Alibaba Cloud Auto Scaling

Product Introduction

Issue: 20190528

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. **You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.**
2. **No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company, or individual in any form or by any means without the prior written consent of Alibaba Cloud.**
3. **The content of this document may be changed due to product version upgrades , adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.**
4. **This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults " and "as available" basis. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity , applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, special, or punitive damages, including lost profits arising from the use**

or trust in this document, even if Alibaba Cloud has been notified of the possibility of such a loss.

5. By law, all the content of the Alibaba Cloud website, including but not limited to works, products, images, archives, information, materials, website architecture, website graphic layout, and webpage design, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of the Alibaba Cloud website, product programs, or content shall be used, modified , reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates . The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates).
6. Please contact Alibaba Cloud directly if you discover any errors in this document.

Generic conventions

Table -1: Style conventions

Style	Description	Example
	This warning information indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	 Danger: Resetting will result in the loss of user configuration data.
	This warning information indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	 Warning: Restarting will cause business interruption. About 10 minutes are required to restore business.
	This indicates warning information, supplementary instructions, and other content that the user must understand.	 Notice: Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructions, best practices, tips, and other content that is good to know for the user.	 Note: You can use Ctrl + A to select all files.
>	Multi-level menu cascade.	Settings > Network > Set network type
Bold	It is used for buttons, menus , page names, and other UI elements.	Click OK.
`Courier font`	It is used for commands.	Run the `cd / d C :/ windows` command to enter the Windows system folder.
Italics	It is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	It indicates that it is a optional value, and only one item can be selected.	`ipconfig` *[-all|-t]*

Style	Description	Example
{} or {a|b}	It indicates that it is a required value, and only one item can be selected.	`swich {stand | slave}`

Contents

1 What is Auto Scaling?

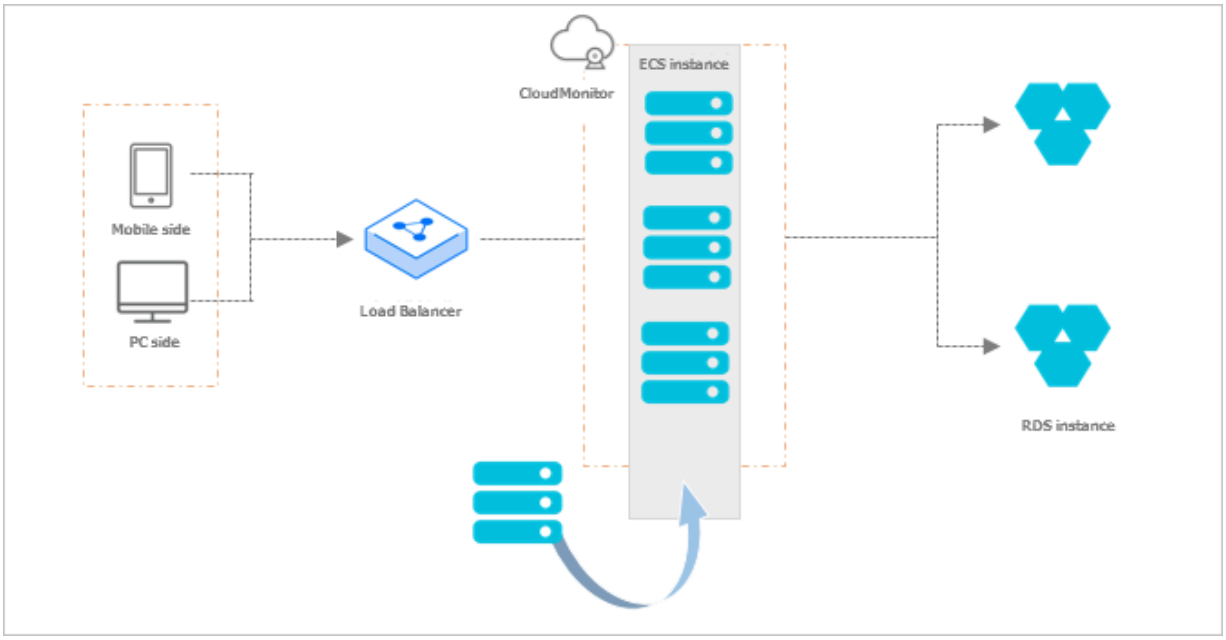
Auto Scaling automatically adjusts the volume of your elastic computing resources to meet your changing business needs. Based on the scaling rules that you set, Auto Scaling automatically adds ECS instances as your business needs grow to ensure that you have sufficient computing capabilities. When your business needs fall, Auto Scaling automatically reduces the number of ECS instances to save on costs.

Scaling-out

When you upgrade your business, Auto Scaling automatically upgrades the underlying resources for you to avoid access delays and excessive resource loads.

You can set CloudMonitor to monitor your ECS instance usage in real time. For example, when CloudMonitor detects that ECS instance vCPU usage exceeds 80% in a scaling group, Auto Scaling elastically scales out your ECS resources based on the scaling rules that you set. This is done by automatically creating a suitable number of ECS instances and automatically adding these ECS instances to the Server Load Balancer instance and RDS instance whitelist. For more details, see *create a scaling group* in Auto Scaling and *monitor Auto Scaling* in CloudMonitor.

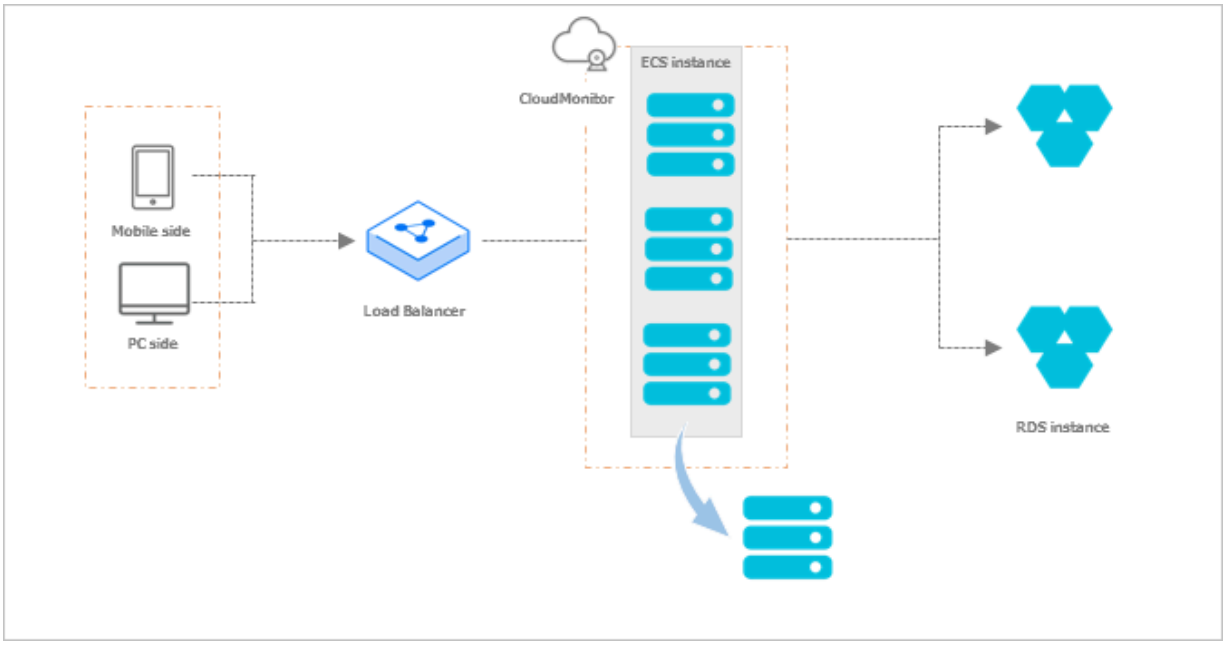
Note:
When Auto Scaling scales out your ECS resources, new ECS instances are automatically created and added to the scaling group by using the active scaling configuration as the template. You can log on to the ECS console to do operations on these ECS instances, such as starting, stopping, and connecting to them.



Scaling-in

When your business needs fall, Auto Scaling automatically releases underlying resources for you to avoid wasting resources.

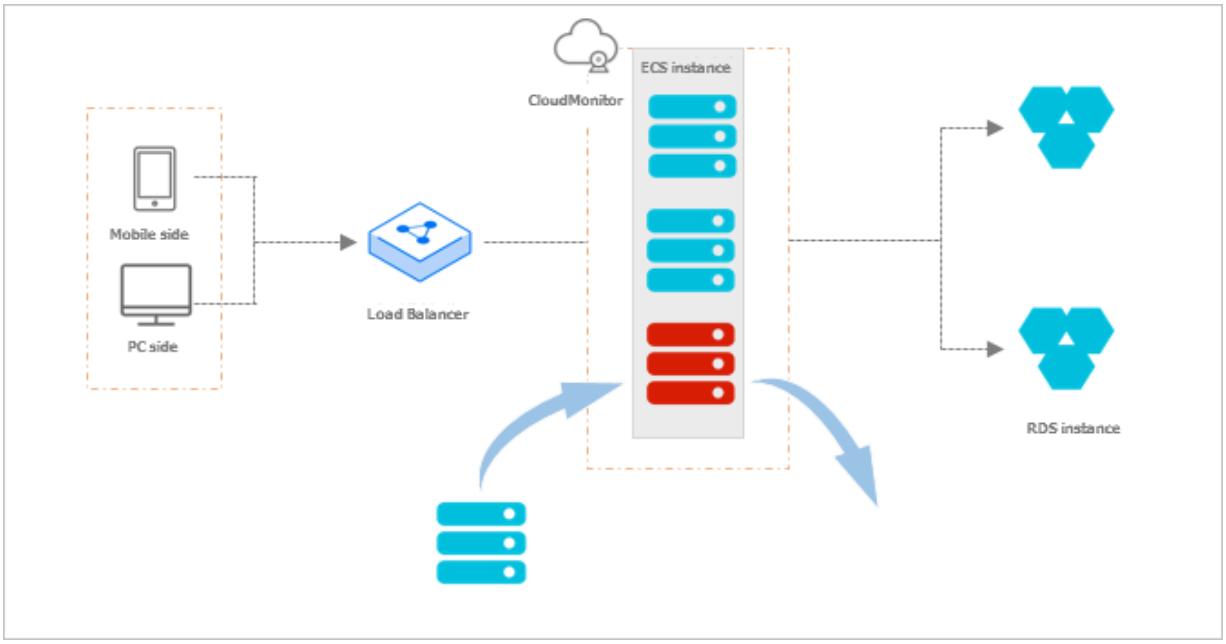
You can set CloudMonitor to monitor your ECS instance usage in real time. For example, when CloudMonitor detects that ECS instance vCPU usage falls below 30% in a scaling group, Auto Scaling elastically scales in your ECS resources based on the scaling rules that you set. This is done by automatically releasing a suitable number of ECS instances and automatically removing these ECS instances from the Server Load Balancer instance and RDS instance whitelist. For more details, see *removal policy* in Auto Scaling and *monitor Auto Scaling* in CloudMonitor.



Flexible recovery

Auto Scaling provides a health check function and automatically monitors the health of ECS instances within scaling groups, so the number of healthy ECS instances in a scaling group does not fall below the minimum value that you set.

When Auto Scaling detects that an ECS instance is not healthy, Auto Scaling automatically releases the unhealthy ECS instance, creates a new ECS instance, and adds the new instance to the Server Load Balancer instance and RDS instance whitelist. For more information, see *remove an unhealthy ECS instance*.



References

- *What is ECS*
- *What is RDS*
- *What is Server Load Balancer*
- *CloudMonitor overview*

2 Benefits

This article introduces the function, product features and scenarios of Auto Scaling.

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.

3 Scaling modes

This article introduces the scaling modes of Auto Scaling.

- 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 the 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.

4 Limits

This article introduces the limits for Auto Scaling.

Auto Scaling has the following 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.

5 History

This article introduces the development history of Auto Scaling.

The development history is as follows:

- August 27, 2015: Auto Scaling was released.
- October 15, 2014: Auto Scaling was beta tested.

6 Glossary

This article explains the related terms of Auto Scaling.

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. During the cool-down time, no other scaling activities can be performed in the same scaling group.

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.