Auto Scaling

User Guide

Issue 13

Date 2025-10-27

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Security Declaration

Vulnerability

Huawei's regulations on product vulnerability management are subject to the *Vul. Response Process.* For details about this process, visit the following web page:
https://www.huawei.com/en/psirt/vul-response-process
For vulnerability information, enterprise customers can visit the following web page:
https://securitybulletin.huawei.com/enterprise/en/security-advisory

Contents

1 AS Group

1.1 Creating an AS Group

Scenarios

An AS group consists of a collection of ECS instances and AS policies that have similar attributes and apply to the same application scenario. An AS group is the basis for enabling or disabling AS policies and performing scaling actions. The pre-configured AS policy automatically adds or deletes instances to or from an AS group, or maintains a fixed number of instances in an AS group.

When creating an AS group, specify an AS configuration for it. Additionally, add one or more AS policies for the AS group.

Creating an AS group involves the configuration of the maximum, minimum, and expected numbers of instances and the associated load balancer.

Notes

ECS types available in different AZs may vary. When creating an AS group, choose an AS configuration that uses an ECS type available in the AZs used by the AS group.

- If the ECS type specified in the AS configuration is not available in any of the AZs used by the AS group, the following situations will occur:
  - If the AS group is disabled, it cannot be enabled.
  - If the AS group is enabled, its status will become abnormal when instances are added to it.
- If the ECS type specified in the AS configuration is only available in certain AZs used by the AS group, the ECS instances added by a scaling action are only deployed in the AZs where that ECS type is available. As a result, the instances in the AS group may not be evenly distributed.

Procedure

1. Log in to the management console.

2. Click  in the upper left corner to select a region and a project.
3. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
4. Click **Create AS Group**.
5. Configure parameters, such as **Name**, **Max. Instances**, **Min. Instances**, and **Expected Instances**. **Table 1-1** describes the key AS group parameters.

Table 1-1 AS group parameters

Parameter	Description	Example Value
Region	A region is where the AS group is deployed. Resources in different regions cannot communicate with each other over internal networks. For lower network latency and faster access to your resources, select the region nearest to your target users.	-
AZ	An AZ is a physical location where resources use independent power supply and networks. AZs are physically isolated but interconnected through an internal network. • If you require high availability, buy servers in different AZs. • If you require low network latency, buy servers in the same AZ.	-
Multi-AZ Scaling Policy	This parameter can be set to **Balanced** or **Sequenced**. • **Balanced**: When scaling out an AS group, the system preferentially distributes ECS instances evenly across AZs used by the AS group. If it fails in the target AZ, it automatically selects another AZ based on the sequenced policy. • **Sequenced**: When scaling out an AS group, the system distributes ECS instances to the AZs according to the order in which AZs are specified. **NOTE** This parameter needs to be configured when two or more AZs are selected.	Balanced
Name	Specifies the name of the AS group to be created. The name contains 1 to 64 characters and consists of only letters, digits, underscores (_), and hyphens (-).	-

Parameter	Description	Example Value
Max. Instances	Specifies the maximum number of ECS instances in an AS group.	1
Expected Instances	Specifies the expected number of ECS instances in an AS group. After an AS group is created, you can change this value, which will trigger a scaling action. The number of expected instances cannot be smaller than the minimum number of instances or greater than the maximum number of instances.	0
Min. Instances	Specifies the minimum number of ECS instances in an AS group.	0
AS Configuration	Specifies the required AS configuration for the AS group. An AS configuration defines the specifications of the ECS instances to be added to an AS group. The specifications include the ECS image and system disk size. You need to create the required AS configuration before creating an AS group.	-
VPC	Provides a network for your ECS instances. All ECS instances in the AS group are deployed in this VPC.	-
Subnet	You can select up to five subnets. The AS group automatically binds all NICs to the created ECS instances. The first subnet is used by the primary NIC of an ECS instance by default, and other subnets are used by extension NICs of the ECS instance.	-

Parameter	Description	Example Value
Load Balancing	This parameter is optional. A load balancer automatically distributes traffic across all ECS instances in an AS group to balance their loads. It improves the fault tolerance of your applications and expands application capabilities. **NOTE** • Up to six load balancers can be added to an AS group. • After multiple load balancers are added to an AS group, multiple services can be concurrently listened to, thereby improving service scalability. If **ELB health check** is selected for **Health Check Method**, when any one of the listeners detects that an ECS instance becomes unhealthy, AS will replace it with a functional one. If you select **Elastic load balancer**, configure the following parameters: • **Load Balancer** • **Backend ECS Group** • **Backend Port**: specifies the port on which a backend ECS listens for traffic. • **Weight**: determines the portion of requests a backend ECS processes compared to other backend ECSs added to the same listener. • For more information about load balancing, see ***Elastic Load Balance User Guide***.	-

Parameter	Description	Example Value
Instance Removal Policy	Controls which instances are first to be removed during scale in. If specified conditions are met, scaling actions are triggered to remove instances by following the removal policy you choose. There are four instance removal policies for you to choose from: • **Oldest instance created from oldest AS configuration**: The oldest instance created from the oldest configuration is removed from the AS group first. • **Newest instance created from oldest AS configuration**: The newest instance created from the oldest configuration is removed from the AS group first. • **Oldest instance**: The oldest instance is removed from the AS group first. • **Newest instance**: The latest instance is removed from the AS group first. **NOTE** Manually added ECS instances are the last to be removed. If AS does remove a manually added instance, it only removes the ECS instance from the AS group. It does not delete the instance. If multiple manually added ECS instances must be removed, AS preferentially removes the earliest-added instance first.	Oldest instance created from oldest AS configuration
EIP	If **EIP** has been selected in the AS configuration of the AS group, an EIP is automatically bound to the ECS instance added by a scaling action to the AS group. If you select **Release**, the EIP bound to an ECS instance is released when the instance is removed from the AS group. Otherwise, the system unbinds the EIP from the instance, but does not release it when the instance is removed from the AS group.	-
Data Disk	If **Data Disk** is configured in the AS configuration used by the AS group, a data disk will be automatically created and attached to the ECS instances added during a scaling action to the AS group. If you select **Release**, the data disks attached to an ECS instance will be deleted when the instance is removed from the AS group. Otherwise, the system detaches the data disks from the ECS instance, but does not release them when the instance is removed from the AS group.	-

Parameter	Description	Example Value
Health Check Method	If an ECS instance fails a health check, AS replaces it with a new one. Health check methods: • **ECS health check**: checks the ECS instance status. If an ECS instance is stopped or deleted, it is considered unhealthy. This method is selected by default. Using this method, the AS group periodically checks the status of each ECS instance. If an ECS instance is unhealthy, AS removes the ECS instance from the AS group.	-
Health Check Interval	Specify the interval between health checks. You can set a health check interval (10 seconds, 1 minute, 5 minutes, 15 minutes, 1 hour, or 3 hours) based on service requirements.	5 minutes
Enterprise Project	Specifies the enterprise project to which the AS group belongs. If an enterprise project is configured for an AS group, ECSs created in this AS group also belong to this enterprise project. If you do not specify an enterprise project, the **default** enterprise project will be used. **NOTE** • Value **default** indicates the default enterprise project. Resources that are not allocated to any enterprise projects under your account are displayed in the default enterprise project. • Enterprise project is an upgraded version of IAM. It allocates and manages resources of different projects.	-

Parameter	Description	Example Value
Tag	If you have many resources of the same type, you can use tags to manage your resources. You can identify specified resources quickly using the tags allocated to them. Each tag contains a key and a value. You can specify the key and value for each tag. • Key – The key must be specified. – The key must be unique to the AS group. – The key can include up to 36 characters. It cannot contain non-printable ASCII characters (0–31) or the following characters: =*<>\,|/ • Value – The value is optional. – A key can have only one value. – The value can include up to 43 characters. It cannot contain non-printable ASCII characters (0–31) or the following characters: =*<>\,|/	-
Agency	Optional. When your ECS resources created in the AS group need to be shared with other accounts, or need to be delegated to professional personnel or team for management, the tenant administrator creates an agency in IAM and grants the ECS management permissions to the personnel or team. The delegated account can log in to the cloud system and switch to your account to manage resources. You do not need to share security credentials such as passwords with the account, ensuring the security of your account. If you have created an agency in IAM, select the agency from the drop-down list.	-

6. After setting the parameters, click **Create Now**.
7. (Optional) Add an AS policy.
   a. Locate the row containing the target AS group and click **View AS Policy** in the **Operation** column.
   b. (Optional) Click **Add AS Policy** to create an AS policy for the newly created AS group. For more details about AS policies, see **Overview**.

1.2 (Optional) Adding a Load Balancer to an AS Group

Elastic Load Balance (ELB) automatically distributes incoming traffic across multiple backend servers based on configured forwarding policies. ELB expands the service capabilities of applications and improves their availability by eliminating single points of failure (SPOFs).

If ELB functions are required, perform the operations provided in this section to add a load balancer to your AS group. The load balancer added to an AS group distributes application traffic to all instances in the AS group when an instance is added to or deleted from the AS group.

Only a created load balancer can be bound to an AS group, and the AS group and load balancer must be in the same VPC. For details about how to create a load balancer, see *Elastic Load Balance User Guide*. To add a load balancer for an AS group, perform the following operations:

- When creating an AS group, configure parameter **Load Balancing** to add a load balancer. For details, see **Creating an AS Group**.
- If an AS group has no scaling action ongoing, modify parameter **Load Balancing** to add a load balancer. For details, see **Modifying an AS Group**.

1.3 Changing the AS Configuration for an AS Group

Scenarios

If you need to change the specifications of ECS instances in an AS group, changing the AS configuration used by the AS group is an easy way to help you get there.

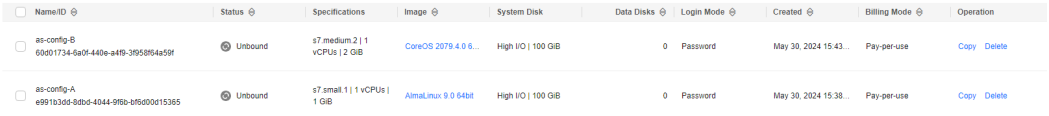
Effective Time of New AS Configuration

After you change the AS configuration for an AS group, the new AS configuration will not be used until any ongoing scaling actions are complete.

For example, if there is a scaling action ongoing for an AS group, and you change the AS configuration of the AS group from **as-config-A** to **as-config-B**, **as-config-A** is still used for the instances that are being added in the ongoing scaling action.

as-config-B will take effect in the next scaling action.

Figure 1-1 Changing the AS configuration



Procedure

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.

3. Click the name of the AS group for which you want to change the AS configuration. On the **Basic Information** page, click **Change Configuration** to the right of **Configuration Name**.

   You can also locate the row containing the target AS group and choose **More** > **Change Configuration** in the **Operation** column.
4. In the displayed **Change AS Configuration** dialog box, select another AS configuration to be used by the AS group.
5. Click **OK**.

1.4 Enabling an AS Group

Scenarios

You can enable an AS group to automatically scale capacity in or out.

After an AS group is enabled, its status changes to **Enabled**. AS monitors the AS policy and triggers a scaling action for AS groups only in **Enabled** state. After an AS group is enabled, AS triggers a scaling action to automatically add or remove ECS instances if the number of ECS instances in the AS group is different from the expected number of instances.

- Only AS groups in the **Disabled** state can be enabled.
- Only AS groups in the **Abnormal** state can be forcibly enabled. You can choose **More** > **Forcibly Enable** to enable an abnormal AS group. Forcibly enabling an AS group does not have adverse consequences.
- After you create an AS group and add an AS configuration to an AS group, the AS group is automatically enabled.

Enabling an AS Group

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
3. In the AS group list, locate the row containing the target AS group and click **Enable** in the **Operation** column. Alternatively, you can also click the AS group name and then **Enable** in the upper right corner of the page to enable the AS group.
4. In the **Enable AS Group** dialog box, click **OK**.

Forcibly Enabling an AS Group

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
3. In the AS group list, locate the row containing the target AS group and select **Forcibly Enable** from the **More** drop-down list in the **Operation** column. Alternatively, you can also click the AS group name and then **Forcibly Enable** in the upper right corner of the page to enable the AS group.
4. In the **Forcibly Enable AS Group** dialog box, click **OK**.

1.5 Disabling an AS Group

Scenarios

If you need to stop an instance in an AS group for configuration or upgrade, disable the AS group before performing the operation. This prevents the instance from being deleted in a health check. When the instance status is restored, you can enable the AS group again.

If a scaling action keeps failing and being retried (the failure cause can be viewed on the page) for an AS group, use either of the following methods to stop the action from being repeated:

- Disable the AS group. Then, after the scaling action fails, it will not be retried. Enable the AS group again when the environment recovers or after replacing the AS configuration.
- Disable the AS group and change the expected number of instances to the number of existing instances. Then after the scaling action fails, the scaling action will not be retried.

After an AS group is disabled, its status changes to **Disabled**. AS does not automatically trigger any scaling actions for a **Disabled** AS group. When an AS group has an in-progress scaling action, the scaling action does not stop immediately after the AS group is disabled.

You can disable an AS group when its status is **Enabled** or **Abnormal**.

Procedure

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
3. In the AS group list, locate the row containing the target AS group and click **Disable** in the **Operation** column. Alternatively, you can also click the AS group name and then **Disable** in the upper right corner of the page to disable the AS group.
4. In the **Disable AS Group** dialog box, click **OK**.

1.6 Modifying an AS Group

Scenarios

You can modify an AS group if needed. The settings of the following parameters can be changed: **Name**, **Max. Instances**, **Min. Instances**, **Expected Instances**, **Health Check Method**, **Health Check Interval**, **Instance Removal Policy**, **Cooldown Period**, and **Multi-AZ Scaling Policy**.

NOTE

Changing the value of **Expected Instances** will trigger a scaling action. AS will automatically increase or decrease the number of instances to the value of **Expected Instances**.

If the AS group is not enabled, contains no instance, and has no scaling action ongoing, you can modify **Subnet** configurations. If an AS group has no scaling action ongoing, you can modify its **AZ** and **Load Balancing** configurations.

Procedure

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
3. In the AS group list, locate the AS group you want to modify and choose **More** > **Modify** in the **Operation** column.

   You can also click the AS group name to switch to the **Overview** page, and click **Modify** in the upper right corner.
4. In the **Modify AS Group** dialog box, modify related data, for example, the expected number of instances.
5. Click **OK**.

1.7 Deleting an AS Group

Scenarios

You can delete an AS group when it is no longer needed.

- If an AS group is not required during a specified period, you are advised to disable it but not delete it.
- For an AS group that has instances or ongoing scaling actions, if you attempt to forcibly delete the AS group and remove and delete the instances in the AS group, the AS group enters the deleting state, rejects new scaling requests, waits until the ongoing scaling action completes, and removes all instances from the AS group. Then, the AS group is automatically deleted. Instances automatically created are removed and deleted, but instances manually added are only removed out of the AS group. During this process, other operations cannot be performed in the AS group.

  CAUTION

  Forcibly deleting an AS group may not delete ECS instances in the group.

- When an AS group is deleted, its AS policies and the alarm rules generated based on those AS policies will be automatically deleted.

Procedure

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
3. In the AS group list, locate the row containing the target AS group and choose **More** > **Delete** in the **Operation** column.

4. In the displayed **Delete AS Group** dialog box, click **OK**.

2 AS Configuration

2.1 Creating an AS Configuration

An AS configuration defines the specifications of the ECS instances to be added to an AS group. The specifications include the ECS image and system disk size.

Scenarios

- When you create an AS group, create an AS configuration or use an existing AS configuration.
- Create the required AS configuration on the **Instance Scaling** page.
- Change the AS configuration on the AS group details page.

Methods

- Create an AS configuration from an existing ECS instance.

  If you create an AS configuration from an existing ECS instance, the vCPU, memory, image, disk, and ECS type are the same as those of the selected ECS instance by default. For details, see **Creating an AS Configuration from an Existing ECS**.
- Create an AS configuration from a new specifications template.

  If you have special requirements on the ECS instances used for scaling out, use a new specifications template to create the AS configuration. For details, see **Creating an AS Configuration from Scratch**.

2.2 Creating an AS Configuration from an Existing ECS

Scenarios

You can use an existing ECS instance to rapidly create an AS configuration. In such a case, the parameter settings, such as the ECS type, vCPUs, memory, image, and disk settings (including the size, encryption, key and type) in the AS configuration are the same as those of the selected ECS instance by default.

Procedure

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
3. Click **Create AS Configuration**.
4. Set the parameters for the AS configuration. **Table 2-1** lists the AS configuration parameters.

Table 2-1 AS configuration parameters

Parameter	Description	Example Value
Region	Select the region where the AS configuration is created.	-
Name	Enter a name for the AS configuration.	-
Configuration Template	Choose **Use existing ECS** > **Select ECS**. In such a case, the parameter settings, such as the ECS type, vCPUs, memory, image, and disk settings (including the size, type, encryption, and key) in the AS configuration are the same as those of the selected ECS instance by default.	Use existing ECS
EIP	An EIP is a static public IP address bound to an ECS in a VPC. Using the EIP, the ECS provides services externally. The following options are provided: • Do not use An ECS without an EIP cannot access the Internet. However, it can still be used as a service ECS or deployed in a cluster on a private network. • Automatically assign An EIP with a dedicated bandwidth is automatically assigned to each ECS. The bandwidth size is configurable. **NOTE** If you select **Automatically assign**, specify **EIP Type**, **Billed By**, and **Bandwidth**.	Automatically assign

Parameter	Description	Example Value
Login Mode	An ECS can be authorized using a key pair or a password. • Key pair A key pair is used for ECS login authentication. If you select this mode, create or import a key pair first. **NOTE** If you use an existing key, make sure that you have saved the key file locally. Without the key, you will not be able to log in to your ECS. • Password The initial password of user **root** (for Linux) or user **Administrator** (for Windows) is used for authentication. You can log in to an ECS instance using a username and password combination.	Admin@123
Advanced Settings	This allows you to configure **User Data** and **ECS Group**. You can select **Do not configure** or **Configure now**.	-
User Data	Injects user data automatically into an ECS when the ECS starts for the first time. This configuration is optional. If this function is enabled, user data will be automatically injected into an ECS upon its first startup. For details, see **Elastic Cloud Server User Guide**. The following two methods are available: • **As text**: allows you to enter the user data in the text box below. • **As file**: allows you to inject a script file or other files when you create an ECS. **NOTE** • For Linux, if you use password authentication, this function is not supported. • If the selected image does not support user data injection, this function is not supported.	-
ECS Group	An ECS group allows you to create ECSs on different hosts to improve service reliability. For details, see **Managing ECS Groups**.	-

5. Click **Create Now**.

6. If you want to use the newly created AS configuration, add it to the AS group. For details, see **Changing the AS Configuration for an AS Group**.

2.3 Creating an AS Configuration from Scratch

Scenarios

If you have special requirements on the ECS instances for scaling out, use a new specifications template to create the AS configuration. In such a case, ECS instances that have the specifications specified in the template will be added to the AS group in scaling actions.

Procedure

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
3. Click **Create AS Configuration**.
4. Set parameters for the AS configuration. **Table 2-2** lists the AS configuration parameters.

Table 2-2 AS configuration parameters

Parameter	Description	Example Value
Region	Select the region where the AS configuration is created.	-
Name	Enter a name for the AS configuration.	-
Configuration Template	Select **Create new template**. If this option is selected, configure parameters, such as the vCPUs, memory, image, disk, and ECS type, to create a new AS configuration.	Create new template
CPU Architecture	Both x86 and Kunpeng CPU architectures are available: • **x86**: The x86-based CPU architecture uses Complex Instruction Set Computing (CISC). • **Kunpeng**: The Kunpeng-based CPU architecture uses Reduced Instruction Set Computing (RISC). **NOTE** This parameter is displayed only when both x86-based and Kunpeng-based ECSs are available in the current region.	x86

Parameter	Description	Example Value
Specifications	The public cloud provides various ECS types for different application scenarios. For more information, see **Elastic Cloud Server User Guide**. Configure the ECS specifications, including vCPUs, memory, image type, and disk, based on the ECS type.	Memory-optimized
Image	• Public image A public image is a standard, widely used image. It contains an OS and preinstalled public applications and is available to all users. If a public image does not contain the applications or software you need, you can use the public image to create a cloud server and then deploy required software as needed. • Private image A private image is an image available only to the user who created it. It contains an OS, preinstalled public applications, and the user's private applications. Using a private image to create ECSs frees you from configuring multiple ECSs repeatedly. • Shared image A shared image is a private image shared by another public cloud user.	Public image

Parameter	Description	Example Value
Disk	Includes system and data disks. • **System Disk** **Common I/O**: uses Serial Advanced Technology Attachment (SATA) drives to store data. **High I/O**: uses serial attached SCSI (SAS) drives to store data. **General Purpose SSD**: uses solid state disk (SSD) drives to store data. **Extreme SSD**: uses enhanced solid state disk (ESSD) drives to store data. **Ultra-high I/O**: uses solid state disk (SSD) drives to store data. If a full-ECS image is used, the system disk is restored using the disk backup. On the console, you can only change the volume type and size. In addition, the volume cannot be smaller than the disk backup. **NOTE** Different ECS flavors support different disk types. The supported disk types will be displayed on the management console. • **Data Disk** You can create multiple data disks for an ECS instance and encrypt them. In addition, you can use a data disk image to export data. (Optional) Set encryption parameters. These parameters are available only when **Encryption** is selected. The encryption parameters are as follows: **Agency Name**: specifies the name of the agency that is used to grant EVS the permissions needed to obtain KMS keys for disk encryption and decryption. When **Agency Name** is displayed as **EVSAccessKMS**, KMS permissions have been granted to EVS. **KMS Encryption**: specifies how to obtain a KMS key. – **Select an existing key**: Select a KMS key from the **KMS Key Name** drop-down list.	**Common I/O** for **System Disk**

Parameter	Description	Example Value
	– **Enter a key ID**: Select a KMS key using the key ID. (Optional) **KMS Key Name**: specifies the name of the key used to encrypt EVS disks. This parameter is displayed only when **KMS Encryption** is set to **Select an existing key**. You can select an existing key, or click **Create KMS Key** and create a KMS key on the KMS console. The default value is **evs/default**. **KMS Key ID**: specifies the ID of the key used to encrypt data disks. If the image you selected is a full-ECS image, you can change the volume type and size and encryption attributes of the data disk restored using the disk backup. Ensure that the disk is at least as big as the disk backup. The encryption attributes can only be modified if the disk backup is in the same region as the disk.	
Security Group	Controls ECS access within or between security groups by defining access rules. ECSs added to a security group are protected by the access rules you define.	-
EIP	An EIP is a static public IP address bound to an ECS in a VPC. Using the EIP, the ECS provides services externally. The following options are provided: • **Do not use**: An ECS without an EIP cannot access the Internet. However, it can still be used as a service ECS or deployed in a cluster on a private network. • **Automatically assign**: An EIP with a dedicated bandwidth is automatically assigned to each ECS. You can set the bandwidth size. **NOTE** If you select **Automatically assign**, you need to specify **EIP Type**, **Billed By**, and **Bandwidth**.	Automatically assign

Parameter	Description	Example Value
Bandwidth	Specifies the bandwidth size in Mbit/s. **NOTE** • This parameter is available only when **EIP** is set to **Automatically assign**. • If you select **Dedicated**, you can select **Bandwidth** or **Traffic** for **Billed By**. • Shared bandwidths can be billed only by bandwidth. You can select a shared bandwidth to which the EIP is to be added.	100
Login Mode	An ECS can be authorized using a key pair or a password. • Key pair A key pair is used for ECS login authentication. If you select this mode, create or import a key pair first. **NOTE** If you use an existing key, make sure that you have saved the key file locally. Without the key, you will not be able to log in to your ECS. • Password The initial password of user **root** (for Linux) or user **Administrator** (for Windows) is used for authentication. You can log in to an ECS instance using a username and password combination.	Admin@123
Advanced Settings	This allows you to configure **ECS Group** and **User Data**. You can select **Do not configure** or **Configure now**.	-

Parameter	Description	Example Value
User Data	Injects user data automatically into an ECS when the ECS starts for the first time. This configuration is optional. If this function is enabled, user data will be automatically injected into an ECS upon its first startup. For details, see ***Elastic Cloud Server User Guide***. The following methods are available: • **As text**: allows you to enter the user data in the text box below. • **As file**: allows you to inject script files or other files when you create an ECS. **NOTE** • For Linux, if you use password authentication, this function is not supported. • If the selected image does not support user data injection, this function is not supported.	-
ECS Group	An ECS group allows you to create ECSs on different hosts to improve service reliability.	-

5. Click **Create Now**. The system displays a message indicating that the AS configuration is successfully created.
6. If you want to use the newly created AS configuration, add it to the AS group. For details, see **Creating an AS Group** or **Changing the AS Configuration for an AS Group**.

Notes on Multiple Flavors in an AS Configuration

AS configuration supports multiple flavors to minimize the probability of capacity expansion failures due to insufficient or unavailable flavors and ensure that capacity expansion succeeds during peak hours.

A maximum of 10 flavors can be selected for an AS configuration.

Applicable Scenario

- No special requirement for the instance flavors created in the AS group
- Requiring higher success ratio and low latency of creating instances in an AS group
- Requiring instances with high specifications
- Services that are stateless and can be horizontally scaled

The AS group sorts multiple flavors in either of the following ways:

- Sequenced: During AS group expansion, flavors are used based on the sequence they are selected. When the first flavor is insufficient or the instance fails to be created due to other reasons, the system attempts to create an instance of the second flavor, and so on.
- Cost-centered: During AS group expansion, the flavor with the minimum cost comes first. When creating an instance in an AS group, the system selects the flavor with the minimum cost. If the instance cannot be created, the system selects one with the minimum cost from the remained flavors, and so on.

2.4 Copying an AS Configuration

Scenarios

You can copy an existing AS configuration.

When copying an AS configuration, you can modify parameter settings, such as the configuration name, ECS specifications, and image of the existing AS configuration to rapidly add a new AS configuration.

Procedure

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
3. Click the **AS Configurations** tab, locate the row containing the target AS configuration, and click **Copy** in the **Operation** column.
4. On the **Copy AS Configuration** page, modify parameter settings, such as **Name**, **Specifications**, and **Image**, and configure the ECS login mode based on service requirements.
5. Click **OK**.

2.5 Deleting an AS Configuration

Scenarios

When you no longer need an AS configuration, you can delete it as long as the AS configuration is not used by an AS group. You can delete a single AS configuration or delete them in batches.

Procedure

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
3. Click the **AS Configurations** tab page, locate the row containing the target AS configuration, and click **Delete** in the **Operation** column to delete this AS configuration. You can also select multiple AS configurations to be deleted

and click **Delete** in the upper part of the AS configuration list to delete them all at once.

3 AS Policy

3.1 Overview

AS policies can trigger scaling actions to adjust bandwidth or the number of instances in an AS group. An AS policy defines the conditions for triggering a scaling action and the operation to be performed in a scaling action. When the trigger condition is met, a scaling action is triggered automatically.

NOTE

If multiple AS policies are applied to an AS group, a scaling action is triggered as long as any of the AS policies is invoked, provided that the AS policies do not conflict with each other.

The number of instances in the AS group will never exceed the specified maximum and minimum numbers of instances.

Constraints

A maximum of 10 AS policies can be created for an AS group.

AS supports the following policies:

- Alarm policy: AS automatically adjusts the number of ECS instances in an AS group or sets the number of instances to the configured value when an alarm is generated for a configured metric, such as the CPU usage.
- Scheduled policy: AS automatically increases or decreases the number of ECS instances in an AS group or sets the number of instances to the configured value at a specified time.
- Periodic policy: AS automatically increases or decreases the number of ECS instances in an AS group or sets the number of instances to the configured value at a configured interval, such as daily, weekly, and monthly.

Resource Adjustment Modes

- Dynamic

  AS adjusts the number of instances or bandwidth when an alarm policy is triggered.

This mode is suitable for scenarios where workloads are unpredictable. Alarm policies are used to trigger scaling actions based on real-time monitoring data (such as CPU usage) to dynamically adjust the number of instances in the AS group.

- Planned

  AS adjusts the number of instances or bandwidth when a periodic or scheduled policy is triggered.

  This mode is suitable for scenarios where workloads are periodic.

- Manual

  You can change the size of an AS group manually. You can either add or remove instances to or from the AS group, or modify the expected number of instances of the AS group.

3.2 Creating an AS Policy

Scenarios

You can create different types of AS policies. In an AS policy, you can define the conditions for triggering a scaling action and what operation to be performed. When the conditions are met, AS automatically triggers a scaling action to adjust the number of instances in the AS group.

This section describes how to create alarm-based, scheduled, or periodic AS policy for an AS group.

Creating an Alarm Policy

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
3. Locate the row containing the target AS group and click **View AS Policy** in the **Operation** column.
4. On the **AS Policies** page, click **Add AS Policy**.
5. Set the parameters listed in **Table 3-1**.

Table 3-1 AS policy parameters

Parameter	Description	Example Value
Policy Name	Specifies the name of the AS policy to be created.	as-policy-p6g5
Policy Type	Select **Alarm**.	Alarm

Parameter	Description	Example Value
Alarm Rule	Specifies whether a new alarm rule is to be created (**Create**) or an existing alarm rule will be used (**Use existing**). For details about how to use an existing alarm rule, see **Setting Monitoring Alarm Rules**. If you choose to create an alarm, system monitoring and custom monitoring are supported. • System monitoring requires the parameters in **Table 3-2**. • Custom monitoring requires the parameters in **Table 3-3**.	-

Parameter	Description	Example Value
Scaling Action	Specifies the scaling action to be executed and the number or percentage of instances. The following scaling action options are available: • Add Adds instances to an AS group when the scaling action is performed. • Reduce Removes instances from an AS group when the scaling action is performed. • Set to Maintains a fixed number of instances in an AS group.	• Add 1 instance • Add 10% instances The number of instances to be added is 10% of the current number of instances in the AS group. If the product of the current number of instances and the percentage is not an integer, AS automatically rounds the value up or down: – If the value is greater than 1, it is rounded down. For example, value 12.7 is rounded down to 12. – If the value is greater than 0 but less than 1, it is rounded up to 1. For example, value 0.67 is rounded up to 1. For example, suppose that there are 10 instances in an AS group, and the scaling action is **Add 15% instances**. When the AS policy is triggered, AS calculates the number of instances to be added is 1.5 and rounds 1.5 down to 1. After the scaling action is complete, there are 11 instances in the AS group.

Parameter	Description	Example Value
Cooldown Period	To prevent an alarm-based policy from being repeatedly triggered by the same event, you can set a cooldown period. A cooldown period (in seconds) is the period of time between two scaling actions. AS recounts the cooldown period after a scaling action is complete. During the cooldown period, AS denies all scaling requests triggered by alarm-based policies. Scaling requests triggered manually or by scheduled or periodic policies are not affected. For example, suppose that the cooldown period is set to 300 seconds (5 minutes), and a scheduled policy is specified to trigger a scaling action at 10:32, and a previous scaling action triggered by an alarm policy ends at 10:30. Any alarm-triggered scaling action will then be denied during the cooldown period from 10:30 to 10:35, but the scaling action scheduled for 10:32 will still take place. If the scheduled scaling action ends at 10:36, a new cooldown period starts at 10:36 and ends at 10:41.	300s

Parameter	Description	Example Value
	NOTE • If a scaling action is triggered by an AS policy, the cooldown period is whatever configured for that AS policy. • If a scaling action is triggered by manually changing the expected number of instances or by other actions, the cooldown period is whatever configured for the AS group. The default cooldown period is 300 seconds. • When an AS group scales out, scale-in requests triggered manually or by scheduled or periodic policies will be executed immediately after the scale-out is complete, without being affected by the cooldown period. • When an AS group scales in, scale-out requests triggered manually or by scheduled or periodic policies will be executed immediately after the scale-in is complete, without being affected by the cooldown period.	

Table 3-2 System monitoring parameters

Parameter	Description	Example Value
Rule Name	Specifies the name of the alarm rule.	as-alarm-7o1u
Monitoring Type	Specifies the type of monitoring metrics, which can be **System monitoring** or **Custom monitoring**. Select **System monitoring**.	System monitoring
Trigger Condition	Select monitoring metrics supported by AS and set alarm conditions for the metrics.	CPU Usage Max. >70%
Monitoring Interval	Specifies the interval at which the alarm status is updated based on the alarm rule.	5 minutes

Parameter	Description	Example Value
Consecutive Occurrences	Specifies the number of sampling points when an alarm is triggered. If **Consecutive Occurrences** is set to **n**, the sampling points of the alarm rule are the sampling points in n consecutive sampling periods. The alarm rule status does not change to **Alarm** unless all sampling points breach the threshold configured by the alarm rule.	3

Table 3-3 Custom monitoring parameters

Parameter	Description	Example Value
Rule Name	Specifies the name of the alarm rule.	as-alarm-7o1u
Monitoring Type	Select **Custom monitoring**. Custom monitoring meets monitoring requirements in various scenarios.	Custom monitoring
Resource Type	Specifies the name of the service for which the alarm rule is configured.	AGT.ECS
Dimension	Specifies the metric dimension of the alarm rule.	instance_id
Instance	Specifies the resources to which the alarm rule applies.	-
Trigger Condition	Select monitoring metrics supported by AS and set alarm conditions for the metrics.	CPU Usage Max. >70%
Monitoring Interval	Specifies the interval at which the alarm status is updated based on the alarm rule.	5 minutes
Consecutive Occurrences	Specifies the number of sampling points when an alarm is triggered. If **Consecutive Occurrences** is set to **n**, the sampling points of the alarm rule are the sampling points in n consecutive sampling periods. The alarm rule status does not change to **Alarm** unless all sampling points breach the threshold configured by the alarm rule.	3

6. Click **OK**.

The newly added AS policy is displayed on the **AS Policies** tab. In addition, the AS policy is in **Enabled** state by default.

Creating a Scheduled or Periodic Policy

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
3. Locate the row containing the target AS group and click **View AS Policy** in the **Operation** column.
4. On the **AS Policies** page, click **Add AS Policy**.
5. Configure the parameters listed in **Table 3-4**.

Table 3-4 Parameter description

Parameter	Description	Example Value
Policy Name	Specifies the name of the AS policy to be created.	as-policy-p6g5
Policy Type	Select **Scheduled** or **Periodic** for expanding resources at a specified time. If you select **Periodic**, you are required to configure two more parameters: • Period – Day – Week – Month • Time Range Specifies the time range during which the AS policy can be triggered.	-
Time Zone	The default value is **GMT +08:00**. GMT+08:00 is 8:00 hours ahead of Greenwich Mean Time.	GMT+08:00
Triggered At	Specifies the time at which the AS policy is triggered.	-

Parameter	Description	Example Value
Scaling Action	Specifies an action and the number of instances. The following scaling action options are available: • Add Adds instances to an AS group when the scaling action is performed. • Reduce Removes instances from an AS group when the scaling action is performed. • Set to Sets the expected number of instances in an AS group to a specified value.	• Add 1 instance • Add 10% instances The number of instances to be added is 10% of the current number of instances in the AS group. If the product of the current number of instances and the percentage is not an integer, AS automatically rounds the value up or down: – If the value is greater than 1, it is rounded down. For example, value 12.7 is rounded down to 12. – If the value is greater than 0 but less than 1, it is rounded up to 1. For example, value 0.67 is rounded up to 1. For example, suppose that there are 10 instances in an AS group, and the scaling action is **Add 15% instances**. When the AS policy is triggered, AS calculates the number of instances to be added is 1.5 and rounds 1.5 down to 1. After the scaling action is complete, there are 11 instances in the AS group.

Parameter	Description	Example Value
Cooldown Period	To prevent an alarm-based policy from being repeatedly triggered by the same event, you can set a cooldown period. A cooldown period (in seconds) is the period of time between two scaling actions. AS recounts the cooldown period after a scaling action is complete. During the cooldown period, AS denies all scaling requests triggered by alarm-based policies. Scaling requests triggered manually or by scheduled or periodic policies are not affected. For example, suppose that the cooldown period is set to 300 seconds (5 minutes), and a scheduled policy is specified to trigger a scaling action at 10:32, and a previous scaling action triggered by an alarm policy ends at 10:30. Any alarm-triggered scaling action will then be denied during the cooldown period from 10:30 to 10:35, but the scaling action scheduled for 10:32 will still take place. If the scheduled scaling action ends at 10:36, a new cooldown period starts at 10:36 and ends at 10:41.	300s

Parameter	Description	Example Value
	NOTE • If a scaling action is triggered by an AS policy, the cooldown period is whatever configured for that AS policy. • If a scaling action is triggered by manually changing the expected number of instances or by other actions, the cooldown period is whatever configured for the AS group. The default cooldown period is 300 seconds. • When an AS group scales out, scale-in requests triggered manually or by scheduled or periodic policies will be executed immediately after the scale-out is complete, without being affected by the cooldown period. • When an AS group scales in, scale-out requests triggered manually or by scheduled or periodic policies will be executed immediately after the scale-in is complete, without being affected by the cooldown period.	

6. Click **OK**.

   The newly added AS policy is displayed on the **AS Policies** tab. In addition, the AS policy is in **Enabled** state by default.

NOTE

If you have created scheduled or periodic AS policies that are invoked at the same time, AS will execute the one created later. This constraint does not apply to alarm-triggered AS policies.

3.3 Managing AS Policies

Scenarios

An AS policy specifies the conditions for triggering a scaling action as well as the operation that will be performed. If the conditions are met, a scaling action is triggered automatically.

This section describes how to manage an AS policy, including modifying, enabling, disabling, executing, and deleting an AS policy.

Modifying an AS Policy

If a particular AS policy cannot meet service requirements, you can modify the parameter settings of the policy.

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
3. Locate the row containing the target AS group and click **View AS Policy** in the **Operation** column. On the displayed page, locate the row containing the target AS policy and choose **More** > **Modify** in the **Operation** column.
4. In the displayed **Modify AS Policy** dialog box, modify the parameters and click **OK**.

Enabling an AS Policy

An AS policy can trigger scaling actions only when it and the AS group are both enabled. You can enable one or more AS policies for an AS group as required.

- Before enabling multiple AS policies, ensure that the AS policies do not conflict with one another.
- An AS policy can be enabled only when its status is **Disabled**.

Locate the row containing the target AS group and click **View AS Policy** in the **Operation** column. On the displayed page, locate the row containing the target AS policy and click **Enable** in the **Operation** column. To concurrently enable multiple AS policies, select these AS policies and click **Enable** in the upper part of the AS policy list.

Disabling an AS Policy

If you do not want a particular AS policy to trigger any scaling actions within a specified period of time, you can disable it.

- If all of the AS policies configured for an AS group are disabled, no scaling action will be triggered for this AS group. However, if you manually change the value of **Expected Instances**, a scaling action will still be triggered.
- You can disable an AS policy only when its status is **Enabled**.

Locate the row containing the target AS group and click **View AS Policy** in the **Operation** column. On the displayed page, locate the row containing the target AS policy and click **Disable** in the **Operation** column. To concurrently disable multiple AS policies, select these AS policies and click **Disable** in the upper part of the AS policy list.

Manually Executing an AS Policy

You can make the number of instances in an AS group reach the expected number of instances immediately by manually executing an AS policy.

- You can manually execute an AS policy if the scaling conditions configured in the AS policy are not met.
- You can manually execute an AS policy only when the AS group and AS policy are both in **Enabled** state.

Locate the row containing the target AS group and click **View AS Policy** in the **Operation** column. On the displayed page, locate the row containing the target AS policy and click **Execute Now** in the **Operation** column.

NOTE

If **Policy Type** is set to **Alarm** and **Alarm Policy Type** to **Refined scaling**, the scaling policy cannot be executed immediately.

Deleting an AS Policy

You can delete an AS policy that will not be used for triggering scaling actions.

An AS policy can be deleted even when the scaling action triggered by the policy is in progress. Deleting the AS policy does not affect a scaling action that has already started.

Locate the row containing the target AS group and click **View AS Policy** in the **Operation** column. On the displayed page, locate the row containing the target AS policy and choose **More** > **Delete** in the **Operation** column.

To concurrently delete multiple AS policies, select these AS policies and click **Delete** in the upper part of the AS policy list.

4 Scaling Action

4.1 Dynamic Scaling

Before using AS to perform scaling actions, you must specify how to perform the scaling actions to dynamically expand resources.

If the demands change frequently, you can configure alarm-based policies to scale resources. When the conditions for invoking an AS policy are met, AS automatically changes the expected number of instances to trigger a scaling action to scale up or down resources. For details about how to create an alarm policy, see **Creating an AS Policy**.

Consider a train ticket booking application. If the CPU usage of the instances that run the application goes up to 90%, an instance needs to be added to ensure that services run properly. If the CPU usage drops down to 30%, an instance needs to be deleted to prevent resource waste. To meet the requirements, you can configure two alarm policies. One policy is used to add one instance if the maximum CPU usage exceeds 90%. For details, see **Figure 4-1**. The other policy is used to remove an instance if the minimum CPU usage drops below 30%. For details, see **Figure 4-2**.

Figure 4-1 Alarm policy 01

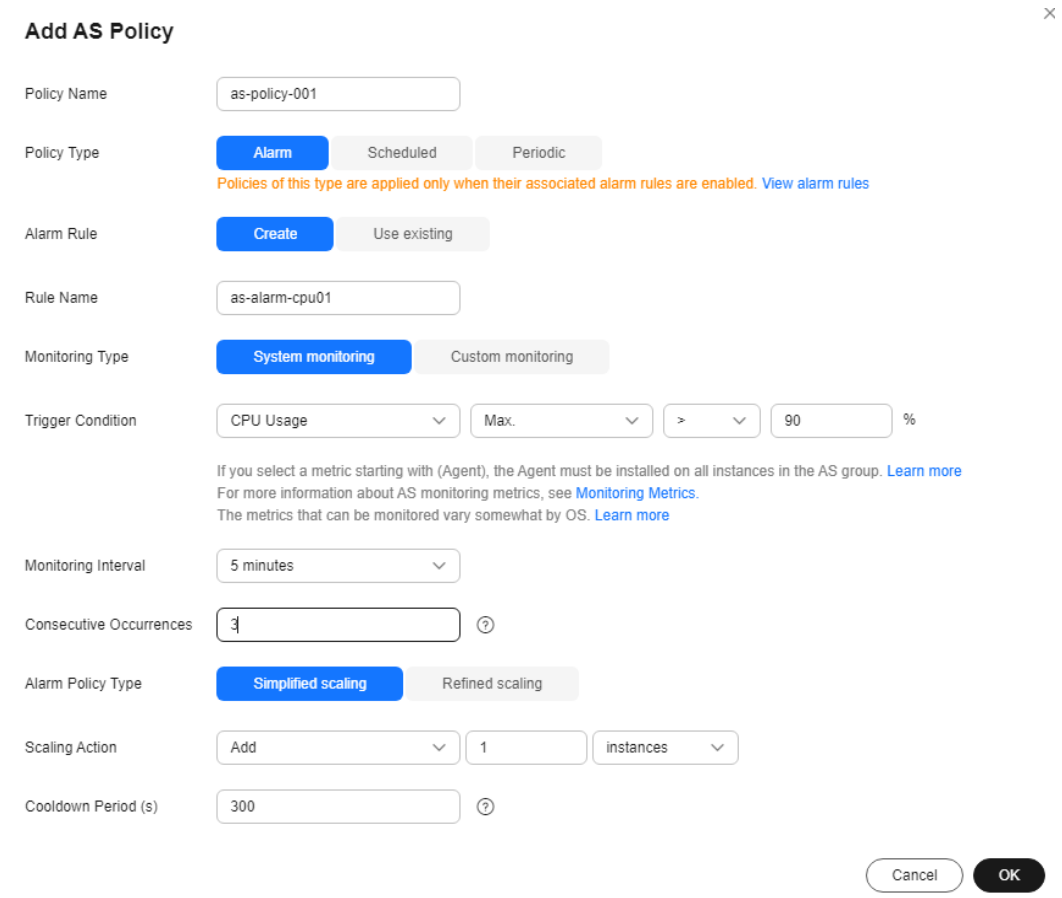
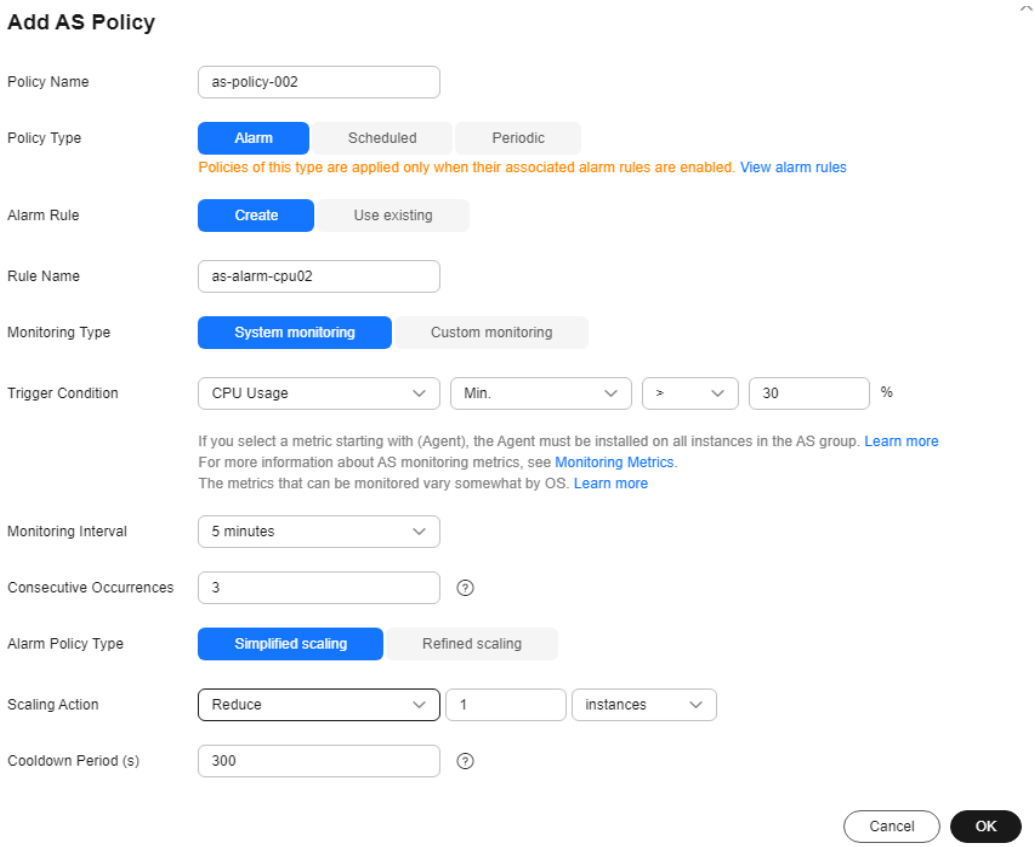


Figure 4-2 Alarm policy 02



4.2 Scheduled Scaling

To satisfy demands that change regularly, you can configure a scheduled or periodic scaling policy to adjust resources at specified times or intervals.

For details about how to create a scheduled or periodic policy, see **Creating an AS Policy**.

Take an online course selection web application as an example. This application is frequently used when a semester starts and seldom used during other parts of the year. You can configure two scheduled policies to scale resources at the beginning of each semester. The first policy is used to add an instance when the course selection starts, and the second policy is used to remove an instance when the course selection ends.

4.3 Manual Scaling

Scenarios

You can change the size of an AS group manually. You can either add or remove instances to or from the AS group, or modify the expected number of instances of the AS group.

Procedure

Adding an instance to an AS group

Before you add an ECS instance to an AS group, ensure that the conditions below are met.

Table 4-1 Conditions for manually adding an instance to an AS group

Item	Condition
AS group	• The AS group is in the **Enabled** status. • The AS group does not have ongoing scaling actions. • The number of instances to be added plus the expected number of instances cannot exceed the maximum number of instances of the AS group.
ECS instance	• The instance to be added is not a member of another AS group. • The ECS instance is in the same VPC as the AS group.

NOTE

- A maximum of 10 instances can be added to an AS group at a time.
- If the AS group has an attached load balancer, the instances will be associated with the load balancer.

To add ECS instances to an AS group, perform the following steps:

1. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
2. Click the **AS Groups** tab and then the name of the target AS group.
3. On the AS group details page, click the **Instances** tab and then **Add**.
4. Select the instances to be added and click **OK**.

Removing an instance from an AS group

You can remove an instance from an AS group, update the instance or fix an instance fault, and add the instance back to the AS group. After the instance is removed from the AS group, it no longer processes any application traffic.

For example, you can change AS configuration for an AS group at any time. New instances will be created using the new configuration, but existing instances in the AS group are not affected. To update the existing instances, you can stop them so that they can be replaced automatically. You can also remove the instances from the AS group, update them, and then add them back to the AS group.

When you remove ECS instances from an AS group, consider the restrictions below.

Table 4-2 Constraints on manually removing an instance from an AS group

Item	Constraint
AS group	• The AS group is in the **Enabled** status. • The AS group does not have ongoing scaling actions.
ECS instance	• The instances are in the **Enabled** lifecycle status. • The ECS instances are not used by SDRS.

NOTE

- A maximum of 50 instances can be removed from to an AS group at a time.
- If the number of instances you are removing decreases the number of instances in the AS group below the minimum number of instances allowed, AS launches new instances to maintain the expected capacity.
- If you remove instances from an AS group that has an associated load balancer, the instances will be dissociated from the load balancer.

To remove ECS instances from an AS group, perform the following steps:

1. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
2. Click the **AS Groups** tab and then the name of the target AS group.
3. Click the **Instances** tab, locate the row containing the desired instance, and click **Remove** or **Remove and Delete** in the **Operation** column.

   To remove multiple instances from the AS group, select the check boxes in front of them and click **Remove** or **Remove and Delete**.

   To remove all instances from the AS group, select the check box on the left of **Name** and click **Remove** or **Remove and Delete**.

   NOTE

   - If the ECS instances you want to remove were automatically added to the AS group, they are billed on a pay-per-use basis by default. You can:
     - Remove the instances from the AS group by choosing **Remove**.
     - Remove the ECS instances from the AS group and delete them by choosing **Remove and Delete**.
   - If the instances were manually added to the AS group, they can only be removed. They cannot be removed and deleted.

Changing the expected number of instances

Manually change the expected number of instances to add or reduce the number of instances in an AS group for expanding resources.

For details, see **Modifying an AS Group**.

4.4 Configuring an Instance Removal Policy

When instances are automatically removed from your AS group, the instances that are not in the currently used AZs will be removed first. Then the instance removal policy you select will be applied.

AS supports the following instance removal policies:

- **Oldest instance**: The oldest instance is removed from the AS group first. Use this policy if you want to upgrade instances in an AS group to a new ECS type. You can gradually replace instances of the old type with instances of the new type.
- **Newest instance**: The newest instance is removed from the AS group first. Use this policy if you want to test a new AS configuration but do not want to keep it in production.
- **Oldest instance created from oldest AS configuration**: The oldest instance created from the oldest configuration is removed from the AS group first. Use this policy if you want to update an AS group and phase out the instances created from a previous AS configuration.
- **Newest instance created from oldest AS configuration**: The newest instance created from the oldest configuration is removed from the AS group first.

NOTE

Manually added ECS instances are the last to be removed. If AS does remove a manually added ECS instance, it only removes the instance. It does not delete the instance. If multiple manually added ECS instances must be removed, AS preferentially removes the earliest-added instance first.

4.5 Viewing a Scaling Action

Scenarios

This section describes how to check whether a scaling action has been performed and how to view scaling action details.

Viewing Monitoring Data

The following steps illustrate how to view scaling actions of an AS group.

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
3. Click the **AS Groups** tab and then the name of the target AS group.
4. Click the **Monitoring** tab and view scaling actions. On the **Monitoring** page, you can view changes in the number of instances and metrics such as CPU Usage.

Viewing Historical Scaling Actions

The following steps illustrate how to view the historical records of scaling actions of an AS group.

1. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
2. Click the **AS Groups** tab and then the name of the target AS group.
3. Click the **Scaling Actions** tab. This page displays historical scaling actions of an AS group, including instance scaling and load balancer migration.

   Scaling Action ID, **Status**, **Scaling Action Type**, **Description**, **Start Time**, and **End Time** of scaling actions are displayed. Click before the scaling action ID to view the resource name, status, and failure cause. You can also use the filtering function in the upper right corner to view scaling actions in a specified period.

4.6 Managing Lifecycle Hooks

Lifecycle hooks enable you to flexibly control addition and removal of ECS instances in AS groups and manage the lifecycle of ECS instances in AS groups. **Figure 4-3** shows the instance lifecycle when no lifecycle hook is added to an AS group.

Figure 4-3 Instance lifecycle when no lifecycle hook is added to an AS group

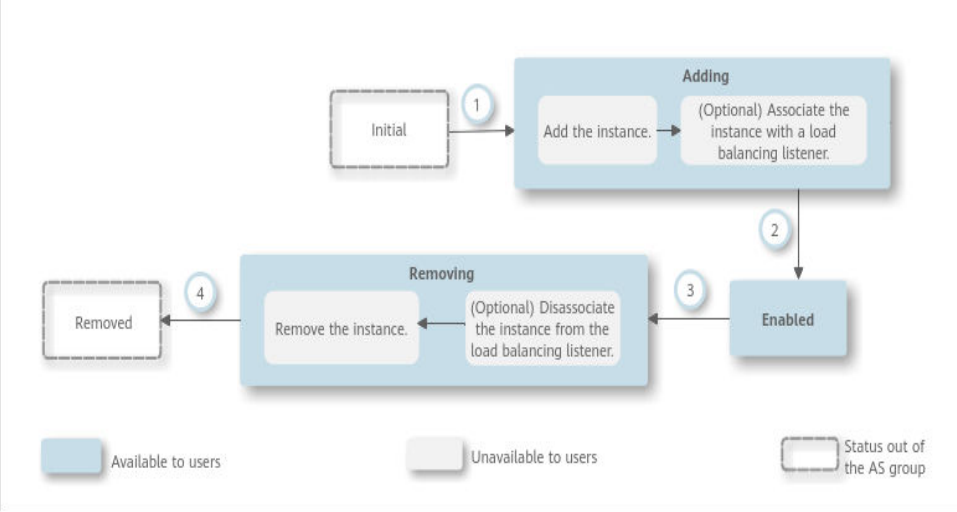
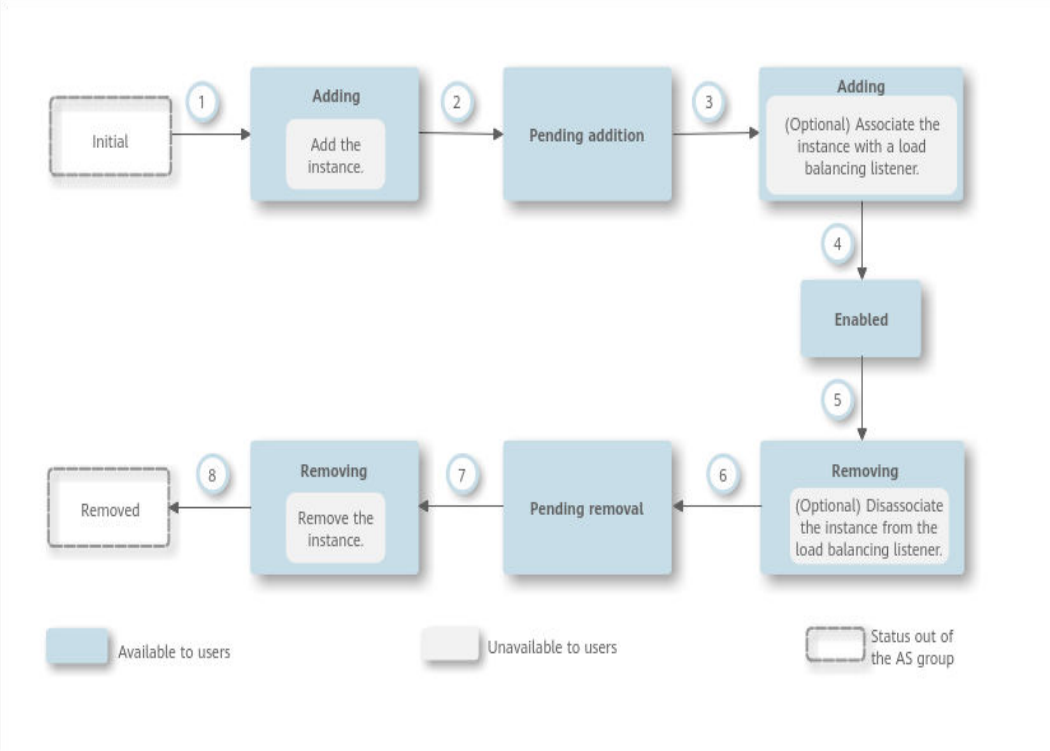


Figure 4-4 shows the instance lifecycle when a lifecycle hook is added to an AS group.

Figure 4-4 Instance lifecycle when a lifecycle hook is added to an AS group



When the AS group scales in or out, the added lifecycle hooks are triggered, the scaling action is suspended, and the instance being added or removed is put into a wait state, as shown in 2 and 6 in **Figure 4-4**. During this period of time, you can perform some custom operations on the instance. For example, you can install or configure software on an instance being added to the AS group. A suspended scaling action will be resumed if either of the following occurs:

- The timeout duration ends.

  Assume that you have set the timeout period to 3,600s by referring to section **Table 4-3**. The suspended scaling action will be automatically resumed if the timeout duration (3,600s) ends.
- A callback action is performed to move the instance out of the wait state. For details, see **Performing a Callback Action**.

Application Scenarios

- Instances newly added to an AS group need to be initialized before they are bound to a load balancer listener. Initialization means the software is installed and configured and the instance is fully ready to accept traffic.
- To remove an instance from an AS group, it needs to be first unbound from the load balancer listener, stop accepting new requests, and finish processing any accepted requests.
- Before instances are removed from an AS group, you may need to back up data or download logs.
- Other scenarios where custom operations need to be performed.

How Lifecycle Hooks Work

After you add lifecycle hooks to an AS group, they work as follows:

- Adding an ECS instance to an AS group

  When an instance is initialized and added to an AS group, a lifecycle hook of the **Instance adding** type is automatically triggered. The instance enters the **Pending addition** state, that is, the instance is suspended by the lifecycle hook. If you have configured a notification object, the system sends a message to the object. After receiving the message, you can perform custom operations, for example, installing software on the instance. The instance remains in a wait state until you complete the custom operations and perform a callback action (see **Performing a Callback Action**) or the timeout duration ends. After the instance moves out of a wait state, the specified default callback action will take place.

  - **Continue**: The instance will be added to the AS group.
  - **Abandon**: The instance will be deleted and a new instance will be created.

  If you have configured multiple **Instance adding** lifecycle hooks, all of them will be triggered when an instance is added to the AS group. If the default callback action of any lifecycle hook is **Abandon**, the instance will be deleted and a new instance will be created. If the default callback action of all lifecycle hooks is **Continue**, the instance is added to the AS group after suspension by the last lifecycle hook is complete.

- Removing an instance from an AS group

  When an instance is removed from an AS group, the instance enters the **Removing** state. After a lifecycle hook is triggered, the instance enters the **Pending removal** state. The system sends messages to the configured notification object. After receiving the message, you can perform custom operations, such as uninstalling software and backing up data. The instance remains in the wait state until you finish the custom operations and perform the default callback operation or the timeout duration ends. After the instance moves out of a wait state, the specified default callback action will take place.

  - **Continue**: The instance is removed from the AS group.
  - **Abandon**: The instance is removed from the AS group.

  If you have configured multiple lifecycle hooks, and the default callback action of all lifecycle hooks is **Continue**, the instance will be removed from the AS group until suspension by the remaining lifecycle hooks time out. If the default callback action of any lifecycle hook is **Abandon**, the instance will be directly removed from the AS group.

Constraints

- You can add, modify, or delete a lifecycle hook when the AS group does not perform a scaling action.
- Up to five lifecycle hooks can be added to one AS group.

Adding a Lifecycle Hook

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
3. Click the name of the AS group to which the lifecycle hook is to be added. On the AS group details page, click the **Lifecycle Hooks** tab and then **Add Lifecycle Hook**.
4. In the displayed **Add Lifecycle Hook** dialog box, set the parameters listed in **Table 4-3**.

Table 4-3 Parameter description

Parameter	Description	Example Value
Hook Name	Specifies the lifecycle hook name. The name can contain letters, digits, underscores (_), and hyphens (-), and cannot exceed 32 characters.	we12_w
Hook Type	Specifies the lifecycle hook type. The value can be **Instance adding** or **Instance removal**. **Instance adding** puts the instance being added to an AS group to the **Pending addition** state. **Instance removal** puts the instance being removed from an AS group to **Pending removal** state.	Instance adding

Parameter	Description	Example Value
Default Callback Action	Specifies the action that the system takes when an instance moves out of a wait state. The default callback action for an **Instance adding** lifecycle hook can be **Continue** or **Abandon**: • **Continue**: If multiple lifecycle hooks are configured for the AS group, and the default callback action of all the hooks is **Continue**, the system will continue to add the instance to the AS group until the all lifecycle hooks time out. • **Abandon**: If multiple lifecycle hooks are configured for the AS group, and the default callback action of one lifecycle hook is **Abandon**, the system will delete the instance and create another one without waiting for the remaining lifecycle hooks to time out. The default callback action for an **Instance removal** lifecycle hook can be **Continue** or **Abandon**: • **Continue**: If only one lifecycle hook is configured for the AS group, the system will remove the instance from the AS group. If multiple lifecycle hooks are configured for the AS group, and the default callback actions of all the hooks are **Continue**, the system will continue to remove the instance from the AS group until all lifecycle hooks time out. • **Abandon**: If multiple lifecycle hooks are configured for the AS group, and the default callback action of one lifecycle hook is **Abandon**, the system will continue to remove the instance from the AS group without waiting for the remaining lifecycle hooks to time out.	Continue
Timeout Duration (s)	Specifies the amount of time for the instances to remain in a wait state. The value ranges from 60s to 86,400s. You can extend the timeout duration or perform a **Continue** or **Abandon** action before the timeout duration ends. For more information about callback actions, see **Performing a Callback Action**.	3600

Parameter	Description	Example Value

Topic	Specifies a notification object for a lifecycle hook. For details, see "Creating a Topic" in **Simple Message Notification User Guide**. When an instance is suspended by the lifecycle hook, the system sends a notification to the object. This notification contains the basic instance information, your custom notification content, and the token for controlling lifecycle actions. An example notification is as follows:
	{
	"service": "AutoScaling",
	"tenant_id": "93075aa73f6a4fc0a3209490cc57181a",
	"lifecycle_hook_type": "INSTANCE_LAUNCHING",
	"lifecycle_hook_name": "test02",
	"lifecycle_action_key": "4c76c562-9688-45c6-b685-7fd732df310a",
	"notification_metadata": "xxxxxxxxxxxx",
	"scaling_instance": {
	"instance_id": "89b421e4-5fa6-4733-bf40-6b07a8657256",
	"instance_name": "as-config-kxeg_RM6OCREY",
	"instance_ip": "192.168.0.202"
	},
	"scaling_group": {
	"scaling_group_id": "fe376277-50a6-4e36-bdb0-685da85f1a82",
	"scaling_group_name": "as-group-wyz01",
	"scaling_config_id": "16ca8027-b6cc-45fc-af2d-5a79996f685d",
	"scaling_config_name": "as-config-kxeg"
	}
	}
		-

Notification Message	After a notification object is configured, the system sends your custom notification to the object.	-

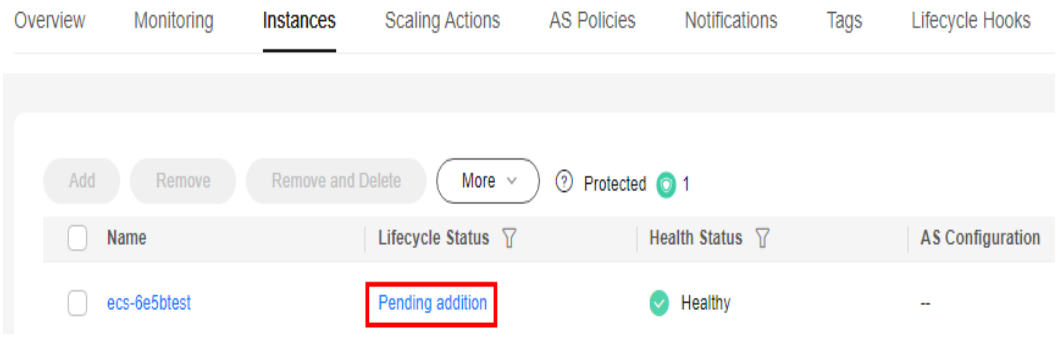
5. Click **OK**.

   The added lifecycle hook is displayed on the **Lifecycle Hooks** page.

Performing a Callback Action

1. On the **AS Groups** page, click the name of the target AS group.
2. On the displayed page, click the **Instances** tab.
3. Locate the instance that has been suspended by the lifecycle hook and click **Pending addition** or **Pending removal** in the **Lifecycle Status** column.

Figure 4-5 Performing a callback action

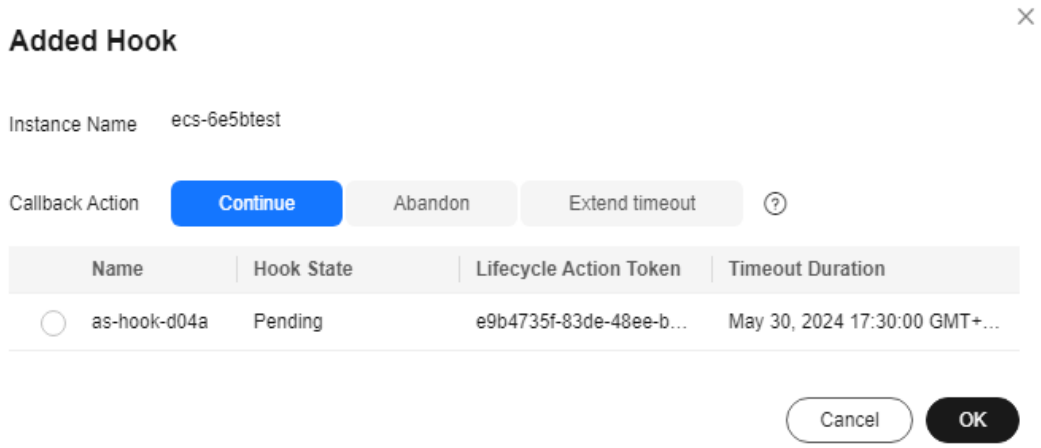


NOTE

Callback actions can only be performed on instances that have been suspended by a lifecycle hook.

4. In the displayed **Added Hook** dialog box, view the suspended instance and all the lifecycle hooks, and perform callback actions on lifecycle hooks.

Figure 4-6 Added Hook dialog box



Callback actions include:

- **Continue**
- **Abandon**
- **Extend timeout**

If you have performed custom operations before the timeout duration ends, select **Continue** or **Abandon** to complete the lifecycle actions. For details about **Continue** and **Abandon**, see **Table 4-3**. If you need more time for custom operations, select **Extend timeout** to extend the timeout duration. Then, the timeout duration will be extended by 3600 seconds each time.

Modifying a Lifecycle Hook

On the **Lifecycle Hooks** page, locate the target lifecycle hook and click **Modify** in the **Operation** column, see **Table 4-3** for parameters. You can modify the parameter except **Hook Name**, such as **Hook Type**, **Default Callback Action**, and **Timeout Duration**.

Deleting a Lifecycle Hook

On the **Lifecycle Hooks** page, locate the target lifecycle hook and click **Delete** in the **Operation** column.

4.7 Configuring Instance Protection

Scenarios

To control whether an instance can be removed automatically from an AS group, use instance protection. Once configured, when AS automatically scales in the AS group, the instance that is protected will not be removed.

Prerequisites

Instance protection does not protect instances from the following:

- Health check replacement if the instance fails health checks
- Manual removal

  NOTE

  - Instance protection does not protect unhealthy instances because such instances cannot provide services.
  - By default, instance protection does not take effect on the ECSs that are newly created in or added to an AS group.
  - If an instance is removed from an AS group, its instance protection setting is lost.

Enabling Instance Protection

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
3. Click the name of the target AS group.
4. Click the **Instances** tab. Select one or more instances and choose **Enable Instance Protection** from the **More** drop-down list. In the displayed **Enable Instance Protection** dialog box, click **Yes**.

Disabling Instance Protection

1. Log in to the management console.
1. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
2. Click the name of the target AS group.
3. Click the **Instances** tab. Select one or more instances and choose **Disable Instance Protection** from the **More** drop-down list. In the displayed **Disable Instance Protection** dialog box, click **Yes**.

4.8 Putting an Instance Into the Standby State

If you want to stop distributing traffic to some instances in your AS group but do not want to remove them from the AS group, you can put the instances on

standby. You can put one or more instances in your AS group on standby, and then stop or restart these instances without worrying about they are removed from the AS group.

Application Scenarios

You cannot control the lifecycle of ECS instances in an AS group. The AS group removes unhealthy instances and does not allow you to stop or restart these instances. As a result, some ECS functions are unavailable. For example, you cannot reset the password, or reinstall or change the OS of these instances.

By putting ECS instances into standby state, you can control their lifecycle and perform operations on them as needed, such as stopping the instances. This facilitates management of instances in your AS group and is helpful in a number of different scenarios.

- If you want to change the OS of an ECS added by a scaling action or stop the ECS, you can set the ECS to standby mode. Then you can perform all operations supported by the ECS service. After completing the operations, cancel standby mode for the ECS.

  For example, you can change the AS configuration for your AS group at any time. This configuration will be used by any instance that is created in the AS group. However, the AS group does not update instances that are running. You can stop these instances, and the AS group will replace them. Alternatively, you can set the instances to standby mode, update software on them, and then cancel standby mode for them.
- If an instance in your AS group associated with a load balancer becomes faulty, you can set the instance to standby mode, after which the load balancer will no longer distribute access traffic to the instance. Then you can log in to the instance, locate and rectify the fault, and restart the instance. After the instance recovers, cancel standby mode for the instance to receive traffic again.

Working Rules

- Set instances to standby mode.

After you set an instance to standby mode, the instance will be automatically unbound from the load balancer associated with the AS group. The instance is still in the AS group, but no health check will be performed on the instance. In this case, load on other instances will increase. To reduce load on other instances and ensure proper service running, you can select **Add the same number of new instances to the AS group** when setting the instance to standby mode.

NOTE

- An instance can be set to standby mode only when the instance is enabled and the AS group has no ongoing scaling action.
- Scaling actions will not remove standby instances from the AS group.
- You can manually remove standby instances from the AS group.

- Cancel standby mode for instances.

After you cancel standby mode for an instance, it will be in running state and receive traffic again. If a load balancer is associated with the AS group, the

instance will be bound to the load balancer. After the instance starts running properly, health check will be performed on it again.

NOTE

Standby mode can be canceled for an instance only when the instance is in standby mode and the AS group has no ongoing scaling action.

Setting Instances to Standby Mode

1. Log in to the management console.
2. Click **Service List**.
3. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
4. Click the name of the target AS group.
5. Click the **Instances** tab. Select one or more instances, click **More**, and select **Set to Standby** in the drop-down list. In the displayed dialog box, select **Add the same number of new instances to the AS group** as you need and click **Yes**.

Canceling Standby Mode for Instances

1. Log in to the management console.

1. Click **Service List**.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
3. Click the name of the target AS group.
4. Click the **Instances** tab. Select one or more instances, click **More**, and select **Cancel Standby** in the drop-down list. In the displayed dialog box, click **Yes**.

5 Bandwidth Scaling

5.1 Creating a Bandwidth Scaling Policy

Scenarios

You can automatically adjust your purchased EIP bandwidth and shared bandwidth using a bandwidth scaling policy. This section describes how to create a bandwidth scaling policy.

When creating a bandwidth scaling policy, you need to configure basic information. The system supports three types of bandwidth scaling policies: alarm-based, scheduled, and periodic.

The basic information for creating a bandwidth scaling policy includes the policy name, resource type, policy type, and trigger condition.

Creating an Alarm-based Bandwidth Scaling Policy

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Bandwidth Scaling**.
3. Click **Create Bandwidth Scaling Policy**.
4. Set parameters, such as the policy name, policy type, and trigger condition. For details, see **Table 5-1**.

Table 5-1 Alarm policy parameters

Parameter	Description	Example Value
Region	Select the region where the bandwidth scaling policy is applied.	-
Policy Name	Enter a name for the bandwidth scaling policy. The name consists of only letters, digits, underscores (_), and hyphens (-).	-

Parameter	Description	Example Value
Resource Type	Select the type of the bandwidth to be adjusted. You can select **EIP** or **Shared bandwidth**.	EIP
EIP	Select the EIP whose bandwidth needs to be scaled. **NOTE** Only bandwidths of pay-per-use EIPs can be scaled.	-
Policy Type	Select **Alarm**.	Alarm
Alarm Rule	You can use an existing alarm rule or create a new one. Alternatively, click **Create Alarm Rule** on the right side of the **Rule Name** parameter and create an alarm rule on the **Alarm Rules** page. For details, see **Creating an Alarm Rule**. To create an alarm rule, configure the following parameters: • Rule Name Enter a name for the new alarm rule, for example, **as-alarm-7o1u**. • Trigger Condition Select a monitoring metric and trigger condition. **Table 5-2** lists the supported monitoring metrics. An example value is **Outbound Traffic Avg. > 100 bit/s**. • Monitoring Interval Specifies the monitoring period for the metric, for example, 5 minutes. • Consecutive Occurrences Specifies the number of consecutive periods in which the triggering condition is met for triggering a scaling action.	-

Parameter	Description	Example Value
Scaling Action	Specifies the execution action in the AS policy. The following scaling action options are available: • Add When a scaling action is triggered, the bandwidth is increased. • Reduce When a scaling action is triggered, the bandwidth is decreased. • Set to The bandwidth is set to a fixed value. **NOTE** The step (minimum unit for bandwidth adjustment) varies depending on the bandwidth value range. The bandwidth will be automatically adjusted to the nearest value according to the actual step. • If the bandwidth is less than or equal to 300 Mbit/s, the default step is 1 Mbit/s. • If the bandwidth ranges from 300 Mbit/s to 1,000 Mbit/s, the default step is 50 Mbit/s. • If the bandwidth is greater than 1,000 Mbit/s, the default step is 500 Mbit/s.	-
Cooldown Period	A cooldown period (in seconds) is the period of time between two scaling actions. AS recounts the cooldown period after a scaling action is complete. During the cooldown period, AS denies all scaling requests triggered by alarm-based policies. Scaling requests triggered manually or by scheduled or periodic policies are not affected.	300s

Table 5-2 Monitoring metrics supported by the alarm policy

Metric	Description
Inbound Bandwidth	Indicates the network rate of inbound traffic.
Inbound Traffic	Indicates the network traffic going into the cloud platform.
Outbound Bandwidth	Indicates the network rate of outbound traffic.
Outbound Traffic	Indicates the network traffic going out of the cloud platform.
Outbound Bandwidth Usage	Indicates the usage of network rate of outbound traffic in the unit of percentage.

5. After setting the parameters, click **Create Now**.

   The newly created bandwidth scaling policy is displayed on the **Bandwidth Scaling** page and is in **Enabled** state by default.

Creating an Alarm Rule

When creating an alarm-based bandwidth scaling policy, you can click **Create Alarm Rule** to the right of **Rule Name** to create an alarm rule. To do so, perform the following operations:

1. Click **Create Alarm Rule** to the right of **Rule Name** to switch to the **Alarm Rules** page of Cloud Eye.
2. On the **Alarm Rules** page, click **Create Alarm Rule** in the upper right corner.
3. Set parameters based on **Figure 5-1** and **Table 5-3**. For more information about how to set alarm rules, see ***Cloud Eye User Guide***.

Figure 5-1 Creating an alarm rule

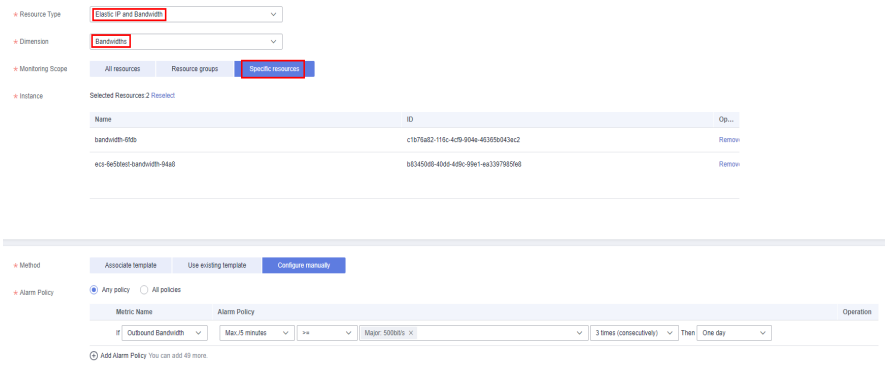


Table 5-3 Key parameters for creating an alarm rule

Parameter	Description	Example Value
Name	Specifies the name of the alarm rule.	alarm-bandwidth
Description	(Optional) Provides supplementary information about the alarm rule.	-
Enterprise Project	Specifies the enterprise project the alarm rule belongs to. Only users with the enterprise project permissions can view and manage the alarm rule.	default
Resource Type	Specifies the name of the service to which the alarm rule applies. Set this parameter to **Elastic IP and Bandwidth**.	Elastic IP and Bandwidth

Parameter	Description	Example Value
Dimension	Specifies the item of the monitored service. Bandwidth scaling adjusts the bandwidth. Therefore, set this parameter to **Bandwidths**.	Bandwidths
Monitoring Scope	Specifies the resources to which the alarm rule applies. Set this parameter to **Specific resources**. Search for the bandwidth using its name or ID. The name and ID can be obtained from the EIP details.	Specific resources
Method	There are three options: **Associate template**, **Use existing template**, and **Configure manually**. **NOTE** After an associated template is modified, the policies contained in this alarm rule to be created will be modified accordingly.	Configure manually
Alarm Policy	Specifies the alarm policy for triggering the alarm rule. Set this parameter as required. For details about the monitoring metrics, see **Table 5-2**.	-

4. After setting the parameters, click **Create**.
5. On the **Create Bandwidth Scaling Policy** page, click  to the right of **Rule Name**, and select the created alarm rule.

Alternatively, you can create an alarm rule on the Cloud Eye console page before creating a bandwidth scaling policy. Ensure that the EIP bandwidth you want to adjust is configured for the alarm rule. After the alarm rule is created, you can select the rule when creating a bandwidth scaling policy.

Creating a Scheduled or Periodic Bandwidth Scaling Policy

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Bandwidth Scaling**.
3. Click **Create Bandwidth Scaling Policy**.
4. Set parameters, such as the policy name, resource type, policy type, and trigger condition. For details, see **Table 5-4**.

Table 5-4 Scheduled or periodic policy parameters

Parameter	Description	Example Value
Region	Select the region where the bandwidth scaling policy is applied.	-

Parameter	Description	Example Value
Policy Name	Enter a name for the bandwidth scaling policy. The name consists of only letters, digits, underscores (_), and hyphens (-).	as-policy-p6g5
Resource Type	Select the type of the bandwidth to be adjusted. You can select **EIP** or **Shared bandwidth**.	EIP
EIP	Select the EIP whose bandwidth needs to be scaled. This parameter is mandatory when **Resource Type** is set to **EIP**. **NOTE** Only bandwidths of pay-per-use EIPs can be scaled.	-
Shared Bandwidth	Specifies the shared bandwidth to be scaled. This parameter is mandatory when **Resource Type** is set to **Shared bandwidth**.	-
Policy Type	Specifies the policy type. You can select a scheduled or periodic policy. If you select **Periodic**, you are required to configure two more parameters: ● Time Range Specifies the time range during which the AS policy can be triggered. ● Period – Day – Week – Month	-
Triggered At	Specifies the time at which the AS policy is triggered.	-

Parameter	Description	Example Value
Scaling Action	Specifies the action to be performed. The following scaling action options are available: • Add When a scaling action is triggered, the bandwidth is increased. • Reduce When a scaling action is triggered, the bandwidth is decreased. • Set to The bandwidth is set to a fixed value. **NOTE** The step (minimum unit for bandwidth adjustment) varies depending on the bandwidth value range. The bandwidth will be automatically adjusted to the nearest value according to the actual step. • If the bandwidth is less than or equal to 300 Mbit/s, the default step is 1 Mbit/s. • If the bandwidth ranges from 300 Mbit/s to 1,000 Mbit/s, the default step is 50 Mbit/s. • If the bandwidth is greater than 1,000 Mbit/s, the default step is 500 Mbit/s.	-
Cooldown Period	A cooldown period (in seconds) is the period of time between two scaling actions. AS recounts the cooldown period after a scaling action is complete. During the cooldown period, AS denies all scaling requests triggered by alarm-based policies. Scaling requests triggered manually or by scheduled or periodic policies are not affected.	300s

5. After setting the parameters, click **Create Now**.

5.2 Viewing Details About a Bandwidth Scaling Policy

Scenarios

You can view details about a bandwidth scaling policy, including its basic information and execution logs. Policy execution logs record details about policy execution.

Procedure

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Bandwidth Scaling**.
3. On the **Bandwidth Scaling** page, click the name of a bandwidth scaling policy to go to the page showing its basic information and view its details.

You can view basic information about the scaling policy, including **Policy Type**, **Trigger Condition**, and **Scaling Action**.

Viewing Execution Logs of a Bandwidth Scaling Policy

In the **Policy Execution Logs** area on the bandwidth scaling policy details page, you can view the policy execution logs. Policy execution logs record the execution status, execution time, original value, and target value of a bandwidth scaling policy.

5.3 Managing a Bandwidth Scaling Policy

Scenarios

You can adjust the bandwidth through a bandwidth scaling policy.

This section describes how to manage bandwidth scaling policies, including enabling, disabling, modifying, deleting, and immediately executing a bandwidth scaling policy.

NOTE

The bandwidth scaling policy configured for a released EIP still occupies the policy quota. Only the account and its IAM users with the global permissions can manage the bandwidth scaling policy. IAM users with permissions to some enterprise projects cannot manage the policy.

Enabling a Bandwidth Scaling Policy

A bandwidth scaling policy can be enabled only when its status is **Disabled**.

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Bandwidth Scaling**.
3. In the bandwidth scaling policy list, locate the row containing the target policy and click **Enable** in the **Operation** column.
4. In the displayed **Enable Bandwidth Scaling Policy** dialog box, click **Yes**.

Disabling a Bandwidth Scaling Policy

A bandwidth scaling policy can be disabled only when its status is **Enabled**.

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Bandwidth Scaling**.
3. In the bandwidth scaling policy list, locate the row containing the target policy and click **Disable** in the **Operation** column.
4. In the displayed **Disable Bandwidth Scaling Policy** dialog box, click **Yes**.

NOTE

After a bandwidth scaling policy is disabled, its status changes to **Disabled**. AS does not automatically trigger any scaling action based on a **Disabled** bandwidth scaling policy.

Modifying a Bandwidth Scaling Policy

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Bandwidth Scaling**.
3. In the bandwidth scaling policy list, locate the row containing the target policy and click the policy name to switch to its details page.

   Click **Modify** in the upper right corner of the page.

   You can also locate the row containing the target policy, click **More** in the **Operation** column, and select **Modify**.
4. Modify parameters. You can modify parameters of a bandwidth scaling policy, such as **Policy Name**, **EIP**, **Policy Type**, **Scaling Action**, and **Cooldown Period**.
5. Click **OK**.

NOTE

A bandwidth scaling policy which is being executed cannot be modified.

Deleting a Bandwidth Scaling Policy

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Bandwidth Scaling**.
3. In the bandwidth scaling policy list, locate the row containing the target policy, click **More** in the **Operation** column, and select **Delete**.
4. In the displayed **Delete Bandwidth Scaling Policy** dialog box, click **Yes**.

   You can also select one or more scaling policies and click **Delete** above the list to delete one or more scaling policies.

NOTE

- You can delete a bandwidth scaling policy when you no longer need it. If you do not need it only during a specified period of time, you are advised to disable rather than delete it.
- A bandwidth scaling policy can be deleted only when it is not being executed.

Executing a Bandwidth Scaling Policy

By executing a bandwidth scaling policy, you can immediately adjust the bandwidth to that configured in the bandwidth scaling policy, instead of having to wait until the trigger condition is met.

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Bandwidth Scaling**.

3. In the bandwidth scaling policy list, locate the row that contains the target policy and click **Execute Now** in the **Operation** column.
4. In the displayed **Execute Bandwidth Scaling Policy** dialog box, click **Yes**.

You can also go to the bandwidth scaling policy details page and click **Execute Now** in the upper right corner.

NOTE

- A bandwidth scaling policy can be executed only when the policy is enabled and no other bandwidth scaling policy is being executed.
- Executing a bandwidth scaling policy does not affect automatic adjustment of the bandwidth when the trigger condition of the policy is met.
- If **Policy Type** is set to **Alarm** and **Alarm Policy Type** to **Refined scaling**, the bandwidth scaling policy cannot be executed immediately.

6 AS Group and Instance Monitoring

6.1 Health Check

Health Check Methods

A health check removes unhealthy instances from an AS group. Then, AS adds new instances to the AS group so that the number of instances is the same as the expected number. There are two types of AS group health checks.

- **ECS health check**: checks the ECS instance status. If an ECS instance is stopped or deleted, it is considered unhealthy. **health check** is the default health check method for an AS group. The AS group periodically uses the check result to determine the status of each ECS instance in the AS group. If an ECS instance is unhealthy, AS removes the ECS instance from the AS group.
- **ELB health check**: determines the ECS instance status using a load balancing listener. If the AS group uses load balancers, the health check method can also be **ELB health check**.

  If you add multiple load balancers to an AS group, an ECS instance is considered healthy only when all load balancers detect that the ECS instance is healthy. If any load balancer detects that an ECS instance is unhealthy, the ECS instance will be removed from the AS group.

In both the **ECS health check** and **ELB health check** methods, AS removes unhealthy ECS instances from the AS group. Whether a removed ECS instance will be deleted depends on how the instance was added to the AS group.

Table 6-1 Instance removal and deletion rules

Instance Type	Description	Billing Mode	Removed If Unhealthy	Deleted When Removed
Automatically added instances	Instances automatically created and added to an AS group in a scaling action	Pay-per-use By default, this type of instances is billed on a pay-per-use basis.	Yes	Yes
		Yearly/Monthly The billing mode of an automatically added instance can be manually changed from pay-per-use to yearly/monthly.	Yes	No
Manually added instances	Instances manually created and added to an AS group	Pay-per-use	Yes	No
		Yearly/Monthly	Yes	No

Constraints

- Even when an AS group is disabled, AS still checks the health of instances in the AS group, but does not remove unhealthy instances.
- AS does not check the health of instances in standby state.

6.2 Configuring Notifications for an AS Group

Scenarios

After the SMN service is provisioned, you can get notifications about AS group changes, such as successful or failed instance additions or removals or AS group exceptions.

To configure notifications for an AS group, you need to specify a notification event and topic. You need to create a notification topic on the SMN console. When the notification scenario matched with the notification topic appears, the AS group sends a notification to the subscribers.

A maximum of five notifications can be configured for an AS group.

Procedure

1. Log in to the management console.
1. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
2. Click the AS group name. On the AS group details page, click the **Notifications** tab and then click **Add Notification**.
3. Set the parameters listed in **Table 6-2**.

   Table 6-2 Parameter description

Parameter	Description	Example Value
Event	When at least one of the following conditions is met, SMN sends a notification to the user: • Instance addition succeeded • Instance removal succeeded • Errors occurred in an AS group • Instance addition failed • Instance removal failed	-
Topic	Select an existing topic. For details about how to create a topic, see **Creating a Topic**.	-

4. Click **OK**.

6.3 Recording AS Operations

Scenarios

AS can work together with Cloud Trace Service (CTS) to record resource operations. CTS can record operations performed on the management console, operations performed by calling APIs, and operations triggered within the cloud system.

If you have enabled CTS, when a call is made to the AS API, the operation will be reported to CTS which will then deliver the operation record to a specified OBS bucket for storage. With CTS, you can record AS operation logs for query, audit, and backtracking.

Obtaining AS Information in CTS

After you enable CTS, whenever an AS API is called, the operation is recorded in a log file. On the **Cloud Trace Service** console, you can view operation records for the last 7 days. For details, see **Viewing CTS Traces in the Trace List**. To store operation records for a longer period, you can transfer them to Object Storage Service (OBS) buckets.

Table 6-3 lists the AS operations that can be recorded by CTS.

Table 6-3 AS operations that can be recorded by CTS

Operation	Resource Type	Trace Name
Creating an AS group	scaling_group	createScalingGroup
Modifying an AS group	scaling_group	modifyScalingGroup
Deleting an AS group	scaling_group	deleteScalingGroup
Enabling an AS group	scaling_group	enableScalingGroup
Disabling an AS group	scaling_group	disableScalingGroup
Performing operations on an AS group	scaling_group	operateScalingGroup
Creating an AS configuration	scaling_configuration	createScalingConfiguration
Deleting an AS configuration	scaling_configuration	deleteScalingConfiguration
Deleting AS configurations in a batch	scaling_configuration	batchDeleteScalingConfiguration
Creating an AS policy	scaling_policy	createScalingPolicy
Modifying an AS policy	scaling_policy	modifyScalingPolicy
Deleting an AS policy	scaling_policy	deleteScalingPolicy
Enabling an AS policy	scaling_policy	enableScalingPolicy
Disabling an AS policy	scaling_policy	disableScalingPolicy
Executing an AS policy	scaling_policy	executeScalingPolicy

Operation	Resource Type	Trace Name
Performing operations on an AS policy	scaling_policy	operateScalingPolicy
Enabling AS policies in a batch	scaling_policy	batchEnableScalingPolicies
Disabling AS policies in a batch	scaling_policy	batchDisableScalingPolicies
Removing an instance	scaling_instance	removeInstance
Removing instances in batches	scaling_instance	batchRemoveInstances
Adding instances in batches	scaling_instance	batchAddInstances
Performing operations on instances in batches	scaling_instance	batchOperateInstance
Enabling instance protection in a batch	scaling_instance	batchProtectInstances
Disabling instance protection in a batch	scaling_instance	batchUnprotectInstances
Putting instances into standby in a batch	scaling_instance	batchEnterStandbyInstances
Configuring a notification	scaling_notification	putScalingNotification
Deleting a notification	scaling_notification	deleteScalingNotification
Creating a lifecycle hook	scaling_lifecycle_hook	createLifecycleHook
Modifying a lifecycle hook	scaling_lifecycle_hook	modifyLifecycleHook

Operation	Resource Type	Trace Name
Deleting a lifecycle hook	scaling_lifecycle_hook	deleteLifecycleHook

6.4 Viewing CTS Traces in the Trace List

Scenarios

Cloud Trace Service (CTS) records operations performed on cloud service resources. A record contains information such as the user who performed the operation, IP address, operation content, and returned response message. These records facilitate security auditing, issue tracking, and resource locating. They also help you plan and use resources, and identify high-risk or non-compliant operations.

What Is a Trace?

A trace is an operation log for a cloud service resource, tracked and stored by CTS. Traces record operations such as adding, modifying, or deleting cloud service resources. You can view them to identify who performed operations and when for detailed tracking.

Constraints

- You can only query operation records of the last seven days on the CTS console. They are automatically deleted upon expiration and cannot be manually deleted. To store them for longer than seven days, configure transfer to Object Storage Service (OBS) or Log Tank Service (LTS) so that you can view them in OBS buckets or LTS log groups.
- After creating, modifying, or deleting a cloud service resource, you can query management traces on the CTS console 1 minute later and query data traces 5 minutes later.
- Data traces are not displayed in the trace list of the new version. To view them, you need to go to the old version.

Viewing Real-Time Traces in the Trace List of the New Edition

Step 1 Log in to the **CTS console**.

Step 2 In the navigation pane, choose **Trace List**.

Step 3 In the time range drop-down list above the trace list, select a desired query time range: **Last 1 hour**, **Last 1 day**, or **Last 1 week**. You can also select **Custom** to specify a custom time range within the last seven days.

Step 4 The search box above the trace list supports advanced queries. Combine one or more filters to refine your search.

Table 6-4 Trace filtering parameters

Parameter	Description
Trace Name	Name of a trace. The entered value is case-sensitive and requires an exact match. Fuzzy matching is not supported. For details about the operations that can be audited for each cloud service, see **Supported Services and Operations**. Example: **updateAlarm**
Trace Source	Cloud service name abbreviation. The entered value is case-sensitive and requires an exact match. Fuzzy matching is not supported. Example: **IAM**
Resource Name	Name of a cloud resource involved in a trace. The entered value is case-sensitive and requires an exact match. Fuzzy matching is not supported. If the cloud resource involved in the trace does not have a resource name or the corresponding API operation does not involve the resource name parameter, leave this field empty. Example: **ecs-name**
Resource ID	ID of a cloud resource involved in a trace. The entered value is case-sensitive and requires an exact match. Fuzzy matching is not supported. Leave this field empty if the resource has no resource ID or if resource creation failed. Example: *{VM ID}*
Trace ID	Value of the **trace_id** parameter for a trace reported to CTS. The entered value requires an exact match. Fuzzy matching is not supported. Example: **01d18a1b-56ee-11f0-ac81-******1e229**
Resource Type	Type of a resource involved in a trace. The entered value is case-sensitive and requires an exact match. Fuzzy matching is not supported. For details about the resource types of each cloud service, see **Supported Services and Operations**. Example: **user**
Operator	User who triggers a trace. Select one or more operators from the drop-down list. If the value of **trace_type** in a trace is **SystemAction**, the operation is triggered by the service and the trace's operator may be empty.

Parameter	Description
Trace Status	Select one of the following options from the drop-down list: • **normal**: The operation succeeded. • **warning**: The operation failed. • **incident**: The operation caused a fault that is more serious than a normal failure, for example, causing other faults.

Step 5 On the **Trace List** page, you can also export and refresh the trace list, and customize columns to display.

- Enter any keyword in the search box and press **Enter** to filter desired traces.
- Click **Export** to export all traces in the query result as an .xlsx file. The file can contain up to 5,000 records.
- Click to view the latest information about traces.
- Click to customize the information to be displayed in the trace list. If **Auto wrapping** is enabled ( ), excess text will move down to the next line; otherwise, the text will be truncated. By default, this function is disabled.

Step 6 (Optional) On the **Trace List** page of the new edition, click **Old Edition** in the upper right corner to switch to the **Trace List** page of the old edition.

----End

Viewing Traces in the Trace List of the Old Edition

Step 1 Log in to the **CTS console**.

Step 2 In the navigation pane, choose **Trace List**.

Step 3 Each time you log in to the CTS console, the new edition is displayed by default. Click **Old Edition** in the upper right corner to switch to the trace list of the old edition.

Step 4 In the upper right corner of the page, set a desired query time range: **Last 1 hour**, **Last 1 day**, or **Last 1 week**. You can also click **Customize** to specify a custom time range within the last seven days.

Step 5 Set filters to search for your desired traces.

Table 6-5 Trace filtering parameters

Parameter	Description
Trace Type	Select **Management** or **Data**. • Management traces record operations performed by users on cloud service resources, including creation, modification, and deletion. • Data traces are reported by OBS and record operations performed on data in OBS buckets, including uploads and downloads.

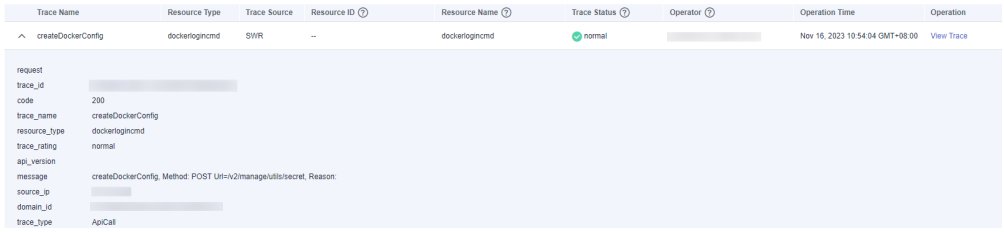
Parameter	Description
Trace Source	Select the name of the cloud service that triggers a trace from the drop-down list.
Resource type	Select the type of the resource involved in a trace from the drop-down list. For details about the resource types of each cloud service, see **Supported Services and Operations**.
Search By	Select one of the following options: • **Resource ID**: ID of the cloud resource involved in a trace. Leave this field empty if the resource has no resource ID or if resource creation failed. • **Trace name**: name of a trace. For details about the operations that can be audited for each cloud service, see **Supported Services and Operations**. • **Resource name**: name of the cloud resource involved in a trace. If the cloud resource involved in the trace does not have a resource name or the corresponding API operation does not involve the resource name parameter, leave this field empty.
Operator	User who triggers a trace. Select one or more operators from the drop-down list. If the value of **trace_type** in a trace is **SystemAction**, the operation is triggered by the service and the trace's operator may be empty.
Trace Status	Select one of the following options: • **Normal**: The operation succeeded. • **Warning**: The operation failed. • **Incident**: The operation caused a fault that is more serious than a normal failure, for example, causing other faults.

Step 6 Click **Query**.

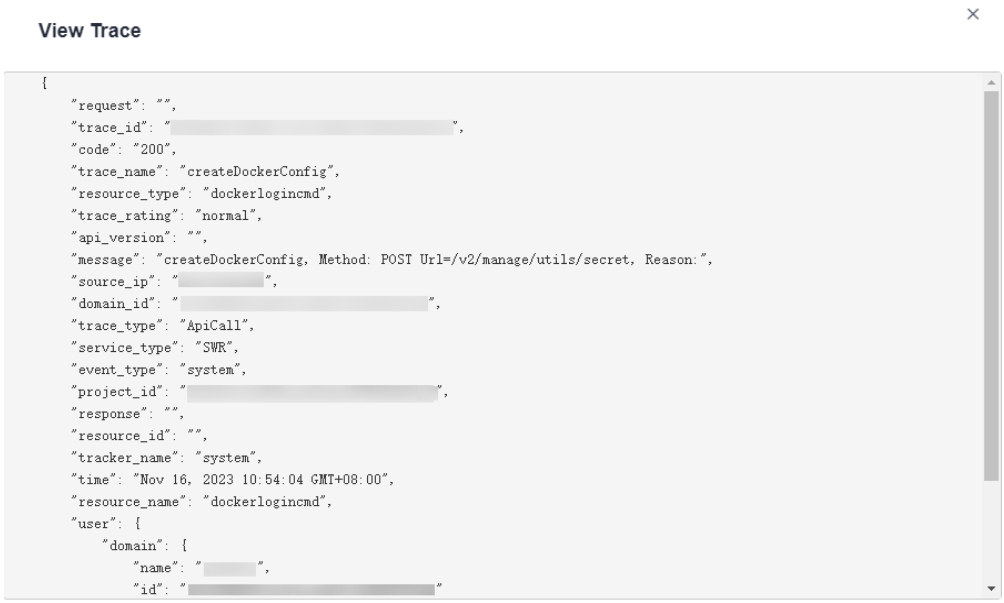
Step 7 On the **Trace List** page, you can also export and refresh the trace list.

- Click **Export** to export all traces in the query result as a CSV file. The file can contain up to 5,000 records.
- Click to view the latest information about traces.

Step 8 Click on the left of a trace to expand its details.



Step 9 Click **View Trace** in the **Operation** column. The trace details are displayed.



Step 10 (Optional) On the **Trace List** page of the old edition, click **New Edition** in the upper right corner to switch to the **Trace List** page of the new edition.

----End

Helpful Links

- For details about the key fields in the trace structure, see **Trace Structure** and **Example Traces**.

6.5 Adding Tags to AS Groups and Instances

Scenarios

If you have many resources of the same type, you can use tags to manage resources flexibly. You can identify specified resources quickly using the tags allocated to them.

Using a tag, you can assign custom data to each AS group. You can organize and manage AS groups, for example, classify AS group resources by usage, owner, or environment.

Each tag contains a key and a value. You can specify the key and value for each tag. A key can be a category associated with certain values, such as usage, owner, and environment.

For example, if you want to distinguish between the test environment and production environment, you can allocate a tag with the key **environment** to each AS group. For the test environment, the key value is **test** and for the production environment, the key value is **production**. You are advised to use one or more groups of consistent tags to manage your AS group resources.

After you allocate a tag to an AS group, the system will automatically add the tag to the instances automatically created in the AS group. If you add a tag to an AS group or modify the tag, the new tag will be added to the ECSs automatically created in the AS group. Creating, deleting, or modifying the tag of an AS group will have no impact on the ECSs in the AS group.

Restrictions of Using Tags

You must observe the following rules when using tags:

- By default, each AS group can have a maximum of 10 tags added to it.
- Each tag contains a key and a value.
- You can set the tag value to an empty character string.
- If you delete an AS group, all tags of it will also be deleted.

Adding Tags

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
3. Click the AS group name. On the AS group details page, click the **Tags** tab and then click **Add Tag**.
4. Set the parameters listed in **Table 6-6**.

Table 6-6 Tag naming rules

Parameter	Requirement	Example Value
Tag Key	• It cannot be empty. • Each key must be unique to the AS group. • A key can contain a maximum of 36 characters, including digits, letters, underscores (_), hyphens (-), and Unicode characters from \u4e00 to \u9fff.	Organization

Parameter	Requirement	Example Value
Tag Value	• It can be an empty string. • Only one tag value can be added to a tag key. • A tag value can contain a maximum of 43 characters, including digits, letters, underscores (_), periods (.), hyphens (-), and Unicode characters from \u4e00 to \u9fff.	Apache

5. Click **OK**.

Modifying or Deleting Tags of an AS Group

1. Log in to the management console.
1. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
2. Click the AS group name. On the **Overview** page, click the **Tags** tab.
3. Locate the row that contains the tag and click **Edit** or **Delete** in the **Operation** column.

   After clicking **Edit**, configure required parameters. For details, see **Table 6-6**.

   After you click **Delete**, the added tag will be deleted.

6.6 Monitoring Metrics

Description

This section describes the monitoring metrics reported by AS to Cloud Eye and defines the namespace for the metrics. You can use Cloud Eye to query monitoring metrics and alarms of AS.

Namespace

SYS.AS

Items in Alarm Policies

You can monitor key metrics of AS. Then you can handle exceptions in a timely manner. For details about the items in an alarm policy, see the following table.

Items in an alarm policy for metrics

Item	Description	Example Value
Metric	Specifies the name of an AS metric.	CPU Usage

Item	Description	Example Value
Statistic	Specifies the metric value type. There are six types of values: **Max.**, **Mini.**, **Avg.**, **Raw data**, **Variance**, and **Sum**. • **Max.** is the highest value observed during a rollup period. • **Min.** is the lowest value observed during a rollup period. • **Avg.** is the value calculated by averaging raw data during a rollup period. • **Raw data** indicates the metric data that is not processed or converted. • **Variance** indicates the difference between each data point in the original value and the average value within a rollup period. • **Sum** is the sum of raw data during a rollup period.	Raw data
Operator	Specifies the operator used to compare the metric value and the threshold. Supported operators: >, >=, <, and <=	>
Threshold	Specifies the threshold of the metric value.	70

AS Metrics

Table 6-7 lists the AS metrics supported by Cloud Eye.

Table 6-7 AS metrics

Metric ID	Metric	Description	Value Range	Unit	Conversion Rule	Monitored Object (Dimension)	Monitoring Interval (Raw Data)
cpu_util	CPU Usage	CPU usage of an AS group Formula: Total CPU usage of all ECS instances in an AS group/Number of ECS instances in the AS group	≥0	%	N/A	instance_id	5 minutes

Metric ID	Metric	Description	Value Range	Unit	Conversion Rule	Monitored Object (Dimension)	Monitoring Interval (Raw Data)
mem_util	Memory Usage	Memory usage of an AS group Formula: Total memory usage of all ECS instances in an AS group/Number of ECS instances in the AS group **NOTE** This metric is unavailable if the image has no VMTools installed.	≥0	%	N/A	instance_id	5 minutes
instance_num	Number of Instances	Number of available ECS instances in an AS group Formula: Total number of ECS instances in **Enabled** state in the AS group	≥0	count	N/A	instance_id	5 minutes
network_incoming_bytes_rate_inband	Inband Incoming Rate	Number of incoming bytes per second on an ECS in an AS group Formula: Total inband incoming rate of all ECS instances in an AS group/Number of ECS instances in the AS group	≥0	byte/s	1024 (IEC)	instance_id	5 minutes

Metric ID	Metric	Description	Value Range	Unit	Conversion Rule	Monitored Object (Dimension)	Monitoring Interval (Raw Data)
network_outgoing_bytes_rate_inband	Inband Outgoing Rate	Number of outgoing bytes per second on an ECS in an AS group Formula: Total inband outgoing rate of all ECS instances in an AS group/Number of ECS instances in the AS group	≥0	byte/s	1024 (IEC)	instance_id	5 minutes
disk_read_bytes_rate	Disks Read Rate	Number of bytes read from an AS group per second Formula: Total disk read rate of all ECS instances in an AS group/Number of ECS instances in the AS group	≥0	byte/s	1024 (IEC)	instance_id	5 minutes
disk_write_bytes_rate	Disks Write Rate	Number of bytes written to an AS group per second Formula: Total disk write rate of all ECS instances in an AS group/Number of ECS instances in the AS group	≥0	byte/s	1024 (IEC)	instance_id	5 minutes
disk_read_requests_rate	Disks Read Requests	Number of disk read requests sent to an AS group per second Formula: Total number of disk read requests sent to all ECS instances in an AS group/Number of ECS instances in the AS group	≥0	r/s	N/A	instance_id	5 minutes

Metric ID	Metric	Description	Value Range	Unit	Conversion Rule	Monitored Object (Dimension)	Monitoring Interval (Raw Data)
disk_write_requests_rate	Disks Write Requests	Number of disk write requests sent to an AS group per second Formula: Total number of disk write requests sent to all ECS instances in an AS group/Number of ECS instances in the AS group	≥0	r/s	N/A	instance_id	5 minutes
cpu_usage	(Agent) CPU Usage	Agent CPU usage of an AS group Formula: Total (Agent) CPU usage of all ECS instances in an AS group/ Number of ECS instances in the AS group	0-100	%	N/A	instance_id	1 minute
mem_usedPercent	(Agent) Memory Usage	(Agent) Memory usage of an AS group Formula: Total (Agent) memory usage of all ECS instances in an AS group/Number of ECS instances in the AS group	0-100	%	N/A	instance_id	1 minute
load_average1	(Agent) 1-Minute Load Average	Average CPU load of all ECSs in an AS group in the last 1 minute	≥0	-	N/A	instance_id	1 minute
load_average5	(Agent) 5-Minute Load Average	Average CPU load of all ECSs in an AS group in the last 5 minutes	≥0	-	N/A	instance_id	1 minute

Metric ID	Metric	Description	Value Range	Unit	Conversion Rule	Monitored Object (Dimension)	Monitoring Interval (Raw Data)
load_average15	(Agent) 15-Minute Load Average	Average CPU load of all ECSs in an AS group in the last 15 minutes	≥0	-	N/A	instance_id	1 minute
gpu_usage_gpu	(Agent) GPU Usage	(Agent) GPU usage of an AS group Formula: Total (Agent) GPU usage of all ECS instances in an AS group/ Number of ECS instances in the AS group	0-100	%	N/A	instance_id	1 minute
gpu_usage_mem	(Agent) GPU Memory Usage	(Agent) GPU memory usage of an AS group Formula: Total (Agent) GPU memory usage of all ECS instances in an AS group/Number of ECS instances in the AS group	0-100	%	N/A	instance_id	1 minute

NOTE

Monitoring metrics are classified into metrics with Agent and without Agent. For some OSs, you need to install the Agent to obtain the corresponding monitoring metrics. In this case, select the monitoring metrics with Agent, for example, (Agent) Memory Usage.

NOTE

OSs determine whether the **Memory Usage**, **Inband Outgoing Rate**, and **Inband Incoming Rate** metrics are supported. For details, see ***Elastic Cloud Server User Guide***.

Before using Agent monitoring metrics, make sure that the Agent has been installed on the instances in the AS group. For details, see **How Do I Install the Agent on the Instances in an AS Group to Use Agent Monitoring Metrics?**

Dimension

Key	Value
instance_id	AS group ID You can obtain the value by referring to **Querying Instances in an AS Group**.

6.7 Viewing Monitoring Metrics

Scenarios

The cloud platform provides Cloud Eye to help you obtain the status of your ECS instances. This section describes how to view details of AS group metrics to obtain information about the status of the ECS instances in the AS group.

Prerequisites

The ECS instance is running properly.

NOTE

- Monitoring metrics such as **CPU Usage** and **Disks Read Rate** are available only when there is at least one instance in an AS group. If not, only the **Number of Instances** metric is available.
- Monitoring data is not displayed for a stopped, faulty, or deleted ECS instance. After such an ECS instance restarts or recovers, the monitoring data is available.

Viewing Monitoring Metrics on Auto Scaling

1. Log in to the management console.
2. Under **Compute**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**. Then click the **AS Groups** tab.
3. On the **AS Groups** page, find the AS group to view monitoring data and click its name.
4. Click the **Monitoring** tab to view the monitoring data.

   You can view data of the last one, three, 12, or 24 hours, or last 7 days. If you want to view data for a longer time range, click **View details** to go to the **Cloud Eye** page, hover your mouse over a graph, and click  .

Viewing Monitoring Metrics on Cloud Eye

1. Log in to the management console.
2. Click  in the upper left corner to select a region and a project.
3. Under **Management & Governance**, select **Cloud Eye**.
4. In the navigation pane on the left, choose **Cloud Service Monitoring** > **Auto Scaling**.

5. Locate the row that contains the target AS group and click **View Metric** in the **Operation** column to view monitoring data.

   You can view data of the last one, three, 12, or 24 hours, or last 7 days. Hover your mouse over a graph and click  to view data for a longer time range.

NOTE

It can take a period of time to obtain and transfer the monitoring data. Therefore, wait for a while and then check the data.

6.8 Setting Monitoring Alarm Rules

Scenarios

Alarm rules allow you to customize monitored objects and notification policies and obtain the status of your ECS instances at any time.

Procedure

1. Log in to the management console.
2. Click  in the upper left corner and select the desired region and project.
3. Under **Management & Governance**, select **Cloud Eye**.
4. In the navigation pane, choose **Alarm Management** > **Alarm Rules**.
5. On the **Alarm Rules** page, click **Create Alarm Rule** to create an alarm rule for the AS service or modify an existing alarm rule of the AS service.
6. After setting the parameters, click **Create**.

   NOTE

   - For more information about how to set alarm rules, see *Cloud Eye User Guide*.
   - You can create alarm rules on the Cloud Eye console to dynamically expand resources.

7 Quota Management

What Is a Quota?

A quota limits the quantity of a resource available to users, thereby preventing spikes in the usage of the resource. For example, an AS group quota limits the number of AS groups that can be created per account. You can also request for an increased quota if your existing quota cannot meet your service requirements.

This section describes how to view the used AS quotas and the total quotas in a specified region.

How Do I View My Quotas?

1. Log in to the **management console**.
2. Click  in the upper left corner and select the desired region and project.
3. In the upper right corner of the page, choose **Resources** > **My Quotas**.

   The **Quotas** page is displayed.
4. View the used and total quota of each type of resources on the displayed page.

   If a quota cannot meet service requirements, apply for a higher quota.

How Do I Apply for a Higher Quota?

1. Log in to the **management console**.
2. In the upper right corner of the page, choose **Resources** > **My Quotas**.

   The **Quotas** page is displayed.
3. Click **Increase Quota** in the upper right corner of the page.
4. On the **Create Service Ticket** page, configure parameters as required.

   In the **Problem Description** area, fill in the content and reason for adjustment.
5. After all necessary parameters are configured, select **I have read and agree to the Ticket Service Protocol and Privacy Statement** and click **Submit**.

8 Permissions Management

8.1 Creating a User and Granting AS Permissions

Scenarios

IAM can help you implement fine-grained permissions control over your AS resources. With IAM, you can:

- Create IAM users for employees based on your enterprise's organizational structure. Each IAM user will have their own security credentials for accessing AS resources.
- Grant users only the permissions required to perform a given task based on their job responsibilities.
- Use **IAM** to entrust a Huawei Cloud account or cloud service to perform efficient O&M on your AS resources.

If your Huawei Cloud account does not require individual IAM users, skip this section.

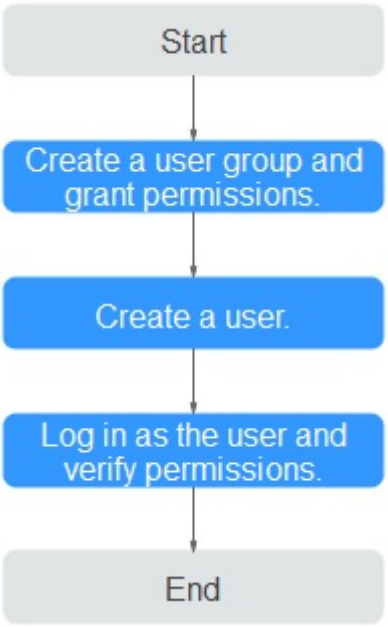
This section describes the procedure for granting permissions. **Figure 8-1** shows the process flow.

Prerequisites

Before granting permissions to user groups, you should learn about the permissions (**Permissions Management**) supported by AS and choose policies or roles based on service requirements. To grant permissions for other services, learn about all **system-defined permissions** supported by IAM.

Process Flow

Figure 8-1 Process for granting AS permissions



1. **Create a user group and assign permissions to it**.

   Create a user group on the IAM console, and assign the ASReadOnlyAccess permissions to the group.

2. **Create an IAM user and add it to the user group**.

   Create a user on the IAM console and add the user to the group created in **1**.

3. **Log in** and verify permissions.

   Log in to the **management console** using the IAM user, switch to a region where the permissions take effect, and verify the permissions.

   - Choose **Service List** > **Auto Scaling**. Then, click **Create AS Group** on the AS console. If a message appears indicating that you have insufficient permissions to perform the operation, the ASReadOnlyAccess policy is in effect.
   - Choose another in the **Service List**. If a message appears indicating that you have insufficient permissions to access the service, the ASReadOnlyAccess policy is in effect.

8.2 AS Custom Policies

Scenarios

Custom policies can be created to supplement the system-defined policies of AS. For the actions that can be added to custom policies, see **Permissions Policies and Supported Actions**.

You can create custom policies in either of the following ways:

- Visual editor: Select cloud services, actions, resources, and request conditions. This does not require knowledge of policy syntax.
- JSON: Edit JSON policies from scratch or based on an existing policy.

For operation details, see **Creating a Custom Policy**. The following section contains examples of common AS custom policies.

Example Custom Policies

- Example 1: Grant permissions to remove instances from AS groups and create AS configurations.

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "as:instances:delete",
                "as:configs:create"
            ]
        }
    ]
}
```

- Example 2: Grant permissions to deny AS group deletion.

  A policy with only "Deny" permissions must be used together with other policies. If the permissions assigned to a user contain both "Allow" and "Deny", the "Deny" permissions take precedence over the "Allow" permissions.

  Assume that you need to grant the permissions of the **AutoScaling FullAccess** policy to a user but want to prevent the user from deleting AS groups. You can create a custom policy for denying AS group deletion, and attach this policy together with the **AutoScaling FullAccess** policy to the user. As an explicit deny in any policy overrides any allows, the user can perform all operations on AS excepting deleting AS groups. The following is an example of a deny policy:

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Action": [
                "as:groups:delete"
            ],
            "Effect": "Deny"
        }
    ]
}
```

