Auto Scaling

Getting Started

Issue 13

Date 2026-01-04

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 Scaling Your Website Dynamically

Scenarios

When service requirements change frequently and irregularly, you can configure alarm-based AS policies in your application system to adjust resources dynamically. When the conditions of a policy are met, the system automatically changes the expected number of instances to trigger a scaling action to add or remove instances. This reduces the workload of manual resource adjustments caused by service changes or peak hours, helping you save resource and labor costs.

Assume that you are an e-commerce vendor that runs a website on an ECS. If you plan to launch a big sales promotion at 00:00 on a certain day but the ECS is unable to handle service requests during peak hours, you can address this issue according the guidance of this section.

Procedure

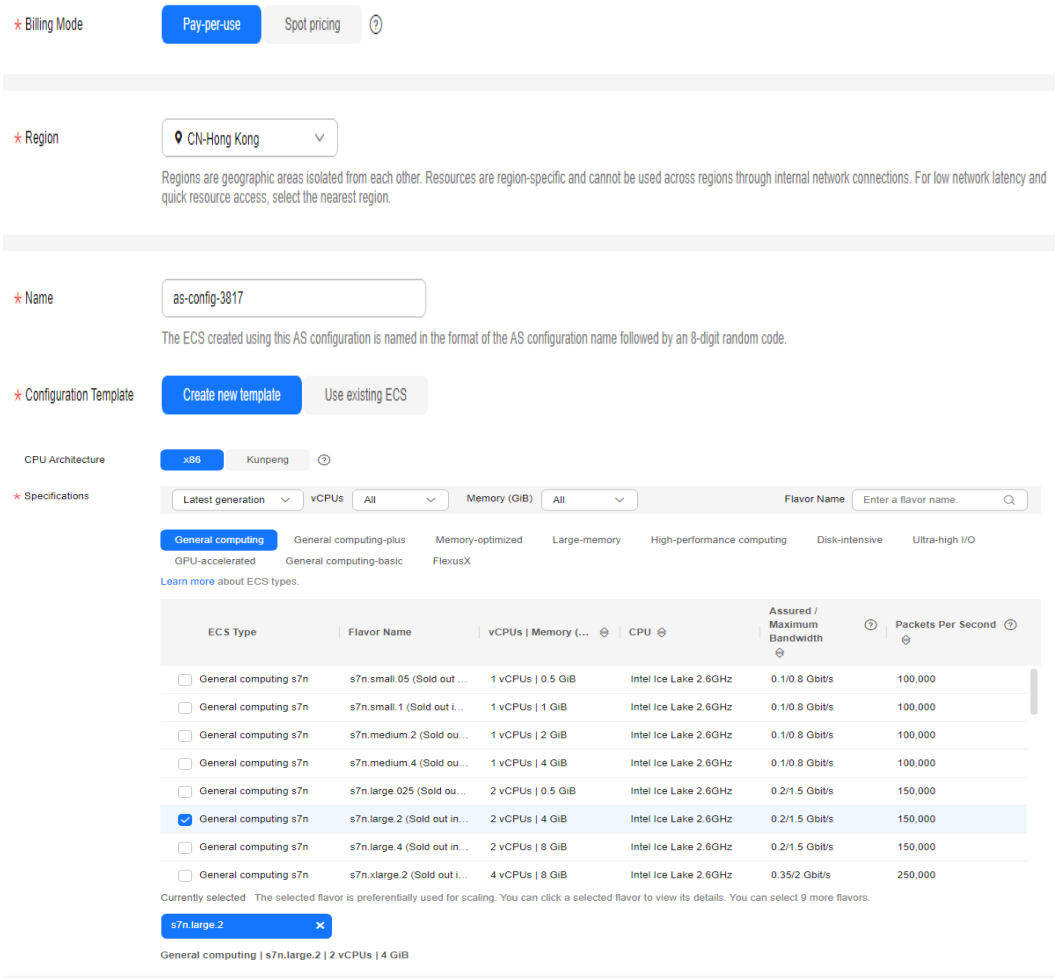
Step	Description
Step 1: Create an AS Configuration	Configure the specifications, image, and disks for new ECS instances.
Step 2: Create an AS Group	Configure the maximum, minimum, and expected number of instances and other parameters.
Step 3: Add an AS Policy	Create an AS policy to adjust service resources.
(Optional) Step 4: Manually Add Instances	Add instances to the AS group when the group is enabled, no scaling action is in progress in it, and the number of instances in it is less than the maximum allowed.
(Optional) Step 5: View Monitoring Data	View monitoring data changes during the scaling action.

Step 1: Create an AS Configuration

In this step, create an AS configuration using the example settings. For more information, see **AS Configuration**.

1. Log in to the **AS console**.
2. Click **Create AS Configuration** and set parameters.

Figure 1-1 Creating an AS configuration



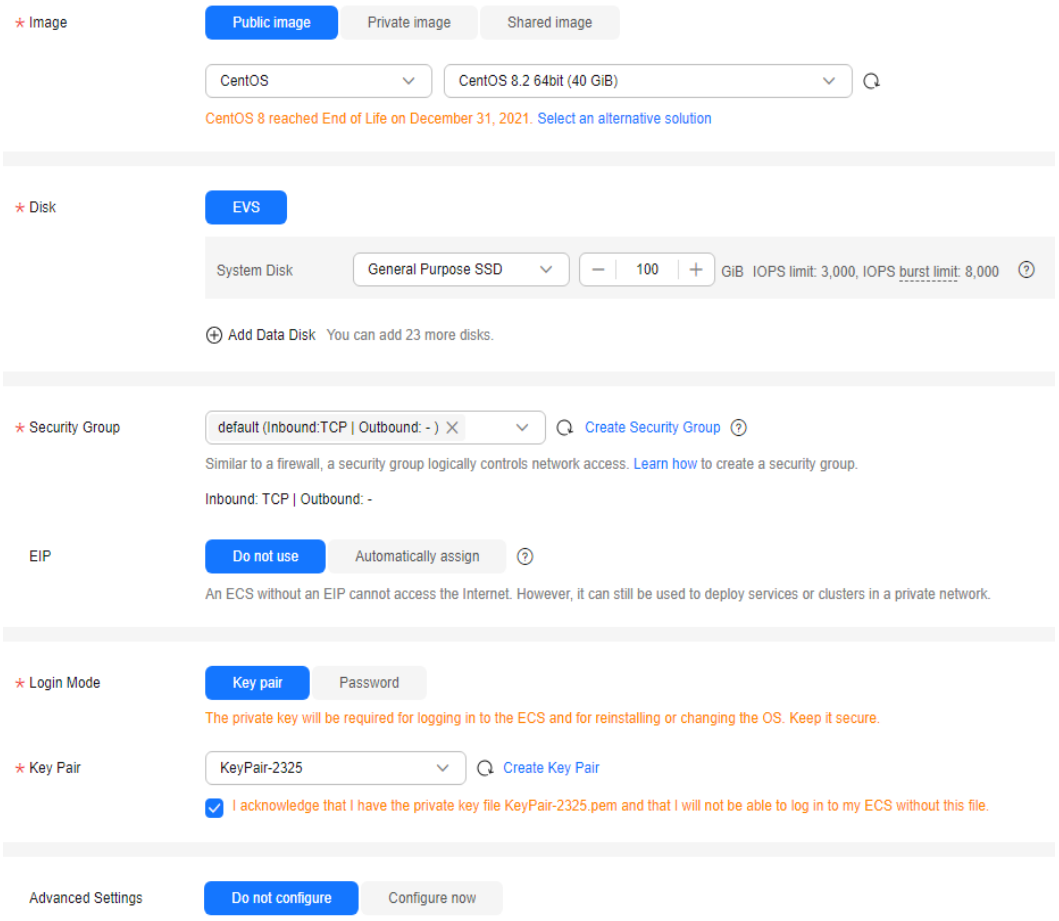


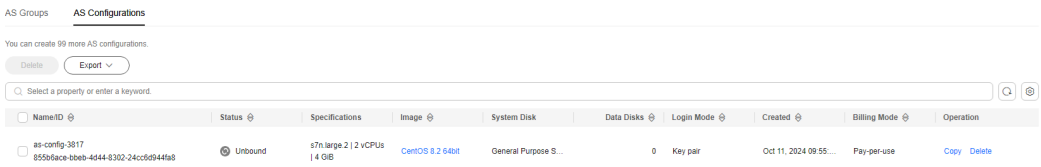
Table 1-1 Parameters for creating an AS configuration

Parameter	Example	Description
Billing Mode	**Pay-per-use**	Resources will be billed based on the usage duration. You can provision or delete resources at any time. For more information, see **Billing Overview**.
Region	**CN-Hong Kong**	For low network latency and fast resource access, select the region nearest to your target users. For more information, see **Region and AZ**.
Name	**as-config-3817**	Enter a name for the AS configuration.
Configuration Template	**Create new template**	Configure the specifications, image, disks, and other parameters for the instance.

Parameter	Example	Description
CPU Architecture	**x86**	x86 uses Complex Instruction Set Computing (CISC).
Specifications	**s7n.xlarge.2**	Select a flavor based on service requirements. For more information, see **A Summary List of x86 ECS Specifications**.
Image	**CentOS 8.2 64bit(40GiB)**	The example is a free public Linux image provided by Huawei Cloud.
Disk	**General Purpose SSD**, **100 GiB**	This is a system disk that will be created together with an instance. They are used to store the OS of each instance.
Security Group	**default**	Use the default security group.
EIP	**Do not use**	If the instances need to access the Internet, you can configure EIPs for them.
Login Mode	**Key pair**	Set **Login Mode** to **Key pair**.
Key Pair	**KeyPair-2325**	Use an existing or create a new key pair. Ensure that you have obtained the private key.
Advanced Settings	**Do not configure**	-

3. Click **Create Now**.
4. Click **Back to AS Configuration List** to view the new AS configuration.

Figure 1-2 Viewing an AS configuration



Step 2: Create an AS Group

In this step, create an AS group using the example settings. For more information, see **AS Group**.

1. Click **Create AS Group** and set parameters.

Figure 1-3 Creating an AS group

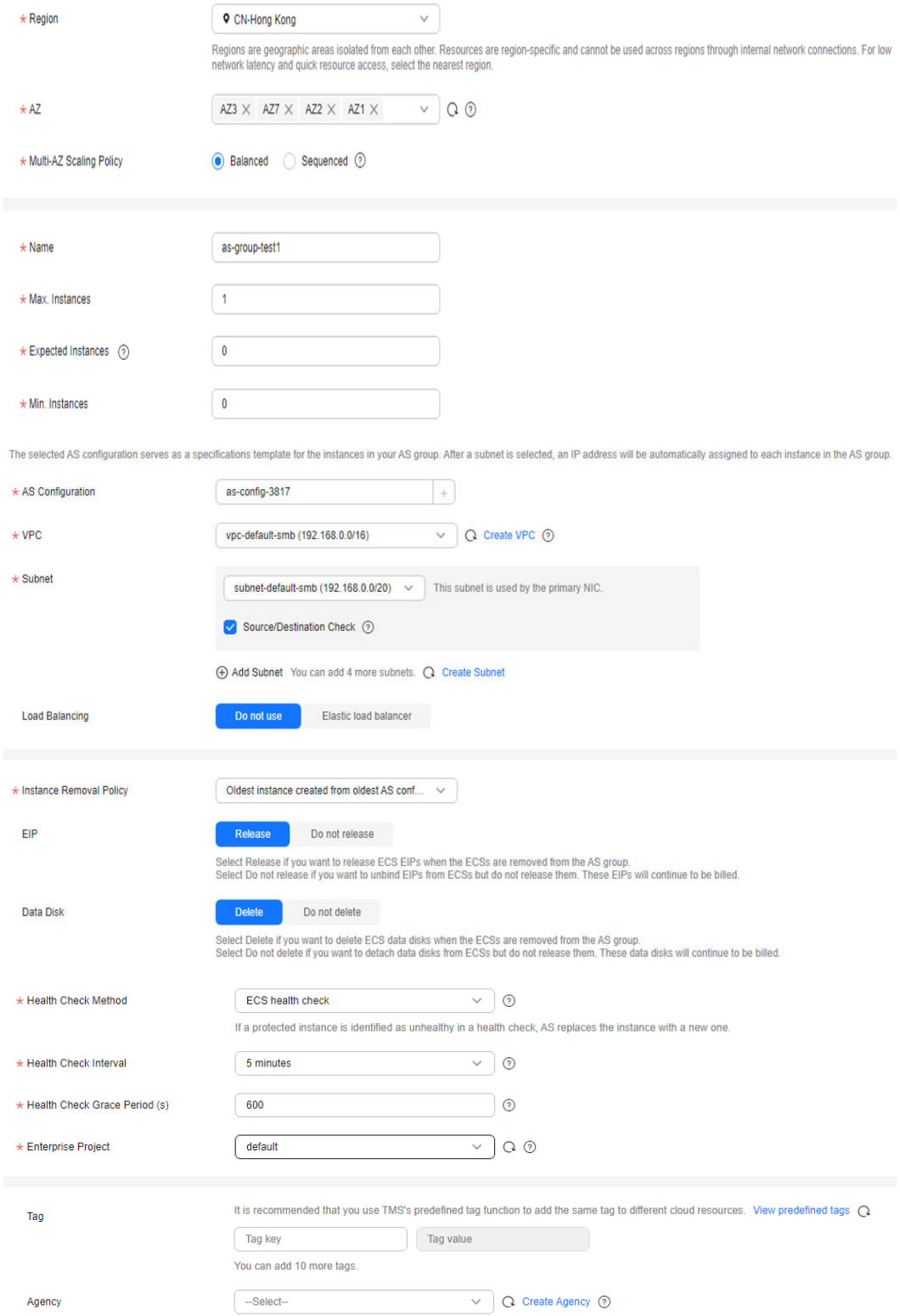


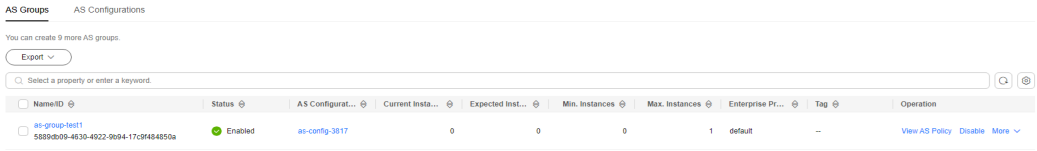
Table 1-2 Parameters for creating an AS group

Parameter	Example	Description
Region	**CN-Hong Kong**	For low network latency and fast resource access, select the region nearest to your target users. For more information, see **Region and AZ**.
AZ	**AZ1**, **AZ2**, **AZ3**, **AZ7**	AZs are physically isolated but interconnected over an intranet.
Multi-AZ Scaling Policy	**Balanced**	This policy ensures that the number of instances in each of the selected AZs is balanced.
Name	**as-group-test1**	Enter a name for the AS group.
Max. Instances	**1**	Specify the maximum number of instances in the AS group.
Expected Instances	**0**	Specify the desired number of instances in the AS group.
Min. Instances	**0**	Specify the minimum number of instances in the AS group.
AS Configuration	**as-config-3817**	Select the AS configuration created in **Step 1: Create an AS Configuration**.
VPC	**vpc-default-smb**	Use the default VPC and subnet. For more information, see **VPC Network Planning Suggestions**.
Subnet	**subnet-default-smb**	
Load Balancing	**Do not use**	This parameter is optional. For more information, see **Creating an AS Group**.
Instance Removal Policy	**Oldest instance created from oldest AS configuration**	With this policy, instances that use the oldest AS configuration are removed from the AS group first.
EIP	**Release**	With this option, when an instance is removed from an AS group, its EIP will be released.

Parameter	Example	Description
Data Disk	**Delete**	With this option, when an instance is removed from the AS group, all data disks attached to it will be deleted
Health Check Method	**ECS health check**	With this method, AS checks whether instances are running. If an instance fails the health check, AS removes it from the AS group.
Health Check Interval	**5 minutes**	Specify the interval between health checks.
Health Check Grace Period (s)	**600**	Specify how long AS must wait before checking the health status of an instance after the instance is enabled in an AS group.
Enterprise Project	**default**	Specify the enterprise project where the AS group is managed. Instances in this AS group are also managed under the same project.

2. Click **Create Now**.
3. Click **Back to AS Group List** to view the new AS group.

Figure 1-4 Viewing an AS group



Step 3: Add an AS Policy

1. In the AS group list, locate the AS group and click **View AS Policy** in the **Operation** column. Then, click **Add AS Policy**.
2. Configure the parameters for adding an AS policy.

Figure 1-5 Parameters for adding an AS policy

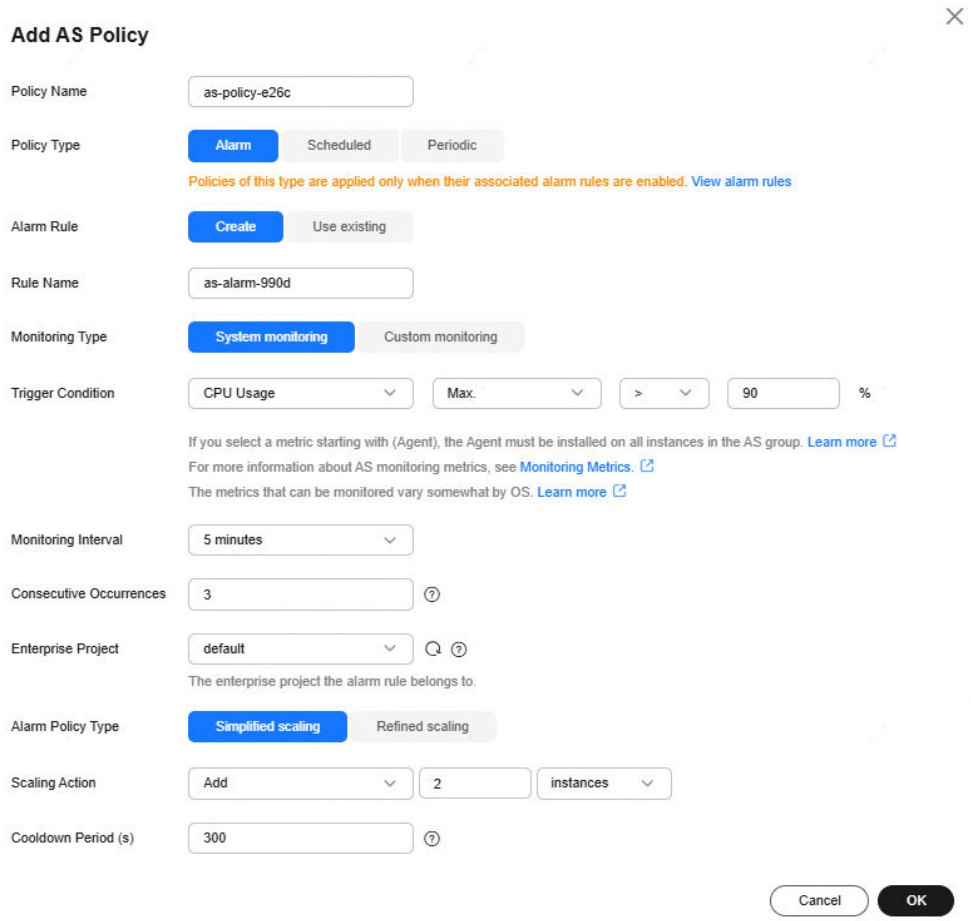


Table 1-3 Alarm-based AS policy parameters

Parameter	Example	Description
Policy Name	**as-policy-e26c**	Enter a name for the AS policy.
Policy Type	**Alarm**	Set **Policy Type** to **Alarm**.
Alarm Rule	**Create**	You can choose either to create a new alarm rule or use an existing one.
Rule Name	**as-alarm-43bd**	Enter a name for the alarm rule.
Monitoring Type	**System monitoring**	You can choose either **System monitoring** or **Custom monitoring**.
Trigger Condition	**CPU Usage Max. > 90%**	-
Monitoring Interval	**5 minutes**	-

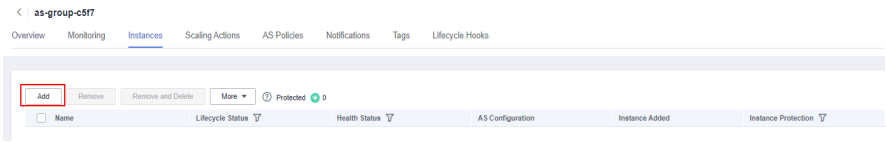
Parameter	Example	Description
Consecutive Occurrences	**3**	-
Enterprise Project	**default**	The default enterprise project is **default**.
Alarm Policy Type	**Simplified scaling**	-
Scaling Action	**Add 2 instances**	Specify an action and the number or percentage of instances. The value can be **Add**, **Reduce**, or **Set to**.
Cooldown Period (s)	**300**	To prevent an alarm-based policy from being repeatedly triggered by the same event, you can set a cooldown period.

3. Click **OK**.

(Optional) Step 4: Manually Add Instances

1. Click the name of the AS group. Click the **Instances** tab.
2. Select the instance to be added and click **Add**.

Figure 1-6 Manually adding instances



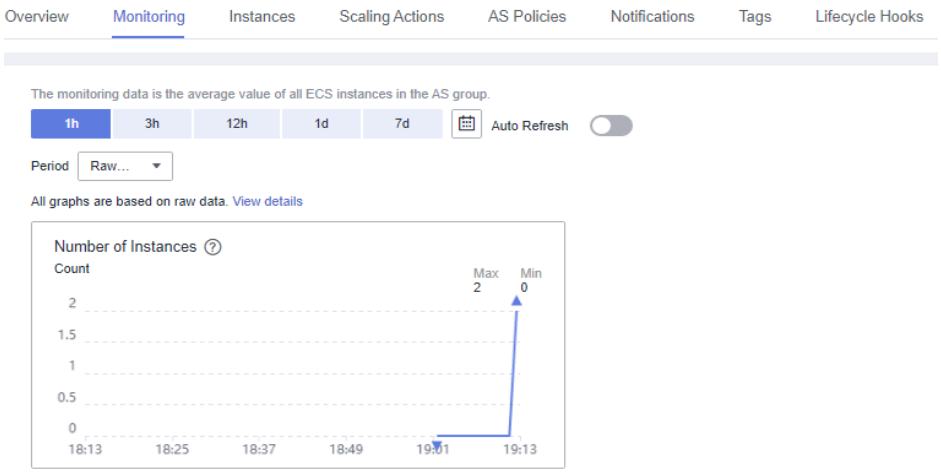
NOTE

Instances can be added to an AS group only when the group is enabled, no scaling action is in progress in it, and the number of instances in it is less than the maximum allowed.

(Optional) Step 5: View Monitoring Data

Click the name of the AS group and then click the **Monitoring** tab. You can view monitoring data changes of the AS group's scaling actions. AS can monitor CPU status in real time and perform scaling actions based on configured policies. This meets service requirements and reduces costs.

Figure 1-7 Viewing monitoring data



2 Scaling Your Website on a Schedule

Scenarios

When service requirements change regularly, you can configure AS policies in your application system to adjust resources on a scheduled basis. When the conditions of a policy are met, the system automatically changes the expected number of instances to trigger a scaling action to add or remove instances. This reduces the workload of manual resource adjustments during peak hours, helping you save resource and labor costs.

Assume that you are an e-commerce vendor that runs a website on an ECS. If you plan to launch a big sales promotion at 00:00 on a certain day but the ECS is unable to handle service requests during peak hours, two ECSs need to be automatically added to the application system at about 23:30 on the previous day, you can address this issue according the guidance of this section.

Procedure

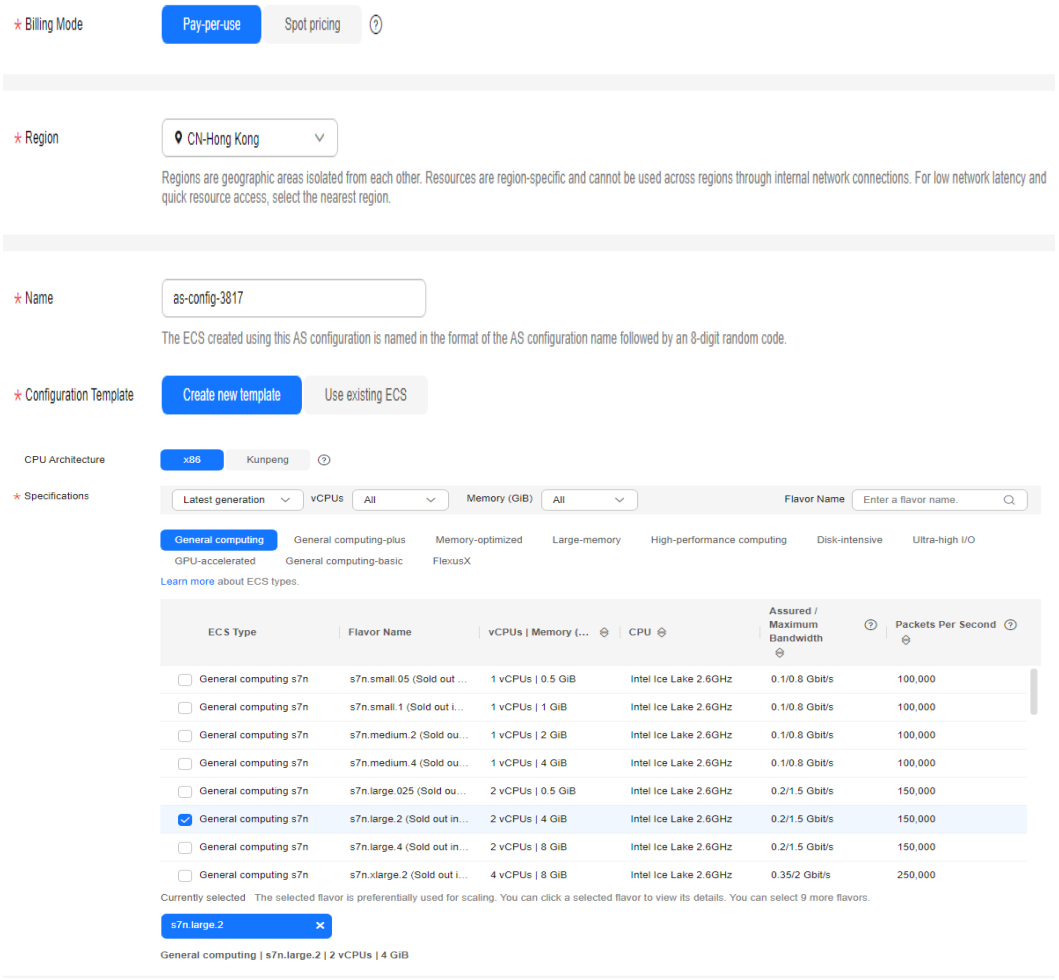
Step	Description
Step 1: Create an AS Configuration	Configure the specifications, image, and disks for new ECS instances.
Step 2: Create an AS Group	Configure the maximum, minimum, and expected number of instances and other parameters.
Step 3: Add an AS Policy	Create an AS policy to adjust service resources.
(Optional) Step 4: Manually Add Instances	Add instances to the AS group when the group is enabled, no scaling action is in progress in it, and the number of instances in it is less than the maximum allowed.
(Optional) Step 5: Viewing the Number of Instances	After a scaling action is triggered, check whether the number of instances in the AS group is the same as the expected number of instances.

Step 1: Create an AS Configuration

In this step, create an AS configuration using the example settings. For more information, see **AS Configuration**.

1. Log in to the **AS console**.
2. Click **Create AS Configuration** and set parameters.

Figure 2-1 Creating an AS configuration



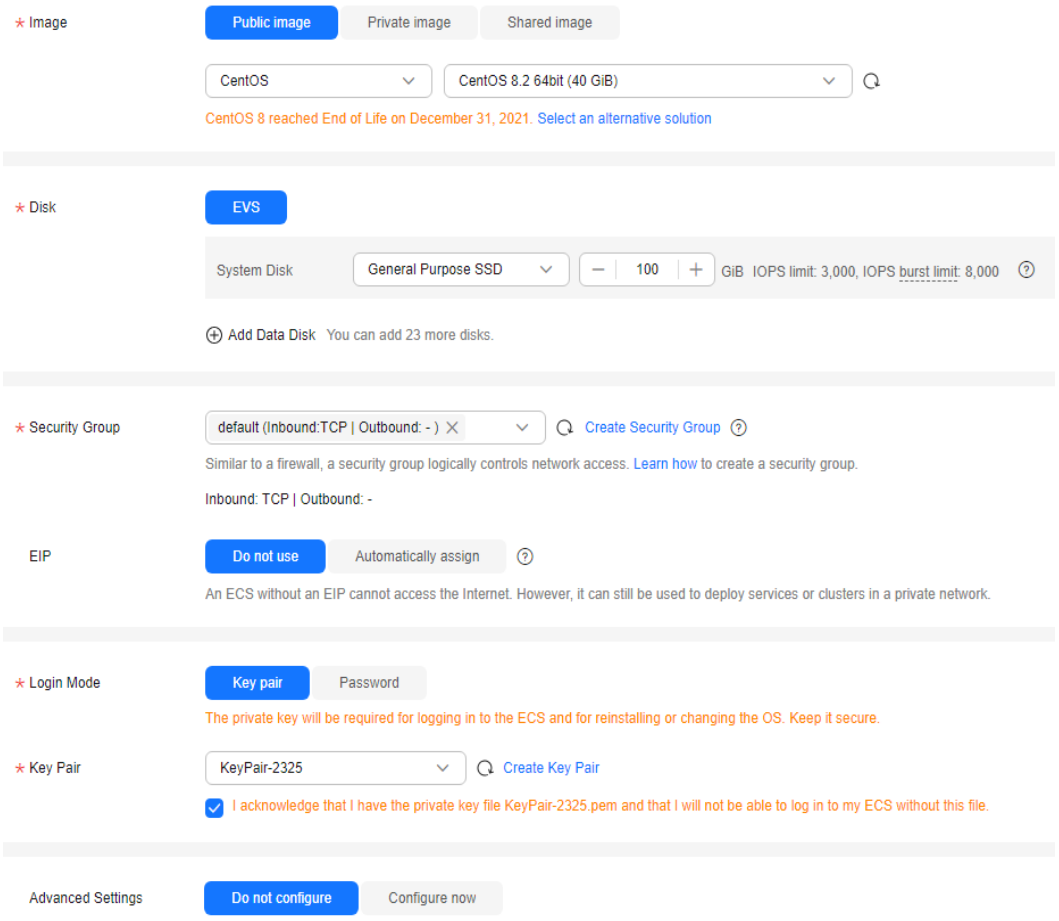


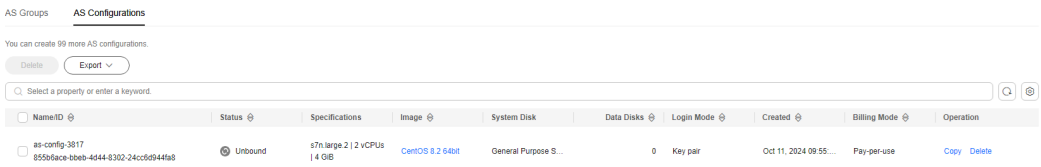
Table 2-1 Parameters for creating an AS configuration

Parameter	Example	Description
Billing Mode	**Pay-per-use**	Resources will be billed based on the usage duration. You can provision or delete resources at any time. For more information, see **Billing Overview**.
Region	**CN-Hong Kong**	For low network latency and fast resource access, select the region nearest to your target users. For more information, see **Region and AZ**.
Name	**as-config-3817**	Enter a name for the AS configuration.
Configuration Template	**Create new template**	Configure the specifications, image, disks, and other parameters for the instance.

Parameter	Example	Description
CPU Architecture	**x86**	x86 uses Complex Instruction Set Computing (CISC).
Specifications	**s7n.xlarge.2**	Select a flavor based on service requirements. For more information, see **A Summary List of x86 ECS Specifications**.
Image	**CentOS 8.2 64bit(40GiB)**	The example is a free public Linux image provided by Huawei Cloud.
Disk	**General Purpose SSD**, **100 GiB**	This is a system disk that will be created together with an instance. They are used to store the OS of each instance.
Security Group	**default**	Use the default security group.
EIP	**Do not use**	If the instances need to access the Internet, you can configure EIPs for them.
Login Mode	**Key pair**	Set **Login Mode** to **Key pair**.
Key Pair	**KeyPair-2325**	Use an existing or create a new key pair. Ensure that you have obtained the private key.
Advanced Settings	**Do not configure**	-

3. Click **Create Now**.
4. Click **Back to AS Configuration List** to view the new AS configuration.

Figure 2-2 Viewing an AS configuration



Step 2: Create an AS Group

In this step, create an AS group using the example settings. For more information, see **AS Group**.

1. Click **Create AS Group** and set parameters.

Figure 2-3 Creating an AS group

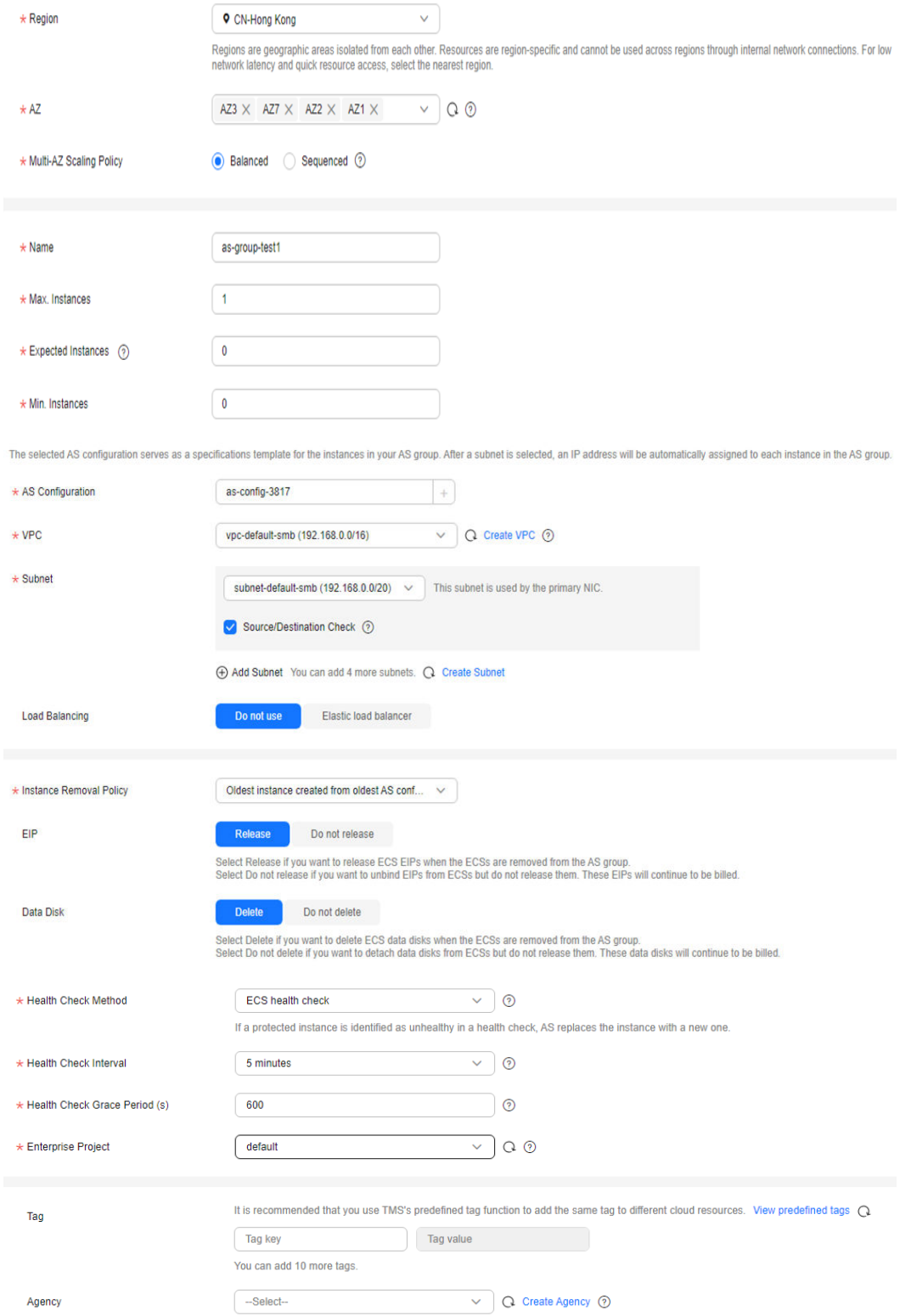


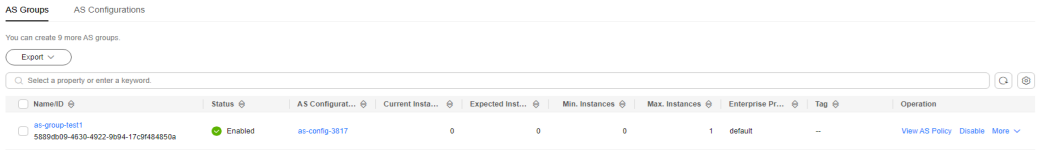
Table 2-2 Parameters for creating an AS group

Parameter	Example	Description
Region	**CN-Hong Kong**	For low network latency and fast resource access, select the region nearest to your target users. For more information, see **Region and AZ**.
AZ	**AZ1**, **AZ2**, **AZ3**, **AZ7**	AZs are physically isolated but interconnected over an intranet.
Multi-AZ Scaling Policy	**Balanced**	This policy ensures that the number of instances in each of the selected AZs is balanced.
Name	**as-group-test1**	Enter a name for the AS group.
Max. Instances	**1**	Specify the maximum number of instances in the AS group.
Expected Instances	**0**	Specify the desired number of instances in the AS group.
Min. Instances	**0**	Specify the minimum number of instances in the AS group.
AS Configuration	**as-config-3817**	Select the AS configuration created in **Step 1: Create an AS Configuration**.
VPC	**vpc-default-smb**	Use the default VPC and subnet. For more information, see **VPC Network Planning Suggestions**.
Subnet	**subnet-default-smb**	
Load Balancing	**Do not use**	This parameter is optional. For more information, see **Creating an AS Group**.
Instance Removal Policy	**Oldest instance created from oldest AS configuration**	With this policy, instances that use the oldest AS configuration are removed from the AS group first.
EIP	**Release**	With this option, when an instance is removed from an AS group, its EIP will be released.

Parameter	Example	Description
Data Disk	**Delete**	With this option, when an instance is removed from the AS group, all data disks attached to it will be deleted
Health Check Method	**ECS health check**	With this method, AS checks whether instances are running. If an instance fails the health check, AS removes it from the AS group.
Health Check Interval	**5 minutes**	Specify the interval between health checks.
Health Check Grace Period (s)	**600**	Specify how long AS must wait before checking the health status of an instance after the instance is enabled in an AS group.
Enterprise Project	**default**	Specify the enterprise project where the AS group is managed. Instances in this AS group are also managed under the same project.

2. Click **Create Now**.
3. Click **Back to AS Group List** to view the new AS group.

Figure 2-4 Viewing an AS group



Step 3: Add an AS Policy

1. In the AS group list, locate the AS group and click **View AS Policy** in the **Operation** column. Then, click **Add AS Policy**.
2. Configure the parameters for adding an AS policy.

Figure 2-5 Parameters for adding an AS policy

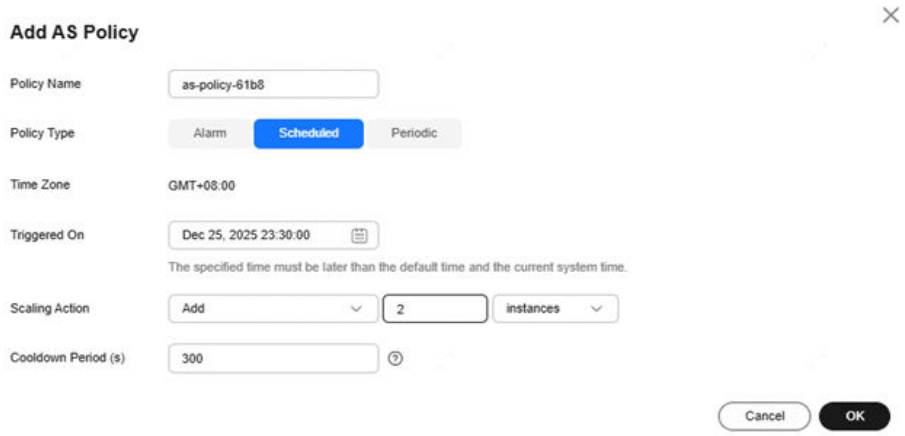


Table 2-3 Scheduled AS policy parameters

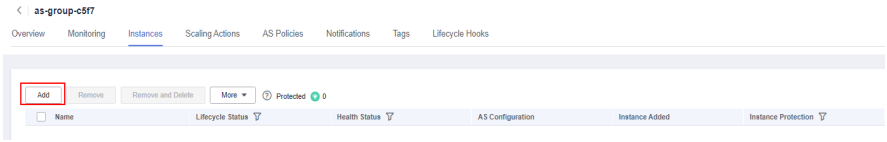
Parameter	Example	Description
Policy Name	**as-policy-61b8**	Enter a name for the AS policy.
Policy Type	**Scheduled**	Set **Policy Type** to **Scheduled policy**.
Time Zone	**GMT+08:00**	The default value is **GMT+08:00**. **GMT+08:00** indicates the time is 8 hours ahead of the Greenwich Mean Time (GMT).
Triggered On	**2025/12/03 23:30:00**	Specify the time when the AS policy will be triggered. The example is for reference only. You can change it as needed.
Scaling Action	**Add 2 instances**	Specify an action and the number or percentage of instances. If two instances need to be added, set the action to **Add two instances**. The value can be **Add**, **Reduce**, or **Set to**.
Cooldown Period (s)	**300**	A cooldown period (in seconds) is the period of time between two scaling actions.

3. Click **OK**.

(Optional) Step 4: Manually Add Instances

1. Click the name of the AS group. Click the **Instances** tab.
2. Select the instance to be added and click **Add**.

Figure 2-6 Manually adding instances



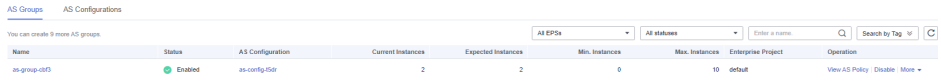
NOTE

Add instances to the AS group when the group is enabled, no scaling action is in progress in it, and the number of instances in it is less than the maximum allowed.

(Optional) Step 5: Viewing the Number of Instances

- Before the scaling action is triggered, the AS group has only two instances. Both the number of current instances and the expected number of instances are **2**.

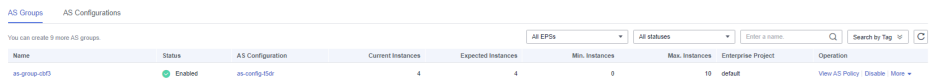
  Figure 2-7 Before the scaling action is triggered

- After the scaling action is triggered, the AS group automatically adds two instances based on the AS policy. The number of instances is **4**, and the expected number of instances is **4**.

  Two instances are added to the application system. In total, there will be four instances processing services, meeting service requirements during peak hours.

  Figure 2-8 After the scaling action is triggered

3 Scaling Your EIP Bandwidth Periodically

Scenarios

If your services are bandwidth-intensive and the required bandwidth changes periodically, you can configure AS policies in your application system to adjust the bandwidth periodically. When the conditions of a bandwidth scaling policy are met, the system automatically adjusts the bandwidth. This reduces manual adjustment workloads during peak hours and helps you save resource and labor costs.

For example, a live streaming website broadcasts the most watched movies at 20:00 for two hours every day. The EIP bandwidth is 5 Mbit/s, which cannot handle the traffic peak. You need to adjust the bandwidth to 10 Mbit/s at 20:00 and adjust it back to 5 Mbit/s at 22:00 every day. You can follow the steps in this section to adjust it.

Procedure

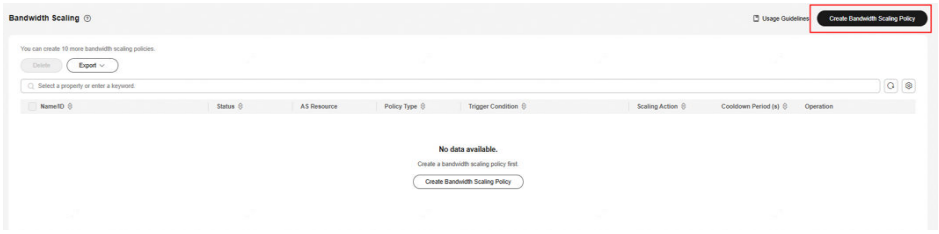
Step	Description
Step 1: Create Bandwidth Scaling Policies	Set parameters such as the period, triggering time, and effective time.
Step 2: Check Policy Execution Logs	After the policy execution time arrives, check the execution results.
Step 3: Check the EIP Bandwidth	Check the EIP bandwidth changes before and after a scaling policy is executed.
(Optional) Step 4: Execute a Bandwidth Scaling Policy Immediately	When a traffic burst occurs, you can execute the bandwidth scale-out policy adjust the bandwidth immediately.

Step 1: Create Bandwidth Scaling Policies

Only parameters involved in the example are described here and the values are for reference only. For details, see **Creating a Bandwidth Scaling Policy**.

1. Log in to the **AS console**.
2. In the navigation pane on the left, click **Bandwidth Scaling**.
3. Click **Create Bandwidth Scaling Policy**.

Figure 3-1 Creating a bandwidth scaling policy



4. Configure the bandwidth scaling policy.

Figure 3-2 Configuring a bandwidth scaling policy

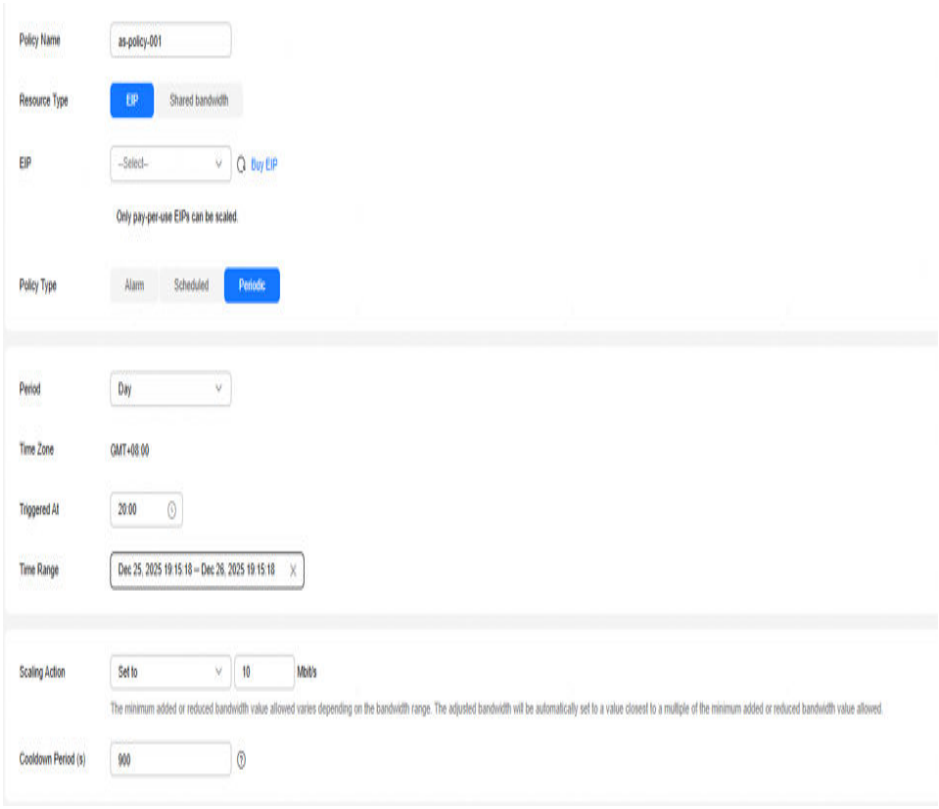


Table 3-1 Parameters of a bandwidth scaling policy

Parameter	Example	Description
Policy Name	**as-policy-001**	Enter a name for the bandwidth scaling policy.
Resource Type	**EIP**	Select the type of the bandwidth to be adjusted. You can select **EIP** or **Shared bandwidth**.

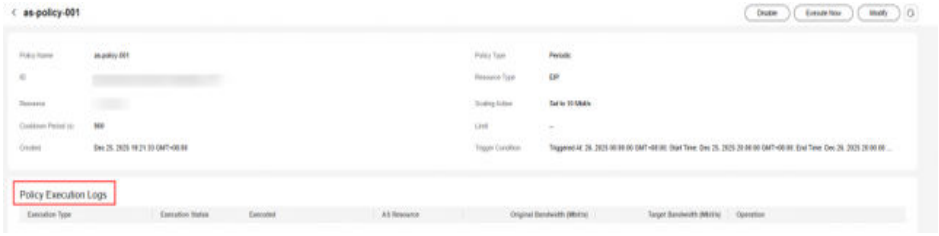
Parameter	Example	Description
EIP	-	Select the EIP whose bandwidth needs to be scaled. This parameter is available when **Resource Type** is set to **EIP**.
Policy Type	**Periodic**	The bandwidth will be adjusted periodically.
Period	**Day**	The value can be **day**, **week**, or **month**.
Time Zone	**GMT+08:00**	The default value is **GMT+08:00**. **GMT+08:00** indicates the time is 8 hours ahead of the Greenwich Mean Time (GMT).
Triggered At	**20:00**	Specify the time when the AS policy will be triggered. **NOTE** The value of **Triggered At** must fall in that of **Time Range**.
Time Range	**Custom**	Specify the time range during which the AS policy can be triggered.
Scaling Action	**Set to 10 Mbit/s**	Specify the action to be performed. The value can be **Add**, **Reduce**, or **Set to**.
Cooldown Period (s)	**900**	A cooldown period (in seconds) is the period of time between two scaling actions.

5. Click **Create Now**.
6. Repeat **Step 3** to **Step 5** to create another bandwidth scaling policy. Set **Policy Name** to **as-policy-002**, **Policy Type** to **Periodic**, **Period** to **Day**, **Triggered At** to **22:00**, and **Scaling Action** to **Set to 5 Mbit/s**

Step 2: Check Policy Execution Logs

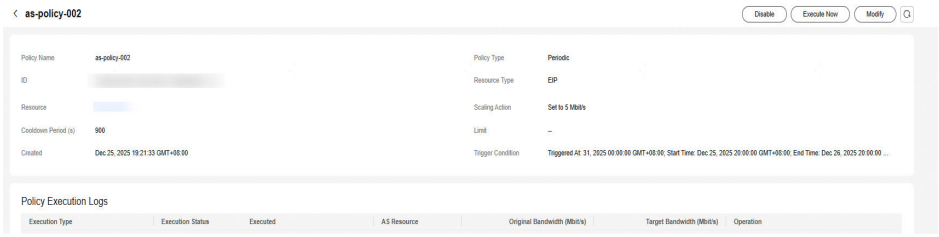
1. After the policy execution time arrives, click the bandwidth scaling policy name **as-policy-001** to go to the policy details page.
2. Checking policy execution logs in the **Policy Execution Logs** area. You can see the bandwidth specified in the policy **as-policy-001** is changed to 10 Mbit/s at 20:00 every day.

Figure 3-3 Execution logs of the bandwidth scaling policy as-policy-001



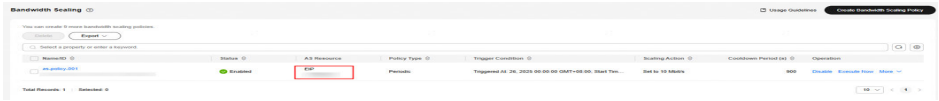
3. Repeat 1 and 2 for the bandwidth scaling policy **as-policy-002** to check the policy execution logs. You can see the bandwidth is adjusted to 5 Mbit/s at 22:00 every day.

Figure 3-4 Execution logs of the bandwidth scaling policy as-policy-002



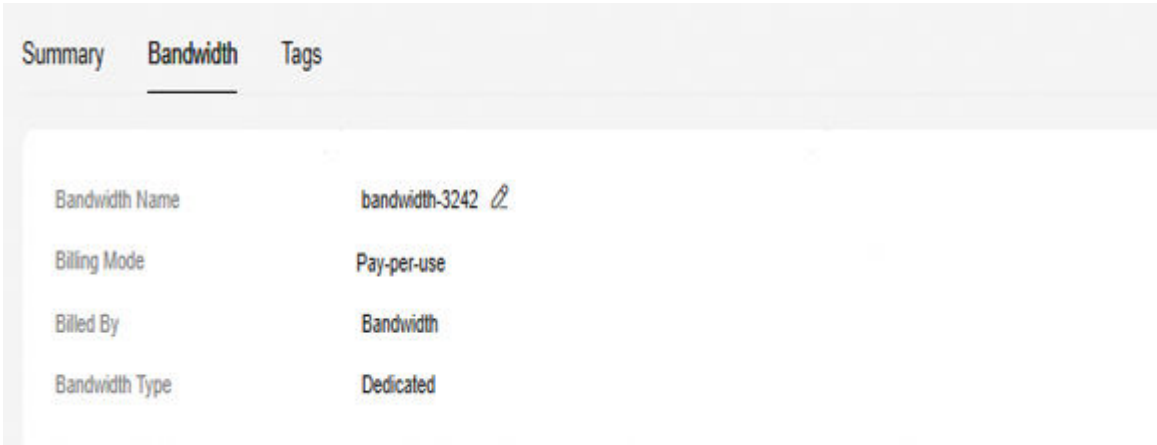
Step 3: Check the EIP Bandwidth

1. On the **Bandwidth Scaling** page, click the EIP to go to its details page.



2. On the **Bandwidth** tab, you can check the bandwidth. From 20:00 to 22:00, the bandwidth is 10 Mbit/s. At other times, the bandwidth is 5 Mbit/s.

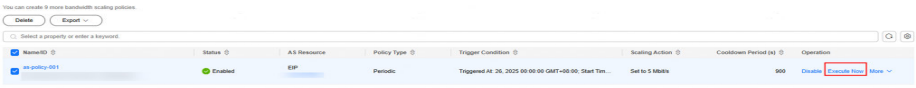
Figure 3-5 EIP details



(Optional) Step 4: Execute a Bandwidth Scaling Policy Immediately

1. On the **Bandwidth Scaling** page, locate the bandwidth scaling policy and click **Execute Now** in the **Operation** column.

Figure 3-6 Executing a bandwidth scaling policy immediately



2. Check the policy execution by referring to **Step 2: Check Policy Execution Logs**.

   NOTE

   When you detect a sudden service load peak, you can execute the scaling bandwidth policy as-policy-001 immediately to increase the bandwidth. After the peak hours, execute the bandwidth scaling policy as-policy-002 immediately to decrease the bandwidth back to 5 Mbit/s. This way, you can adjust a bandwidth anytime you want.

