ALIBABA CLOUD

Alibaba Cloud

Auto Scaling

Instance Management

Document Version: 20220310

Alibaba Cloud

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.

2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company or individual in any form or by any means without the prior written consent of Alibaba Cloud.

3. The content of this document may be changed because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.

4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides this document based on the "status quo", "being defective", and "existing functions" of its products and services. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, special, or punitive damages, including lost profits arising from the use or trust in this document (even if Alibaba Cloud has been notified of the possibility of such a loss).

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

6. Please directly contact Alibaba Cloud for any errors of this document.

Document conventions

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

Table of Contents

1.Overview

This topic describes instance management in Auto Scaling.

Instance management operations

In Auto Scaling, you can manually add or remove instances of Alibaba Cloud services, set instance status such as Protected and Standby in a scaling group, and query instances.

Impacts of instance interruption

Auto Scaling uses instances of other Alibaba Cloud services. You can manage these instances in their consoles. For example, you can delete Elastic Compute Service (ECS) instances that Auto Scaling automatically created in the ECS console. However, the deletion operation interrupts the instances. If instances are interrupted due to operations or system failures, Auto Scaling responds to the instance interruption as described in the following table.

Resource	Instance management operation	Processing method
ECS	A user deletes an ECS instance from a scaling group by using the ECS console or calling API operations.	If the Health Check feature is enabled, Auto Scaling checks whether an ECS instance is healthy based on the status of the ECS instance. Auto Scaling removes the ECS instance if the ECS instance is considered unhealthy. The internal IP address of the ECS instance that is removed from the scaling group cannot be automatically deleted from the whitelist of the associated ApsaraDB RDS instance. If the difference between the total number of ECS instances that is specified by Total Capacity and the number of ECS instances that you determine to remove is less than the value of the MinSize parameter, Auto Scaling automatically creates ECS instances to maintain the required minimum number in the scaling group. For more information about the Health Check feature, see Instance health check.
ECS	A user revokes the ECS API permissions granted to Auto Scaling.	Auto Scaling rejects all scaling activity requests.
SLB	A user manually removes an ECS instance from a Server Load Balancer (SLB) instance by using the SLB console or calling API operations.	Auto Scaling does not automatically detect this operation or handle this exception. The ECS instance remains in the scaling group. When a scale-in activity is triggered, Auto Scaling releases the ECS instance if the instance meets the removal policy.
SLB	A user manually deletes an SLB instance or disables the Health Check feature for an SLB instance by using the SLB console or calling API operations.	Auto Scaling does not add ECS instances to the scaling group that is associated with the SLB instance. Auto Scaling removes ECS instances from the scaling group if a scaling task triggers a scale-in rule or the ECS instances are considered unhealthy after a health check is performed.

Resource	Instance management operation	Processing method
SLB	An SLB instance becomes unavailable due to reasons such as overdue payments or system failures.	All scaling activities fail except for scale-in activities that are manually executed.
SLB	A user revokes the SLB API permissions granted to Auto Scaling.	Auto Scaling rejects all scaling activity requests for the scaling group that is associated with the SLB instance.
ApsaraDB RDS	A user manually removes the IP address of an ECS instance from the whitelist of the associated ApsaraDB RDS instance by using the ApsaraDB RDS console or calling API operations.	Auto Scaling does not automatically detect this operation or handle this exception. The ECS instance remains in the scaling group. When a scale-in activity is triggered, Auto Scaling releases the ECS instance if the instance meets the removal policy.
ApsaraDB RDS	A user manually deletes an ApsaraDB RDS instance by using the ApsaraDB RDS console or calling API operations.	Auto Scaling does not add ECS instances to the scaling group that is associated with the ApsaraDB RDS instance. Auto Scaling removes ECS instances from the scaling group if a scaling task triggers a scale-in rule or the ECS instances are considered unhealthy after a health check is performed.
ApsaraDB RDS	An ApsaraDB RDS instance becomes unavailable due to reasons such as overdue payments or system failures.	All scaling activities fail except for scale-in activities that are manually executed.
ApsaraDB RDS	A user revokes the ApsaraDB RDS API permissions granted to Auto Scaling.	Auto Scaling rejects all scaling activity requests for the scaling group that is associated with the ApsaraDB RDS instance.

2.Instance management in a scaling group

2.1. ECS instance lifecycle

This topic describes how to manage the lifecycle of ECS instances in a scaling group and lists all the possible service states of the instances. When Auto Scaling manages the lifecycle of ECS instances, it checks whether ECS instances are healthy, and removes or even releases unhealthy instances.

Methods for managing the lifecycle of ECS instances

ECS instances in a scaling group are categorized into automatically created and manually added instances based on how the instances are added to the scaling group. The following table describes the methods for managing the lifecycle of automatically created and manually added instances in a scaling group.

Type	Description	Management method
Automatically created ECS instances	Instances that are automatically created based on the instance configuration source of a scaling group	Auto Scaling manages the whole lifecycle of an ECS instance. During a scale-out event, Auto Scaling automatically creates ECS instances. During a scale-in event, Auto Scaling stops and releases ECS instances.
Manually added ECS instances	Instances that are manually created and then added to a scaling group	The management method depends on whether the scaling group is enabled to manage the lifecycle of the ECS instances: • If the scaling group is enabled to manage the lifecycle of the ECS instances, the ECS instances are stopped and released during a scale-in event. • If the scaling group is not enabled to manage the lifecycle of the ECS instances, the ECS instances are removed from the scaling group, but are not released during a scale-in event. **Note** Subscription instances can be manually added to a scaling group, but their lifecycle cannot be managed by the scaling group.

> **Note** The lifecycle of an ECS instance starts when the instance is created and ends when it is released. The instance lifecycle is different from the process during which an ECS instance is added to and removed from a scaling group. For more information, see Instance lifecycle.

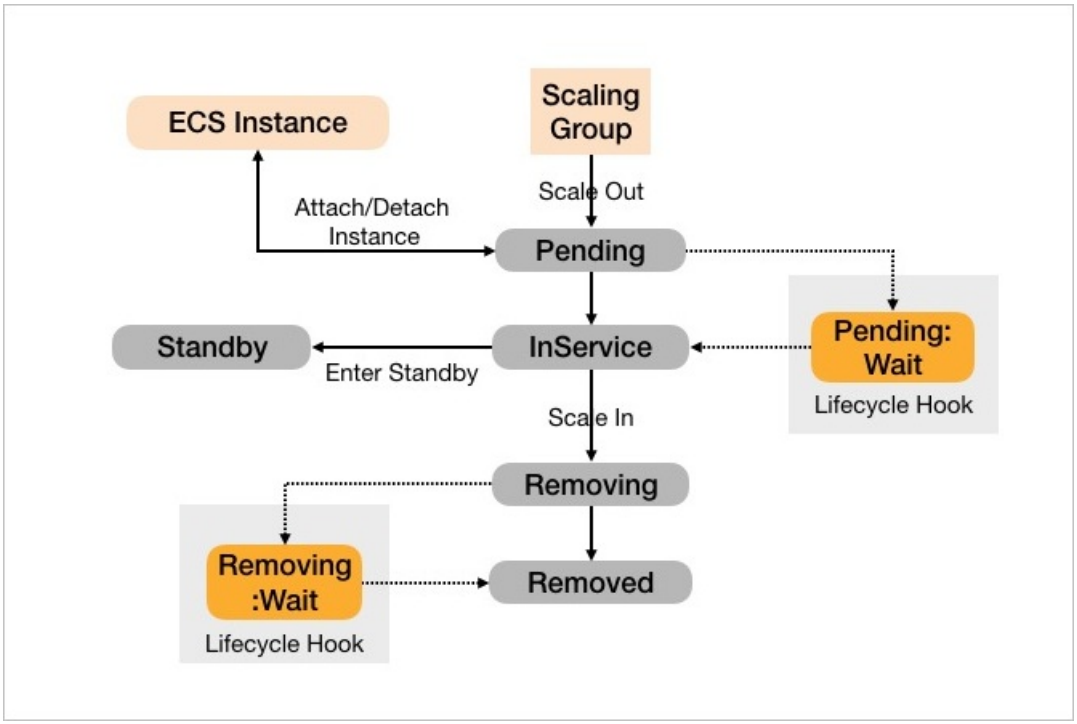
Instance service states

The following table lists all the service states through which an ECS instance in a scaling group may transition from when the instance is added to the scaling group to when the instance is removed from the scaling group.

Service state	Corresponding operation	Description
Pending (Pending)	• Execute a scaling rule. A scaling rule can be executed manually or by using event-triggered or scheduled tasks. • Manually add an ECS instance to a scaling group.	The ECS instance is being added to a scaling group. During this process, the ECS instance is added to the backend server groups of the associated Server Load Balancer (SLB) instances and the whitelists of the associated ApsaraDB RDS instances.
Adding:Wait (Pending:Wait)	创建生命周期挂钩.	If a lifecycle hook that applies to scale-out events is created for a scaling group, the ECS instance is put into the wait state when it is being added to the scaling group. The instance remains in the wait state until the lifecycle hook times out. You can perform customized operations on the instance during the timeout period, such as binding a secondary elastic network interface (ENI) and adding the instance to the whitelist of an ApsaraDB RDS instance.
In Service (InService)	None	The ECS instance is added to a scaling group and can provide services normally.
Standby (Standby)	Put an ECS instance into the Standby state.	The ECS instance stops providing services and its weight as a backend server of an SLB instance is set to zero. The SLB instance stops forwarding traffic to the ECS instance and Auto Scaling does not manage the lifecycle of the instance. You must manually manage the lifecycle of the instance. You can troubleshoot and update the image of the ECS instance that is in the Standby state, and then put the instance into service again.

Service state	Corresponding operation	Description
Protected (Protected)	Put an ECS instance into the Protected state.	The ECS instance can provide services normally. However, Auto Scaling does not manage the lifecycle of the instance. You must manually manage the lifecycle of the instance.
Stopped (Stopped)	Put an ECS instance into the Stopped state.	The ECS instance is stopped and put out of service. When the instance is stopped, its vCPUs, memory, and public IP address are reclaimed. You are no longer charged for these resources. However, you are still charged for other resources such as disks and elastic IP addresses (EIPs). During scale-out events, Auto Scaling preferentially starts ECS instances that are in the Stopped state in a scaling group. **Note** If you want to stop an ECS instance in a scaling group, make sure that the instance reclaim mode is set to Shutdown and Reclaim Mode when you create the scaling group.
Removing (Removing)	• Execute a scaling rule. A scaling rule can be executed manually or by using event-triggered or scheduled tasks. • Manually remove or delete an ECS instance.	The ECS instance is being removed from a scaling group. During this process, the instance is removed from the backend server groups of the associated SLB instances and from the whitelists of the associated ApsaraDB RDS instances.
Removing:Wait (Removing:Wait)	创建生命周期挂钩.	If a lifecycle hook that applies to scale-in events is created for a scaling group, the ECS instance is put into the wait state when it is being removed from the scaling group. The instance remains in the wait state until the lifecycle hook times out. You can perform customized operations on the instance during the timeout period, such as copying logs and clearing data.

The following figure shows the transitions between service states of an ECS instance in a scaling group.



Instance health check

You can enable or disable the health check feature when or after you create a scaling group. For more information, see Create a scaling group and Modify a scaling group.

After the health check feature is enabled, Auto Scaling manages the lifecycle of ECS instances in the scaling group and checks the status of these instances on a regular basis. If Auto Scaling detects that an ECS instance is not in the Running state, the instance is considered unhealthy.

> **Note** The running states of an ECS instance are not its service states in a scaling group. The running states refer to all possible states of an ECS instance from when the instance is created to when the instance is released.

The following section describes how unhealthy ECS instances are removed from a scaling group:

- If the instances are automatically created by Auto Scaling, or are manually added to the scaling group and their lifecycle is managed by the scaling group, Auto Scaling removes and releases these instances.
- If the instances are manually added to the scaling group and their lifecycle is not managed by the scaling group, Auto Scaling removes these instances from the scaling group but does not release them.

The removal of unhealthy ECS instances is not subject to the minimum number of instances in a scaling group. Therefore, the number of instances in the scaling group may be less than the minimum number of instances after instances are removed. Auto Scaling will automatically create a corresponding number of ECS instances to maintain the minimum number.

> **Warning** Make sure that you have sufficient balance in your account. If you have overdue payments in your account, pay-as-you-go and preemptible instances are stopped or released. For information about status changes of ECS instances in a scaling group due to overdue payments, see 欠费说明.

2.2. View ECS instances

This topic describes how to view ECS instances that have been added to a scaling group.

Procedure

1. Log on to the Auto Scaling console.
2. In the top navigation bar, select a region.
3. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
4. In the upper part of the page, click the **Instances** tab.
5. View the ECS instances.
   - The **Auto Created** tab shows ECS instances that are automatically created by Auto Scaling. You can change the status of an instance, remove it from the scaling group, or release it. If an automatically created instance is considered unhealthy, the instance is removed from the scaling group and released.
   - The **Manually Added** tab shows ECS instances that are manually added to the scaling group. You can change the status of an instance or remove it from the scaling group. If a manually added instance is not in the **Running** state, the instance is considered unhealthy and removed from the scaling group. Whether manually added ECS instances are released when they are removed from a scaling group depends on their management mode:
     - If the lifecycle of the instances is not managed by the scaling group, the instances are removed from the scaling group but not released.
     - If the lifecycle of the instances is managed by the scaling group, the instances are removed from the scaling group and released.

Related information

- Put an ECS instance into the Standby state
- Put an ECS instance into the Protected state
- Remove an ECS instance from the Protected state
- Manually add an ECS instance to a scaling group
- Manually remove or delete an ECS instance

2.3. Manually add an ECS instance to a scaling group

This topic describes how to manually add an ECS instance to a scaling group. You can add existing ECS instances to a scaling group to take full advantage of the computing resources.

Prerequisites

The following table describes the requirements that the ECS instances must meet to be manually added to a scaling group.

Item	Prerequisite

The ECS instances to be manually added to a scaling group	• The instances are located in the same region as the scaling group.
	• The instances are not added to any other scaling group.
	• The instances are in the **Running** state.
	• The network type of the instances is classic network or Virtual Private Cloud (VPC). Take note of the following items:
	  ○ When the network type of the scaling group is classic network, only ECS instances of the classic network type can be added to the scaling group.
	  ○ When the network type of the scaling group is VPC, only ECS instances in the same VPC as the scaling group can be added to the scaling group.

The scaling group	• The scaling group is in the **Enabled** state. • The scaling group does not have ongoing scaling activities.

Context

The active scaling configuration of a scaling group does not affect whether ECS instances can be manually added to the scaling group. You can manually add ECS instances without waiting for the cooldown period to expire. For more information, see Cooldown time.

Typically, Auto Scaling scales out sufficient ECS instances based on your configurations. 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 maximum number of instances specified for the scaling group, the number of actually created instances may be less than what you specified. In these cases, check the configuration of the scaling group to troubleshoot the issue. If the issue persists, submit a ticket.

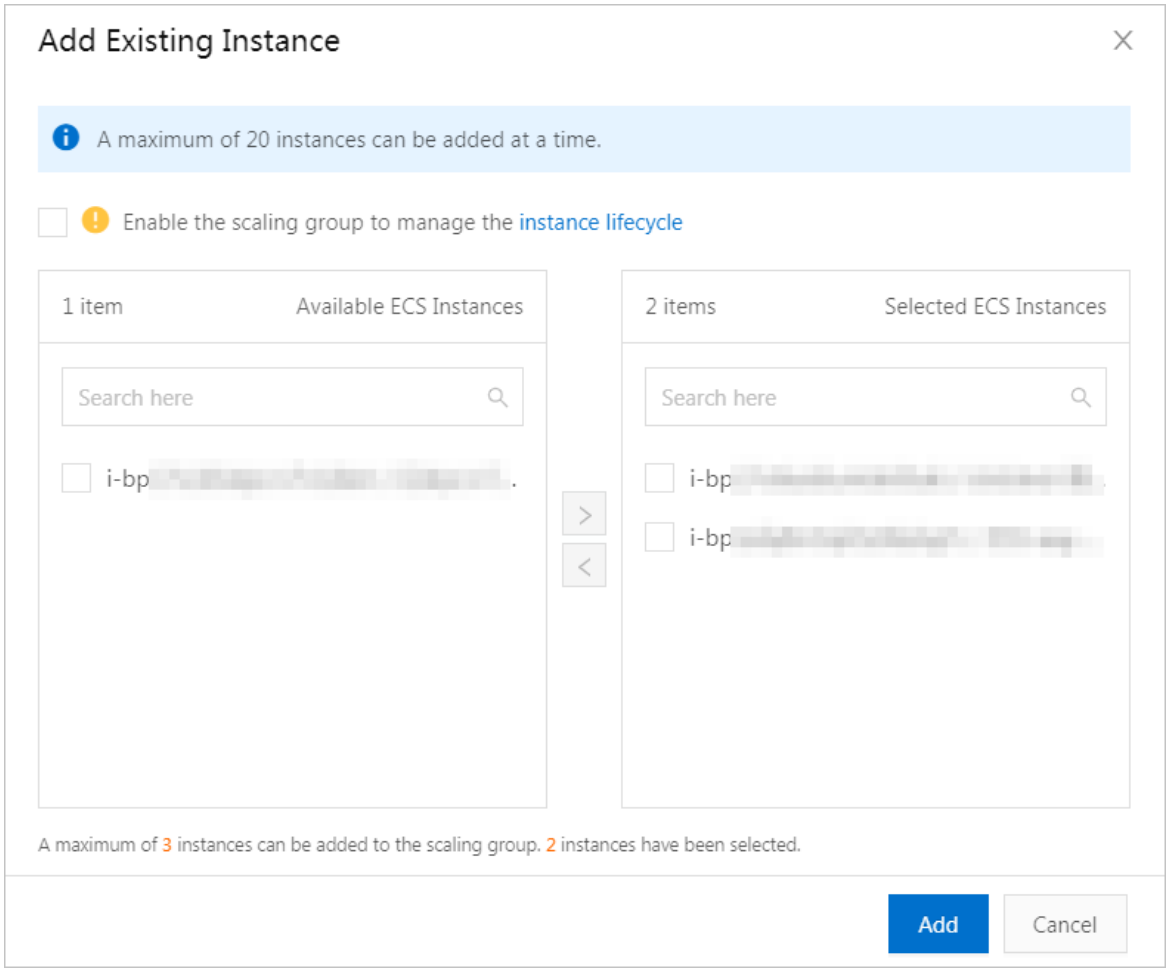
Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
5. In the upper part of the page, click the **Instances** tab.
6. Click the **Manually Added** tab.
7. Click **Add Existing Instance**.
8. In the Add Existing Instance dialog box, select available ECS instances in the left-side list, click **>** to

add the selected ECS instances to the right-side list, and then click **Add**.

If you select **Enable the scaling group to manage the instance lifecycle**, ECS instances that are manually added to the scaling group will be removed and released during scale-in events. Subscription instances can be manually added to a scaling group, but their lifecycle cannot be managed by the scaling group.

> **Warning** Make sure that you have sufficient balance in your account. If you have overdue payments in your account, pay-as-you-go and preemptible instances are stopped or released. For information about status changes of ECS instances in a scaling group due to overdue payments, see 欠费说明.



Related information

- AttachInstances

2.4. Manually remove or delete an ECS instance

This topic describes how to manually remove or delete an ECS instance that is no longer needed in a scaling group.

Prerequisites

- The scaling group is in the **Enabled** state.
- The scaling group has no scaling activities in progress.

Context

You can manually remove an ECS instance from a scaling group without the need to wait for the cooldown period to expire.

You cannot remove an ECS instance from a scaling group if the number of ECS instances in the scaling group after the removal is less than the minimum number.

A scaling activity may fail to be executed after it is triggered. You can check the execution result in the details about the scaling activity. For more information, see View the details of a scaling activity.

Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
5. In the upper part of the page, click the **Instances** tab.
6.
7. Use one of the following methods to remove one or more ECS instances from a scaling group:

   When you delete an ECS instance, the instance is removed from the scaling group and released. Whether you can delete an ECS instance that is manually added to a scaling group is subject to the management mode of the instance. If the lifecycle of the instance is not managed by the scaling group, you can only remove the instance from the scaling group but cannot delete the instance. If the lifecycle of the instance is managed by the scaling group, you can remove the instance from the scaling group and delete the instance.

   - Find the ECS instance that you want to remove from the scaling group and choose > **Remove from Scaling Group** in the **Actions** column.
   - Select multiple ECS instances that you want to remove from the scaling group and click **Remove from Scaling Group** in the lower part of the Instances page.
   - Find the ECS instance that you want to delete and choose > **Delete Instance** in the **Actions** column.
   - Select multiple ECS instances that you want to delete and click **Delete Instance** in the lower part of the Instances page.
8. Specify whether to automatically disassociate the ECS instance from the SLB instance and RDS instance and click **OK**.

2.5. Rebalance the distribution of ECS instances

If ECS instances are not evenly distributed across zones due to insufficient resources, you can use the Rebalance Distribution feature to evenly distribute the ECS instances.

Prerequisites

- The network type of the scaling group is **VPC**.
- The multi-zone scaling policy of the scaling group is **Balanced Distribution Policy**.
- The scaling group is associated with multiple vSwitches that are distributed across at least two zones.

Context

A maximum of 20 ECS instances can be replaced during a single rebalancing activity. When the rebalancing activity is executed, Auto Scaling first creates new ECS instances, and then stops and releases existing ECS instances to ensure that the ECS instances are evenly distributed across multiple zones. This does not affect the performance or availability of applications.

Auto Scaling allows the number of ECS instances to exceed 10% of the maximum number of instances for a short period of time. This occurs if the number of ECS instances in a scaling group approaches or reaches the maximum number of instances, but you must continue the balancing activity. If 10% of the maximum number of instances in the scaling group is not an integer, the decimal can be rounded up to one. The situation may last for a while until the distribution of ECS instances is balanced. It typically takes 1 to 6 minutes.

For example, the maximum number of instances in a scaling group is 15. A value of 10% indicates that the number of instances is 1.5. Then, the number of instances that a scaling group can exceed for a short period of time is 2.

Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
5. In the upper part of the page, click the **Instances** tab.
6. Click the **Auto Created** tab.
7. Click **Rebalance Distribution**.
8. Read the confirmation items and click **Confirm Execution**.

Result

The message `The rebalancing task has been assigned` appears in the upper-right corner of the page. The newly created instances are displayed in the instance list. After a period of time, the newly

created instances enter the InService state, and some existing ECS instances are released. However, the total number of ECS instances in the scaling group remains unchanged.

2.6. Put an ECS instance into the Standby state

This topic describes how to put an ECS instance that is not needed for the moment into the Standby state. After an ECS instance is put into the Standby state, the SLB weight value of the ECS instance changes to zero. If an ECS instance is in the Standby state, Auto Scaling does not check its health status or release it.

Context

After an ECS instance is put into the Standby state:

- If an SLB instance is associated with the scaling group to which the ECS instance belongs, the SLB weight value of the ECS instance changes to zero.
- The ECS instance stays in the Standby state until you manually remove it from the Standby state.
- Auto Scaling stops managing the lifecycle of the ECS instance. You must manually manage the lifecycle of the ECS instance.
- If a scale-in event is triggered, Auto Scaling will not remove the ECS instance.
- When the ECS instance is stopped or restarted, its health check status remains unchanged.
- To release the ECS instance, you must first remove it from the scaling group.
- If you delete the scaling group, the ECS instance is automatically removed from the Standby state and released.
- You can stop the ECS instance or modify its configurations. For example, you can perform the following operations:
  - Stop an instance
  - Restart an instance
  - Upgrade or downgrade instance configurations
  - Change the operating system
  - Initialize a cloud disk
  - Migrate an ECS instance from the classic network to a VPC

Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
5. In the upper part of the page, click the **Instances** tab.
6. Select the source of an ECS instance.

- To select an ECS instance that is automatically created, click the **Auto Created** tab.
- To select an ECS instance that is manually added, click the **Manually Added** tab.

7. Use one of the following methods to put one or more ECS instances into the Standby state:
   - Find the ECS instance that you want to put into the Standby state and click **Switch to Standby** in the **Actions** column.
   - Select multiple ECS instances that you want to put into the Standby state and click **Switch to Standby** in the lower part of the ECS Instances page.
8. In the message that appears, click **OK**.

Related information

- EnterStandby

2.7. Move an ECS instance out of the Standby state

This topic describes how to move an ECS instance out of the Standby state. You can move an instance out of the Standby state to reuse it.

Context

After an ECS instance in a scaling group is moved out of the Standby state:

- The ECS instance enters the In Service state.
- If a Server Load Balancer (SLB) instance is associated with the scaling group to which the ECS instance belongs, the ECS instance is added to the backend server group of the associated SLB instance again. By default, the SLB weight value of the instance is 50.
- When the ECS instance is stopped or restarted, its health status is updated.
- When a scale-in event is triggered in the scaling group, the ECS instance can be removed from the scaling group.

> **Note** If the lifecycle of the ECS instance is managed by the scaling group, the ECS instance is released. Otherwise, the ECS instance can still run normally. For more information about the lifecycle management of ECS instances, see ECS instance lifecycle.

Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
5. In the upper part of the page, click the **Instances** tab.
6. Select the source of an ECS instance.

- To select an ECS instance that is automatically created, click the **Auto Created** tab.
- To select an ECS instance that is manually added, click the **Manually Added** tab.

7. Use one of the following methods to move one or more ECS instances out of the Standby state:
   - Find the ECS instance that you want to move out of the Standby state and click **Remove from Standby** in the **Actions** column.
   - Select multiple ECS instances that you want to move out of the Standby state and click **Remove from Standby** in the lower part of the Instances page.
8. In the message that appears, click **OK**.

Related information

- ExitStandby

2.8. Put an ECS instance into the Protected state

This topic describes how to put an ECS instance into the Protected state to prevent the ECS instance from being automatically removed from a scaling group. After an ECS instance is put into the Protected state, the SLB weight value of the ECS instance remains unchanged. If an ECS instance is in the Protected state, Auto Scaling does not check its health status or release it.

Context

After an ECS instance is put into the Protected state:

- If an SLB instance is associated with the scaling group to which the ECS instance belongs, the SLB weight value of the ECS instance remains unchanged.
- The ECS instance stays in the Protected state until you manually remove it from the Protected state.
- If a scale-in event is triggered, Auto Scaling will not remove the ECS instance. To release the ECS instance, you must remove the ECS instance from the Protected state and then remove it from the scaling group.
- When the ECS instance is stopped or restarted, its health check status remains unchanged.

Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
5. In the upper part of the page, click the **Instances** tab.
6. Select the source of an ECS instance.
   - To select an ECS instance that is automatically created, click the **Auto Created** tab.
   - To select an ECS instance that is manually added, click the **Manually Added** tab.

7. Use one of the following methods to put one or more ECS instances into the Protected state:
   - Find the ECS instance that you want to put into the Protected state and click **Switch to Protected** in the **Actions** column.
   - Select multiple ECS instances that you want to put into the Protected state and click **Switch to Protected** in the lower part of the ECS Instances page.
8. In the message that appears, click **OK**.

Related information

- SetInstancesProtection

2.9. Remove an ECS instance from the Protected state

This topic describes how to remove an ECS instance from the Protected state. After an ECS instance is removed from the Protected state, Auto Scaling continues to manage the lifecycle of the ECS instance.

Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
5. In the upper part of the page, click the **Instances** tab.
6. Select the source of an ECS instance.
   - To select an ECS instance that is automatically created, click the **Auto Created** tab.
   - To select an ECS instance that is manually added, click the **Manually Added** tab.
7. Use one of the following methods to remove one or more ECS instances from the Protected state:
   - Find the ECS instance that you want to remove from the Protected state and click **Remove from Protected** in the **Actions** column.
   - Select multiple ECS instances that you want to remove from the Protected state and click **Remove from Protected** in the lower part of the ECS Instances page.
8. In the message that appears, click **OK**.

Related information

- SetInstancesProtection

2.10. Put an ECS instance into the Stopped state

If the instance reclaim mode of a scaling group is set to Shutdown and Reclaim Mode, you can put Elastic Compute Service (ECS) instances in the scaling group into the Stopped state. During scale-out, Auto Scaling preferentially starts ECS instances that are in the Stopped state.

Prerequisites

- The network type of the scaling group is virtual private cloud (VPC).
- The instance reclaim mode of the scaling group is set to Shutdown and Reclaim Mode.
- The ECS instance that you want to stop is automatically created.

Context

After an ECS instance in a scaling group is stopped, the instance stops providing services, and the billing for some resources of the instance is also stopped. Therefore, you do not need to enable Economical Mode for a pay-as-you-go instance in the ECS console. After an ECS instance is put into the Stopped state, the vCPUs, memory, and public IP address of the instance are released. You are no longer billed for these resources. However, you are still billed for other resources such as disks and elastic IP addresses (EIPs).

Procedure

1. Log on to the Auto Scaling console.
2. In the left-side navigation pane, click **Scaling Groups**.
3. In the top navigation bar, select a region.
4. Find a scaling group and go to the details page of the scaling group. You can use one of the following methods to go to the details page:
   - Click the ID of the scaling group in the **Scaling Group Name/ID** column.
   - Click **Details** in the **Actions** column of the scaling group.
5. In the upper part of the page, click the **Instances** tab.
6. Click the **Auto Created** tab.
7. Find the ECS instance that you want to stop and choose > **Switch to Stopped** in the **Actions** column.
8. In the message that appears, click **OK**.

3.SLB instance

3.1. Use SLB in Auto Scaling

You can associate Server Load Balancer (SLB) instances with a scaling group. The SLB instances distribute traffic to multiple ECS instances in the scaling group. This improves the performance of the scaling group.

Prerequisites

- You have at least one SLB instance in the **Active** state. For more information, see Create a CLB instance.
- The SLB instance and the scaling group are in the same region.
- The SLB instance and the scaling group are in the same VPC if their network type is VPC.
- If the network type of the SLB instance is classic network, the network type of the scaling group is VPC, and the backend server group of the SLB instance contains VPC-type ECS instances, the ECS instances and the scaling group must be in the same VPC.
- At least one listener is configured on the SLB instance. For more information, see Listener overview.
- Health check is enabled on the SLB instance. For more information, see Configure health checks.

Context

SLB allows multiple ECS instances in a region to share the service load by using the IP address of an SLB instance. These ECS instances act as a high-performance and high-availability application service pool. SLB distributes and controls traffic by using SLB instances, listeners, and backend servers. For more information, see What is CLB?

After an SLB instance is associated with a scaling group, all ECS instances including those automatically and manually created in the scaling group are added to the backend server group of the SLB instance. The SLB instance distributes traffic to the ECS instances based on traffic distribution policies and health check policies. This improves resource availability.

> **Note** Each ECS instance in the backend server group of an SLB instance has a default load balancing weight of 50. You can adjust the weight of an ECS instance. For more information, see Change the weight of a backend server.

Procedure

The following section describes the procedure to associate an SLB instance with a scaling group in the Auto Scaling console. For more information about other configurations of a scaling group, see Create a scaling group.

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. Go to the page for associating SLB instances with a scaling group.
   - To create a scaling group to associate with SLB instances, click **Create**.
   - To modify a scaling group that is not associated with SLB instances, find the scaling group and click **Edit** in the **Actions** column.

5. Optional. Configure the **Network Type** parameter.

   The network type of a scaling group cannot be changed after the scaling group is created.

6. Configure the **Associate SLB Instance** parameter.

   i. Select the SLB instances to be associated with the scaling group.

   You can only associate a limited number of SLB instances with a scaling group. For more information, see 使用限制. If your SLB instance does not appear in the drop-down list, check whether your SLB instance meets the prerequisites.

   ii. Select backend server groups for the SLB instance.

   You can select the default server group and vServer groups. For more information, see Backend server overview.

   - The default server group contains the ECS instances that receive requests forwarded by a listener. If no vServer group or primary and secondary server group is specified for the listener, the listener forwards all requests to the ECS instances in the default server group.
   - You can select vServer groups if you want to forward requests to different backend servers, or forward requests based on domain names or URLs.

7. Configure the remaining settings.

Related information

References

- CreateScalingGroup
- AttachLoadBalancers
- DetachLoadBalancers
- AttachVServerGroups
- DetachVServerGroups