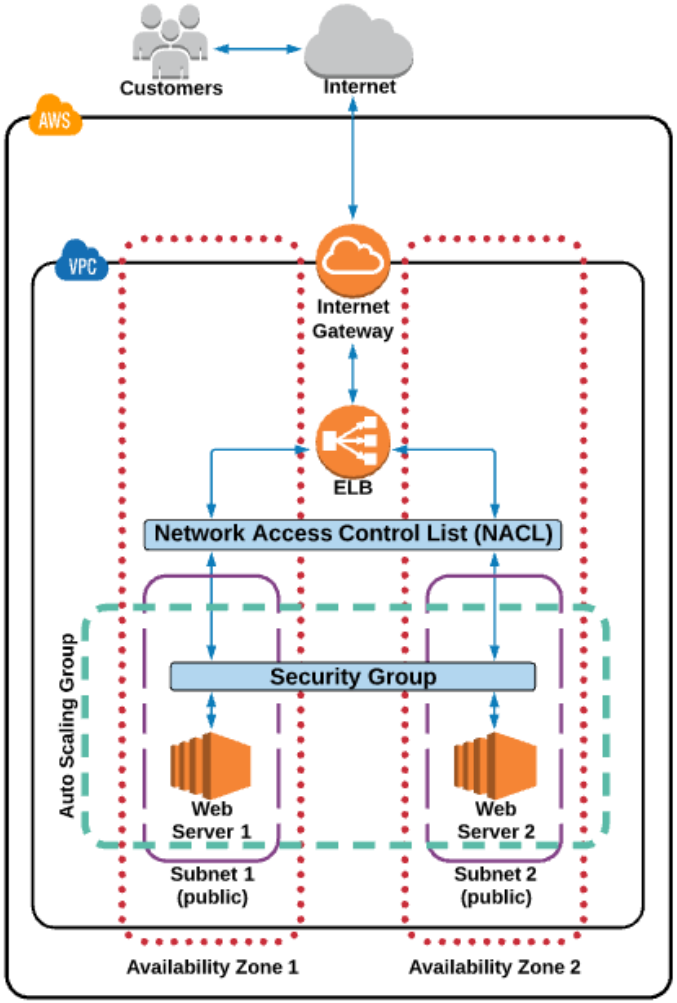
AutoSCALING

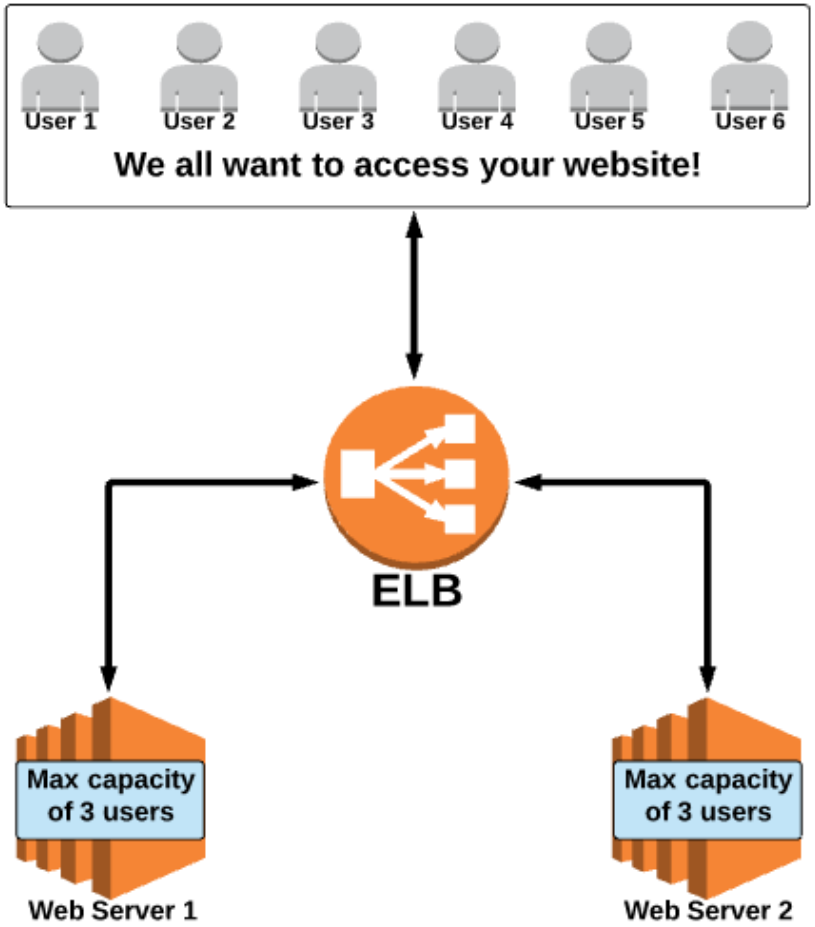
Auto Scaling

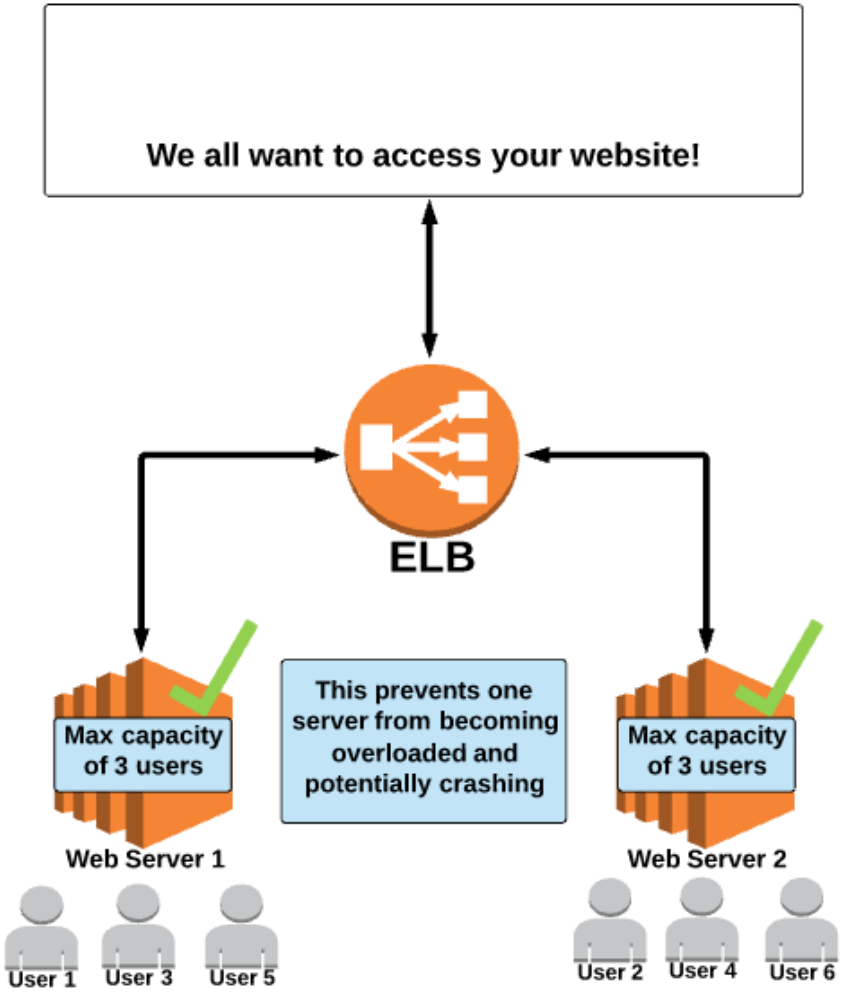
What is Auto Scaling?

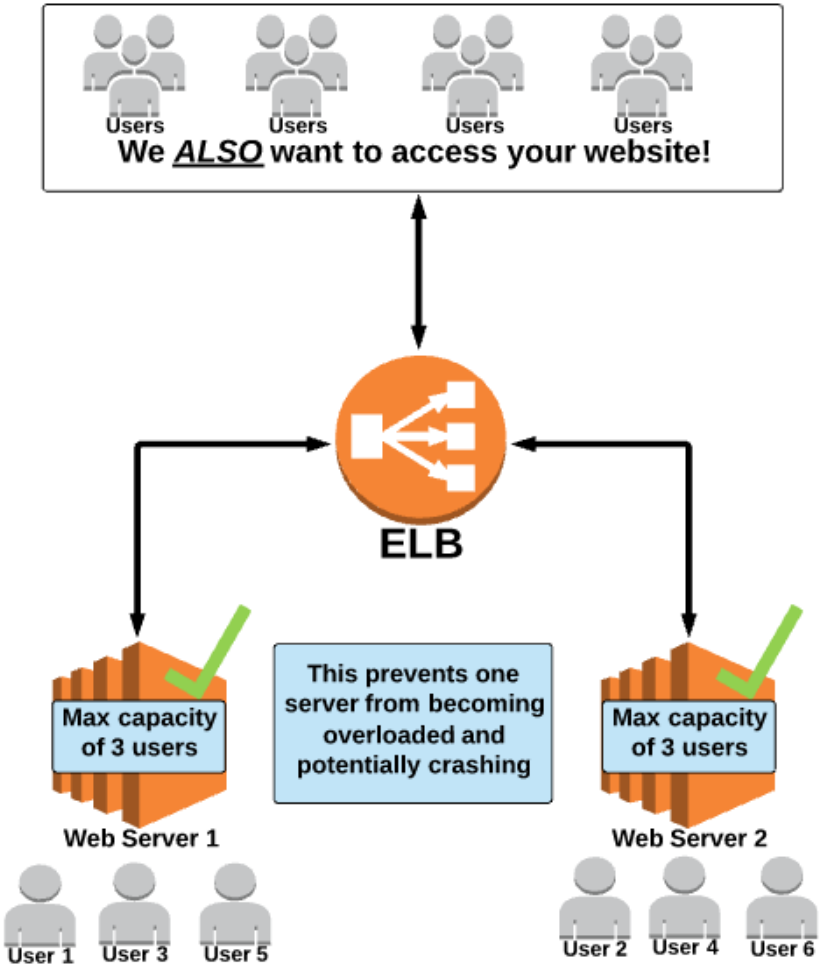
"Essentials" Definition:
Auto Scaling automates the process of adding (***scaling up***) OR removing (***scaling down***) EC2 instances ***based on traffic demand*** for your application.

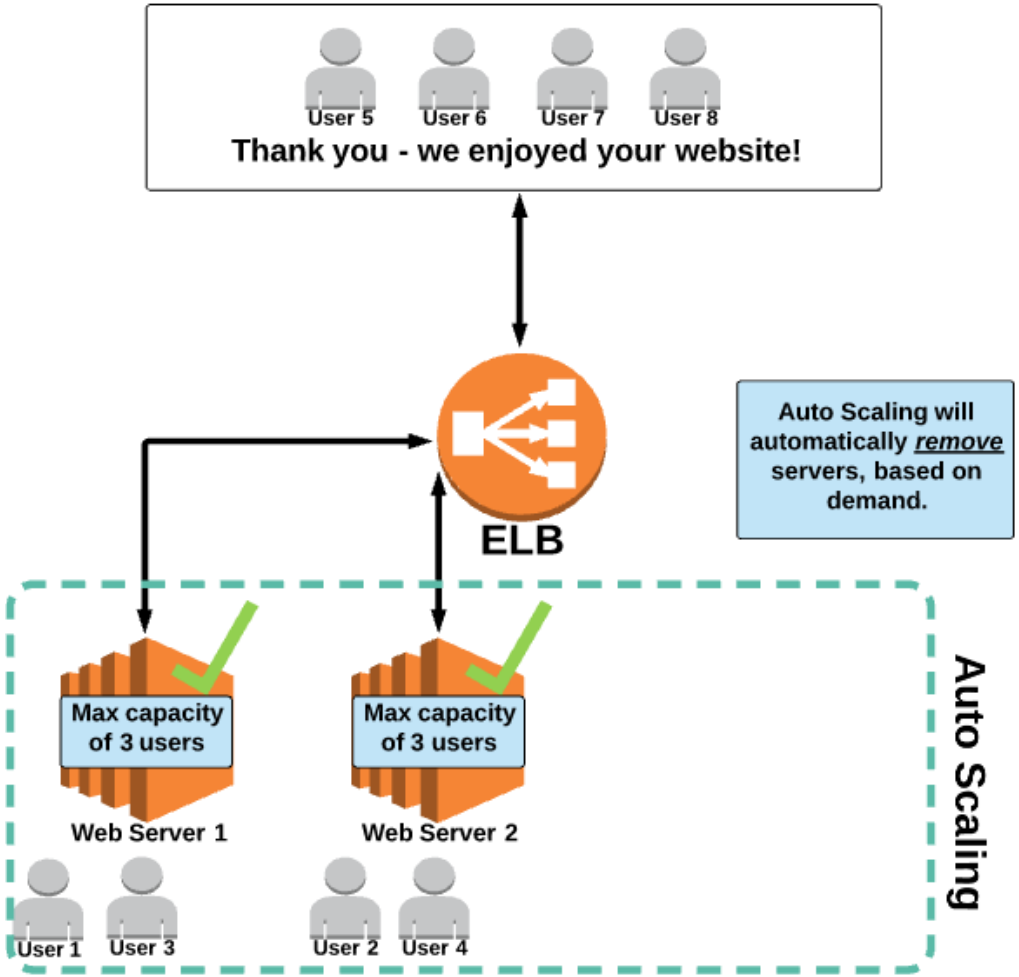
AWS Definition:
"Auto Scaling helps you ensure that you have the correct number of Amazon EC2 instances available to ***handle the load for your application***. You create collections of EC2 instances, called ***Auto Scaling groups***. You can specify the minimum number of instances in each Auto Scaling 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."











Auto Scaling Components

Launch Configuration

The EC2 template used when Auto Scaling needs to add an additional server to your *Auto Scaling Group*.

Auto Scaling Group

All the rules and settings that govern when an EC2 server is automatically added or removed.

Pricing/Cost Overview:

How are you charged for using ELB?

(1) ***Auto Scaling is FREE to use! However...***

(2) ***You will be charged for the resources that Auto Scaling provisions.***
(i.e. any EC2 instances A.S. provisions that go beyond the Free Tier allotment)

Basic Steps:

(1) Select an AMI
(2) Select an Instance Type
(3) Create Launch Configuration:
-Give the Launch Configuration a name.
-Make sure that a public IP address will be assigned.
-(optional) For this demonstration we are going to include the *bash script* to install the Apache web server software:

```
#!/bin/bash
yum update -y
yum install -y httpd
service httpd start
```

(4) Select/Add storage type
(5) Configure Security Group:
-Make sure you select a SG group that has the ports open that you need.
(6) Review & Create

Creating an Auto Scaling Group:

Basic Steps:

(1) Create an Auto Scaling Group from an existing launch configuration:
-Select the launch configuration you want to use.

(2) Configure Auto Scaling group details:
-Give the group a name.
-Select the number of instances with which the group should START.
-Select the VPC and subnets in which you want Auto Scaling to provision instances.
-Open ***ADVANCED DETAILS***:
-Check the box to include Load Balancing.
-Select the ELB you want to use.
-Configure the health checks.

(3) Configure scaling policies:
-Select "Use scaling policies to adjust the capacity of this group".
-Choose the number of instances you want to scale between. This is the MIN and MAX number of instances that Auto Scaling can scale between.
-Create and "Execute Policy" for both the "Increase" and "Decrease Group Size" sections. These are the metric thresholds that govern when Auto Scaling adds or removes instances. CPU utilization the most commonly used metric.

(4) Configure notifications:
-Select an SNS topic to send notifications to whenever Auto Scaling launches/terminates OR fails to launch/terminate an instance.

(5) Configure add (use if you like but not required)

(6) Review & Create