Auto Scaling

Quick Start

Issue 01

Date 2024-10-12

HUAWEI

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are the property 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 Cloud 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.

Contents

1 Creating an AS Group Quickly

Scenarios

Auto Scaling (AS) automatically adjusts resources based on your service requirements and preset AS policies, helping you save resources and labor costs.

AS is available for free, but you pay for the instances (cloud servers) automatically added to the AS group and the resources used by the instances, such as EIPs, disks, and images.

This section walks you through the process of creating an AS configuration and an AS group, which are two critical steps for using AS.

Procedure

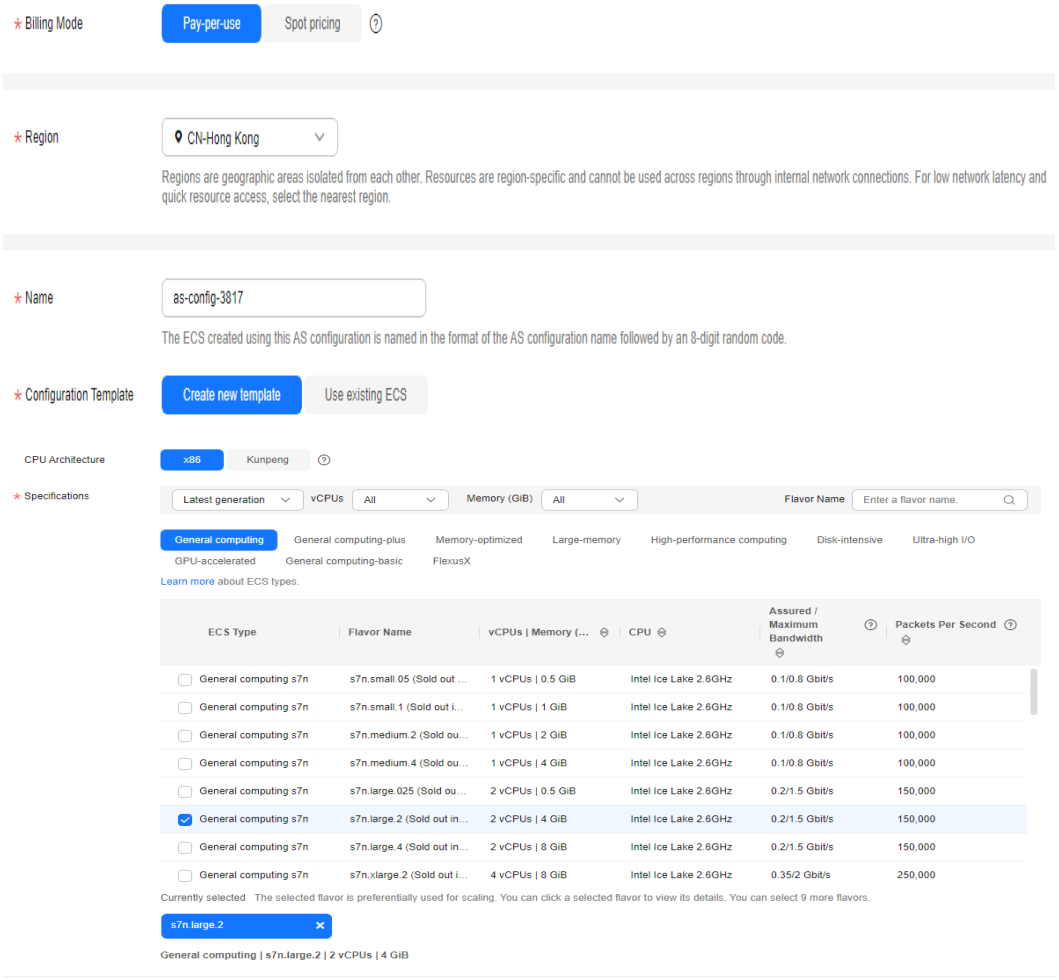
Step	Description
Step 1: Create an AS Configuration	Specify the specifications, image, and disk settings for the instances that AS creates for you.
Step 2: Create an AS Group	Configure scaling limits for the AS group by specifying the maximum, minimum, and desired group size.
Step 3: Create an AS Policy	Create an AS policy to adjust service resources.

Step 1: Create an AS Configuration

In this step, you create an AS configuration using the example settings. For details about how to create an AS configuration, see **Creating an AS Configuration**.

1. Log in to the console and go to the **Create AS Configuration** page.
2. Configure the AS configuration settings.

Figure 1-1 Page for creating an AS configuration



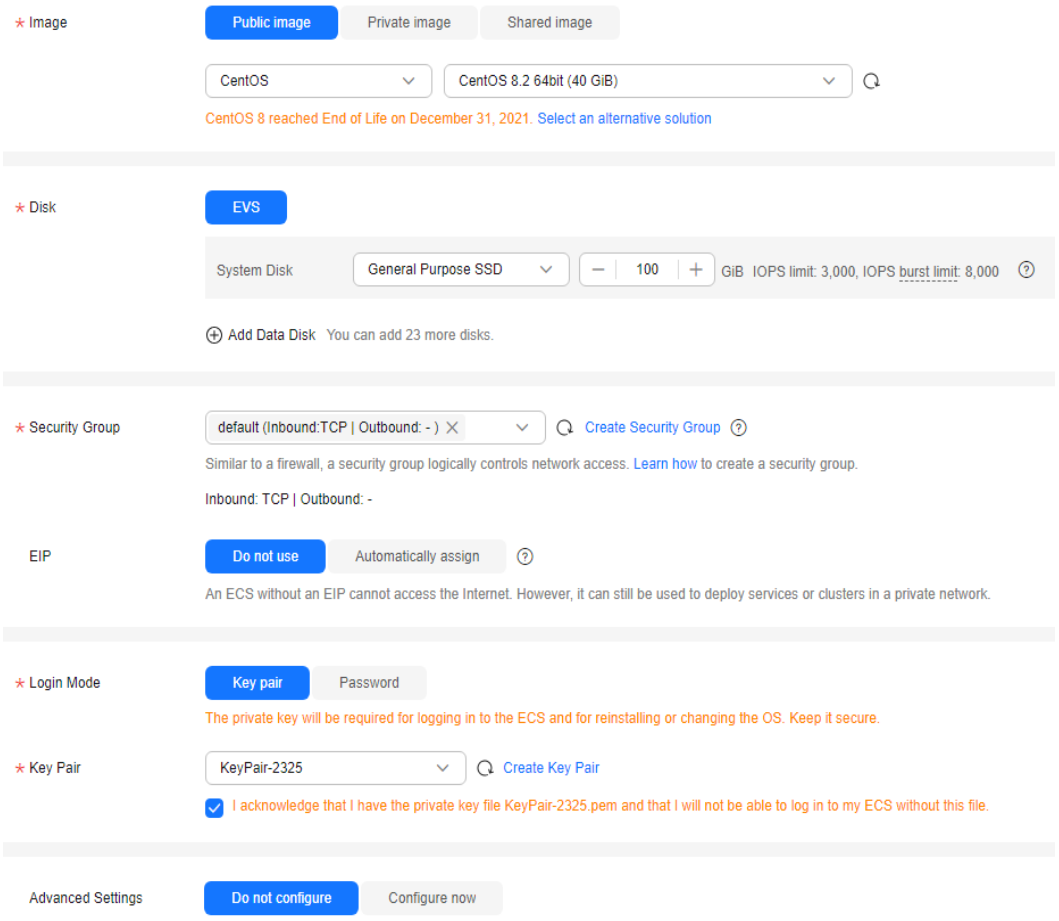


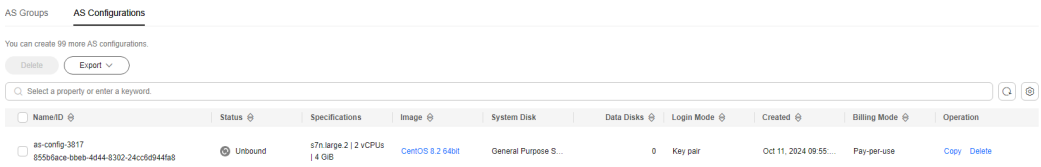
Table 1-1 Parameters for creating an AS configuration

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 details, see **ECS Billing Overview**.
Region	CN-Hong Kong	For low network latency and quick resource access, select the region nearest to your target users. For details, see **Region and AZ**.
Name	as-config-3817	Enter a name for the AS configuration.
Configuration Template	Create new template	Specify the specifications, image, and disk settings for the instances AS creates.

Parameter	Example	Description
CPU Architecture	x86	x86 uses Complex Instruction Set Computing (CISC).
Specifications	s7n.xlarge.2	Select a flavor one based on service requirements. For more information, see **A Summary List of x86 ECS Specifications**.
Image	CentOS 8.2 64bit (40 GiB)	The example is a free public Linux image provided by Huawei Cloud.
Disk	General Purpose SSD, 100 GiB	Specify the specifications of the system disk for instances AS creates.
Security Group	default	Use the default security group.
EIP	Do not use	If the instances in the AS group need to access the Internet, you can configure EIPs for the instances.
Login Mode	Key pair	A key pair for logging in to instances.
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 created AS configuration.

Figure 1-2 Viewing the AS configuration



Step 2: Create an AS Group

In this step, you create an AS group using the example settings. For details about how to create an AS group, see **AS Groups**.

1. Log in to the console and go to the **Create AS Group** page.
2. Configure the AS group settings.

Figure 1-3 Page for creating an AS group

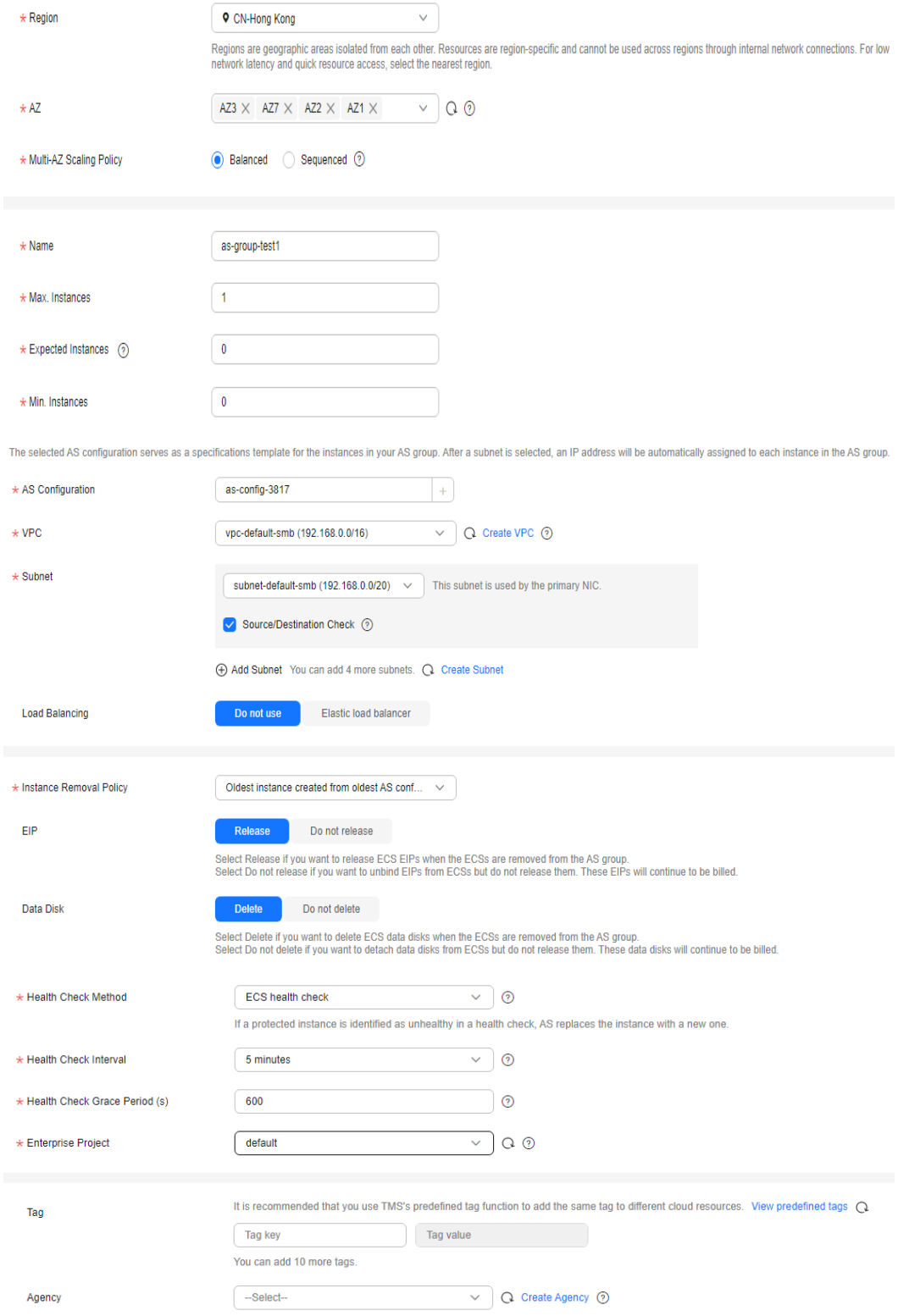


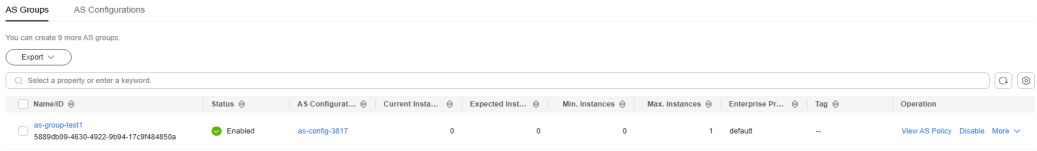
Table 1-2 Parameters for creating an AS group

Parameter	Example	Description
Region	CN-Hong Kong	For low network latency and quick resource access, select the region nearest to your target users. For details, see **Region and AZ**.
AZ	AZ1, AZ2, AZ3, AZ7	AZs are physically isolated but interconnected over a high-speed 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 group size.
Expected Instances	0	Specify the desired initial group size.
Min. Instances	0	Specify the minimum group size.
AS Configuration	as-config-3817	Select the AS configuration created in step 1.
VPC	vpc-default-smb	Use the default VPC and subnet. For details, see **VPC and Subnet Planning**.
Subnet	subnet-default-smb	
Load Balancing	Do not use	This parameter is optional. For details, see **Adding a Load Balancer to 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 the instance is removed from the AS group, all data disks attached to the instance 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.
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.

3. Click **Create Now**.
4. Click **Back to AS Group List** to view the created AS group.

Figure 1-4 Viewing the AS group



Step 3: Create an AS Policy

In this step, you create AS scaling policies to adjust service resources.

Operation Type	Refer To
Dynamic scaling	**Scaling Your Website Dynamically**
Scheduled scaling	**Scaling Your Website on a Schedule**

2 Creating an AS Group Quickly

If you are using AS for the first time, following the wizard-based process is an easy way to create an AS group, AS configuration, and AS policy.

Prerequisites

- You have created the required VPCs, subnets, security groups, and load balancers.
- You have obtained the key pair for logging in to the instances added in a scaling action if key authentication is used.

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 Group**.
4. Set basic information about the AS group, such as **Name**, **Max. Instances**, **Min. Instances**, and **Expected Instances**. Table 2-1 lists the parameters.

Table 2-1 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.	N/A

Parameter	Description	Example Value
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 among 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 expanding ECSs in an AS group, the system selects the target AZ based on the order in which AZs are selected. **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 (-).	-
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.	0
Min. Instances	Specifies the minimum number of ECS instances in an AS group.	0
VPC	Provides a network for your ECS instances. All ECS instances in the AS group are deployed in this VPC.	-

Parameter	Description	Example Value
Subnet	You can select up to five subnets. The AS group automatically binds all NICs to the created ECSs. 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 instance.	-
Load Balancing	This parameter is optional. A load balancer automatically distributes traffic across all instances in an AS group to balance their service load. It improves the fault tolerance of your applications and expands application service 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 instance becomes unhealthy, AS will replace the faulty instance 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 newest instance is removed from the AS group first. **NOTE** • Removing instances will preferentially ensure that the remaining instances are load balanced in AZs. • Manually added ECS instances are the last to be removed. If AS does remove a manually added instance, it only removes the instance from the AS group. It does not delete the instance. If multiple manually added instances must be removed, AS preferentially removes the earliest-added instance first.	-
EIP	If **EIP** has been selected in the AS configuration for an AS group, an EIP is automatically bound to the ECS instance added to the AS group. If you select **Release**, the EIP bound to an 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.	-

Parameter	Description	Example Value
Health Check Method	If an ECS instance fails a health check, AS replaces it with a new one. There are two health check methods: • **ECS health check**: checks ECS instance health status. If an instance is stopped or deleted, it is considered to be unhealthy. This method is selected by default. Using this method, the AS group periodically checks the running status of each instance. If an instance is unhealthy, AS removes the instance from the AS group. • **ELB health check**: determines ECS instance running status using a load balancing listener. This health check method is only available if a load balancer is configured for the AS group. An instance is considered to be healthy only when all associated listeners detect it as healthy. If a listener detects that the instance is unhealthy, AS removes the instance from the AS group.	-
Health Check Interval	Specifies the length of time between health checks. You can set a health check interval, such as 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.	-
Advanced Settings	Configure notifications. You can select **Do not configure** or **Configure now**.	-

Parameter	Description	Example Value
Notification	Results of scaling actions are sent to you based on the functions provided by the **Simple Message Notification (SMN)** service. • **Notification Conditions**: When at least one of the following conditions is met, SMN sends a notification to you: – Instance creation succeeds – Instance removal succeeds – Errors occur in an AS group – Instance creation fails – Instance removal fails • **Send Notification To**: Select an existing topic. For details about how to create a topic, see ***Simple Message Notification User Guide***.	-
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: =*<>\,|/	-

5. Click **Next**.
6. On the displayed page, you can use an existing AS configuration or create an AS configuration.

7. Click **Next**.
8. (Optional) Add an AS policy to an AS group.

   On the displayed page, click **Add AS Policy**.

   Configure the required parameters, such as the **Policy Type**, **Scaling Action**, and **Cooldown Period**.

   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.
9. Click **Create Now**.
10. Check the AS group, AS configuration, and AS policy information. Click **Submit**.
11. Confirm the creation result and go back to the **AS Groups** page as prompted.

    After the AS group is created, its status changes to **Enabled**.

3 Scaling Your Website Dynamically

4 Scaling Your Website on a Schedule

5 Wizard-based Process of Using AS

Figure 5-1 illustrates the wizard-based process of using AS.

Figure 5-1 Wizard-based process of using AS