

Auto Scaling

User Guide (Kuala Lumpur Region)

Issue 04
Date 2021-10-30



Copyright © Huawei Technologies Co., Ltd. 2022. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.

Trademarks and Permissions



HUAWEI 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.

Contents

1 Service Overview.....	1
1.1 What Is Auto Scaling?.....	1
1.2 AS Advantages.....	2
1.3 Instance Lifecycle.....	7
1.4 Constraints.....	11
1.5 Region and AZ.....	12
1.6 AS and Other Services.....	13
1.7 Permissions Management.....	15
1.8 Basic Concepts.....	18
2 Best Practices.....	19
2.1 Setting Up an Automatically Scalable Discuz! Forum.....	19
3 Quick Start.....	22
3.1 Wizard-based Process of Using AS.....	22
3.2 Creating an AS Group Quickly.....	22
4 AS Management.....	29
4.1 AS Group.....	29
4.1.1 Creating an AS Group.....	29
4.1.2 (Optional) Adding a Load Balancer to an AS Group.....	33
4.1.3 Changing the AS Configuration for an AS Group.....	34
4.1.4 Enabling an AS Group.....	34
4.1.5 Disabling an AS Group.....	35
4.1.6 Modifying an AS Group.....	36
4.1.7 Deleting an AS Group.....	36
4.2 AS Configuration.....	37
4.2.1 Creating an AS Configuration.....	37
4.2.2 Creating an AS Configuration from an Existing ECS Instance.....	37
4.2.3 Creating an AS Configuration from a New Specifications Template.....	40
4.2.4 Copying an AS Configuration.....	45
4.2.5 Deleting an AS Configuration.....	45
4.3 AS Policy.....	45
4.3.1 Overview.....	45
4.3.2 Creating an AS Policy.....	46

4.3.3 Managing AS Policies.....	55
4.4 Scaling Action.....	56
4.4.1 Dynamic Scaling.....	56
4.4.2 Scheduled Scaling.....	57
4.4.3 Manual Scaling.....	57
4.4.4 Configuring an Instance Removal Policy.....	58
4.4.5 Viewing a Scaling Action.....	59
4.4.6 Managing Lifecycle Hooks.....	59
4.4.7 Configuring Instance Protection.....	65
4.5 Bandwidth Scaling.....	66
4.5.1 Creating a Bandwidth Scaling Policy.....	66
4.5.2 Viewing Details About a Bandwidth Scaling Policy.....	71
4.5.3 Managing a Bandwidth Scaling Policy.....	72
4.6 AS Group and Instance Monitoring.....	74
4.6.1 Health Check.....	74
4.6.2 Configuring Notification for an AS Group.....	74
4.6.3 Recording AS Resource Operations.....	75
4.6.4 Adding Tags to AS Groups and Instances.....	79
4.6.5 Monitoring Metrics.....	80
4.6.6 Viewing Monitoring Metrics.....	85
4.6.7 Setting Monitoring Alarm Rules.....	86
4.7 Permissions Management.....	86
4.7.1 Creating a User and Granting AS Permissions.....	86
4.7.2 AS Custom Policies.....	87
5 FAQs.....	89
5.1 General.....	89
5.1.1 What Are Restrictions on Using AS?.....	89
5.1.2 Must I Use AS Together With ELB and Cloud Eye?.....	90
5.1.3 Is AS Billed?.....	90
5.1.4 Will an Abrupt Change in Monitoring Metric Values Trigger an Unnecessary Scaling Action?.....	90
5.1.5 How Many AS Policies and AS Configurations Can I Create and Use?.....	90
5.1.6 Can AS Automatically Scale Up or Down vCPUs, Memory, and Bandwidth of ECSs?.....	90
5.1.7 What Is an AS Quota?.....	90
5.1.8 How Do I Fix the Error "The key pair does not exist" When I Connect to an Instance?.....	91
5.2 AS Group.....	91
5.2.1 What Can I Do If the AS Group Fails to Be Enabled?.....	91
5.2.2 How Can I Handle an AS Group Exception?.....	94
5.2.3 What Operations Will Be Suspended If an AS Group Is Disabled?.....	97
5.2.4 Can I Use an ECS Instance ID to Learn What AS Group the Instance Is In?.....	97
5.3 AS Policy.....	97
5.3.1 How Many AS Policies Can I Enable?.....	97
5.3.2 What Are the Conditions to Trigger an Alarm-based AS Policy?.....	97

5.3.3 What Is a Cooldown Period and Why Is It Required?.....	98
5.3.4 Can AS Scale Capacity Based on Custom Monitoring by Cloud Eye?.....	98
5.3.5 What Monitoring Metrics for an AS Group Will Be Affected If VM Tools Are Not Installed on the Instances in the Group?.....	98
5.3.6 What Can I Do If an AS Policy Fails to Be Enabled?.....	98
5.3.7 How Do I Install the Agent on Instances in an AS Group to Use Agent Monitoring Metrics?.....	99
5.4 Instance.....	100
5.4.1 How Do I Prevent Instances Manually Added to an AS Group from Being Automatically Removed?	100
5.4.2 When There Are Multiple Flavors Included in the AS Configuration of an AS Group, What Order Are the Flavors Selected in During Scale-out Events?.....	101
5.4.3 When an Instance Is Removed from an AS Group and Deleted, Is the Application Data Saved?.....	102
5.4.4 Can I Add Yearly/Monthly ECS Instances to an AS Group?.....	102
5.4.5 Can the System Automatically Delete the Instances Added Based on an AS Policy When They Are Not Required?.....	102
5.4.6 What Is the Expected Number of Instances?.....	103
5.4.7 How Do I Delete an ECS Instance Created in a Scaling Action?.....	103
5.4.8 Will a Yearly/Monthly ECS Instance Be Deleted If It Becomes Unhealthy?.....	104
5.4.9 How Do I Handle Unhealthy Instances in an AS Group?.....	104
5.4.10 Why Instances in an AS Group Keep Failing Health Checks and Getting Deleted and Recreated?	105
5.4.11 How Do I Prevent Instances from Being Automatically Removed from an AS Group?.....	105
5.4.12 Why Is an Instance that Was Removed from an AS Group and Deleted Still Displayed in the ECS List?.....	105
5.5 Others.....	105
5.5.1 How Can I Automatically Deploy My Application on an Instance?.....	105
5.5.2 How Does Cloud-Init Affect the AS Service?.....	106
5.5.3 Why Can't I Use a Key File to Log In to an ECS?.....	106
5.5.4 Do I Need to Configure an EIP in an AS Configuration When a Load Balancer Has Been Enabled for an AS Group?.....	107
5.5.5 How Do I Enable Automatic Initialization of EVS Disks on Instances that Have Been Added to an AS Group During Scaling Actions?.....	107
A Change History.....	111

1 Service Overview

1.1 What Is Auto Scaling?

AS Introduction

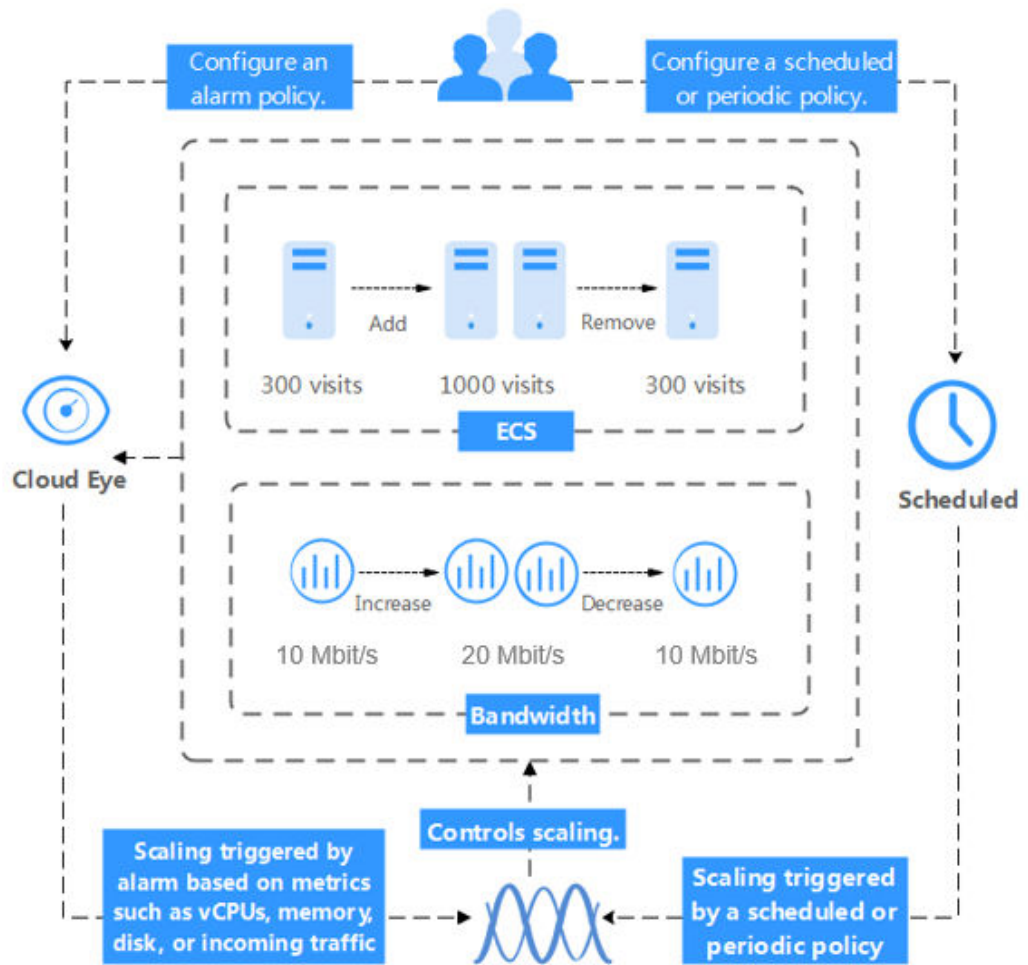
Auto Scaling (AS) helps you automatically scale Elastic Cloud Server (ECS) and bandwidth resources to keep up with changes in demand based on pre-configured AS policies. It allows you to add ECS instances or increase bandwidths to handle increases in load and also save money by removing resources that are sitting idle.

Architecture

AS allows you to scale ECS instances and bandwidths.

- **Scaling control:** You can configure AS policies, configure metric thresholds, and schedule when different scaling actions are taken. AS will trigger scaling actions on a repeating schedule, at a specific time, or when the configured thresholds are reached.
- **Policy configuration:** You can configure alarm-based, scheduled, and periodic policies as needed.
- **Alarm-based policies:** You can configure scaling actions to be taken when alarm metrics such as vCPU, memory, disk, and inbound traffic reaches the thresholds.
- **Scheduled policies:** You can schedule scaling actions to be taken at a specific time.
- **Periodic policies:** You can configure scaling actions to be taken at scheduled intervals, a specific time, or within a particular time range.
- When Cloud Eye generates an alarm for a monitoring metric, for example, CPU usage, AS automatically increases or decreases the number of instances in the AS group or the bandwidths.
- When the configured triggering time arrives, a scaling action is triggered to increase or decrease the number of ECS instances or the bandwidths.

Figure 1-1 AS architecture



Accessing AS

The public cloud provides a web-based service management platform. You can access AS using HTTPS-compliant application programming interfaces (APIs) or the management console.

- Calling APIs
Use this method if you are required to integrate AS on the public cloud into a third-party system for secondary development. For details, see *Auto Scaling API Reference*.
- Management console
Use this method if you do not need to integrate AS with a third-party system. After registering on the public cloud, log in to the management console and select **Auto Scaling** from the service list on the homepage.

1.2 AS Advantages

AS automatically scales resources to keep up with service demands based on pre-configured AS policies. With automatic resource scaling, you can enjoy reduced

costs, improved availability, and high fault tolerance. AS is used for following scenarios:

- Heavy-traffic forums: The traffic on a popular forum is difficult to predict. AS dynamically adjusts the number of ECS instances based on monitored ECS metrics, such as vCPU and memory usage.
- E-commerce: During big promotions, E-commerce websites need more resources. AS automatically increases ECS instances and bandwidths within minutes to ensure that promotions go smoothly.
- Live streaming: A live streaming website may broadcast popular programs from 14:00 to 16:00 every day. AS automatically scales out ECS and bandwidth resources during this period to ensure a smooth viewer experience.

Automatic Resource Scaling

AS adds ECS instances and increases bandwidths for your applications when the access volume increases and removes unneeded resources when the access volume drops, ensuring system stability and availability.

- Scaling ECS Instances on Demand
AS scales ECS instances for applications based on demand, improving cost management. ECS instances can be scaled dynamically, on a schedule, or manually:
 - Dynamic scaling
Dynamic scaling allows scale resources in response to changing demand using alarm-based policies. For details, see [Dynamic Scaling](#).
 - Scheduled scaling
Scheduled scaling helps you to set up your own scaling schedule according to predictable load changes by creating periodic or scheduled policies. For details, see [Scheduled Scaling](#).
 - Manual scaling
You can either manually change the expected number of instances of your AS group, or add or remove instances to or from the AS group. For details, see [Manual Scaling](#).

Consider a train ticket booking application running on the public cloud. The load of the application may be relatively low during Q2 and Q3 because there are not many travelers, but relatively high during Q1 and Q4. Traditionally, there are two ways to plan for these changes in load. The first option is to provision enough servers so that the application always has enough capacity to meet demand, as shown in [Figure 1-2](#). The second option is to provision servers according to the average load of the application, as shown in [Figure 1-3](#). However, these two options may waste resources or be unable to meet demand during peak seasons. By enabling AS for this application, you have a third option available. AS helps you scale servers to keep up with changes in demand. This allows the application to maintain steady, predictable performance without wasting money on any unnecessary resources, as shown in [Figure 1-4](#).

Figure 1-2 Over-provisioned capacity

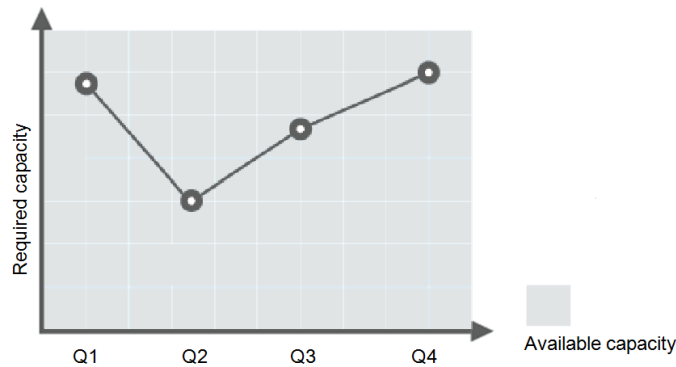


Figure 1-3 Insufficient capacity

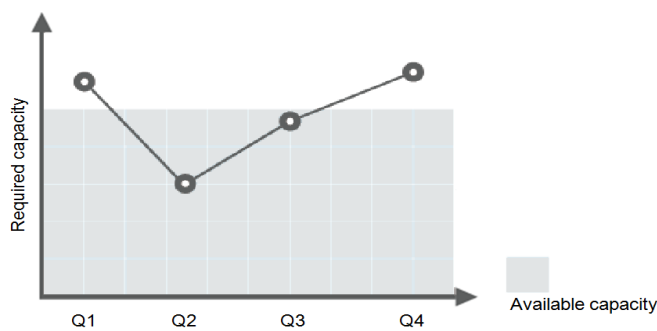
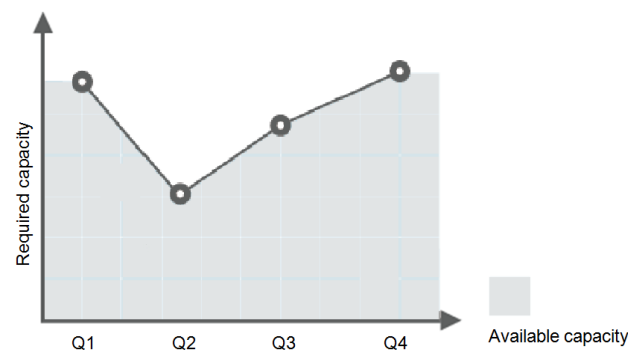


Figure 1-4 Auto-scaled capacity



- **Scaling Bandwidth on Demand**

AS adjusts bandwidth for an application based on demand, reducing bandwidth costs.

There are three types of scaling policies you can use to adjust the IP bandwidth on demand:

- Alarm-based policies

You can configure triggers based on metrics such as outbound traffic and bandwidth. When the system detects that the triggering conditions are met, the system automatically adjusts the bandwidth.

- Scheduled policies

The system automatically increases, decreases, or adjusts the bandwidth to a fixed value on a fixed schedule.

- Periodic policies

The system periodically adjusts the bandwidth based on a configured periodic cycle.

For example, you can use an alarm-based policy to regulate the bandwidth for a live streaming website.

For a live streaming website, service load is difficult to predict. In this example, the bandwidth needs to be dynamically adjusted between 10 Mbit/s and 30 Mbit/s based on metrics such as outbound traffic and inbound traffic. AS can automatically adjust the bandwidth to meet requirements. You just need to select the relevant EIP and create two alarm policies. One policy is to increase the bandwidth by 2 Mbit/s when the outbound traffic is greater than X bytes, with the limit set to 30 Mbit/s. The other policy is to decrease the bandwidth by 2 Mbit/s when the outbound traffic is less than X bytes, with the limit set to 10 Mbit/s.

- Evenly Distributing Instances by AZ

To reduce the impact of power or network outage on system stability, AS attempts to distribute ECS instances evenly across the AZs that are used by an AS group.

A region is a geographic area where resources used by ECS instances are located. Each region contains multiple Availability Zones (AZs) where resources use independent power supplies and networks. AZs are physically isolated from one another but interconnected through an intranet. AZs are engineered to be isolated from failures in other AZs. They provide cost-effective, low-latency network connections to other AZs in the same region.

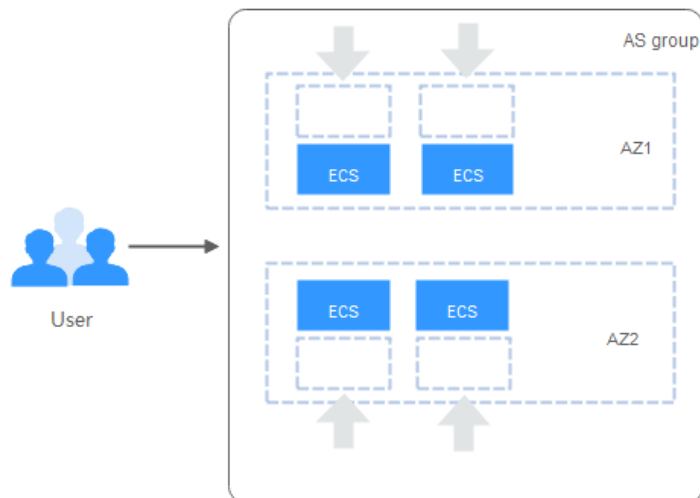
An AS group can contain ECS instances in one or more AZs within a region. During scale the capacity of an AS group, AS attempts to evenly distribute ECS instances across AZs used by the AS group based on the following rules:

Evenly distributing new instances to balanced AZs

AS attempts to evenly distribute ECS instances across the AZs used by an AS group. To do it, AS adds new instances to the AZ with the fewest instances.

Consider an AS group containing four instances that are evenly distributed in the two AZs used by the AS group. If a scaling action is triggered to add four more instances to the AS group, AS adds two to each AZ.

Figure 1-5 Evenly distributing instances

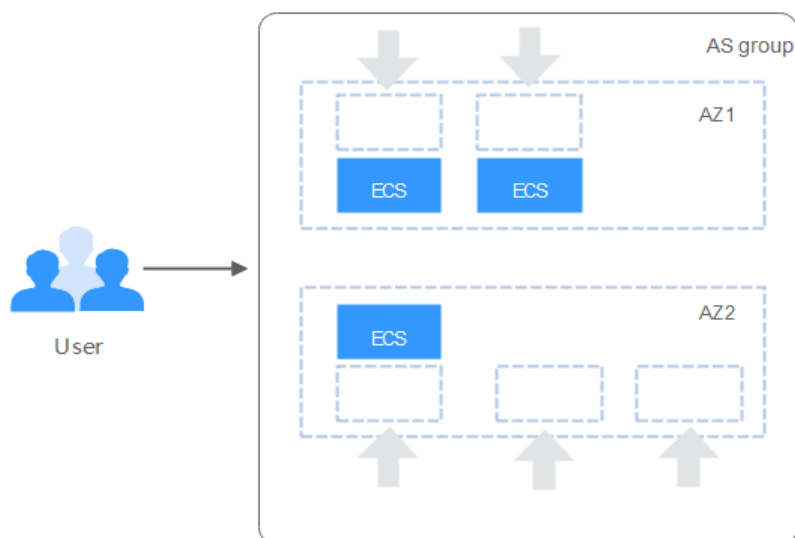


Re-balancing instances across AZs

After you have manually added or removed instances to or from an AS group, the AS group can become unbalanced between AZs. AS compensates by re-balancing the AZs during the next scaling action.

Consider an AS group containing three instances that are distributed in AZ 1 and AZ 2, with two in AZ 1 and one in AZ 2. If a scaling action is triggered to add five more instances to the AS group, AS adds two to AZ 1 and three to AZ 2.

Figure 1-6 Re-balancing instances



Enhanced Cost Management

AS enables you to use ECS instances and bandwidths on demand by automatically scaling resources for your applications, eliminating waste of resources and reducing costs.

Higher Availability

AS ensures that you always have the right amount of resources available to handle the fluctuating load of your applications.

Using ELB with AS

Working with ELB, AS automatically scales ECS instances based on changes in demand while ensuring that the load of all the instances in an AS group stays balanced.

After ELB is enabled for an AS group, AS automatically associates a load balancing listener with any instances added to the AS group. Then, ELB automatically distributes traffic to all healthy instances in the AS group through the listener, which improves system availability. If the instances in the AS group are running a range of different types of applications, you can bind multiple load balancing listeners to the AS group to listen to each of these applications, improving service scalability.

High Fault Tolerance

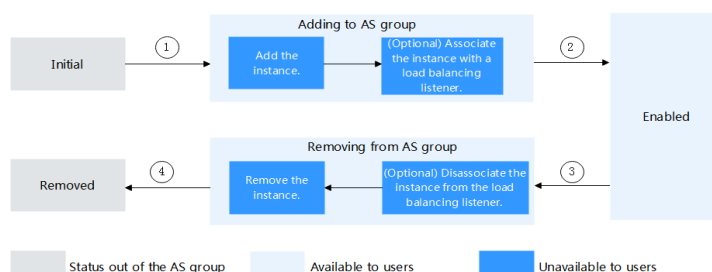
AS monitors instances in an AS group, and replaces any unhealthy instances it detects with new ones.

1.3 Instance Lifecycle

The ECS instances in an AS group transitions through different statuses from the moment it is created to its removal.

The instance status changes as shown in [Figure 1-7](#) if you have not added a lifecycle hook to the AS group.

Figure 1-7 Instance lifecycle



When trigger condition 2 or 4 is met, the system autonomously puts instances into the next status.

Table 1-1 Instance statuses

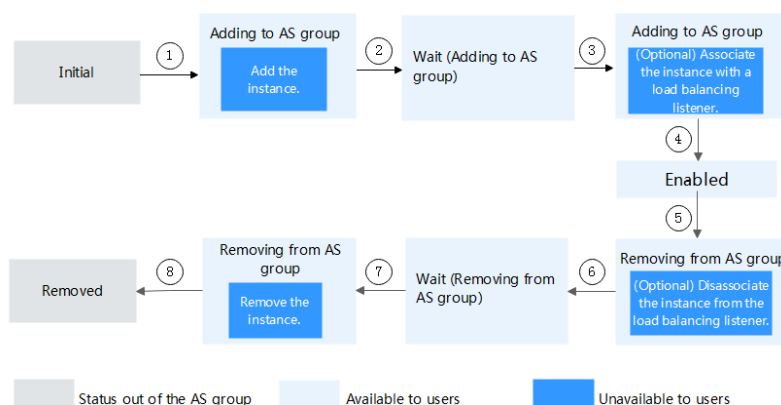
Status	Sub-status	Status Description	Trigger Condition
Initial	None	The instance has not been added to the AS group.	The instance status will be changed to Adding to AS group when either of the following occurs:
Adding to AS group	Add the instance.	When trigger condition 1 is met, AS adds the instance to expand the AS group capacity.	<ul style="list-style-type: none"> You manually increase the expected number of instances of the AS group. The system automatically expands the AS group capacity. You manually add instances to the AS group.
	(Optional) Associate the instance with a load balancing listener.	When trigger condition 1 is met, AS associates the created instance with the load balancing listener.	
Enabled	None	The instance is added to the AS group and starts to process service traffic.	The instance status is changed from Enabled to Removing from AS group when any of the following occurs:
Removing from AS group	(Optional) Disassociate the instance from the load balancing listener.	When trigger condition 3 is met, the AS group starts to reduce resources and disassociate the instance from the load balancing listener.	<ul style="list-style-type: none"> You manually decrease the expected number of instances of the AS group. The system automatically removes instances in a scaling action. A health check shows that an enabled instance is unhealthy, and the system removes it from the AS group. You manually remove instances from the AS group.
	Remove the instance.	After the instances are unbound from the load balancing listener, they are removed from the AS group.	
Removed	None	The instance lifecycle in the AS group ends.	None

When an ECS instance is added to an AS group manually or through a scaling action, it goes through the **Adding to AS group**, **Enabled**, **Removing from AS group** statuses, and is finally removed from the AS group.

If you have added a lifecycle hook to the AS group, the instance statuses change as shown in [Figure 1-8](#). When a scale-out or scale-in event occurs in the AS group, the required instances are suspended by the lifecycle hook and remain in the wait

status until the timeout period ends or you manually call back the instances. You can perform custom operations on the instances when they are in the wait status. For example, you can install or configure software on an instance before it is added to the AS group or download log files from an instance before it is removed.

Figure 1-8 Instance lifecycle



Under trigger conditions 2, 4, 6, or 8, the system automatically changes the instance status.

Table 1-2 Instance statuses

Status	Sub-status	Status Description	Trigger Description
Initial	None	The instance has not been added to the AS group.	The instance status is changed to Adding to AS group when either of the following occurs: <ul style="list-style-type: none"> You manually increase the expected number of instances of an AS group. The system automatically adds instances to the AS group in a scaling action. You manually add instances to the AS group.
Adding to AS group	Add an instance.	When trigger condition 1 is met, AS adds the instance to expand the AS group capacity.	

Status	Sub-status	Status Description	Trigger Description
Wait (Adding to AS group)	None	The lifecycle hook suspends the instance that is being added to the AS group and puts the instance into waiting state.	The instance status is changed from Wait (Adding to AS group) to Adding to AS group when either of the following operations is performed:
Adding to AS group	(Optional) Associate the instance with a load balancing listener.	When trigger condition 3 is met, AS associates the instance with the load balancing listener.	<ul style="list-style-type: none"> The default callback action is performed. You manually perform the callback action.
Enabled	None	The instance is added to the AS group and starts to process service traffic.	The instance status is changed from Enabled to Removing from AS group when any of the following occurs:
Removing from AS group	(Optional) Disassociate the instance from the load balancing listener.	When trigger condition 5 is met, the AS group starts to reduce resources and disassociate the instance from the load balancing listener.	<ul style="list-style-type: none"> You manually decrease the expected number of instances of an AS group. The system automatically removes instances in a scaling action. A health check shows that the instance is unhealthy after being enabled, and the system removes it from the AS group. You manually remove an instance from an AS group.
Wait (Removing from AS group)	None	The lifecycle hook suspends the instance that is being removed from the AS group and sets the instance to be in waiting state.	The instance status is changed from Wait (Removing from AS group) to Removing from AS group when either of the following occurs:
Removing from AS group	Remove the instance.	When trigger condition 7 is met, AS removes the instance from the AS group.	<ul style="list-style-type: none"> The default callback action is performed. You manually perform the callback action.
Removed	None	The instance lifecycle in the AS group ends.	None

1.4 Constraints

AS has the following constraints:

- Only applications that are stateless and can be horizontally scaled can run on instances in an AS group.

NOTE

- A stateless process or application can be understood in isolation. There is no stored knowledge of or reference to past transactions. Each transaction is made as if from scratch for the first time.
ECS instances where stateless applications are running do not store data that needs to be persisted locally.
Think of stateless transactions as a vending machine: a single request and a response.
- Stateful applications and processes, however, are those that can be returned to again and again. They are performed with the context of previous transactions and the current transaction may be affected by what happened during previous transactions.
ECS instances where stateful applications are running store data that needs to be persisted locally.
Stateful transactions are performed repeatedly, such as online banking or e-mail, which are performed with the context of previous transactions.
- AS can release ECS instances in an AS group automatically, so the instances cannot be used to save application status information (such as session statuses) or related data (such as database data and logs). If the application status or related data must be saved, you can store the information on separate servers.
- AS does not support capacity expansion or deduction of instance vCPUs and memory.
- AS resources must comply with quota requirements listed in [Table 1-3](#).

Table 1-3 Quotas

Item	Description	Default
AS group	Maximum number of AS groups per region per account	10
AS configuration	Maximum number of AS configurations per region per account	100
AS policy	Maximum number of AS policies per AS group	10
Instance	Maximum number of instances per AS group	300
Bandwidth scaling policy	Maximum number of bandwidth scaling policies per region per account	10

1.5 Region and AZ

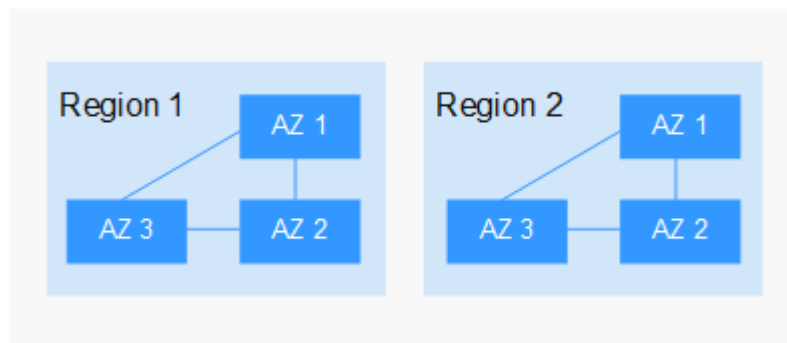
Concept

A region and availability zone (AZ) identify the location of a data center. You can create resources in a specific region and AZ.

- A region is a physical data center, which is completely isolated to improve fault tolerance and stability. The region that is selected during resource creation cannot be changed after the resource is created.
- An AZ is a physical location where resources use independent power supplies and networks. A region contains one or more AZs that are physically isolated but interconnected through internal networks. Because AZs are isolated from each other, any fault that occurs in one AZ will not affect others.

Figure 1-9 shows the relationship between regions and AZs.

Figure 1-9 Regions and AZs



Selecting a Region

Select a region closest to your target users for lower network latency and quick access.

Selecting an AZ

When deploying resources, consider your applications' requirements on disaster recovery (DR) and network latency.

- For high DR capability, deploy resources in different AZs within the same region.
- For lower network latency, deploy resources in the same AZ.

Regions and Endpoints

Before you use an API to call resources, specify its region and endpoint. For more details, see [Regions and Endpoints](#).

1.6 AS and Other Services

AS can work with other cloud services to meet your requirements for different scenarios.

Figure 1-10 shows the relationships between AS and other services.

Figure 1-10 Relationships between AS and other services

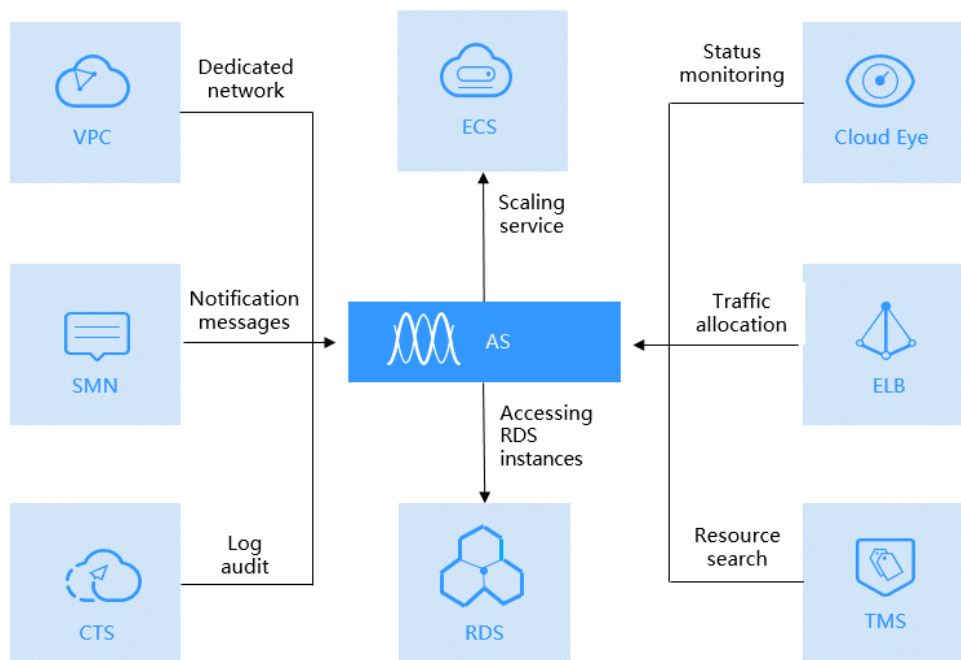


Table 1-4 Related services

Service	Description	Interaction	Reference
Elastic Load Balance (ELB)	After ELB is configured, AS automatically associates ECS instances to a load balancer listener when adding ECSs, and unbinds them when removing the instances. For AS to work with ELB, the AS group and load balancer must be in the same VPC.	AS distributes traffic to all ECSs in an AS group.	(Optional) Adding a Load Balancer to an AS Group
Cloud Eye	If an alarm-triggered policy is configured, AS triggers scaling actions when an alarm condition specified in Cloud Eye is met.	AS scales resources based on ECS instance status monitored by Cloud Eye.	AS Metrics
ECS	ECS instances added in a scaling action can be managed and maintained on the ECS console.	AS automatically adjusts the number of ECS instances.	Dynamically Expanding Resources and Scheduled Scaling
Virtual Private Cloud (VPC)	AS automatically adjusts the bandwidths of EIPs assigned in VPCs and also shared bandwidths.	AS automatically adjusts the bandwidth.	Creating a Bandwidth Scaling Policy
Simple Message Notification (SMN)	If you enable the SMN service, the system sends you notifications about the status of your AS group in a timely manner.	Message notification	Configuring Notification for an AS Group

Service	Description	Interaction	Reference
Cloud Trace Service (CTS)	With CTS, you can record AS operation logs for view, audit, and backtracking.	Log audit	Recording AS Resource Operations
Tag Management Service (TMS)	If you have multiple resources of the same type, TMS enables you to manage these resources more easily.	Tags	Marking AS Groups and Instances
Relational Database Service (RDS)	The prerequisites for directly accessing an RDS DB instance from a scaled instance are as follows: <ul style="list-style-type: none"> The scaled instance and the destination RDS DB instance must be in the same VPC. The scaled instance must be allowed by the security group to access RDS DB instances. 	The scaled instances can access RDS DB instances.	Connecting to a DB Instance Through a Private Network

1.7 Permissions Management

If you need to assign different permissions to employees in your enterprise to access your AS resources, Identity and Access Management (IAM) is a good choice for fine-grained permissions management. IAM provides identity authentication, permissions management, and access control, helping you securely access to your resources.

With IAM, you can create IAM users and assign permissions to the users to control their access to specific resources. For example, you can assign permissions to allow

some software developers to use AS resources but disallow them to delete or perform any high-risk operations on the resources.

If your account does not need individual IAM users for permissions management, skip this section.

IAM can be used free of charge. You pay only for the resources in your account. For more information about IAM, see IAM Service Overview.

AS Permissions

By default, new IAM users do not have any permissions assigned. You need to add them to one or more groups and attach policies or roles to these groups so that these users can inherit permissions from the groups and perform specified operations on cloud services.

When you grant AS permissions to a user group, set **Scope** to **Region-specific projects** and then select projects for the permissions to take effect. If you select **All projects**, the permissions will take effect for the user group in all region-specific projects. When accessing AS, the users need to switch to a region where they have been authorized to use this service.

You can grant users permissions by using roles and policies.

- **Roles:** A type of coarse-grained authorization mechanism that defines permissions related to user responsibilities. This mechanism provides only a limited number of service-level roles for authorization. When using roles to grant permissions, you need to also assign other roles on which the permissions depend to take effect. However, roles are not an ideal choice for fine-grained authorization and secure access control.
- **Policies:** A type of fine-grained authorization mechanism that defines permissions required to perform operations on specific cloud resources under certain conditions. This mechanism allows for more flexible policy-based authorization, meeting requirements for secure access control. For example, you can grant AS users only the permissions for managing a certain type of ECSs. Most policies define permissions based on APIs. For the API actions supported by AS, see .

Table 1-5 lists all the system policies supported by AS.

Table 1-5 System-defined permissions supported by AS

Policy Name	Description	Category	Dependency
AutoScalingFullAccess	All operation permissions on all AS resources	System-defined policy	None
AutoScalingReadOnlyAccess	Read-only permissions on all AS resources	System-defined policy	None

Policy Name	Description	Category	Dependency
AutoScaling Administrator	All operation permissions on all AS resources	System role	The ELB Administrator , CES Administrator , Server Administrator , and Tenant Administrator roles need to be assigned in the same project.

Table 1-6 lists the common operations supported by each system-defined policy of AS. Select the policies as required.

Table 1-6 Common operations supported by each system-defined policy of AS

Operation	AutoScaling FullAccess	AutoScaling ReadOnlyAccess	AutoScaling Administrator
Creating an AS group	√	x	√
Modifying an AS group	√	x	√
Querying details about an AS group	√	√	√
Deleting an AS group	√	x	√
Creating an AS configuration	√	x	√
Creating an AS policy	√	x	√
Creating a bandwidth scaling policy	√	x	√

Helpful Links

- [What Is IAM?](#)

- [Creating a User and Granting AS Permissions](#)

1.8 Basic Concepts

AS Group

An AS group consists of a collection of ECS instances that apply to the same scenario. It is the basis for enabling or disabling AS policies and performing scaling actions.

AS Configuration

An AS configuration is a template specifying specifications for the ECS instances to be added to an AS group. The specifications include the ECS type, vCPUs, memory, image, login mode, and disk.

AS Policy

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

Scaling Action

A scaling action adds instances to or removes instances from an AS group. It ensures that the expected number of instances are running in the AS group by adding or removing instances when the triggering condition is met, which improves system stability.

Cooldown Period

To prevent an alarm-based policy from being triggered repeatedly by the same event, configure a cooldown period. A cooldown period specifies how long any alarm-triggered scaling action will be disallowed after a previous scaling action is complete. This cooldown period does not apply to scheduled or periodic scaling actions.

For example, if you set the cooldown period to 300 seconds (5 minutes), and there is a scaling action scheduled for 10:32, but a previous scaling action was complete at 10:30, any alarm-triggered scaling actions will be denied during the cooldown period from 10:30 to 10:35, but the scheduled scaling action will still be triggered at 10:32. If the scheduled scaling action ends at 10:36, a new cooldown period starts at 10:36 and ends at 10:41.

Bandwidth Scaling

AS automatically adjusts a bandwidth based on the scaling policies you configured. AS can only adjust the bandwidths of EIPs and share bandwidths that are billed on a pay-per-use basis.

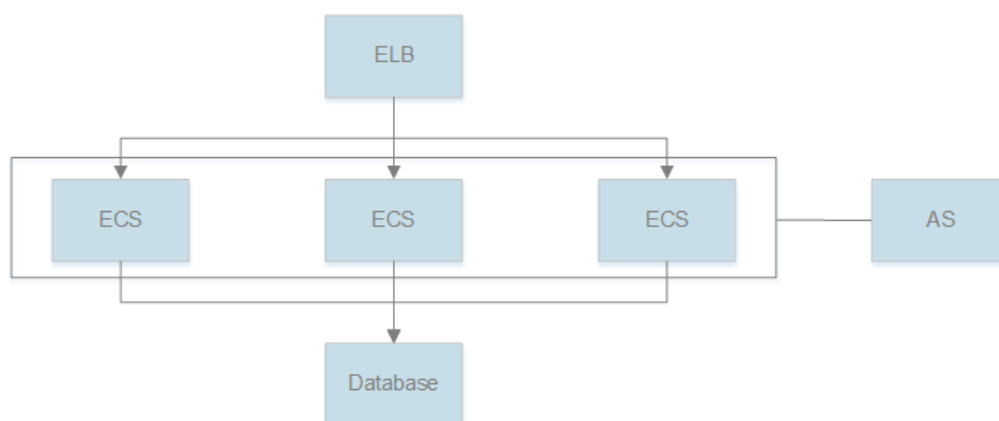
2 Best Practices

2.1 Setting Up an Automatically Scalable Discuz! Forum

Overview

AS automatically adds instances to an AS group for applications and removes unneeded ones on demand. You do not need to prepare a large number of extra ECS instances for an expected marketing activity or unexpected peak hours. By eliminating the need to deploy those extra instances, AS ensures system reliability and reduces your operating costs.

This section describes how to use services, such as AS, ECS, ELB, and VPC to deploy a web service that can be automatically scaled in and out, for example, a Discuz! forum.



Prerequisites

1. A VPC, subnet, security group, and EIP are available.
2. A load balancer and listener have been created. The VPC obtained in **1** is selected during the load balancer creation.

Procedure

Create an ECS and install a MySQL database.

You can create a relational database using the Relational Database Service (RDS) service provided by the cloud platform, or create an ECS and install the database there. In this section, we will install a MySQL database on a newly created ECS.

1. Use the created VPC, security group, and EIP for the ECS you create. For instructions about how to create an ECS, see *Elastic Cloud Server User Guide*.
2. When the status of the ECS changes to **Running**, use Xftp or Xshell to log in to the ECS through its EIP, and install and configure a MySQL database.

Create an ECS and deploy a Discuz! forum on it.

1. Create an ECS but do not bind an EIP to it. For instructions about how to create an ECS, see *Elastic Cloud Server User Guide*.
2. Unbind the EIP from the ECS where the MySQL database is installed and bind the EIP to the ECS where the Discuz! forum is to be deployed.

You can access the MySQL database through a private network, so the EIP bound to the ECS where the MySQL database is installed can be unbound and then bound to the ECS where the Discuz! forum is to be deployed. This improves resource utilization. For detailed operations, see *Virtual Private Cloud User Guide*. After binding the EIP, you can access the ECS from the Internet and install various environments, such as PHP and Apache.

3. Deploy the forum.

For instructions about how to deploy the Discuz! forum, see officially released Discuz! documentation. When configuring parameters, configure the private IP address of the ECS where the MySQL database is installed for the database server, and use the username and password authorized for remotely accessing the ECS where the MySQL database is installed to access the MySQL database. After the configuration is complete, you can unbind the EIP from the ECS where the forum is deployed to reduce resource usage.

Create a private image.

Use the ECS where the Discuz! forum is deployed to create a private image. This private image is used to create the ECSs that will be used for capacity expansion.

1. Only a stopped ECS can be used to create a private image, so stop the ECS where the Discuz! forum is deployed before creating a private image. For detailed operations, see *Elastic Cloud Server User Guide*.
2. Use the ECS to create a private image. For details, see *Image Management Service User Guide*.

Create an AS group.

An AS group consists of a collection of ECS instances, AS configurations, 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. You must create an AS group to automatically add or remove ECS instances to match changes in traffic to the Discuz! forum.

For instructions about how to create an AS group, see [Creating an AS Group](#). During the configuration, use the created VPC, subnet, load balancer, and listener.

Create an AS configuration.

The AS configuration lists the basic specifications of the ECSs to be automatically added to the AS group in a scaling action.

During the configuration, select the private image you created in the preceding step. Configure other parameters based on service requirements.

Manually add the ECS to the AS group.

On the page providing details about the AS group, click the **Instances** tab and then **Add** to add the ECS where the Discuz! forum is deployed to the AS group. You can enable instance protection on this ECS so that it will not be automatically removed from the AS group.

Create an AS policy.

An AS policy specifies the conditions for triggering a scaling action. After you create an AS policy for the AS group, AS automatically increases or decreases the number of instances based on the policy.

You can configure an alarm-based AS policy. When Cloud Eye generates an alarm for a monitoring metric, such as vCPU usage, AS automatically increases or decreases the number of instances in the AS group. If traffic fluctuations are predictable, you can also configure a scheduled or periodic AS policy.

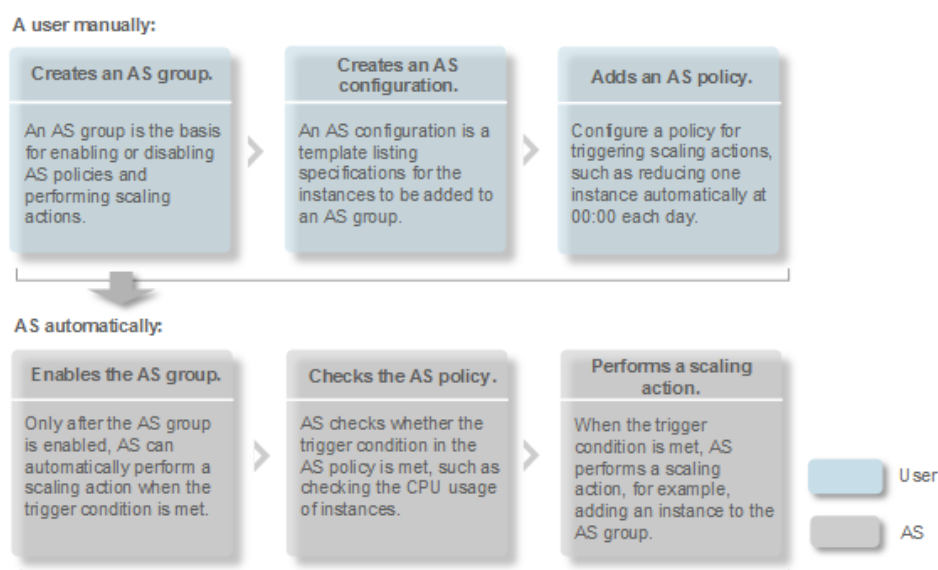
After an AS policy is created and enabled, if a triggering condition is met, the AS group scales in or out as needed.

3 Quick Start

3.1 Wizard-based Process of Using AS

Figure 3-1 illustrates the wizard-based process of using AS.

Figure 3-1 Wizard-based process of using AS



3.2 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 pairs for logging in to the instances added by a scaling action if key authentication is used.

Procedure

- Log in to the management console.
- Under **Computing**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
- Click **Create AS Group**.
- Set basic information about the AS group, such as **Name**, **Max. Instances**, **Min. Instances**, and **Expected Instances**. [Table 3-1](#) lists the parameters.

Table 3-1 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
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. <ul style="list-style-type: none"> If you require high availability, buy servers in difference AZs. If you require low network latency, buy servers in the same AZ. 	N/A
Name	Specifies the name of the AS group to be created. The name can contain up to 64 characters. It can include letters, digits, underscores (_), and hyphens (-).	N/A
Max. Instances	Specifies the maximum number of ECS instances in an AS group.	0
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.	1

Parameter	Description	Example Value
VPC	Provides a network for your ECS instances. All ECS instances in the AS group are deployed in this VPC.	N/A
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.	N/A
Load Balancing	<p>This parameter is optional. A load balancer automatically distributes traffic across all instances in an AS group to balance their service load. It improves fault tolerance of your applications and expands application service capabilities.</p> <p>NOTE</p> <ul style="list-style-type: none"> • 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 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. 	N/A

Parameter	Description	Example Value
Instance Removal Policy	<p>Specifies the priority for removing instances from an AS group. If specified conditions are met, scaling actions are triggered to remove instances. AS supports the following instance removal policies:</p> <ul style="list-style-type: none"> ● 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. <p>NOTE</p> <ul style="list-style-type: none"> ● 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. 	N/A
EIP	<p>If EIP has been selected in an 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.</p>	N/A

Parameter	Description	Example Value
Health Check Method	<p>When a health check detects an unhealthy ECS instance, AS replaces the faulty with a new one. There are two health check methods:</p> <ul style="list-style-type: none"> • 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 evaluates the running status of each instance based on the health check results. If the health check results show that an instance is unhealthy, AS removes the instance from the AS group. 	N/A
Health Check Interval	<p>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 the site requirements.</p>	5 minutes
Enterprise Project	<p>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.</p> <p>NOTE</p> <ul style="list-style-type: none"> • 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. 	N/A
Advanced Settings	<p>Configure notifications. You can select Do not configure or Configure now.</p>	N/A

Parameter	Description	Example Value
Notification	<p>Results of scaling actions are sent to you based on the functions provided by the Simple Message Notification (SMN) service.</p> <ul style="list-style-type: none"> • Notification Conditions: When at least one of the following conditions is met, SMN sends a notification to you: <ul style="list-style-type: none"> - 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 <i>Simple Message Notification User Guide</i>. 	N/A
Tag	<p>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.</p> <p>Each tag contains a key and a value. You can specify the key and value for each tag.</p> <ul style="list-style-type: none"> • Key <ul style="list-style-type: none"> - 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 <ul style="list-style-type: none"> - 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: =*<>\, / 	N/A

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**.

4 AS Management

4.1 AS Group

4.1.1 Creating an AS Group

Scenarios

An AS group consists of a collection of 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 again later.
 - 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. Under **Computing**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
3. Click **Create AS Group**.
4. Set parameters, such as **Name**, **Max. Instances**, **Min. Instances**, and **Expected Instances**. [Table 4-1](#) describes the key parameters to be configured.

Table 4-1 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
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. <ul style="list-style-type: none"> • If you require high availability, buy servers in difference AZs. • If you require low network latency, buy servers in the same AZ. 	N/A
Multi-AZ Extension Policy	This parameter can be set to Load-balanced or Sequenced . <ul style="list-style-type: none"> • Load-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 AZ selected according to the order in which AZs are specified. <p>NOTE This parameter needs to be configured when two or more AZs are selected.</p>	Load-balanced
Name	Specifies the name of the AS group to be created. The name can contain up to 64 characters. It can include letters, digits, underscores (_), and hyphens (-).	N/A

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.	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.	N/A
VPC	Provides a network for your ECS instances. All ECS instances in the AS group are deployed in this VPC.	N/A
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 instance.	N/A
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 fault tolerance of your applications and expands application service capabilities. NOTE <ul style="list-style-type: none"> 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 it with a new one. 	N/A

Parameter	Description	Example Value
Instance Removal Policy	<p>Controls which instances are first to be removed during scale in. If specified conditions are met, scaling actions are triggered to remove instances. You can choose from any of the following instance removal policies:</p> <ul style="list-style-type: none"> ● 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. <p>NOTE</p> <ul style="list-style-type: none"> ● AS always balances instances across AZs first, regardless of which removal policy is used. ● 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 instance. If multiple manually added instances must be removed, AS preferentially removes the earliest-added instance first. 	Oldest instance created from oldest AS configuration
Health Check Method	<p>When a health check detects an unhealthy ECS instance, AS replaces it with a new one. You can choose from either of the following health check methods:</p> <ul style="list-style-type: none"> ● 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 evaluates the running status of each instance based on the health check results. If the health check results show that 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 the AS group uses a load balancing listener. If a load balancing listener detects that an ECS instance is unhealthy, AS removes the instance from the AS group. 	N/A

Parameter	Description	Example Value
Advanced Settings	Configure notifications and tags. You can select Do not configure or Configure now .	N/A

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**.

4.1.2 (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](#).

4.1.3 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.

Procedure

1. Log in to the management console.
2. Under **Computing**, 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 AS 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**.

4.1.4 Enabling an AS Group

Scenarios

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

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 instances if the number of 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 **Computing**, 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. You can also click the AS group name and then **Enable** to the right of **Status** on the **Basic Information** page to enable the AS group.
4. In the **Enable AS Group** dialog box, click **Yes**.

4.1.5 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 **Elastic Cloud Server** 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 **Computing**, 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. You can also click the AS group name and then **Disable** to the right of **Status** on the **Basic Information** page to disable the AS group.
4. In the **Disable AS Group** dialog box, click **Yes**.

4.1.6 Modifying an AS Group

Scenarios

You can modify an AS group if needed. The values of the following parameters can be changed: **Name**, **Max. Instances**, **Min. Instances**, **Expected Instances**, **Health Check Method**, **Health Check Interval**, **Instance Removal 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 **Computing**, 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**.

4.1.7 Deleting an AS Group

Scenarios

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

- If an AS group is not required during a specified period of time, 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.
- When an AS group is deleted, its AS policies and the alarm rules generated based those AS policies will be automatically deleted.

Procedure

1. Log in to the management console.
2. Under **Computing**, 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 **Yes**.

4.2 AS Configuration

4.2.1 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 a new 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 instance by default. For details, see [Creating an AS Configuration from an Existing ECS Instance](#).
- Create an AS configuration from a new specifications template.
If you have special requirements on the ECSs for resource expansion, use a new specifications template to create the AS configuration. For details, see [Creating an AS Configuration from a New Specifications Template](#).

4.2.2 Creating an AS Configuration from an Existing ECS Instance

Scenarios

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

Procedure

1. Log in to the management console.

2. Under **Computing**, 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 4-2](#) lists the AS configuration parameters.

Table 4-2 AS configuration parameters

Parameter	Description	Example Value
Region	A region is where an AS configuration resides.	N/A
Name	Specifies the name of an AS configuration.	N/A
Configuration Template	Select Use specifications of an existing ECS > Select ECS . The ECS type, vCPUs, memory, image, and disk information in the AS configuration are the same as those of the selected ECS by default.	Use specifications of an 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: <ul style="list-style-type: none"> • 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. <p>NOTE If you select Automatically assign, specify Type, Billed By, and Bandwidth.</p>	Automatically assign
Bandwidth Type	You can select Dedicated or Shared . <ul style="list-style-type: none"> • Dedicated: The bandwidth can be used by only one EIP. • Shared: The bandwidth can be used by multiple EIPs. <p>NOTE</p> <ul style="list-style-type: none"> • 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 only be billed by bandwidth. You can select which shared bandwidth the EIP is added to. 	Shared

Parameter	Description	Example Value
Login Mode	<p>An ECS can be authorized using a key pair or a password.</p> <ul style="list-style-type: none"> Key pair Keys are used for authenticating the users who attempt to log in to ECS instances. If you select this mode, create or import a key pair on the Key Pair page. <p>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 instance.</p> <ul style="list-style-type: none"> 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	<p>This allows you to configure User Data. You can select Do not configure or Configure now.</p>	N/A
User Data	<p>Enables an ECS to automatically inject user data when the ECS starts for the first time. This configuration is optional. If this function is enabled, the ECS automatically injects user data during its first startup.</p> <p>For details, see <i>Elastic Cloud Server User Guide</i>.</p> <p>The following two methods are available:</p> <ul style="list-style-type: none"> As text: allows you to enter the user data in the text box below. As file: allows you to inject a script or other files when you create an ECS instance. <p>NOTE</p> <ul style="list-style-type: none"> 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. 	-

- Click **Create Now**.
- 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](#).

4.2.3 Creating an AS Configuration from a New Specifications Template

Scenarios

If you have special requirements on the ECS instances for resource expansion, 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 **Computing**, 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 4-3](#) lists the AS configuration parameters.

Table 4-3 AS configuration parameters

Parameter	Description	Example Value
Region	A region is where an AS configuration resides.	N/A
Name	Specifies the name of the AS configuration to be created.	N/A
Configuration Template	Select Create a new specifications 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 a new specifications template
CPU Architecture	Both x86 and Kunpeng CPU architectures are available: <ul style="list-style-type: none"> • x86: The x86-based CPU architecture uses Complex Instruction Set Computing (CISC). • Kunpeng: The Kunpeng-based CPU architecture uses Reduced Instruction Set Computing (RISC). <p>NOTE This parameter is displayed only when both x86-based and Kunpeng-based ECSs are available in the current region.</p>	x86

Parameter	Description	Example Value
Specifications	<p>The public cloud provides various ECS types for different application scenarios. For more information, see <i>Elastic Cloud Server User Guide</i>.</p> <p>Configure the ECS specifications, including vCPUs, memory, image type, and disk, according to the ECS type.</p>	Memory-optimized ECS
Image	<ul style="list-style-type: none"> • 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. You can configure the applications or software in the public image 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 user. 	Public image

Parameter	Description	Example Value
Disk	<p>Includes system disks and data disks.</p> <ul style="list-style-type: none"> 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. Ultra-high I/O: uses solid state disk (SSD) drives to store data. <p>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.</p> <ul style="list-style-type: none"> Data Disk You can create multiple data disks for an ECS instance. In addition, you can specify a data disk image for exporting data. <p>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.</p>	<p>Common I/O for System Disk</p>
Security Group	<p>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.</p>	<p>N/A</p>

Parameter	Description	Example Value
EIP	<p>An EIP is a static public IP address bound to an ECS in a VPC. Using the EIP, the ECS provides services externally.</p> <p>The following options are provided:</p> <ul style="list-style-type: none"> ● 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. <p>NOTE If you select Automatically assign, you need to specify Type, Billed By, and Bandwidth.</p>	Automatically assign
Bandwidth	<p>You can select Dedicated or Shared.</p> <ul style="list-style-type: none"> ● Dedicated: The bandwidth can be used by only one EIP. ● Shared: The bandwidth can be used by multiple EIPs. <p>NOTE</p> <ul style="list-style-type: none"> ● 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. ● The shared bandwidth can be billed only by bandwidth. You can select a shared bandwidth to which the EIP is to be added. 	Shared

Parameter	Description	Example Value
Login Mode	<p>An ECS can be authorized using a key pair or a password.</p> <ul style="list-style-type: none"> Key pair Keys are used for authenticating the users who attempt to log in to ECS instances. If you select this mode, create or import a key pair on the Key Pair page. <p>NOTE If you use an existing key, make sure that you have saved the key file locally.</p> <ul style="list-style-type: none"> 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	<p>This parameter allows you to configure ECS Group and User Data.</p> <p>You can select Do not configure or Configure now.</p>	N/A
User Data	<p>Enables an ECS to automatically inject user data when the ECS starts for the first time. This configuration is optional. After this function is enabled, the ECS automatically injects user data during its first startup.</p> <p>For details, see <i>Elastic Cloud Server User Guide</i>.</p> <p>The following methods are available:</p> <ul style="list-style-type: none"> 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. <p>NOTE</p> <ul style="list-style-type: none"> 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. 	-

5. Click **Create Now**.

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](#).

4.2.4 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 **Computing**, 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**.

4.2.5 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 batches.

Procedure

1. Log in to the management console.
2. Under **Computing**, 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.

4.3 AS Policy

4.3.1 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 that will be performed. When the triggering 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.

AS supports the following policies:

- Alarm policy: AS automatically adjusts the number of 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 CPU Usage.
- Scheduled policy: AS automatically increases or decreases the number of 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 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
AS allows you to adjust resources by manually adding instances to an AS group, removing instances from an AS group, or changing the expected number of instances.

4.3.2 Creating an AS Policy

Scenarios

You can manage instances in an AS group through AS policies. This section describes how to create an AS policy.

Creating an Alarm Policy

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

2. Locate the row containing the target AS group and click **View AS Policy** in the **Operation** column.
3. On the **AS Policies** page, click **Add AS Policy**.
4. Set the parameters listed in [Table 4-4](#).

Table 4-4 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
Alarm Rule	<p>Specifies whether a new alarm rule is to be created (Create) or an existing alarm rule will be used (Use existing).</p> <p>For details about how to use an existing alarm rule, see Setting Monitoring Alarm Rules.</p> <p>If you choose to create an alarm, system monitoring and custom monitoring are supported.</p> <ul style="list-style-type: none"> • System monitoring requires the parameters in Table 4-5. • Custom monitoring requires the parameters in Table 4-6. 	N/A

Parameter	Description	Example Value
Scaling Action	<p>Specifies an action and the number or percentage of instances.</p> <p>The following scaling action options are available:</p> <ul style="list-style-type: none"> • 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. 	<ul style="list-style-type: none"> • 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: <ul style="list-style-type: none"> – 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. <p>For example, 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.</p>

Parameter	Description	Example Value
<p>Cooldown Period</p>	<p>To prevent an alarm policy from being repeatedly invoked by the same event, you can set a cooldown period.</p> <p>A cooldown period specifies how long any alarm-triggered scaling action will be disallowed after a previous scaling action is complete.</p> <p>The cooldown period does not apply to scheduled or periodic scaling actions. However, AS recounts the cooling duration after a scheduled or periodic scaling action is complete.</p> <p>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.</p> <p>NOTE</p> <ul style="list-style-type: none"> • 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. 	<p>300</p>

Table 4-5 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
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 4-6 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
Monitored Object	Specifies the resources to which the alarm rule applies.	N/A
Trigger Condition	Select monitoring metrics supported by AS and set alarm conditions for the metrics.	CPU Usage Max. >70%

Parameter	Description	Example Value
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

5. Click **OK**.

The newly added AS policy is displayed on the **AS Policy** 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 **Computing**, 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 4-7](#).

Table 4-7 Parameter description

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

Parameter	Description	Example Value
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: <ul style="list-style-type: none">• Interval<ul style="list-style-type: none">- One day- One week- One month• Time Range Specifies a time range during which the AS policy can be triggered.	N/A
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 a time at which the AS policy is triggered.	N/A

Parameter	Description	Example Value
Scaling Action	<p>Specifies an action and the number of instances.</p> <p>The following scaling action options are available:</p> <ul style="list-style-type: none"> • 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. 	<ul style="list-style-type: none"> • 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: <ul style="list-style-type: none"> - 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. <p>For example, there are 10 instances in an AS group, and the scaling action is Add 15% instances. When the AS policy is invoked, 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.</p>

Parameter	Description	Example Value
Cooldown Period	<p>To prevent an alarm policy from being repeatedly invoked by the same event, you can set a cooldown period.</p> <p>Specifies how long any alarm-triggered scaling action will be disallowed after a previous scaling action is complete.</p> <p>The cooldown period does not apply to scheduled or periodic scaling actions. However, AS recounts the cooling duration after a scheduled or periodic scaling action is complete.</p> <p>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.</p> <p>NOTE</p> <ul style="list-style-type: none"> • 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. 	300

6. Click **OK**.

The newly added AS policy is displayed on the **AS Policy** 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.

4.3.3 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 **Computing**, 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.

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.

4.4 Scaling Action

4.4.1 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%. The other policy is used to remove an instance if the minimum CPU usage drops below 30%.

4.4.2 Scheduled Scaling

To satisfy demands that change regularly, you can configure a scheduled or periodic policy to scale resources at specified time or periodically. 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.

4.4.3 Manual Scaling

Scenarios

You can manually add or remove instances to or from an AS group, or changing the expected number of instances.

Procedure

Adding instances to an AS group

If an AS group is enabled and has no ongoing scaling action, and the current number of instances is less than the maximum, you can manually add instances to the AS group.

Before adding instances to an AS group, ensure that the following conditions are met:

- The instances are not in other AS groups.
- The instances are in the same VPC as the AS group.
- After the instances are added, the total number of instances is less than or equal to the maximum number of instances allowed.
- Up to 10 instances can be added at a time.

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

1. Click the **AS Groups** tab and then the name of the target AS group.
2. On the AS group details page, click the **Instances** tab and then **Add**.
3. Select the instances to be added and click **OK**.

Removing instances 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. An instance removed from the AS group no longer carries any application traffic.

You can modify the AS configuration for an AS group at any time, but the new configuration will not be applied to any instances that are running. To apply the

new configuration, stop an instance, and the system will replace it with a new one that has the specifications specified in the new configuration. You can also remove the instance from the AS group, update the instance, and add then instance back to the AS group.

There are some restrictions on instance removal:

- The AS group cannot have a scaling action in progress, the instances must be enabled, and the total number of instances after removal cannot be less than the minimum number of instances specified.
- Instances can be removed from an AS group and deleted only if the AS group has no scaling action ongoing, and the instances are automatically created and enabled, and are not used by Storage Disaster Recovery Service (SDRS).
- Instances manually added to an AS group can only be removed. They cannot be removed and deleted.
- A maximum of 10 instances can be removed at a time.

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

1. Click the **AS Groups** tab and then the name of the target AS group.
2. Click the **Instances** tab, locate the row containing the target instance, and click **Remove** or **Remove and Delete** in the **Operation** column.
3. To delete multiple instances from an AS group, select the check boxes in front of them and click **Remove** or **Remove and Delete**.

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

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](#).

4.4.4 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. Besides, AS will check whether instances are evenly distributed in the currently used AZs. If the load among AZs is unbalanced, AS balances load among AZs when removing instances. If the load among AZs is balanced, AS removes instances following the pre-configured instance removal policy.

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

4.4.5 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 Scaling Actions

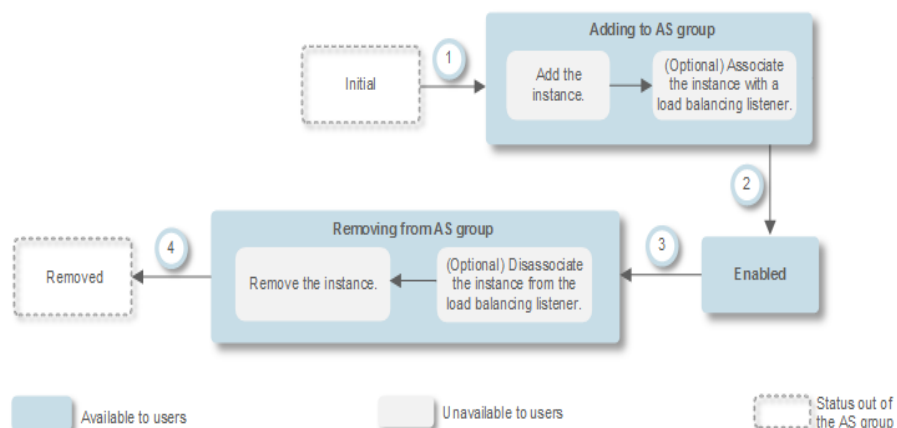
The following steps illustrate how to view scaling actions of an AS group.

1. Log in to the management console.
2. Click the **AS Groups** tab and then the name of the target AS group.

4.4.6 Managing Lifecycle Hooks

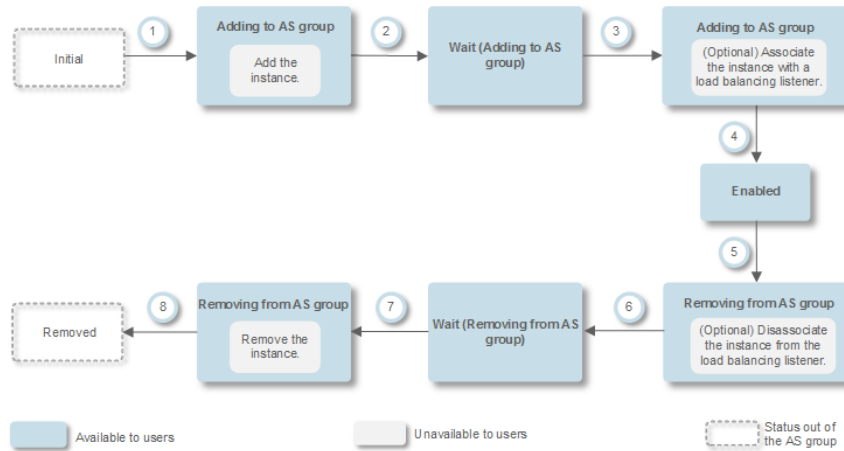
Lifecycle hooks enable you to flexibly control addition and removal of ECS instances in AS groups and manage the lifecycle of ECS instances in AS groups. [Figure 4-1](#) shows the instance lifecycle when no lifecycle hook is added to an AS group.

Figure 4-1 Instance lifecycle when no lifecycle hook is added to an AS group



[Figure 4-2](#) shows the instance lifecycle when a lifecycle hook is added to an AS group.

Figure 4-2 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 4-2](#). 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 instance has been in the wait state longer than the timeout period.
- A callback action is performed to end the instance wait state.

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, stops accepting new requests, and finishes 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 **Wait (Adding to AS group)** 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, 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 from AS group** state. After a lifecycle hook is triggered, the instance enters the **Wait (Removing from AS group)** 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 **Computing**, 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 4-8](#).

Table 4-8 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 an instance that is being added to an AS group to Wait (Adding to AS group) state. Instance removal puts an instance that is being removed from an AS group to Wait (Removing from AS group) state.	Instance adding

Parameter	Description	Example Value
Default Callback Action	<p>Specifies the action that the system takes when an instance moves out of a wait state.</p> <p>The default callback action for an Instance adding lifecycle hook can be Continue or Abandon:</p> <ul style="list-style-type: none"> • 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. <p>The default callback action for an Instance removal lifecycle hook can be Continue or Abandon:</p> <ul style="list-style-type: none"> • 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)	<p>Specifies the amount of time for the instances to remain in a wait state. The value ranges from 300s to 86400s. The default value is 3600s.</p> <p>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.</p>	3600

Parameter	Description	Example Value
Notification Topic	Specifies a notification object for a lifecycle hook. For details, see "Creating a Topic" in <i>Simple Message Notification User Guide</i> . 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: <pre>{ "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" } }</pre>	N/A
Notification Message	After a notification object is configured, the system sends your custom notification to the object.	N/A

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 **Wait (Adding to AS group)** or **Wait (Removing from AS group)** in the **Lifecycle Status** column.

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.

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 4-8](#). 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 4-8](#) 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.

4.4.7 Configuring Instance Protection

Scenarios

To control whether a particular 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 **Computing**, 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 **Computing**, 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**.

4.5 Bandwidth Scaling

4.5.1 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 **Computing**, 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 4-9](#).

Table 4-9 Alarm policy parameters

Parameter	Description	Example Value
Region	Specifies the region where the AS group resides.	N/A
Policy Name	Specifies the name of the bandwidth scaling policy. The name consists of only letters, digits, underscores (_), and hyphens (-).	N/A
EIP	Specifies the public network IP address whose bandwidth needs to be scaled.	N/A

Parameter	Description	Example Value
Policy Type	Select Alarm .	Alarm
Alarm Rule	<p>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.</p> <p>To create an alarm rule, configure the following parameters:</p> <ul style="list-style-type: none"> ● Rule Name Specifies the name of the new alarm rule, for example, as-alarm-7o1u. ● Trigger Condition Select a monitoring metric and trigger condition based on the metric. Table 4-10 lists the supported monitoring metrics. An example value is Outbound Traffic Avg. > 100 bit/s. ● Monitoring Interval Specifies the period for the metric, for example, 5 minutes. ● Consecutive Occurrences Specifies the number of consecutive times, for example, one time, for triggering a scaling action during a monitoring period. 	N/A

Parameter	Description	Example Value
Scaling Action	<p>Specifies the execution action in the AS policy. The following scaling action options are available:</p> <ul style="list-style-type: none"> • 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. <p>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.</p> <ul style="list-style-type: none"> • 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 1000 Mbit/s, the default step is 50 Mbit/s. • If the bandwidth is greater than 1000 Mbit/s, the default step is 500 Mbit/s. 	N/A
Cooldown Period	Specifies a period of time in the unit of second after each scaling action is complete. During the cooldown period, scaling actions triggered by alarms will be denied. Scheduled and periodic scaling actions are not restricted.	300s

Table 4-10 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.

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 and [Table 4-11](#). For more information about how to set alarm rules, see *Cloud Eye User Guide*.

Table 4-11 Key parameters for creating an alarm rule

Parameter	Description	Example Value
Name	Specifies the name of the alarm rule.	alarm-bandwidth
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
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 resources by bandwidth name or ID, which can be obtained on the page providing details about the target EIP.	Specific resources
Method	Select the method of creating an alarm. Set this parameter to Create manually .	Create 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 4-10 .	N/A

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, create your desired alarm rule on the **Cloud Eye** page before creating a bandwidth scaling policy. Ensure that the specific resources selected during alarm rule creation are the bandwidth of the EIP selected for the bandwidth scaling policy to be created. 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 **Computing**, 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 4-12](#).

Table 4-12 Scheduled or periodic policy parameters

Parameter	Description	Example Value
Region	Specifies the region where the AS group resides.	N/A
Policy Name	Specifies the name of the bandwidth scaling policy. The name consists of only letters, digits, underscores (_), and hyphens (-).	as-policy-p6g5
EIP	Specifies the public network IP address whose bandwidth needs to be scaled. This parameter is mandatory when Resource Type is set to EIP .	N/A
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: <ul style="list-style-type: none"> • Time Range Specifies a time range during which the AS policy can be triggered. • Interval <ul style="list-style-type: none"> - One day - One week - One month 	N/A
Triggered At	Specifies a time at which the AS policy is triggered.	N/A

Parameter	Description	Example Value
Scaling Action	<p>Specifies the action to be performed.</p> <p>The following scaling action options are available:</p> <ul style="list-style-type: none"> • 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. <p>NOTE</p> <p>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.</p> <ul style="list-style-type: none"> • 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 1000 Mbit/s, the default step is 50 Mbit/s. • If the bandwidth is greater than 1000 Mbit/s, the default step is 500 Mbit/s. 	N/A
Cooldown Period	<p>Specifies a period of time in the unit of second after each scaling action is complete. During the cooldown period, scaling actions triggered by alarms will be denied. Scheduled and periodic scaling actions are not restricted.</p>	300s

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

4.5.2 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. This section describes how to create a bandwidth scaling policy.

Procedure

1. Log in to the management console.
2. Under **Computing**, 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.

4.5.3 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 corresponding to a released EIP still occupies the policy quota. Only the account and its IAM users with the global permission can manage the AS bandwidth 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 **Computing**, 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 **Computing**, 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 **Computing**, 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 the following parameters of a bandwidth scaling policy: **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 **Computing**, 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 **Computing**, 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.

4.6 AS Group and Instance Monitoring

4.6.1 Health Check

A health check removes abnormal 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 check.

- **ECS health check:** checks ECS instance running status. If an instance is stopped or deleted, it is considered as abnormal. **ECS health check** is the default health check mode for an AS group. The AS group periodically uses the check result to determine the running status of every instance in the AS group. If the health check results show that 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. 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 to be healthy only when all load balancers detect that the instance status is healthy. If any load balancer detects that an instance is unhealthy, the instance will be removed from the AS group.

In both **ECS health check** and **ELB health check** methods, AS removes unhealthy instances from AS groups. However, the removed instances are processed differently in the following two scenarios:

For instances automatically added to an AS group during scaling actions, AS removes and deletes them. For instances manually added to an AS group, AS only removes them from the AS group.

When an AS group is disabled, checking instance health status continues. However, AS will not remove instances.

4.6.2 Configuring Notification for an AS Group

Scenarios

After the SMN service is provisioned, you can promptly send AS group information, such as successful instance increasing, failed instance increasing, successful instance decreasing, failed instance decreasing, or AS group exception to the user using the SMN service. This helps the user learn the AS group status.

To configure notification for an AS group, you need to specify a notification event and topic. You can configure up to five notifications for an AS group. The notification topic is pre-configured on the SMN console. When the live network

complies with the notification scenario that matches the notification topic, the AS group sends a notification to the user.

Procedure

1. Log in to the management console.
1. Under **Computing**, 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. On the AS group details page, click the **Notifications** tab and then click **Add Notification**.
3. Set the parameters listed in [Table 4-13](#).

Table 4-13 Parameter description

Parameter	Description	Example Value
Event	When at least one of the following conditions is met, SMN sends a notification to the user: <ul style="list-style-type: none"> • Instance creation succeeds • Instance removal succeeds • Errors occur in an AS group • Instance creation fails • Instance removal fails 	N/A
Topic	Select a created topic. For details about how to create a topic, see "Creating a Topic" in Simple Message Notification User Guide.	-

4. Click **OK**.

4.6.3 Recording AS Resource Operations

Scenarios

AS can use the 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 operations associated with AS for later query, audit, and backtrack operations.

Obtaining AS Information in CTS

After you enable CTS in the application system, the system logs the API calling operations performed on AS resources. On the **Cloud Trace Service** console, you can view operation records for the last 7 days. To obtain more operation records,

you can enable the Object Storage Service (OBS) and synchronize operation records to the OBS in real time.


Table 4-14 list the AS operations that can be recorded by CTS.


Table 4-14 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
Creating an AS configuration	scaling_configuration	createScalingConfiguration
Deleting an AS configuration	scaling_configuration	deleteScalingConfiguration
Deleting AS configurations in batches	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
Removing an instance	scaling_instance	removeInstance

Operation	Resource Type	Trace Name
Removing instances in batches	scaling_instance	batchRemoveInstances
Adding instances in batches	scaling_instance	batchAddInstances
Enabling instance protection in a batch	scaling_instance	batchProtectInstances
Disabling instance protection in a batch	scaling_instance	batchUnprotectInstances
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
Deleting a lifecycle hook	scaling_lifecycle_hook	deleteLifecycleHook

Viewing Audit Logs

1. Log in to the management console.
2. Click  in the upper left corner to select a region and a project.
3. Click **Service List**. Choose **Management & Deployment > Cloud Trace Service**.
4. Click **Trace List** in the navigation pane on the left.
5. You can use filters to query traces. The following filters are available:
 - **Trace Source, Resource Type, and Search By**
Select a filter criterion from the drop-down list.
When you select **Trace name** for **Search By**, you need to select a specific trace name.
When you select **Resource ID** for **Search By**, you need to select or enter a specific resource ID.
When you select **Resource name** for **Search By**, you need to select or enter a specific resource name.

- **Operator:** Select a specific operator (at user level rather than tenant level).
 - **Trace Status:** Available options include **All trace statuses**, **normal**, **warning**, and **incident**. You can only select one of them.
 - **Time range:** In the upper right corner of the page, you can query traces in the last one hour, last one day, last one week, or within a customized period of time.
6. Click  to the left of the required trace to expand its details.
 7. Locate the required trace and click **View Trace** in the **Operation** column. A dialog box is displayed, showing the trace content.

CTS Log Entries

Each log entry consists of a trace in JSON format. A log entry indicates an AS API request, including the requested operation, the operation date and time, operation parameters, and information about the user who sends the request. The user information is obtained from the Identity and Access Management (IAM) service.

The following example shows CTS log entries for the **CreateScalingPolicy** action:

```
{
  "time": "2016-12-15 15:27:40 GMT+08:00",
  "user": {
    "name": "xxxx",
    "id": "62ff83d2920e4d3d917e6fa5e31ddebe",
    "domain": {
      "name": "xxx",
      "id": "30274282b09749adbe7d9cabeebcbe8b"
    }
  },
  "request": {
    "scaling_policy_name": "as-policy-oonb",
    "scaling_policy_action": {
      "operation": "ADD",
      "instance_number": 1
    },
    "cool_down_time": "",
    "scheduled_policy": {
      "launch_time": "2016-12-16T07:27Z"
    },
    "scaling_policy_type": "SCHEDULED",
    "scaling_group_id": "ec4051a7-6fbd-42d2-840f-2ad8cdabee34"
  },
  "response": {
    "scaling_policy_id": "6a38d488-664b-437a-ade2-dc45f96f7f4c"
  },
  "code": 200,
  "service_type": "AS",
  "resource_type": "scaling_policy",
  "resource_name": "as-policy-oonb",
  "resource_id": "6a38d488-664b-437a-ade2-dc45f96f7f4c",
  "source_ip": "10.190.205.233",
  "trace_name": "createScalingPolicy",
  "trace_rating": "normal",
  "trace_type": "ConsoleAction",
  "api_version": "1.0",
  "record_time": "2016-12-15 15:27:40 GMT+08:00",
  "trace_id": "f627062b-c297-11e6-a606-eb2c0f48bec5"
}
```

4.6.4 Adding Tags to AS Groups and Instances

Scenarios

If you have many resources of the same type, you can use a tag 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 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:

- 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 a Tag to an AS Group

1. Log in to the management console.
2. Under **Computing**, 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 4-15](#).

Table 4-15 Tag naming rules

Parameter	Requirement	Example Value
Tag Key	<ul style="list-style-type: none"> The value cannot be empty. An AS group has a unique key. A key can contain a maximum of 36 characters, including digits, letters, underscores (_), hyphens (-), and Unicode characters from \u4e00 to \u9fff. 	Organization
Tag Value	<ul style="list-style-type: none"> The value can be an empty character string. A key can have only one value. 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 **Computing**, 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 4-15](#).

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

4.6.5 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 generated by AS.

Namespace

SYS.AS

Monitoring metrics

[Table 4-16](#) lists the AS metrics supported by Cloud Eye.

Table 4-16 AS metrics

Metric ID	Metric	Description	Value Range	Monitored Object	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 Unit: Percent	≥0%	AS group	5 minutes
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 Unit: Percent NOTE This metric is unavailable if the image has no VM Tools installed.	≥0%	AS group	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	AS group	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 rates of all ECS instances in an AS group/Number of ECS instances in the AS group Unit: Byte/s	≥0 Byte/s	AS group	5 minutes

Metric ID	Metric	Description	Value Range	Monitored Object	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 rates of all ECS instances in an AS group/Number of ECS instances in the AS group Unit: Byte/s	≥0 Byte/s	AS group	5 minutes
disk_read_bytes_rate	Disks Read Rate	Number of bytes read from an AS group per second Formula: Total disks read rates of all ECS instances in an AS group/Number of ECS instances in the AS group Unit: Byte/s	≥0 Byte/s	AS group	5 minutes
disk_write_bytes_rate	Disks Write Rate	Number of bytes written to an AS group per second Formula: Total disks write rates of all ECS instances in an AS group/Number of ECS instances in the AS group Unit: Byte/s	≥0 Byte/s	AS group	5 minutes
disk_read_requests_rate	Disks Read Requests	Number of read requests per second sent to an ECS disk in an AS group Formula: Total disks read rates of all ECS instances in an AS group/Number of ECS instances in the AS group Unit: Request/s	≥0 request/s	AS group	5 minutes

Metric ID	Metric	Description	Value Range	Monitored Object	Monitoring Interval (Raw Data)
disk_write_requests_rate	Disks Write Requests	Number of write requests per second sent to an ECS disk in an AS group Formula: Total disks write rates of all ECS instances in an AS group/Number of ECS instances in the AS group Unit: Request/s	≥ 0 request/s	AS group	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 Unit: Percent	0-100 %	AS group	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 Unit: Percent	0-100 %	AS group	1 minute
load_ave rage1	(Agent) 1-Minute Load Average	Average CPU load of all ECSs in an AS group in the last 1 minute Unit: none	≥ 0	AS group	1 minute
load_ave rage5	(Agent) 5-Minute Load Average	Average CPU load of all ECSs in an AS group in the last 5 minutes Unit: none	≥ 0	AS group	1 minute

Metric ID	Metric	Description	Value Range	Monitored Object	Monitoring Interval (Raw Data)
load_ave rage15	(Agent) 15- Minute Load Average	Average CPU load of all ECSs in an AS group in the last 15 minutes Unit: none	≥0	AS group	1 minute
gpu_usa ge_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 Unit: Percent	0-100 %	AS group	1 minute
gpu_usa ge_mem	(Agent) Video 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 Unit: Percent	0-100 %	AS group	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.

Dimension

Key	Value
AutoScalingGroup	AS group ID

4.6.6 Viewing Monitoring Metrics

Scenarios

The cloud platform provides Cloud Eye to help you obtain the running 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 instance. After such an instance restarts or recovers, the monitoring data is available.


Viewing Monitoring Metrics on Auto Scaling


1. Log in to the management console.
2. On the **AS Groups** page, find the AS group to view monitoring data and click its name.
3. 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 & Deployment**, 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.

4.6.7 Setting Monitoring Alarm Rules

Scenarios

Setting alarm rules allows you to customize the monitored objects and notification policies and determine the running status of your ECS instances at any time.

Procedure

1. Log in to the management console.
2. Under **Management & Deployment**, select **Cloud Eye**.
3. In the navigation pane, choose **Alarm Management > Alarm Rules**.
4. 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.
5. 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.

4.7 Permissions Management

4.7.1 Creating a User and Granting AS Permissions

Scenarios

This section describes how to use IAM to implement fine-grained permissions control for 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 only the permissions required for users to perform a specific task.
- Use to entrust an account or cloud service to perform efficient O&M on your AS resources.

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

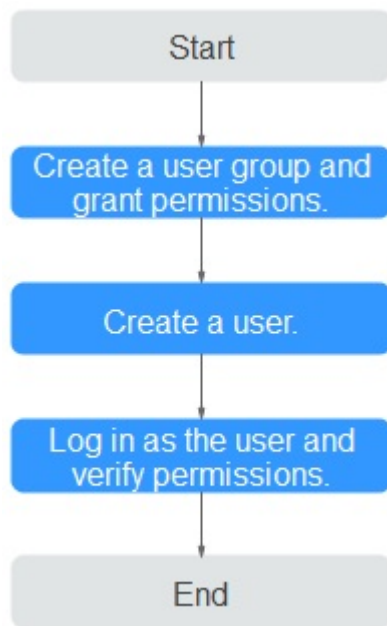
This section describes the procedure for granting permissions (see [Figure 4-3](#)).

Prerequisites

Learn about the permissions (see [Permissions Management](#)) supported by AS and choose policies or roles according to your requirements. For the permissions of other services, see Permission Description.

Process Flow

Figure 4-3 Process for granting AS permissions



1. Create a user group and assign permissions to it.
Create a user group on the IAM console, and attach the **AutoScaling ReadOnlyAccess** policy 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 AS console as the created user, and verify the user's permissions for AS.
 - 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 **AutoScaling ReadOnlyAccess** policy has already taken effect.
 - Choose any other service in the **Service List**. If a message appears indicating that you have insufficient permissions to access the service, the **AutoScaling ReadOnlyAccess** policy has already taken effect.

4.7.2 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" in *Auto Scaling API Reference*.

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: Allowing users to remove instances from an AS group and create AS configurations

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

- Example 2: Denying AS group deletion

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

The following method can be used if you need to assign permissions of the **AutoScaling FullAccess** policy to a user but you want to prevent the user from deleting AS groups. Create a custom policy for denying AS group deletion, and attach both policies to the group to which the user belongs. Then, the user can perform all operations on AS except deleting AS groups. The following is an example of a deny policy:

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

5 FAQs

5.1 General

5.1.1 What Are Restrictions on Using AS?

Only applications that are stateless and horizontally scalable can run on instances in an AS group. ECS instances in an AS group can be released automatically by AS, so they cannot be used to save application status information (such as session statuses) or related data (such as database data and logs).

If the application status or related data must be saved, you can store the information on separate servers.

[Table 5-1](#) lists the AS quotas.

Table 5-1 Quotas

Item	Description	Default
AS group	Maximum number of AS groups per region per account	10
AS configuration	Maximum number of AS configurations per region per account	100
AS policy	Maximum number of AS policies per AS group	10
Instance	Maximum number of instances per AS group	300
Bandwidth scaling policy	Maximum number of bandwidth scaling policies per region per account	10

5.1.2 Must I Use AS Together With ELB and Cloud Eye?

AS can work independently or in conjunction with ELB and Cloud Eye.

Cloud Eye does not require additional fees and is enabled by default. ELB is not enabled by default, but you can enable it manually if needed, for example, if distributed clusters are required.

5.1.3 Is AS Billed?

AS is free, but you will be billed for pay-per-use instances it automatically creates. EIPs used by the instances are also billed. When the capacity of an AS group is reduced, the automatically created instances will be removed from the AS group and deleted. After they are deleted, they no longer generate any fees. Instances manually added, however, are only removed from the AS group. They are not deleted, so they are still be billed. If you do not need these instances, unsubscribe from them on the ECS console.

For example, if two instances are created when an AS group scales out, but then an hour later, the AS group scales back in, the two instances are removed from the AS group and deleted, and you will be billed for that one hour of use.

5.1.4 Will an Abrupt Change in Monitoring Metric Values Trigger an Unnecessary Scaling Action?

No. Monitoring data used by AS is from Cloud Eye. The monitoring period can be set to 5 minutes, 20 minutes, or 1 hour, so an abrupt change in monitoring metric values will not impact scaling actions.

In addition, AS allows you to configure a cooldown period to prevent unnecessary scaling actions caused by frequently reported alarms. You can customize the cooldown period as needed.

5.1.5 How Many AS Policies and AS Configurations Can I Create and Use?

You can create up to 10 AS groups and 100 AS configurations by default. An AS group can use 1 AS configuration and 10 AS policies at a time.

If the default quotas do not meet your service requirements, contact the customer service.

5.1.6 Can AS Automatically Scale Up or Down vCPUs, Memory, and Bandwidth of ECSs?

Currently, AS can automatically scale up or down bandwidths, and scale in or out ECS instances in an AS group.

5.1.7 What Is an AS Quota?



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, click  .
The **Service Quota** 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.

5.1.8 How Do I Fix the Error "The key pair does not exist" When I Connect to an Instance?

A key pair is specific to each user. If the key pair of a user who is under the same account as you is configured for an AS configuration, you cannot connect the instances scaled out using that AS configuration.

If you want to connect to these instances without being restricted by the key pair permission, password authentication needs to be configured as the login mode.

5.2 AS Group

5.2.1 What Can I Do If the AS Group Fails to Be Enabled?

The possible causes and solutions are as follows:

- Issue description: Insufficient ECS, EVS disk or EIP quota.
Possible cause: insufficient quota
Handling method: Increase the quota or delete unnecessary resources, and then enable the AS group.
- Issue description: The VPC or subnet does not exist.
Possible cause: The VPC service encounters an exception or resources have been deleted.
Handling method: Wait until the VPC service recovers, or modify parameters of the VPC and subnet in the AS group, and then enable the AS group.
- Issue description: The ELB listener or backend ECS group does not exist, and the load balancer is unavailable.
Possible cause: The ELB service encounters an exception or resources have been deleted.
Handling method: Wait until the ELB service recovers, or modify load balance parameters in the AS group, and then enable the AS group.

- Issue description: The number of backend ECSs that you add to the ELB listener exceeds the upper limit.
Possible cause: If classical load balancer is used by an AS group, instances added to the AS group are automatically added to the ELB listener. A maximum of 300 backend ECSs can be added to an ELB listener.
Handling method: Remove the backend ECSs that are both not required and not in the AS group from the listener. Then enable the AS group.
- Issue description: The image used by the AS configuration, the flavor, or the key pair does not exist.
Possible cause: Resources have been deleted.
Handling method: Change the AS configuration for the AS group and then enable the AS group.
- Issue description: The notification subject configured for your lifecycle hook does not exist.
Possible cause: The AS group adds a lifecycle hook, while its configured notification subject has been deleted before the scaling action starts. If the notification subject is deleted after the scaling action starts, an AS group exception will occur in the next scaling action.
Handling method: Change the notification subject used by the lifecycle hook or delete the lifecycle hook. Then enable the AS group.
- Issue description: The subnet you select does not have enough private IP addresses.
Possible cause: Private IP addresses in the subnet used by the AS group have been used up.
Handling method: Modify the subnet information and enable the AS group.
- Issue description: The ECS resources of this type in the selected AZ have been sold out.
Possible cause: ECSs of this type have been sold out or are not supported in the AZ selected for the AS group. ECSs of this type are the ECS flavor selected in the AS configuration.
Handling method: Change the AS configuration for the AS group and then enable the AS group. If there is no instance in the AS group, you can also change the AZ for the AS group and then enable the AS group.
- Issue description: The selected specifications and the disk do not match.
Possible cause: The ECS type in the AS configuration does not match the disk type, leading to the ECS creation failure.
Handling method: Change the AS configuration for the AS group and then enable the AS group.
- Issue description: The selected specifications and the image do not match.
Possible cause: The ECS type in the AS configuration does not match the image, leading to the ECS creation failure.
Handling method: Change the AS configuration for the AS group and then enable the AS group.
- Issue description: Storage resources of this type have been sold out in the selected AZ.

Possible cause: Storage resources of this type have been sold out or are not supported in the AZ selected for the AS group. Storage resources of this type refer to the system and data disk types selected for the AS configuration.

Handling method: Change the AS configuration for the AS group and then enable the AS group. If there is no instance in the AS group, you can also change the AZ for the AS group and then enable the AS group.

- Issue description: The shared bandwidth defined in the AS configuration does not exist.

Possible cause: Resources have been deleted.

Handling method: Use a newly purchased or an existing shared bandwidth to create an AS configuration. Then change the AS configuration for the AS group and enable the AS group.

- Issue description: The number of EIPs bound to the shared bandwidth specified in the AS configuration exceeds the limit.

Possible cause: A maximum of 20 EIPs can be bounded to a shared bandwidth.

Handling method: Apply for a higher EIP quota, remove unnecessary EIPs from the shared bandwidth, or change another AS configuration for the AS group. Then enable the AS group.

- Issue description: The DeH selected in your AS configuration does not exist. Change the AS configuration.

Possible cause: Resources have been deleted.

Handling method: Use a newly purchased or an existing DeH to create an AS configuration. Then change the AS configuration for the AS group and enable the AS group.

- Issue description: No DeH is available. Ensure that there are available DeH resources.

Handling method: Rectify the DeH fault and restore the DeH to the available state, or enable the automatic placement attribute for the DeH, and then enable the AS group again. You can also use a newly purchased DeH to create an AS configuration, change the AS configuration for the AS group, and enable the AS group.

- Issue description: The DeH selected in your AS configuration does not have sufficient capacity.

Handling method: You can delete unnecessary ECSs from the DeH and enable the AS group again. You can also use a newly purchased DeH to create an AS configuration, change the AS configuration for the AS group, and enable the AS group.

- Issue description: No DeH is available in the AZ selected for your AS group.

Handling method: Purchase a DeH in the AZ, use it to create an AS configuration, change the AS configuration for the AS group, and enable the AS group. If there is no instance in the AS group, change the AZ for the AS group and then enable the AS group.

- Issue description: The DeH selected in your AS configuration does not support this type of ECS. Change the AS configuration.

Handling method: Select the ECS specifications supported by the DeH, create an AS configuration, change the AS configuration for the AS group, and then enable the AS group again.

- Issue description: A system error has occurred.
Possible cause: An error has occurred in the AS service, peripheral service, or network.
Handling method: Try again later or contact technical support.
- Issue description: The specification defined in the AS configuration is unavailable.
Handling method: Change specifications by creating an AS configuration as prompted by the error message and use this AS configuration for the AS group. Then enable the AS group.
- Issue description: The selected AS configuration cannot be used by the AS group.
Handling method: Create an AS configuration as prompted by the error message and use this AS configuration for the AS group. Then enable the AS group.
- Issue description: AS group expansion fails.
Possible cause: Your account is in arrears or the balance is insufficient.
Handling method: Top up your account and enable the AS group.

5.2.2 How Can I Handle an AS Group Exception?

The handling method depends on the reported possible cause.

- Issue description: Insufficient ECS, EVS disk or EIP quota.
Possible cause: insufficient quota
Handling method: Increase the quota or delete unnecessary resources, and then enable the AS group.
- Issue description: The VPC or subnet does not exist.
Possible cause: The VPC service encounters an exception or resources have been deleted.
Handling method: Wait until the VPC service recovers, or modify parameters of the VPC and subnet in the AS group, and then enable the AS group.
- Issue description: The ELB listener or backend ECS group does not exist, and the load balancer is unavailable.
Possible cause: The ELB service encounters an exception or resources have been deleted.
Handling method: Wait until the ELB service recovers, or modify load balance parameters in the AS group, and then enable the AS group.
- Issue description: The number of backend ECSs that you add to the ELB listener exceeds the upper limit.
Possible cause: If classical load balancer is used by an AS group, instances added to the AS group are automatically added to the ELB listener. A maximum of 300 backend ECSs can be added to an ELB listener.
Handling method: Remove the backend ECSs that are both not required and not in the AS group from the listener. Then enable the AS group.
- Issue description: The image used by the AS configuration, the flavor, or the key pair does not exist.

Possible cause: Resources have been deleted.

Handling method: Change the AS configuration for the AS group and then enable the AS group.

- Issue description: The notification subject configured for your lifecycle hook does not exist.

Possible cause: The AS group adds a lifecycle hook, while its configured notification subject has been deleted before the scaling action starts. If the notification subject is deleted after the scaling action starts, an AS group exception will occur in the next scaling action.

Handling method: Change the notification subject used by the lifecycle hook or delete the lifecycle hook. Then enable the AS group.

- Issue description: The subnet you select does not have enough private IP addresses.

Possible cause: Private IP addresses in the subnet used by the AS group have been used up.

Handling method: Modify the subnet information and enable the AS group.

- Issue description: The ECS resources of this type in the selected AZ have been sold out.

Possible cause: ECSs of this type have been sold out or are not supported in the AZ selected for the AS group. ECSs of this type are the ECS flavor selected in the AS configuration.

Handling method: Change the AS configuration for the AS group and then enable the AS group. If there is no instance in the AS group, you can also change the AZ for the AS group and then enable the AS group.

- Issue description: The selected specifications and the disk do not match.

Possible cause: The ECS type in the AS configuration does not match the disk type, leading to the ECS creation failure.

Handling method: Change the AS configuration for the AS group and then enable the AS group.

- Issue description: The selected specifications and the image do not match.

Possible cause: The ECS type in the AS configuration does not match the image, leading to the ECS creation failure.

Handling method: Change the AS configuration for the AS group and then enable the AS group.

- Issue description: Storage resources of this type have been sold out in the selected AZ.

Possible cause: Storage resources of this type have been sold out or are not supported in the AZ selected for the AS group. Storage resources of this type refer to the system and data disk types selected for the AS configuration.

Handling method: Change the AS configuration for the AS group and then enable the AS group. If there is no instance in the AS group, you can also change the AZ for the AS group and then enable the AS group.

- Issue description: The shared bandwidth defined in the AS configuration does not exist.

Possible cause: Resources have been deleted.

Handling method: Use a newly purchased or an existing shared bandwidth to create an AS configuration. Then change the AS configuration for the AS group and enable the AS group.

- Issue description: The number of EIPs bound to the shared bandwidth specified in the AS configuration exceeds the limit.

Possible cause: A maximum of 20 EIPs can be bounded to a shared bandwidth.

Handling method: Apply for a higher EIP quota, remove unnecessary EIPs from the shared bandwidth, or change another AS configuration for the AS group. Then enable the AS group.

- Issue description: The DeH selected in your AS configuration does not exist. Change the AS configuration.

Possible cause: Resources have been deleted.

Handling method: Use a newly purchased or an existing DeH to create an AS configuration. Then change the AS configuration for the AS group and enable the AS group.

- Issue description: No DeH is available. Ensure that there are available DeH resources.

Handling method: Rectify the DeH fault and restore the DeH to the available state, or enable the automatic placement attribute for the DeH, and then enable the AS group again. You can also use a newly purchased DeH to create an AS configuration, change the AS configuration for the AS group, and enable the AS group.

- Issue description: The DeH selected in your AS configuration does not have sufficient capacity.

Handling method: You can delete unnecessary ECSs from the DeH and enable the AS group again. You can also use a newly purchased DeH to create an AS configuration, change the AS configuration for the AS group, and enable the AS group.

- Issue description: No DeH is available in the AZ selected for your AS group.

Handling method: Purchase a DeH in the AZ, use it to create an AS configuration, change the AS configuration for the AS group, and enable the AS group. If there is no instance in the AS group, change the AZ for the AS group and then enable the AS group.

- Issue description: The DeH selected in your AS configuration does not support this type of ECS. Change the AS configuration.

Handling method: Select the ECS specifications supported by the DeH, create an AS configuration, change the AS configuration for the AS group, and then enable the AS group again.

- Issue description: A system error has occurred.

Possible cause: An error has occurred in the AS service, peripheral service, or network.

Handling method: Try again later or contact technical support.

- Issue description: The specification defined in the AS configuration is unavailable.

Handling method: Change specifications by creating an AS configuration as prompted by the error message and use this AS configuration for the AS group. Then enable the AS group.

- Issue description: The selected AS configuration cannot be used by the AS group.
Handling method: Create an AS configuration as prompted by the error message and use this AS configuration for the AS group. Then enable the AS group.
- Issue description: AS group expansion fails.
Possible cause: Your account is in arrears or the balance is insufficient.
Handling method: Top up your account and enable the AS group.

5.2.3 What Operations Will Be Suspended If an AS Group Is Disabled?

If an AS group is disabled, new scaling actions will not happen, but any scaling actions already in progress will continue. Scaling policies will not trigger any scaling actions. Even if you manually change the number of expected instances, no scaling action will be triggered even though the number of actual instances is not equal to that of expected instances.

Health checks continue to be performed but will not remove any instances.

5.2.4 Can I Use an ECS Instance ID to Learn What AS Group the Instance Is In?

No.

To obtain details about an AS group and the instances in the group, perform the following operations:

- Step 1** Log in to the management console. Choose **Compute > Auto Scaling > Instance Scaling**.
- Step 2** On the **AS Groups** tab page, click the name of the target AS group.
- Step 3** Click the **Instances** tab to view the instances in the AS group.

----End

5.3 AS Policy

5.3.1 How Many AS Policies Can I Enable?

You can enable one or more AS policies as required.

5.3.2 What Are the Conditions to Trigger an Alarm-based AS Policy?

Alarms will be triggered by metrics of CPU Usage, Memory Usage, Inband Incoming Rate, Inband Outgoing Rate, Disk Read Rate, Disk Write Rate, Disk Read Requests, and Disk Write Requests. These alarms will in turn trigger the policy to scale instances in or out in the AS group.

5.3.3 What Is a Cooldown Period and Why Is It Required?

A cooldown period specifies how long any alarm-triggered scaling action will be disallowed after a previous scaling action is complete. This cooldown period does not work for scheduled or periodic scaling actions.

Before an instance is put into use after it is added to the AS group, it takes 2 to 3 minutes to execute the configuration script to install and configure applications. The time varies depending on many factors, such as the instance specifications and startup scripts. If an instance is put into use without cooldown, the system will keep adding instances until the load decreases. As the new instances take over services, the system will detect that the load is too low and start removing instances from the AS group. A cooldown prevents the AS group from repeatedly triggering unnecessary scaling actions.

For example:

When a traffic peak occurs, an alarm policy is triggered and AS automatically adds an instance to the AS group to help handle the increased load. However, it takes time for the instance to start. After the instance is started, it takes time to receive requests from ELB. During this period, alarms may continue to be triggered and instances may continue to be added. If you set a cooldown time, after an instance is started, AS stops adding new instances in response to the alarms until the specified period of time (300 seconds by default) passes. That way the newly started instance has time to start processing application traffic. If an alarm is triggered again after the cooldown period elapses, AS starts another instance and the cooldown period starts up again.

5.3.4 Can AS Scale Capacity Based on Custom Monitoring by Cloud Eye?

Yes. AS can scale capacity based on custom monitoring by Cloud Eye.

5.3.5 What Monitoring Metrics for an AS Group Will Be Affected If VM Tools Are Not Installed on the Instances in the Group?

If VM Tools have not been installed on ECS instances, Cloud Eye can monitor metrics Outband Incoming Rate and Outband Outgoing Rate. However, it cannot monitor metrics Memory Usage, Inband Incoming Rate, and Inband Outgoing Rate, which reduces data accuracy of CPU usage.

If the ECS instances are I/O-optimized, Cloud Eye cannot monitor metrics Disk Usage, Inband Incoming Rate, and Inband Outgoing Rate of instances no matter VM Tools are installed or not.

If VM Tools are not installed on ECS instances, AS cannot obtain the memory usage, inband incoming rate, and inband outgoing rate.

5.3.6 What Can I Do If an AS Policy Fails to Be Enabled?

- Description: The alarm rule does not exist.
Possible cause: The alarm rule used in the alarm policy is deleted.

Handling method: Change the alarm rule used in the alarm policy and enable the AS policy again.

- Description: The triggering time of the periodic policy falls outside the effective time range of the policy.

Possible cause: The periodic policy has expired.

Handling method: Change the start time and end time of the periodic policy and enable the policy again.

- Description: The triggering time of the scheduled policy must be later than the current time.

Possible causes: The triggering time of the scheduled policy has expired.

Handling method: Change the triggering time of the scheduled policy and enable the policy again.

- Description: A system error has occurred.

Handling method: Try again later or contact technical support.

5.3.7 How Do I Install the Agent on Instances in an AS Group to Use Agent Monitoring Metrics?

Scenarios

If an AS group uses an alarm-based scaling policy, you can install the Agent on the instances to better monitor the AS group. Compared with basic monitoring, Agent monitoring (OS monitoring) provides proactive, fine-grained, system-level monitoring of instances. This section describes how to install the Agent on instances.

Procedure

1. Log in to the management console and click **Elastic Cloud Server** under **Computing**.
The **Elastic Cloud Server** page is displayed.
2. Create an ECS and install the Agent.
For more information, see Agent Installation and Configuration.
3. After installing the Agent plug-in, log in to the Cloud Eye console, choose **Server Monitoring** > **Elastic Cloud Server**, and ensure that the Agent is running and collecting the monitoring metrics.
4. Add the AK/SK pair to the **conf.json** configuration file of the ECS.
 - a. Click your username, choose **My Credentials** > **Access Keys**, and obtain the AK/SK pair.
 - If you have created an access key, obtain the **AccessKey** value and the **SecretKey** value in the **credentials.csv** file you obtained when you create the access key.
 - If no access keys are available, click **Create Access Key** to create one. Save the **credentials.csv** file and obtain the **AccessKey** value and the **SecretKey** value in it.
 - b. Log in to the ECS and use **cd /usr/local/telescope/bin** to go to the Agent installation directory.

- c. Run `vi conf.json` to open the configuration file and enter the obtained AK/SK pair.

```
{  
  "InstanceId": "fa07b[REDACTED]4cd0-9aa6-e5c791569e3a",  
  "ProjectId": "050b1[REDACTED]572f8cc01f3740bed5",  
  "AccessKey": "MK8NR3[REDACTED]7FUMJB",  
  "SecretKey": "sPHiTB8[REDACTED]N4wWw3YCNwcUFqj",  
  "RegionId": "cn-north-1"  
}
```

If the Agent was installed during ECS creation, the AK/SK pair was added during user data injection. You only need to check the AK/SK pair in this step.

- d. Press **Esc** and enter `:wq` to save the configuration and exit.
5. Go to the **Image Management Service** page and create a private image **ces-agent-test** for this ECS. For details, see [Creating a Private Image](#).
6. Go to the **Auto Scaling** page and use the private image created in step 5 to create an AS configuration.
Click **Private image** for **Image**, select **ces-agent-test** from the drop-down list, and set other parameters as required.
7. Use the AS configuration created in step 6 to create an AS group.
8. Add an AS policy for the AS group. Set **Policy Type** to **Alarm** and **Trigger Condition** to an Agent monitoring metric, such as **(Agent) Memory Usage**.
9. Manually add the ECS created in step 2 to the AS group.
10. Perform the following operations to check whether the Agent monitoring has taken effect:
 - Verify that the Agent monitoring metric is displayed on the **Monitoring** tab of the AS group details page.
 - When the alarm threshold is reached, verify that the alarm policy has triggered scaling actions and instance are added or removed.
 - The Agent monitoring data is collected from all instances that are automatically added to the AS group.

5.4 Instance

5.4.1 How Do I Prevent Instances Manually Added to an AS Group from Being Automatically Removed?

If you have manually added N instances into an AS group and do not want these instances to be removed automatically, you can use either of the following methods to do this:

Method 1

Perform following configurations in the AS group:

- Set the minimum number of instances in the AS group to N or a larger value.

- Set **Instance Removal Policy** to **Oldest instance created from oldest AS configuration** or **Newest instance created from oldest AS configuration**.

Based on the scaling rules, the manually added instances are not created based on the AS configuration used by the AS group. The instances automatically added using the AS configuration are removed at first. The manually added instances would not be removed until all of the automatically added instances have been removed first. Finally, since you have set the minimum number of instances to N or a larger value, those instances cannot be removed.

Note: If the instances manually added are stopped or if they malfunction, they are marked as unhealthy and removed from the AS group. This is because health checks ensure that all instances in the AS group are healthy.

Method 2

Enable instance protection for the N instances.

You can enable instance protection for these instances at the same time. When the AS group scales in, protected instances will not be removed from the AS group as long as they do not fail health checks. Instances that fail health check will be removed even if they are protected.

5.4.2 When There Are Multiple Flavors Included in the AS Configuration of an AS Group, What Order Are the Flavors Selected in During Scale-out Events?

The order depends on the AZs and multiple-AZ scaling policy used by the AS group. This section describes how flavors are selected during scale-out events in these the following two different scenarios:

Single AZ

If the AS group use only one AZ, new instances will be created in the AZ by following the configured flavor selection policy which determines the flavor priority order.

- **Sequenced:** In this policy, the flavors are used in the order they were selected. For example, if you selected flavors 2, 3, and 1, and in that sequence, the system attempts to use flavor 2 to create instances at first. If flavor 2 is not available in the current AZ or an instance fails to be created due to other reasons, the system follows up with flavor 3 and again with flavor 1.
- **Cost-center:** In this policy, the least expensive flavor is used first. For example, if you select flavors 1, 2, and 3 in sequence, where flavor 1 is the most expensive, followed by flavor 3, and flavor 2 is the least expensive, the system preferentially scales out the AS group by starting with flavor 2, then flavor 3, and then flavor 1.

Multiple AZs

If the AS group spans two or more AZs, you need to configure the multi-AZ extension policy. The system will scale out the AS group by following the scaling policy you configured using the flavor with the highest priority.

- **Balanced:** In this policy, the system preferentially ensures that instances are evenly distributed across the AZs specified in the AS group. If it fails in the target AZ, the system selects another AZ based on the sequenced policy. For example, if:
 - You selected AZ 1, AZ 2, and AZ 3 in sequence.
 - You selected flavors 1, 2, and 3, where the flavor priority order is 2, 3, and 1.
 - AZ 1, AZ 2, and AZ 3 have, respectively, 3, 2, and 3 instances.

AZ 2 has fewer instances. The system will try to create instances in AZ 2 first using flavor 2 (the one with the highest priority). If this attempt succeeds, the scaling action is complete. If this attempt fails, the system will try to create instances in AZ 2 using flavor 3, and then flavor 1. If all attempts fail in AZ 2, the system will try to fulfill the required capacity by starting with flavor 2, then flavor 3, and then flavor 1 in the AZ 1 based on the Sequenced policy. If all attempt fails in AZ 1, the system will try in AZ 3.

- **Sequenced:** In this policy, the target AZ is chosen based on the order in which the AZs are selected. For example, if:
 - You have selected AZ 1, AZ 2, and AZ 3 in sequence.
 - You selected flavors 1, 2, and 3, where the flavor priority order is 2, 3, and 1.

Regardless of how instances are distributed in three AZs, the system will try to use flavor 2 to create instances in AZ 1. If this attempt fails, the system will try flavor 3 and then flavor 1 in AZ 1. If all attempts fail in AZ 1, the system will do that again in AZ 2, and then AZ 3.

 **NOTE**

The flavor priority order is determined by the flavor selection policy in the AS configuration. For details, see [Single AZ](#).

5.4.3 When an Instance Is Removed from an AS Group and Deleted, Is the Application Data Saved?

No. You must ensure that instances in the AS group do not store application status information or other important data, such as sessions, databases, and logs, or the data will be lost when AS automatically releases them. If you want to store your application status, you can store it on an independent server (such as an ECS) or database (such as an RDS database).

5.4.4 Can I Add Yearly/Monthly ECS Instances to an AS Group?

Yes. Currently, AS automatically creates pay-per-use ECS instances by default, but you can also add yearly/monthly or pay-per-use instances manually if needed.

5.4.5 Can the System Automatically Delete the Instances Added Based on an AS Policy When They Are Not Required?

Yes. They can be automatically deleted if an AS policy has been added to trigger scaling actions to delete the instances.

5.4.6 What Is the Expected Number of Instances?

The expected number of instances refers to the number of ECS instances that are expected to run in an AS group. It is between the minimum number of instances and the maximum number of instances. You can manually change the expected number of instances or change it based on the scheduled, periodic, or alarm policies.

You can set this parameter when creating an AS group. If this value is greater than 0, a scaling action is performed to add the required number of instances after the AS group is created. You can also change this value manually or by scaling policies after the AS group is created.

If you manually change this value, the current number of instances will be inconsistent with the expected number, and a scaling action will be performed to bring the number of instances in line with the expected number.

If a scaling policy is triggered to add two instances to an AS group, the system will increase the expected number of instances by 2. Then, a scaling action is performed to add two instances so that the number of instances in the AS group is the same as the expected number.

5.4.7 How Do I Delete an ECS Instance Created in a Scaling Action?

Handling Methods

Method 1

1. Log in to the management console.
2. Under **Computing**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
3. Click the AS group name on the **AS Groups** page.
4. On the AS group details page, click the **Instances** tab.
5. Locate the row that contains the instance and click **Remove and Delete** in the **Operation** column.

NOTE

To delete multiple instances, select the check boxes in front of them and click **Remove and Delete**.

Method 2

1. Log in to the management console.
2. Under **Computing**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
3. Click the AS group name on the **AS Groups** page.
4. On the AS group details page, click the **AS Policies** tab.
5. Click **Add AS Policy**. In the displayed **Add AS Policy** dialog box, add an as policy to remove instances as needed or maintain a specified number of instances.

Method 3

1. Log in to the management console.
1. Under **Computing**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
2. Click the AS group name on the **AS Groups** page.
3. On the AS group details page, click **Modify** in the upper right corner.
4. In the displayed **Modify AS Group** dialog box, change the value of **Expected Instances**.

5.4.8 Will a Yearly/Monthly ECS Instance Be Deleted If It Becomes Unhealthy?

No. If a yearly/monthly ECS becomes unhealthy, it will be removed from the AS group, but will not be deleted.

5.4.9 How Do I Handle Unhealthy Instances in an AS Group?

Normally, you do not need to handle unhealthy instances because AS periodically checks the health status of instances in an AS group. When an AS group is enabled, unhealthy instances are removed and new instances are created to ensure that the expected number of instances are running in the AS group. When an AS group is disabled, AS keeps performing health checks on instances, but does not remove instances.

It should be noted that if ELB health check is selected, ELB sends heartbeat messages to instances through an intranet. To ensure that the ELB health check can be performed properly, ensure that your instances can be accessed through that intranet. To check this, perform the following steps:

1. In the **Listener** area, locate the row containing the target listener and click **View** in the **Health Check** column. A dialog box is displayed.
 - **Health Check Protocol:** Ensure that the protocol has been configured and port has been enabled for the ECS instance to be checked.
 - **Check Path:** If HTTP is used for the health check, ensure that the health check path for the instance is correct.
2. Confirm that there is no software such as firewall on the instance blocking the source IP address used for performing the health check.
3. Confirm that the rules of instance security groups and network ACL allow access from 100.125.0.0/16, and configure the protocol and port used for health check. Obtain the health check protocol and port from the dialog box displayed in step 1.
 - If the default type of health check is used, service ports of the instances must be enabled.
 - If the health check port is different from service ports of the instances, communication between the service ports and health check port must be enabled.
4. If the issue persists, contact technical support.

5.4.10 Why Instances in an AS Group Keep Failing Health Checks and Getting Deleted and Recreated?

The rules of security group that the instances are in must allow access from the 100.125.0.0/16 network segment over the protocol and port used by ELB for health checks, or the health checks will fail. As a result, the instances will be deleted and created again and again.

5.4.11 How Do I Prevent Instances from Being Automatically Removed from an AS Group?

You can enable instance protection for in-service instances in an AS group. After the configuration, the protected in-service instances will not be removed during scale-in events. You can also modify the minimum number of instances for an AS group and use an instance removal policy to ensure that the AS group always has some in-service instances.

Unhealthy instances are removed from an AS group and new instances are created automatically. Do not stop or delete instances that have been added to an AS group on the ECS console as they will be marked as unhealthy and automatically removed from the AS group. Even when an AS group is disabled, AS still checks the status of instances in the AS group, but does not remove unhealthy instances.

5.4.12 Why Is an Instance that Was Removed from an AS Group and Deleted Still Displayed in the ECS List?

If an automatically added instance is protected, it is removed out of the AS group but not deleted, so that it can still be used by other services.

An instance that is being used by other services are protected generally. For example, an instance is used by IMS for creating a private image, or used by SDRS.

5.5 Others

5.5.1 How Can I Automatically Deploy My Application on an Instance?

To enable automatic application deployment on instances automatically added to an AS group, create a private image with the application preinstalled and automatic startup settings preconfigured. Create an AS configuration with the private image, and then change the AS configuration used by the AS group to the one you created. Your application will be automatically deployed on instances that are automatically added to the AS group. The procedure is as follows:

1. Install the application on the instance you will use to create a private image, and configure the application to automatically start at boot.
2. Create a private image using the instance. For details, see *Image Management Service User Guide*.

3. Create an AS configuration using the private image. For details, see [Creating an AS Configuration](#). Ensure that the image in the AS configuration is the private image created in [2](#).
4. Go to the page that shows the details about your AS group.
5. Click **Change Configuration** to the right of **Configuration Name**. In the displayed dialog box, select the AS configuration created in [3](#) and click **OK**.

After new instances are added to the AS group in the next scaling action, you can check whether your application has been installed on the instances. If you encounter any problems, contact technical support.

5.5.2 How Does Cloud-Init Affect the AS Service?

Cloud-Init is an open-source cloud initialization program, which initializes specified custom configurations, such as the hostname, key pair, and user data, of a newly created ECS. When creating an AS configuration, you can choose an image with Cloud-Init or Cloudbase-Init preinstalled for ECS initialization.

If Cloud-Init or Cloudbase-Init is not installed in the private image specified in the AS configuration of an AS group, the following issues can occur on the ECSs created in a scaling action:

- On a Windows image, the system will display a message indicating that the password for logging in to the ECS could not be viewed. In such a case, you can log in to the ECS using the image password. If you forgot the image password, use the password resetting function available on the **Elastic Cloud Server** page to reset the password.
- On a Linux image, the ECS cannot be logged in using the password or key pair configured during ECS creation. In such a case, you can log in to the ECS only using the image password or key pair. If you forgot the image password or key, use the password resetting function available on the **Elastic Cloud Server** page to reset the password.
- On a private image, user data injection fails.

To avoid these issues, confirm that the private image specified in the AS configuration has Cloud-Init or Cloudbase-Init installed. If the program was not installed, use a private image with the program installed to create an AS configuration, and replace the AS configuration of the AS group with the newly created one. The procedure is as follows:

- a. Log in to the management console.
- b. Under **Computing**, click **Auto Scaling**. In the navigation pane on the left, choose **Instance Scaling**.
- c. Click the **AS Configurations** tab.
- d. Click **Create AS Configuration** and select a private image with Cloud-Init or Cloudbase-Init installed to create a desired AS configuration.
- e. Change the AS configuration of the AS group to the newly created one.

5.5.3 Why Can't I Use a Key File to Log In to an ECS?

Issue Description

When I used a key file to attempt to log in to an instance in an AS group, the login failed.

Possible Causes

The image specified in the AS configuration of the AS group is a private image, on which Cloud-Init has not been installed.

In this case, it would not be possible to customize the ECS configuration. As a result, you can log in to the ECS only using the original image password or key pair.

Handling Method

1. Check whether the ECS needs to be logged in to.
 - If yes, use the original image password or key pair to log in to this ECS. The original image password or key pair is the OS password or key pair configured when the private image was created.
 - If no, go to step 2.
2. Change the AS configuration for the AS group. For details, see [Changing the AS Configuration for an AS Group](#).

NOTE

Make sure that Cloud-Init or Cloudbase-Init has been installed on the image specified in the new AS configuration. For how to install Cloud-Init or Cloudbase-Init, see *Image Management Service User Guide*.

After the AS configuration is changed, you can use the key file to log in to the new ECSs that are added to the AS group during scaling actions. You do not need to use the original image password or key pair to log in to these new ECSs anymore.

5.5.4 Do I Need to Configure an EIP in an AS Configuration When a Load Balancer Has Been Enabled for an AS Group?

No. If you have enabled a load balancer for an AS group, you do not have to configure an EIP in the AS configuration. The system automatically associates instances in the AS group to the load balancer. These instances will provide services via the EIP bound to the load balancer.

5.5.5 How Do I Enable Automatic Initialization of EVS Disks on Instances that Have Been Added to an AS Group During Scaling Actions?

Scenarios

After an ECS instance is created, you need to manually initialize EVS disks attached to the instance before using them. If multiple instances are added to the AS group, you must initialize the EVS disks on each instance, which takes a while.

This section describes how to configure a script to enable automatic initialization of EVS disks, including disk partitioning and attachment of specified directories. The script can only be used to initialize one EVS disk.

This section uses CentOS 6.5 as an example. For how to configure automatic initialization of EVS disks on other OSs, see the relevant OS documentation.

Procedure

1. Log in to the instance as user **root**.
2. Run a command to switch to the directory where the script will be stored:

```
cd /script directory
```

For example:

```
cd /home
```

3. Run the following command to create the script:

```
vi script name
```

For example:

```
vi fdisk_mount.sh
```

4. Press **i** to enter editing mode.

The following script is used as an example to show how to implement automatic initialization of one data disk:

```
#!/bin/bash
bash_scripts_name=fdisk_mount.sh
ini_path=/home/fdisk.ini
disk=
size=
mount=
partition=

function get_disk_from_ini()
{
disk=`cat $ini_path|grep disk| awk -F '=' '{print $2}'`
if [ $disk = "" ]
then
echo "disk is null in file,exit"
exit
fi
result=`fdisk -l $disk | grep $disk`
if [ $result = 1 ]
then
echo "disk path is not exist in linux,exit"
exit
fi
}

function get_size()
{
size=`cat $ini_path| grep size|awk -F '=' '{print $2}'`
if [ $size = "" ]
then
echo "size is null,exit"
exit
fi
}

function make_fs_mount()
{
mkfs.ext4 -T largefile $partition
if [ $? -ne 0 ]
then
echo "mkfs disk failed,exit"
exit
fi
}
```

```
dir=`cat $ini_path|grep mount |awk -F '=' '{print $2}'`
if [ $dir = "" ]
then
    echo "mount dir is null in file,exit"
    exit
fi

if [ ! -d $dir ]
then
    mkdir -p $dir
fi

mount $partition $dir
if [ $? -ne 0 ]
then
    echo "mount disk failed,exit"
    exit
fi

echo "$partition $dir ext3 defaults 0 0" >> /etc/fstab
}

function remove_rc()
{
    cat /etc/rc.local | grep $bash_scripts_name
    if [ $? ne 0 ]
    then
        sed -i '/'$bash_scripts_name'/d' /etc/rc.local
    fi
}

##### start #####
##1. Check whether the configuration file exists.
if [ ! -f $ini_path ]
then
    echo "ini file not exist,exit"
    exit
fi

##2. Obtain the device path for the specified disk from the configuration file.
get_disk_from_ini

##3. Obtain the size of the size partition from the configuration file.
get_size

##4. Partition the disk.
fdisk $disk <<EOF
n
p
1
1
$size
w
EOF
partition=`fdisk -l $disk 2>/dev/null | grep "^/dev/[xsh].*d" | awk '{print $1}'`

##5. Format the partition and attach the partition to the specified directory.
make_fs_mount

##6. Change startup items to prevent re-execution of the scripts.
remove_rc

echo 'SUCESS'
```

5. Press **Esc**, enter **:wq**, and press **Enter** to save the changes and exit.
6. Run the following command to create the configuration file:
vi fdisk.ini

7. Press **i** to enter editing mode.

The drive letter, size, and mount directory of the EVS disk are configured in the configuration file. You can change the settings based on the following displayed information.

```
disk=/dev/xdev  
size=+100G  
mount=/opt/test
```

8. Press **Esc**, enter **:wq**, and press **Enter** to save the changes and exit.

9. Run the following command to open configuration file **rc.local**:

```
vi /etc/rc.local
```

10. Press **i** to add the following content to **rc.local**:

```
/home/fdisk_mount.sh
```

After **rc.local** is configured, the EVS disk initialization script will be automatically executed when the ECS starts.

11. Press **Esc**, enter **:wq**, and press **Enter** to save the changes and exit.

12. Create a private image using an ECS.

13. Create an AS configuration.

When you specify the AS configuration information, select the private image created in the preceding step and select an EVS disk.

14. Create an AS group.

When you configure the AS group, select the AS configuration created in the preceding step.

After the AS group is created, EVS disks of new instances added to this AS group in scaling actions will be automatically initialized.

A Change History

Released On	Description
2022-04-12	This issue is the first official release.