

Systems Operations on AWS - Lab 3L - Using Auto Scaling (Linux)

3 hours

Free

★★★★★ [Rate Lab](#)



© 2020 Amazon Web Services, Inc. and its affiliates. All rights reserved.
This work may not be reproduced or redistributed, in whole or in part,
without prior written permission from Amazon Web Services, Inc.
Commercial copying, lending, or selling is prohibited.

Errors or corrections? Email us at aws-course-feedback@amazon.com.

Other questions? Contact us at <https://aws.amazon.com/contact-us/aws-training/>

In this lab, you will be tasked with creating a new Amazon Machine Image (AMI) from an existing EC2 instance, and use that machine as

In this lab, you will be tasked with creating a new Amazon Machine Image (AMI) from an existing EC2 instance, and use that machine as the basis for defining a system that will scale automatically under increasing loads.

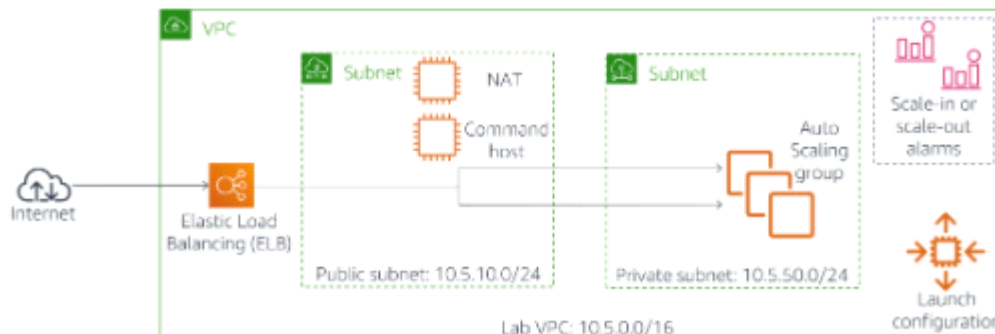
Duration

This lab will require approximately **45 minutes** to complete.

⚠ This is the Linux version of this lab. It means that you will be connecting to an Amazon EC2 Linux instance via SSH/PuTTY. If you instead wish to use RDP to connect to an Amazon EC2 Windows instance, **please use the Windows version of this lab.**

Scenario

In this lab, you will create the scalable web server system shown in the following diagram:



Objectives

After completing this lab, you will be able to:

- Create a new Amazon Machine Image (AMI) by using the Amazon Command Line Interface (CLI).
- Use Auto Scaling to scale up the number of servers available for a specific task when other servers are experiencing heavy load.

- Use Auto Scaling to scale up the number of servers available for a specific task when other servers are experiencing heavy load.

The following components are created for you as a part of the lab environment:

- Amazon VPC
- Public Subnets
- Private Subnets
- Amazon EC2 - Command Host (*in the public subnet*), you will log in to this instance to create a few of your AWS assets.

You will create the following components for this lab:

- Amazon EC2 - Web Server
- Amazon Machine Image (AMI)
- Auto Scaling Launch Configuration
- Auto Scaling Group
- Auto Scaling Policies
- ELB

Accessing the AWS Management Console

Start Lab

Start Lab

1. At the top of your screen, launch your lab by clicking **Start Lab**

This will start the process of provisioning your lab resources. An estimated amount of time to provision your lab resources will be displayed. You must wait for your resources to be provisioned before continuing.

i If you are prompted for a token, use the one distributed to you (or credits you have purchased).

2. Open your lab by clicking **Open Console**

This will open an AWS Management Console sign-in page.

3. On the Sign-in page, configure:

- **IAM user name:** `awsstudent`
- **Password:** Paste the value of **Password** located to the left of these instructions.
- Click **Sign In**

⚠ Please do not change the Region unless instructed.

Common login errors

Error: You must first log out

Amazon Web Services Sign In

You must first log out before logging into a different AWS account.

To logout, [click here](#)

If you see the message, **You must first log out before logging into a different AWS account:**


- Click **click here**
- Close your browser tab to return to your initial Qwiklabs window
- Click **Open Console** again

Task 1: Create a New Amazon Machine Image for Amazon EC2 Auto Scaling


In this task, you will launch a new EC2 instance and then create a new AMI based on that running instance. You will use the AWS CLI tools on the Command Host to perform all of these operations.

The following instructions vary slightly depending on whether you are using Windows or Mac/Linux.

Windows Users: Using SSH to Connect

 These instructions are for Windows users only.


If you are using Mac or Linux, [skip to the next section](#).

4. In the **AWS Management Console**, on the **Services**  menu, click **EC2**.
5. In the left navigation pane, click **Instances**.

-
5. In the left navigation pane, click **Instances**.
 6. Select the **Command Host**.
 7. Copy the **IPv4 Public IP** from the Description in the lower pane.
 8. To the left of the instructions you are currently reading, click **Download PPK**.
 9. Save the file to the directory of your choice.

You will use PuTTY to SSH to Amazon EC2 instances.

If you do not have PuTTY installed on your computer, [download it here](#)

10. Open PuTTY.exe
11. Configure your PuTTY session:
 - **Host Name:** Paste the **Public IPv4** value you copied to your clipboard earlier in the lab
 - In the **Connection** list, expand  **SSH**
 - Click **Auth** (don't expand it)
 - Click **Browse**
 - Browse to and select the PPK file that you downloaded
 - Click **Open** to select it
 - Click **Open**

12. When prompted for **login as:**, enter: `ec2-user`



This will connect to your EC2 instance.

13. [Windows Users: Click here to skip ahead to the next task.](#)

Mac  and Linux  Users

Mac and Linux Users

These instructions are for Mac/Linux users only. If you are a Windows user, [skip to the next task](#).

14. In the **AWS Management Console**, on the **Services**  menu, click **EC2**.
15. In the left navigation pane, click **Instances**.
16. Select the **Command Host**.
17. Copy the **IPv4 Public IP** from the Description pane.
18. To the left of the instructions you are currently reading, click  **Download PEM**.
19. Save the file to the directory of your choice.
20. Copy this command to a text editor:

```
chmod 400 KEYPAIR.pem  
  
ssh -i KEYPAIR.pem ec2-user@EC2PublicIP
```

- Replace **KEYPAIR.pem** with the path to the PEM file you downloaded.
 - Replace **EC2PublicIP** with the **IPv4 Public IP** value you copied to your clipboard earlier in the lab
21. Paste the command into the Terminal window and run it.
 22. Type **yes** when prompted to allow the first connection to this remote SSH server.

Because you are using a key pair for authentication, you will not be

Because you are using a key pair for authentication, you will not be prompted for a password.

Create A New EC2 Instance

Now that you are logged in to CommandHost, you will use the AWS CLI to create a new instance that hosts a web server.

23. Inspect the script `UserData.txt` that was installed for you as part of the CommandHost.

```
more UserData.txt
```

This script performs a number of initialization tasks, including updating all installed software on the box and installing a small PHP web application that you can use to simulate a high CPU load on the instance. Near the bottom of the script, you will see the following lines:

```
find -wholename /root/*.history -wholename /home/*.history  
-exec rm -f {} \;  
find / -name 'authorized_keys' -exec rm -f {} \;  
rm -rf /var/lib/cloud/data/scripts/*
```

These lines erase any history or security information that may have accidentally been left on the instance when the image was taken.

24. Copy the **KEYNAME**, **AMIID**, **SUBNETID** and **HTTPACCESS** shown to the left of the instructions you are currently reading and paste it into relevant sections of the below script.

```
aws ec2 run-instances --key-name KEYNAME --instance-type  
t3.micro --image-id AMIID --user-data file:///home/ec2-  
user/UserData.txt --security-group-ids HTTPACCESS --subnet-id  
SUBNETID --associate-public-ip-address --tag-specifications
```



```
user/UserData.txt --security-group-ids HTTPACCESS --subnet-id
SUBNETID --associate-public-ip-address --tag-specifications
'ResourceType=instance,Tags=
[{Key=Name,Value=WebServerBaseImage}]' --output text --query
'Instances[*].InstanceId'
```

The output of this command will provide you with an **InstanceId**. This value is referred to as **new-instance-id** in subsequent steps and should be replaced appropriately.

25. Use the **aws ec2 wait instance-running** command to monitor this instance's status. Replace *NEW-INSTANCE-ID* with the value you copied previously.

```
aws ec2 wait instance-running --instance-ids NEW-INSTANCE-ID
```

Wait for the command to return to a prompt, before proceeding to the next step.

26. Your instance should have started a new web server. To test that the web server was installed properly, obtain the public DNS name. Copy the output of this value (minus quotation marks). This value is referred to as **public-dns-address** in the next procedure. Replace *NEW-INSTANCE-ID* with the value you copied previously.

```
aws ec2 describe-instances --instance-id NEW-INSTANCE-ID --
query
'Reservations[0].Instances[0].NetworkInterfaces[0].Association.P
```

27. Use a Web browser to navigate to the following page.



- It could take a few minutes for the web server to be installed. Please wait for five minutes before trying other steps.

Do not click on **Start Stage** at this stage. Replace *PUBLIC DNS*

- It could take a few minutes for the web server to be installed. Please wait for five minutes before trying other steps.
- Do not click on *Start Stress* at this stage. Replace *PUBLIC-DNS-ADDRESS* with the value you copied in the last step.

```
http://PUBLIC-DNS-ADDRESS/index.php
```

If your web server does not appear to be running, check with your instructor.

Create a Custom AMI

In this procedure, you will create a new AMI based on that instance you just created.

28. Use the **aws ec2 create-image** command to create a new AMI based on this instance. Replace *NEW-INSTANCE-ID* with the value you copied previously.

```
aws ec2 create-image --name WebServer --instance-id NEW-INSTANCE-ID
```

● By default, **create-image** will restart the current instance before creating the AMI, in order to ensure the integrity of the image on the file system. While your AMI is being created, proceed to the next section.

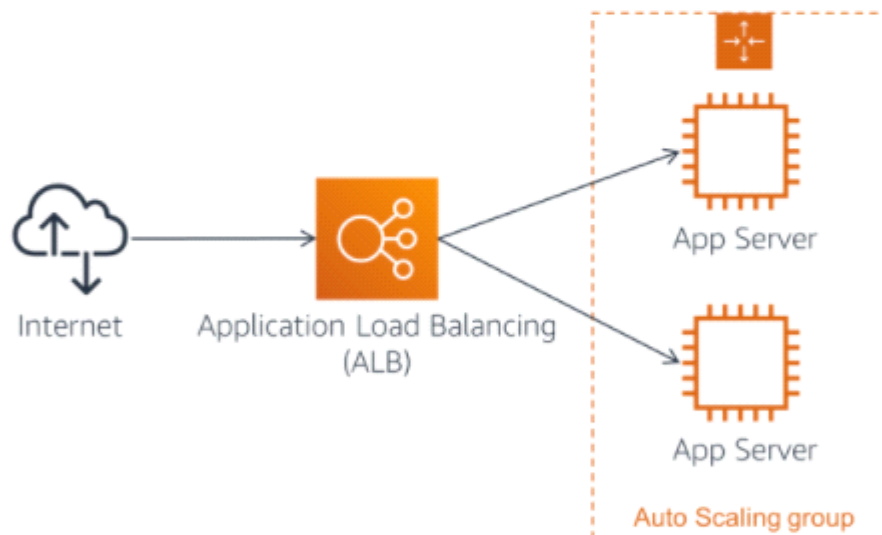
Task 2: Create an Auto Scaling Environment

Environment

In this section, you will create a load balancer that pools a group of EC2 instances under a single DNS address. You will use Auto Scaling to create a dynamically scalable pool of EC2 instances based on the image that you created in the previous section. Finally, you will create a set of alarms that will scale out or scale in the number of instances in your load balancer group whenever the CPU performance of any machine within the group exceeds or falls below a set of specified thresholds.

The following task can be performed using either the AWS CLI or the Management Console. For purposes of simplicity, you will implement this task using the Management Console.

Create an Application Load Balancer



29. On the **Services** menu, click **EC2**.

29. On the **Services** menu, click **EC2**.

30. In the left navigation pane, click **Load Balancers** (you might need to scroll down to find it).

31. Click **Create Load Balancer**

32. Under **Application Load Balancer**, click **Create**

33. For **Name**, enter: `webserverloadbalancer`

34. Scroll down to the **Availability Zones** section, then:

- For **VPC**, select: **Lab VPC**

You will now specify which *subnets* the Load Balancer should use. It will be an Internet-facing load balancer, so you will select both Public Subnets.

35. Click the **first** displayed Availability Zone, then click the **Public Subnet 1** displayed underneath.

36. Click the **second** displayed Availability Zone, then click the **Public Subnet 2** displayed underneath.

You should now have two subnets selected: **Public Subnet 1** and **Public Subnet 2**. (If not, go back and try the configuration again.)

37. Click **Next: Configure Security Settings**

A warning is displayed, which recommends using HTTPS for improved security. This is good advice but is not necessary for this lab.

38. Click **Next: Configure Security Groups**

39. Select ☒ **HTTPAccess** and ensure it is the only item selected.

39. Select ☒ **HTTPAccess** and ensure it is the only item selected.

40. Click **Next: Configure Routing**

Target Groups define where to *send* traffic that comes into the Load Balancer. The Application Load Balancer can send traffic to multiple Target Groups based upon the URL of the incoming request, such as having requests from mobile apps going to a different set of servers. Your web application will use only one Target Group.

41. For **Name**, enter: `webserver-app`

42. Expand ► **Advanced health check settings**.

The Application Load Balancer automatically performs *Health Checks* on all instances to ensure that they are responding to requests. The default settings are recommended, but you will make them slightly faster for use in this lab.

43. Configure these values:

- **Path** `/index.php`
- **Healthy threshold:** `2`
- **Interval:** `10`

This means that the Health Check will be performed every 10 seconds and if the instance responds correctly twice in a row, it will be considered healthy.

44. Click **Next: Register Targets**

Targets are the individual instances that will respond to requests from the Load Balancer. You do not have any web application instances yet, so you can skip this step.

45. Click **Next: Review**

45. Click **Next: Review**

46. Review the settings and click **Create** then **Close**

Create a Launch Configuration

The launch configuration will be used by your Auto Scaling group to specify which AMI to use to create new EC2 instances. For this example, you will launch the AMI that you created previously, which automatically configures itself as a web server when it is launched.

47. In the left navigation pane, click **Launch Configurations**.

48. Click **Create launch configuration**

49. On the **Choose AMI** step, click the **My AMIs** tab.

50. In the row for **WebServer**, click **Select**

51. On the **Choose Instance Type** step, make sure the **t3.micro** instance is selected.

52. Click **Next: Configure details**

53. On the **Configure details** step, configure:

- **Name:** WebServerLaunchConfiguration
- **Monitoring:** Select ☒ **Enable CloudWatch detailed monitoring**

54. Click **Next: Add Storage**

You will use the default storage settings.

55. Click **Next: Configure Security Group**

56. On the **Configure Security Group** step, click  **Select an existing**

56. On the **Configure Security Group** step, click ☒ **Select an existing security group**.

57. Select ☒ **HTTPAccess**.

58. Click **Review**

You may receive a warning that your security group will not enable you to SSH into the instance. You may safely ignore this warning because you will not need to SSH into the instances in your Auto Scaling group.

Click **Continue**

59. Click **Create launch configuration**

60. Select ☒ the acknowledgement check box, then click

Create launch configuration

61. Click **Create an Auto Scaling group using this launch configuration**

Create an Auto Scaling Group

Your Auto Scaling group will create a minimum number of Amazon EC2 instances that will reside behind your load balancer. In subsequent procedures, you will also add scale-out and scale-in policies that increase or decrease the number of running instances in reaction to alarms triggered by Amazon CloudWatch.

You should already be on the **Create Auto Scaling Group** page.

62. Configure these settings:

- **Group name:** WebServersASGroup
- **Group size:** Start with **2** instances.
- **Network:** Lab VPC

- **Group size:** Start with **2** instances.
- **Network:** Lab VPC
- **Subnet:** Private Subnet 1 and Private Subnet 2

🗨 You may receive a warning message that no public IP addresses will be assigned to your instances because you are hosting them outside of your default VPC. This is fine because the public load balancer will be responsible for directing traffic to your instances.

63. Expand ► **Advanced Details**, specify the following settings:

- **Load Balancing:** Select ☒ **Receive traffic from one or more load balancers**
- **Target Groups:** Click `webserver-app`
- **Monitoring:** Select ☒ **Enable CloudWatch detailed monitoring**

64. Click **Next: Configure scaling policies**

65. Select ☒ **Use scaling policies to adjust the capacity of this group.**

66. Specify scaling between **2** and **4** instances.

You will configure Auto Scaling to target 45% CPU Utilization. If average CPU Utilization exceeds this target, additional instances will be launched. If average CPU Utilization falls below this target, instances *might* be terminated.

67. For **Target value**, enter: `45`

68. Click on **Next: Configure Notifications**

69. Click on **Next: Configure Tags**

70. Configure these settings:

- **Key:** `Name`
- **Value:** `WebApp`

- **Key:** Name
- **Value:** WebApp

71. Click **Review**

72. Review your configuration, and then click **Create Auto Scaling group**

73. On the **Auto Scaling group creation status** page, click **Close**

Verifying the Auto Scaling configuration

In this task, you will verify that both the Auto Scaling configuration and the load balancer are working by accessing a pre-installed script on one of your servers that will consume CPU cycles, thus triggering the scale out alarm.

74. In the left navigation pane, click **Instances**.

75. Verify that two new instances labelled **WebApp** are being created as part of your Auto Scaling group.

76. Wait for the two new instances to complete initialization before you proceed to the next step. Observe the **Status Checks** field for the instances until it shows that both status checks have completed successfully.

77. In the left navigation pane, click **Target Groups**, and then select ☒ your target group (**webserver-app**).

78. On the **Targets** tab in the lower half of your screen, verify that two instances are being created. Keep refreshing this list until the **Status** of these instances changes to **healthy**.

You can now test the web application by accessing it via the Load Balancer.

Balancer.

79. In the left navigation pane, click **Load Balancers** and then select ☒ **webserverloadbalancer**.
80. On the **Description** tab below, copy the **DNS name** value (which should look like **webserverloadbalancer-xxxxxxxxxx.xx-xxxx-x.elb.amazonaws.com**). This value will be referred to as **load-balancer-url** in the next step.
81. Open a new web browser tab and paste the URL into the address bar, then press Enter.
82. On the web page, click **Start Stress**.

This will call the application **stress** in the background, causing the CPU utilization on the instance that serviced this request to spike to 100%.

83. In the left navigation pane of the management console, click **Auto Scaling Groups**.
84. Select ☒ **WebServerASGroup**.
85. Select the **Activity History** tab for your Auto Scaling group. After a few minutes, you should see your Auto Scaling group add a new instance.

💡 This is because CloudWatch detected that the average CPU utilization of your Auto Scaling group exceeded 45%, and your scale-up policy has been triggered in response.

Lab Complete

Lab Complete

Congratulations! You have completed the lab.

End Lab

Follow these steps to close the console, end your lab, and evaluate the experience.

86. Return to the AWS Management Console.

87. On the navigation bar, click **awsstudent@<AccountNumber>**, and then click **Sign Out**.

88. Click  **End Lab**

89. Click 

90. (Optional):

- Select the applicable number of stars ☆
- Type a comment
- Click **Submit**
 - 1 star = Very dissatisfied
 - 2 stars = Dissatisfied
 - 3 stars = Neutral
 - 4 stars = Satisfied
 - 5 stars = Very satisfied

- 5 stars = Very satisfied

You may close the dialog if you don't want to provide feedback.