Lab description

This lab introduces the basics of Auto Scaling in Amazon Web Services. The Amazon Web Services (AWS) Auto Scaling service automatically adds or removes compute resources allocated for your cloud application, in response to changes in demand. For applications configured to run on a cloud infrastructure, scaling is an important part of cost control and resource management.

Scaling is the ability to increase or decrease the compute capacity of your application either by changing the number of servers (horizontal scaling) or by changing the size of the servers (vertical scaling).

Auto Scaling helps you maintain application availability and allows you to scale your Amazon EC2 capacity up or down automatically according to the defined conditions. You can use Auto Scaling to help ensure that you are running your desired number of Amazon EC2 instances. Auto Scaling can also automatically increase the number of Amazon EC2 instances during demand spikes to maintain performance and decrease capacity during lulls to reduce costs. Auto Scaling is well suited to applications that have stable demand patterns, or that experience hourly, daily, or weekly variability in usage.

In this lab, you will create an Auto Scaling Group and place it behind a Network Load Balancing (NLB). Don't worry if you don't fully understand all the components yet. Each one will be discussed in greater detail as you create and configure them. At the end of this lab you will have an Auto Scaling Group with several web server instances behind a Network Load Balancer.

Lab Objectives

Upon completion of this lab you will be able to:

- Configure Auto Scaling to automatically launch EC2 instances using conditions described by CloudWatch alarms
- Create and configure a Network Load Balancer
- Utilize Auto Scaling and a Network Load Balancer to ensure the availability of compute resources
- Build an elastic cluster by integrating Auto Scaling with an Elastic Load Balancer
- Perform end-to-end testing of the system and understand how to diagnose issues

Lab Prerequisites

Although this is a beginner-level lab, you should be familiar with AWS basics including:

- Using the AWS Management Console
- EC2 (launching an instance, connecting to an instance via SSH)
- Conceptual understanding of CloudWatch, and EC2 Security Groups (firewall rules)