

Auto Scaling

User Guide

Issue 01
Date 2020-11-05



Copyright © Huawei Technologies Co., Ltd. 2020. 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 AS?.....	1
1.2 AS Advantages.....	3
1.3 Lifecycle.....	7
1.4 Use Restrictions.....	10
1.5 Regions and AZs.....	11
1.6 AS and Other Services.....	13
1.7 Basic Concepts.....	15
2 Best Practices.....	17
2.1 Setting Up an Automatically Scalable Discuz! Forum Website.....	17
3 Quick Start.....	20
3.1 Wizard-based Process of Using AS.....	20
3.2 Creating an AS Group Quickly.....	20
4 AS Management.....	26
4.1 AS Group.....	26
4.1.1 Creating an AS Group.....	26
4.1.2 (Optional) Adding a Load Balancer to an AS Group.....	30
4.1.3 Replacing AS Configuration in an AS Group.....	30
4.1.4 Enabling an AS Group.....	31
4.1.5 Disabling an AS Group.....	31
4.1.6 Modifying an AS Group.....	32
4.1.7 Deleting an AS Group.....	33
4.2 AS Configuration.....	33
4.2.1 Creating an AS Configuration.....	33
4.2.2 Using an Existing ECS to Create an AS Configuration.....	34
4.2.3 Using a New Specifications Template to Create an AS Configuration.....	37
4.2.4 Copying an AS Configuration.....	43
4.2.5 Deleting an AS Configuration.....	44
4.3 AS Policy.....	44
4.3.1 Overview.....	44
4.3.2 Creating an AS Policy.....	45
4.3.3 Managing AS Policies.....	54

4.4 Scaling Action.....	55
4.4.1 Dynamically Expanding Resources.....	55
4.4.2 Expanding Resources as Planned.....	57
4.4.3 Manually Expanding Resources.....	57
4.4.4 Configuring an Instance Removal Policy.....	59
4.4.5 Viewing a Scaling Action.....	59
4.4.6 Managing Lifecycle Hooks.....	60
4.4.7 Configuring Instance Protection.....	66
4.5 Bandwidth Scaling.....	67
4.5.1 Creating a Bandwidth Scaling Policy.....	67
4.5.2 Viewing Details About a Bandwidth Scaling Policy.....	73
4.5.3 Managing a Bandwidth Scaling Policy.....	73
4.6 AS Group and Instance Monitoring.....	75
4.6.1 Health Check.....	75
4.6.2 Configuring Notification for an AS Group.....	76
4.6.3 Recording AS Resource Operations.....	77
4.6.4 Adding Tags to AS Groups and Instances.....	80
4.6.5 Monitoring Metrics.....	82
4.6.6 Viewing Monitoring Metrics.....	86
4.6.7 Setting Monitoring Alarm Rules.....	87
5 FAQs.....	88
5.1 General.....	88
5.1.1 What Are Restrictions on Using AS?.....	88
5.1.2 Are ELB and Cloud Eye Mandatory for AS?.....	89
5.1.3 Is AS Billed?.....	89
5.1.4 Does an Abrupt Change on Monitoring Indicator Values Cause an Incorrect Scaling Action?.....	89
5.1.5 How Many AS Policies and AS Configurations Can I Create and Use?.....	89
5.1.6 Can AS Automatically Scale Up or Down vCPUs, Memory, and Bandwidth of ECSs?.....	89
5.1.7 What Is the AS Quota?.....	89
5.1.8 Why is a message displayed indicating that the key pair does not exist and the operation is discontinued when several users under the same account operate AS resources?.....	90
5.2 AS Group.....	90
5.2.1 What Can I Do If the AS Group Fails to Be Enabled?.....	90
5.2.2 How Can I Handle an AS Group Exception?.....	90
5.2.3 What Operation Will Be Suspended After An AS Group Is Disabled?.....	93
5.3 AS Policy.....	93
5.3.1 How Many AS Policies Can Be Enabled?.....	93
5.3.2 What Are the Conditions to Trigger an Alarm in the AS Policy?.....	93
5.3.3 What Is a Cooldown Period? Why Is It Required?.....	94
5.3.4 Can AS Scale Capacity Based on Custom Monitoring of Cloud Eye?.....	94
5.3.5 What Will Monitoring Metrics for an AS Group Be Affected If VM Tools Are Not Installed on ECSs?.....	94
5.3.6 What Can I Do If an AS Policy Fails to Be Enabled?.....	94

5.3.7 How Can I Install the Agent Plug-in on the Instances in an AS Group to Use Agent Monitoring Metrics?.....	95
5.4 Instance.....	98
5.4.1 How Do I Prevent Instances Manually Added to an AS Group from Being Removed Automatically?	98
5.4.2 What Are the Sequence of Selecting Flavors in Multi-Flavor AS Configuration?.....	99
5.4.3 Will the Application Data on an Instance Be Retained After the Instance Is Removed from an AS Group and Deleted?.....	100
5.4.4 Can I Add ECSs Charged in Yearly/Monthly Mode?.....	100
5.4.5 Can Instances That Have Been Added Based on an AS Policy Be Automatically Deleted When They Are Not Required?.....	101
5.4.6 What Is the Expected Number of Instances?.....	101
5.4.7 How Do I Delete an ECS Created in a Scaling Action?.....	101
5.4.8 Will a Yearly/Monthly ECS Be Deleted When the ECS Becomes Faulty?.....	102
5.4.9 How Should I Handle Abnormal Instances in an AS Group?.....	102
5.4.10 What Can I Do If Instances in an AS Group Frequently Fail in Health Checks and Are Deleted and Then Created Repeatedly?.....	103
5.4.11 How Do I Prevent ECSs from Being Removed from an AS Group Automatically?.....	103
5.4.12 Why Is an Instance Removed and Deleted from an AS Group Still Displayed in the ECS List?.....	103
5.5 Others.....	103
5.5.1 What Can I Do to Enable My Application to Be Automatically Deployed on an Instance?.....	103
5.5.2 How Does Cloud-Init Affect the AS Service?.....	104
5.5.3 How Can I Run Existing Services on an Instance Newly Added to an AS Group?.....	105
5.5.4 Why Cannot I Use a Key File to Log In to an ECS?.....	105
5.5.5 Do I Need to Configure an EIP in an AS Configuration When A Load Balancer Has Been Enabled in an AS Group?.....	105
5.5.6 How Can I Enable Automatic Initialization of EVS Disks of Instances That Have Been Added in a Scaling Action to an AS Group?.....	106
A Change History.....	109

1 Service Overview

1.1 What Is AS?

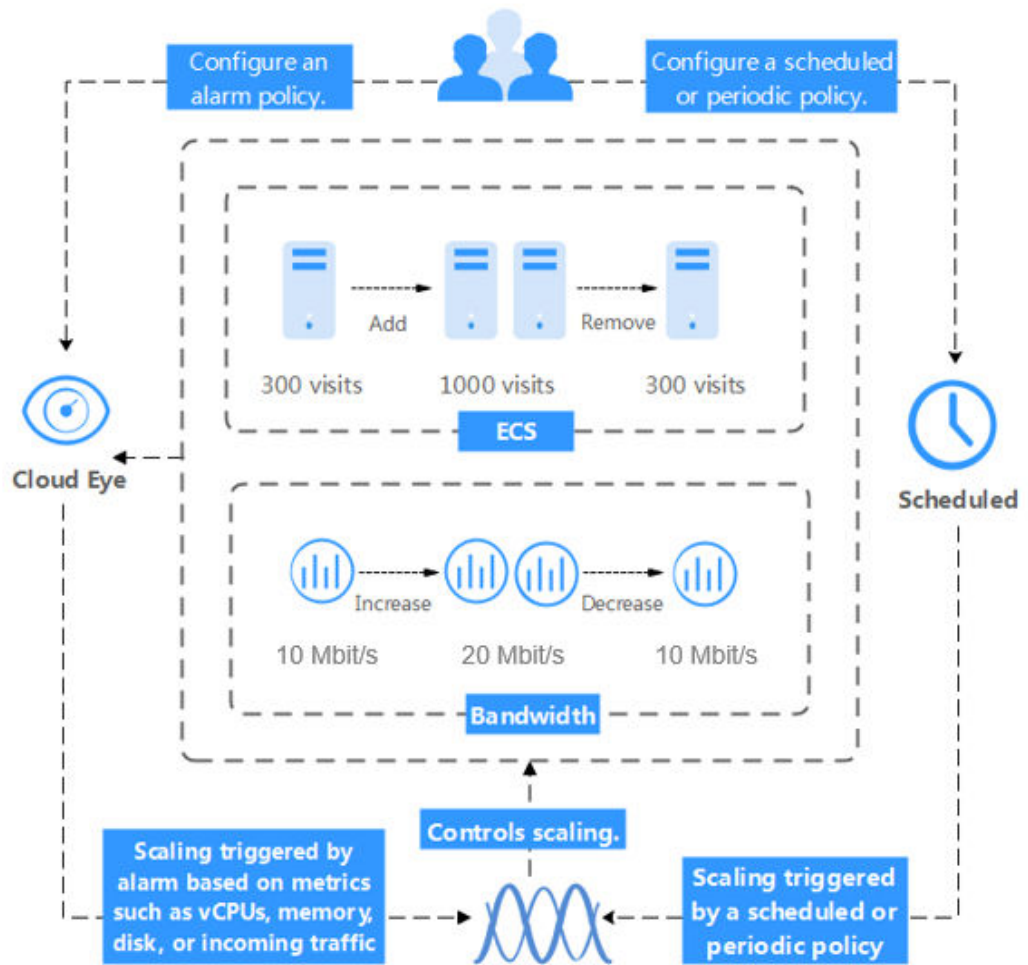
Auto Scaling (AS) automatically adjusts resources to keep up with changes in demand based on pre-configured AS policies. This allows your applications to maintain steady, predictable performance at optimal costs while also freeing you from the cumbersome task of repeatedly, manually adjusting resources to respond to changes. AS automatically adjusts ECS and bandwidth resources.

Architecture

AS allows you to adjust the number of ECSs in an AS group and EIP bandwidth.

- **Scaling control:** You can configure AS policies and set metric thresholds and the execution time of scaling actions. AS will trigger scaling actions when monitoring metrics reach the thresholds or the specified time or period arrives.
- **Policy configuration:** You can configure alarm, scheduled, and periodic policies based on service requirements.
- **Alarm:** You can set a alarm monitoring metric such as vCPU, memory, disk, and inbound traffic.
- **Scheduled:** You can set a scheduled policy by configuring the triggering time.
- **Periodic:** You can set a periodic policy by configuring the interval, triggering time, and time range.
- When Cloud Eye generates an alarm for a monitoring metric, such as CPU usage, AS automatically increases or decreases the number of instances in the AS group or the EIP bandwidth.
- When the configured triggering time arrives, a scaling action is triggered to increase or decrease the number of ECS instances or the bandwidth.

Figure 1-1 AS architecture



How to Access

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

- Calling APIs
Use this method if you are required to integrate the AS service on the public cloud platform into a third-party system for secondary development. For detailed operations, see [Auto Scaling API Reference](#).
- Management console
Use this mode if you are not required to integrate the AS service with a third-party system.
After registering on the public cloud platform, log in to the management console and click **Auto Scaling** under **Computing** on the homepage.

1.2 AS Advantages

AS automatically adjusts service resources to keep up with your demand based on pre-configured AS policies. It has the advantages of automatic resource adjustment, reduced cost, improved availability and high fault tolerance. AS applies to the following scenarios:

- Heavy-traffic forum: Service load changes of a heavy-traffic forum website are difficult to predict. AS dynamically adjusts the number of ECSs based on monitored ECS metrics, such as **vCPU Usage** and **Memory Usage**.
- E-commerce: During large-scale promotions, E-commerce websites need more resources. AS automatically increases ECSs and bandwidth to ensure that promotions go smoothly.
- Live streaming: A live streaming website broadcasts popular programs from 14:00 to 16:00 every day. AS automatically increases ECSs and bandwidth during this period to ensure a smooth viewer experience.

Automatic Resource Adjustment

AS adds instances and increases bandwidth for your applications when the access volume increases and reduces extra resources from the system when the access volume drops, ensuring stable system running.

- Scaling ECSs on Demand

AS adjusts resources for an application system based on demand, thereby enhancing cost management. Resources are adjusted in the following ways:

- Dynamic resource adjustment

AS adjusts resources when an alarm policy is triggered. For details, see [Dynamically Expanding Resources](#).

- Planned resource adjustment

AS adjusts resources when a periodic or scheduled policy is triggered. For details, see [Expanding Resources as Planned](#).

- Manual resource adjustment

Resources are adjusted if you manually change the expected number of instances, or add instances to or remove instances from an AS group. For details, see [Manually Expanding Resources](#).

Consider a train ticket-buying web application running on the public cloud. The load of the application is relatively low during Q2 and Q3 because there aren't many travelers, but is relatively high during Q1 and Q4. The common solution is to provision servers according to the maximum or average load of the application of the application, as shown in [Figure 1-2](#) and [Figure 1-3](#). However, these two solutions may waste resources or cannot meet demand during peak seasons. After you enable AS for the application, AS automatically adjusts the number of servers to keep up with changes in demand. This allows the application to maintain steady, predictable performance at optimal costs, as shown in [Figure 1-4](#).

Figure 1-2 Over-provisioned servers

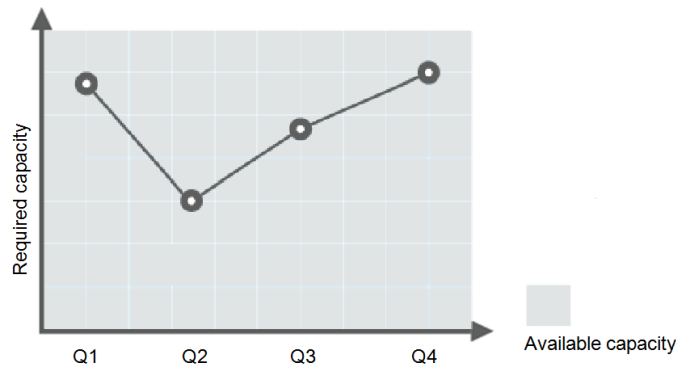


Figure 1-3 Insufficient servers

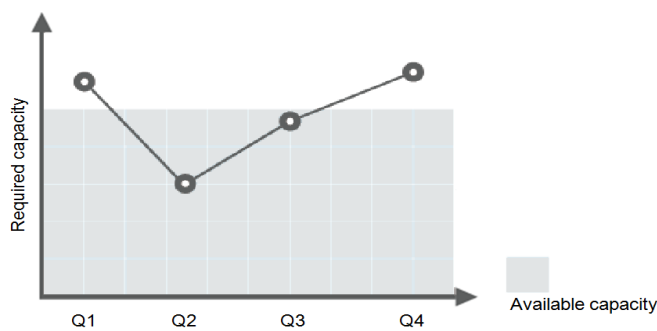
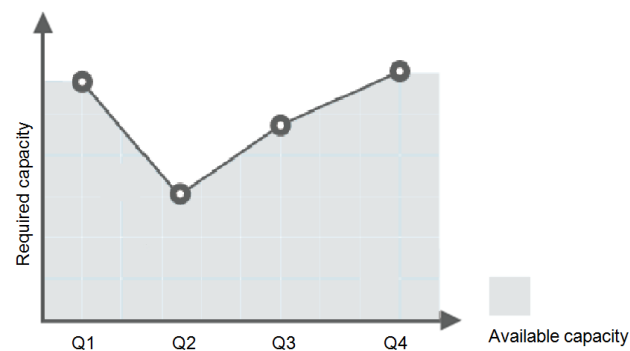


Figure 1-4 Auto-scaled capacity



- **Scaling Bandwidth on Demand**

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

You can select the following scaling policies to adjust the IP bandwidth on demand:

- **Alarm Policy**

You can set the alarm triggering conditions such as outbound traffic and bandwidth. When the system detects that the triggering conditions are met, the system automatically adjusts the bandwidth.

- **Scheduled Policy**

The system can automatically increase, decrease, or adjust the bandwidth to a fixed value at a fixed time according to the scheduled policy.

- Periodic Policy

The system can periodically adjust the bandwidth based on the periodic policy, reducing the workload of manually setting the bandwidth.

The following uses the alarm policy as an example.

Service load changes of a live streaming website in different time periods are difficult to predict. 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 only need to select the target EIP and create two alarm policies. One policy is to add 2 Mbit/s when the outbound traffic is greater than *xxx* bytes, with the limit set to 30 Mbit/s. The other policy is to reduce 2 Mbit/s when the outbound traffic is less than *xxx* bytes, with the limit set to 10 Mbit/s.

- Evenly Distributing Instances by AZ

Instances are evenly distributed in different AZs to reduce the impact of power and network outage on system stability.

A region is a geographic area where resources used by your ECSs are located. Each region contains multiple AZs where resources use independent power supplies and networks. AZs are physically isolated from one another but interconnected through an intranet. Each AZ provides cost-effective and low-latency network connections that are unaffected by faults in other AZs.

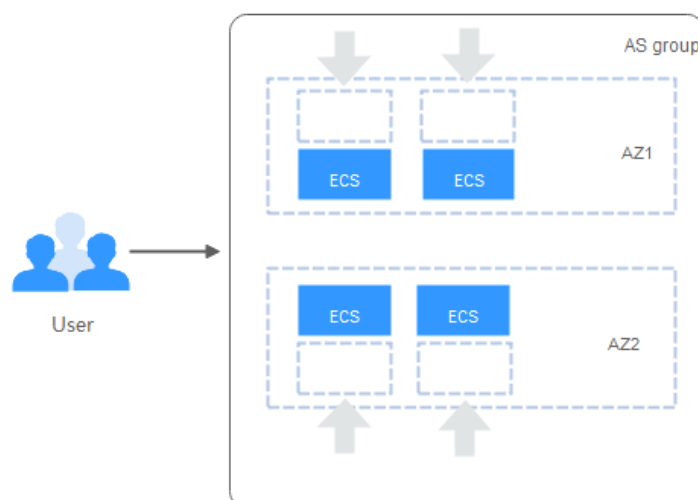
An AS group can contain ECS instances from one or more AZs in a region. To adjust resources, AS evenly distributes ECS instances across AZs based on the following rules:

Evenly distributing new instances to balanced AZs

AS can evenly distribute ECS instances among the AZs used by an AS group. To do it, AS moves new instances to the AZs with the fewest instances.

For example, four instances are evenly distributed in two AZs used by an 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

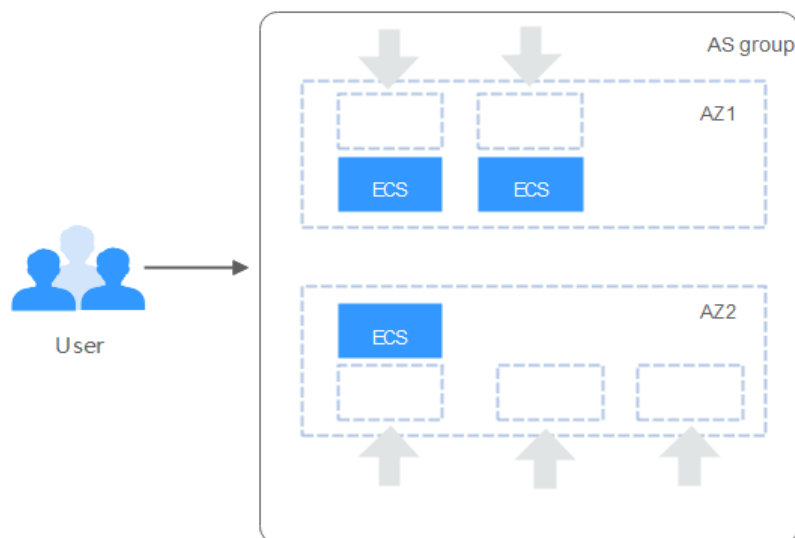


Re-balancing instances in unbalanced AZs

The load of AZs may become unbalanced if you manually add or remove instances in an AS group. The next scaling action will preferentially re-balance the instances in the AZs.

For example, three instances are distributed in AZ 1 and AZ 2 used by an AS group, 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 instances and bandwidth on demand by automatically adjusting resources in the system, eliminating waste of resources and reducing costs.

Improved Availability

AS ensures proper resources for applications. Working with ELB, AS automatically associates a load balancing listener with any ECS instances newly added to the AS group and balances access traffic on all the instances of an AS group through the listener.

Using ELB with AS

Working with ELB, AS automatically increases or decreases ECS instances based on changes in demand while ensuring that the load of all instances is balanced in the AS group.

After ELB is enabled in an AS group, AS automatically associates a load balancing listener with any instances newly added to the AS group. Then, ELB automatically distributes access traffic to all instances in the AS group through the listener, improving system availability. If the instances in the AS group are running various 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

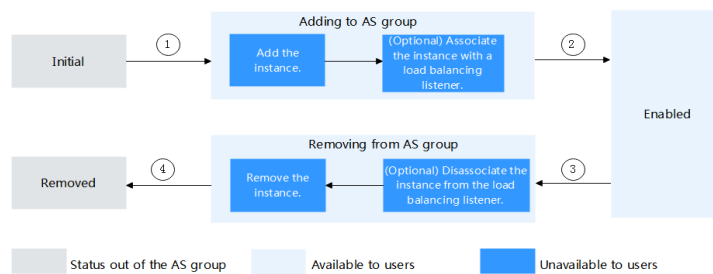
AS monitors instance status in an AS group. After detecting an unhealthy instance, AS replaces it with a new one.

1.3 Lifecycle

The lifecycle of an instance in an AS group starts when it is created and ends when it is removed from the AS group.

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

Figure 1-7 Instance lifecycle



In trigger conditions 2 and 4, a scaling action is automatically triggered to change the instance status.

Table 1-1 Instance status

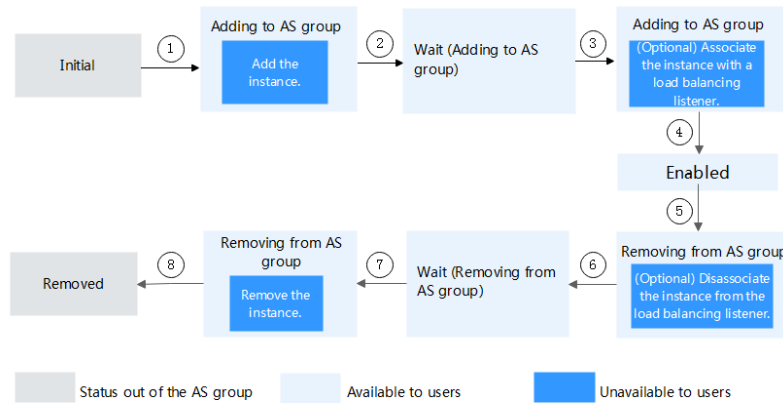
Status	Sub-status	Status Description	Trigger Condition
Initial	N/A	The instance has not been added to an AS group.	The status of an instance is changed to Adding to AS group when either of the following operations is performed: <ul style="list-style-type: none"> When you manually change the expected number of instances or a scaling condition is met, a scaling action is triggered to expand resources. 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.	
	(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.	

Status	Sub-status	Status Description	Trigger Condition
Enabled	N/A	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 operations are performed:
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"> • When you manually change the expected number of instances or a scaling condition is met, a scaling action is triggered to reduce resources. • When a health check shows that an enabled instance is unhealthy, the instance is removed from the AS group. • You manually remove an instance from an AS group.
	Remove the instance.	After the instance is unbound from the load balancing listener, it is removed from the AS group.	
Removed	N/A	The instance lifecycle in the AS group ends.	N/A

Instances are added to an AS group manually or through a scaling action. Then, they go through the **Adding to AS group**, **Enabled**, and **Removing from AS group** statuses, and are finally removed from the AS group.

If you have not added a lifecycle hook to the AS group, the instance lifecycle changes as shown in [Figure 1-8](#). When the AS group is performing a scaling action, instances are suspended by the lifecycle hook and remain in the waiting state until the timeout period ends or the user manually calls back the instances. You can perform desired operations during the waiting. For example, you can install or configure software on a newly added instance or download log files from an instance before it is removed.

Figure 1-8 Instance lifecycle



In trigger conditions 2, 4, 6, and 8, a scaling action is automatically triggered to change the instance status.

Table 1-2 Instance status

Status	Sub-status	Status Description	Trigger Description
Initial	N/A	The instance has not been added to an AS group.	The status of an instance is changed to Adding to AS group when either of the following operations is performed: <ul style="list-style-type: none"> • When you manually change the expected number of instances or a scaling condition is met, a scaling action is triggered to expand resources. • 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.	
Wait (Adding to AS group)	N/A	The lifecycle hook suspends the instance that is being added to the AS group and sets the instance to be in 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: <ul style="list-style-type: none"> • The default callback action is performed. • You manually perform the callback action.
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.	

Status	Sub-status	Status Description	Trigger Description
Enabled	N/A	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 operations are performed:
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"> • When you manually change the expected number of instances or a scaling condition is met, a scaling action is triggered to reduce resources. • When a health check shows that an enabled instance is unhealthy, the instance is removed from the AS group. • You manually remove an instance from an AS group.
Wait (Removing from AS group)	N/A	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 operations is performed:
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	N/A	The instance lifecycle in the AS group ends.	N/A

1.4 Use Restrictions

AS has the following restrictions:

- 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.
ECSs where are stateless applications 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.
ECSs where stateful applications are running store data that needs to be persisted locally.
You can think of stateful transactions as online banking or e-mail, which are performed with the context of previous transactions.
- AS automatically releases ECS instances. Therefore, the instances in AS groups cannot be used to save application status information (such as session statuses) and 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 Quota list

Category	Description	Default Value
AS group	Maximum number of AS groups that you can create	10
AS configuration	Maximum number of AS configurations that you can create	100
AS policy	Maximum number of AS policies that can be added to an AS group	10
Instance	Maximum number of instances that can be added to an AS group	300
Bandwidth scaling policy	Maximum number of bandwidth scaling policies that you can create	10

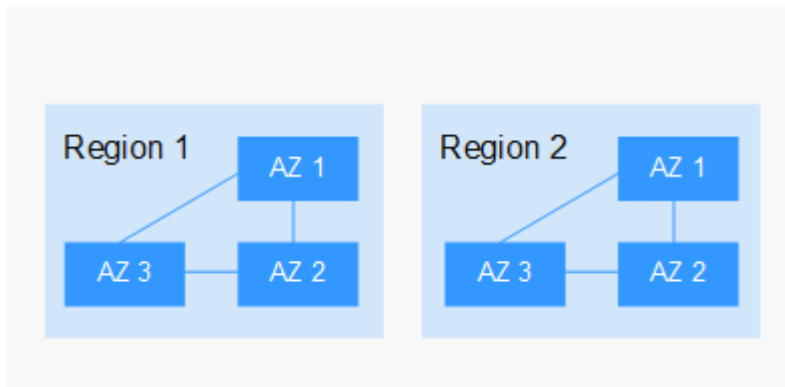
1.5 Regions and AZs

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

- Regions are divided from the dimensions of geographical location and network latency. Public services, such as Elastic Cloud Server (ECS), Elastic Volume Service (EVS), Object Storage Service (OBS), Virtual Private Cloud (VPC), Elastic IP (EIP), and Image Management Service (IMS), are shared within the same region. Regions are classified as universal regions and dedicated regions. A universal region provides universal cloud services for common tenants. A dedicated region provides services of the same type only or for specific tenants.
- An AZ contains one or multiple physical data centers. Each AZ has independent cooling, fire extinguishing, moisture-proof, and electricity facilities. Within an AZ, computing, network, storage, and other resources are logically divided into multiple clusters. AZs within a region are interconnected using high-speed optical fibers to allow you to build cross-AZ high-availability systems.

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

Figure 1-9 Regions and AZs



How to Select a Region?

You are advised to select a region close to you or your target users. This reduces network latency and improves access rate.

How to Select an AZ?

When determining whether to deploy resources in the same AZ, consider your applications' requirements on disaster recovery (DR) and network latency.

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

Regions and Endpoints

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

1.6 AS and Other Services

In addition to scaling resources, AS can work with other cloud services to meet your requirements in various scenarios.

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

Figure 1-10 Relationships between AS and other services

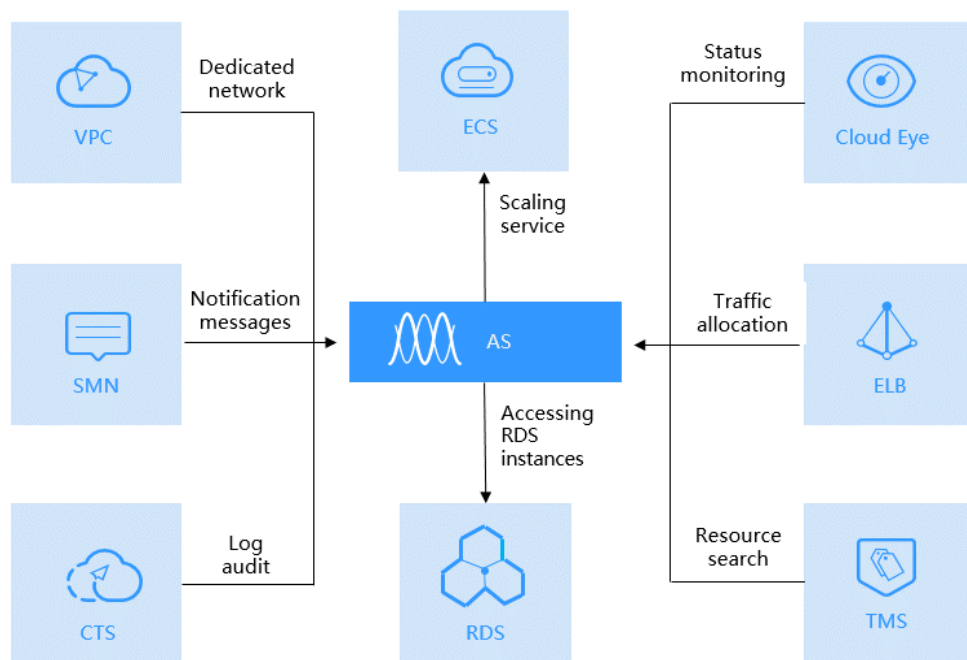


Table 1-4 Related services

Service Name	Description	Function	Reference
Elastic Load Balance (ELB)	After ELB is configured, AS automatically binds ECSs to a load balancer listener when adding ECSs, and unbinds ECSs from the load balancer listener when removing ECSs.	AS distributes traffic to all ECSs in an AS group.	(Optional) Adding a Load Balancer to an AS Group

Service Name	Description	Function	Reference
Cloud Eye	If an alarm-triggered policy is configured, AS triggers scaling actions when an alarm triggering condition specified in Cloud Eye is met.	AS scales resources based on ECS status monitored by Cloud Eye.	AS metrics
Elastic Cloud Server (ECS)	ECSs added in a scaling action can be managed and maintained on the ECS console.	AS automatically adjusts the number of ECSs.	Dynamically Expanding Resources and Expanding Resources as Planned
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
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 easily.	Tags	Marking AS Groups and Instances

Service Name	Description	Function	Reference
Relational Database Service	<p>The prerequisites for directly accessing an RDS DB instance from a scaled instance are as follows:</p> <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 Basic Concepts

AS Group

An AS group consists of a collection of 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 listing specifications for the 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 trigger 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 number of instances in the application system is the same as the expected number of instances.

Cooldown Period

To prevent the alarm policy from being frequently triggered, you must set the cooldown period. 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, the cooldown period is set to 300 seconds, a scheduled policy is specified to trigger a scaling action at 10:32, and a previous scaling action triggered by a alarm policy ends at 10:30. Any alarm-triggered scaling action will be denied during the cooldown period from 10:30 to 10:35, but scheduled scaling actions will still be triggered at 10:32. If the scheduled scaling action ends at 10:36, a new cooldown period starts from 10:36 and ends at 10:41.

Bandwidth Scaling

AS automatically adjusts the bandwidth based on the bandwidth scaling policy you configured. AS can only adjust the bandwidth of pay-per-use EIPs and shared bandwidths, and cannot adjust yearly/monthly bandwidths.

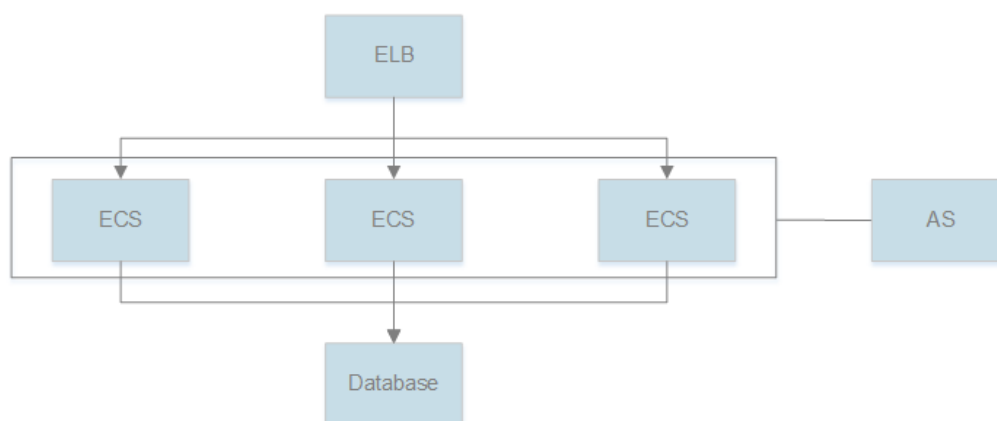
2 Best Practices

2.1 Setting Up an Automatically Scalable Discuz! Forum Website

Overview

AS automatically adds instances to an AS group for applications when necessary and removes extra ones when unnecessary. In this way, you do not need to prepare a large number of ECSs for an expected marketing activity or unexpected peak hours, thereby ensuring system reliability and reducing system 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, the 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 the MySQL database on it.

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 on it. In this section, install the MySQL database on a newly created ECS.

1. Use the created VPC, security group, and EIP to create the ECS. 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 the MySQL database.

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

1. Create an ECS. 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. Therefore, 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 the binding, you can access the ECS through the Internet and install environments, such as PHP and Apache, on the ECS.

3. Deploy the forum.

For instructions about how to deploy the Discuz! forum, see officially released Discuz! documents. When configuring parameters, configure the private IP address of the ECS where the MySQL database is installed for the database server; 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, 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 ECSs for automatically expanding capacity in the AS group.

1. Only a stopped ECS can be used to create a private image. Therefore, 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 ECSs, 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 increase or decrease the number of ECSs to match the Discuz! forum traffic change.

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 have 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 AS policy.

You can configure an alarm-triggered 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. To suit predictable traffic needs, you can also configure a scheduled or periodic AS policy to expand resources.

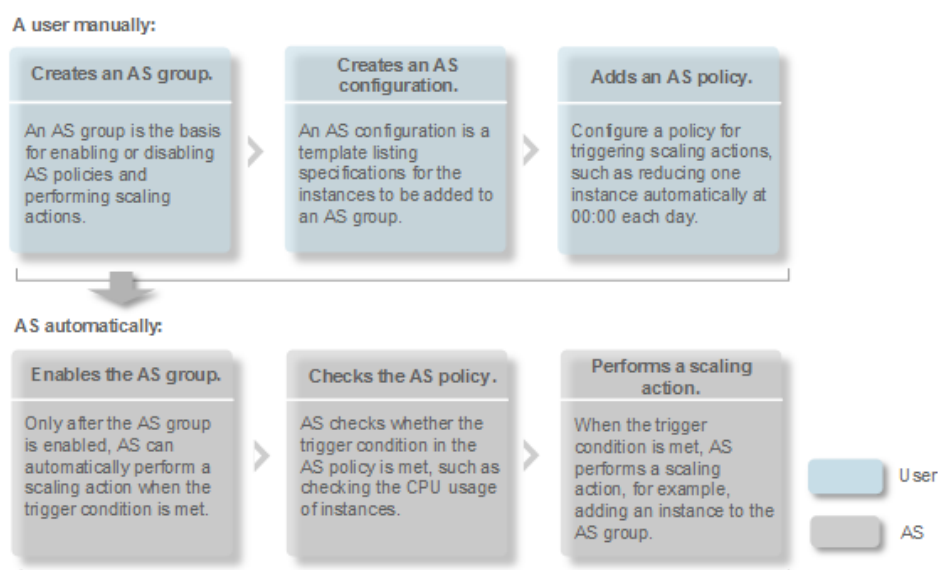
After an AS policy is created and enabled and when the trigger condition is met, the AS group scales up or down.

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 use AS for the first time, it is recommended that you follow the wizard-based process 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 an AS group resides.	N/A
AZ	An AZ is a physical region where resources use independent power supply and networks. AZs are physically isolated but interconnected through an internal network.	N/A
Name	Specifies the name of the AS group to be created. The name contains 1 to 64 characters and consists of only letters, digits, underscores (_), and hyphens (-).	N/A
Max. Instances or Min. Instances	Specifies the maximum or minimum number of ECSs in an AS group.	1
Expected Instances	Specifies the expected number of ECSs in an AS group. After an AS group is created, you can change this value, which will trigger a scaling action.	0
VPC	Provides a network for your ECSs. All ECSs in an AS group belong to the same VPC.	N/A
Subnet	You can select a maximum of five subnets. The AS group automatically binds all NICs to the created ECSs. The first subnet is used by the primary NIC of the ECS by default, and other subnets are used by extension NICs of the ECS.	N/A

Parameter	Description	Example Value
Load Balancing	<p>This parameter is optional. A load balancer automatically distributes access traffic to all instances in an AS group to balance their service load. It enables higher levels of fault tolerance in 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 service scalability. If ELB health check is selected for Health Check Method, when any one of the listeners detects that an instance becomes faulty, AS will replace the faulty instance with a functional one.	N/A
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 based on the oldest configuration is removed from the AS group first.• Newest instance created from oldest AS configuration: The latest instance created based on 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">• Removing instances will preferentially ensure that the remaining instances are evenly distributed in AZs.• A manually added ECS is removed in the lowest priority. AS does not delete a manually added ECS when removing it. If multiple manually added ECSs must be removed, AS preferentially removes the earliest-added ECS.	N/A

Parameter	Description	Example Value
Health Check Method	<p>When a health check detects a faulty ECS, AS removes the faulty ECS from the AS group and adds a new one. The health check is implemented using any of the following methods:</p> <ul style="list-style-type: none"> ECS health check: checks ECS running status. If an ECS is stopped or deleted, it is considered as abnormal. This method is selected by default. Using this method, the AS group periodically determines the running status of each ECS based on the health check result. If the health check result shows that an ECS is faulty, AS removes the ECS from the AS group. 	N/A
Health Check Interval	<p>Specifies the health check period for an AS group. You can set a proper health check interval, such as 10 seconds, 1 minute, 5 minutes, 15 minutes, 1 hour, and 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 Simple Message Notification User Guide. 	N/A
Tag	<p>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. 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 value cannot be empty. - An AS group has a unique key. - The value consists of at most 36 characters. It cannot contain the following characters: =*<>_ / ● Value <ul style="list-style-type: none"> - The value can be an empty character string. - A key can have only one value. - The value consists of at most 43 characters. It cannot contain 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 that which is 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 that which is 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 supported by different AZs may vary. When creating an AS group, you must choose a proper AS configuration according to the ECS type supported by the AZs used by the AS group.

- If the ECS type specified in the AS configuration is supported by none of the AZs used by the AS group, the following situations will occur:
 - If the AS group is disabled, it cannot be enabled.
 - If the AS group is enabled, its status will become abnormal when instances are added to it.
- If the ECS type specified in the AS configuration is supported only by certain AZs used by the AS group, the ECSs added by a scaling action are distributed only in the AZs supporting the ECS type. 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 an AS group resides.	N/A
AZ	An AZ is a physical region where resources use independent power supply and networks. AZs are physically isolated but interconnected through an internal network.	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 expanding ECSs in an AS group, the system preferentially distributes ECSs evenly among AZs used by the AS group. If it fails in the target AZ, it automatically selects another AZ based on the sequenced policy.● Sequenced: When expanding ECSs in an AS group, the system selects the target AZ based on the order in which AZs are selected. NOTE This parameter needs to be configured when two or more AZs are selected.	Load-balanced
Name	Specifies the name of the AS group to be created. The name contains 1 to 64 characters and consists of only letters, digits, underscores (_), and hyphens (-).	N/A
Max. Instances or Min. Instances	Specifies the maximum or minimum number of ECSs in an AS group.	1
Expected Instances	Specifies the expected number of ECSs in an AS group. After an AS group is created, you can change this value, which will trigger a scaling action.	0

Parameter	Description	Example Value
AS configuration	Specifies the required AS configuration for the AS group. An AS configuration defines the specifications of the ECSs 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 ECSs. All ECSs in an AS group belong to the same VPC.	N/A
Subnet	You can select a maximum of five subnets. The AS group automatically binds all NICs to the created ECSs. The first subnet is used by the primary NIC of the ECS by default, and other subnets are used by extension NICs of the ECS.	N/A
Load Balancing	This parameter is optional. A load balancer automatically distributes access traffic to all instances in an AS group to balance their service load. It enables higher levels of fault tolerance in 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 faulty, AS will replace the faulty instance with a functional one. 	-

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 based on the oldest configuration is removed from the AS group first. • Newest instance created from oldest AS configuration: The latest instance created based on 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"> • Removing instances will preferentially ensure that the remaining instances are evenly distributed in AZs. • A manually added ECS is removed in the lowest priority. AS does not delete a manually added ECS when removing it. If multiple manually added ECSs must be removed, AS preferentially removes the earliest-added ECS. 	Oldest instance created from oldest AS configuration
Health Check Method	<p>When a health check detects a faulty ECS, AS removes the faulty ECS from the AS group and adds a new one. The health check is implemented using any of the following methods:</p> <ul style="list-style-type: none"> • ECS health check: checks ECS running status. If an ECS is stopped or deleted, it is considered as abnormal. This method is selected by default. Using this method, the AS group periodically determines the running status of each ECS based on the health check result. If the health check result shows that an ECS is faulty, AS removes the ECS from the AS group. • ELB health check: determines ECS running status using a load balancing listener. This health check method is available only when the AS group uses a load balancing listener. When a load balancing listener detects that an ECS is faulty, AS removes the ECS from the AS group. 	N/A

5. Click **Next**. On the **Add AS Configuration** page, you can choose to use an existing AS configuration or create one. For details, see [Using an Existing ECS](#)

[to Create an AS Configuration](#) and [Using a New Specifications Template to Create an AS Configuration](#).

6. Click **Create Now**.
7. Check the AS group and AS configuration information. and click **Submit**.
8. (Optional) Add AS policies. For details, see [Creating an AS Policy](#).

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.

AS supports only created load balancers. 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 Replacing AS Configuration in an AS Group

Scenarios

If you need to change the ECS specifications in an AS group, you need to change the AS configuration.

Effective Time of New AS Configuration

If the AS group has an in-progress scaling action, the new AS configuration will take effect only for instances scaled in later scaling actions.

For example, the current AS configuration of the AS group is **as-config-A**, and the new AS configuration is **as-config-B**. The instance configuration in the current scaling action is still **as-config-A**.

The instance configuration will change to **as-config-B** in the next scaling action.

Figure 4-1 Changing the AS configuration

Name	Status	Specifications	Image	System Disk	Data Disks	Login Mode	Created	Billing Mode	Operation
<input checked="" type="checkbox"/> as-config-B	Unbound	kc1.large2 2 vCPUs 4 GB	CentOS 8.0 64bit with ARM	High I/O 40 GB	0	Password	Oct 09, 2020 15:17:52 ...	Pay-per-use	Copy Delete
<input type="checkbox"/> as-config-A	Unbound	s2.small1 1 vCPUs 1 GB	CentOS 7.6 64bit	High I/O 40 GB	0	Password	Oct 09, 2020 14:42:33 ...	Pay-per-use	Copy Delete

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 target AS group. 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 increase or decrease instances.

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

When you are required 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 restores, enable the AS group again.

If a scaling action, for example, creating an instance or EVS disk, consistently fails (the failure cause can be viewed on the **Elastic Cloud Server** page) and retries in an AS group, use either of the following methods to stop the retry:

- Disable the AS group. Then, the scaling action that is being performed fails and will not retry. 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. After the scaling action fails and ends, the scaling action will not retry.

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 as 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. Then, AS automatically increases or decreases 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 **ELB** 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 row containing the target AS group, click the AS group name to switch to the **Basic Information** page, and click **Modify** in the upper right corner.

You can also locate the row containing the target AS group and choose **More > Modify** in the **Operation** column.

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 an instance or ongoing scaling action, 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. Manually added instances are only removed out of the AS group, while the instances automatically created in a scaling action are removed and deleted. During the preceding process, you are not allowed to perform other operations in the AS group.
- After an AS group is deleted, its AS policies and the alarm rules generated based on the AS policies configured for the AS group 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**.
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 ECSs to be added to an AS group. The specifications include the ECS image and system disk size.

Scenarios

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

- Change the AS configuration on the AS group details page.

Methods

- Using an existing ECS to create an AS configuration
When you create an AS configuration using an existing ECS, the vCPU, memory, image, disk, and ECS type are the same as those of the selected ECS by default. For details, see [Using an Existing ECS to Create an AS Configuration](#).
- Using a new specifications template to create an AS configuration
If you have special requirements on the ECSs for resource expansion, use a new specifications template to create the AS configuration. For details, see [Using a New Specifications Template to Create an AS Configuration](#).

4.2.2 Using an Existing ECS to Create an AS Configuration

Scenarios

You can use an existing ECS 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 ECS 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 and click 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

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. The bandwidth size is configurable. <p>NOTE If you select Automatically assign, you need to specify Type, Bandwidth Type, Billed By, and Bandwidth.</p>	Automatically assign
Bandwidth Type	<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 In this mode, keys are used for authenticating the users who attempt to log in to target ECSs. 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. Otherwise, logging in to the ECS will fail.</p>• Password In this mode, the initial password of user root (for Linux) or user Administrator (for Windows) is used for authentication. You can log in to an ECS using the username and its initial password.	Admin@123
Advanced Settings	<p>This allows you to configure User Data Injection.</p> <p>You can select Do not configure or Configure now.</p>	N/A

Parameter	Description	Example Value
User Data Injection	<p>Enables the 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 upon its first startup.</p> <p>For details, see Elastic Cloud Server User Guide.</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. If you select As file, the system automatically injects the files into a specified directory when creating an ECS.<ul style="list-style-type: none">– For Linux, specify the path for storing the injected file, for example /etc/foo.txt.– For Windows, the injected file is automatically stored in the root directory of disk C. You only need to specify the file name, such as foo. The file name can contain only letters and digits. <p>NOTE</p> <ul style="list-style-type: none">• For Linux, if you use the password authentication mode, the user data injection function is unavailable.• If the selected image does not support user data injection, the user data injection function is unavailable.	-

5. After setting the parameters, click **Create Now**.
6. If you want to use the newly created AS configuration, add it to the AS group. For details, see [Replacing AS Configuration in an AS Group](#).
7. (Optional) Enable the AS group.
If the AS group is in **Disabled** state, enable it. For details, see [Enabling an AS Group](#).

4.2.3 Using a New Specifications Template to Create an AS Configuration

Scenarios

If you have special requirements on the ECSs for resource expansion, use a new specifications template to create the AS configuration. In such a case, ECSs

meeting specifications of 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	The following two types of 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).	x86
Specifications	The public cloud provides various ECS types for different application scenarios. For more information, see Elastic Cloud Server User Guide . Configure the ECS specifications, including vCPUs, memory, image type, and disk, according to the ECS type.	Memory-optimized ECS

Parameter	Description	Example Value
Image	<ul style="list-style-type: none"> <li data-bbox="647 300 1174 533"> <p>• 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.</p> <li data-bbox="647 546 1174 810"> <p>• 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 removes the need to configure multiple ECSs repeatedly.</p> <li data-bbox="647 824 1174 922"> <p>• Shared image A shared image is a private image shared by another public cloud user.</p> 	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. In addition, you can specify a data disk image for exporting data. If the image you selected is of the full-ECS image type, you can change the volume type and size and encryption attributes of the data disk restored using the disk backup. Ensure that the volume size is greater than or equal to the disk backup size, and the encryption attributes can be modified only if the disk backup of the full-ECS image locates in the target region.	Common I/O for System Disk
Security Group	Controls ECS access within or between security groups by defining access rules. ECSs added to a security group are protected by the access rules you define.	N/A

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 In this mode, keys are used for authenticating the users who attempt to log in to target ECSs. 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. Otherwise, logging in to the ECS will fail.</p>• Password In this mode, the initial password of user root (for Linux) or user Administrator (for Windows) is used for authentication. You can log in to an ECS using the username and its initial password.	Admin@123
Advanced Settings	<p>This parameter allows you to configure ECS Group and User Data Injection. You can select Do not configure or Configure now.</p>	N/A

Parameter	Description	Example Value
User Data Injection	<p>Enables the 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 upon its first startup.</p> <p>For details, see Elastic Cloud Server User Guide.</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. If you select As file, the system automatically injects the files into a specified directory when creating an ECS.<ul style="list-style-type: none">– For Linux, specify the path for storing the injected file, for example /etc/foo.txt.– For Windows, the injected file is automatically stored in the root directory of disk C. You only need to specify the file name, such as foo. The file name can contain only letters and digits. <p>NOTE</p> <ul style="list-style-type: none">• For Linux, if you use the password authentication mode, the user data injection function is unavailable.• If the selected image does not support user data injection, the user data injection function is unavailable.	-

5. Click **Create Now**.
6. If you want to use the newly created AS configuration, add it to the AS group. For details, see [Replacing AS Configuration in 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 use an AS configuration, you can delete it. An AS configuration can be deleted only when it is not used by any AS group. You can delete an AS configuration or multiple AS configurations in a batch.

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 in batches.

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 condition to trigger a scaling action and the operation to be performed in a scaling action. When the trigger condition is met, the system automatically triggers a scaling action.

NOTE

When multiple AS policies are applied to an AS group, a scaling action is triggered as long as one of the AS policies is triggered, 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

Parameter	Description	Example Value
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"> – Rounds down the value that is greater than 1. For example, value 12.7 is rounded down to 12. – Rounds up the value that is greater than 0 and less than 1 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
Cooldown Period	<p>To prevent the alarm policy from being frequently triggered, you must set the cooldown period.</p> <p>A cooldown period specifies a period of time in the unit of second after each scaling action is complete.</p> <p>During the cooldown period, scaling actions triggered by alarms will be denied. Scheduled and periodic scaling actions are not restricted.</p> <p>For example, the cooldown period is set to 300 seconds, a scheduled policy is specified to trigger a scaling action at 10:32, and a previous scaling action triggered by a alarm policy ends at 10:30. Any alarm-triggered scaling action will be denied during the cooldown period from 10:30 to 10:35, but scheduled scaling actions will still be triggered at 10:32. If the scheduled scaling action ends at 10:36, a new cooldown period starts from 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 that which is 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 that which is configured for the AS group. The default cooldown period is 300 seconds.	300

Table 4-5 System monitoring parameters

Parameter	Description	Example Value
Alarm Rule Name	Specifies the name of the alarm rule.	as-alarm-7o1u

Parameter	Description	Example Value
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 Occurrences is set to n , the sampling points of the alarm rule are the sampling points in n consecutive sampling periods. Only if all the sampling points meet the threshold configured for the alarm rule will the alarm rule status be refreshed as the Alarm status.	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%
Monitoring Interval	Specifies the interval at which the alarm status is updated based on the alarm rule.	5 minutes

Parameter	Description	Example Value
Consecutive Occurrences	Specifies the number of sampling points when an alarm is triggered. If Occurrences is set to n , the sampling points of the alarm rule are the sampling points in n consecutive sampling periods. Only if all the sampling points meet the threshold configured for the alarm rule will the alarm rule status be refreshed as the Alarm status.	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
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

Parameter	Description	Example Value
Time Zone	The default value is GMT +08:00 . GMT+08:00, Beijing, China time, is 8:00 hours ahead 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"> • Rounds down the value that is greater than 1. For example, value 12.7 is rounded down to 12. • Rounds up the value that is greater than 0 and less than 1 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
Cooldown Period	<p>To prevent the alarm policy from being frequently triggered, you must set the cooldown period.</p> <p>Specifies a period of time after each scaling action is complete.</p> <p>During the cooldown period, scaling actions triggered by alarms will be denied. Scheduled and periodic scaling actions are not restricted.</p> <p>For example, the cooldown period is set to 300 seconds, a scheduled policy is specified to trigger a scaling action at 10:32, and a previous scaling action triggered by a alarm policy ends at 10:30. Any alarm-triggered scaling action will be denied during the cooldown period from 10:30 to 10:35, but scheduled scaling actions will still be triggered at 10:32. If the scheduled scaling action ends at 10:36, a new cooldown period starts from 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 that which is 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 that which is 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 triggered 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 triggered operation. If the conditions are met, a scaling action is triggered to perform the required operation.

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

Modifying an AS Policy

Modify parameter settings of an AS policy if it cannot meet service requirements.

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

Disable a specified AS policy if you do not want it to trigger any scaling action within a specified period of time.

- If all AS policies of 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

Perform this operation to make the number of instances in an AS group reach the expected number of instances immediately.

- 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

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 based on the AS policy is in progress. Deleting the AS policy does not adversely affect the in-progress scaling action.

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 Dynamically Expanding Resources

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

If your demands change frequently, you can also configure alarm policies for dynamically expanding or reducing resources. When the conditions for triggering an AS policy are met, AS automatically changes the expected number of instances for triggering a scaling action to scale up or down resources. For details about how to create an alarm policy, see [Creating an AS Policy](#).

For example, for a web application that allows users to purchase train tickets, when 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. When the CPU usage goes down to 30%, an instance needs to be deleted to prevent resource waste. To meet the requirements, you can configure two alarm policies. The trigger condition of the first policy is that the maximum CPU usage becomes greater than 90% and the action is to add one instance. For details, see [Figure 4-2](#). In the second alarm policy, the trigger condition is that the minimum vCPU usage is less than 30% and the action is to reduce an instance. For details, see [Figure 4-3](#).

Figure 4-2 Alarm policy 01

Add AS Policy

Policy Name	<input type="text" value="as-policy-001"/>
Policy Type	<input checked="" type="radio"/> Alarm <input type="radio"/> Scheduled <input type="radio"/> Periodic
Alarm Rule	<input checked="" type="radio"/> Create <input type="radio"/> Use existing
Rule Name	<input type="text" value="as-alarm-cpu01"/>
Trigger Condition	<input type="text" value="CPU Usage"/> <input type="text" value="Max."/> <input type="text" value=">"/> <input type="text" value="90"/> % <small>To determine if an OS supports metrics Memory Usage, Inband Outgoing Rate, and Inband Incoming Rate, see Elastic Cloud Server User Guide.</small>
Monitoring Interval	<input type="text" value="5 minutes"/>
Consecutive Occurrences [?]	<input type="text" value="3"/>
Scaling Action	<input type="text" value="Add"/> <input type="text" value="1"/> <input type="text" value="instances"/>
Cooldown Period (s) [?]	<input type="text" value="300"/>

Figure 4-3 Alarm policy 02

Add AS Policy

Policy Name:

Policy Type: Alarm Scheduled Periodic

Alarm Rule:

Rule Name:

Trigger Condition: %

To determine if an OS supports metrics Memory Usage, Inband Outgoing Rate, and Inband Incoming Rate, see [Elastic Cloud Server User Guide](#).

Monitoring Interval:

Consecutive Occurrences [?]:

Scaling Action:

Cooldown Period (s) [?]:

4.4.2 Expanding Resources as Planned

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 in other periods of the semester. You can configure two scheduled policies at the beginning of each semester. The first policy triggers a scaling action to add an instance when the course selection starts, and the second policy triggers a scaling action to reduce an instance when the course selection ends, meeting students' requirements as well as reducing cost.

4.4.3 Manually Expanding Resources

Scenarios

Expand resources by manually adding instances to an AS group, removing instances 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 must be in the same VPC as the AS group.
- After instances are added, the total number of instances is less than or equal to the maximum number of instances allowed.
- A batch operation can be performed on a maximum of 10 instances 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 rectify the instance fault, and add the instance to the AS group again. The instance removed out of the AS group does not carry application traffic any more.

For example, you can modify the AS configuration for an AS group at any time. However, the AS group does not update instances that are running. In such an event, terminate the instance and replace it in the AS group. Alternatively, remove the instance out of the AS group, update the instance software, and add the instance to the AS group again.

Restrictions on instance removal are as follows:

- The AS group does not have a scaling action that is being performed, the instances are enabled, and the total number of instances after removal is not less than the minimum number of instances allowed.
- 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 in Storage Disaster Recovery Service (SDRS).
- Instances manually added to an AS group can only be removed, and cannot be removed and deleted.
- A batch operation can be performed on a maximum of 10 instances 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 **Instance 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 replace old instances by new instances in an AS group.
- **Newest instance:** The latest instance is removed from the AS group first. Use this policy if you want to test a new AS configuration and do not want to retain it.
- **Oldest instance created from oldest AS configuration:** The oldest instance created based on the oldest configuration is removed from the AS group first. Use this policy if you want to update an AS group and delete the instances created based on early AS configurations gradually.
- **Newest instance created from oldest AS configuration:** The latest instance created based on the oldest configuration is removed from the AS group first.

NOTE

A manually added ECS is removed in the lowest priority. AS does not delete a manually added ECS when removing it. If multiple manually added ECSs must be removed, AS preferentially removes the earliest-added ECS.

4.4.5 Viewing a Scaling Action

Scenarios

To check whether a scaling action is performed or view scaling action details, perform the operations described in this section.

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 creation and removal of ECS instances in AS groups and manage the lifecycle of ECS instances in AS groups. **Figure 4-4** shows the instance lifecycle when no lifecycle hook is added to the AS group.

Figure 4-4 Instance lifecycle statuses when no lifecycle hook is added to the AS group

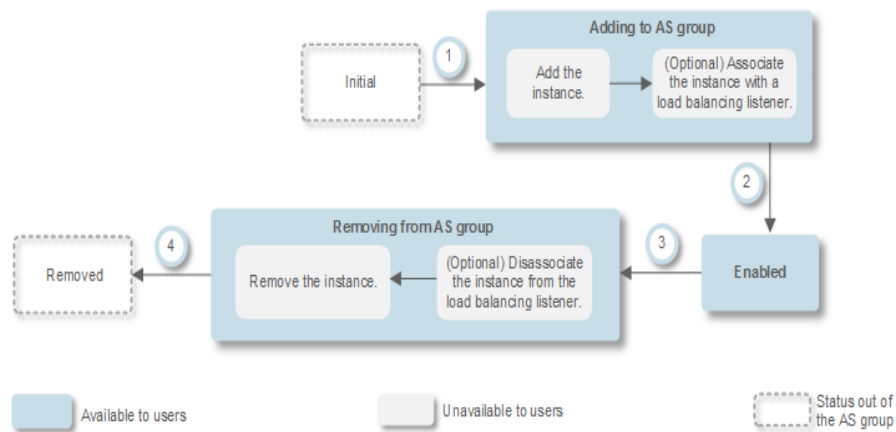
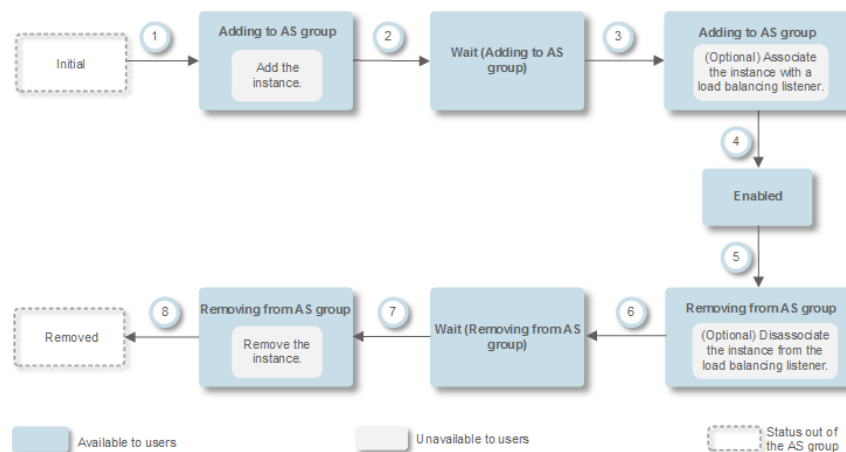


Figure 4-5 shows the instance lifecycle when a lifecycle hook is added to the AS group.

Figure 4-5 Instance lifecycle statuses when a lifecycle hook is added to the AS group



When the AS group performs a scaling action and triggers the lifecycle hook, the scaling action is suspended and the instance that is being added to or removed from the AS group is set to waiting state, as shown in 2 and 6 in **Figure 4-5**.

During this period, you can perform some custom operations on the instance. For example, you can install or configure software on an instance to be added to the AS group. The suspension of a scaling action will be ended in either of the following scenarios:

- The time when the instance stays in waiting state is longer than the timeout duration.
- A callback operation is performed to stop the instance waiting state.

Application Scenarios

- Instances newly added to an AS group can be bound to a load balancer only after initialization (software installation or configuration) has been performed on the instances and services start running properly.
- Instances can be removed from an AS group only after they are unbound from the load balancer and have finished processing ongoing requests.
- Before instances are removed from an AS group, data needs to be backed up and logs need to be downloaded.
- Other scenarios where custom operations need to be performed

Working Rules

After added to an AS group, a lifecycle hook works as follows:

- Adding an ECS instance to an AS group
After an instance is added to an AS group and initialized, 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. After finishing the custom operations, you can perform a callback operation to end the instance waiting state. Alternatively, you can wait until the timeout duration ends, when the system automatically ends the instance waiting state. After the instance waiting state ends, two types of **Default Callback Action** are available, **Continue** and **Abandon**.
 - **Continue**: The instance in waiting state will be added to the AS group.
 - **Abandon**: The instance in waiting state 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 one 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. After finishing the custom operations, you can perform the default callback operation or wait for the system to end the instance waiting state when the timeout duration expires. After the instance waiting state ends, two types of **Default Callback Action** are available, **Continue** and **Abandon**.

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

If there are multiple lifecycle hooks, **Continue** allows suspension by other lifecycle hooks to time out. The instance will be removed from the AS group only when the status of all lifecycle hooks is **Continue**. If the **Default Callback Action** of any lifecycle hook is **Abandon**, the instance will be directly removed from the AS group.

Use Restrictions

- 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**.
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 is a string of 1 and 32 characters and can contain letters, digits, underscores (_), and hyphens (-).	we12_w
Hook Type	Specifies the lifecycle hook type. The value can be Instance adding or Instance removal . Instance adding sets an instance that is being added to an AS group to Wait (Adding to AS group) state. Instance removal sets an instance that is being removed from an AS group Wait (Removing from AS group) state.	Instance adding

Parameter	Description	Example Value
Default Callback Action	<p>Specifies the default system action when the waiting duration of an instance expires.</p> <p>When an instance is being added to an AS group, the default callback action can be Continue or Abandon:</p> <ul style="list-style-type: none"> • Continue: When only one lifecycle hook is available, the instance will continue to be added to the AS group. When multiple lifecycle hooks are available, the instance will continue to be added to the AS group only when all lifecycle hooks change to Continue state. • Abandon: The system deletes the instance and creates a new one only when one lifecycle hook is in Abandon state, regardless of whether one or more lifecycle hooks are available. <p>When an instance is being removed from an AS group, the default callback action can be Continue or Abandon:</p> <ul style="list-style-type: none"> • Continue: When only one lifecycle hook is available, the instance will be removed from the AS group. When multiple lifecycle hooks are available, the instance will be removed from the AS group only when all lifecycle hooks change to Continue state. • Abandon: The system removes the instance from the AS group when one lifecycle hook is in Abandon state, regardless of whether one or more lifecycle hooks are available. 	Continue
Timeout Duration (s)	<p>Specifies the instance waiting duration by default. The value ranges from 300s to 86,400s.</p> <p>You can prolong the timeout duration or perform the Continue or Abandon operation before the timeout duration expires. For more information about callback actions, see Performing a Callback Action.</p>	3600

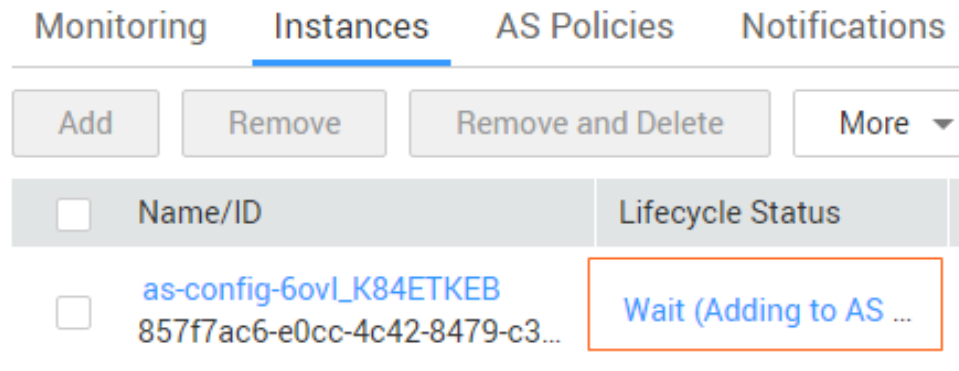
Parameter	Description	Example Value
Notification Topic	<p>Specifies a notification object for a lifecycle hook. For details, see "Configuring a Topic Policy" in <i>Simple Message Notification User Guide</i>. After the configuration, when an instance is suspended by the lifecycle hook, the system sends a notification to the object. This notification contains the basic instance information, your customized notification content, and the token for controlling lifecycle operations. An example notification is as follows:</p> <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 customized 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, as shown in [Figure 4-6](#).

Figure 4-6 Performing a callback operation

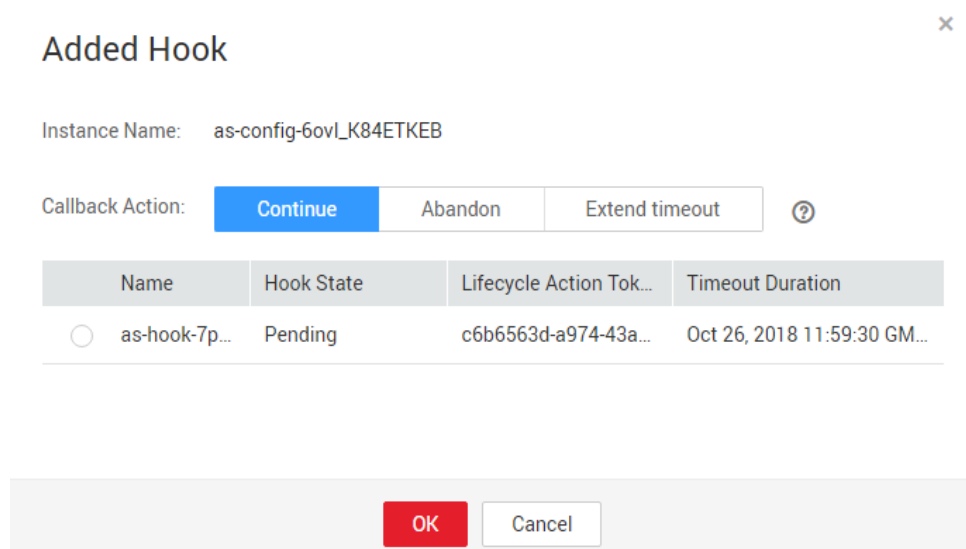


NOTE

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

- In the displayed **Added Hook** dialog box, view the suspended instance and all the lifecycle hooks, and perform callback actions on lifecycle hooks, as shown in **Figure 4-7**.

Figure 4-7 Added Hook dialog box



Callback actions include:

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

If you have performed customized operations before the timeout duration expires, select **Continue** or **Abandon** to complete the lifecycle operations. For details about **Continue** and **Abandon**, see **Table 4-8**. If you require more time for customizing operations, select **Extend timeout** to prolong the timeout

duration. Then, the instance waiting duration will be prolonged by 3600 sections 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

Configure instance protection if you want to specify one or more ECSs not to be automatically removed from an AS group. After the configuration, when AS automatically reduces the number of ECSs in an AS group, the in-service ECSs with instance protection enabled will not be removed.

Prerequisites

In the following scenarios, ECSs will still be removed from the AS group even if instance protection is enabled:

- The ECS is not healthy according to the health check.
- The ECS is manually removed from the AS group.

NOTE

- Instances in the abnormal health status cannot provide services. The AS group preferentially ensures that all instances in the group are normal. Therefore, instance protection cannot protect abnormal instances.
- By default, instance protection does not take effect on the ECSs that are newly created in or added to an AS group.
- Instance protection becomes invalid immediately when the target ECS is removed from the AS group.

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**.
3. Click the name of the target AS group.
4. Click the **Instances** tab. Select one or more ECSs and choose **Enable Instance Protection** from the **More** drop-down list. In the displayed **Enable Instance Protection** dialog box, click **Yes**.

You can also locate the row containing the target ECS and click **Enable Instance Protection** in the **Operation** column. Then, in the **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**.
2. Click the name of the target AS group.
3. Click the **Instances** tab. Select one or more ECSs and choose **Disable Instance Protection** from the **More** drop-down list. In the displayed **Disable Instance Protection** dialog box, click **Yes**.

You can also locate the row containing the target ECS and click **Disable Instance Protection** in the **Operation** column. Then, in the **Disable Instance Protection** dialog box, click **Yes**.

4.5 Bandwidth Scaling

4.5.1 Creating a Bandwidth Scaling Policy

Scenarios

You can automatically adjust the your purchased EIP bandwidth and shared bandwidth using a bandwidth scaling policy. This section describes how to create an 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 [Creating a Bandwidth Scaling Policy](#).

Table 4-9 Alarm policy parameters

Parameter	Description	Example Value
Region	Specifies the region where the AS group resides.	N/A

Parameter	Description	Example Value
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. NOTE Currently, only pay-per-use EIPs can be scaled. Yearly/monthly EIPs cannot be used to create AS groups.	N/A
Policy Type	Select Alarm .	Alarm
Alarm Rule	You can use an existing alarm rule or create a new one. Alternatively, click Create Alarm Rule on the right side of the Alarm parameter and create an alarm rule on the Alarm Rules page. For details, see Creating an Alarm Rule . To create an alarm rule, configure the following parameters: <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.	300

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** on the right of **Rule Name** to create an alarm rule. To do so, perform the following operations:

1. Click **Create Alarm Rule** on the right of **Rule Name** to switch to the **Alarm Rules** page of Cloud Eye.
2. On the **Alarm Rules** page, click **Create Alarm Rule** in the upper right corner.
3. Set parameters based on [Figure 4-8](#) and [Table 4-11](#). For more information about how to set alarm rules, see [Cloud Eye User Guide](#).

Figure 4-8 Creating an alarm rule

The screenshot displays the 'Create Alarm Rule' configuration interface. Key elements include:


- Resource Type:** Elastic IP and Bandwidth...
- Dimension:** Bandwidths
- Monitoring Scope:** Specific resources
- Method:** Create manually
- Alarm Policy:** Outbound Bandwidth, Max., 5 minutes, 3 consecutive, ≥, 500 bit/s
- Alarm Severity:** Major
- Alarm Notification:** (toggle)

A line graph shows bandwidth usage in bit/s over time, with a peak reaching 500 bit/s. The graph is labeled 'ecs-transitvpc-band...'.

Table 4-11 Key parameters for creating an alarm rule

Parameter	Description	Example Value
Name	Specifies the name of the alarm rule.	alarm-bandwidth

Parameter	Description	Example Value
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

- After setting the parameters, click **Create**.
- On the **Create Bandwidth Scaling Policy** page, click  on 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

- Log in to the management console.
- Under **Computing**, click **Auto Scaling**. In the navigation pane on the left, choose **Bandwidth Scaling**.
- Click **Create Bandwidth Scaling Policy**.
- 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

Parameter	Description	Example Value
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 . NOTE Currently, only pay-per-use EIPs can be scaled. Yearly/monthly EIPs cannot be used to create AS groups.	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
Scaling Action	Specifies the action to be performed. The following scaling action options are available: <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. 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. <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

Parameter	Description	Example Value
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.	300

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 an 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. You can access the bandwidth scaling policy details page by referring to [Procedure](#). 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.

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 running status. If an ECS 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 result shows that an ECS is faulty, AS removes the ECS from the AS group.
- **ELB health check:** determines ECS 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, the ECS is considered normal only when all load balancers detect that the ECS

status is normal. If any load balancer detects that the ECS is abnormal, the ECS will be removed from the AS group.

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

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 a 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 **Notification** tab and then click **Add Notification**.
3. Set the parameters listed in [Table 4-13](#).

Table 4-13 Parameter description

Parameter	Description	Example Value
Send Notification To	Select a created topic. For how to create a topic, see <i>Simple Message Notification User Guide</i> .	f123

Parameter	Description	Example Value
Notification Conditions	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 creation fails• Instance removal succeeds• Instance removal fails• Errors occur in an AS group	N/A

4. Click **Save**.

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

Operation	Resource Type	Trace Name
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
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

Operation	Resource Type	Trace Name
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.
 - **Start time and end time**: You can specify the time period to query traces.
6. Click  on 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**.
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 **Basic Information** 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 & Dimension	Monitoring Interval (Raw Data)
cpu_util	CPU Usage	CPU usage of an AS group Formula: Total CPU usage of all ECSs in an AS group/Number of ECSs in the AS group Unit: Percent	≥0%	Object: AS group Dimension: AutoScalingGroup	5 minutes

Metric ID	Metric	Description	Value Range	Monitored Object & Dimension	Monitoring Interval (Raw Data)
mem_util	Memory Usage	Memory usage of an AS group Formula: Total memory usage of all ECSs in an AS group/Number of ECSs in the AS group Unit: Percent NOTE This metric is unavailable if the image has no VM Tools installed.	≥0%	Object: AS group Dimension: AutoScalingGroup	5 minutes
network_outgoing_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 ECSs in an AS group/ Number of ECSs in the AS group Unit: Byte/s	≥0 Byte/s	Object: AS group Dimension: AutoScalingGroup	5 minutes
instance_num	Inband Outgoing Rate	Number of outgoing bytes per second on an ECS in an AS group Formula: Total inband outgoing rates of all ECSs in an AS group/ Number of ECSs in the AS group Unit: Byte/s	≥0	Object: AS group Dimension: AutoScalingGroup	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 ECSs in an AS group/Number of ECSs in the AS group Unit: Byte/s	≥0 Byte/s	Object: AS group Dimension: AutoScalingGroup	5 minutes

Metric ID	Metric	Description	Value Range	Monitored Object & Dimension	Monitoring Interval (Raw Data)
disk_write_bytes_rate	Disks Write Rate	Number of bytes written to an AS group per second Formula: Total disks write rates of all ECSs in an AS group/Number of ECSs in the AS group Unit: Byte/s	≥0 Byte/s	Object: AS group Dimension: AutoScalingGroup	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 ECSs in an AS group/Number of ECSs in the AS group Unit: Request/s	≥0 request/s	Object: AS group Dimension: AutoScalingGroup	5 minutes
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 ECSs in an AS group/Number of ECSs in the AS group Unit: Request/s	≥0 request/s	Object: AS group Dimension: AutoScalingGroup	5 minutes
cpu_usage	(Agent) CPU Usage	Agent CPU usage of an AS group Formula: Total Agent CPU usage of all ECSs in an AS group/Number of ECSs in the AS group Unit: Percent	0-100 %	Object: AS group Dimension: AutoScalingGroup	1 minute

Metric ID	Metric	Description	Value Range	Monitored Object & Dimension	Monitoring Interval (Raw Data)
mem_usedPercent	(Agent) Memory Usage	Agent memory usage of an AS group Formula: Total Agent memory usage of all ECSs in an AS group / Number of ECSs in the AS group Unit: Percent	0-100 %	Object: AS group Dimension: AutoScalingGroup	1 minute
load_average1	(Agent) 1-Minute Load Average	Average CPU load of all ECSs in an AS group in the last 1 minute Unit: none	≥0	Object: AS group Dimension: AutoScalingGroup	1 minute
load_average5	(Agent) 5-Minute Load Average	Average CPU load of all ECSs in an AS group in the last 5 minutes Unit: none	≥0	Object: AS group Dimension: AutoScalingGroup	1 minute
load_average15	(Agent) 15-Minute Load Average	Average CPU load of all ECSs in an AS group in the last 15 minutes Unit: none	≥0	Object: AS group Dimension: AutoScalingGroup	1 minute
gpu_usage_gpu	(Agent) GPU Usage	Agent GPU usage of an AS group Formula: Total Agent GPU usage of all ECSs in an AS group / Number of ECSs in the AS group Unit: Percent	0-100 %	Object: AS group Dimension: AutoScalingGroup	1 minute

Metric ID	Metric	Description	Value Range	Monitored Object & Dimension	Monitoring Interval (Raw Data)
gpu_usage_mem	(Agent) Video Memory Usage	Agent GPU memory usage of an AS group Formula: Total Agent GPU memory usage of all ECSs in an AS group / Number of ECSs in the AS group Unit: Percent	0-100 %	Object: AS group Dimension: AutoScalingGroup	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 ECSs. This section describes how to view details of AS group metrics to obtain information about the status of the ECSs in the AS group.


Prerequisites

The ECS is running properly.



 NOTE

- Monitoring metrics such as **CPU Usage** and **Disks Read Rate** are available only when there is at least one instance in an AS group. If not, only the **Number of Instances** metric is available.
- Monitoring data is not displayed for a stopped, faulty, or deleted ECS. After such an ECS 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 1, 3, and 12 hours. 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**, click **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 1, 3, and 12 hours. 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 ECS alarm rules allows you to customize the monitored objects and notification policies and determine the running status of your ECSs at any time.

Procedure

1. Log in to the management console.
2. Under **Management & Deployment**, click **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 **Finish**.

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.

5 FAQs

5.1 General

5.1.1 What Are Restrictions on Using AS?

Only applications that are stateless and can be horizontally scaled can run on instances in an AS group. AS automatically releases ECS instances. Therefore, the instances in AS groups cannot be used to save application status information (such as session statuses) and 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 service resource quotas.

Table 5-1 Quota list

Category	Description	Default Value
AS group	Maximum number of AS groups that you can create	10
AS configuration	Maximum number of AS configurations that you can create	100
AS policy	Maximum number of AS policies that can be added to an AS group	10
Instance	Maximum number of instances that can be added to an AS group	300
Bandwidth scaling policy	Maximum number of bandwidth scaling policies that you can create	10

5.1.2 Are ELB and Cloud Eye Mandatory for AS?

AS can work independently or work together with ELB and Cloud Eye.

Cloud Eye does not require additional fees and is enabled by default. You can enable the ELB service when required. For example, if distributed clusters are required, you can enable the ELB service.

5.1.3 Is AS Billed?

AS is free of charge. The pay-per-use instances automatically created in an AS group are billed. 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 be deleted. After the deletion, these instances are no longer billed. The instances manually added are removed from the AS group but still billed. If you do not need the instances, unsubscribe instances on the ECS console.

For example, two instances are created when an AS group scales out. The AS group scales in after an hour. The two instances are removed from the AS group and are billed for the one-hour usage.

5.1.4 Does an Abrupt Change on Monitoring Indicator Values Cause an Incorrect Scaling Action?

No. Monitoring data used by AS is from Cloud Eye. The monitoring interval of Cloud Eye can be set to 5 minutes, 20 minutes, or 1 hour. Therefore, an abrupt change of monitoring indicator values will not cause an incorrect scaling action.

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

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 supports 1 AS configuration and 10 AS policies at a time.

If the default configurations fail to meet your service requirements, contact the administrator.

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

Currently, AS can only scale up bandwidth.

5.1.7 What Is the AS Quota?



What Is Quota?

Quotas are enforced for service resources on the platform to prevent unforeseen spikes in resource usage. Quotas can limit the number or amount of resources

available to users, for example, how many AS groups you can create. You can apply for increasing quotas if necessary.

This section describes how to view the usage of each type of AS resource 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 Why is a message displayed indicating that the key pair does not exist and the operation is discontinued when several users under the same account operate AS resources?

A key pair cannot be used by multiple users. If the key pair of another user under the same account is configured in the AS configuration, the AS configuration cannot be used to manually provision resources.

If users need to perform operations on others' AS configuration resources without being restricted by the key pair permission, use password authentication for instances.

5.2 AS Group

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

See section "How Can I Handle an AS Group Exception?"

5.2.2 How Can I Handle an AS Group Exception?

The handling method varies depending on the possible cause.

- Issue description: Insufficient quota for ECSs, EVS disks, or EIPs.
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 Operation Will Be Suspended After An AS Group Is Disabled?

After an AS group is disabled, the group will not automatically any trigger scaling actions, but the on-going scaling action will continue. Scaling policies will not trigger any scaling actions. After you manually change the number of expected instances, no scaling action is triggered although the number of actual instances is not equal to that of expected instances.

The health check continues to check the health status of the instances but does not remove the instances.

5.3 AS Policy

5.3.1 How Many AS Policies Can Be Enabled?

Enable one or more AS policies as required.

5.3.2 What Are the Conditions to Trigger an Alarm in the 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 increase or decrease instances in the AS group.

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

A cooldown period is a period of time 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.

Before an instance is added to the AS group, it requires 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. Therefore, if an instance is put into use without cooldown after started, the system will continuously increase instances until the load decreases. After the new instances take over services, the system detects that the load is too low and decreases instances in the AS group. A cooldown prevents the AS group from repeatedly triggering unnecessary scaling actions.

The following uses an example to introduce the cooling principles:

When a traffic peak occurs, an alarm policy is triggered. In this case, AS automatically adds an instance to the AS group to help handle the added demands. However, it takes several minutes for the instance to start. After the instance is started, it takes a certain period of time to receive requests from ELB. During this period, alarms may be triggered continuously. As a result, an instance is added each time an alarm is triggered. If you set a cooldown time, after an instance is started, AS stops adding new instances according to the alarm policy until the specified period of time (300 seconds by default) passes. Therefore, 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 takes effect again.

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

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

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

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

If the ECSs are of I/O-optimized type, Cloud Eye cannot monitor the disk usage, inband incoming rate, and inband outgoing rate metrics of ECSs, regardless of whether VM tools are installed.

If VM tools are not installed on ECSs, 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 effective time of 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 Can I Install the Agent Plug-in on the Instances in an AS Group to Use Agent Monitoring Metrics?

Scenarios

When the scaling policy of an AS group is alarm-triggered, you can use Agent monitoring metrics to trigger scaling actions. Compared with basic monitoring, Agent monitoring provides system-level, proactive, and fine-grained monitoring services for instances. Before using Agent monitoring metrics, make sure that the Agent plug-in has been installed on the instances in the AS group. For details, see this section.

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 plug-in.

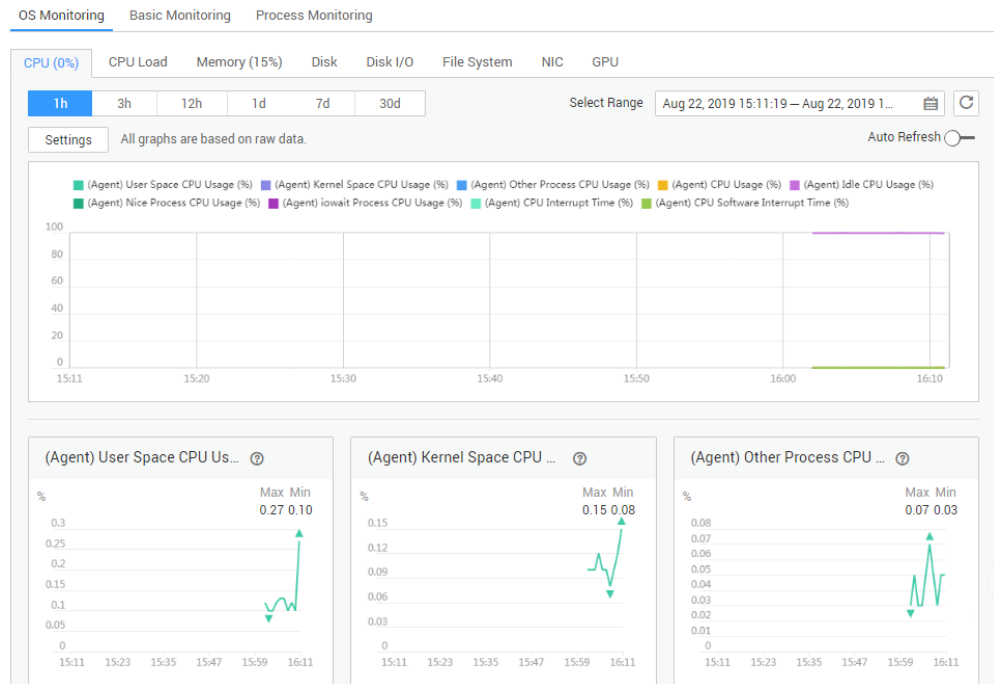
To install the plug-in, see .

3. After installing the Agent plug-in, log in to the Cloud Eye console, choose **Server Monitoring** > **Elastic Cloud Server**, and ensure that the plug-in is running and that the Agent monitoring metrics can be collected.

Figure 5-1 Checking the plug-in status

<input type="checkbox"/>	Name/ID	Private IP Address	ECS Status	Agent Status	Monitoring Status
<input type="checkbox"/>	ecs-ec78 fa07b856-2edd-4cd0-9aa6-...	192.168....	➔ Runn...	➔ Running	<input checked="" type="checkbox"/>

Figure 5-2 Viewing Agent monitoring metrics



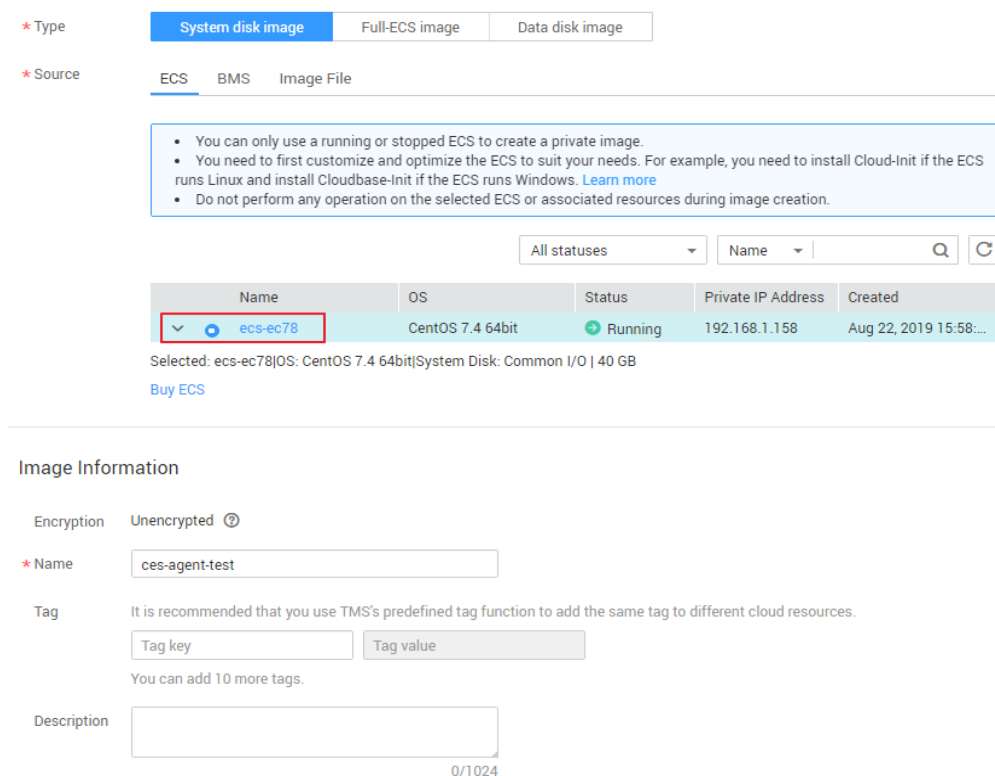
4. Add the AK/SK to the ECS **conf.json** configuration file.
 - a. Click the username, choose **My Credentials > Access Keys**, obtain the AK/SK.
 - If you have obtained the access key, obtain the AK/SK in the **credentials.csv** file saved when you created **Access Keys**.
 - If **Access Keys** is not available, click **Create Access Key** to create one. Then, save the **credentials.csv** file and obtain the AK/SK in it.
 - b. Log in to the ECS and run the **cd /usr/local/telescope/bin** command to go to the Agent installation directory.
 - c. Run the **vi conf.json** command to open the configuration file and enter the obtained AK/SK.

```
{
  "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 has been installed during ECS creation, the AK/SK has been added during user data injection. You only need to check the AK/SK in this step.

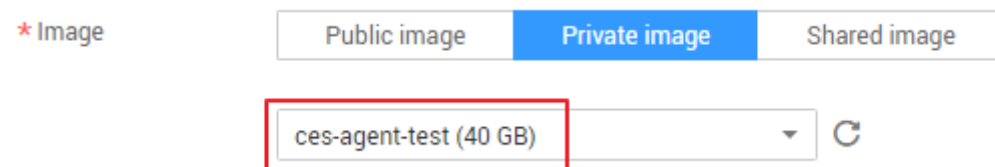
- d. Press **Esc** and enter **:wq** to save and exit the configuration file.
5. Go to the **Image Management Service** page and use this ECS to create a private image. For details, see [Creating a Private Image](#).

Figure 5-3 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.

Figure 5-4 Selecting the private image



Click **Private image** for **Image**, select **ces-agent-test** from the drop-down list, and set other parameters as required.

7. Create an AS group and associate the AS configuration created in step 6 with the 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**.

Figure 5-5 Selecting a trigger condition

The screenshot shows the 'Add AS Policy' dialog box with the following configuration:

- Policy Name:** as-policy-stk5
- Policy Type:** Alarm (selected), Scheduled, Periodic
- Alarm Rule:** Create (selected), Use existing
- Rule Name:** as-alarm-agent-test
- Monitoring Type:** System monitoring (selected), Custom monitoring
- Trigger Condition:** CPU Usage (selected), Max., >, %
- Monitoring Interval:** (Agent) CPU Usage (selected), (Agent) Memory Usage, (Agent) 1-Minute Load Average, (Agent) 5-Minute Load Average, (Agent) 15-Minute Load Average
- Consecutive Occurrences:** ?
- Scaling Action:** 1 instances

Buttons: OK, Cancel

9. Manually add the ECS created in step 2 to the AS group.
10. Perform the following operations to check whether the Agent monitoring metric takes effect:
 - Verify that the Agent monitoring metric is displayed on the **Monitoring** tab of the page providing details about the AS group.
 - When the alarm threshold is reached, verify that the alarm policy is triggered on the **Scaling Actions** tab of the page providing details about the AS group and that instances are added for capacity expansion.
 - The Agent monitoring data is available for 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 Removed Automatically?

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 ensure this:

Method 1

Perform following configurations in the AS group:

- Set the minimum number of instances in the AS group to N or greater than N.
- 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 do not correspond to any AS configuration (because they are not created using the AS configuration). Therefore, the instances automatically created using the AS configuration are removed at first. Only when such instances are removed, the instances manually added are removed. Since you have set the minimum number of instances to N or greater than N, the instances manually added are not selected.

Note: If the instances manually added are stopped or they malfunction, they are regarded as unhealthy and removed from the AS group. This is because health check ensures that instances in the AS group must be healthy.

Method 2

Enable instance protection for the N instances.

You can enable instance protection for the N instances at the same time. When an AS group reduces the capacity, protected instances will not be removed from the AS group. Note: Instances that fail to pass a health check will still be removed from the AS group.

5.4.2 What Are the Sequence of Selecting Flavors in Multi-Flavor AS Configuration?

When multiple flavors are selected in an AS configuration, the sequence varies according to the extension policies of a single AZ and multiple AZs. This section describes the sequence of selecting flavors in a single AZ and multiple AZs.

Single AZ

If only one AZ is selected for an AS group, all instances in the AS group are created in the AZ. If multiple flavors are selected in the AS configuration, there are two flavor selection policies:

- **Sequenced:** During AS group expansion, flavors are used based on the sequence they are selected. For example, you have selected flavors 2, 3, and 1 in sequence. The system selects flavor 2 at first. If flavor 2 is insufficient or an instance fails to be created due to other reasons, the system selects flavor 3. Flavor 1 is used only when flavor 2 and 3 cannot be used.
- **Cost-center:** During AS group expansion, the flavor with the minimum cost comes first. For example, you select flavors 1, 2, and 3 in sequence. The costs of flavors 1, 3, and 2 are in descending order. The system preferentially selects flavor 2 (with the minimum cost) to create an instance. When flavor 2 fails, select flavor 3. When flavor 3 also fails, flavor 1 is used.

Multiple AZs

When two or more AZs are selected for an AS group, you need to configure the **Multi-AZ Extension Policy** (load-balanced or sequenced). When you select

different multi-AZ extension policies, the sequence of creating instances is also different. The sequence is described as follows:

- **Load-balanced:** When expanding an AS group, preferentially ensure that ECSs are evenly distributed in AZs. If it fails in the target AZ, the system automatically selects another AZ based on the sequenced policy. The following is an example of selecting AZs and flavors:

You have selected AZ 1, AZ 2, and AZ 3 in sequence and flavors 1, 2, and 3. The priority sequence of the flavors is 2, 3, and 1. AZ 1, AZ 2, and AZ 3 have 3, 2, and 3 instances respectively. According to the load-balanced policy, AZ 2 has fewer instances and is therefore preferentially selected to create instances. Use flavor 2 to create an instance in AZ 2. If the instance is successfully created, the scaling action is successful. If flavor 2 fails, use flavor 3, and so on. If all of them fail, the instance cannot be created in AZ 2. If instances cannot be created using load-balanced policy, select other AZs based on the sequenced policy and try to create instances using flavors 2, 3, and 1 in AZ 1. If ECSs still cannot be created in AZ 1, AZ 3 is selected and the sequence of selecting flavors is also 2, 3, and 1.

- **Sequenced:** When expanding the ECS capacity, the target AZ is used based on the order in which AZs are selected. The following is an example of selecting AZs and flavors:

You have selected AZ 1, AZ 2, and AZ 3 in sequence and flavors 1, 2, and 3. The priority sequence of the flavors is 2, 3, and 1. No matter whether instances in three AZs are evenly distributed, the system creates instances in sequence, that is, AZ 1, AZ 2, and AZ 3. Use flavor 2 to create an instance in AZ 1. If this fails, use flavor 3. Flavor 1 is used when both flavors 2 and 3 fail. If all three flavors fail in AZ 1, AZ 2 is selected. The sequence of flavors is also 2, 3, and 1. Similarly, if AZ 2 also fails, AZ 3 is selected. The sequence of flavors is also 2, 3, and 1.

NOTE

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

5.4.3 Will the Application Data on an Instance Be Retained After the Instance Is Removed from an AS Group and Deleted?

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

5.4.4 Can I Add ECSs Charged in Yearly/Monthly Mode?

Yes. Currently, AS automatically creates pay-per-use ECSs by default. In addition, you can manually add ECSs charged in yearly/monthly or pay-per-use mode.

5.4.5 Can Instances That Have Been Added Based on an AS Policy Be Automatically Deleted When They Are Not Required?

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

5.4.6 What Is the Expected Number of Instances?

The expected number of instances refers to the number of ECSs 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 ECSs 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 ECSs is not consistent with the expected number. A scaling action is performed to adjust the number of ECSs to the expected number.

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

5.4.7 How Do I Delete an ECS 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 target 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 target 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 target 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 reduce instances as needed or set the number of instances to a specified value.

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 target 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 Be Deleted When the ECS Becomes Faulty?

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

5.4.9 How Should I Handle Abnormal Instances in an AS Group?

In normal cases, you do not need to handle instances in abnormal state because the AS service periodically checks the health status of instances in an AS group. When an AS group is enabled, abnormal instances are removed and new instances are created to ensure that the number of expected instances is the same as current instances. When an AS group is disabled, checking instance health status continues. However, AS will not remove instances.

It should be noted that if the ELB health check mode is selected, ELB sends heartbeat messages to backend ECSs through an intranet. Therefore, to ensure that the ELB health check can run properly, ensure that your ECS can be accessed through an intranet. Perform the following steps to check this:

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 to be checked.
 - **Check Path**: If HTTP is used for the health check, ensure that the health check path for backend ECSs is correct.
2. Check whether software (such as firewall) on the ECS masks the source IP address performing the health check.
3. Check whether the rules of backend ECS security groups and network ACL allow access by 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 health check mode is used, service ports of the backend ECSs must be enabled.
 - If the health check port is different from service ports of the ECSs, communication between the service ports of the ECSs and health check port must be enabled.
4. If the issue persists, contact technical support.

5.4.10 What Can I Do If Instances in an AS Group Frequently Fail in Health Checks and Are Deleted and Then Created Repeatedly?

The security group rule of the instance must allow communication with the 100.125.0.0/16 network, and the protocol and port number must be the same as those used by ELB for health checks. Otherwise, the health check will fail. As a result, the instances will be deleted and created again and again.

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

You can enable instance protection for in-service ECSs in an AS group. After the configuration, when AS automatically reduces the number of ECSs in an AS group, the in-service ECSs with instance protection enabled will not be removed. You can also set the minimum number of instances for an AS group and the instance removal policy to ensure that an AS group has some in-service ECSs.

Unhealthy ECSs are removed from an AS group and new ECSs are created. Therefore, do not stop or delete ECSs that have been added to an AS group on the ECS console. The stopped or deleted ECSs are considered unhealthy and automatically removed from the AS group. Even when an AS group is disabled, AS checks the status of ECSs in the AS group. In this case, however, unhealthy ECSs will not be removed from the AS group.

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

If an instance automatically added to an AS group is protected, it is only 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 storage DR.

5.5 Others

5.5.1 What Can I Do to Enable My Application to Be Automatically Deployed on an Instance?

To enable automatic application deployment on instances automatically added to an AS group, you need to create a private image which contains application

software and automatic startup settings. When creating an AS group, select the private image you have created for the AS configuration. In this way, applications will be automatically deployed on instances that are added to the AS group. The procedure is as follows:

1. Before creating a private image, install the application and set it to automatically start upon system startup on the ECS which you will use to create the private image.
2. Create a private image. For details, see [Image Management Service User Guide](#).
3. Create an AS configuration. For details, see [Creating an AS Configuration](#). Ensure that the image in the AS configuration is the private image created in [2](#).
4. Click the **AS Groups** tab and then click the name of the target AS group.
5. Click **Change Configuration** on 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 the required application has been installed on the instances. If any issue occurs, 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 customized configurations, such as the hostname, key pair, and user data, of a newly created ECS. When creating an AS configuration, you can use Cloud-Init to initialize the ECS.

If Cloud-Init or Cloudbase-Init is not installed in the private image specified in an AS configuration for an AS group, the following cases occur on the instance created in a scaling action:

- If an ECS created using a Windows private image without Cloudbase-Init installed is used, the system will display a message indicating that the password for logging in to the ECS cannot be obtained when you obtain the password. In such a case, log in to the ECS using only the image password. If you forget the image password, use the password resetting function available on the **Elastic Cloud Server** page to reset the password.
- If an ECS created using a Linux private image without Cloud-Init installed is used, the ECS cannot be logged in using the password or key pair configured during ECS creation. In such a case, log in to the ECS using only the image password or key pair. If you forget the image password or key, use the password resetting function available on the **Elastic Cloud Server** page to reset the password.
- If an ECS created using a private image without Cloud-Init or Cloudbase-Init installed is used, user data injection fails.

To prevent the preceding issues from occurring, check whether the private image in the AS configuration has Cloud-Init or Cloudbase-Init installed before using the AS service. Delete the AS configurations that use the private images without Cloud-Init or Cloudbase-Init installed. Use the private images with Cloud-Init or Cloudbase-Init installed to create new AS configurations. 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 and query the AS configuration list.
- 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 in the AS group to the newly created one.

5.5.3 How Can I Run Existing Services on an Instance Newly Added to an AS Group?

5.5.4 Why Cannot 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 in the AS configuration of the AS group is your private one, and the Cloud-Init tool had not been installed when you created the private image.

If the Cloud-Init tool had not been installed when you created a private image, you would fail to customize the ECS configuration. In such a case, you can log in to the ECS only using the original image password or key pair.

Handling Method

1. Check whether the ECS must 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. Modify the AS configuration of the AS group.

NOTE

Make sure that the Cloud-Init or Cloudbase-Init tool has been installed in the image of the modified AS configuration. For instructions about how to install the Cloud-Init or Cloudbase-Init tool, see *Image Management Service User Guide*.

After the AS configuration is modified, you can use the key file to log in to the new ECSs that are added when the AS action is performed in the AS group. In such a case, you do not need to use the original image password or key pair to log in to the new ECSs any more.

5.5.5 Do I Need to Configure an EIP in an AS Configuration When A Load Balancer Has Been Enabled in an AS Group?

No. If you have enabled a load balancer in an AS group, you do not have to configure an EIP in the AS configuration. The system automatically adds instances

in the AS group to the load balancer. These instances will provide services via the EIP bound to the load balancer.

5.5.6 How Can I Enable Automatic Initialization of EVS Disks of Instances That Have Been Added in a Scaling Action to an AS Group?

Scenarios

After an ECS is created, EVS disks attached to the ECS must be initialized. If multiple ECSs are added to the AS group in scaling actions, you must manually initialize the EVS disks of each ECS, which requires a long period of time.

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

This section uses CentOS 6.5 as an example. For details about how to configure DHCP 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 storing the script:

```
cd /script directory
```

An example is as follows:

```
cd /home
```

3. Run the following command to create a script:

```
vi script name
```

An example is as follows:

```
vi fdisk_mount.sh
```

4. Press **i** to go to the script editing page.

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 go to the file editing page.
The drive letter, size, and 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 the **rc.local** file:

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

After the **rc.local** file 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 ECSs added in scaling actions to this AS group will be automatically initialized based on the configuration in the private image.

A Change History

Released On	Description
2020-11-03	This issue is the first official release.