Alibaba Cloud Auto Scaling

User Guide

Issue: 20190218

MORE THAN JUST CLOUD | 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 due to product version upgrades , adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.**
4. **This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults " and "as available" basis. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity , applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, special, or punitive damages, including lost profits arising from the use**

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

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

Generic conventions

Table -1: Style conventions

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

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

Contents

1 Usage notes

1.1 ECS instance lifecycle

This topic introduces the life cycle management of ECS instances.

There are two types of ECS instances that are added to scaling groups: automatically created and manually added instances.

Automatically created ECS instances

Automatically created ECS instances are automatically created according to scaling configuration and rules.

Auto Scaling manages the lifecycle of this ECS instance type. It creates ECS instances during scale-up and stops and releases them during scale-down.

Manually added ECS instances

Manually added ECS instances are manually attached to a scaling group.

Auto Scaling does not manage the lifecycle of this ECS instance type. When this ECS instance is removed from a scaling group, either manually or as the result of a scaling -down activity, Auto Scaling does not stop or release the instance.

Instance status

During its lifecycle, an ECS instance is in one of the following states: Pending:

- The ECS instance is being added to a scaling group. For example, Auto Scaling creates the ECS instance and adds it to the Server Load Balancer instance, or to the RDS access whitelist.
- InService: The ECS instance has been added to a scaling group and is functioning correctly.
- Removing: The ECS instance is being removed from a scaling group.

Instance health status

An ECS instance may be in the following health conditions:

- Healthy
- Unhealthy

ECS instances are regarded as unhealthy when they are not Running(`Running`). Auto Scaling automatically removes unhealthy ECS instances from scaling groups.

- Auto Scaling only stops and releases automatically created unhealthy instances.
- It does not stop and release manually added ones.

1.2 Cool-down time

This topic introduces the cool-down time in Auto Scaling.

Cool-down time refers to a period during which Auto Scaling cannot execute any new scaling activity after another scaling activity is executed successfully in a scaling group. Cool-down time is described as follows:

- During cool-down time, the scaling activity requests from CloudMonitor alarm tasks are rejected. Other tasks, such as manually executed scaling rules and scheduled tasks, can immediately trigger scaling activities without waiting for the cool-down time to expire. See *Create a scaling group*.
- During cool-down time, only the corresponding scaling group is locked. Scaling activities set for other scaling groups can be executed. See *Create a scaling rule*.

 Note:
After a scaling group is re-enabled after being disabled, the cool-down time is no longer in effect. For example, if a scaling activity is completed at 12:00 PM and the cool-down time is 15 minutes, the scaling group is then disabled and re-enabled, and the cool-down time is no longer in effect. If a request for triggering a scaling activity is sent at 12:03 PM from the CloudMonitor, the requested scaling activity is executed immediately.

Cool-down time rules

After the scaling group successfully performs the scaling activity, Auto Scaling starts to calculate the cool-down time. If multiple ECS instances are added to or removed from the scaling group in a scaling activity, the cool-down time is calculated since the last instance is added to or removed from the scaling group. See *Examples*. If no ECS instance is successfully added to or removed from the scaling group during the scaling activity, the cool-down time is not calculated.

Within the cool-down time, the scaling group rejects the scaling activity request triggered by the CloudMonitor alarm tasks. However, the scaling activity triggered by

other types of tasks (manually executing tasks, scheduled tasks) can be performed immediately, bypass cooling time.

If you disable a scaling group and then enable the scaling group again, the cool-down time becomes invalid. See *Example 2*.

Note:
The cool-down time only locks the scaling activities in the same scaling group. It does not affect the scaling activities in other scaling groups.

Examples

Example 1

You have a scaling group asg-uf6f3xewn3dvz4bsy7r1. The default cool-down time is 10 minutes, and a scaling rule add3 exists in the scaling group with a cool-down time of 15 minutes.

After a scaling activity is successfully performed based on add3, three ECS instances are added. The cool-down time is calculated since the third instance is added to the scaling group. Within 15 minutes, scaling activity requests triggered by the alarm task from CloudMonitor are rejected.

Example 2

You have a scaling group asg-m5efkz67re9x7a571bjh. The default cool-down time is 10 minutes, and a scaling rule remove1 exists in the scaling group without cool-down time set.

A scaling activity is successfully performed based on remove1 at 18:00. One ECS instance is decreased. Normally, scaling activity requests triggered by the alarm task from CloudMonitor are rejected before 18:10. Disable the scaling group, and then enable the scaling group again at 18:05. The cool-down time becomes invalid. The scaling group accepts the scaling activity requests triggered by the alarm task from CloudMonitor from 18:05 to 18:10.

1.3 Scaling group status

This topic introduces the status of the scaling group.

A scaling group has three statuses: Active, Inactive, and Deleting.

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

1.4 Scaling activity process

This topic introduces the scaling activity process.

A scaling activity’s lifecycle starts with determining the scaling group’s health status and boundary conditions and ends with enabling the cool-down time.

Automatic scaling

Scaling up

1. Determine the scaling group’s health status and boundary conditions.
2. Allocate the activity ID and execute the scaling activity.
3. Create ECS instances.
4. Modify Total Capacity.
5. Allocate IP addresses to the created ECS instances.
6. Add the ECS instances to the RDS access whitelist.
7. Launch the ECS instances.
8. Attach the ECS instances to the Server Load Balancer and set the weight to 0. Wait 60s and then set the weight to 50.
9. Complete the scaling activity, and enable the cool-down time.

Scaling down

1. Determine the scaling group’s health status and boundary conditions.
2. Allocate the activity ID and execute the scaling activity.
3. The Server Load Balancer stops forwarding traffic to the ECS instances. Wait 60s and then remove the ECS instances from the Server Load Balancer.
4. Disable the ECS instances.

5. Remove the ECS instances from the RDS access whitelist.
6. Release the ECS instances.
7. Modify Total Capacity.
8. Complete the scaling activity, and enable the cool-down time.

Manually add or remove an existing ECS instance

Manually add an existing ECS instance

1. Determine the scaling group’s health status and boundary conditions, and check the ECS instance’s status and type.
2. Allocate the activity ID and execute the scaling activity.
3. Add the ECS instance.
4. Modify Total Capacity.
5. Add the ECS instance to the RDS access whitelist.
6. Attach the ECS instance to the Server Load Balancer and set the weight to 0. Wait 60ss and then set the weight to 50.
7. Complete the scaling activity, and enable the cool-down time.

Manually remove an existing instance

1. Determine the scaling group’s health status and boundary conditions.
2. Allocate the activity ID and execute the scaling activity.
3. The Server Load Balancer stops forwarding traffic to the ECS instance.
4. Wait 60s and then remove the ECS instance from the Server Load Balancer.
5. Remove the ECS instance from the RDS access whitelist.
6. Modify Total Capacity.
7. Remove the ECS instance from the scaling group.
8. Complete the scaling activity, and enable the cool-down time.

1.5 Scaling activity status

This topic introduces the status of scaling activity in Auto Scaling.

A scaling activity is in the Rejected status if the request for execution is rejected.

A scaling activity is in the In Progress status if it is being executed.

After a scaling activity is completed, there are three possible states:

- Successful(`Successful`): The scaling activity has successfully added or removed the ECS instances to or from the scaling group as specified by the MaxSize value or the MinSize value adjusted by the scaling rule.

  Note:
  When an ECS instance is successfully added to a scaling group, it has been created and added to the Server Load Balancer instance and the RDS access whitelist. If any of the above steps fail, the ECS instance is considered “failed”.
- Warning(`Warning`): The scaling activity fails to add or remove at least one ECS instance to or from the scaling group as specified by the MaxSize value or the MinSize value adjusted by the scaling rule.
- Failed(`Failed`): The scaling activity fails to add or remove any ECS instance to or from the scaling group as specified by the MaxSize value or the MinSize value adjusted by the scaling rule.

Example

A scaling rule is defined to be added five ECS instances. The existing Total Capacity of the scaling group is three ECS instances, and the MaxSize value is five ECS instances. When the scaling rule is executed, Auto Scaling adds only two ECS instances as specified by the MaxSize value. After the scaling activity is completed, there are three possible states:

- Successful: Two ECS instances are created successfully and correctly added to the Sever Load Balancer instance and the RDS access whitelist.
- Warning: Two ECS instances are created successfully, but only one is correctly added to the Sever Load Balancer instance and the RDS access whitelist. The other one failed, and is rolled back and released.
- Failed: No ECS instances are created. Or two ECS instances are created successfully, but neither are added to the Server Load Balancer instance or the RDS access whitelist. Both are rolled back and released.

1.6 Instance rollback resulting from a scaling activity failure

This topic describes the instance rollback resulting from a scaling activity failure.

When a scaling activity fails to add one or more ECS instances to a scaling group, the failed ECS instances are rolled back. The scaling activity is not rolled back.

For example, if a scaling group has 20 ECS instances, out of which 19 instances are added to the Server Load Balancer instance, only the one ECS instance that fails to be added is automatically released.

Auto Scaling uses Alibaba Cloud’s Resource Access Management (RAM) service to adjust the resources of ECS instances through ECS Open APIs. Therefore, API usage fees apply.

1.7 Remove an unhealthy ECS instance

This topic introduces how to remove an unhealthy ECS instance.

After an ECS instance has been successfully added to a scaling group, the Auto Scaling service regularly scans its status. If the ECS instance is not in the Running(`Running`) status, Auto Scaling removes the ECS instance from the scaling group.

- If the ECS instance was created automatically, Auto Scaling immediately removes and releases it.
- If the ECS instance was added manually, Auto Scaling immediately removes it, but does not stop or release it.

The removal of unhealthy ECS instances is not restricted by the MinSize value. If, due to removal, the number of ECS instances (Total Capacity) in the scaling group is smaller than the MinSize value, Auto Scaling automatically adds ECS instances to the group until the number of instances reaches the MinSize value.

1.8 Notification

This topic introduces the Notification in Auto Scaling.

A text message or email is sent when a scaling activity meets either of the following conditions:

- The scaling activity is triggered by a scheduled task, CloudMonitor alarm task, or health check.
- An ECS instance has been created or released.

A text message or email is sent to corresponding scaling activity when the preceding conditions are met.

1.9 Forced intervention

This topic introduces the forced intervention.

Auto Scaling does not prevent users from performing forced interventions, such as deleting automatically created ECS instances from the ECS console. Auto Scaling handles forced interventions in the following ways:

Resource	Forced intervention types	Solutions
ECS	An ECS instance is deleted from a scaling group through the ECS console or API.	Auto Scaling determines if the ECS instance is in an unhealthy state through *health check*, and if so, removes the instance from the scaling group. The ECS instance's intranet IP address is not automatically deleted from the RDS access whitelist. When the number of ECS instances (Total Capacity) in the scaling group is smaller than the MinSize value, Auto Scaling automatically adds ECS instances to the group until the number of instances reaches the MinSize value.
ECS	The ECS OpenAPI permissions are revoked from Auto Scaling.	All scaling activity requests are rejected.
Server Load Balancer	An ECS instance is removed from a Server Load Balancer instance by force through the Server Load Balancer console or API.	Auto Scaling does not automatically detect this action or handle such an exception. The ECS instance remains in the scaling group, but is released if it was selected according to the removal policy of a scale-down activity.

Resource	Forced intervention types	Solutions
Server Load Balancer	A Server Load Balancer instance is deleted (or its health check function is disabled) by force through the Server Load Balancer console or API.	No ECS instance is added to the scaling group that has been added to the Server Load Balancer instance. Scaling tasks can trigger scaling rules to remove ECS instances from the scaling group . ECS instances deemed unhealthy by the health check function can also be removed.
Server Load Balancer	A Server Load Balancer instance becomes unavailable (due to overdue payment or a fault).	All scaling activities fail, except for activities that are manually triggered to remove ECS instances.
Server Load Balancer	The Server Load Balancer API permissions are revoked from Auto Scaling.	Auto Scaling rejects all scaling activity requests for the scaling groups added to the Server Load Balancer instance.
RDS	The IP address of an ECS instance is removed from an RDS whitelist through the RDS console or API.	Auto Scaling does not detect this action automatically or handle such an exception. The ECS instance remains in the scaling group. If this instance is selected according to the removal policy of a scale-down activity, it is released.

Resource	Forced intervention types	Solutions
RDS	An RDS instance is deleted by force through the RDS console or API.	The scaling group that configured the RDS instance will no longer add ECS instances. No ECS instance is added to the scaling group that has been added to this RDS instance. Scaling tasks can trigger scaling rules to remove ECS instances from the scaling group. ECS instances determined to be unhealthy by the health check function can also be removed.
RDS	An RDS instance becomes unavailable (due to overdue payment or a fault).	All scaling activities fail except for those manually triggered to remove ECS instances.
RDS	The RDS API permission s are revoked from Auto Scaling.	Auto Scaling rejects all scaling activity requests for the scaling groups added to the RDS instance.

1.10 Quantity restrictions

This topic introduces the relevant quantity restrictions of Auto Scaling.

At present, the quantity limits of Auto Scaling are as follows:

- You can create up to 50 scaling groups under an account in one region.
  - Up to 10 scaling configurations can be created for a scaling group.
  - Up to 50 scaling rules can be created for a scaling group.
  - Up to 6 event notifications can be created for a scaling group.
  - Up to 6 lifecycle hooks can be created for a scaling group.
- You can create up to 20 scheduled tasks.
- You can select up to 10 instance types for a scaling configuration.

1.11 Considerations

This topic introduces the considerations about Auto Scaling.

Scaling rules

When you run and compute a scaling rule, the system can automatically adjust the number of ECS instances according to the MaxSize value and the MinSize value of the scaling group. For example, if the number of ECS instances is set to 50 in the scaling rule, but the MaxSize value of the scaling group is set to 45, we compute and run the scaling rule with 45 ECS instances.

Scaling activity

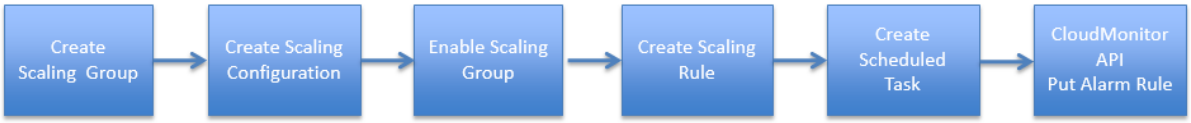
- Only one scaling activity can be executed at a time in a scaling group.
- A scaling activity cannot be interrupted. For example, if a scaling activity to add 20 ECS instances is being executed, it cannot be forced to terminate when only five instances have been created.
- When a scaling activity fails to add or remove ECS instances to or from a scaling group, the system maintains the integrity of ECS instances rather than the scaling activity. That is, the system rolls back ECS instances, not the scaling activity. For example, if the system has created 20 ECS instances for the scaling group, but only 19 ECS instances are added to the Server Load Balancer instance, the system only releases the failed ECS instance.
- Since Auto Scaling uses Alibaba Cloud’s Resource Access Management (RAM) service to replace ECS instances through ECS API, the rollback ECS instance is still charged.

Cool-down time

- During the cool-down time, only scaling activity requests from CloudMonitor alarm tasks are rejected by the scaling group. Other tasks, such as manually executed scaling rules and scheduled tasks, can immediately trigger scaling activities without waiting for the cool-down time to expire.
- The cool-down time starts after the last ECS instance is added to or removed from the scaling group by a scaling activity.

1.12 Operation procedure

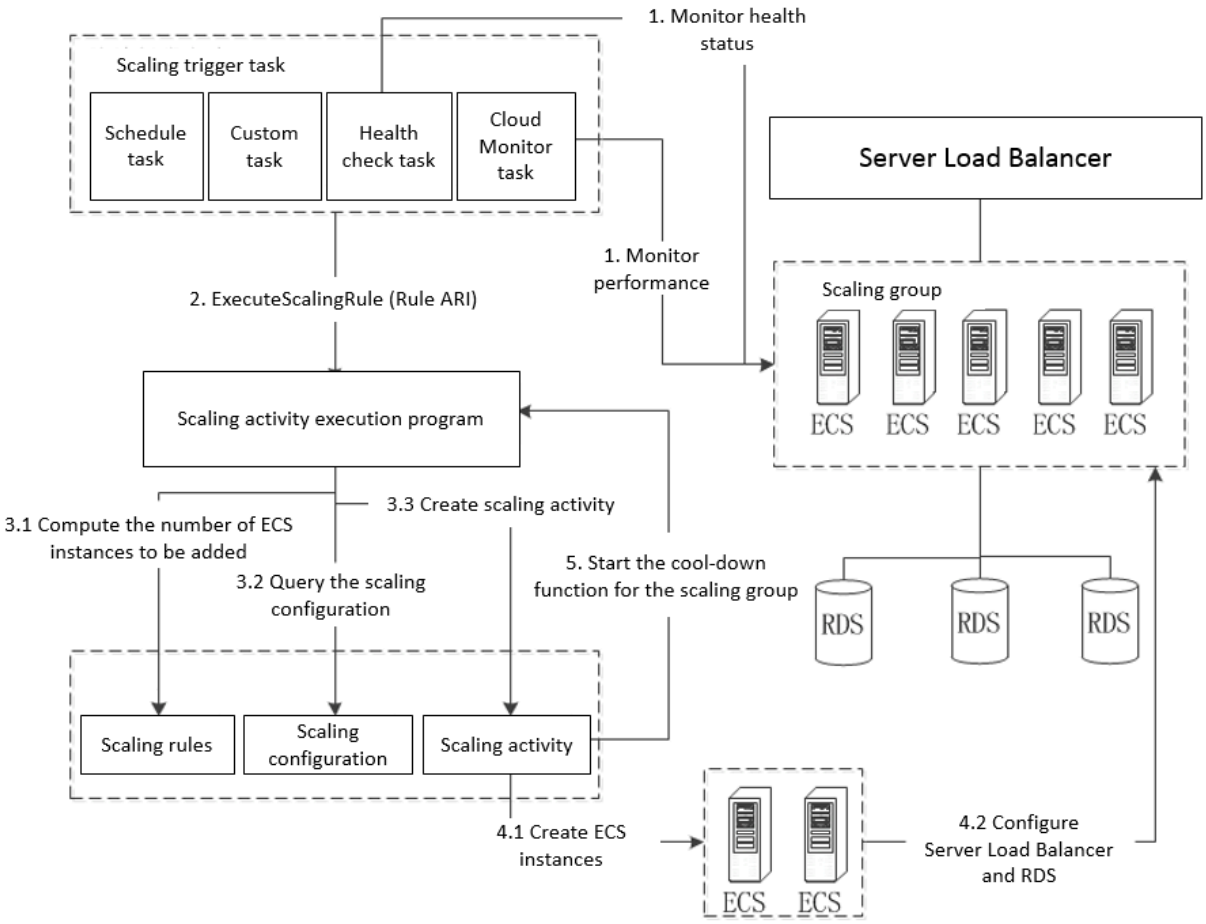
This topic introduces the steps to create a complete Auto Scaling solution.



1. Create a scaling group (`CreateScalingGroup`). Configure the minimum and maximum number of ECS instances in the scaling group, and select the associated Server Load Balancer and RDS instances.
2. Create scaling configuration (`CreateScalingConfiguration`). Configure the ECS instances attributes for Auto Scaling, such as Image ID and Instance Type.
3. Enable the scaling group with the scaling configuration created in Step 2 (`EnableScalingGroup`).
4. Create a scaling rule (`CreateScalingRule`). For example, add N ECS instances.
5. Create a scheduled task (`CreateScheduledTask`). For example, to trigger the scaling rule created in Step 4 at 12:00 AM.
6. Create an alarm task (CloudMonitor API `PutAlarmRule`). For example, to add 1 ECS instance when the average (it can also be max or min) CPU usage is greater than or equal to 80%.

1.13 Workflow

This topic introduces the workflow of Auto Scaling.



After a scaling group, scaling configuration, scaling rule, and scaling trigger task are created, the system executes the process as follows (in this example, we add ECS instances):

1. The scheduled task triggers the request for executing a scaling rule at the specified time.
   - The CloudMonitor task monitors the performance of ECS instances in the scaling group in real time and triggers the request for executing a scaling rule based on the configured alarm rules. For example, when the average CPU usage of all ECS instances in the scaling group exceeds 60%.
   - The task triggers a scaling activity according to the trigger condition.
   - The custom task triggers the request for executing a scaling rule based on the monitoring system and alarm rules. For example, the number of online users or the job queue.

- Health check tasks regularly check the health status of the scaling group and its ECS instances. If an ECS instance is found to be unhealthy (not in the Running status), the health check task triggers a request to remove the ECS instance from the group.

2. The system triggers a scaling activity through the ExecuteScalingRule interface and specifies the scaling rule to be executed by its unique Alibaba Cloud resource identifier (ARI) in this interface.

   If a custom task needs to be executed, you must have the ExecuteScalingRule interface called in your program.

3. The system obtains information about the scaling rule, scaling group, and scaling configuration based on the scaling rule ARI entered in Step 2 and creates a scaling activity.

   a. The system uses the scaling rule ARI to query the scaling rule and scaling group , computes the number of ECS instances to be added, and configures the Server Load Balancer and RDS instances.

   b. According to the scaling group, the system queries the scaling configuration to determine the correct parameters (CPU, memory, bandwidth) to use when creating new ECS instances.

   c. The system creates scaling activity based on the number of ECS instances to be added, the ECS instance configuration, and the Server Load Balancer and RDS instance configurations.

4. During the scaling activity, the system creates ECS instances and configures Server Load Balancer and RDS instances.

   a. The system creates the specified number of ECS instances based on the instance configuration.

   b. The system adds the intranet IP addresses of the created ECS instances to the whitelist of the specified RDS instance and adds the created ECS instances to the specified Server Load Balancer instance.

5. After the scaling activity is completed, the system starts the cool-down function for the scaling group. The cool-down time must elapse before the scaling group can execute any new scaling activity.

2 Scaling configurations

2.1 ECS instance templates

Auto Scaling can automatically create ECS instances based on preconfigured templates and add them to scaling groups as the needs of your business grow. Currently, ECS instance templates include the following two types: *custom scaling configurations* and *launch templates*.

Custom scaling configurations

Scaling configurations allow you to create ECS instance launch templates for scaling groups. When creating a scaling configuration, you can set ECS instance parameters, such as instance types, image types, storage size, and the SSH key pairs that are used to log on to the ECS instances. You can also modify an existing scaling configuration to meet your business needs.

 Note:
A scaling configuration requires a scaling group. You need to create at least one scaling group before creating a scaling configuration. There is a limit to the number of scaling configurations that you can create in a scaling group. For more information, see *Quantity restrictions*.

Currently, you can perform the following operations on a scaling configuration:

- *Create a scaling configuration*
- *View a scaling configuration*
- *Delete scaling configuration*

Launch templates

Launch templates are a tested feature of ECS. You can use existing launch templates to configure scaling groups. For more information, see *Launch templates*.

 Note:
Launch templates allow you to immediately enable a scaling group after it is created.

Differences between custom scaling configurations and launch templates

Item	Custom scaling configuration	Launch template
Parameters supported by scaling groups	All parameters of custom scaling configurations.	Certain parameters of launch templates are not supported. Instances created by launch templates may have missing configurations.
Parameter verification	Supported. You cannot create a custom scaling configuration when required parameters such as image type are missing. In this case, ECS instance creation failures will not occur.	Launch templates do not verify the parameters. All parameters are optional. Therefore, instance may fail to be created if required parameters such as image type are not specified.
Configuration procedure	You must first create a scaling group.	You do not need to create a scaling group.
Creation method	Can only be created in a scaling group.	Can be created on the buy page of ECS, the Launch Templates page, and the Instance Settings page.
Modification	You must manually modify scaling configurations. All modifications are irreversible. However, you can create multiple scaling configurations based on different needs.	Cannot be modified. You can create templates based on your needs.
Multiple instance types	Supported. Applied in scenarios where performance rather than a specified instance type is required, with a bigger chance of successfully scaling out.	Unsupported.

For more information, see *Use custom scaling configurations to create scaling groups* and *Use launch templates to create scaling groups*.

2.2 Create a scaling configuration

This topic describes how to create a scaling configuration.

The process of creating a scaling configuration is similar to creating an ECS instance, but some configuration items are not supported, such as region. Follow the console to perform operations. Each configuration on the page is provided with a brief description. For more information about the meaning and usage, see *create an instance by using the wizard*.

Follow these steps to create a scaling configuration:

1. Log on to the *Auto Scaling console*, and click Manage in the Actions column.
2. Go to the Scaling Configuration page, and click Create Scaling Configuration.
3. On the Basic Configuration page, configure the billing method, instance, image, storage, public network bandwidth, and security group, and then click Next: System Configuration.

    Note:
   Auto Scaling only supports *Pay-As-You-Go instances* and *preemptible instances*, and only public images, custom images, and shared images are supported.
4. On the System Configuration page, configure the tab (optional), logon credential, instance name (optional), and advanced options (optional), and then click Confirm.

    Note:
   Advanced options include instance RAM roles and user-defined data. You can configure the advanced options only for VPC-connected scaling groups.
5. On the Confirm page, check the selected configurations, enter the scaling configuration names, and then click Create.
6. In the Created successfully dialog box, click Enable Configuration. If you do not want to use the scaling configuration now, you can close the dialog box.

After successful creation, you can view and select the scaling configurations in the scaling configuration list.

2.3 View a scaling configuration

Scaling configurations have two life cycle status. Active: The scaling group uses the scaling configuration in active status to create ECS instances. Inactive: Inactive scaling configurations are still in a scaling group, but are not used to create ECS instances. You can select a scale configuration according to your business needs. The scaling group automatically creates ECS instances only based on the Active scaling configuration.

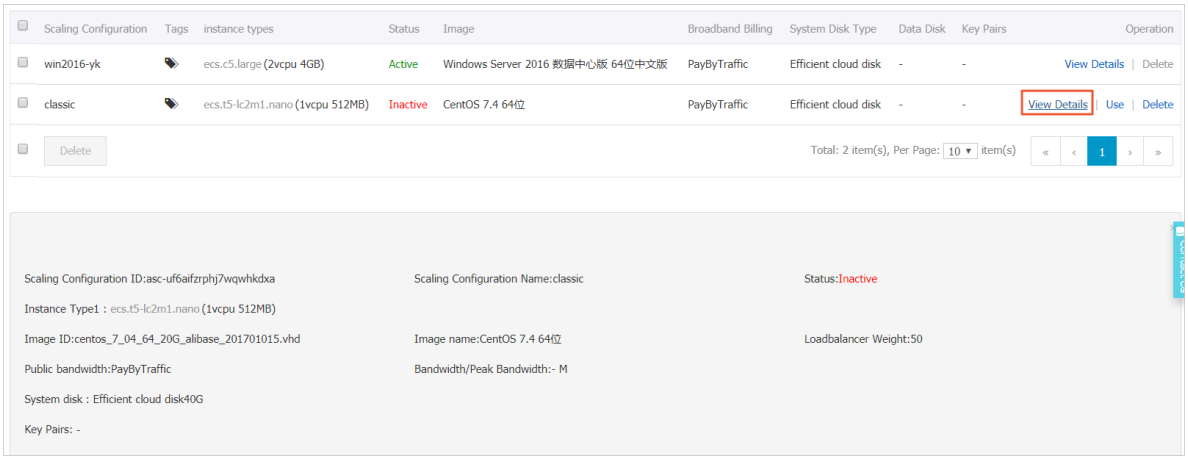
Context

The scaling configuration include much content, and you may forget the specific configuration items. Therefore, Auto Scaling also provides the function to view the details of the scaling configurations for you to learn each scaling configuration at any time, and to select a template for the ECS instance.

Follow these steps to view the details of a scaling configuration and select a scaling configuration:

Procedure

1. Log on to the *Auto Scaling console*, and click Manage in the Actions column after a scaling group.
2. Go to the Scale Configuration page, and click View Details in the Actions column after a specified scaling configuration.



3. After you confirm the configuration, click Select to enable the scaling configuration.

Note:

After you select a scaling configuration, the other scaling configurations are in the Inactive status.

Result

After you select a scaling configuration, when the scale-up conditions are met, the corresponding scaling group automatically creates ECS instances based on this scaling configuration.

2.4 Delete scaling configuration

You can delete scaling configuration.

Prerequisites

Before you delete a scale configuration, make sure the following conditions are met, otherwise, the deletion fails:

- The scaling configuration to be deleted is in the Inactive status.
- If any ECS instances still used for a scaling group and are created according to the scaling configuration, the scaling configuration cannot be deleted.

Procedure

1. Log on to the *Auto Scaling* console, and click Manage in the Actions column.
2. Go to the Scaling Configuration page, and click Delete in the Actions column after the specified scaling configuration.

   You can also check the box at the left of the scale configuration, and then click Delete to delete multiple scaling configurations.
3. In the Delete Scaling Configuration dialog box, click OK.

3 Use custom scaling configurations to create scaling groups

You must create a scaling group before using Auto Scaling to reallocate resources.

A scaling group is a collection of ECS instances that are applied to the same scenario . You can set scaling group parameters, such as the maximum and minimum number of instances and cooldown time. You can also associate the ECS instances with SLB instances and RDS instances for easy management.

> Note:
> There is a limit to the number of scaling groups that you can create under one account. For more information, see *Quantity restrictions*.

For information about using launch templates to create scaling groups, see *Use launch templates to create scaling groups*.

Procedure

1. Log on to the *Auto Scaling console*.
2. On the Scaling Groups page, click Create Scaling Group.
3. Configure the scaling group.
   a. Enter a name in the Scaling Group Name field.
   b. Enter a number in the Maximum Instances field.

      > Note:
      > When the number of ECS instances exceeds the upper limit, Auto Scaling automatically removes instances to make the number of instances in the scaling group match the upper limit.

   c. Enter a number in the Minimum Instances field.

      > Note:
      > When the number of ECS instances drops below the lower limit, Auto Scaling automatically adds instances to make the number of instances in the scaling group match the lower limit.

   d. Enter a number in the Default Cooldown Time (Seconds) field.

> **Note:**
> This parameter specifies the cooldown time of a scaling activity. For more information, see *Cool-down time*.

e. Specify the Removal Policy .

> **Note:**
> This parameter specifies the policy for removing ECS instances when the number of instances in the scaling group exceeds the upper limit. For more information, see *Removal policies*.

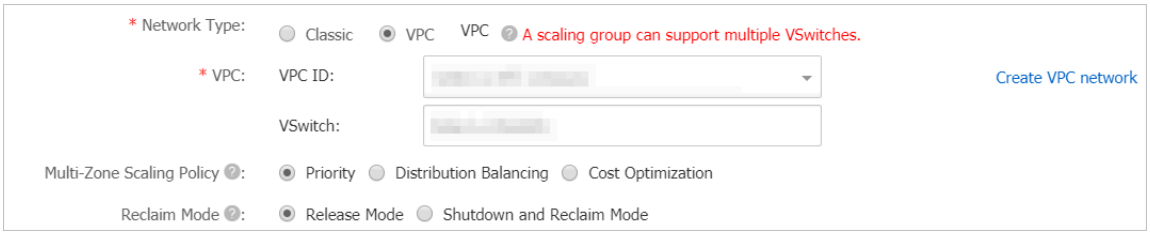
f. Select an Instance Configuration Source. This example uses Custom Scaling Configuration.

g. Select a Network Type. You must set the following parameters if you select VPC:

   A. Specify a VPC ID and a VSwitch.

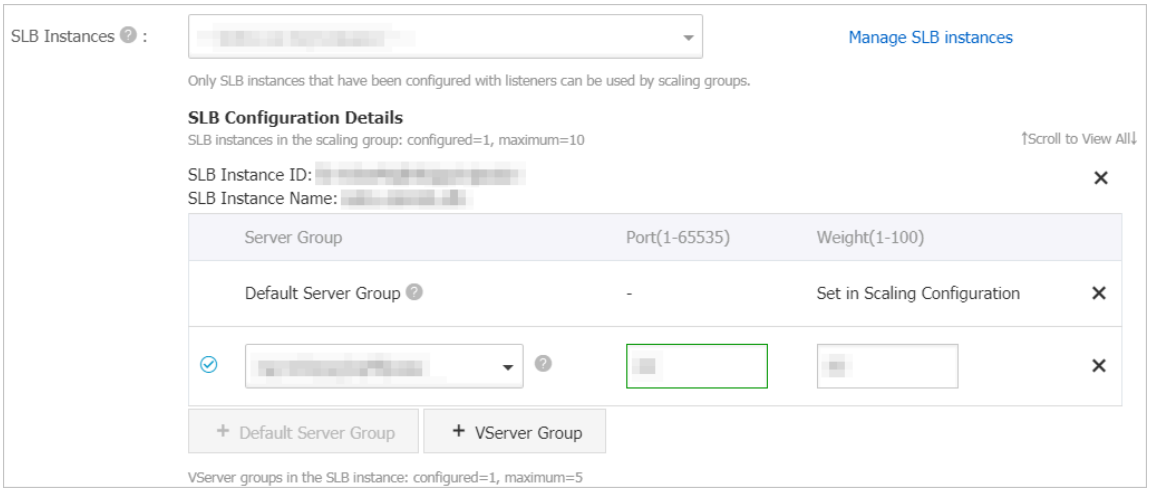
   B. Specify the Multi-Zone Scaling Policy. For more information, see *Multi-zone scaling policy*.

   C. Specify the Reclaim Mode. For more information, see *Reclaim mode*.



h. (Optional) Click SLB Instances to associate the scaling group with SLB instances.

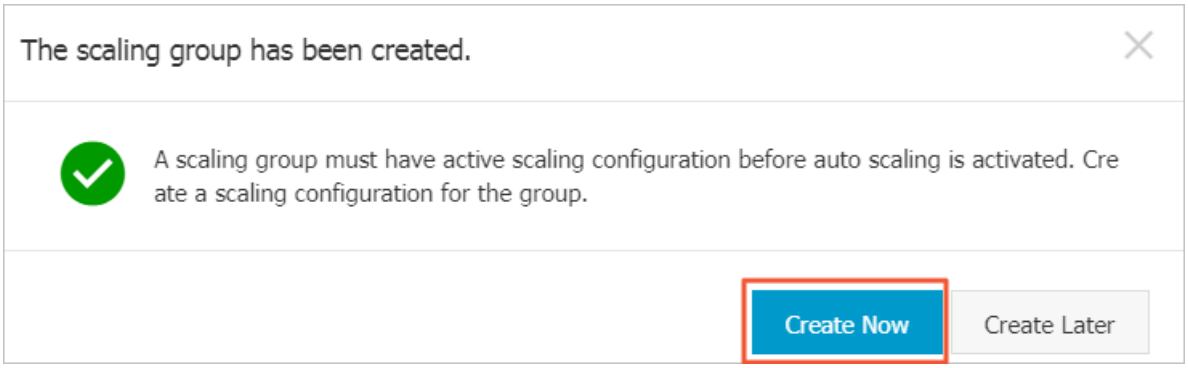
> **Note:**
> A scaling group can be associated with up to five SLB instances at the same time. You can also select the *default server group* or *VServer group(s)* of a SLB instance for the scaling group. You can select up to five VServer groups for a scaling group at the same time. For more information, see *Use Server Load Balancer (SLB) in Auto Scaling*.



i. (Optional) Add RDS Instances. Currently, only RDS databases are supported.

> **Note:**
> You can only add RDS instances in the region where the scaling group is created. Auto Scaling automatically adds the internal IP addresses of the newly created ECS instances to the whitelist of the RDS instances to allow communication between the ECS and RDS instances.

4. Click OK.
5. Click Create Now to create an ECS instance template that is used to create new ECS instances.



> **Note:**
> For more information about scaling configurations, see *Create a scaling configuration*.

6. In the Apply Scaling Configuration dialog box that appears, click Confirm.

Multi-zone scaling policy

Policy name	Description
Priority	Scales out ECS instances based on the specified VSwitch. This policy allows Auto Scaling to use a secondary VSwitch to create ECS instances when the primary VSwitch cannot create ECS instance in its region.
Distribution balancing	Evenly distributes ECS instances in the specified zones when multiple VSwitches are specified. You can reallocate ECS instances to make them evenly distributed when the ECS instances are unevenly distributed in the zone due to certain issues such as insufficient ECS resources. Note: This policy only takes effect when you have specified multiple VSwitches.
Cost optimization	This policy has the following benefits when the network type of the scaling group is VPC: · Ensures business stability when preemptible instances are selected for the scaling configuration. · Reduces costs when multiple instance types are selected for the scaling configuration. Creates an instance based on vCPU billing rates. Instances with the lowest vCPU rate are given priority. · Creates the specified type of preemptible instance when multiple preemptible instance types are specified for the scaling configuration. · Automatically creates Pay-As-You-Go instances when all types of preemptible instances are unavailable.

Reclaim mode

Name	Description
Release Mode	Automatically releases ECS instances based on your scheduled tasks or event-triggered tasks during a scale in event. Creates new ECS instances and adds them to the scaling group based on your scheduled tasks or event-triggered tasks during a scale out event.

Name	Description
Shutdown and Reclaim Mode	Increases scaling efficiency: · Automatically created ECS instances will be stopped during a scale in event. CPUs and memory of stopped instances will not be billed. However, cloud disks including system disks and data disks, EIP addresses, and bandwidth will still be billed. Public IP addresses will be reclaimed and then reassigned when the ECS instances are restarted. EIP addresses will be reserved. All stopped ECS instances are added to an instance pool. · ECS instances in the stopped instance pool will be restarted first when a scale out event starts. If the number of these instances is insufficient, Auto Scaling creates new instances. Note: · This mode is supported only by scaling groups that consist of VPC-connected ECS instances. · This mode is not supported by instances with local disks, such as d1, d1ne, ga1, gn5, i1, and i2 instances. · We cannot guarantee that all stopped instances in the stopped instance pool can be successfully restarted when Auto Scaling scales out instances. The stopped ECS instances will be released if they cannot be restarted. Auto Scaling creates new ECS instances to ensure that the scaling activity is successful. · You cannot modify a scaling group when it is set to the shutdown and reclaim mode.

4 Use launch templates to create scaling groups

You must create a scaling group before using Auto Scaling to reallocate resources.

A scaling group is a collection of ECS instances that are applied to the same scenario . You can set scaling group parameters, such as the maximum and minimum number of instances and cooldown time. You can also associate the ECS instances with SLB instances and RDS instances for easy management.

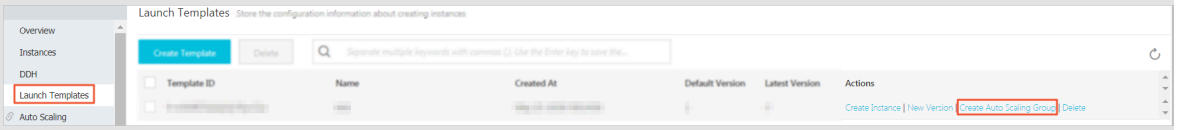
> **Note:**
> There is a limit to the number of scaling groups that you can create under one account. For more information, see *Quantity restrictions*.

For information about using custom scaling configurations to create scaling groups, see *Use custom scaling configurations to create scaling groups*.

Procedure

1. Log on to the *Auto Scaling console*.
2. On the Scaling Groups page, click Create Scaling Group.

   > **Note:**
   > You can also create scaling groups on the Launch Templates page in the *ECS console*.

3. Set scaling group parameters.
   a. Enter a name in the Scaling Group Name field.
   b. Enter a number in the Maximum Instances field.

      > **Note:**
      > When the number of ECS instances exceeds the upper limit, Auto Scaling automatically removes instances to make the number of instances in the scaling group match the upper limit.

   c. Enter a number in the Minimum Instances field.

      > **Note:**

When the number of ECS instances drops below the lower limit, Auto Scaling automatically adds instances to make the number of instances in the scaling group match the lower limit.

d. Enter a number in the Default Cooldown Time (Seconds) field.

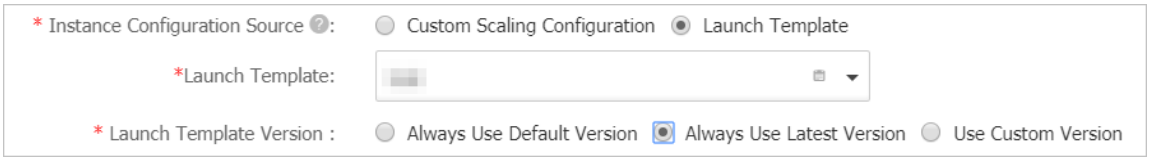
Note:
This parameter specifies the default cooldown time of a scaling activity . For more information, see *Cool-down time*.

e. Specify the Removal Policy.

Note:
This parameter specifies the policy for removing ECS instances when the number of instances in the scaling group exceeds the upper limit. For more information, see *Removal policies*.

f. Select an Instance Configuration Source. This example uses the Launch Template.

Note:
You can manage launch template versions. For more information, see *Manage launch template versions*.



g. Select a Network Type. You must set the following parameters if you select VPC:

A. Specify a VPC ID and a VSwitch.

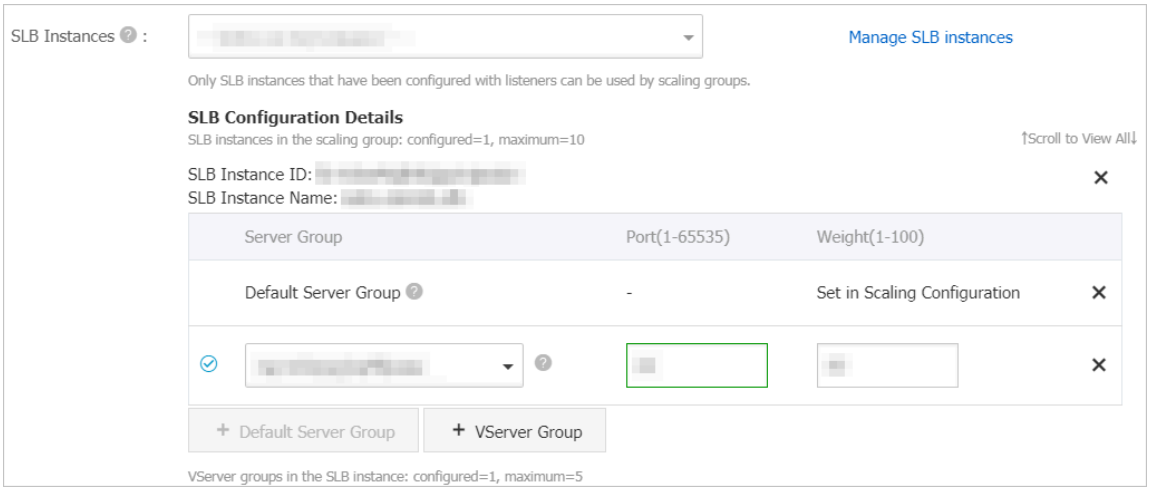
B. Specify the Multi-Zone Scaling Policy. For more information, see *Multi-zone scaling policy*.

C. Specify the Reclaim Mode. For more information, see *Reclaim mode*.

h. (Optional) Click SLB Instances to associate the scaling group with SLB instances.

Note:
A scaling group can be associated with up to five SLB instances at the same time. You can also select the *default server group* or *VServer group(s)* of a SLB instance for the scaling group. You can select up to five VServer groups for a scaling

group at the same time. For more information, see *Use Server Load Balancer (SLB) in Auto Scaling*.

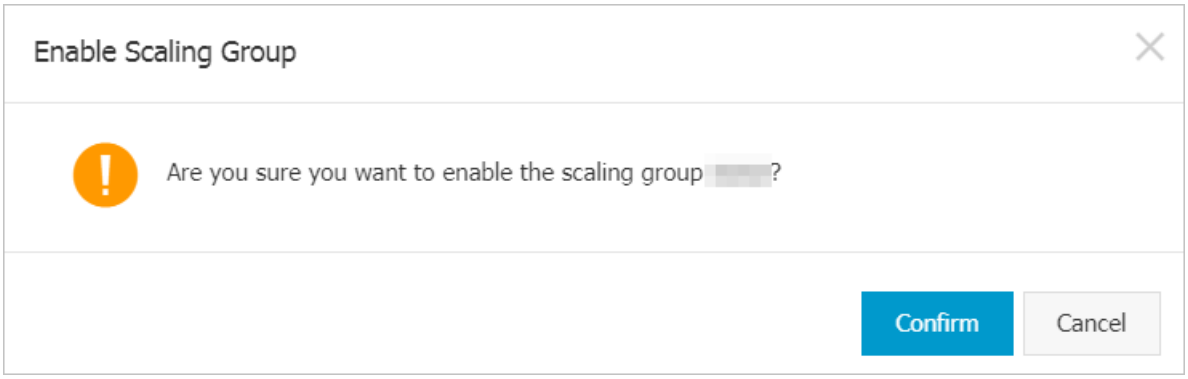


i. (Optional) Add RDS Instances. Currently, only RDS databases are supported.

> **Note:**
> You can add RDS instances only in the region where the scaling group is created. Auto Scaling automatically adds the internal IP addresses of the newly created ECS instances to the whitelist of the RDS instances to allow communication between the ECS and RDS instances.

4. Click OK.
5. In the Enable Scaling Group dialog box that appears, click Confirm.

> **Note:**
> After you have created the scaling group, you cannot modify the network configuration, including the VPC and VSwitch. Keep the network configuration of the launch template consistent with that of the scaling group when changing the version of the launch template, or the version modification will fail.



Manage launch template versions

Management policy	Description
Always Use Default Version	This policy requires the scaling group to always use the default launch template to create ECS instances.
Always Use Latest Version	This policy requires the scaling group to always use the latest launch template to create ECS instances.
Use Custom Version	This policy requires the scaling group to use a specified launch template version to create ECS instances.  Note: Note: If this policy is selected, the scaling group automatically sets the network configuration based on the network configuration of the launch template.

Multi-zone scaling policy

Policy	Description
Priority	Add or remove ECS instances based on the specified VSwitch. This policy allows Auto Scaling to create ECS instances in the zone of the secondary VSwitch when Auto Scaling fails to create ECS instances in the zone of the primary VSwitch.
Distribution Balancing	Evenly distributes ECS instances in the specified zones when multiple VSwitches are specified. You can reallocate ECS instances to make them evenly distributed when the ECS instances are unevenly distributed in the zone due to certain issues such as insufficient ECS resources.  Note: This policy takes effect only when you have specified multiple VSwitches.

Policy	Description
Cost Optimization	This policy has the following benefits when the network type of the scaling group is VPC: · Ensures business stability when preemptible instances are selected for the scaling configuration. · Reduces costs when multiple instance types are selected for the scaling configuration. Creates an instance based on vCPU billing rates. Instances with the lowest vCPU rate are given priority. · Creates preemptible instances of specified types first when both preemptible instances and Pay-As-You-Go instances are available. · Automatically creates Pay-As-You-Go instances when all types of preemptible instances are unavailable.

Reclaim modes

Mode	Description
Release Mode	Automatically releases ECS instances based on your scheduled tasks or event-triggered tasks during a scale in event. Creates new ECS instances and adds them to the scaling group based on your scheduled tasks or event-triggered tasks during a scale out event.

Mode	Description
Shutdown and Reclaim Mode	Increases scaling efficiency: · Automatically created ECS instances will be stopped during a scale in event. CPUs and memory of stopped instances will not be billed. However, cloud disks including system disks and data disks, EIP addresses, and bandwidth will still be billed. Public IP addresses will be reclaimed and then reassigned when the ECS instances are restarted. EIP addresses will be reserved. All stopped ECS instances are added to an instance pool. · ECS instances in the stopped instance pool will be restarted first when a scale out event starts. If the number of these instances is insufficient, Auto Scaling creates new instances.  Note: · This mode is supported only by scaling groups that consist of VPC-connected ECS instances. · This mode is not supported by instances with local disks, such as d1, d1ne, ga1, gn5, i1, and i2 instances. · We cannot guarantee that all stopped instances in the stopped instance pool can be successfully restarted during a scale out event. The stopped ECS instances will be released if they cannot be restarted. Auto Scaling creates new ECS instances to ensure that the scaling activity is successful. · You cannot modify a scaling group when it is set to the shutdown and reclaim mode.

5 Realize Auto Scaling

5.1 Create a scaling rule

This topic introduces the definition and creation steps of scaling rules.

After *creating a scaling group*, you successfully enable a scaling group. If you want to scale up or scale down ECS resources, you must create scaling rules.

What is a scaling rule

A scaling rule defines specific scaling actions; for example, adding or removing ECS instances. Currently, the following three scaling rules are supported:

- Change to N instances: After you perform the scaling rule, the number of instances in service is changed to N.
- Add N instances: After you perform the scaling rule, the number of instances in service increases by N.
- Decrease N instances: After you perform the scaling rule, the number of instances in service is reduced by N.

 Note:
The number of scaling rules that can be created within a scaling group is limited. See *quantity restrictions*.

After you perform a scaling rule, if the actual number of instances in service in the scaling group is greater than the MaxSize or less than the MinSize of instances, Auto Scaling automatically changes the number of instances to ensure that the scaling result does not exceed the limit.

Examples

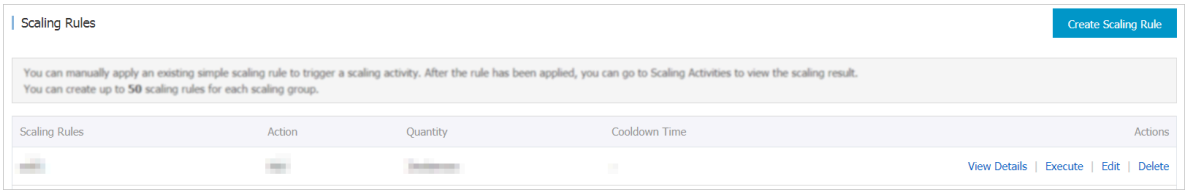
- You have a scaling group asg-bp19ik2u5w7esjcucu28. The MaxSize is three, and the scaling rule add3 is to add three instances. If the current number of instances in service is two, when you perform the scaling rule add3, only 1 ECS instance is added.
- You have a scaling group asg-bp19ik2u5w7esjcucu28. The MinSize is two, and the scaling rule reduce2 is to reduces two instances. If the current number of instances

in service is three, when you perform the scaling rule reduce2, only 1 ECS instance is reduced.

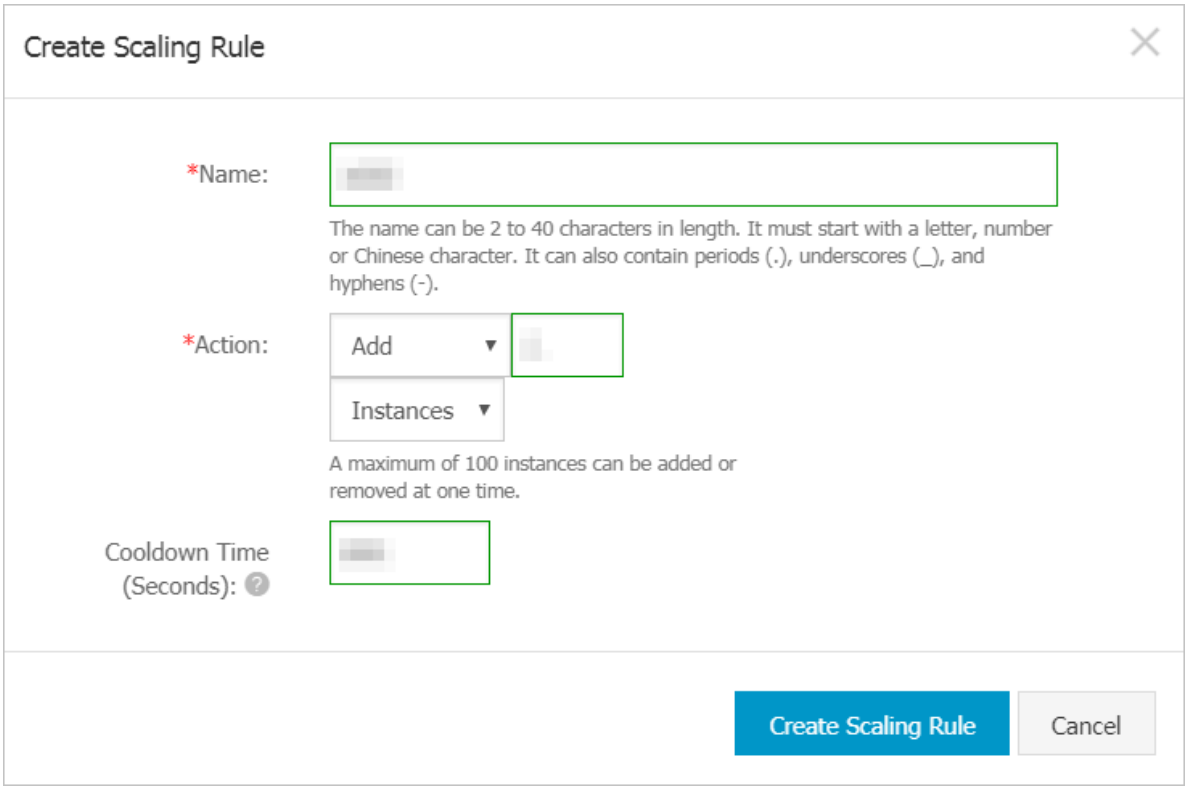
Procedure

Following these steps to create a scaling rule:

1. On the Scaling Groups page, click Manage in the Actions list next to the target scaling group.
2. Go to the Scaling Rules page, click Create Scaling Rule.



3. In the Create Scaling Rule dialog box, specify the rule name, rule, and the cool-down time, and then click Create Scaling Rule.



Note:

The cool-down time is an optional item. If you leave it empty, the cool-down time of the scaling group applies by default.

5.2 Execute a scaling rule

This topic introduces how to perform a scaling rule.

After *creating a scaling rule*, you have successfully created a scaling rule, and then you can perform a scaling rule to scale up or scale down ECS resources.

Limits

If you need to perform a scaling rule, note the following:

- The status of the scaling group which the scaling rule belongs to must be `Enabled`.
- No scaling activities in progress are in the scaling group which the scaling rule belongs to.
- For all regions and all scaling groups, the number of ECS instances to be scaled for each account is limited. See *quantity restrictions*.

Currently, you can perform a scaling rule in three ways:

- *Through scheduled tasks*.
- *Through alarm tasks*.
- *By performing manually*.

Through scheduled tasks

Select a scaling rule when you *create a scheduled task*. Auto Scaling automatically performs the scaling rule at the specified time point.

Through alarm tasks

Select an alarm trigger rule when you create an alarm task. Auto Scaling automatica lly performs the scaling rule when the alarm is triggered.

By performing manually

When no scaling activities in progress are in the scaling group, you can skip *cool-down time* by performing the scaling rule manually. Follow these steps to manually perform a scaling rule:

1. In the Scaling Rules page, click Perform in the Actions column.
2. In the Perform Scaling Rule dialog box, click OK.
3. If the scaling rule is performed successfully, a prompt appears in the upper-right corner of the page.

If the scaling rule fails to be performed, an error prompt appears.

4. You can go to the Scaling Activities page to view the result of scaling rule performing.

5.3 Create a lifecycle hook

This topic describes the definition of lifecycle hook and how to create a lifecycle hook.

You have followed the steps described in *Execute a scaling rule* to execute scaling rules to scale ECS instances. However, these ECS instances are only configured with basic settings. To use these instances in complex business, you may need to perform custom actions before enabling these ECS instances. To complete this task, you can use lifecycle hooks.

What is lifecycle hook

You can create lifecycle hooks for a scaling group. When a scaling group with lifecycle hooks performs scaling activities, the instances to be added to or removed from the scaling group will be put to the Wait status. Lifecycle hooks only take effect on ECS instances that are automatically added to or removed from the scaling groups. Manually added or removed ECS instances are not affected.

 Note:
The maximum number of lifecycle hooks that you can create for a scaling group is limited. For more information, see *Restrictions*.

Examples

For example, you have created scaling group sg-yk201808201449. The minimum number of instances that the scaling group must contain is 0. The scaling group has one lifecycle hook for scaling activities. Currently, the scaling group does not have any ECS instances.

Change the minimum number of instances to 1 for the scaling group. After the modification, a scaling activity is triggered because the number of instances that the scaling group contains does not meet the minimum requirement. An ECS instance is then automatically added to the scaling group. However, since the scaling group has a lifecycle hook, the status of the ECS instance will not change to InService immediately. Instead, its status changes to Adding:Wait.

During the lifecycle hook timeout period, you can log on to the ECS instance, and install applications or perform custom actions.

Features

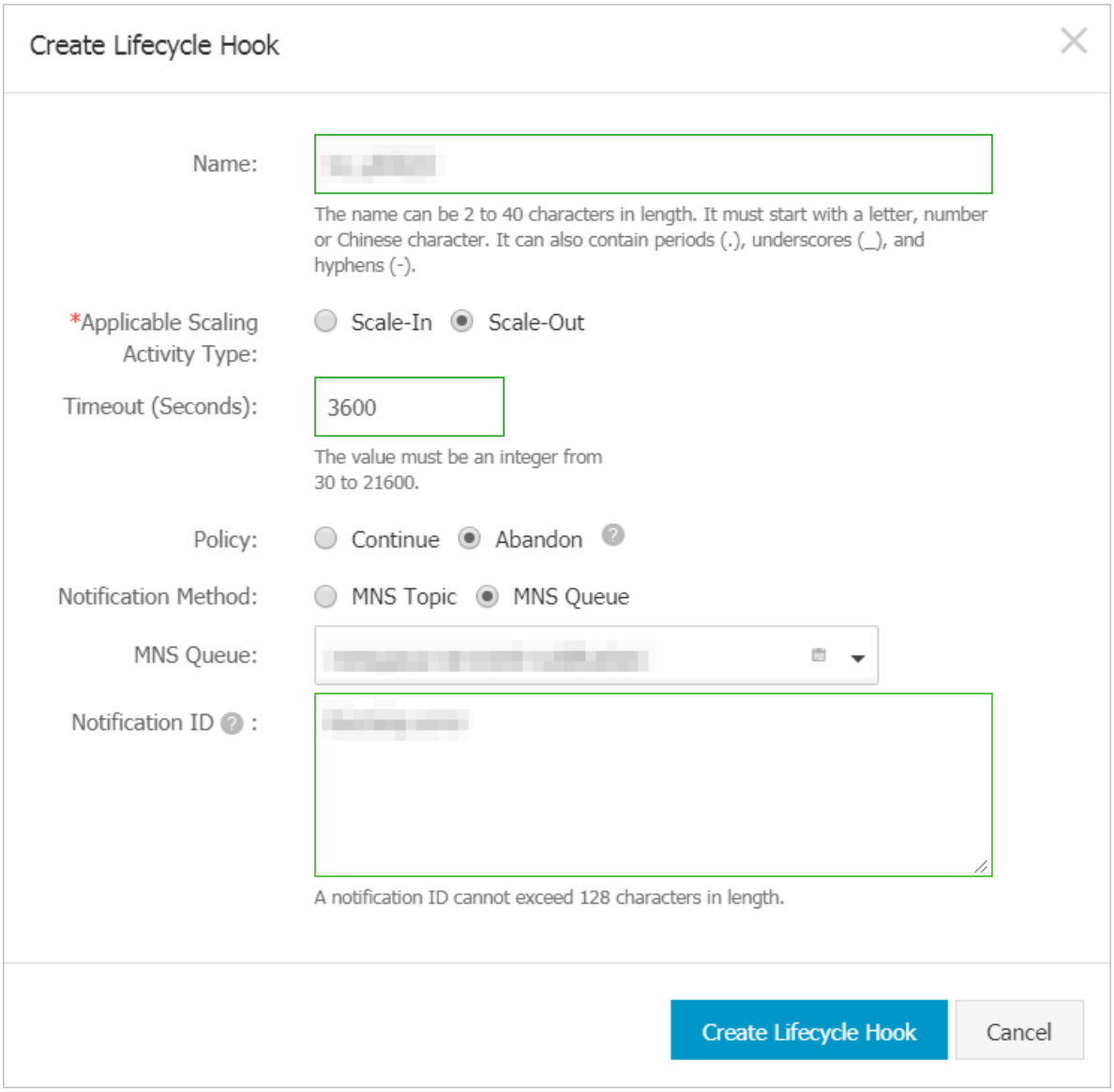
The scaling group has the following features while its ECS instance is put into the Wait status by the lifecycle hook:

- The scaling group does not perform other scaling activities.
- You can perform custom actions during the lifecycle hook timeout period. For example, you can initialize the configuration of the ECS instance or obtain the ECS instance data.
- You can delete the corresponding lifecycle hook to resume the scaling activity.
- You can also call the *CompleteLifecycleAction* or *DeleteLifecycleHook* interface to resume the scaling activity.

Procedure

Follow these steps to create a lifecycle hook:

1. Log on to the *Auto Scaling console*.
2. On the Scale Groups page, click Manage in the Actions column for the target scaling group.
3. Go to the Lifecycle Hooks page, click Create Lifecycle Hook.
4. In the Create Lifecycle Hook dialog box, set the Name, Applicable Scaling Activity Type, Timeout, Policy, Notification Method, MNS Topic/Queue, and Notification ID then click Create Lifecycle Hook.



Lifecycle hook properties

The following table describes the lifecycle hook properties and examples.

Property	Description	Example
Name	The lifecycle hook name must be 2 to 40 characters in length. It must start with an English letter, number, or Chinese character. The name can contain periods (.), underscores (_), and hyphens (-). After you have set the lifecycle name, you can no longer change it.	hz_yk0626

Property	Description	Example
Applicable Scaling Activity Type	The type of scaling activities.	Scale-In
Timeout	The lifecycle hook timeout period. During this period, the instances remain in the Wait status. The value must be an integer from 30 to 21,600 seconds.	600
Policy	Available policies include Continue and Abandon. · Continue: Continues the scaling activity when the current lifecycle action ends. · Abandon: Releases the created ECS instances if the scaling activity type is scale-out. Removes the ECS instances if scaling activity type is scale-in.	Continue
Notification Method	The available notification methods include MNS Topic and MNS Queue. After you select a notification method, you must select the specific MNS topic or queue.	MNS Topic
Notification ID	The notification ID is sent to you with notifications so that you can easily manage the notifications by ID.	General information

5.4 Removal policies

This article introduces the removal policies

There are two types of removal policies: default policy and custom policy.

Default removal policy

This policy first performs level-1 instance screening on the ECS instances created according to the oldest scaling configuration (OldestScalingConfiguration), and then performs level-2 screening on the oldest ECS instances (OldInstances).

- This policy first selects the ECS instances created according to the oldest scaling configuration (OldestScalingConfiguration) of the scaling group, and then selects the oldest ECS instance (OldestInstance) from these ECS instances. If more than one oldest ECS instance is found, one of them is selected at random and removed from the scaling group.
- Manually added ECS instances are not first selected for removal because they are not associated with any scaling configuration.
- If all ECS instances associated with the scaling configuration have been removed , but more instances still need to be removed from the scaling group, this policy selects the instance that was manually added earliest.

Custom release policy

You can set multiple policies to select and remove ECS instances successively from the scaling group.

Release policy types

- OldestInstance: This policy selects the ECS instance that was created earliest. As level-1 screening, the policy selects the earliest ECS instance, either created manually or automatically.
- NewestInstance: This policy selects the ECS instance that was created most recently. As level-1 screening, the policy selects the newest ECS instance, either created manually or automatically.
- OldestScalingConfiguration: This policy selects the instance created according to the oldest scaling configuration and skips over manually added instances. However , if all ECS instances associated with scaling configurations have been removed, but more instances still need to be removed from the scaling group, this policy randomly selects a manually added ECS instance (an instance not associated with any scaling configuration).

6 Scheduled tasks

6.1 Create a scheduled task

This topic introduces the definition and creation of scheduled tasks.

You can create up to 20 scheduled tasks according to input parameters.

What is a scheduled task

A scheduled task is a default task that performs a specified scaling rule at a specified time. Thus, it automatically scales up or scales down the computing resources to meet business needs and control costs. You can also specify the repetition cycle for scheduled tasks to respond to the business changes with flexible rules.

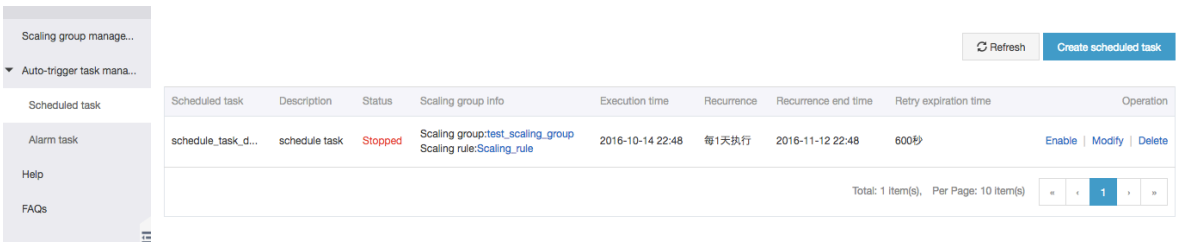
> **Note:**
> The number of scheduled tasks that can be created under one account is limited. See *quantity restrictions*.

Since only one scaling activity can exist in a scaling group at a time, the scheduled task also provides automatic retry function to guarantee the scheduled task results in case of single scaling rule performing failure. If more than one scheduled tasks to be performed exist in one minute, Auto Scaling performs the most recently created scheduled task.

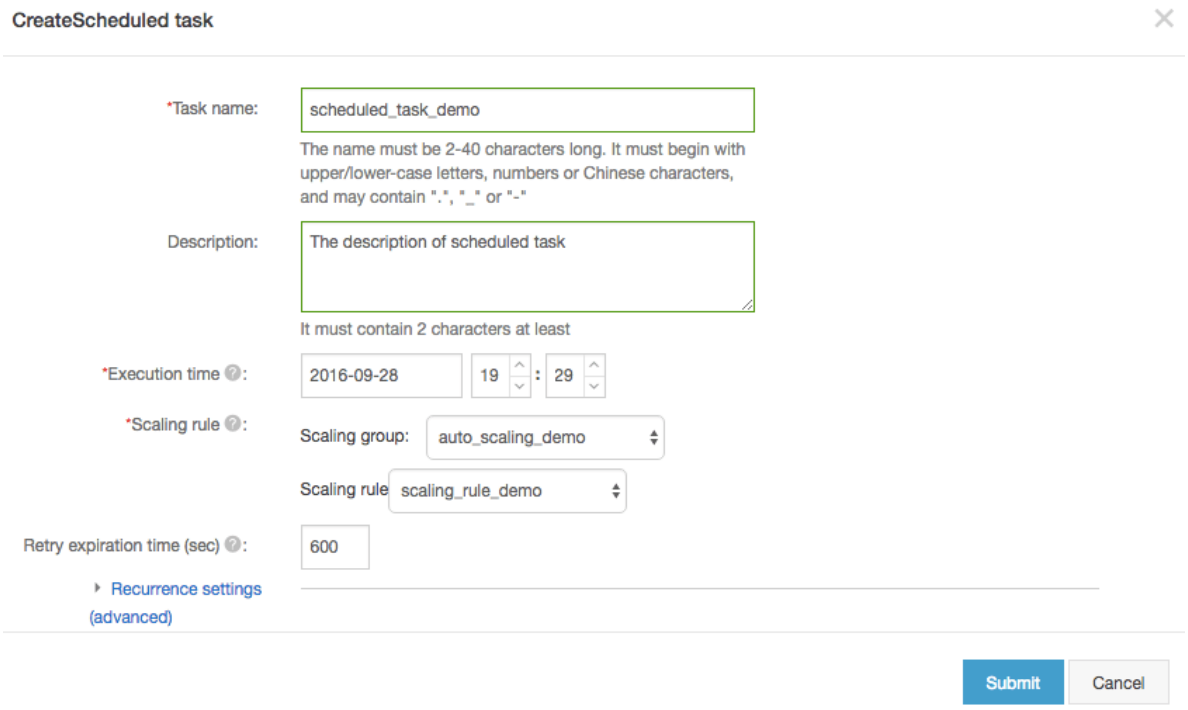
Procedure

Following these steps to create a scheduled task:

1. Log on to the *Auto Scaling console*.
2. Select Auto-Trigger Tasks, go to the Scheduled Tasks page, and click Create Scheduled Task.



3. In the Create Scheduled Task dialog box, specify the task name, time to perform, scaling rule, retry expiration time (optional), and repetition cycle (optional). You can also add a description for later viewing. Click Submit.



Note:

For the attributes of scheduled tasks, see *scheduled task attributes*.

Scheduled task attribute

Name	Description	Example
Task name	The name must consist of 2-40 characters. It must begin with a lower-case letter, number, or a Chinese character. It can contain ".", "_", or "-".	st-yk201808301442
Description	Describes the purpose, function, and other information of the scheduled task.	The PV is large at the beginning of a month. Add three instances.
Time to perform	Time to trigger the scheduled task	00:00, September 2, 2018

Name	Description	Example
Scaling rule	The name of the scaling rule, which specifies the scaling action to perform when the task is triggered.	add3
Retry expiration time	The time range is 0 seconds ~ 21,600 seconds (6 hours). If the scaling action fails to be performed at the specified time, Auto Scaling continues to perform the scheduled task within the retry expiration time.	600
Repetition cycle	The repetition cycle to perform the scheduled task. It can be on a daily, weekly, and monthly basis. If different requirements are needed, you can use the *Cron expressions*.	By month Perform on the second to third day each month.
Repetition ending time	The time to stop repeated performing of the scheduled task	00:00, September 30, 2018

Cron expressions

The Cron expressions use the UTC + 0 time zone. Eight hours should be added when you convert it into the system local time in China. In addition, the time of the first Cron expression performing must be less than the repetition ending time, otherwise, the scheduled task fails to be created.

A Cron expression is a string separated by spaces. It is divided into five to seven fields . Currently, the Auto Scaling scheduled tasks support five-field Cron expressions, incllluding minutes, hours, days, months, and weeks. The range of values are shown in the following table.

Field	Required	Valid values
Minutes	Yes	[0, 59]
Hours	Yes	[0, 23]

Field	Required	Valid values
Days	Yes	[1, 31]
Months	Yes	[1, 12]
Weeks	Yes	[0, 7]; Sunday = 0 or 7

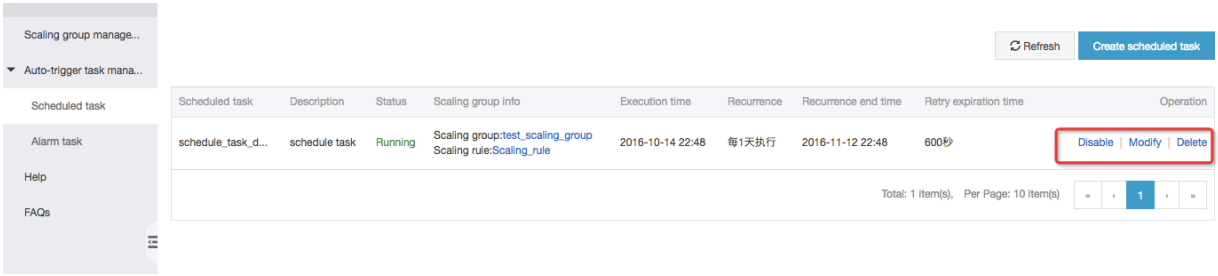
You can enter multiple values in a field:

- Specify multiple values using a comma (,). For example, 1, 3, 4, 7, 8.
- Specify the range of values using "-". For example, 1-6. The result is the same as 1, 2 , 3, 4, 5, 6.
- Specify any possible values using an asterisk (*). For example, an asterisk in the hour field represents each hour, and the result is the same as 0-23.
- Specify the interval frequency using a slash (/). For example, 0-23/2 in the hour field indicates performing every 2 hours. Slashes (/) can be used with asterisks (*). For example, */3 in the hour field indicates performing every 3 hours.

6.2 Modify, disable, or delete a scheduled task

You can modify, disable, or delete a scheduled task.

Modifies the attributes of a scheduled task, queries the details of a scheduled task, or deletes a specified scaling rule.



7 Alarm tasks

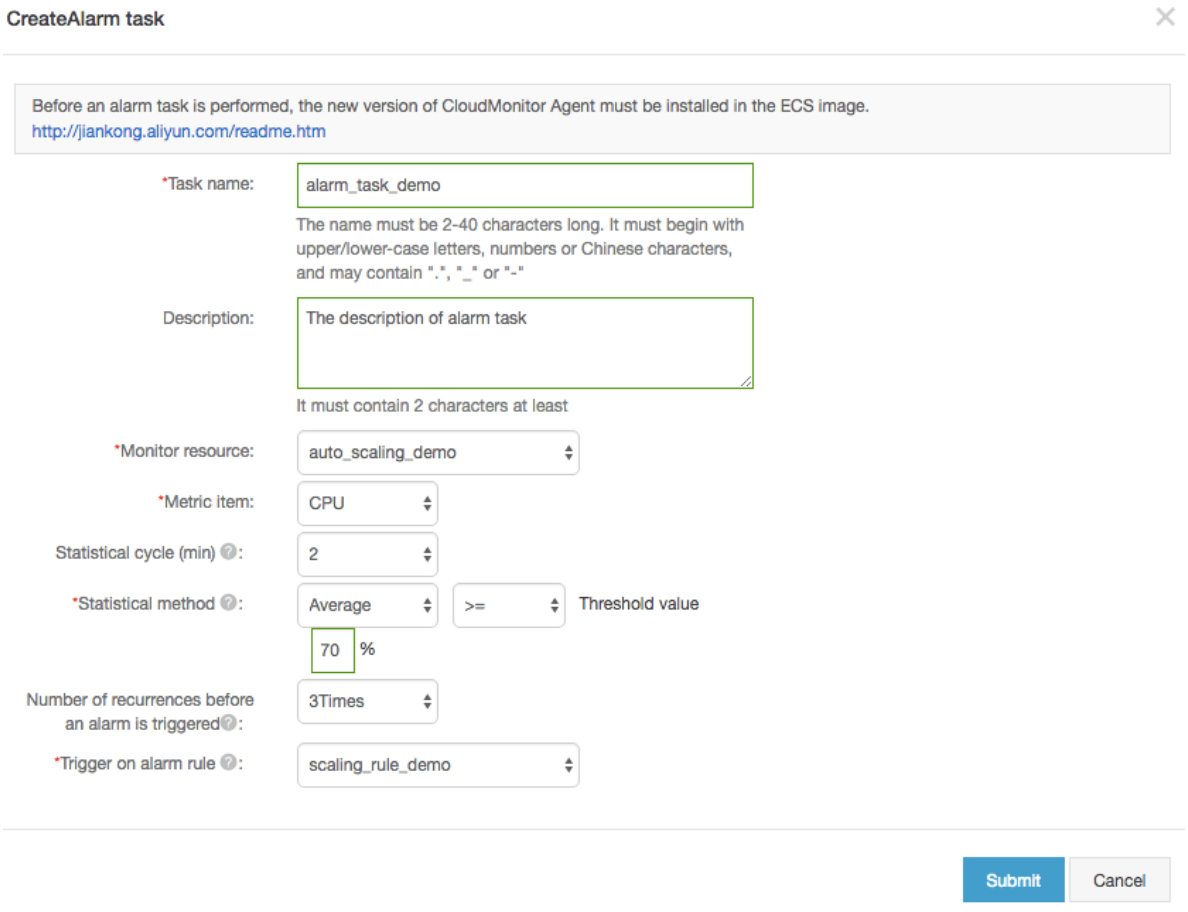
7.1 Create an alarm task

This topic describes how to create an alarm task.

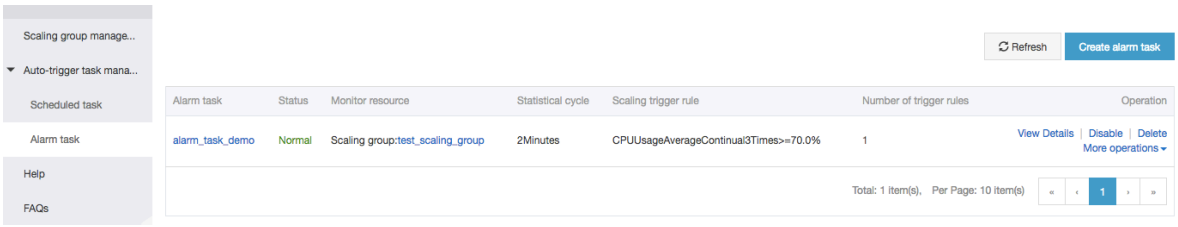
There are currently two types of alarm tasks: System Metric Alarm Task and Custom Metric Alarm Task.

Create a System Metric Alarm Task

1. Log on to the *Auto Scaling console*.
2. Select Auto-Trigger Tasks > Alarm Tasks, and then click Create Alarm Task.



3. In the dialog box, enter the custom information, for example:



The information of the alarm task in the preceding figure is defined as follows:

- test_cpu_alarm is the task name, and cpu utilization is the task description.
- classic is a monitoring resource, that is, the scaling group monitored by the alarm task.
- `System Monitoring` is the monitoring type.
- CPU (CPU utilization) is the metric.
- The data is collected and checked every minute to determine whether the alarm is triggered.
- An average value greater than or equal to 50% is the statistical method, which is repeated three times. It means that when average value of the CPU utilizatio n in one minute exceeds the threshold of 50%, and the statistical methods are satisfied three times consecutively, the alarm is triggered.
- Scaling rule add1 is an alarm trigger rule, indicating that when an alarm is triggered, the alarm rule add1 is performed, that is, one instance is added.

Create a Custom Metric Alarm Task

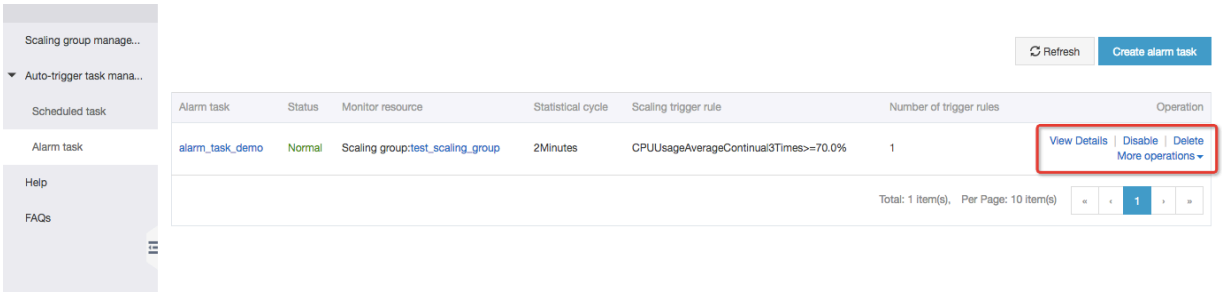
The process of creating a Custom Metric Alarm Task is similar to creating a System Metric Alarm Task. The only difference is that, the metrics of the System Metric Alarm Task are collected by the CloudMonitor for the users, and the Custom Metric Alarm Task requires the users to report the metrics to the CloudMonitor themselves.

When you create a Custom Metric Alarm Task, the custom metrics that have been reported must exist, that is, the Time Sequence. You can then set alarm rules for this Time Sequence.

Before the Custom Metric Alarm Task is created in the preceding figure, a custom monitoring data stream (Time Sequence) has been pushed to the CloudMonitor. The application group which the Time Sequence belongs to is 54504, the metric name is testMetric, and the dimension information is age=10.

7.2 View, modify, or delete an alarm task

This topic describes how to view, modify, and delete alarm tasks.



View the metric details

After successfully creating an alarm task, you can see the alarm task in the alarm task list.

1. Log on to the *Auto Scaling console*.
2. Select System Monitoring to view the system monitoring alarm tasks you created.
3. Select Custom Monitoring to view the custom alarm tasks you created.
4. Click the name of the alarm task to go to the details page, on which you can view the data of the corresponding metrics of the alarm tasks.

Modify alarm tasks

You can modify the alarm tasks on the alarm task list page, and you can also go to the details page of the alarm task to modify the alarm rules.

Modifying an alarm task is divided into two parts: modifying the basic information of the alarm task and modifying the trigger rule for the alarm rule.

Modifying basic information includes modifying the task name, metrics, statistical period, statistical method, times of repetition, and so on. We recommend that you do not modify the metrics of the alarm task, because modifying it means monitoring different indexes. At this time, creating a new alarm task for a new index is a better way.

Delete alarm tasks

You can delete an alarm task in the Actions column on the Alarm Tasks page.

8 Manual scaling

8.1 Add an ECS instance

This topic describes how to add an ECS instance to a scale group.

Prerequisites

If you add an ECS instance manually, the instance must meet the following conditions:

- The ECS instance is in the same region as the scaling group.
- The ECS instance is not in any other scaling group.
- The ECS instance is Running.
- The ECS instance can be the classic type or VPC, but has the following restrictions:
  - If the scaling group is the classic type, only classic type instances can be added.
  - If the scaling group is the VPC type, only instances belonging to the same VPC can be added.

To add an ECS instance, the scaling group must meet the following conditions:

- The scaling group is active.
- The scaling group is not executing any scaling activity.

Note:
When no scaling activity is being executed for the scaling group, adding an ECS instance is executed directly without waiting for the cool-down time. A successful return indicates that the Auto Scaling service will shortly execute the scaling activity, but does not mean that the scaling activity will be successfully executed. Use the returned ScalingActivityID to check the scaling activity status. If the number of ECS instances to be added by the scaling rule plus the number of existing ECS instances in the scaling group (Total Capacity) exceeds the MaxSize value, the operation fails. Manually added ECS instances are not associated with the active scaling configuration in the scaling group. If you have any problem, *open a ticket*.

Context

For detailed ECS instance configuration, see *create a scaling configuration*.

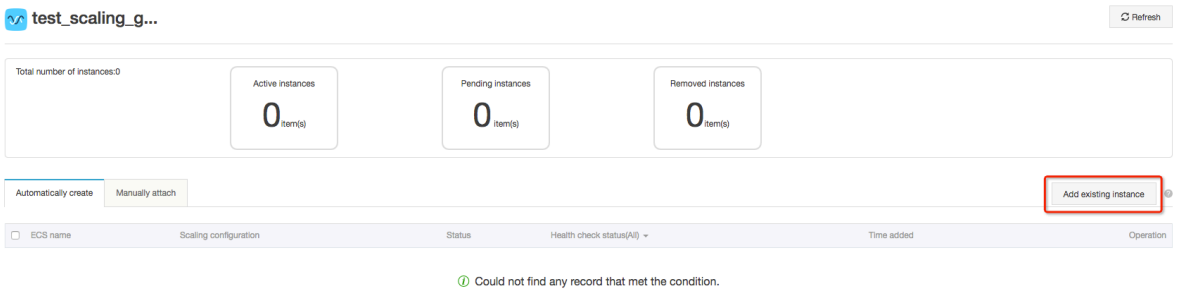
You can add ECS instances in two ways:

- *Perform a scaling rule* to automatically create one or more ECS instances. Automatically created ECS instances automatically meet the current scaling configuration. You do not have to worry about specification limitations.
- Manually add one or more ECS instances. The manually added ECS instance configuration is not associated with the current scaling configuration.

You can skip the *cool-down time* if you add ECS instances manually. The procedure of adding ECS instances manually is shown as follows.

Procedure

1. Log on to the *Auto Scaling console*.
2. On the Scale Groups page, click Manage in the Actions column next to the specified scaling group.
3. Go to the ECS Instances page, click Add Existing Instance.



4. Select the available ECS instances from the list on the left, click > to to the instances to the scaling group, and then click OK.
5. Go to the Manually Add page to view the result.

> **Note:**
> If the page does not refresh automatically, click Refresh in the upper-right corner of the page.

8.2 Remove an ECS instance

You can remove an ECS instance from a specified scaling group.

When an automatically created ECS instance is removed from a scaling group, the instance is stopped and released.

When a manually added ECS instance is removed from a scaling group, the instance is not stopped or released.

The operation will succeed under the following conditions:

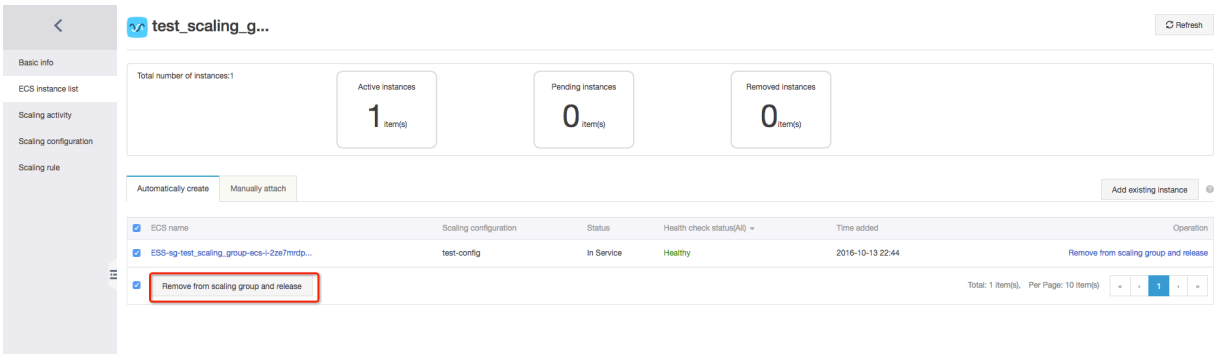
- The scaling group is active.
- The scaling group is not executing any scaling activity.

When no scaling activity is being executed for the scaling group, removing an ECS instance is executed directly without waiting for the cool-down time.

A successful return indicates that the Auto Scaling service will shortly execute the scaling activity, but it does not mean that the scaling activity will be successfully executed. Use the returned ScalingActivityID to check the scaling activity status.

If the number of existing ECS instances in the scaling group (Total Capacity) minus the number of ECS instances to be removed is less than the MinSize value, the operation fails.

Example



9 Maintain Auto Scaling

9.1 Edit a lifecycle hook

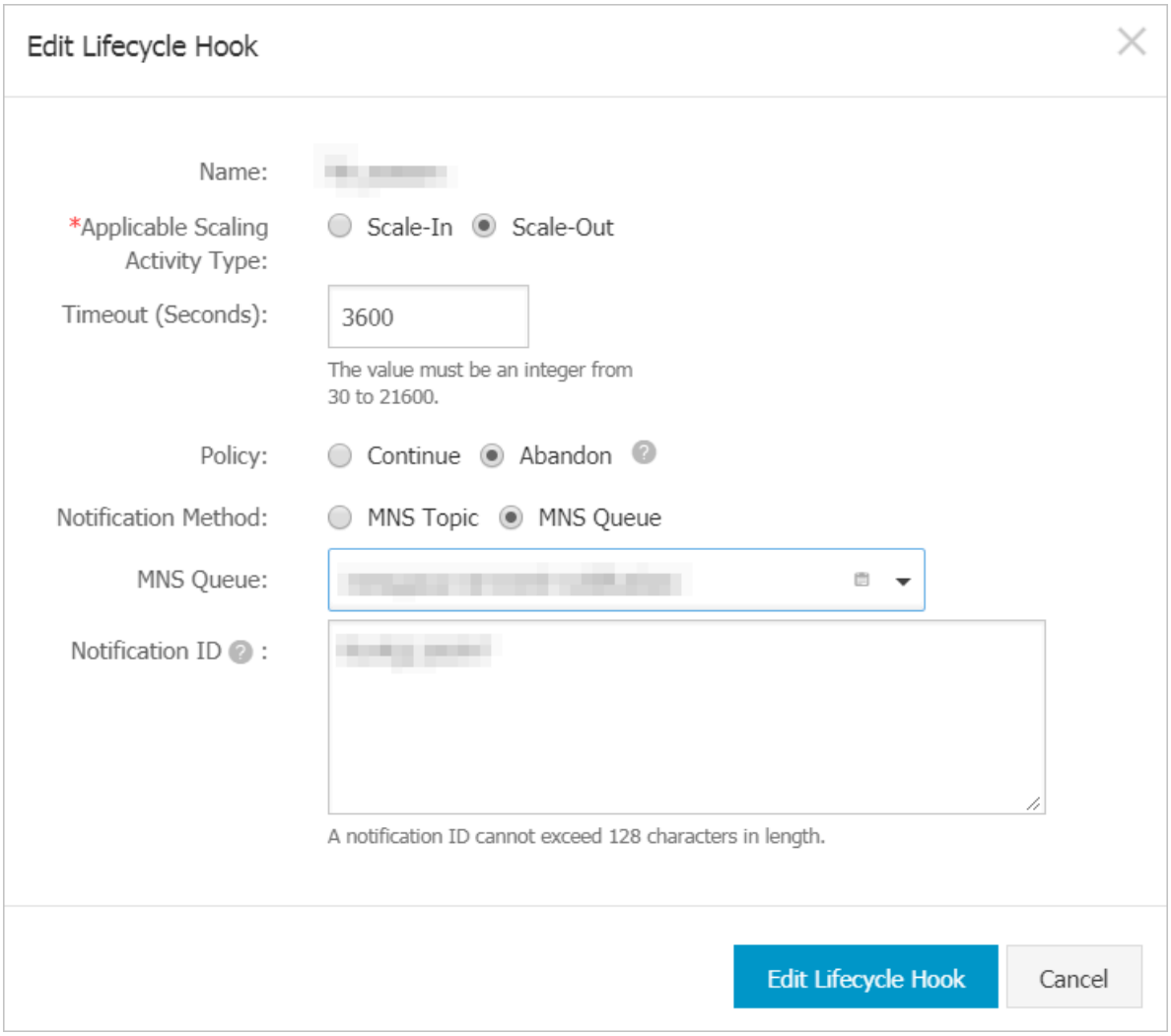
This topic describes how to edit a lifecycle hook.

Context

After you *create a lifecycle hook*, when the lifecycle hook no longer meets your requirements, you do not need to delete it and create a new one. You can change the properties of the lifecycle hook to meet your requirements.

Procedure

1. Log on to the *Auto Scaling console*.
2. On the Scale Groups page, click Manage in the Actions column for the target scaling group.
3. On the Lifecycle Hooks page, click Edit in the Actions column for the target lifecycle hook.
4. In the Modify Lifecycle Hook dialog box, change the properties of the lifecycle hook, and then click Edit Lifecycle Hook.



Note:

- You can change all properties of a lifecycle hook, except its name.
- For more information about lifecycle hook properties, see Lifecycle hook properties under *Create a lifecycle hook*.

9.2 Delete a lifecycle hook

This topic describes how to delete a lifecycle hook.

Context

The maximum 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.

Procedure

1. Log on to the *Auto Scaling console*.
2. On the Scale Groups page, click Manage in the Actions column for the target scaling group.
3. On the Lifecycle Hooks page, click Delete in the Actions column for the target lifecycle hook.

   Note:
   You can also select the check box to the left of the Lifecycle Hook Name field, and then click Delete to delete multiple lifecycle hooks at the same time.
4. In the Delete Lifecycle Hook dialog box, click OK to delete the lifecycle hook.

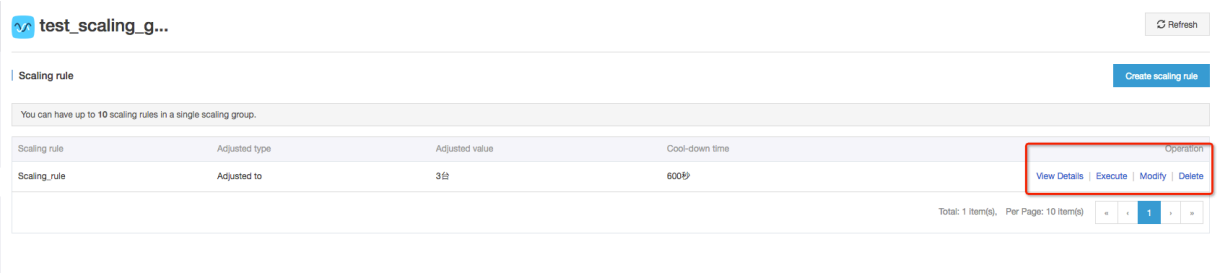
   Note:
   Deleting a lifecycle hook also changes the Wait status of the ECS instance that has been paused by the lifecycle hook.

9.3 Modify, query, or delete a scaling rule

This article introduces the steps to modify a scaling rule.

Context

After *creating a scaling rule*, you can modify, delete, or query a scaling rule. You can view the details of the scaling rule before you modify it.



Procedure

1. Go to the Scaling Rules page, and click View Details in the Actions column after the scaling rule to be modified.
2. After you confirm the scaling rule to be modified, click Modify.
3. In the Modify Scaling Rule dialog box, modify the attributes as needed, and then click Modify Scaling Rule.

Note:
For the attributes of a scaling rule, see *what is a scaling rule*.

9.4 Modify a scaling group

This topic describes how to modify a scaling group.

You can modify the attributes of a scaling group based on your actual needs after it is created.

Note:
If you specify the Maximum Instances or Minimum Instances parameter and the number of instances exceeds or drops below the limit, Auto Scaling automatically adds or removes instances to make sure that the number of instances is valid.

Procedure

Follow these steps to modify the attributes of a scale group:

1. On the Scaling Groups page, click Edit in the Actions column.
2. In the Edit Scaling Group dialog box that appears, modify the attributes as needed.
   a. Enter a name in the Scaling Group Name field.
   b. Enter a number in the Maximum Instances field.

      Note:
      If the specified number exceeds the upper limit, Auto Scaling automatically removes instances to make the number of instances equal to the upper limit.
   c. Enter a number in the Minimum Instances field.

      Note:
      If the specified number drops below the lower limit, Auto Scaling automatically adds instances to make the number of instances equal to the lower limit.
   d. Enter a number in the Default Cooldown Time (Seconds) field.

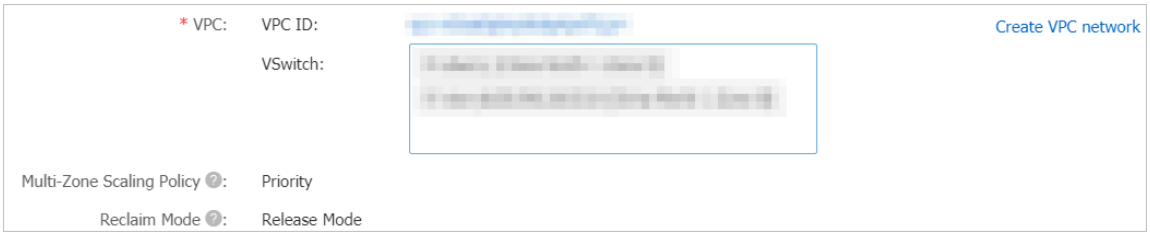
      Note:
      This parameter specifies the default scaling activity cooldown time. For more information, see *Cool-down time*.

e. Configure a Removal Policy.

> **Note:**
> This parameter specifies the policy to remove instances when the number of instances in a scaling group exceeds the upper limit. For more information, see *Removal policies*.

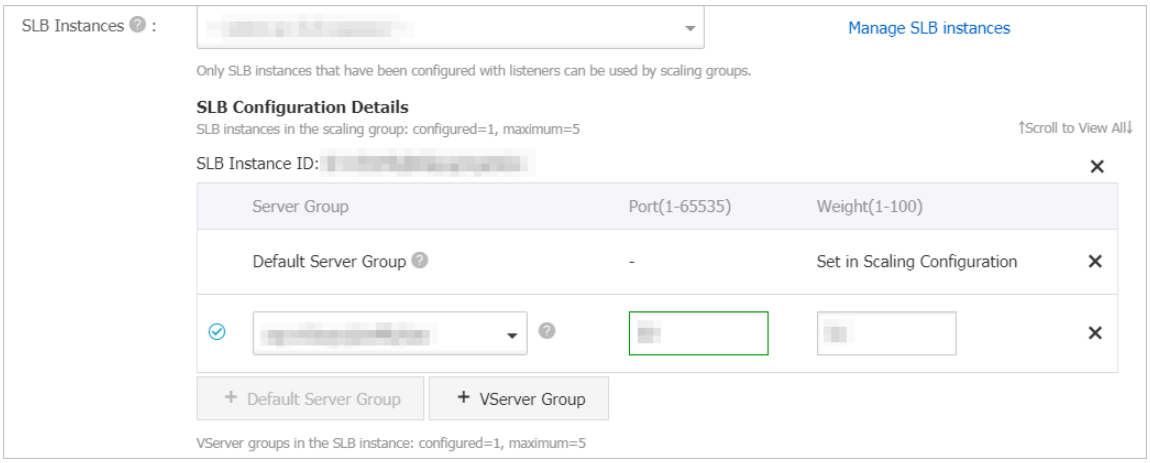
f. Select an Instance Configuration Source.

g. (Optional) Once selected, you cannot modify the Network Type of the scaling group. If the Network Type of the scaling group that you need to modify is VPC, then you can change the VSwitch. However, you cannot change the Multi-Zone Scaling Policy or the Reclaim Mode .



h. (Optional) Select SLB Instances.

> **Note:**
> A scaling group can be associated with up to five SLB instances at the same time. You can also select the *default server group* or *VServer group(s)* of a SLB instance for the scaling group. You can select up to five VServer groups for a scaling group at the same time. For more information, see *Use Server Load Balancer (SLB) in Auto Scaling*.



i. (Optional) Select RDS Instances. Currently, only RDS databases are supported.

> Note:
> You can only add RDS instances that are in the same region where the scaling group is created. After ECS instances are added to the scaling group, Auto Scaling automatically adds the internal IP addresses of the ECS instances to the RDS whitelist to allow internal communication between the ECS and the RDS instances.

9.5 Delete a scaling group

This article describes the steps to delete a scaling group.

Context

You can delete a scaling group if you no longer need it.

> Note:
> Deleting a scaling group also deletes its scaling configurations and scaling rules. If the scaling group includes ECS instance in the Running status, Auto Scaling stops the ECS instance first, removes all manually added instances, and releases all automatically created instances.

Procedure

1. On the Scaling Groups page, click Delete in the Actions column next to the scaling group to be deleted.
2. In the Delete Scaling Group dialog box, click Confirm.
3. On the Scaling Groups page, click Refresh to confirm that the deletion has completed.

10 Move ECS instance to Standby

This topic describes how to move ECS instance to Standby.

Auto Scaling allows you to set the Standby status for one or more ECS instances. After an ECS instance is in the Standby status, you can upgrade or maintain the ECS instance. Meanwhile, we do not either perform health check for the specified instance or release it.

Features

- If an ECS instance is set to the Standby status:
  - It is not in service until you resume the ECS instance.
  - Its lifecycle is controlled by you rather than Auto Scaling service.
  - The weight of the ECS instance is deregistered to zero if the scaling group has Server Load Balancer instances attached.
  - You can *stop* instance, *restart* instance, or do other maintenance operations, such *upgrade the instance configurations*, *change the operating system*, *reinitialize the cloud disk*, or *migrate from the classic network to a VPC*.
  - It is not removed from the scaling group whenever a scaling event happens.
  - The health status is not updated even the specified instance is stopped or restarted.
  - It must be removed from the scaling group before you release the instance.
  - It is resumed for a short while when you delete the related scaling group and then it is release along with the scaling group.
- If an ECS instance is back to the in service status:
  - It handles application traffic actively again.
  - The weight of the ECS instance is set to a predefined value if the scaling group has Server Load Balancer instances attached.
  - The health status is updated if the specified instance is stopped or restarted.
  - Its lifecycle is controlled by Auto Scaling service rather than you.

Move to Standby

1. Log on to the *Auto Scaling console*.
2. Select a region, such as China East 2 (Shanghai).

3. Find and click the target scaling group.
4. In the left-side navigation pane, click ECS instances.
5. Find and click the target ECS instance, click Move to Standby.

Remove from Standby

1. Log on to the *Auto Scaling console*.
2. Select a region, such as China East 2 (Shanghai).
3. Find and click the target scaling group.
4. In the left-side navigation pane, click ECS instances.
5. Find and click the target ECS instance, click Remove from Standby.

API operations

- Move to Standby: *EnterStandby*
- Remove from Standby: *ExitStandby*

References

- *What is Server Load Balancer*
- *Remove an unhealthy ECS instance*

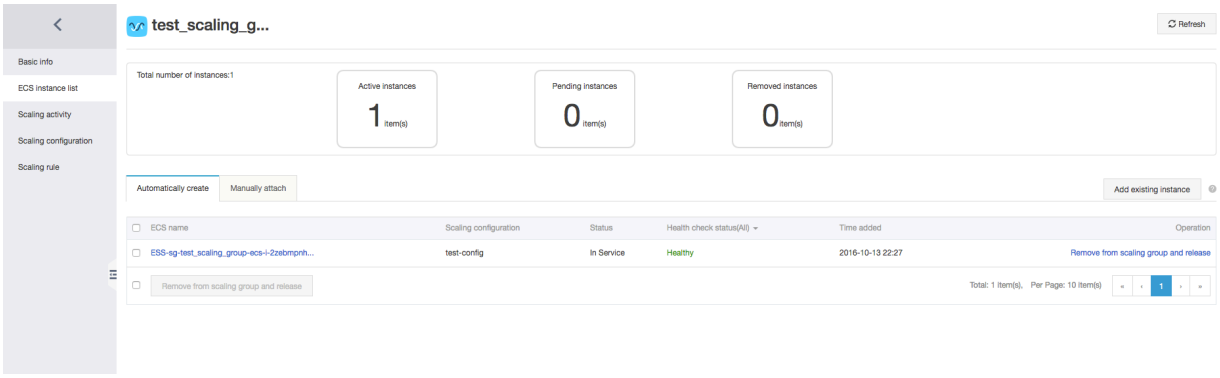
11 Query the ECS instance list

This article describes how to query the ECS instance list.

ECS instances not in the Running(`Running`) status are regarded as unhealthy. Auto Scaling automatically removes unhealthy ECS instances from the scaling groups. Automatically created ECS instances are created by the Auto Scaling service based on scaling configuration and rules. Manually added ECS instances are manually added to a scaling group, not created by the Auto Scaling service.

Example

The example is shown as follows.

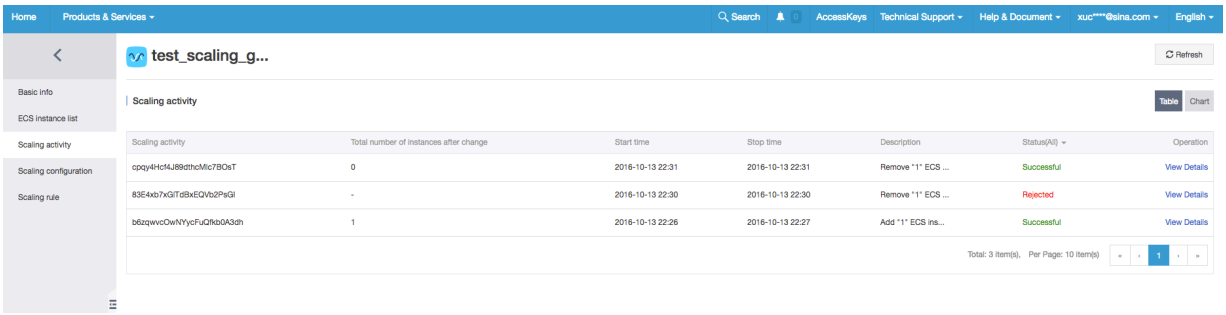


12 View scaling activities

This article shows how to view scaling activities.

This operation queries the information of scaling activities performed in the last 30 days.

Example



13 Use Server Load Balancer (SLB) in Auto Scaling

You can associate a scaling group with SLB instances to distribute traffic to multiple ECS instances in a scaling group, improving the performance of the scaling group.

Overview

SLB allows multiple ECS instances in a region to use the same SLB instance IP address to share the service load. These ECS instances act as a high-performance and highly available application service pool. This means that SLB allocates and controls traffic by using SLB instances, listeners, and backend servers. For more information, see *What is Server Load Balancer*.

Prerequisites

Make sure that the following requirements are met before you attach SLB instances to a scaling group:

- You have one or more SLB instances in the Running status. If you do not have a running SLB instance, *create an SLB instance* first.
- The SLB instance and the scaling group must be in the same region.
- The SLB instance and the scaling group must be in the same VPC network if their network type is VPC.
- If the network type of the SLB instance is classic, the network type of the scaling group is VPC, and the backend server group of the SLB instance contains VPC-connected ECS instances, then the ECS instances and the scaling group must be in the same VPC network.
- You must configure at least one listener for the SLB instance. For more information about listeners, see *Listener overview*.
- You must enable health check for the SLB instance. For more information, see *Configure health check*.

Manage SLB instances in the Auto Scaling console

> **Note:**
> This section describes how to manage SLB instances in the Auto Scaling console. For more information, see *Use custom scaling configuration to create a scaling group*.

1. Log on to the *Auto Scaling console* .

2. On the Scaling Groups page, use one of the following methods to add SLB instances:
   - Click Create Scaling Group in the upper-right corner when you create a scaling group.
   - Click Edit in the Actions column when you modify a scaling group.
3. Select a Network Type.

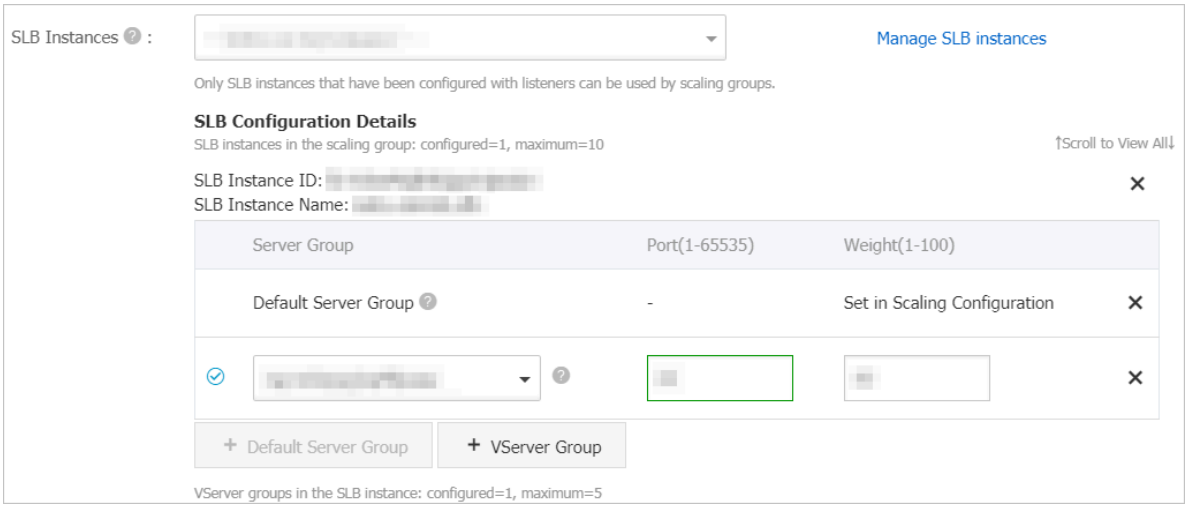
> Note:
> Once selected, you cannot change the network type.

4. Associate the scaling group with an SLB instance.

> Note:
> You can associate a scaling group with up to five SLB instances at the same time. Your SLB instances may not be displayed if they do not meet the requirements described in *prerequisites*. Click Manage SLB instances to view and update the SLB instances in the SLB console.

5. Select a server group for the ECS instances in the scaling group.

> Note:
> You can also select the *default server group* or *VServer group(s)* for each SLB instance in the scaling group. You can select up to five VServer groups for a scaling group at the same time.



6. Configure the remaining settings as needed.

Call API operations to manage SLB instances

When you call *CreateScalingGroup* to create a scaling group, you can use `LoadBalancerIds` to associate the scaling group with SLB instances and use `VServerGroup` to set the attributes of the VServer group.

To modify a scaling group, you can call `AttachLoadBalancers` or `DetachLoadBalancers` to associate or disassociate the scaling group from SLB instances, and call `AttachVServerGroups` or `DetachVServerGroups` to add or remove VServer groups.

Load balancing effect

After the scaling group is associated with SLB instances, all ECS instances will be added to the backend server group of the SLB instances. The SLB instances distribute traffic to ECS instances based on traffic distribution and health check policies. This improves resource availability.

> **Note:**
> The weight of these ECS instances is 50 by default, you can adjust the weight on the corresponding SLB instances.

14 Scaling groups