Amazon S3 Section

Section introduction

- Amazon S3 is one of the main building blocks of AWS
- It's advertised as "infinitely scaling" storage

- Many websites use Amazon S3 as a backbone
- Many AWS services use Amazon S3 as an integration as well

- We'll have a step-by-step approach to S3
- The CCP exam requires "deeper" knowledge about S3

S3 Use cases

- Backup and storage
- Disaster Recovery
- Archive
- Hybrid Cloud storage
- Application hosting
- Media hosting
- Data lakes & big data analytics
- Software delivery
- Static website

Nasdaq stores 7 years of data into S3 Glacier

Sysco runs analytics on its data and gain business insights

Amazon S3 Overview - Buckets

- Amazon S3 allows people to store objects (files) in "buckets" (directories)
- Buckets must have a **globally unique name (across all regions all accounts)**
- Buckets are defined at the region level
- S3 looks like a global service but buckets are created in a region
- Naming convention
  - No uppercase
  - No underscore
  - 3-63 characters long
  - Not an IP
  - Must start with lowercase letter or number

Amazon S3 Overview - Objects

- Objects (files) have a Key

- The key is the **FULL** path:
  - s3://my-bucket/my_file.txt
  - s3://my-bucket/my_folder1/another_folder/my_file.txt

- The key is composed of prefix + object name
  - s3://my-bucket/my_folder1/another_folder/my_file.txt

- There's no concept of "directories" within buckets (although the UI will trick you to think otherwise)

- Just keys with very long names that contain slashes ("/")

# Amazon S3 Overview – Objects (continued)

- Object values are the content of the body:
  - Max Object Size is 5TB (5000GB)
  - If uploading more than 5GB, must use "multi-part upload"

- Metadata (list of text key / value pairs – system or user metadata)

- Tags (Unicode key / value pair – up to 10) – useful for security / lifecycle

- Version ID (if versioning is enabled)

S3 Security

- User based
  - IAM policies - which API calls should be allowed for a specific user from IAM console
- Resource Based
  - Bucket Policies - bucket wide rules from the S3 console - allows cross account
  - Object Access Control List (ACL) – finer grain
  - Bucket Access Control List (ACL) – less common
- **Note:** an IAM principal can access an S3 object if
  - the user IAM permissions allow it <u>OR</u> the resource policy ALLOWS it
  - <u>AND</u> there's no explicit DENY
- **Encryption:** encrypt objects in Amazon S3 using encryption keys
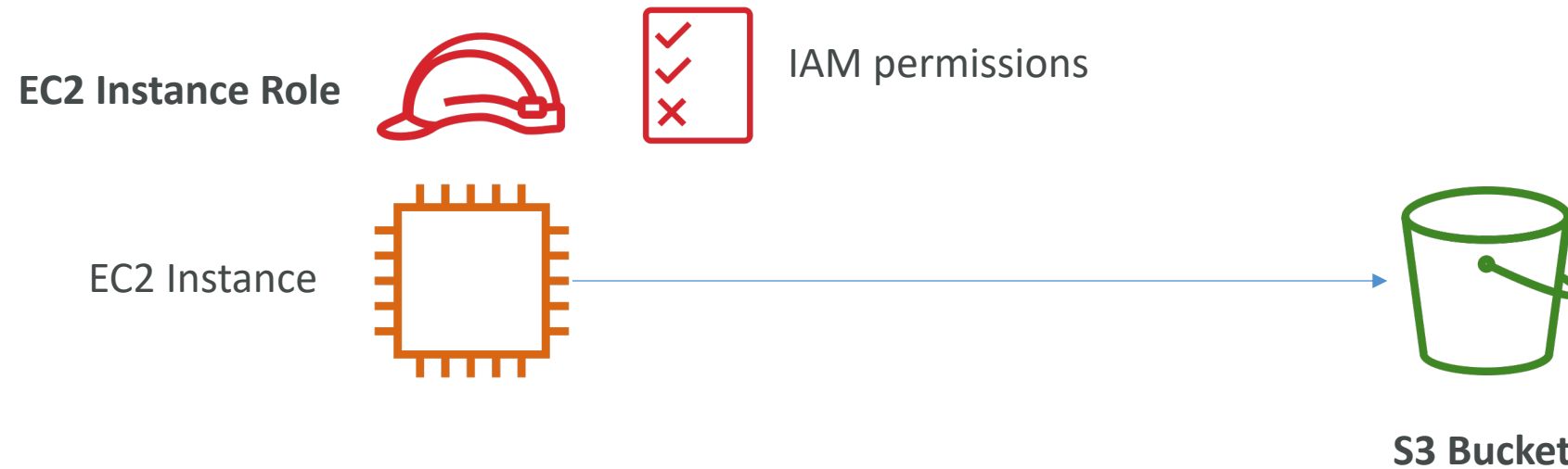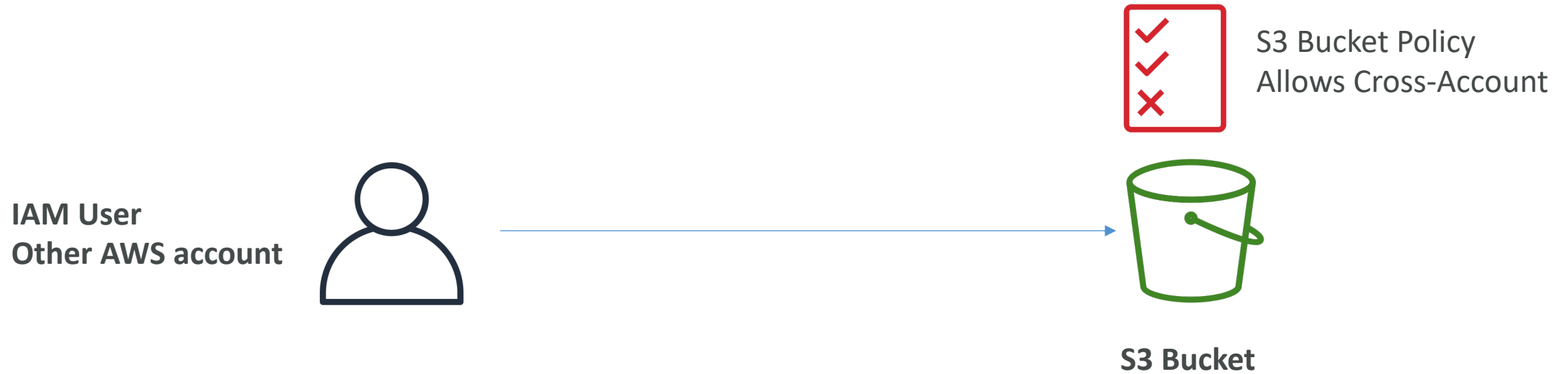
Example: Public Access - Use Bucket Policy

S3 Bucket Policy
Allows Public Access

Anonymous www website visitor
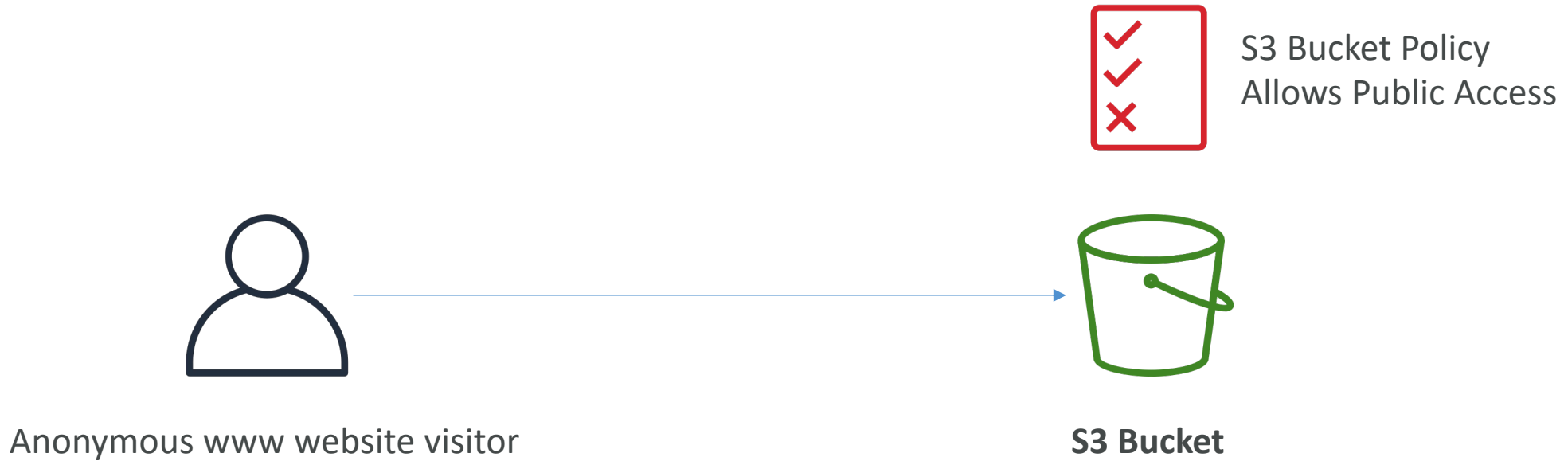
S3 Bucket

Example: User Access to S3 – IAM permissions

IAM Policy

IAM User

S3 Bucket

Example: EC2 instance access - Use IAM Roles

EC2 Instance Role

IAM permissions

EC2 Instance

S3 Bucket

Advanced: Cross-Account Access – Use Bucket Policy

**IAM User
Other AWS account**

S3 Bucket Policy
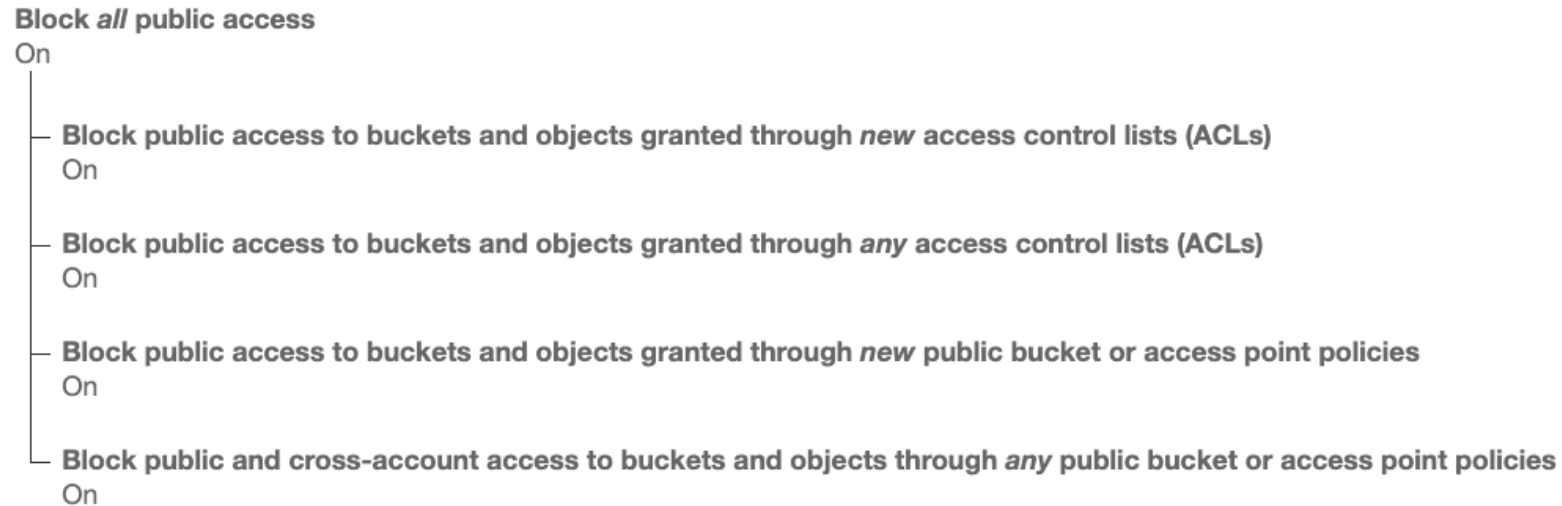Allows Cross-Account

S3 Bucket

S3 Bucket Policies

- JSON based policies
  - Resources: buckets and objects
  - Actions: Set of API to Allow or Deny
  - Effect: Allow / Deny
  - Principal: The account or user to apply the policy to

- Use S3 bucket for policy to:
  - Grant public access to the bucket
  - Force objects to be encrypted at upload
  - Grant access to another account (Cross Account)

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "PublicRead",
            "Effect": "Allow",
            "Principal": "*",
            "Action": [
                "s3:GetObject"
            ],
            "Resource": [
                "arn:aws:s3:::examplebucket/*"
            ]
        }
    ]
}
```

Bucket settings for Block Public Access

Block *all* public access
On

- **Block public access to buckets and objects granted through *new* access control lists (ACLs)**
On

- **Block public access to buckets and objects granted through *any* access control lists (ACLs)**
On

- **Block public access to buckets and objects granted through *new* public bucket or access point policies**
On

- **Block public and cross-account access to buckets and objects through *any* public bucket or access point policies**
On

- **These settings were created to prevent company data leaks**
- If you know your bucket should never be public, leave these on
- Can be set at the account level

S3 Websites

- S3 can host static websites and have them accessible on the www

- The website URL will be:
  - \<bucket-name>.s3-website-\<AWS-region>.amazonaws.com

  OR
  - \<bucket-name>.s3-website.\<AWS-region>.amazonaws.com

- If you get a 403 (Forbidden) error, make sure the bucket policy allows public reads!

Amazon S3 - Versioning

- You can version your files in Amazon S3

- It is enabled at the **bucket level**

- Same key overwrite will increment the "version": 1, 2, 3….

- It is best practice to version your buckets
  - Protect against unintended deletes (ability to restore a version)
  - Easy roll back to previous version

- Notes:
  - Any file that is not versioned prior to enabling versioning will have version "null"
  - Suspending versioning does not delete the previous versions

S3 Access Logs

- For audit purpose, you may want to **log all access to S3 buckets**

- Any request made to S3, from any account, authorized or denied, will be logged into another S3 bucket

- That data can be analyzed using data analysis tools…

- Very helpful to come down to the root cause of an issue, or audit usage, view suspicious patterns, etc…

requests

My-bucket

Log all requests

Logging Bucket
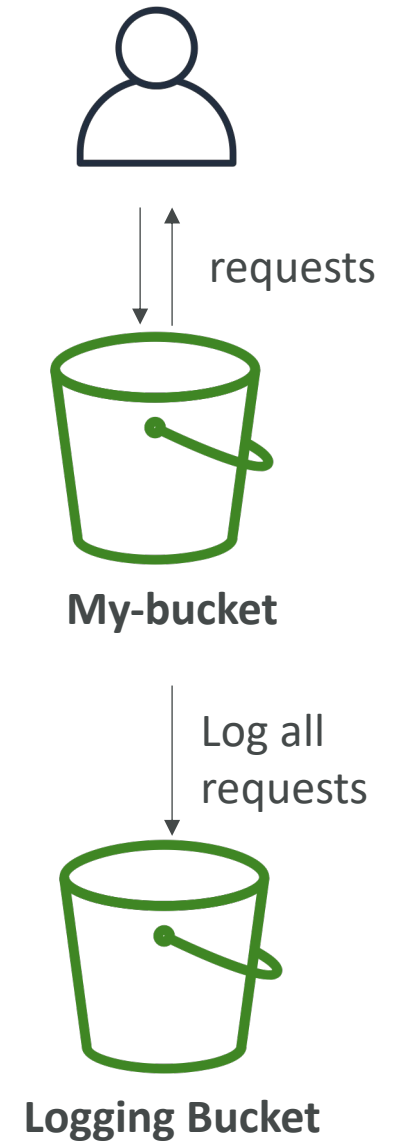
S3 Replication (CRR & SRR)

- **Must enable versioning** in source and destination
- Cross Region Replication (CRR)
- Same Region Replication (SRR)
- Buckets can be in different accounts
- Copying is asynchronous
- Must give proper IAM permissions to S3

eu-west-1 → Asynchronous replication → us-east-1
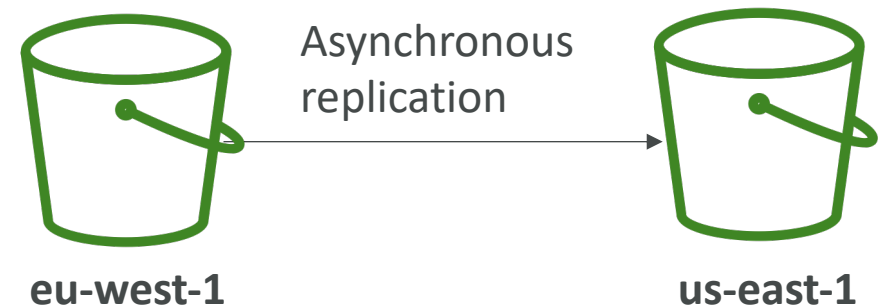
- <u>CRR - Use cases</u>: compliance, lower latency access, replication across accounts
- <u>SRR – Use cases</u>: log aggregation, live replication between production and test accounts

S3 Storage Classes

- Amazon S3 Standard - General Purpose
- Amazon S3 Standard-Infrequent Access (IA)
- Amazon S3 One Zone-Infrequent Access
- Amazon S3 Glacier Instant Retrieval
- Amazon S3 Glacier Flexible Retrieval
- Amazon S3 Glacier Deep Archive
- Amazon S3 Intelligent Tiering

- Can move between classes manually or using S3 Lifecycle configurations

S3 Durability and Availability

- Durability:
  - High durability (99.999999999%, 11 9's) of objects across multiple AZ
  - If you store 10,000,000 objects with Amazon S3, you can on average expect to incur a loss of a single object once every 10,000 years
  - Same for all storage classes

- Availability:
  - Measures how readily available a service is
  - Varies depending on storage class
  - Example: S3 standard has 99.99% availability = not available 53 minutes a year

S3 Standard – General Purpose

- 99.99% Availability
- Used for frequently accessed data
- Low latency and high throughput
- Sustain 2 concurrent facility failures

- Use Cases: Big Data analytics, mobile & gaming applications, content distribution…

S3 Storage Classes – Infrequent Access

- For data that is less frequently accessed, but requires rapid access when needed
- Lower cost than S3 Standard

- Amazon S3 Standard-Infrequent Access (S3 Standard-IA)
  - 99.9% Availability
  - Use cases: Disaster Recovery, backups

- Amazon S3 One Zone-Infrequent Access (S3 One Zone-IA)
  - High durability (99.999999999%) in a single AZ; data lost when AZ is destroyed
  - 99.5% Availability
  - Use Cases: Storing secondary backup copies of on-premise data, or data you can recreate

www.datacumulus.com

Amazon S3 Glacier Storage Classes

- Low-cost object storage meant for archiving / backup
- Pricing: price for storage + object retrieval cost

- **Amazon S3 Glacier Instant Retrieval**
  - Millisecond retrieval, great for data accessed once a quarter
  - Minimum storage duration of 90 days
- **Amazon S3 Glacier Flexible Retrieval** (formerly Amazon S3 Glacier):
  - Expedited (1 to 5 minutes), Standard (3 to 5 hours), Bulk (5 to 12 hours) – free
  - Minimum storage duration of 90 days
- **Amazon S3 Glacier Deep Archive** – for long term storage:
  - Standard (12 hours), Bulk (48 hours)
  - Minimum storage duration of 180 days

S3 Intelligent-Tiering

- Small monthly monitoring and auto-tiering fee
- Moves objects automatically between Access Tiers based on usage
- There are no retrieval charges in S3 Intelligent-Tiering

- *Frequent Access tier (automatic):* default tier
- *Infrequent Access tier (automatic)*: objects not accessed for 30 days
- *Archive Instant Access tier (automatic):* objects not accessed for 90 days
- *Archive Access tier (optional):* configurable from 90 days to 700+ days
- *Deep Archive Access tier (optional):* config. from 180 days to 700+ days
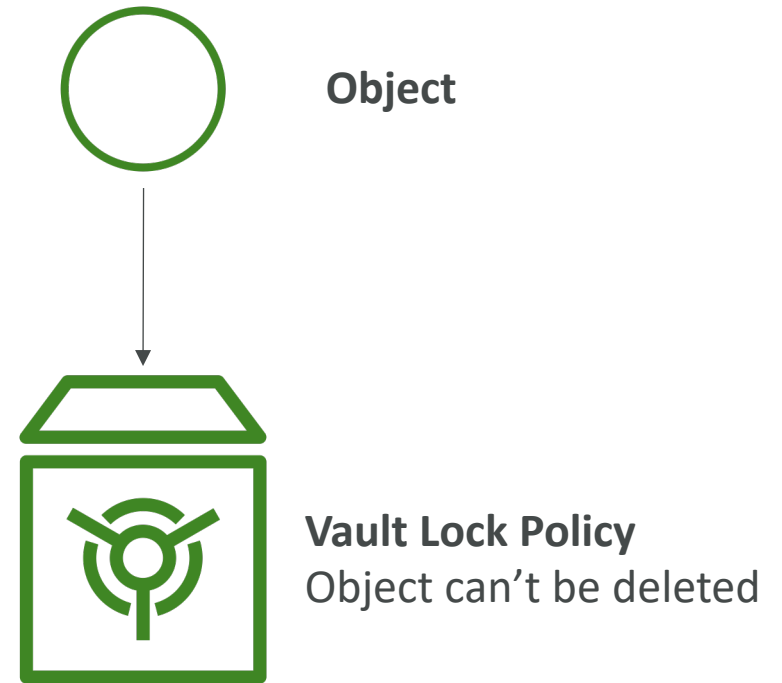
S3 Storage Classes Comparison

	Standard	Intelligent-Tiering	Standard-IA	One Zone-IA	Glacier Instant Retrieval	Glacier Flexible Retrieval	Glacier Deep Archive
Durability	99.999999999% == (11 9's)						
Availability	99.99%	99.9%	99.9%	99.5%	99.9%	99.99%	99.99%
Availability SLA	99.9%	99%	99%	99%	99%	99.9%	99.9%
Availability Zones	>= 3	>= 3	>= 3	1	>= 3	>= 3	>= 3
Min. Storage Duration Charge	None	None	30 Days	30 Days	90 Days	90 Days	180 Days
Min. Billable Object Size	None	None	128 KB	128 KB	128 KB	40 KB	40 KB
Retrieval Fee	None	None	Per GB retrieved	Per GB retrieved	Per GB retrieved	Per GB retrieved	Per GB retrieved

https://aws.amazon.com/s3/storage-classes/

S3 Storage Classes – Price Comparison Example: us-east-1

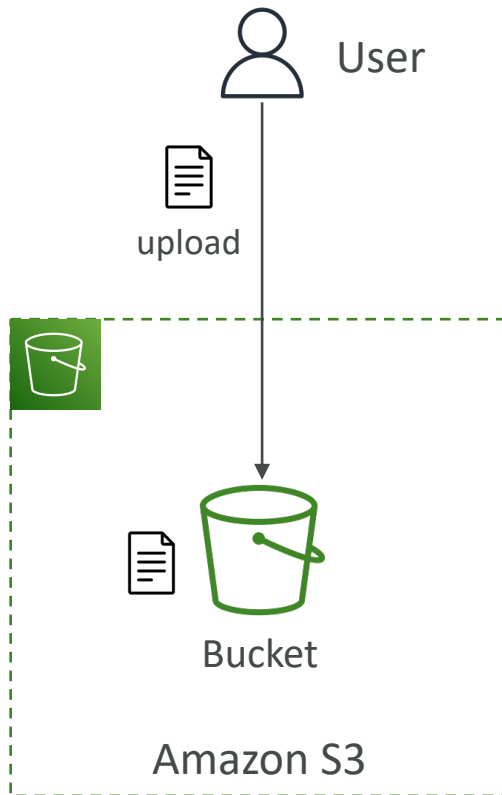|	Standard	Intelligent-Tiering	Standard-IA	One Zone-IA	Glacier Instant Retrieval	Glacier Flexible Retrieval	Glacier Deep Archive
Storage Cost (per GB per month)	$0.023	$0.0025 - $0.023	%0.0125	$0.01	$0.004	$0.0036	$0.00099
Retrieval Cost (per 1000 request)	**GET:** $0.0004 **POST:** $0.005	**GET:** $0.0004 **POST:** $0.005	**GET:** $0.001 **POST:** $0.01	**GET:** $0.001 **POST:** $0.01	**GET:** $0.01 **POST:** $0.02	**GET:** $0.0004 **POST:** $0.03 **Expedited:** $10 **Standard:** $0.05 **Bulk:** free	**GET:** $0.0004 **POST:** $0.05 **Standard:** $0.10 **Bulk:** $0.025
Retrieval Time	Instantaneous					**Expedited** (1 – 5 mins) **Standard** (3 – 5 hours) Bulk (5 – 12 hours)	**Standard** (12 hours) **Bulk** (48 hours)
Monitoring Cost (pet 1000 objects)		$0.0025					

https://aws.amazon.com/s3/pricing/

S3 Object Lock & Glacier Vault Lock

- S3 Object Lock
  - Adopt a WORM (Write Once Read Many) model
  - Block an object version deletion for a specified amount of time

- Glacier Vault Lock
  - Adopt a WORM (Write Once Read Many) model
  - Lock the policy for future edits (can no longer be changed)
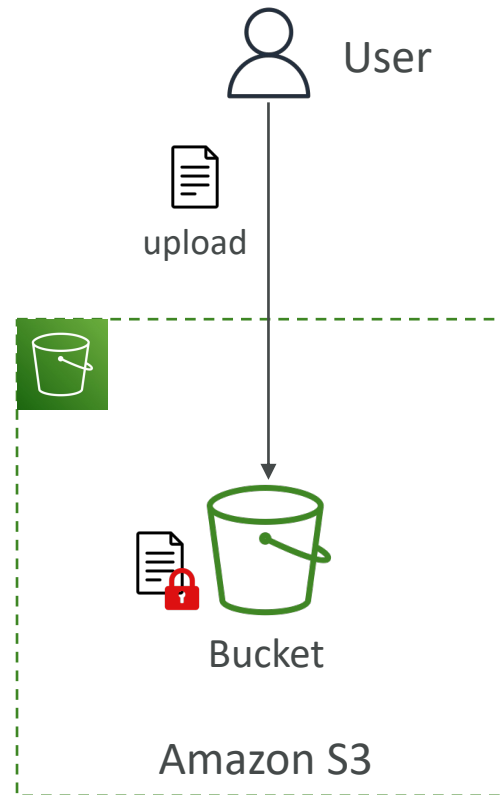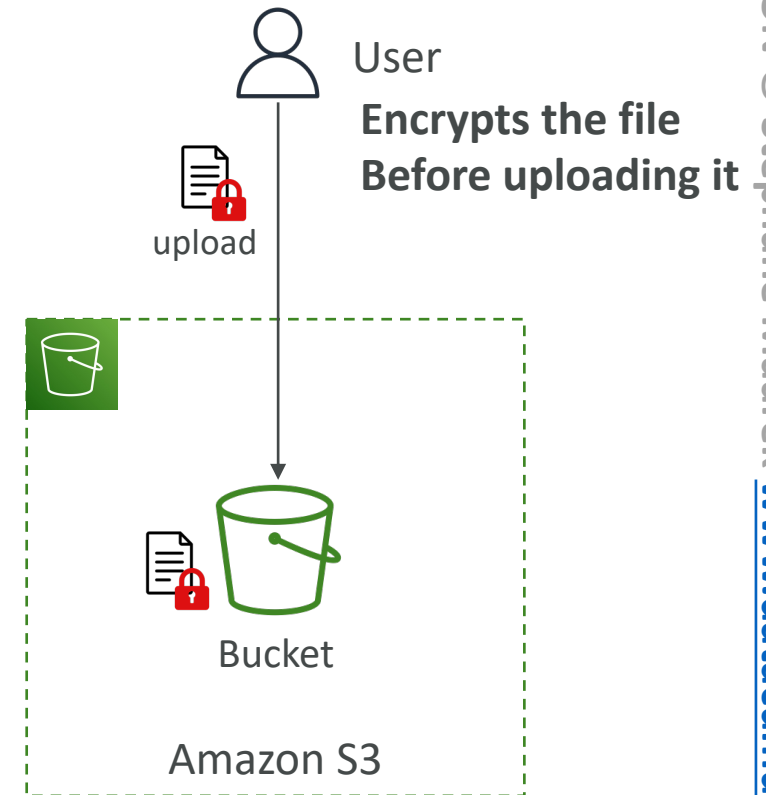 - Helpful for compliance and data retention

Object

Vault Lock Policy
Object can't be deleted

www.datacumulus.com

S3 Encryption

No Encryption

User

upload

Bucket

Amazon S3

Server-Side Encryption

User

upload

Bucket

Amazon S3

Client-Side Encryption

User

Encrypts the file Before uploading it

upload

Bucket

Amazon S3

Server encrypts the file after receiving it

# Shared Responsibility Model for S3

aws

- Infrastructure (global security, durability, availability, sustain concurrent loss of data in two facilities)

- Configuration and vulnerability analysis

- Compliance validation

- S3 Versioning
- S3 Bucket Policies
- S3 Replication Setup
- Logging and Monitoring
- S3 Storage Classes
- Data encryption at rest and in transit

AWS Snow Family

- Highly-secure, portable devices to **collect and process data at the edge,** and **migrate data into and out of AWS**

- Data migration:
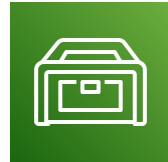
  Snowcone

  Snowball Edge

  Snowmobile

- Edge computing:

  Snowcone

  Snowball Edge

Data Migrations with AWS Snow Family

	Time to Transfer		
	100 Mbps	**1Gbps**	**10Gbps**
10 TB	12 days	30 hours	3 hours
100 TB	124 days	12 days	30 hours
1 PB	3 years	124 days	12 days

Challenges:
- Limited connectivity
- Limited bandwidth
- High network cost
- Shared bandwidth (can't maximize the line)
- Connection stability

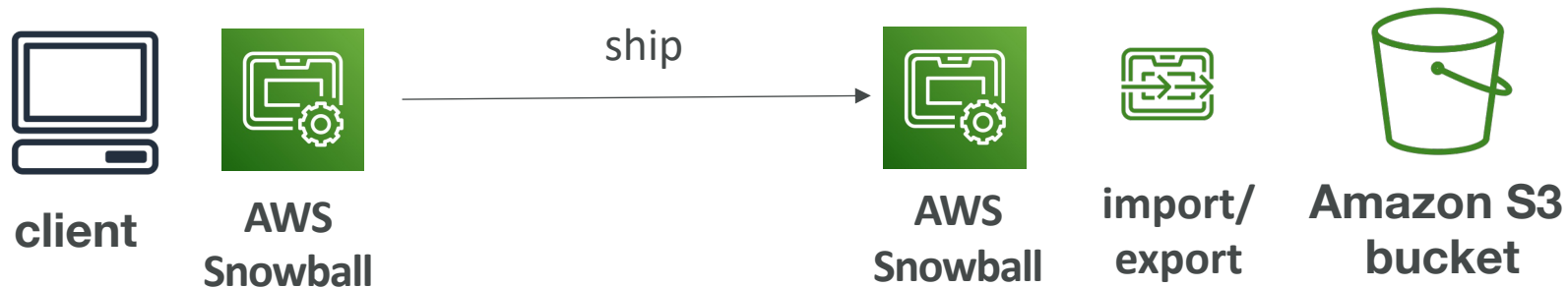AWS Snow Family: offline devices to perform data migrations

If it takes more than a week to transfer over the network, use Snowball devices!

Diagrams

- Direct upload to S3:

client → www: 10Gbit/s → **Amazon S3 bucket**

- With Snow Family:

client **AWS Snowball** → ship → **AWS Snowball** **import/export** **Amazon S3 bucket**

Snowball Edge (for data transfers)

- Physical data transport solution: move TBs or PBs of data in or out of AWS
- Alternative to moving data over the network (and paying network fees)
- Pay per data transfer job
- Provide block storage and Amazon S3-compatible object storage
- **Snowball Edge Storage Optimized**
  - 80 TB of HDD capacity for block volume and S3 compatible object storage
- **Snowball Edge Compute Optimized**
  - 42 TB of HDD capacity for block volume and S3 compatible object storage
- Use cases: large data cloud migrations, DC decommission, disaster recovery

AWS Snowcone

- **Small, portable computing, anywhere, rugged & secure, withstands harsh environments**
- Light (4.5 pounds, 2.1 kg)
- Device used for edge computing, storage, and data transfer
- **8 TBs of usable storage**
- Use Snowcone where Snowball does not fit (space-constrained environment)
- Must provide your own battery / cables

- Can be sent back to AWS offline, or connect it to internet and use **AWS DataSync** to send data

AWS Snowmobile

- Transfer exabytes of data (1 EB = 1,000 PB = 1,000,000 TBs)
- Each Snowmobile has 100 PB of capacity (use multiple in parallel)
- High security: temperature controlled, GPS, 24/7 video surveillance
- **Better than Snowball if you transfer more than 10 PB**

AWS Snow Family for Data Migrations

Snowcone

Snowball Edge

Snowmobile

	Snowcone	Snowball Edge Storage Optimized	Snowmobile
Storage Capacity	8 TB usable	80 TB usable	< 100 PB
Migration Size	Up to 24 TB, online and offline	Up to petabytes, offline	Up to exabytes, offline
DataSync agent	Pre-installed		
Storage Clustering		Up to 15 nodes	

Snow Family – Usage Process

1. Request Snowball devices from the AWS console for delivery

2. Install the snowball client / AWS OpsHub on your servers

3. Connect the snowball to your servers and copy files using the client

4. Ship back the device when you're done (goes to the right AWS facility)

5. Data will be loaded into an S3 bucket

6. Snowball is completely wiped

What is Edge Computing?

- Process data while it's being created on **an edge location**
  - A truck on the road, a ship on the sea, a mining station underground...

- These locations may have
  - Limited / no internet access
  - Limited / no easy access to computing power

- We setup a **Snowball Edge / Snowcone** device to do edge computing

- Use cases of Edge Computing:
  - Preprocess data
  - Machine learning at the edge
  - Transcoding media streams

- Eventually (if need be) we can ship back the device to AWS (for transferring data for example)

Snow Family – Edge Computing

- Snowcone (smaller)
  - 2 CPUs, 4 GB of memory, wired or wireless access
  - USB-C power using a cord or the optional battery
- Snowball Edge – Compute Optimized
  - 52 vCPUs, 208 GiB of RAM
  - Optional GPU (useful for video processing or machine learning)
  - 42 TB usable storage
- Snowball Edge – Storage Optimized
  - Up to 40 vCPUs, 80 GiB of RAM
  - Object storage clustering available
- All: Can run EC2 Instances & AWS Lambda functions (using AWS IoT Greengrass)
- Long-term deployment options: 1 and 3 years discounted pricing

AWS OpsHub

- Historically, to use Snow Family devices, you needed a CLI (Command Line Interface tool)
- Today, you can use **AWS OpsHub** (a software you install on your computer / laptop) to manage your Snow Family Device
  - Unlocking and configuring single or clustered devices
  - Transferring files
  - Launching and managing instances running on Snow Family Devices
  - Monitor device metrics (storage capacity, active instances on your device)
  - Launch compatible AWS services on your devices (ex: Amazon EC2 instances, AWS DataSync, Network File System (NFS))

https://aws.amazon.com/blogs/aws/aws-snowball-edge-update/
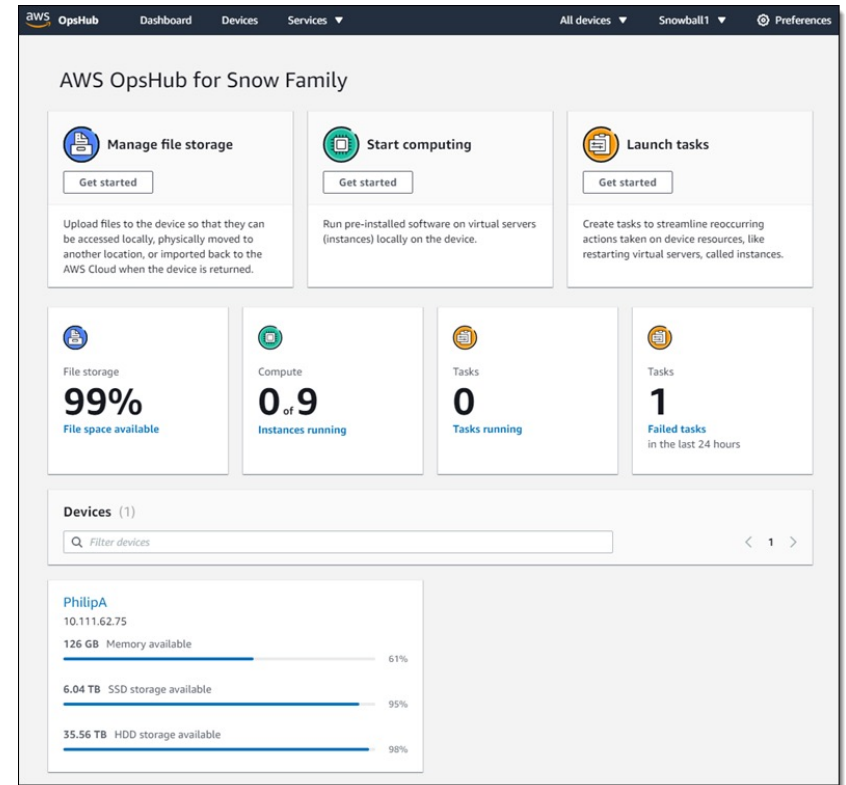
Hybrid Cloud for Storage

- AWS is pushing for "hybrid cloud"
  - Part of your infrastructure is on-premises
  - Part of your infrastructure is on the cloud

- This can be due to
  - Long cloud migrations
  - Security requirements
  - Compliance requirements
  - IT strategy

- S3 is a proprietary storage technology (unlike EFS / NFS), so how do you expose the S3 data on-premise?
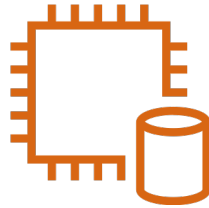
- AWS Storage Gateway!
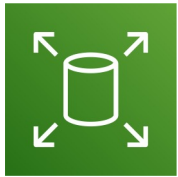
AWS Storage Gateway

- Bridge between on-premise data and cloud data in S3

- Hybrid storage service to allow on-premises to seamlessly use the AWS Cloud

- Use cases: disaster recovery, backup & restore, tiered storage

- Types of Storage Gateway:
  - File Gateway
  - Volume Gateway
  - Tape Gateway

- No need to know the types at the exam

Files

Volumes

Tapes

AWS Storage Gateway

Amazon EBS

S3

Glacier

Amazon S3 – Summary

- **Buckets vs Objects:** global unique name, tied to a region
- **S3 security:** IAM policy, S3 Bucket Policy (public access), S3 Encryption
- **S3 Websites:** host a static website on Amazon S3
- **S3 Versioning:** multiple versions for files, prevent accidental deletes
- **S3 Access Logs:** log requests made within your S3 bucket
- **S3 Replication:** same-region or cross-region, must enable versioning
- **S3 Storage Classes:** Standard, IA, 1Z-IA, Intelligent, Glacier, Glacier Deep Archive
- **S3 Lifecycle Rules:** transition objects between classes
- **S3 Glacier Vault Lock / S3 Object Lock:** WORM (Write Once Read Many)
- **Snow Family:** import data onto S3 through a physical device, edge computing
- **OpsHub:** desktop application to manage Snow Family devices
- **Storage Gateway:** hybrid solution to extend on-premises storage to S3