Auto Scaling

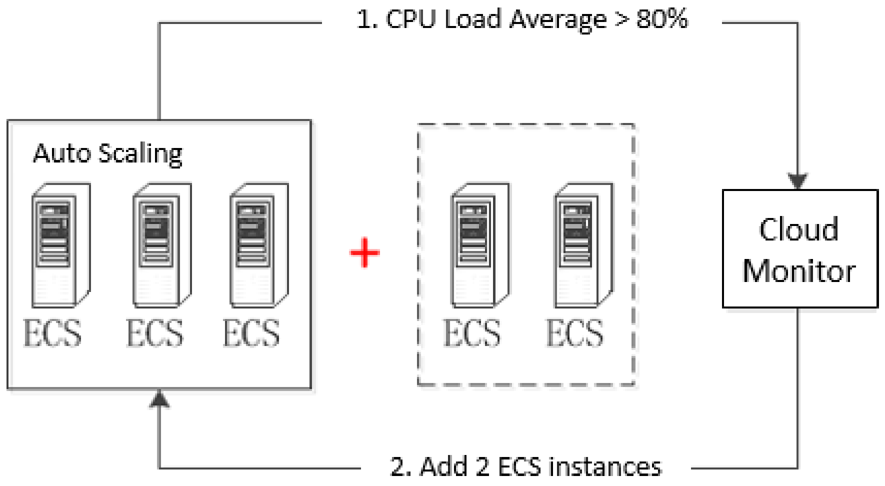
Product Introduction

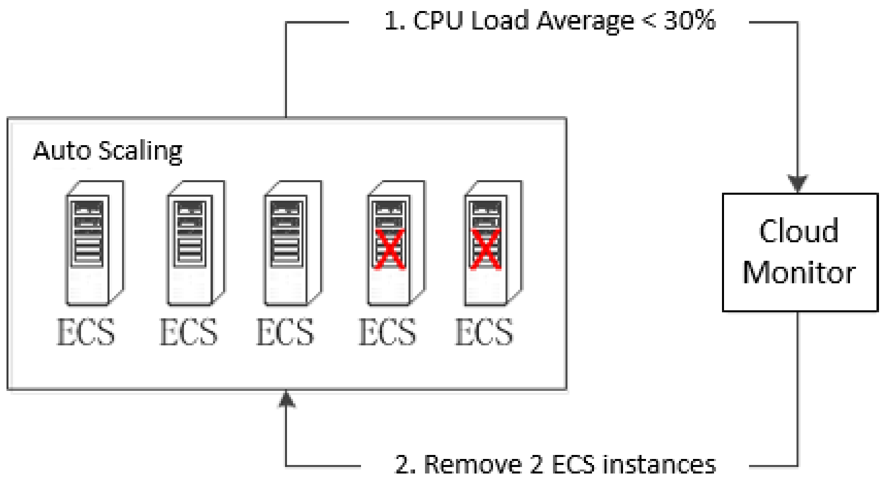
MORE THAN JUST CLOUD | Alibaba Cloud

Product Introduction

What is Auto Scaling

Auto Scaling enables you to dynamically scale your computing capacity up or down to meet the workload of your ECS instances according to scaling policies you specify, and reduces the need of manual provision. It monitors your resources, automatically adds or removes capacity in real time as demands change, so as to guarantee availability and save costs.





Dynamic scaling out

When demand spikes, Auto Scaling automatically adds computing resources to the scaling group.

Dynamic scaling in

When demand decreases, Auto Scaling automatically releases ECS resources to save costs.

Automatic replacing

Auto Scaling monitors ECS instances, automatically replaces the unhealthy instances with healthy ones to maintain availability.

Benefits

Overview

Automatically add or remove ECS instances when demand on your application increases or decreases.

Automatically configure the ECS instances of Sever Load Balancer.

Supports configure the ApsaraDB for RDS whitelist.

Features

On demand: Adjust resources to fit the demand curve in real time. You do not have to worry about your computing capacity when demand surges.

Automated: Automatically create and release ECS instances based on policies you specify. Configure the Server Load Balancer and RDS whitelists with no manual operation.

Flexible: You can setup scheduled scaling, dynamic scaling based on targets monitored, scaling fixed number of instances, and automated replacing of unhealthy instances. It also can use external monitoring systems through APIs.

Intelligent: Can be applied to complicated scenarios.

Scenarios

Video sharing: Workload surges during holidays and festivals. Computing resources have to be scaled out automatically in real time.

Video streaming: Demand curve is difficult to predict manually. Computing resources have to be scaled out based on CPU usage, workload, or bandwidth.

Gaming: Demand increasing starts at 12:00 and lasts from 18:00 to 21:00, scheduled scaling is needed.

Scaling modes

Scheduled scaling: You tell Auto Scaling to perform a scaling operation at specified times. For example, scaling up at 13:00 every day.

Dynamic scaling: Auto Scaling dynamically scales up and down by tracking targets. You select a metric and set a target value. Auto Scaling creates the CloudMonitor alarms that trigger the scaling policy. The scaling policy adds or removes capacity as required to keep the metric at, or close to, the specified target value.

Capacity maintaining: You setup the **MinSize** to maintain the minimum number of **running** healthy instances in the scaling group.

Customized target tracking: Uses API to manually scale based on metrics from your own monitoring system.

- Manually run scaling policy.
- Manually add or remove ECS instances.
- Automatically adjust the number of your ECS instances to lie between the MinSize

and MaxSize you setup.

Health check: Automatically release instances with status other than **running** according to policies you specify.

Multimode: Combine multiple scaling modes when demand of your application is hard to predict. For example, you setup to scale out 20 ECS instances during 13:00 ~ 14:00 everyday, but the actual demand may need more instances, then you can use this scheduled scaling together with other scaling modes to better follow the demand changes.

Limits

Applications deployed in the ECS instances for Auto Scaling must be stateless and scalable.

Auto Scaling automatically releases ECS instances, so the application status (such as sessions) or data (such as databases and logs) must not be saved in the ECS instances. If necessary, you can save this kind of data in independent state servers, databases (such as RDS), or centralized log storage (such as Log Service).

The instances added by Auto Scaling cannot be automatically added to ApsaraDB for Memcache whitelist, you must do it manually.

Auto Scaling cannot scale the specifications of your instances, such as CPU, RAM, and bandwidth.

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

Development history

Development history

2015-08-27: Auto Scaling was released.

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

Glossary

Auto Scaling

Auto Scaling is a management service that allows users to automatically adjust elastic computing resources according to application demand and scaling policies you specify. It automatically creates ECS instances when demand peaks to improve capacity, and release them when demand decreases to save costs.

Scaling group

A scaling group is a collection of ECS instances with similar configuration applying to a scenario. You can setup the minimum and maximum number of ECS instances, Server Load Balancer, and RDS for the scaling group.

Scaling configuration

Scaling configuration defines the specifications of ECS instances used to scale.

Scaling rule

A scaling rule specifies the scaling operation, such as whether, when, and how to create or release ECS instances.

Scaling activity

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

Scaling trigger task

Tasks that can trigger scaling rules, such as the scheduled task or CloudMonitor alarm task.

Cool-down time

The time Auto Scaling waited for the previous scaling activity to complete before resuming scaling activities.

Remarks

- A scaling group includes settings of 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 does not lead to the deletion the scheduled tasks.
- CloudMonitor alarm tasks are independent of the scaling group. Deleting the scaling group does not lead to the deletion of the CloudMonitor alarm tasks.