Tutorial: Set Up a Scaled and Load-Balanced Application

September 29, 2016

You can attach a load balancer to your Auto Scaling group. The load balancer automatically distributes incoming traffic across the instances in the group. For more information about the benefits of using Elastic Load Balancing with Auto Scaling, see Using a Load Balancer With an Auto Scaling Group.

This tutorial attaches a load balancer to an Auto Scaling group when you create the group. To attach a load balancer to an existing Auto Scaling group, see Attaching a Load Balancer to Your Auto Scaling Group.

Prerequisites

- (Optional) Create an IAM role that grants your application the access to AWS that it needs.
- Launch an instance; be sure to specify the IAM role (if you created one) and specify any configuration scripts that you need as user data. Connect to the instance and customize it. For example, you can install software and applications and copy data. Test your application on your instance to ensure that your instance is configured correctly. Create a custom Amazon Machine Image (AMI) from your instance. You can terminate the instance if you no longer need it.
- Create a load balancer. Elastic Load Balancing supports two types of load balancers: Classic Load Balancers and Application Load Balancers. You can create either type of load balancer to attach to your Auto Scaling group. For more information, see the Elastic Load Balancing User Guide.

  With a Classic Load Balancer, instances are registered with the load balancer, and with an Application Load Balancer, instances are registered as targets with a target group. When you plan to use your load balancer with an Auto Scaling group, you don't need to register your EC2 instances with the load balancer or target group. After you attach a load balancer or target group to your Auto Scaling group, Auto Scaling registers your instances with the load balancer or target group when it launches them.

Getting Started with Elastic Load Balancing

Step 1: Select a Load Balancer Type

Elastic Load Balancing supports two types of load balancers: Application Load Balancers and Classic Load Balancers. For this tutorial, you create an Application Load Balancer.

To create an Application Load Balancer

1. Open the Amazon EC2 console at https://console.aws.amazon.com/ec2/.
2. On the navigation bar, choose a region for your load balancer. Be sure to select the same region that you selected for your EC2 instances.
3. On the navigation pane, under LOAD BALANCING, choose Load Balancers.
4. Choose Create Load Balancer.
5. Choose Application Load Balancer, and then choose Continue.

Step 2: Configure Your Load Balancer and Listener

On the Configure Load Balancer page, complete the following procedure.

To configure your load balancer and listener

1. For Name, type a name for your load balancer.

    The name of your Application Load Balancer must be unique within your set of Application Load Balancers for the region, can have a maximum of 32 characters, can contain only alphanumeric characters and hyphens, and must not begin or end with a hyphen.

2. For Scheme, keep the default value, internet-facing.

Basic Configuration

To configure your load balancer, provide a name, select a scheme, specify one or more listeners, and select a network. The default configuration is an Internet-facing load balancer in the selected network with a listener that receives HTTP traffic on port 80.

Name ⓘ	my-load-balancer	
Scheme ⓘ	◉ internet-facing	
	○ internal	

3. For Listeners, keep the default, which is a listener that accepts HTTP traffic on port 80.

Listeners

A listener is a process that checks for connection requests, using the protocol and port that you configured.

Load Balancer Protocol	**Load Balancer Port**	
HTTP ⌄	80	⊗

Add listener

4. For VPC, select the same VPC that you used for your EC2 instances.

Availability Zones

Specify the Availability Zones to enable for your load balancer. The load balancer routes traffic to these Availability Zones only. You can specify only one subnet per Availability Zone. You must sp from at least two Availability Zones to increase the availability of your load balancer.

VPC ⓘ	Choose a VPC ⌄

5. For Available subnets, select at least two public subnets using their add icons. The subnets are moved under Selected subnets. Note that you can select only one subnet per Availability Zone. If you select a subnet from an Availability Zone where there is already a selected subnet, this subnet replaces the currently selected subnet for the Availability Zone.

Available subnets

Actions	Availability Zone	Subnet ID	Subnet CIDR	Name
➕	us-west-2a	subnet-65ea5f08	10.0.0.0/24	
➕	us-west-2b	subnet-7ad90a22	10.0.2.0/24	

Selected subnets

Actions	Availability Zone	Subnet ID	Subnet CIDR	Name
➖	us-west-2b	subnet-6bea5f06	10.0.1.0/24	

6. Choose Next: Configure Security Settings.
7. For this tutorial, you are not using a secure listener. Choose Next: Configure Security Groups.

Step 3: Configure a Security Group for Your Load

Balancer

The security group for your load balancer must allow it to communicate with registered targets on both the listener port and the health check port. On the Configure Security Groups page, complete the following procedure.

To configure a security group for your load balancer

1. Choose Create a new security group.
2. Type a name and description for the security group, or keep the default name and description. This new security group contains a rule that allows traffic to the load balancer listener port that you selected on the Configure Load Balancer page.

Assign a security group:	⦿ Create a **new** security group
	○ Select an **existing** security group
Security group name:	my-load-balancer-group
Description:	for my load balancer

Type ⓘ	Protocol ⓘ	Port Range ⓘ	Source ⓘ	
HTTP ▾	TCP	80	Anywhere ▾ 0.0.0.0/0	✖

3. Choose Next: Configure Routing.

Step 4: Configure Your Target Group

Create a target group, which is used in request routing. The default rule for your listener routes requests to the to registered targets in this target group. The load balancer checks the health of targets in this target group using the health check settings defined for the target group. On the Configure Routing page, complete the following procedure.

To configure your target group

1. For Target group, keep the default, New target group.
2. For Name, type a name for the new target group.
3. Keep Protocol as HTTP and Port as 80.

Target group

Target group ⓘ	New target group	⌄
Name ⓘ	my-targets	
Protocol ⓘ	HTTP	⌄
Port ⓘ	80	

4. For Health checks, keep the default protocol and ping path.

Health checks

Protocol ⓘ	HTTP	⌄
Path ⓘ	/	

5. Choose Next: Register Targets.

Step 5: Register Targets with Your Target Group

On the Register Targets page, complete the following procedure.

To register targets with the target group

1. For Instances, select one or more instances.
2. Keep the default port, 80, and choose Add to registered.

3. If you need to remove an instance that you selected, for Registered instances, select the instance and then choose Remove.
4. When you have finished selecting instances, choose Next: Review.

Step 6: Create and Test Your Load Balancer

Before creating the load balancer, review the settings that you selected. After creating the load balancer, verify that it's sending traffic to your EC2 instances.

To create and test your load balancer

1. On the Review page, choose Create.
2. After you are notified that your load balancer was created successfully, choose Close.
3. On the navigation pane, under LOAD BALANCING, choose Target Groups.
4. Select the newly created target group.

5. On the Targets tab, verify that your instances are ready. If the status of an instance is `initial`, it's probably because the instance is still in the process of being registered, or it has not passed the minimum number of health checks to be considered healthy. After the status of at least one instance is `healthy`, you can test your load balancer.

6. On the navigation pane, under LOAD BALANCING, choose Load Balancers.

7. On the Description tab, copy the DNS name of the load balancer (for example, my-load-balancer-1234567890.us-west-2.elb.amazonaws.com). Paste the DNS name into the address field of an Internet-connected web browser. If everything is working, the browser displays the default page of your server.

Configure Scaling and Load Balancing Using the AWS

Management Console

Complete the following tasks to set up a scaled and load-balanced application when you create your Auto Scaling group.

Tasks

- [Create or Select a Launch Configuration](#)
- [Create an Auto Scaling Group](#)
- [(Optional) Verify that Your Load Balancer is Attached to Your Auto Scaling Group](#)

Create a Launch Configuration

To create a new launch configuration, use the following procedure.

To create a launch configuration

1. Open the Amazon EC2 console at https://console.aws.amazon.com/ec2/.
2. On the navigation bar at the top of the screen, select the region that you used when creating your load balancer.
3. On the navigation pane, under **Auto Scaling**, choose **Launch Configurations**. If you don't have any Auto Scaling resources, you see a welcome page; choose **Create Auto Scaling group**.
4. Choose **Create launch configuration**.
5. On the **Choose AMI** page, select your custom AMI.
6. On the **Choose Instance Type** page, select a hardware configuration for your instance, and then choose **Next: Configure details**.
7. On the **Configure Details** page, do the following:
    a. For **Name**, type a name for your launch configuration.
    b. (Optional) To securely distribute credentials to your EC2 instance, select your IAM role.

    c. (Optional) If you need to connect to an instance in a nondefault VPC, for **Advanced Details, IP Address Type**, choose **Assign a public IP address to every instance**.

    d. (Optional) To specify user data or a configuration script for your instance, for **Advanced Details, User data**, paste your configuration script.

    e. Choose **Skip to review**.

8. On the **Review** page, choose **Edit security groups**. Follow the instructions to choose an existing security group, and then choose **Review**.

9. On the **Review** page, choose **Create launch configuration**.

10. On the **Select an existing key pair or create a new key pair** page, select one of the listed options. Select the acknowledgment check box, and then choose **Create launch configuration.**

    Warning

    Do not choose **Proceed without a key pair** if you need to connect to your instance.

11. The **Launch configuration creation status** page displays the status of your newly created launch configuration. Choose **Create an Auto Scaling group using this launch configuration.**

Create an Auto Scaling Group

Use the following procedure to continue where you left off after selecting or creating your launch configuration.

To create an Auto Scaling group

1. On the **Configure Auto Scaling group details** page, do the following:

    a. For **Group name**, type a name for your Auto Scaling group.

    b. For **Group size**, type the initial number of instances for your Auto Scaling group.

    c. If you selected an instance type for your launch configuration that requires a VPC, such as a T2 instance, you must select a VPC for **Network**. Otherwise, if your account supports EC2-Classic and you

selected an instance type that doesn't require a VPC, you can select either `Launch into EC2-Classic` or a VPC.

d. If you selected a VPC in the previous step, select one or more subnets from **Subnet**. If you selected EC2-Classic instead, select one or more Availability Zones from **Availability Zone(s).**

e. For **Advanced Details**, select `Receive traffic from Elastic Load Balancer(s)` and then do one of the following:

- [Classic load balancers] Select your load balancer from **Load Balancers.**
- [Application load balancers] Select your target group from **Target Groups.**

f. For Advanced Details, select `Receive traffic from Elastic Load Balancer(s)` and then select your load balancer from **Load Balancers.**

g. (Optional) To use Elastic Load Balancing health checks, choose **ELB** for **Advanced Details**, **Health Check Type**.

h. Choose **Next: Configure scaling policies.**

2. On the **Configure scaling policies** page, select **Keep this group at its initial size**, and then choose **Review**.

   If you want to configure scaling policies for your Auto Scaling group, see [Scaling Based on Metrics](#).

3. Review the details of your Auto Scaling group. You can choose **Edit** to make changes. When you are finished, choose **Create Auto Scaling group**.

(Optional) Verify that Your Load Balancer is Attached to Your Auto Scaling Group

To verify that your load balancer is attached to your Auto Scaling group

1. Select your Auto Scaling group.
2. On the Details tab, Load Balancers shows any attached load balancers and Target Groups shows any attached target groups.
3. On the Details tab, Load Balancers shows any attached load balancers.

4. On the Activity History tab, the Status column shows you the status of your Auto Scaling instances. While an instance is launching, its status is `In progress`. The status changes to `Successful` after the instance is launched.

5. On the Instances tab, the Lifecycle column shows the state of your Auto Scaling instances. After an instance is ready to receive traffic, its state is `InService`.

    The Health Status column shows the result of the health checks on your instances.

Configure Scaling and Load Balancing Using the AWS CLI

Complete the following tasks to set up a scaled and load-balanced application.

Tasks

- [Create a Launch Configuration](#)
- [Create an Auto Scaling Group with a Load Balancer](#)

Create a Launch Configuration

If you already have a launch configuration that you'd like to use, skip this step.

To create the launch configuration

Use the following [create-launch-configuration](#) command:

```
aws autoscaling create-launch-configuration --launch-configuration-name my-lc \
--image-id ami-514ac838 --instance-type m1.small
```

Create an Auto Scaling Group with a Load Balancer

You can attach an existing load balancer to an Auto Scaling group when you create the group.

To create an Auto Scaling group with an attached Classic Load Balancer

Use the following [create-auto-scaling-group](#) command with the `--load-balancer-names` option to create an Auto Scaling group with an attached Classic Load Balancer:

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-lb-asg \
--launch-configuration-name my-lc \
--availability-zones "us-west-2a" "us-west-2b" \
--load-balancer-names "my-lb" \
--max-size 5 --min-size 1 --desired-capacity 2
```

To create an Auto Scaling group with an attached target group

Use the following [create-auto-scaling-group](#) command with the `--target-group-arns` option to create an Auto Scaling group with an attached target group:

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-lb-asg \
--launch-configuration-name my-lc \
--vpc-zone-identifier "subnet-41767929" \
--vpc-zone-identifier "subnet-b7d581c0" \
--target-group-arns "arn:aws:elasticloadbalancing:us-west-2:123456789012:targetgroup/my-targets/1234567890123456" \
--max-size 5 --min-size 1 --desired-capacity 2
```