Designing Applications and Architectures in AWS

(Aligned with AWS Solution Architect Associate Certification)

©Simplilearn. All rights reserved.

simplilearn

Compute and Related Features

©Simplilearn. All rights reserved.

A Day in the Life of a Cloud Architect

You are a cloud architect for an IT company with multiple cloud-based applications.

The organization has encountered routing challenges that have resulted in significant financial losses. Hence, you have been asked to ensure that such issues are proactively prevented in the future.

By the end of this lesson, you will have a concise understanding of crafting routing requests, creating AMI images, and exploring the types of load balancers.

©Simplilearn. All rights reserved.

Learning Objectives

By the end of this lesson, you will be able to:

- Create an AMI image for fast deployment
- Launch and connect to a Windows instance to enable disaster recovery
- Deploy various load balancers to improve scalability
- Create a routing request in ALB to enable path-based routing

©Simplilearn. All rights reserved.

TECHNOLOGY

Introduction to Amazon EC2

©Simplilearn. All rights reserved.

simplilearn

Elastic Compute Cloud (EC2)

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides scalable computing capacity in the Amazon Web Services (AWS) cloud.

©Simplilearn. All rights reserved.

Benefits of EC2

- Elastic Web-scale Computing
- Flexible Cloud Hosting Services
- AWS Integration
- Reliability and Security
- Low Cost
- Completely Controlled

- EC2 increases or decreases the storage capacity in minutes.
- It launches thousands of server instances simultaneously.

©Simplilearn. All rights reserved.

Benefits of EC2

- Elastic Web-scale Computing
- Flexible Cloud Hosting Services
- AWS Integration
- Reliability and Security
- Low Cost
- Completely Controlled

- EC2 launches numerous operating systems, instance types, and software in minutes.
- It allows the users to choose the memory, CPU, instance storage, and boot partition size that is best for their applications and OS.

©Simplilearn. All rights reserved.

Benefits of EC2

- Elastic Web-scale Computing
- Flexible Cloud Hosting Services
- AWS Integration
- Reliability and Security
- Low Cost
- Completely Controlled

EC2 is integrated with other AWS products, such as Amazon S3, Amazon RDS, and Amazon SQS, to provide a complete IT architecture solution.

©Simplilearn. All rights reserved.

Benefits of EC2

- Elastic Web-scale Computing
- Flexible Cloud Hosting Services
- AWS Integration
- Reliability and Security
- Low Cost
- Completely Controlled

- Amazon EC2 provides a highly reliable environment where replacement instances can be quickly and consistently deployed.
- Users can easily create secure and robust networks to run their Amazon EC2 instances using the Virtual Private Cloud (VPC).

©Simplilearn. All rights reserved.

Benefits of EC2

- Elastic Web-scale Computing
- Flexible Cloud Hosting Services
- AWS Integration
- Reliability and Security
- Low Cost
- Completely Controlled

- AWS charges the users in seconds. They only pay for what they use.
- The rates are lower than the existing on-premise infrastructure.

©Simplilearn. All rights reserved.

Benefits of EC2

- Elastic Web-scale Computing
- Flexible Cloud Hosting Services
- AWS Integration
- Reliability and Security
- Low Cost
- Completely Controlled

- The users have complete control over their instances. They have root access to all the instances.
- The users can use web service APIs to stop their instance while keeping the data on their boot partition and resume it.

©Simplilearn. All rights reserved.

Amazon EC2 Storage

Amazon EC2 offers flexible, cost-effective, and simple data storage options for instances.

The storage options include the following:

- Amazon EBS with EC2
- Amazon EC2 instance store
- Amazon EFS with EC2
- Amazon S3 with EC2

©Simplilearn. All rights reserved.

Amazon EC2 Storage: Linux Instances

The following image depicts the relationship between the storage options and Linux instances:

©Simplilearn. All rights reserved.

Instances

An instance is a virtual server in the AWS cloud. Amazon EC2 offers various instance types, allowing users to select the CPU, memory, storage, and networking resources required to run their applications.

The instance types are:

1. General Purpose
2. Compute Optimized
3. Memory Optimized
4. Accelerated Computing
5. Storage Optimized

©Simplilearn. All rights reserved.

General Purpose Instances

An instance is a virtual server in the AWS cloud. Amazon EC2 offers various instance types, allowing users to select the CPU, memory, storage, and networking resources required to run their applications.

- These instances are useful for applications like web servers and code repositories that require such resources in equal parts.
- The examples are Mac, T4g, T3, T3a, T2, M6g, M6i, M6a, M5, M5a, M5n, M5zn, M4, and A1 instances.

©Simplilearn. All rights reserved.

Compute Optimized Instances

Compute optimized instances are suitable for compute-intensive applications that benefit from high-performance processors.

- These instances are useful for batch processing workloads, dedicated gaming servers and ad server engines, scientific modeling, and so on.
- The examples are C7g, C6g, C6gn, C6i, C6a, Hpc6a, C5, C5a, C5n, and C4 instances.

©Simplilearn. All rights reserved.

Memory Optimized Instances

Memory optimized instances are intended to provide rapid performance for workloads that process large data sets in memory.

- These instances are useful for databases, electronic design automation (EDA) workloads, real-time analytics, and real-time caching servers.
- The examples are R6g, R6i, R5, R5a, R5b, R5n, R4, X2gd, X2idn, X2iezn, X1e, X1, high memory, and z1d instances.

©Simplilearn. All rights reserved.

Accelerated Computing Instances

Accelerated computing instances use hardware accelerators or co-processors to execute tasks more effectively than software running on CPUs.

- These instances are useful for performing tasks such as floating-point number calculations, graphic processing, data pattern matching, and so on.
- The examples are P4, P3, P2, DL1, Trn1, Inf1, G5, G5g, G4dn, G4ad, G3, F1, and VT1 instances.

©Simplilearn. All rights reserved.

Storage Optimized Instances

Storage optimized instances are designed to provide applications with tens of thousands of low-latency, random I/O operations per second (IOPS).

- These instances are useful for workloads that demand high sequential read and write access to large data sets on local storage.
- The examples are Im4gn, Is4gen, I4i, I3, I3en, D2, D3, D3en, and H1 instances.

©Simplilearn. All rights reserved.

Amazon Machine Images (AMI)

An Amazon Machine Image (AMI) is an AWS-supported and maintained image that contains the information required to launch an instance.

- Users must specify an AMI when they launch an instance.
- Users can launch multiple instances from a single AMI when they require multiple instances with the same configuration.
- Users can use different AMIs to launch instances when they require instances with different configurations.

©Simplilearn. All rights reserved.

Amazon Machine Images (AMI)

AMI

AMI is a virtual instance that includes:

- A template for the root volume of the instance
- Launch permissions to control AMI launch instances
- A block device mapping that specifies volumes to attach to the instance

©Simplilearn. All rights reserved.

AMI Lifecycle

The following diagram represents the AMI lifecycle:

- After creating and registering an AMI, users can use it to launch new instances.
- Users can copy an AMI within the same AWS Region or to different AWS Regions.
- When an AMI is no longer required, users can deregister it.

©Simplilearn. All rights reserved.

Amazon Linux AMI

It is a Linux image offered by AWS that is both officially supported and continuously maintained.

The following are the features of the Amazon Linux AMI:

Updates regularly to include the latest components

Offers components at no additional charge to Amazon EC2 users

Allows repository access to MySQL, PostgreSQL, Python, Ruby, and Tomcat

Provides a secure, stable, and high-performance execution environment

Includes packages that enable easy integration with AWS services

©Simplilearn. All rights reserved.

AMI Distribution

An AMI can be kept private, shared with a specific list of AWS accounts, or made public.

©Simplilearn. All rights reserved.

AMI Types

An AMI can be categorized as either an EBS-backed AMI or an instance store-backed AMI.

Characteristic	Amazon EBS-backed AMI	Amazon instance store-backed AMI
Boot time for an instance	Usually less than 1 minute	Usually less than 5 minutes
Size limit for a root device	64 TiB	10 GiB
Root device volume	EBS volume	Instance store volume
Data persistence	By default, the root volume is deleted when the instance terminates. Data on any other EBS volumes persists after instance termination by default.	Information stored on instance store volumes remains intact solely throughout the duration of the instance's existence.

©Simplilearn. All rights reserved.

Creating an AMI Image

Duration:10 min

Problem Statement:

You have been assigned a task to create an AMI image from an EC2 instance.

ASSISTED PRACTICE

©Simplilearn. All rights reserved.

Assisted Practice: Guidelines

Steps to be followed are:

1. Create an AMI image

©Simplilearn. All rights reserved.

Automating App Installation on EC2 Instance

Duration:10 min

Problem Statement:

You have been assigned a task to automate application installation during the EC2 instance launch in AWS using User Data.

ASSISTED PRACTICE

©Simplilearn. All rights reserved.

Assisted Practice: Guidelines

Steps to be followed are:

1. Install apps when launching an instance

©Simplilearn. All rights reserved.

Placement Groups

©Simplilearn. All rights reserved.

Placement Groups

These help the users to deploy a group of interdependent instances to handle their workload more effectively.

There are three types of placement groups based on the type of workload:

- Cluster placement groups
- Partition placement groups
- Spread placement groups

©Simplilearn. All rights reserved.

Placement Groups

The general rules and limitations of placement groups are as follows:

- A maximum of 500 placement groups per account can be created in each Region.
- The name of a placement group must be unique within the AWS account for the Region.
- The placement groups cannot be merged.

©Simplilearn. All rights reserved.

Placement Groups

The general rules and limitations of placement groups are as follows:

- An instance can only be launched in one placement group at a time.
- Users cannot launch Dedicated Hosts in placement groups.
- A zonal Reserved Instance cannot be used to explicitly reserve capacity in a placement group.

©Simplilearn. All rights reserved.

Cluster Placement Groups

It is a logical grouping of instances inside a single Availability Zone.

Instances deployed in the same cluster placement group share a high-bisection bandwidth network segment and have a higher TCP or IP traffic throughput limit per flow.

©Simplilearn. All rights reserved.

Cluster Placement Groups

The rules and limitations of cluster placement groups are as follows:

- A cluster placement group cannot be deployed across multiple Availability Zones.
- The maximum network throughput speed of traffic between two instances is limited by the slower of the two instances.
- The cluster placement group instances can use up to 10 Gbps for single-flow traffic.

©Simplilearn. All rights reserved.

Cluster Placement Groups

The rules and limitations of cluster placement groups are as follows:

- Users can launch multiple instance types into a cluster placement group.
- The network traffic to the internet and on-premises resources via an AWS Direct Connect connection is limited to 5 Gbps.
- All available instance aggregate bandwidth can be used for traffic to and from Amazon S3 buckets across the same Region.

©Simplilearn. All rights reserved.

Partition Placement Groups

A partition placement group distributes the instances across logical partitions so that groups of instances in one partition do not share the same underlying hardware as groups of instances in other partitions.

Large distributed and replicated workloads can use it, such as Hadoop, Cassandra, and Kafka.

©Simplilearn. All rights reserved.

Partition Placement Groups

The rules and limitations of partition placement groups are as follows:

- A partition placement group can accommodate a maximum of seven partitions per Availability Zone.
- Amazon EC2 offers no guarantee that instances are evenly allocated across all partitions.
- A partition placement group containing Dedicated Instances can only have two partitions.
- Capacity Reservations cannot be used to reserve capacity in a partition placement group.

©Simplilearn. All rights reserved.

Spread Placement Groups

A spread placement group distributes a small group of instances across different underlying hardware to avoid correlated failures.

The placement groups can distribute instances among hosts or racks. The users can only use host-level spread placement groups with AWS Outposts.

©Simplilearn. All rights reserved.

Spread Placement Groups

The rules and limitations of spread placement groups are as follows:

- A rack-spread placement group can support a maximum of seven instances per Availability Zone.
- The spread placement groups are not supported for Dedicated Instances.
- The host-level spread placement groups are only accessible for AWS Outpost placement groups.
- Capacity Reservations cannot be used to reserve capacity in a spread placement group.

©Simplilearn. All rights reserved.

EC2 Instance Metadata

©Simplilearn. All rights reserved.

EC2 Instance Metadata

- Instance metadata is the data about an instance that users can use to configure or manage it. It's divided into sections like hostname, events, and security groups.
- Instance metadata allows users to access user data specified when the instance was launched.

Note

- Metadata is information about the EC2 instance.
- User data is the launch script of the EC2 instance.

©Simplilearn. All rights reserved.

EC2 Instance Metadata

Users can access instance metadata from a running instance using one of the following methods:

01 **Instance Metadata Service Version 1 (IMDSv1):** A request or response method

02 **Instance Metadata Service Version 2 (IMDSv2):** A session-oriented method

©Simplilearn. All rights reserved.

simplilearn

EC2 Instance Metadata

To view all the categories of instance metadata from within a running instance, use the following URL:

IMDSv1	IMDSv2
curl http://169.254.169.254/latest/meta-data/	TOKEN=`curl -X PUT "http://169.254.169.254/latest/api/token" -H "X-aws-ec2-metadata-token-ttl-seconds: 21600"` \ && curl -H "X-aws-ec2-metadata-token: $TOKEN" -v http://169.254.169.254/latest/meta-data/

©Simplilearn. All rights reserved.

Instance Metadata Categories

The different categories of instance metadata are:

ami-id	instance-type	public-hostname
ami-launch-index	local-hostname	public-ipv4
ami-manifest-path	local-ipv4	public-keys/
block-device mapping/	mac	reservation-id
hostname	metrics/	security-groups
iam/	network/	services/
instance-action	placement/	spot/
instance-id	profile	tags/

©Simplilearn. All rights reserved.

Instance Metadata Service Version 2

Instance Metadata Service Version 2 (IMDSv2) uses session-oriented requests.

With session-oriented requests:

- Users generate a session token that specifies the session duration, which can range from one second to six hours.
- Users can further request to use the same session token during the specified duration.
- Users need to create a new session token to use for subsequent queries once the time limit has passed.

©Simplilearn. All rights reserved.

IAM Roles

An IAM role is included with an instance profile on Amazon EC2. The IAM console automatically creates an instance profile when users create an IAM role, giving it the same name as the role it belongs to.

©Simplilearn. All rights reserved.

Connection Between IAM and EC2

The connection between IAM Roles and EC2 instances by applying necessary permissions:

©Simplilearn. All rights reserved.

Connection Between IAM and EC2

©Simplilearn. All rights reserved.

EC2 Instance Purchasing Options

Amazon EC2 provides the following purchasing options to enable users to optimize their costs based on their needs:

©Simplilearn. All rights reserved.

Spot Instance Interruptions

Spot Instances can be launched on spare EC2 capacity at substantial savings in exchange for returning them when Amazon EC2 requires the capacity back.

©Simplilearn. All rights reserved.

Reasons for Interruption

1. Capacity
2. Price
3. Constraints

©Simplilearn. All rights reserved.

Hibernate Interrupted Spot Instances

When an Amazon EC2 Spot Instance is hibernated, the following happens:

- When the instance gets a signal from Amazon EC2, the agent instructs the operating system to enter a state of hibernation.
- The root volume retains the instance memory (RAM).
- Other than Elastic IP addresses, instance storage volumes and public IP addresses are not maintained.

©Simplilearn. All rights reserved.

Using IAM Roles to Access S3 Bucket

Duration:10 min

Problem Statement:

You have been assigned a task to securely access Amazon S3 (Simple Storage Service) buckets from an EC2 instance using IAM (Identity and Access Management) roles.

ASSISTED PRACTICE

©Simplilearn. All rights reserved.

Assisted Practice: Guidelines

Steps to be followed are:

1. Create an IAM role
2. Connect IAM Profile to EC2
3. Validate access to the S3 bucket

©Simplilearn. All rights reserved.

Starting and Accessing a Windows Instance on EC2

Duration:10 min

Problem Statement:

You have been assigned a task to launch a Windows-based EC2 instance on Amazon Web Services (AWS) and securely access it using the Remote Desktop Protocol (RDP).

ASSISTED PRACTICE

©Simplilearn. All rights reserved.

Assisted Practice: Guidelines

Steps to be followed are:

1. Launch an EC2 instance
2. Access the EC2 instance via RDP

©Simplilearn. All rights reserved.

Hibernation

©Simplilearn. All rights reserved.

Hibernation: Overview

The following diagram shows an overview of the hibernation process:

©Simplilearn. All rights reserved.

Hibernation: Overview

The changes that occur while hibernating a running instance are as follows:

- Amazon EC2 signals the operating system to perform hibernation.
- The instance moves to the stopped state.
- EBS volumes remain attached to it.
- The instance store volumes remain attached to the instance, but the data is lost.

©Simplilearn. All rights reserved.

Hibernation: Overview

The changes that occur while hibernating a running instance are as follows:

- It gets migrated to a new underlying host computer when it starts again.
- It boots up, and the operating system reads in the contents of the RAM from the EBS root volume once it starts.
- It retains its private IPv4 addresses and any IPv6 addresses.
- It gets assigned to a new IPV4 address by Amazon EC2 once it starts again.

©Simplilearn. All rights reserved.

Hibernation: Overview

The changes that occur while hibernating a running instance are as follows:

- While hibernating the instance with EC2-Classic, an Elastic IP address is disassociated from the instance.
- While hibernating a **ClassicLink** instance, it gets unlinked from the VPC to which it was linked.

©Simplilearn. All rights reserved.

Hibernation for EC2 Instances

The user will launch EC2 instances by setting them up as desired, hibernating them, and then bringing them back to life when the user needs them.

The hibernation process stores the in-memory state of the instance along with its private and elastic IP addresses, allowing it to pick up exactly where it left off.

©Simplilearn. All rights reserved.

Hibernation for EC2 Instances

The feature is presently available, and users can use it on freshly launched M3, M4, M5, C3, C4, C5, R3, R4, and R5 instances running Amazon Linux 1.

It applies to on-demand instances and a few others that are running with reserved instance coverage.

©Simplilearn. All rights reserved.

Hibernation for EC2 Instances

The encryption ensures proper protection for sensitive data when copied from memory to the EBS volume.

The AMI used to launch the instance must be encrypted.

©Simplilearn. All rights reserved.

Hibernation for EC2 Instances

The encryption ensures proper protection for sensitive data when copied from memory to the EBS volume.

A user only pays for the EBS volumes and Elastic IP Addresses associated with the instance while it is in hibernation; there are no additional hourly fees.

©Simplilearn. All rights reserved.

Important Points for Hibernating an Instance

Instance Type: Users can enable hibernation and use it on freshly launched instances.

Root Volume Size: The root volume must have free space equal to the amount of RAM on the instance for the hibernation to succeed.

©Simplilearn. All rights reserved.

Important Points for Hibernating an Instance

Operating Systems: The newest Amazon Linux 1 AMIs are configured for hibernation. Users need to create an encrypted AMI, using one AMI as a base.

Modifications: Users cannot modify the instance size or type while it is in hibernation but can modify the user data and the EBS Optimization setting.

©Simplilearn. All rights reserved.

Important Points for Hibernating an Instance

Pricing: While the instance is in hibernation, users only need to pay for the EBS storage and any Elastic IP addresses attached to the instance.

Performance: The time to hibernate depends on the memory size of the instance, the amount of in-memory data to be saved, and the throughput of the root EBS volume.

©Simplilearn. All rights reserved.

Limitations

The limitations regarding hibernating an instance are:

- While hibernating an instance, the data on any instance store volumes is lost.
- Users can not hibernate an instance that has more than 150 GB of RAM.
- A user may not be able to connect to a new instance launched from a snapshot if the snapshot is created from an instance that is hibernating.

©Simplilearn. All rights reserved.

Limitations

The limitations regarding hibernating an instance are:

- Users can't change the instance type or size of an instance when hibernation is enabled.
- Users can't hibernate an instance that is in an auto scaling group or used by Amazon ECS.
- Users can't hibernate an instance configured to boot in UEFI mode.

©Simplilearn. All rights reserved.

Limitations

The limitations regarding hibernating an instance are:

- If the user is working on an instance that has been hibernating, it may not resume after the user tries to start it.
- AWS does not support keeping an instance hibernated for more than 60 days.
- AWS constantly updates its platform, which can conflict with existing hibernated instances.

©Simplilearn. All rights reserved.

Updates on AWS

AWS is working on support for Amazon Linux 2, Ubuntu, Windows Server 2008 R2, Windows Server 2012, Windows Server 2012 R2, and Windows Server 2016, along with the SQL Server variants of the Windows AMIs.

Microsoft® SQL Server®

Ubuntu

©Simplilearn. All rights reserved.

EBS

©Simplilearn. All rights reserved.

Introduction to Storage

AWS offers scalable, reliable, and secure cloud services with faster data access.

©Simplilearn. All rights reserved.

Storage Units

AWS provides a variety of options, including object storage, file storage, block storage, and data migration. Some of the examples are as follows:

EBS

EFS

S3

FSX

©Simplilearn. All rights reserved.

EBS Overview

Amazon Elastic Block Store (Amazon EBS) provides block-level storage volumes for use with EC2 instances.

©Simplilearn. All rights reserved.

Benefits of EBS

The benefits of Amazon Elastic Block Store (Amazon EBS) are as follows:

©Simplilearn. All rights reserved.

EBS Snapshot

Amazon EBS provides the ability to create snapshots (backups) of any EBS volume and write a copy of the data in the volume to Amazon S3, where it is stored redundantly in multiple Availability Zones.

- When performing incremental backups, back up only the blocks on the device that have changed after the most recent snapshot is saved.
- When a snapshot is deleted, only the data exclusive to that snapshot is removed.

©Simplilearn. All rights reserved.

EBS Optimized Instances

An Amazon EBS-optimized instance employs an optimized configuration stack and offers extra dedicated capacity for Amazon EBS I/O.

©Simplilearn. All rights reserved.

Amazon EBS vs. Instance Store

The following table shows the different characteristics between the EBS-backed and Amazon Instance store-backed volumes:

Characteristics	Amazon EBS	Instance Store
Boot time	Usually less than 1 minute	Usually less than 5 minutes
Root device volume	Amazon EBS volume	Instance store volume
Upgrading	EBS-backed instances can be upgraded for instance type, Kernel, RAM disk, and user data	Instance store-backed instances cannot be upgraded
AMI creation	AMI can be easily created using a single command	AMI creation requires AMI tools and needs to be executed from within the running instance
Volume attachment	EBS volumes can be attached as additional volumes when the instance is being launched and even when the instance is up and running	Instance store volumes can be attached as additional volumes only when the instance is being launched and not when the instance is up and running

©Simplilearn. All rights reserved.

Supported Instance

The supported instance types include the dedicated bandwidth to Amazon EBS and the typical maximum aggregate throughput that can be achieved on that connection with a streaming read workload.

Instance size	Maximum bandwidth (Mbps)	Maximum throughput (MB/s, 128 KiB I/O)	Maximum IOPS (16 KiB I/O)
c1.xlarge	1,000	125	8,000
c3.xlarge	500	62.5	4,000
g2.2xlarge	1,000	125	8,000
i2.xlarge	500	62.5	4,000
m1.large	500	62.5	4,000
r3.xlarge	500	62.5	4,000

©Simplilearn. All rights reserved.

Amazon EBS Volume

An Amazon EBS volume is a long-lasting, block-level storage device that can be attached to instances. You can use a volume attached to an instance, just like a physical hard disk drive.

©Simplilearn. All rights reserved.

EBS Volume Types

EBS Volume

Amazon EBS volume types can be listed into :

- Solid state drives (SSD)
- Hard disk drives (HDD)
- Previous Generations
- Magnetic Volumes

©Simplilearn. All rights reserved.

Solid State Drives (SSD)

SSD-backed volumes are optimized for transactional workloads with low I/O size and frequent read or write operations, where the primary performance attribute is IOPS.

©Simplilearn. All rights reserved.

General Purpose SSD

The General-purpose SSD drive is used to balance performance and price. It can be used for most of the workload. There are two types of General-purpose SSD volumes: GP2 and GP3:

Volume type	gp3	gp2
Volume size	1 GiB – 16 TiB	
Use cases	• Lower latency apps • Development and test environments	
Durability	99.8% – 99.9% durability	
Amazon EBS Multi-attach	Not supported	
Max IOPS per volume	16,000	
Max throughput per volume	1,000 MiB/s	250 MiB/s
Boot volume	Supported	

©Simplilearn. All rights reserved.

Provisioned IOPS SSD

Provisioned IOPS SSD volumes are used for high-performance, low-latency, high-throughput, and mission-critical workloads. Three types of Provisioned IOPS SSD volumes are io2 Block Express, io2, and io1. The io2 Block Express volume type is only supported with R5b instances.

Volume type	io2 Block Express	io2	io1
Volume size	4 GiB – 64 TiB	4 GiB – 16 TiB	
Use cases	• Sub-milliseconds latency • More than 64,000 IOPS or 1,000 MiB of throughput	• Workloads that require sustained IOPS performance or more than 16,000 IOPS • I/O intensive database workload	
Durability	99.999% durability	99.999% durability	99.8% – 99.9% durability
EBS Multi-attach	Supported		
Max IOPS per volume	256,000	64,000	
Max throughput per volume	4,000 MiB/s	1,000 MiB/s	
Boot volume	Supported		

©Simplilearn. All rights reserved.

Hard Disk Drives (HDD)

For heavy streaming applications that demand higher throughput, hard disk drives are ideal. Throughput is the key performance factor for SSDs.

©Simplilearn. All rights reserved.

Throughput Optimized HDD

They are low-cost HDD volumes with defined performance. They are designed for sequential workloads like big data processing and log processing. The volume size differs from 500GB to 16TB.

Volume type	st1
Volume size	125 GiB – 16 TiB
Use cases	• Big data • Data warehouses • Log processing
Durability	99.8% – 99.9% durability
Amazon EBS Multi-attach	Not supported
Max IOPS per volume	500
Max throughput per volume	500 MiB/s
Boot volume	Not supported

©Simplilearn. All rights reserved.

Get Maximum Performance

- The EBSIOBalance-percent and EBSByteBalance-percent metrics determine whether the instances are scaled efficiently or not.
- The cases with a low balance % regularly are candidates for sizing up.
- The downsizing is appropriate in cases where the balance percentage never falls below 100%.

©Simplilearn. All rights reserved.

EBS Security

©Simplilearn. All rights reserved.

EBS Encryption

An AWS KMS key is a logical representation of an encryption key.

The KMS key also includes data about the key, such as its ID, description, and creation date. EBS encrypts your volume with a data key using the industry-standard AES-256 algorithm.

©Simplilearn. All rights reserved.

EBS Encryption

It works when the snapshot is encrypted.

EBS encrypted

- Amazon EC2 sends a **GenerateDataKeyWithoutPlaintext** request to AWS KMS, specifying the KMS key to choose volume encryption.
- If the volume is encrypted using the same KMS key as the snapshot, AWS KMS uses the same data key as the snapshot and encrypts it under that same KMS key.

©Simplilearn. All rights reserved.

EBS Encryption

It works when the snapshot is unencrypted.

EBS unencrypted

- Amazon EC2 sends a **CreateGrant** request to AWS KMS.
- Amazon EC2 sends a **GenerateDataKeyWithoutPlaintext** request to AWS KMS, specifying the KMS key to choose for volume encryption.
- AWS KMS generates a new data key, encrypts it under the KMS key volume encryption, and sends the encrypted data key to Amazon EBS to be stored with the volume metadata.
- Amazon EC2 sends a decrypt request to AWS KMS to get the encryption key to encrypt the volume data.

©Simplilearn. All rights reserved.

EBS Encryption

It works when the snapshot is unencrypted.

EBS unencrypted

- When you attach the encrypted volume to an instance, Amazon EC2 sends a CreateGrant request to AWS KMS.
- When you attach the encrypted volume to an instance, Amazon EC2 sends a decrypt request to AWS KMS, specifying the encrypted data key.
- AWS KMS decrypts the encrypted data key and sends the decrypted data key to Amazon EC2.
- Amazon EC2 uses the plaintext data key in hypervisor memory to encrypt disk I/O to the volume.

©Simplilearn. All rights reserved.

Elastic Load Balancer

©Simplilearn. All rights reserved.

What Is Load Balancing?

Load balancing refers to the distribution of network traffic across multiple servers or instances of virtual machines that host the application.

©Simplilearn. All rights reserved.

Amazon Elastic Load Balancing (ELB)

ELB is a load balancing service offered by AWS.

It distributes incoming application traffic across multiple targets, such as Amazon EC2 instances, containers, and Lambda functions.

©Simplilearn. All rights reserved.

Importance of Load Balancing

Load balancing is needed to:

01	Prevent traffic overload on any server
02	Improve application responsiveness
03	Increase availability of applications
04	Avoid single point of failure in servers

©Simplilearn. All rights reserved.

Benefits of Amazon ELB

Amazon ELB offers the following benefits:

©Simplilearn. All rights reserved.

Types of Amazon Load Balancers

Amazon ELB offers the following types of load balancers:

©Simplilearn. All rights reserved.

Application Load Balancer

Functions of Application Load Balancer:

01	It is used for load balancing of HTTP and HTTPS traffic.
02	It routes traffic to targets within Amazon Virtual Private Cloud (Amazon VPC) based on the content of the request.
03	It operates on layer 7.

©Simplilearn. All rights reserved.

Network Load Balancer

Functions of Network Load Balancer:

01	It is used for load balancing of TCP, UDP, and TLS traffic.
02	It routes traffic to targets within Amazon Virtual Private Cloud (Amazon VPC) regardless of the content of the request.
03	It operates on layer 4.

©Simplilearn. All rights reserved.

Classic Load Balancer

Functions of Classic Load Balancer:

01	It provides basic load balancing across multiple Amazon EC2 instances.
02	It is best suited for applications built on the EC2-Classic network.
03	It operates on both layer 4 and layer 7.

©Simplilearn. All rights reserved.

What Is AWS Lambda?

AWS Lambda is a serverless compute service that allows users to run code without provisioning or managing servers.

It executes the code only when needed and scales automatically, from a few requests per day to thousands per second.

©Simplilearn. All rights reserved.

Terminologies in AWS Lambda

The terminologies used in context with AWS Lambda are:

01	02	03
Function	**Qualifier**	**Runtime**

A function is a resource that contains a code to process events and runtime to pass requests between Lambda and the function code.

©Simplilearn. All rights reserved.

Terminologies in AWS Lambda

The terminologies used in context with AWS Lambda are:

01	02	03
Function	**Qualifier**	**Runtime**

A qualifier is used to specify a version or an alias for a Lambda function.

©Simplilearn. All rights reserved.

Terminologies in AWS Lambda

The terminologies used in context with AWS Lambda are:

01	02	03
Function	**Qualifier**	**Runtime**

Runtime allows the function code written in different languages to run in the same base execution environment. Users can choose a runtime based on the programming language of their code.

©Simplilearn. All rights reserved.

Terminologies in AWS Lambda

The terminologies used in context with AWS Lambda are:

04	05	06
Event	Concurrency	Trigger

An event is a JSON formatted document that contains data for a function to process. It is converted to an object and passed to the function code.

©Simplilearn. All rights reserved.

Terminologies in AWS Lambda

The terminologies used in context with AWS Lambda are:

04	05	06
Event	Concurrency	Trigger

Concurrency is the number of requests that a function is serving at any given time. Users can configure their functions to limit their concurrency.

©Simplilearn. All rights reserved.

Terminologies in AWS Lambda

The terminologies used in context with AWS Lambda are:

04	05	06
Event	Concurrency	Trigger

A trigger is a resource that invokes a Lambda function. It can be an AWS service, an application, or an event.

©Simplilearn. All rights reserved.

Gateway Load Balancer

A Gateway Load Balancer operates at the third layer of the Open Systems Interconnection (OSI) model. It is used to deploy and manage a fleet of virtual appliances that support the GENEVE protocol.

Note

Users can add or remove targets from the Load Balancer without disrupting the overall flow of requests.

©Simplilearn. All rights reserved.

Gateway Load Balancer

Components of Gateway Load Balancer are:

01 It listens to all IP packets across all ports and forwards traffic to the target group that's specified in the listener rule using the GENEVE protocol on port 6081.

02 Elastic Load Balancing scales the load balancer as traffic to the application changes over time.

03 Elastic Load Balancing can scale to most of the workloads automatically.

©Simplilearn. All rights reserved.

Gateway Load Balancer: States

A Gateway Load Balancer can be in one of the following states:

01

Provisioning:
Being set up

02

Active:
Fully set up and ready to route traffic

03

Failed:
Could not be set up

©Simplilearn. All rights reserved.

HTTP 503: Service Unavailable Error

The **HTTP 503: Service unavailable error** indicates that either the load balancer or the registered instances are causing the error.

©Simplilearn. All rights reserved.

HTTP 503: Service Unavailable Error

The following are a few causes and solutions for the HTTP 503: Service unavailable error:

Cause 1:
Insufficient capacity in the load balancer to handle the request.

Solution 1:
This should be a transient issue and should not last more than a few minutes. If it persists, go to the AWS Support Center for assistance.

©Simplilearn. All rights reserved.

HTTP 503: Service Unavailable Error

The following are a few causes and solutions for the HTTP 503: Service unavailable error:

Cause 2:
There are no registered instances.

Solution 2:
A user needs to register at least one instance in every Availability Zone, where the load balancer is configured to respond. Verify this by looking at the **HealthyHostCount** metrics in CloudWatch. If a user is unable to ensure that an instance is registered in each Availability Zone, enabling cross-zone load balancing is recommended.

©Simplilearn. All rights reserved.

HTTP 503: Service Unavailable Error

The following are a few causes and solutions for the HTTP 503: Service unavailable error:

Cause 3:
There are no healthy instances.

Solution 3:
A user needs to ensure that they have healthy instances in every Availability Zone, where their load balancer is configured to respond. This can be verified by looking at the **HealthyHostCount** metric.

©Simplilearn. All rights reserved.

HTTP 503: Service Unavailable Error

The following are a few causes and solutions for the HTTP 503: Service unavailable error:

Cause 4:
The surge queue is full.

Solution 4:
A user needs to ensure that their instances have sufficient capacity to handle the request rate. This can be verified by looking at the **SpillOverCount** metric.

©Simplilearn. All rights reserved.

Setting Up a Web Server on an EC2 Instance

Duration:10 min

Problem Statement:

You have been assigned a task to demonstrate the process of setting up a web server on AWS, including creating a VPC, configuring essential components, and launching an EC2 instance with a basic web server.

ASSISTED PRACTICE

©Simplilearn. All rights reserved.

Assisted Practice: Guidelines

Steps to be followed are:

1. Create a VPC, a subnet, and an internet gateway
2. Create Route tables
3. Create an EC2 web server instance

©Simplilearn. All rights reserved.

Using a Classic Load Balancer to Distribute Traffic

Duration:10 min

Problem Statement:

You have been assigned a task to demonstrate the process of creating a classic Load Balancer in AWS EC2 and deploying it to multiple instances in different availability zones.

ASSISTED PRACTICE

©Simplilearn. All rights reserved.

Assisted Practice: Guidelines

Steps to be followed are:

1. Create a security group
2. Launch instances with different availability zones
3. Create the Classic Load Balancer
4. Deploy the Classic Load Balancer to an EC2 instance

©Simplilearn. All rights reserved.

Configuring an Application Load Balancer

Duration:10 min

Problem Statement:

You have been assigned a task to demonstrate the process of setting up and testing an Application Load Balancer in AWS.

ASSISTED PRACTICE

©Simplilearn. All rights reserved.

Assisted Practice: Guidelines

Steps to be followed are:

1. Create a target group
2. Launch EC2 instances
3. Configure the target group
4. Create a Load Balancer
5. Test the Load Balancer

©Simplilearn. All rights reserved.

Creating Route Requests in ALB

Duration:10 min

Problem Statement:

You have been assigned a task to set up routing requests in Amazon Web Services (AWS) using an Application Load Balancer (ALB).

ASSISTED PRACTICE

©Simplilearn. All rights reserved.

Assisted Practice: Guidelines

Steps to be followed are:

1. Set up the prerequisites for EC2
2. Create a routing request in ALB

©Simplilearn. All rights reserved.

Creating a Network Load Balancer

Duration:10 min

Problem Statement:

You have been assigned a task to create a Network Load Balancer (NLB) in the Amazon Web Services (AWS) environment.

ASSISTED PRACTICE

©Simplilearn. All rights reserved.

Assisted Practice: Guidelines

Steps to be followed are:

1. Set up the Network Load Balancer

©Simplilearn. All rights reserved.

ELB Logs

©Simplilearn. All rights reserved.

Load Balancer Access Logs

Following are the classifications of access logs for different Load Balancers:

- Application Load Balancer
- Network Load Balancer
- Classic Load Balancer

- Access logging is an optional feature of Elastic Load Balancing that is disabled by default.
- Each log contains information such as the time the request was received, the client's IP address, latencies, request paths, and server responses.
- These access logs can be used to analyze traffic patterns and troubleshoot issues.

©Simplilearn. All rights reserved.

Load Balancer Access Logs

Following are the classifications of access logs for different Load Balancers:

- Application Load Balancer
- Network Load Balancer
- Classic Load Balancer

- Access logging is an optional feature of Elastic Load Balancing that is disabled by default.
- It captures the logs and stores them in the Amazon S3 bucket that the users specify as compressed files once the access logging is enabled.
- Access logging can be disabled at any time.

©Simplilearn. All rights reserved.

Load Balancer Access Logs

Following are the classifications of access logs for different Load Balancers:

- Application Load Balancer
- Network Load Balancer
- Classic Load Balancer

- There is no additional charge for access logs.
- The users are charged storage costs for Amazon S3, but not charged for the bandwidth used by Elastic Load Balancing to send log files to Amazon S3.

©Simplilearn. All rights reserved.

Access Log Files

Elastic Load Balancing publishes a log file for each Load Balancer node every 5 minutes. The file names of the access logs use the following format:

- bucket: The name of the S3 bucket
- prefix: The prefix (logical hierarchy) in the bucket
- aws-account-id: The AWS account ID of the owner
- region: The Region of a Load Balancer and S3 bucket
- yyyy/mm/dd: The date that the log was delivered

©Simplilearn. All rights reserved.

Access Log Files

Elastic Load Balancing publishes a log file for each Load Balancer node every 5 minutes. The file names of the access logs use the following format:

- load-balancer-id: The resource ID of the Load Balancer
- end-time: The date and time that the logging interval ended
- ip-address: The IP address of the Load Balancer node that handled the request
- random-string: A system-generated random string

©Simplilearn. All rights reserved.

Auto Scaling

©Simplilearn. All rights reserved.

Auto Scaling

Amazon EC2 Auto Scaling helps the users to maintain application availability.

According to the conditions defined by the users, EC2 Auto Scaling allows them to automatically add or remove EC2 instances.

©Simplilearn. All rights reserved.

Auto Scaling Groups

A collection of EC2 instances is called an Auto Scaling group.

The users can specify the minimum number of instances in each group, and Auto Scaling ensures that the group never goes below the minimum size.

©Simplilearn. All rights reserved.

Auto Scaling

The difference between the load balancer and Auto Scaling are:

Load Balancer	Auto Scaling
Elastic load balancer only monitors each instance's health, distributes traffic, and connects each request to the right target groups.	The user's ability to specify policies based on predefined criteria is enabled by auto scaling, which enables multiple instances to work parallely.

©Simplilearn. All rights reserved.

Auto Scaling Features

Built-in scaling strategies

Automatic resource discovery

Predictive scaling

Unified scaling

Smart scaling policies

©Simplilearn. All rights reserved.

Auto Scaling Planning

Here are the four categories of auto scaling planning:

©Simplilearn. All rights reserved.

Auto Scaling Benefits

- Better fault tolerance
- Increased application availability
- Lower costs

Amazon EC2 Auto Scaling can determine the health of an instance. It can terminate the instance and replace it with a new one.

©Simplilearn. All rights reserved.

Auto Scaling Benefits

Better fault tolerance

Increased application availability

Lower costs

Amazon EC2 Auto Scaling ensures that the application always has the right amount of computing and proactively provisions capacity with predictive scaling.

©Simplilearn. All rights reserved.

Auto Scaling Benefits

Better fault tolerance

Increased application availability

Lower costs

Amazon EC2 Auto Scaling adds instances only when needed, and it can scale across purchase options to optimize performance and cost.

©Simplilearn. All rights reserved.

Fleet Management with Auto Scaling

If the user's application runs on Amazon EC2 instances, the user has what is known as a fleet.

- The users can automate the management of a user's fleet
- It is easy to set up.

©Simplilearn. All rights reserved.

Functions of Auto Scaling

The three main functions that Auto Scaling performs to automate fleet management for EC2 instances are as follows:

Monitoring the health of running instances

Replacing impaired instances automatically

Balancing capacity across Availability Zones

©Simplilearn. All rights reserved.

Monitoring the Health of Running Instances

- Auto Scaling monitors the health of all instances that are placed within an **Auto Scaling group** at regular intervals.
- If the instance is connected to an **Elastic Load Balancing** load balancer, it can also perform ELB health checks.

©Simplilearn. All rights reserved.

Replacing Impaired Instances Automatically

- When an impaired instance fails a health check, Auto Scaling terminates it and replaces it with a new one automatically.

- If the user is using an Elastic Load Balancing load balancer, Auto Scaling detaches the impaired instance from the load balancer before provisioning a new one and attaches it back to the load balancer.

©Simplilearn. All rights reserved.

Balancing Capacity Across Availability Zones

Balancing resources across **Availability Zones** is the best practice for well-architected applications, as this greatly increases aggregate system availability.

Auto Scaling automatically balances EC2 instances across zones when one configures multiple zones in the Auto Scaling group settings.

©Simplilearn. All rights reserved.

Configuring the Manual and Dynamic Scaling

Duration:15 min

Problem Statement:

You have been assigned a task to understand how to configure manual and dynamic scaling for your application using Amazon Web Services (AWS) tools and services.

ASSISTED PRACTICE

©Simplilearn. All rights reserved.

Assisted Practice: Guidelines

Steps to be followed are:

1. Set up predefined auto scaling group
2. Set up EC2 Auto Scaling with Load Balancer

©Simplilearn. All rights reserved.

Setting Up Auto Scaling Using Launch Templates

Duration:15 min

Problem Statement:

You have been assigned a task to create an Auto Scaling group using a launch template.

ASSISTED PRACTICE

©Simplilearn. All rights reserved.

Assisted Practice: Guidelines

Steps to be followed are:

1. Create a launch template in EC2
2. Create a launch configuration
3. Create an Auto Scaling group

©Simplilearn. All rights reserved.

Key Takeaways

- Amazon EC2 provides a highly reliable environment where replacement instances can be quickly and consistently deployed.
- The users can launch multiple instances from a single AMI when they require multiple instances with the same configuration.
- AWS Lambda executes the code only when needed and scales automatically, from a few requests per day to thousands per second.
- Auto Scaling performs EC2 health checks at regular intervals, and if the instance is connected to an Elastic Load Balancer, it can also perform ELB health checks.

©Simplilearn. All rights reserved.

Scaling of EBS Volume for a Linux VM

Duration: 30 mins

Project agenda: Perform vertical scaling of EBS volume for a Linux VM

Description: Your company is experiencing business growth where solution deployment is happening with limited resources. In this case, the vertical scalability feature of AWS can be used to create a cost-optimised architecture.

Perform the following:

1. Create an EC2 Instance
2. Identify the EBS volume
3. Create a snapshot
4. Create a new volume
5. Detach the existing volume from the EC2 instance
6. Attach a new volume to the EC2 instance

LESSON-END PROJECT

©Simplilearn. All rights reserved.

TECHNOLOGY

Thank You

©Simplilearn. All rights reserved.

simplilearn