AWS EC2 Services - Summary Notes

1. Amazon EC2 Overview

Amazon Elastic Compute Cloud (EC2) is a central part of AWS's cloud computing platform, providing scalable computing capacity in the cloud. It allows users to launch virtual machines (called instances), configure networking and security, and manage storage. EC2 helps eliminate the need to invest in hardware upfront, allowing applications to scale up or down easily.

2. EC2 Instance Types

Amazon EC2 provides different instance types to meet specific use cases:

- General Purpose (e.g., t2, t3, m5): Balanced compute, memory, and networking resources. Suitable for web servers and development environments.
- Compute Optimized (e.g., c5, c6g): Ideal for compute-intensive tasks like high-performance web servers, scientific modeling, or machine learning inference.
- Memory Optimized (e.g., r5, x1e): Designed for memory-intensive applications like high-performance databases and in-memory caches.
- Storage Optimized (e.g., i3, d2): Suitable for workloads requiring high IOPS and throughput, such as NoSQL databases or data warehousing.
- Accelerated Computing (e.g., p4, inf1): Use hardware accelerators like GPUs or FPGAs for tasks such as machine learning training and inference.

3. EC2 Key Features

- Elastic IPs: Static IPv4 addresses designed for dynamic cloud computing. Useful when instances fail and need replacement.
- Security Groups: Virtual firewalls that control inbound and outbound traffic to instances.
- AMIs (Amazon Machine Images): Templates containing the software configuration (OS, application server, and applications) required to launch an instance.
- EBS (Elastic Block Store): Persistent block storage that can be attached to EC2 instances.
- Auto Scaling: Automatically increases or decreases instance count based on demand.
- Elastic Load Balancer (ELB): Distributes incoming traffic across multiple targets (instances) to ensure high availability and fault tolerance.

4. EC2 Pricing Models

AWS EC2 Services - Summary Notes

AWS offers several flexible pricing options:

- On-Demand Instances: Pay per hour/second with no long-term commitment. Best for short-term or unpredictable workloads.
- Reserved Instances: Up to 75% cost savings for 1- or 3-year commitments. Suitable for steady-state usage.
- Spot Instances: Purchase unused capacity at discounts up to 90%. Best for fault-tolerant, flexible applications.
- Savings Plans: Commitment-based model for compute usage (EC2, Fargate, Lambda) for significant savings with flexibility.

5. EC2 Storage Options

- Amazon EBS: Block-level storage volumes for persistent data. Supports SSD and HDD types optimized for different use cases.
- Instance Store: Temporary block storage physically attached to the host. Good for ephemeral data.
- Amazon EFS (Elastic File System): Scalable, elastic file storage for use across multiple EC2 instances, ideal for shared access or lift-and-shift workloads.

6. Networking in EC2

- Virtual Private Cloud (VPC): Isolated network environment where you can define IP ranges, subnets, route tables, and gateways.
- Subnets: Logical subdivisions of a VPC to organize resources and control access.
- Elastic IPs: Public IPs associated with your AWS account that can be remapped to different instances.
- Security Groups: Control traffic to and from EC2 instances at the instance level.
- Network ACLs (NACLs): Operate at the subnet level to provide stateless filtering of traffic.

7. Monitoring & Management

- Amazon CloudWatch: Monitors EC2 metrics such as CPU utilization, disk I/O, and network traffic. Allows for custom alarms and dashboards.
- AWS CloudTrail: Logs API calls made in your AWS account, useful for security and compliance auditing.
- EC2 Dashboard: Web-based interface in the AWS Management Console for managing instances, AMIs, volumes, and security settings.

8. EC2 Auto Scaling & Load Balancing

- Auto Scaling: Automatically adjusts the number of instances based on demand. Helps maintain performance while minimizing cost.

- Elastic Load Balancing (ELB): Automatically distributes incoming traffic across multiple instances in different availability zones. Supports health checks and integrates with Auto Scaling.

9. EC2 Best Practices

- Use IAM roles for secure instance access to other AWS services without storing credentials.

- Regularly patch and update your operating systems and applications.

- Monitor instances using CloudWatch to detect anomalies.

- Enable detailed monitoring and logging for troubleshooting and auditing.

- Use Auto Scaling and ELB to ensure high availability and fault tolerance.

- Regularly back up EBS volumes and AMIs.