Multi-tier solutions

Considerations
Access patterns should be considered based on who and what needs access and how they will access services and architectures.

- Peek or steady connections?
- Large volume of data traffic?
- Is scaling required?

High-Availability | Fault Tolerance

High Availability
Design for minimal downtime.

Fault tolerance
Having an architecture that requires zero downtime. Higher cost due to replica and redundancy. More complexity.

Disaster Recovery
Evaluate recovery time RTO (Recovery Time Objective) and RPO (Recovery Point Objective)
Active/Passive and Active/Active
Multi-site, Back-up/Restore

Terminology

- **Reliability / Resiliency**
  The ability of an application to avoid and recover from failure
- **Availability**
  The percentage of time that an application is performing as expected. Poor performance = low availability
- **Loose coupling**
  Every component is independent and does not depend on a specific component. One to many relationship of components

Simple Queue Service (SQS)

- **Standard Queue**
  Unlimited Throughput
  At-least one delivered, occasionally more than one
  Best-effort ordering
- **FIFO Queue**
  3'000 messages per second (300 batches of 10 messages)
  Quota increase via Support Ticket
  Exactly once Processed
  First-in, First-out ordering
- **Pricing**
  1M monthly requests is free

Decoupling

Components remaining autonomous and unaware of each other.

Synchronous
At least two components which must both be available. Example can be Elastic Load balancer. To properly function both EC2 must be available but not aware of each other.

Asynchronous
Communication can be achieved even if one component goes down. For example SQS handling communication between components.

Resiliance storage

- Understand data durability and scenario suited for each service
- How data consistency and how calls to the services are impacted
- Understand access-pattern and type of access (fast read, fast write, user based, file size)
- Hybrid scenarios, which services fullfill this requirement

Availability SLA

- **99.0%**
  3 days and 15 hours
- **99.9%**
  5 hours and 45 minutes
- **99.99%**
  1 hour
- **99.999%**
  5 minutes

Example

- **SLA of 99.9%** -> Front-end with Back-end and SQL Database.
  - 4 EC2 instance and 2 SQL Databases
  - 1 EFS and 1 ELB
- SLA for EC2 is 90%. 4 EC2 will be 10%*10%*10%*10% = 0.01% = 99.99%
- Then we have ELB for ensuring Elasticity (99.99%)
- EC2 is 99.99% * ELB 99.99% = 99.998%
- EFS is 99.99% * 99.998% = 99.97%
- RDS is 99.95% * 99.97% = **99.92%**

Elastic Container Service (ECS)

- **ECR**
  Elastic Container Registry to upload Docker images
- **AWS Copilot**
  CLI tool used to build, release and operate containers in ECS
- **Pricing**
  AWS Fargate goes by consumption (vCPU, memory)
  EC2 goes only by EC2 usage
  EC2 on AWS Outposts same as EC2
- Integrated with Elastic Load Balancer

Cloud Native Applications

Considerations
Applications which depends on services that are Cloud native and cannot be replicated on-premises. (Dynamo DB, SQS, S3)
- Dynamo DB (**99.999%** Avail.) Global Tables
- S3 (**99.9%** Avail.)
- Lambda (**99.5%** Avail.)

Multi-region can achieve **99.999%**

Cost Alerts

You can use two different methods to set comsumption and budget alerts:
- Cloud watch alert
- Budget in Billing section
  - **Cost budget** - monitor amount
  - **Usage budget** - monitor usage of types
  - **Reservation budget** - monitor reserved instances
  - **Saving plans budget** - monitor utilization of Saving Plans

Important: AWS does not shut-down if you exceed the budget

CloudTrail

Audit tool which allows you to see actions on AWS Account. The log is retained for **90 days**.

- Event history to show all events
- You can filter

A custom Trail requires an **S3 bucket** to store the events.

VPC Networking

Subnet
It is inside a VPC and it is on *one Availability Zone* and cannot be moved to a different one.
- An EC2 instance exists in one Aval. Zone and in a Subnet
- Private and Public do not have to be on same Zone but *beware on Failure/Latency*

VPC
A VPC has a Region scope. Two VPC connect via *Transit Gateway*.

Elastic IP (EIP)
EIP assigned to a Public IP address, will be replaced by EIP. Amazon one is tight to Region, you cannot choose it. Customer (BYOD) Public IP can be transferred into EIP.
Global Accelerator gives you 2 Public IPv4, are not region specific.

Trusted Advisor

Can be used to find Service Limits
- **Service Limits**
  You can check the limits of a specific service
- **Support case**
  Can be used to increase the service limit of a specific service you are using

IAM Creation

When you create a User in IAM
- Set a username (***choose one or more below***)
  - Choose "Access Key if you need for CLI"
  - ***After creation, Access key*** cannot be retrieved twice, you must create a new one and delete existing
  - Choose "Password" if AWS Console login
- By default ***no permissions*** are added
- Create a Group, Add Policies to the Group and add User to the Group
- Inside ***Account settings*** you can change Password policy including permission to reset password

Root User must have *MFA* enabled, from IAM Dashboard
- MFA device
- Security Key (RSA generator)
- Other hardware

Here you also get Security Reccomendations for your IAM users
- MFA is required
- Root access keys should be removed

AWS CLI

Available for: Windows, MAC and Linux
Login:
- `aws configure` (will store ~/.aws/config)
  - accesskey secret region
- `aws ec2 allocate-address` (allocate EIP to EC2)

Private vs Public Subnet

Public
- Associated with a Route Table that has a route pointing to a Internet Gateway

Private
- Associate dwith a Route Table that has a route pointing to a NAT Gateway (used to pass out to internet)

Remeber: Route table association to a Subnet is how to switch the association between NAT or Internet Gateway

Direct Connect or VPN

Direct Connect
- Low-latency but more expensive than VPN
- Bypasses the internet
- ***Dedicated*** ((1-10GB) or ***Hosted*** (50MB-10GB)

VPN
- IPSec encryption
- ***Private Gateway*** to a VPC
- T***ransit Gateway*** (Virtual router with multiple VPN)

Disaster recovery approaches

Backup - Restore
- Protects against data loss
- This is only for Low Priority use cases. Cheapest

Pilot Light
- Replication to another region
- Kept off until disaster happens

Warm standby
- Smaller caled version of production in another region
- Contrary to pilot light, here you keep the environment on but lighter

Multi-site active - active
- Specular copy into another region
- Running all the time
- Load balanced via Route 53

Route 53

Active-Active
- ***Weighted Resource Records*** - Route 53 - Round Robin Load balancing between two regions
- Weighted resource requires the "weight" you want to distribute, i.e. 50/50%
- Use dnschecker.com to test weight of route 53

Active-Passive
- Secondary region service only when primary region fails
- ***Pilot light*** - Secondary region is lean, less services, cheaper
- ***Warm standby*** - near replica of primary environment, mirroring would be the strategy
- ***Failover records*** - Sends only to one region until it becomes unhealthy

Health Check

Health check can be used as an alternative to ELB for health check. You define an endpoint to monitor the health of your services.
- You will need a A Record for each Health check record

Load Balancer options

Application Load Balancer (only layer 7)
- Used to load balance HTTP and HTTPS traffic
- Provides advanced routing logic

Network Load Balancer
- Ultra-high performance
- TLS offloading
- Central certificate deployment
- Ultra-low latency because it operates at connection level

Gateway Load Balancer
- Used to load balance third party appliances

Target Group
- Group multiple targets together
- You can have a listener rule which target a specific group

Identify elastic and scalable services

Scalable and Elastic by nature

- AWS Lambda (*15 minutes* max run)
- EC2 is not scalable
- Choose the right EC2 instance type (remember the family)
- Which AWS is the most scalable for your needs?
- When Lambda, or EC2 or ECS?
- CloudWatch Alarm will trigger scale

High Performing Storage

How and where to store the data

- EBS scale only by changing the volume size manually
- EFS scales automatically when space is full
- Know bounds of capacity of your solution
- EBS low in latency (know different types)
- S3 CLI and APIs calls
- S3 accelerator and cloudfront options
- S3 multi-part upload. S3 CP perform automatically multi-part Important for file higher than 100MB

High-performing Database

Scalable Databases

- Choose suitable DB for Performances
- Know differences between RDS/Aurora
- Scaling strategies between various DB
- Knows various Replicas in RDS based on DB Type
- ElasticCache (patterns, scenarios)
- Knowledge of Kinesis and DataLake patterns

High Performing Network

How to select high performing networks

- Read about VPN and AWS Direct Connect
- AWS Transit Gateway for multiple VPC, use cases and network peering
- AWS VPN Cloud Hub
- AWS Private Link, VPC Peering
- Route 53, policies and routing techniques
- Data-transfer services (AWS DataSync, Snow family, RDS migration and similar)

Overview

Which Layers

- Compute
  - Important to understand which is the best fit. Which migration patterns (i.e. lift-shit)
  - EC2 scaling and troubleshooting, concurrencies
- Database
  - Replica, Accelerators, ElasticCache, Offloads
- Networking
  - Understand Networking use cases, mitigate issues, multi-region, multi-az
  - Route 53, VPC options
- Storage
  - Which services for Block, File and other data
  - Knowledge of Snapshots, Backup, Replication

AWS Architect Associate

High-Performing Architectures

Design Principles

- Go Global
- Think serverless, managed services to reduce support
- Use new techonologies
- Experiment often with new technologies
- Right tool for the task

ECS Performances

- Long running applications, batch processing and micro-services
- Choose EC2 instance types to host your containers
- **AWS Fargate** - It manages the EC2 instances for you. You don't manage infrastructure

Lambda (Function as a Service)

- Back-end processing no longer than **15 minutes**
- Stream processing of data (IoT), Data processing
- **Triggers**
  - DynamoDB, SQS, SMS, API Endpoints
- You choose only Memory, vCPU is given
- Auto-scale and Fault tolerant and rerun last work

S3 Data Retrieval options

S3 Glacier Instant
- Milliseconds response
- Rarely used

S3 Glacier Flexible
- Expedited - within 1-5 minutes
- Standard - within 3-5 hours
- Free Bulk - within 5-12 hours

S3 Deep Archive
- Bulk - within 48 hours
- Standard - within 12 hours

You can query data before retrieval, wihitout actually retrieving it

- Object storage with shared access
- Low latency and high throughput
- **Encryption**
  - Server side Encryption managed by AWS
  - Client side Encryption, Client encrypt before sending
- **Access**
  - By default everything is private

S3 Transfer Acceleration minimize latency

EC2 Performances

- **General Purpose (T and M)**
  Balanced level of resources for standard workloads
- **Compute Optimized (C)**
  For high compute needs, data lakes, extra compute power
- **Memory Optimized (R and X2)**
  Memory intensive workload, memory database
- **Accellerating Compute (P and G)**
  Instance with GPU and extra compute power
- **Storage Optimized (L and D)**
  Heavy storage needs
- **Bare Metal**
  Support custom licensing and customized design
- **EC2 Autoscaling**
  - *Metrics Based*
    Different poerformance criteria to scale-up/down
  - *Schedule Based*
    At a certain time of day/year
  - *Health Based*
    Monitor the instance and scale when one instance has issues

EFS Volume Types

- Shared File Shared HD

Elastic
- Grow and Shrink based on Data

Available
- Stored in multiple AZs on Same Region

Performance
- General Purpose - Random and Sequential I/O
- Max I/O - Scale higher for parallel multiple access

Throughput
- Bursting - The available throughput is based on File size. Adjust dynamically
- Provisioned - We decide upfront which throughput we need all the time

EBS Volume Types

- A volume can be attached only to 1 instance
- Multi-attach is a new version for multiple EC2 (no boot, same AZ all EC2)
- Replicated on multiple AZs in same Region

Magnetic Volumes
- Cheapest, for infrequent data access
- 1GB to 1TB

General Purpose SSD
- Default type
- 10'000 IOPS - 160MB/s

Provisioned IOPS SSD
- Intensive IOPS systems
- 32'000 IOPS - 500MB/s

HHD - Backed
- Large I/O sizes
- ETL, Map Reduces, Data warehouse

RDS Performances

- Allow **multi-AZs solutions**
  - Creates another AZ instance for standby and switch if primary fails
- **Read replicas** (read-only connection)
  - MySQL, MariaDB, PostgreSQL, Oracle and SQL Server
- **Encryption**
  - TLS - in transit
  - KMW - at rest encryption
- Backup (automatically done) - allow restore
- Snapshot (on-demand) - allow restore
- **Instance type**
  - General (T and M), Memory optimized (R and X)
- Storage type
  - General
  - Provisioned IOPS - for performances

DynamoDB

- **Reason for choice**
  - Flexible structure, can query only by keys
  - Low latency
- **Global tables**
  - Store the data in multiple ***Regions***
- **Encryption**
  - KMS or CMK
  - Activated by Table when creating a new Table
- **Capacity**
  - Provision - auto-scale based on operation, pay what you provision, mor expensive
  - On-demand - pay only for operations

Route 53

DNS Solution to an IP Address

Private DNS to setup friendly name internally to a VPC.

Traffic Flow - allow to redirect users to the closest region.

- ***Latency*** routing - closest region to user with less latency
- ***Geographic*** routing - closest region to user
- ***Health*** routing - based on region's health
- ***Round*** robin - user routed to the next available region

Regions and AZs

Regions

- Geographical areas
- Isolated between each-other, made of multiple AZs
- Choice: Laws, User location, Data location, Cost

Availability Zone

- Geographically separated between each-other
- Failover to another AZ in the same region
- Redundant power and redundant internet provider

Local Zones

- In large cities away from an AWS region
- Connected into the closest Region
- Subset of services is available (EC2, EBS, RDS)

Edge Locations

Global network outside of a Region. CloudFront is a CDN hosted on hedge locations that deliver static content.

Georestriction restrict content to certain region/location.

Lambda ad Edge you can deploy Lambda Function on Edge Locations.

Various VPN Solutions

AWS Site-to-Site

- IPSec VPN to remote network. Virtual Private Gateway or Transit Gateway to connect to your network

AWS Client VPN

- Client based VPN. Establish TLS private VPN to a VPC

AWS VPN CloudHub

- More than one Remote network, it manages multiple Site-to-Site

VPC Peering

- Connect two VPC together
- ***PrivateLink*** - connect a specific service(s) between VPC

Domain Segments

- Design secure access to AWS services
- Design secure application tiers
- Select appropriate data security options

IAM - Users, Groups and Role. How you create, manage and use in each scenario

How to secure credentials, Root users and various techniques and best practices.

CIA triad

- **Confidentiality**
  Only authorized people can access specific data
  ACL, Encryption are sample techniques
- **Integrity**
  Data has not been modified in unauthorized way
  You need to know if data has been modified
- **Availability**
  Who is supposed to have access to data, should have it

Credentials Types

- **Root user**
  Only one, created together with the Account
- **IAM Principal**
  Any entity that can perform actions and has permissions
  - Policies define what a Principal can do
  - Can be a User or a Role or a Technical Account
  - By default it has **NO PERMISSIONS**

Policy

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "ReadWriteTable",
            "Effect": "Allow",
            "Action": [
                "dynamodb:BatchGetItem",
                "dynamodb:GetItem",
                "dynamodb:Query",
                "dynamodb:Scan",

"dynamodb:BatchWriteItem",
                "dynamodb:PutItem",
                "dynamodb:UpdateItem"
            ],
            "Resource":
"arn:aws:dynamodb:*:*:table/SampleTable"
        }
    ]
}
```

Secure Application Tiers

- SecurityGroups and ACL to secure each tier of your application
- Public and Private subnet differences. Routing via VPC and route tables
- Protection from external threats, AWS Shield, WAF and similar services

Secure Data

- REST and in-transit security options
- VPN, Direct connect, S3, End user Public internet
- S3 vs EBS protection mechanisms
- Options of Encryption
- KMS options and how services handles KMS

Lock-down Root User

- **MFA**
  First step enable MFA to double secure authentication of Root user
- **Don't use it**
  A IAM User should be created with specific policies to accomplish actions
- **Delete Root Access Keys**
  Because they can bypass MFA via CLI, for example
- Password Complexity Policy

Policy

- **Effect**
  Allow or Deny an Action on a Service
- **Service**
  The service affected by the Policy
- **Action**
  The operation the policy allows or denies on a service
  *You can use wildcards * to target multiple*
- **Resource**
  The specific service instance you want to target
  *You can use wildcards * to target multiple*
- **Condition**
  For example MFA, IP Range, Time
- **Principal**
  The Principal, used by Role, which assume to policy effect

A *Deny Policy* will **always override** a previous Policy that allows an Action. So Deny take precedence over Allow.

You can **simulate** a Policy to verify the outcome of it.

An **inline-policy** is attached to a Principal, Group or User. An **inline-policy** has the same lifecycle of the Principal itself.

CloudTrail

CloudTrail logs actions for limited time of **90 days** by default. Log goes into S3 which is Money/Space used. Up to **15 minutes** before files go to S3 Bucket.

A Trail can track:

- **Management Events**
  - Configuration changes into AWS
  - Reading resources
  - Logging into console
  - Assuming a role
- **Data Events**
  - Actions on objects in S3
  - Lambda executions

AWS Config

- Track what a service configuration looks like in a point in time
- It is a **Time Machine** on a specific configuration on your Account(s)
- Uses **S3** buckets to track changes
- Can send **SNS** events to react over a change

Public Subnet

- **CIDR Calculation**
  - x.x.x.x/32 - 1 IP address
  - x.x.x.x/24 (2^8) - x.x.x.0 - x.x.x.255
  - x.x.x.x/16 (2^16) - x.x.0.0 - x.x.255.255
  - x.x.x.x/8 (2^24) - x.0.0.0 - x.255.255.255
  - x.x.x.x/0 (2^32) - 0.0.0.0 - 255.255.255.255
- Example
  - CIDR VPC 10.97.224.0/20
  - You have 32-20 = 12 bits which is 2^12 = 4096 IP
- **Internet Gateway** will expose the Subnet to internet and make it **Public**
  - It is attached to the VPC
  - In the **route table** we associate Internet Gateway to make Subnet Public

ENi (Elastic Network Interface)

ENI provides 3 ways of attaching to an EC2 instance:

- **Hot attach**
  When the ENI is attached to a Running EC2 instance
- **Warm attach**
  When the EC2 instance is stopped
- **Cold attach**
  When the EC2 instance is being launched

CloudWatch

- Aggregates logs file from multiple sources, also CloudTrail logs.
- In order to get CloudTrail logs, you need to configure an **IAM Role so that CloudTrail** can stream logs into CloudWatch.
- You can trigger an **Alarm** if a CloudTrail event trigger with specific logic/metrics. Alarms take up to **15 min**.

Athena

- Search content in files inside S3 buckets
- It uses SQL query language, which requires a **structure** in your data

VPC Endpoint | ACL

VCP Endpoint are used to allow traffic between services, by traversing private link via VPC endpoint.

ACL (Network Control Access List)

- Can be applied to multiple Subnet
- Stateless
- Rules are ordered and applied ascending so a second rule with DENY will not override a previous rule
- A Rule can **ALLOW** or **DENY**

Security Group

- Applied to an instance level
- Stateful
- Order of rules doesn't matter
- A Rule only **ALLOW**

Internet Check Steps for EC2

- An Internet gateway must be attached to the VPC
- Associate Route Table to Public Subnet
- Create a Route to the Internet Gateway
- Allow traffic using a Security Group
- EC2 must have a Public IP

Data at Rest

Two different ways to protect data:

- **Access permissions**
  - Bucket policies
  - User policies
  - Access Control List
- **Encryption**
  - Access to key to encrypt

S3 Access

When you create a Bucket, the creator becomes owner via **ACL**.

- I can create a new Policy and assign it to the Bucket for specific Objects, Folders and Wildcards

```
{
  "Id": "Policy1659076348174",
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "Stmt1659076327136",
      "Action": "s3:*",
      "Effect": "Allow",
      "Resource": "arn:aws:s3:::s3-
encrypted-raf/test/*",
      "Principal": {
        "AWS": [
          "arn:aws:iam::925885294484:user/
raffaeu-devops"
        ]
      }
    }
  ]
}
```

KMS (Keys Management Service)

Allows you to create keys. You can create custom CMK to encrypt content.

- A **Key Administrator** can change the key, can use but cannot read it by default
- You can add *external Principals* to read or manage the key
- A **Policy** is created to interact with the key, for example Read, Manage and so on

If you want to **Encrypt an existing EBS** you need to stop the EC2 and convert an EBS into an encrypted one

S3 Encryption

You can use CMK to encrypt content. **Remember, who access needs also permissions to the CMK!!** Otherwise the user won't even see the uploaded content

- **AES-256**
  Managed by AWS and customer cannot access it
- **AWS-KMS**
  Managed by the customer via KMS

Also, **existing objects are not affected** if you change Encryption.

Data in Transit

TLS (Transport Layer Security)

- TLS is the protocol, HTTPS is the way we carry data

Configure TLS

- Not interconnected to AWS

Application Load balancer

- Force Clients to connect to ALB

ACM (AWS Certificate Manager) is used to generate a TLS Certificate and assign it to an ALB

Data Versioning

Prevent accidental delition or override of data.

- Applies to all S3 objects within a Bucket
- If you delete an object, you'll see as a version as **Deleted**
  - You can Delete the Deleted version, to restore it

Replication

A **Bucket** is only in one **Region** by default.

- You can setup a new Bucket to replicate data to
- **Beware**, if you are using **KMS**, also the destination Bucket needs to be able to access the Encryption mechanism
- If **Versioning** is in place, destination Bucket also needs Versioning

Lifecycle management

Different S3 classes have different level of redundancy. Default is **Standard**.

Lifecycle rotate the storage class of objects to a cheaper class, or delete them, based on age.

- **Lifecycle rules** trigger the action registered inside the Lifecycle management of your S3
- You can choose the nr. of days before triggering an action, a transition

Beware **less than 128KB** file size do not trigger Lifecycle rules!!

Cost Effective Storage

- Understand the different storage solution keys
  - Right size EBS
  - Choose appropriate volume type for EBS
- Identify ways to monitor and track Costs
  - How to use CloudWatch to monitor usage
  - Trusted advisor for underutilized resources
- Beware of snapshots (***they are not free***)
- S3 logic to store and access objects
- EBS Lifecycle Management (DLM), very important
  - EBS Backup mechanisms

Cost Optimized Network

- Consider API costs and Network traffic costs
  - Reduce costs via CloudFront, for example
- How to reduce data transfer costs
  - AWS Regions data transfer (VPC Peering is more expensive than AWS Transit Gateway)
  - On-premise to cloud (Direct connect is more expensive)
- EC2 connectivity
  - RDP, SSH vs AWS System Manager
  - ***Session Manager*** is the cheapest and safest
- NAT Gateway and Transit Gateway differences in costs

S3 Data Retrieval options

S3 Glacier Instant
- Milliseconds response
- Rarely used

S3 Glacier Flexible
- Expedited - within 1-5 minutes
- Standard - within 3-5 hours
- Free Bulk - within 5-12 hours

S3 Deep Archive
- Bulk - within 48 hours
- Standard - within 12 hours

You can query data before retrieval, wihitout actually retrieving it

- Object storage with shared access
- Low latency and high throughput
- **Encryption**
  - Server side Encryption managed by AWS
  - Client side Encryption, Client encrypt before sending
- **Access**
  - By default everything is private

S3 Transfer Acceleration minimize latency

Cost Effective Compute and DB

- Understand pricing options (saving plans, on-demand ...)
- Select RDS scaling strategies
  - Read replica, caching mechanisms
- Always opt for ***managed services*** when possibles (Lambda, Fargate, DynamoDB)

S3 Storage Classes

Standard
- By default when creating a new object
- *Most expensive*
- For objects ***accessed frequently***, charges by Object Size
- Multiple AZs

Standard IA (Infrequent Access)
- Cheaper than standard
- Retrieval is more expensive
- Multiple AZs

Intelligent tiering
- Look at access pattern (Standard or AI)
- Small fee for transition objects

One zone IA
- Single AZ so it is cheaper than others

Glacier
- Lowest costs but long time retrieval
- Size does not affect

Lifecycle rules
- Can expire and delete objects
- Can transition Object classes
- Can be applied to ***Prefix*** or to ***Tags***

EBS Storage Classes

EBS
- Provision immediate size that you provision, it is all allocated and charged upfront
- **SSD** - involve frequent Read/Write
- **HDD** - throughput more important that IOPS
- **Cost based**
  a. Cold HDD
  b. Thorughput optimized HDD
  c. General Purpose SSD
  d. Provisioned IOPS SSD

Snapshot
Charge only the space used, not the entire EBS space

EC2 cost effective

Payment Types
- You are charged by time, shut-down when not use it
- EBS is charged **always**
- **On-Demand**
  Charged by hour/second, no contract, no upfront commitment
- **Reserved (up to 72%)**
  Cost saving. 1 or 3 years commitment. You can sell on Marketplace. Pay all upfront or by month
  - *Standard* or *Convertible*
- **Spot (up to 90%)**
  Bid of unusued capacity of AWS. 1 Spot, ***Spot Fleets*** (several) or ***Spot Blocks*** (only guarantee time)
- **Scheduled Reserved**
  Only for ***Periodical*** jobs. Discount applies to a specific timeframe

Snapshot
Charge only the space used, not the entire EBS space

Saving Plans
Can be applied to multiple services, not only EC2

RDS Pricing

- Instance type and size because RDS is under an EC2 instance(s)
- Database Storage
- Only Aurora is serverless
- Data transfers between AZs and Regions

Types
- **General Purpose**
  M4 and M5 good balanced instance
- **Memory Optimized**
  R4 and R5 good for high memory intensive
- **Burstable Performance**
  Baseline CPU that you can burst above

Payments
- On-deman
- Reserved save up to 69% with 1 or 3 years contract

Storage
- General Purpose
- Provisioned IOPS SSD (more expensive)
- Magnetic is cheaper but not suggested

EC2 sizing

Burstable vs Fixed
- Fixed is reserving power for you (i.e. M5)
- Burstable you give baseline that can span (i.e. T3)
  - Better for intermitted workload

Tools for right sizing
- ***Amazon CloudWatch*** to metrics CPU and other hardware
- ***AWS Cost Explorer***
  Optimization reccomendations
- ***Trusted Advisor***
  Best practice tool, reccomendations on your setup

AWS Serverless Alternatives

Compute
- AWS Lambda
- AWS Fargate

Data stores
- DynamoDB
- Aurora

API Proxy
- Amazon API Gateway

Integration
- Amazon SNS
- Amazon SQS

Storage
- Amazon S3

DynamoDB Pricing

Charging
- ***On-demand***
  Charged for data reads and writes
- ***Provisioned***
  You buy the read and write capacity. If you pass over, connections will drops

Capacity Units
- Write capacity unit (***WCU***) = 1KB/s
- Read capacity unit (***RCU***) = beware on Transaction consistency level = 4KB/s

Additional considerations
- DynamoDB auto-scaling
- Reserve 1 to 3 years RCU and WCU
- Global secondary indexes have extra costs
- Global DynamoDB Table have extra costs because it goes to another region
- Backups is not free

Factors that affect costs

- Some Regions have higher costs than other for the same Service
- EC2 size and type and generation
- S3 storage class have highest impact on cost

Useful Tips

1. EC2 Payment based on scenario
2. Use Reservation for RDS and DynamoDB
3. TAG everything in order to proper breakdown billings
4. SCP (Service Control Policies) to restrict features
5. CloudWatch, monitor everything to check spare capacity

Optimize data transfers

AWS Snowcone

- Edge Computing and Data Transfer
- HDD = 8TB
- SSD = 14TB

AWS Snowball Edge Storage Optimized

- Data Transfer and Edge Computing
- HDD = 80TB
- SSD = 1TB

AWS Snowball Edge Compute Optimized

- Edge Computing and Data Transfer
- HDD = 42TB
- SSD = 7.68TB

AWS Snowmobile

- Data transfer
- HDD = 100PB

Storage cost effective

Cached Volumes

- Minimize the needs of expand on-premise storages. Frequent access and low-latency

File Gateway

- Its primary use is for creating file shares between on-premise and AWS

Stored Volumes Gateway

- It is used primarily to backing up your on-premise data into Amazon

Tape Gateway

- It is used to move data from a tape, straight into Amazon

Efficient usage of TAGS

You can use TAGs for:

- *Cost allocation*
  You can create associations between TAGs and cost of a specific Business Unit
- *Automation*
  You can identify resources that can be subject to automation, for example turn-off during week-end
- *Operations Support*
  You can associate resources and map workflow of support
- *Access Control*
  You can restrict access via IAM by using TAGs
- *Security Risk Management*
  You can tag resources that holds sensitive data

AWS Framework Pillars

Operational Excellence

The ability to support development and run workloads effectively, gain insight into their operation, and constinuously improve supporting processes and procedures to deliver business value.

- Perform operations as code
- Make frequest, small, reversible changes
- Refine operations procedures frequently
- Learn from all operational failures

Security

The ability to protect data, systems, and assets to take advantage of cloud technologies to improve your security.

- Implement a strong identity fundation
- Enable traceability
- Apply security at all layers
- Automate security best practices
- Protect data in transit
- Keep people away from data
- Prepare for security events

Reliability

The ability of a workload to perform its intended function correctly and consistently when it's expected to.

- Automatically recover from failure
- Test recovery procedures
- Scale horizontally to increase aggregate workload
- Stop guessing capacity
- Manage change in automation

Performance Efficiency

The ability to use computing resources efficinelty to meet system requirements and to maintain that efficiency as demand changes.

- Democratize advanced technologies
- Go global in minutes
- Use serverless architectures
- Experiment more often
- Consider mechanical simpathy

Cost Optimization

The ability to run systems to deliver business value at the lowest price point.

- Implement cloud financial management
- Adopt a comsumption model
- Measure overall efficiency
- Stop sp[ending money on heavy lifting
- Analyze and attribute expediture

Sustainability

It addresses the long-term environmental, economic, and societal impact of your business activities.

- Understand your impact
- Establish sustainability goals
- Maximize utilization
- Anticipate and adopt new and more efficient hardware
- Use managed services
- Reduce the downstream impact of your cloud workloads