

Lecture 3: AWS Compute

Maharishi International University

Department of Computer Science

M.S. Thao Huy Vu

Maharishi International University - Fairfield, Iowa



All rights reserved. No part of this slide presentation may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying or recording, or by any information storage and retrieval system, without permission in writing from Maharishi International University (MIU).

Agenda

- Overview of AWS Compute Services
- Amazon EC2 (Elastic Compute Cloud)
- Security Group
- Amazon Lightsail

Overview of AWS Compute Services



Overview of AWS Compute Services

- Virtual Machine (EC2)
- Lightweight and Simplified servers (Lightsail)
- Serverless Computing (Lambda)
- Containers (ECS and EKS)

Overview of AWS Compute Services

- Cloud Compute: Servers, Storage, APIs
- Benefits
 - Auto-scaling
 - Pricing options
 - Integration with other services
 - Security: isolation, compliance certifications, built-in encryption
 - Performance metrics: CPU/GPU
 - Network enhancements

Overview of AWS Compute Services

- Choose the right Compute option:
 - **Computational Power:** Ensure the compute option meets your CPU, memory, storage, and GPU needs.
 - **Budget:** Align with cost constraints, considering pricing models like Spot Instances or serverless billing.
 - **Development Complexity:** Choose based on ease of development, deployment, and required expertise.
 - **Operational Control:** Select the level of control needed over the infrastructure, from full control (e.g., EC2) to minimal (e.g., Lambda).
 - **Scalability and Elasticity:** Assess how well the option handles scaling for fluctuating workloads.
 - **Workload Requirements:** Match the compute service to specific application needs, such as event-driven tasks, real-time processing, or batch jobs.

Amazon EC2 (Elastic Compute Cloud)

- **Virtual Private Server (VPS):** Provides fully customizable cloud-based virtual servers.
- **Network of Servers:** Build distributed systems or load-balanced networks with multiple EC2 instances.
- **Flexibility:** Wide range of instance types, operating systems, and configurations to suit various workloads.
- **Scalability:** Scale resources up or down dynamically with Auto Scaling and Load Balancing.
- **Cost-Effective:** Pay-as-you-go pricing with multiple cost-saving options (e.g., Spot Instances, Reserved Instances).

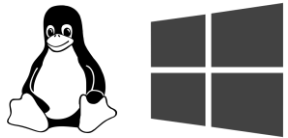
Computer and EC2 Instance



Computer



EC2 Instance



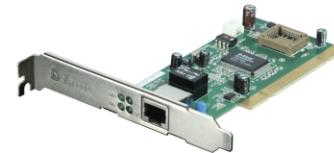
Operating System
AMIs
(Linux or Windows)



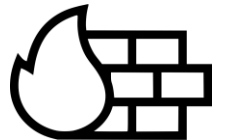
CPU & RAM
Instance Type



Hard Drive
EBS



Network Adapter
ENI



Firewall
Security Groups

Amazon EC2 (Elastic Compute Cloud)

- Pricing model
 - **On-Demand** Instances: Pay by the second, with no long-term commitments or upfront payments
 - **Reserved** Instances: Up to 75% discount compared to On-Demand, based on a one-year or three-year commitment
 - **Spot** Instances: Purchase unused EC2 capacity at significant discounts but with the possibility of termination by AWS if it needs the capacity back.
 - **Savings** Plans: Flexible pricing model that provides savings of up to 72% on specified usage in exchange for a commitment to a consistent amount of usage (measured in \$/hour) for a 1 or 3-year period.
 - Dedicated hosts: Dedicated physical servers.
 - Dedicated instances: instances running on dedicated hardware.

EC2 - Database

- **Use Managed Database Services (e.g., Amazon RDS):** These services offer instance types optimized for databases, providing automated backups, scaling, and maintenance, reducing operational overhead.
- **Use EC2 for Databases Only When Full Control is Needed:** Choose EC2 for hosting databases when you require custom configurations, non-standard database engines, or advanced tuning unavailable in managed services.

Amazon EC2 (Elastic Compute Cloud)

- Key features
 - Instance Types
 - Amazon Machine Images (AMIs)
 - Elastic IP Addresses
 - Security Group: Inbound, Outbound rules
 - Elastic Network Interface
 - Instance Store

Instance Types

- **Hardware Configuration:** Instance types define the underlying hardware, including CPU, GPU, RAM, network, and disk I/O performance.
- **Capabilities and Families:** Grouped into families optimized for compute, memory, storage, or graphics, catering to specific workloads.
- **GPU Instances:** NVIDIA GPU-accelerated instances are ideal for machine learning and parallel computing tasks.
- **Pricing Variations:** Costs vary by region; for example, instances may be priced higher in the USA compared to Japan.

EC2 instance type name

- *<family><generation>.<size>*
- Family
 - T: general purpose
 - P: Accelerated computing (GPU)
- Generation: a number represents the newer generations
- Size: Determines the vCPU count, memory or storage. E.g. **nano**, **micro**, **small**, large, xlarge...

General-purpose and memory-optimized instance type

- **General-purpose:**

- Balance of compute, memory, and networking resources.
- Use for Web servers and code repositories.
- Start with **T** and **M**: **t2.medium**; **m6i.large**

- **Memory-optimized**

- Process large amounts of data in memory such as caching.
- Start with **R** and **X**: **r5.large**; **x2gd.medium**

EC2 Pricing factors

- Instance Type and Size
- Pricing model
- Region
- OS
- Storage
- Data transfer: Inbound traffic is free, but outbound and inter-region transfer are charged

IP Addressing

An IP address is the EC2 instance address on the network.

Private IP Address:

- EC2 instance receives the private IP from the subnet.
- All EC2 instances (all devices in the network) have a private IP address.
- Private IP addresses allow instances to communicate with resources in the same network.
- No cost

Public IP address:

- All EC2 Instances can be launched with or without a public IP address.
- Public IP addresses are required for the instance to communicate with the Internet.
- It is dynamic as it is changed when restarting the instance
- No cost

Elastic IP

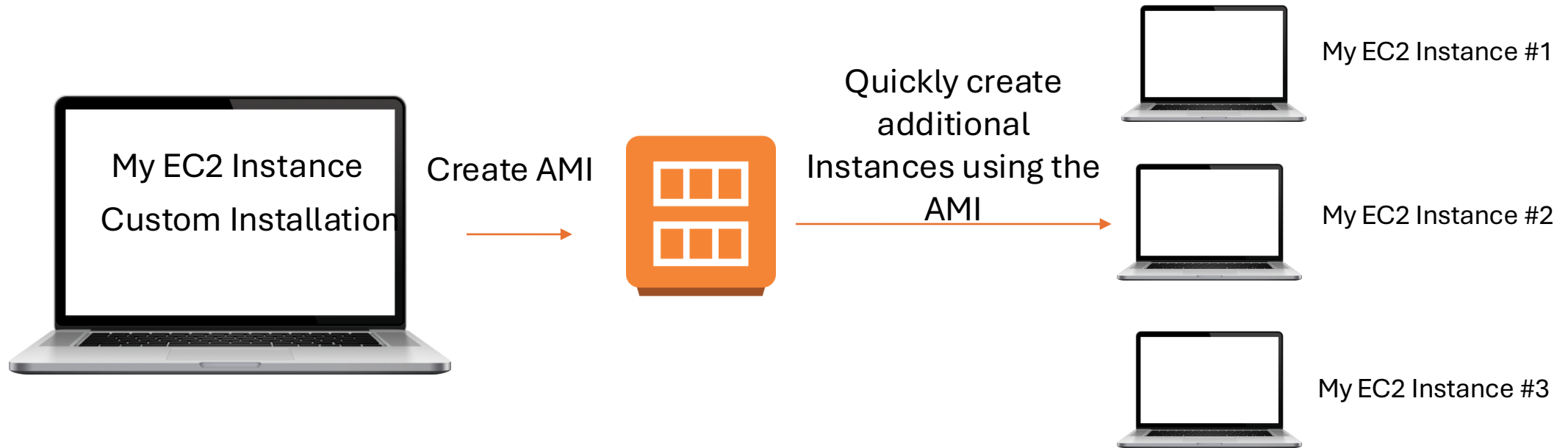
- **Static Public IPv4 Address:** A fixed IP address that remains the same across instance restarts.
- **Instance Attachment:** Can be attached to one instance at a time but reassigned as needed.
- **Regional Limit:** Up to 5 Elastic IPs per region by default.
- **Cost:** Priced at \$0.005 per hour (~\$4 per month). The first one is free if it is attached to a running instance.

Amazon Machine Image - AMI

- **Definition:** A predefined template containing the necessary information to launch an EC2 instance.
- **Reusable Image:** Captures the current state of an EC2 instance, allowing reinstallation or duplication across other instances.
- **Contents:**
 - **Operating System (OS):** The base system (e.g., Linux, Windows).
 - **Software Packages:** Pre-installed applications or dependencies.
 - **Settings:** Custom configurations, such as application code or environment setups.

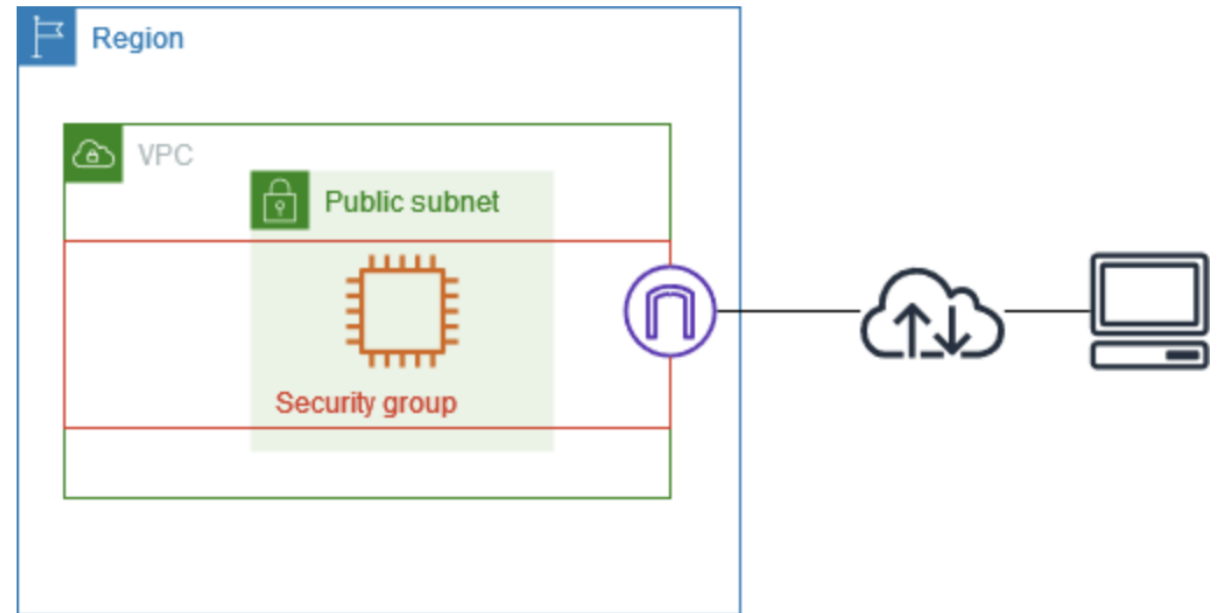


Amazon Machine Image - AMI



Security Groups

- **A virtual firewall** for EC2 instances to control inbound and outbound traffic
- **Inbound Rules:** Specify the type of incoming traffic to reach the instance
- **Outbound Rules:** Specify the type of outgoing traffic to leave the instance



Inbound/Outbound Rules

☐ Name Group ID Group Name VPC ID

☐ sg-b63da5fa default vpc-af9b48d5

Description Inbound Outbound Tags

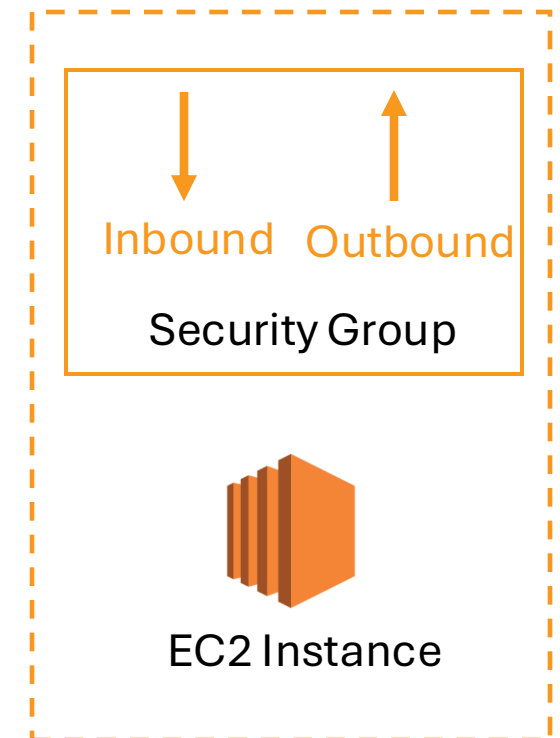
Edit

Type ⓘ	Protocol ⓘ	Port Range ⓘ	Source ⓘ	Description ⓘ
All traffic	All	All	your IP	

Description Inbound Outbound Tags

Edit

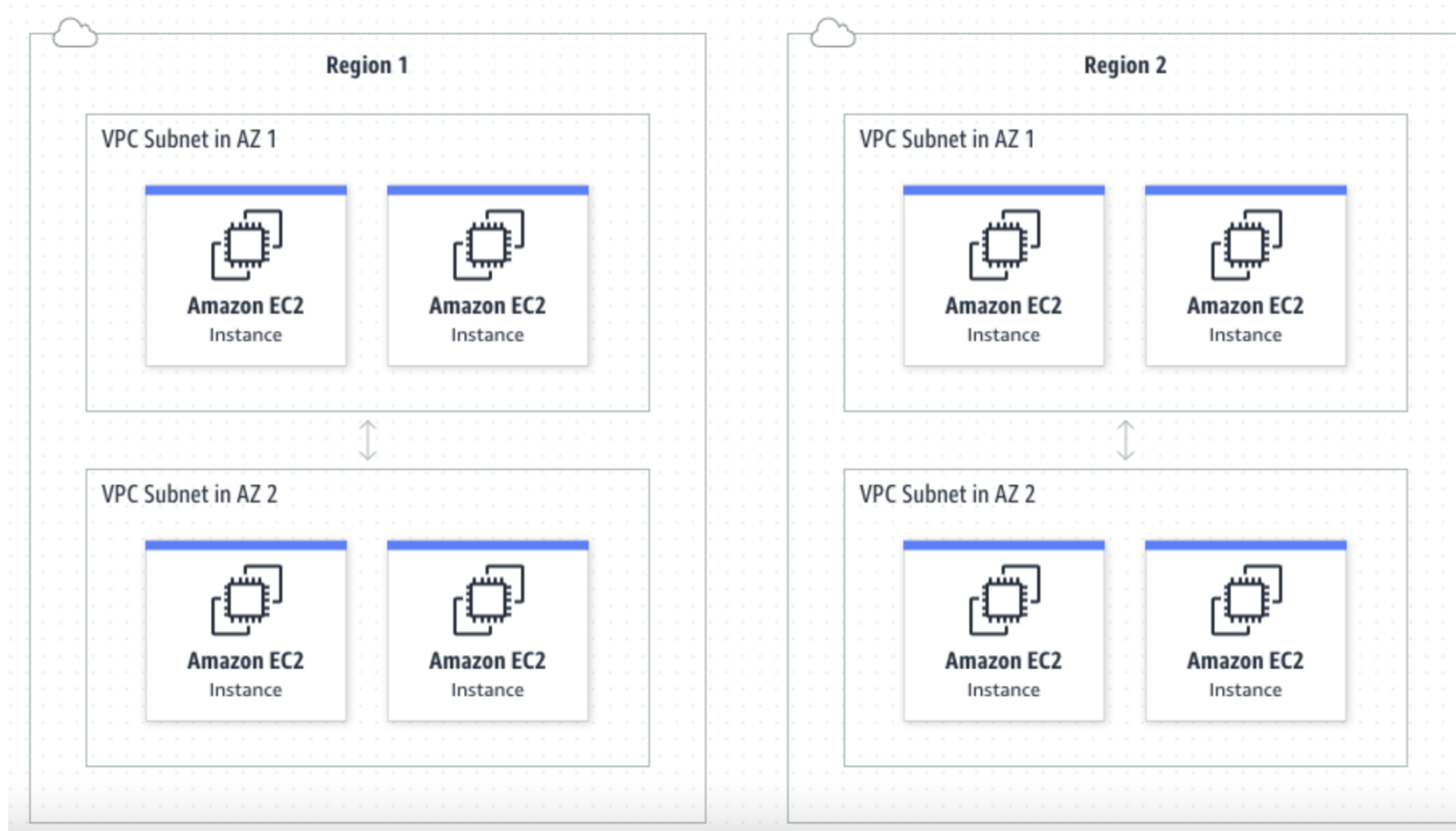
Type ⓘ	Protocol ⓘ	Port Range ⓘ	Destination ⓘ	Description ⓘ
All traffic	All	All	0.0.0.0/0	



Security group

- **Free of Charge:** There is no cost associated with creating or using Security Groups.
- **Rule Limitations:** Supports up to **120 total rules**, with a maximum of **60 rules per direction** (inbound and outbound).
- **Stateful Rules:** Security Groups are stateful, meaning if an inbound rule allows traffic, the response traffic is automatically allowed without an explicit outbound rule.
- **Applies at Instance Level:** Security Groups are attached to EC2 instances, and multiple instances can share the same Security Group.
- **Default Settings:** All inbound traffic is denied, and all outbound traffic is allowed.

Virtual Private Network (VPC)



Amazon EC2 (Elastic Compute Cloud)

- Demo

☰

[EC2](#) > [Instances](#) > Launch an instance

Launch an instance [Info](#)

Amazon EC2 allows you to create virtual machines, or instances, that run on the AWS Cloud. Quickly get started by following the simple steps below.

Name and tags [Info](#)

Name

[Add additional tags](#)

▼ Application and OS Images (Amazon Machine Image) [Info](#)

▼ Summary

Number of instances [Info](#)

[Software Image \(AMI\)](#)

Amazon Linux 2023 AMI 2023.4.2...[read more](#)
ami-0eb9d67c52f5c80e5

[Virtual server type \(instance type\)](#)

t2.micro

[Firewall \(security group\)](#)

New security group

[Storage \(volumes\)](#)

1 volume(s) - 8 GiB

Elastic Load Balancer (ELB)

- **Distributes** incoming web **traffic** (visitors to a web site) and equally across multiple EC2 instances running the same app.
- **Fault tolerance:** Seamlessly providing the required amount of load-balancing capacity needed to route application traffic.
- **Reliability:** Prevent one server from being overloaded while another server can handle more visitors.
- Handle **millions** of transactions in a second.
- **Types of Load Balancer:** Application Load Balancer, Network Load Balancer, or Classic Load Balancer
- **Configuration:** Health checks, SSL/TLS
- Integration with Auto Scaling Group

7 Layers of the OSI Model

Application

- End User layer
- HTTP, FTP, IRC, SSH, DNS

Presentation

- Syntax layer
- SSL, SSH, IMAP, FTP, MPEG, JPEG

Session

- Synch & send to port
- API's, Sockets, WinSock

Transport

- End-to-end connections
- TCP, UDP

Network

- Packets
- IP, ICMP, IPSec, IGMP

Data Link

- Frames
- Ethernet, PPP, Switch, Bridge

Physical

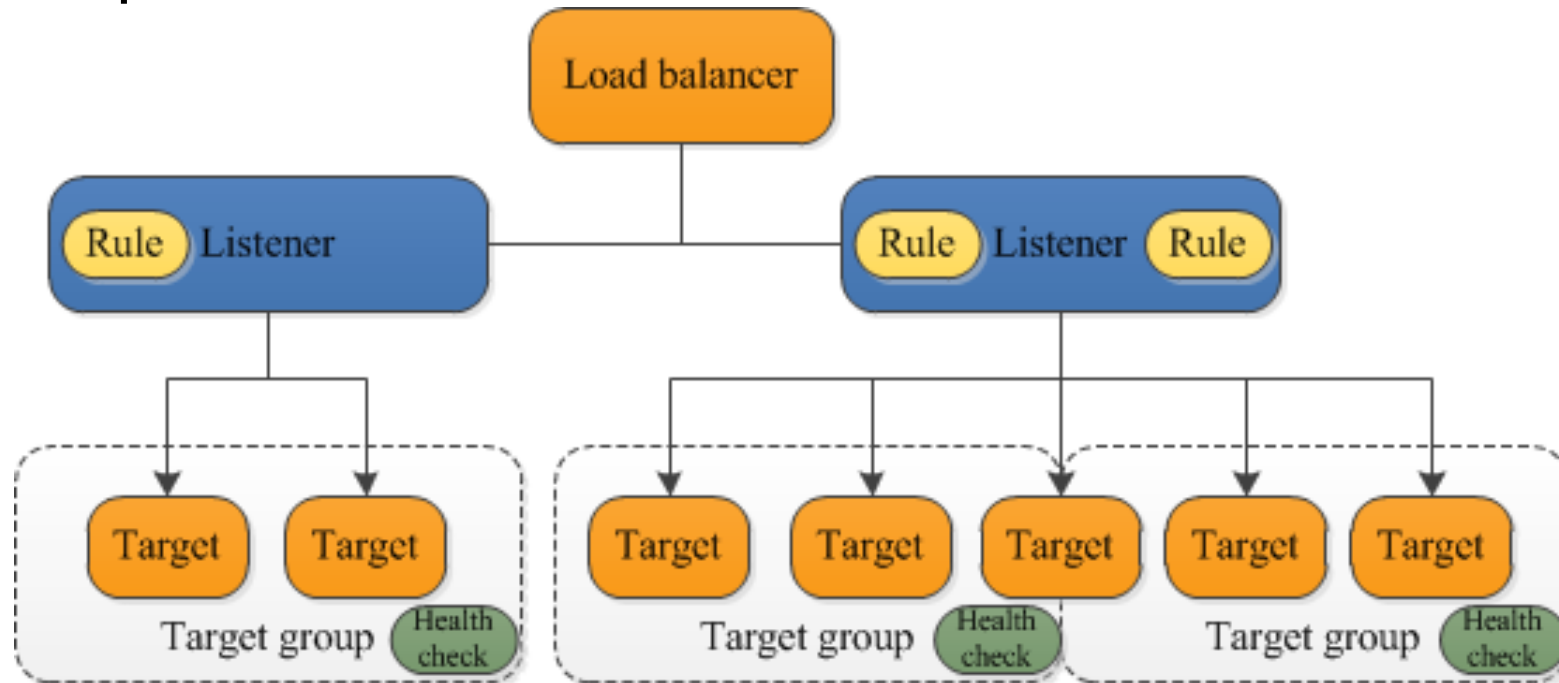
- Physical structure
- Coax, Fiber, Wireless, Hubs, Repeaters

Types of ELB

- **Application Load Balancer (ALB):**
 - Based on the content of the request (layer 7).
 - Support path-based, host-based, query string, parameter-based, and source IP-based routing.
 - Support instances, IP addresses, Lambda, and containers as targets.
- **Network Load Balancer (NLB)**
 - Based on IP protocol data (layer 4).
 - Provide ultra-high performance and low latency.
 - Support TCP/UDP and static IP addresses as targets.
- **Gateway Load Balancer:**
 - Operates at Layer 3 (Network).
 - Deploy, scale, and manage virtual appliances, such as firewalls, intrusion detection and prevention systems, and deep packet inspection systems.
- **Classic Load Balancer (CLB):** Old generation, not recommended for new apps. Performs routing at Layer 4 and Layer 7.

Application Load Balancer (ALB)

- A single point to distribute traffic across multiple EC2 in multiple AZ in single region.
- Can have up to 50 listeners.



ALB Listener

- Check connections based on protocol and port (HTTP:80, HTTPS:443). HTTPS requires certificates. Use ACM to store certificates.
- Have up to 100 rules

The screenshot displays the AWS Management Console interface for configuring an ALB listener. At the top, there's a 'Create Load Balancer' button and an 'Actions' dropdown. Below this is a search bar and a table listing load balancers. The table has columns for Name, DNS name, State, VPC ID, Availability Zones, Type, and Created At. One load balancer, 'my-alb', is listed with a state of 'provisioning'. Below the table, the 'Load balancer: my-alb' section is active, showing tabs for Description, Listeners, Monitoring, Integrated services, and Tags. The 'Listeners' tab is selected, showing a table with columns for Listener ID, Security policy, SSL Certificate, and Rules. A single listener is listed with the protocol 'HTTP : 80' and a default rule forwarding to 'my-tg'. Buttons for 'Add listener', 'Edit', and 'Delete' are visible above the listener table.

Name	DNS name	State	VPC ID	Availability Zones	Type	Created At
my-alb	my-alb-967328916.us-east-1...	provisioning	vpc-063dae80fe38de125	us-east-1b, us-east-1c	application	May 23, 2021 at 1:03:47 PM ...

Load balancer: my-alb

Description | **Listeners** | Monitoring | Integrated services | Tags

A listener checks for connection requests using its configured protocol and port, and the load balancer uses the listener rules to route requests to targets. You can add, remove, or update listeners and listener rules.

[Add listener](#) [Edit](#) [Delete](#)

Listener ID	Security policy	SSL Certificate	Rules
<input type="checkbox"/> HTTP : 80 arn...bfe3f9eb5b0c8ddd	N/A	N/A	Default: forwarding to my-tg View/edit rules

ALB Listener Rule

- Determine how ALB routes the request to its targets based on priority, conditions, actions.

Listener rules (3) [Info](#)

[Rule limits](#)[Actions ▼](#)[Add rule](#)

Traffic received by the listener is routed according to the default action and any additional rules. Rules are evaluated in priority order from the lowest value to the highest value.



<input type="checkbox"/>	Name tag	Priority ▲	Conditions (If)	Actions (Then)	ARN	Tags
<input type="checkbox"/>	finance-app	1	Path Pattern is /finance	Redirect to HTTPS://myfinance.com:443/? <ul style="list-style-type: none">Status code: HTTP_301	ARN	1 tag
<input type="checkbox"/>	hr-app	2	Path Pattern is /hr	Forward to target group <ul style="list-style-type: none">my-tg : 1 (100%)Target group stickiness: Off	ARN	1 tag
<input type="checkbox"/>	Default	Last (default)	If no other rule applies	Forward to target group <ul style="list-style-type: none">my-tg : 1 (100%)Target group stickiness: Off	ARN	0 tags

ALB Listener Rule - Conditions

- Host Header: Domain of the request. E.g. host=api.example.com
- Path Pattern: URL path. E.g. /api/*; /images/*
- HTTP Header: e.g. User-Agent = mobile
- HTTP Request Method: POST, GET, PATCH...
- Query String: e.g. ?language=en
- Source IP

Edit condition [Rule limits](#)

Rule condition types

Route traffic based on the condition type of each http-request-method and source-ip. Each rule can

Path ▲

Include one of each

Host header

Path ✓

HTTP request method

Source IP

Include one or more of each

HTTP header

Query string

ALB Listener Actions

- Forward: Send traffic to one or more target groups.
- Redirect: Redirect the request to another URL or protocol.
- Fixed Response: Respond with a fixed HTTP status code and optional message or body.

Actions

Action types

Routing actions

☒ Forward to target groups

☐ Redirect to URL

☐ Return fixed response

Forward to target group [Info](#)

Choose a target group and specify routing weight or [Create target group](#).

Target group

Select a target group ▼

[Add target group](#)

You can add up to 4 more target groups.



Weight

1

0-999

Percent

100%


ALB Target Groups

- A logical grouping of targets (e.g. EC2, ECS tasks, Lambda) that receive traffic from the ALB based on rules.
- Key features:
 - **Targets:** EC2 instances, ECS tasks, or IP addresses... that receive traffic.
 - **Health Checks:** ALB monitors target health and stops traffic to unhealthy targets, ensuring reliability.
 - **Routing:** Enables traffic routing based on listener rules, such as path patterns, hostnames, or query parameters.
 - **Port Configuration:** Each target group maps to a specific port on the targets.

Target Group

Choose a target type

☒ Instances

- Supports load balancing to instances within a specific VPC.
- Facilitates the use of [Amazon EC2 Auto Scaling](#)  to manage and scale your EC2 capacity.

☐ IP addresses

- Supports load balancing to VPC and on-premises resources.
- Facilitates routing to multiple IP addresses and network interfaces on the same instance.
- Offers flexibility with microservice based architectures, simplifying inter-application communication.
- Supports IPv6 targets, enabling end-to-end IPv6 communication, and IPv4-to-IPv6 NAT.

☐ Lambda function

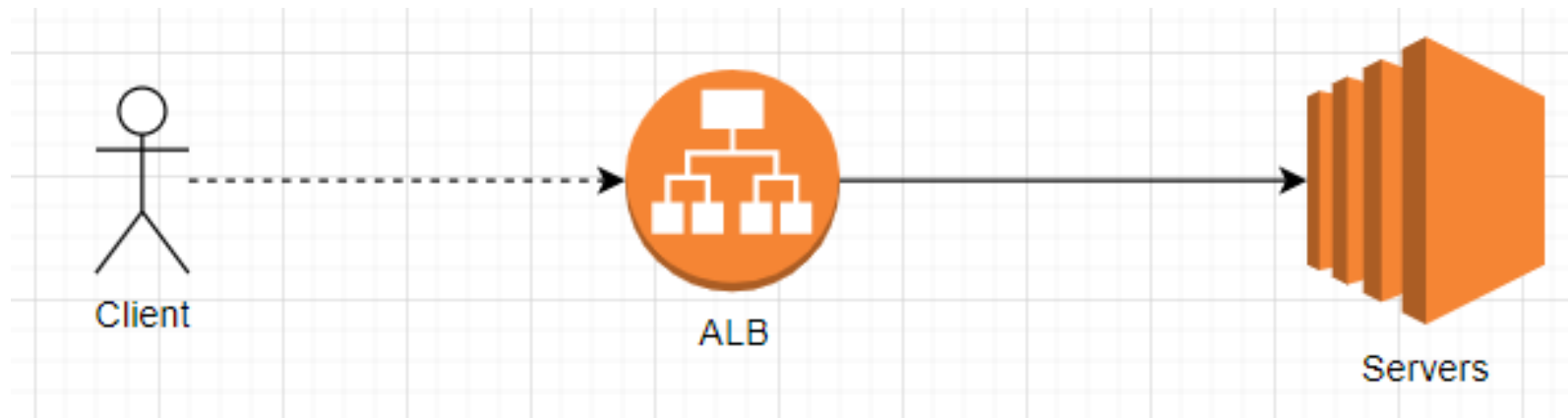
- Facilitates routing to a single Lambda function.
- Accessible to Application Load Balancers only.

☐ Application Load Balancer

- Offers the flexibility for a Network Load Balancer to accept and route TCP requests within a specific VPC.
- Facilitates using static IP addresses and PrivateLink with an Application Load Balancer.

ALB Target Groups

- **Routing:** A target group forwards requests to one or more registered targets (e.g., EC2 instances) based on the specified protocol and port.
- **Two Connections:**
 - **Client to ALB:** The first connection is from the client to the Application Load Balancer.
 - **ALB to Target:** The second connection is from the ALB to the registered targets (servers) in the target group.



ALB Target Groups

- **Resource Management:**

- Target groups typically consist of **multiple resources** running the same application, often provisioned by **Auto Scaling Groups (ASG)** to dynamically scale based on traffic.

- **Health Checks:**

- Performed on all targets within a target group to ensure availability and reliability.

- **Association with ALB:**

- Resources are associated with an ALB through their target group.
- A target group can initially have no resources but must be configured with the appropriate target type (**IP**, **instance**, or **Lambda**).

ALB – How it works

- **Listener Matching:** ALB determines the appropriate listener based on:
 - Protocol: HTTP, HTTPS, or TCP.
 - Port: The port configured for the listener.
- **Rule Evaluation (Within the Matched Listener):**
 - The ALB evaluates the rules within the matched listener in **priority order** (lowest to highest).
 - The **first rule** that matches the conditions is applied.
 - If no rules match, the listener's **default rule** is used.

ALB Example

- A website (`example.com`) for users.
- An API (`api.example.com`) with two versions (v1 and v2) for developers.
- A secure admin interface (`admin.example.com`).

ALB - Example


ALB


- └ Listener 80 (HTTP)
 - └ Redirect to HTTPS (Port 443)
- └ Listener 443 (HTTPS)
 - └ Rule 1: Host = api.example.com AND Path = /v1/* → Target Group: API-v1
 - └ Rule 2: Host = api.example.com AND Path = /v2/* → Target Group: API-v2
 - └ Rule 3: Host = admin.example.com → Target Group: Admin
 - └ Rule 4: Host = example.com → Target Group: Website
 - └ Default Rule → Return 404

Amazon Lightsail

- Pre-configured environments for fast deployment such as WordPress, e-commerce, web apps, development stacks, and more.
- Flat-rate monthly pricing, starting with affordable plans.
- Easily upgrade to more robust AWS services if needed such as migrating to EC2.
- e.g:
 - Hosting websites and blogs.
 - Testing environment

Amazon Lightsail


**Linux/Unix**
16 blueprints


**Microsoft Windows**
3 blueprints


Select a blueprint


Apps + OS


OS Only


**WordPress**
4.9.6


**LAMP Stack**
5.6.36


**Node.js**
10.1.0


**Joomla**
3.8.8


**Magento**
2.2.4


**MEAN**
3.6.5

**Drupal**
8.5.3-1

**GitLab CE**
10.8.1

**Redmine**
3.4.5

**Nginx**
1.14.0

**Plesk Hosting Stack on Ubuntu**
17.8.11

Amazon Lightsail

Select a size

Sort by

Price per month ▼



\$5

USD per month

512 MB Memory
2 vCPUs Processing
20 GB SSD Storage
1 TB Transfer

First 90 days free



\$7

USD per month

1 GB Memory
2 vCPUs Processing
40 GB SSD Storage
2 TB Transfer

First 90 days free



\$12

USD per month

2 GB Memory
2 vCPUs Processing
60 GB SSD Storage
3 TB Transfer

First 90 days free



\$24

USD per month

4 GB Memory
2 vCPUs Processing
80 GB SSD Storage
4 TB Transfer



\$44

USD per month

8 GB Memory
2 vCPUs Processing
160 GB SSD Storage
5 TB Transfer



\$84

USD per month

16 GB Memory
4 vCPUs Processing
320 GB SSD Storage
6 TB Transfer



\$164

USD per month

32 GB Memory
8 vCPUs Processing
640 GB SSD Storage
7 TB Transfer



\$384

USD per month

64 GB Memory
16 vCPUs Processing
1,280 GB SSD Storage
8 TB Transfer

Largest plan

LightSail vs. EC2

Feature	AWS Lightsail	Amazon EC2
Performance	Good for small to medium workloads.	Scalable to handle small to enterprise-grade workloads.
Pricing Model	Flat, predictable monthly pricing.	Pay-as-you-go with granular billing based on usage.
Customization	Limited customization.	Full control over configurations, networking, and storage.
Integration	Limited AWS service integration.	Seamlessly integrates with all AWS services.
Use Cases	Small websites, blogs, simple applications.	Enterprise applications, large-scale databases, machine learning, and more.
Scaling	Manual vertical scaling only.	Supports both vertical and horizontal scaling with Auto Scaling.

References

- <https://docs.aws.amazon.com/>
- ChatGPT: <https://chatgpt.com/>
- Google AI: <https://gemini.google.com/app>
- <https://dev.to/aws-builders/all-you-need-to-know-about-aws-compute-services-17gf>