Nội dung

CloudNut
Learn Clouds in a Nutshell

- High Availibility and Scalibility
- Load Balancer
- Auto Scaling Group
- Design HA Architecture

Scalibility and High Availibility

Scalibility

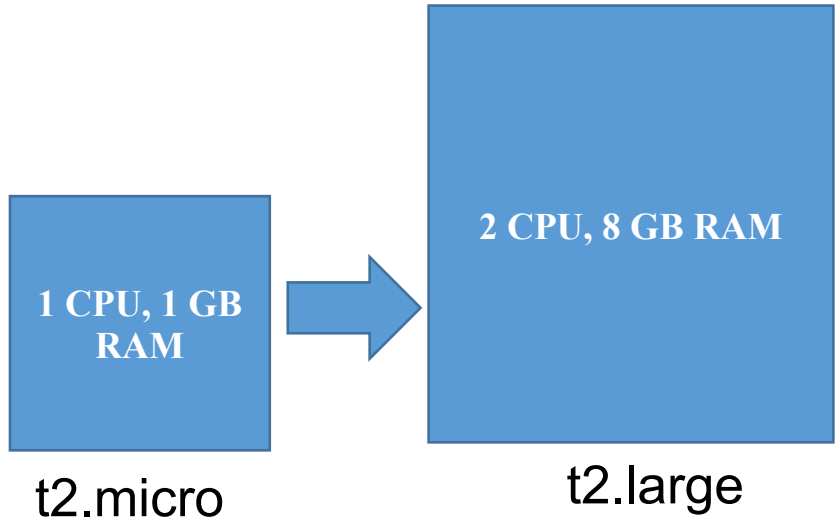
CloudNut
Learn Clouds in a Nutshell

- Tính Scalibility được định nghĩa là hệ thống có thể tự co giãn, thay đổi về công suất để đáp ứng sự thay đổi của tải, traffic
- Có 2 loại scalability:
  - Theo chiều dọc: Vertical Scalibility
  - Theo chiều ngang: Horizontal Scalibility

Vertical Scaling

CloudNut
Learn Clouds in a Nutshell

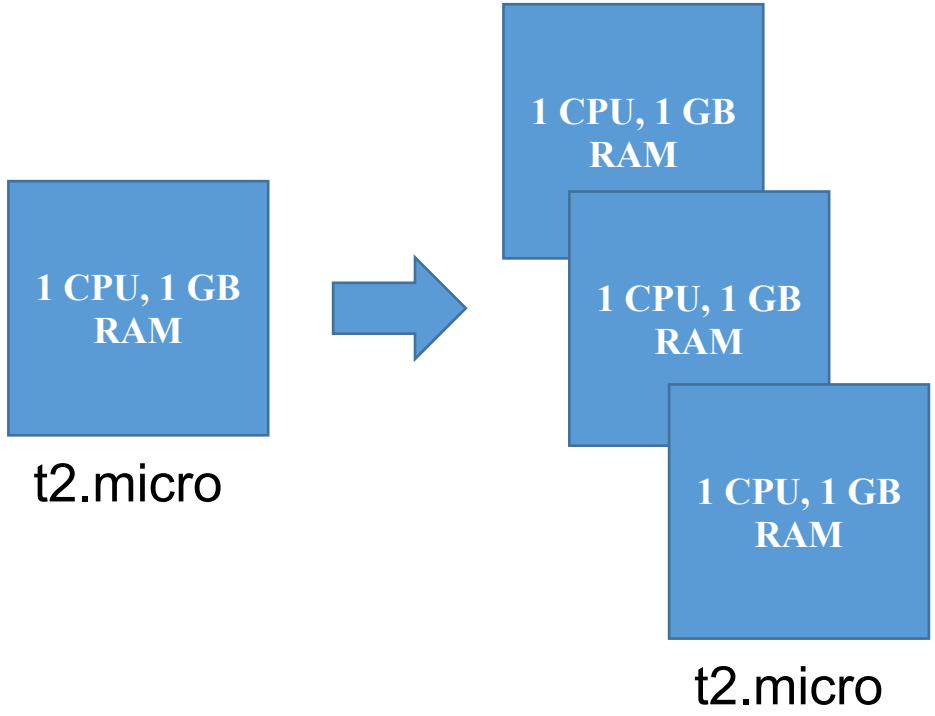
- Vertical Scaling là nhằm tăng công suất của VM, Instances (Tăng RAM, CPU)
- Ví dụ:
  - Thay đổi Instance type của EC2 Instance từ t2.micro (1CPU, 1GB RAM) sang t2.large (2CPU, 8GB RAM)
- Use cases:
  - Cho hệ thống không phân tán (non-distributed system: Database)
  - Các hệ thống kết nối chặt với nhau (Strictly Coupling)



Horizontal Scaling

CloudNut
Learn Clouds in a Nutshell

- Horizontal Scaling là tăng công suất dựa vào cách tăng số lượng VM/Instances
- Ví dụ:
  - Tăng số lượng VM/Instances từ 1 lên 3 Instances
- Use cases:
  - Các ứng dụng theo mô hình mới, microservices
  - Hệ thống phân tán, loose-coupling system



High Availibility (HA)

CloudNut
Learn Clouds in a Nutshell

- Tính có sẵn cao (High Availibility) được hiểu là Hệ thống hoạt động đúng chức năng (available) và hiệu năng (performance) ở mức độ cao (High Level) trong một khoảng thời gian
- Ví dụ:
  - Tính Availibility của Hệ thống là 99% trên năm => Thời gian Downtime của hệ thống là 1% ~ 3.65 ngày
  - High Availibility có thể lên tới 99.99%
- Để đạt được tính có sẵn cao cần một số tiêu chí sau:
  - Hệ thống không có Single Point of Failure (SPOF)
  - Các thành phần phải được Redundancy (Tính dư thừa)

High Availibility (HA)

CloudNut
Learn Clouds in a Nutshell

- Một số tiêu chí thiết hệ hệ thống có HA trong AWS
  - Đảm bảo hệ thống được triển khai trên ít nhất 2 Availibility Zone
  - Đảm bảo hệ thống có cơ chế Auto Scaling (Horizontal Scaling) ở nhiều thành phần nhất có thể
  - Tầng Database có thể sử dụng RDS Multi-AZ hoặc Cluster

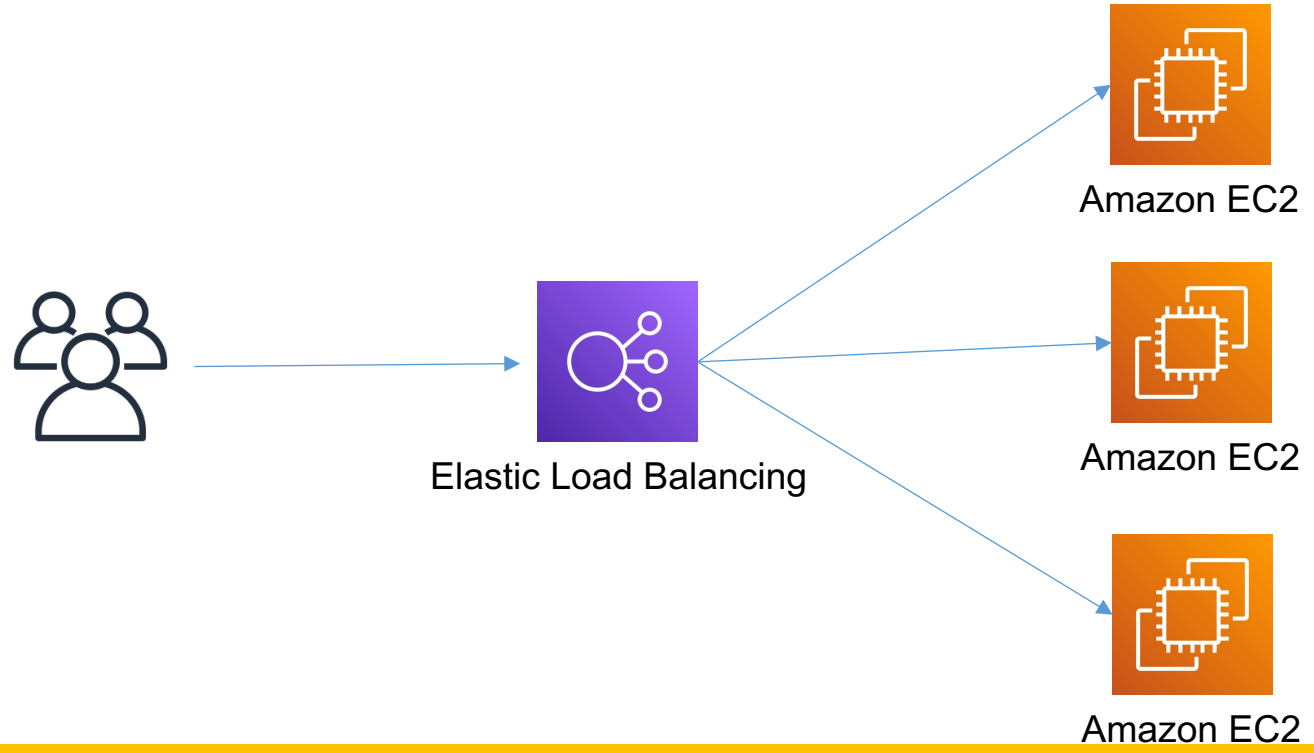
CloudNut
Learn Clouds in a Nutshell

Load Balancer

Load Balancer

CloudNut
Learn Clouds in a Nutshell

- Bộ cân bằng tải - Load Balancer (LB) là thành phần hoạt động như một proxy dùng để chuyển tiếp (forward) các yêu cầu tới các Servers đứng sau nó (Downstream Servers)



Tính năng Load Balancer

CloudNut
Learn Clouds in a Nutshell

- Cân bằng tải cho các Downstream Servers
- Cung cấp một điểm truy cập duy nhất cho Clients (Single Point of Access DNS)
- Tự động xử lý sự cố khi Downstream Servers gặp vấn đề (Ngắt traffic)
- Health Check các Downstream Servers
- Xử lý SSL (SSL Termination)
- Tính có sẵn cao do các Servers của LB nằm trên nhiều Availibilty Zone khác nhau
- Tách luồng Traffic từ Internet (Internet Facing) và nội bộ hệ thống (Internal Facing)

Elastic Load Balancer

CloudNut
Learn Clouds in a Nutshell

- Elastic Load Balancer là Load Balancer quản lý bởi AWS (Managed LB)
  - Managed by AWS
  - High Availibility, Scalibility
- ELB có thể tích hợp được với các dịch vụ của AWS như EC2, ASG, Route53…

Elastic Load Balancer

Các loại Elastic Load Balancer

CloudNut
Learn Clouds in a Nutshell

- AWS cung cấp 4 loại Load Balancers khác nhau
- Lựa chọn các LB thế hệ mới để hỗ trợ tính năng cao cấp hơn

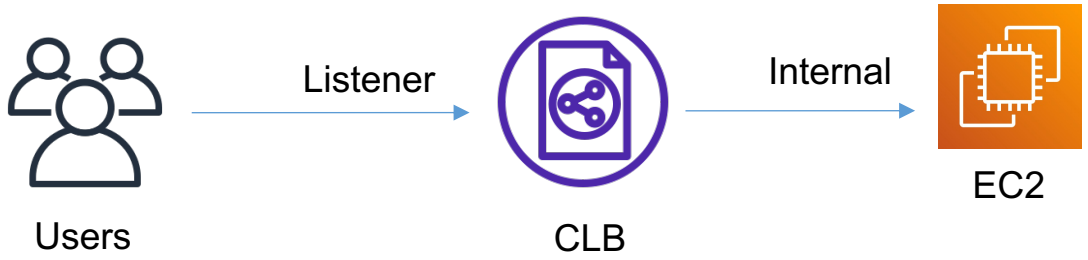
LB type	Classic LB	Application LB	Network LB	Gateway LB
Release	2009	2016	2017	2020
Supported Protocols	HTTP/HTTPS, TCP, SSL (secure TCP)	HTTP, HTTPS, WebSocket	TCP, SSL (secure TCP), UDP	Working at layer 3 (Network Layer) – IP protocol

Classic Load Balancer (CLB)

CloudNut
Learn Clouds in a Nutshell

- Hỗ trợ TCP (layer 4), HTTP and HTTPS (Layer 7)
- Health Check sử dụng TCP hoặc HTTP
- DNS name (fixed)

  XXX.region.elb.amazonaws.com



Application Load Balancer (ALB)

CloudNut
Learn Clouds in a Nutshell

- ALB hoạt động ở Layer 7 (HTTP/HTTPS)
- Phân tải giữa nhiều Target Group (Nhóm các Server cùng chức năng) khác nhau
- Hỗ trợ HTTP/2 và WebSocket
- Hỗ trợ redirects (Từ HTTP sang HTTPS)

Application Load Balancer (ALB)

CloudNut
Learn Clouds in a Nutshell

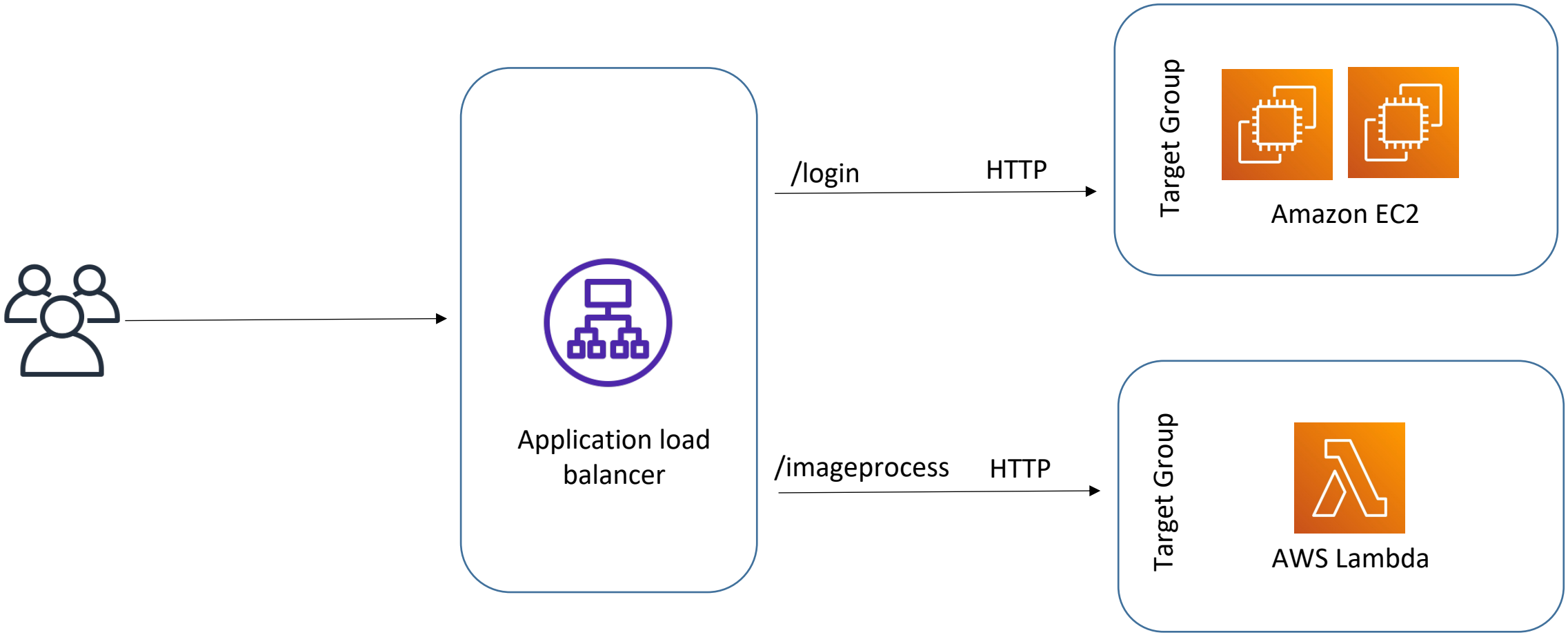
- Định tuyến tới **Target Group** dựa vào:
  - Path (example.com/users & example.com/posts)
  - Hostname (one.example.com & other.example.com)
  - Query String, Headers (example.com/users?id=123&order=false)
- ALB phù hợp cho các ứng dụng triển khai Microservice, Containerization (Docker, Kubernetes)
- Một ứng dụng có thể cần nhiều CLB trong khi đó sử dụng ALB chỉ cần 1 là đủ

ALB – Target Groups

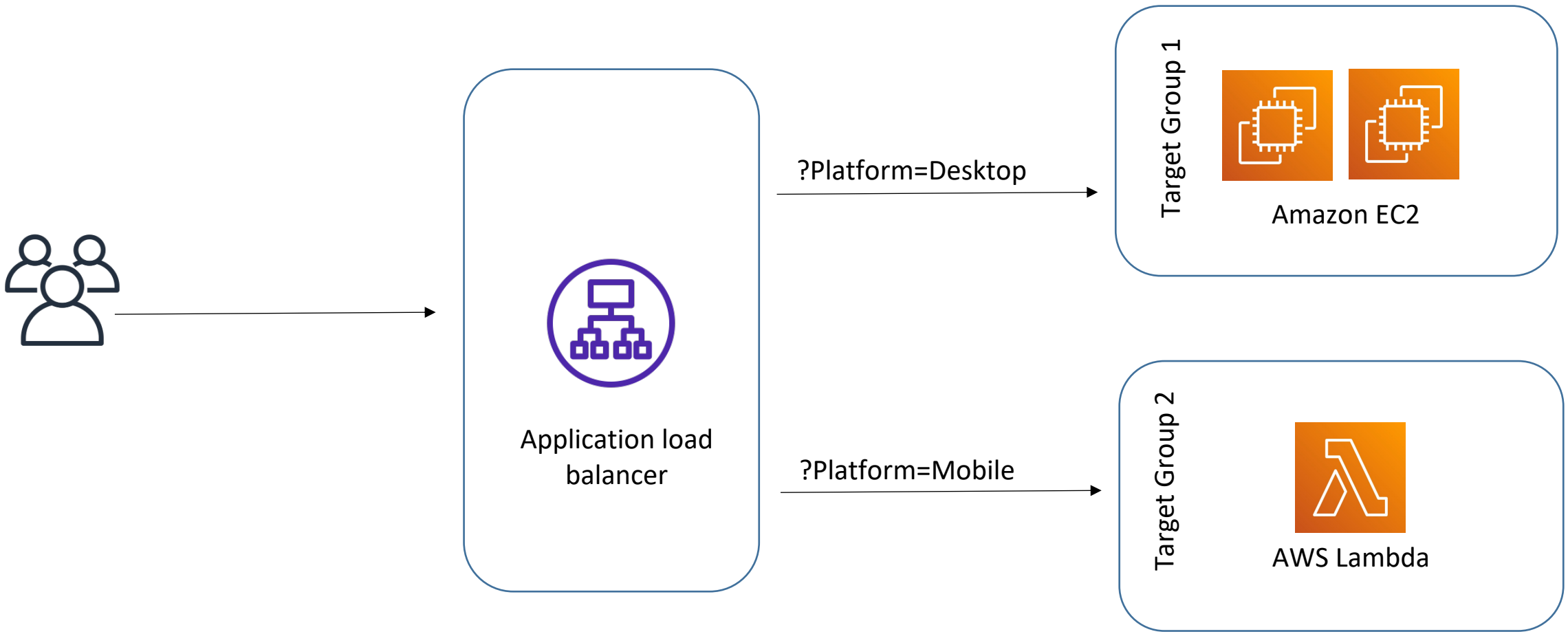
CloudNut
Learn Clouds in a Nutshell

- Target Groups là đơn vị xử lý các request được forward từ ALB
- Target Groups có thể là:
  - EC2 instances (có thể được quản lý bởi Auto Scaling Group)
  - ECS tasks
  - Lambda function
  - IP addresses
- ALB có thể định tuyến tới nhiều **Target Groups** khác nhau

Path Routing



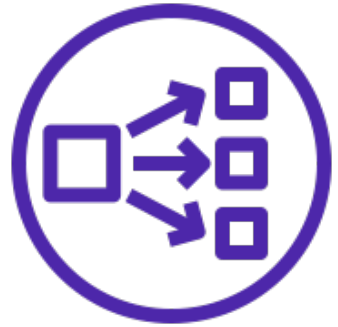
Query Strings/Parameters Routing



Network Load Balancer (NLB)

CloudNut
Learn Clouds in a Nutshell

- NLB hoạt động ở Layer 4 (TCP + UDP)
  - Forward UDP/TCP traffics to instances
  - Độ trễ thấp, xử lý hàng triệu request per second



- Elastic IP có thể gắn vào NLB (Phục vụ cho các y/c cần IP tĩnh, cố định)
- Không được hỗ trợ với AWS Free Tier

NLB – Target Groups

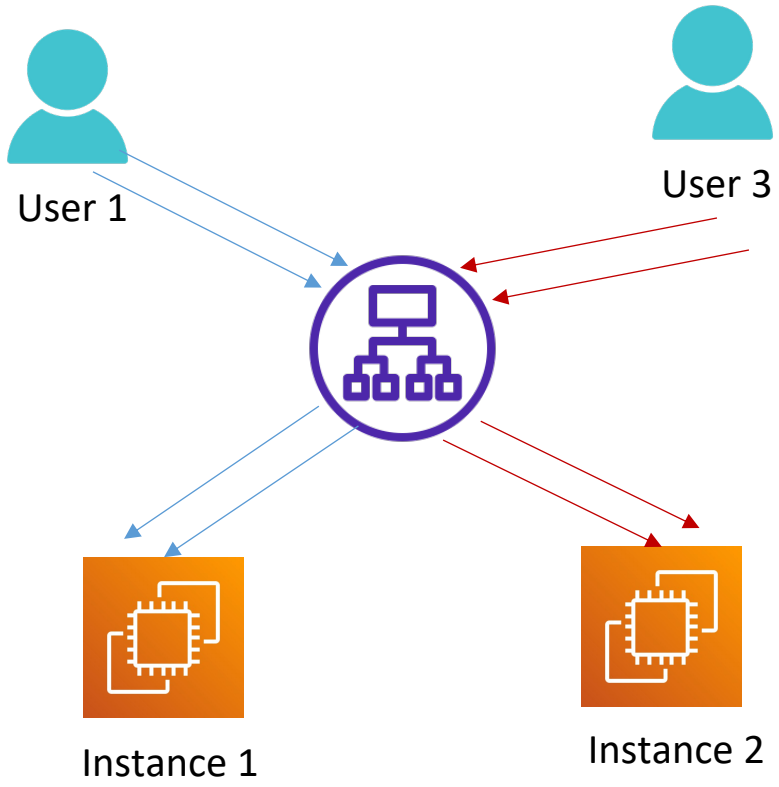
CloudNut
Learn Clouds in a Nutshell

- NLB Target Groups có thể là :
  - EC2 instances
  - IP addresses (Must be private IP)
  - Application Load Balancer

Sticky Session (Session Affinity)

CloudNut
Learn Clouds in a Nutshell

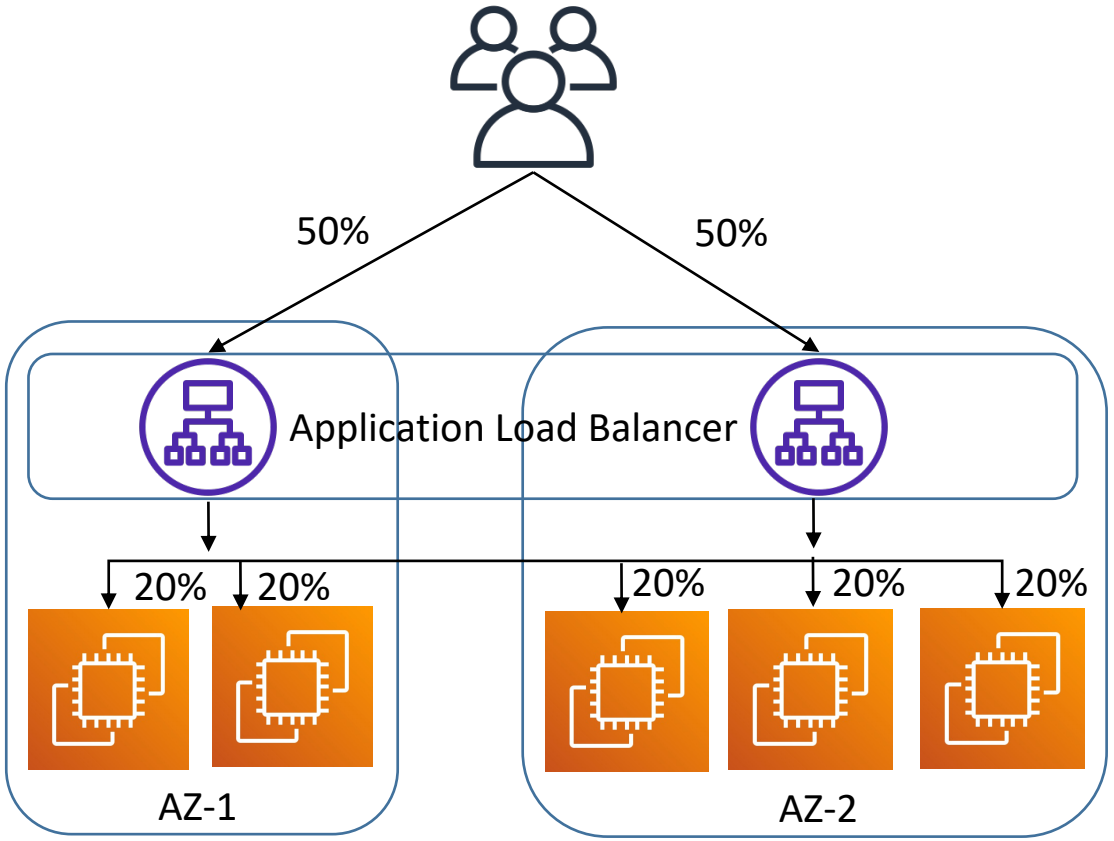
- Sử dụng Sticky Session để định tuyến cố định các Request từ cùng một User sẽ được chuyển tới tới cùng một Server Downstream
- Tính năng hỗ trợ trên CLB & ALB
- Sử dụng cho Stateful Application
- Sử dụng Sticky Session sẽ không đảm bảo tải ở các Server Downstream được cân bằng



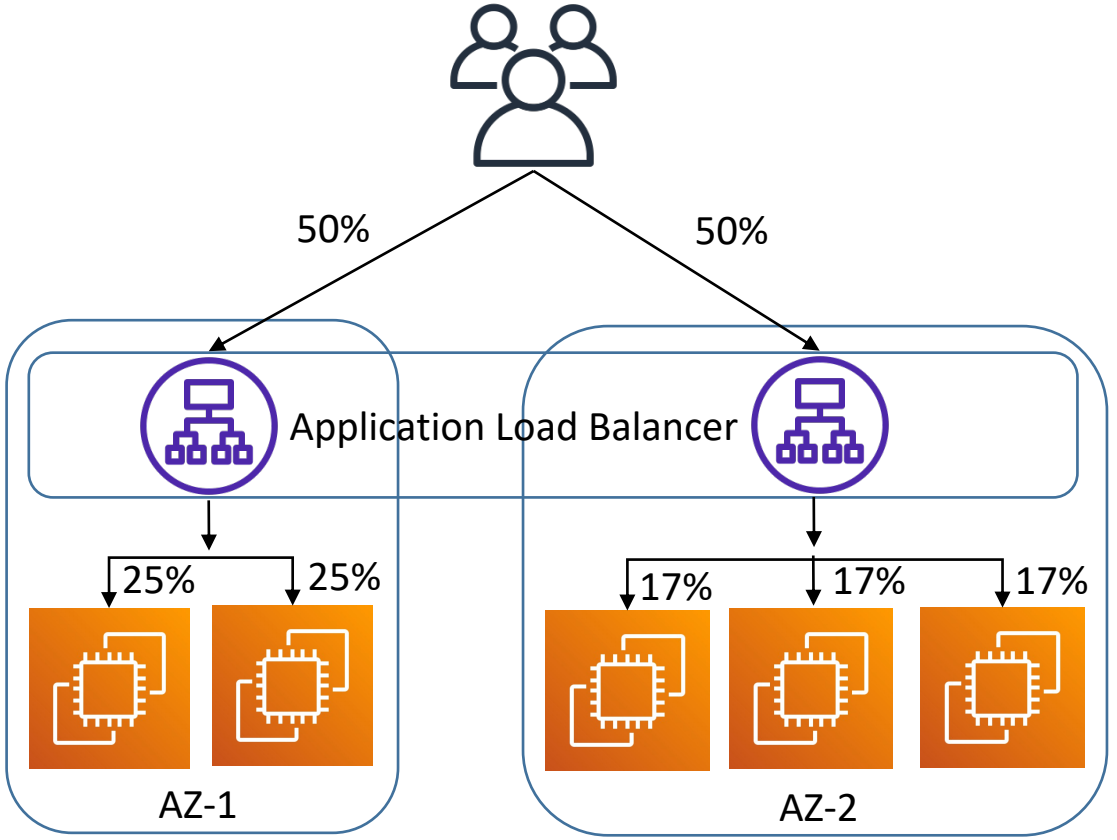
Cross-Zone Load Balancing

CloudNut
Learn Clouds in a Nutshell

- Set Cross-Zone Loab Balacing



- Không set Cross-Zone Load Balancing



Cross-Zone Load Balancing (cont.)

CloudNut
Learn Clouds in a Nutshell

Characters	Application LB	Network LB	Classic LB
Enable by default	Yes (cannot disable)	No	No
Chi phí Data Transfer (Inter AZ)	No	Yes	Yes

SSL Termination

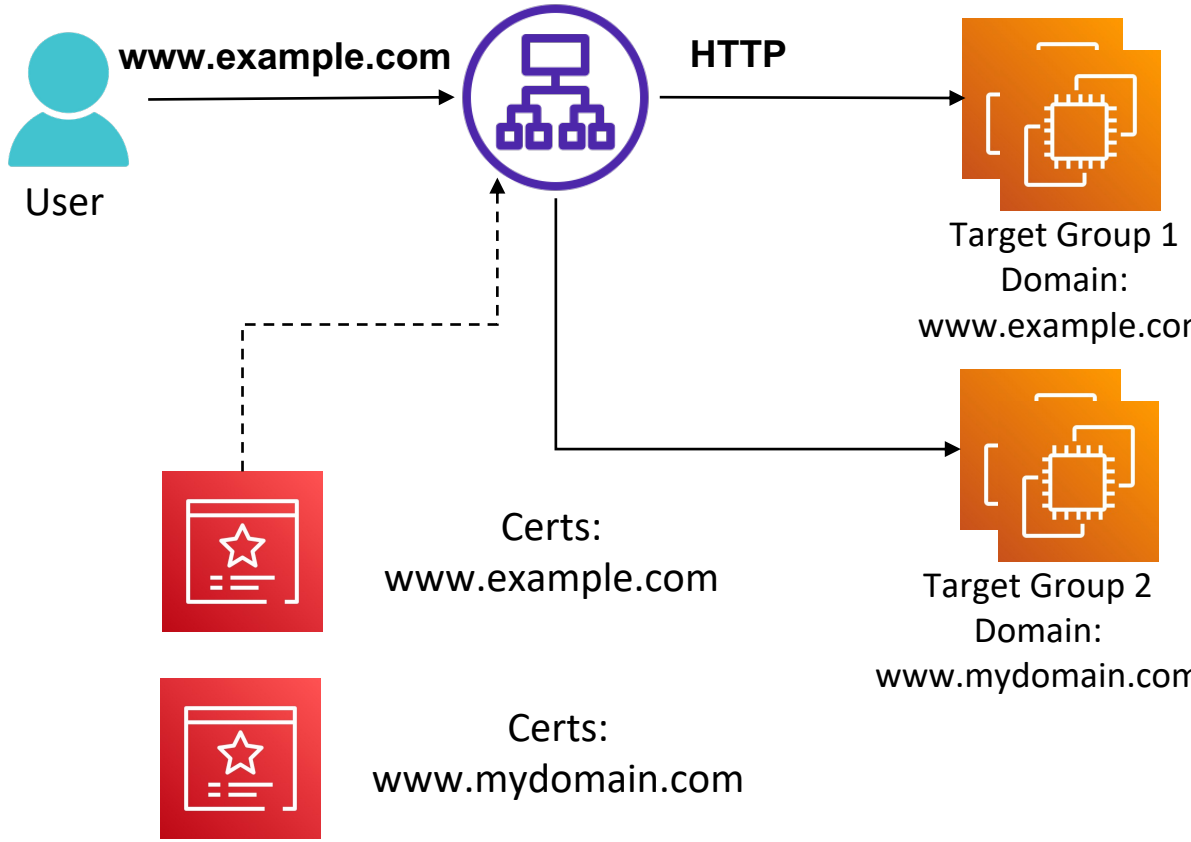
CloudNut
Learn Clouds in a Nutshell

- LB sử dụng chứng chỉ X.509 certificates
- SSL certs được quản lý trên ACM (Amazon Certificate Manager)
- SSL certs có thể được tạo ra trên ACM hoặc upload từ users mua của 3rd party

SSL – Server Name Indicator (SNI)

CloudNut
Learn Clouds in a Nutshell

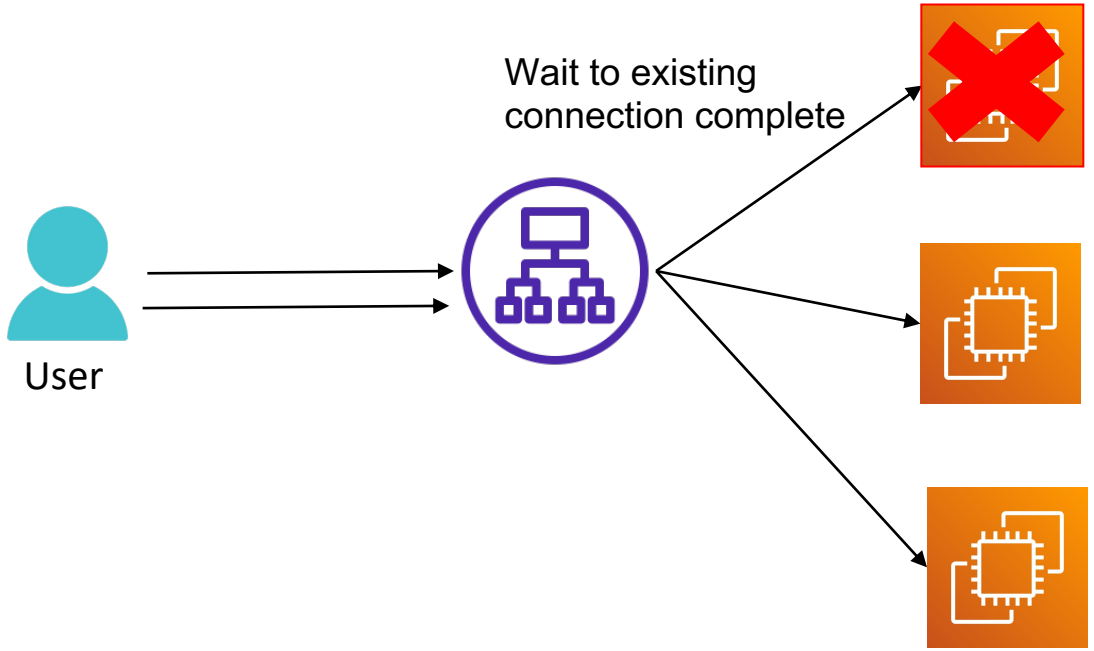
- SNI cho phép hỗ trợ nhiều Web Servers với nhiều SSL Certs tương ứng
- Clients phải chỉ định Hostname Server để thực hiện quá trình Handshaking
- Lưu ý:
  - Chỉ hỗ trợ cho các LB thế hệ mới (ALB, NLB)
  - Không hỗ trợ CLB



Connection Draining

CloudNut
Learn Clouds in a Nutshell

- Khoảng thời gian đợi hoàn thành các request đang được xử lý (In-Processing Requests) khi Instance bị Unhealthy, Failed Health Check
- Dừng việc gửi Request tới các Instances này
- Draining period khoảng từ 1 ~ 3600s (mặc định 300s)
- Có thể set giá trị về 0



CloudNut
Learn Clouds in a Nutshell

Auto Scaling Group

Auto Scaling Group (ASG)

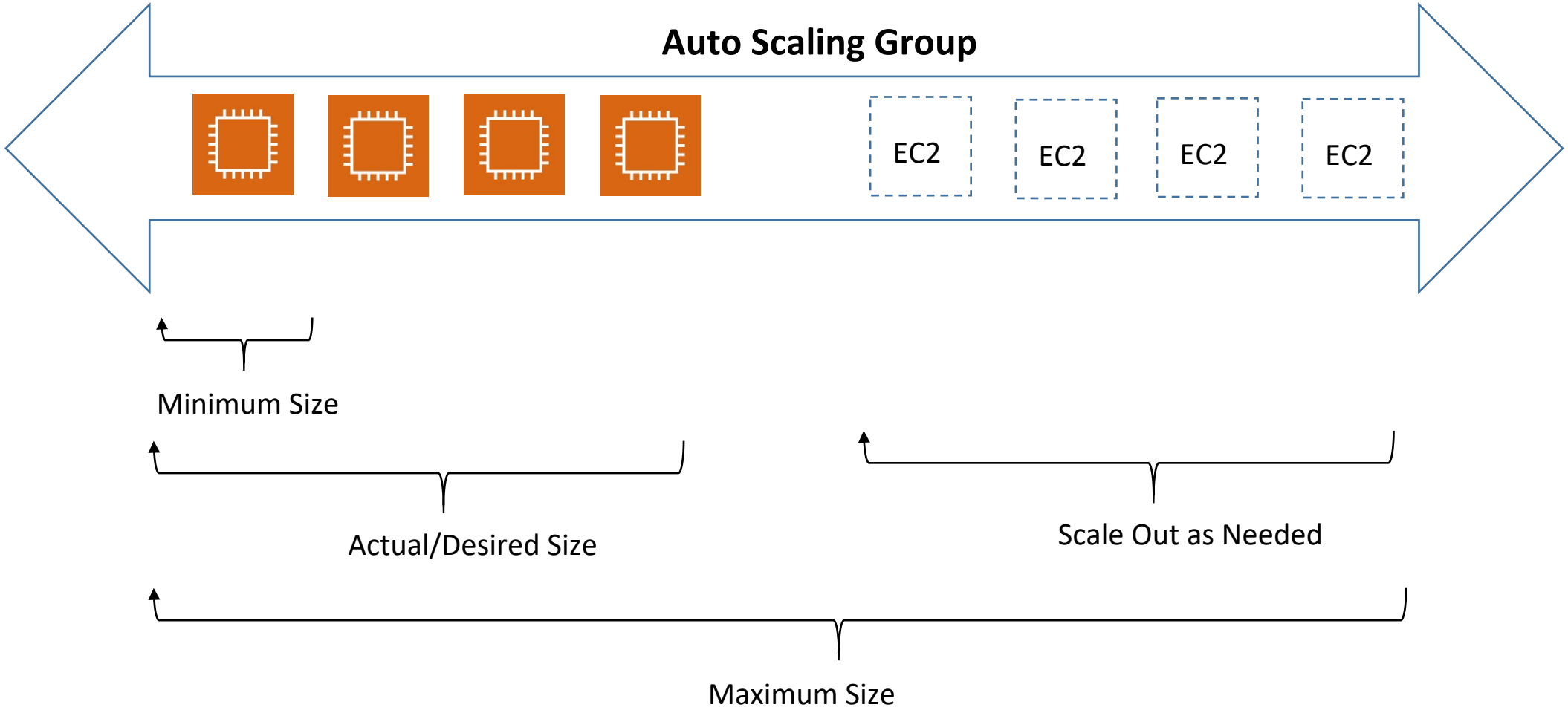
CloudNut
Learn Clouds in a Nutshell

- ASG cho phép tăng (Scale Out) hoặc giảm (Scale In) số lượng EC2 Instances để đáp với sự thay đổi của tải
- Tận dụng tính linh hoạt trong việc cấp phát tài nguyên của Public Cloud
- Tính năng của ASG:
  - Scale Out (Tăng số lượng EC2 Instance)
  - Scale In (Xoá/Giảm số lượng EC2 Instance)
  - Đảm bảo số lượng Min/Max Instance running
  - Tự động thêm các Instances vào LB

Auto Scaling Group

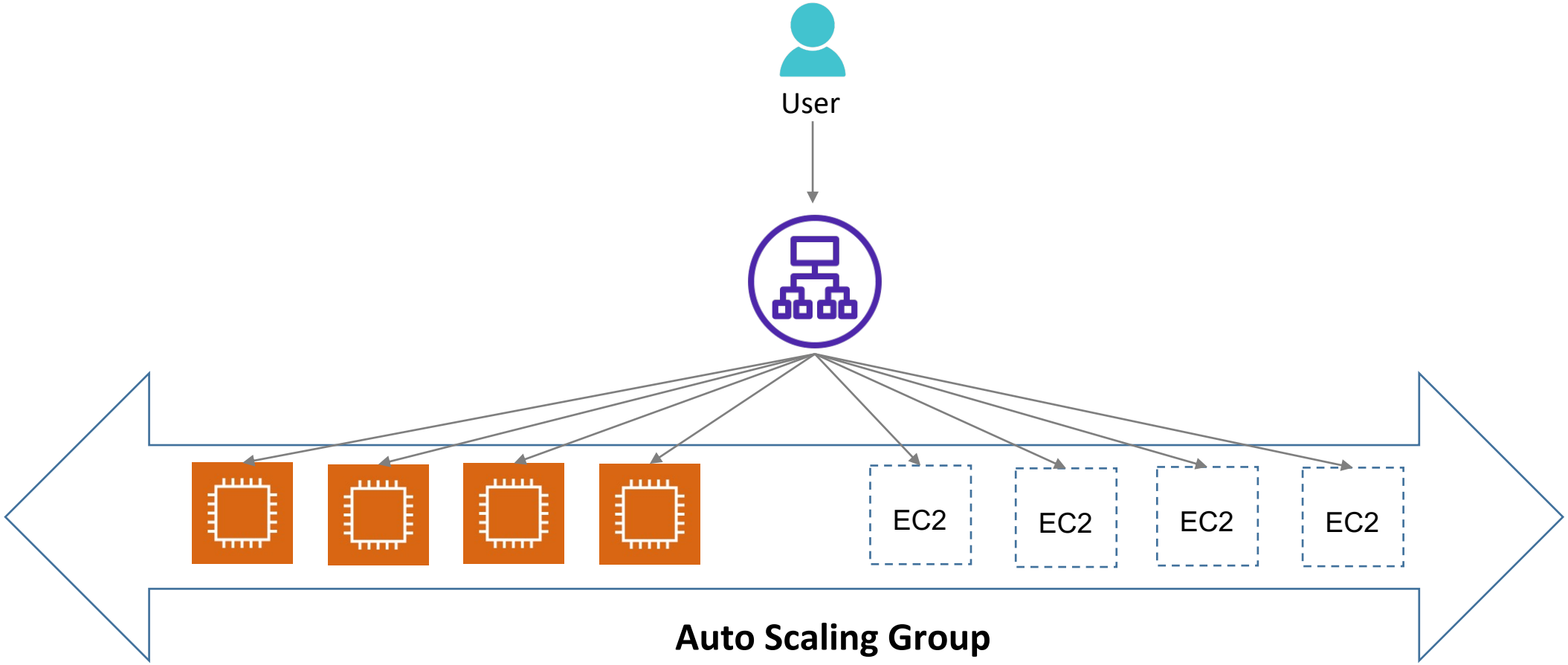
Auto Scaling Group in AWS

CloudNut
Learn Clouds in a Nutshell



ASG with LB

CloudNut
Learn Clouds in a Nutshell



ASG Component

CloudNut
Learn Clouds in a Nutshell

- **Groups:** Tập các EC2 Instance thực hiện cùng chức năng (Logical Group)
- **Configuration Template**
  - **Groups** sử dụng **Lanch Template** hoặc **Launch Configuration (LC)** để định nghĩa tham số cấu hình của EC2 Instance (AMI, Instance type, Volume, Keypair, SGs….)
- **Scaling Options**
  - Các cách thực hiện việc Scale ASG
  - Ví dụ: Dựa vào một số điều kiện (Dynamic Scaling) hoặc dựa vào việc lập lịch (Schedule)

Scaling Options

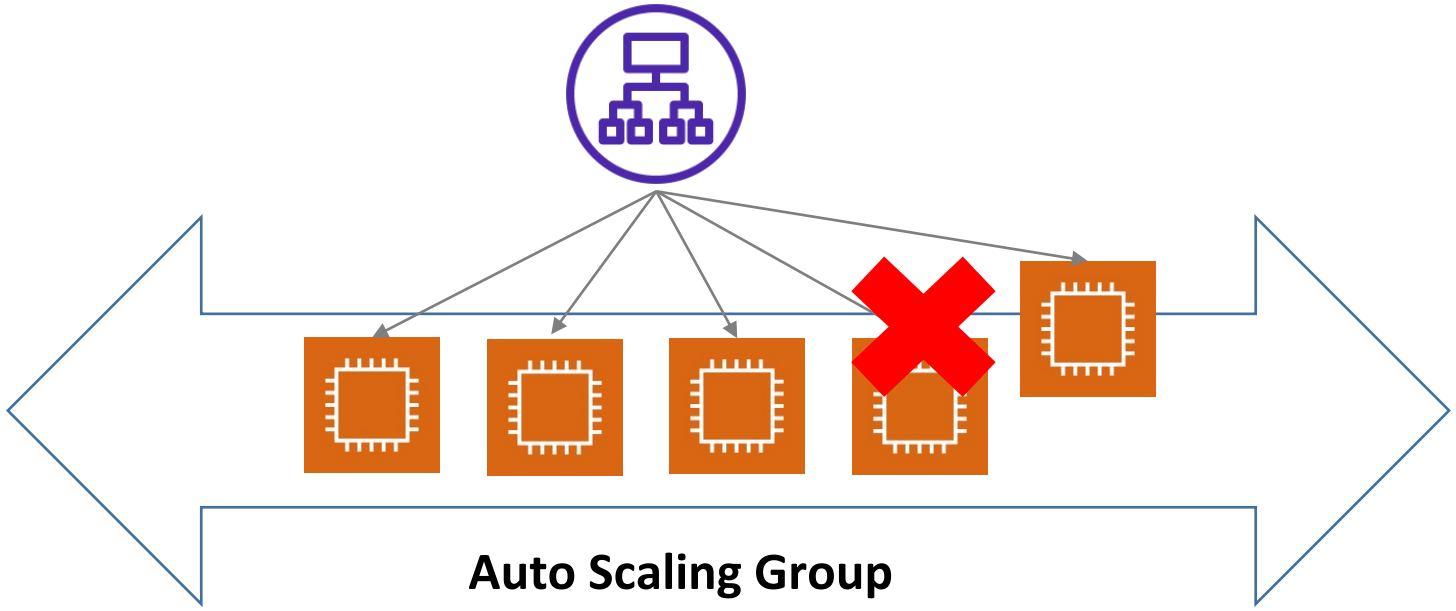
CloudNut
Learn Clouds in a Nutshell

- Duy trì số lượng Instances mọi thời điểm (Maintain a fixed number of instances)
- Scale thủ công (Scale Manually)
- Scale dựa vào việc lập lịch (Scale based on Schedule
- Scale dựa vào nhu cầu về tải (Scale based on deman
- Scale dựa vào tiên đoán (Use predictive scaling)

Maintain a fixed number of instances

CloudNut
Learn Clouds in a Nutshell

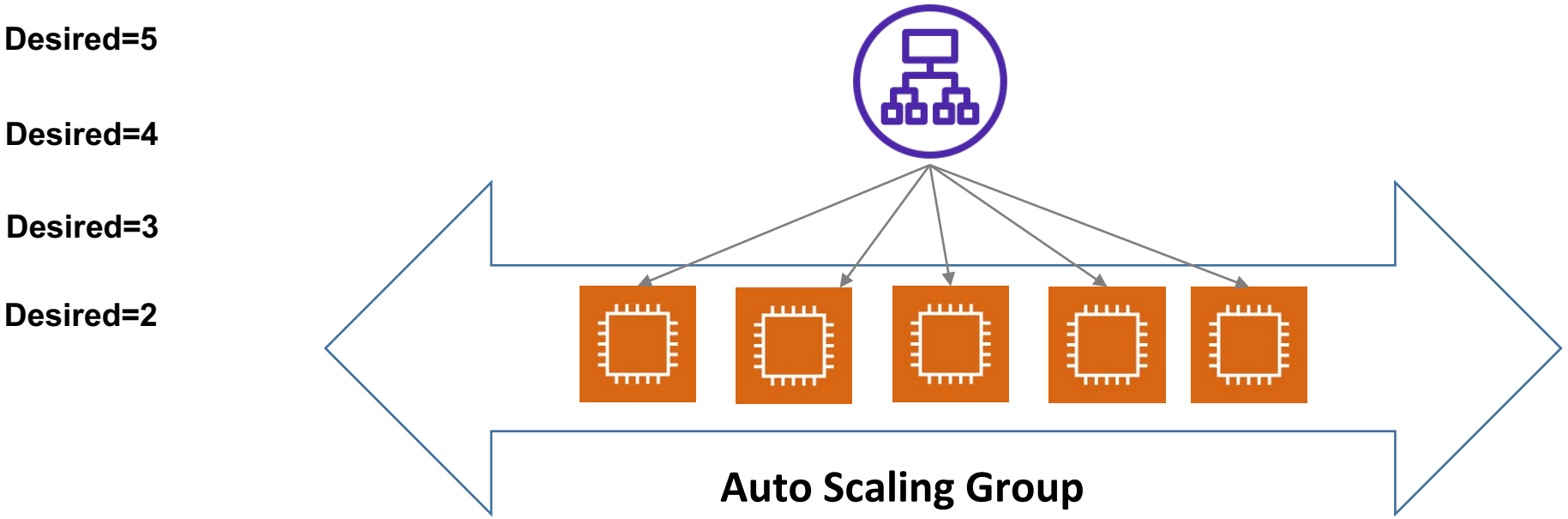
- Thiết lập số lượng Instance Min=Max=Desired
- ASG thực hiện việc gửi Healthcheck định kỳ tới các Instances trong ASG
- Khi ASG phát hiện ra Unhealthy Instance, nó sẽ Terminate và thay thế bằng một Instance khác



Scale Manually

CloudNut
Learn Clouds in a Nutshell

- Chỉ định số lượng Instances Min < Desired < Max cho ASG



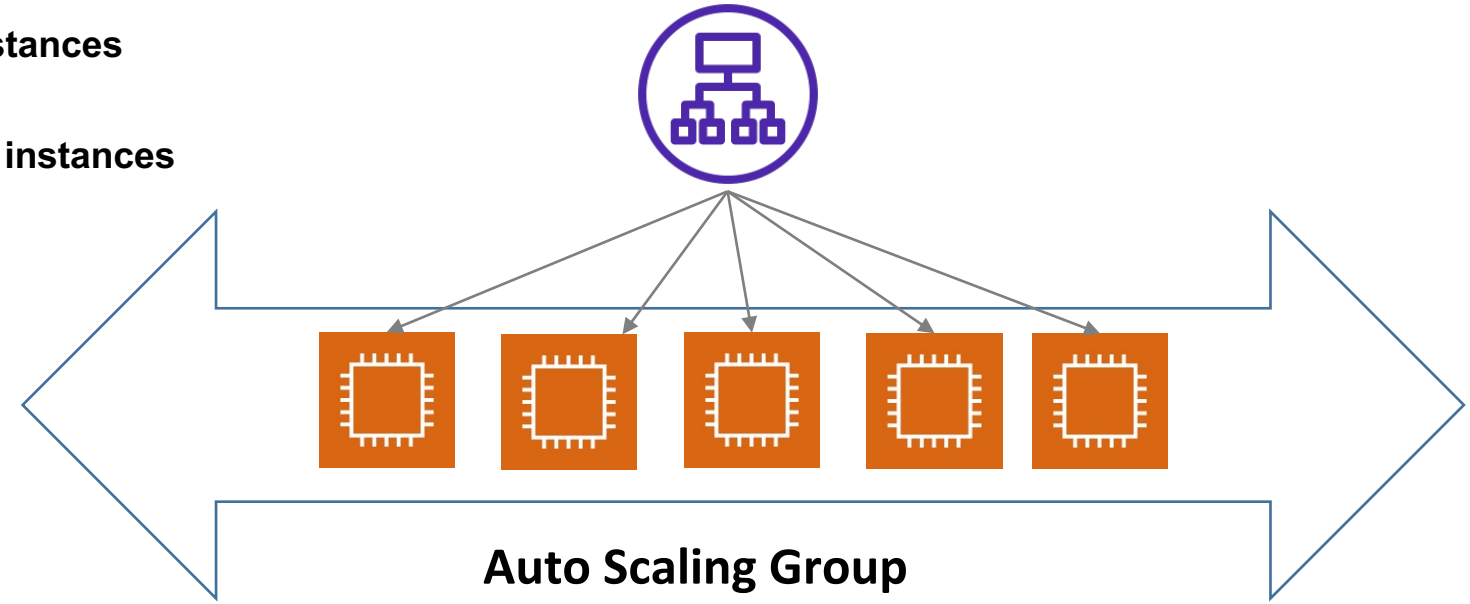
Scale Based On Schedule

CloudNut
Learn Clouds in a Nutshell

- Thực hiện hành động Scale cho ASG theo lịch đặt trước (Scheduler)

19:00 Scale Up 2 Instances

08:00 Scale Down 2 instances

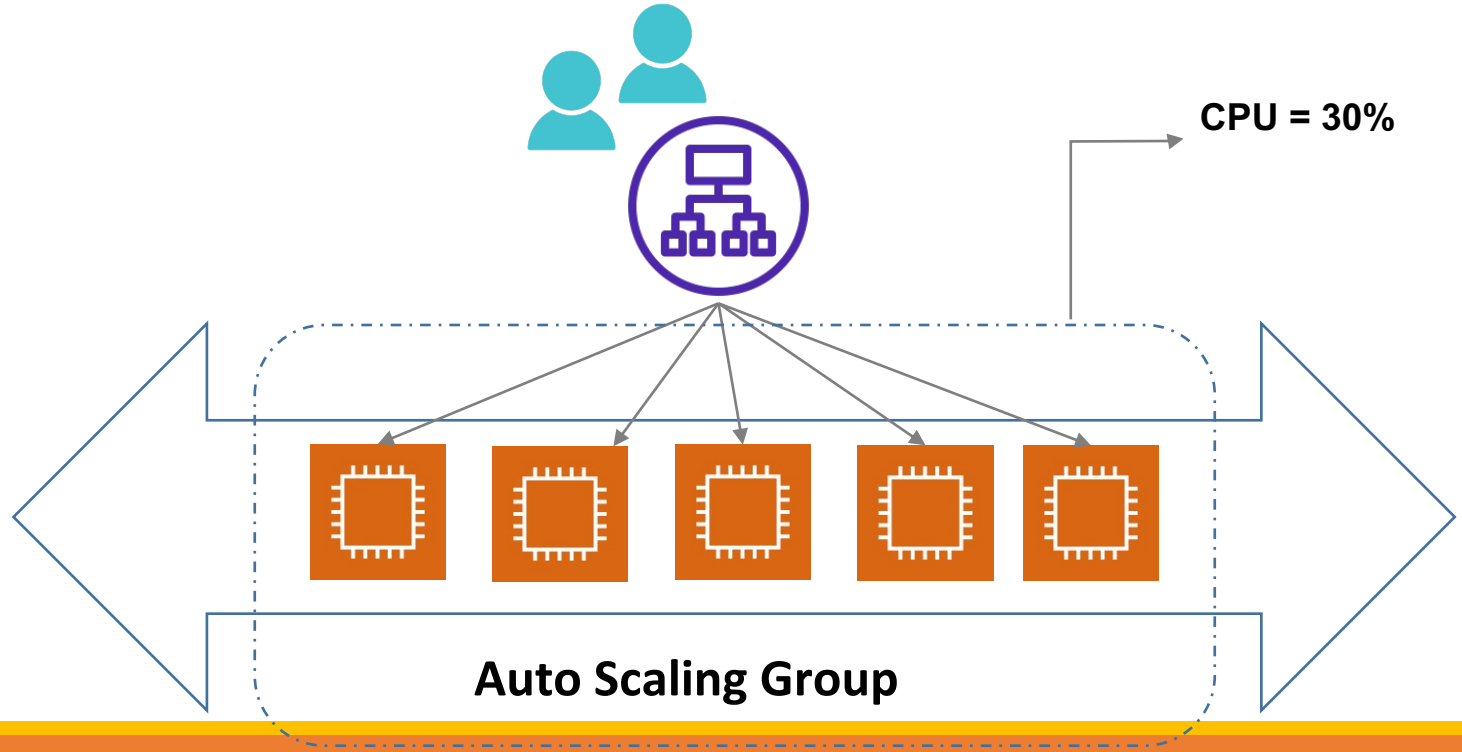


Scale Based On Demand

CloudNut
Learn Clouds in a Nutshell

- Thiết lập số lượng Instances Min < Desired < Max
- Cho phép việc Scale ASG để đáp ứng lại việc thay đổi về nhu cầu công suất capacity

Policy
- CPU > 80% Scale Up 2 Instances
- CPU < 45% Scale Down 1 instances

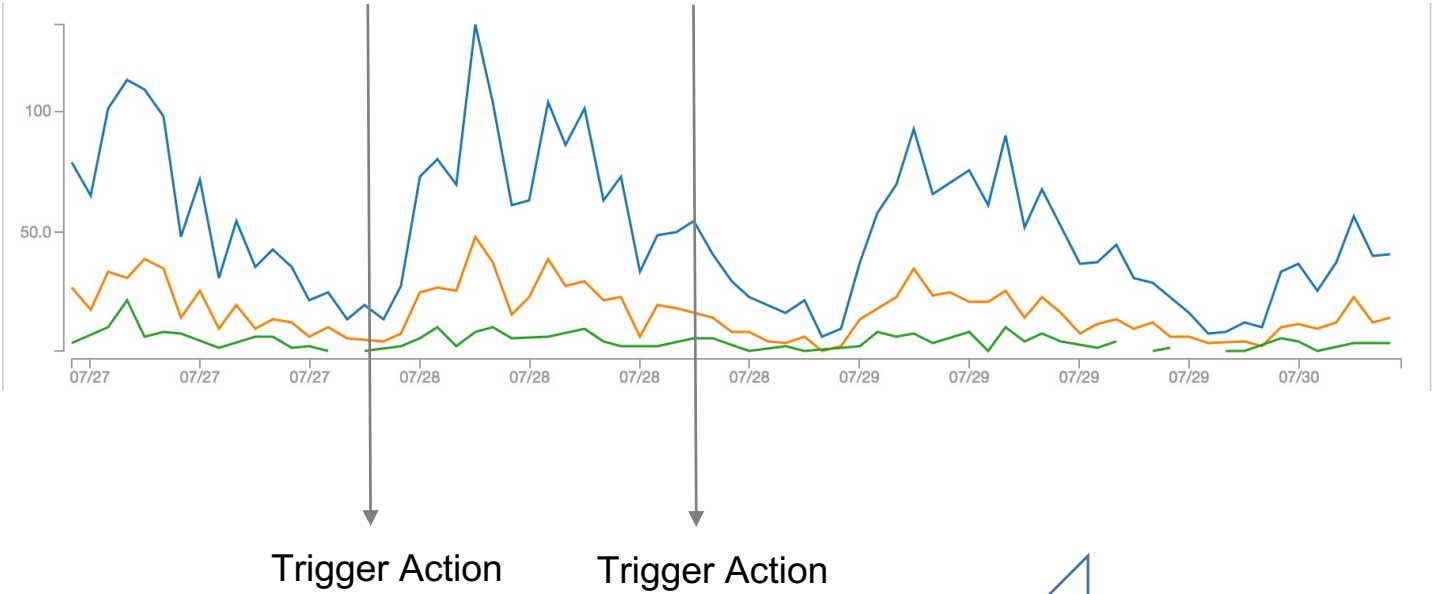


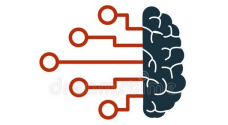
Predictive Scaling

CloudNut
Learn Clouds in a Nutshell

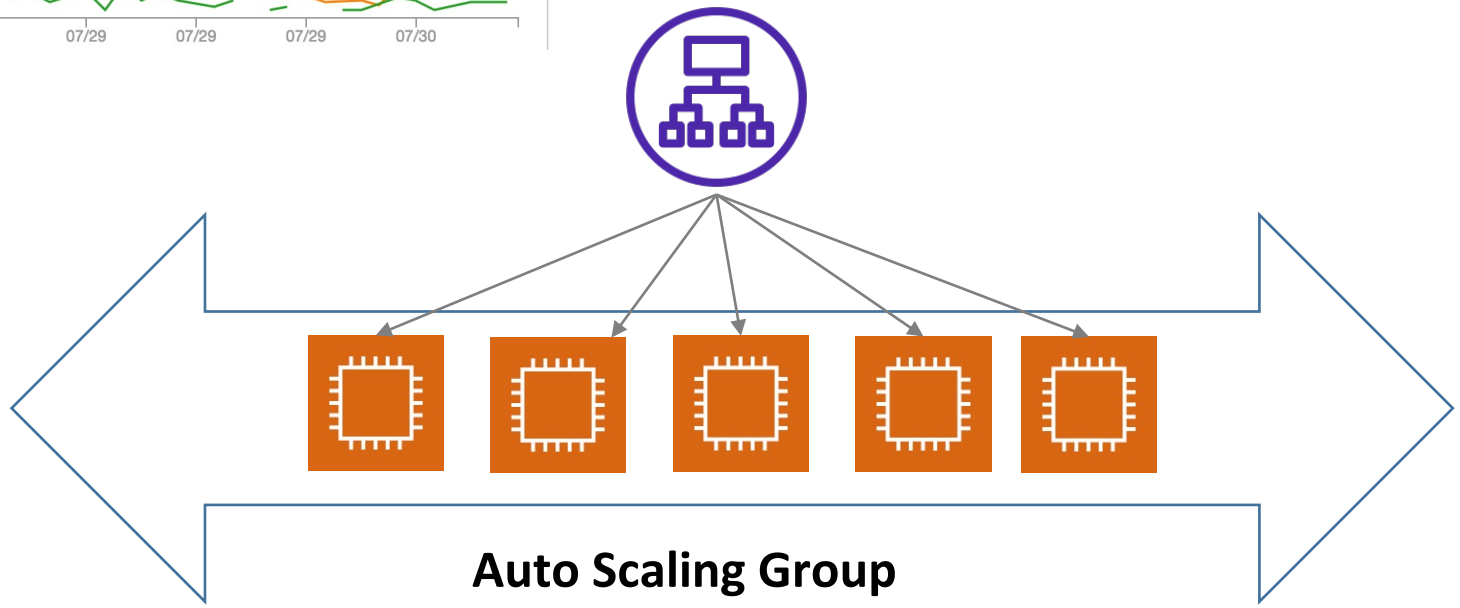
- Sử dụng ML để dự đoán công suất cần thiết
- Dữ liệu được thu thập từ CW phục vụ cho xây dựng ML model
- Scaling ASG sẽ được thực hiện trước sự kiện thay đổi về tải

Predictive Scaling (cont.)





MACHINE LEARNING



Exam Tips

CloudNut
Learn Clouds in a Nutshell

- Scaling Options
  - Maintain number of instances at all times
  - Scale Manually
  - Scale based on Scheduler
  - Scale based on demand
  - Use predictive scaling

CloudNut
Learn Clouds in a Nutshell

Design HA Architecture

Plan for failure

CloudNut
Learn Clouds in a Nutshell

- Werner Vogels (CTO AWS): “Everything fails all the time”
- Luôn phải chuẩn bị sẵn sàng cho sự cố

Exam Tips

CloudNut
Learn Clouds in a Nutshell

- Luôn luôn thiết kế hệ thống có sự dự phòng cho sự cố (Plan for failure)
- Sử dụng MultiAZ hoặc Regions
- Hiểu được sự khác biệt giữa MultiAZ và Read Replica của RDS
- Hiểu được sự khác biệt Scale In và Scale Out
- Luôn cân nhắc về Chi phí (Cost)

HA Architecture