AWS Cloud Practitioner Certification Notes

# 1. Introduction to AWS

## 3 types of deployment

1. **Cloud-based deployment**: run all from the cloud
2. **On-premises deployment**: Deploy resources by using virtualization and resource management tools.
3. **Hybrid deployment**: Connect cloud-based resources to on-premises infrastructure.

## 6 Benefits of cloud computing

1. Trade upfront expense for variable expense
2. Stop spending money to run and maintain data centers
3. Stop guessing capacity
4. Benefit from massive economies of scale
5. Increase speed and agility
6. Go global in minutes

# 2. Cloud Computing

## 5 EC2 Instance Types

**1. General purpose Instances**

General purpose instances provide **a balance of compute, memory, and networking resources**. You can use them for a variety of workloads, such as **application servers, gaming servers, backend servers for enterprise applications, and small and medium databases**.

**2 Compute-Optimized Instances**

Compute-optimized instances are ideal **for high-performance web servers, compute-intensive applications servers, and dedicated gaming servers**. You can also use compute optimized instances for batch processing workloads that require processing many transactions in a single group.

**3. Memory-Optimized Instances**

Memory optimized instances are designed to deliver fast performance for workloads that process large datasets in memory. Memory optimized instances enable you to run workloads with high memory needs and receive great performance.

**4. Accelerated-Computing Instances**

Accelerated computing instances use hardware accelerators, or coprocessors, to perform some functions more efficiently than is possible in software running on CPU, e.g., floating-point number calculations, graphics processing, and data pattern matching. Accelerated computing instances are ideal for workloads such as graphics applications, game streaming, and application streaming.

**5. Storage-Optimized Instances**

Storage optimized instances are designed for workloads that require high, sequential read and write access to large datasets on local storage. Examples of workloads suitable for storage optimized instances include distributed file systems, data warehousing applications, and high-frequency online transaction processing (OLTP) systems.

### 5 Amazon EC2 Pricing Models

With Amazon EC2, you pay only for the compute time that you use.

**1.** **On-Demand**: ideal for short-term, irregular workloads that cannot be interrupted. No upfront costs or minimum contracts apply. The instances run continuously until you stop them, and you pay for only the compute time you use. Sample use cases for On-Demand Instances include developing and testing applications and running applications that have unpredictable usage patterns. On-Demand Instances are not recommended for workloads that last a year or longer because these workloads can experience greater cost savings using Reserved Instances.

**2.** **Amazon EC2 Savings Plans**: AWS offers Savings Plans for several compute services, including Amazon EC2. Amazon EC2 Savings Plans enable you to reduce your compute costs by committing to a consistent amount of compute usage for a 1-year or 3-year term. This term commitment results in savings of up to 66% over On-Demand costs. **AWS Cost Explorer** can analyze your Amazon EC2 usage over the past 7, 30, or 60 days. **AWS Cost Explorer** also provides customized recommendations for Savings Plans.

**3.** **Reserved Instances**: Reserved Instances are a billing discount applied to the use of On-Demand Instances in your account. You can purchase **Standard Reserved and Convertible Reserved Instances** for a 1-year or 3-year term and Scheduled Reserved Instances for a 1-year term. You realize greater cost savings with the 3-year option.

**4.** **Spot Instances**: Spot Instances are ideal for workloads with flexible start and end times, or that can withstand interruptions. Spot Instances use unused Amazon EC2 computing capacity and offer you cost savings at up to 90% off On-Demand prices.

5. **Dedicated Hosts**: Dedicated Hosts are physical servers with Amazon EC2 instance capacity that is fully dedicated to your use. You can use your existing per-socket, per-core, or per-VM software licenses to help maintain license compliance. You can purchase On-Demand Dedicated Hosts and Dedicated Hosts Reservations. Of all the Amazon EC2 options that were covered, **Dedicated Hosts are the most expensive**.

## 4 Settings for AWS EC2 Auto Scaling

Diagram

Description automatically generated

1. **Minimum Capacity**: your application needs a minimum of a single Amazon EC2 instance to run
2. **Desired Number**: defaults to minimum capacity if not set
3. **Maximum Amazon EC2 Instances**: the maximum number of EC2 Instances

## Elastic Load Balancing

Elastic Load Balancing is the AWS service that automatically distributes incoming application traffic across multiple resources, such as Amazon EC2 instances. A load balancer acts as a single point of contact for all incoming web traffic to your Auto Scaling group.

## 2 Types of Messaging and Queuing Services

1. **Amazon Simple Notification Service (SNS):** Amazon Simple Notification Service (Amazon SNS) is a publish/subscribe service. Using Amazon SNS topics, a publisher publishes messages to subscribers. In Amazon SNS, subscribers can be web servers, email addresses, AWS Lambda functions, or several other options.
2. **Amazon Simple Queue Service (SQS):** a message queuing service that can send, store, and receive messages between software components, without losing messages or requiring other services to be available. In Amazon SQS, an application sends messages into a queue. A user or service retrieves a message from the queue, processes it, and then deletes it from the queue.

## AWS Lambda – Serverless Computing

AWS Lambda is one serverless compute option. Lambda's a service that allows you to upload your code into what's called a Lambda function. Configure a trigger and from there, the service waits for the trigger. When the trigger is detected, the code is automatically run in a managed environment, an environment you do not need to worry too much about because it is automatically scalable, highly available and all the maintenance in the environment itself is done by AWS. If you have one or 1,000 incoming triggers, Lambda will scale your function to meet demand. Lambda is designed to run code under 15 minutes, so this isn't for long running processes like deep learning.

## Amazon Elastic Container Service (Amazon ECS)

Amazon Elastic Container Service (Amazon ECS) is a highly scalable, high-performance container management system that enables you to run and scale containerized applications on AWS.

Amazon ECS supports Docker containers.. With Amazon ECS, you can use API calls to launch and stop Docker-enabled applications.

## Amazon Elastic Kubernetes Service (Amazon EKS)

Amazon Elastic Kubernetes Service (Amazon EKS) is a fully managed service that you can use to run Kubernetes on AWS.

## AWS Fargate

Both Amazon ECS and Amazon EKS can run on top of EC2. But if you don't want to even think about using EC2s to host your containers because you either don't need access to the underlying OS or you don't want to manage those EC2 instances, you can use a compute platform called **AWS Fargate**. Fargate is a serverless compute platform for ECS or EKS. **When using AWS Fargate, you do not need to provision or manage servers**. AWS Fargate manages your server infrastructure for you. You can focus more on innovating and developing your applications, and you pay only for the resources that are required to run your containers.

# Global Infrastructure & Reliability

### AWS Regions

Throughout the globe, AWS builds Regions to be closest to where the business traffic demands. Locations like Paris, Tokyo, Sao Paulo, Dublin, Ohio. Inside each Region, we have multiple data centers that have all the compute, storage, and other services you need to run your applications. Each Region can be connected to each other Region through a high-speed fiber network, controlled by AWS.

Regional data sovereignty is part of the critical design of AWS Regions. With data being subject to the local laws and statutes of the country where the Region lives. So with that understanding, that your data, your application, lives and runs in a Region, one of the first decisions you get to make is which Region do you pick?

#### 4 business factors that go into choosing a Region:

1. Compliance
2. Proximity
3. Feature Availability
4. Pricing

### AWS Availability Zones

AWS calls a single data center or a group of data centers, an Availability Zone or AZ. Each Availability Zone is one or more discrete data centers with redundant power, networking, and connectivity. When you launch an Amazon EC2 instance, it launches a virtual machine on a physical hardware that is installed in an Availability Zone. This means each AWS Region consists of multiple isolated and physically separate Availability Zones within a geographic Region.

**And as a best practice with AWS, we always recommend you run across at least two Availability Zones in a Region. This means redundantly deploying your infrastructure in two different AZs.**

Many of the AWS services run at the Region level, meaning they run synchronously across multiple AZs without any additional effort on your part. Take the ELB. This is a regional construct. It runs across all Availability Zones, communicating with the EC2 instances that are running in a specific Availability Zone. Regional services are already highly available at no additional cost of effort on your part.

## Edge Locations

CDNs are commonly used, and on AWS, we call our CDN **Amazon CloudFront**. Amazon CloudFront is a service that helps deliver data, video, applications, and APIs to customers around the world with low latency and high transfer speeds. Amazon CloudFront uses what are called Edge locations, all around the world, to help accelerate communication with users, no matter where they are.

A picture containing application

Description automatically generated

## 3 take-aways about global infrastructure and reliability:

1. Regions are geographically isolated areas, where you can access services needed to run your enterprise.
2. Regions contain Availability Zones, that allow you to run across physically separated buildings, tens of miles of separation, while keeping your application logically unified. Availability Zones help you solve high availability and disaster recovery scenarios, without any additional effort on your part,
3. AWS Edge locations run Amazon CloudFront to help get content closer to your customers, no matter where they are in the world.

## 5 Ways to Provision AWS Resources:

1. **AWS Management Console**: Use the AWS UI
2. **AWS CLI Interface**: The CLI allows you to make API calls using the terminal on your machine. This is different than the visual navigation style of the Management Console. Writing commands using the CLI makes actions scriptable and repeatable.
3. **Software Development Kits**: The SDKs allow you to interact with AWS resources through various programming languages. This makes it easy for developers to create programs that use AWS without using the low-level APIs, as well as avoiding that manual resource creation.
4. **AWS Elastic Beanstalk**: a service that helps you provision Amazon EC2-based environments. You can provide your application code and desired configurations to the AWS Elastic Beanstalk service, which then takes that information and builds out your environment for you.
5. **AWS CloudFormation**: AWS CloudFormation is an infrastructure as code tool that allows you to define a wide variety of AWS resources in a declarative way using JSON or YAML text-based documents called CloudFormation templates.

# Networking

## Connectivity to AWS

A **VPC, or Virtual Private Cloud**, is essentially your own private network in AWS. A VPC allows you to define your private IP range for your AWS resources, and you place things like EC2 instances and ELBs inside of your VPC.

Now you don't just go throw your resources into a VPC and move on. You place them into different subnets. Subnets are chunks of IP addresses in your VPC that allow you to group resources together. Subnets, along with networking rules we will cover later, control whether resources are either publicly or privately available.

## Internet Gateway

To allow traffic from the public internet to flow into and out of your VPC, you must attach what is called an **Internet gateway, or IGW, to your VPC.**

Graphical user interface, diagram, application

Description automatically generated

### Virtual Private Gateway

Next, let's talk about a VPC with all internal private resources. We don't want just anyone from anywhere to be able to reach these resources. So, we don't want an internet gateway attached to our VPC. Instead, we want a private gateway that only allows people in if they are coming from an approved network, not the public internet. This private doorway is called a virtual private gateway, and it allows you to create a VPN connection between a private network, like your on-premises data center or internal corporate network to your VPC.

Timeline

Description automatically generated

## AWS Direct Connect

**AWS Direct Connect** allows you to establish a completely private, dedicated fiber connection from your data center to AWS. You work with a Direct Connect partner in your area to establish this connection, because like my magic doorway, AWS Direct Connect provides a physical line that connects your network to your AWS VPC.

Diagram

Description automatically generated

## Subnets and Network Access Control Lists

You have a gateway on the VPC that only permits traffic in or out of the VPC. But that only covers perimeter, and that's only one part of network security that you should be focusing on as part of your IT strategy.

### 6 layers of AWS security:

1. network hardening,
2. application security,
3. user identity,
4. authentication and authorization,
5. distributed denial-of-service or DDoS prevention,
6. data integrity, encryption, much more.

### 4 mechanisms of AWS network security:

1. **Gateways**: Internet & Virtual Private
2. **Subnets**: the only technical reason to use subnets in a VPC is to control access to the gateways. The public subnets have access to the internet gateway; the private subnets do not. But subnets can also control traffic permissions
3. **Network ACLs**: every packet that crosses the subnet boundaries gets checked against a **network** **access control list or network ACL (Stateless)**. This check is to see if the packet has permissions to either leave or enter the subnet based on who it was sent from and how it's trying to communicate.
4. **Security Groups**: To secure in-network traffic AWS uses security groups. Every EC2 instance, when it's launched, automatically comes with a security group. And by default, the security group does not allow any traffic into the instance at all. All ports are blocked; all IP addresses sending packets are blocked. Security groups are configured to allow specific traffic in and, by default, all traffic is allowed out.

### NACLS vs Security Groups

If NACLs are a passport control, a security group is like the doorman at your building, the building being the EC2 Instance, in this case. The doorman will check a list to ensure that someone is allowed to enter the building but won't bother check the list on the way out. With security groups, you allow specific traffic in and by default, all traffic is allowed out.

The key difference between a security group and a network ACL is the security group is stateful, meaning, as we talked about, it has some kind of a memory when it comes to who to allow in or out, and the network ACL is stateless, which remembers nothing and checks every single packet that crosses its border regardless of any circumstances.

## 2 Global Networking Service

1. **Route 53:**  is AWS's domain name service, or DNS, and it's highly available and scalable. Route 53 can direct traffic to different endpoints using several different routing policies, such as
   * latency-based routing,
   * geolocation DNS,
   * geoproximity,
   * and weighted round robin.
2. **Amazon CloudFront**. If you remember, we talked about Edge locations earlier in the course, these locations are serving content as close to customers as possible, and one part of that, is the content delivery network, or CDN. For those who need reminding, a CDN is a network that helps to deliver edge content to users based on their geographic location.