AWS SysOps Administrator – Associate Level

Deployment and Provisioning

```
#include <iostream>
#include <cstdlib>
using namespace std;
int main()
{
    cout << "Hello, world!" << endl;
    return 0;
}
```

Learning Objectives

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

- Select appropriate load balancers for your application
- Assign EBS volumes to instances
- Deploy **ALBs** with EC2 instances for traffic management

EC2 Instance Lab

Duration: 10 Min.

Problem Statement:

Create an AWS EC2 instance assigning it to an IAM role.

ASSISTED PRACTICE

Assisted Practice: Guidelines

Steps to create an AWS EC2 instance:

1. Log in to your AWS lab
2. Click on **IAM Roles** in **Services**
3. Create an IAM role
4. Select EC2 from **Services**
5. Select Ubuntu machine
6. Assign your IAM role to the instance
7. Launch the instance

TECHNOLOGY

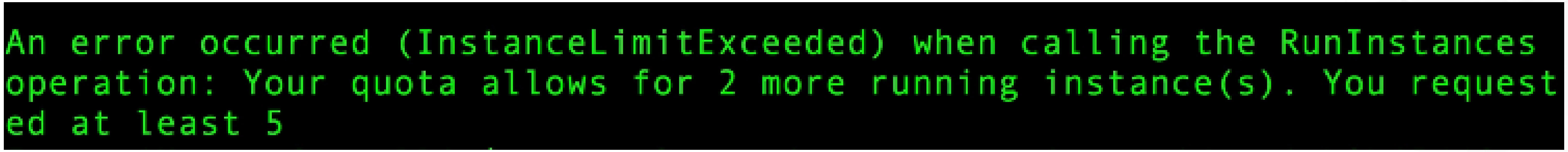
EC2, ELB, and IOPS

EC2 Launch Issues

Two major issues that can occur while launching or creating an EC2 instance are given below:

● **InstanceLimitExceeded** error:

- This error occurs when you have reached the limit of the number of instances that you are allowed to launch within a region.
- By default, this limit is set to 20 by AWS.



```
An error occurred (InstanceLimitExceeded) when calling the RunInstances
operation: Your quota allows for 2 more running instance(s). You request
ed at least 5
```

EC2 Launch Issues

InsufficientInstanceCapacity error:

- This error means that AWS is out of the number of a type of instances (on-demand) that you have requested to launch.
- However, this is a rare issue that can occur and can be solved by requesting less instances, selecting another type of instance, changing zones, and purchasing reserved instances.
- Example: This error can occur if you request more than twenty **t2.micro** instances at once.

EBS Volumes and IOPS

Elastic Block Store (EBS) is a storage volume that can be attached with an EC2 instance.

- These volumes appear similar to disk space on the instances.
- These volumes can be used to create file systems and databases, run operating systems, and perform other functions.
- SSD-backed storage is a type of EBS volume used quite often.
- SSD can be used to run operating systems and databases which are majorly I/O-intensive tasks.

EBS Volumes and IOPS

gp2 and **io1** are two types of EBS SSD volumes.

- IOPS stands for Input/Output Operations Per Second and is used to provide standard values to the performance capacity of the volume.
- **gp2** stands for General Purpose which is mostly used as boot volumes.
- **io1** is the Provisioned IOPS used for I/O-intensive tasks, databases, and latency-sensitive workloads.
- IOPS capability depends on the size of the volume:
    - **gp2** volumes: 3 IOPS/GB up to 16k IOPS
    - **io1** volumes: 50 IOPS/GB up to 64k IOPS

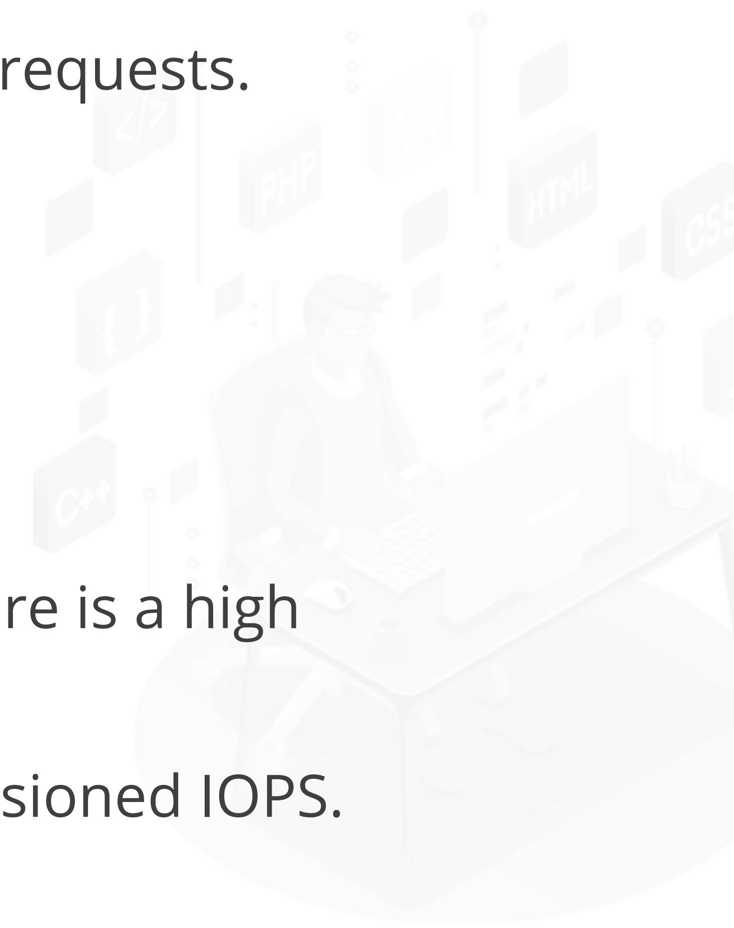
EBS Volumes and IOPS

IOPS issues:

- There are cases when a user might reach the IOPS limit or exceed the number of requests.
- If the limit is reached, the user starts getting requests queuing.
- The application becomes slow depending on its sensitivity to IOPS.

Solutions:

1. User can increase the size of the volume. However, if it is already 5.2TB, then there is a high possibility that it has reached the 16k IOPS limit.
2. If more than 16k IOPS is needed, it is advised to change the volume type to Provisioned IOPS.



Elastic Load Balancers

A load balancer helps in distributing requests on multiple servers or instances for efficient working of the application and mitigating response delays. The types of load balancers are given below:

Application load balancer

- It works at the application layer of the OSI model.

Network load balancer

- It works at the transport layer of the OSI model.

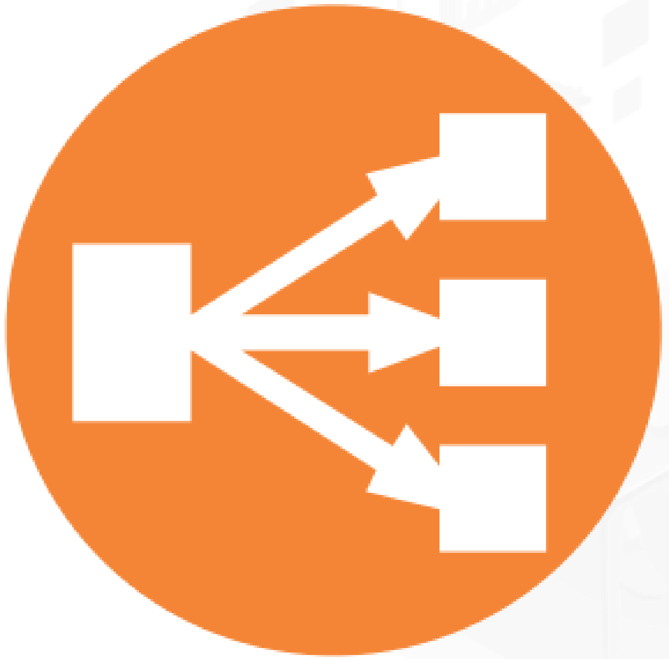
Classic load balancer

- This is a legacy load balancer.

Application Load Balancer

It is best suited for the application layer for load balancing HTTP and HTTPS traffic.

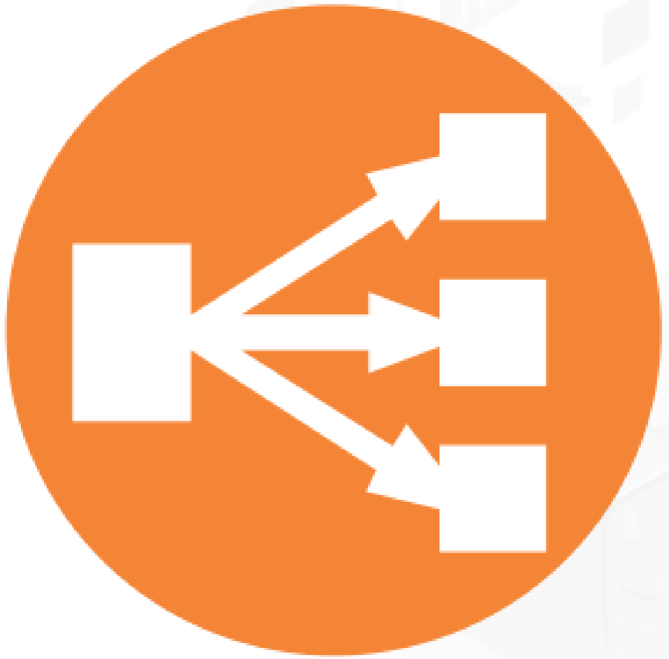
- It can be used for advanced routing and sending requests to determined servers.
- It can identify the required request for the determined server and can route the requests to the right servers if the sent request is not for the specified server.
- It performs all routing tasks using HTTP packets and headers.



Network Load Balancer

It is best suited for load balancing TCP traffic where high performance is required.

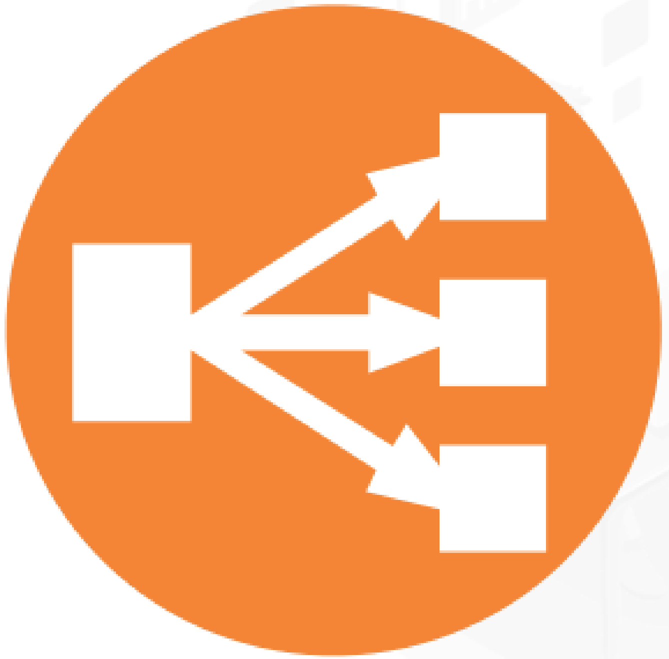
- It is capable of handling millions of requests per second.
- It maintains the lowest latency compared to all other load balancers.
- It is majorly used for production servers where low latency is of utmost importance.
- It is the most expensive load balancer.



Classic Load Balancer

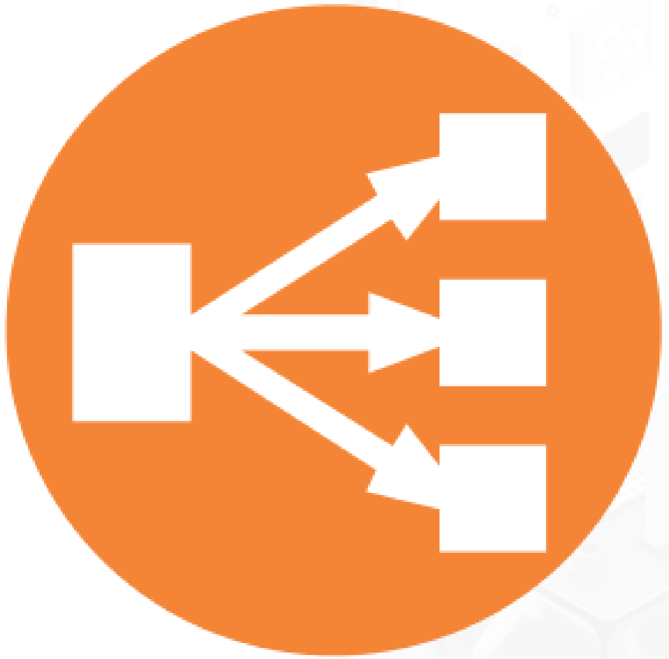
It is a legacy load balancer and can be used on both application and transport layers.

- It provides basic features on layer 7 like X-forwarded and sticky sessions.
- It is rarely used and is not recommended for modern applications.
- It can be strictly used at layer 4 for an application that relies purely on TCP protocol.



Pre-Warming a Load Balancer

- Application load balancers scale automatically to adapt to your workload.
- This changes the IP addresses that the client connects to, as new ALBs are brought into service.
- A network load balancer creates a static IP address in each subnet.
- This keeps the firewall rules simple, as the client only needs to enable a single IP address for each subnet.
- This is done using AWS elastic IP addresses.
- Moreover, keeping an ALB behind an NLB reduces the task of choosing one or other LBs and gives the benefits of both the load balancers.



Load Balancer and Static IP

Pre-warming process is used to make a load balancer scale up if the traffic suddenly increases on the application.

- Example: An e-commerce company’s marketing team plans to announce a sale on a public holiday and estimates that there will be five times more traffic on the website.
- To avoid any downtime in this sudden increase in the number of requests, AWS can pre-warm the ELB and configure it to the appropriate level of capacity required to handle requests.

- AWS needs to know the following data to pre-warm the load balancers:

1. Start and end dates of the high-performance capacity
2. Expected request rate per second
3. Size of a typical request

ELB Error Messages

ELB Error Messages

4XX and 5XX are the major error code that can occur in ELB operations.

- Any unsuccessful request generates 4XX and 5XX errors.
- 4XX error message indicates that there is an error on the client side.
- 5XX error message indicates the issue on the server side.

ELB Error Messages

- 400 indicates that it is a malformed request such as an incorrect header and is not per HTTP and HTTPS standards.
- 401 indicates that the user doesn’t have access to the webpage.
- 403 indicates that the request is forbidden and the url is blocked.
- 460 indicates the client’s timeout period is short and the load balancer doesn’t have time to respond.
- 463 indicates that the load balancer has received an X-Forwarded-For request header with more than thirty IP addresses.

ELB Error Messages

- 500 indicates an internal server error such as a configuration issue with ELB.

- 502 indicates bad gateway in cases when an application server has closed the connection or sent a malformed response.

- 503 indicates that the service is unavailable which means there are no registered target or web servers.

- 504 indicates gateway timeout which means the application is not responding due issues with web servers or databases.

- 561 indicates that the load balancer is not getting a response from the ID provider to authenticate a user.

ELB Cloudwatch Metrics

ELB publishes metrics to Cloudwatch for the load balancer and also for the backend instances.

- The metrics help to verify a system's performance.
- Metrics are gathered in an interval of sixty seconds.
- User can also create a Cloudwatch alarm for a specific action.
- Example: User can create a Cloudwatch alarm to send an email if the metrics reach the limit.

ELB Cloudwatch Metrics

Cloudwatch metrics can be categorized based on the operations:

Overall Health and **Performance Metrics**

Overall Health:

- It checks the overall performance and status of the system.
- It includes issues like:

1. **BackendConnectionError:** Number of unsuccessful backend connections to instances

1. **HealthyHostCount:** Number of healthy registered instances

1. **UnhealthyHostCount:** Number of unhealthy host count with issues in services

1. **HTTPCode_Backend_2XX_4XX_5XX**

ELB Cloudwatch Metrics

Performance Metrics:

They deal with checks like:

1. **Latency:** Number of seconds taken for a registered instance to respond or connect

1. **RequestCount:** Number of requests completed or connections made during a specified interval

1. **SurgeQueueLength:**
- Number of pending requests
- It’s for classic load balancers and has the maximum queue size of 1024
- Any additional request is rejected

1. **SpilloverCount:**
- Number of requests rejected
- This metric is for classic load balancers

Application Load Balancer

Duration: 10 Min.

Problem Statement:

You are given a project to deploy an application load balancer.

Assisted Practice: Guidelines

Steps to deploy an application load balancer:

1. Select a load balancer
2. Select a security group
3. Configure targets
4. Select an instance

TECHNOLOGY

Systems Manager

Systems Manager

AWS Systems Manager is a tool that provides visibility and control of the entire AWS infrastructure to the user.

- It integrates with Cloudwatch which allows user to view the dashboard, operational date, or reporting bugs.
- It also includes Run command to automate operational tasks such as security patching.
- It also organizes the inventory by grouping resources by application or environment.

Run Command

Run command allows the user to run predefined commands on one or more EC2 instances.

Some of the basic tasks that can be executed using the Run command:

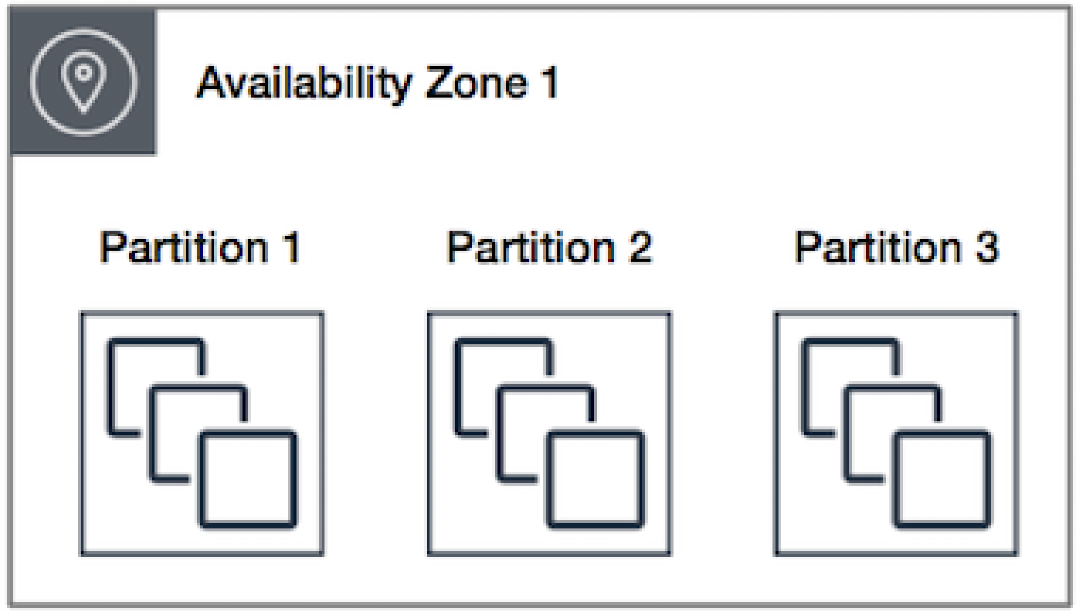
1. Stop, restart, terminate, and resize instances
2. Attach or detach an instance
3. Create snapshots
4. DynamoDB backup
5. Apply updates and system patches
6. Run scripts

Placement Groups

Placement groups help users control how the instances are deployed.

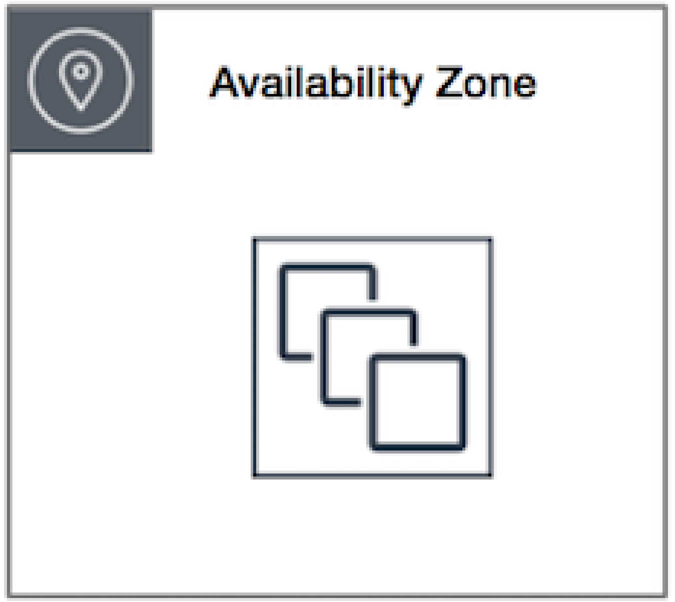
- Placement groups help in getting low latency, high network throughput, and high computing power.
- There are three types of placement groups:

1. Cluster: All instances are created in one availability zone
2. Partition: Instances are created in segments called partitions with each present in a different rack with separate power and network resources
3. Spread: Every instance has a different rack and an independent power and network setup



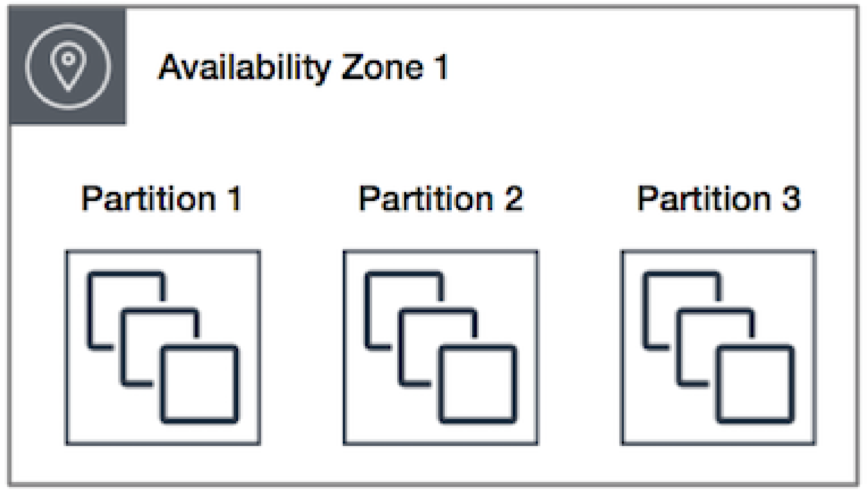
Cluster Placement Group

- A cluster placement group is a grouping of instances in one availability zone.
- User can span peered VPCs in the same region.
- Instances have a throughput limit of 10 Gbps for TCP/IP traffic.
- Instances are placed in the same high bisection bandwidth segment of the network.
- Cluster placement groups are recommended for applications that benefit from low network latency, high network throughput, or both.
- It is recommended to have a single launch request for all instances and also to keep same instance types in one cluster.
- There are, however, chances of reaching the instance limit when trying to add more instances.



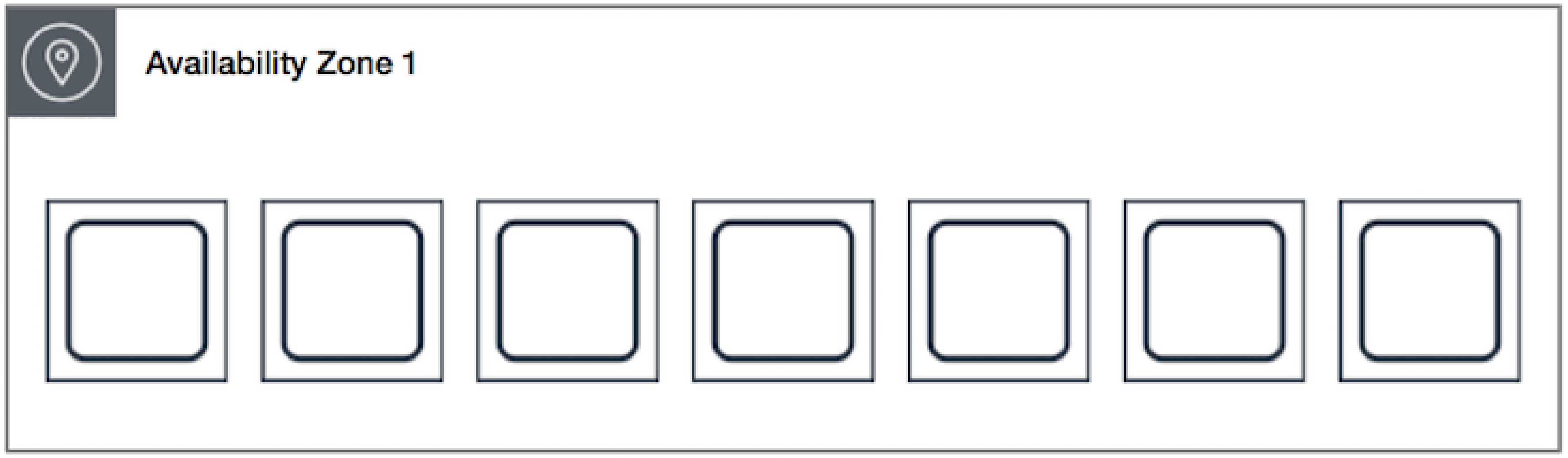
Partition Placement Group

- In partition placement groups, Amazon EC2 divides each group into segments called partitions.
- Each partition within a placement group has a separate set of racks.
- Each rack has a separate network and power source.
- It allows user to isolate and mitigate the impact of hardware failure.
- It can be used to deploy large workloads like HDFS and Cassandra across distinct racks.
- By default, AWS distributes instances across partitions. However, one can also decide where the instances should be launched.
- If there is insufficient unique hardware to fulfill the request during instance startup, the request fails.



Spread Placement Group

- A spread placement group is a one-instance-per-rack arrangement with distinct power and network sources for each instance.
- It is used for applications having a small number of critical instances that should be kept separate from each other.
- Launching instances reduces the risk of simultaneous failures.
- A spread placement group can span multiple availability zones in the same region with a maximum of seven running instances per availability zone per group.



Key Takeaways

- Elastic Block Store (EBS) is a storage volume that can be attached with an EC2 instance.
- IOPS capability depends on the size of the volume.
- A load balancer helps in distributing requests on multiple servers or instances.
- **Overall Health** and **Performance Metrics** are the two Cloudwatch metrics.
- Systems Manager organizes an inventory by grouping resources by application or environment.

Applying a Load Balancer

Problem Statement:
Create and apply load balancer on EC2 instances.

Background of the problem statement:
As a cloud architect, you are responsible for designing, installing, and maintaining the DevOps infrastructure in your organization. As the festive season draws near, there is a high possibility of an increase in traffic on the website. Hence, you are required to set up a load balancer that navigates requests to the determined servers based on the request header.

LESSON-END PROJECT