ALIBABA CLOUD

# Alibaba Cloud

## Auto Scaling

## Scaling Group

Document Version: 20220317

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.Scaling group

## 1.1. Scaling group status

This topic introduces the status of the scaling group.

| Status | API indicator |
| --- | --- |
| Creating | Inactive |
| Created | Inactive |
| Enabling | Inactive |
| Running | Active |
| Disabling | Inactive |
| Stopped | Inactive |
| Deleting | Deleting |

## 1.2. Cooldown time

This topic describes the cooldown time in Auto Scaling.

Cooldown time refers to a period during which Auto Scaling cannot execute new scaling activities after a scaling activity is complete. You can use one of the following methods to configure the cooldown time:

- Configure the cooldown time in a scaling group. The default value cannot be empty. For more information, see Create a scaling group.
- Configure the cooldown time in a scaling rule. If no cooldown time is specified, the default value is used. For more information, see Create a scaling rule.

> **Note** If you configure the cooldown time in both a scaling group and scaling rule, Auto Scaling preferentially uses the cooldown time configured in the scaling rule.

### Cooldown time rules

Within the cooldown time, Auto Scaling rejects all scaling activity requests triggered by event-triggered tasks from Cloud Monitor. However, scaling activities triggered by other types of tasks such as manually triggered tasks and scheduled tasks are not limited by the cooldown time. These tasks are immediately executed.

Auto Scaling starts to calculate the cooldown time after a scaling activity is complete. If multiple ECS instances are added to or removed from a scaling group during a scaling activity, the cooldown time is calculated since the time when the last instance is added to or removed from the scaling group. For more information, see Example 1. If no ECS instance is added to or removed from a scaling group during a scaling activity, the cooldown time is not calculated.

If you disable a scaling group and then enable it again, the cooldown time becomes invalid. For more information, see Example 2.

> **Note** The cooldown time applies only to scaling activities in the same scaling group. It does not affect scaling activities in other scaling groups.

## Examples

- Example 1

  You have a scaling group named asg-uf6f3xewn3dvz4bs****. The default cooldown time is 10 minutes. A scaling rule named add3 exists in the scaling group and a cooldown time of 15 minutes is specified in the scaling rule.

  After a scaling activity triggered by the add3 scaling rule is complete, three ECS instances are added to the scaling group. The cooldown time is calculated since the time when the third instance is added to the scaling group. Within 15 minutes, scaling activity requests triggered by event-triggered tasks from Cloud Monitor are rejected.

- Example 2

  You have a scaling group named asg-m5efkz67re9x7a57****. The default cooldown time is 10 minutes. A scaling rule named remove1 exists in the scaling group and no cooldown time is specified in the scaling rule.

  After a scaling activity triggered by the remove1 scaling rule is complete at 18:00, one ECS instance is removed from the scaling group. Typically, scaling activity requests triggered by event-triggered tasks from Cloud Monitor are rejected before 18:10. If you disable the scaling group and then enable it again at 18:05, the cooldown time becomes invalid. If a scaling activity request is triggered by an event-triggered task from Cloud Monitor from 18:05 to 18:10, Auto Scaling accepts and executes the scaling activity for the scaling group.

# 1.3. Expected number of instances

This topic describes the Expected Number of Instances feature and how to specify the expected number of instances in a scaling group to execute parallel scaling activities.

## Overview

To enable the Expected Number of Instances feature, you must set the **Expected Number of Instances** parameter when you create a scaling group. You can change the specified expected number of instances in the scaling group after the scaling group is created. After you specify the expected number of instances in a scaling group, when the number of instances in the scaling group is not equal to the expected number of instances, Auto Scaling automatically adjusts the number of instances to ensure that the number of instances in the scaling group is maintained at the expected number.

> **Note** You can enable the Expected Number of Instances feature only when you create a scaling group. You cannot enable the Expected Number of Instances feature for an existing scaling group for which the Expected Number of Instances is disabled.

If you do not enable the Expected Number of Instances feature, you may encounter the following problems:

- You must manually retry failed scaling activities.

- Auto Scaling rejects requests to execute new scaling activities in the scaling group. The utilization of the scaling group is not maximized. For example, if the health check feature triggers a scaling activity in a scaling group while the scaling group has an ongoing scaling activity, Auto Scaling rejects the scaling activity triggered by the health check feature.

Before you use the Expected Number of Instances feature, we recommend that you understand the terms described in the following table.

| Term | Description |
| --- | --- |
| stable instance | An instance that is in the **In Service**, **Protected**, or **Standby** state in a scaling group. |
| parallel scaling activity | A scaling activity that can be executed in parallel with other ongoing scaling activities in a scaling group. Parallel scaling activities can be triggered by the following actions:<br>• Execute a scaling rule manually or by using a scheduled task.<br>• Manually add or remove instances.<br>• Perform checks on the expected number of instances, instance health, minimum number of instances, and maximum number of instances. |
| non-parallel scaling activity | A scaling activity that cannot be executed in parallel with other ongoing scaling activities. A scaling activity that is not a parallel activity is a non-parallel scaling activity. Non-parallel scaling activities can be triggered by the following actions:<br>• Execute a scaling rule by using event-triggered tasks.<br>• Manually execute a rebalancing task on the **Instances** tab.<br>• Automatically supplements preemptible instances.<br>**Note** When you create a scaling group, you must set **Scaling Policy** to **Cost Optimization Policy** and turn on **Enable Supplemental Preemptible Instances**. |

## Limits

- You cannot disable the Expected Number of Instances feature for a scaling group after you enable the feature.
- Scaling groups for which an expected number of instances is specified cannot have parallel scaling activities and non-parallel scaling activities at the same time.
- If you do not specify an expected number of instances when you create a scaling group, you cannot edit the Expected Number of Instances parameter for the scaling group.
- You must set Expected Number of Instances to a value that ranges from the value of Minimum Number of Instances to the value of Maximum Number of Instances in a scaling group.

## Differences between scaling groups with an expected number of instances and scaling groups without an expected number of instances

The following table describes the differences between scaling groups for which an expected number of instances is specified and scaling groups for which an expected number of instances is not specified.

| Item | With an expected number of instances | Without an expected number of instances |
| --- | --- | --- |
| Maintenance costs | Automatic maintenance: Auto Scaling automatically scales instances to maintain the expected number of instances in a scaling group. If a scaling activity fails in the scaling group, Auto Scaling automatically retries the failed scaling activity. | Manual maintenance: You need to manually adjust the number of instances in a scaling group. If a scaling activity fails in the scaling group, you need to manually retry the failed scaling activity. |
| Execution results of scaling activities | The execution results of scaling activities vary based on how the scaling rules are executed.<br>• If you manually execute a scaling rule or if you use a scheduled task to execute the scaling rule in a scaling group, Auto Scaling does not immediately trigger the scaling activity. Instead, Auto Scaling automatically changes the number of expected instances. During the check for the expected number of instances, Auto Scaling calculates the difference between the number of stable instances and the number of expected instances and then scales instances based on the difference.<br>• If you manually add instances to or remove instances from a scaling group, manually delete instances, or use event-triggered tasks to execute a scaling rule, Auto Scaling immediately triggers the scaling activity and automatically changes the number of expected instances. | If you manually execute a scaling rule, use an event-triggered task or scheduled task to execute the scaling rule, manually add instances to or remove instances from a scaling group, or manually delete instances, Auto Scaling immediately triggers the scaling activity to add or remove instances. |
| Parallel execution | Scaling groups can have parallel scaling activities. For more information, see Examples of parallel scaling activities. | A scaling group can have only one ongoing scaling activity at a time. The scaling activity may take a long time. When the scaling activity is being executed, you cannot adjust the number of instances in the scaling group. |

## Rules for changes to the expected number of instances

You can specify the expected number of instances. You can also manually trigger scaling activities to change the expected number of instances. The changes take effect based on how the scaling activities are triggered.

| Scaling activity type | Method to trigger a scaling activity | Scaling activity result | New expected number of instances | Examples |
| --- | --- | --- | --- | --- |
| | Manually execute the scaling rule. | Only the expected number of instances is changed. You must wait for Auto Scaling to perform a check on the expected number of instances to determine whether to add or remove instances. | Number of stable instances ± Number of added or removed instances | Scenario:<br>• Current expected number of instances: 3.<br>• Current number of stable instances: 2.<br>• The scaling rule specifies to create four ECS instances.<br>Result:<br>The expected number of instances is changed to 6, but the ECS instances are not immediately created. You must wait for Auto Scaling to perform a check on the expected number of instances to determine whether to add or remove instances. |
| | Execute a scaling rule by using scheduled tasks | Only the expected number of instances is changed. You must wait for Auto Scaling to perform a check on the expected number of instances to determine whether to add or remove instances. | Number of stable instances ± Number of added or removed instances | Scenario:<br>• Current expected number of instances: 3.<br>• Current number of stable instances: 2.<br>• The scaling rule specifies to create four ECS instances.<br>Result:<br>The expected number of instances is changed to 6, but the ECS instances are not immediately created. You must wait for Auto Scaling to perform a check on the expected number of instances to determine whether to add or remove instances. |

| Scaling activity type | Method to trigger a scaling activity | Scaling activity result | New expected number of instances | Examples |
| --- | --- | --- | --- | --- |
| Parallel scaling activity | Manually add instances | The instances are immediately added, and the expected number of instances is changed. | Current expected number of instances + Number of manually added instances | Scenario:<br>• Current expected number of instances: 3.<br>• Current number of stable instances: 2.<br>• Manually add four ECS instances to the scaling group.<br>Result:<br>Four ECS instances are added to the scaling group. The number of stable instances is changed to 6, and the expected number of instances is changed to 7. |
| | Manually remove ECS instances | The instances are immediately removed, and the expected number of instances is changed. | Current expected number of instances - Number of manually removed instances | Scenario:<br>• Current expected number of instances: 3.<br>• Current number of stable instances: 2.<br>• Manually remove one ECS instance from the scaling group.<br>Result:<br>One instance is removed from the scaling group. The number of stable instances is changed to 1, and the expected number of instances is changed to 2. |

| Scaling activity type | Method to trigger a scaling activity | Scaling activity result | New expected number of instances | Examples |
| --- | --- | --- | --- | --- |
| | Perform checks on the minimum and maximum numbers of instances | - | To be manually specified | Scenario:<br>• Current maximum number of instances: 5.<br>• Current minimum number of instances: 0.<br>• Current expected number of instances: 3.<br>• Current number of stable instances: 2.<br>• Manually change the maximum number of instances to one.<br>Result:<br>The maximum number failed to be changed. You must change the expected number of instances at the same time. |
| | Perform health checks on instances | The instances are immediately removed from the scaling group. | The expected number of instances remains unchanged | Scenario:<br>• Current expected number of instances: 3.<br>• Current number of stable instances: 2.<br>• One ECS instance is considered unhealthy.<br>Result:<br>The expected number of instances remains unchanged. The unhealthy instance is removed from the scaling group, and the number of stable instances is changed to 1. Auto Scaling detects the difference between the expected number of instances and the number of stable instances, and then automatically performs a check on the expected number of instances to trigger the scaling activity to create two ECS instances. |
| | | | | |

| Scaling activity type | Method to trigger a scaling activity | Scaling activity result | New expected number of instances | Examples |
| --- | --- | --- | --- | --- |
| | Perform a check on the expected number of instances | The instances are immediately added to or removed from the scaling group. | The expected number of instances remains unchanged | Scenario:<br>• Current expected number of instances: 3.<br>• Current number of stable instances: 2.<br>Result:<br>The expected number of instances remains unchanged. Auto Scaling detects the difference between the expected number of instances and the number of stable instances, and then automatically performs a check on the expected number of instances to create one ECS instance. |
| Non-parallel scaling activity | Execute a scaling rule by using event-triggered tasks | The instances are immediately added to or removed from the scaling group, and the expected number of instances is changed. | Number of stable instances ± Number of added or removed instances | Scenario:<br>• Current expected number of instances: 3.<br>• Current number of stable instances: 2.<br>• The add3 scaling rule specifies to create four ECS instances.<br>Result:<br>A scaling activity is triggered to create four instances, and the expected number of instances is changed to 6. |

## Examples of parallel scaling activities

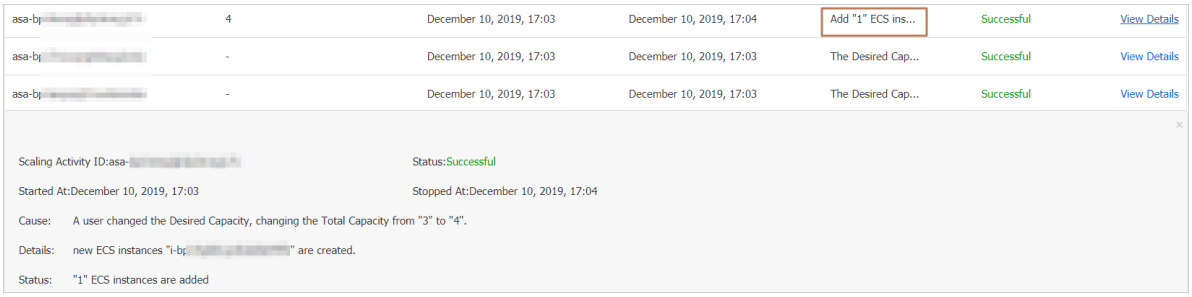
After you specify the expected number of instances, Auto Scaling can execute parallel scaling activities at the same time.

- Example 1: Manually execute scaling rules in a consecutive manner

  Scenario:

  - Expected number of instances: 3.
  - Number of stable instances: 3.
  - The add3 scaling rule specifies to create three ECS instances.
  - The add1 scaling rule specifies to create one ECS instance.
  - Manually execute the add3 scaling rule and then immediately execute the add1 scaling rule.

Result: After the add3 scaling rule is executed, the expected number of instances is changed from 3 to 6. The add1 scaling rule is immediately executed, and the expected number of instances is changed from 6 to 4. After the parallel scaling activity is executed, one instance is created and the number of stable instances in the scaling group is changed to 4.

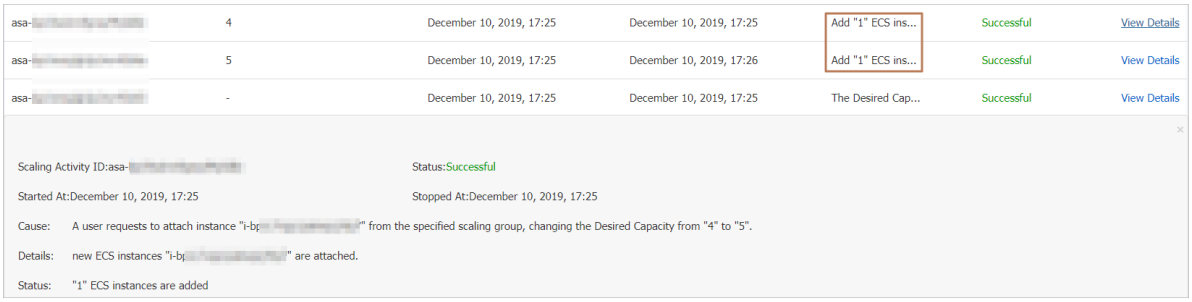



- Example 2: Manually execute a scaling rule and manually add instances

  Scenario:

  - Expected number of instances: 3.
  - Number of stable instances: 3.
  - The add1 scaling rule specifies to create one ECS instance.
  - Manually execute the add1 scaling rule and then immediately add an existing ECS instance.

  Result: After the add1 scaling rule is executed, the expected number of instances is changed from 3 to 4. One existing ECS instance is immediately added to the scaling group, and the expected number of instances is changed from 4 to 5. After the parallel scaling activity is complete, one ECS instance is created, one existing ECS instance is added, and the number of stable instances in the scaling group is changed to 5.



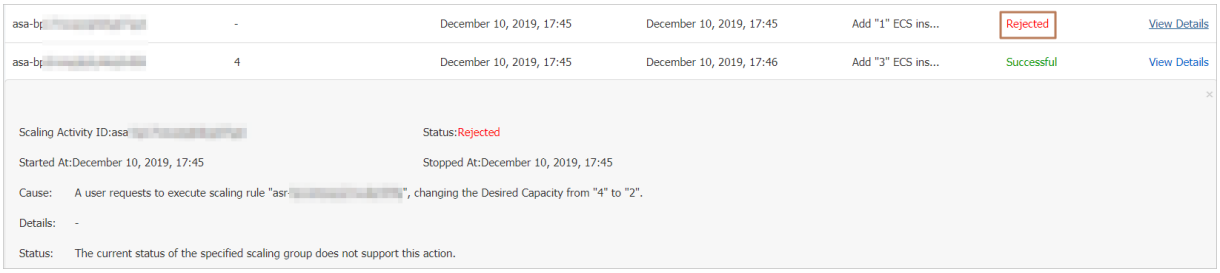

## Examples of non-parallel scaling activities

After you specify the expected number of instances, Auto Scaling cannot execute parallel and non-parallel scaling activities at the same time. Example:

Scenario:

- Expected number of instances: 1.
- Number of stable instances: 1.
- The scaling rule associated with an event-triggered task specifies to create three ECS instances.
- The add1 scaling rule specifies to create one ECS instance.
- After the event-triggered task is triggered, execute the add1 scaling rule immediately.

Result: After a scaling activity is triggered by the event-triggered task, the expected number of instances is changed from 1 to 4. The scaling activity triggered by the event-triggered event is a non-parallel scaling activity and cannot be immediately executed. The request to execute the add1 scaling rule is rejected, and the expected number of instances remains unchanged. After the non-parallel scaling activity is complete, three ECS instances are created, and the number of stable instances in the scaling group is changed to 4.




# 1.4. Create a scaling group

A scaling group is a collection of Elastic Compute Service (ECS) instances or elastic container instances that serve the same purpose. When you create a scaling group, you can specify the minimum and maximum numbers of instances, templates that are used for scale-out, and policies that are used for scale-in. You can use a scaling group to manage a group of instances based on your business requirements.

## Prerequisites

- Auto Scaling is activated. This prerequisite must be met if this is the first time you use Auto Scaling. For more information, see 管理弹性伸缩服务关联角色.
- A launch template is created if you want to use a launch template to create ECS instances. For more information, see Create a launch template.
- The following requirements are met if you want to associate a Classic Load Balancer (CLB) instance with a scaling group:
  - You have one or more CLB instances in the **Active** state. For more information, see Create a CLB instance.
  - The CLB instance and the scaling group reside in the same region.
  - The CLB instance is associated with at least one listener. For more information, see Listener overview.
  - Health check is enabled for the CLB instance. For more information, see Configure health checks.
  - The network types of the CLB instance and the scaling group are subject to the following limits:
    - The CLB instance and the scaling group must be in the same virtual private cloud (VPC) if their network type is VPC.
    - If the network type of the CLB instance is classic network, the network type of the scaling group is VPC, and the backend server group of the CLB instance contains ECS instances of the VPC type, the ECS instances and the scaling group must be in the same VPC.

      > **Note** In other cases, no limit is applied to the network type when you associate a scaling group with a CLB instance.

- The following requirements are met if you want to associate an Application Load Balancer (ALB) server group with a scaling group:

  - The network type of the scaling group is virtual private cloud (VPC). The scaling group and the ALB server group that you want to associate are in the same VPC.
  - The ALB server group is in the Available state.
- The following requirements are met if you want to associate an ApsaraDB RDS instance with a scaling group:
  - You have one or more ApsaraDB RDS instances that are in the **Running** state. For more information, see What is ApsaraDB RDS?
  - The scaling group and the ApsaraDB RDS instance are in the same region.

## Context

You can create only a limited number of scaling groups in a region. To view the quota or request a quota increase, go to the Quota Center.

## Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. In the upper-left corner of the Scaling Groups page, click **Create**.
5. Configure the following parameters for the scaling group and click **OK**.

| Parameter | Description |
| --- | --- |
| **Scaling Group Name** | The name of a scaling group must be 2 to 64 characters in length and can contain letters, digits, periods (.), underscores (_), and hyphens (-). The name must start with a letter or a digit. |
| **Type** | The type of instances that provide computing power in the scaling group. Auto Scaling scales instances based on the Type parameter. Valid values:<br>○ ECS<br>○ ECI |

| Parameter | Description |
| --- | --- |
| Instance Configuration Source | Auto Scaling creates instances based on the Instance Configuration Source parameter. Valid values:<br>◦ **Launch Template**: A launch template contains information such as a key pair, a RAM role, an instance type, and the network configurations. A launch template does not contain passwords. Launch Template is available only if you set **Type** to **ECS**.<br>If you use Launch Template, you must configure the Select Launch Template parameter and the Select Template Version parameter. To make the template fit various needs, you can use the **Extend Launch Template Configurations** parameter to select multiple instance types. For more information about how to specify weights for instance types, see Use performance metrics to measure Auto Scaling.<br>◦ **Select Existing Instance**: Select an existing instance. Then, Auto Scaling extracts the basic configurations of the instance to create a default scaling configuration.<br>If you set **Type** to **ECS**, the basic configurations that are extracted from the selected ECS instance include the instance type, network type, security group, and base image. The instance logon password and tags are not extracted. The base image is the image used by the existing instance. The base image does not include instance data such as application data. If you want to include all system configurations of the instance and instance data in the scaling configuration, create a custom image for the instance and use the custom image to update the image of the scaling configuration. For more information, see Replace the image in a single scaling configuration.<br>◦ **Create from Scratch**: Do not specify a template that is used to automatically create instances. After you create a scaling group, create a scaling configuration or specify a launch template.<br>**Note** When you create a scaling group that contains ECS instances created in the ECS console, the instance configurations and network type of the instances are automatically populated. We recommend that you use the default settings. |
| Tag | You can add tags to help you find and manage scaling groups. For more information, see Overview.<br>**Note** The tags that you add apply only to the scaling group. If you want to add tags to an instance in the scaling group, specify the tags in the scaling configuration or in the launch template. |
| | If you need to remove more than one instance from a scaling group, Auto Scaling removes instances based on the Instance Removing Policy parameter. If multiple instances meet the conditions of the policy, a random instance is removed. The Instance Removing Policy parameter is available only if you set **Type** to **ECS**. |

| Parameter | Description |
| --- | --- |
| Instance Removing Policy | The Instance Removing Policy parameter contains **Filter First** and **Then Remove from Results**. Specify different values for the two parameters.<br>The following part describes the values.<br>**Note** If you set **Type** to **ECI**, instances that are created based on the earliest scaling configuration are removed by default. Then, Auto Scaling removes the instances that are created at the earliest point in time from the results obtained based on the Filter First parameter.<br>○ **Earliest Instance Created Using Scaling Configuration**: Auto Scaling removes instances that are created based on the earliest scaling configuration or launch template. No scaling configuration or launch template is associated with manually added instances. Therefore, manually added instances are not removed first. If all instances with which the earliest scaling configuration and launch template are associated are removed, and more instances need to be removed from the scaling group, manually added instances are removed at random.<br>**Note** Scaling Configuration in **Earliest Instance Created Using Scaling Configuration** specifies the instance configuration source that contains the scaling configuration and launch template.<br>The version of a launch template does not indicate the sequence in which the template is added. For example, you use the lt-foress V2 template to create a scaling group, and then you replace the template with the lt-foress V1 template when you modify the scaling group. In this scenario, the scaling group considers the lt-foress V2 launch template as the earliest template.<br>○ **Earliest Created Instance**: Auto Scaling removes the instances that are created at the earliest point in time.<br>○ **Most Recent Created Instance**: Auto Scaling removes the instances that are created at the latest point in time.<br>○ **No Policy**: This value is available only for the **Then Remove from Results** parameter. If you select No Policy, Auto Scaling does not remove instances from the results obtained based on the Filter First parameter.<br>If Auto Scaling removes instances based on the **Earliest Instance Created Using Scaling Configuration** value, you can select one of the following values for the Then Remove from Results parameter:<br>○ **No Policy**: Auto Scaling does not remove instances from the results obtained based on the Filter First parameter.<br>○ **Earliest Created Instance**: Auto Scaling removes the instances that are created at the earliest point in time from the results obtained based on the Filter First parameter.<br>○ **Most Recent Created Instance**: Auto Scaling removes the instances that are created at the latest point in time from the results obtained based on the Filter First parameter.<br>**Note** The manner in which instances are removed from scaling groups is also affected by the value of the **Scaling Policy** parameter. For more information about how to remove instances from scaling groups, see Configure a combination policy for removing instances. |

| Parameter | Description |
| --- | --- |
| **Suspended Processes** | You can suspend processes before you perform specific operations. For example, you can suspend the health check process before you stop an instance. This way, the instance is not removed from the scaling group if the health check fails. You can suspend the following processes for a scaling group:<br>○ **Scale-out**: If you suspend this process, Auto Scaling rejects all scale-out requests.<br>○ **Scale-in**: If you suspend this process, Auto Scaling rejects all scale-in requests.<br>○ **Health Check**: If you suspend this process, Auto Scaling suspends the health check process and does not remove unhealthy instances.<br>○ **Scheduled Task**: When the execution time of a scheduled task arrives, the scaling rules that are associated with the task are not triggered.<br>○ **Event-triggered Task**: When an event-triggered task enters the alert state, the scaling rules that are associated with the task are not triggered.<br>For more information, see Suspend a scaling process and Resume a scaling process. |
| **Enable Deletion Protection for Scaling Group** | After you enable this feature, you cannot delete the scaling group in the Auto Scaling console or by calling API operations. |
| **Health Check for Instances** | After you enable this feature, Auto Scaling checks the status of instances on a regular basis. If an instance is not in the Running state, the instance is considered unhealthy and is removed from the scaling group. For more information, see ECS instance lifecycle. |
| **Minimum Number of Instances** | If the number of instances in a scaling group is less than the minimum number of instances allowed, Auto Scaling automatically creates instances until the number of instances in the scaling group reaches the minimum number. |
| **Maximum Number of Instances** | If the number of instances in a scaling group is greater than the maximum number of instances allowed, Auto Scaling automatically creates instances until the number of instances in the scaling group does not exceed the maximum number. |
| **Expected Number of Instances** | If you specify an expected number of instances, Auto Scaling automatically maintains the specified number of instances. For more information, see Expected number of instances.<br>**Note** The Expected Number of Instances parameter is available only when you create a scaling group. The Expected Number of Instances parameter is unavailable when you edit an existing scaling group. |

| Parameter | Description |
| --- | --- |
| **Default Cooldown Time (Seconds)** | Specifies the default cooldown time of a scaling group. Unit: seconds. During the cooldown time, Auto Scaling rejects all scaling requests of event-triggered tasks. Scaling activities that are triggered by other types of tasks such as scheduled tasks and manually executed tasks are not subject to the cooldown time and can be immediately executed. |
| **Network Type** | The **Expansion and Contraction Strategy**, **Instance Reclaim Mode**, and **Associated ALB Server Group** parameters are available only if you set Network Type to VPC.<br><br>**Note** When you create a scaling group that contains ECS instances created in the ECS console, the instance configurations and network type of the instances are automatically populated. We recommend that you use the default settings.<br><br>A scaling group and instances in the scaling group must belong to the same network type. For example, if a scaling group resides in a VPC, the instances in the scaling group must also reside in the VPC. If a scaling group resides in the classic network, the instances in the scaling group must also reside in the classic network.<br><br>**Note** After you create a scaling group, you cannot change the network type of the scaling group. |

| Parameter | Description |
| --- | --- |
| Expansion and Contraction Strategy | The Expansion and Contraction Strategy parameter is available only if you set **Type** to **ECS** and **Network Type** to **VPC**. Valid values:<br>○ **Priority Policy**: Instances are preferentially created in the zone where the vSwitch that has the highest priority resides. Auto Scaling preferentially scales instances in the zone where the vSwitch that has the highest priority resides. If the scaling fails, Auto Scaling attempts to scale instances in the zone where the vSwitch that has the next highest priority resides.<br>**Note** If you set **Type** to **ECI**, the default value of Expansion and Contraction Strategy is Priority Policy.<br>○ **Balanced Distribution Policy**: This policy is valid only if the scaling group is associated with multiple vSwitches that are distributed across more than two zones. Auto Scaling evenly distributes instances across the zones where the vSwitches reside based on this policy. If instances are not evenly distributed across multiple zones due to insufficient resources, you can use Balanced Distribution Policy to re-distribute instances across zones. For more information, see Rebalance the distribution of ECS instances.<br>○ **Cost Optimization Policy**: This policy is valid only if you specify multiple instance types in the scaling configuration. When a scale-out activity is triggered, Auto Scaling preferentially creates ECS instances that have the lowest vCPU price. When a scale-in activity is triggered, Auto Scaling preferentially removes ECS instances that have the highest vCPU price. If you select Preemptible Instance as the billing method in the scaling configuration, Auto Scaling preferentially creates preemptible instances. If preemptible instances cannot be created due to insufficient resources, Auto Scaling creates pay-as-you-go instances.<br>If you select **Cost Optimization Policy**, you must configure the following parameters:<br>○ **Minimum Pay-as-you-go Instances**: the minimum number of pay-as-you-go ECS instances in the scaling group. Default value: 0. If the number of pay-as-you-go ECS instances in the scaling group is less than the value of Minimum Pay-as-you-go Instances, Auto Scaling preferentially creates pay-as-you-go instances.<br>○ **Percentage of Pay-as-you-go Instances**: the percentage of pay-as-you-go ECS instances among all automatically created instances. Default value: 70%. When you calculate this percentage, the pay-as-you-go ECS instances do not include the minimum number of pay-as-you-go ECS instances that you specified for the scaling group.<br>○ **Lowest Cost Instance Types**: the number of instance types that have the lowest price. Default value: 1. This parameter is valid only if multiple instance types are specified in the scaling configuration. Auto Scaling evenly creates preemptible ECS instances of the instance types that have the lowest prices.<br>○ **Enable Supplemental Preemptible Instances**: After you enable the Supplemental Preemptible Instances feature, Auto Scaling automatically creates preemptible instances five minutes before the existing instances are reclaimed. |

| Parameter | Description |
| --- | --- |
| Instance Reclaim Mode | Instance Reclaim Mode is available only if you set **Type** to **ECS** and **Network Type** to **VPC**. Valid values:<br>○ **Release Mode**: Instances that are removed from the scaling group are released. Resources of these instances are not retained. When a scale-out activity is triggered, Auto Scaling creates instances and adds the instances to the scaling group.<br>**Note** If you set **Type** to **ECI**, instances that are removed from the scaling group are released by default.<br>○ **Shutdown and Reclaim Mode**: ECS instances that are removed from the scaling group are stopped and enter the Economical Mode state. Some resources of the ECS instances are retained, and you are charged for these resources. When a scale-out activity is triggered, Auto Scaling preferentially adds the stopped ECS instances. After all stopped ECS instances are added, Auto Scaling determines whether to continue to create new ECS instances based on your scale-out requirements. This mode improves scaling efficiency.<br>**Notice**<br>■ Your data stored on instances may be lost when the instances are reclaimed. To prevent data loss, do not store application data or logs on instances.<br>■ Stopped instances may be released due to the following reasons:<br>■ If the total number of instances in a scaling group exceeds the maximum number of instances allowed for the scaling group after you reduce the maximum number, Auto Scaling preferentially releases the ECS instances that are in the Stopped state.<br>■ If stopped instances fail to be added to a scaling group due to insufficient resources or overdue payments, the instances are released.<br>■ You can enable Economical Mode only for pay-as-you-go instances in the Stopped state. For more information, see the "Prerequisites", "Application resources", and "Trigger effects" sections in Economical mode. |
| VPC | Select an existing VPC.<br>**Note** When you create a scaling group that contains ECS instances created in the ECS console, the instance configurations and network type of the instances are automatically populated. We recommend that you use the default settings. |

| Parameter | Description |
| --- | --- |
| **Select vSwitch** | After you select a VPC, you must select a vSwitch. Each vSwitch resides in a single zone. To deploy instances across multiple zones, you must specify multiple vSwitches in different zones. We recommend that you select multiple zones to increase the success rate of scale-out.<br><br>**Note** When you create a scaling group that contains ECS instances created in the ECS console, the instance configurations and network type of the instances are automatically populated. We recommend that you use the default settings. |
| **Add Existing Instance** | The Add Existing Instance parameter is available only if you set **Type** to **ECS** and set **Instance Configuration Source** to **Launch Template** or **Select Existing Instance**.<br><br>If you specify an expected number of instances and then add existing instances, the expected number of instances automatically increases. For example, when you create a scaling group, you set Expected Number of Instances to 1 and add two existing instances. After the scaling group is created, two existing instances are added to the scaling group, and the expected number of instances is three.<br><br>You can select **Enable the scaling group to manage the instance lifecycle**.<br><br>◦ If a scaling group manages the lifecycle of instances, the instances are automatically released when the instances are manually removed from the scaling group or are considered unhealthy.<br>◦ If a scaling group does not manage the lifecycle of instances, the instances are not automatically released when the instances are removed from the scaling group.<br><br>**Note** You can add subscription instances to a scaling group. However, the lifecycle of the subscription instances cannot be managed by the scaling group. |

| Parameter | Description |
| --- | --- |
| **Associate CLB (Formerly Known as SLB) Instance** | After you associate a CLB instance with a scaling group, the instances that you add to the scaling group are automatically added as the backend servers of the CLB instance. Then, the CLB instance forwards requests to the instances.<br>You can specify a server group to which you want to add instances. Valid values:<br>◦ Default server group: the group of instances that are used to receive requests. If you do not specify a vServer group or a primary/secondary server group for a listener, requests are forwarded to the instances in the default server group.<br>◦ vServer group: If you want to forward requests to different backend servers or configure domain name- or URL-based routing methods, you can use vServer groups.<br>If you specify the default server group and multiple vServer groups at the same time, the instances are added to these server groups.<br>**Note** You can associate only a limited number of CLB instances and vServer groups with a scaling group. To view the quota or request a quota increase, go to the Quota Center. |
| **Associated ALB Server Group** | The Associated ALB Server Group parameter is available only if you set **Network Type** to **VPC**. After you associate an ALB server group with a scaling group, the instances that you add to the scaling group are automatically added as the backend servers of the ALB server group to process access requests that are forwarded by the ALB instance. You must specify the port number and weight for each backend server. By default, the weight is 50. If you increase the weight of a server group, the number of access requests that are forwarded to the server group increases. If you set the weight to 0, no access requests are forwarded to the server group.<br>If you associate multiple ALB server groups with the same scaling group, all instances that you add to the scaling group are added to these server groups.<br>**Note** You can associate only a limited number of ALB server groups with a scaling group. To view the quota or request a quota increase, go to the Quota Center. |
| **Associate RDS Instance** | The Associate RDS Instance parameter is available only if you set **Type** to **ECS**. After you associate an ApsaraDB RDS instance with a scaling group, the internal IP addresses of ECS instances that you add to the scaling group are automatically added to the whitelist of the ApsaraDB RDS instance to allow internal communication.<br>**Note** You can associate only a limited number of ApsaraDB RDS instances with a scaling group. To view the quota or request a quota increase, go to the Quota Center. |

| Parameter | Description |
| --- | --- |
| **Set Notification Receiving** | When a scaling activity succeeds, fails, or is rejected, Auto Scaling sends notifications to you by using SMS messages, internal messages, or emails. For more information, see Set notification receiving. |

6. In the **Create Scaling Group** dialog box, click **OK**.
   You can view the scaling group on the Scaling Groups page. The scaling group is in the **Stopped** state.

## What's next

1. If you set **Instance Configuration Source** to **Create from Scratch** when you create the scaling group, you must create and enable a scaling configuration or specify a launch template. The steps that you must perform to create a scaling configuration vary based on the value of the Type parameter.
   - For more information about how to create a scaling configuration if you set Type to ECS, see Create a scaling configuration (Elastic Compute Service).
   - For more information about how to create a scaling configuration if you set Type to ECI, see Create a scaling configuration (Elastic Container Instance).
2. Enable the scaling group. For more information, see Enable a scaling group.

   You can only scale scaling groups that are in the Enabled state.

## Related information

### References

- CreateScalingGroup

# 1.5. Enable a scaling group

This topic describes how to enable a scaling group. Auto Scaling can scale ECS instances for a scaling group only when the scaling group is in the Enabled state.

## Prerequisites

- The scaling group is in the **Disabled** state.
- One of the following methods is used to specify an instance configuration source for the scaling group:
  - Specify a launch template.
  - Create and enable a scaling configuration.

## Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find the scaling group and choose > **Enable** in the **Actions** column.

   The scaling group enters the **Enabled** state.

## Related information

- EnableScalingGroup

# 1.6. Modify a scaling group

This topic describes how to modify a scaling group. You can modify the parameters of a scaling group based on your business requirements.

## Context

The following limits apply when you modify a scaling group:

- You cannot change the network type of a scaling group.
- If the network type of a scaling group is **VPC**, you can change the vSwitch, but cannot change the reclaim mode.

## Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find the scaling group that you want to modify and click **Edit** in the **Actions** column.
5. Modify the parameters of the scaling group.

   For more information about the parameters of a scaling group, see Create a scaling group. When you modify the parameters of a scaling group, the following limits apply:

   - If the Maximum Number of Instances or Minimum Number of Instances parameters are modified, the number of existing ECS instances in the scaling group may fall outside the specified range. In this case, Auto Scaling automatically adds or removes ECS instances to ensure that the number of ECS instances in the scaling group is within the specified range.
   - If the Associate RDS Instance parameter is modified, and ECS instances exist in the scaling group, select **When you associate RDS instances with or disassociate RDS instances from the scaling group, existing ECS instances in the scaling group are added to or removed from the whitelists of the RDS instances.** based on your business requirements.
     - If you select this option, the existing ECS instances in the scaling group are automatically added to the whitelist of the associated RDS instances, or removed from the whitelist of the disassociated RDS instances.
     - If you do not select this option, the whitelist of the RDS instances remains unchanged.
   - If you want to modify the Associate CLB (Formerly Known as SLB) Instance parameter, and the ECS instances exist in the scaling group, select **When you associate *SLB* instances with or disassociate *SLB* instances from the scaling group, existing ECS instances in the scaling group are added to or removed from the server groups of the *SLB* instances.**
     - If you select this option, the existing ECS instances in the scaling group are automatically added to the associated SLB server groups or removed from the disassociated SLB server groups.
     - If you do not select this option, the SLB server groups remain unchanged.
6. In the message that appears, click **OK**.

## Related information

- ModifyScalingGroup

# 1.7. Disable a scaling group

This topic describes how to disable a scaling group that is no longer needed.

## Prerequisites

The scaling group is in the **Enabled** state.

## Context

If a scaling activity is being executed when you disable the scaling group, the scaling activity continues until it is complete. Scaling requests that are made after the scaling group is disabled are rejected.

## Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find the scaling group and choose > **Disable** in the **Actions** column.
5. In the **Disable Scaling Group** message, click **OK**.
   The scaling group enters the **Disabled** state.

## Related information

- DisableScalingGroup

# 1.8. Resume a scaling process

You can resume a suspended scaling process. This allows the scaling group to continue the specified process based on its functional logic. This topic describes how to resume a scaling process.

## Context

After a scaling process is resumed, the process takes into account the changes made during the suspension. For example, the expected number of instances in a scaling group changes when the scale-out process is suspended for the scaling group. After the scale-out process is resumed, Auto Scaling performs checks on the expected number of instances to add instances, and the expected number of instances changes based on the check.

| Scaling process | Effect of a resumed process |
| --- | --- |
| Scale-out | The scaling group resumes scale-out processes, such as processes to manually add instances and to check the expected and minimum numbers of instances.<br>If the Expected Number of Instances feature is enabled and the expected number of instances increases during a suspension, Auto Scaling performs checks on the expected number of instances to add instances. |

| Scaling process | Effect of a resumed process |
| --- | --- |
| Scale-in | The scaling group resumes scale-in processes, such as processes to manually remove instances and to check the expected and maximum numbers of instances.<br>If the Expected Number of Instances feature is enabled and the expected number of instances decreases during a suspension, Auto Scaling performs checks on the expected number of instances to remove instances. |
| Health checks | The scaling group resumes health check processes and automatically removes unhealthy instances. |
| Scheduled tasks | If a scheduled task is not due or is being retried, the scaling rule that is associated with the task is triggered. |
| Event-triggered tasks | If an event-triggered task enters the alert state, the scaling rule that is associated with the task is triggered. |

## Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. On the Scaling Groups page, find the scaling group that you want to manage and click **Edit** in the **Actions** column.
5. In the **Edit Scaling Group** dialog box, select one or more processes that you want to resume from the **Suspend Process** drop-down list.
6. Click **OK**.

   If the scaling strength of your scaling group is not strong, a message that prompts you to confirm the resume operation appears. You can confirm the operation or go back to the Edit Scaling Group dialog box to change the parameter settings based on your business requirements.

   > **Note** If the scaling strength of your scaling group is not strong, the scaling activities in the scaling group may fail. To prevent scaling failures, we recommend that you go back to the Edit Scaling Group dialog box to increase the scaling strength of your scaling group.

   - If you do not want to change the parameter settings, click **Continue**.
   - If you want to change the parameter settings, click **Back to Modify**. After you change the parameter settings, click **OK**.
7. In the message that shows the operation is successful, click **OK**.

## Result

On the Scaling Groups page, find the scaling group that you edited and click **Details** in the **Actions** column. On the **Basic Information** tab, you can view the processes that are resumed.

## Related information

- ResumeProcesses

# 1.9. Set the deletion protection status for a scaling group

After deletion protection is enabled for a scaling group, the scaling group cannot be deleted by using the Auto Scaling console or by calling API operations. This topic describes how to set the deletion protection status for a scaling group.

## Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find the scaling group and choose > **Set Deletion Protection** in the **Actions** column.
5. In the dialog box that appears, set the deletion protection status for the scaling group.
   - To enable deletion protection, turn on **Deletion Protection**.
   - To disable deletion protection, turn off **Deletion Protection**.
6. Click **OK**.

## Related information

- SetGroupDeletionProtection
- ModifyScalingGroup

# 1.10. Delete a scaling group

This topic describes how to delete a scaling group that is no longer needed.

## Prerequisites

Deletion protection is disabled for the scaling group. If deletion protection is enabled for the scaling group, disable deletion protection before you delete the scaling group. For more information, see Set the deletion protection status for a scaling group.

## Context

If you delete a scaling group, its scaling configurations and scaling rules are also deleted. If ECS instances in the Running state exist in the scaling group, Auto Scaling stops the ECS instances first. Then, Auto Scaling removes all manually added ECS instances and releases all automatically added ECS instances.

## Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find the scaling group and click **Delete** in the **Actions** column.
5. In the message that appears, click **OK**.

## Result

The scaling group enters the **Deleting** state. After the scaling group is deleted, it is no longer displayed on the Scaling Groups page.

## Related information

- DeleteScalingGroup

# 1.11. FAQ about scaling groups

This topic provides answers to commonly asked questions about scaling groups.

- FAQ about scaling group configurations
  - Must Auto Scaling be used with Alibaba Cloud services such as SLB, Cloud Monitor, and RDS?
  - How many ECS instances can I add to a scaling group?
  - Can I add existing ECS instances to a scaling group?
  - Can I add existing subscription ECS instances to a scaling group?
  - Can I add an ECS instance to multiple scaling groups?
  - Can I control how Auto Scaling removes ECS instances from a scaling group?
  - After I disable a scaling group, does Auto Scaling release ECS instances that are automatically added to the scaling group?
- FAQ about scaling configurations
  - Can I specify multiple ECS instance types for a scaling configuration?
  - Can I configure 8-core or 16-core ECS instances for a scaling group?
  - How do I set the capacity of data disks for ECS instances that are automatically created by Auto Scaling?
  - If Auto Scaling uses images from Alibaba Cloud Marketplace to create ECS instances, how do I make sure that the ECS instances can be created?
  - Can I buy multiple images from Alibaba Cloud Marketplace at a time?
  - If Alibaba Cloud Marketplace no longer provides an image that is used to create ECS instances in Auto Scaling, how do I make sure that ECS instances can be created with the image?
  - Why is an error of unsupported Alibaba Cloud Marketplace images returned when Auto Scaling creates an ECS instance?
  - Can images of different regions use the same product code?
  - I purchased 100 images with the same product code. Can I use them in any region?
- FAQ about scaling activities
  - What information do I need to provide when I submit a ticket about Auto Scaling?
  - Why is a resource error returned when Auto Scaling creates an ECS instance?
  - How do I avoid scale-out failures due to insufficient inventory of an instance type?
  - Why is an ECS instance for which I enabled release protection automatically released from a scaling group?
  - How do I prevent Auto Scaling from automatically releasing an ECS instance that I manually added to a scaling group?

- When Auto Scaling automatically adds or removes ECS instances to or from a scaling group, does Auto Scaling automatically add or remove the IP addresses of the ECS instances to or from the whitelists of the RDS instances or ApsaraDB for Memcache instances associated with the scaling group?
- How do I prevent Auto Scaling from automatically removing an ECS instance that I manually added to a scaling group?
- After ECS instances are removed from scaling groups and released, is the data of the ECS instances retained?
- How do I delete ECS instances that are automatically created by Auto Scaling in a scaling group?
- When Auto Scaling automatically creates multiple ECS instances, will Auto Scaling attempt to recreate instances that fail to be created?

- FAQ about ECS instances in scaling groups
  - For an ECS instance that is automatically created by Auto Scaling, how do I obtain its password and log on to it?
  - Why am I unable to use the password set in the custom image to log on to an ECS instance that is automatically created by Auto Scaling?
  - How do I synchronize data to ECS instances in a scaling group?
  - Why is IP Address 127.0.0.1 deleted from the /etc/hosts file of ECS instances that are automatically created by Auto Scaling?
- FAQ about SLB instances associated with scaling groups
  - What does an SLB instance do after it is associated with a scaling group?
  - How do SLB instances work with scaling groups?
  - After an ECS instance is added to a scaling group, does Auto Scaling automatically add the ECS instance to an SLB instance as a backend server?
  - Can Auto Scaling add an automatically created ECS instance to multiple SLB instances as a backend server?
  - After an ECS instance is added to an SLB instance as a backend server, can I change the weight of the ECS instance?
  - After I associate a public SLB instance with a scaling group, do I need to set the public bandwidth for ECS instances when I create a scaling configuration for the scaling group?
  - Why is a health check error about the SLB instance returned when I create a scaling group?
  - How does an SLB instance determine whether a newly created ECS instance can process requests?
  - Why does the timeout period of a Layer-7 HTTP listener of an SLB instance exceed 60 seconds?
- FAQ about RDS instances associated with scaling groups
  - What does an RDS instance do after it is associated with a scaling group?
  - How do RDS instances work with scaling groups?

## Must Auto Scaling be used with Alibaba Cloud services such as SLB, Cloud Monitor, and RDS?

No, Auto Scaling is an open and flexible service and can be used alone without the need to combine Server Load Balancer (SLB), ApsaraDB RDS (RDS), or Cloud Monitor. We recommend that you use Auto Scaling with these Alibaba Cloud services to improve performance. You can deploy Auto Scaling with SLB or RDS and use Cloud Monitor to trigger scaling activities in Auto Scaling. For information about these Alibaba Cloud services, see the following topics:

- SLB: What is CLB?
- RDS: What is ApsaraDB RDS?
- Cloud Monitor: Event-triggered task overview

## How many ECS instances can I add to a scaling group?

For information about the limits related to Auto Scaling, see 使用限制.

## Can I add existing ECS instances to a scaling group?

Yes, you can add existing ECS instances to a scaling group. Make sure that the ECS instances meet the following conditions:

- The ECS instances and the scaling group must be in the same region. For more information, see Regions and zones.
- The ECS instances are in the **Running** state. For more information, see Instance lifecycle.
- The ECS instances are not added to other scaling groups.

## Can I add existing subscription ECS instances to a scaling group?

Yes, you can add existing subscription or pay-as-you-go instances to a scaling group. You can add existing ECS instances to a scaling group, including subscription and pay-as-you-go instances. Additionally, Auto Scaling can automatically create pay-as-you-go and preemptible instances.

## Can I add an ECS instance to multiple scaling groups?

No, you can add an ECS instance only to a single scaling group.

## Can I control how Auto Scaling removes ECS instances from a scaling group?

Yes, you can control how Auto Scaling removes ECS instances from a scaling group. You can set a removal policy for a scaling group to preferentially remove the ECS instances created based on the Earliest Instance Created Using Scaling Configuration, Earliest Created Instance, or Most Recent Created Instance option. For more information, see Create a scaling group.

## After I disable a scaling group, does Auto Scaling release ECS instances that are automatically added to the scaling group?

No, Auto Scaling does not automatically release ECS instances after you disable a scaling group by using the Auto Scaling console or calling an API operation. For more information, see Disable a scaling group.

## Can I specify multiple ECS instance types for a scaling configuration?

Yes, you can specify multiple ECS instance types for a scaling configuration. This increases the success rate of creating ECS instances. However, make sure that the number of ECS instance types specified for a scaling configuration does not exceed the upper limit. For more information, see 使用限制.

## Can I configure 8-core or 16-core ECS instances for a scaling group?

Yes, you can configure 8-core or 16-core ECS instances for a scaling group. If the current available instance types cannot meet your requirements,submit a ticket to request the use of more ECS instance types.

## How do I set the capacity of data disks for ECS instances that are automatically created by Auto Scaling?

You can specify the capacity of data disks in the Storage section when you create a scaling configuration. For more information, see Create a scaling configuration (Elastic Compute Service).

## If Auto Scaling uses images from Alibaba Cloud Marketplace to create ECS instances, how do I make sure that the ECS instances can be created?

If you want Auto Scaling to create N instances that use the same Alibaba Cloud Marketplace image, you must purchase the image N times from Alibaba Cloud Marketplace in advance.

## Can I buy multiple images from Alibaba Cloud Marketplace at a time?

No, you can buy only a single image from Alibaba Cloud Marketplace at a time.

## If Alibaba Cloud Marketplace no longer provides an image that is used to create ECS instances in Auto Scaling, how do I make sure that ECS instances can be created with the image?

We recommend that you use other images available in Alibaba Cloud Marketplace.

## Why is an error of unsupported Alibaba Cloud Marketplace images returned when Auto Scaling creates an ECS instance?

If you specify an Alibaba Cloud Marketplace image that you have not purchased in the scaling configuration, the following error is returned when an ECS instance is created based on the scaling configuration:

```
 Fail to create Instance into scaling group("The specified image is from the image market.
You have not bought it or your quota has been exceeded.").
```

Auto Scaling cannot create ECS instances by using Alibaba Cloud Marketplace images that you have not purchased. To use Alibaba Cloud Marketplace images to create ECS instances in Auto Scaling, you must purchase the images from Alibaba Cloud Marketplace in advance.

## Can images of different regions use the same product code?

Yes, images of different regions can use the same product code as long as the images are supported in the regions.

## I purchased 100 images with the same product code. Can I use them in any region?

No, you cannot use an image in all regions. Alibaba Cloud Marketplace images are region-specific. If you want to use an image in a specific region, purchase the image for that region.

## What information do I need to provide when I submit a ticket about Auto Scaling?

When you submit a ticket, provide the ID of the scaling activity or relevant logs to facilitate troubleshooting. The ID of a scaling activity is specified by the ScalingActivityId parameter.

For information about how to query scaling activities, see View the details of a scaling activity.

## Why is a resource error returned when Auto Scaling creates an ECS instance?

If the following errors are returned, ECS resources may be insufficient. We recommend that you change the zone for the scaling group and try again.

```
Fail to create Instance into scaling group("The resource is out of usage").
```

```
Fail to create Instance into scaling group(The specified region is in resource control, please try later.").
```

## How do I avoid scale-out failures due to insufficient inventory of an instance type?

When you create a scaling group, specify multiple zones by selecting vSwitches in different zones. When you create a scaling configuration for the scaling group, select multiple ECS instance types. Then, if an instance type is unavailable in one zone, Auto Scaling automatically selects another instance type with sufficient inventory in that zone. If all instance types are unavailable in that zone, Auto Scaling switches to another zone. For more information, see Create a scaling group and Create a scaling configuration (Elastic Compute Service).

## Why is an ECS instance for which I enabled release protection automatically released from a scaling group?

Auto Scaling can automatically release an ECS instance created during a scale-out event even if you enabled release protection by using the ECS console or calling the ModifyInstanceAttribute operation.

To prevent an ECS instance from being automatically released, you must put the ECS instance into the protected state. For more information, see Put an ECS instance into the Protected state.

## How do I prevent Auto Scaling from automatically releasing an ECS instance that I manually added to a scaling group?

To prevent an ECS instance from being automatically released, you must put the ECS instance into the protected state. For more information, see Put an ECS instance into the Protected state.

## When Auto Scaling automatically adds or removes ECS instances to or from a scaling group, does Auto Scaling automatically add or remove the IP addresses of the ECS instances to or from the whitelists of the RDS instances or ApsaraDB for Memcache instances associated with the scaling group?

Auto Scaling can automatically add or remove the IP addresses of ECS instances to or from the whitelists of RDS instances, but not the ApsaraDB for Memcache instances.

## How do I prevent Auto Scaling from automatically removing an ECS instance that I manually added to a scaling group?

Assume that you want to prevent Auto Scaling from automatically removing the 100 ECS instances that you manually added to a scaling group. Perform the following operations to configure the scaling group:

- Set the minimum number of ECS instances in the scaling group to a value greater than or equal to 100.
- Set the removal policy to **Earliest Instance Created Using Scaling Configuration** for the first step.

Manually added ECS instances are not created based on scaling configurations. Therefore, the preceding settings ensure that Auto Scaling preferentially selects, removes, and releases automatically created ECS instances. Auto Scaling selects and removes manually added ECS instances only after all automatically created ECS instances are removed. When Auto Scaling removes a manually created ECS instance, Auto Scaling does not release the ECS instance.

> **Note** To prevent Auto Scaling from automatically removing a manually added ECS instance, do not stop the ECS instance. If the ECS instance is stopped, Auto Scaling considers that the ECS instance is unhealthy and automatically removes it.

## After ECS instances are removed from scaling groups and released, is the data of the ECS instances retained?

No, Auto Scaling does not retain any data of ECS instances after the instances are removed and released. Therefore, do not store application status information or important data, such as sessions, databases, and logs, on ECS instances that belong to a scaling group. If the status information of your applications must be stored, we recommend that you store the status information on an independent state server or in a database or service. For example, you can store the status information on an independent ECS instance, in RDS, or in Log Service.

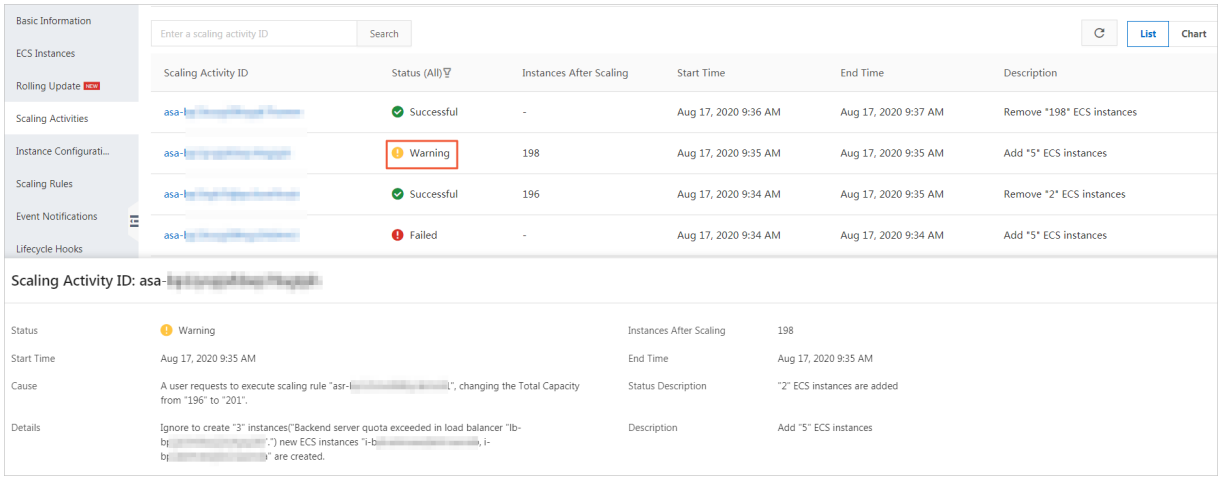
## How do I delete ECS instances that are automatically created by Auto Scaling in a scaling group?

Log on to the Auto Scaling console, find the ECS instances that you want to delete, and then delete them. For more information, see Manually remove or delete an ECS instance.

## When Auto Scaling automatically creates multiple ECS instances, will Auto Scaling attempt to recreate instances that fail to be created?

No, Auto Scaling will not attempt to recreate instances that fail to be created. Auto Scaling does not guarantee that all ECS instances are created in a scaling activity. If some instances fail to be created during a scaling activity, Auto Scaling considers that the scaling activity is complete without attempting to recreate the failed instances. You can view the status of a scaling activity in the Auto Scaling console. For more information, see View the details of a scaling activity.

Assume that Auto Scaling attempts to automatically create 20 ECS instances. 19 instances are created, and one instance fails to be created. Then, Auto Scaling adds the 19 created instances to the scaling group but does not attempt to recreate the failed one. The scaling activity is complete but is in the **Warning** state.




## For an ECS instance that is automatically created by Auto Scaling, how do I obtain its password and log on to it?

If the ECS instance is a Linux instance and a Secure Shell (SSH) key pair is set in the scaling configuration for creating the ECS instance, you can use the SSH key pair to log on to the ECS instance. Auto Scaling does not allow you to set a unified password for automatically created ECS instances. For Linux ECS instances, we recommend that you set an SSH key pair in the scaling configuration.

If you do not set an SSH key pair in the scaling configuration, you must reset the password for the ECS instance in the ECS console. The new password takes effect after the ECS instance is restarted. Then, you can use the new password to log on to the ECS instance.

## Why am I unable to use the password set in the custom image to log on to an ECS instance that is automatically created by Auto Scaling?

For security reasons, Auto Scaling does not use the password set in the custom image as the logon password of automatically created ECS instances. We recommend that you set an SSH key pair in the scaling configuration for creating an ECS instance. Then, you can use the SSH key pair to log on to the ECS instance.

If you do not set an SSH key pair in the scaling configuration, you must reset the password for the ECS instance in the ECS console. The new password takes effect after the ECS instance is restarted. Then, you can use the new password to log on to the ECS instance.

## How do I synchronize data to ECS instances in a scaling group?

You can configure a custom image in the scaling configuration for creating ECS instances. In this way, you can pass custom data to the created ECS instances. To synchronize data to or from a running ECS instance, we recommend that you install rsync on the ECS instance and use rsync to transfer data.

## Why is IP Address 127.0.0.1 deleted from the */etc/hosts* file of ECS instances that are automatically created by Auto Scaling?

If you use a custom image where the */etc/hosts* file is modified to create ECS instances, the file is restored to default settings when Auto Scaling automatically creates ECS instances from the custom image. Therefore, your modifications to the /etc/hosts file are lost. To retain the modifications to the */etc/hosts* file, add a script to the *rc.local* file to check whether the corresponding information exists in the */etc/hosts* file and add the information if the information does not exist.

## What does an SLB instance do after it is associated with a scaling group?

The SLB instance forwards traffic to ECS instances in the scaling group based on the routing method. By associating an SLB instance with a scaling group, you can improve the performance and availability of your applications. For more information about SLB, see What is CLB?.

## How do SLB instances work with scaling groups?

You can specify an SLB instance for one or more scaling groups. Then, Auto Scaling automatically adds ECS instances in the scaling groups to the SLB instance as backend servers. By default, the weight of an ECS instance added to an SLB instance as a backend server is 50. For more information about SLB backend servers, see Add a default backend server.

## After an ECS instance is added to a scaling group, does Auto Scaling automatically add the ECS instance to an SLB instance as a backend server?

Yes, Auto Scaling automatically adds the ECS instance as a backend server to the SLB instance associated with the scaling group. You must associate the SLB instance with the scaling group in advance.

## Can Auto Scaling add an automatically created ECS instance to multiple SLB instances as a backend server?

Yes, Auto Scaling can add an automatically created ECS instance to multiple SLB instances associated with the scaling group. Only a limited number of SLB instances can be associated with a scaling group. For more information, see 使用限制.

If you want to associate more SLB instances with a scaling group,submit a ticket.

## After an ECS instance is added to an SLB instance as a backend server, can I change the weight of the ECS instance?

Yes, you can change the weight of an ECS instance after it is added to an SLB instance as a backend server. For more information, see Change the weight of a backend server.

An SLB instance balances loads among its backend servers based on the ratio of their weights, instead of the weight values. Assume that you have two ECS instances and their weights are both 50. If you change both the weights to 100, the SLB instance balances loads among the two ECS instances in the same way. This is because the weight ratio remains unchanged and is still 1:1. Typically, ECS instances in a scaling group provide the same service and are of the same instance type. Therefore, their weights are the same. By default, the weight of an ECS instance added to an SLB instance as a backend server is 50.

## After I associate a public SLB instance with a scaling group, do I need to set the public bandwidth for ECS instances when I create a scaling configuration for the scaling group?

No, public bandwidth is an optional setting for ECS instances. However, we recommend that you set the public bandwidth to at least 1 Mbit/s when you create a scaling configuration. This facilitates the management of ECS instances.

## Why is a health check error about the SLB instance returned when I create a scaling group?

The following error is returned if health check is disabled for the SLB instance:

```
The current health check type of load balancer "xxxx" does not support this action.
```

Make sure that health check is enabled for SLB instances that are associated with scaling groups. For more information, see Configure health checks.

## How does an SLB instance determine whether a newly created ECS instance can process requests?

After a newly created ECS instance is added to a scaling group, the SLB instance associated with the scaling group checks whether the corresponding ports on the ECS instance can respond to requests. The SLB instance forwards requests to the new ECS instance only when the ports on the ECS instance can respond to requests.

## Why does the timeout period of a Layer-7 HTTP listener of an SLB instance exceed 60 seconds?

Problem: The timeout period of an HTTP listener is about 60 seconds on an SLB instance. However, the timeout period exceeds 60 seconds for all HTTP requests, or the SLB instance directly returns a 504 error code for an HTTP request.

Cause: You can set the timeout period of an HTTP listener to make sure that an HTTP request is responded within the specified period. However, the total timeout period depends on the number of ECS instances configured for the SLB instance.

If an HTTP request times out on the first ECS instance, the SLB instance forwards the HTTP request to the second ECS instance. The same rule applies to the remaining ECS instances until all ECS instances are polled or until the request is processed by an ECS instance. Assume that three ECS instances are added as backend servers to an SLB instance. The total HTTP timeout period reaches about 180 seconds.

Other services may also affect the HTTP timeout period of an SLB instance. We recommend that you do not rely on the HTTP timeout period configured on the SLB instance to monitor the processing of HTTP requests. Instead, we recommend that you configure the timeout period in applications deployed on ECS instances.

## What does an RDS instance do after it is associated with a scaling group?

RDS is a stable, reliable, and scalable online database service. By associating an RDS instance with a scaling group, you can back up data of the scaling group to the RDS instance based on custom backup policies. This improves the security and reliability of the data and allows you to restore the data when it is deleted by mistake. For more information about RDS, see What is ApsaraDB RDS?.

## How do RDS instances work with scaling groups?

After an RDS instance is associated with a scaling group, Auto Scaling automatically adds the internal IP addresses of newly created ECS instances in the scaling group to the whitelist of the RDS instance. This allows the ECS instances to access the RDS instance over the internal network. You can associate multiple RDS instances with a scaling group. For more information about whitelists of RDS instances, see Use a database client or the CLI to connect to an ApsaraDB RDS for MySQL instance.

# 2.Scaling configuration

## 2.1. Create a scaling configuration (Elastic Compute Service)

This topic describes how to create a scaling configuration for a scaling group that contains Elastic Compute Service (ECS) instances. Auto Scaling uses the scaling configuration as a template to automatically create ECS instances in a scaling group after Auto Scaling triggers scale-out activities based on your configurations, such as scheduled tasks.

### Prerequisites

- A scaling group that contains ECS instances is created. If you select Intelligent Configuration as the instance configuration mode and you do not set the Instance Type parameter when you create a scaling configuration for a scaling group, make sure that the scaling group resides in a virtual private cloud (VPC). For more information, see Create a scaling group.
- A security group is created. If your scaling group resides in a VPC, make sure that the security group and the scaling group are in the same VPC. For more information, see Create a security group.

### Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. 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.
5. In the upper part of the page, click the **Instance Configuration Sources** tab.
6. Click the **Scaling Configurations** tab.
7. Click **Create Scaling Configuration**.
8. In the Basic Configurations step, configure the following parameters and click **Next: System Configurations**.

| Parameter | Description | References |
| --- | --- | --- |

| Parameter | Description | References |
| --- | --- | --- |
| **Billing Method** | Valid values:<br>◦ **Pay-As-You-Go**: Auto Scaling allocates and releases resources on demand. You are charged only for the resources that you use. You do not need to purchase resources in advance.<br>◦ **Preemptible Instance**: The market price of a preemptible instance fluctuates based on the supply and demand of the instance type. Compared with pay-as-you-go instances, preemptible instances incur lower costs but may be automatically reclaimed. You can use preemptible instances to reduce costs in specific business scenarios. | ◦ Pay-as-you-go<br>◦ Overview |

| Parameter | Description | References |
| --- | --- | --- |
| **Instance Configuration Mode** | Different instance types are provided to meet diverse requirements. Auto Scaling provides the following instance configuration modes:<br>○ **Intelligent Configuration**: Intelligent Configuration allows you to specify only the number of vCPUs, memory size, instance family, and maximum price. The system selects an instance type with the lowest price based on your configurations, to implement more intelligent and effective scale-out. Intelligent Configuration is only applicable to scaling groups that reside in VPCs. Intelligent Configuration also reduces the chance that the scale-out activity fails due to insufficient inventory of instance types.<br>○ **Instance Type**: You can specify multiple instance types. If there is insufficient inventory in one instance type, Auto Scaling evaluates the instance types that you specified and creates instances of the type that has sufficient inventory. This helps increase the success rate of scaling activities.<br>**Note** You can configure weights for different instance types based on performance metrics such as the number of vCPUs. For more information, see Use performance metrics to measure Auto Scaling.<br>If you select a burstable instance, Enable Unlimited Mode for Burstable Instances is not selected by default. You can select Enable Unlimited Mode for Burstable Instances based on your business requirements. For more information about burstable instances, see Overview.<br>Auto Scaling allows you to configure **Intelligent Configuration** and **Instance Type**. If you configure both parameters, you cannot configure weights for instance types. In this case, Auto Scaling uses the instance types that you specified for the Instance Type parameter to execute scaling activities. If there is insufficient inventory in the instance types that you specified, Auto Scaling uses the instance type that is matched by using the Intelligent Configuration parameter and whose price is the lowest to create ECS instances. | Instance family |

| Parameter | Description | References |
| --- | --- | --- |
| **Image** | An image can provide the data that is required to create ECS instances, such as the system and application environments, and related software configurations. | Image overview |
| **Storage** | Select a system disk or one or more data disks for ECS instances to store data.<br>If the disk category that you specify is not supported by the instance type or zone, the scale-out activity fails. Auto Scaling provides system disks and data disks of different categories to improve the success rate of scale-out. After you configure multiple disk categories, Auto Scaling selects disk categories based on the specified order. | Elastic Block Storage devices |
| **Public IP Address** | Assign IPv4 addresses to ECS instances to connect the instances to the Internet. If you select Assign Public IP Address, you must specify a billing method for public bandwidth. | ○ IP addresses of ECS instances within VPCs<br>○ IP addresses of ECS instances within the classic network<br>○ Public bandwidth |
| **Security Group** | A security group is a virtual firewall that controls access to ECS instances. | ○ Overview<br>○ Create a security group |

9. In the System Configurations (Optional) step, configure the following parameters and click **Next: Preview**.

| Parameter | Description | References |
| --- | --- | --- |
| **Tags** | Tags can be used to identify resources. You can use tags to classify ECS instances and related resources for easy management and search. | ○ Overview<br>○ Create or bind a tag |
| **Resource Group** | Resource groups enable you to categorize your resources by purpose, permissions, or region. This way, you can manage the resources across accounts and projects. | Resource groups |

| Parameter | Description | References |
| --- | --- | --- |
| **Logon Credentials** | The supported logon credentials vary based on the OS type.<br>○ Linux: You can select an SSH key pair when you create a scaling configuration. You can also configure logon credentials after you create ECS instances.<br>○ Windows: You can configure logon credentials only after you create ECS instances. | ○ Overview<br>○ Reset the logon password of an instance |
| **Instance Name** | This parameter specifies the name of the ECS instance. If you do not set this parameter, a default name is used. | None |
| **Host** | This parameter specifies the hostname of the ECS instance. If you do not set this parameter, a default name is used. | Set rules for generating sequential and unique hostnames |
| **RAM Role** | You can bind RAM roles to ECS instances. Then, you can use Security Token Service (STS) tokens to call API operations of other Alibaba Cloud services. This ensures the security of your AccessKey pair and helps you perform fine-grained permission control and management by using the RAM roles.<br>**Note** You can select RAM roles only for scaling configurations in a VPC-type scaling group. | ○ Overview<br>○ Attach an instance RAM role |
| **User Data** | User data can be used to configure the startup behavior of an ECS instance or pass data to the ECS instance, such as to automatically obtain software resource packages, activate services, and print logs. You must write a custom script and pass data to the script by using the User Data feature.<br>**Note** You can specify user data only for scaling configurations in a VPC-type scaling group. | ○ Overview of ECS instance user data<br>○ Manage the user data of Linux instances |

| Parameter | Description | References |
| --- | --- | --- |
| **Private Pool** | This parameter can be used to ensure the availability of resources. Auto Scaling uses resources from the associated private pool to create ECS instances. After an elasticity assurance or a capacity reservation is created, the system generates a private pool to reserve capacity for a specific number of instances that contain specific attributes. Valid values:<br>◦ Open: The system uses the resources that are allocated to open private pools. If all resources in open private pools are used, the system uses the resources that are allocated to public pools.<br>◦ Do Not Use: The system uses the resources that are allocated to public pools instead of the resources that are allocated to private pools to create instances.<br>◦ Target: The system uses the resources that are allocated to a specific pool or an open private pool to create instances. If all resources that are allocated to a specific pool or an open private pool are used, no instances can be created. | Overview |
| **Dedicated Host** | You must specify a dedicated host for Auto Scaling to create ECS instances to prevent resource contention and ensure security compliance. A dedicated host provides exclusive physical resources for the ECS instances. | ◦ What is DDH?<br>◦ Create a dedicated host |

10. In the Preview step, check your configurations, specify a name for the scaling configuration, and then click **Create**.
11. In the Created dialog box, click **Enable Configuration**.

## Related information

### References

- Create an instance by using the wizard
- CreateScalingConfiguration

# 2.2. Create a scaling configuration (Elastic Container Instance)

This topic describes how to create a scaling configuration for a scaling group that contains elastic container instances. Auto Scaling uses the scaling configuration as a template to automatically create elastic container instances in a scaling group after Auto Scaling triggers scale-out activities based on your configurations, such as scheduled tasks.

## Prerequisites

- A scaling group that contains elastic container instances is created. For more information, see Create a scaling group.
- A security group is created. For more information, see Create a security group.

## Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. 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.
5. In the upper part of the page, click the **Instance Configuration Sources** tab.
6. Click the **Scaling Configurations** tab.
7. Click **Create Scaling Configuration**.
8. In the Basic Configurations step, configure the following parameters and click **Next: Other Settings**.

| Parameter | Description | References |
| --- | --- | --- |
| **Billing Method** | ○ **Pay-As-You-Go**: Auto Scaling allocates and releases resources on demand. You are charged only for the resources that you use. You do not need to purchase resources in advance.<br>○ **Preemptible Instance**: The market price of a preemptible instance fluctuates based on the supply and demand of the instance type. Compared with pay-as-you-go instances, preemptible instances incur lower costs but may be automatically reclaimed. You can use preemptible instances to reduce costs in specific business scenarios. | ○ Billing<br>○ Create a preemptible instance |
| **Configuration Information** | This parameter specifies the region, virtual private cloud (VPC), and vSwitch based on which you want to create the elastic container instance. The configuration information is automatically populated. The elastic container instance and the scaling group to which the instance belongs must be in the same region, VPC, and vSwitch. | None |
| **Security Group** | A security group is a virtual firewall that manages access to elastic container instances. | Configure a security group |

| Parameter | Description | References |
| --- | --- | --- |
| **Container Group Configuration** | A container group is a set of containers that can be scheduled to the same host. The lifecycle of a container group is determined based on all containers in the group. These containers share the network and storage resources of the container group. A container group can be used in a similar manner as a pod in Kubernetes. | ○ Create a custom temporary storage space<br>○ Use an image cache to create an elastic container instance |
| **Container Configuration** | A container is a lightweight and executable standalone software package. A container is also the running entity of an image. | ○ Specify the number of vCPUs and memory size to create an elastic container instance<br>○ Configure startup commands and arguments for a container<br>○ Use probes to perform health checks on a container |

9. Configure the following parameters and click **Configuration confirmation**.

| Parameter | Description | References |
| --- | --- | --- |
| **EIP** | Elastic IP addresses (EIPs) are public IP addresses that can be purchased and managed based on the requirements of your business scenario. You can enable Internet access for an elastic container instance by associating an EIP with the instance. | Connect to the Internet |
| **Mirror Warehouse Access Credentials** | This parameter is required if your container uses private images such as non-Alibaba Cloud Container Registry images and non-Docker Hub images. This parameter allows the system to pull images from the image repository. | None |
| **Label** | Each label is a key-value pair and can be used to identify elastic container instances. You can use labels to manage elastic container instances by group to perform searches and batch operations. | Use labels to manage elastic container instances |
| **Resource Group** | You can add Elastic Container Instance resources that are used for different purposes to specific resource groups. Then, you can grant the administrator permissions to RAM users, to allow the users to perform fine-grained control on the resources in resource groups. | Use resource groups to manage the permissions of a RAM user |

| Parameter | Description | References |
| --- | --- | --- |
| **Configuration Name** | The name specifies the scaling configuration. If you do not set this parameter, the scaling configuration ID is used. | None |

10. Click **Create Order**.
11. In the dialog box that is displayed, click **Enable**.

### Related information

#### References

- Use an NGINX image to create an elastic container instance
- Use a CentOS image to create an elastic container instance
- Pull an image from Docker Hub to create an elastic container instance

# 2.3. Apply a scaling configuration

This topic describes how to apply a scaling configuration. You can create multiple scaling configurations for a scaling group and apply one.

## Context

After you apply a scaling configuration, the scaling configuration takes effect and enters the **Active** state. Only one scaling configuration can be in the Active state in a scaling group. After a scaling configuration is applied, if you apply another scaling configuration, the previous scaling configuration enters the **Inactive** state.

## Procedure

1. Log on to the Auto Scaling console.
2. In the top navigation bar, select a region.
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 **Instance Configuration Sources** tab.
5. In the upper-left corner of the Instance Configuration Source page, click **Create Scaling Configuration**.
6. Find the scaling configuration that you want to apply and click **Apply** in the **Actions** column.
7. In the message that appears, click **OK**.

## Result

After you apply a scaling configuration, Auto Scaling creates ECS instances based on the scaling configuration when a scale-out event is triggered.

# 2.4. Modify a scaling configuration

This topic describes how to modify a scaling configuration. You can modify the parameters of a scaling configuration after it is created.

## Context

After the scaling configuration is modified, Auto Scaling creates ECS instances based on the new settings of the scaling configuration. However, ECS instances that are created based on the original settings of the scaling configuration are not affected.

## Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. 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.
5. In the upper part of the page, click the **Instance Configuration Sources** tab.
6. In the upper-left corner of the Instance Configuration Source page, click the **Scaling Configurations** tab.
7. Find the scaling configuration that you want to modify and click **Edit** in the **Actions** column.
8. In the Basic Configurations step, configure the following parameters and click **Next: System Configurations**.

| Parameter | Description | References |
| --- | --- | --- |
| **Billing Method** | Valid values:<br>○ **Pay-As-You-Go**: Auto Scaling allocates and releases resources on demand. You are charged only for the resources that you use. You do not need to purchase resources in advance.<br>○ **Preemptible Instance**: The market price of a preemptible instance fluctuates based on the supply and demand of the instance type. Compared with pay-as-you-go instances, preemptible instances incur lower costs but may be automatically reclaimed. You can use preemptible instances to reduce costs in specific business scenarios. | ○ Pay-as-you-go<br>○ Overview |

| Parameter | Description | References |
| --- | --- | --- |
| **Instance Configuration Mode** | Different instance types are provided to meet diverse requirements. Auto Scaling provides the following instance configuration modes:<br>○ **Intelligent Configuration**: Intelligent Configuration allows you to specify only the number of vCPUs, memory size, instance family, and maximum price. The system selects an instance type with the lowest price based on your configurations, to implement more intelligent and effective scale-out. Intelligent Configuration is only applicable to scaling groups that reside in VPCs. Intelligent Configuration also reduces the chance that the scale-out activity fails due to insufficient inventory of instance types.<br>○ **Instance Type**: You can specify multiple instance types. If there is insufficient inventory in one instance type, Auto Scaling evaluates the instance types that you specified and creates instances of the type that has sufficient inventory. This helps increase the success rate of scaling activities.<br>**Note** You can configure weights for different instance types based on performance metrics such as the number of vCPUs. For more information, see Use performance metrics to measure Auto Scaling.<br>If you select a burstable instance, Enable Unlimited Mode for Burstable Instances is not selected by default. You can select Enable Unlimited Mode for Burstable Instances based on your business requirements. For more information about burstable instances, see Overview.<br>Auto Scaling allows you to configure **Intelligent Configuration** and **Instance Type**. If you configure both parameters, you cannot configure weights for instance types. In this case, Auto Scaling uses the instance types that you specified for the Instance Type parameter to execute scaling activities. If there is insufficient inventory in the instance types that you specified, Auto Scaling uses the instance type that is matched by using the Intelligent Configuration parameter and whose price is the lowest to create ECS instances. | Instance family |

| Parameter | Description | References |
| --- | --- | --- |
| **Image** | An image can provide the data that is required to create ECS instances, such as the system and application environments, and related software configurations. | Image overview |
| **Storage** | Select a system disk or one or more data disks for ECS instances to store data.<br>If the disk category that you specify is not supported by the instance type or zone, the scale-out activity fails. Auto Scaling provides system disks and data disks of different categories to improve the success rate of scale-out. After you configure multiple disk categories, Auto Scaling selects disk categories based on the specified order. | Elastic Block Storage devices |
| **Public IP Address** | Assign IPv4 addresses to ECS instances to connect the instances to the Internet. If you select Assign Public IP Address, you must specify a billing method for public bandwidth. | ◦ IP addresses of ECS instances within VPCs<br>◦ IP addresses of ECS instances within the classic network<br>◦ Public bandwidth |
| **Security Group** | A security group is a virtual firewall that controls access to ECS instances. | ◦ Overview<br>◦ Create a security group |

9. In the System Configurations (Optional) step, configure the following parameters and click **Next: Preview**.

| Parameter | Description | References |
| --- | --- | --- |
| **Tags** | Tags can be used to identify resources. You can use tags to classify ECS instances and related resources for easy management and search. | ◦ Overview<br>◦ Create or bind a tag |
| **Resource Group** | Resource groups enable you to categorize your resources by purpose, permissions, or region. This way, you can manage the resources across accounts and projects. | Resource groups |

| Parameter | Description | References |
| --- | --- | --- |
| **Logon Credentials** | The supported logon credentials vary based on the OS type.<br>○ Linux: You can select an SSH key pair when you create a scaling configuration. You can also configure logon credentials after you create ECS instances.<br>○ Windows: You can configure logon credentials only after you create ECS instances. | ○ Overview<br>○ Reset the logon password of an instance |
| **Instance Name** | This parameter specifies the name of the ECS instance. If you do not set this parameter, a default name is used. | None |
| **Host** | This parameter specifies the hostname of the ECS instance. If you do not set this parameter, a default name is used. | Set rules for generating sequential and unique hostnames |
| **RAM Role** | You can bind RAM roles to ECS instances. Then, you can use Security Token Service (STS) tokens to call API operations of other Alibaba Cloud services. This ensures the security of your AccessKey pair and helps you perform fine-grained permission control and management by using the RAM roles.<br>**Note** You can select RAM roles only for scaling configurations in a VPC-type scaling group. | ○ Overview<br>○ Attach an instance RAM role |
| **User Data** | User data can be used to configure the startup behavior of an ECS instance or pass data to the ECS instance, such as to automatically obtain software resource packages, activate services, and print logs. You must write a custom script and pass data to the script by using the User Data feature.<br>**Note** You can specify user data only for scaling configurations in a VPC-type scaling group. | ○ Overview of ECS instance user data<br>○ Manage the user data of Linux instances |

| Parameter | Description | References |
| --- | --- | --- |
| **Private Pool** | This parameter can be used to ensure the availability of resources. Auto Scaling uses resources from the associated private pool to create ECS instances. After an elasticity assurance or a capacity reservation is created, the system generates a private pool to reserve capacity for a specific number of instances that contain specific attributes. Valid values:<br>○ Open: The system uses the resources that are allocated to open private pools. If all resources in open private pools are used, the system uses the resources that are allocated to public pools.<br>○ Do Not Use: The system uses the resources that are allocated to public pools instead of the resources that are allocated to private pools to create instances.<br>○ Target: The system uses the resources that are allocated to a specific pool or an open private pool to create instances. If all resources that are allocated to a specific pool or an open private pool are used, no instances can be created. | Overview |
| **Dedicated Host** | You must specify a dedicated host for Auto Scaling to create ECS instances to prevent resource contention and ensure security compliance. A dedicated host provides exclusive physical resources for the ECS instances. | ○ What is DDH?<br>○ Create a dedicated host |

10. Check you configurations, enter a name for the scaling configuration, and then click **Modify**.

## Related information

- ModifyScalingConfiguration

# 2.5. Delete a scaling configuration

This topic describes how to delete a scaling configuration that you no longer need to maintain a sufficient quota.

## Prerequisites

Before you delete a scaling configuration, make sure that the following conditions are met:

- The scaling configuration is in the **Inactive** state.
- The scaling group does not contain ECS instances that are created based on the scaling configuration.

## Procedure

1. Log on to the Auto Scaling console.

2. In the top navigation bar, select a region.
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 **Instance Configuration Sources** tab.
5. Click the **Scaling Configurations** tab.
6. Find the scaling configuration that you want to delete and click **Delete** in the **Actions** column.
7. In the message that appears, click **OK**.

### Related information

- DeleteScalingConfiguration

# 2.6. Export scaling configurations

This topic describes how to export scaling configurations from a scaling group. You can export scaling configurations to back them up on a local disk or import them to other scaling groups.

### Procedure

1. Log on to the Auto Scaling console.
2. In the top navigation bar, select a region.
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 **Instance Configuration Sources** tab.
5. In the upper-left corner of the Instance Configuration Source page, click the **Scaling Configurations** tab.
6. Click **Export Scaling Configurations**.

   The scaling configurations are exported in the *.csv* format. Example: *ess_scalingConfiguration_list_cn-hangzhou_2019-10-14.csv*.

# 2.7. Import scaling configurations

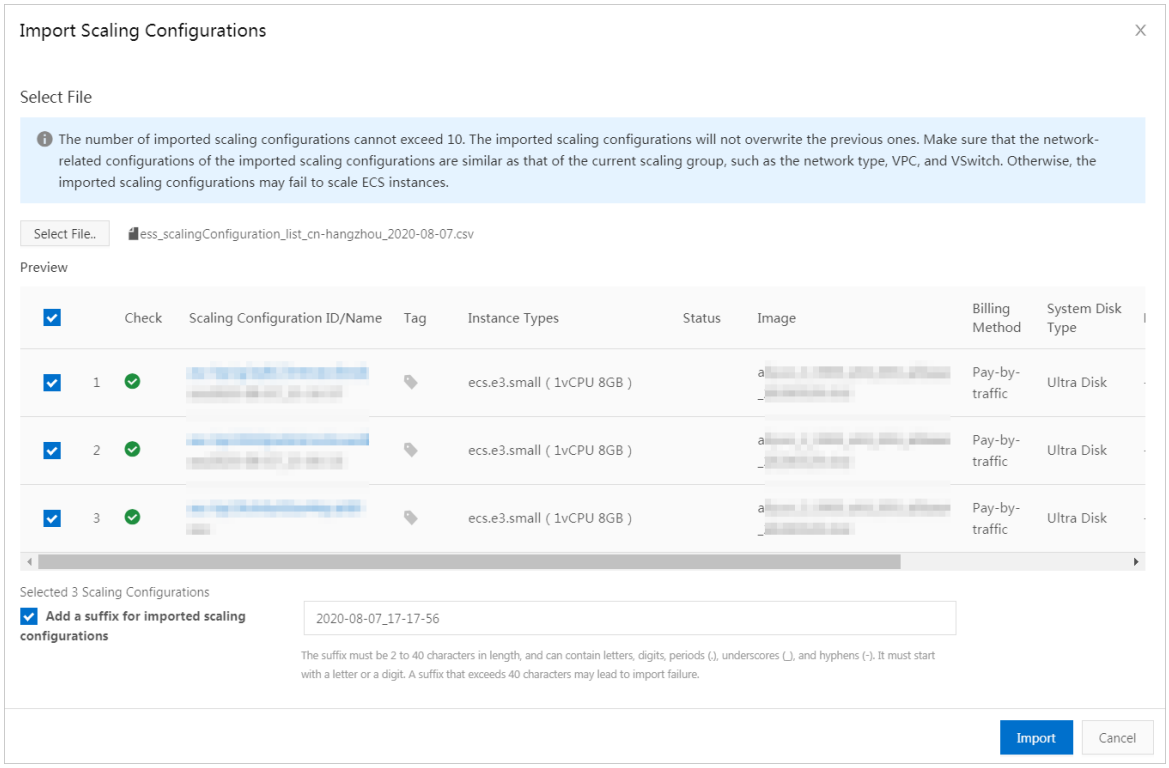
This topic describes how to import scaling configurations of a scaling group to another scaling group. This improves the efficiency in creating scaling configurations.

### Prerequisites

- Scaling configurations are exported as a .csv file. For more information, see Export scaling configurations.
- The two scaling groups have the same network type.
- The two scaling groups are in the same VPC if their network type is VPC.

### Procedure

1. Log on to the Auto Scaling console.
2. In the top navigation bar, select a region.
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 **Instance Configuration Sources** tab.
5. In the upper-left corner of the Instance Configuration Source page, click the **Scaling Configurations** tab.
6. Click **Import Scaling Configurations**.
7. In the **Import Scaling Configurations** dialog box, import the scaling configurations.




   i. Click **Select File**. Navigate to and select the *.csv* file to be imported.
   ii. Select the scaling configurations that you want to import.

   > **Note** If you cannot select the scaling configurations in the **Preview** section, the possible cause is that the source and destination scaling groups have different network types or are not in the same VPC. Check the network configurations of the source and destination scaling groups.

   iii. (Optional)Select **Add a suffix for imported scaling configurations** to make sure that the scaling configuration names are unique in the scaling group.
   iv. Click **Import**.
8. Click **Close**.

# 2.8. Replace the image in a single scaling configuration

This topic describes how to manually replace the image in a scaling configuration. An image provides the data required to create ECS instances, such as the system and application environments, and related software configurations. An image in a scaling configuration is frequently updated to meet business requirements.

## Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. 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.
5. In the upper part of the page, click the **Instance Configuration Sources** tab.
6. In the upper-left corner of the page, click the **Scaling Configurations** tab.
7. Find the scaling configuration for which you want to replace the image. Use one of the following methods to open the **Edit Image** dialog box based on the status of the scaling configuration:
   - If the scaling configuration is in the Active state, click **Edit Image** in the **Actions** column.
   - If the scaling configuration is in the Inactive state, click the icon and select **Edit Image** in the **Actions** column.
8. Select an image type and an image, and click **OK**.

   Supported image types include **Public Image**, **Custom Image**, and **Shared Image**.

   After the image is replaced, you can view the **Image** column corresponding to the scaling configuration. If the name of the new image is displayed, the image is updated.

# 2.9. Replace images in multiple scaling configurations at a time

The image update feature applies to scenarios where applications are frequently published. You can create an image update task in the Auto Scaling console. When the task is executed, Auto Scaling automatically creates a custom image from the source instance and replaces images in the scaling configurations of the scaling group with the created custom image.

## Prerequisites

Image update tasks are executed by using Operation Orchestration Service (OOS). You must authorize OOS to perform operations on related resources. Make sure that at least one of the following requirements is met:

- The current account has permissions to perform operations on resources of ECS and Auto Scaling.

- A RAM role is created for OOS and is granted permissions to perform operations on resources of ECS and Auto Scaling. For more information, see Configure RAM permissions for OOS.

> **Note** We recommend that you attach the AliyunECSFullAccess and AliyunESSFullAccess policies to grant permissions to the RAM role.

## Context

When an image update task is executed, Auto Scaling automatically creates a custom image from the source instance. You are charged based on the size of snapshots created from the image. For more information, see Snapshots.

## Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. 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.
5. In the upper part of the page, click the **Instance Configuration Sources** tab.
6. Click the **Update Image Tasks** tab.
7. Click **Update Image in Scaling Configuration**.
8. Configure the parameters for the image update task.

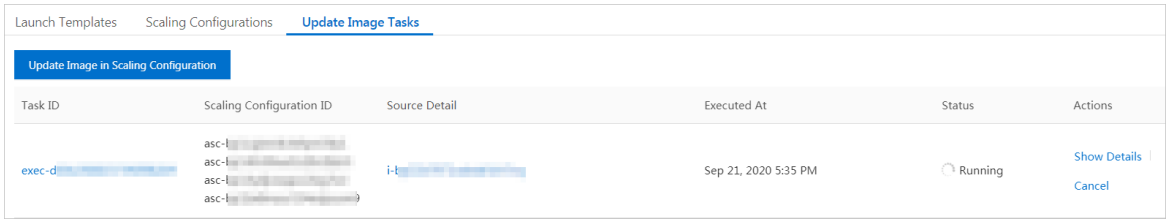
   The following table describes the parameters.

| Parameter | Description |
| --- | --- |
| **Instance** | Select an ECS instance. Auto Scaling creates a custom image from the instance. The custom image is used to replace the images in scaling configurations. The custom image is created only from the system disk of the instance. |
| **Scaling Configuration ID** | Select one or more scaling configurations for which you want to update images. |

| Parameter | Description |
| --- | --- |
| **Executed At** | Specify the time when to execute the task.<br>○ **Now**: The task is executed immediately after it is created.<br>○ **Scheduled**: The task is to be executed at the specified time. You must specify the execution time, in minutes.<br>○ **Periodic**: Specify the recurrence period, expiration time, and start time of the image update task. Assume that you configure the following settings for an image update task on August 17, 2020:<br>■ **Recurrence**: Monthly<br>■ Execute from Day 21 to Day 25 of Each Month<br>■ **Start Time**: 02:00<br>■ **Expired At**: 00:00 on August 26, 2020<br>The image update task is executed once at 02:00 every day from August 21, 2020 to August 25, 2020. |
| **Permission Source** | Select the permission source for OOS to perform operations on related resources.<br>○ **Use Existing Permissions of Current Account**: The permissions of the current account are used.<br>○ **Specify RAM Role and Use Permissions Granted to This Role**: You must select a RAM role to be assumed by OOS and use the permissions granted to this role.<br>**Note** If you set Executed At to **Scheduled** or **Period**, you can only specify a RAM role to grant permissions to OOS. |

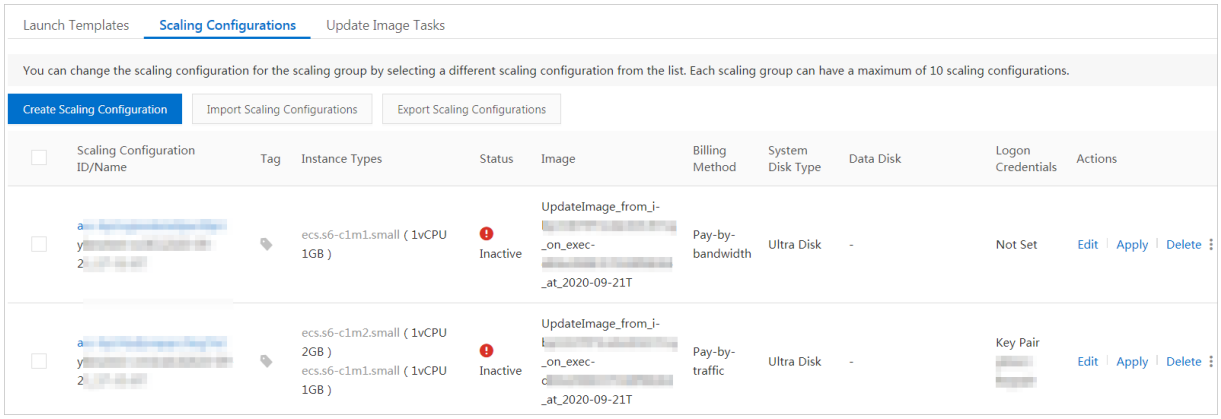
9. Click **OK**.

   The image update task is automatically executed at the specified time. You can view the execution status and source instance of this task in the task list.




## Result

After the task is executed, you can view the execution result on the **Scaling Configurations** tab. If the image names in the scaling configurations are in the `UpdateImage_from_<source ECS instance ID>_on_<image update task ID>` format, the images in the scaling configurations are updated.




# 2.10. Rolling update

Rolling update tasks can be used to update the configurations of multiple Elastic Compute Service (ECS) instances at the same time. You can use rolling update tasks to update the images of, execute scripts on, and install Operation Orchestration Service (OOS) packages on running ECS instances in batches in a scaling group.

## Prerequisites

- OOS is available in the region where the scaling group resides.

  > **Note** OOS is not available in the China (Qingdao), China (Ulanqab), China (Heyuan), US (Silicon Valley), and UAE (Dubai) regions. The rolling update feature is also not available in these regions.

- No scaling activities are in progress in the scaling group.
- New images are prepared before you update images for ECS instances in a scaling group.
- Scripts are prepared before you execute the scripts on ECS instances in a scaling group.
- OOS packages are created before you install the OOS packages on ECS instances in a scaling group. For more information, see Manage custom software on multiple ECS instances.

## Context

The following table describes the types of tasks that are supported by the rolling update feature.

| Task type | Description |
| --- | --- |

| Task type | Description |
| --- | --- |
| Image update | You can execute this type of task to update the OSs of ECS instances in a scaling group.<br>**Note** The image update task has the following impacts on the instance configuration source of the scaling group:<br>• If the instance configuration source is a scaling configuration, the image in the scaling configuration that is in the Enabled state is automatically updated. The images in the scaling configurations that are in the Disabled state in the scaling group are not updated.<br>• If the instance configuration source is a launch template, the image in the launch template is not automatically updated. You need to manually update the image in the launch template. |
| Script execution | You can execute this type of task to perform O&M operations. Examples:<br>• View and update specific system configurations such as the disk size.<br>• Install common software such as Apache.<br>• Deploy business code. |
| OSS package installation | You can execute this type of task to install and uninstall software in batches. |

## Limits

- You can execute rolling update tasks only for ECS instances that are in the **In Service** state.
- You can execute only one rolling update task at a time.

## Create and execute a rolling update task

1. Log on to the Auto Scaling console.
2. In the top navigation bar, select a region.
3. In the left-side navigation pane, click **Scaling Groups**.
4. 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.
5. In the upper part of the page, click the **Rolling Update** tab.
6. Click **Create Execution Task**.
7. In the **Create Execution Task** dialog box, configure the following parameters and click **Create Task**.

| Parameter | Description |
| --- | --- |
| **Task Description** | The description of the rolling update task. |

| Parameter | Description |
| --- | --- |
| **Task Type** | ◦ **Update Image**: You can create this type of task to replace images of ECS instances. You can select a public image, a custom image, a shared image, or an image from Alibaba Cloud Marketplace. The ECS instances are restarted during the update. You must set the following parameters:<br>▪ **Image for Update**: the image that is used to execute the rolling update task<br>▪ **Image for Rollback**: the image that is used to execute the rollback task<br>**Note** By default, the image that you used when you created the rolling update task is automatically selected. You can select other images.<br>◦ **Script Execution**: You can create this type of task and use Cloud Assistant to execute scripts without stopping ECS instances. You must set the following parameters:<br>▪ In the **Script Type** section, select one of the following script types:<br>▪ **Linux Shell**, such as **echo hello** and **hostname**<br>▪ **Windows Bat**, such as **dir c:**<br>▪ **Windows PowerShell**, such as **Get-Services**<br>▪ **Script for Execution**: the script that is used to execute update tasks.<br>▪ **Script for Rollback**: the script that is used to execute rollback tasks<br>**Note** By default, the script that you used when you created the rolling update task is automatically used. You can modify the script based on the requirements of your business scenario.<br>◦ **Install OOS Package**: You can create this type of task to install OOS packages. Select the name and version of the package that you want to install. |
| **Execution Batch** | The number of batches that are required to execute the task. The ECS instances are classified into multiple batches, and the task is executed in batches. Each batch contains at least one ECS instance. For example, if there are 10 ECS instances in the In Service state in a scaling group and the value of the Execution Batch parameter is set to 2, the task is executed in two batches. |

| Parameter | Description |
| --- | --- |
| **Suspension Policy** | ○ **Without Suspension**: Auto Scaling executes the task without interruptions.<br>○ **Suspend First Batch**: Auto Scaling suspends the task after the first batch is complete. You must manually continue the task. After you manually continue the task, the task is not suspended in subsequent batches.<br>○ **Suspend Each Batch**: Auto Scaling suspends the task each time after a batch is complete. You must manually continue the task after each batch is complete. |

8. View the impacts of the execution and click **OK** if you confirm the information.
   Then, the rolling update task is automatically executed. After the rolling update task starts to be executed, it has the following impacts on the scaling group and ECS instances in the scaling group:
   - Auto Scaling suspends the scale-in and scale-out processes during the rolling update. After the rolling update task is complete, the scale-in and scale-out processes are resumed. If you suspend the scale-in and scale-out processes before the task is executed, the processes remain suspended. This ensures that the process status remains unchanged before and after the task is executed.
   - Auto Scaling changes the state of ECS instances to **Standby** in batches and restores the instances to the **In Service** state after the execution is complete.

     > **Note** If the scaling group is associated with Server Load Balancer (SLB) instances, the SLB weight values of the ECS instances that are in the **Standby** state are set to zero and no business traffic is received.

9. Perform the following operations based on the execution status of the rolling update task:
   - If the suspension policy of the rolling update task is **Suspend First Batch** or **Suspend Each Batch**, the task enters the **Pending (Batch Suspension)** state once or multiple times. Confirm whether the ECS instances that are updated meet your business requirements. Perform the following operations after you confirm the information:
     a. Click **Continue** in the **Actions** column.
     b. In the **Continue Execution Task** dialog box, click **OK**.
   - If the rolling update task fails to be executed, the task enters the **Pending (Failure Suspension)** state. To continue this task, perform the following operations:
     a. Click **Details** in the **Actions** column.

b. Find an ECS instance that is in the Failure state, and click **Retry**, **Skip**, or **Cancel** in the **Actions** column.
   - Click **Retry** to execute the rolling update for this ECS instance again.
   - Click **Skip** to execute the rolling update for the subsequent ECS instance. The state of the current ECS instance is changed to **Success**.

     > **Notice** You must manually remove the ECS instance that is skipped from the **Standby** state.

   - Click **Cancel** to execute the rolling update for the subsequent ECS instance. The state of the current ECS instance is changed to **Failure**.

- If you want to cancel the rolling update task, click **Cancel** in the **Actions** column.

  > **Notice** After you cancel the rolling update task, you must manually resume the suspended scaling processes and remove the ECS instances that are in the Failure state or are being updated in the current batch from the **Standby** state.

- For more information, see Roll back a rolling update task.

## Roll back a rolling update task

To restore the configurations of ECS instances when an error occurs, you can roll back a rolling update task that is in the Pending (Batch Suspension) or Pending (Failure Suspension) state. You can also roll back the most recent rolling update task that is executed. You cannot roll back a rollback task.

> **Note** Before you roll back a rolling update task that is in the Pending state, the rolling update task is canceled. Instances that are updated are rolled back.

1. In the execution task list, find the rolling update task that you want to roll back and click **Rollback** in the **Actions** column.
2. In the **Create Rollback Task** dialog box, configure the following parameters.

| Parameter | Description |
| --- | --- |
| **Task Description** | The description of the rollback task. |
| **Task Type** | The task type is the same as the task type of the rolling update task. You cannot change the task type.<br>- If the task type is **Image Update**, the image that you used when you created the rolling update task is automatically selected. You can select other images based on the requirements of your business scenario.<br>- If the task type is **Script Execution**, the script that you specified when you created the rolling update task is automatically used. You can modify the script based on the requirements of your business scenario.<br>- If the task type is **Install OOS Package**, the version of the OOS package that you specified when you created the rolling update task is automatically selected. You can select other versions based on the requirements of your business scenario. You cannot select other packages. |

| Parameter | Description |
| --- | --- |
| **Execution Batch** | The number of batches that are required to execute the task. The ECS instances are classified into multiple batches, and the task is executed in batches. Each batch contains at least one ECS instance. For example, if a scaling group has 10 ECS instances in the In Service state, and Execution Batch is set to 2, the task is executed in two batches. |
| **Suspension Policy** | ○ **Without Suspension**: Auto Scaling executes the task without interruptions.<br>○ **Suspend First Batch**: Auto Scaling suspends the task after the first batch is complete. You must manually continue the task.<br>○ **Suspend Each Batch**: Auto Scaling suspends the task each time after a batch is complete. You must manually continue the task after each batch is complete. |

3. Click **Create Task**.
4. View the impacts of the rollback task and click **OK** if you confirm the information. Then, the rollback task is automatically executed.

## View the details of a rolling update task

You can view the details of a rolling update task and retry or skip the task that is created for an ECS instance.

1. In the execution task list, find the task whose details you want to view and click **Details** in the **Actions** column.
2. View basic task information.

   The basic task information contains the task status and the task type. If the task type is **Script Execution**, click **Script Details** to view the details of the script.
3. View the execution instance list.

   ECS instances that are in various states are displayed.

   ○ If you do not update an ECS instance, you can skip or cancel the task that is created for the instance.

   ○ If you fail to update an ECS instance, you can retry, skip, or cancel the task that is created for the instance in the **Actions** column.

   ○ If the task type is **Script Execution**, click **View** in the **Result** column to view the execution result of the script.

   The following section describes the differences between retrying, skipping, and canceling a task that is created for an ECS instance:

   ○ Click **Retry** to execute the rolling update for this ECS instance again.

   ○ Click **Skip** to execute the rolling update for the subsequent ECS instance. The state of the current ECS instance is changed to **Success**.

   > **Notice** You must manually remove the ECS instance that is skipped from the **Standby** state.

- Click **Cancel** to execute the rolling update for the subsequent ECS instance. The state of the current ECS instance is changed to **Failure**.

## Related information

- Update images and execute scripts
- Use Alibaba Cloud CLI to execute rolling update tasks

# 3.Scaling rule

## 3.1. Overview

The purpose of a scaling rule is determined by its type. A scaling rule can be used to trigger a scaling activity or set the minimum and maximum numbers of ECS instances for a scaling group. This topic describes the types and limits of scaling rules.

### Rule types

The following table describes the scaling rule types supported by Auto Scaling.

| Type | Purpose | Description |
| --- | --- | --- |
| Step scaling rule | A step scaling rule is used to trigger a scaling activity. | When you create a step scaling rule, you must associate an event-triggered task with the step scaling rule and configure a set of step scaling policies that are triggered based on the condition of the event-triggered task. Step scaling rules are similar to simple scaling rules, except that each simple scaling rule defines only one scaling policy. You can use a step scaling rule to precisely control the number of ECS instances to be scaled in a scaling group. |
| Predictive scaling rule | A predictive scaling rule is used to intelligently set the minimum and maximum numbers of ECS instances for a scaling group. | After a predictive scaling rule is executed in a scaling group, Auto Scaling analyzes historical monitoring data of the scaling group to predict the values of specified metrics by using machine learning. Auto Scaling can also automatically create scheduled tasks to help set the optimal minimum and maximum numbers of ECS instances for the scaling group.<br>However, the boundary values you set may not reflect the actual requirements. If the minimum number of ECS instances exceeds the actual requirements, you may purchase excess computing resources. If the maximum number of ECS instances cannot meet the actual requirements, the service stability may be affected due to insufficient computing resources. After a predictive scaling rule is executed, Auto Scaling can obtain historical monitoring data that is generated in a period of at least the past 24 hours. Based on the historical data, Auto Scaling predicts the values of monitored metrics over the next 48 hours by using machine learning. Then, Auto Scaling can calculate the number of ECS instances that are required by the scaling group per hour and provide the number as the predicted value. Predictions are updated once a day, and 48 prediction tasks are created for each of the next 48 hours.<br>**Note** A predictive scaling rule changes only the minimum and maximum numbers of ECS instances for a scaling group, but does not scale ECS instances in the scaling group. |

| Type | Purpose | Description |
| --- | --- | --- |
| Target tracking scaling rule | A target tracking scaling rule is used to trigger a scaling activity. | When you create a target tracking scaling rule, you must select a Cloud Monitor metric and set a target value. Auto Scaling automatically calculates the number of ECS instances required to keep the metric close to the target value, and scales ECS instances accordingly.<br>**Note** After you create a target tracking scaling rule in a scaling group, Auto Scaling automatically creates an event-triggered task and associates this task with the target tracking scaling rule. When the metric of the scaling group reaches the target value, the event-triggered task is triggered to execute the associated target tracking rule. If the event-triggered task is no longer needed, you must delete the associated target tracking scaling rule. Then, Auto Scaling automatically deletes the event-triggered task. |
| Simple scaling rule | A simple scaling rule is used to trigger a scaling activity. | You can specify a simple scaling rule to increase or decrease the number of ECS instances in a scaling group by or to a specific number. |

A predictive scaling rule can be used together with a target tracking scaling rule or a simple scaling rule. When a predictive scaling rule is used with a target tracking scaling rule, we recommend that you set the same metric and target value for both rules. Otherwise, the number of instances in the scaling group may frequently change due to the difference in metrics.

## Limits

The following limits apply when you use scaling rules:

- You can create only a limited number of scaling rules for a scaling group. For more information, see 使用限制.
- After a scaling rule is executed, the actual number of ECS instances in the scaling group may fall outside of the specified range. In this case, Auto Scaling automatically adjusts the number of ECS instances to be added or removed to ensure that the number of ECS instances in the scaling group is within the specified range. Examples:
  - You have a scaling group named asg-bp19ik2u5w7esjcu****. The scaling group can contain up to three instances. One of its scaling rules named add3 specifies that three instances are added during a scaling activity. The scaling group contains two instances. If the add3 scaling rule is executed, only one ECS instance is added to the scaling group.
  - You have a scaling group named asg-bp19ik2u5w7esjcu****. The scaling group must contain at least two instances. One of its scaling rules named reduce2 specifies that two instances are removed during a scaling activity. The scaling group contains three instances. If the reduce2 scaling rule is executed, only one ECS instance is removed from the scaling group.
- If you have overdue payments within your account, all scaling rules fail to be executed.

  **Notice** Make sure that you have sufficient balance within your account to ensure the service availability of Auto Scaling.

# 3.2. Create a scaling rule

The purpose of a scaling rule is determined by its type. A scaling rule can be used to trigger a scaling activity or set the minimum and maximum numbers of Elastic Compute Service (ECS) instances for a scaling group. This topic describes how to create a scaling rule.

## Context

Auto Scaling supports four types of scaling rules. For information about the purposes and limits of scaling rules, see Overview

You can create only a limited number of scaling rules for a scaling group. For more information, see 使用限制.

## Procedure

1. Log on to the Auto Scaling console.
2. 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.
3. In the upper part of the page, click the **Scaling Rules and Activities** tab, and then click the **Scaling Rules** tab.
4. Click **Create Scaling Rule**.
5. In the Create Scaling Rule dialog box, configure parameters for the scaling rule.
   i. Enter a scaling rule name.
   ii. Select a scaling rule type and configure the related parameters.

> **Note** You cannot change the type of a scaling rule after the scaling rule is created.

For information about parameters for different types of scaling rules, see Parameter description for a step scaling rule, Parameter description for a predicative scaling rule, Parameter description for a target tracking scaling rule, or Parameter description for a simple scaling rule.

- Parameters for a step scaling rule

| Parameter | Description |
| --- | --- |
| **Monitoring Type** | Select a monitoring type based on the event-triggered task to be associated with the scaling rule.<br>■ **System Monitoring**: Select this option to associate a system monitoring event-triggered task with the scaling rule.<br>■ **Custom Monitoring**: Select this option to associate a custom monitoring event-triggered task with the scaling rule. |

| Parameter | Description |
| --- | --- |
| **Start Time** | Specify an event-triggered task to be associated with the scaling rule. The condition of the event-triggered task are used as the reference for triggering step scaling operations.<br><br>For example, you can select an event-triggered task that generates an alert when the average CPU utilization is greater than or equal to 80% three times in a row.<br><br>If no event-triggered tasks are available, you can create one. For more information, see Create an event-triggered task.<br><br>**Note** If an event-triggered task is created when you create a step scaling rule, the event-triggered task automatically uses the current step scaling rule as the trigger rule and the current scaling group as the monitoring resource. |
| **Operation** | Specify step scaling operations based on the condition of the event-triggered task. To set a step scaling operation, you must specify the size of the alert breach and the corresponding operation. When the event-triggered task generates an alert, Auto Scaling performs a step scaling operation based on the size of the alert breach. You must set at least one step scaling operation.<br><br>Examples:<br><br>■ Adds two ECS instances when the average CPU utilization is greater than or equal to 80% but less than 90%.<br>■ Adds three ECS instances when the average CPU utilization is greater than or equal to 90%.<br><br>The size of the alert breach for each step scaling operation must be set based on the condition of the event-triggered task. For example, if the condition is that the average CPU utilization is greater than or equal to 80% three times in a row, the lower limit of the alert breach for the initial step scaling operation must be greater than or equal to 80%.<br><br>Step scaling rules support the same operations as simple scaling rules, including changing to N instances, adding N instances, removing N instances, adding instances by N%, and removing instances by N%. |
| **Instance Warmup Time** | The instance warmup period. Unit: seconds. Auto Scaling adds instances in the warmup state to a scaling group, but does not report monitoring data to Cloud Monitor during the warmup period. When Auto Scaling calculates the number of instances to be scaled based on the Cloud Monitor metric, Auto Scaling does not count instances in the warmup state as existing instances in the scaling group. Otherwise, the metric value may change when the warmup period expires.<br><br>Assume that the instance warmup period is 300 seconds and two ECS instances are added to the scaling group. Within 300 seconds after the two ECS instances are created, Auto Scaling does not take the two instances into account when Auto Scaling calculates the average CPU utilization of ECS instances in the scaling group. |

- Parameters for a predictive scaling rule

> **Note** A scaling group can have only one predictive scaling rule.

| Parameter | Description |
| --- | --- |
| **Reference Existing Target Tracking Scaling Rule** | Specifies whether to reference an existing target tracking scaling rule. If a target tracking scaling rule exists in the scaling group, you can use the values of **Metric Type** and **Target Value** specified in the target tracking scaling rule. |
| **Select a rule** | This parameter appears when you select **Reference Existing Target Tracking Scaling Rule**. After you select a target tracking scaling rule, the values of **Metric Type** and **Target Value** specified in the target tracking scaling rule apply to the predictive scaling rule. |
| **Metric Type** | The name of a Cloud Monitor metric to be monitored. Valid values:<br>■ **(ECS) Average CPU Utilization**. Unit: %.<br>■ **(ECS) Average Inbound Internal Traffic**. Unit: KB/min.<br>■ **(ECS) Average Outbound Internal Traffic**. Unit: KB/min. |
| **Target Value** | The target value of the metric to be monitored by Cloud Monitor. The predictive scaling rule calculates an appropriate predicted value based on multiple factors such as the target value.<br>If you change the target value, existing prediction tasks of the current scaling group are cleared. New prediction tasks are created within an hour. |
| **Predicative Mode** | The prediction mode for the scaling group. Valid values:<br>■ **Predict Only**: produces predictions but does not create prediction tasks.<br>■ **Predict and Scale**: produces predictions and creates prediction tasks.<br>We recommend that you first select **Predict Only** and change it to **Predict and Scale** if you confirm that the predictions can meet your requirements. You can view the prediction results in the details of the scaling rule. For more information, see View the prediction effect of a predictive scaling rule. |
| **Preset Max Capacity** | The maximum number of ECS instances in the scaling group. This parameter is used together with **Maximum Capacity Handling Method**.<br>The default value is the maximum number of instances in the current scaling group. |

| Parameter | Description |
| --- | --- |
| **Maximum Capacity Handling Method** | The action to be taken on the predicted value when it exceeds the preset maximum capacity. Valid values:<br>■ **Predicted Capacity Overwrites Preset Max Capacity**: uses the predicted value as the maximum value for prediction tasks.<br>■ **Preset Max Capacity Overwrites Predicted Capacity**: uses the preset maximum capacity as the maximum value for prediction tasks.<br>■ **Predicted Capacity with Additional Ratio**: increases the predicted value with a specified ratio before the system compares it with the preset maximum capacity. When you select this option, the **Increase Ratio** parameter appears. This allows you to set the ratio.<br>Default value: **Predicted Capacity Overwrites Preset Max Capacity** |
| **Increase Ratio** | The ratio of the increment to the predicted value. This parameter appears when you set **Maximum Capacity Handling Method** to **Predicted Capacity with Additional Ratio**. The current predicted value increases with this ratio, and the predicted value after increase is used as the maximum value for prediction tasks. For example, if the current predicted value is 100 and increases with a ratio of 10%, the maximum value for prediction tasks is 110.<br>Valid values: 0 to 100. Default value: 0. |
| **Pre-launch Time** | The buffer time before prediction tasks are executed. By default, all scheduled tasks that are automatically created for a predictive scaling rule are executed on the hour. You can set a buffer time to execute prediction tasks ahead of schedule so that resources can be prepared in advance.<br>Valid values: 0 to 60. Default value: 0. |

- Parameters for a target tracking scaling rule

| Parameter | Description |
| --- | --- |
| **Reference Existing Predictive Scaling Rule** | Specifies whether to reference an existing target tracking scaling rule. If a predictive scaling rule exists in the scaling group, you can specify whether to inherit the metric and target value from the predictive scaling rule. |
| **Metric Type** | The name of a Cloud Monitor metric to be monitored. Valid values:<br>■ **(ECS) Average CPU Utilization**<br>■ **(ECS) Average Inbound Internal Traffic**<br>■ **(ECS) Average Outbound Internal Traffic**<br>■ **(ECS) Average Inbound Public Traffic**<br>■ **(ECS) Average Outbound Public Traffic** |
| **Target Value** | The target value of the Cloud Monitor metric. The target tracking scaling rule keeps the Cloud Monitor metric close to the target value. |
| **Instance Warmup Time** | The instance warmup period. Unit: seconds. Auto Scaling adds instances in the warmup state to a scaling group, but does not report monitoring data to Cloud Monitor during the warmup period. When Auto Scaling calculates the number of instances to be scaled based on the Cloud Monitor metric, Auto Scaling does not count instances in the warmup state as existing instances in the scaling group. Otherwise, the metric value may change when the warmup period expires. |
| **Disable Scale-in** | Specifies whether to disable scale-in events. This parameter affects the number of automatically created event-triggered tasks.<br>■ If this parameter is selected, only one event-triggered task for scale-out events is automatically created and associated with the target tracking scaling rule. Then, the target tracking scaling rule cannot remove ECS instances from the scaling group.<br>■ If this parameter is not selected, two event-triggered tasks are automatically created and associated with the target tracking scaling rule. One task is used for scale-out events, and the other task is used for scale-in events. |

- Parameters for a simple scaling rule

| Parameter | Description |
| --- | --- |
| **Operation** | The operation to be executed when the scaling rule is triggered. Valid values:<br>■ Change to N Instances: When the scaling rule is executed, the number of instances in the scaling group is changed to N. A maximum of 500 instances can be added to or removed from a scaling group at a time.<br>■ Add N Instances: When the scaling rule is executed, N instances are added to the scaling group. A maximum of 500 instances can be added to a scaling group at a time.<br>■ Add Instances by N%: When the scaling rule is executed, N% of the number of existing instances in the scaling group are added. A maximum of 500 instances can be added to a scaling group at a time.<br>■ Remove N Instances: When the scaling rule is executed, N instances are removed from the scaling group. A maximum of 500 instances can be removed from a scaling group at a time.<br>■ Remove Instances by N%: When the scaling rule is executed, N% of the number of existing instances in the scaling group are removed. A maximum of 500 instances can be removed from a scaling group at a time. |
| **Cooldown Time** | Optional. The cooldown period. Unit: seconds. If this parameter is not specified, the cooldown period of the scaling group is used. For more information, see Cooldown time. |

6. Click **OK** to create the scaling rule.

## Related information

- CreateScalingRule

# 3.3. Execute a scaling rule

This topic describes how to manually or automatically execute a scaling rule to scale ECS instances.

## Prerequisites

- You have no overdue payments within your account.

  > **Notice** If you have overdue payments within your account, all scaling activities fail to be executed. Make sure that you have sufficient balance within your account to ensure the service availability of Auto Scaling.

- The scaling group to which the scaling rule belongs is in the **Enabled** state.
- Scaling activities of the scaling group to which the scaling rule belongs meet the following requirements:
  - If the expected number of instances feature is disabled, no scaling activities are in progress in the scaling group.

- If the expected number of instances feature is enabled, no non-parallel scaling activities are in progress in the scaling group. For more information, see Expected number of instances.

## Context

For information about the limits on the number of ECS instances in a scaling group, see 使用限制.

## Manually execute a scaling rule

If you want to temporarily scale ECS instances, you can manually execute a scaling rule.

> **Note** If the scaling group has no scaling activities in progress, you can immediately execute the scaling rule without the need to wait for the cooldown time to expire. Within the cooldown time, Auto Scaling rejects only scaling activity requests triggered by event-triggered tasks from Cloud Monitor.

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. 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.
5. In the upper part of the page, click the **Scaling Rules and Activities** tab, and then click the **Scaling Rules** tab.
6. Find the scaling rule and click **Execute** in the **Actions** column.
7. In the message that appears, click **OK**.

## Execute a scaling rule by using a scheduled task

If your business uses ECS instances on a regular basis, you can create a scheduled task to execute a scaling rule. Auto Scaling automatically executes the scaling rule at the specified point in time. For information about how to create a scheduled task, see Create a scheduled task.

## Execute a scaling rule by using an event-triggered task

If your business does not use ECS instances on a regular basis, you can create an event-triggered task to execute a scaling rule. When the specified condition is met, Auto Scaling automatically executes the scaling rule. For information about how to create an event-triggered task, see Create an event-triggered task. For more information about event-triggered tasks, see Event-triggered task overview.

> **Note** A target tracking scaling rule can be triggered only by the associated event-triggered task. For more information, see Create a scaling rule.

# 3.4. Modify a scaling rule

This topic describes how to modify a scaling rule. You can modify the parameters of a scaling rule based on your actual needs after it is created.

### Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. 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.
5. In the upper part of the page, click the **Scaling Rules and Activities** tab, and then click the **Scaling Rules** tab.
6. Find the scaling rule that you want to modify and click **Edit** in the **Actions** column.
7. Modify the parameters and click **OK**.

   For more information, see Create a scaling rule.

> **Note** You cannot modify the **Rule Type** parameter for a scaling rule or the **Start Time** parameter for a step scaling rule.

# 3.5. Delete a scaling rule

This topic describes how to delete a scaling rule. If you no longer need a scaling rule, you can delete it.

### Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. 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.
5. In the upper part of the page, click the **Scaling Rules and Activities** tab, and then click the **Scaling Rules** tab.
6. Find the target scaling rule and click **Delete** in the **Actions** column.
7. Click **OK**.

# 3.6. View the prediction effect of a predictive scaling rule

You can check whether the prediction based on a predictive scaling rule meets your expectations.

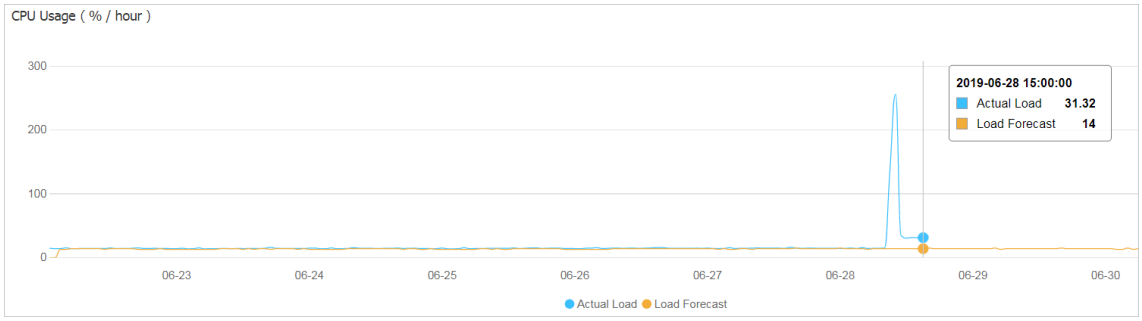
### Procedure

1. Log on to the Auto Scaling console.

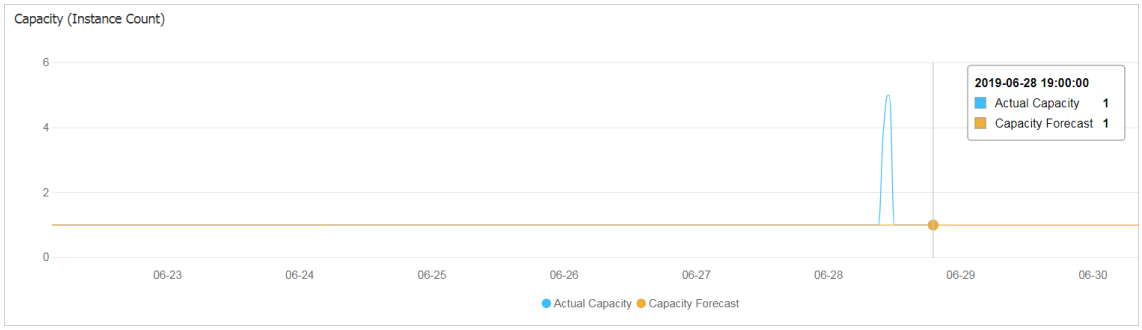
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
5. In the upper part of the page, click the **Scaling Rules and Activities** tab, and then click the **Scaling Rules** tab.
6. Find the predictive scaling rule and click its ID in the Scaling Rule ID/Name column.

   The scaling rule details page shows multiple metrics to help you evaluate the prediction. You can enable **Predict and Scale** after you confirm that the prediction meets your expectations. For more information, see Modify a scaling rule.

   - Compare the actual and predicted CPU utilization to evaluate the prediction accuracy.




   - Compare the actual and predicted numbers of ECS instances to evaluate the prediction accuracy.



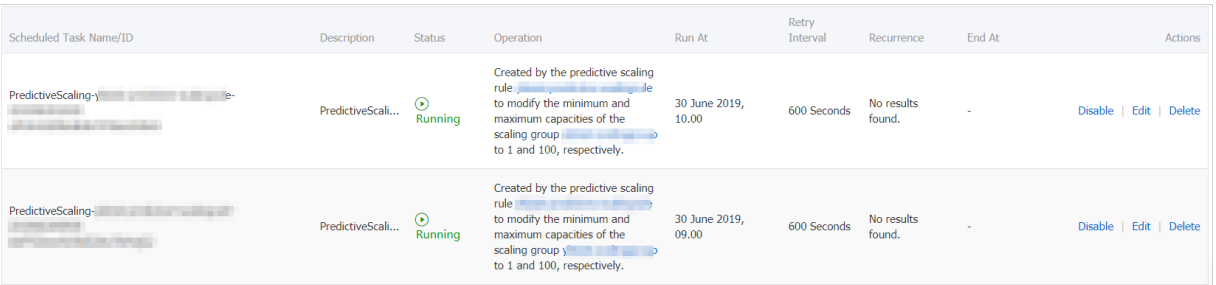

   - Check whether the scheduled plans generated from the prediction meet your expectations.

Scheduled Plans Generated from Forecast | Refresh

| Start Time | Min Capacity | Max Capacity |
|---|---|---|
| 28 June 2019, 19.00 | 1 | 100 |
| 28 June 2019, 20.00 | 1 | 100 |
| 28 June 2019, 21.00 | 1 | 100 |
| 28 June 2019, 22.00 | 1 | 100 |
| 28 June 2019, 23.00 | 1 | 100 |
| 29 June 2019, 00.00 | 1 | 100 |
| 29 June 2019, 01.00 | 1 | 100 |

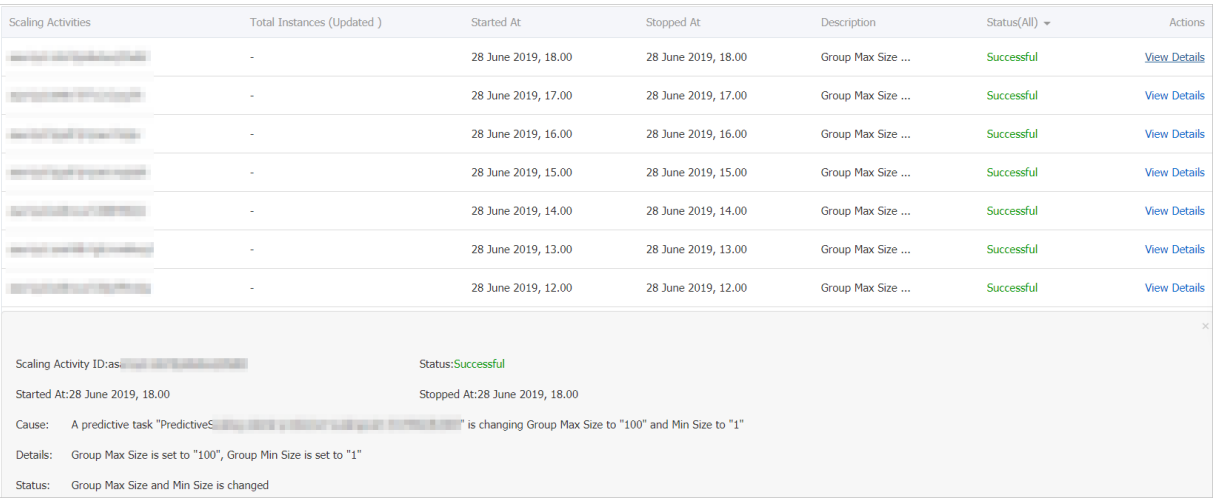
## What's next

If **Predict and Scale** is enabled, Auto Scaling automatically creates scheduled tasks for the predictive scaling rule based on the scheduled plans generated from the prediction. You can view the details of the scheduled tasks on the Scheduled Tasks page. These tasks are named in the following format: PredictiveScaling-Scaling rule name-Execution time.



| Scheduled Task Name/ID | Description | Status | Operation | Run At | Retry Interval | Recurrence | End At | Actions |
|---|---|---|---|---|---|---|---|---|
| PredictiveScaling-y[redacted]e-[redacted] | PredictiveScali... | Running | Created by the predictive scaling rule [redacted]le to modify the minimum and maximum capacities of the scaling group [redacted]p to 1 and 100, respectively. | 30 June 2019, 10.00 | 600 Seconds | No results found. | - | Disable \| Edit \| Delete |
| PredictiveScaling-[redacted] | PredictiveScali... | Running | Created by the predictive scaling rule [redacted] to modify the minimum and maximum capacities of the scaling group y[redacted]p to 1 and 100, respectively. | 30 June 2019, 09.00 | 600 Seconds | No results found. | - | Disable \| Edit \| Delete |

These scheduled tasks change the minimum and maximum numbers of ECS instances in the scaling group and are deleted after they are executed. You can view the details of the scaling activities triggered by these scheduled tasks on the Scaling Activities page.



| Scaling Activities | Total Instances (Updated ) | Started At | Stopped At | Description | Status(All) | Actions |
|---|---|---|---|---|---|---|
| [redacted] | - | 28 June 2019, 18.00 | 28 June 2019, 18.00 | Group Max Size ... | Successful | View Details |
| [redacted] | - | 28 June 2019, 17.00 | 28 June 2019, 17.00 | Group Max Size ... | Successful | View Details |
| [redacted] | - | 28 June 2019, 16.00 | 28 June 2019, 16.00 | Group Max Size ... | Successful | View Details |
| [redacted] | - | 28 June 2019, 15.00 | 28 June 2019, 15.00 | Group Max Size ... | Successful | View Details |
| [redacted] | - | 28 June 2019, 14.00 | 28 June 2019, 14.00 | Group Max Size ... | Successful | View Details |
| [redacted] | - | 28 June 2019, 13.00 | 28 June 2019, 13.00 | Group Max Size ... | Successful | View Details |
| [redacted] | - | 28 June 2019, 12.00 | 28 June 2019, 12.00 | Group Max Size ... | Successful | View Details |

Scaling Activity ID:as[redacted] Status:Successful

Started At:28 June 2019, 18.00 Stopped At:28 June 2019, 18.00

Cause: A predictive task "PredictiveS[redacted]" is changing Group Max Size to "100" and Min Size to "1"

Details: Group Max Size is set to "100", Group Min Size is set to "1"

Status: Group Max Size and Min Size is changed

# 4.Lifecycle hooks

## 4.1. Overview

A lifecycle hook is a tool used to manage the lifecycle of Elastic Compute Service (ECS) instances in a scaling group. When Auto Scaling triggers a scaling activity, a lifecycle hook can be triggered to put the ECS instances involved in the scaling activity into the wait state. This provides a period of time for you to perform custom operations until the lifecycle hook times out.

### What is a lifecycle hook?

A lifecycle hook can suspend a scaling activity that is triggered by Auto Scaling and provides a period of time for you to perform custom operations. You can specify the following parameters for a lifecycle hook:

- Applicable Scaling Activity Type. For example, if you set this parameter to Scale-out Event, the lifecycle hook puts ECS instances into the wait state only during scale-out events.
- Timeout Period. During the timeout period, you can perform custom operations. You can delete a lifecycle hook or call the CompleteLifecycleAction operation to terminate the wait state of a scaling activity ahead of schedule.
- Notification Method. For example, you can use Message Service (MNS) queues or topics to help you perform custom operations in a timely manner or use Operation Orchestration Service (OOS) templates to automatically execute tasks.

  > **Note** You are charged for the MNS service. For more information, visit the Pricing tab of the Message Service page.

- Execution Policy. This parameter specifies the next action to take after the wait state ends. The action is to continue or reject a scaling activity.

  > **Note** Only when **Notification Method** in a lifecycle hook is set to OOS Template, the next action to take is determined by the execution result of the specified OOS template, instead of by the Execution Policy parameter value. If the execution is successful, the scaling activity continues. If the execution fails, the scaling activity that is a scale-out event is rolled back and the scaling activity that is a scale-in event continues.

### Scenarios

You can use lifecycle hooks in the following scenarios. After ECS instances are put into the wait state, you can perform custom operations on them before the instances provide services.

- ECS instances that are created during scale-out events are not suitable to immediately provide services to clients. For example, if ECS instances are required to be added to an ApsaraDB instance or to be bound with secondary elastic network interfaces (ENIs) or if applications deployed on ECS instances require some time to start to provide services, you can use lifecycle hooks.
- During scale-in events, ECS instances are not suitable to be immediately removed. For example, before ECS instances are removed, you must back up data for the instances, copy logs from the instances, or wait until the instances finish processing all client requests.

For more information, see Use lifecycle hooks to ensure service availability and Overview of best practices for lifecycle hooks and OOS templates.

### Limits

- You can create only a limited number of lifecycle hooks for a scaling group. For more information, see 使用限制.
- When a scaling activity is triggered in a scaling group, whether other scaling activities in the scaling group can be executed is determined by the settings of the expected number of instances.
  - If the expected number of instances is not specified for the scaling group, Auto Scaling rejects other scaling activities.
  - If the expected number of instances is specified for the scaling group, Auto Scaling can execute other scaling activities only when the ongoing scaling activity is a parallel scaling activity. For information about how to determine a parallel scaling activity, see Terms.

# 4.2. Modify a lifecycle hook

This topic describes how to modify a lifecycle hook. You can modify the parameters of a lifecycle hook after it is created.

### Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. 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.
5. In the upper part of the page, click the **Lifecycle Hook** tab.
6. Find the lifecycle hook and click **Modify** in the **Actions** column.
7. Modify the parameters for the lifecycle hook and click **OK**.

   You cannot change the name of the lifecycle hook. For more information about other parameters, see the "Parameter description" section in 创建生命周期挂钩.

# 4.3. Delete a lifecycle hook

This topic describes how to delete a lifecycle hook. The number of lifecycle hooks that you can create for a scaling group is limited. To create new lifecycle hooks when the upper limit is reached, you can delete any lifecycle hooks that you no longer need.

### Context

If an ECS instance has been put to the Wait state by a lifecycle hook, deleting the lifecycle hook removes the ECS instance from the Wait state.

### Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.

4. You can use either of the following methods to open the details page of a scaling group.
   - In the **Scaling Group Name/ID** column, click a scaling group name.
   - Click **Manage** in the **Actions** column corresponding to a scaling group.
5. In the left-side navigation pane, click **Lifecycle Hooks**.
6. Use one of the following methods to delete one or more lifecycle hooks.
   - Find the target lifecycle hook and click **Delete** in the **Actions** column.
   - Select the target lifecycle hooks and click **Delete** under the lifecycle hook list.
7. Click **OK**.