ELB and ASG.

Elastic load Balancing & Autoscaling groups.

Scalability: ability to handle greater loads by adapting.

Vertical
↑ size of the instance.

- non distributed systems eg DB
- limit of hardware on how much it can scale.
- scale up or down

Horizontal
↑ no. of instances

⇒ distributed systems.
-> web apps, modern apps.

- scale out/in ↑ ↓
- ASG & LB (load balancer)

High Availability:
- running your application/system in atleast 2 availability zones
- saves from disasters/ data center loss
- done with - 1) ASG multi AZ
  2) LB in multi AZ

Scalability
- ability to accomodate larger load by making the hardware stronger.
  scale up / scale out

Elasticity
- once a system is scalable elasticity means that there will be some autoscaling so that the system can scale based on the load.
- cloud friendly
- pay per use, match demand, optimize costs

Agility
- need IT resources are click away
- reduced time to make resources available to developers from weeks to minutes.

Load Balancing - servers that forward internet traffic to multiple servers (EC2 instances) downstream.

ELB - Elastic load Balancer - managed by AWS

Why use LB?
- 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 instances.
- Provide SSL termination (HTTPS) for your websites.
- High availability across zones.

Why use ELB?

- AWS managed
- AWS guarantees it will be working.
- AWS takes care of upgrades, maintenance, high availability.
- AWS provides only a few configuration knobs.

- costs less to set up your own lb but maintenance, integrations efforts ↑.

3 Types of LB by AWS

Application LB
- HTTP/HTTPS only
- layer 7.

Network LB
- ultra high performance
- allows for TCP
- layer 4

Classic LB
- layer 4 and 7.
- slowly retiring

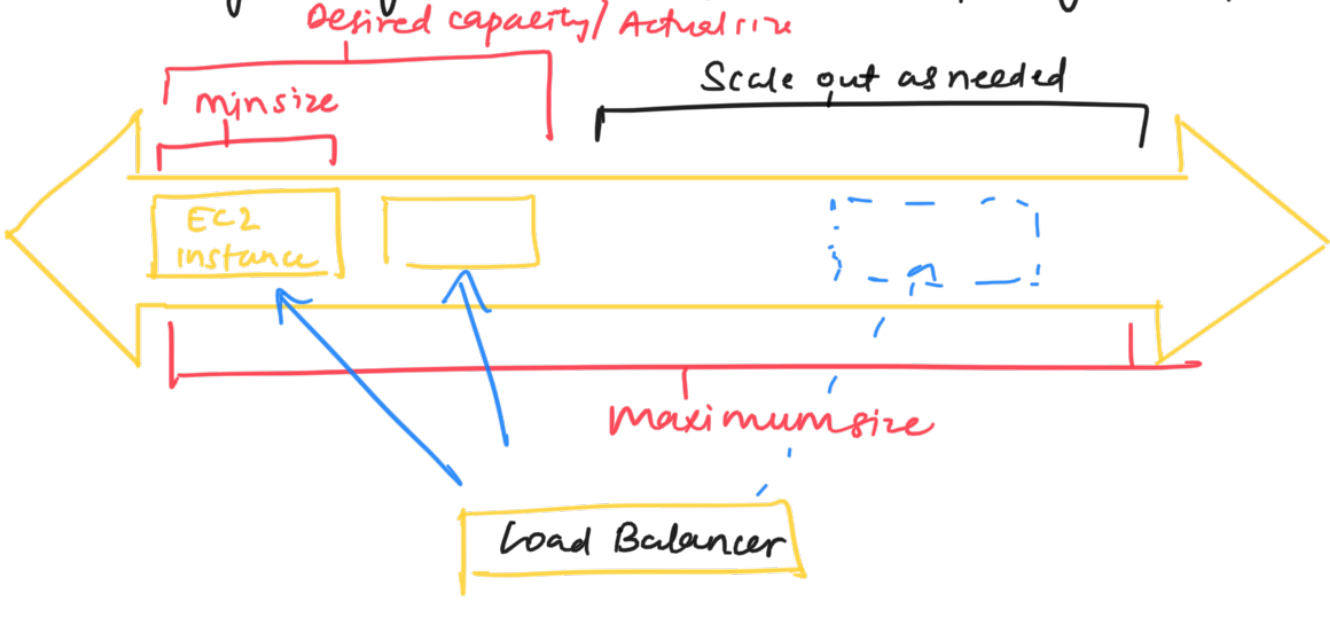
Gateway Load Balancer - added in 2020

AutoScaling Group (ASG)

Goals:

1. Scale out to match increased load
2. Scale in to match decreased load.
3. Ensure we have a min and max no. of machines running.
4. Automatically register a new instance to a(lb) Load balancer
5. Replace unhealthy instances

* Cost savings - only run at an optimal capacity (principle of the cloud)



Scaling Strategies for ASG:

1) Manual scaling: Update the size of ASG manually.

2) Dynamic Scaling: Respond to changing demand
   i) Simple/Step Scaling: eg: when cloudwatch alarm is triggered (CPU > 70% etc), then

add / remove etc.

ii) Targeted Tracking Scaling: eg: keep avg CPU around 40%.

iii) Scheduled Scaling: anticipate based on usage pattern
eg: more traffic on black friday, so scale

3) Predictive Scaling:
- use ML. to predict future traffic ahead of time
- automatically provisions the right no of EC2 instances in advance.
- useful when your load has predictable time based patterns.