01 Why Auto Scaling?

02 What is Auto Scaling?

03 Why ELB?

04 What is ELB?

05 ELB Hands-on

06 How to get Started

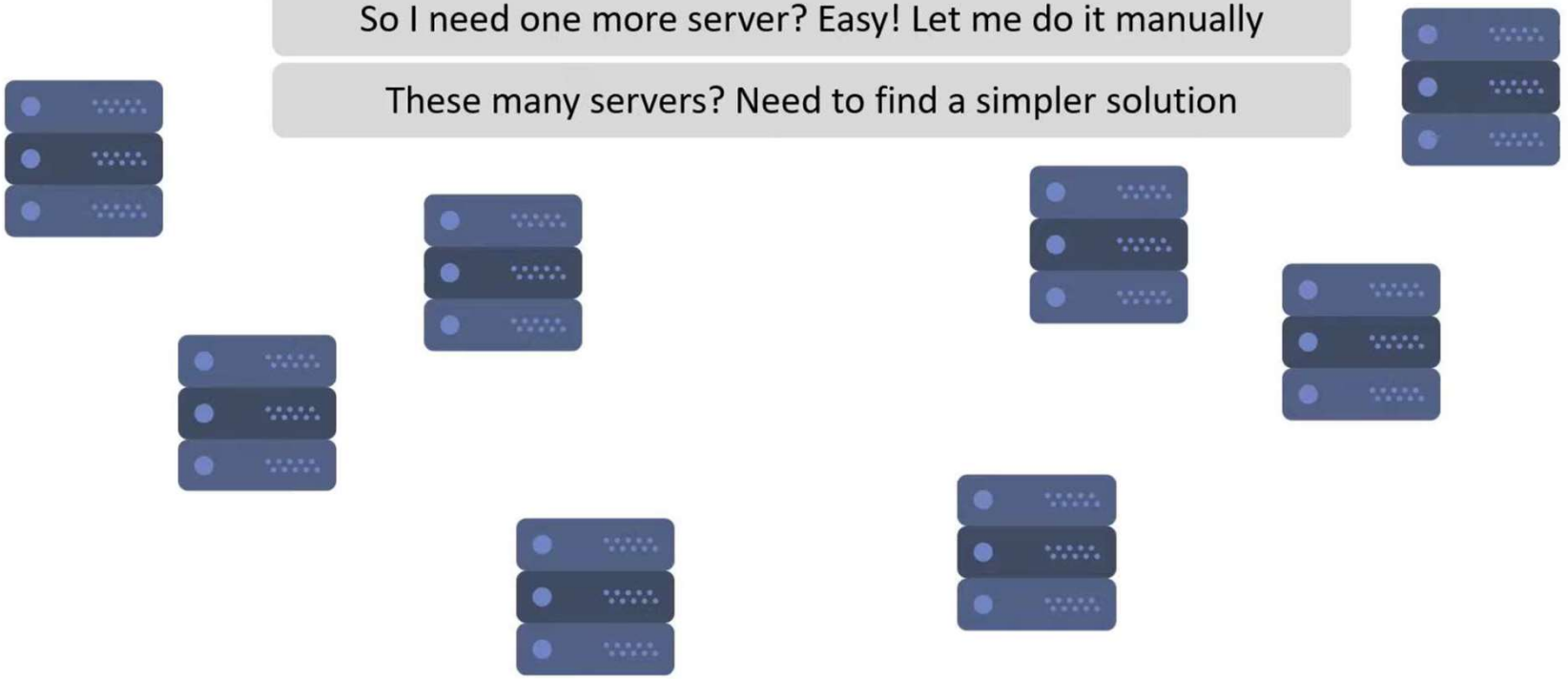
Why Auto Scaling?

What would you do if your website's traffic increased and it needs more servers?

Why Auto Scaling?

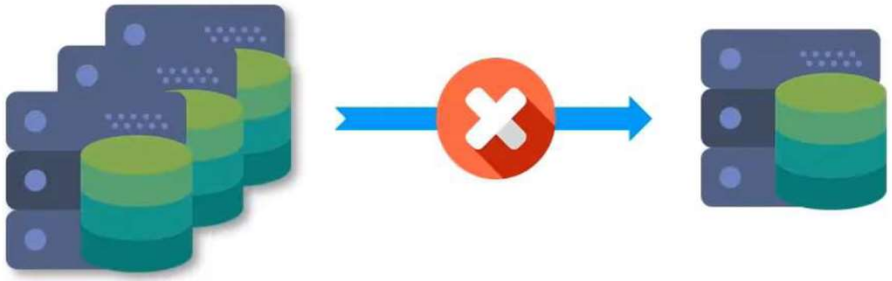
So I need one more server? Easy! Let me do it manually

These many servers? Need to find a simpler solution



Why Auto Scaling?

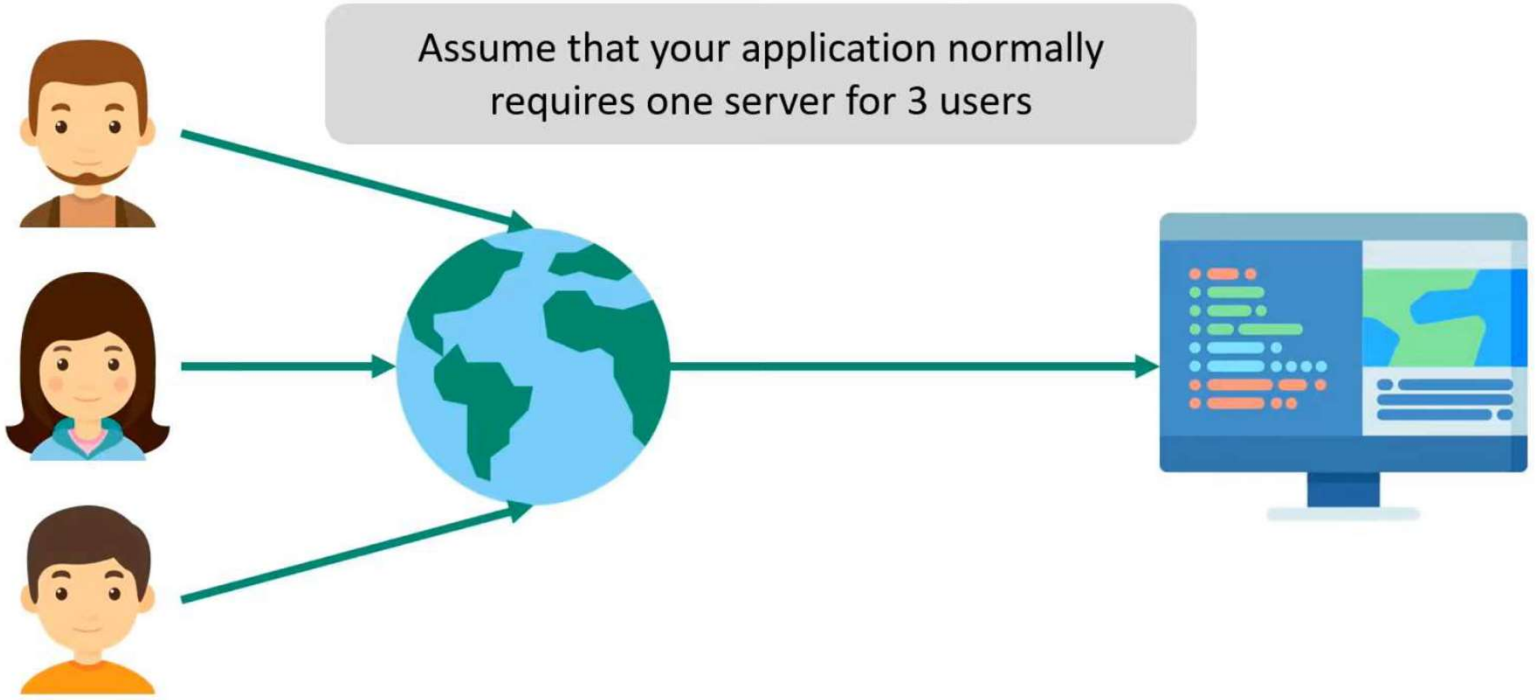
If it's a on-premises set up then we have to do it manually and it would be very hard if they wanted to scale down their infrastructure



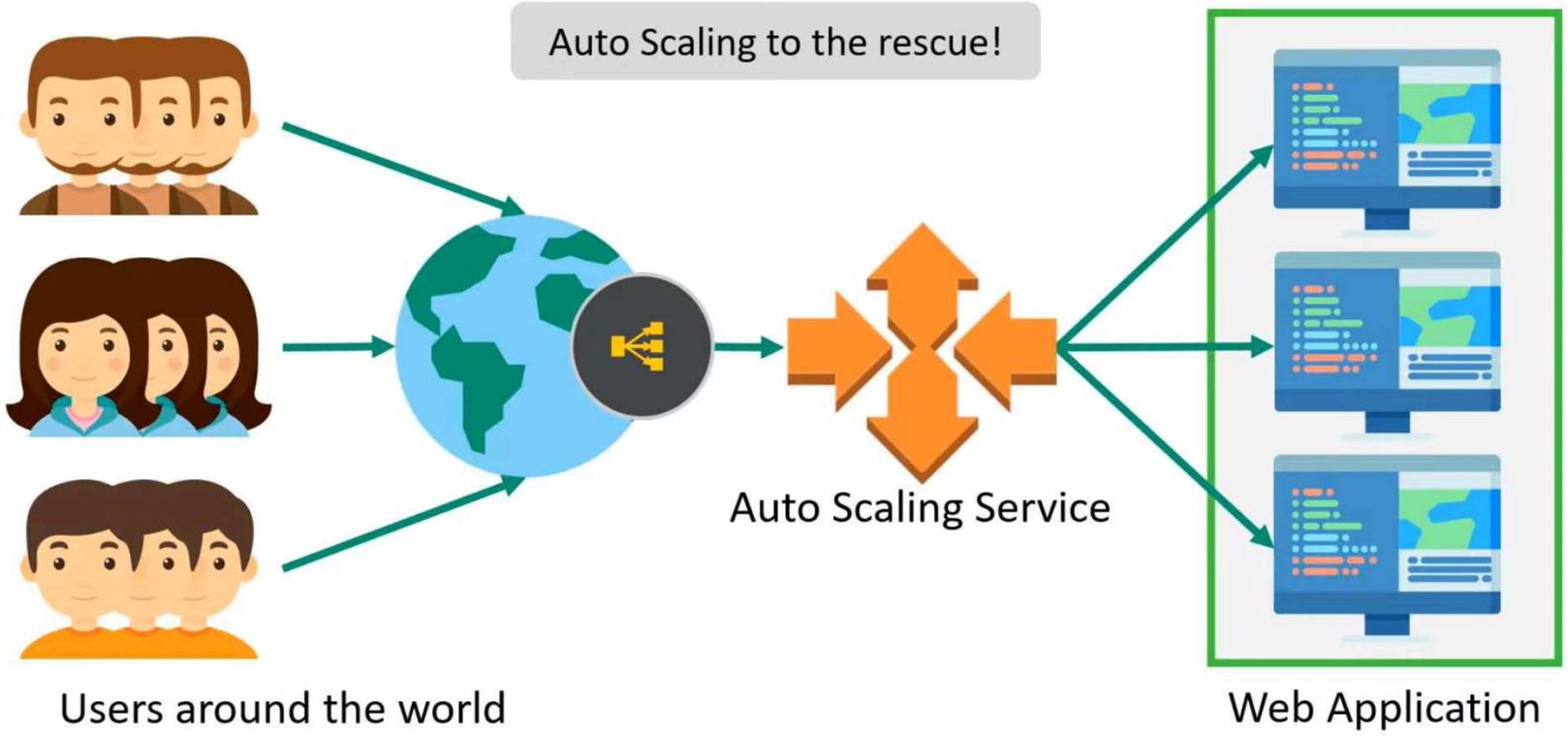
Why Auto Scaling?

Then AWS came with the Idea of Automating the scaling process dynamically!

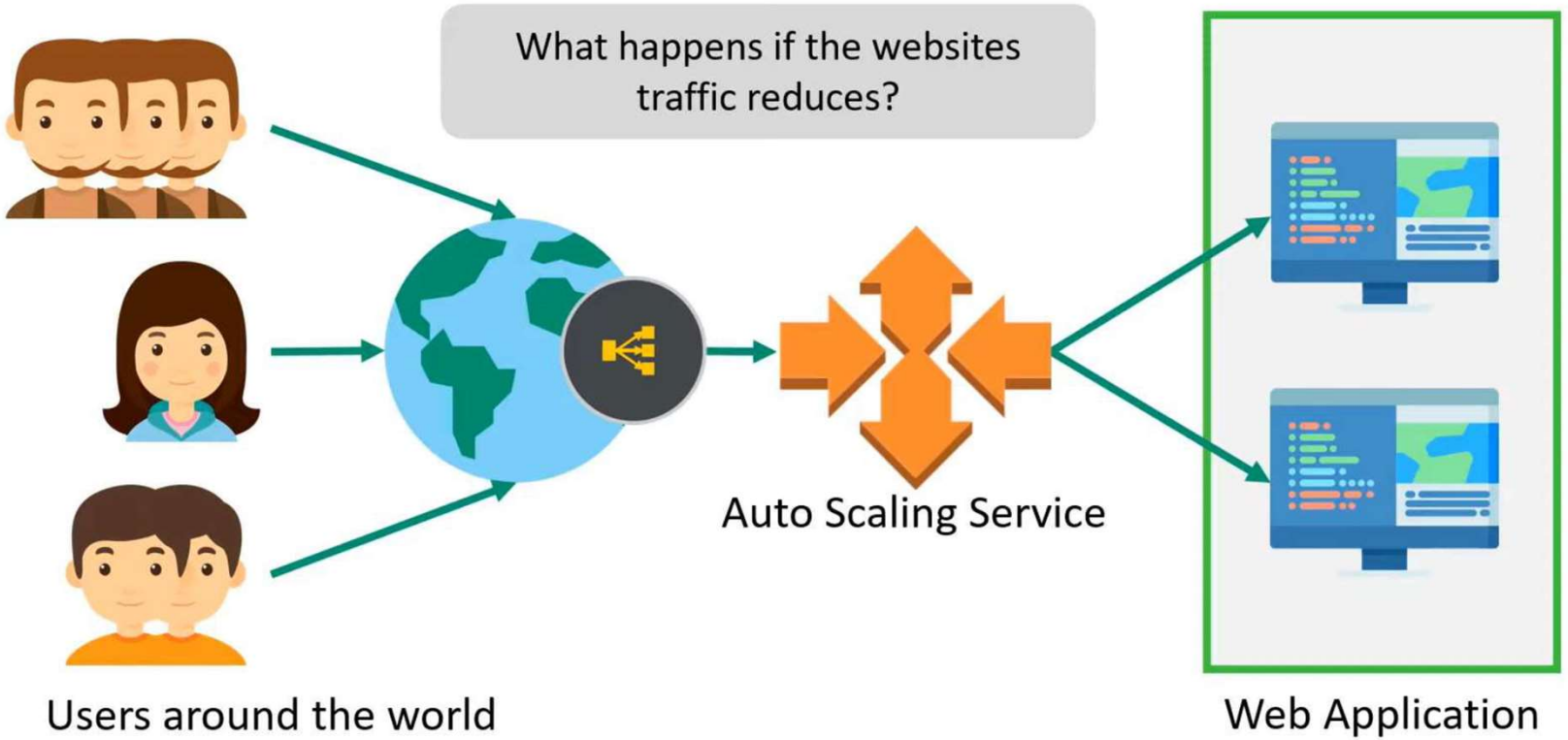
Why Auto Scaling?



Why Auto Scaling?

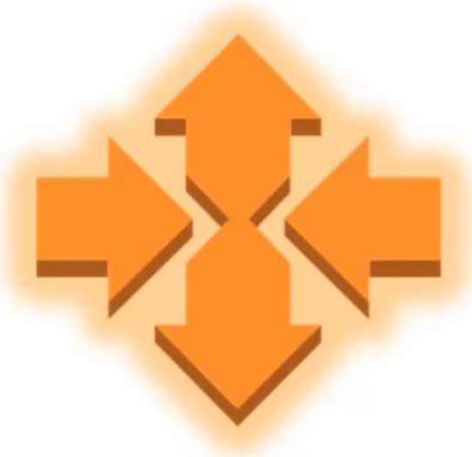


Why Auto Scaling?

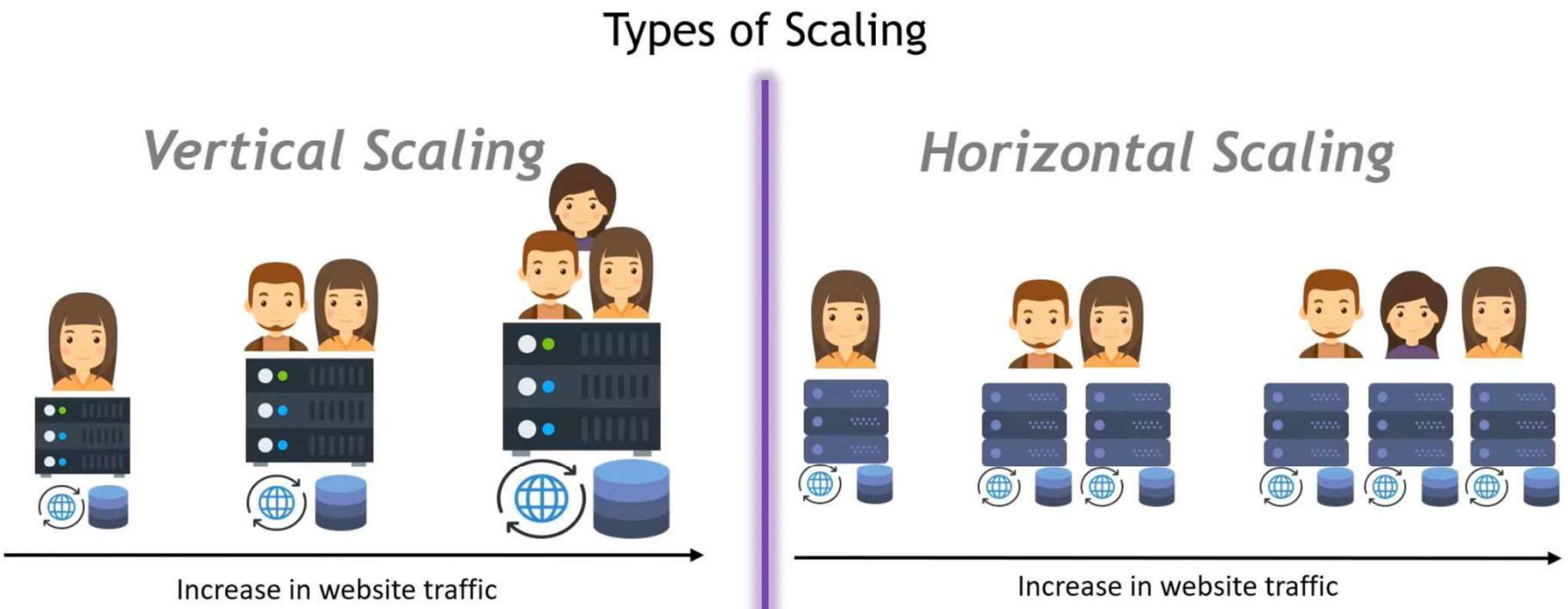


What is Auto Scaling?

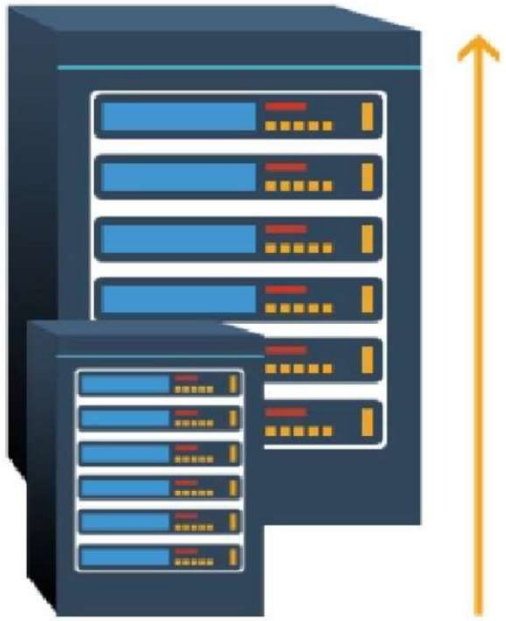
AWS Auto Scaling monitors your applications and adjusts the capacity according to the traffic for providing a steady performance and high availability



What is Auto Scaling?



What is Auto Scaling?



Vertical Scaling
(Scaling up)

Horizontal Scaling
(Scaling out)

To Generate the Traffic manually
We need the below application

apt install stress

stress --cpu 4

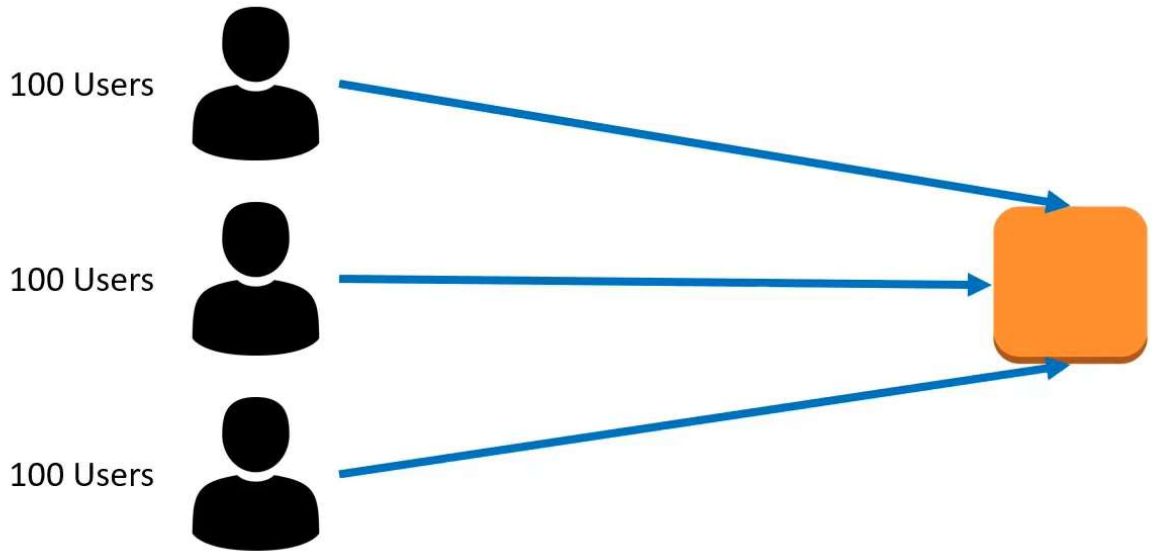
Why ELB?

Why Elastic Load Balancing?

Assume your application runs on multiple EC2 instances, how do you know that the traffic is distributed evenly among these instances?

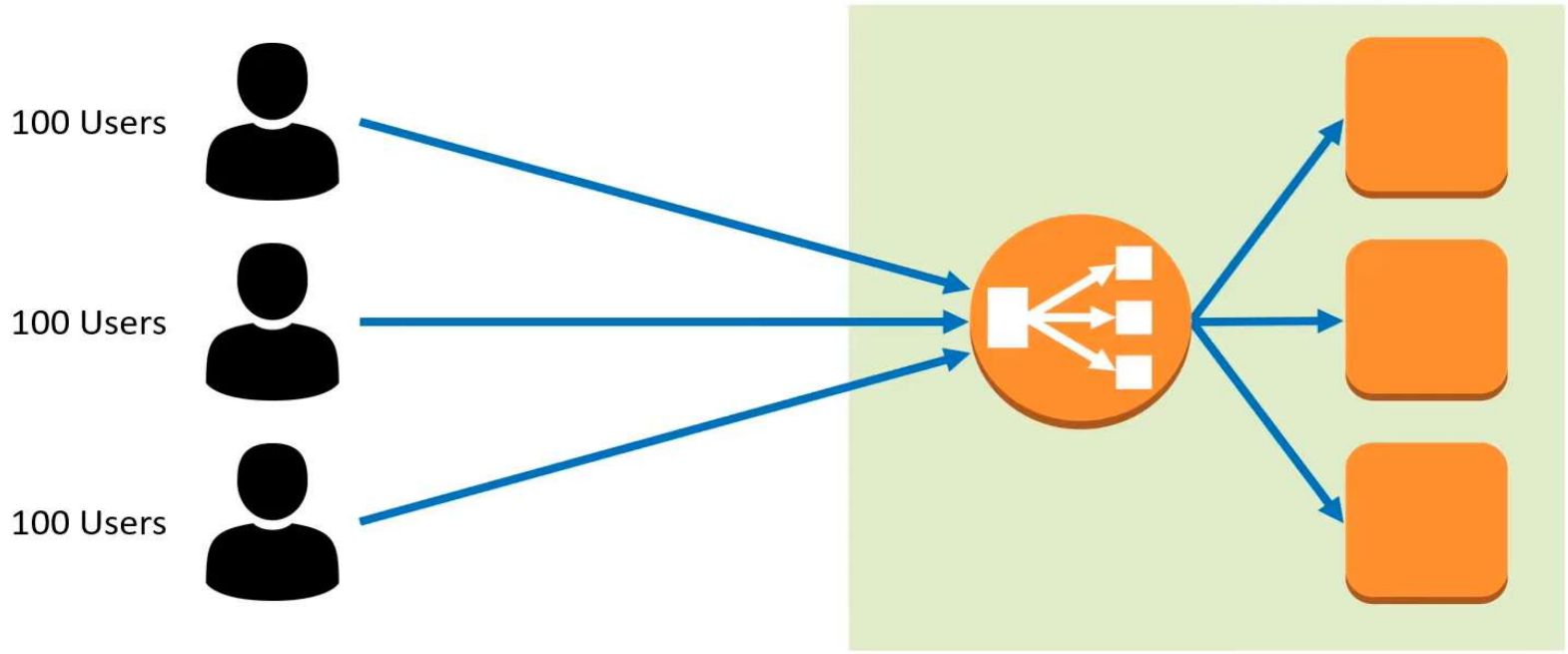
Why Elastic Load Balancing?

For 100 users one instance works fine, what if there are 300 users and still running on one instance?



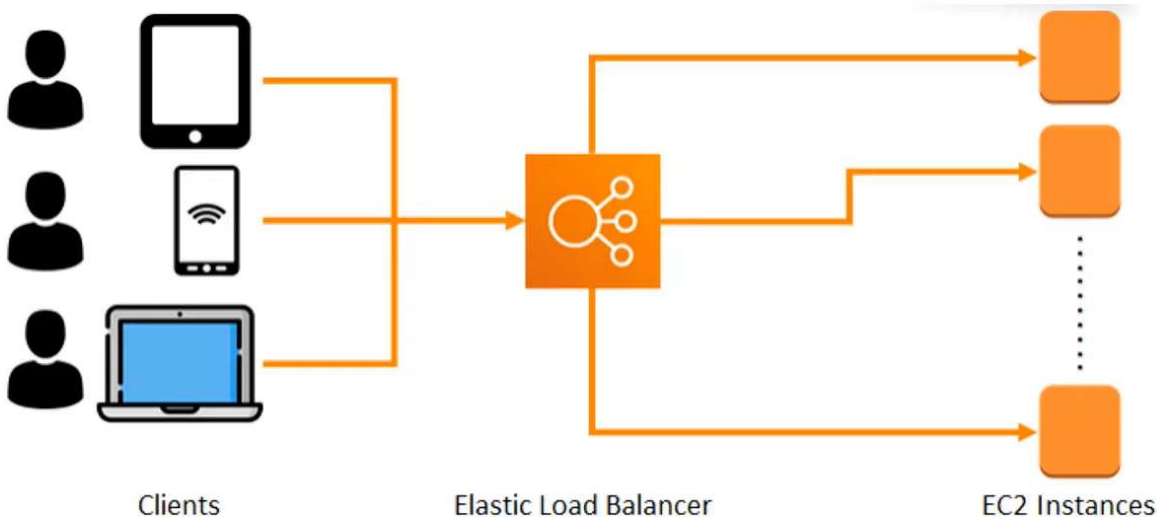
Why Elastic Load Balancing?

Now, ELB comes into play. ELB basically receives all user requests and then equally distributes the workload across all the EC2 instances.

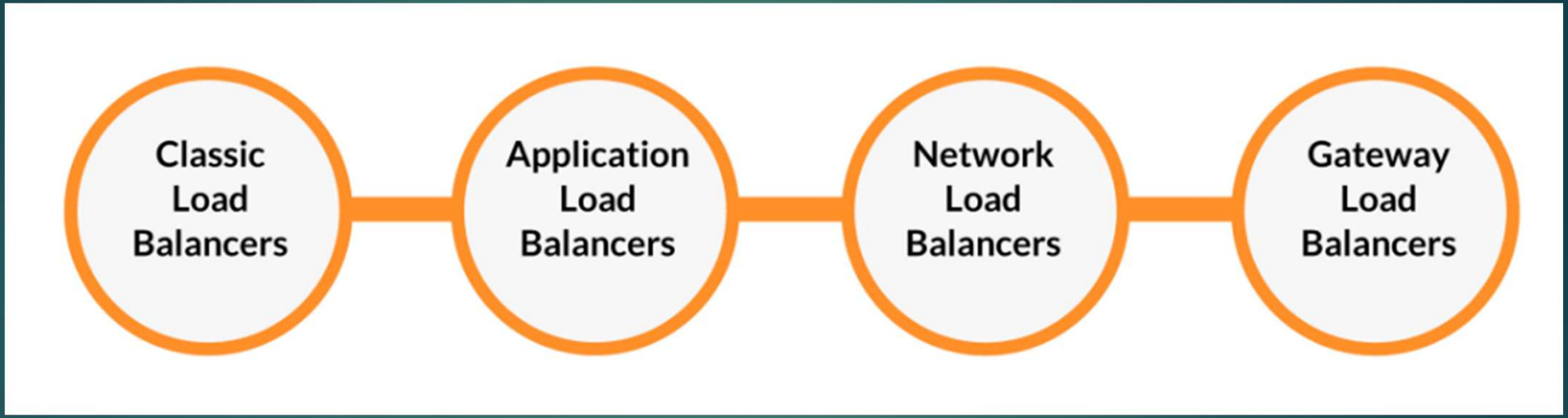


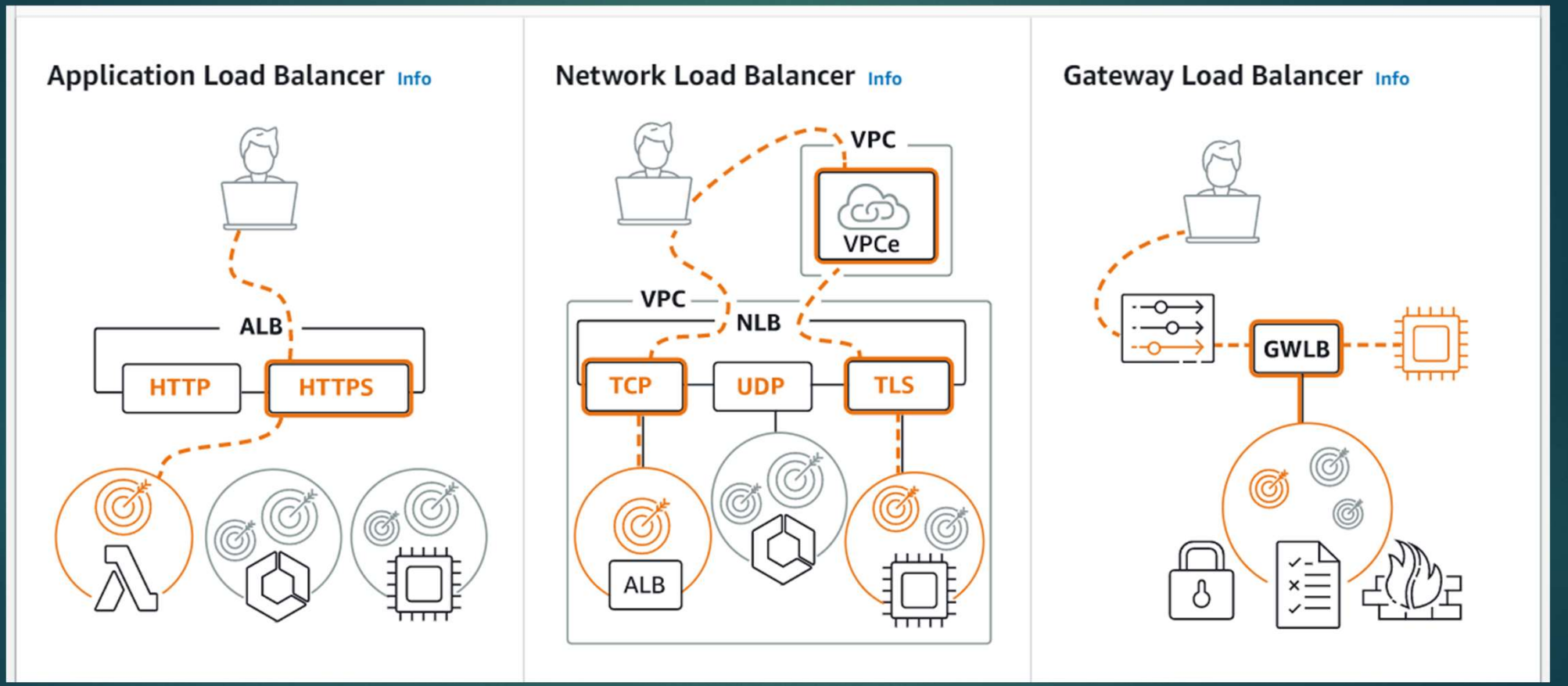
What is ELB?

In the simplest terms, Elastic Load Balancer accepts incoming traffic from its clients and then routes requests to the targets which the client want



Types of Load Balancer:




Application Load Balancer Info
ALB
HTTP
HTTPS
Network Load Balancer Info
VPC
VPCe
VPC
NLB
TCP
UDP
TLS
ALB
Gateway Load Balancer Info
GWLB

Classic Load Balancers

Classic Load Balancers distribute upcoming traffic to different EC2 instances in multiple Availability Zones. During this process, there is a chance of fault tolerance of your application. These Load Balancers detect healthy and unhealthy instances and direct the traffic towards only healthy ones.

Application Load Balancers

- The Load Balancer that distributes the traffic to target groups on the basis of content is called Application Load Balancer.
- Supports web sockets, HTTP, HTTPS, and microservices and container-based applications, including deep integration with EC2 container service.
- Support for path-based and host-based routing. Also, provide routing requests to multiple applications on a single EC2 instance.

Network Load Balancers

- 1. Network load balancer allows to:
- Forward TCP & UDP traffic to your instance
- Handle millions of requests per second
- Less Latency – 100ms (vs 400ms)
- NLB has one static IP per AZ and supports Elastic IP
- Not including in aws free tier

Gateway Load Balancer

•It makes deploying, scaling, and managing your third-party virtual appliances easy.

The AWS Gateway Load Balancer takes care of routing traffic to the appropriate virtual appliance in your network, instead of traffic going directly to virtual appliances. With the Gateway Load Balancer, traffic is routed to healthy virtual appliances and rerouted away from failing ones.