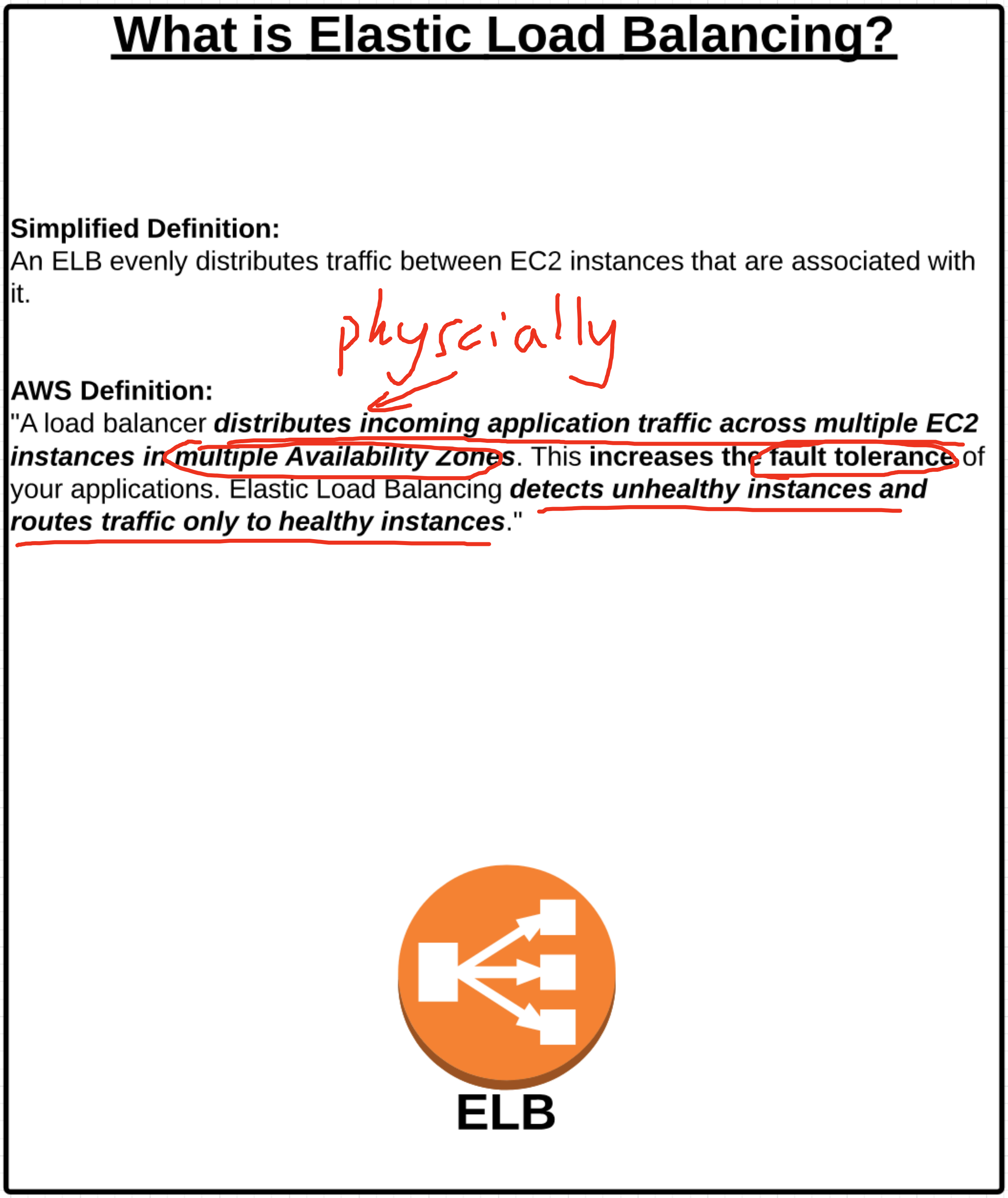
ELB & AS

2019年2月10日 星期日　　上午1:46



What is Elastic Load Balancing?

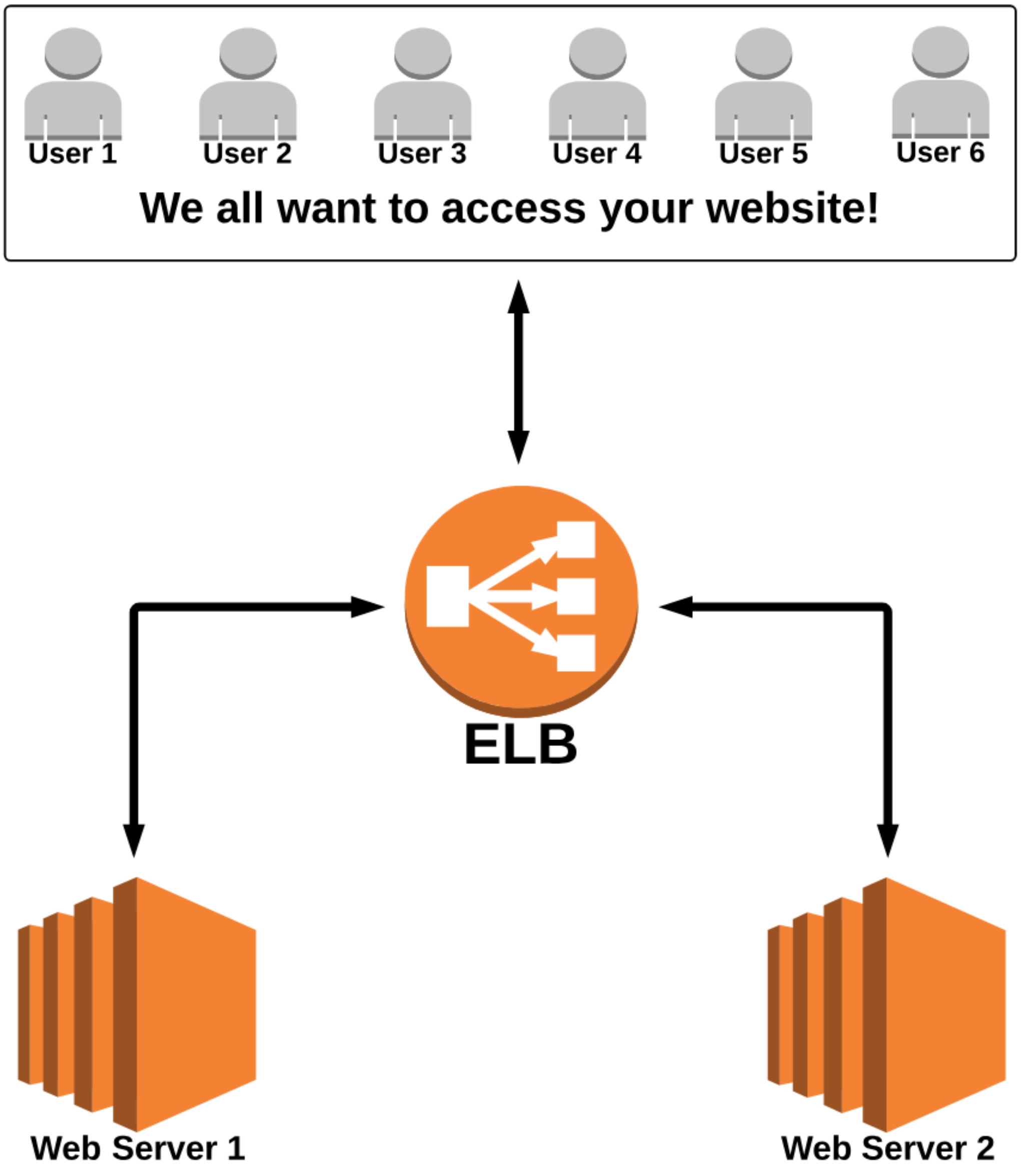
Simplified Definition:
An ELB evenly distributes traffic between EC2 instances that are associated with it.

physcially

AWS Definition:
"A load balancer ***distributes incoming application traffic across multiple EC2 instances in multiple Availability Zones***. This **increases the fault tolerance** of your applications. Elastic Load Balancing ***detects unhealthy instances and routes traffic only to healthy instances***."

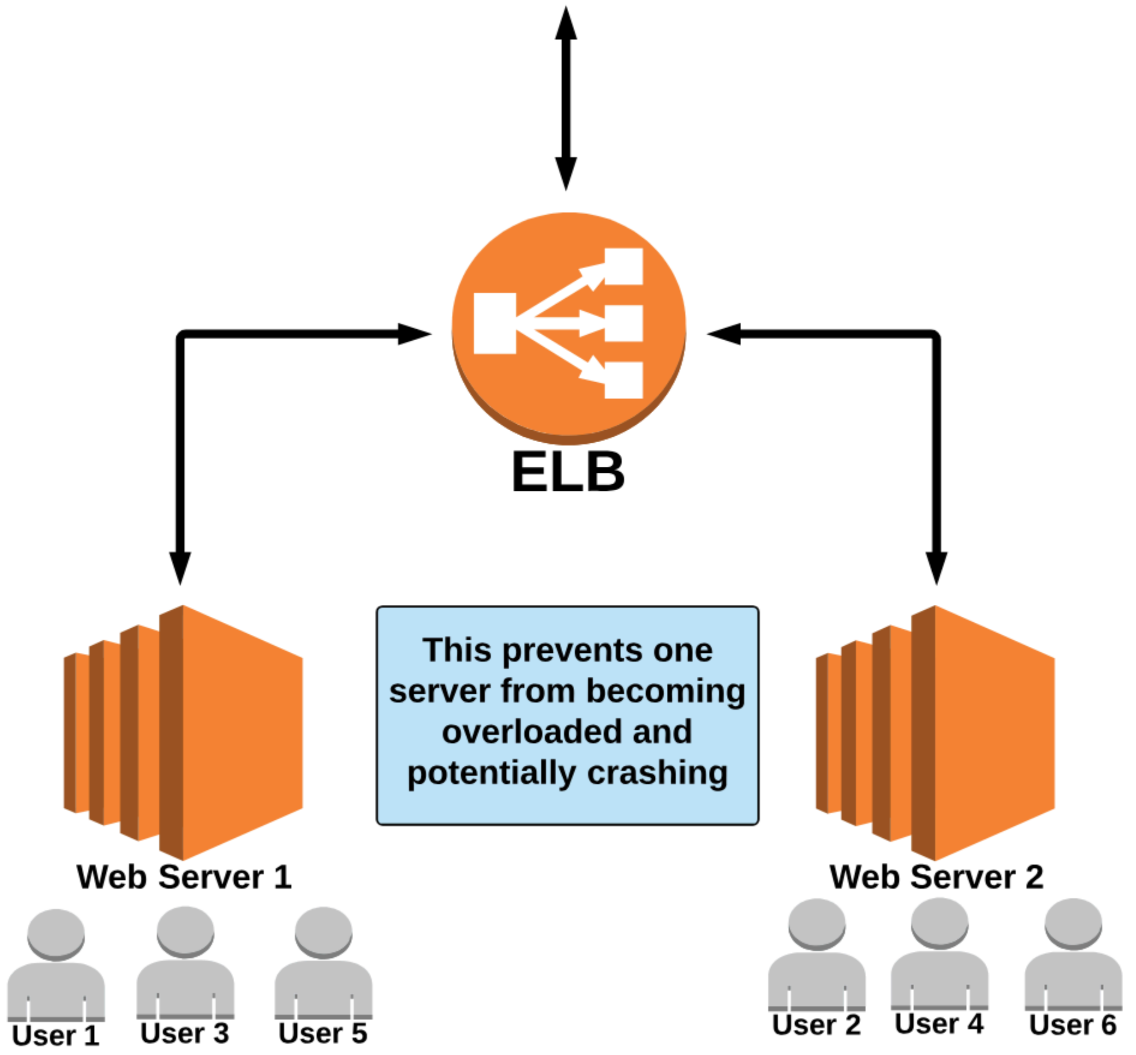
ELB Basics:

ELB Basics:



ELB Basics:

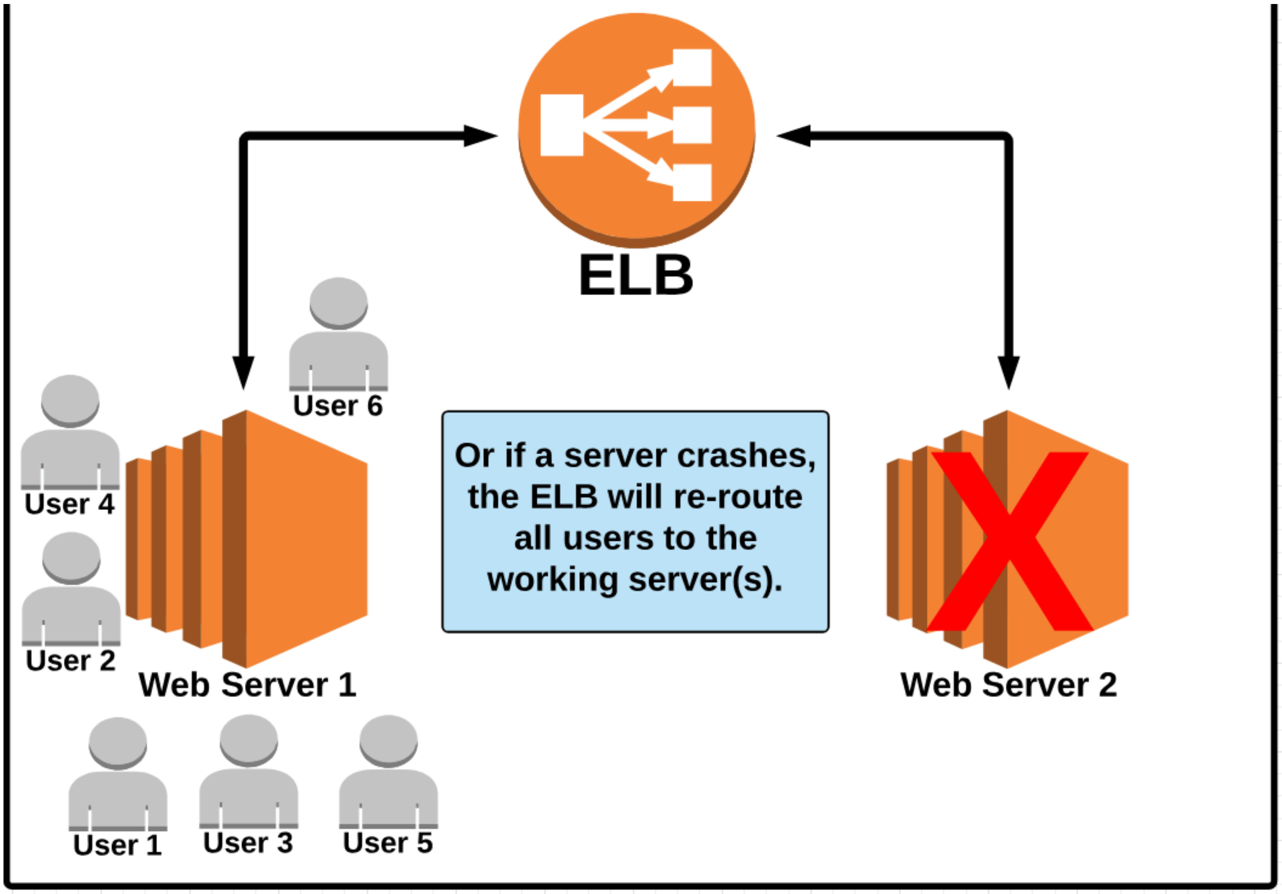
We all want to access your website!



This prevents one server from becoming overloaded and potentially crashing

ELB Basics:

We all want to access your website!



Or if a server crashes, the ELB will re-route all users to the working server(s).

ELB Basics:

Elastic Load Balancing

is a foundational component of

High Availability

and

Fault Tolerance

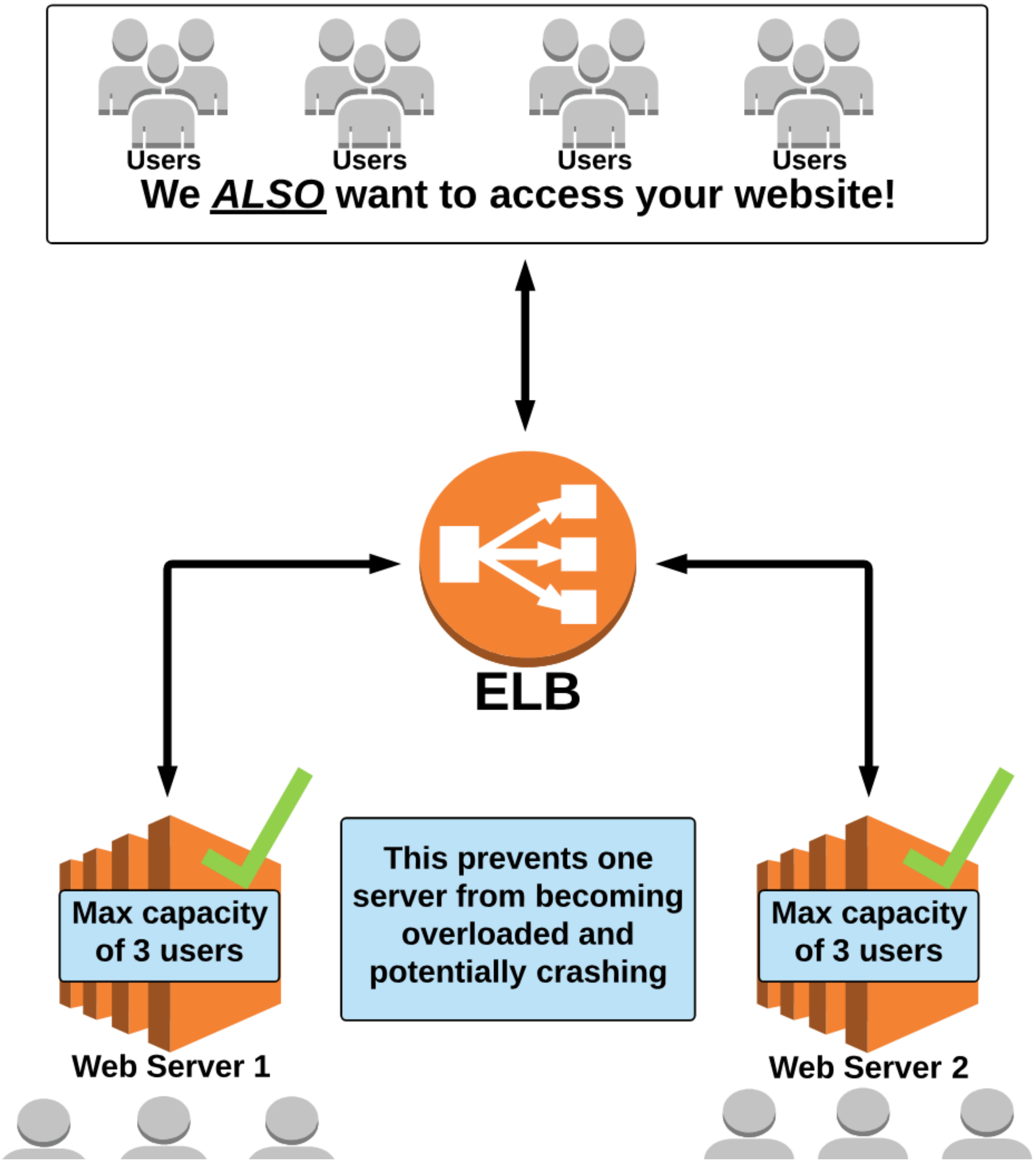
We now know that Elastic Load Balancing can evenly distribute traffic between all active servers - but what happens if demand (traffic) is so high that the active services can't handle it?

What is Auto Scaling?

Simplified Definition:
Auto Scaling automates the process of adding (***scaling up***) OR removing (***scaling down***) EC2 instances ***based on traffic demand*** for your application.

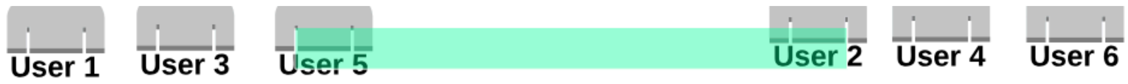
AWS Definition:
"Auto Scaling helps you ensure that you have the correct number of Amazon EC2 instances available to ***handle the load for your application***. You create collections of EC2 instances, called ***Auto Scaling groups***. You can specify the minimum number of instances in each Auto Scaling group, and Auto Scaling ensures that your group never goes below this size. You can specify the maximum number of instances in each Auto Scaling group, and Auto Scaling ensures that your group never goes above this size. If you specify the desired capacity, either when you create the group or at any time thereafter, Auto Scaling ensures that your group has this many instances. If you specify scaling policies, then Auto Scaling can launch or terminate instances as demand on your application increases or decreases."

<u>Auto Scaling Basics:</u>



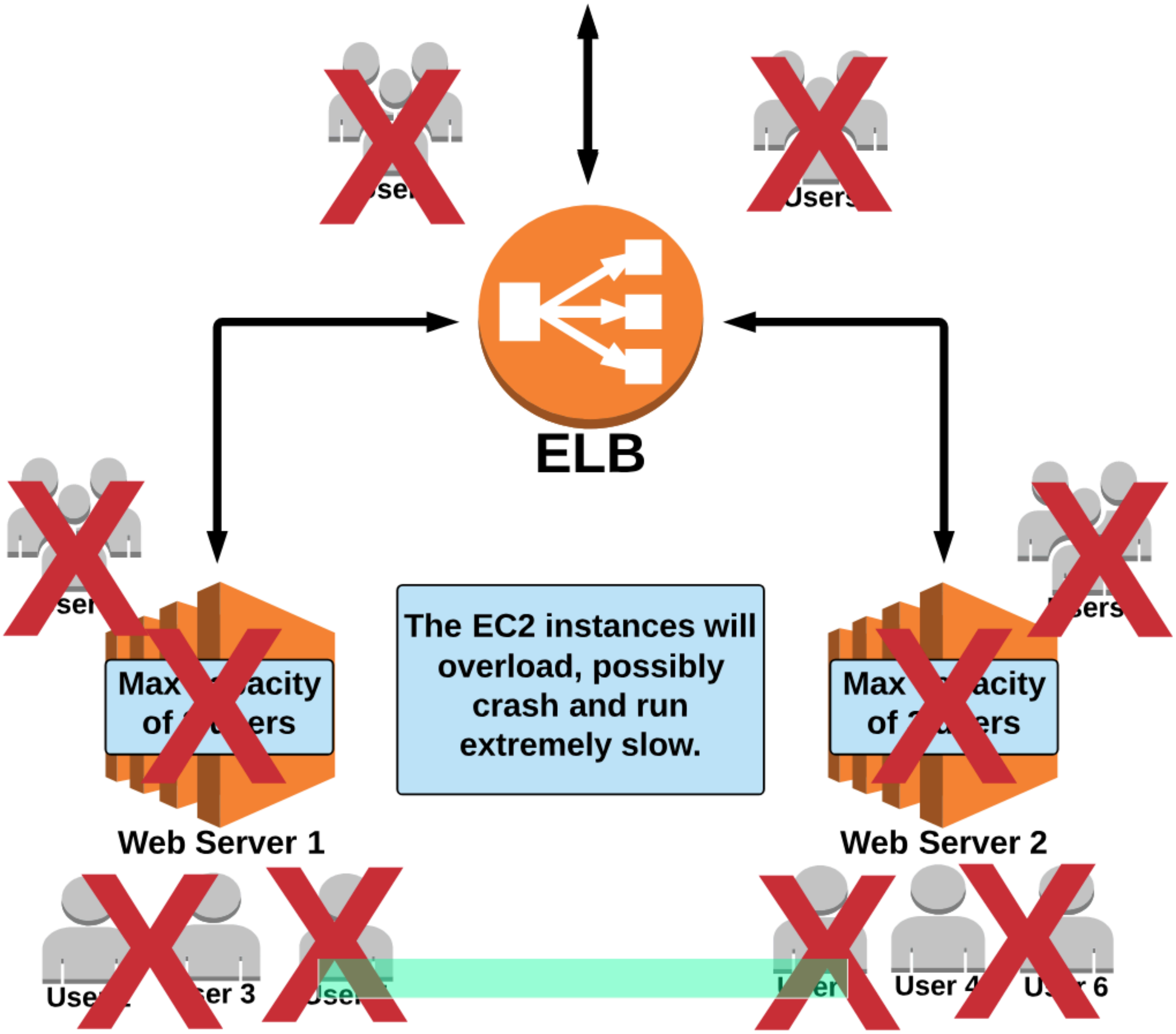
We ***<u>ALSO</u>*** want to access your website!

This prevents one server from becoming overloaded and potentially crashing



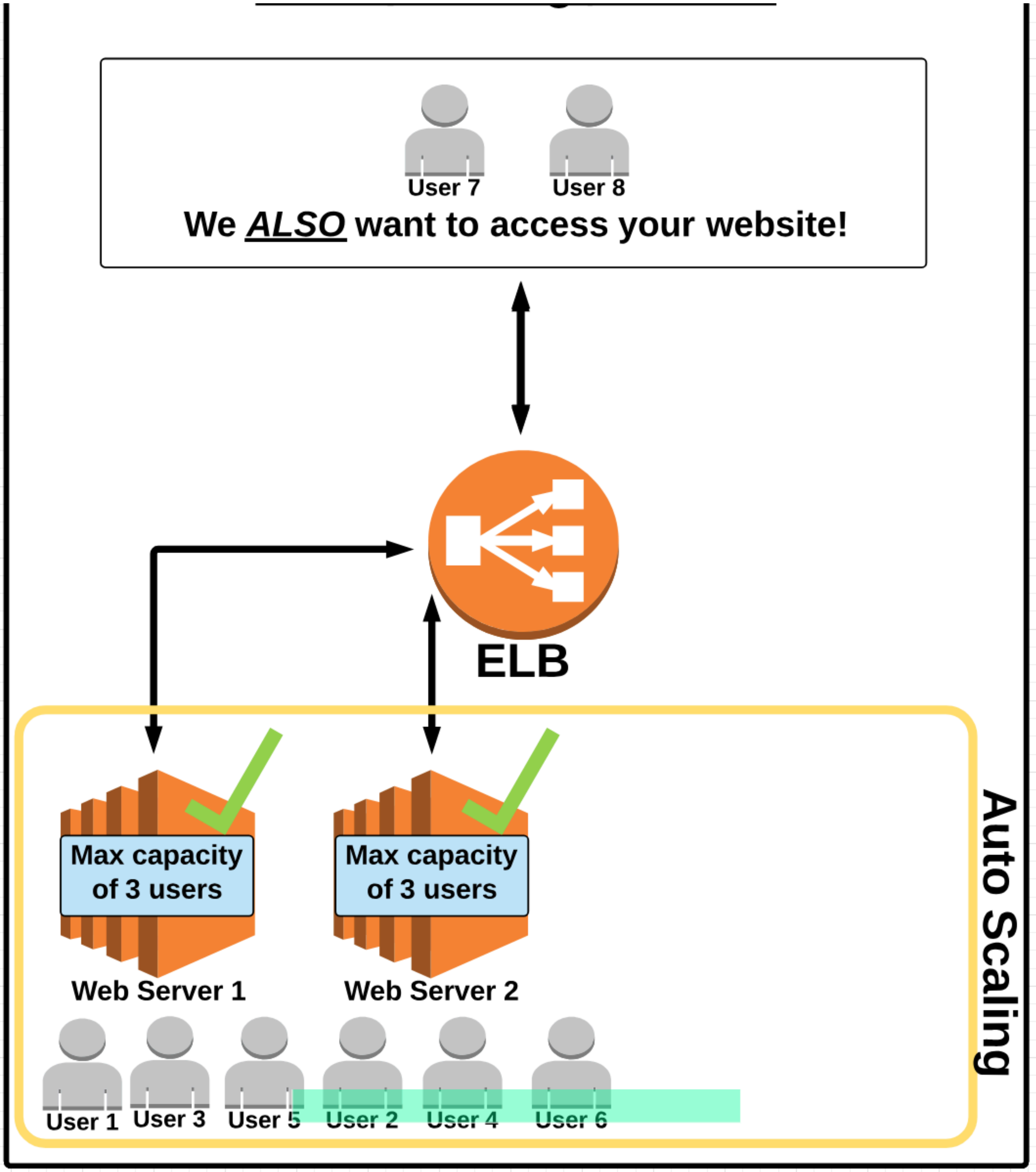
Auto Scaling Basics:

We all want to access your website!



The EC2 instances will overload, possibly crash and run extremely slow.

Auto Scaling Basics:



User 7 User 8

We ***ALSO*** want to access your website!

ELB

Max capacity of 3 users

Web Server 1

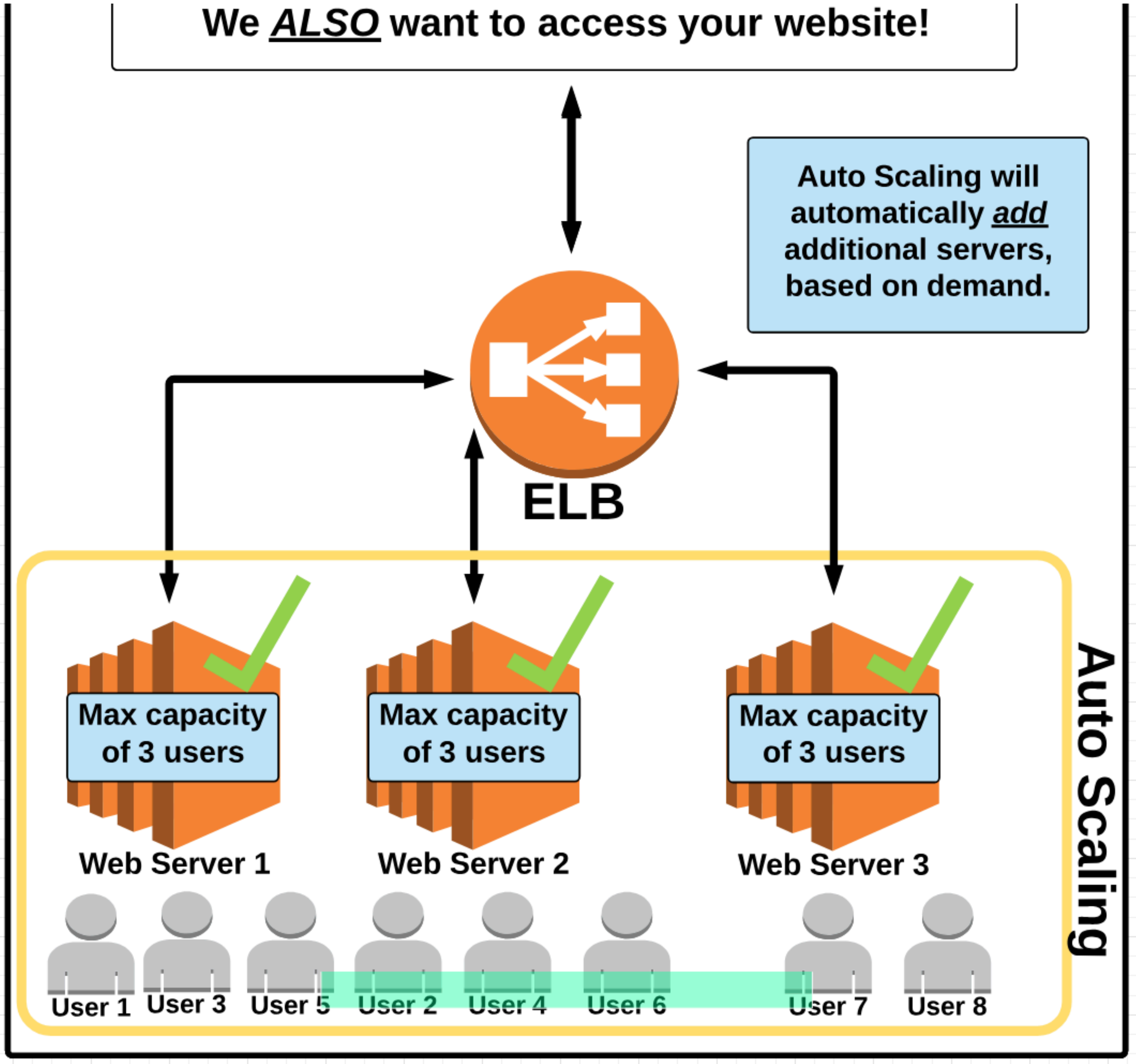
Max capacity of 3 users

Web Server 2

User 1 User 3 User 5 User 2 User 4 User 6

Auto Scaling

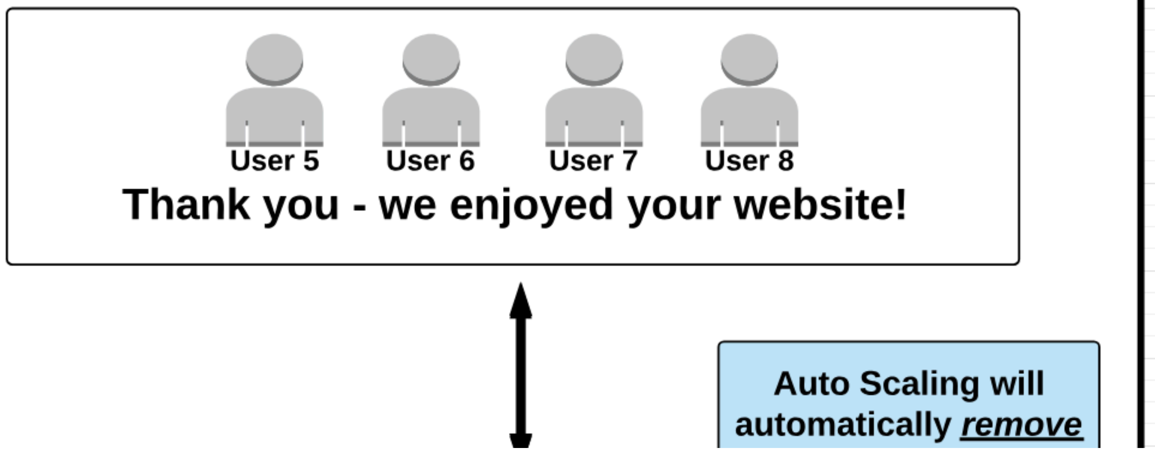
Auto Scaling Basics:



We *ALSO* want to access your website!

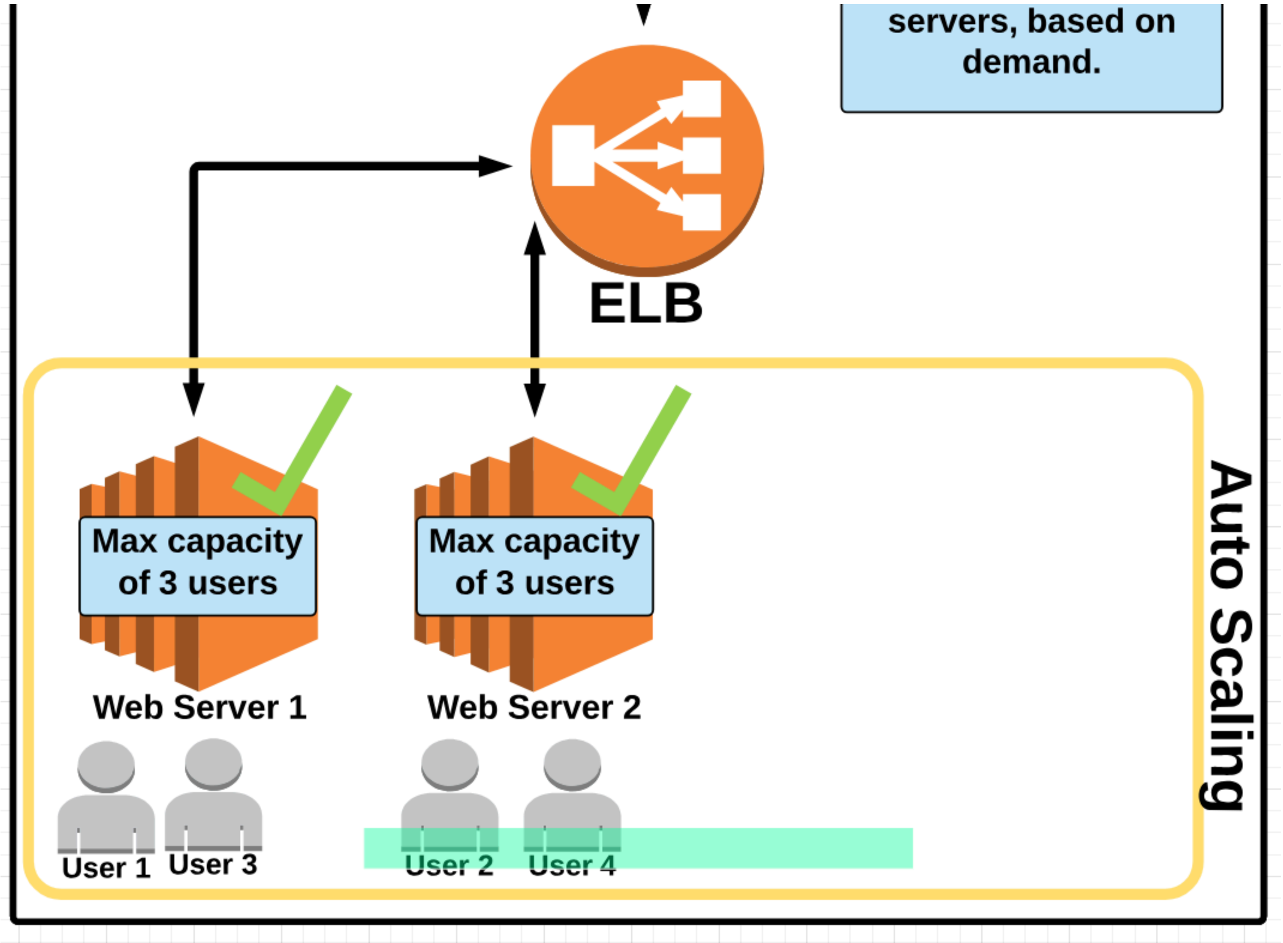
Auto Scaling will automatically *add* additional servers, based on demand.

Auto Scaling Basics:



Thank you - we enjoyed your website!

Auto Scaling will automatically *remove*



servers, based on demand.

Auto Scaling Basics:

Auto Scaling

builds on the benefits of

Elastic Load Balancing

while adding the benefits of

Scalability

and

Elasticity