ALIBABA CLOUD

Alibaba Cloud

Auto Scaling

Quick Start

Document Version: 20211028

Alibaba Cloud

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 because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should 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 this document based on the "status quo", "being defective", and "existing functions" of its products and services. 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 take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, 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 contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, 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 this document 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 directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
Danger	A danger notice 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.
Warning	A warning notice 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 restart an instance.
Notice	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	**Notice:** If the weight is set to 0, the server no longer receives new requests.
Note	A note indicates supplemental instructions, best practices, tips, and other content.	**Note:** You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Table of Contents

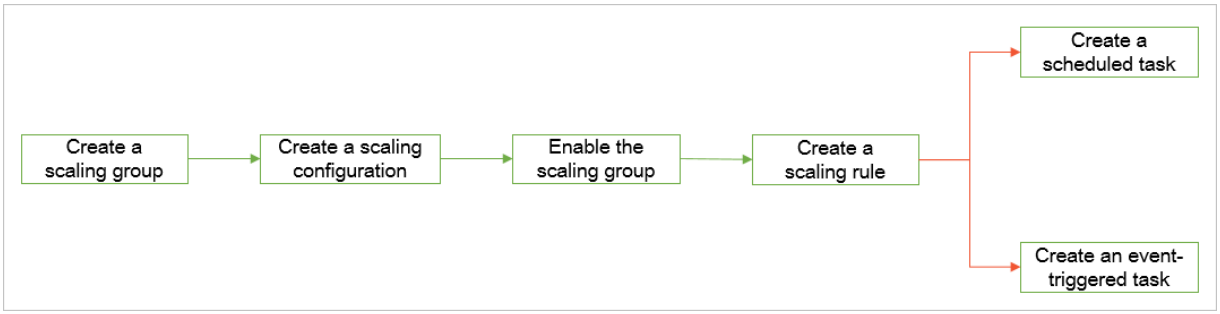
1.Usage processes

This topic describes how to use Auto Scaling to build high-elasticity and high-availability applications.

Auto Scaling provides the following capabilities for applications:

- High elasticity: Auto Scaling automatically creates and releases ECS instances based on your settings without the need for manual intervention.
- High availability: Auto Scaling automatically performs checks on ECS instances, releases stopped instances, and then creates instances.

The following figure shows the general process of using Auto Scaling.



The following section describes the process:

1. Create a scaling group.

   A scaling group is a basic management unit when you use Auto Scaling to manage ECS instances on which your business is deployed. Scaling groups are used to manage ECS instances that are applied to the same scenario and can be associated with multiple SLB instances and ApsaraDB RDS instances. After a scaling group is associated with SLB and ApsaraDB RDS instances, ECS instances that are added to the scaling group are automatically added as backend servers of the associated SLB instances. The internal IP addresses of these instances are automatically added to the whitelists of the associated ApsaraDB RDS instances.

2. Create a scaling configuration.

   A scaling configuration is a template used by Auto Scaling to automatically create ECS instances. You can create multiple scaling configurations for a scaling group. However, only one scaling configuration can be active at a time. For more information, see Overview of instance configuration sources.

   > **Note** If you use a launch template or an existing instance as the configuration source when you create a scaling group, you can directly enable the scaling group without the need to manually create a scaling configuration.

3. Enable the scaling group.

   After you create a scaling configuration for the first time, you are prompted to enable the scaling group. You can also enable the scaling group on the Scaling Groups page. For more information, see Enable a scaling group.

4. Create a scaling rule.

   A scaling rule is used to specify information such as the number of ECS instances to be scaled or intelligently set the boundary values of a scaling group. You can create scaling rules of the corresponding type based on your business needs. For more information, see Scaling rule overview.

5. Create a scaling task.

   After a scaling rule is created, you can use a scaling task to automatically execute the scaling rule. Auto Scaling supports the following types of scaling tasks:

   - Scheduled tasks

     If you can predict the time when your business loads fluctuate, you can use scheduled tasks to automatically scale ECS instances at the specified time. You can set the recurrence for scheduled tasks to meet your periodic requirements for automatic scaling.

   - Event-triggered tasks

     If you want to automatically scale ECS instances based on their running metrics, you can use event-triggered tasks. An event-triggered task dynamically manages ECS instances in a scaling group based on monitoring metrics from Cloud Monitor. For more information, see Event-triggered task overview.

2.Manage the Auto Scaling service linked role

This topic describes how to use the Auto Scaling service linked role to grant access permissions on Alibaba Cloud resources to Auto Scaling.

Prerequisites

If you are a Resource Access Management (RAM) user, make sure that you are granted permissions on Auto Scaling. For more information, see Grant permissions to a RAM user.

The following section shows the permissions to be added:

> **Note** Replace <account ID> with the ID of your Alibaba Cloud account.

```
{
  "Statement": [
    {
      "Action": [
        "ram:CreateServiceLinkedRole"
      ],
      "Resource": "acs:ram:*:<account ID>:role/*",
      "Effect": "Allow",
      "Condition": {
        "StringEquals": {
          "ram:ServiceName": [
            "ess.aliyuncs.com"
          ]
        }
      }
    }
  ],
  "Version": "1"
}
```

Context

The Auto Scaling service linked role (AliyunServiceRoleForAutoScaling) is a RAM role that enables Auto Scaling to access other Alibaba Cloud resources. You can use AliyunServiceRoleForAutoScaling to enable Auto Scaling to access Elastic Compute Service (ECS), Virtual Private Cloud (VPC), ApsaraDB RDS (RDS), Server Load Balancer (SLB), Operation Orchestration Service (OOS), Message Service (MNS), and Cloud Monitor. For more information, see Service-linked roles.

Create AliyunServiceRoleForAutoScaling

When you use Auto Scaling, the system checks whether the AliyunServiceRoleForAutoScaling role is attached to your account. If the AliyunServiceRoleForAutoScaling role is not attached to your account, the system creates the role for your account.

The AliyunServiceRolePolicyForAutoScaling system policy is attached to AliyunServiceRoleForAutoScaling. System policies attached to service linked roles are defined and used by corresponding Alibaba Cloud services. You cannot add, modify, or delete the permissions of service linked roles. You can view policies attached to a RAM role in the RAM role details. For more information, see View the basic information about a RAM role.

Delete AliyunServiceRoleForAutoScaling

If you do not need the AliyunServiceRoleForAutoScaling service linked role for the moment and understand the impacts of not using the role, you can delete it. For example, when you do not need scaling groups to create and manage resources, you can delete the service linked role. For more information, see Delete a RAM role.

> **Note** Before you delete AliyunServiceRoleForAutoScaling, you must delete resources of Auto Scaling in all regions in your current account, including scaling groups, scheduled tasks, and event-triggered tasks. Otherwise, AliyunServiceRoleForAutoScaling cannot be deleted.

After you delete AliyunServiceRoleForAutoScaling, you cannot use Auto Scaling to create or manage resources.

3.Create a scaling group and a scaling configuration

A scaling group is a group of Elastic Compute Service (ECS) instances that can be dynamically scaled based on the preconfigured rule. A scaling configuration is a template that is used by a scaling group to create ECS instances. To enable automatic scaling, you must first create both a scaling group and a scaling configuration. This topic describes how to create a scaling group, create a scaling configuration for the scaling group, and enable the scaling group. The minimum number of ECS instances in a scaling group is one.

Prerequisites

Before you proceed, make sure that the following operations are performed:

- Manage the Auto Scaling service linked role
- Create a security group

> **Note** The security group and the scaling group must be located in the same region.

Context

After a scaling group is created, the region of the scaling group cannot be changed. For more information, see 地域和可用区.

Step 1: Create a scaling group

In this example, a scaling group is created with simple settings. For more information about parameters, see Create a scaling group.

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.

   In this example, the operations are performed on the Scaling Groups page. If you do not have scaling resources in the current region, the welcome page is displayed after you log on to the console. To create scaling resources, follow the on-screen instructions.
3. In the top navigation bar, select a region.
4. In the upper-left corner of the Scaling Groups page, click **Create**.
5. Configure parameters for the scaling group and click **OK**.

   The following table describes the parameters that are used in this example. For information about the parameters that are not described in the following table, use the default values.

Parameter	Example value	Description
Scaling Group Name	MyFirstScalingGroup	None

Parameter	Example value	Description
Instance Configuration Source	**Create from Scratch**	If you create a scaling group from scratch, the scaling group that you created does not have an instance configuration source. You must create a scaling configuration for the scaling group.
Tag	ESS:Documentation	Tags are used to categorize the scaling groups for easy management.
Minimum Number of Instances	1	A scaling group must contain at least one ECS instance. If the number of ECS instances in the scaling group is less than one, an instance is automatically created for the scaling group.
Maximum Number of Instances	3	A maximum of three ECS instances can be created in a scaling group. If the number of ECS instances is greater than three, the extra ECS instances are automatically removed from the scaling group.
Network Type	**Classic Network**	When you create a scaling configuration, you can select only instance types that support the classic network.

6. In the **Create Scaling Group** dialog box, click **OK**.

Step 2: Create a scaling configuration and enable the scaling group

In this example, a scaling configuration is created with simple settings. For more information about parameters, see Create a scaling configuration.

1. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
2. In the upper part of the page, click the **Instance Configuration Sources** tab.
3. In the upper-left corner of the page, click the **Scaling Configurations** tab.
4. Click **Create Scaling Configuration**.
5. Follow the on-screen instructions to create the scaling configuration and click **OK**.

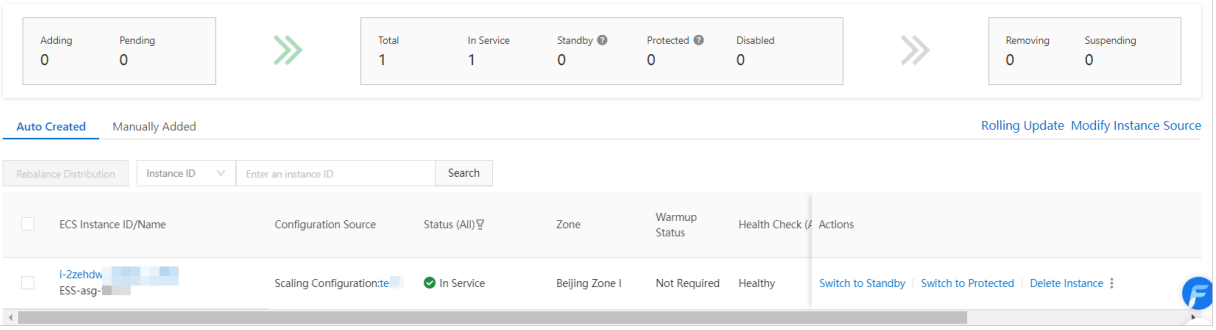
   The following table describes the parameters that are used in this example. For information about the parameters that are not described in the following table, use the default values.

Step	Parameter	Example value	Description
Basic Configurations	**Billing Method**	**Pay-As-You-Go**	Auto Scaling is free of charge. However, you must pay for the ECS instances that are added to the scaling group based on the pricing of ECS. For more information, see Billing overview.
	Instance Type	ecs.mn4.small, an instance type of the shared general-purpose instance family mn4.	Scaling groups in a virtual private cloud (VPC) support more newly released instance types. For more information about how to create a scaling group in a VPC, see Create a scaling group.
	Image	Public image: CentOS 7.6 64-bit	After the instance is started, the operating system and application data of the image are copied to the system disk.
	Public IP Address	Pay-by-traffic. A peak bandwidth of 1 Mbit/s	You are charged for the outbound Internet traffic. The peak bandwidth is 1 Mbit/s.
	Security Group	sg-bp18kz60mefsicfg****	Select the security group that you created.
System Configurations	**Logon Credentials**	**Set Later**	After the ECS instances are created, you can manually set the password for ECS instances.
Preview	**Scaling Configuration Name**	MyFirstScalingConfiguration	None

6. In the **Created** dialog box, click **Enable Configuration**.
7. In the **Enable Scaling Configuration** message, click **OK**.
8. In the **Enable Scaling Group** message, click **OK**.

Result

A scaling group must have at least one ECS instance. Therefore, Auto Scaling automatically creates an ECS instance for the scaling group after the scaling group is enabled. You can view the ECS instance in the ECS instance list. Configurations specified in the scaling configuration are applied to the ECS instance that is automatically created. For more information, see View ECS instances.



What's next

After the scaling group is enabled, you can create scaling rules for the scaling group and create scheduled tasks and event-triggered tasks to automatically scale ECS instances in the scaling group. For more information, see the following topics:

- Automatically add ECS instances
- Automatically remove ECS instances

4.Automatically add ECS instances

Auto Scaling executes a scaling rule to add or remove ECS instances by using scheduled tasks. This topic describes how to automatically add ECS instances to a scaling group. Two ECS instances are added to a scaling group in this topic.

Prerequisites

The following operations are performed:

- Manage the Auto Scaling service linked role
- Create a scaling group and a scaling configuration

Context

The MyFirstScalingGroup scaling group and MyFirstScalingConfiguration scaling configuration are used in this topic to demonstrate how to automatically add ECS instances. The scaling group already contains one ECS instance.

Step 1: Create a scaling rule

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
4. In the upper part of the page, click the **Scaling Rules and Activities** tab, and then click the **Scaling Rules** tab.
5. In the upper-left corner of the Scaling Rules tab, click **Create Scaling Rule**.
6. Configure parameters for the scaling rule and click **OK**.

   The following table describes the parameters used in this example. For parameters that are not described in the following table, use the default values.

Parameter	Example
Rule Name	Add2
Rule Type	Simple Scaling Rule
Operation	Add 2 Instances

Step 2: Create a scheduled task

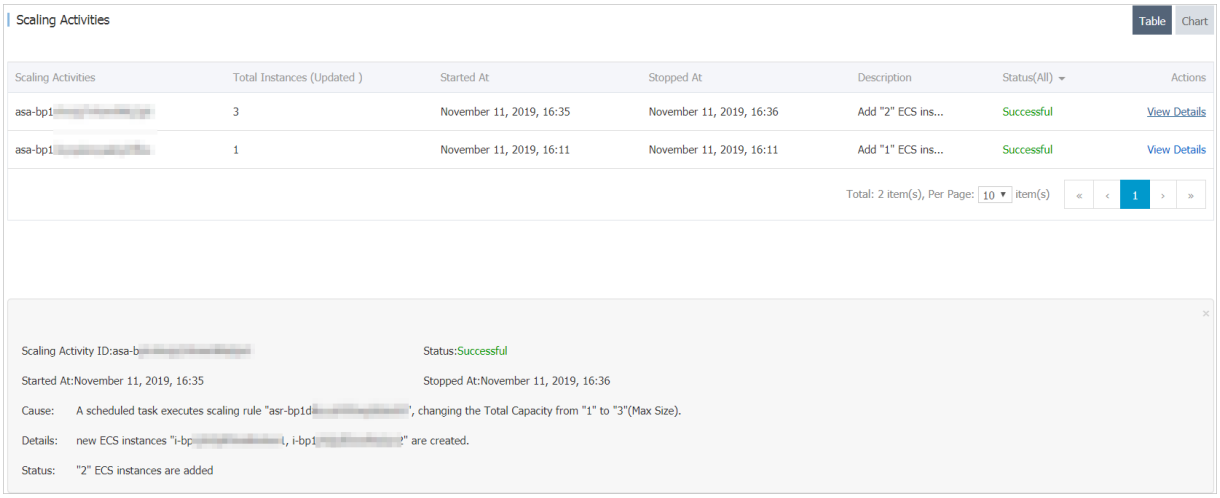
1. In the left-side navigation pane, choose **Scaling Tasks > Scheduled Tasks**.
2. In the upper-left corner of the Scheduled Tasks page, click **Create Scheduled Task**.
3. Configure parameters for the scheduled task and click **OK**.

   The following table describes the parameters used in this example. For parameters that are not described in the following table, use the default values.

Parameter	Example	Description
Task Name	ScheduledScalingOut	None
Description	Add two ECS instances at the scheduled time.	None
Executed At	2019-11-11 16:35	Five minutes after the current time.
Scaling Group	MyFirstScalingGroup	Executes the scheduled task for the MyFirstScalingGroup scaling group.
Scaling Method	Select an existing rule	None
Simple Scaling Rule	Add2	Executes the Add2 scaling rule to add two ECS instances to the scaling group.

Result

At 16:35 on November 11, 2019, the ScheduledScalingOut scheduled task automatically executes the Add2 scaling rule to add two ECS instances to the MyFirstScalingGroup scaling group. For more information, see the Scaling Activities page.



5.Automatically remove ECS instances

Auto Scaling monitors the ECS instances in a scaling group by using event-triggered tasks and automatically executes a scaling rule to add or remove ECS instances when an alert is triggered. This topic describes how to use event-triggered tasks to automatically remove ECS instances from a scaling group.

Prerequisites

The following operations are performed:

- Manage the Auto Scaling service linked role
- Create a scaling group and a scaling configuration
- Automatically add ECS instances

Context

The MyFirstScalingGroup scaling group is used in this topic to demonstrate how to automatically remove ECS instances. The scaling group already contains three ECS instances.

Step 1: Create a scaling rule

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
4. In the upper part of the page, click the **Scaling Rules and Activities** tab, and then click the **Scaling Rules** tab.
5. In the upper-left corner of the Scaling Rules tab, click **Create Scaling Rule**.
6. Configure parameters for the scaling rule and click **OK**.

   The following table describes the parameters used in this example. For parameters that are not described in the following table, use the default values.

Parameter	Example
Rule Name	To1
Rule Type	Simple Scaling Rule
Operation	Change to 1 Instances

Step 2: Create an event-triggered task

1. In the left-side navigation pane, choose **Scaling Tasks > Event-Triggered Tasks**.
2. In the upper-right corner of the Event-Triggered Tasks page, click **Create Event-Triggered Task**.

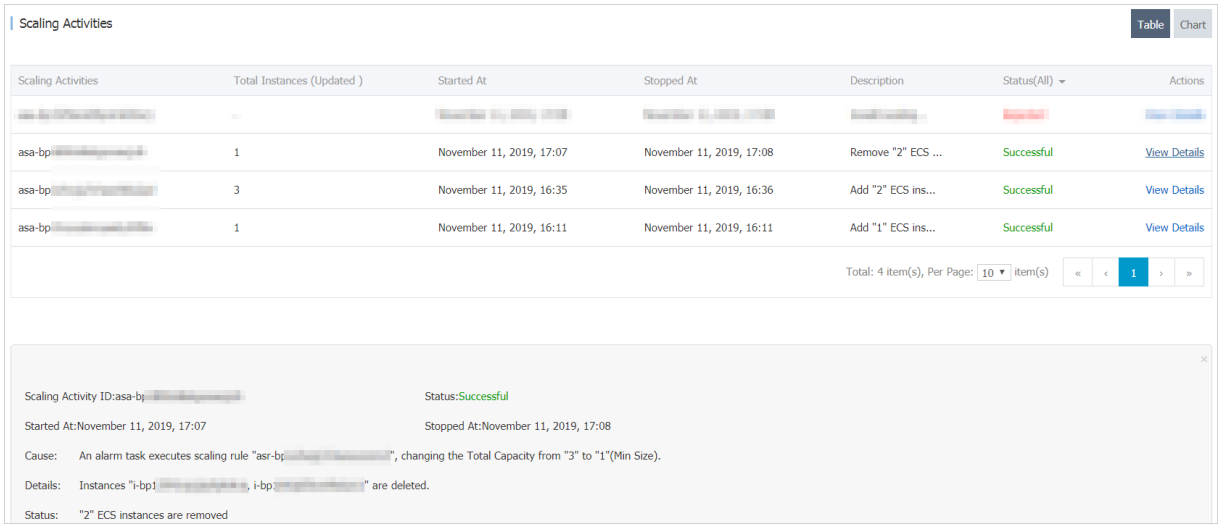
3. Configure parameters for the event-triggered task and click **OK**.

   The following table describes the parameters used in this example. For parameters that are not described in the following table, use the default values.

Parameter	Example	Description
Task Name	EventTriggeredScalingIn	None
Description	Remove ECS instances when the average CPU utilization is less than 10%.	None
Resource Monitored	MyFirstScalingGroup	Monitors metrics of the MyFirstScalingGroup scaling group.
Monitoring Type	System Monitoring	None
Monitoring Metrics	(ECS) CPU Usage	Monitors the CPU utilization of ECS instances in the scaling group.
Reference Period (Minutes)	1	Collects data once every one minute.
Condition	Average <= 10%	If the average CPU utilization of ECS instances in the scaling group is less than or equal to 10%, data is recorded once.
Trigger After	5 Times	If the average CPU utilization of ECS instances in the scaling group is less than or equal to 10% for five times in a row, an alert is triggered.
Triggered Rule	To1	When the alert is triggered, the To1 scaling rule is executed. Auto Scaling adjusts the number of ECS instances in the scaling group to one.

Result

The EventTriggeredScalingIn event-triggered task monitors the average CPU utilization of the three ECS instances in the MyFirstScalingGroup scaling group. The monitoring metric is calculated once every minute. If the average CPU utilization is less than or equal to 10% for five times in a row, an alert is triggered. The EventTriggeredScalingIn event-triggered task automatically executes the To1 scaling rule to adjust the number of ECS instances in the MyFirstScalingGroup scaling group to one. For more information, see the Scaling Activities page.


Scaling Activities
Table
Chart
Scaling Activities
Total Instances (Updated)
Started At
Stopped At
Description
Status(All)
Actions
asa-bp
1
November 11, 2019, 17:07
November 11, 2019, 17:08
Remove "2" ECS ...
Successful
View Details
asa-bp
3
November 11, 2019, 16:35
November 11, 2019, 16:36
Add "2" ECS ins...
Successful
View Details
asa-bp
1
November 11, 2019, 16:11
November 11, 2019, 16:11
Add "1" ECS ins...
Successful
View Details
Total: 4 item(s), Per Page: 10 item(s)
1
Scaling Activity ID:asa-b
Status:Successful
Started At:November 11, 2019, 17:07
Stopped At:November 11, 2019, 17:08
Cause: An alarm task executes scaling rule "asr-b", changing the Total Capacity from "3" to "1"(Min Size).
Details: Instances "i-bp1 , i-bp" are deleted.
Status: "2" ECS instances are removed