backup archive on the appliance and in Amazon S3 or Amazon Glacier. You can keep encryption keys in a separate physical location to provide additional security. Certain storage gateways from APN partners also offer advanced key rotation features.

Durability and Availability

By persisting data to Amazon S3 and Amazon Glacier, storage gateways gain high levels of durability. Both services redundantly store objects on multiple devices across multiple facilities in an Amazon S3 region and are designed to provide 99.999999999% durability of objects over a given year.

To improve availability of the infrastructure deployed in a data center, you can configure storage gateways in high-availability pairs, and you can configure connections from the data center to AWS over multiple redundant paths.

Amazon S3 also comes with an availability SLA. For more information about Amazon S3 availability, see Amazon S3 SLA.

Recovery Process

For the recovery process, you need the presence of a storage gateway in a Disaster Recovery (DR) location. You also need the appropriate Internet connectivity to download the compressed or deduplicated data, decrypt the data locally, and expose the data to a backup server. As a best practice, you should test your recovery procedures periodically.

In addition to using physical DR sites, you can use AWS as a potential DR site because your data is already stored in AWS. For this purpose, you can instantiate AWS Storage Gateway from an AMI directly into an Amazon Virtual Private Cloud (Amazon VPC). Many storage gateways that are provided by APN partners have prebuilt AMIs in AWS Marketplace that allow for fast and easy deployment of the storage gateway appliance in AWS.

Architecture Blueprint 2: Multi-Site Backup and Recovery with Gateways

This architectural blueprint expands on the preceding blueprint. It also addresses backup needs for multiple data centers and remote branch offices in addition to backup infrastructure in the main corporate data center.

Design Pattern

Each remote site utilizes a backup server and a storage gateway that uploads backup archives to Amazon S3. AWS becomes a single, centralized backup store that covers the backup needs of a distributed enterprise.

Storage gateways can be sized to match the needs of a particular site. For example, a 10-person remote sales office will likely need a smaller appliance than a major office with hundreds of employees. The global topology can accommodate a mix of different size storage gateways to achieve appropriate performance at the best cost possible.

The following diagram shows a multi-site architecture with storage gateways placed at each site for fast access to backup data. Each storage gateway stores and retrieves its backup sets from Amazon S3, where the data is encrypted, deduplicated, and compressed. Based on policies that you set and control, such as age-based lifecycle policies, backup sets can be moved in and out of Amazon Glacier for long-term retention at very low cost.

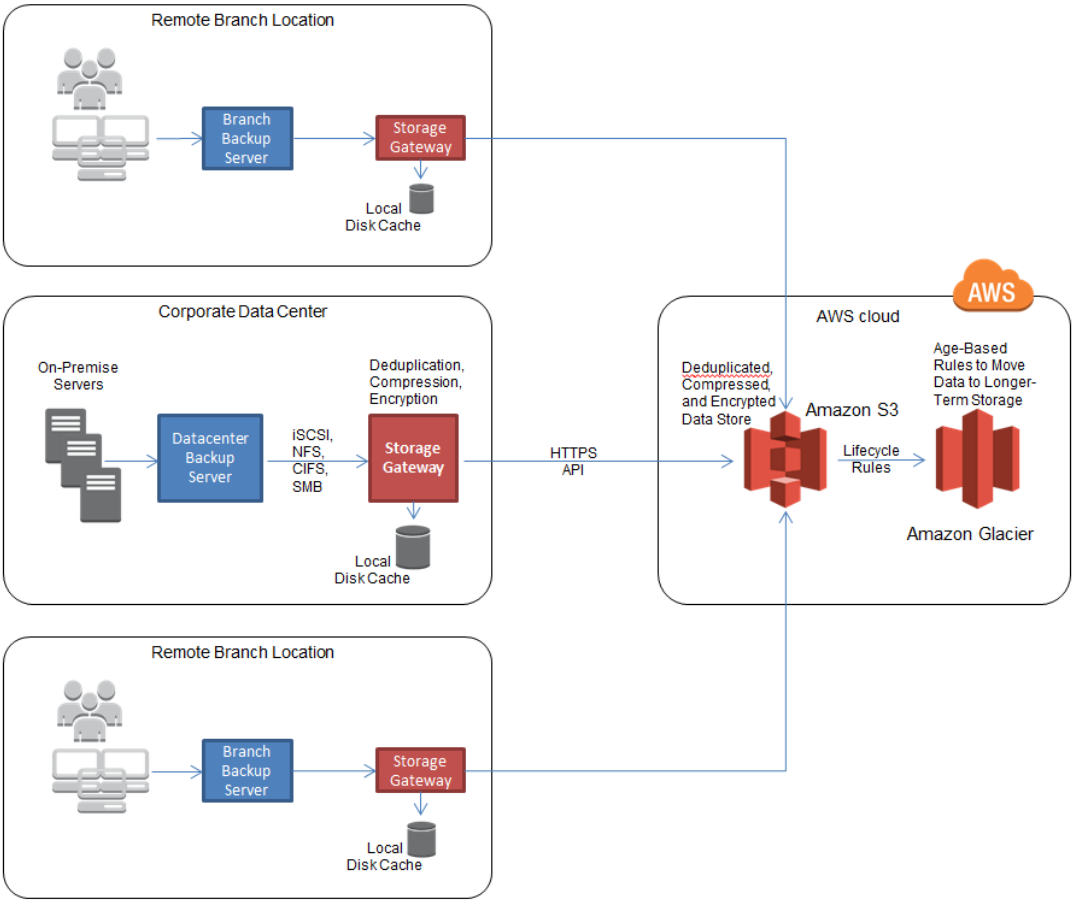

Figure 4: Storage Gateway Integration with AWS

Performance Considerations

Remote sites, such as remote branch offices, often have less available Internet bandwidth, especially when compared to corporate data centers. For these bandwidth-constrained remote sites, deduplication and compression efficiencies become even more important to meet the required RPO.

The backup window is affected by LAN speeds and the storage gateway resources. You should consider the maximum allowable backup window when sizing storage gateways for different remote sites.

Recovery Process

Recovery can occur on a dedicated DR site, inside AWS, or another remote site. Recovery to another remote site is an attractive option, as the necessary prerequisites for recovery are already in place: backup server, storage gateway, and Internet connectivity. In this scenario, you can begin restoration of critical files and applications immediately.

Durability and Availability

Remote sites often have reduced redundancy in terms of backup resources and environmental controls as compared to major corporate data centers. The backup to AWS services increases the durability of backup data and archives, allowing for that data to be mirrored on multiple devices across multiple facilities in an AWS region.

Security Considerations

Physical security is often a major concern in remote sites, particularly in smaller remote branch offices. To mitigate physical security concerns, most storage gateways include options to encrypt the data before it is saved to a local disk cache and persisted to Amazon S3. Many storage gateway products include additional features to manage, rotate, and protect encryption keys.

Architecture Blueprint 3: Direct Endpoint Backup and Recovery

The widespread use and adoption of endpoint devices—laptops, tablets, and smartphones—presents another challenge to IT: how to protect, backup from, and recover to devices within the remote workforce?

In a common scenario, on-premises users create content with their laptops, desktops, and even their tablets or smartphones. That content can then be transparently backed up to the cloud through a storage gateway, which could encrypt, deduplicate, or protect the data with security policies and access control lists. Once stored, files can be shared and collaborated upon throughout the enterprise.

However, remote users typically will not have a gateway through which to protect, backup, share, or recover data. The following sections describe how to create a solution for these users.

Endpoint Backup

Off-premises users and enterprises alike can receive the same benefits that storage gateways provide by using endpoint protection that delivers automated and transparent encryption, data deduplication, bandwidth throttling, and scheduled synchronization.

Sophisticated systems will even allow search, download, and recovery capabilities, as well as providing sharing capabilities for large files.

As shown in the following diagram, your data can be delivered to the AWS cloud with a simple software solution that has the ability to interact with or without a VPN.

Figure 5: Endpoint Protection Used With and Without a VPN

Data that is backed up to Amazon S3 can be seamlessly moved to Amazon Glacier—an extremely low-cost cloud archive storage service—by using data lifecycle policies to archive your data after specified periods of time.

Endpoint Recovery

Recovery to endpoint devices can range from simple file recovery to application-aware, point-in-time recovery to complete deployment of file systems to the remote endpoint. Snapshots of local storage can be regularly scheduled, and endpoint recovery systems can be used for automating snapshot management so that recovery from snapshots are efficient and quickly carried out. We recommend taking these endpoint devices into consideration when planning your disaster recovery architecture and business continuity plan.

In many cases, where virtual desktop infrastructures (VDIs) are employed to give remote users access to desktop environments on remote servers, state files—like configuration, preferences, or other local storage—must be backed up and recovered in order to achieve the lowest recovery-time objective (RTO).

Endpoint Protection

In addition to data backup and recovery, we recommend endpoint protection as a further consideration in your storage and backup plans. Endpoint protection can prevent unauthorized access to data by using various data loss prevention techniques:

- Endpoint devices can be remotely wiped by administrators
- Administrators can remotely wipe or encrypt only confidential or sensitive data, leaving the operating system or other files in place
- Geo-location services can be enabled to help track down lost or stolen devices, which can be essential to protecting data leakage and corporate information

By utilizing these best practices for endpoint backup, recovery, and protection, you can create a data protection strategy that will both minimize enterprise IT costs and maximize productivity.

Cloud-Based Backup Models

Switching to cloud-based backups can free you from tape backup constraints, such as limited storage and media reliability, but it also introduces new aspects for your backup and recovery strategy. For example, instead of physical security for tape vaults, encryption management becomes a key to your data security. Instead of depending on the tape library hardware reliability, your Internet bandwidth and availability become primary factors for backup performance and reliability.

Considerations

The following list shows design considerations for cloud-based backup and recovery solutions on AWS that you should incorporate into your backup and recovery solution planning.

- **Data Classification** — Does your organization have a data classification policy that outlines what type of data can be stored in the cloud? For example, do you have data in your backup sets that requires specific levels of encryption, or do you have data that must remain within a specific geographic region? Consider using specific AWS regions to write your backup data to. Consider how you manage data encryption. Decide which data sets cannot be stored in the cloud based on your data classification.
- **Encryption Key Management** — Can data be encrypted at rest with keys generated by AWS? Do you have a requirement to encrypt your data using your own keys? If you need to manage your own keys, consider using a key management solution such as AWS CloudHSM or AWS Key Management Service, or leverage your current on-premises key management. Use key management built into AWS Storage Gateway to encrypt your backup data.
- **Network Bandwidth** — Determine the type of network connection needed to meet your backup and recovery requirements. Consider going directly across the Internet with HTTPS for endpoint backups that can be taken from anywhere in the world and recovered to any device anywhere. For on-premises to AWS backups using gateway technology, consider using a VPN connection or AWS Direct Connect. Remote branch offices are frequently served well by VPN connections, while medium-to-large data centers greatly benefit from using AWS Direct Connect.
- **Backup and Recovery Methodology** — Will you be backing up snapshots of complete systems or volumes attached to systems? For example, VMware guests are frequently backed up by taking snapshots of data stores. The snapshots can be moved to the AWS cloud simply by placing them onto an AWS Storage Gateway volume. If you use backup software for file-level backup of your systems, then consider using cached volumes on gateways or solutions that provide a unified namespace, such as the CTERA cloud service delivery platform. When possible, consider data replication over traditional backup and recovery methods. For

example, replicate on-premises databases to a pilot-light database on AWS as your main method for backup and disaster recovery. For more details about how to use AWS for disaster recovery, see the Using Amazon Web Services for Disaster Recovery whitepaper.

- **Tape Replacement** — AWS is a great repository for tape replacement. Amazon Storage Gateway has virtual tape library (VTL) functionality built into the ready-to-deploy software appliance. Consider this technology if you need to retain the concepts of tapes within the organization or have existing software products that will interact with an iSCSI-compatible VTL. AWS Storage Gateway technology with VTL functionality will enable you to replace your tape vault with the cloud.
- **Availability of Backup Data Set** — Consider how quickly you might need to access a specific backup data set. For example, do you need to keep certain backup sets on local disk attached to the storage gateway and ready for immediate recovery at local area network speeds? If so, consider keeping the most recent backup data in the cache with an on-premises storage gateway. If it's acceptable to wait a few hours for a backup data set to become available, consider storing backups directly in Amazon Glacier.
- **Accessibility** — Storing your backup data in the AWS cloud can make it accessible from any location with Internet access. Consider where data will be most likely be accessed from, and select the AWS region as your backup location that is closest to the potential recovery location. You must further consider regulatory requirements that might require data to be stored in specific geographies.
- **Import of Existing Backup Data** — AWS provides a cost-effective way of importing large amounts of data from physical media. The AWS Import/Export service accepts physical media in the form of USB drives. The AWS Import/Export processes create a manifest file that describes the media to be received and where to recover the data to. The instructions provided with AWS Import/Export service give detailed instructions about how to ship USB drive to an AWS facility for import. Consider moving large amounts of existing backup data by writing the data first to USB drives on-premises, shipping the drives to AWS, and using AWS Import/Export to load the data in a quick and economic way to your target Amazon S3 bucket or Amazon EBS volumes.

Backup vs. Replication for Recovery

Traditional backup and recovery architectures have been part of the IT disaster recovery strategy for a very long time, and are certainly a proven and well-tested approach to securing data. However, as increasing network bandwidth at affordable cost becomes more available, enterprises are rethinking their backup and recovery strategies, including disaster recovery (DR) strategies. The predominant shift is away from typical backup and recover configurations towards architectures that favor replication of data over backup and recovery, especially for DR.

For your DR strategy, we recommend the following approaches:

- Traditional backup and recovery
- Pilot light on AWS
- Warm standby on AWS
- Multi-Site on AWS

You can use variations and combinations of all four of these models to achieve the right RTO/RPO for your organization. For details about how to use AWS for disaster recovery, see the Using Amazon Web Services for Disaster Recovery whitepaper.

Conclusion

Amazon Web Services provides you with a number of different approaches that enable cloud-based backup and recovery. For example, you can back up your data directly to Amazon S3 and recover the data to an endpoint that is either on-premises or in the cloud. As another option, you can completely replace your on-premises backup and recovery solutions by using storage gateway technologies, such as AWS Storage Gateway and cloud-ready backup software. You can significantly lower the TCO of your backup solution by using low-cost Amazon S3 and Amazon Glacier object storage for long-term retention of your backup data sets.

Further Reading

For additional help, consult the following sources:

Amazon S3 Getting Started Guide:
http://docs.amazonwebservices.com/AmazonS3/latest/gsg/

Amazon EC2 Getting Started Guide:
http://docs.amazonwebservices.com/AWSEC2/latest/GettingStartedGuide/

AWS Partner Directory (for a list of AWS solution providers):
http://aws.amazon.com/solutions/solution/providers/

AWS Security and Compliance Center:
http://aws.amazon.com/security/

AWS Architecture Center:
http://aws.amazon.com/architecture

Using AWS for Disaster Recovery:
http://media.amazonwebservices.com/AWS_Disaster_Recovery.pdf

Notices

Appendix: On-Premises to AWS via AWS Direct Connect

Backup and recovery traffic can be both large in volume as well as time-sensitive. For example, backups need to complete in a predictable amount of time to be relevant for a given business solution or to meet the RPO of a DR solution. A backup and recovery architecture that stores backup data in the AWS cloud needs to give special consideration to network limitations introduced by a wide area link or by moving data directly across the Internet. One way to closely control the network between your on-premises systems and AWS is by the use of AWS Direct Connect.

You can use AWS Direct Connect to tightly control network integration between your on-premises environment and the AWS cloud. As a best practice, you should determine your baseline data-transfer needs that are specific to your on-premises backup and recovery workload. An Amazon Solutions Architect can help you find the right connectivity method, and can also help you determine network bandwidth needs for your WAN connection.

AWS Direct Connect links your internal network to an AWS Direct Connect location over a standard 1 GBit or 10 GBit Ethernet fiber-optic cable. One end of the cable is connected to your router, and the other is connected to an AWS Direct Connect router. With this connection in place, you can create virtual interfaces directly to the Amazon VPC, bypassing Internet service providers in your network path. For time-sensitive backups and high volumes of data, AWS Direct Connect enables you to attain a high

degree of predictability for your cloud-based backup and recovery solution through a dedicated capacity connection. The following diagram shows a typical AWS Direct Connect topology.

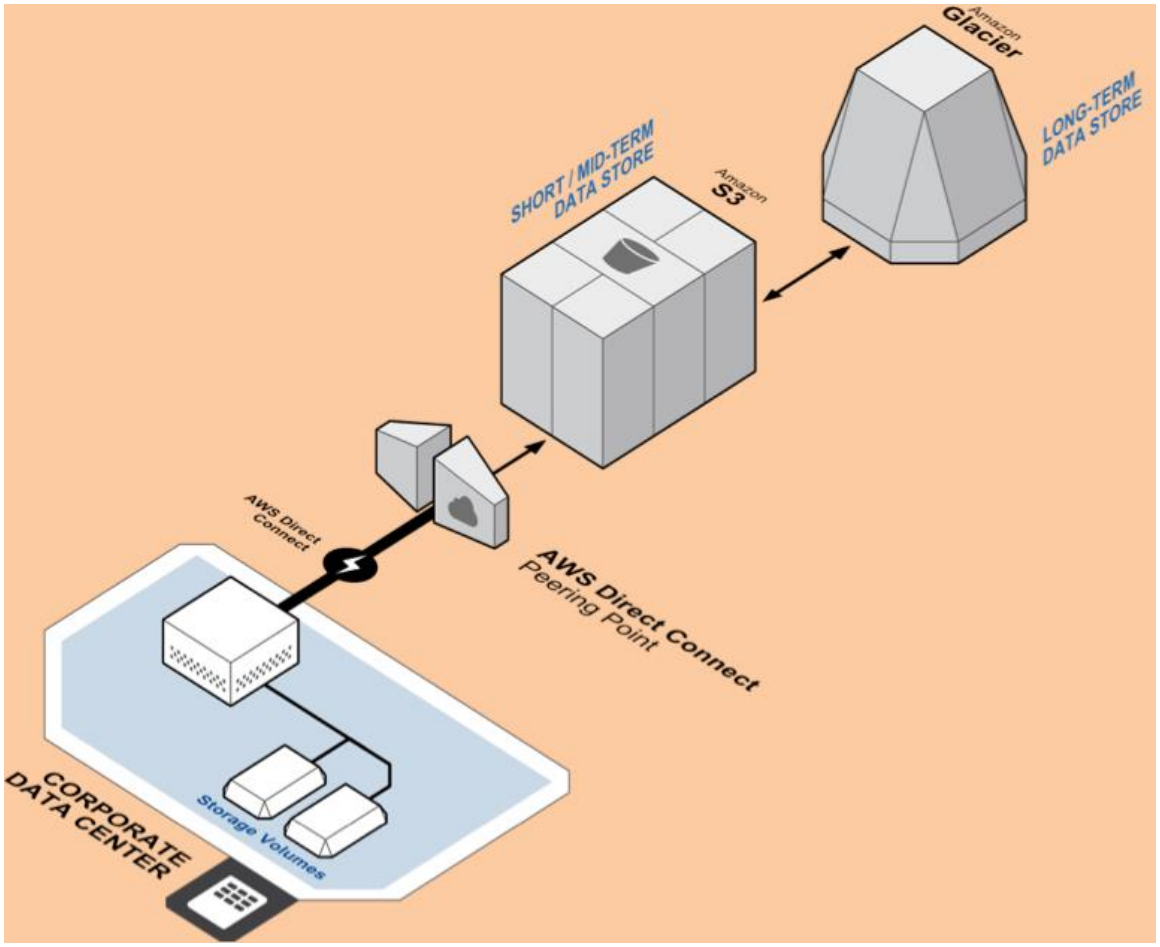

Figure 6: AWS Direct Connect for Backup and Recovery

The diagram shows an on-premises data center with a dedicated capacity AWS Direct Connect link to a peering point in an APN partner location. At the peering point, either a 1 GBit or 10 GBit link is made to the AWS cloud, which then enables fast access to Amazon S3 and Amazon Glacier.

AWS has been growing the number of peering points ever since the inception of AWS Direct Connect. The following table shows the current peering locations that are provided by APN partners.

AWS Direct Connect Location	AWS Region
CoreSite NY1 & NY2	US East (Virginia)

AWS Direct Connect Location	AWS Region
CoreSite One Wilshire & 900 North Alameda, CA	US West (Northern California)
Equinix DC1 – DC6 & DC10	US East (Virginia)
Equinix SV1 & SV5	US West (Northern California)
Equinix SE2 & SE3	US West (Oregon)
Eircom, Clonshaugh	EU West (Ireland)
TelcityGroup, London Docklands	EU West (Ireland)

We recommend that you use AWS Direct Connect to transfer large backup and recovery data sets. However, to control cost we also recommend that you limit the Internet bandwidth used by your backups to only what is necessary to meet your backup window requirements and, for recovery, to meet your RTO/RPO requirements. By doing so, you can reduce network fees that you pay to your Internet Service Provider (ISP), and avoid paying for increased Internet bandwidth commitments or new contracts. In addition, all data transferred over AWS Direct Connect is charged at the reduced AWS Direct Connect data transfer rate rather than Internet data transfer rates, which can greatly reduce your network costs.

Document Revisions

12/12/2014 – Initial Document