Interview questions:

Containerisation At Scale

Here are some theoretical questions that are frequently asked in a DevOps interview. We have provided general guidelines for answering every question. Keeping these guidelines in mind, you can deliver your answers in a much better way in an actual interview.

Q1. What is a Region and an Availability Zone (AZ) in AWS?

Guidelines:

Region

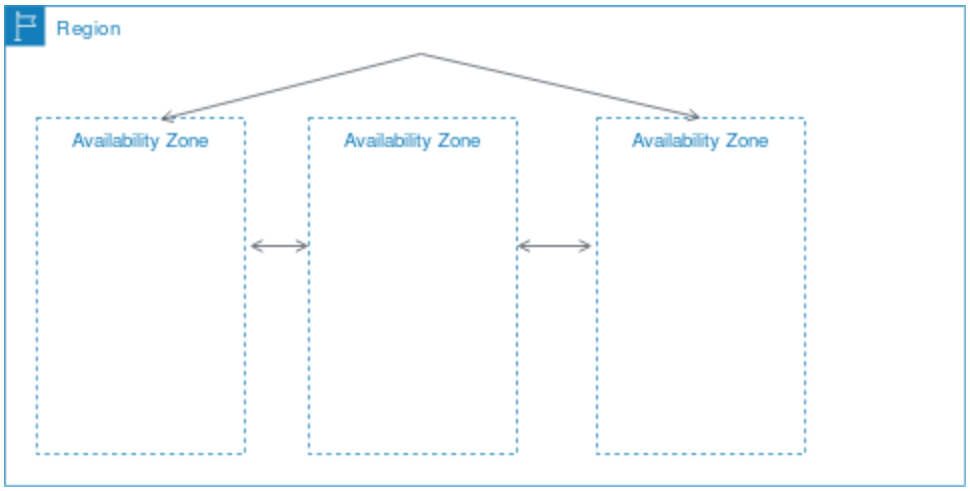
AWS has the concept of a Region, which is a physical location around the world where they cluster data centers. We call each group of logical data centers an Availability Zone. Each AWS Region consists of multiple, isolated, and physically separate AZ's within a geographic area.

Availability Zones

An Availability Zone (AZ) is one or more discrete data centres with redundant power, networking, and connectivity in an AWS Region. All AZs in an AWS Region are interconnected with redundant fibre cables providing high-bandwidth and low-latency between AZs If an application is partitioned across AZs, companies are better isolated and protected from issues such as power outages or any natural

calamity such as lightning strikes, earthquakes, and more. AZs are physically separated by a meaningful distance, many kilometers, from any other AZ.

Multiple Availability Zones in an AWS Region:



AWS Global Infrastructure Map

AWS has 80 Availability Zones across 25 geographic regions

Q2. What is VPC ? List down its major components.

Guidelines:

A VPC is a logically isolated virtual network, spanning an entire AWS Region, where your EC2 instances are launched. A VPC is primarily responsible for isolating your AWS resources from other accounts , routing network traffic to and from your instances and protecting your instances from network intrusion

Major components of VPC that will be created by a user or by AWS as part of a default VPC:

1. VPC CIDR Block
2. Subnet
3. Gateways
4. Route Table
5. Network Access Control Lists (ACLs)
6. Security Group

Q3. What is an IAM ? What is the IAM role ?

Guidelines:

1. AWS Identity and Access Management (IAM) enables you to manage access to AWS services and resources securely. Using IAM, you can create and manage AWS users and groups, and use permissions to allow and deny their access to AWS resources.

2. An IAM role allows a user or resource of AWS to take any action on another resource of AWS. You assign IAM roles to AWS resources or a user. For example, you can create an IAM role with the “AmazonS3fullaccess” policy and attach this role to any EC2 instance. Now, EC2 instances can access s3 buckets (Create, read and write). Policies are in JSON format.

Q4. What is ECR ?

Guidelines:

Amazon Elastic Container Registry (ECR) is a fully managed container registry that makes it easy to store, manage, share, and deploy your container images and artifacts anywhere.
The most crucial aspect of ECR is that AWS IAM handles authentication and authorization for the container registry. Therefore, it is easy to access ECR from all the different services AWS provides (ECS, EKS, CodeBuild, and many more). AWS IAM is not easy to use but allows you to define strict access control to your container registry.

Q5. Explain ECS and its different modes?

Guidelines:

Amazon Elastic Container Service (Amazon ECS) is a fully managed container orchestration service. Basically ECS manages deployment, update, networking, scaling and load balancing of your containers.

ECS provides two modes :

EC2 based
AWS ECS in EC2 mode is just a logical grouping (cluster) of EC2 instances, and all the EC2 instances that are part of an ECS act as Docker host on which containers are launched. With EC2 mode you can choose your own instance types and optimise the billing by using right amount of spot , reserved and on-demand instances.

Fargate Based (Serverless)
Fargate allocates the right amount of compute, eliminating the need to choose instances and scale cluster capacity. You only pay for the resources required to run your containers, so there is no over-provisioning and paying for additional servers. Fargate runs each task or pod in its own kernel providing the tasks and pods their own isolated compute environment. This enables your application to have workload isolation and improved security by design.

Q6. What is an S3 bucket ?

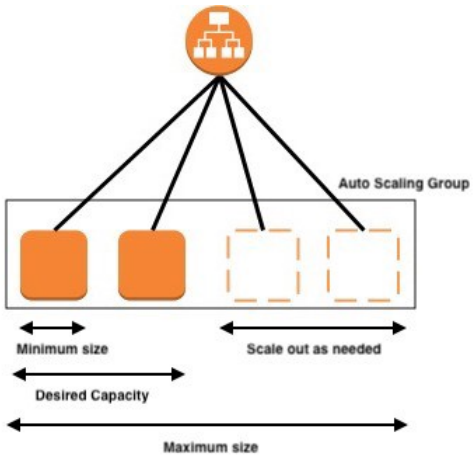
Guidelines:

S3 is an object storage service which is fast, scalable and durable. S3 enables customers to upload, download or store any file or object that is up to 5 TB in size. 3 stands for 3 S, Simple Storage Service.

Q7. What is an auto-scaling group in aws?

Guidelines:

An Auto Scaling group contains a collection of Amazon EC2 instances that are treated as a logical grouping for the purposes of automatic scaling and management. An Auto Scaling group also enables you to use Amazon EC2 Auto Scaling features such as health check replacements and scaling policies. Both maintaining the number of instances in an Auto Scaling group and automatic scaling are the core functionality of the Amazon EC2 Auto Scaling service.



Q8. What are task definitions and service definitions in ECS ?

Guidelines:

A **task definition** is required to run Docker containers in Amazon ECS. The following are some of the parameters you can specify in a task definition:

- The Docker image to use with each container in your task
- How much CPU and memory to use with each task or each container within a task
- The launch type to use, which determines the infrastructure on which your tasks are hosted
- The Docker networking mode to use for the containers in your task
- The logging configuration to use for your tasks
- Whether the task should continue to run if the container finishes or fails
- The command the container should run when it is started
- Any data volumes that should be used with the containers in the task
- The IAM role that your tasks should use

You can define multiple containers in a task definition. The parameters that you use depend on the launch type you choose for the task. Not all parameters are valid.

An **Amazon ECS service** enables you to run and maintain a specified number of instances of a task definition simultaneously in an Amazon ECS cluster. If any of your tasks should fail or stop for any reason, the Amazon ECS service scheduler launches another instance of your task definition to replace it in order to maintain the desired number of tasks in the service.

In addition to maintaining the desired number of tasks in your service, you can optionally run your service behind a load balancer. The load balancer distributes traffic across the tasks that are associated with the service.

Q9. What does the term service discovery mean?

Guidelines:

Service Discovery has the ability to locate a network automatically making it so that there is no need for a long configuration set up process. Service discovery works by devices connecting through a common language on the network allowing devices or services to connect without any manual intervention. (i.e Kubernetes service discovery, AWS service discovery)

There are two types of service discovery: Server-side and Client-side. Server-side service discovery allows clients applications to find services through a router or a load balancer. Client-side service discovery allows clients applications to find services by looking through or querying a service registry, in which service instances and endpoints are all within the service registry.

Amazon ECS creates and manages a registry of service names using the Route 53 Auto Naming API. Names are automatically mapped to a set of DNS records so you can refer to services by an alias, and have this alias automatically resolve to the service's endpoint at runtime. You can specify health check conditions in a service's task definition and Amazon ECS will ensure that only healthy service endpoints are returned by a service lookup.

Refer : What is service discovery ?

Q10. How to setup dynamic port mapping in ECS cluster in EC2 mode

Guidelines:

In the task definition, use bridge mode of networking. Expose required port to the host port 0. Exposing to host port 0 dynamically allocates random ephemeral ports through which communication could get established. Its is crucial to understand that there ports and containers get automatically registed on the application load balancer.

Q11. Explain different placement strategies available on ECS ?

Guidelines:

A *task* ***placement strategy*** is an algorithm for selecting instances for task placement or tasks for termination. Task placement strategies can be specified when either running a task or creating a new service. The task placement strategies can be updated for existing services as well.

Amazon ECS supports the following task placement strategies:

1. Binpack :

Tasks are placed on container instances so as to leave the least amount of unused CPU or memory. This strategy minimizes the number of container instances in use. When this strategy is used and a scale-in action is taken, Amazon ECS will terminate tasks based on the amount of resources that will be left on the container instance after the task is terminated. The container instance that will have the most available resources left after task termination will have that task terminated.

2. random

Tasks are placed randomly.

3. spread

Tasks are placed evenly based on the specified value. Accepted values are instanceId (or host, which has the same effect), or any platform or custom attribute that is applied to a container instance, such as

attribute:ecs.availability-zone. Service tasks are spread based on the tasks from that service. Standalone tasks are spread based on the tasks from the same task group. When this strategy is used and a scale-in action is taken, Amazon ECS will select tasks to terminate that maintains a balance across Availability Zones. Within an Availability Zone, tasks will be selected at random.

Q12. What is the worker process and worker connections in NGINX ?

Guidelines:

NGINX can run multiple worker processes, each capable of processing a large number of simultaneous connections. You can control the number of worker processes and how they handle connections with the following directives:

worker_processes – The number of NGINX worker processes (the default is 1). In most cases, running one worker process per CPU core works well, and we recommend setting this directive to auto to achieve that. There are times when you may want to increase this number, such as when the worker processes have to do a lot of disk I/O.

worker_connections – The maximum number of connections that each worker process can handle simultaneously. The default is 512, but most systems have enough resources to support a larger number. The appropriate setting depends on the size of the server and the nature of the traffic, and can be discovered through testing.

Example Configuration

```
user www www;
worker_processes 2;

error_log /var/log/nginx-error.log info;

events {
    use kqueue;
    worker_connections 2048;
}

...
```

File : /etc/nginx/nginx.conf

Q13. What are different deployment strategies ?

Guidelines: DIfferent deployment strategies.

Q14. How can you expose your application internally and externally on AWS ? Can we use ALB for both purposes?

Guidelines:

Using service discovery we can expose our services internally. Service discovery gives a namespace (private DNS) to service and within VPC the application can be access through that namespace

Yes, load balancers can be used for exposing externally as well as internally.
Internet facing ALBs are used to expose applications to the internet.
We can set up a private load balancer to expose our application internally.

Q15. How can you add a standalone EC2 instance in ECS cluster ?

Guidelines:

Choose an EC2 instance with ECS-Optimised AMI.
Attach IAM role “AmazonEC2COntainerServiceforEC2Role”
Run

```
#!/bin/bash
echo ECS_CLUSTER={cluster_name} >> /etc/ecs/ecs.config
```

Reference:

1. NGINX vs Apache2

2. Tuning Your NGINX Configuration
3. How nginx processes a request
4. How to optimize NGINX performance
5. DIfferent deployment strategies.