**AWS**

**-EC2**

Amazon Elastic Compute Cloud (Amazon EC2) is a web service from Amazon Web Services (AWS) that provides scalable compute capacity in the cloud. It is designed to make web-scale cloud computing easier for developers  
  
Its like renting a computer from Amazon that you can use over the internet. You don’t have to physically buy or set up the computer yourself, but you can use it whenever you need it, and you only pay for the time you use it.  
  
Key Features of EC2:

Scalability: EC2 allows you to scale your compute capacity up or down based on demand. You can launch multiple EC2 instances or terminate them as needed, offering flexibility in resource management.

Customizable Instances: You can configure the virtual machine's specifications like CPU, memory, storage, and network capacity according to your needs by choosing from a range of instance types.

Elasticity: EC2 provides the ability to increase or decrease the instance count according to the current workload, optimizing cost efficiency.

Various Operating Systems: EC2 supports multiple operating systems, including:

Linux distributions (e.g., Ubuntu, CentOS, Red Hat)

Windows Server

Custom AMIs: You can create your own Amazon Machine Image (AMI) from a pre-configured system setup.

Instance Types: EC2 offers different types of instances tailored to specific workloads:

General Purpose Instances (e.g., t2, t3): Balanced in terms of compute, memory, and networking.

Compute Optimized Instances (e.g., c5, c6g): For tasks requiring higher compute performance.

Memory Optimized Instances (e.g., r5, r6g): For applications that require fast performance for large datasets in memory.

Storage Optimized Instances (e.g., i3, d2): Optimized for applications with high I/O operations and require local storage.

Elastic Block Store (EBS): Persistent block-level storage for EC2 instances. Data stored in EBS persists even if the instance stops or is terminated.

Common Use Cases:

Web and Application Hosting: Hosting dynamic websites, web apps, and backend services.

Big Data Processing: Running high-performance computing (HPC) applications for tasks like scientific simulations or large-scale analytics.

Machine Learning: Training and deploying machine learning models using EC2’s powerful GPU instances.

Gaming Servers: Setting up low-latency game servers for multiplayer games.

Development and Testing: Deploying isolated environments for developers to test and deploy software.

Monitoring and Management:

CloudWatch:

AWS CloudWatch provides monitoring for EC2 instances, including CPU utilization, disk I/O, network traffic, etc. You can set up alarms for certain thresholds.

Elastic Beanstalk:

An easy-to-use service for deploying and managing applications, allowing you to focus on code rather than the underlying infrastructure.

Types:

1. General Purpose Instances

These are balanced in terms of compute, memory, and networking resources, making them suitable for a wide variety of applications.

Use Case: Web servers, development and test environments, small to medium databases.

Instance Types:

t3, t3a, t4g (Burstable Performance): They offer a baseline level of CPU performance with the ability to burst to higher levels temporarily.

m6g, m5, m5a, m4 (Fixed Performance): More powerful and consistent performance for applications that need constant computing power.

2. Compute Optimized Instances

These instances are designed for applications that require high computational power relative to memory.

Use Case: High-performance web servers, batch processing, scientific modeling, machine learning inference.

Instance Types:

c6g, c5, c5n, c4: Optimized for compute-intensive tasks with fast CPUs but relatively less memory.

3. Memory Optimized Instances

These instances are designed for workloads that require fast performance for processing large datasets in memory.

Use Case: In-memory databases (e.g., Redis, Memcached), real-time big data analytics, high-performance computing (HPC).

Instance Types:

r6g, r5, r5a, r4: Optimized for memory-intensive applications.

x2idn, x2iedn: Provides the highest memory capacity per vCPU.

u-6tb1.metal, u-12tb1.metal: For workloads like SAP HANA requiring massive memory.

4. Storage Optimized Instances

These are optimized for high, sequential read and write access to large datasets on local storage.

Use Case: NoSQL databases, data warehousing, distributed file systems, high I/O applications.

Instance Types:

i4i, i3, i3en: Optimized for high random I/O performance, great for databases and file systems that require fast disk reads/writes.

d2, d3, d3en: Best for distributed file systems with high sequential read and write access.

5. Accelerated Computing Instances

These instances use hardware accelerators, or coprocessors, to perform functions like floating-point number calculations, graphics processing, and data pattern matching.

Use Case: Machine learning, AI training and inference, video rendering, graphics-intensive applications, scientific simulations.

Instance Types:

p4, p3: GPU instances for deep learning and AI workloads.

g5, g4ad, g4dn: Optimized for machine learning inference and graphics-heavy applications like video processing.

f1: FPGA instances used for specialized hardware acceleration, such as genomics research or high-frequency trading.

6. High Memory Instances

These are EC2 instances with very large memory allocations, specifically designed for memory-bound applications like in-memory databases.

Use Case: Large in-memory databases (e.g., SAP HANA), big data analytics, high-performance computing (HPC).

Instance Types:

u-6tb1.metal, u-9tb1.metal, u-12tb1.metal: These instances are bare metal and provide the highest memory capacities.

Regions and Availability Zones

Regions and Availability Zones(AZs) are key components of AWS infrastructure, including EC2. They play a crucial role in ensuring high availability, redundancy, and low latency for your cloud applications.

AWS Regions

Definition: An AWS Region is a geographical area where AWS has a cluster of data centers. Each region is isolated from others to ensure the highest levels of security and compliance.

Purpose: Regions allow you to run applications closer to your users, reducing latency and meeting specific legal or regulatory requirements (e.g., data residency laws).

Structure: Each AWS Region typically consists of two or more Availability Zones to provide fault tolerance and high availability.

Naming Convention: AWS Regions have specific names like us-east-1 (Northern Virginia), eu-west-1 (Ireland), and ap-southeast-1 (Singapore).

Availability Zones (AZs)

Definition: An Availability Zone is one or more discrete data centers within a region. Each AZ has its own power, cooling, and network infrastructure, but they are connected to other AZs in the same region with low-latency, high-bandwidth networking.

Purpose: AZs provide high availability and fault tolerance for your applications. By deploying applications across multiple AZs, you ensure that even if one data center (or AZ) goes down, your application can continue running smoothly in others.

Redundancy: AZs are isolated to ensure that failures in one AZ do not affect others. They are geographically separated within a region to reduce the risk of simultaneous failure due to physical events like floods or power outages.

High Availability Architecture:

When you distribute your EC2 instances across multiple AZs, AWS automatically balances the load. If one AZ goes offline, traffic can be shifted to instances running in another AZ within the same region.

Services like Amazon RDS (Relational Database Service) can replicate data across AZs to ensure availability even during an outage.

Naming Convention: AZs are identified by a combination of the region name and a letter (e.g., us-east-1a, us-west-2b).

AMI

In Amazon EC2, an AMI (Amazon Machine Image) is a template that contains the software configuration needed to launch an EC2 instance.

Key Features of an AMI:

Reusable Template: AMIs allow you to quickly launch instances with the exact setup you need without configuring them from scratch every time.

Region-Specific: AMIs are region-specific, meaning you can launch an EC2 instance in the same AWS region where the AMI is stored, but you can copy AMIs across regions.

Customizable: You can create your own custom AMI based on an existing EC2 instance. For example, if you configure a web server on one instance and want to replicate that setup, you can create an AMI from it and use it to launch new instances.

Types of AMIs:

AWS Marketplace AMIs: These are pre-built AMIs from third-party vendors that provide applications like security software, monitoring tools, etc. These are often pay-per-use.

Public AMIs: AMIs that are free to use and publicly available. AWS provides several basic AMIs, such as those with a Linux or Windows operating system.

Private AMIs: Custom AMIs that you create and control. These are used within your account and can be shared with specific AWS accounts if needed.

Public and Elastic IP

In Amazon EC2, both Public IP and Elastic IP addresses are used to allow instances to be accessed from the internet, but they differ in how they are allocated, managed, and retained.

Public IP

Definition: A public IP is an automatically assigned IPv4 address that allows your EC2 instance to be reachable from the internet.

Key Characteristics:

Automatically Assigned: When you launch an EC2 instance in a public subnet of a Virtual Private Cloud (VPC), a public IP address is assigned to the instance by default unless you disable this option.

Dynamic: The public IP is associated with the instance while it is running, but once the instance is stopped or terminated, the public IP is released and cannot be reused.

No Additional Cost: Public IP addresses are provided for free, but because they are not static, they aren't ideal for services requiring a consistent IP address (e.g., DNS entries).

Use Case:

Ideal for temporary or non-critical use cases where you don’t need a permanent, static IP address for your instance.

For example, if you’re hosting a short-term application or conducting tests, a public IP would be sufficient.

Elastic IP

Definition: An Elastic IP (EIP) is a static public IPv4 address that you can allocate for use with your AWS account. It is associated with your account rather than a specific instance, and it remains yours until you choose to release it.

Key Characteristics:

Static: Unlike public IPs, Elastic IPs do not change when an instance is stopped or restarted. They remain associated with your AWS account and can be reassigned to different instances as needed.

Controlled by You: You control the Elastic IP; you can associate or disassociate it with different instances as your needs change.

Use Case:

Useful for applications where you need a static, public IP address, such as hosting a web service, DNS settings, or when moving an IP between instances during maintenance.

For example, if you’re running a production website that needs a fixed IP address to ensure consistent routing for your users, an Elastic IP is the right choice.

Hibernate

Hibernate is a feature that allows you to pause your instance, saving its current state to disk, and then later resume it from exactly where it left off. This is similar to how you hibernate a laptop — rather than shutting it down, the system is paused, and when you turn it back on, you can continue working without losing your session.

Key Features of Hibernate:

Preserve RAM State: Unlike a regular stop, hibernate saves the memory (RAM) contents to the disk, so applications running in memory before hibernation can be resumed.

Faster Startup: Since the instance resumes from the saved RAM state, it starts up much faster compared to booting from scratch.

Preserve Instance Attributes: The instance retains the same instance ID, root EBS volume, and data stored in the EBS volume. You won’t lose any in-memory data as it is saved when the instance hibernates.

Billing: During hibernation, you only pay for the storage of the EBS volume, not for the instance’s compute power. However, if you’re using Elastic IP addresses or reserved instance billing, charges for those will still apply.