Elastic Compute Cloud (EC2)

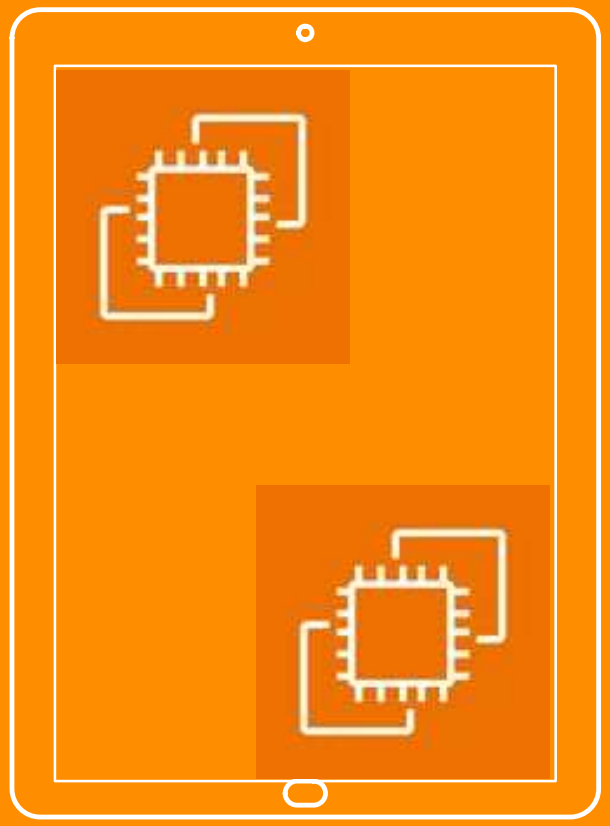
Overview of compute service

TABLE OF CONTENTS

01

Introduction

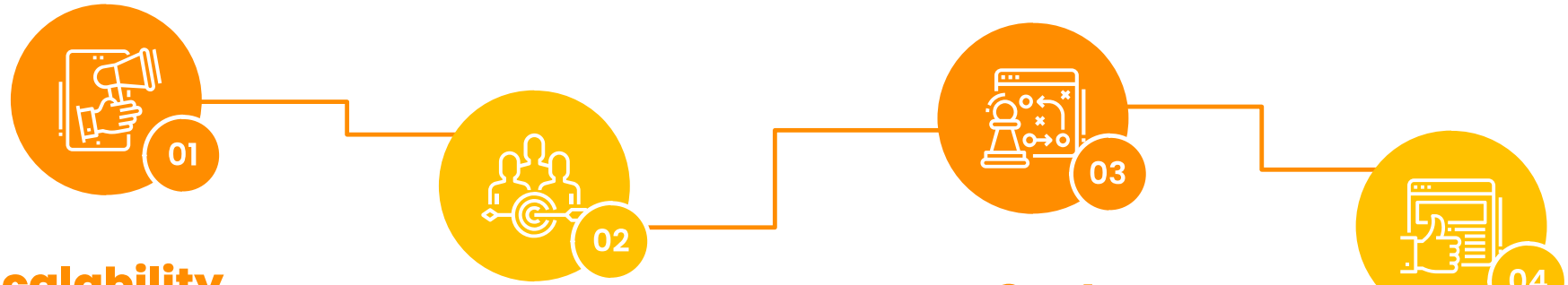
Configuration, User Data, AMI



Introduction

A web service that provides resizable compute capacity in the cloud. It is designed to make web-scale cloud computing easier for developers by allowing them to obtain and configure virtual servers, known as instances, quickly and easily.

Benefits of EC2



Scalability

Easily scale up or down to handle changes in requirements or spikes in popularity.

Flexibility

Wide range of instance types to fit different use cases.

Cost

Pay only for the resources used

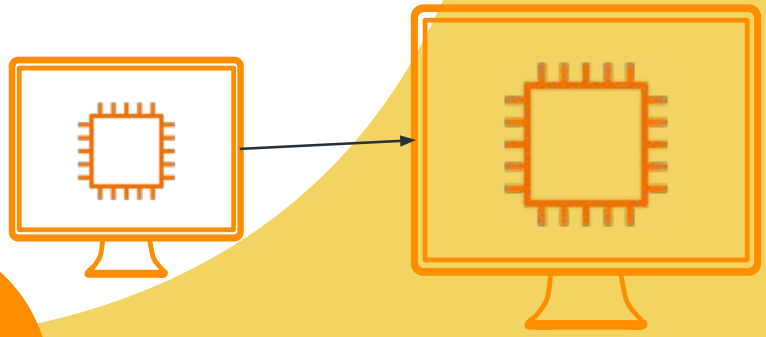
Security

Integrated with AWS security services like IAM and VPC

Scaling

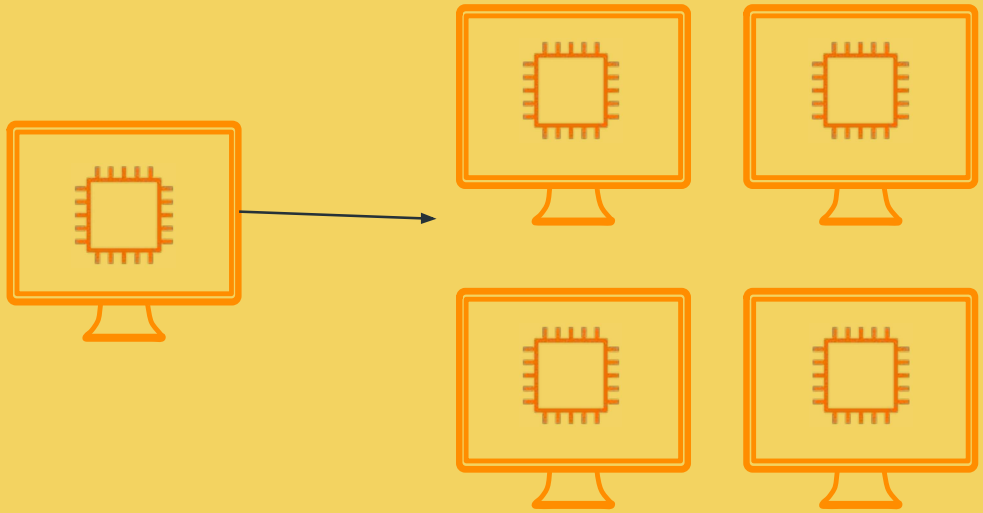
Vertical Scaling

- Increasing or decreasing the resources (CPU, RAM) of a single server or instance.
- Upgrade the hardware of the existing server or instance.



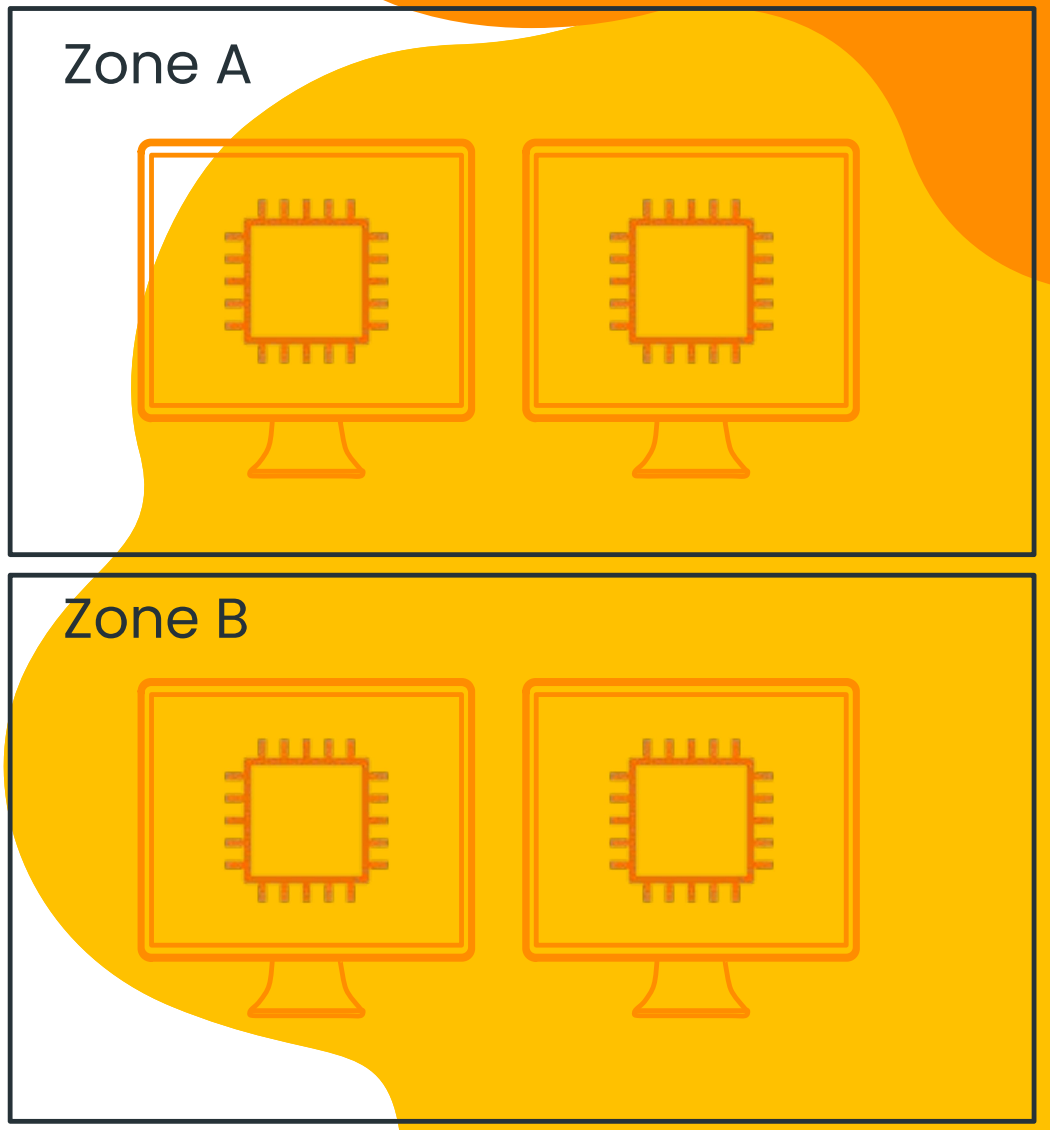
Horizontal Scaling

- Adding or removing instances to/from your system
- Distribute the load across multiple servers or nodes.



High Availability

High Availability (HA) refers to systems that are designed to be operational and accessible without interruption for a very high percentage of time. This is achieved through redundancy, failover mechanisms, and eliminating single points of failure.



Configurations

EC2 Configurations on the console

Terminal Configurations

```
aws ec2 run-instances \

    --image-id ami-0abcdef1234567890 \  # Replace with your chosen AMI ID

    --count 1 \

    --instance-type t2.micro \

    --key-name MyKeyPair \

    --security-group-ids sg-123abc45 \  # Replace with your security group ID

    --subnet-id subnet-6e7f829e \  # Replace with your subnet ID
```

EC2 AMIs

A master image used to create instances (virtual servers) within the Amazon Elastic Compute Cloud (EC2). It contains the information required to launch an instance.

TYPES OF AMI

Public

Provided by AWS or third parties, available to any AWS user

Private

Created and owned by an individual AWS account, not shared by default.

Market Place

Provided by AWS Marketplace vendors, typically include pre-configured software.

User Data

A mechanism to run scripts or commands on your instance at the time of launch. This feature allows you to automate the setup of your instance, such as installing software, configuring settings, or running any initialization tasks. User data is executed by the cloud-init process on the instance when it first starts.

User Data

```
#!/bin/bash

yum update -y

yum install -y httpd

systemctl start httpd

systemctl enable httpd

echo "<h1>Apache Server Installed</h1>" > /var/www/html/index.html
```

02

Instance Types

EC2 Instance Types

1)

General Purpose

For workloads that require a balance of compute, memory, and networking resources.

Use Cases: Web Servers, Microservices, Test and Development Environments

M and T class

2)

Compute Optimized

Suitable for compute-bound applications that benefit from high-performance processors.

Use Cases: High Performance Web Servers, batch processing, Gaming Servers

C Class

3)

Memory Optimized

Designed to deliver fast performance for workloads that process large datasets in memory.

Use Cases: High Performance Databases, Real time Data analytics, In-memory caches

R, U and X class

4)

Storage Optimized

Designed for workloads that require high, sequential read and write access to large datasets on local storage.

Use Cases: NoSQL Database, Data warehousing, Distributed file systems

D, H and I class

5)

Accelerated Computing

They use hardware accelerators, or co-processors, to perform functions such as floating-point number calculations, graphics processing, or data pattern matching more efficiently than software running on general-purpose CPUs.

Use cases: Machine Learning, Graphic Processing, Scientific Computing

EC2 Instance Types

Category	Example	Use Case
General Purpose	t3.micro, m5.large	Web servers, development environments, microservices
Compute Optimized	c5.large, c6g.medium	High-performance web servers, batch processing, gaming servers
Memory Optimized	r5.large, x1e.xlarge	High-performance databases, real-time big data analytics, in-memory caches
Storage Optimized	i3.large, d2.xlarge	NoSQL databases, data warehousing, distributed file systems
Accelerated Computing	p3.2xlarge, g4dn.xlarge	Machine learning, graphics processing, scientific computing

03

Security & Networking

Security group, key pairs, VPC and Subnets

Security Groups

Scope

Instance level

Support

Supports only "allow" rules

Stateful

Stateful like Door Man, asks when entering only not existing

Default Behavior

Denies all inbound traffic by default, allows all outbound traffic by default

Key Pairs

Public Key

Stored by AWS and associated with your EC2 instance.

Private Key

Kept by you and used to access your instances securely.

VPC & Subnet

VPC

Virtual Private Cloud that allows you to launch AWS resources into a virtual network.

Subnet

Segment your VPC for organizational purposes and security.

04

IP Address

IP Addresses and Elastic IPs

IP Addresses

(EC2) instances can have both private and public IP addresses. Understanding the different types of IP addresses and their use cases is critical for configuring and managing network connectivity.

Private IP Address

Used for communication between instances within the same VPC (Virtual Private Cloud). These addresses are not routable over the internet.

Public IP Address

Allow EC2 instances to communicate with the internet. These addresses are routable over the internet.

Static public IPv4 addresses designed for dynamic cloud computing

Elastic IPs

Key Features

Static

Unlike regular public IP addresses, EIPs are static and do not change.

Reassociation

can be reassigned to another instance or network interface in your account

Cost

AWS charges for EIPs when they are not associated with a running instance

Use Cases

Failover Scenarios

Reassign EIPs to standby instances to quickly recover from instance failures.

Consistent Addressing

Maintain the same public IP address even when the underlying instance changes, which is useful for DNS configurations.

10 Minute Break

05

Storage Options

Overview on Storage Options

Storage Options

EBS

Persistent block storage for use with instances.

Instance Store

Temporary block-level storage for instances.

EFS

Scalable file storage for use with EC2.

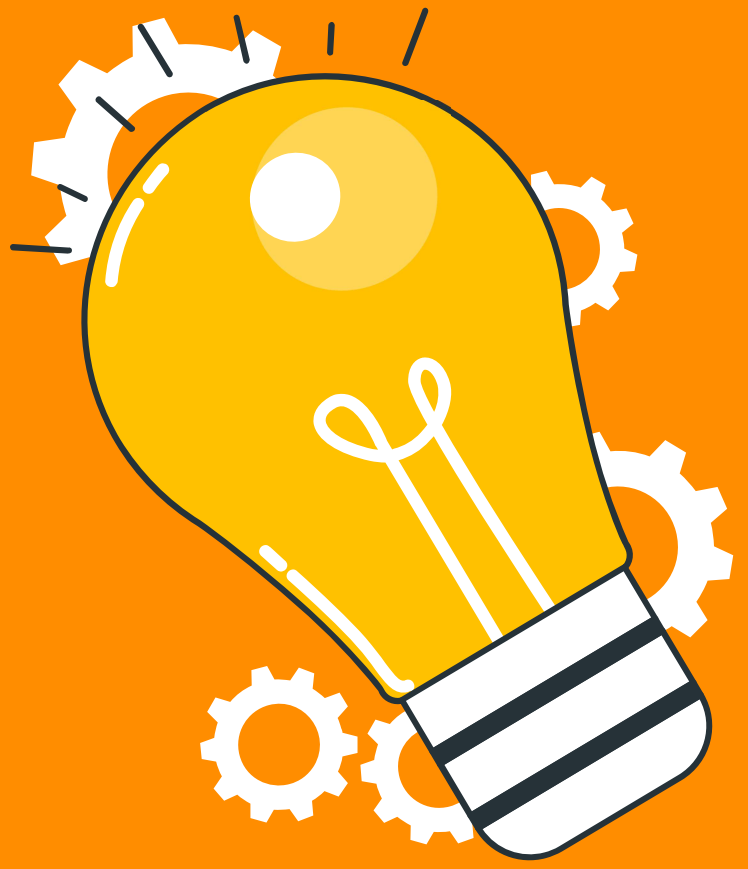
S3

Object storage service

06

Monitoring

Cloudwatch, ELB, & ASG

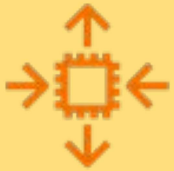


Cloudwatch

Monitor your EC2 instances and other AWS resources in real-time.

ELB

Distributes incoming application traffic across multiple EC2 instances.



ASG

Automatically adjusts the number of EC2 instances in response to changes in demand.

Elastic Load Balancers

ALB

NLB

GLB

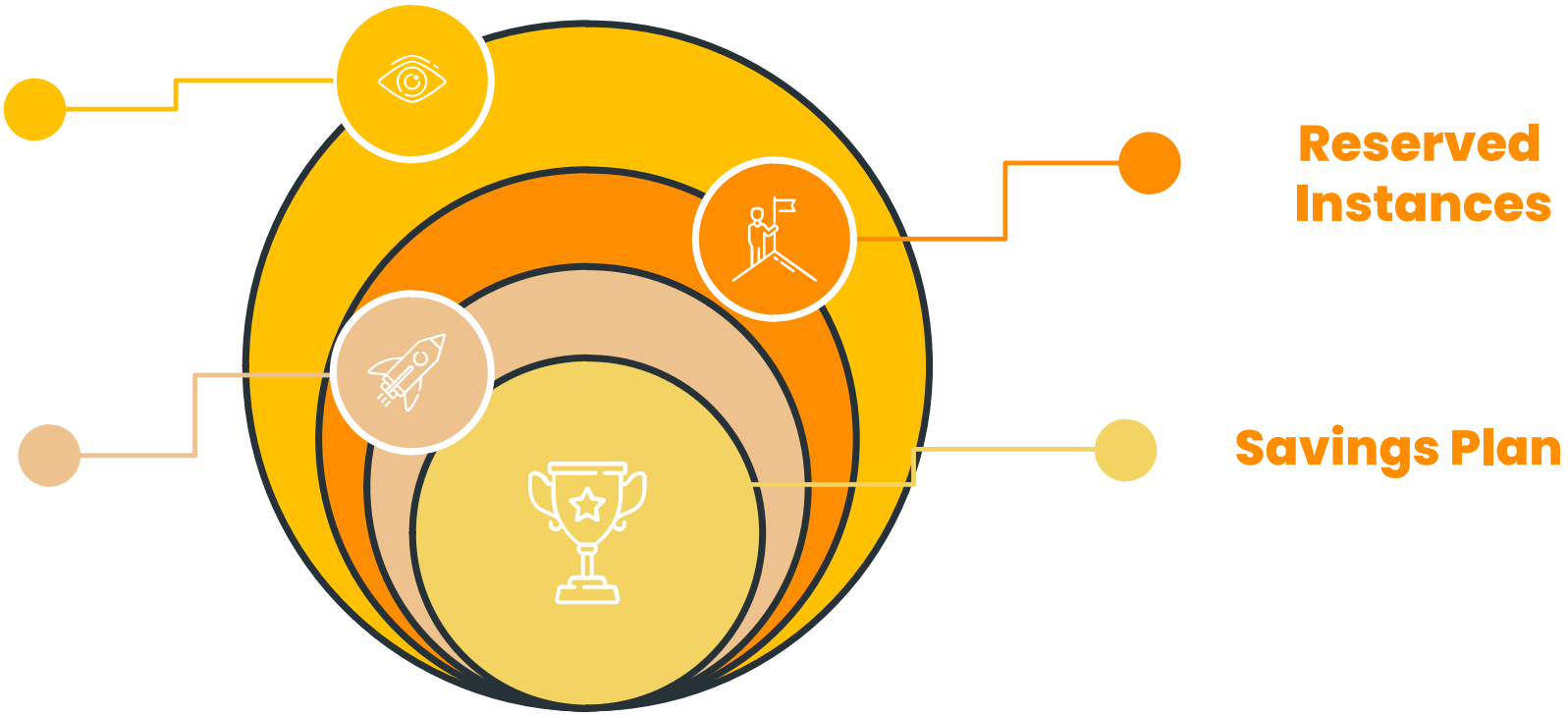
CLB

07

Pricing and Cost

Instance pricing options

Purchasing Options



Purchasing Options	Description	Use Cases	Billing	Commitment	Flexibility
On Demand	Pay for compute capacity by the hour or second with no long-term commitments.	Short-term, spiky, or unpredictable workloads that cannot be interrupted.	Per-second billing (min of 60 sec).	No commitment.	Full.
Reserved	Up to 75% discount compared to On-Demand pricing. Requires a commitment of 1 or 3 years.	Steady-state or predictable usage.	Upfront, Partial, or No Upfront.	1-year or 3-year commitment	Less
Saving Plans	Up to 72% discount compared to On-Demand pricing. Applies to EC2 and Fargate usage.	Steady-state or predictable usage with more flexibility compared to RIs.	Commitment to a consistent amount of usage ($/hour) for a 1-year or 3-year term.	1-year or 3-year commitment.	Flexible across instance families, sizes, AZs, regions, OS, and tenancy

Purchasing Options	Description	Use Cases	Billing	Commitment	Flexibility
Spot Instance	Up to 90% discount Suitable for flexible, fault-tolerant workloads.	Applications with flexible start and end times, or applications	Per-second billing (min of 60 sec).	No commitment.	Most flexible, but can be interrupted.
Dedicated hosts	Physical servers dedicated for your use. Can help you meet compliance requirements	Workloads requiring a dedicated physical server, regulatory requirements.	Hourly, daily, or monthly.	1-year or 3-year commitment	Flexible, but limited to specific use cases.
Dedicated Instances	Instances that run on hardware dedicated to a single customer.	Workloads requiring isolation from other customers.	Hourly billing.	No commitment.	Flexible

08

Best Practices

Recommendations

Security Best Practices

1)

Use IAM roles to control access.

2)

Regularly update your AMIs and patches.

3)

Use security groups effectively.

Performance Best Practices

1)

Choose the right instance type.

2)

Use Auto Scaling.

3)

Monitor with Cloudwatch

Cost Best Practices

1)

Use Reserved and Spot Instances.

2)

Take advantage of Savings Plans.

3)

Regularly review and adjust resources.

THANKS!

Do you have any questions?

Mayamnaizel2013@gmail.com

CREDITS: This presentation template was created by Slidesgo, including icons by **Flaticon**, and infographics & images by **Freepik** and illustrations by **Storyset**