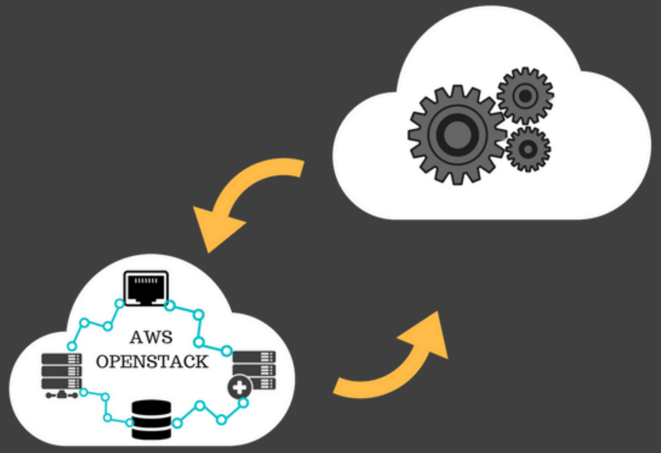
Course and Labs

AWS Design and Automation



Module 3: Design Storage in AWS

Topic 1: Simple Storage Service(s3)

Mohanraj Shanmugam

Course and Labs

Design Storage in AWS

Mohanraj Shanmugam

Traditional Storage Options

- Storage area network (SAN)—Block devices (virtual disk LUNs) on dedicated SANs often provide the highest level of disk performance and durability for both business-critical file data and database storage.
- Direct-attached storage (DAS)—Local hard disk drives or arrays residing in each server provide higher performance than a SAN, but lower durability for temporary and persistent files, database storage, and operating system (OS) boot storage than a SAN.
- Network attached storage (NAS)—NAS storage provides a file-level interface to storage that can be shared across multiple systems. NAS tends to be slower than either SAN or DAS.
- Backup and Archive—Data retained for backup and archival purposes is typically stored on non-disk media

Factors to Choose the Storage

- The below are the factors influence the choosing a Storage option for your application
  - Performance - How fast I can Read and Write
  - Durability – Chances of my data getting lost
  - Availability – How long it is available for me to use
  - Cost – How much cost per GB I pay for
  - Interface – What the interfaces to access the Storage (Via OS, Web or Client)
  - Scalability – How big storage I can add
  - Elasticity – Increase and Decrease when ever I want
  - Persistence – Permanent storage
  - Archival – How long I can store the data whit lowest price
  - Archival Recoverability – Recover when even I want at shortest period possible

AWS Storage Offerings

- Amazon S3 - Scalable and Durable storage in the cloud
- Amazon Glacier - Low-cost archive storage in the cloud
- Amazon EBS - Persistent block storage volumes for Amazon EC2 virtual machines
- Amazon EC2 Instance Storage - Temporary block storage volumes for Amazon EC2 virtual machines
- AWS Import/Export - Large volume data transfer
- AWS Storage Gateway - Integrates on-premises IT environments with cloud storage

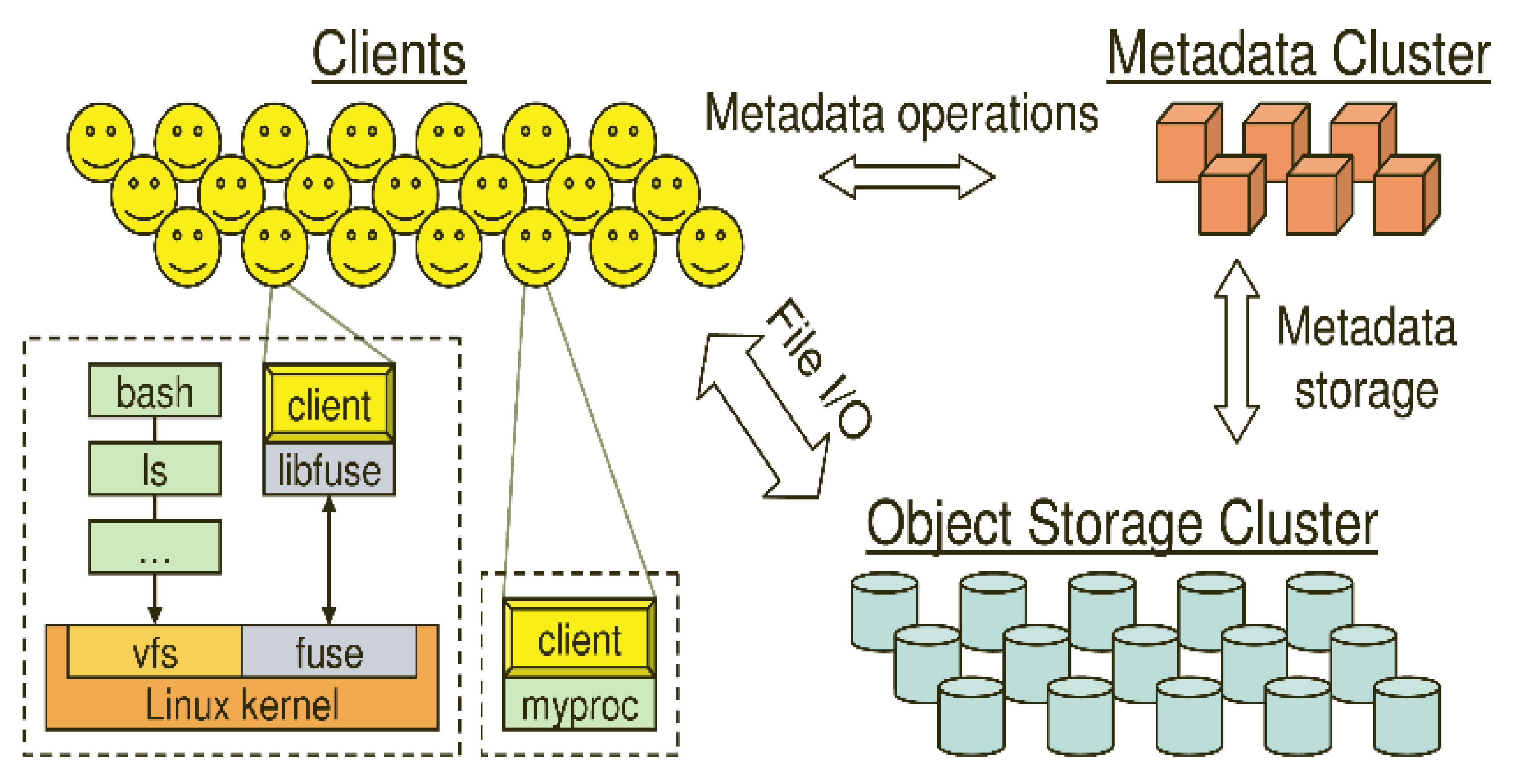
Amazon Simple Storage Service (Amazon S3)

Amazon S3

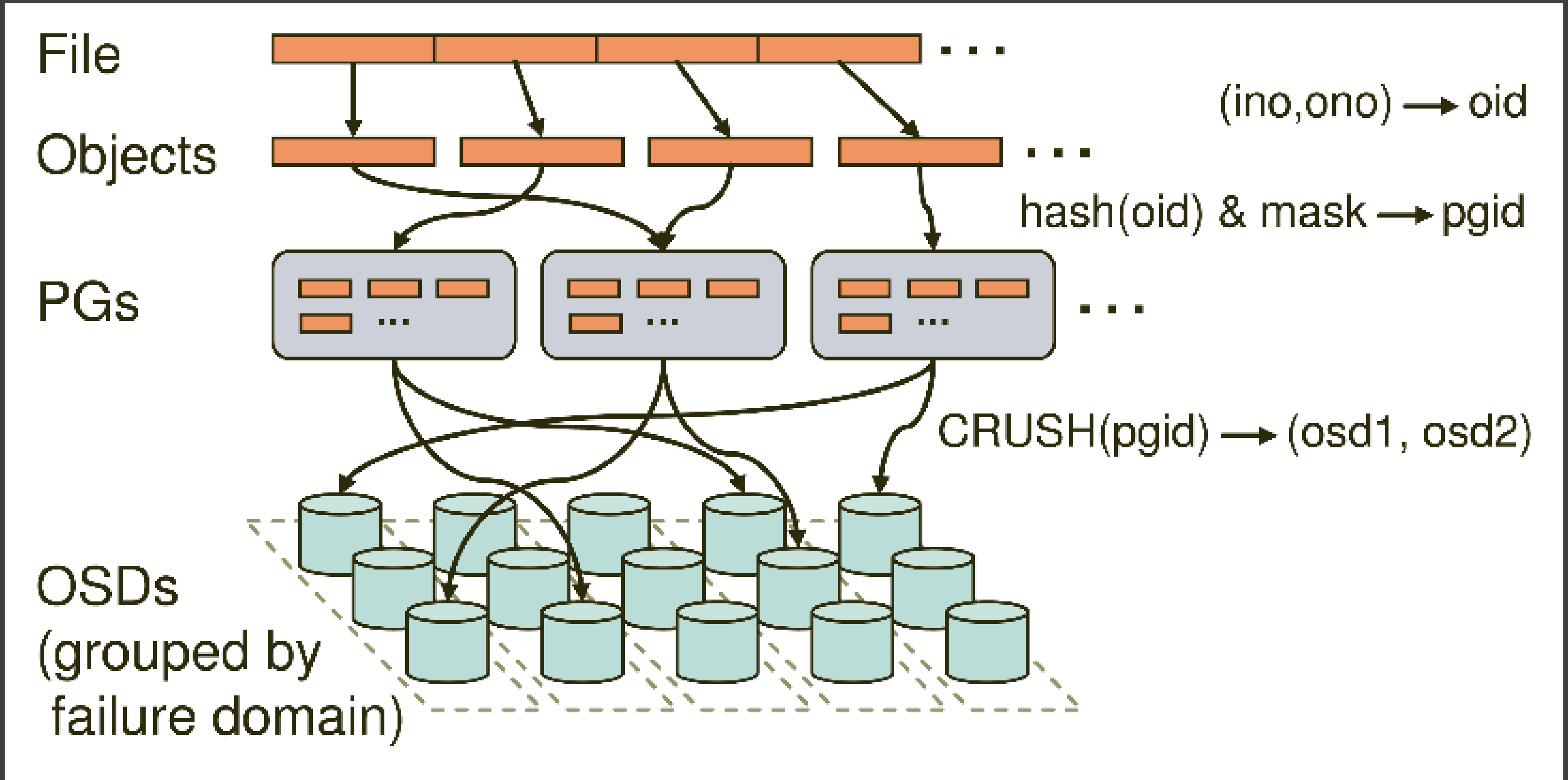
- Amazon Simple Storage Service (Amazon S3) is a object storage for the Internet.
- you can use Amazon S3 to store and retrieve any amount of data at any time, from anywhere on the web.
- You can accomplish these tasks using the simple and intuitive web interface of the AWS Management Console.

Object Storage

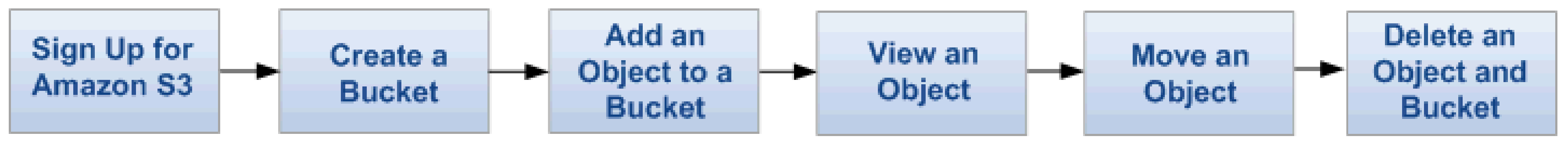
Course and Labs



Object Storage



S3 Lifecycle



- To use Amazon S3, you need an AWS account. If you don't already have one, you'll be prompted to create one when you sign up for Amazon S3. You will not be charged for Amazon S3 until you use it. **To sign up for Amazon S3** Go to http://aws.amazon.com/s3
- Every object in Amazon S3 is stored in a bucket. Before you can store data in Amazon S3, you must create a bucket.
- you are not charged for creating a bucket; you are charged only for storing objects in the bucket and for transferring objects in and out of the bucket.

S3 Storage Classes

- Amazon S3 Standard – General Purpose
  - Amazon S3 Standard offers high durability, availability, and performance object storage for frequently accessed data.
  - Because it delivers low latency and high throughput, Standard is perfect for a wide variety of use cases including cloud applications, dynamic websites, content distribution, mobile and gaming applications, and big data analytics.
  - Lifecycle management offers configurable policies to automatically migrate objects to the most appropriate storage class.
- Key Features:
  - Low latency and high throughput performance
  - Designed for durability of 99.999999999% of objects
  - Designed for 99.99% availability over a given year
  - Backed with the Amazon S3 Service Level Agreement for availability.
  - Supports SSL encryption of data in transit and at rest
  - Lifecycle management for automatic migration of objects

S3 Storage Classes

- Amazon S3 Standard - Infrequent Access
  - Amazon S3 Standard - Infrequent Access (Standard - IA) is an Amazon S3 storage class for data that is accessed less frequently, but requires rapid access when needed.
  - Standard - IA offers the high durability, throughput, and low latency of Amazon S3 Standard, with a low per GB storage price and per GB retrieval fee
  - This combination of low cost and high performance make Standard - IA ideal for long-term storage, backups, and as a data store for disaster recovery.
  - The Standard - IA storage class is set at the object level and can exist in the same bucket as Standard, allowing you to use lifecycle policies to automatically transition objects between storage classes without any application changes.

S3 Storage Classes

- Amazon Glacier
  - Amazon Glacier is a secure, durable, and extremely low-cost storage service for data archiving.
  - You can reliably store any amount of data at costs that are competitive with or cheaper than on-premises solutions. To keep costs low, Amazon Glacier is optimized for data that is rarely accessed and a retrieval time of several hours is suitable.
  - Amazon Glacier supports lifecycle policies for automatic migration between storage classes. Please see the Amazon Glacier page for more details.
- Key Features:
  - Designed for durability of 99.999999999% of objects
  - Supports SSL encryption of data in transit and at rest
  - Vault Lock feature enforces compliance via a lockable policy
  - Extremely low cost design is ideal for long-term archive
  - Lifecycle management for automatic migration of objects

Course and Labs

Compare Storage Classes

	Standard	Standard - IA	Amazon Glacier
Designed for Durability	99.999999999 %	99.999999999%	99.999999999%
Designed for Availability	99.99%	99.9%	N/A
Availability SLA	99.9%	99%	N/A
Minimum Object Size	N/A	128KB*	N/A
Minimum Storage Duration	N/A	30 days	90 days
Retrieval Fee	N/A	per GB retrieved	per GB retrieved**
First Byte Latency	milliseconds	milliseconds	4 hours
Storage Class	object level	object level	object level
Lifecycle Transitions	yes	yes	yes

Amazon S3 Pricing

	Standard Storage	Standard - Infrequent Access Storage †	Glacier Storage
First 1 TB / month	$0.0300 per GB	$0.0125 per GB	$0.007 per GB
Next 49 TB / month	$0.0295 per GB	$0.0125 per GB	$0.007 per GB
Next 450 TB / month	$0.0290 per GB	$0.0125 per GB	$0.007 per GB
Next 500 TB / month	$0.0285 per GB	$0.0125 per GB	$0.007 per GB
Next 4000 TB / month	$0.0280 per GB	$0.0125 per GB	$0.007 per GB
Over 5000 TB / month	$0.0275 per GB	$0.0125 per GB	$0.007 per GB

Amazon S3 Request Pricing

	Pricing
For Requests Not Otherwise Specified Below	
PUT, COPY, POST, or LIST Requests	$0.005 per 1,000 requests
GET and all other Requests	$0.004 per 10,000 requests
Delete Requests	Free †
For Standard – Infrequent Access Requests	
PUT, COPY, or POST Requests	$0.01 per 1,000 requests
GET and all other Requests	$0.01 per 10,000 requests
Lifecycle Transition Requests into Standard – Infrequent Access	$0.01 per 1,000 requests
Data Retrievals	$0.01 per GB
For Glacier Requests	
Glacier Archive and Restore Requests	$0.05 per 1,000 requests
Glacier Data Restores	Free ‡

Amazon S3 Data Transfer Pricing

Pricing	
Data Transfer IN To Amazon S3	
All data transfer in	$0.000 per GB
Data Transfer OUT From Amazon S3 To	
Amazon EC2 in the Northern Virginia Region	$0.000 per GB
Another AWS Region	$0.020 per GB
Amazon CloudFront	$0.000 per GB
Data Transfer OUT From Amazon S3 To Internet	
First 1 GB / month	$0.000 per GB
Up to 10 TB / month	$0.090 per GB
Next 40 TB / month	$0.085 per GB
Next 100 TB / month	$0.070 per GB
Next 350 TB / month	$0.050 per GB
Next 524 TB / month	Contact Us
Next 4 PB / month	Contact Us
Greater than 5 PB / month	Contact Us

Amazon S3 Reduced Redundancy Storage

- Reduced Redundancy Storage (RRS) is an Amazon S3 storage option that enables customers to reduce their costs by storing noncritical, reproducible data at lower levels of redundancy than Amazon S3's standard storage.
- It provides a cost-effective, highly available solution for distributing or sharing content that is durably stored elsewhere, or for storing thumbnails, transcoded media, or other processed data that can be easily reproduced.
- The RRS option stores objects on multiple devices across multiple facilities, providing 400 times the durability of a typical disk drive, but does not replicate objects as many times as standard Amazon S3 storage.
- Reduced Redundancy Storage is:
  - Backed with the Amazon S3 Service Level Agreement for availability.
  - Designed to provide 99.99% durability and 99.99% availability of objects over a given year. This durability level corresponds to an average annual expected loss of 0.01% of objects.
  - Designed to sustain the loss of data in a single facility.

Features of S3

- Cross-Region Replication
  - Cross-region replication (CRR) replicates every object uploaded to your source bucket to a destination bucket in a different AWS region that you choose.
  - The metadata and ACLs associated with the object are also part of the replication.
  - Once you configure CRR on your source bucket, any changes to the data, metadata, or ACLs on the object trigger a new replication to the destination bucket.
- Event Notifications
  - Amazon S3 event notifications can be sent when objects are uploaded to or deleted from Amazon S3.
  - Event notifications can be delivered using Amazon SQS or Amazon SNS, or sent directly to AWS Lambda, enabling you to trigger workflows, alerts, or other processing.
  - Event Trigger work flow are used in many use cases like whenever you upload a media file it should encode it in to multiple formats

Features of S3

- Versioning
  - Amazon S3 allows you to enable versioning so you can preserve, retrieve, and restore every version of every object stored in an Amazon S3 bucket.
  - This allows you to easily recover from both unintended user actions and application failures.
  - By default, requests will retrieve the most recently written version.
  - Older versions of an object can be retrieved by specifying a version in the request.
  - Storage rates apply for every version stored.
  - You can configure lifecycle rules to automatically control the lifetime and cost of storing multiple versions.

Features of S3

- Lifecycle Management
  - Amazon S3 can automatically assign and change cost and performance characteristics as your data evolves.
  - It can even automate common data lifecycle management tasks, including capacity provisioning, automatic migration to lower cost tiers, and regulatory compliance policies and eventual scheduled deletions.
  - Amazon S3 also enables you to automatically migrate your data to lower cost storage as your data ages.
  - You can define rules to automatically migrate Amazon S3 objects to Standard - Infrequent Access (Standard - IA) or Amazon Glacier based on the age of the data.
  - When your data reaches its end of life, Amazon S3 provides programmatic options for recurring and high volume deletions.

Features of S3

- Encryption
  - You can securely upload or download your data to Amazon S3 via the SSL-encrypted endpoints using the HTTPS protocol.
  - Amazon S3 can automatically encrypt your data at rest and gives you several choices for key management.
  - If you choose to have Amazon S3 encrypt your data at rest with server-side encryption (SSE), Amazon S3 will automatically encrypt your data on write and decrypt your data on retrieval
  - When Amazon S3 SSE encrypts data at rest, it uses Advanced Encryption Standard (AES) 256-bit symmetric keys.
- SSE with Amazon S3 Key Management (SSE-S3)
  - With SSE-S3, Amazon S3 will encrypt your data at rest and manage the encryption keys for you.
- SSE with Customer-Provided Keys (SSE-C)
  - With SSE-C, Amazon S3 will encrypt your data at rest using the custom encryption keys that you provide. To use SSE-C
- SSE with AWS KMS (SSE-KMS)
  - With SSE-KMS, Amazon S3 will encrypt your data at rest using keys that you manage in the AWS Key Management Service (KMS). AWS KMS provides an audit trail so you can see who used your key to access which object and when, as well as view failed attempts to access data from users without permission to decrypt the data.
  - Additionally, AWS KMS provides additional security controls to support customer efforts to comply with PCI-DSS, HIPAA/HITECH, and FedRAMP industry requirements.

Features of S3

- Security and Access Management
    - Amazon S3 provides several mechanisms to control and monitor who can access your data as well as how, when, and where they can access it. VPC endpoints allow you to create a secure connection without a gateway or NAT instances.
- Programmatic Access Using the AWS SDKs
    - Amazon S3 is supported by the AWS SDKs for Java, PHP, .NET, Python, Node.js, Ruby, and the AWS Mobile SDK. The SDK libraries wrap the underlying REST API, simplifying your programming tasks.
- Cost Monitoring and Controls
    - Amazon S3 has several features for managing and controlling your costs, including bucket tagging to manage cost allocation and integration with Amazon CloudWatch to receive billing alerts.

Ideal Usage Patterns

- Static web content and media.
    - This content can be delivered directly from Amazon S3, since each object in Amazon S3 has a unique HTTP URL address, or Amazon S3 can serve as an origin store for a content delivery network (CDN), such as Amazon CloudFront.
    - Because of Amazon S3's elasticity, it works particularly well for hosting web content with extremely spiky bandwidth demands.
    - Also, because no storage provisioning is required, Amazon S3 works well for fast growing websites hosting data intensive, user-generated content, such as video and photo sharing sites.

Ideal Usage Patterns

- Host entire static websites.
  - Amazon S3 provides a highly-available and highly scalable solution for websites with only static content, including HTML files, images, videos, and client-side scripts such as JavaScript.
- Data store for computation and large-scale analytics,
  - such as analysing financial transactions, clickstream analytics, and media transcoding.
  - Because of the horizontal scalability of Amazon S3, you can access your data from multiple computing nodes concurrently without being constrained by a single connection.
- Highly durable, scalable, and secure solution for backup and archival of critical data, and to provide disaster recovery solutions for business continuity.
  - Because Amazon S3 stores objects redundantly on multiple devices across multiple facilities, it provides the highly-durable storage infrastructure needed for these scenarios.
  - Amazon S3’s versioning capability is available to protect critical data from inadvertent deletion.

Performance

- Access to Amazon S3 from within Amazon EC2 in the same region is fast.
- Amazon S3 is designed so that server-side latencies are insignificant relative to Internet latencies.
- Amazon S3 is also built to scale storage, requests, and users to support a virtually unlimited number of web-scale applications.
- If you access Amazon S3 using multiple threads, multiple applications, or multiple clients concurrently, total Amazon S3 aggregate throughput will typically scale to rates that far exceed what any single server can generate or consume.

Performance

- To speed access to relevant data, many developers pair Amazon S3 with a database, such as Amazon Dynamo DB or Amazon RDS.
    - Amazon S3 stores the actual information,
    - The database serves as the repository for associated metadata (e.g., object name, size, keywords, and so on).
    - Metadata in the database can easily be indexed and queried, making it very efficient to locate an object’s reference via a database query.
    - This result can then be used to pinpoint and then retrieve the object itself from Amazon S3.

Durability and Availability

- By automatically and synchronously storing your data across both multiple devices and multiple facilities within your selected geographical region, Amazon S3 storage provides the highest level of data durability and availability in the AWS platform.
- Error correction is built-in, and there are no single points of failure.
- Amazon S3 is designed to sustain the concurrent loss of data in two facilities, making it very well-suited to serve as the primary data storage for mission-critical data.
- In fact, Amazon S3 is designed for 99.999999999% (11 nines) durability per object and 99.99% availability over a one-year period.
- In addition to its built-in redundancy, Amazon S3 data can also be protected from application failures and unintended deletions through the use of Amazon S3 versioning.
- You can also enable Amazon S3 versioning with Multi-Factor Authentication (MFA) Delete.
    - With this option enabled on a bucket, two forms of authentication are required to delete a version of an Amazon S3 object: valid AWS account credentials plus a six-digit code (a single-use, time-based password) from a physical token device.

Durability and Availability

- For noncritical data that can be reproduced easily if needed, such as transcoded media or image thumbnails, you can use the Reduced Redundancy Storage (RRS) option in Amazon S3, which provides a lower level of durability at a lower storage cost.
- Objects stored using the RRS option have less redundancy than objects stored using standard Amazon S3 storage.
- In either case, your data is still stored on multiple devices in multiple locations. RRS is designed to provide 99.99% durability per object over a given year.
- While RRS is less durable than standard Amazon S3, it is still designed to provide 400 times more durability than a typical disk drive.

Cost Model

- With Amazon S3, you pay only for what you use and there is no minimum fee.
- Amazon S3 has three pricing components:
  - storage (per GB per month),
  - data transfer in or out (per GB per month), and
  - requests (per n thousand requests per month).

Scalability and Elasticity

- Amazon S3 has been designed to offer a very high level of scalability and elasticity automatically.
- Unlike a typical file system that encounters issues when storing large number of files in a directory, Amazon S3 supports a virtually unlimited number of files in any bucket.
- Also, unlike a disk drive that has a limit on the total amount of data that can be stored before you must partition the data across drives and/or servers, an Amazon S3 bucket can store a virtually unlimited number of bytes.
- You are able to store any number of objects (files) in a single bucket, and Amazon S3 will automatically manage scaling and distributing redundant copies of your information to other servers in other locations in the same region, all using Amazon’s high-performance infrastructure.

Interfaces

- Amazon S3 provides standards-based REST and SOAP web services APIs for both management and data operations.
- These APIs allow Amazon S3 objects (files) to be stored in uniquely-named buckets (top-level folders).
- Each object must have a unique object key (file name) that serves as an identifier for the object within that bucket.
- While Amazon S3 is a web-based object store rather than a traditional file system, you can easily emulate a file system hierarchy (folder1/folder2/file) in Amazon S3 by creating object key names that correspond to the full path name of each file.

Anti-Patterns

- File system—Amazon S3 uses a flat namespace and isn’t meant to serve as a standalone, POSIX-compliant file system. However, by using delimiters (commonly either the ‘/’ or ‘\’ character) you are able construct your keys to emulate the hierarchical folder structure of file system within a given bucket.
- Structured data with query—Amazon S3 doesn’t offer query capabilities: to retrieve a specific object you need to already know the bucket name and key. Thus, you can’t use Amazon S3 as a database by itself. Instead, pair Amazon S3 with a database to index and query metadata about Amazon S3 buckets and objects.
- Rapidly changing data—Data that must be updated very frequently might be better served by a storage solution with lower read / write latencies, such as Amazon EBS volumes, Amazon RDS or other relational databases, or Amazon DynamoDB.
- Backup and archival storage—Data that requires long-term encrypted archival storage with infrequent read access may be stored more cost-effectively in Amazon Glacier.
- Dynamic website hosting—While Amazon S3 is ideal for websites with only static content, dynamic websites that depend on database interaction or use server-side scripting should be hosted on Amazon EC2.

Amazon Glacier

Ideal Usage Patterns

- Organizations are using Amazon Glacier to support a number of use cases.
- These include
  - Archiving offsite enterprise information,
  - Media assets,
  - Research and scientific data,
  - Digital preservation and magnetic tape replacement.

Performance

- Amazon Glacier is a low-cost storage service designed to store data that is infrequently accessed and long lived.
- Amazon Glacier jobs typically complete in 3 to 5 hours.

Durability and Availability

- Amazon Glacier is designed to provide average annual durability of 99.999999999% (11 nines) for an archive.
- The service redundantly stores data in multiple facilities and on multiple devices within each facility.
- To increase durability, Amazon Glacier synchronously stores your data across multiple facilities before returning SUCCESS on uploading archives.
- Unlike traditional systems, which can require laborious data verification and manual repair, Amazon Glacier performs regular, systematic data integrity checks and is built to be automatically self-healing.

Cost Model

- With Amazon Glacier, you pay only for what you use and there is no minimum fee.
- In normal use, Amazon Glacier has three pricing components:
  - storage (per GB per month),
  - data transfer out (per GB per month), and
  - requests (per thousand UPLOAD and RETRIEVAL requests per month).
- Note that Amazon Glacier is designed with the expectation that retrievals are infrequent and unusual, and data will be stored for extended periods of time. You can retrieve up to 5% of your average monthly storage (pro-rated daily) for free each month.
- If you choose to retrieve more than this amount of data in a month, you are charged an additional (per GB) retrieval fee. There is also a pro-rated charge (per GB) for items deleted prior to 90 days.

Scalability and Elasticity

- Amazon Glacier scales to meet your growing and often unpredictable storage requirements.
- A single archive is limited to 4 TBs, but there is no limit to the total amount of data you can store in the service.
- Whether you’re storing petabytes or gigabytes, Amazon Glacier automatically scales your storage up or down as needed.

Interfaces

- There are two ways to use Amazon Glacier, each with its own set of interfaces.
- The Amazon Glacier APIs provide both management and data operations.
- First, Amazon Glacier provides a native, standards-based REST web services interface, as well as Java and .NET SDKs.
- The AWS Management Console or the Amazon Glacier APIs can be used to create vaults to organize the archives in Amazon Glacier.
- You can then use the Amazon Glacier APIs to upload and retrieve archives, monitor the status of your jobs and also configure your vault to send you a notification via Amazon Simple Notification Service (Amazon SNS) when your jobs complete

Interfaces

- Second, Amazon Glacier can be used as a storage class in Amazon S3 by using object lifecycle management to provide automatic, policy-driven archiving from Amazon S3 to Amazon Glacier.
- You simply set one or more lifecycle rules for an Amazon S3 bucket, defining what objects should be transitioned to Amazon Glacier and when. You can specify an absolute or relative time period (including 0 days) after which the specified Amazon S3 objects should be transitioned to Amazon Glacier.
- A new RESTORE operation has been added to the Amazon S3 API, and the retrieval process takes the same 3-5 hours.
- On retrieval, a copy of the retrieved object is placed in Amazon S3 RRS storage for a specified retention period; the original archived object remains stored in Amazon Glacier. For more information on how to use Amazon

Anti-Patterns

- Rapidly changing data—Data that must be updated very frequently might be better served by a storage solution with lower read/write latencies, such as Amazon EBS or a database.
- Real time access—Data stored in Amazon Glacier is not available in real time. Retrieval jobs typically require 3-5 hours to complete, so if you need immediate access to your data, Amazon S3 is a better choice.