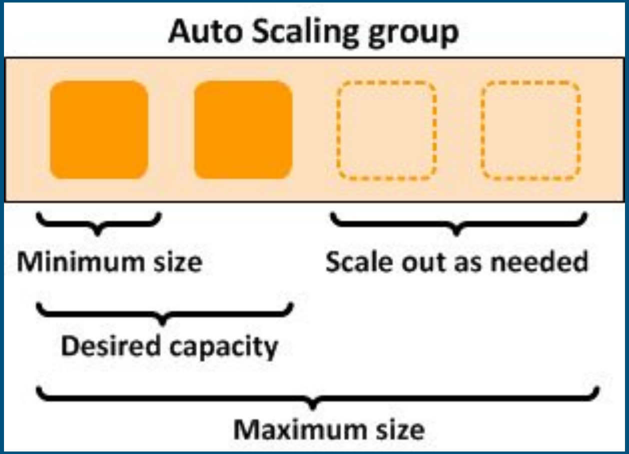
AWS Auto Scaling and Elastic Load Balancing

-By
Kaushik Rajak
Roll no -29

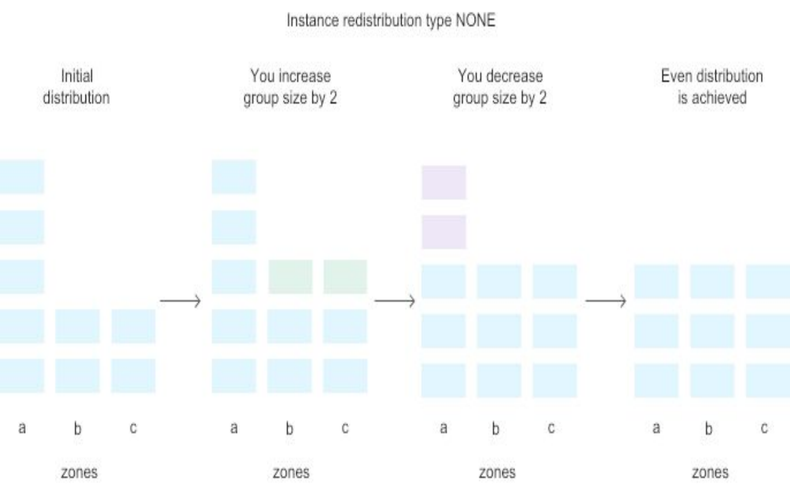
Auto Scaling

- A method used in AWS to automatically increase or decrease the number of computational resources (eg- web server, mainly EC2 instances in aws).
- **Autoscaling group -** A logical group of computational resources(EC2 instances) for the purpose of auto scaling.
- AWS clients can make a autoscaling group and can set its minimum size, desired capacity, maximum size, scale out conditions, scale in conditions etc.



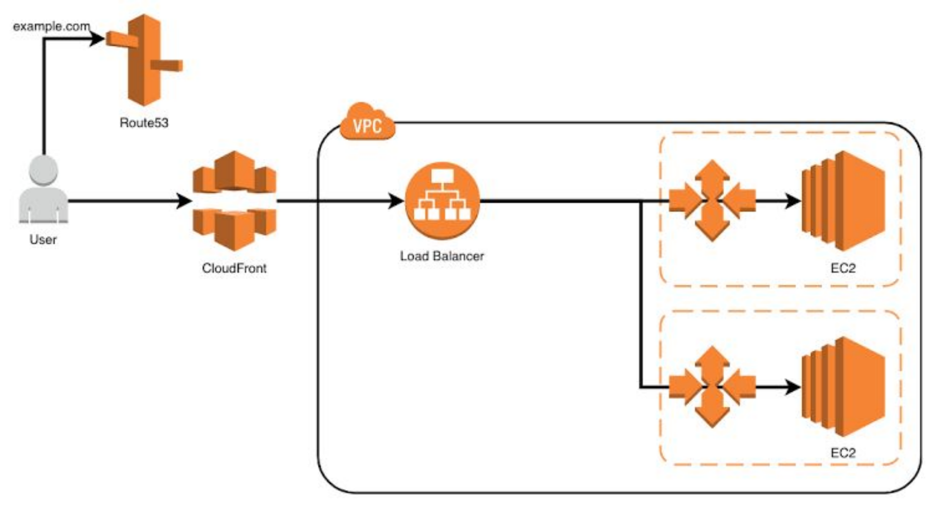
Properties of Auto Scaling :

- It is an horizontal scaling, i.e same type of EC2 instances are added at the time of scale out in a autoscaling group.
- We can define a autoscaling group in a region of AWS i.e a autoscaling group can not be expanded to more than one region
- It trys for maximum evenly distribution of EC2 instances over all the available zones (AZ) in a autoscaling group.


Instance redistribution type NONE
Initial distribution
You increase group size by 2
You decrease group size by 2
Even distribution is achieved
a
b
c
zones
a
b
c
zones
a
b
c
zones
a
b
c
zones

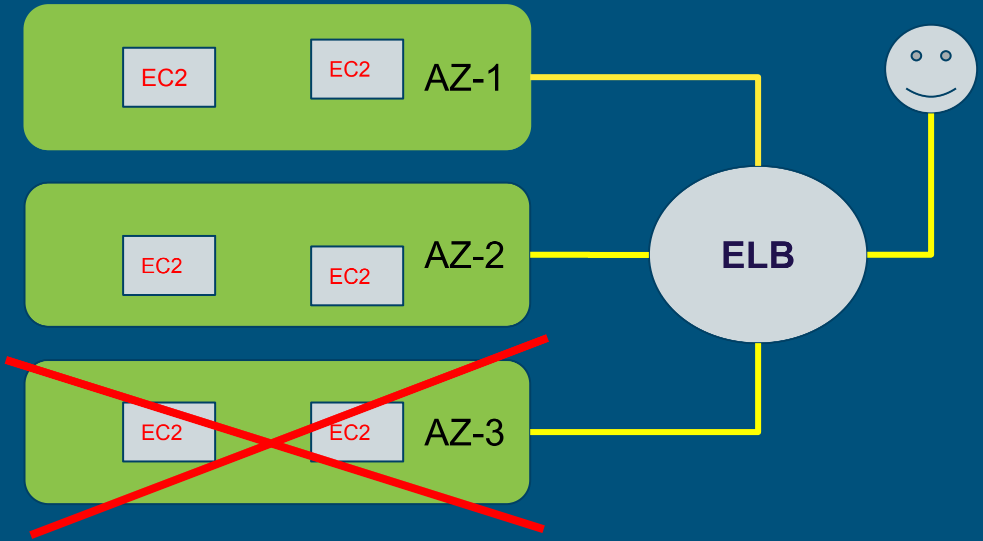
Elastic Load Balancing

- Elastic Load Balancing distributes incoming application or network traffic across multiple targets (EC2 instances) in all autoscaling groups attached to **Elastic load balancer (ELB).**
- It trys to evenly distribute the traffice accross all the available zones or EC2 instances(for cross zone load balancer)
- We can attach security group to ELB , for deciding which type of incaming traffic should pass through the ELB.



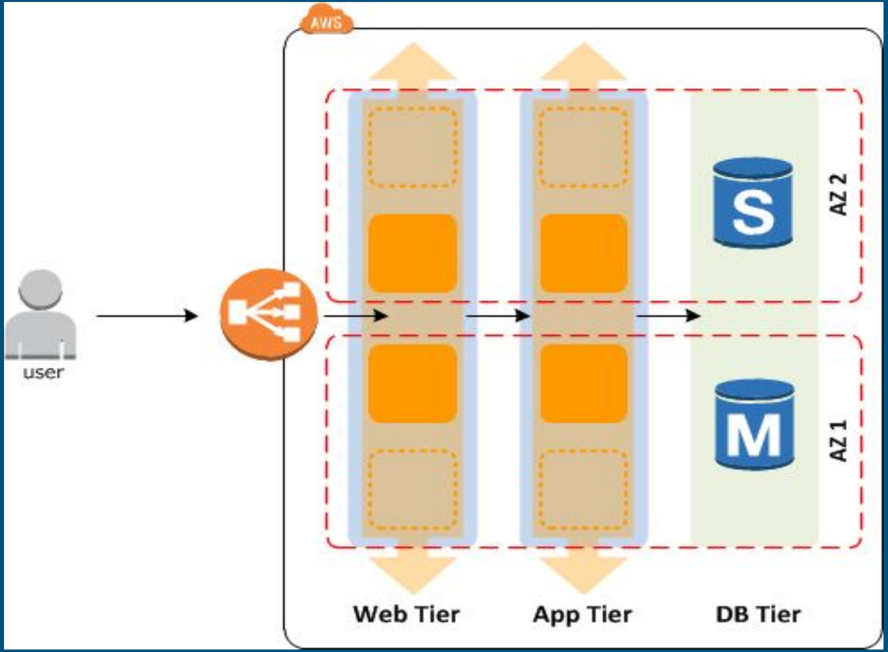
Why Auto Scaling and Load Balancing ?

- **High availability :** Autoscaling ensure high availability by scaling out.
- **High Faltu tolerance :** Due to evenly scaling over all AZs in a autoscaling group if one AZ gets destroyed then also other AZ's instances will be available.
- **Quic response :** Due to load balancing traffic get divided across EC2 instances.
- Better cost management.



Example: Web App Architecture

- Here we have many tiers of application.
- We can define different numbers of autoscaling groups in different tiers.
- Different autoscaling groups can have different scaling policies.
- These will ensure different need for different number of EC2 instances in every tiers to give high availability and high fault tolerance.
- By using ELB we can ensure high response of our web applications.



Thank You