# ELB & ASG Section 7

(Elastic Load Balancing & Auto Scaling Groups)

* High Availability, Scalability

- Scalability means that an application / system can handle greater loads by adapting

- There are two kinds of scalability:-
  ① Vertical Scalability
- ② Horizontal Scalability

- Scalability is linked but diffⁿ to High Availabi

① Vertical Scalability

- Vertical Scalability means increasing the size of the instance.

- forex:- your applⁿ runs on t2.micro Scaling that applⁿ vertically means running it on t2.large.

- Vertical scalability is very common for non-distributed systems, such as database.

- There's usually a limit to how much you can vertically scale (hardware limit)

Horizontal Scalability :-

- Horizontal Scalability means increasing number of instances / system for your appln.

- Horizontal scaling implies distributed systems

- This is very common for web appln / modern applications

## * High Availability

- High availability usually goes hand in hand with horizontal scaling.

- High availability means running your application / system in at least 2 Availability Zone

- The goal of high availability is to survive a data center loss (disaster)

## High Availability & Scalability for EC2

+ Vertical Scaling :- Size up instance

+ Horizontal scaling :- Use no. of instances
  - Auto Scaling Group
  - Load Balancer

+ High Availability :- Run instances for the same appln across A multi AZ.

# Scalability Vs Elasticity (Vs Agility)

- **Scalability** :- ability to accomodate a larger load by making the hardware stronger (scale up), or by adding nodes (scale out)

- **Elasticity** : Once a system is scalable, elasticity means that there will be some "auto-scaling" so that the system can scale based on the load. This is "cloud-friendly" : pay-per use, match demand, optimize cost.

- **Agility** :- (not related to scalability - disaster) new IT resource are only a click away, which means that your reduce the time to make those resources available to your developers from weeks to just minutes.

# Elastic Load Balancing (ELB) Overview

**Load Balancing :→** Load balancers are servers that forward internet traffic to multiple servers (EC2 instances) downstream

**Why use a Load Bdi Balancer?**

- Spread load across multiple downstream instances.
- Expose a single point of access (DNS) to your application.
- Seamlessly handle failures of downstream instances
- Do regular health checks to your instanc
- Provide SSL termination (HTTPs) for your websites.
- High availability across zones.

# What is Elastic Load Balancer?

+ An ELB (Elastic Load Balancer) is a managed load Balancer.
  - AWS guarantees that it will be working.
  - AWS takes care of upgrades, maintenance, high availability.
  - AWS provides only a few configuration knobs.

+ It costs less to setup your own load balancer but it will be lot more effort on your end (maintenance, integrations)

+ 4 kinds of load balancers offered by AWS:

1] Application Load balancer (HTTP/HTTPS only) - Layer 7

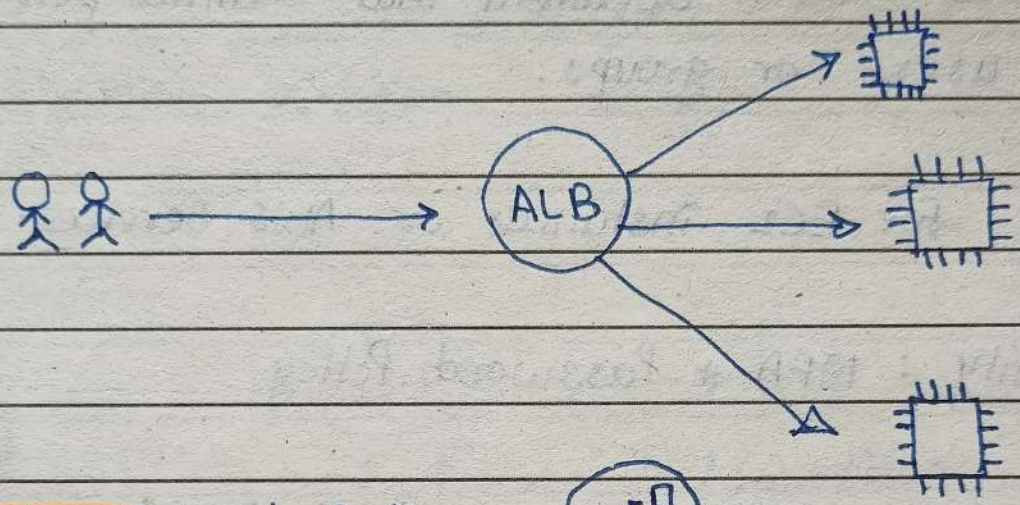
2] Network Load Balancer (ultra-high performance allows for TCP) - Layer 4

3] Gateway Load Balancer - Layer 3

4] Classic Load Balancer (retired in 2023) - Layer 4 & 7

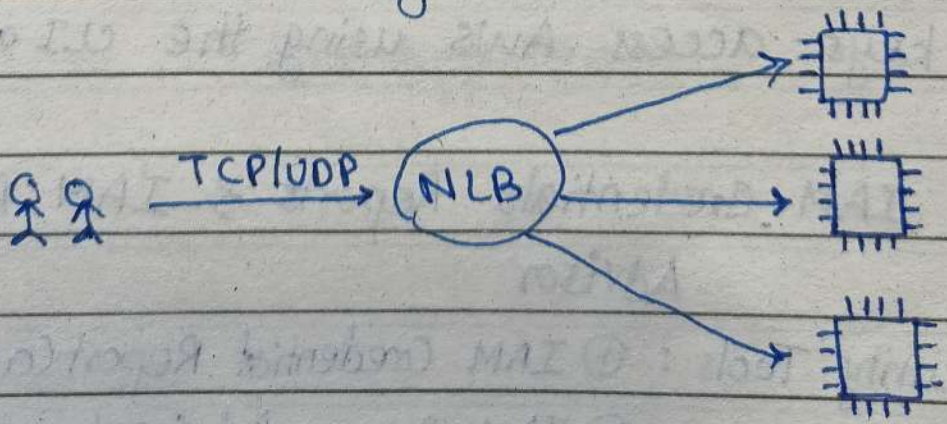
1] Application Load Balancer

- HTTP | HTTPS | gRPC protocols (Layer 7)
- HTTP Routing features
- Static DNS (URL)

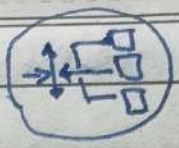



2] Network Load Balancer

- TCP | UDP Protocols (Layer 4)
- High Performance: millions of requests per seconds
- Static IP through Elastic IP



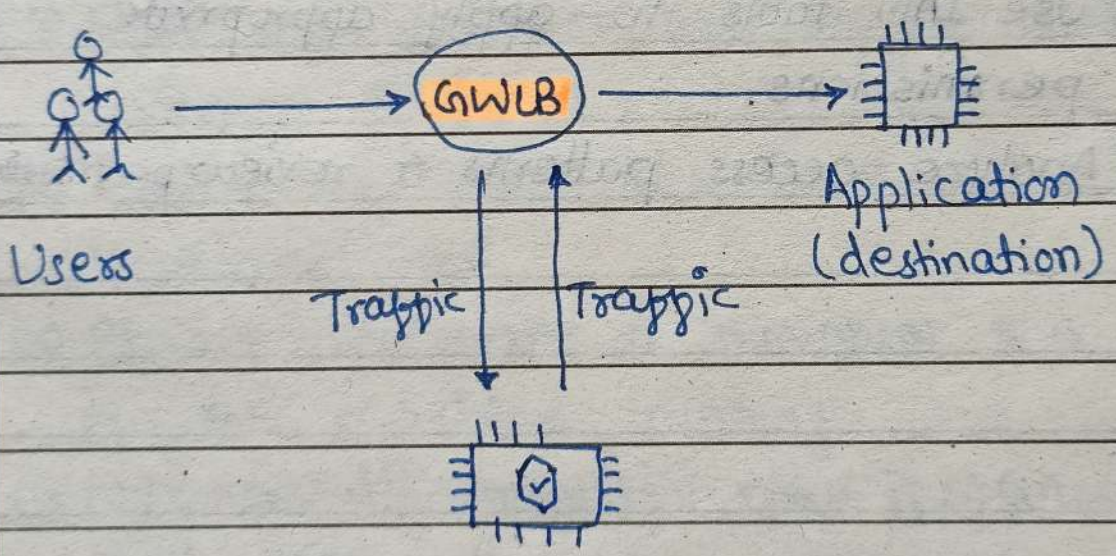
## 3] Gateway Load Balancer



- GENEVE Protocol on IP Packets (Layer 3)

- Route traffic to firewalls that you manage on EC2 instance.

- Intrusion detection.




3rd party security
Virtual Appliances

# Auto Scaling Group (ASG) Overview

- In real life, the load on your websites and application can change

- In the cloud, you can create, & get rid of servers very quickly

- The goal of an Auto Scaling Group (ASG) is to:

① Scale out (add EC2 instances) to match increased load.
② Scale in (remove EC2 instances) to match decreased load.
③ Ensure we have a minimum & a max. no. of machine running
④ Automatically register new instances to a load balancer
⑤ Replace unhealthy instances

- Cost saving:- Only run at an optimal capacity (principle of the cloud)

# Auto scaling Groups - Scaling Strategies

⊙ Manual Testing : Update the size of an ASG Scaling manually.

⊙ Dynamic Scaling :- Respond to changing demand

+ Simple / Step Scaling:
- When a cloudWatch alarm is triggred (ex. CPU > 70%.), then add 2 units

+ Target Tracking Scaling:
ex: I want the average ASG CPU to stay at around 40%.

+ Scheduled Scailing :
- Anticipate a scaling based on known usage patterns.
- ex: Increase the min. capacity to 10 at 5pm on friday.

## ⊙ Predictive scaling :-

- Use Machine learning to predict future traffic ahead of time.

- Automatically provisions the right no: of EC2 instances in advances.

- Useful when your load has predictable time based patterns.

# ELB & ASG Summary

① High Availability vs Scalability (vertical & horizontal) vs Elasticity vs Agility in the cloud.

② Elastic Load Balancers (ELB)

- Distribute traffic across backend EC2 instances, can be Multi-AZ.
- Supports health checks.
- 4 types: Classic (old), ALB (HTTP - L7), NLB (TCP - L4), GWLB (L3)

③ Auto Scaling Group (ASG)

- Implement Elasticity for your application, across multiple AZ.
- Scale EC2 instances based on the demand on your system, replace unhealthy integrated with the ELB.