Auto Scaling

Product Introduction

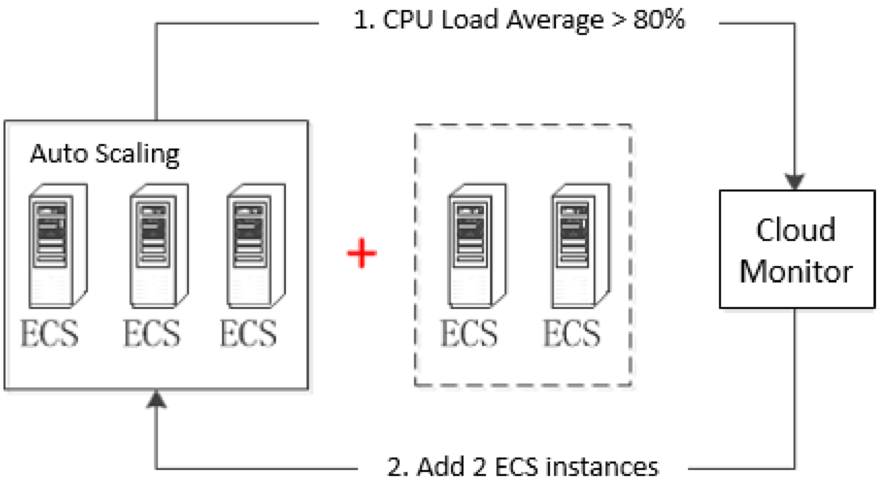
MORE THAN JUST CLOUD | Alibaba Cloud

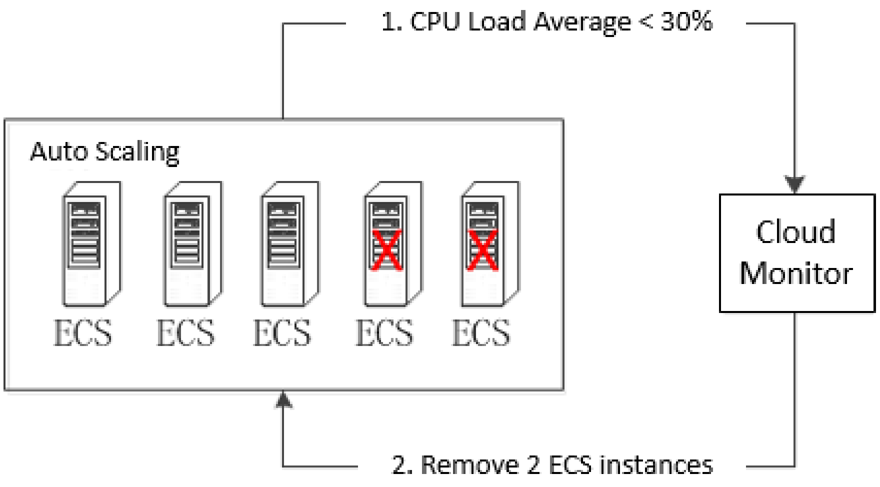
Product Introduction

Product overview

Auto Scaling is a management service that allows users to automatically adjust elastic computing resources according to business needs and policies.

- ECS instances can be added seamlessly to support traffic peaks.
- When traffic returns to a normal condition, ECS instances can be removed automatically to save costs.





Elastic scale-up

During peak traffic times, Auto Scaling will automatically add additional computing resources to the scaling group.

Elastic scale-down

When traffic returns to a normal condition, Auto Scaling will automatically release ECS resources to reduce costs.

Elastic self-healing

When an unhealthy instance is detected, Auto Scaling will automatically replace the instance with a new one in order to maintain uninterrupted service.

Glossary

Auto scaling

Auto Scaling is a management service that allows users to automatically adjust elastic computing resources according to business needs and policies. This way, ECS instances can be added automatically to support traffic peaks, and removed when traffic returns to a normal condition to save costs.

Scaling group

A scaling group is a collection of ECS instances with similar configuration deployed in an application scenario. The minimum and maximum number of ECS instances can be configured for the scaling group. Associated Server Load Balancer instances, RDS instances, and attributes can also be configured.

Scaling configuration

Scaling configuration defines the ECS instances information used for Auto Scaling.

Scaling rule

A scaling rule defines scaling activities, such as adding or removing ECS instances.

Scaling activity

When a scaling rule is triggered, a scaling activity takes place. Scaling activities describe the changes of the ECS instances in a scaling group.

Scaling trigger task

These tasks are used to trigger scaling rules, such as a scheduled task or a CloudMonitor alarm task.

Cool-down time

Once a scaling activity is completed within a scaling group, cool-down time refers to a period of time when no activities are executed.

Remarks

- A scaling group contains scaling configuration, scaling rules, and scaling activities.
- Scaling configuration, scaling rules, and scaling activities are associated with the lifecycle management of a scaling group. Deleting the scaling group also deletes the associated scaling configuration, scaling rules, and scaling activities.
- Scaling trigger tasks include scheduled tasks and CloudMonitor alarm tasks.
- Scheduled tasks are independent of the scaling group. Deleting the scaling group will not delete the scheduled tasks.
- CloudMonitor alarm tasks are independent of the scaling group. Deleting the scaling group does not delete the CloudMonitor alarm tasks.

Product functions

Alibaba Cloud Auto Scaling provides the following functions:

Scales up ECS instances horizontally according to your business needs by automatically adding or removing ECS instances.

Supports Sever Load Balancer configuration. When ECS instances are added or removed, Auto Scaling also adds or removes corresponding ECS instances from Sever Load Balancer instances.

Supports the RDS access whitelist. When ECS instances are added or removed, Auto Scaling will also add or remove their IP addresses from the RDS access whitelist.

Product features

On demand: Allocates resources to where they are needed, removing a user’ s need to predict demand and minimizing the impact of traffic bursts.

Automated: Automatically creates and releases ECS instances based on templates. Configures access whitelists for the Server Load Balancer and RDS services.

Rich: Allows configuration of multiple scaling modes simultaneously, such as timing, dynamic, custom, fixed, and healthy modes. External monitoring systems can be accessed through APIs.

Smart: Suits various complicated scenarios through smart scheduling.

Application scenarios

Video production and hosting: Auto Scaling allows video companies to automatically scale their resources to handle rising demands for popular events and programs.

Video streaming: Companies that cannot predict their business loads can use Auto Scaling to automatically scale their resources based on CPU usage, loads, or bandwidth usage.

Gaming: Auto Scaling can regularly scale up the system at pre-defined intervals. For example, scaling can be started at 12:00 AM and again from 6:00 PM to 9:00 PM, in order to effectively manage peak traffic intervals.

Scaling modes

There are several scaling modes:

Timing mode: Configures scheduled tasks to regularly add or remove ECS instances.

Dynamic mode: Automatically adds or removes ECS instances based on CloudMonitor performance indicators, such as CPU usage.

Fixed quantity mode: The **MinSize** attribute maintains a certain number of healthy ECS instances for real-time use in routine scenarios.

Custom mode: Uses an API to manually scale up or down ECS instances based on the monitoring system.

- Manually execute scaling rules.
- Manually add or remove existing ECS instances.
- After the MinSize or MaxSize attribute is adjusted, Auto Scaling will automatically create or release ECS instances to keep the numbers within the range.

Healthy mode: If ECS instances are not in running state, Auto Scaling will automatically remove or release them.

Multimode: Use of a combination of any of the preceding modes. For example, if you predict a peak time between 1:00 PM and 2:00 PM every day, the timing mode can be set to regularly create 20 ECS instances at that time. If you then determine that your actual load is higher than expected during the predicted peak time, use timing mode and dynamic mode together. Timing mode will create a fixed number of instances at a scheduled time, and if the load increases beyond what the scheduled extra instances can manage, dynamic mode will automatically add more.

Restrictions

Auto Scaling has the following constraints:

Applications deployed in the ECS instances for Auto Scaling must be stateless and horizontally scalable. The application status (for example, session) or data (for example, databases and logs) cannot be saved in the ECS instances. This is because Auto Scaling will automatically release ECS instances. If necessary, the status can be saved into an independent state server, database (for example, RDS), or centralized log storage (for example, Log Service).

Users can create a limited number of scaling groups, scaling configurations, scaling rules, scaling ECS instances, and scheduled tasks.

Development history

2015-08-27: Auto Scaling was released.

2014-10-15: Auto Scaling was beta tested.