



Computing Cloud Service – Auto Scaling



Foreword

- This chapter provides an overview of Auto Scaling (AS) and its basic functions, application scenarios, and usages.



Objectives

- Upon completion of this course, you will:
 - Be familiar with AS concepts, functions, and application scenarios.
 - Be able to create AS groups and bandwidth scaling policies.



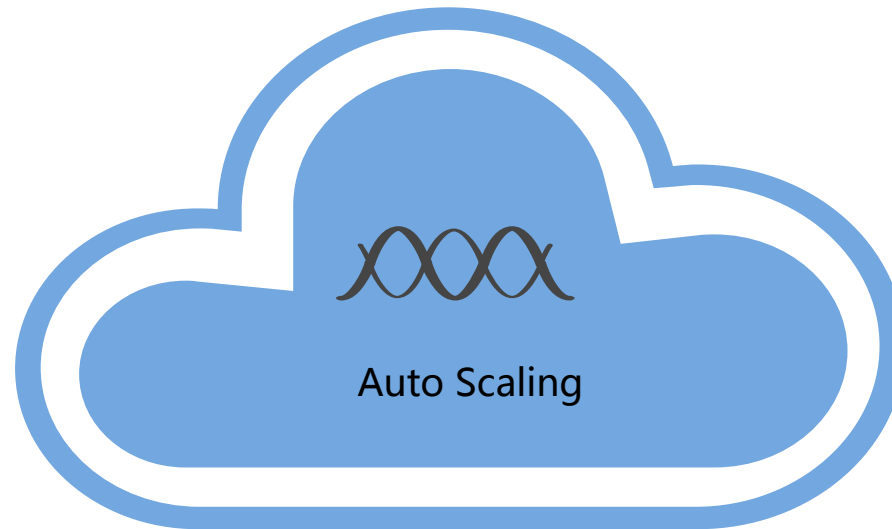
Contents

- 1. Overview**
2. Creating an AS Group
3. Creating a Bandwidth Scaling Policy
4. Usage and Management
5. Related Services



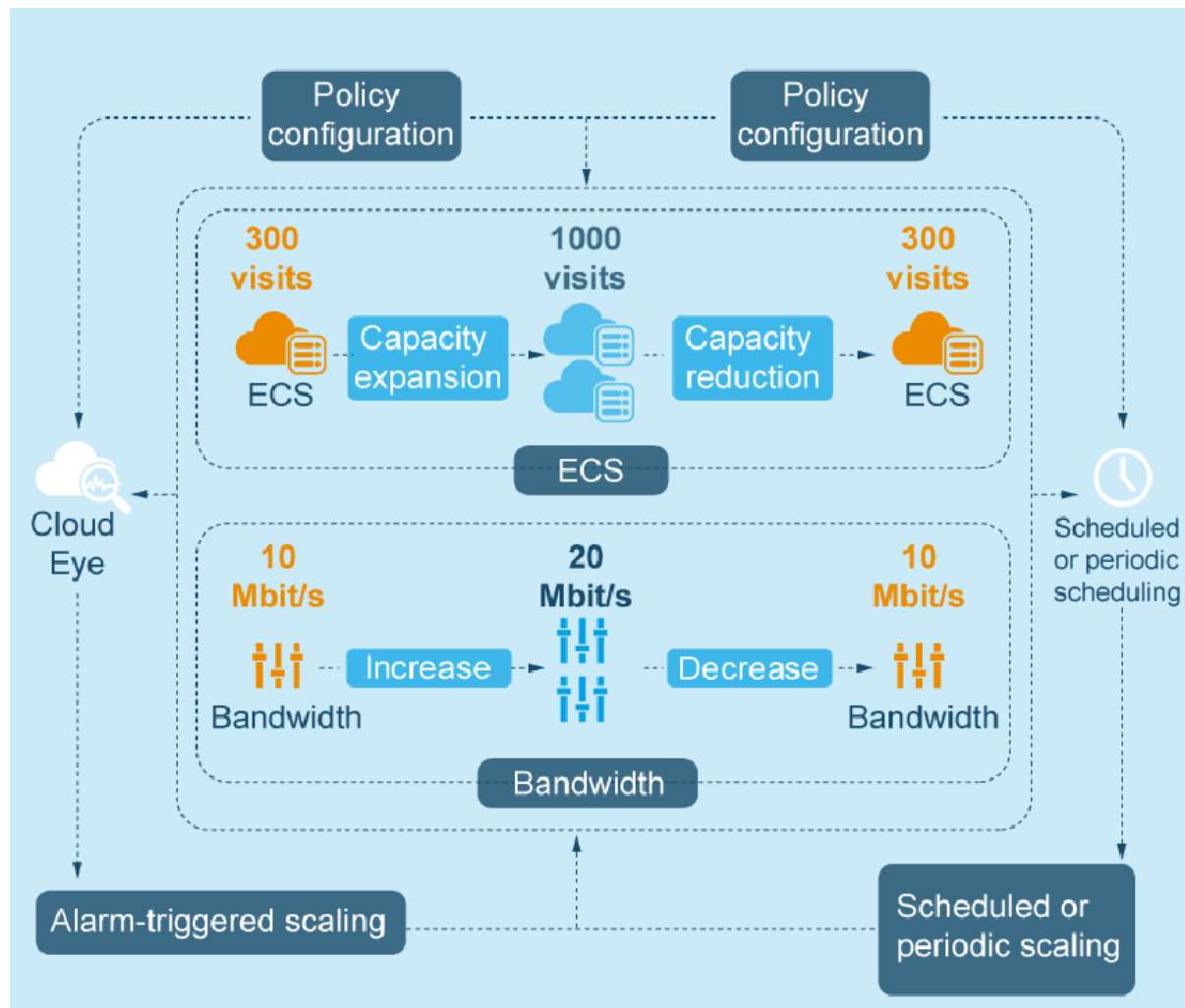
Concepts

- AS automatically adjusts resources based on your service needs and allows you to specify AS configurations and policies as required. These configurations and policies free you from having to repeatedly adjust resources to keep up with service changes and demand spikes, thereby reducing the resources and manpower required.



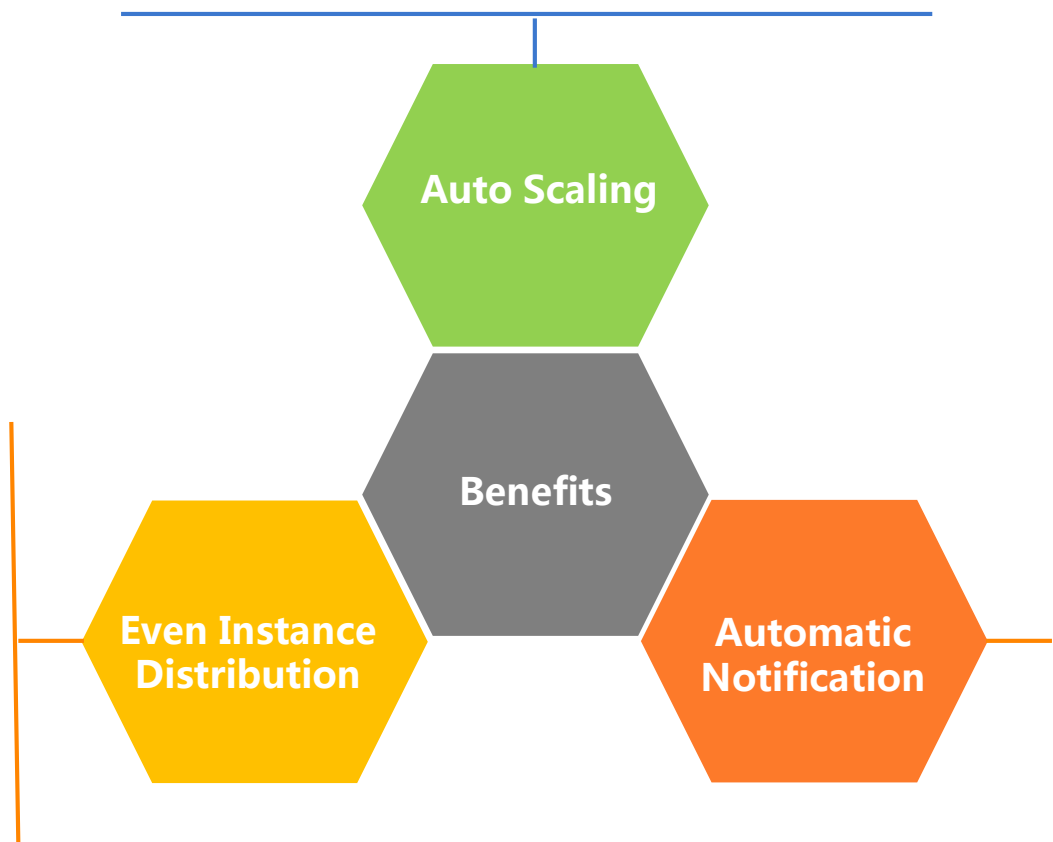


Product Architecture



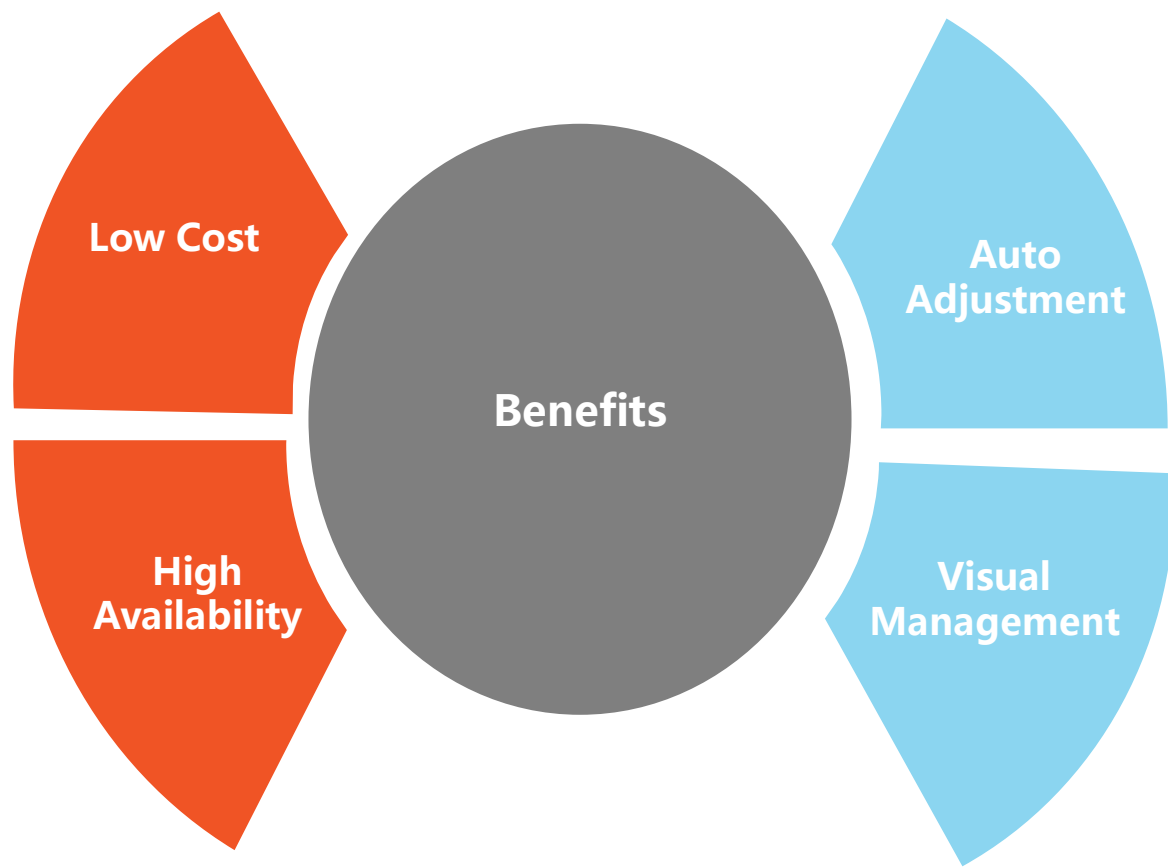


Features





Product Advantages





Application Scenarios

Typical Application Scenario	Description
Web app service	The AS service scales up or down logical servers of common web services, such as enterprise websites, e-commerce platforms, video websites, online education institutions, and mobile apps. Requests from clients are distributed among app servers through load balancing. The AS service scales up or down app servers according to the number of requests. If you enable the bandwidth scaling function, AS will adjust the bandwidth size based on access traffic.
High-performance cluster deployment	The AS service scales up or down distributed backend servers of common web services in real time based on the data volume. The servers include distributed big data computing nodes and data retrieval servers in computing clusters.
Request server deployment	The AS service is used for deploying server clusters that are used to send requests or collect data. These servers are time-effective. The AS service enables quick creation, deployment, and scaling of these servers.



Contents

1. Overview
- 2. Creating an AS Group**
3. Creating a Bandwidth Scaling Policy
4. Usage and Management
5. Related Services



Wizard-based Creation Process





Configuring Parameters - Creating an AS Group

Create AS Group ?

[< Back to AS Group List](#)

1 Specify AS Group Details

2 Add AS Configuration

Region

CN North-Beijing1

AZ ?

AZ1 × AZ2 × AZ3 ×

Name

as-group-sjc7

Max. Instances

1

Expected Instances ?

0



Min. Instances


0



Configuring Parameters - Creating an AS Group

The following information shows the VPC and subnet where the instances created during scaling reside. Load balancing automatically distributes traffic across instances.

VPC   [View VPC](#)

Subnet  [View Subnet](#)


If you select multiple subnets, instances created during scaling will have multiple NICs belonging to different subnets.


Load Balancing ☒ Do not use ☐ Classic load balancer ☐ Enhanced load balancer


Instance Removal Policy

EIP ☒ Release ☐ Do not release

If you select Release, EIPs bound to ECSs are released when the ECSs are removed from the AS group. Otherwise, EIPs will only be unbound from the ECSs.

Health Check Method 

Health Check Interval 

Health Check Grace Period (s) 

Advanced Settings ☒ Do not configure ☐ Configure now

You can configure the notifications and tags.



Parameters for Creating an AS Group

Parameter	Description	Example Value
Max. Instances or Min. Instances	Specifies the maximum or minimum number of instances in an AS group.	10 or 5
Expected Instances	Specifies the expected number of ECS instances in an AS group.	6
AZ	Specifies a physical region where resources use independent power supplies and networks. AZs are physically isolated but interconnected through an internal network.	None
VPC	Specifies the VPC of the ECS network. All ECSs in an AS group belong to the same VPC.	None
Subnet	By default, only ECSs in the same VPC subnet can communicate with each other.	None
Security Group	You can define different access rules for a security group to protect the ECSs that are added to this security group.	None
Load Balancing	This parameter is optional. A load balancer automatically distributes access traffic to all ECSs in an AS group to balance their service load. It enables higher levels of fault tolerance in your applications and expands application service capabilities.	None
Health Check Method	Checks the ECS health status. When the health check detects a faulty ECS, the system removes the faulty ECS from the AS group and adds a new one. The health check supports two modes, respectively ECS health check and ELB health check.	None
Health Check Interval	Specifies the health check period for an AS group.	5 minutes
Instance Removal Policy	Specifies the priority for removing an ECS instance. If required conditions are met, scaling actions are triggered to remove instances.	None
Release EIP on Instance Removal	If the AS configuration of an AS group uses an EIP, the system binds the EIP to the newly created ECS instance when the scaling action is performed. If you select Yes , the EIP bound to the ECS instance is released when the instance is removed from the AS group. Otherwise, the system reserves the EIP.	None



Configuring Parameters - Creating an AS Configuration

☒ Specify AS Group Details 2 Add AS Configuration 3 (Optional) Add AS Policy

1 After the AS group is created, you can change the AS configuration as required.
For fine-grained monitoring of ECS instances, install the Cloud Eye agent in the AS configuration image. [Learn more](#)

AS Configuration

Use existing

Create

Name

as-config-qn08

Configuration Template

Create a new specifications template

Use specifications of an existing ECS

Specifications

Latest generations

vCPUs All

Memory All

Enter a specification name.

General computing

General computing-plus

Memory-optimized

Large-memory

High-performance computing

Ultra-high performance computing

Disk-intensive

Ultra-high I/O

GPU-accelerated

General computing-basic

[Learn more](#) about ECS types.

Flavor name	vCPUs/Memory	Assured/Maximum Bandwidth	PPS
<input checked="" type="radio"/> s2.small.1 (Sold out in cn-north-1c)	1 vCPUs 1 GB	0.1/0.5 Gbps	50 Kpps
<input type="radio"/> s2.medium.2 (Sold out in cn-north-1b, cn-north-1c)	1 vCPUs 2 GB	0.1/0.5 Gbps	50 Kpps
<input type="radio"/> s2.medium.4 (Sold out in cn-north-1c)	1 vCPUs 4 GB	0.1/0.5 Gbps	50 Kpps
<input type="radio"/> s2.large.2 (Sold out in cn-north-1b, cn-north-1c)	2 vCPUs 4 GB	0.2/0.8 Gbps	100 Kpps



Configuring Parameters - Creating an AS Configuration

Image

Public image

Private image

Shared image

--Select an OS--

--Select an OS version--

Disk

EVS

System Disk

Common I/O

-

100

+

GB | IOPS Limit 700, IOPS Burst Limit 2,200 IOPS

+ Add Data Disk

You can add 23 more disks.

Security Group

For your convenience, the security group configuration has been moved to this page. [Learn how](#) to configure a security group.

Sys-default (Inbound:TCP/93...

Create Security Group

Inbound: TCP/9300, 9200, 3389, 22, 443, 80, 8888; UDP/53 | Outbound: UDP/53; TCP/443

EIP

Do not use

Automatically assign

An ECS without an EIP cannot access the Internet. However, it can still be used as a service ECS deployed in a cluster or on a private network.

Login Mode

Key pair

Password

To log in to, reinstall, or change the OS of an ECS, you must have the private key. Keep this key safe.

Key Pair

KeyPair-1589_demo

View Key Pair

☐

I acknowledge that I have the private key file KeyPair-1589_demo.pem and that I will not be able to log in to my ECS without this file.

Advanced Settings

Do not configure

Configure now



Parameters for Creating an AS Configuration

Parameter	Description	Example Value
Configuration Name	Specifies the name of an AS configuration.	None
Configuration Template	Select Create a new specifications template . If this option is selected, configure the parameters, such as ECS Type , vCPU , Memory , Image , and Disk to create an AS configuration.	Create a new specifications template
Specifications	The public cloud provides various ECS types for you to select based on application scenarios.	Memory-optimized
Image	Images are classified into public images, private images, and shared images.	Public image
Disk	The disk, also called the EVS disk, can be a system disk or a data disk. The disk type includes common I/O, high I/O, and ultra-high I/O.	Common I/O
Security Group	Controls ECS access within or between security groups to enhance security protection on ECSs.	None
EIP	Specifies a static public IP address bound to an ECS in a VPC. Using the EIP, the ECS provides services externally. The system provides the following options: Do not use: Without an EIP, the ECS cannot access the Internet and is used only in the private network or in a cluster. Automatically assign: The system automatically assigns an EIP for the ECS. The EIP provides exclusive bandwidth that is configurable.	Automatically assign
Login Mode	A key pair is used for authenticating the ECS. In this mode, create or import a key pair on the Key Pair page.	Key pair
Advanced Settings	This parameter allows you to configure File Injection , User Data Injection , and ECS Group . You can select Do not configure or Configure now .	None



Configuring Parameters - Adding an AS Policy

Add AS Policy

Policy Name

as-policy-qcw4

Policy Type

Alarm

Scheduled

Periodic

Monitoring Type

System monitoring

Custom monitoring

Alarm

Create Alarm Rule

View Alarm Rule

Alarm Name

as-alarm-a15n

Trigger Condition

CPU Usage

Max.

>

%

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

Monitoring Interval

5 minutes

Consecutive Occurrences ?

Scaling Action

Add

1

instances

Cooldown Period (s) ?

900

OK

Cancel



Parameters for Adding an AS Policy

Parameter	Description	Example Value
Policy Name	Specifies the name of an AS policy.	as-policy-p6g5
Policy Type	The value can be Alarm, Scheduled, or Periodic .	Alarm
Monitoring Type	Specifies the alarm monitoring type. The value can be System monitoring or Custom monitoring .	System monitoring
Alarm	<p>You can use an existing alarm rule or create an alarm rule. To create an alarm rule, configure the following parameters:</p> <p>Alarm Name: specifies the name of the new alarm rule, for example, as-alarm-7o1u.</p> <p>Trigger Condition: specifies a metric and condition for triggering a scaling action. For example, when the CPU usage becomes higher than 70%, AS automatically triggers a scaling action.</p> <p>Monitoring Interval: specifies the period for the metric, for example, 5 minutes.</p> <p>Consecutive Occurrences: specifies the number of consecutive times, for example, one time, for triggering a scaling action during a monitoring period.</p>	None
Scaling Action	<p>Specifies an action and the expected number of instances. The following AS action options are available:</p> <p>Add: adds instances to an AS group when the scaling action is performed.</p> <p>Reduce: removes instances from an AS group when the AS action is performed.</p> <p>Set to: sets the expected number of instances in an AS group to a specified value.</p>	Add 1 instance
Cooldown Period	<p>Specifies a period of time after a scaling action starts and before any further scaling actions can be triggered.</p> <p>The cooling duration prevents alarm-triggered scaling actions. Scaling actions triggered at a scheduled time or periodically will not be affected. However, the AS service restarts the cooling duration in seconds after a scheduled or periodic scaling action is performed.</p>	900s



Contents

1. Overview
2. Creating an AS Group
- 3. Creating a Bandwidth Scaling Policy**
4. Usage and Management
5. Related Services



Configuring Parameters - Creating a Bandwidth Scaling Policy

Create Bandwidth Scaling Policy ?

< Back to Bandwidth Scaling Policy List

Region

CN North-Beijing1

Policy Name

as-policy-55q9

EIP

49.4.80.142

View EIP

Bandwidth Size

10 Mbit/s

Policy Type

Alarm

Scheduled

Periodic

Alarm

Create Alarm Rule

View Alarm Rule

Alarm Name

as-alarm-zthe

Trigger Condition

Inbound Bandwidth

Max.

>

bit/s

Monitoring Interval

5 minutes

Consecutive Occurrences ?

Scaling Action

Set to

1

Mbit/s



Parameters for Creating a Bandwidth Scaling Policy

Parameter	Description	Example Value
Region	Specifies the region where the AS group resides.	None
Policy Name	Specifies the name of the bandwidth scaling policy.	None
EIP	Specifies the public network IP address whose bandwidth needs to be scaled.	None
Policy Type	The value can be Alarm , Scheduled , or Periodic .	Alarm
Alarm	<p>You can use an existing alarm rule or create a new one.</p> <p>To create an alarm rule, configure the following parameters:</p> <p>Alarm Name: specifies the name of the new alarm rule, for example, as-alarm-7o1u.</p> <p>Trigger Condition: specifies a monitoring metric and condition for triggering a scaling action. For example, when the CPU usage becomes higher than 70%, AS automatically triggers a scaling action.</p> <p>Monitoring Interval: specifies the interval (such as five minutes) at which the alarm status is updated based on the alarm rule.</p> <p>Consecutive Occurrences: specifies the number of sampling points when an alarm is triggered.</p>	Create Alarm Rule
Scaling Action	<p>Specifies an action and the number/percentage of instances.</p> <p>The following AS action options are available:</p> <p>Add Reduce Set to</p>	Add 1 instance
Limit	Specifies the maximum and minimum bandwidth allowed (Mbit/s).	2000 Mbit/s
Cooldown Period	Specifies a period of time after a scaling action starts and before any further scaling actions can be triggered. The cooling duration prevents alarm-triggered scaling actions. Scaling actions triggered at a scheduled time or periodically will not be affected. However, the AS service restarts the cooling duration in seconds after a scheduled or periodic scaling action is performed.	900s



Contents

1. Overview
2. Creating an AS Group
3. Creating a Bandwidth Scaling Policy
- 4. Usage and Management**
5. Related Services



Management Overview

- AS Groups
- AS Configurations
- Scaling Actions
- Scaling Bandwidth Policies
- AS Group and Instance Monitoring
- Constraints



AS Groups

- An AS group consists of a collection of ECS instances and AS policies that have similar attributes and apply to the same application scenario.
 - Creating an AS group
 - Adding a load balancer to an AS group
 - Adding/Replacing an AS configuration to/in an AS group
 - Enabling an AS group
 - Disabling an AS group
 - Modifying an AS group
 - Deleting an AS group



AS Configurations

- An **AS configuration is a template** listing specifications for the instances that will be added to an AS group.
 - Using an existing ECS to create an AS configuration
 - Using a new specifications template to create an AS configuration
 - Copying an AS configuration
 - Deleting an AS configuration



Resource Expansion

- When service demands increase, you need to expand resources through scaling actions.
- There are three methods for resource expansion:
 - Dynamically expanding resources
 - Expanding resources as planned
 - Manually expanding resources



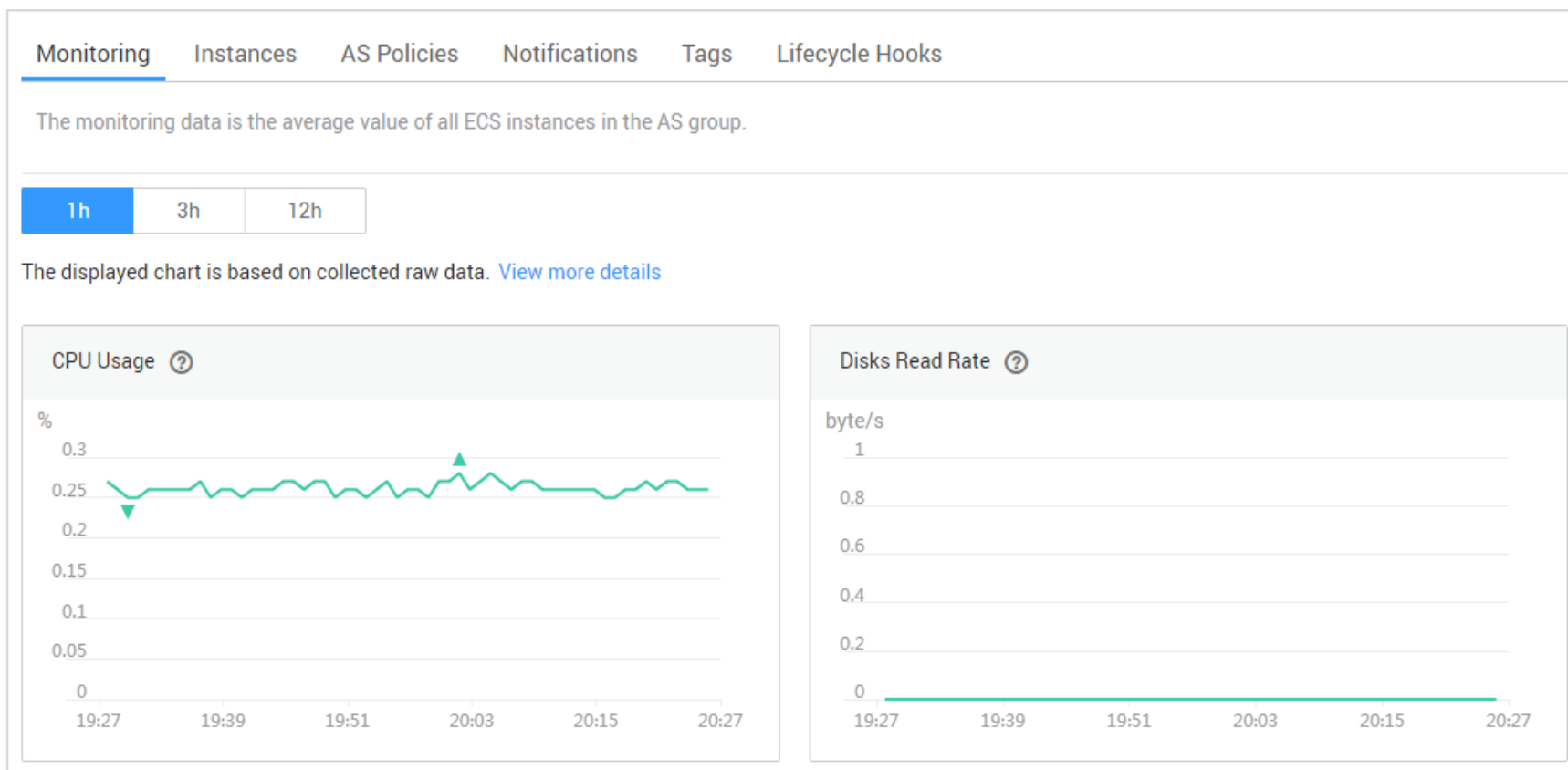
Scaling Actions - Configuring an Instance Removal Policy

- AS supports the following instance removal policies:
 - Oldest instances created based on the oldest configuration
 - Newest instances created based on the oldest configuration
 - Oldest instances
 - Newest instances



Scaling Actions – Viewing a Scaling Action

- On the AS group details page, click the **Monitoring** tab, and click **Diagram** or **Table** to view scaling actions. Below are some diagrams.





Scaling Actions - Adding a Lifecycle Hook

- After a lifecycle hook is added to an AS group, when the AS group performs a scaling action, the lifecycle hook suspends the instance that is being added to or removed from the AS group and sets the instance to the waiting state. During the waiting period, you can perform customized operations on the instance. For example, you can install or configure software on the newly started instance, or download the log file from the instance before the instance terminates.
 - Adding a lifecycle hook
 - Modifying a lifecycle hook
 - Deleting a lifecycle hook
 - Perform a callback action



Scaling Actions - Managing AS Policies

- An AS policy specifies conditions for triggering an AS action. An AS action will be triggered if conditions are met. The AS service allows you to:
 - Create an AS policy
 - Modify an AS policy
 - Delete an AS policy
 - Enable an AS policy
 - Disable an AS policy
 - Manually execute an AS policy



AS Group and Instance Monitoring

A health check removes abnormal instances from an AS group. Then, the AS group creates new instances so that the number of instances is the same as the number before instance removal. There are two types of AS group health checks.

- **ECS health check:** checks the ECS running status. If an ECS is stopped or deleted, it is considered as abnormal. The AS group automatically removes the abnormal instances.
- **ELB health check:** checks the ECS running status based on the health check result obtained using a load balancing listener. After you add multiple elastic load balancers to an AS group, the AS group will remove the ECSs once one of the load balancers detects that the ECSs are abnormal.



Constraints

AS has the following restrictions:

- Only applications that are stateless and can be horizontally scaled can run on ECS instances in an AS group.
- The following table lists the constraints on AS resources.

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	50



Contents

1. Overview
2. Creating an AS Group
3. Creating a Bandwidth Scaling Policy
4. Usage and Management
- 5. Related Services**



Related Services

- Elastic Cloud Server
- Virtual Private Cloud
- Elastic Load Balance
- Simple Message Notification
- Cloud Trace Service
- Cloud Eye



Quiz

1. Which of the following methods does AS use to expand resources?
 - A. Dynamically expanding resources
 - B. Expanding resources as planned
 - C. Manually expanding resources
 - D. Automatically expanding resources
2. Which of the following policies does AS support?
 - A. Alarm policies
 - B. Scheduled policies
 - C. Periodic policies
 - D. Monitoring policies



Quiz

1. Which of the following methods does AS use to expand resources?
 - A. Dynamically expanding resources
 - B. Expanding resources as planned
 - C. Manually expanding resources
 - D. Automatically expanding resources
2. Which of the following policies does AS support?
 - A. Alarm policies
 - B. Scheduled policies
 - C. Periodic policies
 - D. Monitoring policies



Summary

During this chapter, we covered:

- Concepts, functions, and application scenarios of the AS service
- Creation and management of AS groups and bandwidth scaling policies



More Information

Abbreviation	Full Name
AS	Auto Scaling
ELB	Elastic Load Balance

The background of the image shows silhouettes of several groups of business professionals in a modern office environment. They are standing on a highly reflective floor, and their reflections are clearly visible. The entire scene is overlaid with a semi-transparent blue filter. In the center, the text "Thank You" is written in a large, white, sans-serif font, with the website address "www.huawei.com" in a smaller, white, sans-serif font directly below it.

Thank You

www.huawei.com