Alibaba Cloud Auto Scaling

Monitoring

Issue: 20200525

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

Document conventions

Style	Description	Example
	A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	 **Danger:** Resetting will result in the loss of user configuration data.
	A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	 **Warning:** Restarting will cause business interruption. About 10 minutes are required to restart an instance.
	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	 **Notice:** If the weight is set to 0, the server no longer receives new requests.
	A note indicates supplemental instructions, best practices, tips, and other content.	 **Note:** You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings** > **Network** > **Set network type**.
Bold	Bold formatting is used for buttons, menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands.	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`

Style	Description	Example
{} or {a|b}	This format is used for a required value, where only one item can be selected.	switch {active|stand}

Contents

1 Scaling events

1.1 Overview

A scaling activity is triggered when a scaling rule is executed or when an instance is manually added to or removed from a scaling group. After a scaling activity is triggered, the system performs a scale-in or scale-out action. This topic describes the process of a scaling activity, its status, and instance rollback.

Process of the scaling activity when ECS instances are automatically added or removed

When ECS instances are automatically added to a scaling group after a scaling rule is executed:

1. Check the health status and boundary conditions of the scaling group.
2. Assign the activity ID and execute the scaling activity.
3. Create ECS instances.
4. Modify the number of instances of the scaling group.
5. Allocate IP addresses to the added ECS instances.
6. Add the ECS instances to the whitelist of the ApsaraDB for RDS instance.
7. Start ECS instances.
8. Add the ECS instances to the backend server group of the SLB instance, and set the weights of these ECS instances to the values specified by the scaling configuration of the scaling group.
9. The cooldown period starts after the scaling activity is completed.

When ECS instances are automatically removed from a scaling group after a scaling rule is executed:

1. Check the health status and boundary conditions of the scaling group.
2. Assign the activity ID and execute the scaling activity.
3. Remove ECS instances from the backend server group of the SLB instance.
4. Stop the ECS instances.
5. Remove the ECS instances from the whitelist of the ApsaraDB for RDS instance.
6. Release the ECS instances.
7. Modify the number of instances of the scaling group.

8. The cooldown period starts after the scaling activity is completed.

Process of the scaling activity when existing ECS instances are manually added or removed

When existing ECS instances are manually added to a scaling group:

1. Check the health status and boundary conditions of the scaling group, and check the status and types of ECS instances.
2. Assign the activity ID and execute the scaling activity.
3. Add ECS instances to the scaling group.
4. Modify the number of instances of the scaling group.
5. Add the ECS instances to the whitelist of the ApsaraDB for RDS instance.
6. Add the ECS instances to the backend server group of the SLB instance and set the weights of these ECS instances to the values specified by the active scaling configuration of the scaling group.
7. The cooldown period starts after the scaling activity is completed.

When existing ECS instances are manually removed from a scaling group:

1. Check the health status and boundary conditions of the scaling group.
2. Assign the activity ID and execute the scaling activity.
3. The SLB instance stops forwarding traffic to the ECS instances.
4. Wait 60 seconds, and remove the ECS instances from the backend server group of the SLB instance.
5. Remove the ECS instances from the whitelist of the ApsaraDB for RDS instance.
6. Modify the number of instances of the scaling group.
7. Remove the ECS instances from the scaling group.
8. The cooldown period starts after the scaling activity is completed.

Status of a scaling activity

A scaling activity may undergo the status described in the following table.

Status	Description	Example
Rejected	The scaling activity is rejected in the request phase and does not perform the scale-in or scale-out action.	Scenario: • The maximum number of instances in the scaling group is 100. • The scaling group already has 100 ECS instances. • A scaling rule is executed to automatically create 10 ECS instances. Result: The scaling activity fails the condition check and the system rejects the activity. No subsequent processes are followed. After the scaling activity ends, the number of instances in the scaling group is still 100.
In Process	The scaling activity passes the condition check and is in progress. Auto Scaling automatically scales in or out the ECS instances based on the maximum and minimum numbers of instances in the scaling group.	Scenario: • The maximum number of instances in the scaling group is 100. • The scaling group already has 95 ECS instances. • A scaling rule is executed to automatically create 10 ECS instances. Result: The scaling activity passes the condition check and is performed. Only five ECS instances are automatically created. After the scaling activity ends, the number of instances in the scaling group changes to 100.

Status	Description	Example
Successful	The scaling activity is completed , and all target ECS instances are added to or removed from the scaling group.	Scenario: • The maximum number of instances in the scaling group is 100. • The scaling group already has 90 ECS instances. • A scaling rule is executed to automatically create 10 ECS instances. Result: The scaling activity passes the condition check and is performed. After the scaling activity ends, the number of instances in the scaling group changes to 100.

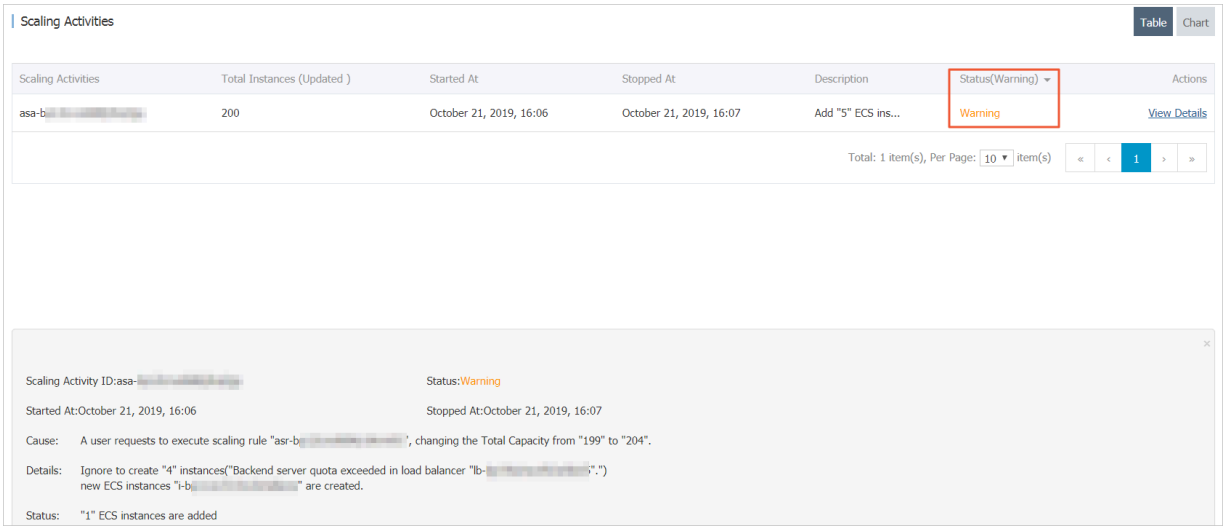
Status	Description	Example
Warning	The scaling activity is completed, and at least one ECS instance is added to or removed from the scaling group, but at least one ECS instance is not added to or removed from the scaling group. An ECS instance is considered to be added to the scaling group if the instance is successfully created, added to the backend server group of the SLB instance, and then added to the whitelist of the ApsaraDB for RDS instance. If any step fails, the instance is not considered to be added to the scaling group. When an instance fails to be added to a scaling group, the instance will be rolled back. For more information, see ECS instance rollback.	Scenario: • The scaling group is associated with an SLB instance. All created ECS instances will be automatically added to the backend server group of the SLB instance. • The quota of the backend servers of the SLB instance is 200. **Note:** For more information, see #unique_5. • The maximum number of instances in the scaling group is 300. • The scaling group already has 199 ECS instances that are added to the backend server group of the SLB instance. • A scaling rule is executed to automatically create five ECS instances. Result: The scaling activity passes the condition check and is performed to create five ECS instances. Because the quota of backend servers is 200, four ECS instances failed to be added to the backend server group, and therefore cannot be added to the scaling group. After the scaling activity ends, only one instance is added to the scaling group. The number of instances in the scaling group is 200.

Status	Description	Example
Failed	The scaling activity is completed, and all target ECS instances fail to be added to or removed from the scaling group.	Scenario: • The instance types specified by the scaling configuration are out of stock in the region where the scaling group resides. • The maximum number of instances in the scaling group is 100. • The scaling group already has 95 ECS instances. • A scaling rule is executed to automatically create five ECS instances. Result: The scaling activity passes the condition check and is performed. The five instances failed to be created due to insufficient resources. After the scaling activity ends, no instances are added to the scaling group. The number of instances in the scaling group is still 95.

ECS instance rollback

When a scaling activity fails to complete, the system prioritizes the integrity of the ECS instances over that of the scaling activity. The system will roll back the ECS instances that failed to be added or removed, but not the scaling activity. Auto Scaling uses Alibaba Cloud Resource Access Management (RAM) to call ECS API operations to create ECS instances. ECS instances that are rolled back still incur fees for the duration of their creation to release.

For example, a scaling group wants to add five ECS instances to the backend server group of an SLB instance. After the five instances are created, only one instance is added to the scaling group, and the other four instances are not added and are automatically released. After the scaling activity ends, the status is **Warning**.



When the ECS instances are rolled back, the scaling group does not reach its expected capacity. This means that the scaling group is unable to provide the required computing power and achieve the required monitoring metrics. In this case, you can use other methods to ensure that the scaling group can provide the required computing power for your business needs. For example, you can manually trigger scaling rules, manually add existing ECS instances to the scaling group, or configure scheduled or monitoring tasks to trigger a scaling activity.

1.2 View the details of a scaling activity

This topic describes how to view the details of a scaling activity. You can view the details of a scaling activity to check the execution result of the activity that is triggered by a scheduled or event-triggered task.

Context

A scaling activity can have the following states: Rejected, In Progress, Successful, Warning, and Failed. For more information, see Status of a scaling activity.

Note:

During a scaling activity, if ECS instances are not all added to a scaling group, Auto Scaling performs rollback on the ECS instances that are not added to the scaling group. After the scaling activity is completed, the scaling activity enters the Warning state. For more information, see ECS instance rollback.

Procedure

1. Log on to the Auto Scaling console.

2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. You can use either of the following methods to open the details page of a scaling group.
   - In the **Scaling Group Name/ID** column, click a scaling group name.
   - Click **Manage** in the **Actions** column corresponding to a scaling group.
5. In the left-side navigation pane, click **Scaling Activities**.
6. Find the target scaling activity and click **View Details** in the **Actions** column.

2 Event notification

2.1 Event notification overview

The event notification feature helps you monitor the scaling activities. It can automatically send messages to CloudMonitor or Message Service (MNS), providing you with timely information on scaling groups to improve automatic management.

Event notification methods

Supported 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 #unique_9.

Message Service offers two service models: 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 features of MNS topics and MNS queues, see Message Service overview.

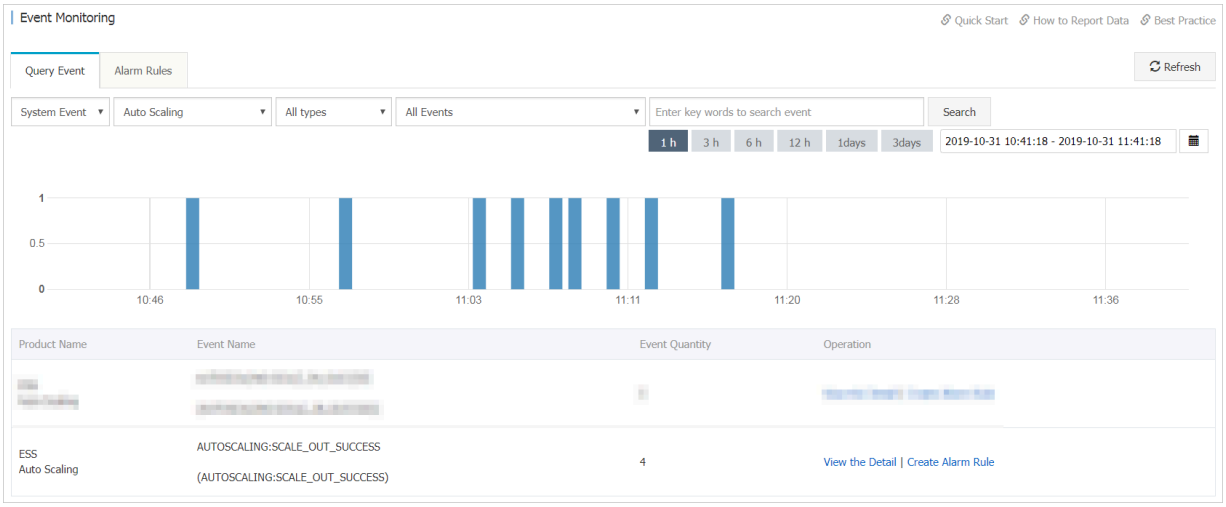
- The MNS queue model supports point-to-point sending and receiving of messages. It is designed to deliver a highly reliable and concurrent consumption model in a point-to-point manner. Each message in a queue can only be consumed by a single consumer.
- The MNS topic model supports one-to-many publishing and subscription of messages. It is designed to publish and subscribe to messages and send notifications 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 the parameter configuration, see Create an event notification.

Example of 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 results of the successful scale-out activity. Two events are displayed in the results, including **The scale-out activities for the specified scaling group are running** and **Successful Scale-Outs**.

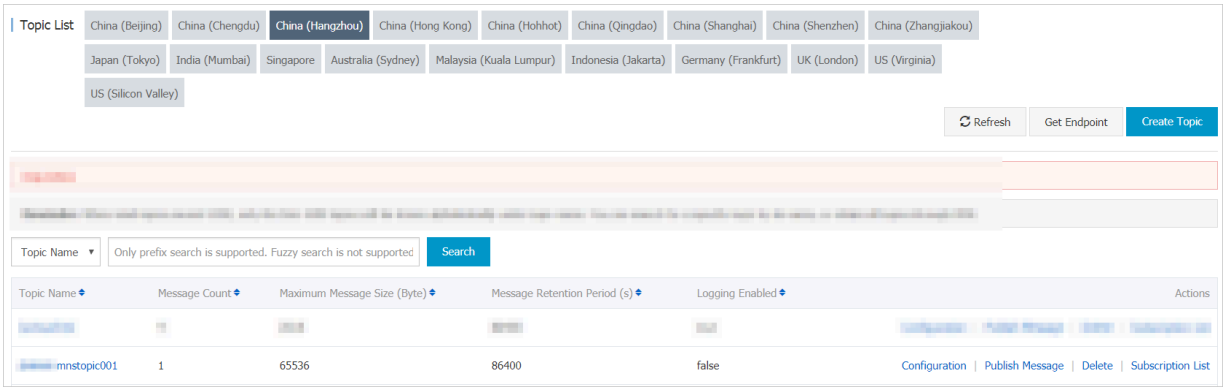


In the CloudMonitor console, you can view the status of scaling groups and create alarm rules to notify multiple alarm contacts through SMS messages and emails. This helps improve operations and maintenance (O&M) efficiency.

Example of 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 specified MNS topic receives an event notification and sends it to its subscribers. The following figure shows the notification results of the successful scale-in activity. The number displayed in the **Message Count** column corresponding to the MNS topic has increased. You can view the subscribers for message details.

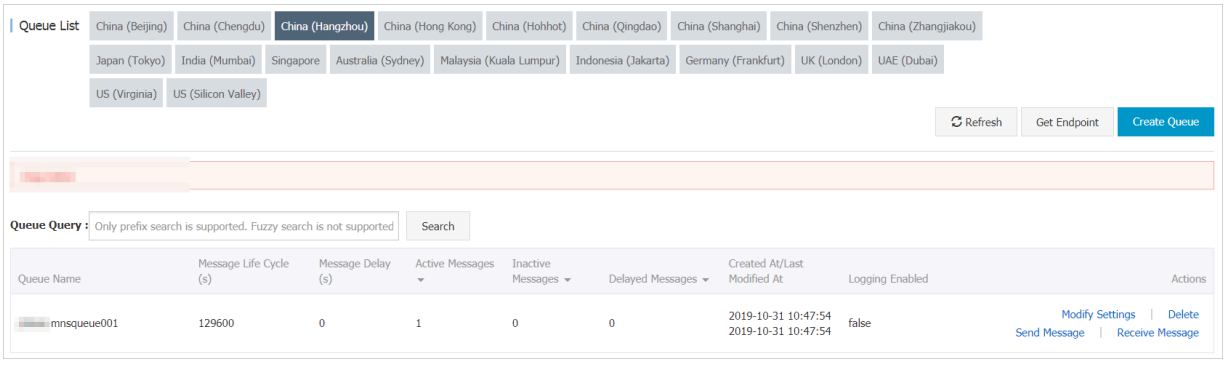
The MNS topic does not allow direct consumption of messages. You must subscribe to the MNS topic through an MNS queue, HTTP request, or email. When the MNS topic receives a message, it pushes the message to subscribers. In this way, multiple subscribers separately consume messages from the same publisher,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



Example of 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 of a scaling group fails, the specified MNS queue receives an event notification and allows you to configure the messages for consumption. The following figure shows the notification results of the failed scale-out activity. The number displayed in the **Active Messages** column corresponding to the MNS queue has increased.

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



2.2 Create an event notification

This topic describes how to create an event notification in a scaling group. After an event of the specified type occurs, Auto Scaling automatically sends a notification to the specified Message Service (MNS) topic, MNS queue, or CloudMonitor.

Prerequisites

If you want Auto Scaling to automatically send messages to an MNS topic or queue, Create a topic or Create a queue in advance. Ensure that the MNS topic or queue belongs to the same region where the scaling group resides.

Context

- Only a limited number of event notifications can be created in a scaling group. For more information, see #unique_12.
- Receivers in a scaling group must be unique. For example, CloudMonitor, the same MNS topic, or the same MNS queue cannot be used for different event notifications in a scaling group.

Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. You can use either of the following methods to open the details page of a scaling group.
   - In the **Scaling Group Name/ID** column, click a scaling group name.
   - Click **Manage** in the **Actions** column corresponding to a scaling group.
5. In the left-side navigation pane, click **Event Notifications**.
6. Click **Create Event Notification**.
7. Configure parameters to create an event notification.

   a) Configure the notification method.

Notification method	Description
CloudMonitor	If a specific event occurs, a notification is sent to CloudMonitor. For more information, see #unique_9.
MNS Topic	If a specific event occurs, a notification is sent to an MNS topic.

Notification method	Description
MNS Queue	If a specific event occurs, a notification is sent to an MNS queue.

b) Configure the event types.

You can select multiple event types.

Event type	Description
Successful Scale-out Event	ECS instances are added to the scaling group.
Successful Scale-in Event	ECS instances are removed from the scaling group.
Failed Scale-out Event	Scale-out events were triggered but ECS instances failed to be added to the scaling group.
Failed Scale-in Event	Scale-in events were triggered but ECS instances failed to be removed from the scaling group.
Rejected Scaling Activity	The scaling group received the scaling request but rejected the request because the triggering conditions are not met.
Start of Scale-out Event	Scale-out events were triggered and ECS instances are being added to the scaling group.
Start of Scale-in Event	Scale-in events were triggered and ECS instances are being removed from the scaling group.
Expiration of Scheduled Task	If you select this type, notifications will be sent on a daily basis for seven days before the scheduled task expires. If you specify the frequency for the scheduled task, the task expiration time is the last time when the task will be executed.

Note:

Scaling activities can be successful or partly successful activities. You can view the details of a scaling activity to check the execution result.

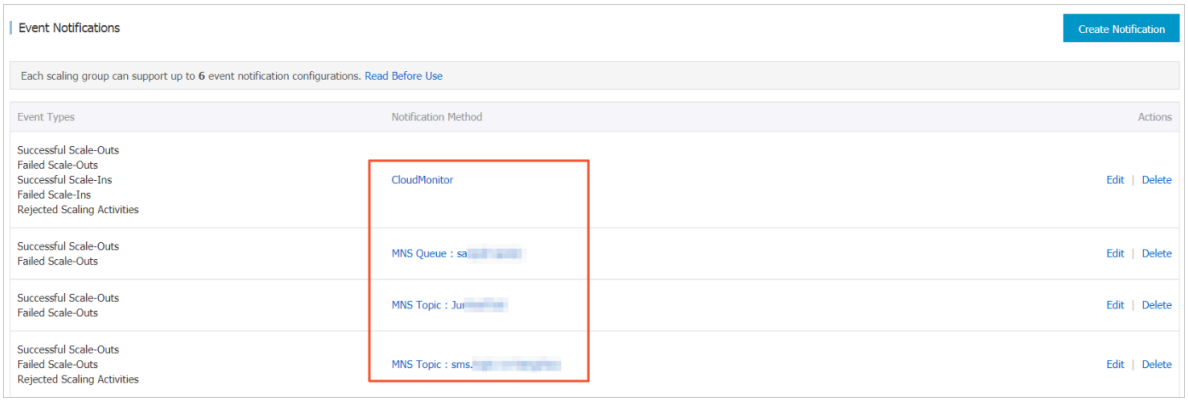
8. Click **Create Event Notification**.

2.3 View an event notification

This topic describes how to view an event notification. In the Auto Scaling console, you can click a link to go the CloudMonitor console or the Message Service (MNS) console to view events or messages.

Procedure

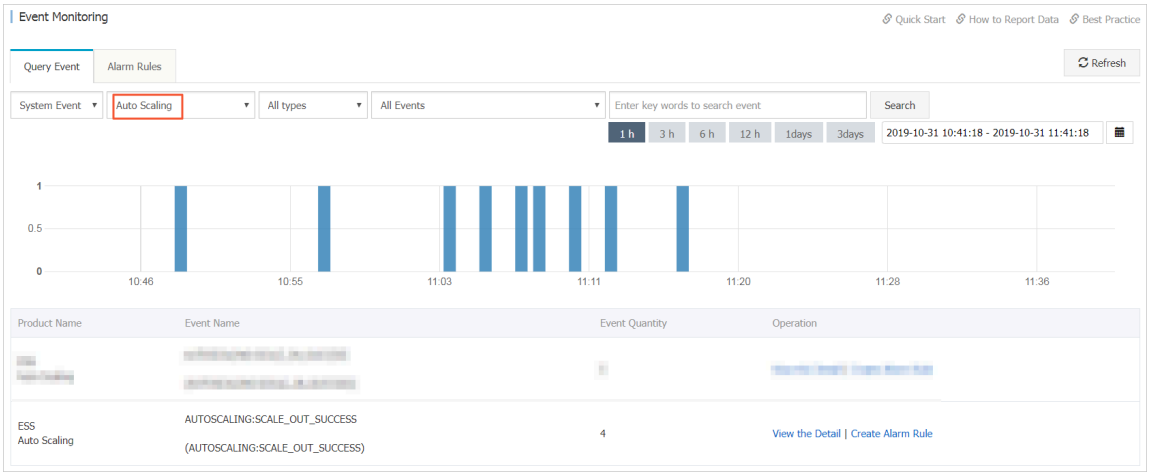
1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. You can use either of the following methods to open the details page of a scaling group.
   - In the **Scaling Group Name/ID** column, click a scaling group name.
   - Click **Manage** in the **Actions** column corresponding to a scaling group.
5. In the left-side navigation pane, click **Event Notifications**.
6. Find the target event notification and click a link in the **Notification Method** column.



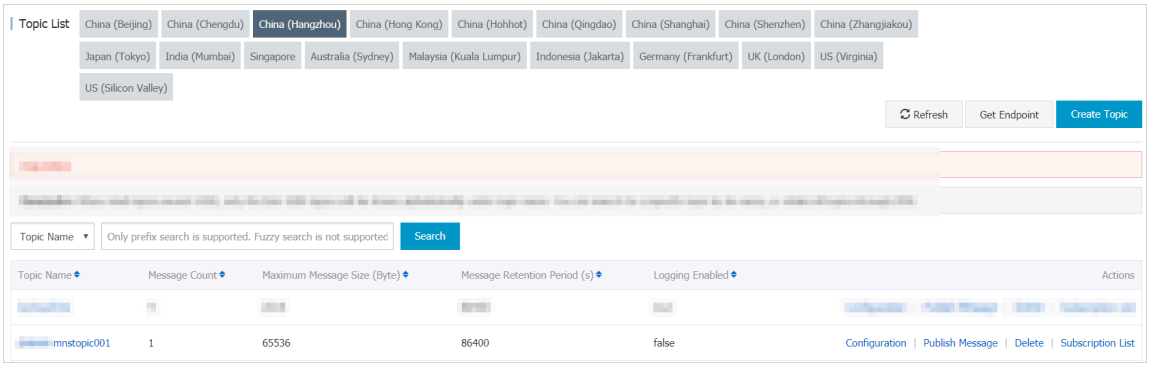
7. View events in CloudMonitor or messages in the MNS topic or queue.

   For more information, see Event notification overview.

   - CloudMonitor: On the **Event Monitoring** page of the CloudMonitor console, select **System Event** and then **Auto Scaling**. All system events related to Auto Scaling appear on the page.

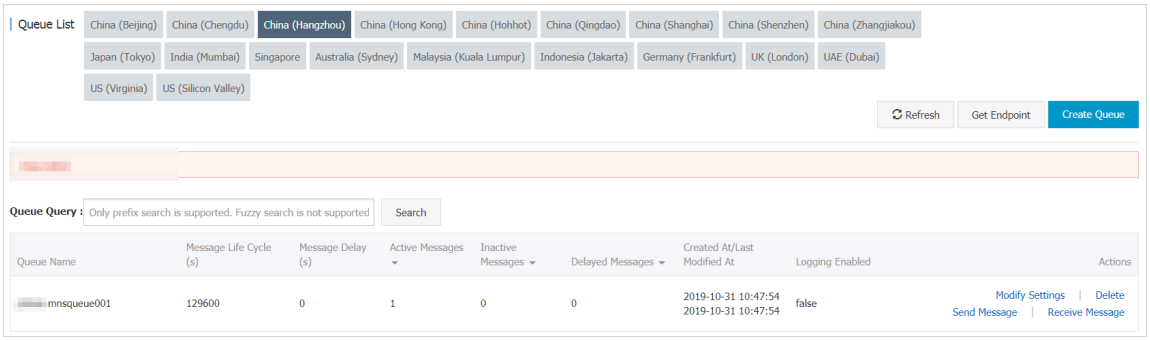


   - MNS topic: On the **Topic List** page of the MNS console, find the target topic and check whether the number in the **Message Count** column increases. If the number increases, new messages are received. You can view the message details on a subscription client.



   - MNS queue: On the **Queue List** page of the MNS console, find the target queue and check whether the number in the **Active Messages** column increases. If the number

increases, new messages are received. Click **Receive Message** in the **Actions** column to view the message details.



2.4 Modify an event notification

This topic describes how to modify an event notification. If event types of an event notification cannot meet your requirements, you can change the event types instead of creating a new event notification. Note that you cannot change the notification method of an event notification.

Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. You can use either of the following methods to open the details page of a scaling group.
   - In the **Scaling Group Name/ID** column, click a scaling group name.
   - Click **Manage** in the **Actions** column corresponding to a scaling group.
5. In the left-side navigation pane, click **Event Notifications**.
6. Find the target event notification and click **Edit** in the **Actions** column.
7. Select event types.

   For more information about event types, see Create an event notification.
8. Click **Edit Notification**.

2.5 Delete an event notification

This topic describes how to delete an event notification. You can delete an event notification if you do not use it any more.

Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. You can use either of the following methods to open the details page of a scaling group.
   - In the **Scaling Group Name/ID** column, click a scaling group name.
   - Click **Manage** in the **Actions** column corresponding to a scaling group.
5. In the left-side navigation pane, click **Event Notifications**.
6. Find the target event notification and click **Delete** in the **Actions** column.
7. Click **OK**.