AWS EC2

What is AWS EC2?

Amazon EC2 (Elastic Compute Cloud) is a web service that provides secure, resizable compute capacity in the cloud. In simple words, EC2 is like a virtual machine (VM) that runs in the cloud.

- Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides secure, resizable compute capacity in the cloud.

- Access reliable, scalable infrastructure on demand. Scale capacity within minutes with SLA commitment of 99.99% availability.

- Provide secure compute for your applications. Security is built into the foundation of Amazon EC2 with the AWS Nitro System.

- Optimize performance and cost with flexible options like AWS Graviton-based instances, Amazon EC2 Spot instances, and AWS Savings Plans.

Key Features:

Feature	Description
Scalability	Easily increase or decrease capacity (number of instances).
Elastic IP	Static IP address for dynamic cloud computing.
Security Groups	Firewall rules that control traffic to your instances
AMI	Amazon Machine Image — preconfigured OS and software for launching EC2.
Instance Types	Different hardware configurations (CPU, RAM, Network)
Key Pairs	Secure login using SSH private-public key

Common EC2 Use Cases:

Hosting websites/applications

Running backend services or APIs

Running test environments

Data processing

Deliver secure, reliable, high-performance, and cost-effective compute infrastructure to meet demanding business needs.

Access the on-demand infrastructure and capacity you need to run HPC applications faster and cost-effectively.

Access environments in minutes, dynamically scale capacity as needed, and benefit from AWS's pay-as-you-go pricing.

Deliver the broadest choice of compute, networking (up to 400 Gbps), and storage services purpose-built to optimize price performance for ML projects

Hands-On: Launch Your First EC2 Instance

Prerequisites:

An [AWS Free Tier account](https://aws.amazon.com/free/)

Web browser

Step-by-Step Guide to Launch EC2:

1. Login to AWS Console → https://console.aws.amazon.com/

2. Go to EC2 Dashboard

3. Click Launch Instance

4. Fill out:

Name: `MyFirstEC2`

AMI: Choose Amazon Linux 2023 (free tier eligible)

Instance Type: `t2.micro` (free tier)

Key Pair: Create new → Download `.pem` file and keep it safe

Network Settings:

Allow SSH (22) and optionally HTTP (80)

Storage: Default 8 GB is fine

5. Click Launch Instance

6. Wait until instance is Running

7. Connect:

Open terminal (Mac/Linux) or use Git Bash (Windows)

Run:

```bash
chmod 400 your-key.pem
ssh -i "your-key.pem" ec2-user@<Public-IP>
```

Exercise: Deploy a Simple Web Server

Once logged in to the EC2 instance:

1. Install Apache HTTP Server:

```bash
sudo yum update -y
sudo yum install httpd -y
```

2. Start Web Server:

```bash
sudo systemctl start httpd
sudo systemctl enable httpd
```

3. Add HTML Page:

```bash
echo "<h1>Hello from EC2!</h1>" | sudo tee /var/www/html/index.html
```

4. Open Web Browser:

Go to: `http://<Your-EC2-Public-IP>`

You should see: `Hello from EC2!`

What is an EC2 Instance Type?

An **EC2 instance type** defines:

- CPU (vCPUs)
- Memory (RAM)
- Storage type and bandwidth
- Network performance

Each type is designed for specific workloads (general use, compute-heavy, memory-heavy, etc.).

Family	Optimized for	Examples
t	**General purpose**	t2.micro, t3a.small
m	**Balanced CPU + Memory**	m5.large, m6i.xlarge
c	**Compute optimized**	c5.large, c6g.xlarge
r	**Memory optimized**	r5.large, r6g.xlarge
g, p, inf	**GPU-based for ML/AI/Graphics**	g4dn.xlarge, p3.2xlarge
i, d	**Storage optimized (fast I/O)**	i3.large, d2.xlarge
h, z	**High memory / high clock speed**	z1d.large

General purpose

General Purpose instances are designed to deliver a balance of compute, memory, and network resources. They are suitable for a wide range of applications, including web servers,

small databases, development and test environments, and more.

Compute optimized

Compute Optimized instances provide a higher ratio of compute power to memory. They excel in workloads that require high-performance processing such as batch processing,

scientific modeling, gaming servers, and high-performance web servers.

Memory optimized

Memory Optimized instances are designed to handle memory-intensive workloads. They are suitable for applications that require large amounts of memory, such as in-memory databases, real-time big data analytics, and high-performance computing.

Storage optimized

Storage Optimized instances are optimized for applications that require high, sequential read and write access to large datasets.

They are ideal for tasks like data warehousing, log processing, and distributed file systems.

Accelerated computing

Accelerated Computing Instances typically come with one or more types of accelerators, such as Graphics Processing Units (GPUs),

Field Programmable Gate Arrays (FPGAs), or custom Application Specific Integrated Circuits (ASICs).

These accelerators offload computationally intensive tasks from the main CPU, enabling faster and more efficient processing for specific workloads.

Model	Description	Use Case
On-Demand	Pay per hour or second, no commitment	Short-term, unpredictable workloads
Reserved	1- or 3-year contract, big discount	Long-term usage (e.g. stable web apps)
Spot	Up to 90% cheaper, can be interrupted by AWS anytime	Fault-tolerant, batch jobs, CI/CD runners
Savings Plans	Commit to spend per hour, flexible across instance types	Long-term flexible workloads

Use Case	Instance Family to Use	Example Type
Small website	General purpose	t2.micro (Free tier)
Web server/app backend	Balanced	m5.large
High CPU: encoding, CI/CD	Compute optimized	c5.large
Memory-heavy: DB, cache	Memory optimized	r5.large
ML training/inference	GPU optimized	g4dn.xlarge, p3
IOPS-heavy DB (NoSQL)	Storage optimized	i3.large
Real-time gaming/render	High clock speed / GPU	z1d.large, g5

EC2 UseCases:

Deliver secure, reliable, high-performance, and cost-effective compute infrastructure to meet demanding business needs.

Access the on-demand infrastructure and capacity you need to run HPC applications faster and cost-effectively.

Access environments in minutes, dynamically scale capacity as needed, and benefit from AWS's pay-as-you-go pricing.

Deliver the broadest choice of computer, networking (up to 400 Gbps), and storage services purpose-built to optimize price performance for ML projects.

How to Choose an EC2 Instance

1. Start with a free-tier or small instance

- t2.micro (1 vCPU, 1GB RAM) is **free tier eligible**
- Good for basic testing and learning

2. Identify your workload

- **Web server** = needs balance → t or m
- **Compute-heavy** = encoding, CI/CD → c
- **Memory-heavy** = databases → r
- **AI/ML** = requires GPU → g, p
- **Fast disk** = high IOPS DB → i, d