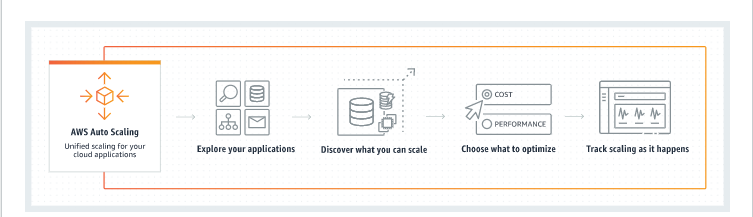
How it works

With AWS Auto Scaling, you configure and manage scaling for your scalable AWS resources through a scaling plan. It lets you choose scaling strategies to define how to optimize your resource utilization. You can optimize for availability, for cost, or a balance of both. Alternatively, you can leverage custom strategies for greater flexibility. Learn more



Benefits

Maintain performance

Right-size multiple resources to meet application performance targets.

Control costs

Automatically remove unneeded resources to avoid overspending.

Simplify scaling

Discover scalable resources and build a scaling plan in just a few steps.

Scale predictively

Use historical load metrics to learn your application’s traffic patterns and proactively scale resources ahead of time. Currently only available for EC2 Auto Scaling groups.

General

Q. What is AWS Auto Scaling?

AWS Auto Scaling is a new AWS service that helps you optimize the performance of your applications while lowering infrastructure costs by easily and safely scaling multiple AWS resources. It simplifies the scaling experience by allowing you to scale collections of related resources that support your application with just a few clicks. AWS Auto Scaling helps you configure consistent and congruent scaling policies across the full infrastructure stack backing your application. AWS Auto Scaling will automatically scale resources as needed to align to your selected scaling strategy, so you maintain performance and pay only for the resources you actually need.

Show less

Q. What are the benefits of AWS Auto Scaling?

AWS Auto Scaling is a fast, easy way to optimize the performance and costs of your applications.

Setup scaling quickly: AWS Auto Scaling provides a unified scaling experience for all of the scalable resources powering your application. You can see the average utilization for all of your scalable resources and quickly define target utilization levels for each group of like resources from a single, intuitive interface.

Make smart scaling decisions: AWS Auto Scaling lets you automate how groups of different resources respond to changes in demand. Easy-to-understand scaling strategies let you choose to optimize availability, costs, or a balance of both. AWS Auto Scaling automatically creates all of the scaling policies and sets targets for you based on your preference.

Automatically maintain performance: AWS Auto Scaling continually monitors resources underlying your application to make sure that they are operating at your desired performance levels. When demand spikes, AWS Auto Scaling automatically increases the capacity of constrained resources so you maintain a high quality of service.

Anticipate costs and avoid overspending: AWS Auto Scaling can help you optimize your utilization and cost efficiencies when consuming AWS services so you only pay for the resources you actually need. When demand drops, AWS Auto Scaling will automatically remove any excess resource capacity so you avoid overspending.

Show less

Q. When should I use AWS Auto Scaling?

You should use AWS Auto Scaling if you have an application that uses one or more scalable resources and experiences variable load. A good example would be an e-commerce web application that receives variable traffic through the day. It follows a standard three tier architecture with Elastic Load Balancing for distributing incoming traffic, Amazon EC2 for the compute layer, and DynamoDB for the data layer. In this case, AWS Auto Scaling will scale one or more EC2 Auto Scaling groups and DynamoDB tables that are powering the application in response to the demand curve.

Show less

Q. How can I get started with AWS Auto Scaling?

AWS Auto Scaling allows you to select your applications based on resource tags or AWS CloudFormation stacks. In just a few clicks, you can create a scaling plan for your application, which defines how each of the resources in your application should be scaled. For each resource, AWS Auto Scaling creates a target tracking scaling policy with the most popular metric for that resource type and keeps it at a target value based on your selected scaling strategy. To set target values for your resource metrics, you can choose from three predefined scaling recommendations that optimize availability, optimize costs, or balance the two. Or, if you prefer, you can define your own target values. AWS Auto Scaling also automatically sets the min/max values for the resources.

Show less

Scaling Options

Q. What are the different ways that I can scale AWS resources?

AWS customers have multiple options for scaling resources. Amazon EC2 Auto Scaling helps you ensure that you have the correct number of Amazon EC2 instances available to handle the load for your application. EC2 Auto Scaling can also detect when an instance is unhealthy, terminate it, and launch an instance to replace it. When you use EC2 Auto Scaling, your applications gain better fault tolerance, availability, and cost management.

To scale a resource other than EC2, you can use the Application Auto Scaling API, which allows you to define scaling policies to automatically scale your AWS resources or schedule one-time or recurring scaling actions. Application Auto Scaling can scale Amazon ECS services, Amazon EC2 Spot fleets, Amazon EMR clusters, Amazon AppStream 2.0 fleets, provisioned read and write capacity for Amazon DynamoDB tables and global secondary indexes, Amazon Aurora Replicas, and Amazon SageMaker endpoint variants.

To configure automatic scaling for multiple resources across multiple services, use AWS Auto Scaling to create a scaling plan for the resources underlying your application. AWS Auto Scaling is also used to create predictive scaling for EC2 resources.

Show less

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

You should use AWS Auto Scaling to manage scaling for multiple resources across multiple services. AWS Auto Scaling lets you define dynamic scaling policies for multiple EC2 Auto Scaling groups or other resources using predefined scaling strategies. Using AWS Auto Scaling to configure scaling policies for all of the scalable resources in your application is faster than managing scaling policies for each resource via its individual service console. It’s also easier, as AWS Auto Scaling includes predefined scaling strategies that simplify the setup of scaling policies. You should also use AWS Auto Scaling if you want to create predictive scaling for EC2 resources.

You should use EC2 Auto Scaling if you only need to scale Amazon EC2 Auto Scaling groups, or if you are only interested in maintaining the health of your EC2 fleet. You should also use EC2 Auto Scaling if you need to create or configure Amazon EC2 Auto Scaling groups, or if you need to set up scheduled or step scaling policies (as AWS Auto Scaling supports only target tracking scaling policies).

EC2 Auto Scaling groups must be created and configured outside of AWS Auto Scaling, such as through the EC2 console, Auto Scaling API or via CloudFormation. AWS Auto Scaling can help you configure dynamic scaling policies for your existing EC2 Auto Scaling groups.

Show less

Q. When should I use AWS Auto Scaling vs. Auto Scaling for individual services?

You should use AWS Auto Scaling to manage scaling for multiple resources across multiple services. AWS Auto Scaling enables unified scaling for multiple resources, and has predefined guidance that helps make it easier and faster to configure scaling. If you prefer, you can instead choose to use the individual service consoles, Auto Scaling API, or Application Auto Scaling API to scale individual AWS services. You should also use the individual consoles or API if you want to setup step scaling policies or scheduled scaling, as AWS Auto Scaling creates target tracking scaling policies only.

Show less

Q. What is Predictive Scaling?

Predictive Scaling is a feature of AWS Auto Scaling that looks at historic traffic patterns and forecasts them into the future to schedule changes in the number of EC2 instances at the appropriate times going forward. Predictive Scaling uses machine learning models to forecast daily and weekly patterns.

Auto Scaling enhanced with Predictive Scaling delivers faster, simpler, and more accurate capacity provisioning resulting in lower cost and more responsive applications. By predicting traffic changes, Predictive Scaling provisions EC2 instances in advance of changing traffic, making Auto Scaling faster and more accurate.

Show less

Q. Which services can I use Predictive Scaling with?

At this time, Predictive Scaling only generates schedules for EC2 instances.

Show less

Q. How can I use Predictive Scaling with target tracking?

Predictive Scaling works with in conjunction with target tracking to make your EC2 capacity changes more responsive to your incoming application traffic. While Predictive Scaling sets up the minimum capacity for your application based on forecasted traffic, target tracking changes the actual capacity based on the actual traffic at the moment. Target tracking works to track the desired capacity utilization levels over varying traffic conditions and addresses unpredicted traffic spikes and other fluctuations. Predictive Scaling and target tracking are configured together by a user to generate a scaling plan.

Show less

Q. What is a scaling plan?

A scaling plan is a collection of scaling instructions for multiple AWS resources. You configure a scaling plan by first selecting all the EC2 resources underlying your application in AWS Auto Scaling. Then you pick the resource utilization metric that you would like to track, such as CPU utilization, and set the value to track, for example 50%. Finally, you select the CloudWatch metric that represents your input traffic flow – you might have to set this up if you haven’t already.

The resource utilization metric and the incoming traffic metric are the key parameters for the scaling plan. The incoming traffic metric is used by Predictive Scaling to generate traffic forecasts. Based on these forecasts, Predictive Scaling then schedules future scaling actions to configure minimum capacity. Dynamic Scaling uses the resource utilization metric and its target value to dynamically change the EC2 capacity for your application over time as traffic varies.

Show less

Q. Can I configure a scaling plan without Predictive Scaling?

Yes, you can configure a scaling plan with only Dynamic Scaling and opt-out of Predictive Scaling. Conversely, you can also enable just Predictive Scaling without configuring Dynamic Scaling.

Show less

Q. How much historical data does Predictive Scaling need to generate the scaling plan?

Predictive Scaling needs up to two weeks of historical data but can generate a predictive scaling schedule with as little as a day's worth of data.

Show less

Q. How much into the future does Predictive Scaling forecast the traffic?

Every 24 hours, Predictive Scaling forecasts traffic 48 hours into the future and schedules capacity changes for those 48 hours.

Show less

Q. Can I configure Predictive Scaling to provision instances before an actual spike in traffic?

Yes, you can optionally configure buffer time to provision instances at some time before a predicted traffic change. This is useful for applications whose EC2 instances need some “warm-up” time before they are ready to serve application traffic .

Show less

Q. How much does Predictive Scaling cost?

As with other Auto Scaling features, Predictive Scaling is free to use. You pay for the resources being utilized for running your applications.

Show less

Q. How is AWS Auto Scaling different than the scaling capabilities for individual services?

The following table provides a comparison of AWS scaling options.

AWS

Auto Scaling

Amazon EC2

Auto Scaling

Auto Scaling

for Other Services

Resources you can scale EC2 Auto Scaling groups

EC2 Spot Fleets

ECS services

DynamoDB provisioned capacity for tables & GSIs

Aurora Replicas EC2 Auto Scaling groups EC2 Spot Fleets

ECS services

DynamoDB provisioned capacity for tables & GSIs

Aurora Replicas

EMR clusters

Appstream 2.0 fleet

Sagemaker endpoint variants

Scaling method Application-wide scaling using a unified interface

One Auto Scaling group at a time One resource at a time

Predictive Scaling Yes (EC2 Only) No No

Automatic discovery of all scalable

resources in your application Yes No No

Ability to scale multiple resources across multiple services with a unified interface Yes No

No

Guidance and recommendations

for setting up scaling policies Yes No No

Ability to create and setup

Auto Scaling groups No Yes

Not applicable

Ability to use Auto Scaling only for

EC2 Fleet Management No Yes Not applicable

Setup intelligent, self-optimizing

target tracking scaling policies\* Yes Yes Yes

Setup scheduled scaling actions No Yes Yes

Setup step scaling policies No Yes

Yes

Configure a scaling policy with different metrics and thresholds for each resource No Yes Yes

\* Recommended versus step scaling policies

Show less

Features

Q. What can I scale with AWS Auto Scaling?

You can use AWS Auto Scaling to setup scaling for the following resources in your application through a single, unified interface:

Amazon EC2 Auto Scaling groups

Amazon Elastic Container Service (ECS) services (currently ECS services cannot be discovered using resource tags)

Amazon EC2 Spot Fleets

Amazon DynamoDB throughput capacity

Aurora replicas for Amazon Aurora

Show less

Q. How does AWS Auto Scaling make scaling recommendations?

AWS Auto Scaling bases its scaling recommendations on the most popular scaling metrics and thresholds used for Auto Scaling. It also recommends safe guardrails for scaling by providing recommendations for the minimum and maximum sizes of the resources. This way you can get started quickly and can then fine tune your scaling strategy over time.

Show less

Q. How do I select an application stack within AWS Auto Scaling?

You can either select an AWS CloudFormation stack or select resources based on common resource tag(s). Please note that currently, ECS services cannot be discovered using tags.

Show less

Q. How does AWS Auto Scaling discover what resources can scale?

AWS Auto Scaling will scan your selected AWS CloudFormation stack or resources with the specified tags to identify the supported AWS resource types that can be scaled. Please note that currently, ECS services cannot be discovered using tags.

Show less

Availability and Pricing

Q. Which regions is AWS Auto Scaling available in?

AWS Auto Scaling is available in Asia Pacific (Mumbai), Asia Pacific (Tokyo), Asia Pacific (Seoul), Asia Pacific (Sydney), Canada (Central), US West (Northern California), Europe (London), Europe (Frankfurt), US East (Virginia), US East (Ohio), US West (Oregon), EU (Ireland), and Asia Pacific (Singapore).

Show less

Q. How much does AWS Auto Scaling cost?

Similar to Auto Scaling on individual AWS resources, AWS Auto Scaling is free to use. AWS Auto Scaling is enabled by Amazon CloudWatch, so service fees apply for CloudWatch and your application resources (such as Amazon EC2 instances, Elastic Load Balancing load balancers, etc.).

Show less