



Amazon EC2 Auto Scaling ▾

Overview

Features

Pricing

Getting Started

Resources

FAQs

Q: What is Amazon EC2 Auto Scaling?

Amazon EC2 Auto Scaling is a fully managed service designed to launch or terminate Amazon EC2 instances automatically to help ensure you have the correct number of Amazon EC2 instances available to handle the load for your application. Amazon EC2 Auto Scaling helps you maintain application availability through fleet management for EC2 instances, which detects and replaces unhealthy instances, and by scaling your Amazon EC2 capacity up or down automatically according to conditions you define. You can use Amazon EC2 Auto Scaling to automatically increase the number of Amazon EC2 instances during demand spikes to maintain performance and decrease capacity during lulls to reduce costs.

[Show less](#)

Q. When should I use Amazon EC2 Auto Scaling vs. AWS Auto Scaling?

You should use [AWS Auto Scaling](#) if you want more guidance on defining your application scaling plan, or if you want to scale multiple resources beyond EC2, such as [Amazon DynamoDB](#) tables and indexes, or [Amazon ECS](#) tasks. At this time, to use AWS Auto Scaling, you must create your applications via [AWS CloudFormation](#) or [AWS Elastic Beanstalk](#). AWS Auto Scaling helps you manage all your scaling policies in one place for your applications making tuning easy and intuitive.

You should use Amazon EC2 Auto Scaling if you only need to scale Amazon EC2 Auto Scaling Groups, or just want to maintain the health of your EC2 fleet.



Amazon EC2 Auto Scaling ▾

Overview

Features

Pricing

Getting Started

Resources

FAQs

Amazon CloudWatch to send alarms to trigger scaling activities and Elastic Load Balancing (ELB) to distribute traffic to your instances within Amazon EC2 Auto Scaling groups. If you have predictable load changes, you can set a schedule through Amazon EC2 Auto Scaling to plan your scaling activities. Amazon EC2 Auto Scaling enables you to run your Amazon EC2 fleet at optimal utilization.

[Show less](#)

Q: What is fleet management and how is it different from dynamic scaling?

If your application runs on Amazon EC2 instances, then you have what's referred to as a 'fleet'. *Fleet management* refers to the functionality that automatically replaces unhealthy instances and maintains your fleet at the desired capacity. Amazon EC2 Auto Scaling fleet management ensures that your application is able to receive traffic and that the instances themselves are working properly. When Auto Scaling detects a failed [health check](#), it can replace the instance automatically.

The *dynamic scaling* capabilities of Amazon EC2 Auto Scaling refers to the functionality that automatically increases or decreases capacity based on load or other metrics. For example, if your CPU spikes above 80% (and you have an alarm setup) Amazon EC2 Auto Scaling can add a new instance dynamically.

[Show less](#)

Q: What is target tracking?



Amazon EC2 Auto Scaling ▾

Overview

Features

Pricing

Getting Started

Resources

FAQs

An Amazon EC2 Auto Scaling group contains a collection of EC2 instances that share similar characteristics and are treated as a logical grouping for the purposes of fleet management and dynamic scaling. For example, if a single application operates across multiple instances, you might want to increase the number of instances in that group to improve the performance of the application, or decrease the number of instances to reduce costs when demand is low. Amazon EC2 Auto Scaling will automatically adjust the number of instances in the group to maintain a fixed number of instances even if a instance becomes unhealthy, or based on criteria that you specify. You can find more information about Amazon EC2 Auto Scaling groups in the [Amazon EC2 Auto Scaling User Guide](#).

[Show less](#)

Q: What happens to my Amazon EC2 instances if I delete my EC2 Auto Scaling Group?

If you have an EC2 Auto Scaling group with running instances and you choose to delete the Amazon EC2 Auto Scaling group, the instances will be terminated and the EC2 Auto Scaling group will be deleted.

[Show less](#)

Q: How do I know when EC2 Auto Scaling is launching or terminating the EC2 instances in an EC2 Auto Scaling group?



Amazon EC2 Auto Scaling ▾

Overview

Features

Pricing

Getting Started

Resources

FAQs

Q: What is a launch configuration?

A launch configuration is a template that an EC2 Auto Scaling group uses to launch EC2 instances. When you create a launch configuration, you specify information for the instances such as the ID of the Amazon Machine Image (AMI), the instance type, a key pair, one or more security groups, and a block device mapping. If you've launched an EC2 instance before, you specified the same information in order to launch the instance. When you create an EC2 Auto Scaling group, you must specify a launch configuration. You can specify your launch configuration with multiple EC2 Auto Scaling groups. However, you can only specify one launch configuration for an EC2 Auto Scaling group at a time, and you can't modify a launch configuration after you've created it. Therefore, if you want to change the launch configuration for your EC2 Auto Scaling group, you must create a launch configuration and then update your EC2 Auto Scaling group with the new launch configuration. When you change the launch configuration for your EC2 Auto Scaling group, any new instances are launched using the new configuration parameters, but existing instances are not affected. You can see the [launch configurations](#) section of the EC2 Auto Scaling User Guide for more details.

[Show less](#)

Q: How many instances can an EC2 Auto Scaling group have?

You can have as many instances in your EC2 Auto Scaling group as your EC2 quota allows.

[Show less](#)



Amazon EC2 Auto Scaling ▾

Overview

Features

Pricing

Getting Started

Resources

FAQs

Q: Can I launch different types of EC2 instances in same EC2 Auto Scaling group?

EC2 Auto Scaling groups optimize for the case when all your instance types are the same. You can use the `AttachInstances` API to attach instances of different types to an Auto Scaling group, and you can also update your launch configuration so that any new instances in the group will be launched with a different instance type. However, this will not affect any of the existing instances.

[Show less](#)

Q: How can I implement changes across multiple instances in an EC2 Auto Scaling group?

You can use AWS CodeDeploy or CloudFormation to orchestrate code changes to multiple instances in your EC2 Auto Scaling group.

[Show less](#)

Q: If I have data installed in an EC2 Auto Scaling group, and a new instance is dynamically created later, is the data copied over to the new instances?

Data is not automatically copied from existing instances to new instances. You can use [lifecycle hooks](#) to copy the data, or an [Amazon RDS](#) database including replicas.

[Show less](#)



Amazon EC2 Auto Scaling ▾

Overview

Features

Pricing

Getting Started

Resources

FAQs

group settings. Amazon EC2 Auto Scaling always launches new instances such that they are balanced between zones as evenly as possible across the entire fleet. What's more, Amazon EC2 Auto Scaling only launches into Availability Zones in which there is available capacity for the requested instance type.

[Show less](#)

Q: What are lifecycle hooks?

Lifecycle hooks let you take action before an instance goes into service or before it gets terminated. This can be especially useful if you are not baking your software environment into an Amazon Machine Image (AMI). For example, launch hooks can perform software configuration on an instance to ensure that it's fully prepared to handle traffic before Amazon EC2 Auto Scaling proceeds to connect it to your load balancer. One way to do this is by connecting the launch hook to an AWS Lambda function that invokes RunCommand on the instance. Terminate hooks can be useful for collecting important data from an instance before it goes away. For example, you could use a terminate hook to preserve your fleet's log files by copying them to an Amazon S3 bucket when instances go out of service.

Visit [lifecycle hooks](#) in our Amazon EC2 Auto Scaling User Guide for more information.

[Show less](#)

Q: What are the characteristics of an “unhealthy” instance?



Amazon EC2 Auto Scaling

Overview

Features

Pricing

Getting Started

Resources

FAQs

Q: Can I suspend health checks (for example, to evaluate unhealthy instances)?

Yes, you can temporarily suspend Amazon EC2 Auto Scaling health checks by using the `SuspendProcesses` API. You can use the `ResumeProcesses` API to resume automatic health checks.

[Show less](#)

Q: Which health check type should I select?

If you are using Elastic Load Balancing (ELB) with your group, you should select an ELB health check. If you're not using ELB with your group, you should select the EC2 health check.

[Show less](#)

Q: Can I use Amazon EC2 Auto Scaling for health checks and to replace unhealthy instances if I'm not using Elastic Load Balancing (ELB)?

You don't have to use ELB to use Auto Scaling. You can use the EC2 health check to identify and replace unhealthy instances.

[Show less](#)



Amazon EC2 Auto Scaling

Overview

Features

Pricing

Getting Started

Resources

FAQs

A volume is attached to a new instance when it is added. Amazon EC2 Auto Scaling doesn't automatically add a volume when the existing one is approaching capacity. You can use the EC2 API to add a volume to an existing instance.

[Show less](#)

Q: What does the term “stateful instances” refer to?

When we refer to a stateful instance, we mean an instance that has data on it, which exists only on that instance. In general, terminating a stateful instance means that the data (or state information) on the instance is lost. You may want to consider using lifecycle hooks to copy the data off of a stateful instance before it's terminated, or enable instance protection to prevent Amazon EC2 Auto Scaling from terminating it.

[Show less](#)

Replacing Impaired Instances

Q: How does Amazon EC2 Auto Scaling replace an impaired instance?

When an impaired instance fails a health check, Amazon EC2 Auto Scaling automatically terminates it and replaces it with a new one. If you're using an Elastic Load Balancing load balancer, Amazon EC2 Auto Scaling gracefully detaches the impaired instance from the load balancer before



Amazon EC2 Auto Scaling

[Overview](#)[Features](#)[Pricing](#)[Getting Started](#)[Resources](#)[FAQs](#)

terminate first. You can configure this through the use of a termination policy. You can also use instance protection to prevent Amazon EC2 Auto Scaling from selecting specific instances for termination when scaling in. If you have data on an instance, and you need that data to be persistent even if your instance is scaled in, then you can use a service like S3, RDS, or DynamoDB, to make sure that it is stored off the instance.

[Show less](#)

Q: How long is the turn-around time for Amazon EC2 Auto Scaling to spin up a new instance at inService state after detecting an unhealthy server?

The turnaround time is within minutes. The majority of replacements happen within less than 5 minutes, and on average it is significantly less than 5 minutes. It depends on a variety of factors, including how long it takes to boot up the AMI of your instance.

[Show less](#)

Q: If Elastic Load Balancing (ELB) determines that an instance is unhealthy, and moved offline, will the previous requests sent to the failed instance be queued and rerouted to other instances within the group?

When ELB notices that the instance is unhealthy, it will stop routing requests to it. However, prior to discovering that the instance is unhealthy, some requests to that instance will fail.

[Show less](#)



Amazon EC2 Auto Scaling

Overview

Features

Pricing

Getting Started

Resources

FAQs

Q: How do I control access to Amazon EC2 Auto Scaling resources?

Amazon EC2 Auto Scaling integrates with [AWS Identity and Access Management \(IAM\)](#), a service that enables you to do the following:

- Create users and groups under your organization's AWS account
- Assign unique security credentials to each user under your AWS account
- Control each user's permissions to perform tasks using AWS resources
- Allow the users in another AWS account to share your AWS resources
- Create roles for your AWS account and define the users or services that can assume them
- Use existing identities for your enterprise to grant permissions to perform tasks using AWS resources

For example, you could create an IAM policy that grants the Managers group permission to use only the *DescribeAutoScalingGroups*, *DescribeLaunchConfigurations*, *DescribeScalingActivities*, and *DescribePolicies* API operations. Users in the Managers group could then use those operations with any Amazon EC2 Auto Scaling groups and launch configurations. With Amazon EC2 Auto Scaling resource-level permissions, you can restrict access to a particular EC2 Auto Scaling group or launch configuration.

For more information, see the [Controlling Access to Your Auto Scaling Resources](#) section of the Amazon EC2 Auto Scaling user guide.

[Show less](#)



Amazon EC2 Auto Scaling

[Overview](#)[Features](#)[Pricing](#)[Getting Started](#)[Resources](#)[FAQs](#)

If your AMI contains a CloudWatch agent, it's automatically installed on EC2 instances when you create an EC2 Auto Scaling group. With the stock Amazon Linux AMI, you need to install it (recommended, via yum).

[Show less](#)

Pricing

Q: What are the costs for using Amazon EC2 Auto Scaling?

Amazon EC2 Auto Scaling fleet management for EC2 instances carries no additional fees. The dynamic scaling capabilities of Amazon EC2 Auto Scaling are enabled by Amazon CloudWatch and also carry no additional fees. Amazon EC2 and Amazon CloudWatch service fees apply and are billed separately.

[Show less](#)

Learn more about Amazon EC2 Auto Scaling pricing
[Visit the pricing page](#)



Amazon EC2 Auto Scaling

Overview








Features

Pricing

Getting Started

Resources

FAQs

-  [Twitter](#)
-  [Facebook](#)
-  [Podcast](#)
-  [Twitch](#)
-  [AWS Blog](#)
-  [RSS News Feed](#)
-  [Email Updates](#)

AWS & Cloud Computing

[What is Cloud Computing?](#)

[What is Caching?](#)

[What is NoSQL?](#)

[What is DevOps?](#)

[What is Docker?](#)

[Products & Services](#)

[Customer Success](#)

[Economics Center](#)

[Architecture Center](#)

[Security Center](#)

[What's New](#)

[Whitepapers](#)

[AWS Blog](#)

[Events](#)

[Sustainable Energy](#)

[Press Releases](#)

[AWS in the News](#)

[Analyst Reports](#)

[Legal](#)

Solutions

[Websites & Website Hosting](#)

[Business Applications](#)

[Backup & Recovery](#)

[Disaster Recovery](#)



Amazon EC2 Auto Scaling

Overview

Features

Pricing

Getting Started

Resources

FAQs

[Oil & Gas](#)

[Automotive](#)

[Blockchain](#)

[Manufacturing](#)

Resources & Training

[Developers](#)

[Java on AWS](#)

[JavaScript on AWS](#)

[Mobile on AWS](#)

[PHP on AWS](#)

[Python on AWS](#)

[Ruby on AWS](#)

[.NET on AWS](#)

[SDKs & Tools](#)

[AWS Marketplace](#)

[User Groups](#)

[Support Plans](#)

[Service Health Dashboard](#)

[Discussion Forums](#)

[FAQs](#)

[Documentation](#)

[Articles & Tutorials](#)

[Quick Starts](#)

Manage Your Account

[Management Console](#)

[Billing & Cost Management](#)

[Subscribe to Updates](#)

[Personal Information](#)



Amazon EC2 Auto Scaling ▾

- Overview
- Features
- Pricing
- Getting Started
- Resources
- FAQs

Language [Bahasa Indonesia](#) | [Deutsch](#) | [English](#) | [Español](#) | [Français](#) | [Italiano](#) | [Português](#) | [Tiếng Việt](#) | [Türkçe](#)
[Русский](#) | [ไทย](#) | [日本語](#) | [한국어](#) | [中文 \(简体\)](#) | [中文 \(繁體\)](#)

[Site Terms](#) | [Privacy](#)

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.