Table of Contents

Section 1

Cloud Computing

- On demand delivery of compute power. Database storage, applications and other IT resources

- Pay as you go pricing
- Right type & size of computing resources you need
- Can access resources instantly

Deployment models of the cloud

1. Private cloud
2. Public cloud
3. Hybrid cloud

Characteristics of Cloud computing

- On-demand self service
- Broad network access
- Multi tenancy & resource sharing
- Rapid elasticity & scalability
- Measured service

Advantages of cloud computing

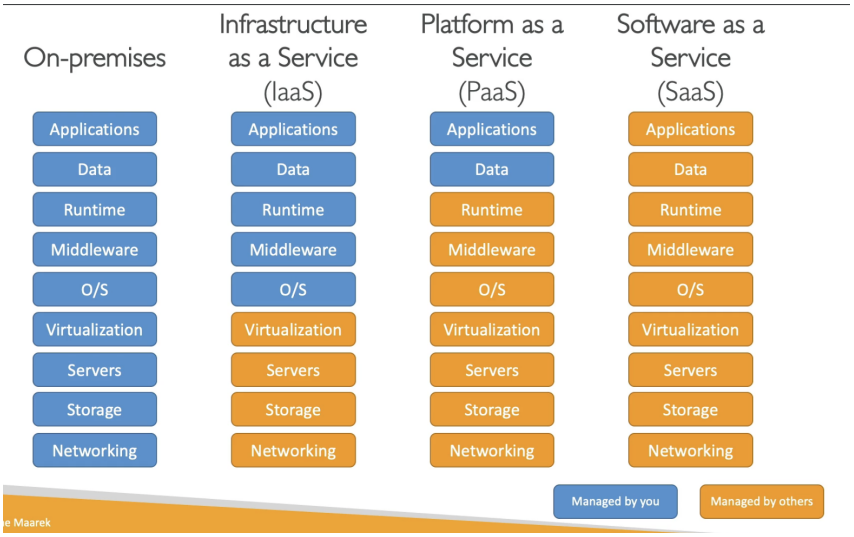
- Trade capex for opex
- Benefit from massive economies of scale
- Increase speed & agility
- Stop spending money running & maintaining data centers
- Go global in minutes: leverage the AWS global infrastructure

Problems solved by the cloud

- Flexibility - Change resource types when needed
- Cost effectiveness - Pay as you go
- Scalability
- Elasticity - Scale in/out according to the demand
- High availability & fault tolerance
- Agility - Rapidly develop, test, launch

Types of Cloud Computing

- Infrastructure as a service (IaaS)
    - Provides basic blocks for IT
    - Networking, computers, data storage
    - Easy to use with traditional on-premises IT
    - Highest level of flexibility
- Platform as a service (PaaS)
    - Removes the need for your organization to manage the underlying infrastructure
- Software as a Service (SaaS)
    - Completed product that is run & managed by the service provider



Pricing of the cloud

- Compute
- Storage
- Data Transfer OUT of the cloud

AWS Global Infrastructure

- AWS Regions
  - A cluster of data centers
  - All around the world
  - How to choose a region:
    - Compliance
    - Proximity
    - Available services
    - Pricing
- AWS Availability Zones (AZ)
  - Usually 3, min is 2, max is 6 in a region
  - Separate from each other
  - Each AZ has one or more discrete data centers
  - Connected with high bandwidth, ultra low latency networking
- AWS Point of presence (Edge locations)
  - 216 POP in 84 cities across 42 countries

Section 2

Identity and Access Management

- Global service
- Root account created by default, shouldn't be used or shared so create IAM users
- JSON document -> IAM policies

IAM Policies

- Inline -> Policy only attached to a user
- Consists of:
    - Version: policy language version (Mandatory)
    - ID: An identifier for the policy (optional)
    - Statement: one or more individual statements (Mandatory)
- Statement consists of:
    - Sid: an identifier for the statement (optional)
    - Effect: whether the statement allows or denies access (Allow, Deny)
    - Principal: account/user/role to which this policy applied to
    - Action: list of actions this policy allows or devices
    - Resources: list of actions this policy allows or denies
    - Condition: conditions for when this policy is in effect (optional)

IAM Password Policy

In AWS one can setup a password policy for higher security

Multi factor Authentication (MFA)

- Protect your root accounts & IAM users

Virtual MFA device

- Google authenticator (Phone)
- Authy (multi device)

Universal 2nd factor (U2F)

- Single key
- Hardware device

Hardware key Fob MFA Device

Hardware key Fob MFA Device for AWS GovCloud (US)

How can users access AWS?

- AWS management console

  Protected by password +MFA

- AWS Command Line Interface (CLI)

  Protected by access keys
  Access key ID - username
  Secret key - password

- AWS Software Developers Kit (SDK)
  - For code, protected by access keys
  - Language specific APIs
  - Embedded within your application
  - Supports
    - SDKs (JS, python, PHP, .NET, Ruby…)
    - Mobile SDKs (Android, iOS,...)
    - IOT Device SDKs (Embedded C, Arduino)
  - Eg: AWS CLI is built on AWS SDK for python

Note - AWS Cloudshell is AWS inbuilt terminal

IAM Roles for Services

- To give permissions to AWS services to perform actions

IAM Security Tools

- IAM Credentials Report (Account level)
  - A report that lists all your account's users and the status of their various credentials
- IAM Access Advisor (user level)
  - Shows the service permissions granted to a user and when those services were last accessed
  - Least privilege principle

IAM Guidelines & Best Practices

- Root account only for AWS account setup
- One physical user - one AWS user
- Assign users to groups & assign permissions to groups
- Use and enforce the use of MFA
- Create and use rules for giving permissions to AWS services

- Use access keys for programmatic access
- For audit - IAM credentials report

Section 3

Elastic Compute Cloud

- Infrastructure as a service
- Sizing and configuration options
    - OS: Linux, MacOS, windows
    - CPU
    - RAM
    - Storage space
        - Network attached (EBS & EFS)
        - Hardware (EC2 Instance store)
    - Network card: Speed of the card, public IP address
    - Firewall rules: security group
    - Bootstrap script: EC2 user data
        - Only runs once at the instance first start
        - Used to automate boot tasks
        - Runs with the user-root

EC2 Instance Types

m5.2xlarge

m: instance class
5: Generation (AWS improves them over time)
2xlarge: Size within the instance class

EC2 Instance Types - General Purpose (T)

- Great for a diversity of workloads such as web servers or code repositories
- Balance between: compute, memory, networking
- Eg: t2.micro

EC2 Instance Types - Compute Optimized (C)

- Great for compute-intensive tasks that require high performance processors
    - Batch processing workloads
    - Media transcoding
    - High performance web servers
    - High performance computing (HPC)
    - Scientific modeling & ML
    - Dedicated gaming servers

EC2 Instance Types - Memory Optimized (R)

- Fast performance for workloads that process large data sets in memory
- High performance, relational/non relational databases
- Distributed web scale cache stores
- In-memory databases optimized for BI
- Applications of big unstructured data

EC2 Instance Types - Storage Optimized (I)

- Great for storage intensive tasks that require high, sequential read and write access to large data sets on local storage
- Use cases:
    - High frequency online transaction processing (OLTP) systems
    - Relational & NoSQL databases
    - Cache for in-memory databases (redis)
    - Data warehousing applications
    - Distributed file systems (I,G,H)

Security Groups

- To control how traffic is allowed into or out of EC2 instances
- Only contain allow rules & act as a firewall on EC2 instances
- Can reference by IP or security group
- They regulate:
    - Access to ports
    - Authorized IP ranges - IPv4 & IPv6
    - Control of inbound and outbound network
- Can be attached to multiple instances
- Locked down to a region/VPC combination
- Lives outside the instance
- Good to maintain one separate security group for SSH access
- If your application is not accessible; then it’s a security group issue
- “Connection refused” means application issue
- All inbound traffic is blocked and all outbound traffic is authorized by default

Classic Ports to know

- 22 = SSH (Secure shell) - log into a linux instance
- 21 = FTP (File transfer protocol) - upload files into a file share
- 22 = SFTP (Secure file transfer protocol) - upload files using SSH
- 80 = HTTP - Access unsecured websites
- 443 = HTTPS - Access secured websites
- 3389 = RDP (Remote desktop protocol) - Log into a windows instance

SSH Summary Table

	SSH	Putty	EC2 Instance

			Connect
Mac	Yes	-	Yes
Linux	Yes	-	Yes
Windows<10	-	Yes	Yes
Windows>10	Yes	Yes	Yes

EC2 Instance Purchasing Options

- On demand Instances
  - Pay for what you use
    - Linux or windows - billing per second after the first min
    - All other OS - billing per hour
    - Has the highest cost but no upfront payment
    - No long-term commitment
- Reserved Instances
  - Upto 75% discount compared to on demand
  - 1 year = +discount | 3 years = +++discount
  - No upfront | partial upfront = +discount | All upfront = +++discount
  - Reserve a specific instance types
  - Types:
    - Convertible reserved instances
      - Can change the EC2 instance type
      - Upto 54% discount
    - Scheduled reserved instances (deprecated)
      - Launch within time window you reserve
      - Fraction of day/week/month
      - Still commitment over 1 to 3 years
- Spot Instances
  - Upto 90% compared to on demand
  - Can lose any point of time is price < spot price
  - Useful for- Batch jobs, data analysis, image processing, any distributes workloads
  - Not suitable for Critical jobs or databases
- Dedicated hosts
  - An amazon EC2 dedicated host is a physical server with EC2 instance
  - Capacity fully dedicated to your use
  - Dedicated hosts can help you address compliance requirements and reduce costs by allowing you to use your existing server - bound software licenses
  - 3 year period
  - More expensive
  - For software that have complicated licensing model (BYOL)
- Dedicated Instances
  - Instances running on dedicated hosts
  - May share hardware with other instances in same account

    - No control over instance placement (can move hardware after start /stop)

Section 4

Storage options for EC2 Instances

EBS Volume (Elastic Block Store)

- It is network drive you can attach to your instances while they run
- Allows your instance to persist data, even after their termination
- They can only be mounted to one instance at a time (at CCP level)
- Bound to a specific availability zone
- Free - tier: 30GB of free EBS storage of type general purpose (SSD) or magnetic/month
- It uses the network to communicate the instance, which means there might be a bit of latency
- It can be detached from an EC2 instance and attached to another one quickly
- To move a volume across, you need to snapshot it
- You get billed for all the provisioned capacity
- You can increase the capacity over time

EBS - delete on termination attribute

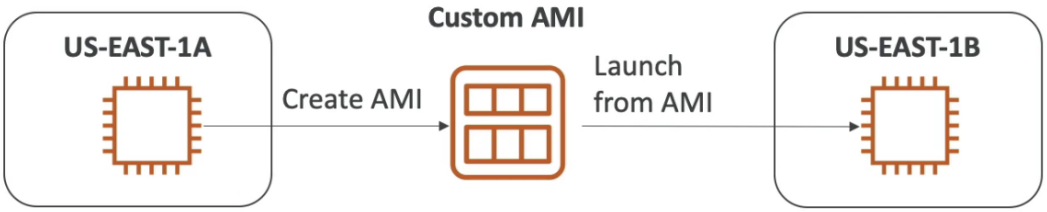
    - By default the root EBS volume
    - By default, any other attached EBS volume is not deleted
    - Use case: preserve root volume when instance is terminated

EBS Snapshots

    - Backup of your EBS volume
    - Not necessary to detach but recommended
    - Can copy snapshots across AZ or region

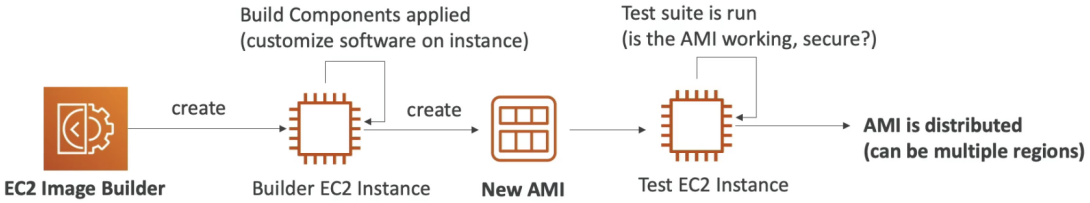
Amazon Machine Image (AMI)

- AMI are a customization of a EC2 instance
    - Can add your own software, config, OS
    - Faster boot time because all software is pre-packaged
- AMI are built for a specific region (and can be copied across regions)
- Can launch EC2 instances from:
    - A Public AMI: AWS provided
    - Own AMI
    - AWS marketplace AMI



EC2 Image Builder

- Used to automate the creation of virtual machines or container images
- Automate the creation, maintain, validate & test EC2 AMIs
- Can be run on a schedule
- Free service (only pay for the resources)



EC2 Instance Store

- High performance storage option for EC2 instances
- Better I/O performance
- It loses its storage if they are stopped
- Good for buffer/cache/scratch data/temporary content
- Risk of data loss if hardware fails
- Backups & replication are our responsibility
- High performance - local EC2 Instance store

Elastic File System (EFS)

- Managed NFS that can be mounted on 100s of EC2
- EFS works with linux EC2 instances in multi AZ
- Highly available, scalable, expensive, pay per use, no capacity planning

Note: EBS vs EFS - Single vs Multiple AZ

EFS Infrequent Access (EFS -IA)

- Storage class that is cost optimized for files not accessed every day
- Upto 92% lower cost compared to standard
- Automatically move your files to EFS-IA
- Enable EFS-IA with a lifecycle policy

Amazon FSx

- Launch 3rd party high-performance file systems on AWS
- Fully managed service

Amazon FSx for windows

- Fully managed, highly reliable, and scalable windows native shared file system
- Built on windows file system
- Supports SMB protocol & windows NTFS
- Integrated with Microsoft active directory
- Can be accessed from AWS or our on-premise infrastructure

Amazon FSx for Lustre

- A fully managed, high - performance, scalable file storage for high performance computing (HPC)
- Linux + Cluster = Lustre
- Machine learning, analytics, video processing, financial modeling
- Scales up to 100s GB/s millions of 10PS, sub-ms latencies

Section 5

ELB & ASG

Scalability

- Scalability means handling greater loads by adapting
- Two kinds of scalability
    - Vertical
        - Means increasing the size of the instance
        - Common for DBs
        - Hardware limit
    - Horizontal
        - Means increasing number of instances
        - Distributed systems
        - Common for web application, eg - EC2
- Scale in & Out

Availability

- Means running in at least 2 AZs
- Goal - survive a data center loss

Elasticity

- Auto scaling so the system can scale based on the load

Agility

- Easily available IT resources to reduce time

Elastic Load Balancer (ELB)

- Load balancing the traffic to multiple servers
- Managed load balancer by AWS
- Handle failures of downstream instances
- High availability across zones
- 3 kinds:
    - **Application Load balancer** (HTTP/HTTPS only) - layer 7
    - **Network Load balancer** (Ultra high performance, allows for TCP) - layer 4
    - **Classic Load balancer** - Layer 4&7
    - **Gateway Load balancer** - new service: Gateway Load Balancer helps you easily deploy, scale, and manage your third-party virtual appliances

Auto Scaling Groups (ASG)

- Scale in/out to match the load & replace unhealthy instances
- Cost savings
- Need to define min & max range
- Can work with ELB

ASG Strategies

- Manual Scaling: manually update the size
- Scheduled Scaling: Anticipate a scaling based on known usage pattern
- Dynamic Scaling:
    - Simple/step scaling: When this then add/remove
    - Target tracking: I want the average ASG CPU to stay at around 40%
- Predictive Scaling: Uses ML to predict future traffic

Section 6

S3

- Store “objects” (files) in “buckets” (directories)
- Buckets = globally unique name but limited to regions
- Objects have a key which is the full path = prefix + object name
- Max object size is 5TB (5000GB)
- >5GB, use multi use upload

Used for

- Infinitely scaling
- Backup & restore

- Disaster recovery
- Archive
- Hybrid cloud storage
- Static websites

S3 Security

- User based
  - IAM policies
- Resource based
  - Bucket policies - Bucket wide rules from the S3 console, allows cross account
  - Object access control list (ACL) - finer grain
  - Bucket ACL - less common
  - An IAM principal can access an S3 object if:
    - The user IAM permissions allow it OR the resource policy allows it
    - AND there's no explicit Deny
  - Encryption via encryption keys

Bucket Policies

- Allows public access
- Can be done via JSON policies

S3 Websites

- Static websites on public www
- Need to make bucket public

S3 - Versioning

- Enabled at bucket level
- Best practice to version buckets
  - Protect against unintended deletes
  - Easy roll back to previous version

S3 Access Logs

- Log all access to S3 buckets in another S3 bucket (logging bucket)

S3 Replication (CRR & SRR)

- Must enable versioning in source & destination
- Copy is async
- CRR - use cases: compliance, lower latency access, replication across accounts
- SRR - use cases: log aggregation, live replication between prod & test accounts

S3 Storage Classes

Durability

- How often you lose a file
- Same for all storage classes

Availability

- Measures how readily available a service is

Different Classes

Parameter	Standard	Intelligent tiering	Standard IA	One zone IA	Glacier	Glacier Deep Archive
Durability	99.999999 999%	99.999999 999%	99.999999 999%	99.999999 999%	99.999999 999%	99.999999 999%
Availability	99.99%	99.99%	99.99%	99.5%	99.99%	99.99%
Use case	Big data analytics	Frequently + less frequently both cases	For less frequently accessed data but required rapid when needed	Storing backup copies or data which you can recreate	Very low cost but high archival time, data is retained for longer time, takes 1 min to 12 hours hours for retrieval	Lowest cost but highest archival time, data is retained for longer time, takes 12-48 hours for retrieval
Minimum storage duration charge	NA	NA	30 Days	30 Days	90 Days	180 Days
Retrieval charge	NA	NA	per GB retrieved	per GB retrieved	per GB retrieved	per GB retrieved

S3 - Moving between storage classes

- Can transition objects between storage classes
- Moving objects can be automated using a lifecycle configuration

S3 Locks

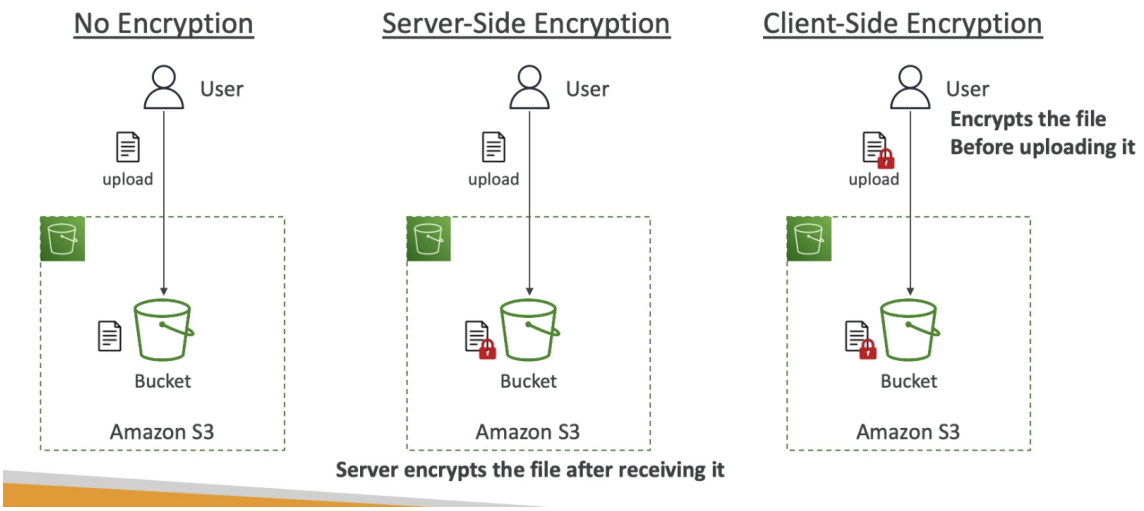
S3 Object Lock

- Adopt a WORM (Write once read many) model
- Block an object version deletion for a specified amount of time

Glacier Vault Lock

- WORM
- Lock the policy for future edits
- Helpful for compliance & data retention

S3 Encryption



AWS Snow Family

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

Two use cases:

Data Migration - Snowcone, Snowball edge, Snow mobile
Edge Computing - Snowcone, snowball edge

Note - if it takes more than a week to transfer over the network, use snowball device

Data Migration

Snowball Edge

- TBs or PBs data in or out
- Pay per data transfer job
- Snowball edge storage optimized
    - 80TB of HDD capacity for block volume & S3 compatible object storage

- Snowball edge compute optimized
    - 42TB of HDD capacity for block volume & S3 compatible object storage
- Use cases: Large data cloud migrations, DC decommission, disaster recovery

Snowcone

- Small, portable computing, anywhere, rugged & secure
- 4.5 pounds/2.1 kg
- Used for edge computing, storage & data transfer
- 8TBs of usage storage
- Can be connected to internet & use AWS datasync to send data

Snowmobile

- Actual Truck!
- 1EB = 1000PB = 1,000,000TB
- 100PB of capacity
- High security, GPS, 24/7 surveillance

Edge Computing

- Process data while it's being created on an edge location (Any location which is far from cloud, doesn't have internet)
- We setup a snowball edge/snow cone device to do edge computing
- Use cases: Preprocess data, ML at the edge, Transcoding media streams

Snowcone

- 2 CPU, 4GB 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 ML)
- 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)

AWS OpsHub

- Software to install to manage snow family devices

Hybrid Cloud Storage

AWS Storage Cloud Native Options

AWS Storage Gateway

- Bridge between on-premises data and cloud data in S3
- Service to allow on-premises to seamlessly use the AWS cloud

Section 7

Databases & Analytics

Database Intro

Relational Databases

- Excel but with links between them
- Can use SQL language to perform queries, lookups

NoSQL Databases

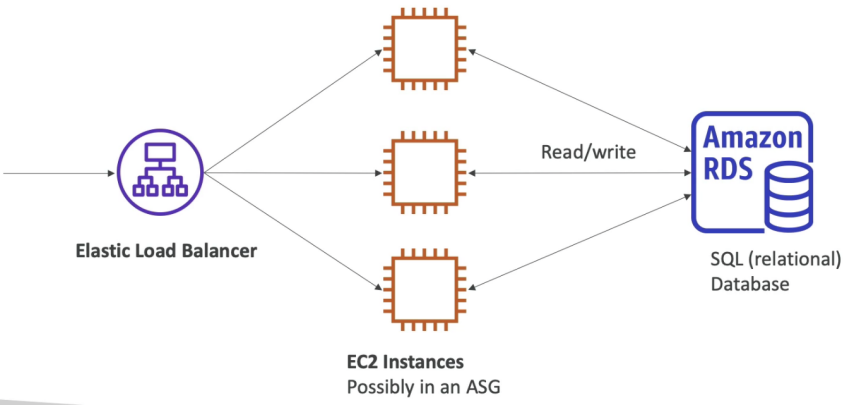
- Purpose built for specific data models and have flexible schemas for building modern applications

Amazon RDS

- SQL

- Create databases in cloud managed by AWS: Postgres, Mysql, MariaDB, Oracle, Microsoft SQL, Aurora (AWS)
- Great in all cases except that you can't SSH into your instances

RDS Solution Architecture

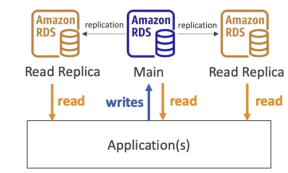


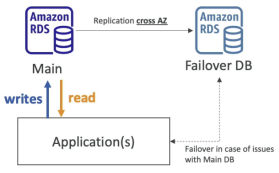
Amazon Aurora

- Not open source
- Postgres & MySQL both supported
- "AWS cloud optimized" = 5 X performance improvements of MySQL over RDS & 3X performance of postgres over RDS
- Automatically grows in increments of 10GB, up to 64TB
- 20% more cost than RDS but more efficient
- Not in free tier

RDS Deployments Options

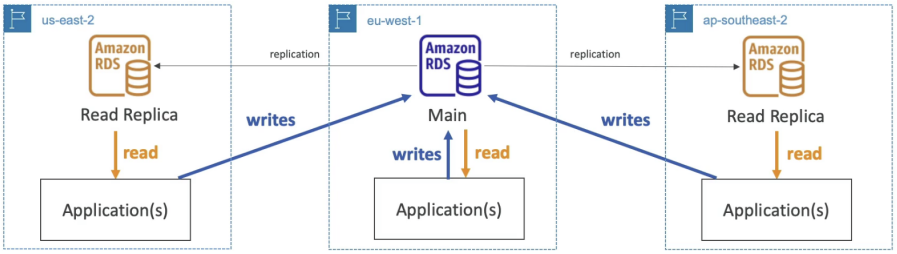
- Read Replicas
  - Scale the read workload of your DB
  - Can create up to 5 read replicas
  - Data is only written to the main DB
- Multi AZ
  - Failover in case of AZ outage
  - Data is read/written to the main database
  - Can have only 1 other AZ as failover





- RDS Deployments: Multi Region (Read replicas)
  - Disaster recovery in the case of region failure

- Local performance for the global reads
- Replication cost



Read replicas, Multi-AZ deployments, and multi-region deployments

Amazon RDS read replicas complement Multi-AZ deployments. While both features maintain a second copy of your data, there are differences between the two:

Multi-AZ deployments	Multi-Region deployments	Read replicas
Main purpose is high availability	Main purpose is disaster recovery and local performance	Main purpose is scalability
Non-Aurora: synchronous replication; Aurora: asynchronous replication	Asynchronous replication	Asynchronous replication
Non-Aurora: only the primary instance is active; Aurora: all instances are active	All regions are accessible and can be used for reads	All read replicas are accessible and can be used for readscaling
Non-Aurora: automated backups are taken from standby; Aurora: automated backups are taken from shared storage layer	Automated backups can be taken in each region	No backups configured by default
Always span at least two Availability Zones within a single region	Each region can have a Multi-AZ deployment	Can be within an Availability Zone, Cross-AZ, or Cross-Region
Non-Aurora: database engine version upgrades happen on primary; Aurora: all instances are updated together	Non-Aurora: database engine version upgrade is independent in each region; Aurora: all instances are updated together	Non-Aurora: database engine version upgrade is independent from source instance; Aurora: all instances are updated together
Automatic failover to standby (non-Aurora) or read replica (Aurora) when a problem is detected	Aurora allows promotion of a secondary region to be the master	Can be manually promoted to a standalone database instance (non-Aurora) or to be the primary instance (Aurora)

Amazon ElastiCache

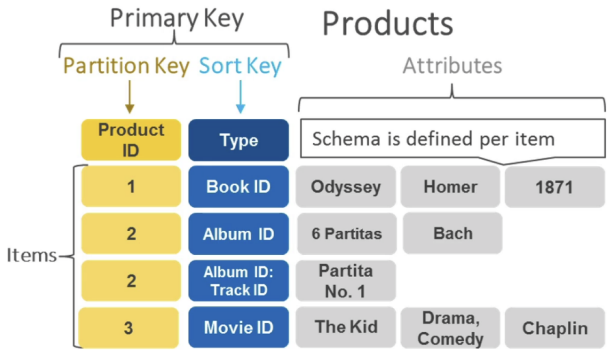
- To get managed redis or memcached
- In-memory databases with high performance, low latency
- Helps **reduce load off databases** for read intensive workloads

---

DynamoDB

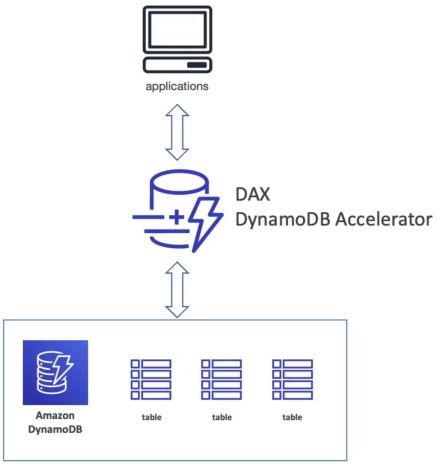
- Fully managed highly available with replication across 3 AZ
- NoSQL

- Scales to massive workloads
- Distributed “serverless” database
- Single digit millisecond latency - low latency retrieval
- Integrated with IAM for security, authorization & administration
- Low cost
- Auto scaling
- It is a key/value pair database



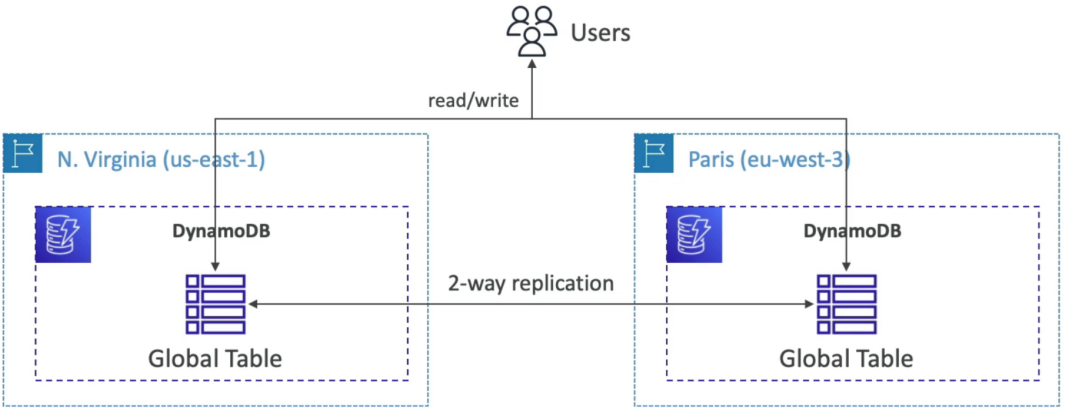
DynamoDB Accelerator - DAX

- Fully managed in-memory cache for DynamoDB
- 10 times performance improvement
- Secure, highly scalable & highly available
- Note: Elasticache can be used for other databases but DAX can only be used for DynamoDB.



DynamoDB - Global tables

- Make a DynamoDB table accessible with low latency in multiple regions
- Active - Active replication



Redshift

- PostgreSQL based
- OLAP (Online Analytical Processing) - **Analytics and data warehousing**
- Load data once every hour
- 10 times better performance
- Columnar storage of data
- Massive parallel Query Execution, highly available
- Has a SQL interface or performing the queries
- BI tools such as AWS Quicksight or Tableau integrated with it

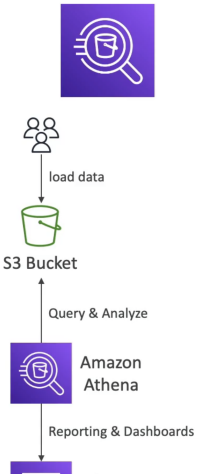
---

EMR (Elastic MapReduce)

- Helps creating **Hadoop Clusters** to analyze/process large amount of data
- Clusters can be made of 1000s of EC2 instances
- Also supports Apache Spark, HBase, Presto, Flink
- EMR takes care of the provisioning/configuration of these instances
- Autoscaling & integrated with spot instances
- Use Cases: Data Processing, M/L, Web indexing, Big data

Athena

- Serverless query service to perform analytics against S3 objects
- Uses standard SQL Language
- Supports CSV, JSON, ORC, Avro and Parquet
- $5 per TB



- Use compressed or columnar data for cost savings
- Use cases: Business Intelligence/ analytics/ reporting, logs
- Exam tip: **Analyze data in S3 serverless SQL**

QuickSight

- Serverless machine learning-powered BI service to create **interactive dashboards**
- Fast, automatic scalable, embeddable
- Per-session pricing

DocumentDB

- It is AWS Implementation of **MongoDB** (NoSQL database)
- MongoDB is used to store, query and index JSON data
- Deployment structure like aurora

Neptune

- Fully managed **graph database** example: social network, wikipedia
- 3 AZs upto 15 read replicas
- Highly connected datasets

Amazon QLDB (Quantum Ledger Database)

- Used to review **history of all the changes made** to your application data
- Serverless, 3 AZs
- Immutable System: no entry removal or modification
- **No decentralized component** (different from Amazon managed blockchain)

Amazon managed Blockchain

- Possible to build applications where multiple **parties can execute transactions** without the need for a trusted, central authority
- It is a managed service to:
    - Join public blockchain networks
    - Or create your own scalable private network

- Compatible with the framework Hyperledger Fabric & Ethereum

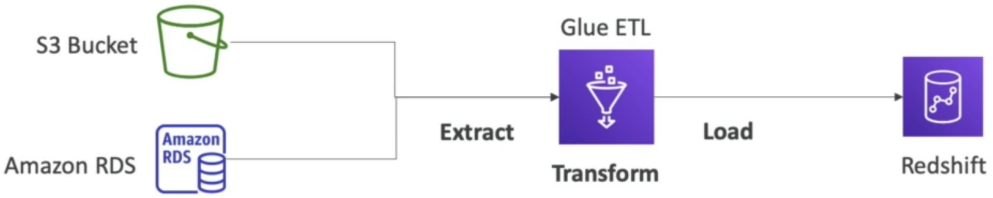
---

Database Migration System (DMS)

- Quickly and securely **migrate databases** to AWS, resilient, self healing
- The source database remains available during the migration
- Supports both homogeneous and heterogeneous migration

AWS Glue

- Useful to prepare & transform data for analytics
- Managed extract, transform and load (**ETL**) service
- Fully serverless



Glue Data Catalog - catalog of datasets (can be used for EMR, Athena, Redshift)

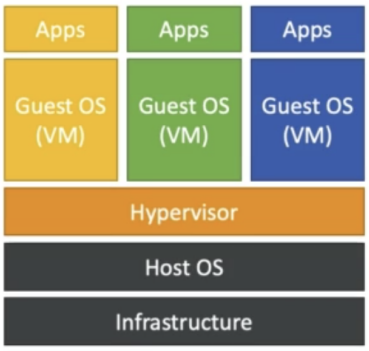
Section 8

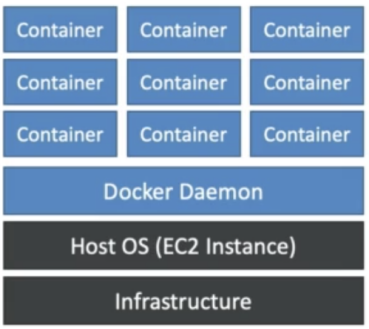
Other Compute Services

Docker

- Docker is a software development platform to deploy apps
- Apps are packaged in containers that can be run on any OS
- Scale containers up & down very quickly
- Docker images are stored in docker repositories

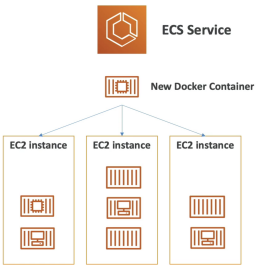
Docker vs Virtual machines





Elastic Container Service (ECS)

- Use to launch docker containers on aws
- We need to create & maintain infrastructure (EC2 instances)
- AWS takes care of starting/stopping of containers

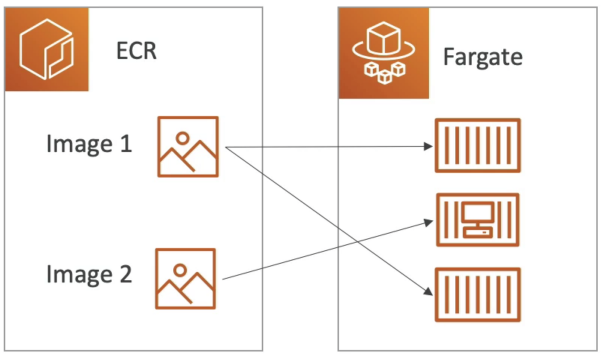


Fargate

- Use to launch docker containers on aws
- No need to create EC2 instances
- Serverless offering

Elastic Container Repository (ECR)

- Private Docker registry on AWS



Serverless

- No need to manage servers just deploy the codes directly
- Automatically scales according to the load

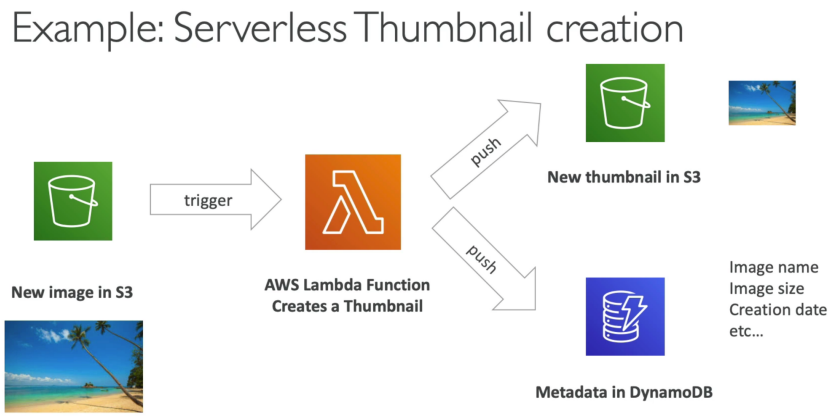
AWS Lambda

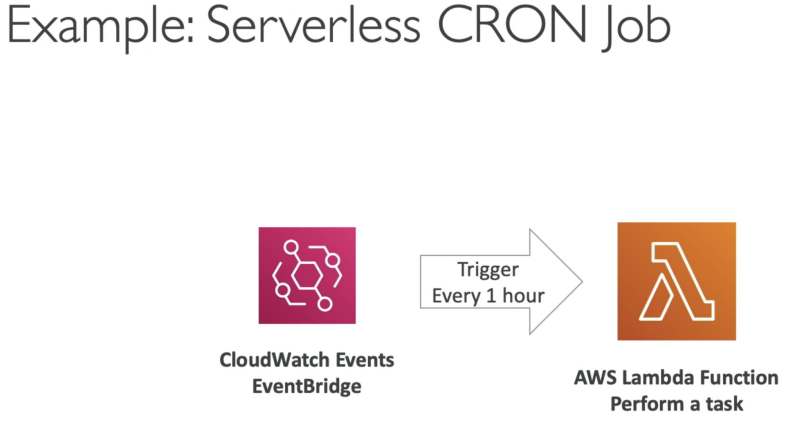
- Functions as a service
- Virtual functions
- Limited by time
- Run on demand
- Scaling is automated

Benefits

- Easy Pricing
- Integrated with whole AWS suite
- Event driven
- Integrated with many programming languages
- Easy monitoring through CloudWatch
- Easy to get more resources per functions
- Increasing RAM will also improve CPU and network

Lambda container Image: Lambda can't run all the docker images but can run a few specific ones which implement the lambda Runtime API
*ECS/Fargate are preferred to run Docker images



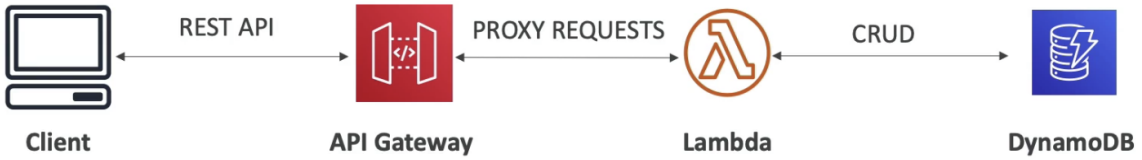


Lambda Pricing

- Pay per calls:
  - First 1,000,000 requests are free
  - $0.20 per 1 million req thereafter
- Pay per duration
  - 400,000 GB sec of compute time per month in free tier
  - $1 for 600,000 GB-seconds thereafter

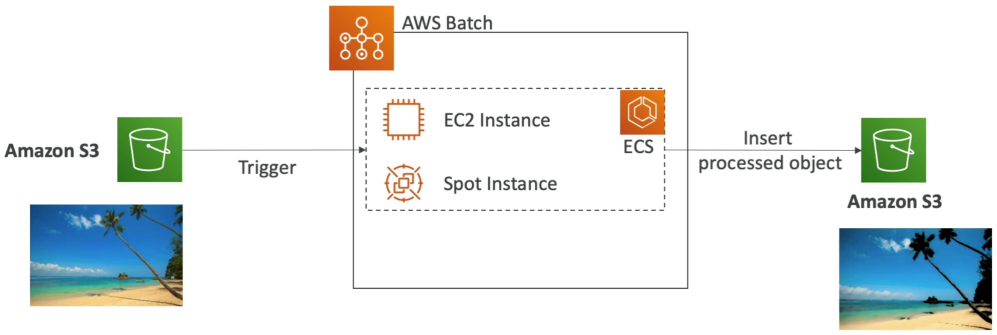
Amazon API Gateway

- Fully managed service for developers to easily create, publish, maintain, monitor and secure APIs
- Serverless & scalable
- Supports REST APIs & WebSocket APIs



AWS Batch

- Fully managed batch processing at any scale.
- Batch = opposite to continuous
- AWS Batch will launch EC2 instances or spot instances we just need to submit & schedule batch jobs
- Batch jobs are defined as docker images and run on ECS
- Helpful for cost optimization



Batch vs Lambda

Batch	Lambda
No time limit	Time limit
Any runtime as long as it's packaged as a Docker image	Limited runtime
Rely on EBS / instance store for disk space	Limited temporary disk space
Relies on EC2	Serverless

Amazon Lightsail

- Great for people with little experience with cloud, alternative to EC2, RDS, ELB, EBS,..
- Virtual servers, storage, databases and networking
- Low & predictable pricing
- Has high availability but no auto scaling, limited aws integrations
- Use cases: simple web applications or websites deployment

Section 9

Deployments & managing Infrastructure at scale

CloudFormation

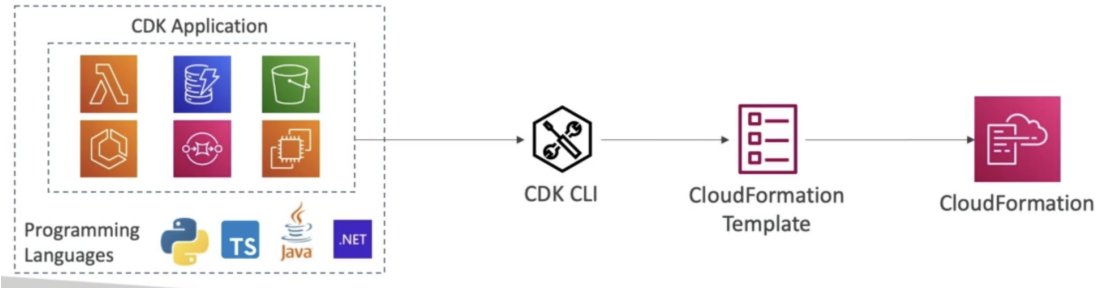
- CloudFormation is a declarative way of outlining your AWS Infrastructure
- Example template:
    - I want a security group
    - I want two EC2 instances using this security group
    - I want an S3 bucket
    - I want a load balancer in front of these machines
- CloudFormation creates these for you in the right order with the exact configuration that you specify

Benefits

- Infrastructure as a code: reviews through code
- Cost: Easy estimation and management of cost, saving strategy
- Productivity: Easy to create/destroy, automated generation of diagram for the templates, declarative programming
- Leverage the existing documentation/templates on the web
- Supports almost all the AWS resources (custom resources also available)

AWS Cloud Development Kit (CDK)

- Define your cloud infrastructure in familiar language (JS, Python, Java)
- Code is then compiled into a CloudFormation template (JSON/YAML)



AWS Elastic Beanstalk

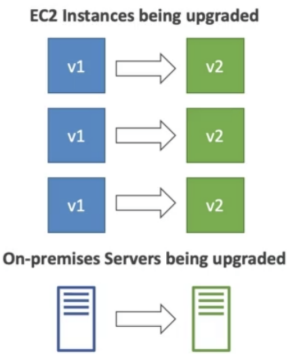
- Platform as a service
- Developer centric view of deploying an application on AWS
- Managed service(just the application code is handled by developer)
- Three architectural models:
    - Single instance development: good for dev
    - LB + ASG: great for production or pre-production web applications
    - ASG only: great for non-web applications (workers etc.)

Health monitoring

- Health agent pushes metrics to CloudWatch
- Checks for app health, publishes health events

AWS CodeDeploy

- Deploy applications automatically
- Works with EC2 Instances
- Hybrid Service
- Servers/Instances must be provisioned and configured ahead of time with the CodeDeploy agent



AWS CodeCommit

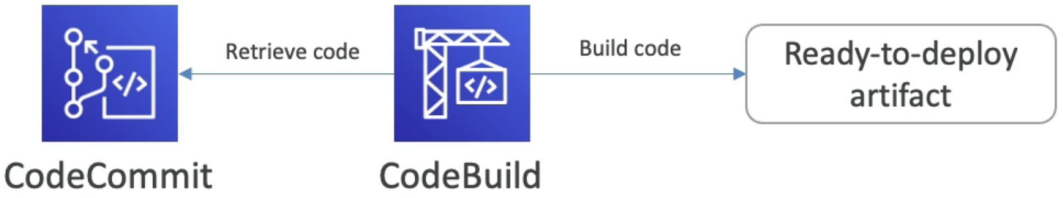
- Before pushing the code to the servers, it needs to be stored somewhere
- CodeCommit
    - Source-control service that hosts Git-based repositories
    - Easy to collaborate
    - Code changes are automatically versioned

Benefits

- Fully managed
- Scalability & availability
- Private, secure and integrated with AWS

AWS CodeBuild

- Code building service - Compiles source code, run tests, and produces packages

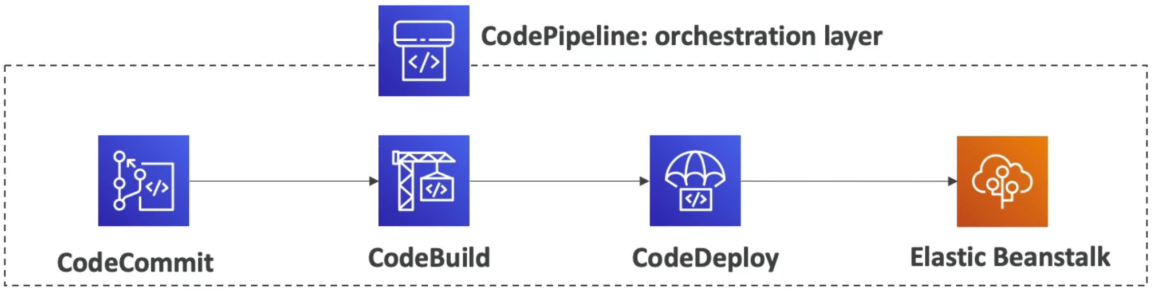


Benefits

- Fully managed, serverless
- Continuously scalable, highly available, Secure
- Pay-as-you-go

AWS CodePipeline

- Orchestrate the different steps to have the code automatically pushed to prod
- Basis of CICD



Benefits

- Fully managed
- Compatible with many services
- Fast delivery & rapid updates

AWS CodeArtifact

- Storing and retrieving code dependencies is called artifact management
- CodeArtifact is a secure, scalable and cost effective artifact management for software development

---

AWS CodeStar

- Unified UI to easily manage software development activities in one place
- Quick way to set up CodeCommit, CodePipeline, CodeBuild, CodeArtifact, Elastic Beanstalk, EC2 etc

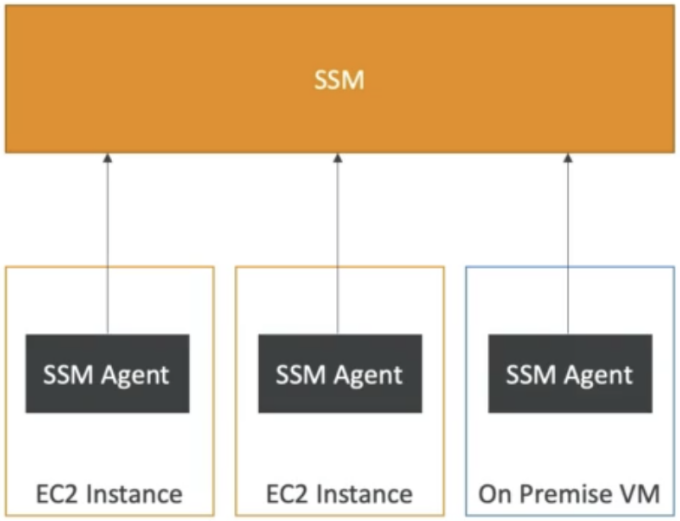
---

AWS Cloud9

- It is a cloud IDE for writing, running and debugging code
- Can be used within a web browser
- Real time code collaboration

AWS Systems Manager (SSM)

- Helps you manage your EC2 and on-premises systems at scale
- Hybrid service
- Suite of 10+ products
- Most important features are:
    - Patching automation
    - Run commands across an entire fleet of servers
    - Store parameter configuration with the SSM parameter store
- Works for both windows and linux



SSM Session Manager

- Allows you to start a secure shell on your EC2 and on-premises servers without SSH access
- Supports windows, macOS, Linux
- Send session log data to S3 or CloudWatch logs

AWS OpsWorks

- Alternative to AWS SSM
- For Managed Chef & Puppet (helps you perform server configuration automatically or repetitive actions)
- Only provision standard AWS resources

Section 10

Leveraging the AWS Global Infrastructure

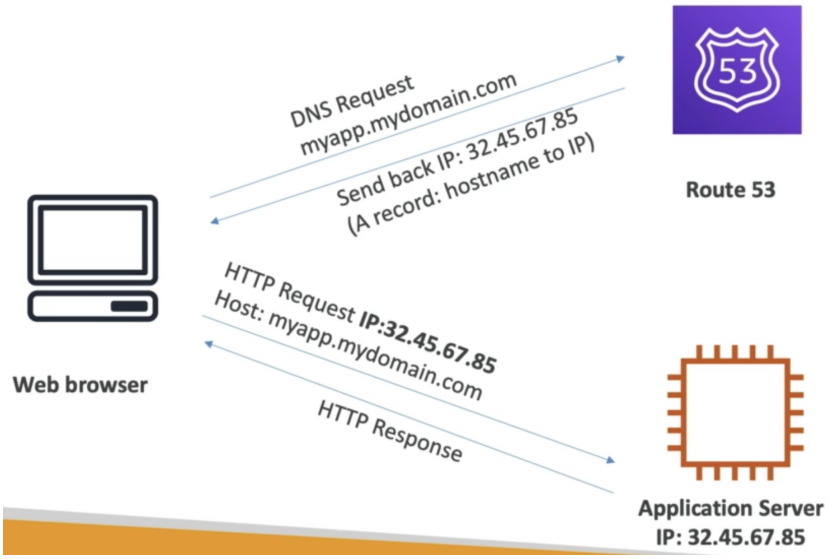
Why Global?

- Decreased latency
- Disaster recovery
- Attack protection

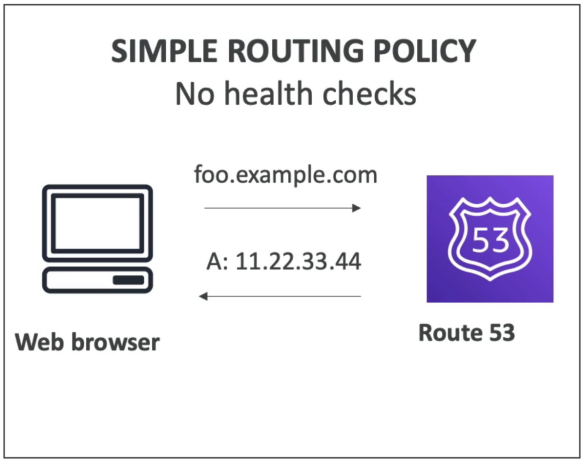
Amazon Route 53

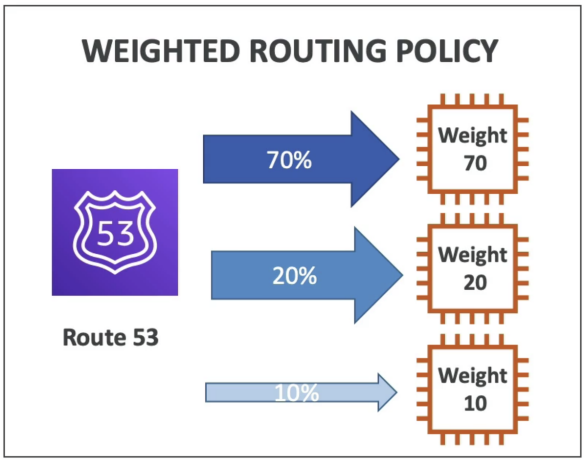
- Managed DNS

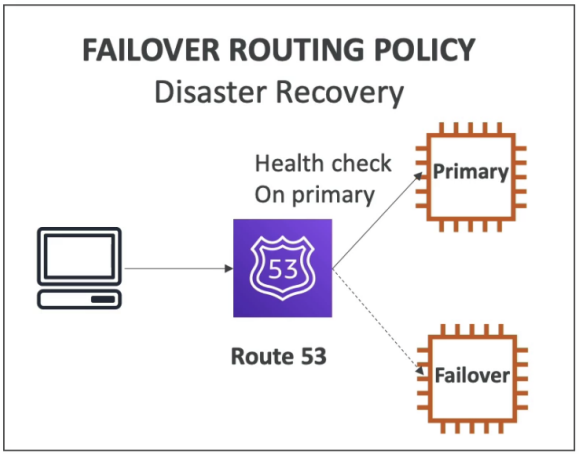
- DNS is a collection of rules and records which helps clients understand how to reach a server through URLs

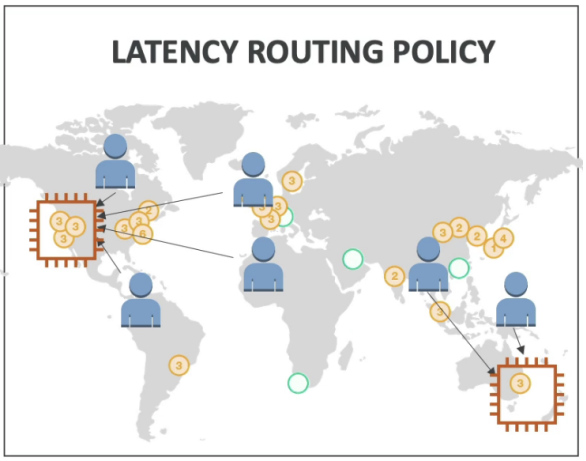


Route 53 Routing Policies









Choosing a routing policy

PDF | Kindle | RSS

When you create a record, you choose a routing policy, which determines how Amazon Route 53 responds to queries:

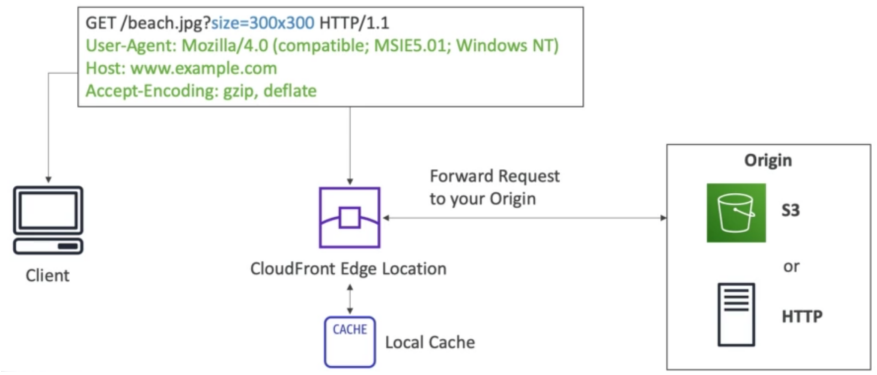
- **Simple routing policy** – Use for a single resource that performs a given function for your domain, for example, a web server that serves content for the example.com website.
- **Failover routing policy** – Use when you want to configure active-passive failover.
- **Geolocation routing policy** – Use when you want to route traffic based on the location of your users.
- **Geoproximity routing policy** – Use when you want to route traffic based on the location of your resources and, optionally, shift traffic from resources in one location to resources in another.
- **Latency routing policy** – Use when you have resources in multiple AWS Regions and you want to route traffic to the region that provides the best latency.
- **Multivalue answer routing policy** – Use when you want Route 53 to respond to DNS queries with up to eight healthy records selected at random.
- **Weighted routing policy** – Use to route traffic to multiple resources in proportions that you specify.

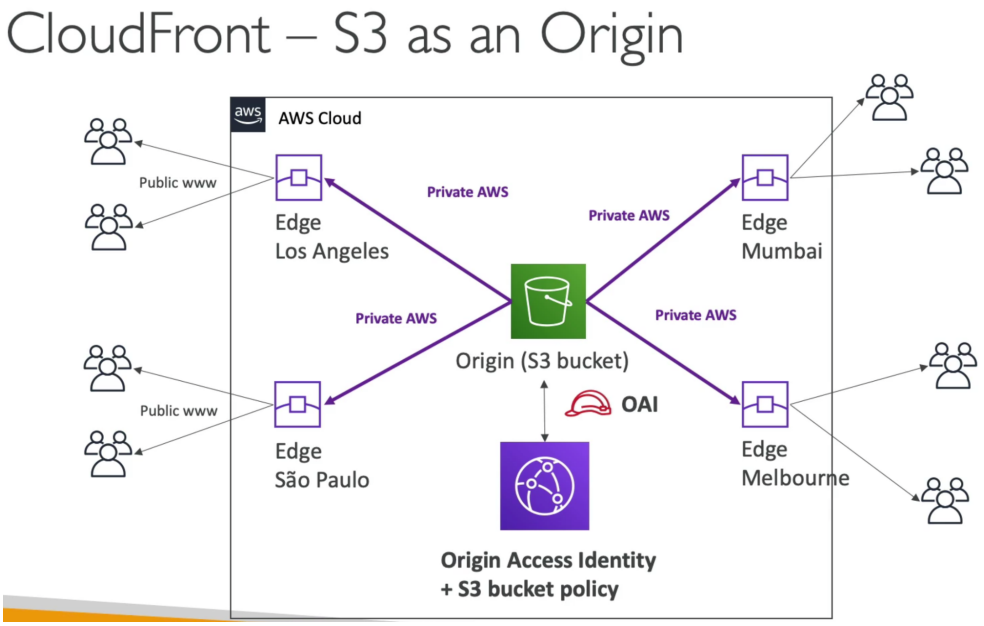
AWS CloudFront

- Content delivery network (CDN)
- Improves read performance, content is cached at the edge
- Improved user experience
- 216 Point of Presence worldwide (edge locations)

Origins

- S3 bucket
- Custom Origin (HTTP)





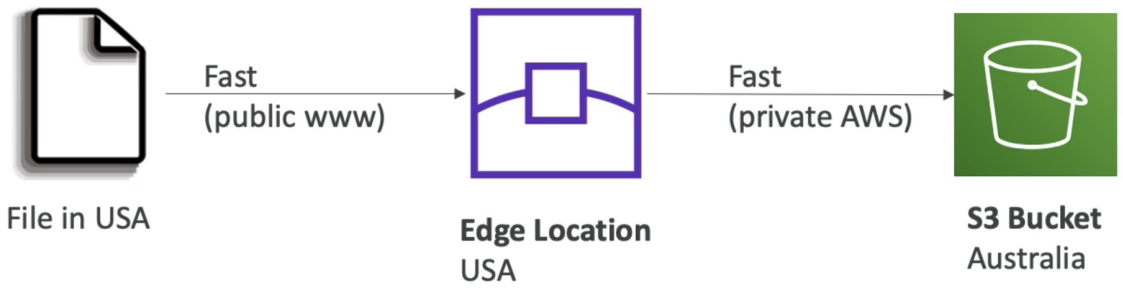
CloudFront vs S3 Cross Region Replication

CloudFront	**S3 CRR**
Global edge network (216 ELs)	Must be setup for each region
Files cached for TTL	Files are updated in near real-time
Great for static content that must be available everywhere	Great for dynamic content that needs to be available at low latency in few regions

---

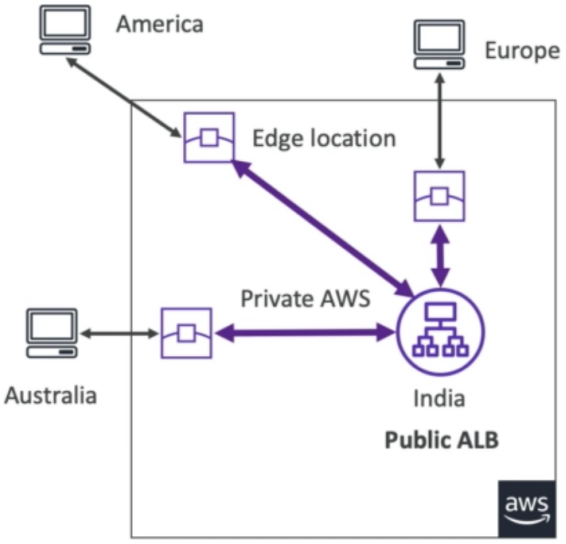
S3 Transfer Acceleration

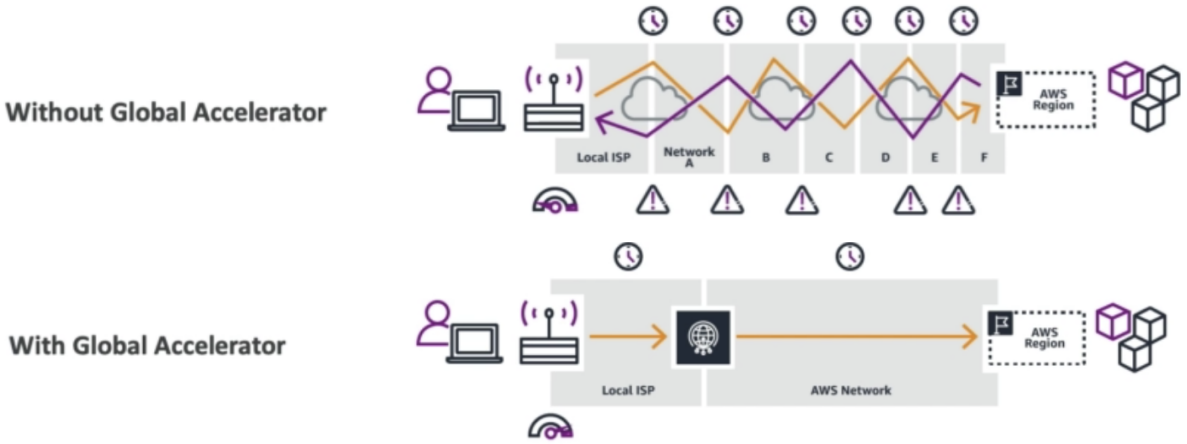
- Increase transfer speed by transferring file to an AWS edge location which will forward the data to the S3 bucket in the target region



AWS Global Accelerator

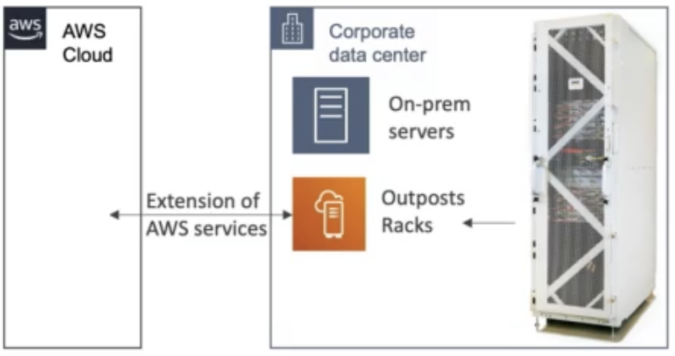
- Improve global application availability & performance using the AWS global network
- Upto 60% improvement
- 2 anycast IP are created for your application & traffic is sent through ELs
- Different from cloudFront because no caching





AWS Outposts

- These are server racks that offer the same AWS infrastructure, services, APIs & tools to build your own applications on-premises just as in the cloud.
- AWS will set up and manage Outposts racks within your on-premises infrastructure and you can start leveraging AWS services on-premises.
- You are responsible for the physical security of racks.



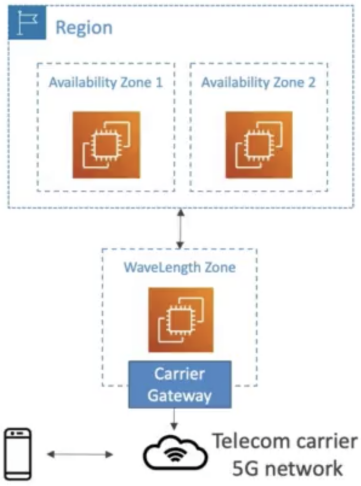
Benefits

- Low latency
- Local data processing
- Data residency

- Easier migration from on-premises to the cloud
- Fully managed service
- EC2, EBS, S3, EKS, ECS, RDS, EMR work with outposts

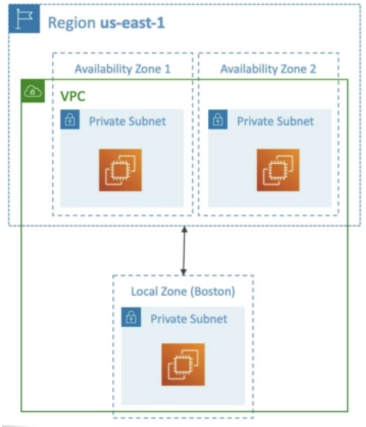
AWS Wavelength

- WaveLength Zones are infrastructure deployments, embedded within the telecommunications providers' datacenters at the edge of 5G networks.
- Brings AWS services to the edge of the 5G networks
- Ultra low latency
- No additional charges
- Use case: smart cities, ML-assisted diagnostics etc.



AWS Local Zones

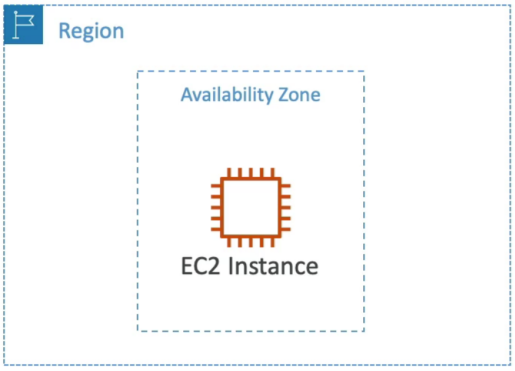
- Places AWS compute, storage, database and selected AWS services closer to end users to run latency sensitive applications



Global Application Architecture

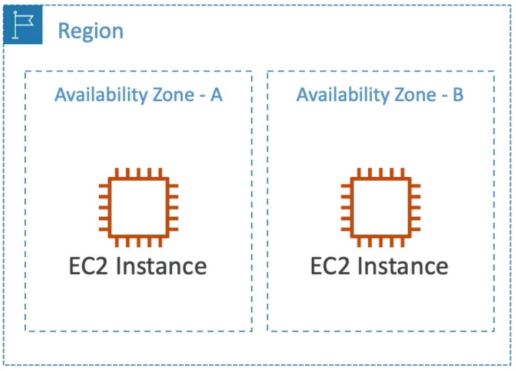
Single Region, **Single** AZ

- High Availability
- Global Latency
- Difficulty



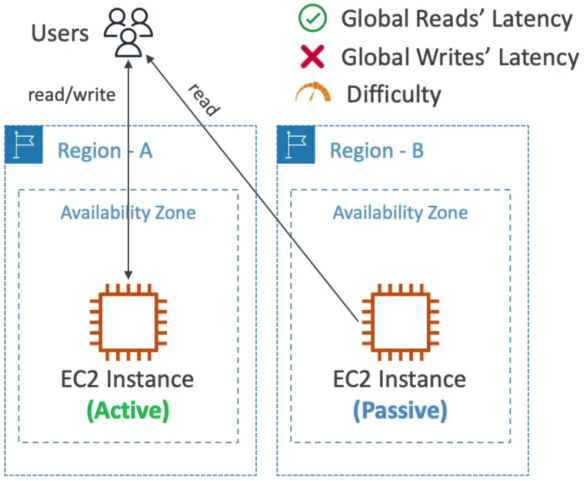
Single Region, **Multi** AZ

- High Availability
- Global Latency
- Difficulty



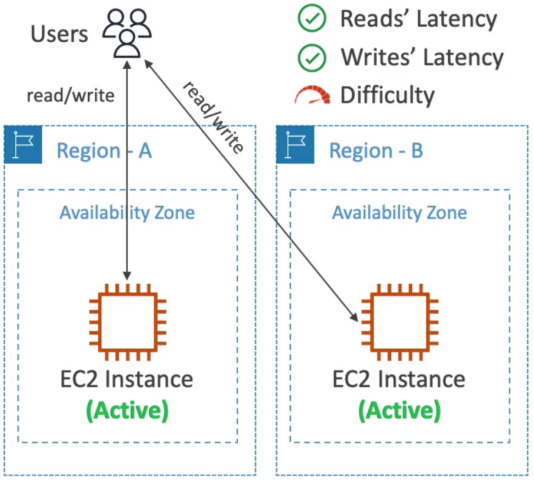
Multi Region, **Active-Passive**

- Global Reads' Latency
- Global Writes' Latency
- Difficulty

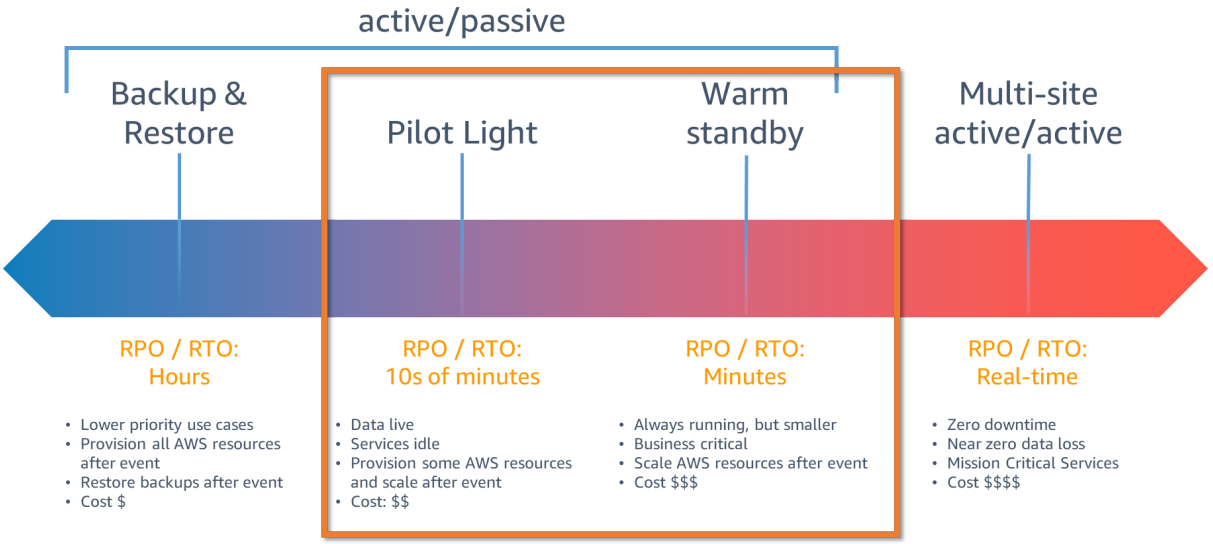


Multi Region, **Active-Active**

- Reads' Latency
- Writes' Latency
- Difficulty



Disaster Recovery Strategies



Section 11

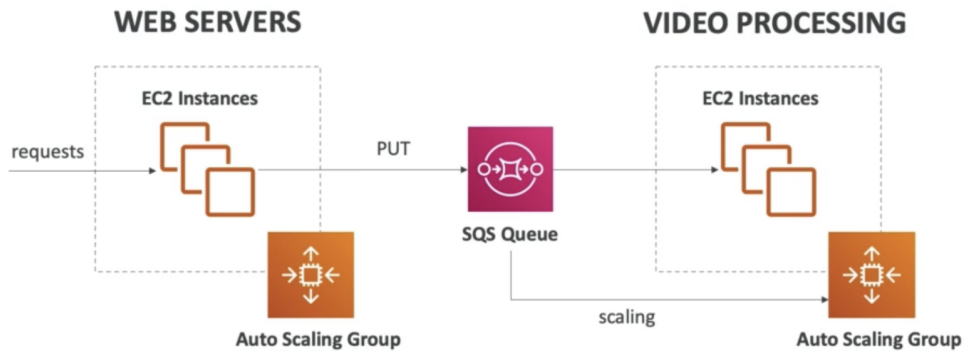
Cloud integrations

There is 2 types of communications:

1. Synchronous: fails in sudden spikes of traffic
2. Asynchronous or decoupled: can scale independently from our application

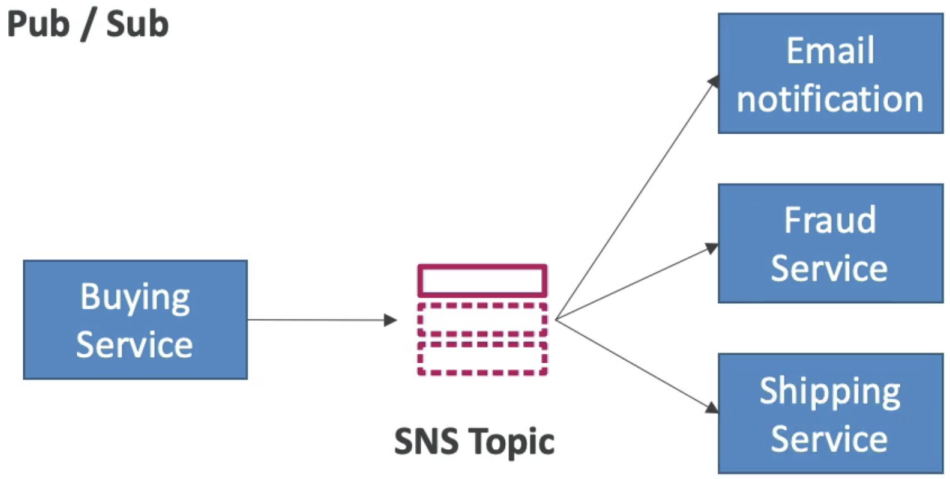
Amazon Simple Queue Service (SQS)

- Oldest AWS service
- Fully managed, used to **decoupled** applications
- No limit on how many msgs can be in the queue
- Msgs can be deleted after they are read by consumers
- Low latency
- Consumers share work to read msgs & scale horizontally



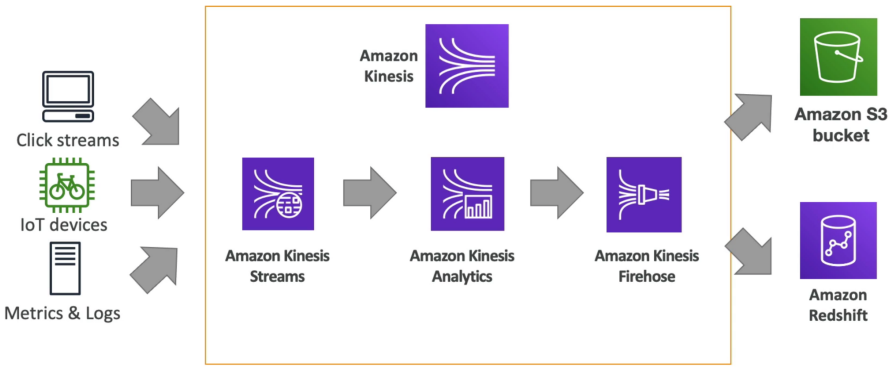
Amazon Simple Notification Service (SNS)

- The "event publishers" will only send messages to one SNS topic.
- Can have as many "event subscribers" as you want to listen to the SNS topic notifications.
- Each subscriber to the topic will get all the messages
- SNS subscribers can be:
    - HTTP/HTTPS (with delivery retries - how many times)
    - Emails, SMS, mobile notifications
    - SQS queues, Lambda functions



Kinesis

- Managed service to collect, process & analyze real-time big data streaming at any scale



Amazon MQ

- When a company migrating to cloud -> use amazon MQ because SQS & SNS are cloud native
- Not serverless
- Doesn't scale as much
- = managed apache ActiveMQ
- Has features of both SQS & SNS

Section 12

Cloud Monitoring

Amazon CloudWatch Metrics

- CloudWatch provides metrics for *every* service in AWS
- Can create CloudWatch dashboards of metrics

Important metrics

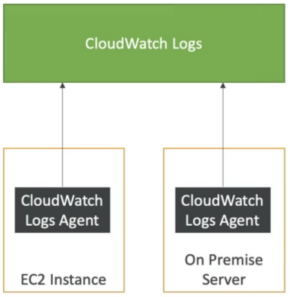
- EC2 Instances: CPU utilization, Status checks, Network (Not RAM)
- EBS Volumes: Disk read/writes
- S3 Buckets: BucketSizeBytes, NumberOfObjects, AllRequests
- Billing: Total estimated charge (only in us-east-1)
- Service limits: How many you have been using a service API

CloudWatch Alarms

- Used to trigger notifications for any metric
- There can be specific alarm actions
- Alarms states: OK, INSUFFICIENT_DATA, ALARM

CloudWatch Logs

- CloudWatch can collect logs from beanstalk, ECS, EC2, CloudWatch log agent, AWS lambda, CloudTrail, Route53



Amazon CloudWatch Events/EventBridge

CloudWatch Events

- Schedule cron jobs (scheduled scripts)

- Event pattern: event rules to react to a service doing something

- Trigger lambda function, send SNS, SQS messages

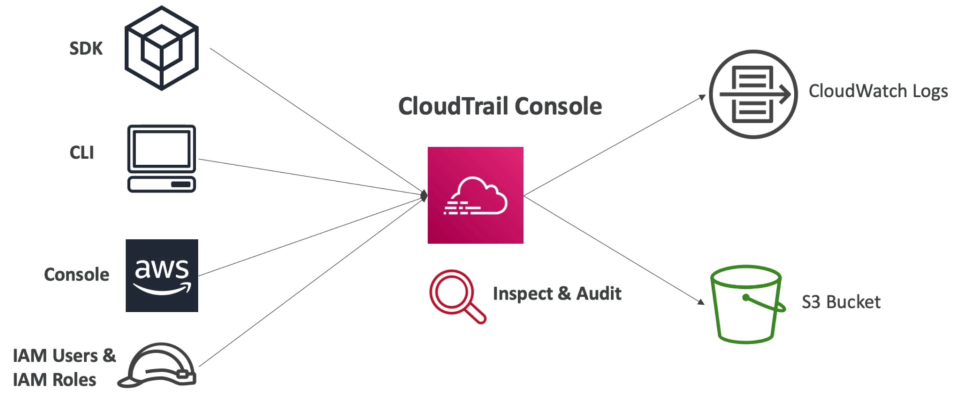
Event Bridge

- Next evolution of CloudWatch Events
- Extra functionality
    - Partner event bus: receive events from SaaS service or applications Zendesk, datalogs
    - Custom event buses: For our own applications
    - Schema Registry: Model event schema

Note: From exam perspective, both of these are same

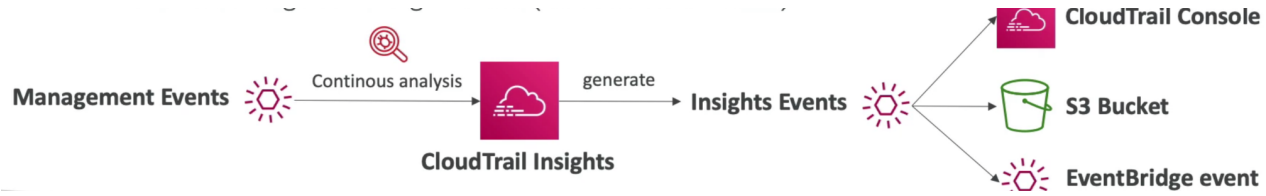
AWS CloudTrail

- Provides governance, compliance and audit for our AWS account
- Enabled by default
- Used to get a history of events/API calls made within our AWS account
- Can put logs from CloudTrail to CloudWatch Logs or S3
- Can be applied to all regions/single region



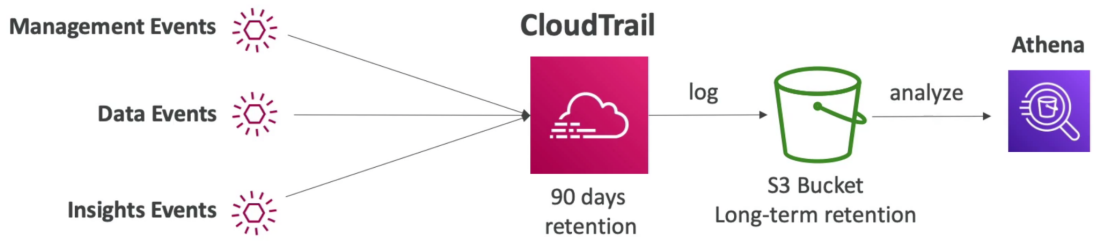
CloudTrail Events

1. Management Events:
    a. Operations that are performed on the resources in our AWS account
    b. By default trails are configured to log management events
    c. Can separate write events from read events
2. Data Events:
    a. By default not logged
    b. Amazon S3 object-level activity
    c. Lambda function execution activity
    d. Can separate write events from read events
3. CloudTrail Insights Events
    a. Enable it to detect unusual activity in your account
    b. Continuously analyzes **write** events to detect unusual patterns



CloudTrail Events Retention

- Events are stored for 90 days in CloudTrail
- To keep them longer, log them to S3 & use athena



AWS X-Ray

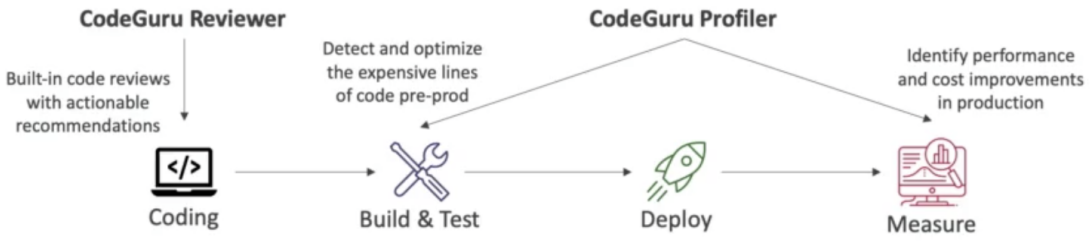
- Visual analysis of our applications

Benefits

- Troubleshooting performance (bottlenecks)
- Understanding dependencies in microservices architecture
- Find errors & exceptions
- Where am I throttled?
- Are we meeting time SLA?
- Identify users that are impacted

Amazon CodeGuru

- An ML powered service for automated code reviews (**CodeGuru Reviewer**) and application performance recommendation during production (**CodeGuru Profiler**)



AWS Status - Service Health Dashboard

- Shows all regions, all services health
- Has an RSS feed to subscribe
- Shows data for each day

AWS Personal Health Dashboard

- AWS Personal Health Dashboard provides **alerts and remediation guidance** when AWS is experiencing event that may impact you (related to the AWS services underlying your AWS resources)

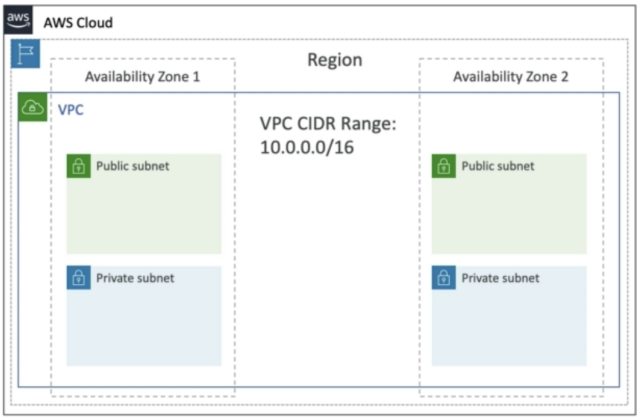
Section 13

VPC & Networking

VPC, Subnets, Private & Public Subnets

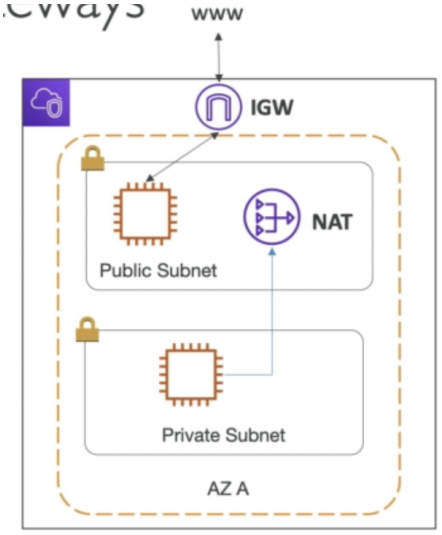
- accessible from internet
- is a private network to deploy your resources
- **Subnets** allows you to partition your network

- **Public Subnet**: accessible from internet
- **Private Subnet**: not accessible from internet
- To define access to the internet and between subnets we use **Route tables**



Internet Gateways & NAT Gateways

- **Internet Gateways** helps our VPC instances connect with the internet
- Public subnets have a route to IGW
- NAT gateways (AWS-managed) & NAT instances (self-managed) allow your instances in your private subnets to access the internet while remaining private

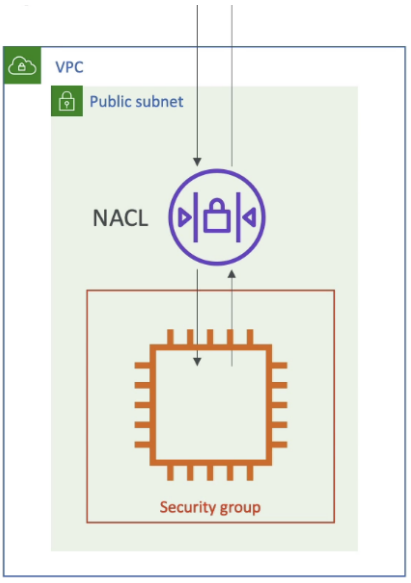


Network Access Control List

- A firewall which controls traffic from and to subnet
- Can have ALLOW and DENY rules
- Rules only include IP addresses

Security Groups

- A firewall which controls traffic from and to an ENI / an EC2 Instance
- Can have ALLOW rules
- Rules include IP addresses and other security groups

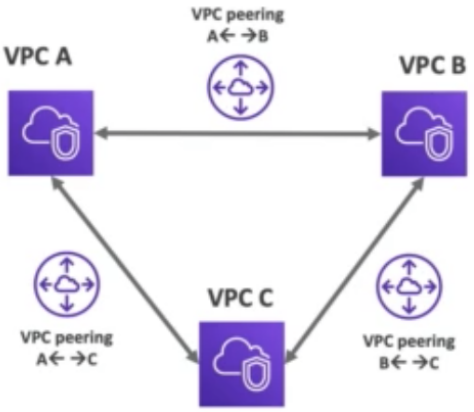


VPC Flow Logs

- Capture information about IP traffic going into your interfaces:
  - VPC Flow Logs
  - Subnet Flow Logs
  - Elastic Network Interface Flow Logs
- Helps to monitor & troubleshoot connectivity issues. Eg: Subnet to internet, subnets to subnets, internet to subnets
- Captures network information from AWS managed interfaces too.
- VPC flow logs data can go to S3 / CloudWatch Logs

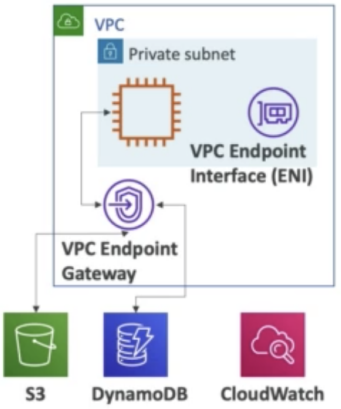
VPC Peering

- Connect two VPS, privately using AWS' network
- Make them behave as if they were in the same network
- Must not have overlapping CIDR
- Must be established for each VPC that need to communicate with one another



VPC Endpoints

- Endpoints allow you to connect to AWS services using a private network instead of the public www network
- Gives enhanced security & lower latency
- VPC Endpoint **Gateway:** S3 & DynamoDB
- VPC Endpoint **Interface**: the rest



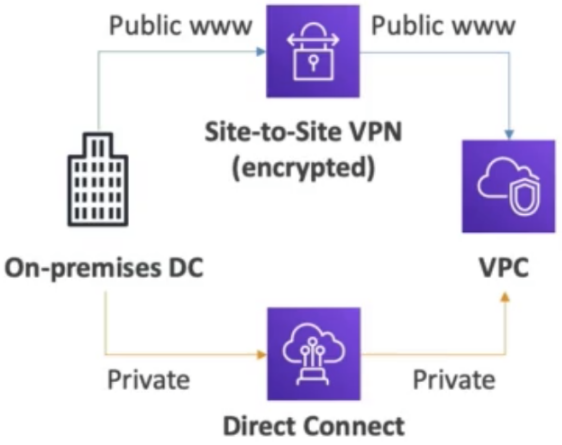
Site to site VPN

- Connect an on-premises VPN to AWS

- Automatically encrypted
- Goes over public network

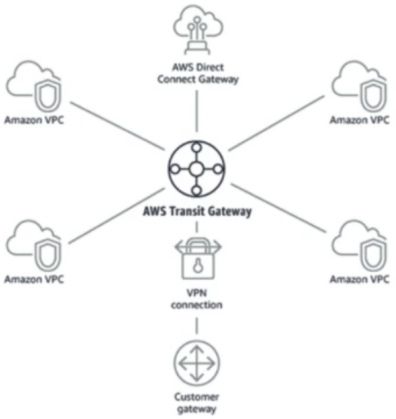
Direct Connect (DX)

- Establish a physical connection between on-premises and AWS
- The connection is private, secure & fast
- Goes over private network



Transit Gateway

- For having transitive peering between thousands of VPC and on-premises, hub-and-spoke (star connection)

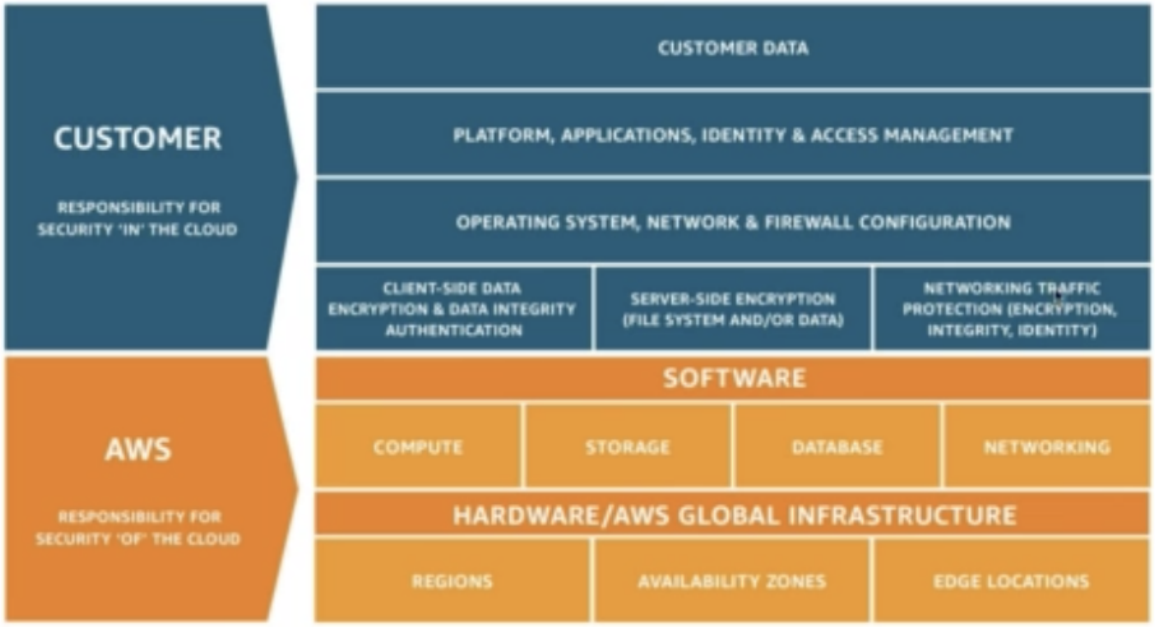


Section 14

Security & Compliance

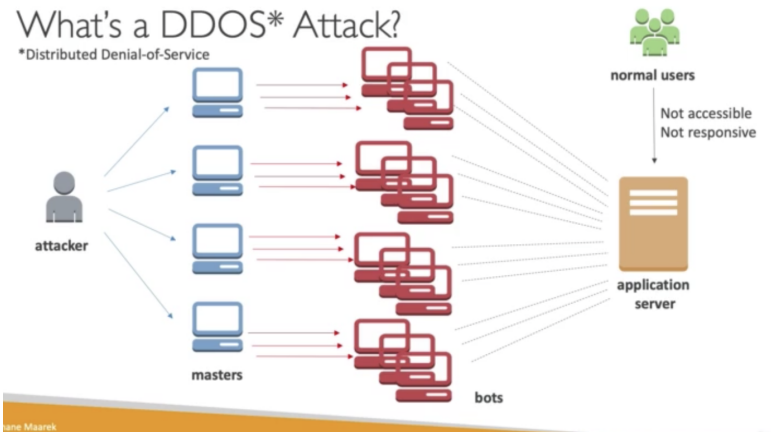
AWS Shared Responsibility Model

- AWS Responsibility - Security of the cloud (Protecting infrastructure, managed services like S3, DynamoDB etc.)
- Customer Responsibility - Security in the cloud (firewall & network config, IAM)
- Shared Controls: Patch management, configuration management, Awareness & Training



DDOS Protection on AWS

- Distributed Denial-of-Service



AWS Shield Standard

- Protects against DDoS attack for your website and applications for all customers at no additional costs, eg attacks: SYN/UDP floods, reflection attacks

AWS Shield Advanced

- 24/7 premium DDoS protection and access to AWS DDoS response team
- $3000 per month per organization
- AWS Shield Advanced provides expanded DDoS attack protection for web applications running on the following resources: **Amazon EC2, Elastic Load Balancing (ELB), CloudFront, Route 53, AWS Global Accelerator**

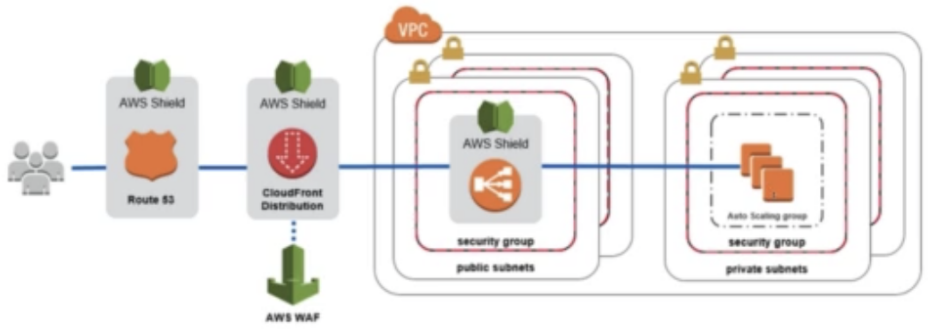
AWS Web Application Firewall (WAF)

- To protect your web applications from common web exploits (layer 7)
- Filter specific requests based on rules
- Deploy on Application Load balancer, API Gateway, CloudFront
- Define Web ACL

CloudFront and Route 53

- Availability protection using global edge network
- Combined with AWS shield, provides attack mitigation at the edge

Be ready to scale - leverage AWS auto scaling



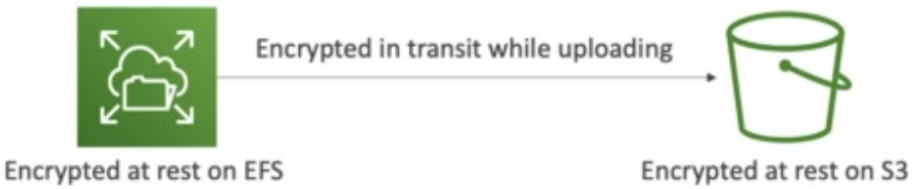
Penetration Testing

- AWS customers are welcome to carry out security assessments on 8 services:
  - Amazon EC2 instances, NAT Gateways, ELB
  - RDS
  - CloudFront
  - Aurora
  - API Gateways
  - Lambda & Lambda edge functions
  - Lightsail resources
  - Elastic Beanstalk environments
- Prohibited Activities
  - DNS zone walking via Amazon Route 53 Hosted Zones
  - DoS, DDoS, Simulated DoS
  - Port flooding, Protocol flooding
  - Request flooding
- For any other contact aws-security-simulated-event@amazon.com

---

Data at rest vs. Data in transit

At rest: data stored or archived on a device
In transit: data in motion



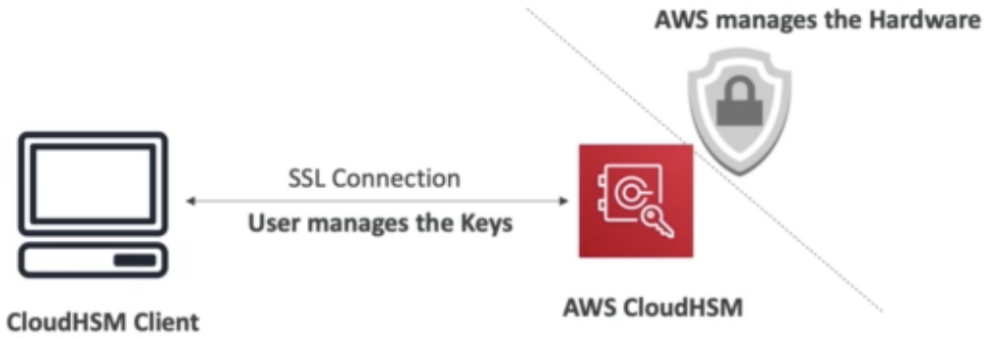
We leverage **encryption keys** to protect data in both states.

AWS KMS (Key Management Service)

- AWS manages the encryption keys for us
- Encryption Opt-in:
    - EBS volumes, S3 buckets, Redshift database, RDS database, EFS drives
- Encryption automatically enabled
    - CloudTrail Logs, S3 Glacier, Storage Gateway

CloudHSM (Hardware Security Model)

- AWS provisions encryption hardware but we manage the keys
- HSM device is tamper resistant, FIPS | 40-2 level 3 compliance



Types of Customer Master Keys: CMK

Customer Managed CMK

- Create, manage and used by the customer
- New key generated every year, old key preserved
- Possibility to bring-your-own-key

AWS managed CMK

- Create, manage and used on the customer's behalf by AWS
- Used by AWS services

AWS owned CMK

- Collection of CMKs that an AWS service owns and manages to use in multiple accounts
- AWS can use those to protect resources in your account (but you can't view the keys)

CloudHSM Keys (Custom keystore)

- Keys generated from your own CloudHSM hardware device
- Cryptographic operations are performed within the CloudHSM cluster

---

AWS Certificate Manager (ACM)

- Easy provision, manage and deploy SSL/TLS certificates
- Used to provide in-flight encryption for websites (HTTPS)
- Supports both public (free) and private TLS
- Automatic TLS certificate renewal

AWS Secrets Manager

- Meant for storing secrets (passwords etc.)
- Capability to force rotation of secrets every X days
- Automatic generation of secrets on rotation (uses lambda)
- Integration with Amazon RDS (MySQL, PostgreSQL, Aurora)
- Encrypted using KMS

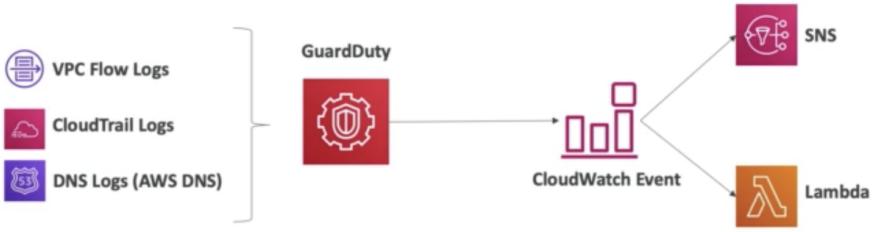
AWS Artifact

- Not really a service
- Portal that provides customers with on-demand access to AWS compliance documentation and AWS agreements
- Artifact reports: AWS ISO certifications, Payments Card Industry, SOC reports
- Artifact Agreements: Allows you to review, accept and track the status of AWS agreements such as BAA (Business Associate Addendum) or the HIPAA (Health Insurance Portability and Accountability Act

---

Amazon GuardDuty

- Intelligent threat discovery to protect AWS account
- Uses **machine learning** algorithms
- Input data includes:
    - CloudTrail Logs
    - VPC Flow Logs
    - DNS Logs
- Can setup CloudWatch Event rules to be notified, these can target Lambda or SNS



Amazon Inspector

- Automated Security assessments for **EC2 instances**
- Analyze the running OS against known vulnerabilities and against unintended network accessibility
- Must be installed on OS in EC2 instances

AWS Config

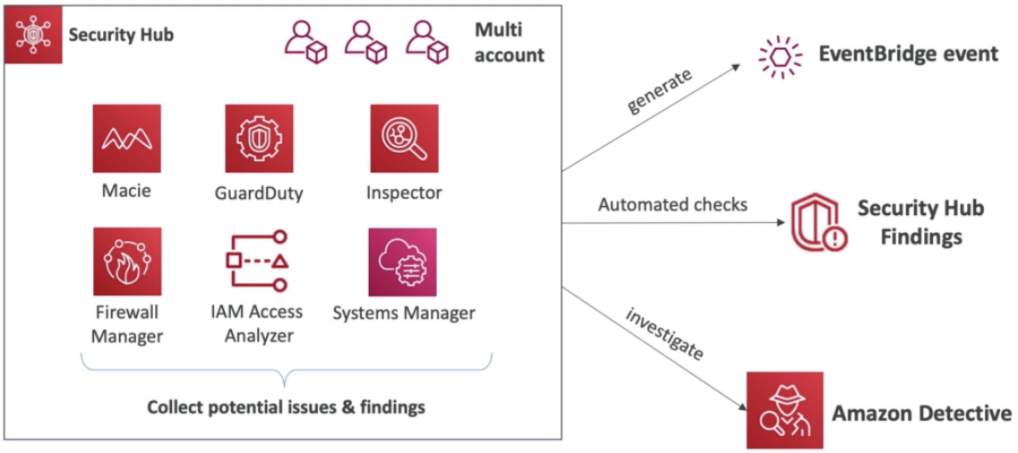
- Helps with **auditing and recording compliance** of your AWS resources
- Helps record **configurations and changes** over time
- Possibility of storing the configuration data into S3
- Can receive alerts
- Per-region service but can be aggregated

Amazon Macie

- Fully managed data security and data privacy service that uses **machine learning and pattern matching** to discover and protect your sensitive data in AWS

AWS Security Hub

- Central security tool to manage security across several AWS accounts and automate security checks
- Integrated dashboards
- Must enable config service



Amazon Detective

- Analyzes, investigates and quickly identifies the root cause of security issues or suspicious activities (using ML and graphs)
- Automatically collects and processes events from VPC Flow logs, CloudTrai, GuardDuty and create unified view

---

AWS Abuse

- Report suspected AWS resources used for abusive or illegal purposes:
  - Spam
  - Port scanning
  - DoS or DDoS attacks
  - Intrusion attempts
  - Hosting objectionable or copyrighted content
  - Distributed malware
- Contact abuse team via forum or email

Root User Privileges

- Account owner, first user, all access
- Lock away your AWS account root user access keys
- Actions only by root user:
  - **Change account settings**
  - View certain tax invoices

- **Close your AWS account**
- Restore IAM user permissions
- **Change/cancel AWS Support plan**
- **Register as a seller in the reserved instance marketplace**
- Configure an Amazon S3 bucket to enable MFA
- Edit or delete and Amazon S3 bucket policy that includes an invalid VPC ID or VPC endpoints ID
- Sign us for GovCloud

Section 15

Machine Learning

Amazon Rekognition

- Find objects, people, text, scenes in images and videos using ML
- Use cases: Labeling, Content Moderation, Text Detection, Face Detection & Analysis, Face search and Verification, Celebrity Recognition, Pathing

---

Amazon Transcribe

- Automatically convert speech to text
- Uses Automatic speech recognition (ASR)
- Use cases: transcribing, closed captioning

Polly

- Turn text into lifelike speech
- Allows you to create applications that talk

Amazon Translate

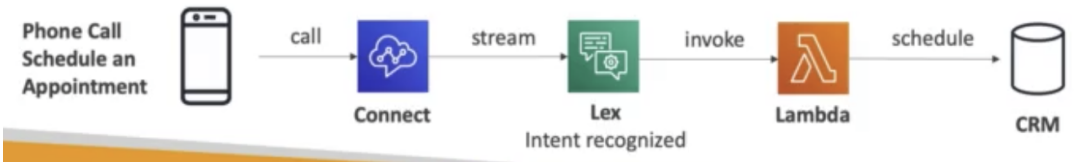
- Natural & accurate language translation
- For international websites

Amazon Lex

- Eg: Alexa
- ASR to convert speech to text
- Natural language understanding to recognize the intent of text, callers
- Helps build chatbots, call center bots

Amazon Connect

- Receive calls, create contact flows, cloud-based virtual contact center
- Can integrate with other CRM systems or AWS
- 80% cheaper than traditional contact center solutions



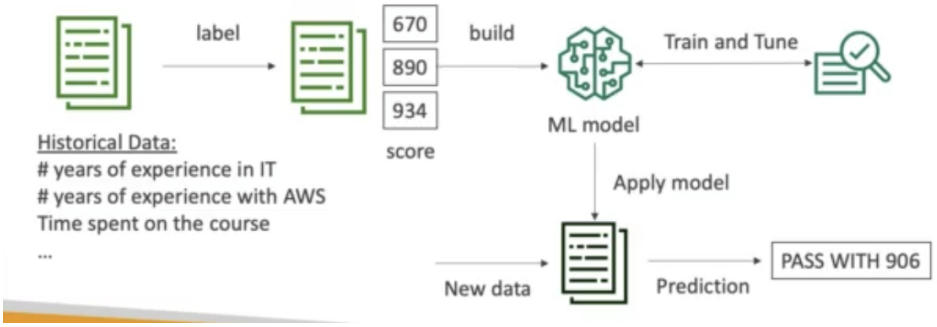
Amazon Comprehend

- For NLP, uses machine learning to find insights and relationships in text
- Fully managed & serverless
- Use case: Analyze customer interactions, create & group article by topics

---

Amazon SageMaker

- Fully managed service for developers/data scientists to build ML models



Amazon Forecast

- Fully managed service that uses ML to deliver highly accurate forecasts
- 50% more accurate than looking at the data, reduces time
- Use case: Product demand planning

Amazon Kendra

- Fully managed document search service
- Ability to manually fine-tune search results

Amazon Personalize

- Fully managed ML-service to build apps with real time personalized recommendations

Section 16

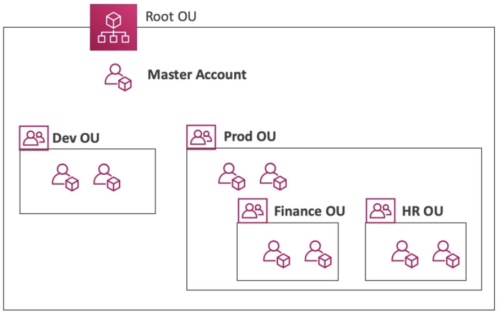
Account management, Billing & Support

AWS Organizations

- Global service
- Manage multiple AWS accounts
- Main account - Master account
- Cost benefits:
    - Consolidated billing
    - Benefits from **aggregated usage**
    - **Pooling of reserved EC2 instances** for optimal savings
- API available to **automate AWS account creation**
- Restrict account privileges using Service control policies (SCP)

Multi Account Strategies

- Create accounts per **department**, per **cost center**, per **dev/test/prod** based on

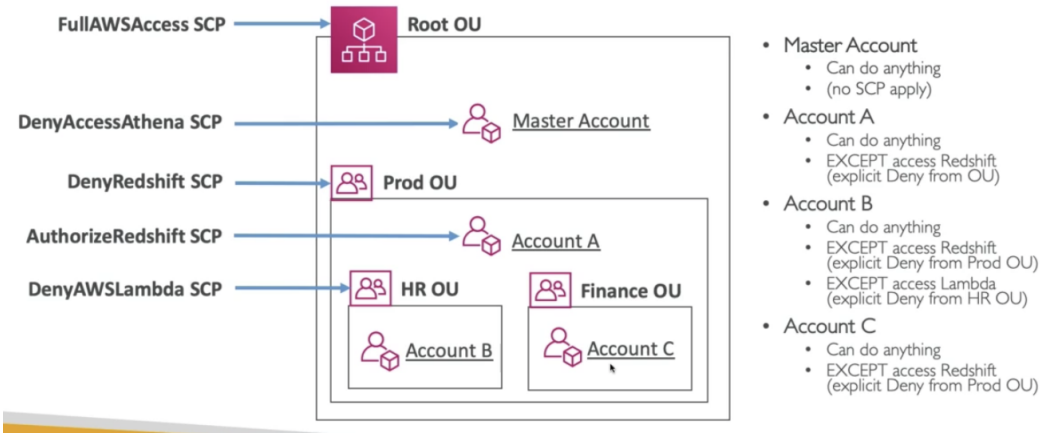


regulatory restrictions (using SCP), for **better resource isolation,** to have **separate per-account service limits,** isolated account for **logging**

- Multi account vs One account Multi VPC - Tradeoff
- Enable CloudTrail on all accounts, send logs to central S3 account
- Send Cloudwatch Logs to central logging account

SCP

- Whitelist or blacklist IAM actions
- Applied at the OU or Account level, doesn't apply to the master account
- Doesn't affect service-linked roles
    - Service-linked roles enable other AWS services to integrate with AWS Organizations and can't be restricted by SCPs
    - SCP must have an explicit allow



Consolidated Billing

- Combined usage - share the volume pricing, reserved instances & savings plans discounts
- One Bill
- Management account can turn off reserved instances discount sharing

---

AWS Control Tower

- Easy way to set up and govern a secure and compliant **multi-account AWS environment** based on best practices
- Benefits:
    - Automate the setup of your env. In a few clicks
    - Automate ongoing policy management using guardrails
    - Detect policy violations and remediate them
    - Monitor compliance through an interactive dashboard
- Runs on top of AWS organizations:
    - Automatically sets up AWS organizations to organize accounts and implement SCPs

---

Pricing Models in AWS

1. Pay as you go
2. Save when you reserve
3. Pay less by using more
4. Pay less as AWS grows

Free services & free tier in AWS

- IAM
- VPC
- Consolidated Billing
- (Elastic Beanstalk
- CloudFormation
- Auto Scaling Group) —-> Pay for the resources created
- Free tier:
    - EC2 t2.micro instance for a year
    - Se, EBS, ELB, AWS Data transfer

Computing Pricing - EC2

- Number of instances
- Instance configuration:
    - Physical capacity
    - Region
    - OS and software
    - Instance type
    - Instance size
- ELB running time and amount of data processed
- Detailed monitoring

On-demand instances

- Minimum of 60 sec
- Pay per second (Linux/windows) or per hour (other)

Reserved instances

- Up to 75% discount compared to on-demand on hourly rate
- 1 or 3 years commitment
- All upfront, partial upfront, no upfront

Spot instances

- Up to 90% discount compared to on-demand on hourly rate
- Bid for unused capacity

Dedicated Host

- On-demand
- Reservation for 1 year or 3 years commitment

Savings plans

- Alternative to save on sustained usage

Computing Pricing - Lambda & ECS

- Lambda
  - Pay per call
  - Pay per duration
- ECS
  - EC2 launch type model: No additional fees, you pay for AWS resources stored and created in your application
- Fargate
  - Fargate launch type model: Pay for vCPU and memory resources allocated to your applications in your containers

---

Storage Pricing - S3

- Storage class: S3 standard, S3 infrequent access, S3 one-zone IA, Se intelligent Tiering, S3 Glacier and S3 Glacier Deep Archive
- Number and size of objects: Price can be tiered (based on volume)
- Number and type of requests
- Data transfer OUT of the S3 region
- Lifecycle transitions
- Similar service: EFS (pay per use, has infrequent access & lifecycle rules)

Storage Pricing - EBS

- Volume type (based on performance)
- Storage volume in GB per month **provisioned**
- IOPS:
  - General purpose SSD: included
  - Provisioned IOPS SSD: Provisioned amount in IOPS
  - Magnetic: Number of requests
- Snapshots:
  - Added data costs per GB per month
- Data transfer:
  - Outbound data transfer are tiered for volume discounts
  - Inbound is free

Database Pricing - RDS

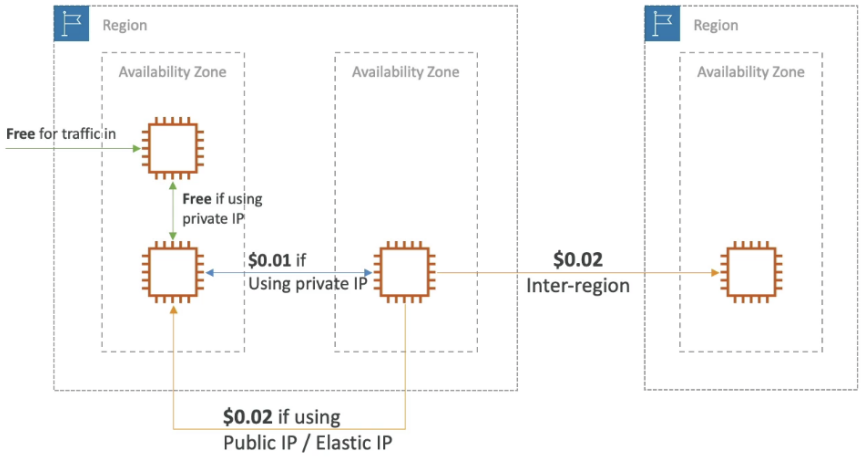
- Per hour billing
- Database characteristics:
    - Engine
    - Size
    - Memory class
- Purchase type:
    - On-demand
    - Reserved instances (1 or 3 years) with required up-front
- Backup storage: There is no additional charge for backup storage up to 100% of your total database storage for a region
- Additional storage (per GB per month)
- Number of input and output requests per month
- Deployment type (storage and I/O are variable):
    - Single AZ
    - Multiple AZs
- Data Transfer:
    - Outbound data transfer are tiered for volume discounts
    - Inbound is free

Content Delivery - CloudFront

- Pricing is different across different geographic regions
- Aggregated for each edge location, then applied to your bill
- Data transfer out (volume discount)
- Number of HTTP/HTTPS requests

Networking costs in AWS per GB

- Use private IP instead of public IP for good savings and better network performance
- Use same AZ for max savings (at the cost of high availability)



Savings Plan

- Commit a certain amount per hour for 1 or 3 years

- Easiest way to setup long-term commitments on AWS
- **EC2 Savings Plan**
    - Up to 72% discount compared to on-demand
    - Commit to usage of individual instance families in a region
    - Regardless of AZ, size, OS or tenancy
    - All upfront, partial upfront, no upfront
- **Compute savings plan**
    - Upto to 66% discount compared to on-demand
    - Regardless of family, region, size, os, tenancy, compute options(EC2, fargate, lambda)
    - Setup from AWS cost explorer console

AWS Compute Optimizer

- Reduce costs and improve performance by recommending optimal AWS resources for your workloads
- Helps you choose optimal configurations & right size
- Uses ML to analyze
- Supported: **EC2 instances, EC2 auto scaling groups, EBS volumes, Lambda functions**
- Lower costs by up to 25%

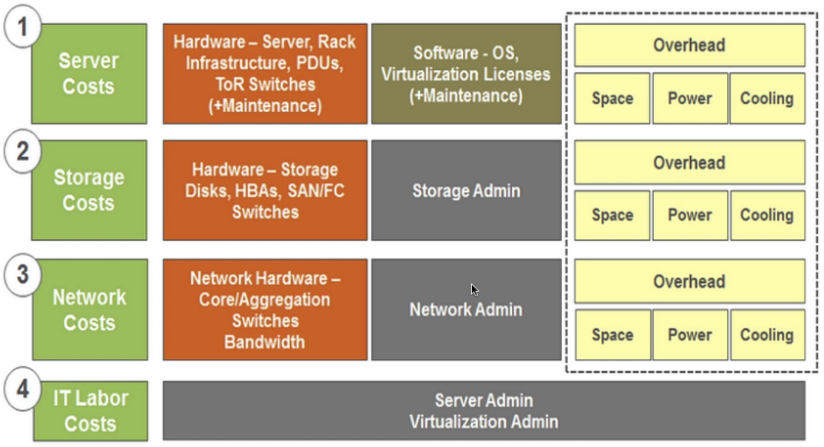
Billing and costing tools

- **Estimating costs in the cloud:**
    - TCO calculator (Deprecated)
    - Simple monthly calculator/pricing calculator
- **Tracking costs in the cloud:**
    - Billing dashboard
    - Cost Allocation Tags
    - Cost and usage Reports
    - Cost explorer
- **Monitoring against costs plans:**
    - Billing alarms
    - Budgets

Estimating

TCO calculator (Deprecated)

- AWS helps you reduce total cost of ownership (TCO) by reducing the need to invest in large capital expenditures and providing a pay-as-you-go model
- Allows you to **estimate the cost savings** when using AWS and provide a detailed set of reports that can be used in executive presentations
- Compare the cost of your applications in an on-premises or traditional hosting environment to AWS: server, storage, network, IT labor



Simple monthly calculator/pricing calculator

Note: deprecated (30th June 2020)

- New name: AWS pricing calculator
- Estimate the cost of your architecture solution

Tracking

AWS Billing Dashboard

- Shows overall cost, free tier

Cost Allocation Tags

- Use cost allocation tags to track your AWS costs on a detailed level
- AWS generated tags
    - Automatically applied to the resource you create
    - Starts with prefix aws
- User-defined tags
    - Starts with prefix user
- Tags are used for organizing resources
- Free naming
- Tags can be used to create **Resource Groups** - manage tags using tag editor

Cost and usage Reports

- Dive deeper into your AWS costs and usage
- Most comprehensive set of AWS cost and usage data available
- It lists AWS usage for each service category as per any tags
- Can be integrated with Athena, Redshift or Quicksight

Cost explorer

- Visualize, understand and manage your AWS costs and usage over time
- Create custom reports
- Analyze data at a high level - across all accounts

- Or monthly, hourly, resource level granularity
- Choose an optimal savings plan
- **Forecast** usage up to 12 month

Monitoring

Billing Alarms in CloudWatch

- Billing data metric is stored in CloudWatch us-east-1
- Billing data are for overall worldwide AWS costs
- Intended a simple alarm (not as powerful as AWS budgets)

AWS Budgets

- Create budget and send alarms when costs exceeds the budget
- 3 types of budgets: Usage, Cost, Reservation
- For Reserved Instances:
  - Track utilization
  - Supports EC2, ElastiCache, RDS, Redshift
- Up to 5 SNS notifications per budget
- Can filter by: Service, linked account, tag, etc,..
- Same options as AWS Cost explorer!
- 2 budgets are free, then $0.02/day/budget

AWS Trusted Advisor

- No need to install anything - high level AWS account assessment
- Analyze your AWS accounts and provides recommendation on 5 categories
  - Cost optimization
  - Performance
  - Security
  - Fault tolerance
  - Service limits

Support Plans

7 CORE CHECKS Basic & Developer Support plan	FULL CHECKS Business & Enterprise Support plan
• S3 Bucket Permissions • Security Groups – Specific Ports Unrestricted • IAM Use (one IAM user minimum) • MFA on Root Account • EBS Public Snapshots • RDS Public Snapshots • Service Limits	• Full Checks available on the 5 categories • Ability to set CloudWatch alarms when reaching limits • Programmatic Access using AWS Support API

Support Plans - Pricing

- **Basic Support Plan**
  - **Customer Service & Communities** - 24X7 access to customer service, documentation, whitepapers and support forums
  - **AWS Trusted Advisor -** Access to the 7 core trusted advisor checks and guidance to provision your resources following best practices to increase performance and improve security
  - **AWS personal Health Dashboard -** A personalized view of the health of AWS services, and alerts when your resources are impacted

- **Developer Support Plan**
  - All basic support plan +
  - Business hours email access to cloud support associates
  - Unlimited cases / 1 primary contact
  - Case severity/ response times:
    - General guidance: < 24 business hours
    - System impaired: < 12 business hours

- **Business Support Plan**
  - Production workloads
  - Trusted advisor - full set of checks + API access
  - 24x7 phone, email and chat access to cloud support engineers
  - Unlimited cases / unlimited contacts
  - Access to Infrastructure event management for additional fee
  - Case severity/ response times:
    - General guidance: < 24 business hours
    - System impaired: < 12 business hours
    - Production system impaired: < 4 hours
    - Production system down: < 1 hours

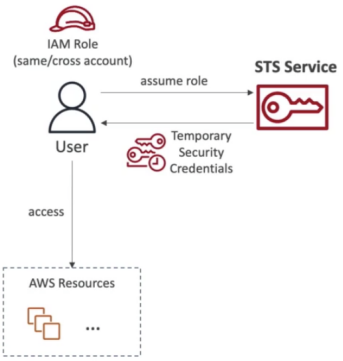
- **Enterprise Support Plan**
  - Intended to be used if you have mission critical workloads
  - All of business support plan +
  - Access to a Technical Account Manager (TAM)
  - Concierge support team (for billing and account best practices)
  - Infrastructure event management, well-architected & operations reviews
  - Case severity/ response times:
    - General guidance: < 24 business hours
    - System impaired: < 12 business hours
    - Production system impaired: < 4 hours
    - Production system down: < 1 hours
    - Business-critical system down: < 15 mins

Section 17

Advanced Identity

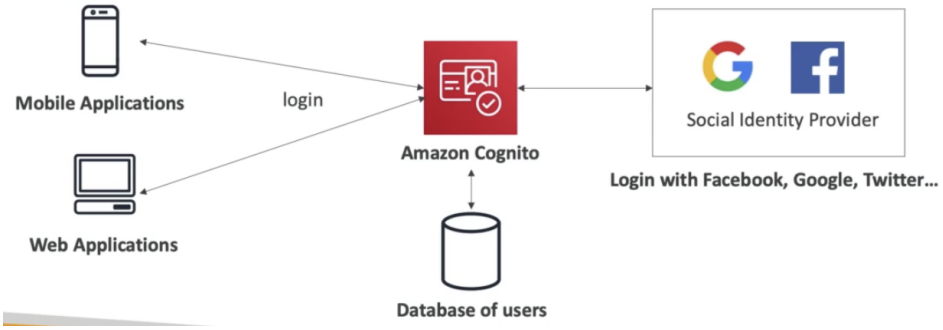
AWS Security Token Service (STS)

- Enables you to create **temporary, limited privileges credentials** to access your AWS resources



Amazon Cognito

- Identity for your web and mobile applications users (potentially millions)



AWS Directory Services

- AWS managed Microsoft Active Directory
  - Create your own AD in AWS
- AD Connector
  - Directory Gateway (proxy) to redirect to on-premise AD
- Simple AD
  - AD-compatible managed directory on AWS
  - Cannot be joined with on-premise AD

Amazon Single Sign-on (SSO)

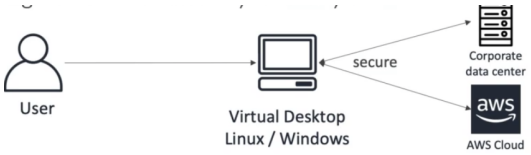
- Centrally managed Single sign-on to access multiple accounts and 3rd party business applications
- Integrated with AWS Organizations

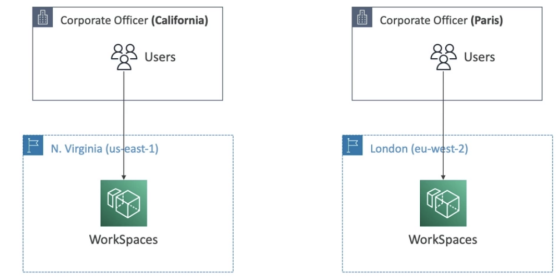
Section 18

Other Services

Amazon Workspaces

- Managed Desktop as a Service (DaaS) solution to easily provision Windows or Linux desktops
- Great to eliminate management of on-premise VDI, scales to thousands of users





Amazon AppStream 2.0

- Desktop Application Streaming Service
- Deliver to any computer, without acquiring, provisioning infrastructure
- App is delivered from within a web browser

Workspaces vs AppStream 2.0

Workspaces:

- Fully managed VDI
- On-demand or always on

AppStream 2.0:

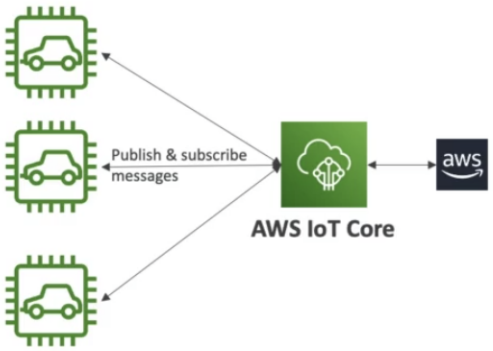
- No need to connect to a VDI
- Works with any device
- Allow to configure an instance type per application type

Amazon Sumerian

- Create & run VR, AR and 3D applications
- Quickly create 3D models with animations

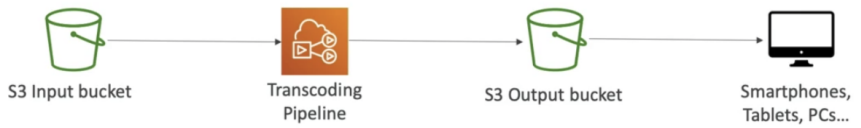
Amazon IoT Core

- Allows you to easily connect IoT devices to AWS cloud
- Serverless, secure & scalable



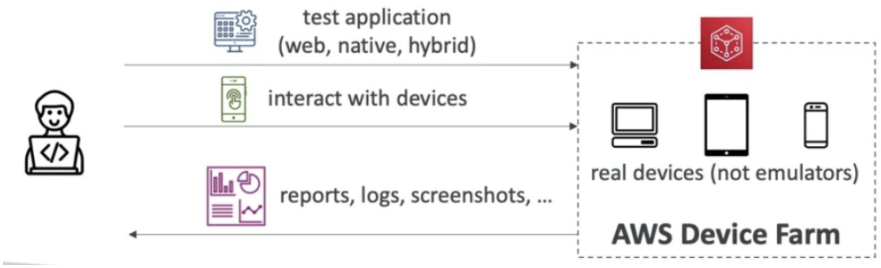
Amazon Elastic Transcoder

- Used to convert media files stored in S3 into media files in the formats required by customer playback devices (phones etc.)



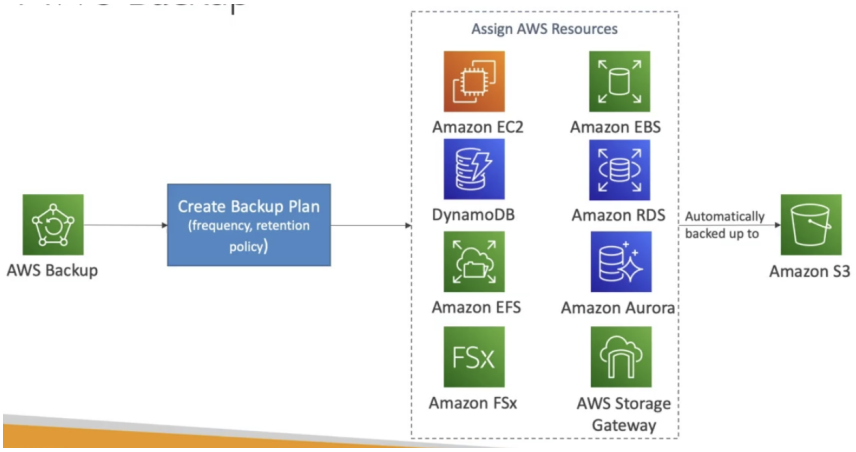
AWS Device Farm

- Fully-managed service that tests your web & mobile apps against desktop browsers, real mobile devices & tablets

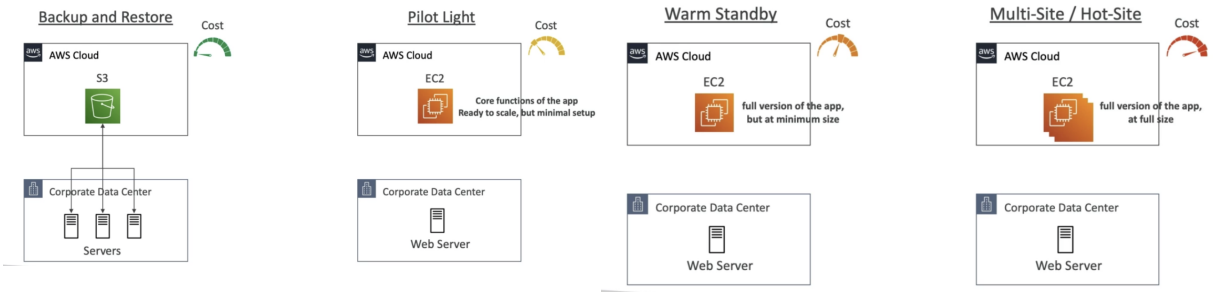


AWS Backup

- Fully-managed service to centrally manage & automate backups across AWS services
- On demand & scheduled

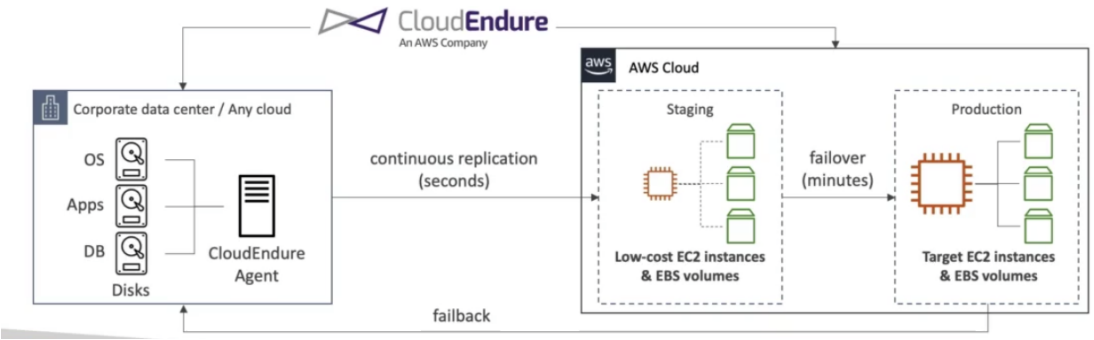


Disaster Recovery Strategies



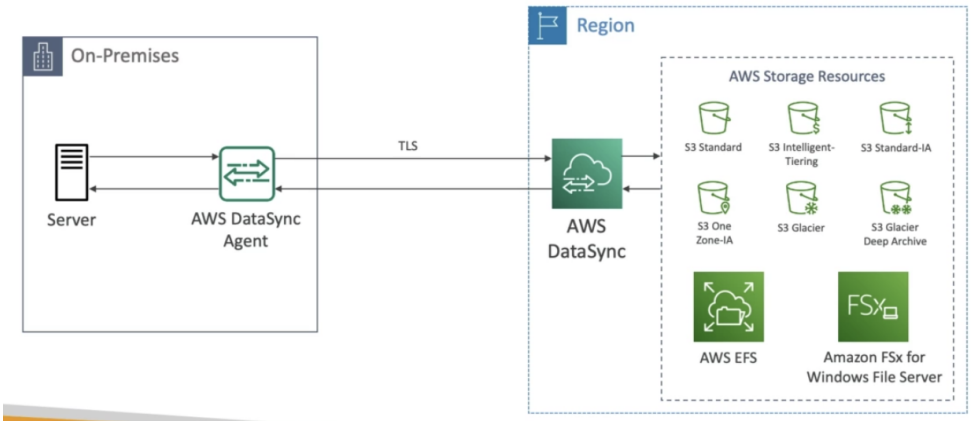
CloudEndure Disaster Recovery

- Quickly & easily recover your physical, virtual and cloud-based servers into AWS



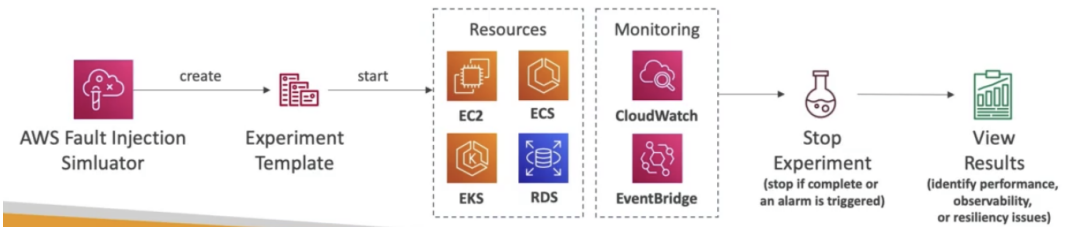
AWS DataSync

- Move large amounts of data from on-premises to AWS
- Replication tasks can be scheduled hourly, daily, weekly
- Replication tasks are **incremental after the first full load**



AWS Fault Injection Simulator (FIS)

- A fully managed service for running fault injection experiments on AWS workloads
- Helps you uncover hidden bugs & performance bottlenecks
- Can use pre-built templates that generate the desired disruptions



Section 19

AWS Architecting & Ecosystem

Guiding Principles

- Stop guessing your capacity needs
- Test systems at production scale
- Automate to make architectural experimentation easier
- Allow for evolutionary architectures
- Drive architectures using data
- Improve through game days (Simulate applications for flash sale days)

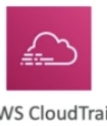
Best Practices - Design Principles

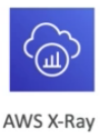
- **Scalability:** vertical & horizontal
- **Disposable resources**
- **Automation:** Serverless, Infrastructure as a service, Auto scaling
- **Loose coupling:** A change or failure in one component should not cascade to other components
- **Services not servers:** Use managed services

Well Architected Framework Pillars

1. Operational Excellence
    - Ability to run and monitor systems to deliver business value and to continually improve supporting processes and procedures.
    - Design principles:
        - **Perform operations as code** - Infrastructure as code
        - **Annotate documentation** - Automate the creation of annotated documentation
        - **Make frequent, small, reversible changes**
        - **Refine operations procedures frequently**
        - **Anticipate failure and learn from it**
    - AWS Services:
        - Prepare

        - Operate





- Evolve

2. Security

- Ability to protect information, systems and assets while delivering business value through risk assessments and mitigation strategies
- Design principles:
  - Implement a strong identity foundation - Principle of least privilege
  - Enable traceability
  - Apply security at all layers
  - Automate security best practices
  - Protect data in transit and at rest
  - Keep people away from data
  - Prepare for security events
- AWS Services:
  - IAM

  - Detective Controls

  - Infrastructure Protection

  - Data Protection

  - Incident Response

3. Reliability
    - Ability of a system to recover from infrastructure or service disruptions, dynamically acquire computing resources to meet demand, and mitigate disruptions such as misconfigurations or transient network issues

    - Design principles:
        - Test recovery procedures
        - Automatically recover from failure
        - Scale horizontally to increase aggregate system availability
        - Stop guessing capacity - Use autoscaling
        - Manage change in automation - use automation to make changes to infrastructure

    - AWS Services:
        - Foundations

        - Change Management

        - Failure Management

4. Performance Efficiency
    - Ability to use computing resources efficiently to meet system requirements, and to maintain that efficiency as demand changes and technologies evolve

    - Design Principles:
        - Democratize advanced technologies
        - Go global in minutes
        - Use serverless architectures
        - Experiment more often
        - Mechanical sympathy - Be aware of all AWS services

    - AWS Services:
        - Selection

- Review

- Monitoring

- Trade Offs

5. Cost Optimization

- Ability to run systems to deliver business value at the lowest price point
- Design Principles:
  - Adopt a consumption mode
  - Measure overall efficiency
  - Stop spending money on data center operations
  - Analyze and attribute expenditure
  - Use managed and application level services to reduce cost of ownership
- AWS Services:
  - Expenditure Awareness

  - Cost effective resources

  - Matching supply and demand

- Optimizing Over time

6. Sustainability
    - The discipline of sustainability addresses the long-term environmental, economic, and societal impact of your business activities.
    - Design Principles:
        - Understand your impact
        - Establish sustainability goals
        - Maximize utilization
        - Anticipate and adopt new, more efficient hardware and software offerings
        - Use managed services
        - Reduce the downstream impact of your cloud workloads

AWS Well- Architected Tool

- Free tool to review your architectures against the 6 pillars well-architected framework and adopt architectural best practices

---

AWS Right Sizing

- Right sizing is the process of matching instances types and sizes to your workload performance and capacity requirements at the lowest possible cost
- Always start small, assess continuously

AWS Ecosystem - Free resources

- AWS blogs
- AWS Forums
- AWS Whitepapers & Guides
- AWS Quick Starts - various templates
- AWS Solutions - Vetted technology solutions for the AWS Cloud

---

AWS Ecosystem - AWS Support

DEVELOPER	• Business hours email access to Cloud Support Associates • General guidance: < 24 business hours • System impaired: < 12 business hours
BUSINESS	• 24x7 phone, email, and chat access to Cloud Support Engineers • Production system impaired: < 4 hours • Production system down: < 1 hour
ENTERPRISE	• Access to a Technical Account Manager (TAM) • Concierge Support Team (for billing and account best practices) • Business-critical system down: < 15 minutes

---

AWS Marketplace

- Digital catalog from independent software vendors
- Eg: Custom AMI, SaaS, Containers
- Goes directly to AWS bill
- Can sell your own solution on it

AWS Training

- AWS Digital (online) and Classroom Training
- AWS private training
- Training & certification for the U.S. Government
- Training and certification for the enterprise
- AWS Academy - helps universities teach AWS

AWS Professional Services & Partner network

- Organization of global team of experts
- They work alongside your team & a chosen member of the APN
- APN Technology Partners - Providing hardware, connectivity and software
- APN Consulting Partners - Professional services firm to help build on AWS
- APN Training Partners - Helping in learning AWS
- AWS Competency Program - AWS Competencies are granted to APN Partners who have demonstrated technical proficiency and proven customer success in specialized solution areas
- AWS Navigate Program - Help partners become better partners

---

Exam Tips

Exam Alert:
You may see use-cases asking you to select one of CloudWatch vs CloudTrail vs Config. Just remember this thumb rule -

Think resource performance monitoring, events, and alerts; think CloudWatch.

Think account-specific activity and audit; think CloudTrail.

Think resource-specific change history, audit, and compliance; think Config.