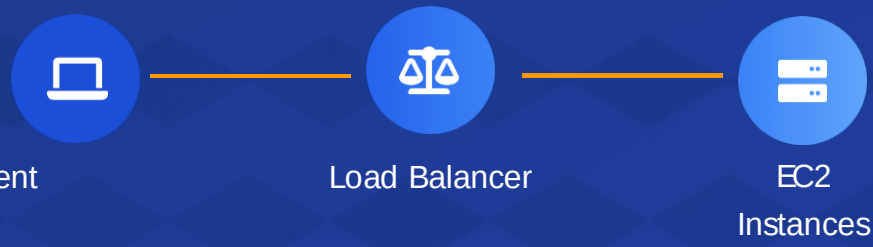


Elastic Load Balancer

(ELB)

AWS Managed

Scalable

High Availability

Dịch vụ cân bằng tải được quản lý hoàn toàn bởi AWS

Phân phối lưu lượng truy cập đến nhiều đích một cách hiệu quả

Hỗ trợ khả năng mở rộng và tính sẵn sàng cao

Scalability & High Availability

Khái niệm

Scalability: Khả năng của ứng dụng/hệ thống có thể xử lý tải lớn hơn bằng cách thích nghi.

High Availability: Đảm bảo ứng dụng/hệ thống luôn hoạt động và có thể truy cập được, thường bằng cách loại bỏ các điểm lỗi đơn lẻ.

Loại Scalability

Vertical Scalability (Scale Up/Down)

Tăng hoặc giảm kích thước của một instance.
Phổ biến cho hệ thống không phân tán, ví dụ: cơ sở dữ liệu.
Dịch vụ như Amazon RDS, ElastiCache có thể mở rộng theo chiều dọc.

Horizontal Scalability (Scale Out/In)

Tăng hoặc giảm số lượng instance/hệ thống cho ứng dụng.
Áp dụng cho hệ thống phân tán.
Phổ biến cho ứng dụng web/ứng dụng hiện đại.

High Availability

Chạy ứng dụng trên ít nhất 2 trung tâm dữ liệu (Availability Zones)

Mục tiêu: Survive sau sự cố mất một trung tâm dữ liệu

HA có thể là passive (RDS Multi-AZ) hoặc active (Horizontal Scaling)

Vertical vs Horizontal Scalability

Vertical Scalability

Scale Up/Down

Tăng hoặc giảm kích thước của một instance.

Scale Up
Tăng kích thước instance

Scale Down
Giảm kích thước instance

Ứng dụng cho:

Hệ thống không phân tán
Cơ sở dữ liệu

Ví dụ:

Amazon RDS
ElastiCache

Horizontal Scalability

Scale Out/In

Tăng hoặc giảm số lượng instance/hệ thống cho ứng dụng.

Scale Out
Tăng số lượng instance

Scale In
Giảm số lượng instance

Ứng dụng cho:

Hệ thống phân tán
Ứng dụng web/ứng dụng hiện đại

Ví dụ:

Amazon EC2
Amazon ECS

Tính năng	Vertical Scalability	Horizontal Scalability
Mô hình	Thang đo theo chiều dọc	Thang đo theo chiều ngang
Khả năng mở rộng	Giới hạn bởi hardware	Không giới hạn (nhiều server)

High Availability cho EC2

Các phương pháp triển khai

Vertical Scaling

Tăng kích thước instance (scale up/down)

Ví dụ: từ t2.nano lên u-12tb1.metal
RAM: 0.5G → 12.3 TB, vCPU: 1 → 448

Horizontal Scaling

Tăng số lượng instance (scale out/in)

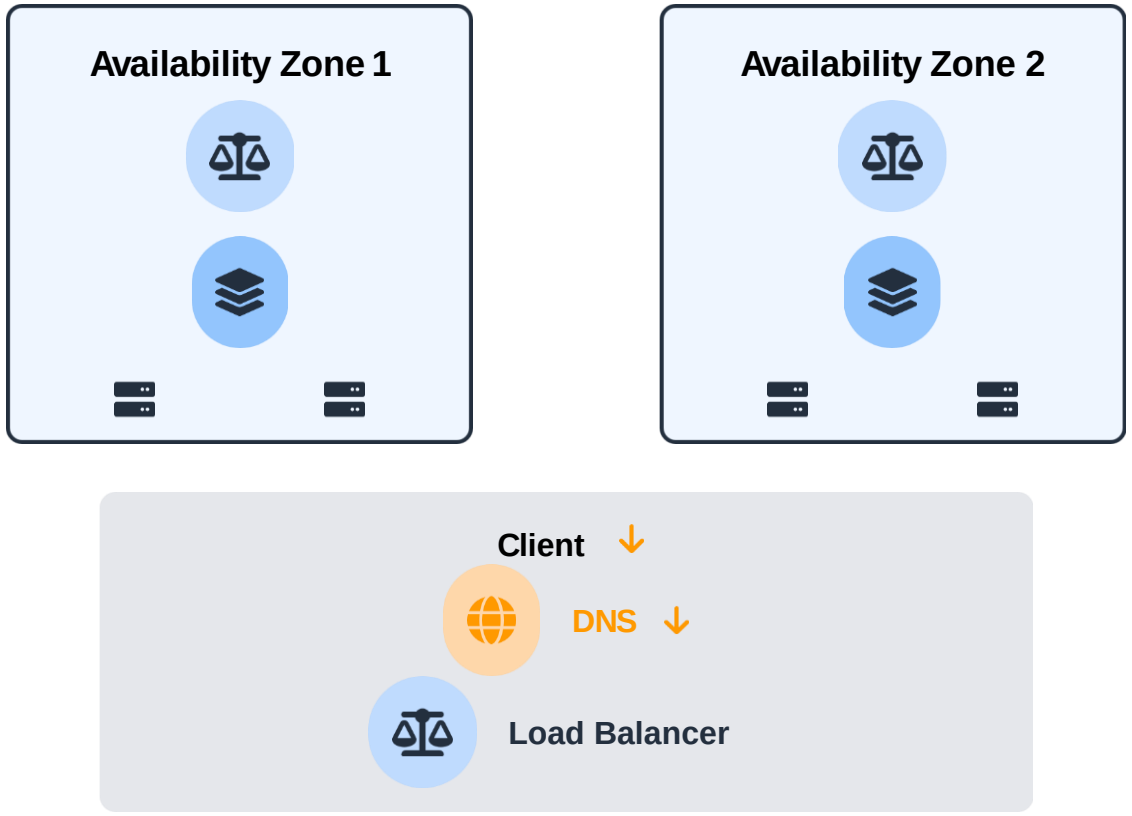
Sử dụng Auto Scaling Group

High Availability

Chạy các instance trên nhiều Availability Zones

Auto Scaling Group multi AZ
Load Balancer multi AZ

Kiến trúc High Availability



- Tăng tính sẵn sàng
- Phân phối tải đều
- Khắc phục lỗi đơn lẻ

Load Balancing là gì?

Khái niệm

Load Balancing là quá trình phân phối lưu lượng truy cập mạng hoặc ứng dụng trên nhiều máy chủ (ví dụ: EC2 instances) để đảm bảo hiệu suất và độ tin cậy tối ưu.

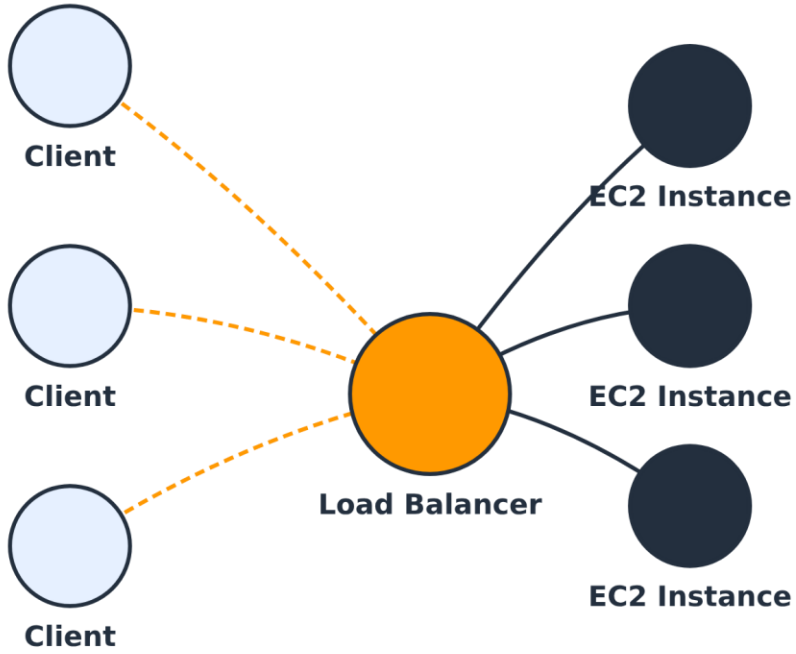
Lợi ích

- **Phân tán tải**: Chia đều lưu lượng truy cập giữa các instance hạ nguồn
- **Điểm truy cập duy nhất (DNS)**: Cung cấp một tên miền duy nhất cho ứng dụng
- **Xử lý lỗi liền mạch**: Tự động chuyển hướng lưu lượng khỏi các instance bị lỗi
- **Health Checks định kỳ**: Thường xuyên kiểm tra tình trạng của các instance

Các tính năng khác

- **SSL termination** (HTTPS)
- **Enforce stickiness** (cookies)
- **High availability** across zones
- **Tách biệt** public/private traffic

Cách Load Balancing hoạt động



Load Balancer đóng vai trò then chốt trong việc xây dựng các ứng dụng có khả năng mở rộng và tính sẵn sàng cao trên AWS.

Tại sao sử dụng Elastic Load Balancer?

Elastic Load Balancer là dịch vụ được quản lý hoàn toàn bởi AWS, mang lại nhiều lợi ích đáng kể cho các ứng dụng.

Managed Load Balancer

- AWS đảm bảo ELB luôn hoạt động ổn định
- Chịu trách nhiệm nâng cấp, bảo trì
- Cung cấp tính sẵn sàng cao

Tích hợp sâu rộng

- EC2, EC2 Auto Scaling Groups, Amazon ECS
- AWS Certificate Manager (ACM)
- CloudWatch cho giám sát
- Route 53, AWS WAF, AWS Global Accelerator

High Availability

- Phân phối lưu lượng truy cập đến nhiều AZ
- Tự động chuyển hướng lưu lượng khỏi các instance bị lỗi
- Đảm bảo ứng dụng luôn hoạt động

Hiệu suất

- Phân tán tải đều giữa các instance hạ nguồn
- Giải mã HTTPS tại Load Balancer
- Hỗ trợ HTTP/2 và WebSocket

ELB là dịch vụ quản lý của AWS, đóng vai trò then chốt trong việc xây dựng ứng dụng có khả năng mở rộng

Health Checks

Health Checks là gì?

Health Checks là tính năng thiết yếu của Load Balancer, giúp đảm bảo rằng lưu lượng truy cập chỉ được gửi đến các instance đang hoạt động bình thường.

Cho phép Load Balancer biết liệu các instance có sẵn sàng để phản hồi yêu cầu hay không.

Cách thức hoạt động

Kiểm tra trên cổng và đường dẫn cụ thể

Ví dụ: Kiểm tra được thực hiện trên cổng 80 và đường dẫn /health

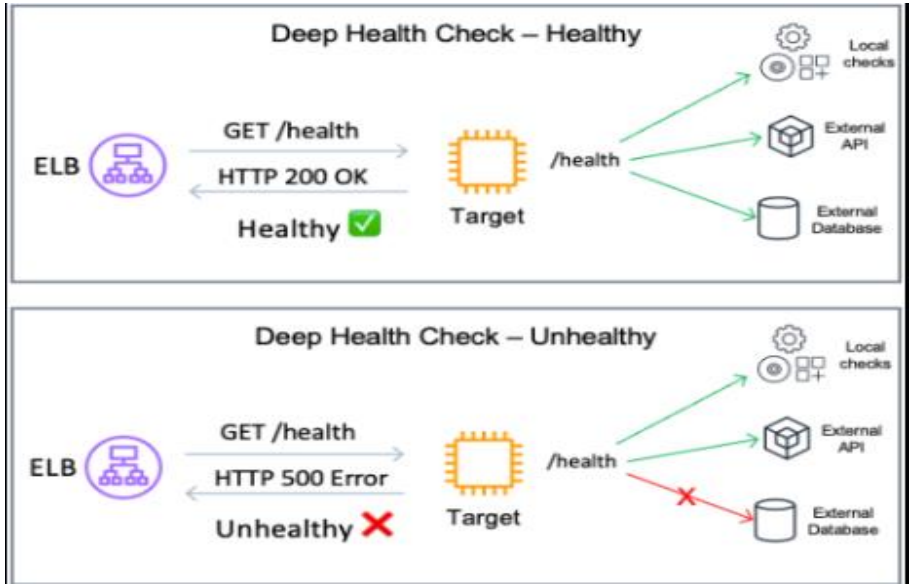
Xác định trạng thái

Nếu phản hồi không phải là mã trạng thái 200 (OK), instance đó sẽ được đánh dấu là unhealthy

Hành động của Load Balancer

Load Balancer sẽ ngừng gửi lưu lượng đến instance không khỏe mạnh

Minh họa Health Checks



Target Group

Health Check Response

Load Balancer Security Groups

Security Groups

Security Groups đóng vai trò quan trọng trong việc kiểm soát lưu lượng truy cập đến và đi từ Load Balancer và các instance backend.

Load Balancer Security Group

- Cho phép lưu lượng truy cập từ bên ngoài (ví dụ: HTTP/HTTPS) đến Load Balancer
- Thường mở cổng 80 (HTTP) và 443 (HTTPS)

Application Security Group

- Chỉ cho phép lưu lượng truy cập từ Security Group của Load Balancer
- Ngăn chặn truy cập trực tiếp từ bên ngoài vào các instance backend
- Thêm bảo mật cho ứng dụng của bạn

Lợi ích

- Kiểm soát chặt chẽ luồng traffic vào và ra khỏi Load Balancer cũng như các instance
- Thêm lớp bảo mật bổ sung cho ứng dụng của bạn
- Đảm bảo chỉ có traffic hợp lệ được phép đi qua hệ thống

Cấu hình Security Groups

Load Balancer SG

- Mở cổng 80 (HTTP)
- Mở cổng 443 (HTTPS)
- Cho phép traffic từ client

SG của Load Balancer

Application SG

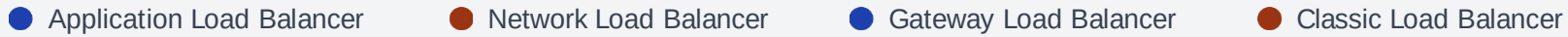
- Cho phép traffic từ LB SG
- Ngăn chặn truy cập trực tiếp

SG của Application

Các loại Load Balancer trên AWS

AWS cung cấp nhiều loại Elastic Load Balancer, mỗi loại được thiết kế để đáp ứng các nhu cầu cụ thể của ứng dụng.

Tính năng	Application Load Balancer (ALB)	Network Load Balancer (NLB)	Gateway Load Balancer (GWLB)	Classic Load Balancer (CLB)
Loại Load Balancer	Layer 7	Layer 4	Layer 3 Gateway + Layer 4	Layer 4/7
Target type	IP, Instance, Lambda	IP, Instance, Application Load Balancer	IP, Instance	Instance
Protocol listeners	HTTP, HTTPS, gRPC	TCP, UDP, TLS	IP	TCP, SSL/TLS, HTTP, HTTPS

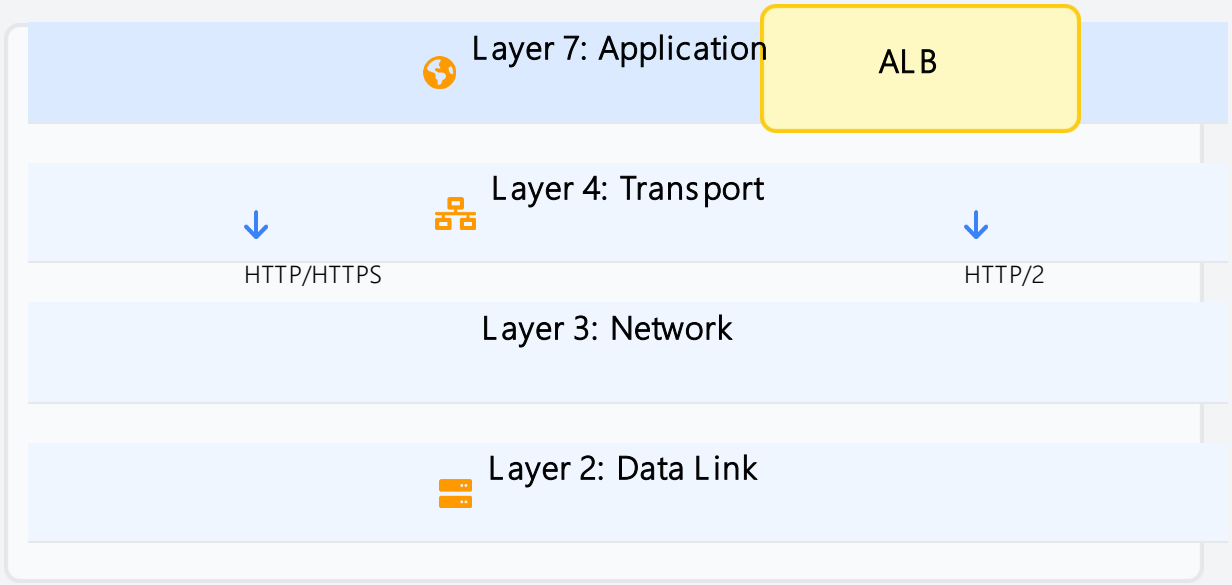


Application Load Balancer (ALB)

Giới thiệu

Application Load Balancer hoạt động ở Layer 7 (HTTP/HTTPS) của mô hình OSI, chuyên xử lý các yêu cầu HTTP/HTTPS.

Lựa chọn lý tưởng cho các ứng dụng web hiện đại và kiến trúc microservices.



Tính năng nổi bật

Cân bằng tải cho nhiều ứng dụng HTTP
Định tuyến đến nhiều Target Group khác nhau

Hỗ trợ HTTP/2 và WebSocket
Kết nối hai chiều hiệu quả

Chuyển hướng (redirects)
Ví dụ: từ HTTP sang HTTPS

Định tuyến nâng cao
Dựa trên Path, Hostname, Query String, Headers

Advanced Routing

Routing dựa trên Path
example.com/users & example.com/posts

Routing dựa trên Hostname
one.example.com & other.example.com

Routing dựa trên Query String, Headers
example.com/users?id=123&order=false

ALB - Advanced Routing

Định tuyến nâng cao

ALB cung cấp khả năng định tuyến nâng cao, cho phép bạn điều hướng lưu lượng truy cập đến các Target Group khác nhau dựa trên nội dung của yêu cầu.

Phù hợp với:

Microservices và ứng dụng dựa trên container (ví dụ: Docker & Amazon ECS)

Tính năng:

- Ánh xạ cổng (port mapping) để chuyển hướng đến một cổng động trong ECS
- Định tuyến dựa trên nhiều tiêu chí khác nhau
- Lincoln giữa các service khác nhau

ALB hoạt động ở Layer 7 (HTTP/HTTPS) nên có thể đọc nội dung yêu cầu để thực hiện định tuyến.

Routing dựa trên Path

Ví dụ:

example.com /users → Target Group 1

example.com /posts → Target Group 2

Path có thể chứa ký tự đại diện như /api/v* để match nhiều path.

Routing dựa trên Hostname

Ví dụ:

one.example.com → Target Group 1

other.example.com → Target Group 2

Useful cho multi-tenant applications hoặc các dịch vụ khác nhau.

Routing dựa trên Query String

Ví dụ:

example.com/users ?id=123 → Target Group 1

example.com/users ?order=false → Target Group 2

Query string là phần optional của URL sau dấu ?.

Routing dựa trên Headers

Ví dụ:

Accept: text/html → Target Group 1

User-Agent: Mobile → Target Group 2

Headers là các trường tùy chọn trong yêu cầu HTTP.

ALB - Target Groups

Target Group là nơi bạn đăng ký các tài nguyên mà Load Balancer sẽ gửi lưu lượng truy cập đến. ALB có thể định tuyến đến nhiều Target Group khác nhau.

EC2 instances

Các phiên bản EC2 (có thể được quản lý bởi Auto Scaling Group)

- Thường dùng cho HTTP
- Có thể tự động điều chỉnh số lượng instance

ECS tasks

Các tác vụ ECS (được quản lý bởi ECS)

- Thường dùng cho HTTP
- Tốt cho ứng dụng container-based

Lambda functions

Yêu cầu HTTP được dịch thành sự kiện JSON

- Serverless computing
- Tự động scaling theo nhu cầu

IP Addresses

Phải là các IP riêng tư (private IPs)

- Linh hoạt trong việc mục tiêu
- Dùng cho các dịch vụ không phải EC2

Lưu ý: Health checks được thực hiện ở cấp Target Group, không phải ở cấp Load Balancer.

Network Load Balancer (NLB)

Đặc điểm chính

Layer 4 (TCP & UDP)
Hoạt động ở tầng vận chuyển của mô hình OSI, chuyển tiếp lưu lượng TCP & UDP.

Hiệu suất cực cao
Xử lý hàng triệu yêu cầu mỗi giây.

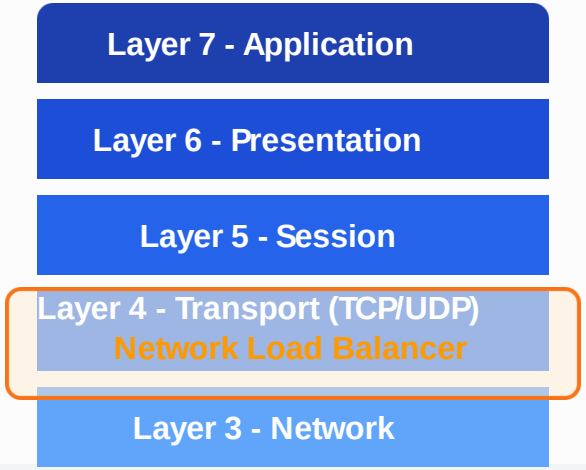
Độ trễ thấp
~100 ms so với ~400 ms của ALB.

IP & Elastic IP

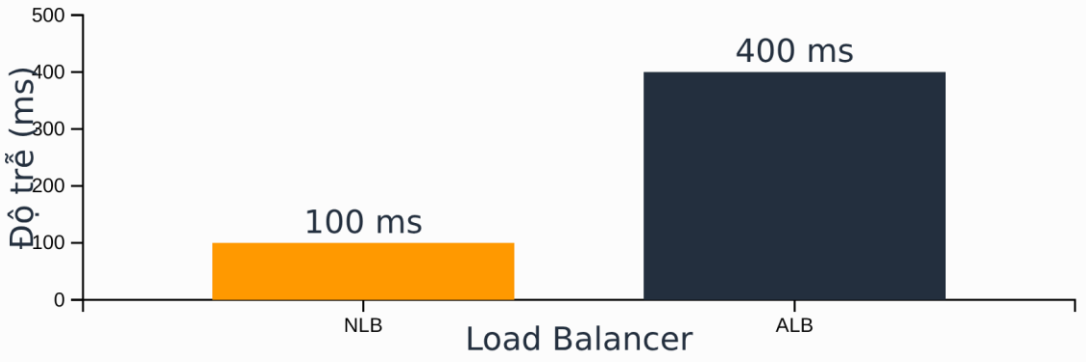
IP tĩnh trên mỗi AZ
Có một IP tĩnh trên mỗi Availability Zone.

Hỗ trợ Elastic IP
Hữu ích cho việc whitelisting IP cụ thể.

Hoạt động ở Layer 4



So sánh độ trễ



Gateway Load Balancer (GWLB)

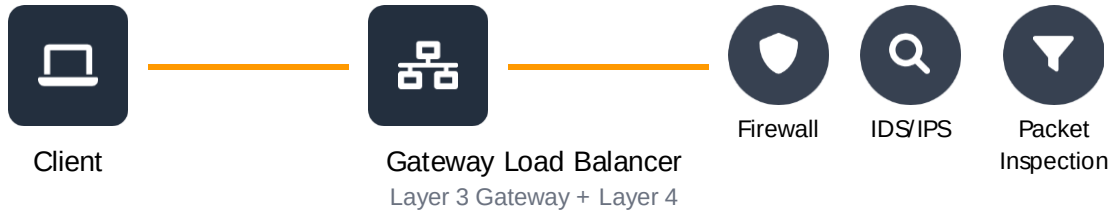
Giới thiệu

Gateway Load Balancer là một loại Load Balancer chuyên biệt, hoạt động ở Layer 3 (lớp mạng), được thiết kế để triển khai, mở rộng và quản lý các thiết bị mạng ảo của bên thứ ba.

Mục đích

- Triển khai, mở rộng và quản lý một nhóm các thiết bị mạng ảo của bên thứ ba trong AWS
- Hỗ trợ nhiều giải pháp bảo mật mạng:
  - Firewalls
  - Intrusion Detection and Prevention Systems (IDS/IPS)
  - Deep Packet Inspection Systems
  - Thao tác payload

Cách hoạt động



Hoạt động

GWLB hoạt động ở Layer 3 (lớp mạng) – xử lý các gói IP.

Kết hợp các chức năng:

- Transparent Network Gateway: Điểm vào/ra duy nhất cho tất cả lưu lượng
- Load Balancer: Phân phối lưu lượng đến các thiết bị ảo của bạn
- Giao thức: Sử dụng GENEVE trên cổng 6081

Target Groups Hỗ trợ

Sticky Sessions (Session Affinity)

Khái niệm

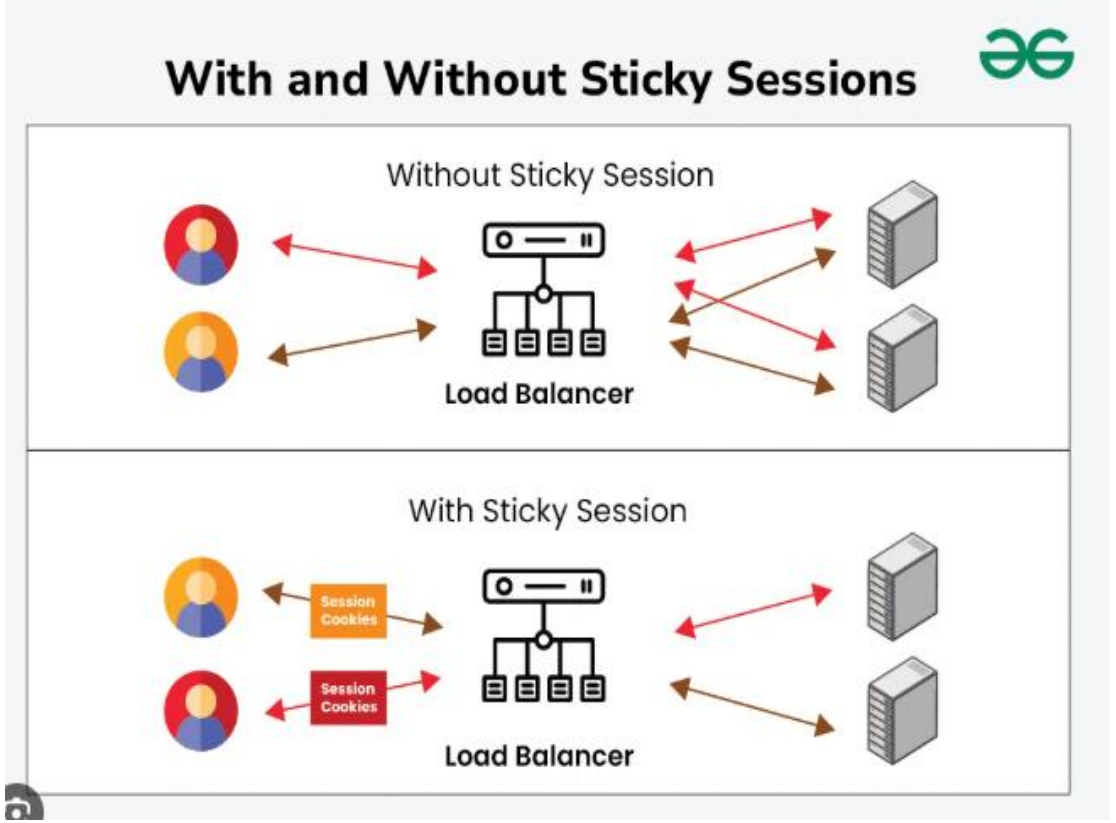
Sticky Sessions (Session Affinity) là tính năng cho phép Load Balancer định tuyến các yêu cầu liên tiếp từ cùng một client đến cùng một instance backend.

Điều này đảm bảo rằng trạng thái phiên của người dùng được duy trì, đặc biệt quan trọng đối với các ứng dụng yêu cầu duy trì trạng thái.

Cách hoạt động

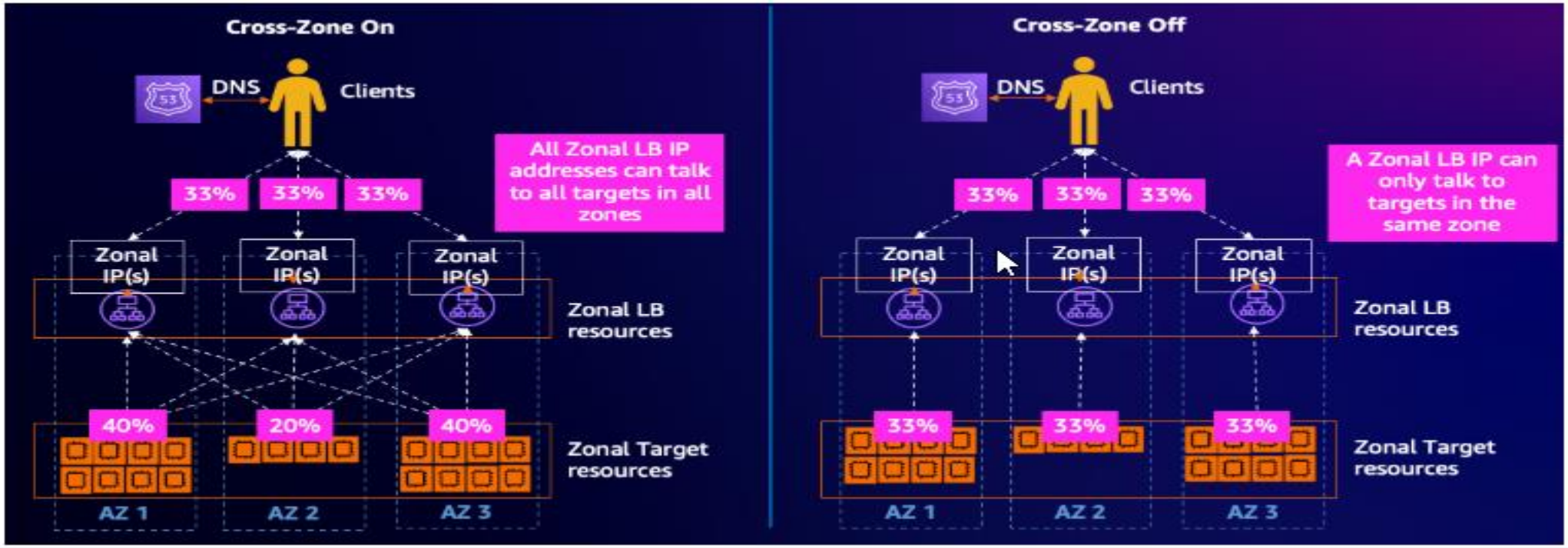
- Sử dụng cookie để theo dõi phiên của client.
- Đối với Classic Load Balancer (CLB) và Application Load Balancer (ALB), cookie có thời gian hết hạn có thể cấu hình.
- Việc bật Sticky Sessions có thể dẫn đến sự mất cân bằng tải giữa các instance backend.

Cơ chế hoạt động



Cross-Zone Load Balancing

Cross-Zone Load Balancing là một tính năng giúp phân phối lưu lượng truy cập đồng đều hơn giữa các Availability Zone (AZ) và các instance backend.



Lợi ích:
Đảm bảo phân phối tải đồng đều hơn, ngay cả khi số lượng instance không đều giữa các AZ.

Rủi ro:
Có thể dẫn đến tình trạng một số AZ bị quá tải trong khi các AZ khác lại ít tải.

Cấu hình Cross-Zone Load Balancing

Trạng thái mặc định và chi phí của Cross-Zone Load Balancing khác nhau tùy thuộc vào loại Load Balancer.

Loại Load Balancer	Trạng thái Mặc Định	Có Thể Tắt	Chi Phí
Application Load Balancer (ALB)	Được bật	Có (ở cấp Target Group)	Không tính phí
Network Load Balancer (NLB)	Được tắt	Không	Phát sinh chi phí
Gateway Load Balancer (GWLB)	Được tắt	Không	Phát sinh chi phí

Application Load Balancer

- Mặc định được bật
- Có thể tắt ở cấp Target Group
- Không tính phí cho dữ liệu truyền tải giữa các AZ

Network & Gateway Load Balancer

- Mặc định bị tắt
- Nếu bật, bạn sẽ phải trả phí cho dữ liệu truyền tải giữa các AZ
- Không bao gồm trong gói miễn phí của AWS

SSL/TLS Termination

Khái niệm SSL/TLS Termination

SSL/TLS Termination là quá trình Load Balancer giải mã lưu lượng truy cập được mã hóa từ client và gửi lưu lượng không mã hóa đến các instance backend.

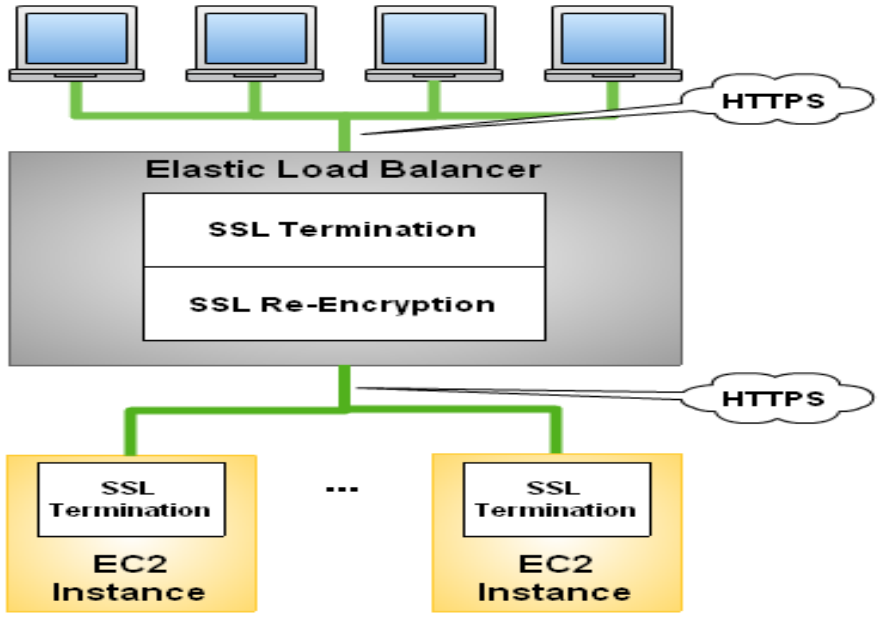
Mã hóa in-transit: Bảo vệ dữ liệu giữa client và Load Balancer

SSL/TLS & Certificate

Chứng chỉ X.509
SSL/TLS server certificate được sử dụng để mã hóa kết nối

AWS Certificate Manager (ACM)
Quản lý chứng chỉ SSL/TLS dễ dàng

HTTPS Listener
Yêu cầu một chứng chỉ mặc định và có thể có danh sách chứng chỉ tùy chọn để hỗ trợ nhiều domain



Phiên bản SSL/TLS

SSL (Secure Sockets Layer)
Phiên bản cũ hơn để mã hóa kết nối

TLS (Transport Layer Security)
Phiên bản mới hơn, hiện đại hơn
Thuật ngữ "SSL" vẫn thường được sử dụng

Server Name Indication (SNI)

Vấn đề giải quyết

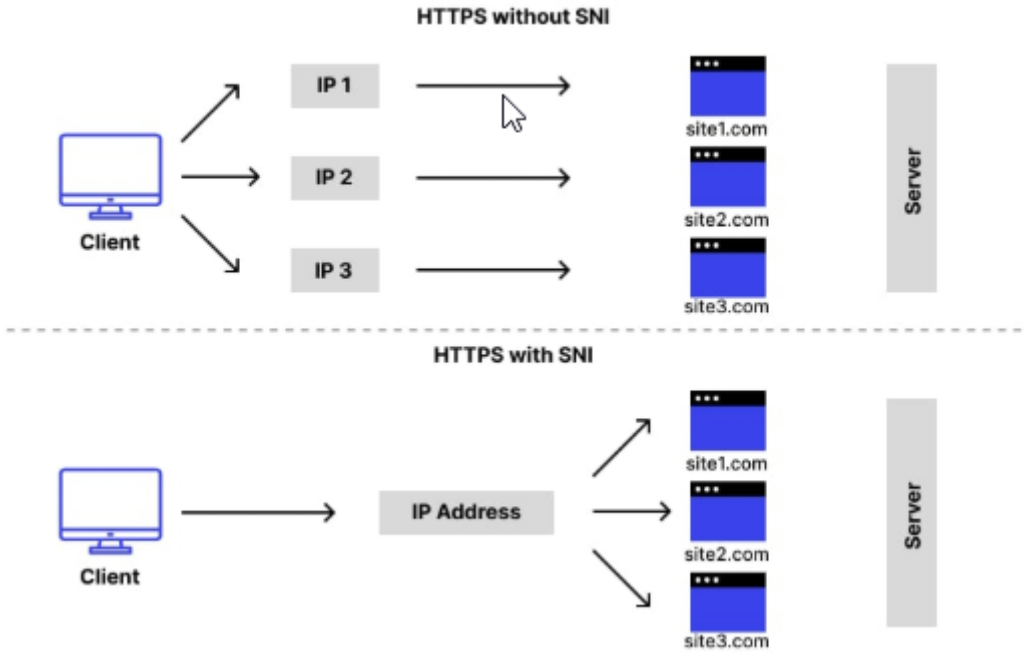
SNI giải quyết vấn đề khi một máy chủ web cần phục vụ nhiều trang web khác nhau, mỗi trang có chứng chỉ SSL riêng, trên cùng một địa chỉ IP.

Trước SNI, chỉ có thể sử dụng một chứng chỉ SSL cho mỗi IP address - giới hạn đối với các máy chủ chia sẻ IP.

Cách hoạt động

- Client gửi tên máy chủ mà nó muốn kết nối trong quá trình bắt tay TLS
- Load Balancer chọn đúng chứng chỉ dựa trên tên máy chủ
- Quá trình kết nối được mã hóa an toàn

SNI hoạt động như thế nào?



Hỗ trợ:

- ALB & NLB (thế hệ mới hơn)
- CLB (thế hệ cũ hơn)

Connection Draining / Deregistration Delay

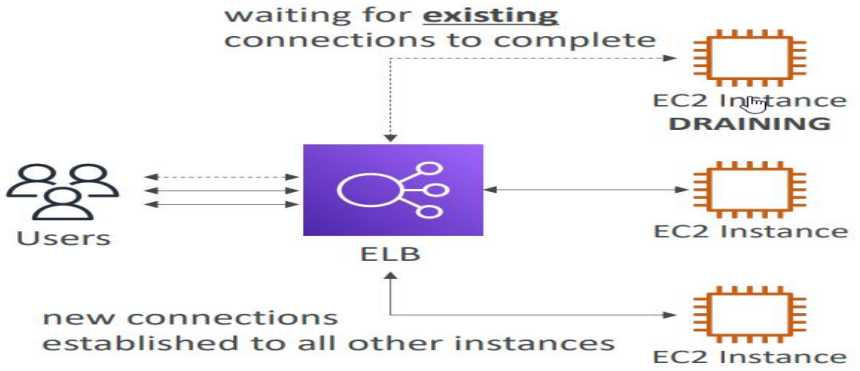
Giới thiệu

Connection Draining: Dành cho Classic Load Balancer (CLB)

Deregistration Delay: Dành cho Application Load Balancer (ALB) và Network Load Balancer (NLB)

Tính năng này đảm bảo rằng các yêu cầu đang xử lý (in-flight requests) được hoàn thành trước khi một instance bị gỡ đăng ký hoặc bị đánh dấu là không khỏe mạnh.

Quá trình Connection Draining



Cấu hình

Thời gian cấu hình
1 đến 3600 giây (mặc định: 300 giây)

Tắt tính năng
Đặt giá trị về 0

Lưu ý
Nên đặt giá trị thấp nếu các yêu cầu của bạn ngắn

So sánh tên gọi

Load Balancer	Tên gọi
Classic Load Balancer (CLB)	Connection Draining
Application Load Balancer (ALB)	Deregistration Delay
Network Load Balancer (NLB)	Deregistration Delay

Tổng kết

Elastic Load Balancer (ELB) là gì?

ELB là dịch vụ quản lý của AWS, đóng vai trò then chốt trong việc xây dựng các ứng dụng có khả năng mở rộng và tính sẵn sàng cao. Phân phối lưu lượng truy cập đồng đều đến nhiều instance, đảm bảo ứng dụng luôn hoạt động ổn định và hiệu quả.

Các loại ELB

- **Application Load Balancer (ALB)**: Layer 7, HTTP/HTTPS
- **Network Load Balancer (NLB)**: Layer 4, TCP/UDP
- **Gateway Load Balancer (GWLB)**: Layer 3, thiết bị mạng ảo
- **Classic Load Balancer (CLB)**: Layer 4/7, thế hệ cũ

Tính năng quan trọng

- **Health Checks**: Đảm bảo lưu lượng chỉ đến instance khỏe mạnh
- **Security Groups**: Kiểm soát luồng traffic
- **Sticky Sessions**: Duy trì phiên làm việc với instance
- **Cross-Zone Load Balancing**: Phân phối tải đều giữa các AZ

Cách sử dụng hiệu quả

- **SSL/TLS Termination**: Mã hóa lưu lượng HTTPS
- **Connection Draining**: Hoàn thành yêu cầu đang xử lý
- **Target Groups**: Nhóm các instance mục tiêu
- **Auto Scaling**: Tự động điều chỉnh số lượng instance

ELB là một thành phần không thể thiếu trong mọi kiến trúc AWS hiện đại, giúp tối ưu hóa hiệu suất, độ tin cậy và khả năng quản lý của ứng dụng.