Alibaba Cloud Auto Scaling

User Guide

Issue: 20190918

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 informatio n, 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

Legal disclaimer..... I
Generic conventions..... I
1 Usage notes..... 1
1.1 ECS instance lifecycle..... 1
1.2 Cool-down time..... 2
1.3 Scaling group status..... 3
1.4 Scaling activity process..... 4
1.5 Scaling activity status..... 5
1.6 Instance rollback resulting from a scaling activity failure..... 6
1.7 Remove an unhealthy ECS instance..... 7
1.8 Notification..... 7
1.9 Forced intervention..... 8
1.10 Quantity limits..... 10
1.11 Considerations..... 11
1.12 Operation procedure..... 12
1.13 Workflow..... 13
2 Scaling configurations..... 16
2.1 ECS instance templates..... 16
2.2 Create a scaling configuration..... 18
2.3 Apply scaling configurations..... 20
2.4 Modify scaling configurations..... 21
2.5 Delete scaling configurations..... 22
2.6 Export scaling configurations..... 23
2.7 Import scaling configurations..... 23
3 Realize Auto Scaling..... 24
3.1 Use custom scaling configurations to create scaling groups..... 24
3.2 Use launch templates to create scaling groups..... 30
3.3 Create a scaling rule..... 36
3.4 Create a lifecycle hook..... 39
3.5 Create a predictive scaling rule..... 43
3.6 Execute a scaling rule..... 46
3.7 Scheduled tasks..... 50
3.7.1 Create a scheduled task..... 50
3.7.2 Manage scheduled tasks..... 53
3.8 Alarm tasks..... 54
3.8.1 Auto Scaling alarm tasks..... 54
3.8.2 System monitoring alarm tasks..... 56
3.8.3 Custom monitoring alarm tasks..... 58
3.8.4 Create event-triggered tasks..... 60
3.8.5 View, modify, or delete an alarm task..... 62

4 Maintain Auto Scaling.....63
4.1 Check the result of a predictive scaling rule.....63
4.2 Edit a lifecycle hook.....65
4.3 Delete a lifecycle hook.....66
4.4 Modify a scaling rule.....67
4.5 Delete a scaling rule.....70
4.6 Move ECS instance to Standby.....71
4.7 Switch to the protected state for an ECS instance.....72
4.8 Remove from the protected state for an ECS instance.....73
4.9 Removal policies.....73
4.10 Change the status of a scaling group.....74
4.11 Modify a scaling group.....75
4.12 Delete a scaling group.....77
5 Manual scaling.....79
5.1 Add ECS instances.....79
5.2 Remove an ECS instance.....81
6 Event notification.....83
6.1 Event notification overview.....83
6.2 Create an event notification.....85
6.3 Manage event notifications.....87
7 Query the ECS instance list.....89
8 View scaling activities.....90
9 Related services.....91
9.1 Use Server Load Balancer (SLB) in Auto Scaling.....91
10 Auto Scaling FAQ.....94
11 Scaling groups.....96

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 limits

This topic describes the quantity limits in Auto Scaling.

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

- A single account can have up to 50 scaling groups in a region.
  - A maximum of 10 scaling configurations can be created for a scaling group.
  - A maximum of 50 scaling rules can be created for a scaling group.
  - A maximum of six event notifications can be created for a scaling group.
  - A maximum of six lifecycle hooks can be created for a scaling group.
  - A maximum of five SLB instances can be associated with a scaling group at the same time.
- A maximum of 20 scheduled tasks can be created under an account.
- A maximum of 10 instance types can be specified in 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

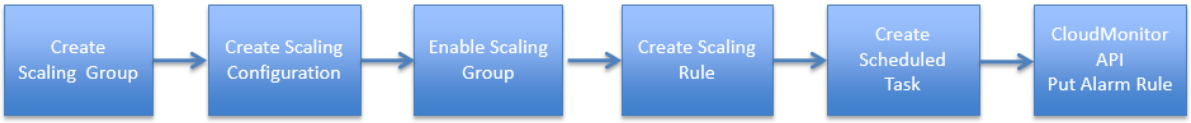
- Only one scaling activity can be executed at a time in a scaling group.
- A scaling activity cannot be interrupted. For example, if a scaling activity to add 20 ECS instances is being executed, it cannot be forced to terminate when only five instances have been created.
- When a scaling activity fails 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

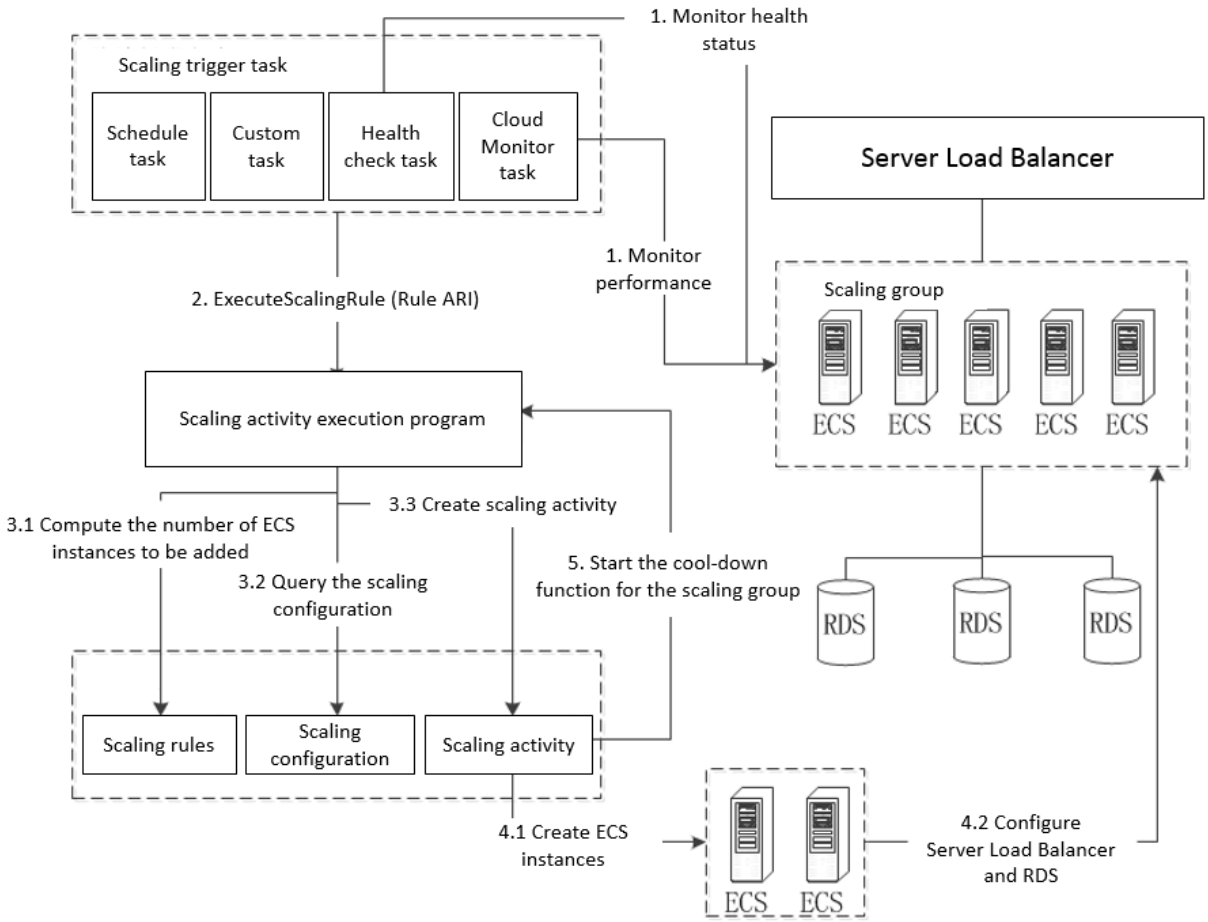
This topic introduces the steps to create a complete Auto Scaling solution.



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:

- #unique_24
- #unique_25
- #unique_26

Launch templates

Launch templates are a tested feature of ECS. You can use existing launch templates to configure scaling groups. For more information, see #unique_27.

> **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 for a scaling group and specify an ECS instance template for automatic scaling.

Background information

The scaling configuration creation process is similar to that of an ECS instance. However, as a scaling configuration is a template for ECS instance to be used during automatic scaling, it has several key differences such as supporting different instance types and not allowing you to configure certain parameters (such as region or resource group). See the actual interface of the Auto Scaling console when using this document. Brief descriptions of each parameter are also displayed on the interface. For more information about parameter descriptions, see #unique_30 .

Preparations

- If you need to create a scaling configuration during the scaling group creation process in the Auto Scaling console, perform operations from step 3.
- If you need to use the API examples provided in this topic, set API access permissions first.

Procedure in the Auto Scaling console

1. Log on to the Auto Scaling console. In the Actions column corresponding to a scaling group, click Manage.
2. In the left-side navigation pane, click Instance Configuration Source. On the tab page that appears, click Create Scaling Configuration.
3. On the Basic Configurations page, configure the billing method, instance, image, storage, public network bandwidth, and security group. Click Next: System Configurations.

> Note:
> In the basic configurations:
>
> - Billing method: Only the Pay-As-You-Go and #unique_33 methods are supported.
> - Instance: Multiple instance types are supported. When the instance inventory of a specific type is insufficient, the instances of the alternate types will be used to improve the scaling success rate.

4. On the System Configurations page, configure the logon credential, tag (optional), instance name (optional), and advanced options (optional). Click Next: Preview.

   Note:
   Advanced options are available only for scaling configurations in a VPC-type scaling group. The options include RAM role and custom data of instances.

5. On the Preview page, check your configurations, enter the scaling configuration name, and click Create.
6. In the Activated dialog box that appears, you can click Enable Configuration, click Create More to create another scaling configuration, or close the dialog box.

API example

A scaling configuration is a template used by a scaling group to elastically create ECS instances. Before you use an API to create a scaling configuration, make sure that the request contains the ID of a scaling group, the ID of a security group to which ECS instances will belong, the ID of an ECS instance image, and the type of the ECS instances to be used.

We recommend that you choose an Alibaba Cloud SDK based on your programming language. Java SDK example:

```
 import   com . aliyuncs . CommonRequ  est ;
 import   com . aliyuncs . CommonResp  onse ;
 import   com . aliyuncs . DefaultAcs  Client ;
 import   com . aliyuncs . IAcsClient ;
 import   com . aliyuncs . exceptions . ClientExce  ption ;
 import   com . aliyuncs . exceptions . ServerExce  ption ;
 import   com . aliyuncs . http . MethodType ;
 import   com . aliyuncs . profile . DefaultPro  file ;
/*
 pom . xml
< dependency >
  < groupId > com . aliyun </ groupId >
  < artifactId > aliyun - java - sdk - core </ artifactId >
  < version > 4 . 0 . 3 </ version >
</ dependency >
*/
 public   class   CommonRpc  {
     public   static   void   main ( String []  args ) {
         DefaultPro  file   profile  =  DefaultPro  file . getProfile
 (" cn - hangzhou ", "< accessKeyI  d >", "< accessSecr  et >");
         IAcsClient   client  =  new   DefaultAcs  Client ( profile );

         CommonRequ  est   request  =  new   CommonRequ  est ();
         request . setMethod ( MethodType . POST );
         request . setDomain (" ess . aliyuncs . com ");
         request . setVersion (" 2014 - 08 - 28 ");
         request . setAction (" CreateScal  ingConfigu  ration ");
         request . putQueryPa  rameter (" ScalingGro  upId ", " asg -
 bp1a4xzjr1  ypd6016356 ");
```

```
        request . putQueryPa  rameter (" SecurityGr  oupId ", " sg -
 bp147qpndp  7iyj08l74h ");
        request . putQueryPa  rameter (" ImageId ", " centos6u5_
 64_20G_ali  aegis_2014  0703 . vhd ");
        request . putQueryPa  rameter (" InstanceTy  pe ", " ecs . t1
 . xsmall ");
        try  {
            CommonResp  onse   response  =  client . getCommonR
 esponse ( request );
            System . out . println ( response . getData ());
        }  catch  ( ServerExce  ption   e ) {
            e . printStack  Trace ();
        }  catch  ( ClientExce  ption   e ) {
            e . printStack  Trace ();
        }
    }
}
```

After you have initiated a call through the Java SDK or other methods, the request body is similar as follows:

```
 https :// ess . aliyuncs . com /? Action = CreateScal  ingConfigu
 ration
& ImageId = centos6u5_  64_20G_ali  aegis_2014  0703 . vhd
& InstanceTy  pe = ecs . t1 . xsmall
& ScalingGro  upId = asg - bp1a4xzjr1  ypd6016356
& SecurityGr  oupId = sg - bp147qpndp  7iyj08l74h
& Version = 2014 - 08 - 28
```

In the request,

- `centos6u5_ 64_20G_ali aegis_2014 0703 . vhd` indicates the ID of the ECS instance image.
- `ecs . t1 . xsmall` indicates the type of the ECS instances.
- `AG6CQdPU8O KdwLjgZcJ2 ea` indicates the ID of the scaling group.
- `sg - 280ih3w` indicates the ID of the scaling group to which the ECS instances will belong.
- `2014 - 08 - 28` indicates the version of the API.

You can also customize attributes such as different instance types and ECS instance disks. For more information about the API, see CreateScalingConfiguration.

2.3 Apply scaling configurations

In a scaling group, you can create multiple scaling configurations, but you can only apply one of these scaling configurations.

Context

A scaling configuration includes many configuration items. You can obtain the overview of a scaling configuration by using the View Details function to ensure that an appropriate ECS instance template is applied.

Procedure

1. Log on to the Auto Scaling Console.
2. On the Scaling Groups page, click Manage in the Actions column.
3. In the left-side navigation pane, select Instance Configuration Source.
4. In the left-side navigation pane, select Instance Configuration Source. On the Scaling Configurations tab, locate the scaling configuration to be applied, and click View Details in the Actions column.
5. If you decide to apply the scaling configuration, click Apply in the Actions column.

> Note:
> After you apply a scaling configuration, the status of the other scaling configurations will switch to Inactive.

Result

After you apply a scaling configuration, ECS instances that are created later will be automatically based on the scaling configuration in a scaling group when the specified scaling condition is met.

2.4 Modify scaling configurations

When business requirements change, you can improve efficiency by modifying the required options of the current scaling configuration rather than creating a new scaling configuration.

Context

For more information about configuration items, see #unique_30.

Procedure

1. Log on to the Auto Scaling Console.
2. On the Scaling Groups page, click Manage in the Actions column.
3. In the left-side navigation pane, select Instance Configuration Source.
4. On the Scaling Configurations tab, click Modify in the Actions column.

5. On the Basic Configurations page, modify the required options, then click Next: System Configurations.
6. On the System Configurations page, modify the required options, then click Next: Preview.
7. On the Preview (Required) page, click Modify.

> Note:
> After you modify a scaling configuration, ECS instances that are created based on the scaling configuration still work as expected.

2.5 Delete scaling configurations

You can delete scaling configurations that are no longer required to release more quotas.

Prerequisites

Before you delete a scaling configuration, ensure the following conditions are met, or the delete operation will fail.

- The status of the scaling configuration must be Inactive.
- In the scaling group, no ECS instance is automatically created based on the scaling configuration.

Procedure

1. Log on to the Auto Scaling Console.
2. On the Scaling Groups page, click Manage in the Actions column.
3. In the left-side navigation pane, select Instance Configuration Source.
4. On the Scaling Configurations tab, locate the required scaling configuration, and click Delete in the Actions column.

   If you decide to delete multiple scaling configurations, select multiple scaling configurations, and click Delete in the Scaling Configuration Name/ID column.
5. In the Delete Scaling Configuration dialog box, click Confirm.

2.6 Export scaling configurations

You can export the scaling configurations of a scaling group to a .csv file. The file can be quickly imported into another scaling group or used as a backup.

Procedure

1. Log on to the Auto Scaling Console.
2. On the Scaling Groups page, click Manage in the Actions column.
3. In the left-side navigation pane, select Instance Configuration Source.
4. Select the Scaling Configurations tab and click Export to download and save a .csv file to your local PC.

2.7 Import scaling configurations

You can import a scaling configuration file to a scaling group to improve the efficiency of creating scaling configurations. However, the network type must be the same for both the source scaling group and the target scaling group.

Procedure

1. Log on to the Auto Scaling Console.
2. On the Scaling Groups page, click Manage in the Actions column.
3. In the left-side navigation pane, select Instance Configuration Source.
4. On the Scaling Configurations tab, click Import.
5. Click Select File to select a .csv file to be imported.
6. In the Preview area, select the required scaling configuration, and click Import.

   Note:

   - You can add a suffix to the name of an imported scaling configuration to avoid duplicate names.
   - If you cannot select a scaling configuration in the Preview area, the network type of the target scaling group may be different from that of the source scaling group.

7. View the importing result, and click OK.

3 Realize Auto Scaling

3.1 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 #unique_43.

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 #unique_44.

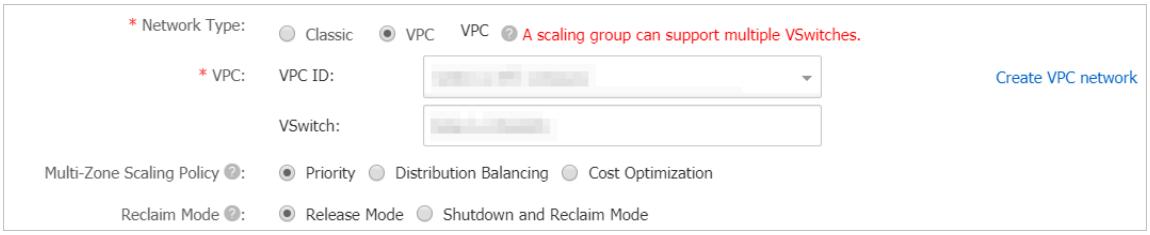
f. Select an Instance Configuration Source. This example uses Custom Scaling Configuration.

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.

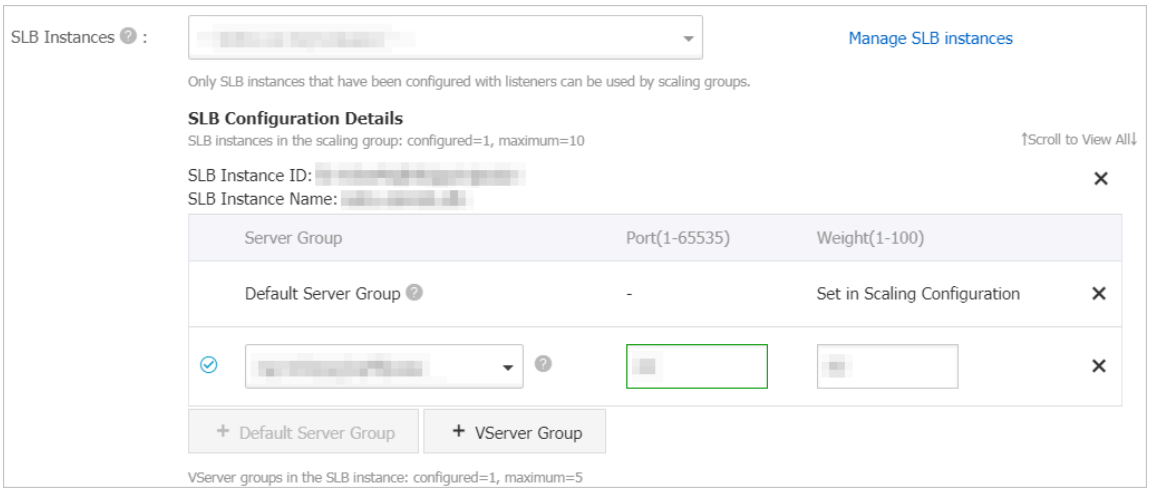
   C. Specify the Reclaim Mode. For more information, see Reclaim mode.



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

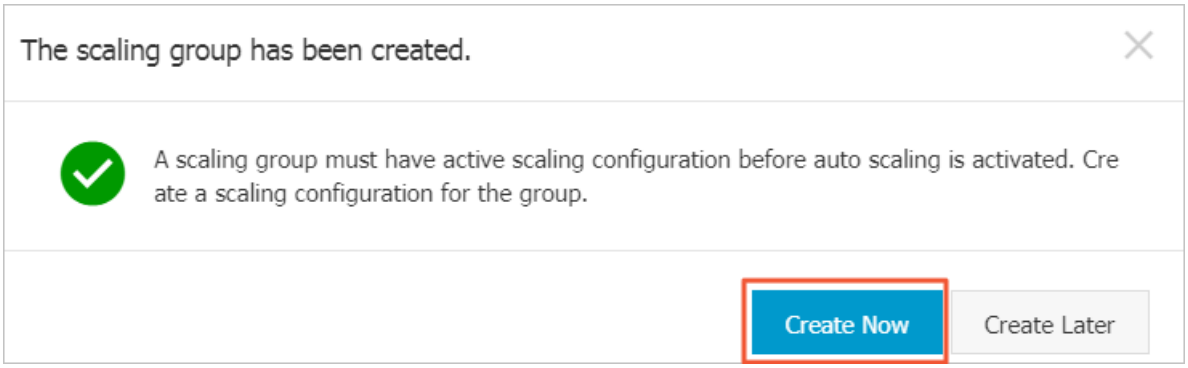
scaling group at the same time. For more information, see Use Server Load Balancer (SLB) in Auto Scaling.



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 #unique_24.

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.

3.2 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.

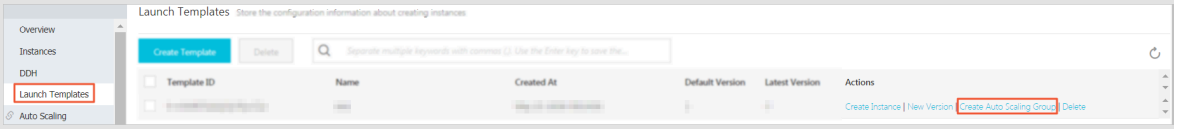
> **Note:**
> There is a limit to the number of scaling groups that you can create under one account. For more information, see Quantity restrictions.

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.

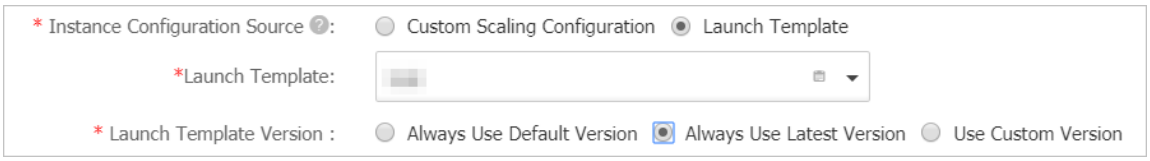
> **Note:**
> This parameter specifies the default cooldown time of a scaling activity . For more information, see #unique_43.

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 #unique_44.

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.

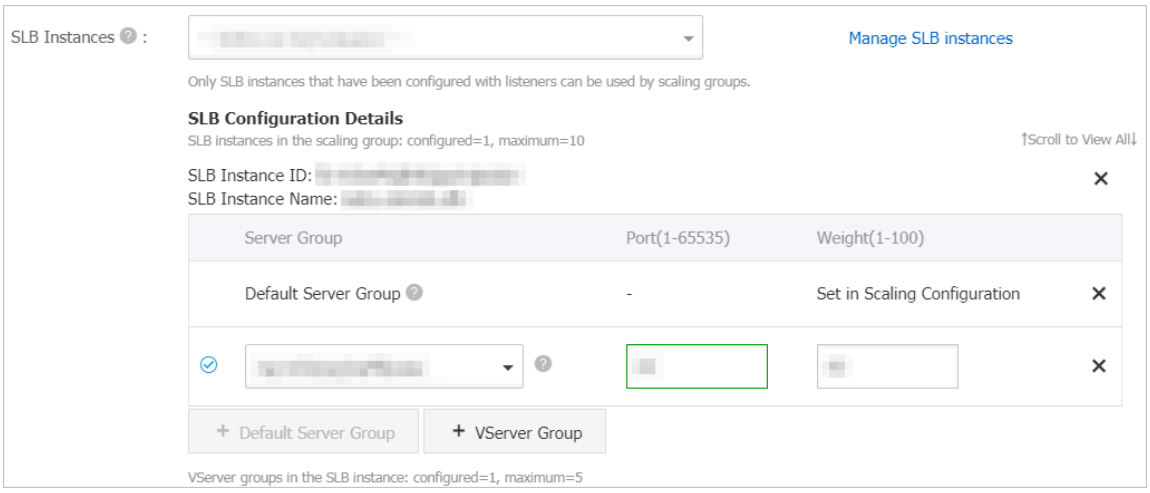
B. Specify the Multi-Zone Scaling Policy. For more information, see Multi-zone scaling policy.

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.

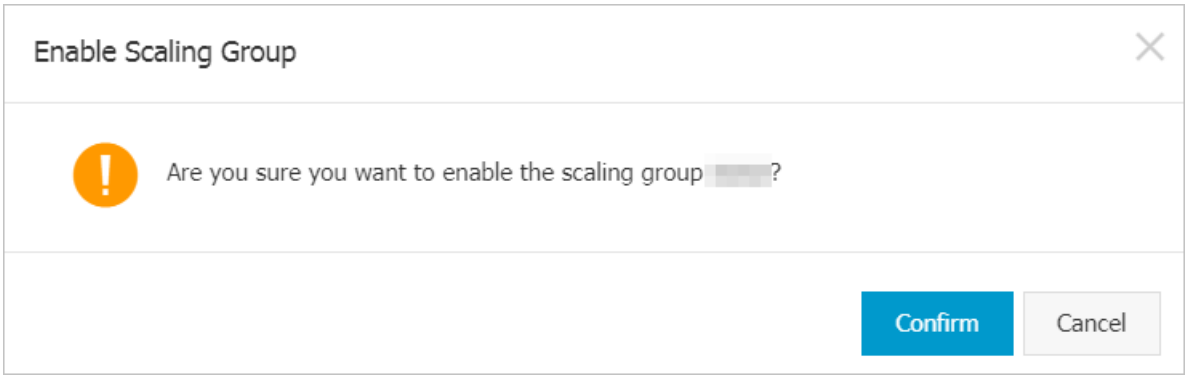


i. (Optional) Add RDS Instances. Currently, only RDS databases are supported.

> **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.

3.3 Create a scaling rule

After creating a scaling group, you must create scaling rules to manage specific scaling actions of the group. This topic describes how to create a scaling rule.

Limits

- There is a limit to the maximum number of scaling rules that you can create in a scaling group. For more information, see Quantity limits.
- Target tracking scaling rules can only be executed by alert tasks that were automatically created.
- After you execute a scaling rule, if the actual number of ECS instances in service in a scaling group is greater than the configured MaxSize or smaller than the configured MinSize of the group, Auto Scaling automatically adds or removes instances to ensure the number of instances is within the configured range. The following examples illustrate this limit:
  - Assume that you have a scaling group named asg-bp19ik2u5w7esjcu****. The group has a MaxSize attribute of three, and has a scaling rule named add3 to add three instances. If the current number of instances in service is two and you execute the scaling rule add3, only one ECS instance is added.
  - Assume that you have a scaling group named asg-bp19ik2u5w7esjcu****. The group has a MinSize attribute of two, and has a scaling rule named reduce2 to remove two instances. If the current number of instances in service is three and you execute the scaling rule reduce2, only one ECS instance is removed.

Procedure

1. Log on to the Auto Scaling console.
2. Click Manage in the Actions column corresponding to a scaling group for which you want to create a scaling rule.
3. In the left-side navigation pane, click Scaling Rules. On the page that appears, click Create Scaling Rule in the upper-right corner.

4. In the Create Scaling Rule dialog box that appears, set the parameters as needed, and then click Create Scaling Rule.

   You can set Rule Type as needed. For more information about these parameters, see Simple scaling rules and Target tracking scaling rules.

   Note:
   When you create a target tracking scaling rule, an alert task associated with the rule is also created. Only this alert task is able to execute the target tracking scaling rule.

Simple scaling rules

A simple scaling rule directly adds or removes a specified number of instances, or adjusts instances in a scaling group to a specified number. The following table describes the simple scaling rule parameters.

Parameter	Description
Rule Name	The name of the scaling rule.
Rule Type	The type of the scaling rule. This parameter cannot be modified after the scaling rule is created.

Parameter	Description
Operation	The operation to be executed when the scaling rule is triggered. The operations include: · Change to N instances: When the scaling rule is executed, the number of instances in the scaling group changes to N. A maximum of 500 instances can be scaled at a time. · 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 at a time. · Add instances by N%: When the scaling rule is executed, N % of the current instances in the scaling group are added. A maximum of 500 instances can be scaled at a time. · 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 at a time. · Remove instances by N%: When the scaling rule is executed , N% of the current instances in the scaling group are removed. A maximum of 500 instances can be scaled at a time.
Cooldown Time	Optional. The cooldown period. Unit: seconds. If this parameter is not specified, the default cooldown period of the scaling group is used. For more information, see Cooldown period.

Target tracking scaling rules

A target tracking scaling rule specifies a target value of a CloudMonitor metric. Auto Scaling automatically calculates the number of instances required to meet that target and scales ECS instances to ensure that the metric value remains close to the target value. The following table describes the parameters of target tracking scaling rules.

Parameter	Description
Rule Name	The name of the scaling rule.
Rule Type	The type of the scaling rule. This parameter cannot be modified after the scaling rule is created.

Parameter	Description
Metric Name	The name of a CloudMonitor metric to be monitored. Valid values: · Average CPU Usage · Average Inbound Internal Traffic · Average Outbound Internal Traffic · Average Inbound Public Traffic · Average Outbound Public Traffic
Target Value	The target value of the CloudMonitor metric. A target tracking scaling rule keeps the CloudMonitor metric value close to the target value.
Warmup Time	The instance warm-up period. Unit: seconds. During this warm-up period, if instances are created by a target tracking scaling rule, the created and started instances do not affect the CloudMonitor metric value. This is to prevent the metric value from being changed multiple times because of scaling activities within the warm-up period.
Disable Scale-in	Indicates whether to disable scale-in. If this parameter is selected, a target tracking scaling rule cannot remove instances from the scaling group.

3.4 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:

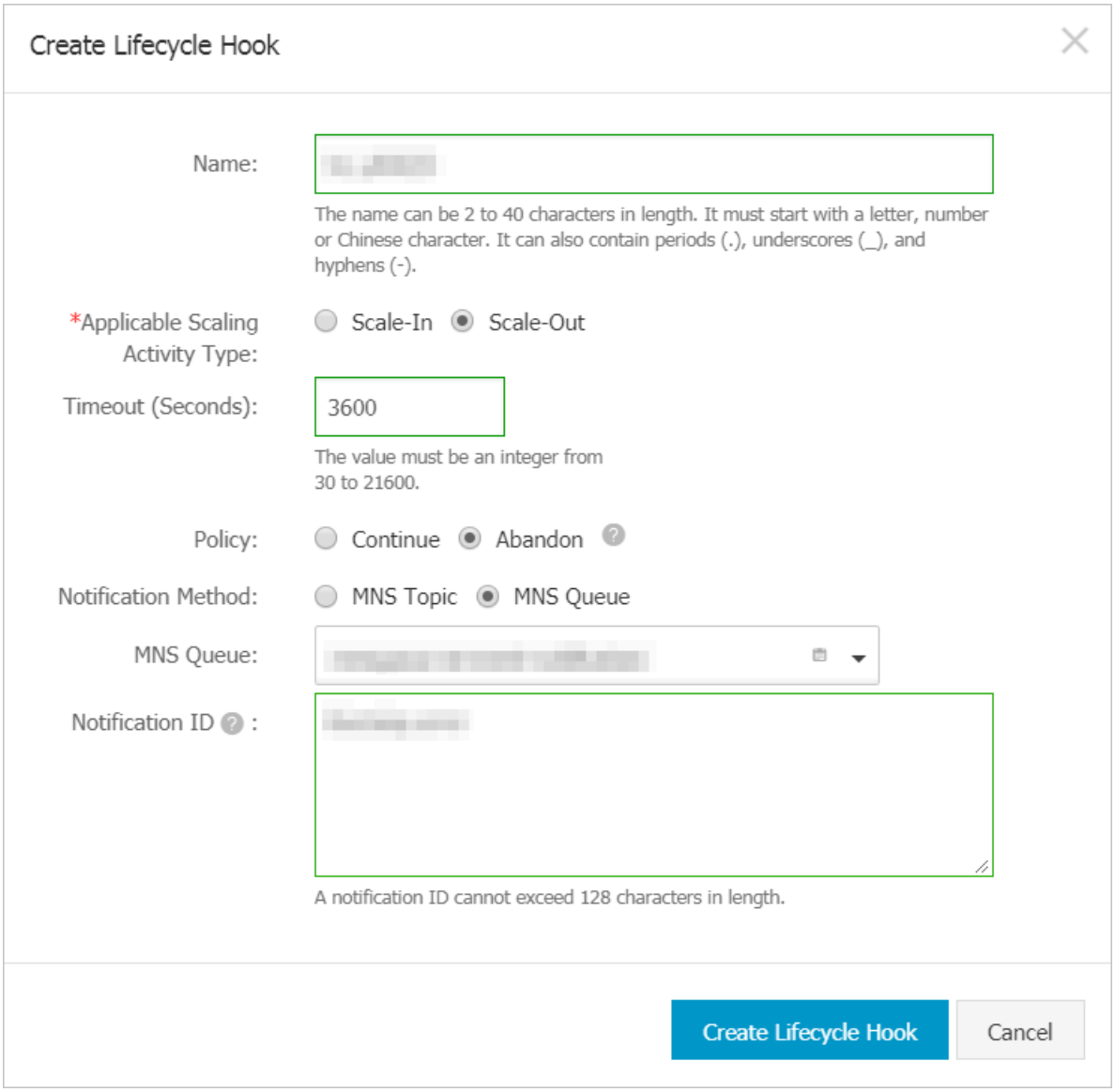
- The scaling group does not perform other scaling activities.
- You can perform custom actions during the lifecycle hook timeout period. For example, you can initialize the configuration of the ECS instance or obtain the ECS instance data.
- You can delete the corresponding lifecycle hook to resume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

Property	Description	Example
Notification ID	The notification ID is sent to you with notifications so that you can easily manage the notifications by ID.	General information

3.5 Create a predictive scaling rule

Predictive scaling rules are used to analyze historical monitoring data in a scaling group and predict the values of monitored metrics by means of machine learning. Scheduled tasks can be automatically created for the predictive scaling rule to set the boundary values of the scaling group intelligently. This topic describes how to create a predictive scaling rule.

Context

When creating a scaling group, you can set its boundary values, namely, the maximum and minimum number of ECS instances for scaling. However, the boundary values you set may not meet the actual requirements. If the minimum number of ECS instances is too large, excessive computing resources may be purchased. If the maximum number of ECS instances is too small, service stability may be affected due to insufficient computing resources.

A predictive scaling rule can obtain historical monitoring data generated in a period of at least the past 24 hours, and then predict the values of monitored metrics in the next 48 hours through machine learning. Then, the number of ECS instances required by the scaling group per hour (predicted value) can be calculated. Forecasts are updated once a day, and 48 forecast tasks are created for each of the next 48 hours. Forecast tasks change the boundary values of the scaling group, but not the actual number of ECS instances in the scaling group.

A predictive scaling rule can be used together with target tracking scaling rules and simple scaling rules. When it is used with a target tracking scaling rule, we recommend that you set the same metrics and target values for the rules, to prevent variation of the number of ECS instances caused by the difference in metrics.

Procedure

1. Log on to the Auto Scaling console.

2. On the Scaling Groups page, locate the row that contains the target scaling group and click Manage in the Actions column.
3. In the left-side navigation pane, click Scaling Rules. On the page that appears, click Create Scaling Rule in the upper-right corner.
4. Set the parameters as needed, and then click Create Scaling Rule.

   A scaling group can have only one predictive scaling rule. The following table describes the parameters of a predictive scaling rule.

Parameter	Description
Rule Name	The name of the scaling rule.
Rule Type	The type of the scaling rule. Select Predictive Scaling Rule.
Reference Existing Target Tracking Scaling Rule	If you select this option, the Select Rule field appears for you to select a target tracking scaling rule.
Select Rule	The target tracking scaling rule for the new predictive scaling rule to reference. This field appears when you select Reference Existing Target Tracking Scaling Rule. After you select a target tracking scaling rule, its Metric Name and Target Value apply to the new predictive scaling rule.
Metric Name	The name of a metric that CloudMonitor will monitor. Valid values: · Average CPU Usage (%) · Average Inbound Internal Traffic (KB/Min) · Average Outbound Internal Traffic (KB/Min)
Target Value	The target value of the selected metric that CloudMonitor will monitor. The predictive scaling rule calculates an appropriate predicted value based on the target value and other factors. If you change the target value, existing forecast tasks of the current scaling group will be cleared, and new forecast tasks will be created within an hour.

Parameter	Description
Predictive Scaling Mode	The forecast mode for the scaling group. Valid values: · Forecast and Scale: produces forecasts and creates forecast tasks. · Forecast Only: produces forecasts but does not create forecast tasks. We recommend that you select Forecast Only first and change it to Forecast and Scale after confirming that the forecasts meet your expectations. You can check the result of the predictive scaling rule on the details page.
Initial Max Capacity	The maximum number of ECS instances in the scaling group. This parameter is used together with Max Capacity Behavior. The default value is the current value of Maximum Instances.
Max Capacity Behavior	The action taken on the predicted value when it exceeds Initial Max Capacity. Valid values: · Predicted Max Capacity Overwrites Initial Max Capacity: uses the predicted value as the maximum value for forecast tasks. · Initial Max Capacity Overwrites Predicted Max Capacity: uses the initial maximum capacity as the maximum value for forecast tasks. · Increase Predicted Max Capacity by Specified Ratio: increases the predicted value with a specified ratio before comparing it with the initial maximum capacity. When you select this option, the Increase Ratio field appears for you to set the ratio. The default value is Predicted Max Capacity Overwrites Initial Max Capacity.
Increase Ratio	The ratio of the increment to the predicted value. This field appears when you set Max Capacity Behavior to Increase Predicted Max Capacity by Specified Ratio. If the predicted value increased with this ratio is greater than the initial maximum capacity, the predicted value after increase is used as the maximum value for forecast tasks. The default value is 0, and the maximum value is 100.

Parameter	Description
Scheduled Task Buffer Time (minutes)	The amount of buffer time ahead of the forecast task execution time. By default, all scheduled tasks that are automatically created for a predictive scaling rule are executed at the beginning of each hour. You can set a buffer time to execute forecast tasks ahead of schedule, so that resources can be prepared in advance. The default value is 0, and the maximum value is 60.

3.6 Execute a scaling rule

This topic describes how to execute scaling rules either manually or automatically to scale ECS instances.

Prerequisites

Before you execute a scaling ruling, note that:

- The status of the scaling group to which the scaling rule belongs is in the Active state.
- The scaling group to which the scaling rule belongs is not undergoing any scaling activities.
- Target tracking scaling rules can only be executed by alert tasks that were automatically created. For more information, see Create a scaling rule.
- There is no limit to the maximum number of ECS instances that a scaling group can have. However, the limits on ECS instance usage apply to Auto Scaling. For more information, see Limits.

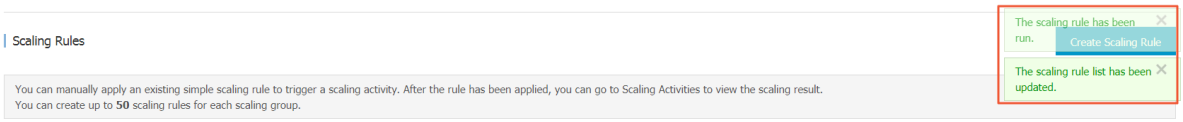
Manually execute a scaling rule

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

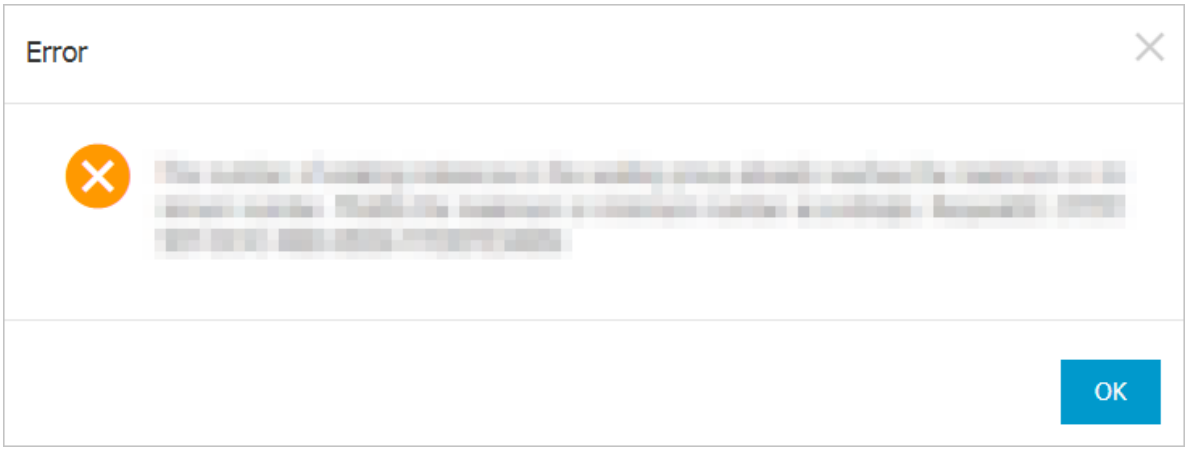
> Note:
> If the scaling group is not undergoing any scaling activities, you can immediately execute the scaling rule without the need to wait for the cooldown period to expire.

1. On the Scaling Rules page, click Execute in the Actions column corresponding to the scaling rule that you want to execute.
2. In the Run Scaling Rule message that appears, click OK.

3. If the scaling rule is executed, a prompt appears in the upper-right corner of the page.



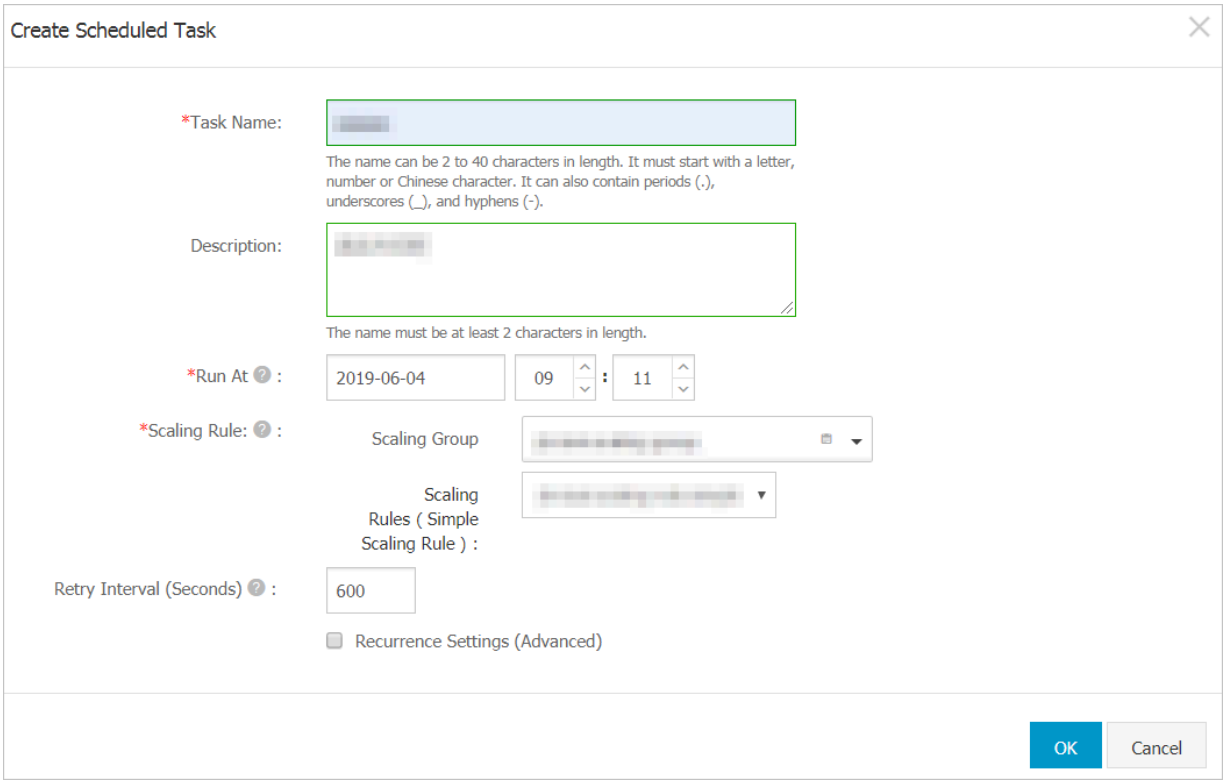
If the scaling rule fails to be executed, an error message appears in the center of the page.



4. You can go to the Scaling Activities page to view the results of the scaling rule execution.

Execute a scaling rule by using a scheduled task

There are services that use ECS instances on a regular basis. For these services, you can specify a scaling rule when you create a scheduled task. Then, Auto Scaling executes this scaling rule at the scheduled points in time.



Execute a scaling rule by using an alert task

There are services that do not use ECS instances on a regular basis. For these services, you can specify a scaling rule when you create an alert task. Then, Auto Scaling automatically executes this scaling rule when the conditions specified in the alert task are met.

Alert tasks include system monitoring alert tasks and custom monitoring alert tasks, which meet monitoring requirements in different scenarios. For more information, see Auto Scaling alert tasks.

Create Event-Triggered Task

Before an event-triggered task can be performed, the latest version of CloudMonitor Agent must be installed on the ECS image.View Help Documentation

*Task Name:

The name can be 2 to 40 characters in length. It must start with a letter, number or Chinese character. It can also contain periods (.), underscores (_), and hyphens (-).

Description:

The name must be at least 2 characters in length.

*Resource Monitored:

Monitoring Type System Monitoring Custom Monitoring

*Metric: CPU

We recommend that you use target tracking scaling rules. A target tracking scaling rule adds or removes capacity to keep the metric close to the target value, and also allows you to use the average value of the selected metric as a monitoring metric. This facilitates easy use and clarity.

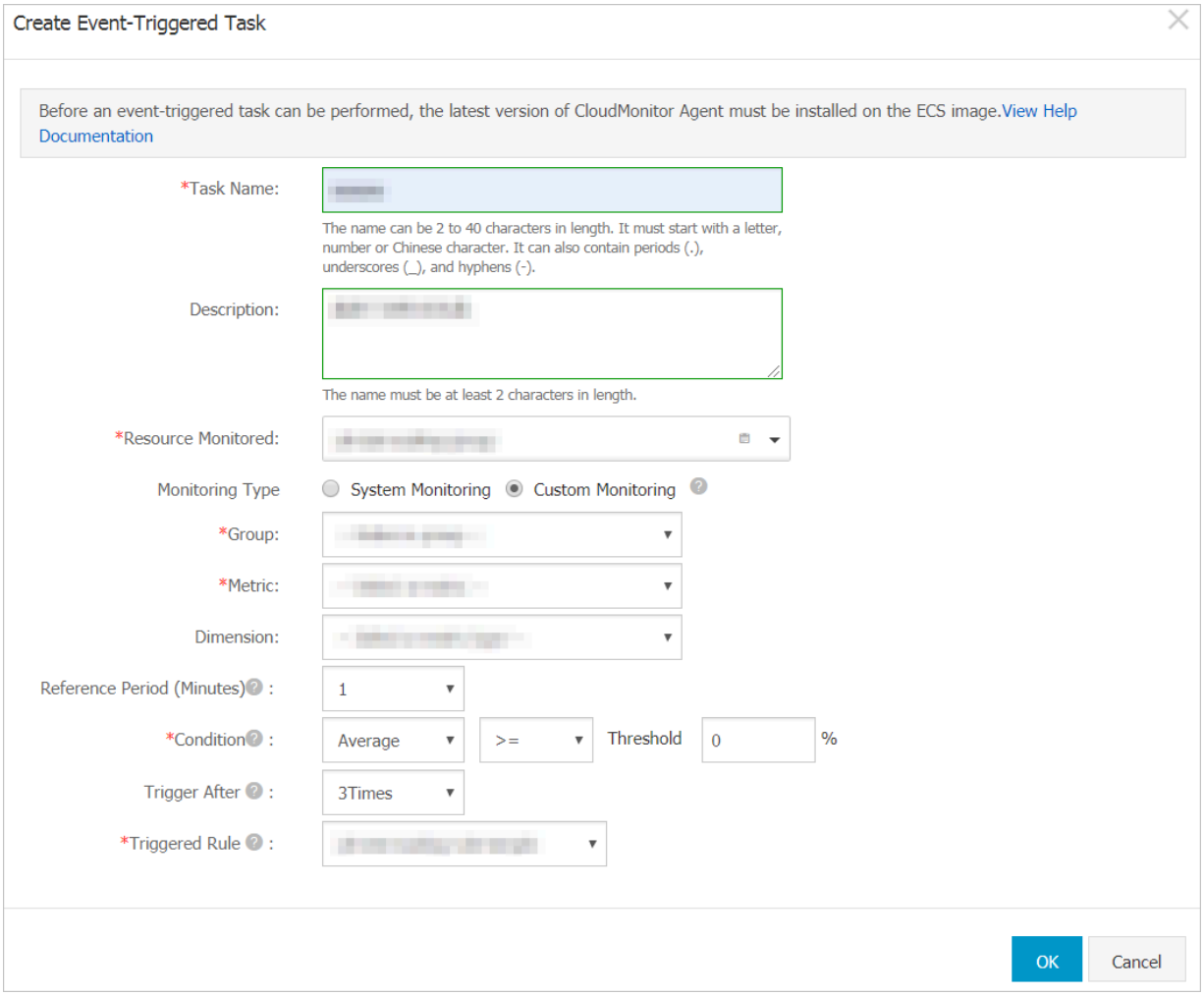
Reference Period (Minutes): 1

*Condition: Average >= Threshold 0 %

Trigger After: 3Times

*Triggered Rule:

OK Cancel



3.7 Scheduled tasks

3.7.1 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.

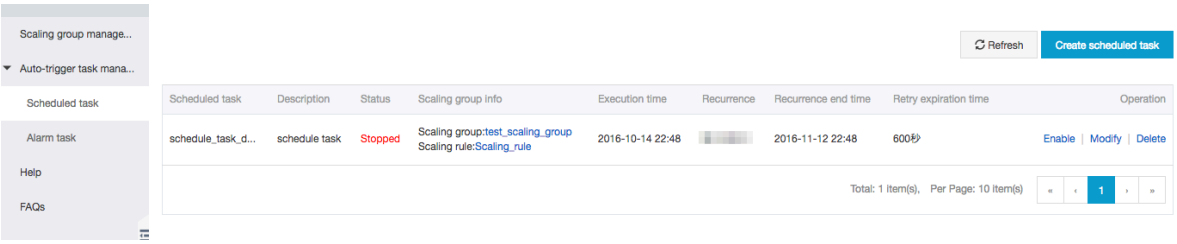
> **Note:**
> The number of scheduled tasks that can be created under one account is limited. See quantity restrictions.

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.

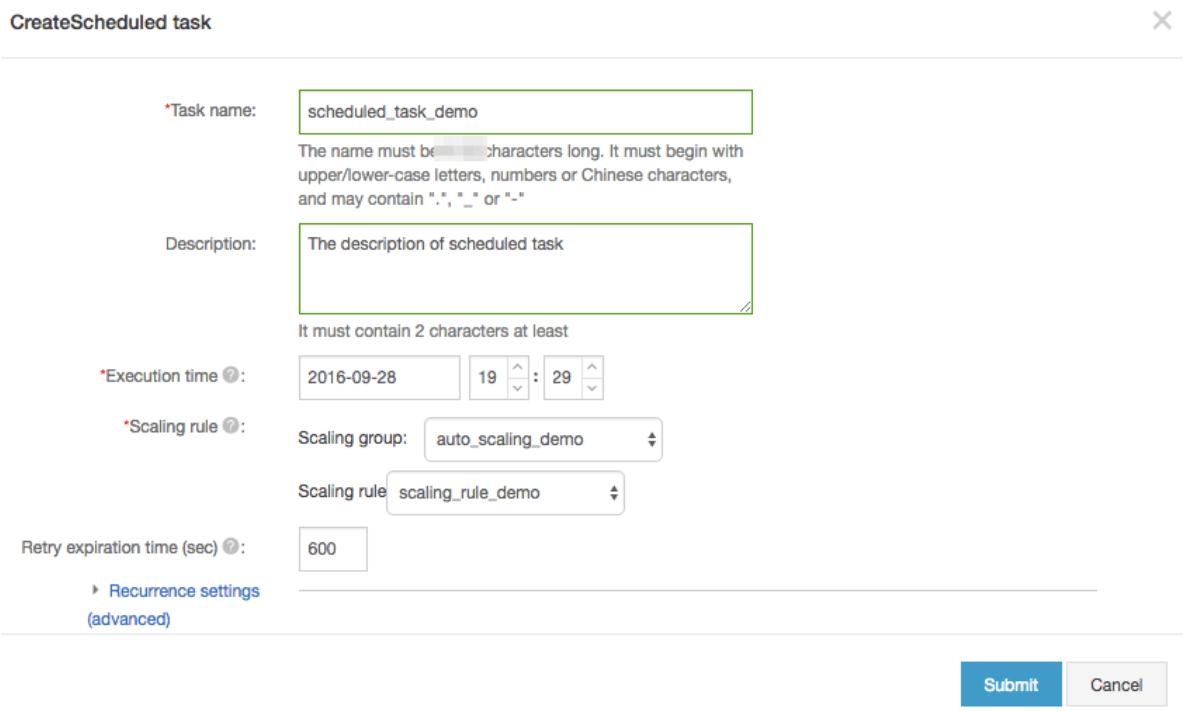
Procedure

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 2-64 characters. It must begin with a lower-case letter, number, or a Chinese character. It can contain ".", "_", or "-".	st-yk20180830****
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
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, inclluding 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]
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.

3.7.2 Manage scheduled tasks

After you create a scheduled task, you can enable, disable, modify, and delete a scheduled task at any time based on business changes.

Enable or disable a scheduled task

After you create a scheduled task, the task is automatically enabled. You can manually disable tasks that are currently not required.

1. Log on to the Auto Scaling Console.

2. Choose Scaling Tasks > Scheduled Tasks.
3. On the Scheduled Tasks page, click Disable in the Actions column.
4. In the Disable Scheduled Task dialog box, click Confirm.

> Note:
> After you disable a scheduled task, in the Actions column, the button changes from Disable to Enable. You can re-enable a disabled scheduled task in the same way you disable the scheduled task.

Modify a scheduled task

When the requirement of a scheduled task changes, such as execution time, scaling rule, or reference period, you can modify the task rather than creating a new one.

> Note:
> This step only describes where to modify a scheduled task. For more information about configuration items, see Create a scheduled task.

1. Log on to the Auto Scaling Console.
2. Choose Scaling Tasks > Scheduled Tasks.
3. On the Scheduled Tasks page, click Edit in the Actions column.
4. In the Edit Scheduled Task dialog box, modify the required options, then click OK.

Delete a scheduled task

You can delete scheduled tasks that are no longer required to release more scheduled task quotas.

1. Log on to the Auto Scaling Console.
2. Choose Scaling Tasks > Scheduled Tasks.
3. On the Scheduled Tasks page, click Delete in the Actions column.
4. In the Delete Scheduled Task dialog box, click Confirm.

3.8 Alarm tasks

3.8.1 Auto Scaling alarm tasks

This topic introduces Auto Scaling alarm tasks.

Auto Scaling alarm tasks integrate some functions of Auto Scaling and CloudMonitor . These tasks allow you to manage scaling groups in a manner similar to Auto Scaling

scheduled tasks. The alarm tasks trigger user-defined scaling rules to execute scaling activities, adjusting the number of instances in scaling groups.

You can use scheduled tasks to execute specific scaling rules at specifics point in time. In scenarios where the time of traffic changes is predictable, scheduled tasks are sufficient to respond to such changes in advance. However, in scenarios where traffic is not predictable or for sudden spikes in traffic, scheduled tasks are insufficie nt to deal with the changes. In this case, alarm tasks provide more flexibility in triggering scaling rules. Alarm tasks can be used to increase the number of instances in a scaling group during peak hours to suffice business requirements, and release instances in the scaling group during off-peak hours to reduce production costs.

Alarm tasks collect measurements from specific metrics in real time. When a measurement meets user-defined alarm conditions, an alarm is triggered and the specified scaling rule is executed. Alarm tasks adjust the number of instances in a scaling group in real time based on business changes, ensuring that monitored metrics stay within a user-defined range.

Auto Scaling alarm tasks allow the dynamic change of the number of instances in a scaling group by monitoring specific metrics. Through the tasks, specified scaling rules are executed in real time based on business changes to adjust the number of instances in a scaling group.

Updated version of Auto Scaling alarm tasks

Auto Scaling alarm tasks have been comprehensively optimized in the scope, method , and response time of monitoring. The features of the updated version allow you to use alarm tasks to manage scaling groups in a more comprehensive and reliable manner.

The new features are as follows:

- You can configure alarm tasks for the system disks, NICs, and TCP connections.
- You can set the data collection interval as short as one minute to perform monitoring functions at a finer granularity.
- You can use the new custom monitoring function, which provides you with a standard way to integrate your own monitoring system with Auto Scaling alarm tasks.

You can use more metrics in the updated version. In addition to the metrics provided in earlier versions, you can add your custom metrics to customize alarm tasks. These

custom metrics enhance the capabilities of Auto Scaling alarm tasks to meet a range of diverse scenarios.

3.8.2 System monitoring alarm tasks

This topic introduces system monitoring alarm tasks.

Metrics in system monitoring alarm tasks are monitoring data collected from ECS instances by CloudMonitor. ECS instances are monitored at the scaling group level, meaning that the measurement of a certain metric in a scaling group is the average value of the metric measurements for all ECS instances in a scaling group. When the number of ECS instances in the scaling group changes, the metrics are also updated accordingly.

Supported metrics

The following table lists the metrics supported by system monitoring alarm tasks.

Metric	Unit	Applicable network
CPU	%	Classic network and VPC
Memory	%	Classic network and VPC
Average system load	None	Classic network and VPC
Internal network outbound traffic	KB/min	Classic network and VPC
Internal network inbound traffic	KB/min	Classic network and VPC
Total TCP connections	N/A	Classic network and VPC
Established TCP connections	N/A	Classic network and VPC
System disk reads measured in bit/s	Bit/s	Classic network and VPC
System disk writes measured in bit/s	Bit/s	Classic network and VPC
System disk read IOPS	Times/s	Classic network and VPC
System disk write IOPS	Times/s	Classic network and VPC
Packets sent by internal network NICs	Packets/s	Classic network and VPC
Packets received by internal network NICs	Packets/s	Classic network and VPC

Metric	Unit	Applicable network
External network outbound traffic	KB/min	Classic network and VPC
External network inbound traffic	KB/min	Classic network and VPC
Packets sent by external network NICs	Packets/s	Classic network
Packets received by external network NICs	Packets/s	Classic network

Notes

- A scaling group can only execute one scaling activity at a time. When a scaling activity is being executed, the scaling group rejects all other scaling activities generated by scaling rules triggered by alarm tasks.
- The cooldown period of a scaling rule affects how the rule is triggered by Auto Scaling alarm tasks. When the cooldown period of a rule has not yet expired, Auto Scaling will not execute that rule. After ECS instances have been added to a scaling group, the systems of the instances are started, configured, and have businesses deployed on them. This process takes several minutes, during which monitoring data for the instances is not recorded. Because of this, you must set an appropriate cooldown period based on your specific business to prevent scaling rules from being triggered repeatedly.
- Each Auto Scaling alarm task has a default cooldown period of one minute and scaling rules cannot be re-triggered during this period.
- You must install the CloudMonitor client to collect system metrics such as memory, load, the number of packets sent by NICs, and the number of TCP connections. When you need to set an alarm task for metrics collected by the CloudMonitor client, the client is automatically installed on all instances belonging to the scaling group associated with the alarm task. At the same time, `CloudMonitor auto installation for newly purchased ECS instances` is automatically enabled on the CloudMonitor console so that the client will also be installed on newly purchased ECS instances.

3.8.3 Custom monitoring alarm tasks

This topic introduces custom monitoring alarm tasks in Auto Scaling.

The monitored objects of a custom monitoring alarm task are the metrics that you choose to be reported to CloudMonitor. In some scenarios, system metrics may not contain the metrics you need. You may have your own monitoring system and be concerned with some metrics related to your specific business. By using custom monitoring alarm tasks, you can import custom metrics specific to your business from your own monitoring system to CloudMonitor to create alarm tasks.

Custom monitoring alarm tasks in Auto Scaling are associated with custom metrics in Alibaba Cloud CloudMonitor. Therefore, you must report custom monitoring data (custom metrics) to CloudMonitor before you can use custom monitoring alarm tasks . CloudMonitor custom monitoring is a service that allows you to define metrics and alarm rules as needed. With this service, you can monitor target metrics, report collected monitoring data to CloudMonitor for processing, and set alarm rules for these metrics.

Report monitoring data to CloudMonitor

CloudMonitor custom monitoring allows you to report monitoring data. You can report time-series data that you have collected to CloudMonitor. The reported data is called a time sequence. CloudMonitor allows you to report data through open APIs, Java SDKs, and the Alibaba Cloud command line interface (CLI). The following section describes how to report monitoring data by using Java SDKs. For more information, see #unique_72.

Before using a Java SDK, you must import the JAR file containing the SDK to a project . If you use Apache Maven to manage a project, you only need to add the following dependency to the project:

```
< dependency >
    < groupId > com . aliyun </ groupId >
    < artifactId > aliyun - java - sdk - core </ artifactId >
    < version > 3 . 2 . 6 </ version >
</ dependency >
< dependency >
    < groupId > com . aliyun . openservic  es </ groupId >
    < artifactId > aliyun - cms </ artifactId >
    < version > 0 . 2 . 4 </ version >
```

```
</ dependency >
```

You can run the following commands to report custom metrics to CloudMonitor:

```
static   String   endPoint      = " https :// metrichub - cms - cn -
hangzhou . aliyuncs . com ";
CMSClient   cmsClient  =  new   CMSClient ( endPoint ,  accAutoSca
lingKey ,  accAutoSca  lingSecret );
CustomMetr  icUploadRe  quest   request  =  CustomMetr  icUploadRe
quest . builder ()
                   . append ( CustomMetr  ic . builder ()
                   . setMetricN  ame (" myCustomMe  tric ")// Set
the   custom   metric   name .
                   . setGroupId ( 54504L )// Set   the   group   ID .
                   .  Settime  ( new   date  () // Set   the   time .
                   . setType ( CustomMetr  ic . TYPE_VALUE )// Set
the   type   to   original   value .
                   . appendValu  e ( MetricAttr  ibute . VALUE ,
number )// Add   an   original   value .  The   key   must   be
original   values .
                   . appendDime  nsion (" key1 ", " value1 ")// Add
a   dimension .
                   . appendDime  nsion (" key2 "," value2 ")
                   . build ())
                . build ();
            CustomMetr  icUploadRe  sponse   response  =  cmsClient .
putCustomM  etric ( request );// Report   data .
```

The preceding example shows how to report a metric to CloudMonitor. When reporting a metric, you must specify the groupId parameter that represents the application group in CloudMonitor. The application group can be a group that you have created in CloudMonitor or a group that does not exist. You can create application groups and view their details on the Application Groups page of the CloudMonitor Console. The reported custom metrics (also called time sequences) are displayed on the Custom Monitoring page.

We recommend that you push custom monitoring data to an existing application group in CloudMonitor to increase the flexibility of CloudMonitor and other functions . An application group in CloudMonitor is a logical group of multiple cloud services. You can also choose to push data to any group regardless of existing groups.

CloudMonitor automatically aggregates reported monitoring data. If you need to report large amounts of data to CloudMonitor, you can also aggregate the data locally before reporting it. For more information, see #unique_72.

Limits

CloudMonitor has the following limits on user-reported monitoring data:

- The QPS of an Alibaba Cloud account is limited to 100.

- A maximum of 100 data records can be reported at a time. The maximum body size is 256 KB.
- The `metricName` field can contain only letters, numbers, and underscores (_). It must start with a letter. If the field starts with a character other than a letter, the character will be replaced with the uppercase letter A. If the field contains characters other than letters, numbers, and underscores (_), the characters are invalid and will be replaced with underscores (_) .
- The `dimensions` field cannot contain equal signs (=), ampersands (&), or commas (,). Such characters are invalid and will be replaced with underscores (_).
- Each key-value pair in the `metricName` and `dimensions` fields can contain a maximum of 64 characters. If the key-value pair exceeds 64 characters, it will be truncated.

3.8.4 Create event-triggered tasks

This section describes how to create and customize event-triggered tasks.

Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, choose Scaling Tasks > Event-Triggered tasks.
3. On the Event-Triggered tasks page, click Create Event-Triggered Task.
4. In the Create Event-Triggered Task dialog box, configure the required options.
   a. Enter the Task Name.
   b. Enter the Description.
   c. You must select a group to monitor from the Resource Monitored field.
   d. Select a Monitoring Type.
      - If you select the System Monitoring option, you must select a monitoring metric. For more information about metrics, see Event-triggered tasks of system monitoring.
      - If you select the Custom Monitoring option, you are required to select the Group, Metric and Metric Type that are predefined in CloudMonitor. For

more information about custom metrics, see Event-triggered tasks of custom monitoring.

e. Select a Reference Period (Minutes).

> Note:
> You can select one the following options, such as 1, 2, 5, and 15. Auto Scaling collects, summarizes, and computes data based on the specified reference period. The smaller the granularity, the more easily an event will be triggered. You can select a rational reference period based on your business requirements.

f. Configure the Condition.

> Note:
> You can set a condition to verify whether the value of a metric exceeds the specified threshold. Assume that the condition is met when the CPU usage rate is higher than 80%.
>
> - Average: For all ECS instances in a scaling group, an event will be triggered when the average CPU usage rate is higher than 80%.
> - Max: In a scaling group, for the ECS instance that has the highest CPU usage rate, an event will be triggered when its usage rage is higher than 80%.
> - Min: In a scaling group, for the ECS instance that has the lowest CPU usage rate, an event will be triggered when its usage rage is higher than 80%.

g. Select a Trigger After.

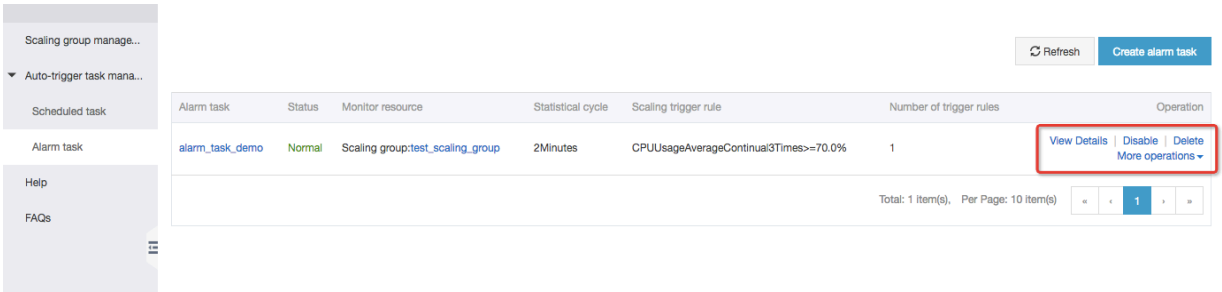
> Note:
> You can select one the following options, such as 1, 2, 3, and 5 times. When the value of a metric exceeds the threshold for specified times, an event is triggered, and the specified scaling rule is applied.

h. Select a Triggered Rule to be applied when the value of a metric meets the specified condition.

5. Click OK.

3.8.5 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.

4 Maintain Auto Scaling

4.1 Check the result of a predictive scaling rule

You can check whether the result calculated based on a predictive scaling rule meets your expectation.

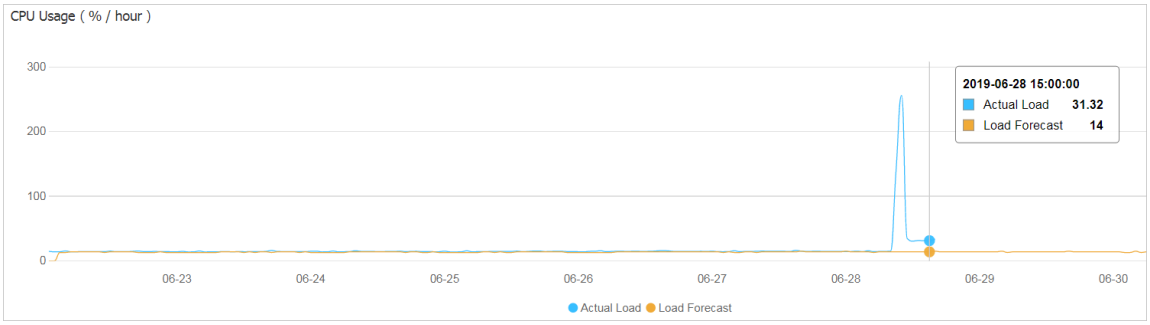
Procedure

1. Log on to the Auto Scaling console.
2. On the Scaling Groups page, locate the row that contains the target scaling group and click Manage in the Actions column.
3. In the left-side navigation pane, click Scaling Rules. On the page that appears, locate the row that contains the target predictive scaling rule and click View Details in the Actions column.

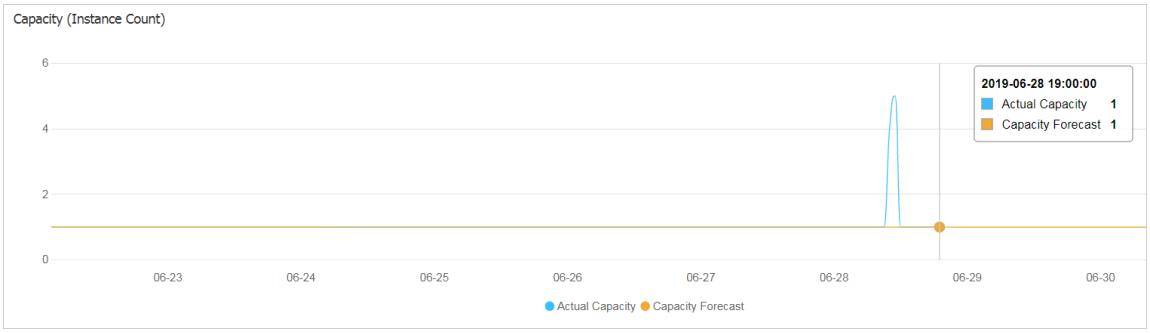
   The Rule Details page shows multiple metrics to help you understand the predicted result. You can enable Forecast and Scale after confirming that the predicted result

meets your expectations. For more information about the operation, see Modify a predictive scaling rule.

- Compare the actual and predicted CPU usage to evaluate the forecast accuracy.



- Compare the actual and predicted number of ECS instances to evaluate the forecast accuracy.



- Check whether the execution results of the scheduled plans generated from forecast 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 you have enabled Forecast and Scale, the system automatically creates forecast tasks for the predictive scaling rule based on the scheduled plans generated from forecast. Forecast tasks are scheduled tasks. You can view details about forecast 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[illegible]e- [illegible]	PredictiveScali...	Running	Created by the predictive scaling rule [illegible]le to modify the minimum and maximum capacities of the scaling group [illegible] to 1 and 100, respectively.	30 June 2019, 10.00	600 Seconds	No results found.	-	Disable | Edit | Delete
PredictiveScaling- [illegible]	PredictiveScali...	Running	Created by the predictive scaling rule [illegible] to modify the minimum and maximum capacities of the scaling group y[illegible] to 1 and 100, respectively.	30 June 2019, 09.00	600 Seconds	No results found.	-	Disable | Edit | Delete

Forecast tasks change the boundary values of the scaling group and are deleted after being successfully executed. You can view details about forecast tasks on the Scaling Activities page.

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

Scaling Activity ID:as[illegible] Status:Successful

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

Cause: A predictive task "PredictiveS[illegible]" 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

4.2 Edit a lifecycle hook

This topic describes how to edit a lifecycle hook.

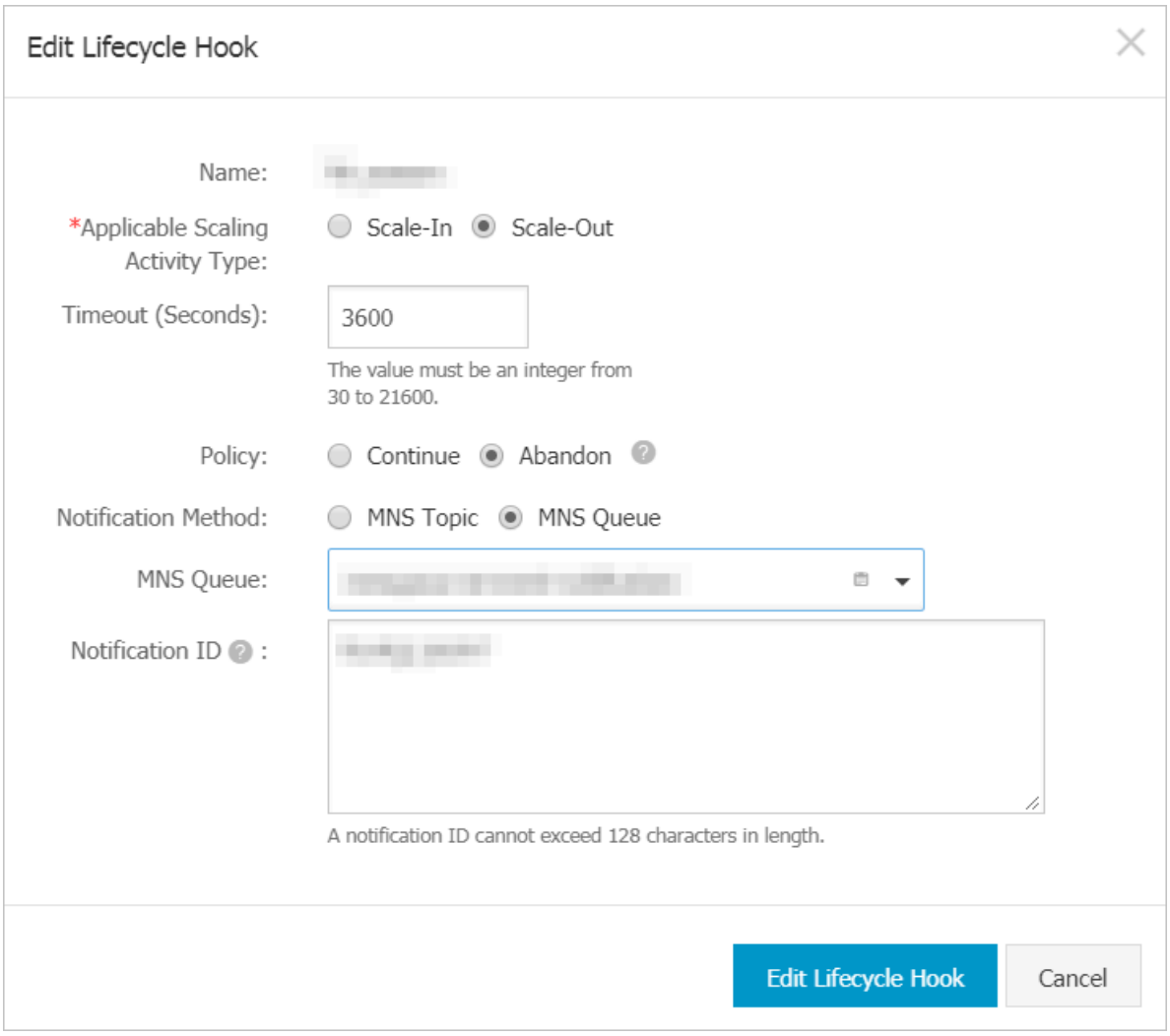
Context

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.

4.3 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.

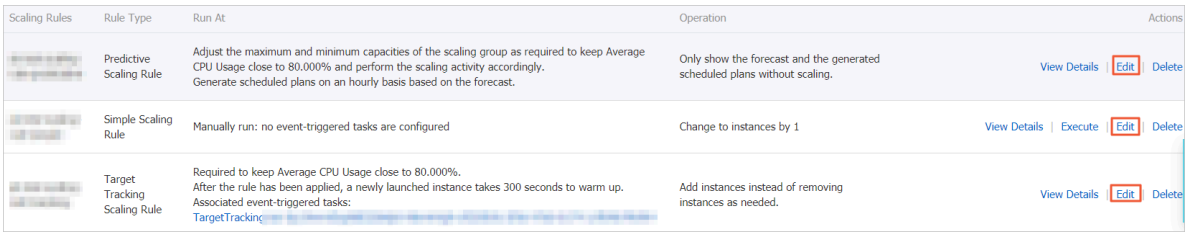
> **Note:**
> Deleting a lifecycle hook also changes the Wait status of the ECS instance that has been paused by the lifecycle hook.

4.4 Modify a scaling rule

After a scaling rule is created, you can modify its attributes as needed to suit your current business requirements. This topic describes how to modify a scaling rule.

Procedure

1. Log on to the Auto Scaling console.
2. On the Scaling Groups page, locate the row that contains the target scaling group and click Manage in the Actions column.
3. In the left-side navigation pane, click Scaling Rules. On the page that appears, locate the row that contains the target scaling rule and click Edit in the Actions column.

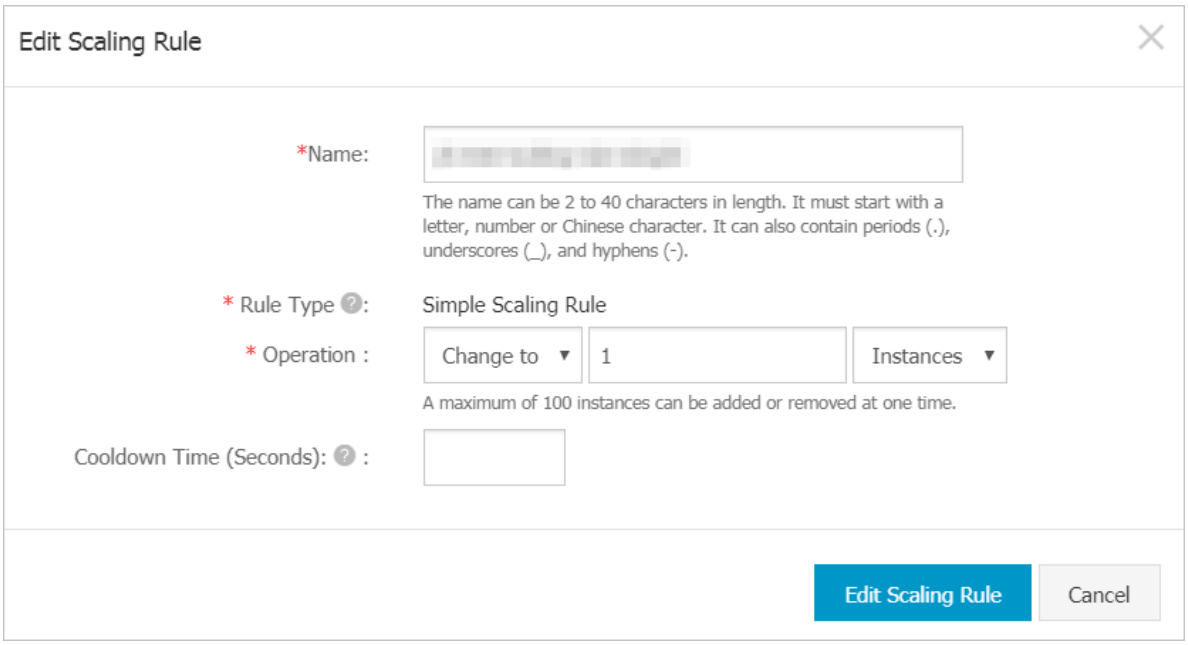


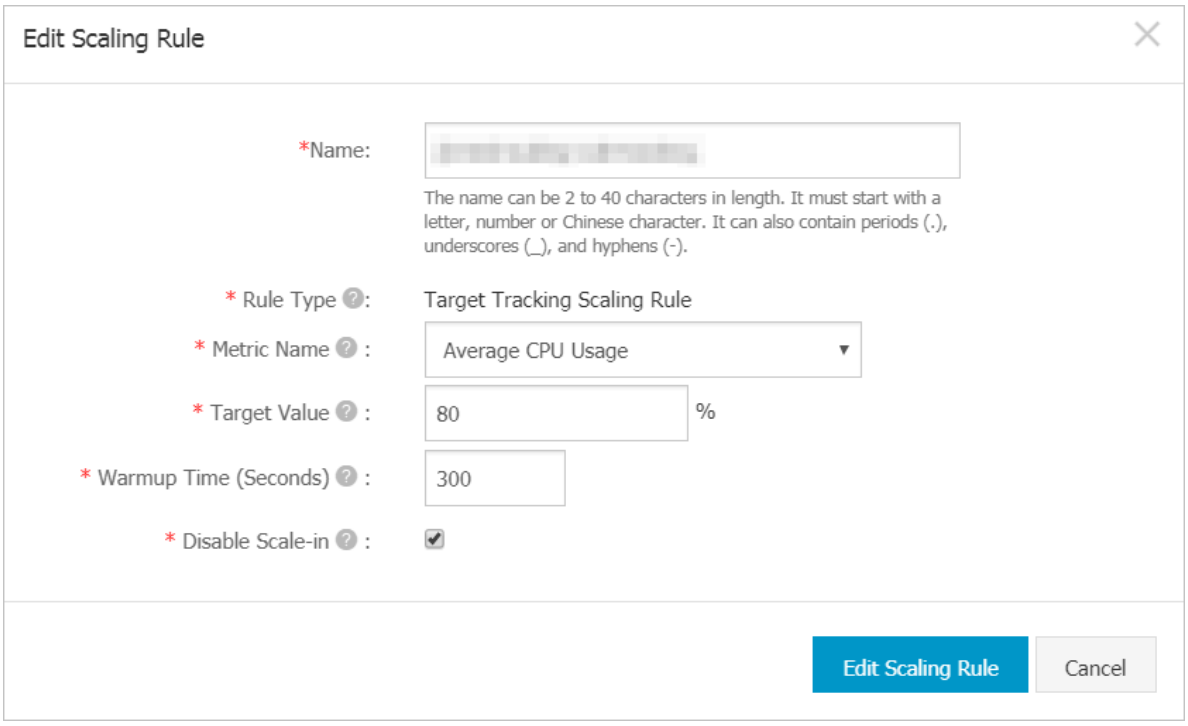
4. Modify the parameters as needed, and then click Edit Scaling Rule.

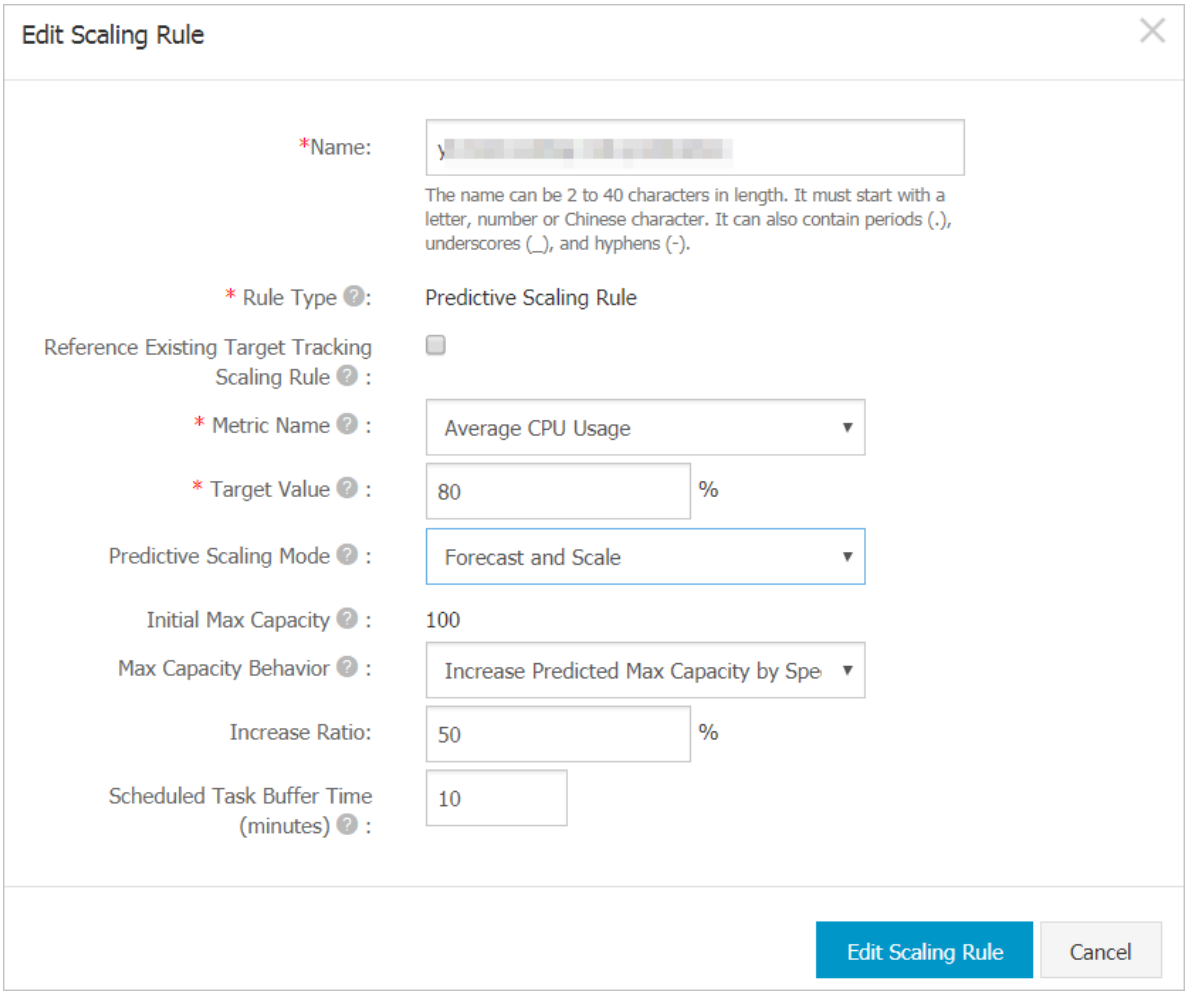
   For more information about these parameters, see parameter tables in Simple scaling rules, Target tracking scaling rules, and Create a predictive scaling rule.

   Note:

Among these parameters, Rule Type of any scaling rule and Initial Max Capacity of a predictive scaling rule cannot be changed.







4.5 Delete a scaling rule

This topic describes how to delete a scaling rule.

Context

You can delete scaling rules that are no longer in use.

Procedure

1. Log on to the Auto Scaling console.
2. On the Scaling Groups page, click Manage in the Actions column corresponding to a specific scaling group.
3. Navigate to the Scaling Rules page. In the Actions column corresponding to the scaling rule to be deleted, click Delete.
4. In the Delete Scaling Rule message that appears, click Confirm.

4.6 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

4.7 Switch to the protected state for an ECS instance

This topic describes how to switch to the protected state for an Elastic Compute Service (ECS) instance.

Procedure

1. Log on to the Auto Scaling console.
2. In the top navigation bar, select a region.
3. Find the scaling group to which the instance belongs and click Manage in the Actions column.
4. In the left-side navigation pane, click ECS Instances.
5. Find the target ECS instance and choose More > Switch to Protected State in the Actions column.
6. Click OK.

Result

An ECS instance in the protected state has the following features:

- If a Server Load Balancer (SLB) instance is attached to the scaling group to which the ECS instance belongs, the load balancing weight of the ECS instance remains unchanged after it is switched to the protected state.
- Auto Scaling cannot automatically release the ECS instance until you remove from the protected state for the ECS instance.
- If a scaling-in event is triggered when the number of ECS instances in a scaling group changes or an alert task detects triggers, Auto Scaling does not automatica lly release the ECS instance in the protected state. You need to remove from the protected state for the ECS instance to release it.
- When the ECS instance is stopped or restarted, its health check status remains unchanged.

4.8 Remove from the protected state for an ECS instance

This topic describes how to remove from the protected state for an Elastic Compute Service (ECS) instance.

Procedure

1. Log on to the Auto Scaling console.
2. In the top navigation bar, select a region.
3. Find the scaling group to which the instance belongs and click Manage in the Actions column.
4. In the left-side navigation pane, click ECS Instances.
5. Find the target ECS instance and click Remove from Protected State in the Actions column.
6. Click OK.

4.9 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).

4.10 Change the status of a scaling group

This topic describes how to change the status of a scaling group.

Context

After you create a scaling group, you can disable scaling groups that are no longer in use, and re-enable them as needed.

Procedure

1. Log on to the Auto Scaling console.
2. On the Scaling Groups page, click More in the Actions column corresponding to the scaling group to be disabled. Choose Disable from the shortcut menu. In the Actions column corresponding to the scaling group to be enabled, click More. Choose Enable from the shortcut menu.
3. Check whether the status in the Status column corresponding to the scaling group has been changed accordingly.

4.11 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.

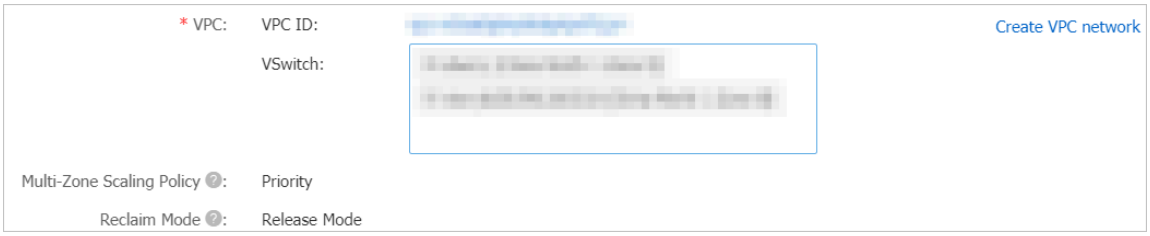
> **Note:**
> This parameter specifies the default scaling activity cooldown time. For more information, see Cool-down time.

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.

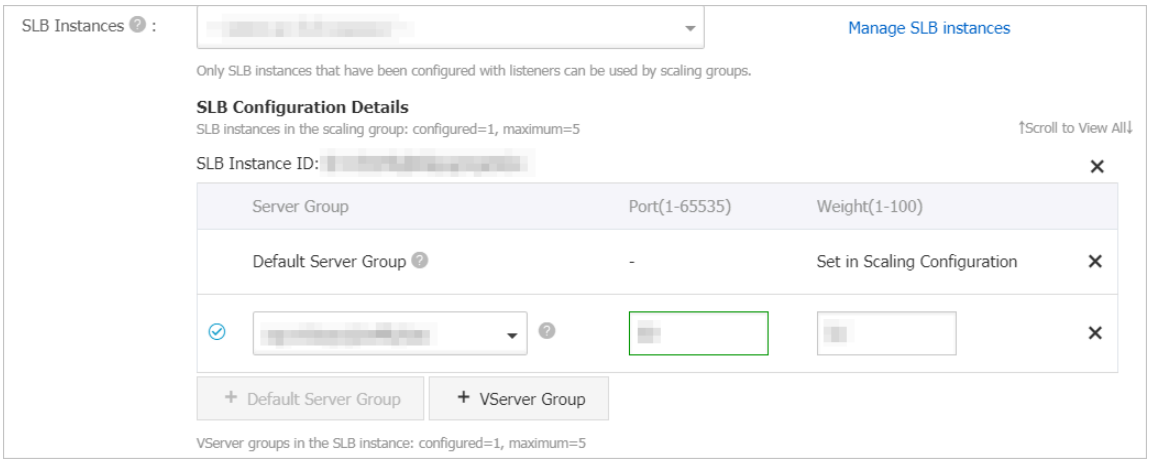
g. (Optional) Once selected, you cannot modify the Network Type of the scaling group. If the Network Type of the scaling group that you need to modify is VPC, then you can change the VSwitch. However, you cannot change the Multi-Zone Scaling Policy or the Reclaim Mode .



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.

4.12 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.

5 Manual scaling

5.1 Add ECS instances

This topic describes how to manually add ECS instances to a scaling group.

Prerequisites

Before manually adding ECS instances, ensure that the instances to be added meet the following conditions:

- The instances are in the same region as the scaling group.
- The instances do not belong to any other scaling groups.
- The instances are in the Running state.
- The instances to be added can be of either the classic network type or VPC type, but you must take note of the following limits:
  - When the scaling group is classic network-connected, only classic network-connected instances can be added to the group.
  - When the scaling group is VPC-connected, only instances in the same VPC as the group can be added to the group.

Before manually adding ECS instances to a scaling group, ensure that the group meet the following conditions:

- The scaling group is in the Enabled state.
- The scaling group is not currently executing any scaling activities.

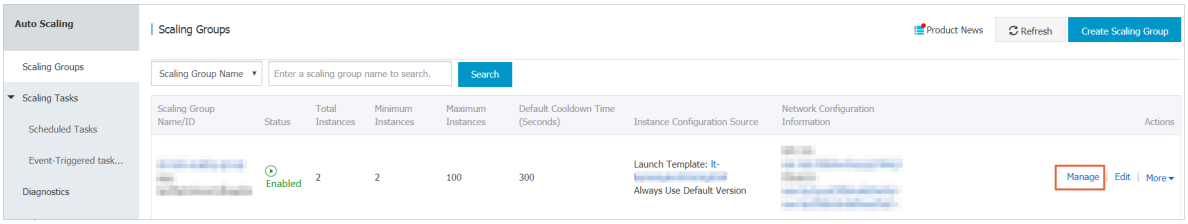
Context

The scaling configuration of the scaling group does not affect whether ECS instances can be manually added. Therefore, you can manually add ECS instances even with in the #unique_43.

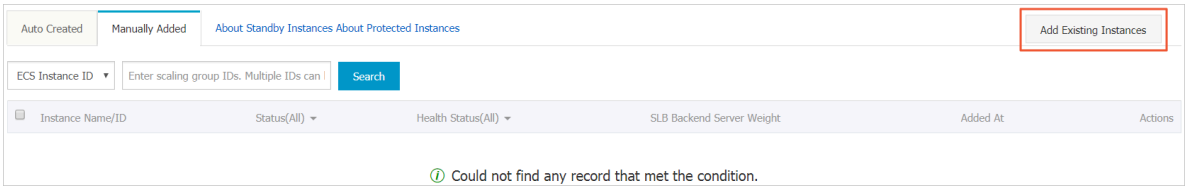
Typically, Auto Scaling scales out ECS instances based on your specifications. However, if the instance inventory is insufficient or the sum of ECS instances to be added and existing ECS instances in the scaling group exceeds the upper limit of the group, the number of actually scaled out instances may be less than what you specified. In these cases, check the scaling group configuration to troubleshoot the issue. If the problem persists, submit a ticket.

Procedure

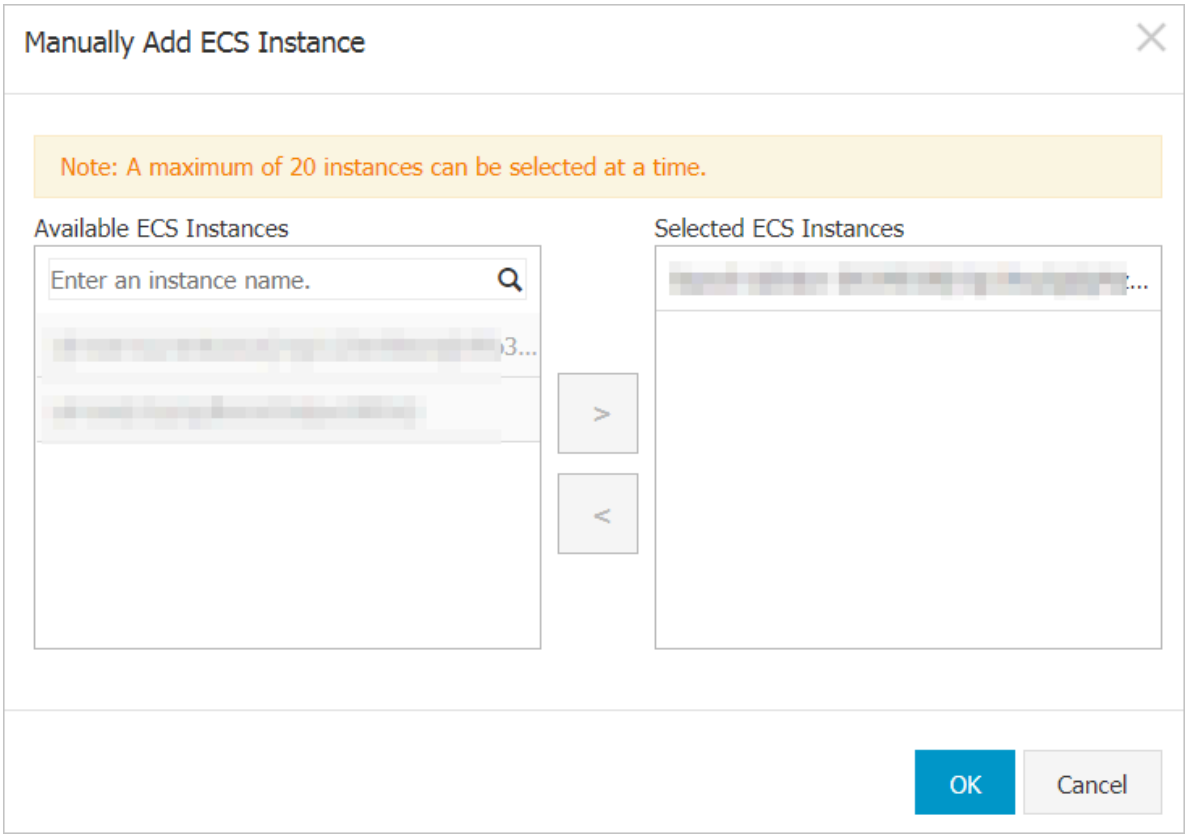
1. Log on to the Auto Scaling console.
2. On the Scaling Groups page, click Manage in the Actions column corresponding to the scaling group to add instances.



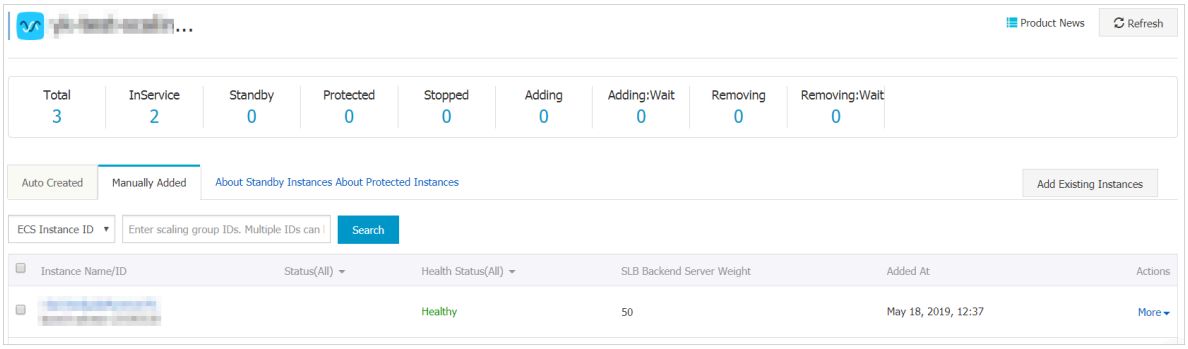
3. In the left-side navigation pane of the displayed page, click ECS Instances. On the page that appears, click Add Existing Instances.



4. In the dialog box that appears, select available ECS instances in the left-side list, click > to add the selected instances to the scaling group, and click OK.



5. Go to the Manually Added tab to view the result.



Note:

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

5.2 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:

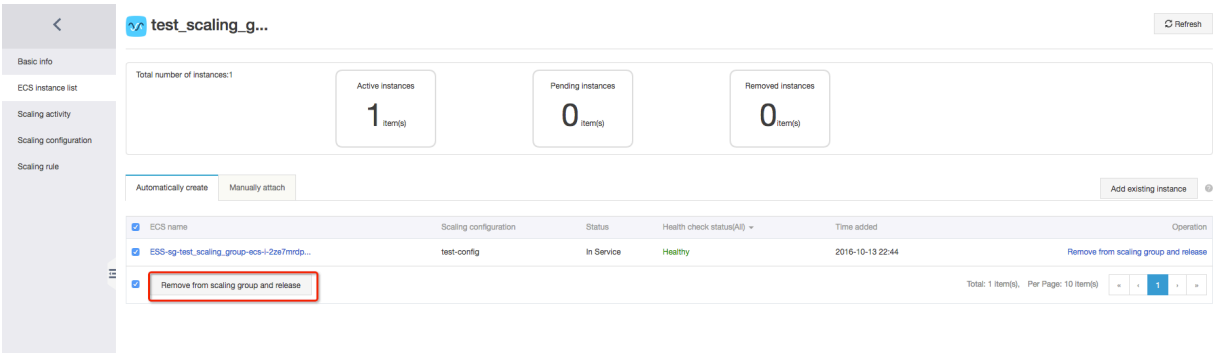
- The scaling group is active.
- The scaling group is not executing any scaling activity.

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


test_scaling_g...
Refresh
Basic info
ECS instance list
Scaling activity
Scaling configuration
Scaling rule
Total number of instances:1
Active instances
1 item(s)
Pending instances
0 item(s)
Removed instances
0 item(s)
Automatically create
Manually attach
Add existing instance
ECS name
Scaling configuration
Status
Health check status(All)
Time added
Operation
ESS-sg-test_scaling_group-ecs-i-2ze7mrdp...
test-config
In Service
Healthy
2016-10-13 22:44
Remove from scaling group and release
Remove from scaling group and release
Total: 1 item(s), Per Page: 10 item(s)
1

6 Event notification

6.1 Event notification overview

The event notification feature is a monitoring method that can automatically send messages to CloudMonitor or Message Service (MNS), providing you with timely information on scaling groups and improving automated management.

Event notification methods

Event notification methods include sending messages to CloudMonitor system events , MNS topics, and MNS queues.

In CloudMonitor, you can query and view statistics on system events of various cloud services, such as Auto Scaling. You can also obtain up-to-date information about scaling groups For more information about the event monitoring feature of CloudMonitor, see Cloud service system event monitoring.

There are two service models in Message Service: MNS topic and MNS queue. Message Service is a distributed message service that helps you easily transfer data and notification messages among distributed components, and build loosely coupled systems. For more information about the functions of MNS topics and MNS queues, see Message Service overview.

- The queue model supports point-to-point sending and receiving of messages. It is designed to deliver a highly reliable and concurrent consumption model in a point-to-point manner. Each message in a queue can only be consumed by a single consumer.
- The topic model supports one-to-many publishing and subscribing of messages. It is designed to provide publishing-subscribing and notification capabilities in a one -to-many manner. The model also allows you to publish messages in various ways.

The following section provides examples of each event notification method. For more information about parameter details, see Create an event notification.

Example: event notifications through CloudMonitor

You have created an event notification in which Notification Method is set to CloudMonitor and Event Types to Successful Scale-Outs and The scale-out activities

for the specified scaling group are running. After a scale-out activity of a scaling group succeeds, CloudMonitor receives an event notification and displays the event. The following figure shows the notification result after the scale-out activity succeeds. Two events are displayed in the results, including The scale-out activities for the specified scaling group are running and The scale-out activities for the specified scaling group are completed.

In the CloudMonitor console, you can view the status of scaling groups and create an alarm rule to notify multiple alarm contacts through SMS messages and emails, improving operation and maintenance efficiency.

Example: event notifications through an MNS topic

You have created an event notification in which Notification Method is set to MNS Topic, and Event Types to Successful Scale-Ins and The scale-in activities for the specified scaling group are running. After a scale-in activity of a scaling group succeeds, the MNS topic receives an event notification and sends it to its subscribers. The following figure shows the notification result after the scale-in activity succeeds. The number displayed in the Messages column corresponding to the MNS topic has increased. You can view the subscribers for message details.

The MNS topic does not allow direct consumption of messages. You must subscribe to the MNS topic through an MNS queue, HTTP request, or email. When the MNS topic receives a message, it pushes the message to subscribers. In this way, multiple subscribers separately consume messages from the same publisher, achieving efficient automated management.

Example: event notifications through an MNS queue

You have created an event notification in which Notification Method is set to MNS Queue, and Event Types to Failed Scale-Outs and The scale-out activities for the specified scaling group are running. After a scale-out activity for a scaling group fails, the MNS queue receives an event notification and allows you to configure the messages for consumption. The following figure shows the notification result after the scale-out activity fails. The number displayed in the Active Message column corresponding to the MNS topic has increased.

You can consume, delay, activate or delete the messages as needed, achieving automated management through event notifications.

6.2 Create an event notification

This topic describes the limits and procedure for creating event notifications.

Limits

- You can create a limited number of event notifications at the same time. For more information, see Quantity limits.
- A receiver in a scaling group must be unique. For example, CloudMonitor, a specific MNS topic, or a specific MNS queue cannot be repeatedly used for different event notifications of a scaling group.
- You must create an MNS topic or MNS queue before using it, and ensure the topic or queue is in the same region as the scaling group.

Create an event notification by using the console

1. Log on to the Auto Scaling console.
2. In the Actions column corresponding to a scaling group, click Manage.
3. In the left-side navigation pane, click Event Notifications.
4. On the Event Notifications page that appears, click Create Notification.
5. In the Create Notification dialog box that appears, set parameters for creating an event notification.

   a. Set Notification Method.

Method	Description
CloudMonitor	If a specific event occurs, an event notification is sent to CloudMonitor. For more information, see Cloud service system event monitoring.
MNS Topic	If a specific event occurs, a message is pushed to an MNS topic.
MNS Queue	If a specific event occurs, a message is pushed to an MNS queue.

   b. Configure Event Types.

   You can select multiple types.

Type	Description
Successful Scale-Outs	An ECS instance is added to the scaling group.

Type	Description
Successful Scale-Ins	An ECS instance is removed from the scaling group.
Failed Scale-Outs	Scale-out activities were triggered but ECS instances failed to be added to the scaling group.
Failed Scale-Ins	Scale-in activities were triggered but ECS instances failed to be removed from the scaling group.
Rejected Scaling Activities	The scaling group received the scaling request but rejected the request because the triggering conditions are not met.
The scale-out activities for the specified scaling group are running	Scale-out activities were triggered and ECS instances are being added to the scaling group.
The scale-in activities for the specified scaling group are running	Scale-in activities were triggered and ECS instances are being removed from the scaling group.
Scheduled Task Expirations	If you select this type, notifications on expiring scheduled tasks of the scaling group will be sent on a daily basis seven days before the task expires. Note: If a scheduled task has a recurring cycle, the task expiration time is the last time the task will be executed.

Note:

A successful scaling activity can be partially or completely successful. To determine whether a scaling activity is partially or completely successful, you can view the scaling detail in the event notification for the activity.

6. Click Create Notification.

6.3 Manage event notifications

This topic describes how to manage event notifications, such as viewing notification method details, modifying notification types, and deleting notifications.

View notification method details

In the Auto Scaling console, you can click links to navigate to the CloudMonitor and Message Service pages. The pages show the notifications received for the relevant services.

1. Log on to the Auto Scaling console.
2. In the Actions column corresponding to a scaling group, click Manage.
3. In the left-side navigation pane, click Event Notifications.
4. On the Event Notifications page that appears, click the link in the Notification Method column corresponding to an event notification.

> Note:
> If the method is CloudMonitor, CloudMonitor is displayed in the column. If the method is MNS Topic or MNS Queue, the topic or queue name is displayed in the column.

Modify event notification types by using the console

> Notice:
> The notification type of a created event notification cannot be modified.

1. Log on to the Auto Scaling console.
2. In the Actions column corresponding to a scaling group, click Manage.
3. In the left-side navigation pane, click Event Notifications.
4. On the Event Notifications page that appears, click Edit in the Actions column corresponding to an event notification.
5. In the Edit Notification dialog box that appears, modify Notification Types.
6. Click Edit Notification.

Modify event notification types by using an API

You can call ModifyNotificationConfiguration to modify notification types of an event notification. Before modifying notification types, you can call DescribeNotificationConfigurations to view event notification details.

Delete an event notification by using the console

1. Log on to the Auto Scaling console.
2. In the Actions column corresponding to a scaling group, click Manage.
3. In the left-side navigation pane, click Event Notifications.
4. On the Event Notifications page that appears, click Delete in the Actions column corresponding to an event notification.
5. In the Delete Notification message that appears, click OK.

Delete an event notification by using an API

You can call DeleteNotificationConfiguration to delete an event notification. Before deleting an event notification, you can call DescribeNotificationConfigurations to view event notification details.

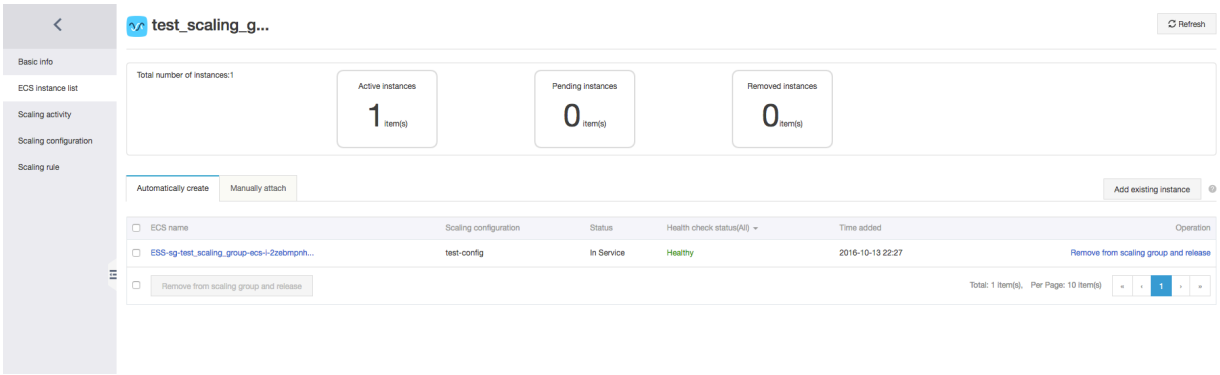
7 Query the ECS instance list

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.



8 View scaling activities

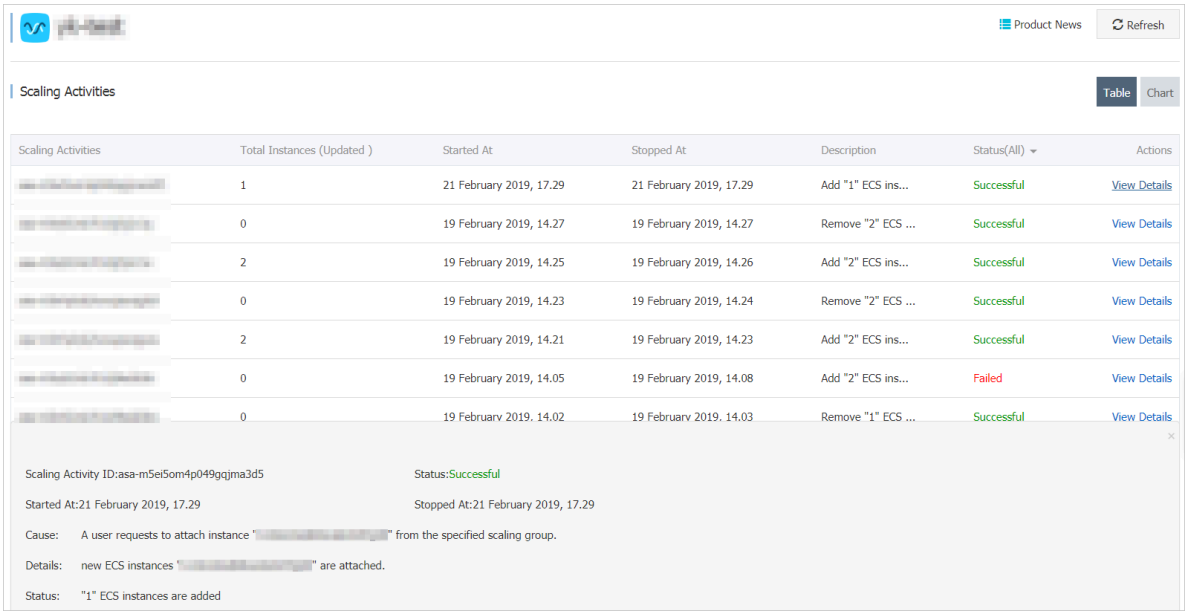
You can view scaling activities in Auto Scaling, to understand the results of the activities triggered by various means, such as scheduled tasks and alarm tasks. This topic describes how to view scaling activities.

Procedure

1. Log on to the Auto Scaling console.
2. On the Scaling Groups page, click Manage in the Actions column corresponding to the scaling group to be viewed.
3. In the left-side navigation pane of the displayed page, click Scaling Activities.

> **Note:**
> You can query information about scaling activities executed within the last 30 days.

4. Click View Details in the Actions column corresponding to a scaling activity to view more information.



You can also call DescribeScalingActivities to view scaling activities.

9 Related services

9.1 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.

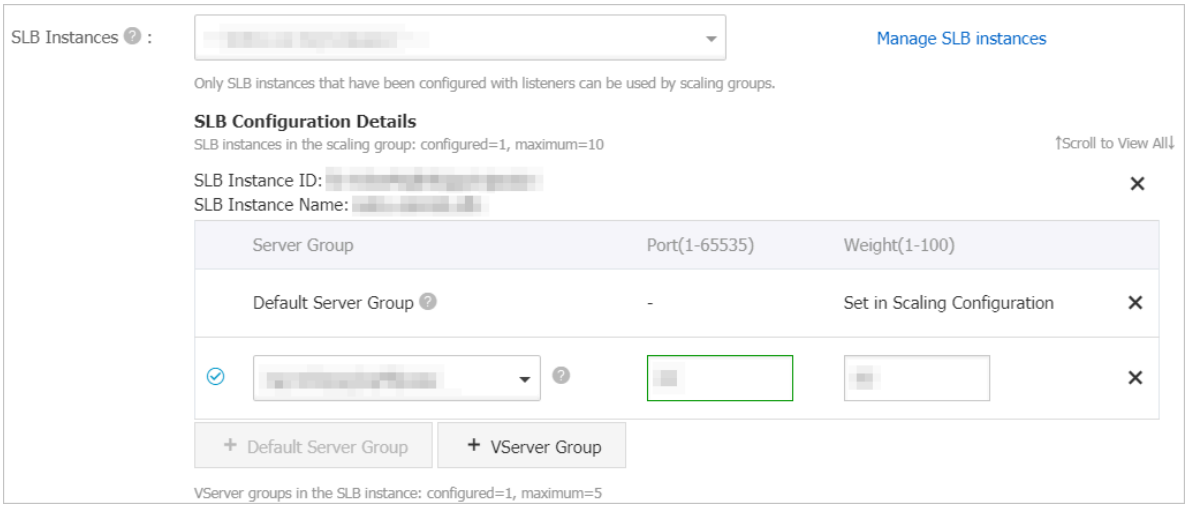
1. Log on to the Auto Scaling console .
2. On the Scaling Groups page, use one of the following methods to add SLB instances:
   - Click Create Scaling Group in the upper-right corner when you create a scaling group.
   - Click Edit in the Actions column when you modify a scaling group.
3. Select a Network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

10 Auto Scaling FAQ

This topic lists FAQs related to Auto Scaling.

- How do I avoid scale-out failures due to an instance type having insufficient inventory?
- Which event takes priority between executing an alert task and executing a scheduled task?
- Can I add an ECS instance to multiple scaling groups?
- Why is an ECS instance in the protected state automatically released from a scaling group?
- How can I prevent Auto Scaling from automatically releasing an ECS instance that I manually added to a scaling group?

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

Specify multiple zones (select VSwitches in different zones) when creating a scaling group, and select multiple Elastic Compute Service (ECS) instance types when creating a scaling configuration. If an instance type is out of inventory in one of the specified zones, Auto Scaling automatically switches to another zone where the instance type has sufficient inventory and executes the scale-out. For more information, see Use custom scaling configurations to create scaling groups or Use launch templates to create scaling groups.

Which event takes priority between executing an alert task and executing a scheduled task?

No priority is given to one task type over the other. At present, only one scaling activity can be executed in a scaling group at a time. If one task triggers a scaling activity earlier than the other task does, the earlier one is executed and the other one is rejected.

If an alert task is rejected and the alert triggers still exist, the alert task is executed after the scheduled task is completed.

You can set the Retry Interval parameter for a scheduled task. This ensures that the scheduled task can be triggered again after it is rejected. For more information, see Create a scheduled task.

Can I add an ECS instance to multiple scaling groups?

No. An ECS instance can be added to only one scaling group.

Why is an ECS instance in the protected state automatically released from a scaling group?

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

To protect the ECS instance from being automatically released, you need to switch to the protected state for the ECS instance in the Auto Scaling console. For more information, see Switch to the protected state for an ECS instance.

How can 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 need to switch to the protected state for the ECS instance in the Auto Scaling console. For more information, see Switch to the protected state for an ECS instance.

11 Scaling groups