ECS Cluster & Infrastructure

1 ECS Cluster

1. An ECS Cluster represents a group of Docker hosts acting as a unified environment for container orchestration
2. It provides the resource pool where ECS can schedule and run containerized applications
3. Creating a cluster is required because ECS needs a centralized control point to manage task placement, resource isolation, and service scaling
4. You can run multiple services and tasks inside an ECS Cluster
5. ECS supports two types of cluster infrastructure for running containers
    1. Amazon EC2
    2. AWS Fargate

2 Infrastructure

1. Amazon EC2 (Self-Managed Infrastructure)

1	Infrastructure Provisioning	You launch and manage EC2 instances
2	Control Level	Full control over OS, instance type, storage, AMIs
3	Launch Time	Slower (wait for EC2 boot and registration)
4	Billing Model	Pay for EC2 uptime; supports On-Demand, Spot, and Savings Plans
5	VPC Selection Timing	Selected during cluster creation
6	Auto Scaling Configuration	ASG is created during cluster creation, but you manage instance scaling policies
7	Scaling Flexibility	Based on EC2 limits — scale EC2 first, then tasks
8	Maintenance Responsibility	You manage patching, scaling, instance health
9	Task Placement	ECS places tasks on EC2 instances based on available resources
10	Container Isolation	Multiple containers share same EC2 instance resources
11	Use Case Fit	Best for apps needing OS-level access or GPU support
12	Suitable For	Long-running apps, OS customization, GPU/SSD needs

2. AWS Fargate (Fully Managed Serverless)

1	Infrastructure Provisioning	AWS provisions infrastructure automatically
2	Control Level	No access to underlying OS or compute layer
3	Launch Time	Faster (AWS handles compute provisioning instantly)
4	Billing Model	Pay per vCPU and memory usage per second for each running task
5	VPC Selection Timing	Selected when creating a task or service
6	Auto Scaling Configuration	ECS handles scaling of Fargate tasks automatically
7	Scaling Flexibility	Directly scales tasks (no infrastructure scaling needed)
8	Maintenance Responsibility	AWS manages all underlying infrastructure
9	Task Placement	ECS directly launches tasks on Fargate compute
10	Container Isolation	Each task runs in isolated environment with its own ENI
11	Use Case Fit	Best for serverless microservices, batch jobs, quick deployments
12	Suitable For	Microservices, fast-scaling APIs, short-lived workloads

3 ECS Anywhere

1. ECS Anywhere is a feature of Amazon ECS that allows you to run ECS tasks on external (non-AWS) machines
2. ECS Anywhere supports the following types of external infrastructure
    1. On-premises servers (physical or virtual machines)
    2. Virtual machines running in other cloud providers (e.g., Azure, GCP)
3. Note: These external machines are not added during cluster creation. Instead, you use a registration command after the cluster is created to link them to ECS using a registration token provided by AWS

4 Cluster-level encryption

1. During ECS cluster creation, you can enable storage encryption using AWS KMS
2. There are two options
    1. Managed Storage Encryption — Encrypts storage volumes across all launch types
    2. Fargate Ephemeral Storage Encryption — Encrypts temporary runtime storage for Fargate tasks
3. Note
    1. Even if encryption is set to false, EBS volumes will still be encrypted if EBS encryption by default is enabled in your account
    2. Even if cluster-level encryption is not enabled, you can still encrypt task or service volumes by configuring encryption settings directly in the task definition