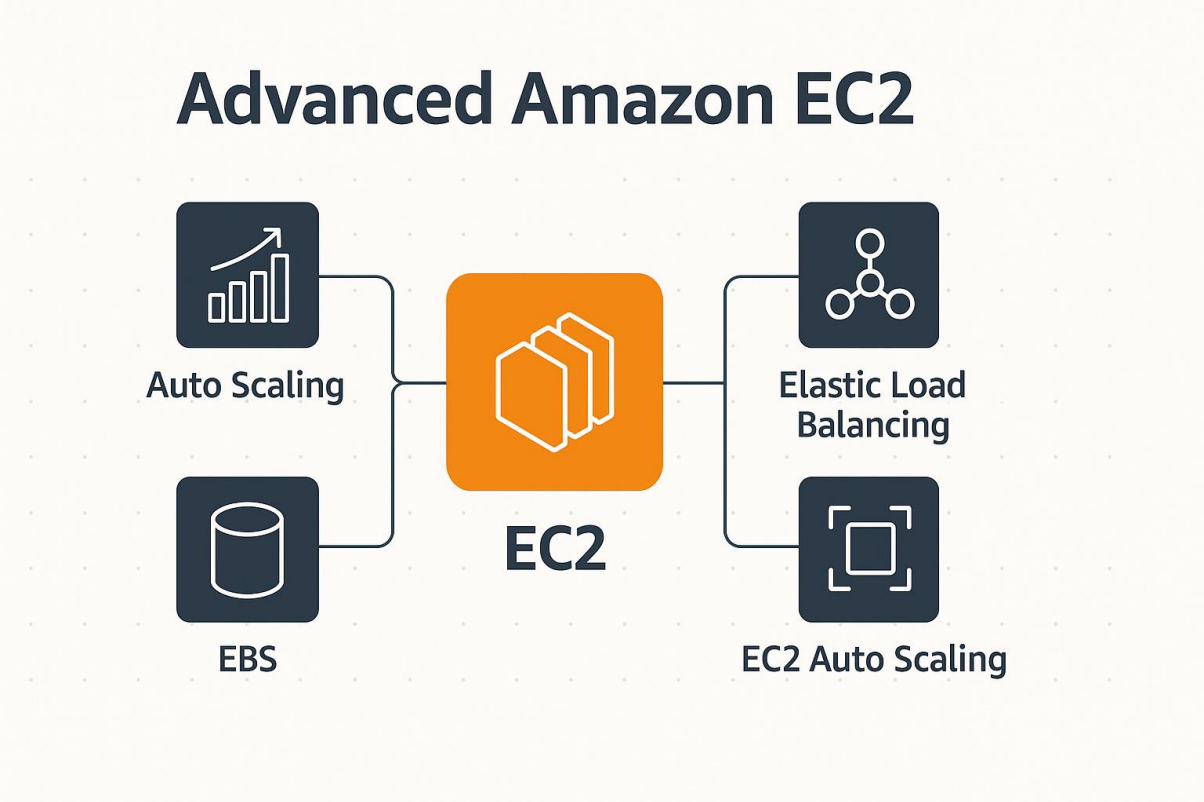
# Mastering Amazon EC2






## 1. What is Amazon EC2?

Amazon EC2 (Elastic Compute Cloud) is a web service that provides resizable compute capacity in the cloud. It's designed to make web-scale cloud computing easier for developers. You can launch virtual servers (instances), configure networking and security, and manage storage.

Core Capabilities:

- Launch and terminate instances on demand.
- Choose AMIs to define OS and software.
- Select instance types based on workload.
- Automate scaling and high availability.
- Integrate with Elastic Load Balancer and Auto Scaling.

---

## 2. EC2 Instance Lifecycle

1. Pending – Instance is being launched.
2. Running – Instance is active.
3. Stopping – Instance is shutting down.
4. Stopped – Instance is off but data is intact.
5. Terminated – Instance is permanently deleted.

---

## 3. EC2 Instance Types (With Use Cases)

General Purpose

- t4g, t3, t2 – Low-cost, burstable performance.
- Use Case: Development, testing, web servers.

Compute Optimized

- c7g, c6g, c5 – High-performance CPU.
- Use Case: Batch processing, game servers, ML inference.

Memory Optimized

- r6g, r5, x2idn – Large memory capacity.
- Use Case: In-memory DBs, real-time analytics.

Storage Optimized

- i4i, d3en, h1 – High IOPS and throughput.
- Use Case: NoSQL DBs, big data workloads.

Accelerated Computing

- p4, inf2, g5 – GPU-based.
- Use Case: AI/ML, video processing, 3D rendering.

---

## 4. Purchasing Options

On-Demand Instances

- No upfront cost. Pay per second.
- Ideal for unpredictable workloads.

Reserved Instances (RIs)

- Commitment (1 or 3 years).
- Up to 75% cost savings.
- Convertible or Standard RIs.

Savings Plans

- Commitment on usage (EC2, Fargate, Lambda).
- More flexibility than RIs.

Spot Instances

- Use spare capacity.
- Up to 90% discount.
- Suitable for fault-tolerant workloads.

Dedicated Hosts

- Physical server for your use.
- Bring Your Own License (BYOL).

---

## 5. Elastic Load Balancing (ELB)

Distributes traffic to multiple EC2s.

Types:

- Application Load Balancer (ALB): HTTP, Web apps.
- Network Load Balancer (NLB): TCP, low latency.
- Gateway Load Balancer (GWLB): Third-party appliances.

Features:

- Health checks.
- Sticky sessions.
- SSL termination.

---

### 6. EC2 Auto Scaling

Auto Scaling ensures availability and cost efficiency.

Components:

- Launch Template/Config
- Auto Scaling Group (ASG)
- Scaling Policies: Target, step, scheduled

Advanced Strategies:

- Predictive scaling
- Lifecycle hooks
- Warm pools

---

### 7. EC2 Storage and Volumes

EBS (Elastic Block Store):

- gp3: General purpose
- io1/io2: High IOPS
- st1/sc1: Throughput and archival

Instance Store:

- High-speed, ephemeral
- Data lost on stop/terminate

EFS (Elastic File System):

- NFS file system

- Scalable across multiple EC2

Amazon FSx:

- Windows File Server, Lustre, NetApp ONTAP

---

## 8. AMIs and Snapshots

Amazon Machine Image (AMI):

- Template to launch instances
- Includes OS, configuration, apps

Snapshots:

- Point-in-time backups of EBS volumes
- Used to create AMIs

Golden AMI Strategy:

- Hardened base image
- Pre-installed apps and security settings

---

## 9. Backup Strategies

EBS Snapshots:

- Manual or scheduled
- Incremental backups

AMI-Based Backups:

- Save complete OS and data state

AWS Backup:

- Centralized backup across services

Cross-region Backup:

- Enable DR and compliance

---

## 10. EC2 Security Best Practices

Key Concepts:

- Security Groups: Instance-level firewall
- NACLs: Subnet-level control
- Key Pairs: SSH authentication
- IAM Roles: Secure access to AWS services
- SSM: Secure shell-free management

Hardening Tips:

- Disable root login
- Regular patching
- Use least privilege IAM policies

---

## 11. Monitoring and Logging

CloudWatch:

- EC2 metrics, custom alarms
- Log agent for file-level monitoring

CloudTrail:

- Record API activity

EC2 Detailed Monitoring:

- 1-minute interval metrics

AWS Config:

- Audit and compliance checks

---

## 12. Server Connectivity & Session Management

Linux:

- SSH with PEM key
- SSM Session Manager

Windows:

- RDP (Remote Desktop Protocol)
- EC2 Connect for browser-based access

Connection Tools:

- Putty, Mobaxterm, VS Code SSH plugin

---

## 13. High Availability and Cost Optimization

High Availability:

- Deploy in multiple AZs
- Use Load Balancer + Auto Scaling
- Elastic IP for static access

Cost Optimization:

- Choose right instance type
- Use Spot and RIs where suitable
- Use Auto Scaling to scale down
- Schedule non-production shutdown

---

## 14. Real-Time Use Cases

1. Web Hosting: Scalable app hosting with Auto Scaling + ALB
2. Batch Jobs: Use Spot Instances for cost-effective processing
3. Gaming: Low-latency game server on EC2
4. CI/CD Runners: Host Jenkins or GitHub runners

5. Dev/Test Environments: Spin up/down quickly
6. AI/ML Training: Use GPU-based instances

---

## 15. EC2 Interview Questions

Basic:

- What is EC2?
- How do you connect to EC2?
- What are the differences between AMI and Snapshot?

Intermediate:

- How does Auto Scaling work?
- Difference between Security Group and NACL?
- What is the difference between instance store and EBS?

Advanced:

- Design a fault-tolerant EC2 architecture.
- How do you implement patch management in EC2?
- How do you maintain golden AMIs?
- How do you monitor 100+ EC2 instances efficiently?

---

## 16. Conclusion

This EC2 document was crafted to empower AWS learners and DevOps professionals with a full-stack understanding of Elastic Compute Cloud. From launching instances to securing and optimizing them, this knowledge is critical in interviews and real-world deployments.