AWS

EC2 – Elastic Compute Cloud

Introduction to Amazon EC2

What is EC2?

- Amazon **EC2 (Elastic Compute Cloud)** is a cloud service that provides scalable virtual servers.
- It lets you run applications on-demand without managing physical hardware.
- You can choose instance types, storage, and networks based on your needs.
- It scales automatically to handle variable workloads efficiently.

Use cases of EC2

- **Hosting web applications** that need scalable, reliable compute resources.
- **Running batch processing or data analytics** jobs on-demand.
- **Deploying microservices and containerized workloads** using ECS or Kubernetes.
- **High-performance computing (HPC)** for simulations, ML training, and scientific workloads.

Benefits of EC2

- **Scalability:** Easily scale instances up or down based on demand.
- **Cost-effective:** Pay only for the compute you use.
- **Flexible configuration:** Wide choice of instance types, OS, storage, and networking.
- **High availability:** Runs in multiple regions and Availability Zones for reliability.

EC2 Instance Types

- General Purpose
- Compute Optimized
- Memory Optimized
- Storage Optimized

General Purpose

- **General Purpose EC2 instances** offer a balanced mix of compute, memory, and networking resources.
- They are suitable for most everyday workloads and applications.
- Popular families include **T-series (burstable)** and **M-series (balanced)**.
- They provide steady performance with flexible configuration options.
- Ideal for web servers, small databases, microservices, and development environments.
- They deliver good performance at a cost-efficient price point.

Compute Optimized

- **Compute Optimized EC2 instances** are designed for compute-intensive workloads.
- They provide high-performance processors with low latency.
- Common families include **C-series** (e.g., C6i, C7g).
- Ideal for high-performance computing, batch processing, and scientific modeling.
- Also suitable for media transcoding, gaming servers, and machine learning inference.
- They deliver maximum compute power at a cost-efficient rate.

Memory Optimized

- Memory Optimized EC2 instances are designed for workloads that require high memory-to-CPU ratios.
- They are ideal for large in-memory databases, real-time big data analytics, and high-performance computing.
- Common families include **R-series**, **X-series**, **High Memory** (u-*), and **z-series** instances.
- These instances offer very high RAM capacity, often scaling into multiple terabytes.
- They provide enhanced networking and optimized performance for memory-intensive operations.
- Use them when your workload depends more on memory throughput and size rather than raw CPU power.

Storage Optimized

- Storage Optimized instances are built for workloads needing high, low-latency, and sequential read/write performance.
- They use **NVMe SSDs** or **HDDs** for extremely fast local storage access.
- Ideal for **NoSQL databases**, **data warehousing**, **distributed file systems**, and **search workloads**.
- Families include **I-series**, **D-series**, and **H-series** instances.
- They deliver high IOPS, high throughput, and consistent performance for data-heavy operations.
- Best suited for applications where storage speed and capacity are more important than compute power.

EC2 Key Concepts

- AMI (Amazon Machine Image)
- Instance Types
- Security Groups
- Key Pairs

AMI (Amazon Machine Image)

- An AMI is a preconfigured template used to launch EC2 instances in AWS.
- It includes the operating system, application server, and any installed software.
- AMIs can be AWS-provided, user-created, or shared across accounts.
- They allow consistent, repeatable deployments of servers.
- You can customize, copy, or version AMIs for different environments.

Instance Types

- EC2 instance types define different combinations of CPU, memory, storage, and networking capacity.
- They are grouped into families like General Purpose, Compute Optimized, Memory Optimized, and Storage Optimized.
- Each type is designed to support specific workload requirements
- Users can choose sizes within each family for scaling resources.
- This flexibility helps optimize performance and cost for different applications.

Security Groups

- Security Groups act as virtual firewalls for your AWS resources, mainly EC2 instances.
- They control **inbound and outbound** traffic based on defined rules.
- Rules allow or deny traffic using protocols, ports, and source/destination IPs.
- Security Groups are **stateful**, meaning return traffic is automatically allowed.
- They can be attached to multiple instances for consistent security management.
- They help ensure controlled, secure access to your cloud environment.

Key Pairs

- A key pair is used to securely connect to EC2 instances.
- It consists of a **public key** stored by AWS and a **private key** kept by you.
- The private key allows SSH or RDP access to your instance.
- It ensures that only authorized users can log in.
- Losing the private key means you cannot access the instance directly.

EC2 Storage Options

- EBS Volumes
- Instance Store
- EBS Volume Types

EBS Volumes

- EBS (Elastic Block Store) volumes provide persistent block storage for EC2 instances.
- They remain available even if the instance stops or terminates (unless using ephemeral storage).
- EBS volumes can be attached, detached, and reattached to any instance in the same Availability Zone.
- They support features like snapshots, encryption, and resizing.
- You can choose volume types based on performance needs (GP, IO, HDD).
- Ideal for databases, file systems, and applications requiring durable storage.

Instance Store

- Instance Store provides **temporary block-level storage** directly attached to the EC2 host machine.
- It offers very **high I/O performance** since it uses local physical disks.
- Data is **not persistent** and is lost if the instance stops, terminates, or fails.
- Suitable for **buffering, caching, and temporary data** workloads.
- Ideal when you need **fast, ephemeral storage** with low latency.

EBS Volume Types

- AWS offers different EBS volume types based on performance and cost needs.
- **gp3** (General Purpose SSD) is the default, balancing price and performance for most workloads.
- **io1/io2** (Provisioned IOPS SSD) provide the highest performance for mission-critical databases.
- **st1** (Throughput Optimized HDD) is ideal for big data and streaming workloads.
- **sc1** (Cold HDD) is the lowest-cost option for infrequently accessed data.
- Each type varies in IOPS, throughput, and durability to fit specific application requirements.

Networking in EC2

- VPC Basics
- Subnets
- Public vs Private IP
- ENI (Elastic Network Interface)

VPC Basics

- A VPC is a virtual network dedicated to your AWS account.
- It allows you to launch AWS resources in an isolated, secure environment.
- You can define **subnets**, **IP ranges**, **route tables**, and **gateways**.
- Security controls like **security groups** and **network ACLs** manage traffic.
- VPCs can be connected to on-premises networks via **VPN or Direct Connect**.
- They provide full control over networking, isolation, and security in the cloud.

Subnets

- A subnet is a segment of a VPC’s IP address range.
- It divides the VPC into smaller, manageable networks.
- Subnets can be **public** (accessible from the internet) or **private** (isolated).
- They help organize resources based on security and functional requirements.
- Route tables control traffic flow between subnets and the internet.
- Using subnets improves network performance, security, and scalability.

Public vs Private IP

- A **Public IP** is reachable over the internet, while a **Private IP** is used within a local network.
- Public IPs allow external access to resources like web servers.
- Private IPs are used for internal communication between instances or devices.
- Public IPs are unique across the internet; private IPs can be reused in different networks.
- Public IPs may incur additional costs, whereas private IPs are free within a VPC.
- Using private IPs with NAT or a gateway allows secure internet access without exposing internal resources.

ENI (Elastic Network Interface)

- An ENI is a virtual network card that can be attached to an EC2 instance.
- It provides one or more private IP addresses, a MAC address, and security group associations.
- ENIs enable instances to have multiple network interfaces for different purposes.
- They can be detached from one instance and attached to another in the same VPC.
- Useful for high availability, network separation, and failover configurations.

Elastic Load Balancing

- Application or Classical Load Balancer
- Network Load Balancer
- Gateway Load Balancer

Application or Classical Load Balancer

- Load balancers distribute incoming traffic across multiple EC2 instances to improve availability and performance.
- **Application Load Balancer (ALB)** works at Layer 7 (HTTP/HTTPS) and supports content-based routing.
- ALBs can route requests based on URL paths, host headers, or query parameters.
- **Classic Load Balancer (CLB)** works at both Layer 4 (TCP) and Layer 7 but with limited features.
- CLBs are simpler and suitable for legacy applications.
- Using load balancers ensures fault tolerance, scalability, and seamless user experience.

Network Load Balancer

- NLB operates at **Layer 4 (TCP/UDP)** for ultra-fast, low-latency traffic distribution.
- It is designed to handle millions of requests per second while maintaining high throughput.
- NLB automatically scales to manage sudden traffic spikes.
- It provides a **static IP** per Availability Zone and supports Elastic IPs.
- Ideal for performance-critical applications requiring extreme reliability and minimal latency.

Gateway Load Balancer

- GWLB operates at **Layer 3 (Network Layer)** to distribute traffic to virtual appliances.
- It is used for deploying, scaling, and managing third-party network appliances like firewalls or intrusion detection systems.
- GWLB provides **transparent traffic handling** with a single entry and exit point.
- It integrates with **VPC routing** to redirect traffic without changing IP addresses.
- Ideal for scalable and resilient network security and monitoring solutions.

Auto Scaling

- ASG Basics
- Scaling Policies
- Benefits of Auto Scaling

ASG Basics

- An ASG automatically manages a group of EC2 instances to match demand.
- It can **scale out** (add instances) or **scale in** (remove instances) based on defined metrics.
- ASGs help maintain application **availability and fault tolerance**.
- Launch configurations or templates define instance type, AMI, and networking settings.
- They ensure cost optimization by running only the required number of instances.

Scaling Policies

- Scaling policies define how an Auto Scaling Group (ASG) adjusts the number of instances.
- **Target tracking** policies automatically maintain a specific metric, like CPU utilization.
- **Step scaling** policies adjust capacity in steps based on metric thresholds.
- **Simple scaling** policies add or remove instances based on a single alarm.
- They help optimize performance, availability, and cost by dynamically adjusting resources.

Benefits of Auto Scaling

- Ensures **high availability** by maintaining the right number of instances.
- Automatically **scales resources** up or down based on demand.
- Helps **optimize costs** by running only the necessary instances.
- Improves **fault tolerance** by replacing unhealthy instances automatically.
- Reduces manual intervention and supports **dynamic workloads efficiently**.

EC2 Monitoring

- CloudWatch Metrics
- Alarms
- EC2 Status Checks

CloudWatch Metrics

- CloudWatch Metrics are data points collected over time to monitor AWS resources.
- They provide insights into resource performance, such as CPU, memory, and disk usage.
- Metrics can trigger **alarms** when thresholds are breached.
- Both AWS services and custom applications can send metrics to CloudWatch.
- They help optimize performance, detect issues, and maintain system reliability.

Alarms

- CloudWatch Alarms monitor AWS resources and trigger actions based on metrics.
- They compare metric values against defined thresholds over a specific period.
- Alarms can **send notifications** via SNS or automatically **scale resources**.
- They help detect issues like high CPU, low disk space, or unhealthy instances.
- Useful for **proactive monitoring** and automating response to changes in resource state.

EC2 Status Checks

- EC2 Status Checks monitor the health of instances and the underlying hardware.
- **System status checks** detect issues with the physical host, networking, or AWS infrastructure.
- **Instance status checks** identify problems within the EC2 instance itself, like OS or boot errors.
- Failed checks can trigger notifications or automated recovery actions.
- Regular status checks help ensure **high availability and reliability** of your instances.

EC2 Best Practices

- Right-sizing
- Use of Spot Instances
- Security Best Practices

Right-sizing

- Right-sizing is the process of matching cloud resources to the actual workload requirements.
- It helps avoid over-provisioning, which wastes cost, and under-provisioning, which affects performance.
- In AWS, it often involves adjusting EC2 instance types, storage, or other resources.
- Monitoring usage metrics helps identify opportunities for optimization.
- Right-sizing improves **cost efficiency, performance, and resource utilization**.

Use of Spot Instances

- Spot Instances allow you to use spare AWS EC2 capacity at **significantly reduced costs**.
- They are ideal for **flexible, fault-tolerant workloads** like batch processing, big data, or testing.
- AWS can **interrupted them with a two-minute notice** if capacity is needed elsewhere.
- They help **optimize costs** while scaling workloads efficiently.

Security Best Practices

- Enable **multi-factor authentication (MFA)** for all accounts.
- Use **least privilege access** by granting only necessary permissions.
- Regularly **rotate credentials and keys** to reduce security risks.
- Monitor activity with **CloudTrail, CloudWatch, and GuardDuty**.
- Encrypt data at rest and in transit for **confidentiality and integrity**.