Amazon EC2

Amazon Elastic Compute Cloud (EC2) is a web service that provides resizable compute capacity in the cloud. It enables businesses to run virtual servers, scale up or down as needed, and pay only for the resources they use. EC2 is an essential service within the AWS ecosystem, allowing users to deploy and manage virtual machines effortlessly.

EC2 Instance Types

Amazon EC2 offers a variety of instance types optimized for different use cases. Here are the main categories:

1. **General Purpose Instances**
   - Balanced CPU, memory, and networking.
   - Examples: t3, t3a, m5, m6g.
   - Use Case: Web servers, small databases, development environments.
2. **Compute Optimized Instances**
   - High-performance processors for compute-heavy applications.
   - Examples: c5, c6g, c7g.
   - Use Case: High-performance computing (HPC), batch processing, gaming servers.
3. **Memory Optimized Instances**
   - Large memory capacity for memory-intensive workloads.
   - Examples: r5, r6g, x2idn, z1d.
   - Use Case: In-memory databases, caching, analytics.
4. **Storage Optimized Instances**
   - High disk throughput and storage capabilities.
   - Examples: i3, i4i, d2, h1.
   - Use Case: Large-scale data processing, big data analytics.
5. **Accelerated Computing Instances**
   - Includes GPUs and FPGAs for AI, ML, and scientific computing.
   - Examples: p4, g4dn, f1.
   - Use Case: Machine learning (ML), video processing, deep learning.

EC2 Pricing Models

Amazon EC2 provides multiple pricing options to suit various use cases and budgets:

1. **On-Demand Instances**
   - Pay-as-you-go pricing with no long-term commitment.
   - Suitable for short-term workloads and unpredictable traffic.
2. **Reserved Instances (RI)**
   - Commit to using instances for 1 or 3 years for significant discounts.
   - Suitable for steady-state workloads.
3. **Spot Instances**
   - Purchase unused EC2 capacity at up to 90% discount.

    - Suitable for fault-tolerant applications, batch processing, and testing.
4. **Savings Plans**
    - Flexible pricing model offering lower prices for consistent usage over time.
    - Provides discounts similar to Reserved Instances but with more flexibility.
5. **Dedicated Hosts**
    - Physical servers dedicated to a single customer for regulatory compliance.
    - Useful for enterprises with strict licensing requirements.

Auto Scaling Groups (ASG)

Auto Scaling Groups (ASG) help manage the number of EC2 instances automatically, ensuring high availability and cost optimization. ASG can:

- **Scale Out**: Add instances during peak demand.
- **Scale In**: Remove instances during low demand.
- **Replace Failed Instances**: Ensure the application runs smoothly by replacing unhealthy instances.
- **Scheduled Scaling**: Increase or decrease capacity based on predictable traffic patterns.

Components of Auto Scaling Groups

1. **Launch Template/Configuration**: Defines the instance type, AMI, and other settings.
2. **Scaling Policies**: Determines how instances scale based on CPU usage, network traffic, or other metrics.
3. **Load Balancer Integration**: Distributes traffic across instances in the group.

Load Balancers in AWS EC2

AWS Elastic Load Balancing (ELB) automatically distributes incoming traffic across multiple instances. There are three main types:

1. **Application Load Balancer (ALB)**
    - Operates at the **application layer (Layer 7)**.
    - Ideal for HTTP/HTTPS traffic.
    - Supports host-based and path-based routing.
2. **Network Load Balancer (NLB)**
    - Operates at the **transport layer (Layer 4)**.
    - Capable of handling millions of requests per second with ultra-low latency.
    - Suitable for TCP/UDP-based applications.
3. **Classic Load Balancer (CLB)**
    - Legacy load balancer operating at both Layer 4 and Layer 7.
    - Recommended for applications requiring simple load balancing.