Lab 8: Working with Elastic Load Balancing and Auto Scaling

East Tennessee State University

CSCI 4417/5417: Introduction to System Administration

Spring 2016

Pramod Nepal

Purpose

The first objective of this lab was to use Elastic Load Balancing to load balance traffic across multiple Amazon Elastic Compute Cloud (EC2) instance in a single Availability Zone. A simple web application was deployed on multiple Amazon EC2 instances and load balancing was observed by viewing the application in the browser. The second objective of this lab was to experiment with Amazon Auto Scaling.

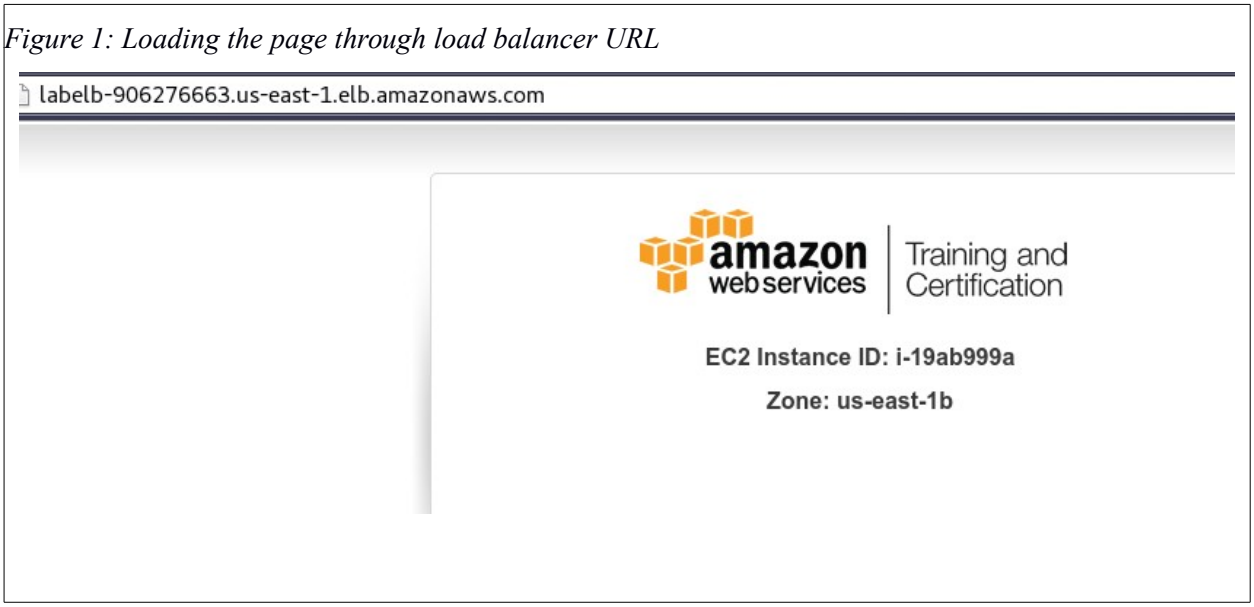
Materials

- Lab Instructions from QuickLabs
- Web browser

Summary Load Balancing Lab

Elastic Load Balancing was used in the lab to load balance traffic across multiple Amazon Elastic Computer Cloud (EC2) instances in a single Availability Zone. For this a simple web application was deployed in both instances and load balancing was observed by loading the URL provided by the Load Balancer in a browser (see Fig. 1).

Figure 1: Loading the page through load balancer URL



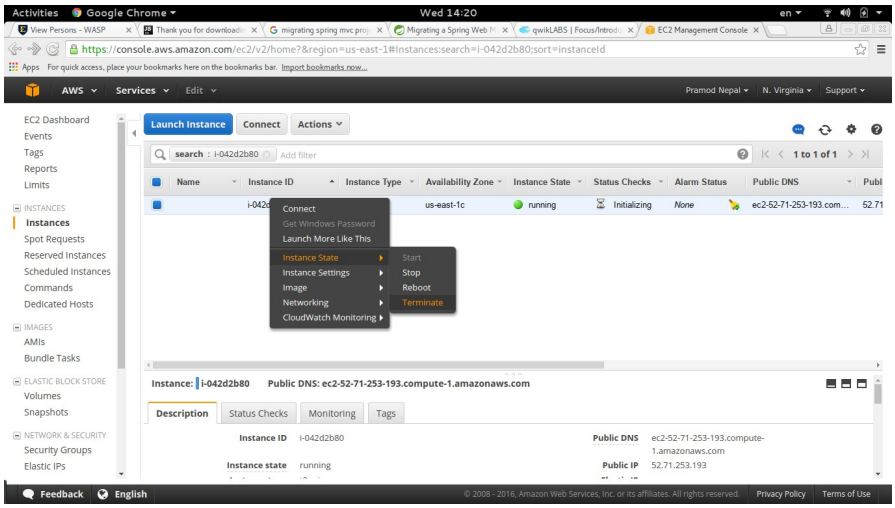
The installation of the web application was automated by a script. The script downloaded the necessary packages and setup the web server. From the observation it was apparent that load or web request from the users was balanced across both instances.

The load balancing spreads the request across multiple servers so that only one instance does not get overloaded. Metric tools like CloudWatch can help monitor basic statistics of the servers. A System Admin then would be able to plan for scaling the applications. Since the application is running at multiple sites, failure of one of the servers would not effect the operation. In this way, the load balancing provides redundancy. A System Admin would also be able to maintain one of the servers by taking it offline while others are handling the client requests. There is a major cost penalty for unavailability of sites for organizations that have their majority of client requests through the web. Therefore through load balancing the organization can ensure their business is operational most of the time.

Summary Auto Scaling Lab

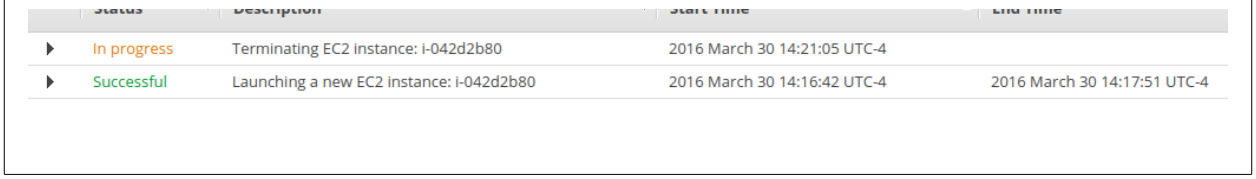
An Auto Scaling group was created using a '64-bit Amazon Linux AMI' instance. The maximum and minimum numbers of instances were set to 1. To test that Auto Scalar maintains the minimum and maximum number of instances a running instance was terminated (see Fig. 2).

Figure 2: Terminating a running instance



After some time another instance started (see Fig. 3) without user intervention.

Figure 3: Automatic start of second instance



Amazon EC2 Auto Scaling allows to create groups so that the configuration and other settings can be applied for a group rather than for individual instances. Different policies can be set so that auto scaling can launch or terminate the instances on demand. An auto scaling group can make sure that the application always has the right amount of capacity to handle the current traffic demands. It makes the application highly available and fault tolerant. With maximum and minimum instances set, right number of instances serve the client requests. The servers can scale down or scale up depending upon the demand. Also cloud providers like Amazon apply cost based on use. Therefore auto scaling helps an organization save cost and resources. Just like load balancing, auto scaling is also a mechanism for fault tolerance and availability. In a traditional scaling setup most organizations either under utilize or over utilize the resources. With auto scaling the resources are used and initialized as per the load. An Administrator can create policies for scaling out and scaling down scenarios. Thus it provides a mechanism for optimum utilization of resources without loss of service.