Auto Scaling

Service Overview

Issue	05
Date	2022-11-15

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Security Declaration

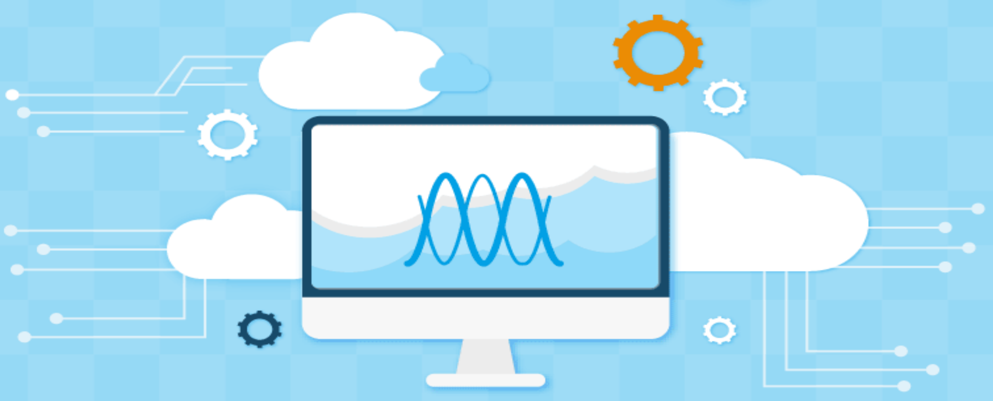
Vulnerability

Huawei's regulations on product vulnerability management are subject to the *Vul. Response Process.* For details about this process, visit the following web page:
https://www.huawei.com/en/psirt/vul-response-process
For vulnerability information, enterprise customers can visit the following web page:
https://securitybulletin.huawei.com/enterprise/en/security-advisory

Contents

1 AS Infographics

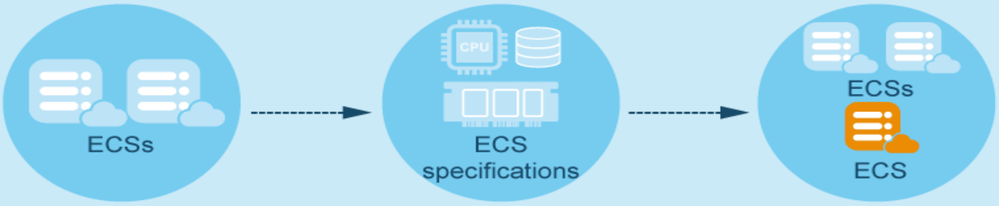
Auto Scaling



01 What Is AS?

Auto Scaling (AS) automatically adjusts resources to keep up with changes in demand based on pre-configured AS policies.

You can specify AS configurations and policies based on service requirements. These configurations and policies free you from having to repeatedly adjust resources to keep up with service changes and spikes in demand. In this way, AS helps you reduce the resources and manpower required.



Create an AS group.

Create an AS configuration and policy.

AS automatically scales out ECSs.

02 Advantages of AS

? During the 11.11 Shopping Festival in China on November 11, it is extremely hard to keep up with the massive increase in demand. Traditional solutions are not up to the task.

Provision as many as servers as required for peak demand. All that capacity will be wasted during off-peak hours.

Provision servers based on average loads of applications. Capacity will be insufficient during peak hours.

AS perfectly resolves this issue.

2 What Is Auto Scaling?

AS Introduction

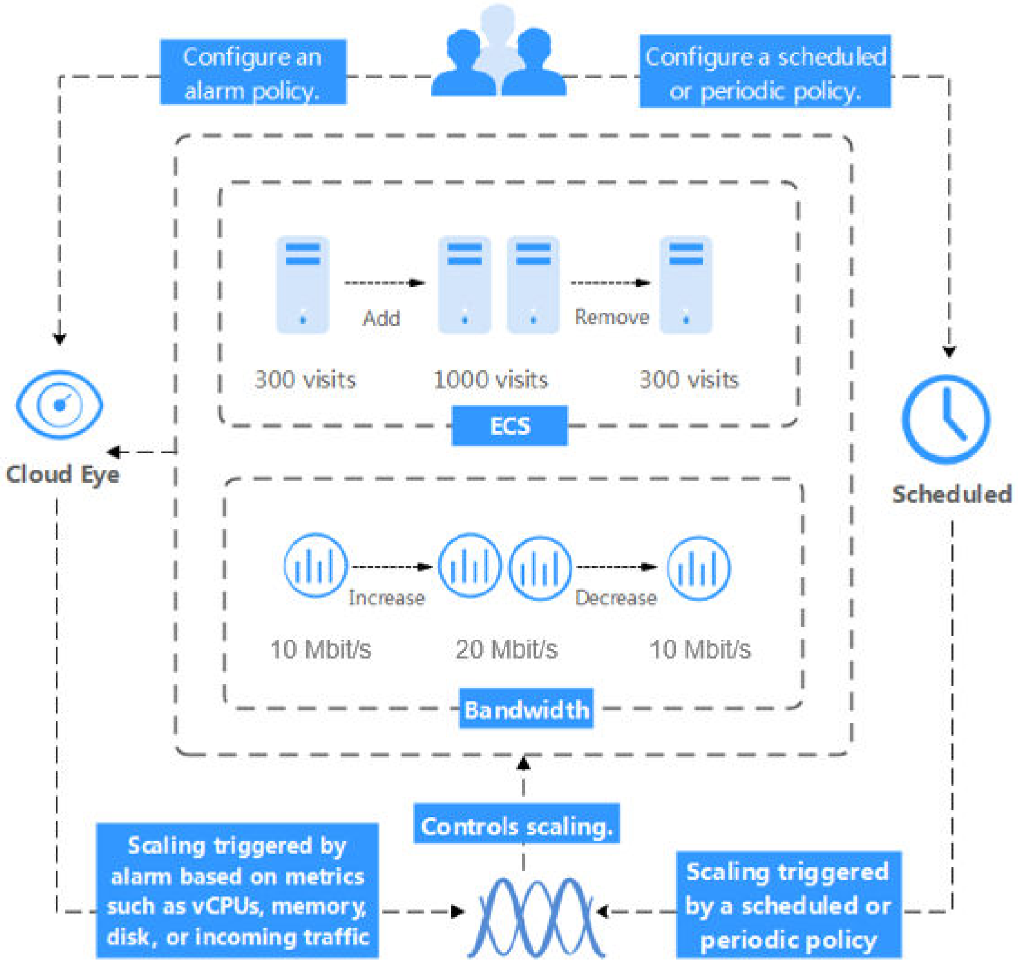
Auto Scaling (AS) helps you automatically scale Elastic Cloud Server (ECS) and bandwidth resources to keep up with changes in demand based on pre-configured AS policies. It allows you to add ECS instances or increase bandwidths to handle load increases and also save money by removing resources that are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

Architecture

AS allows you to scale ECS instances and bandwidths.

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

Figure 2-1 AS architecture


Configure an alarm policy.
Configure a scheduled or periodic policy.
Add
Remove
300 visits
1000 visits
300 visits
ECS
Cloud Eye
Scheduled
Increase
Decrease
10 Mbit/s
20 Mbit/s
10 Mbit/s
Bandwidth
Controls scaling.
Scaling triggered by alarm based on metrics such as vCPUs, memory, disk, or incoming traffic
Scaling triggered by a scheduled or periodic policy

Accessing AS

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

- Calling APIs

  Use this method if you are required to integrate AS on the public cloud into a third-party system for secondary development. For more information, see ***Auto Scaling API Reference***.

- Management console

  Use this method if you do not need to integrate AS with a third-party system.

  After registering on the public cloud, log in to the management console and select **Auto Scaling** from the service list on the homepage.

3 AS Advantages

AS automatically scales resources to keep up with service demands based on pre-configured AS policies. With automatic resource scaling, you can enjoy reduced costs, improved availability, and high fault tolerance. AS is used for the following scenarios:

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

Automatic Resource Scaling

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

- Scaling ECS Instances on Demand

  AS scales ECS instances for applications based on demand, improving cost management. ECS instances can be scaled dynamically, on a schedule, or manually:

  - Dynamic scaling

    Dynamic scaling allows scale resources in response to changing demand using alarm-based policies.

  - Scheduled scaling

    Scheduled scaling helps you set up your scaling schedule according to predictable load changes by creating periodic or scheduled policies.

  - Manual scaling

    You can either manually change the expected number of instances of your AS group, or add or remove instances to or from the AS group.

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

Figure 3-1 Over-provisioned capacity

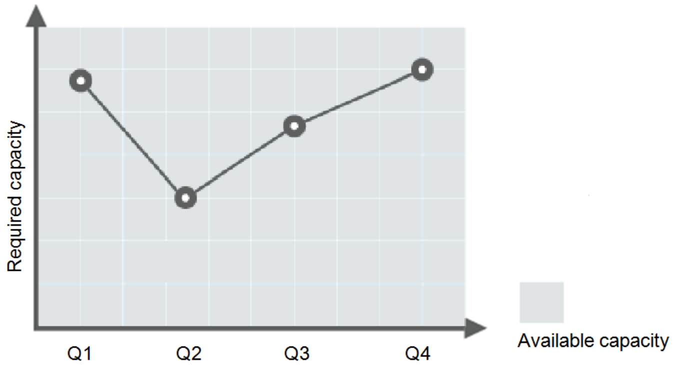


Figure 3-2 Insufficient capacity

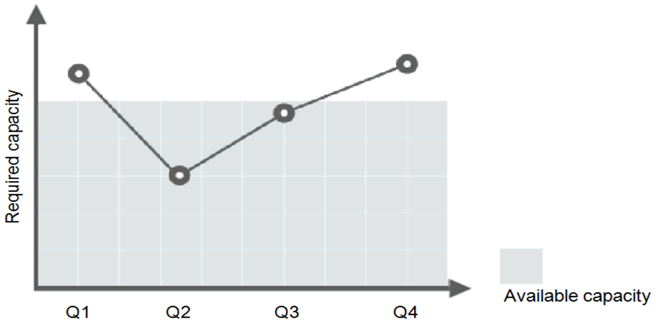
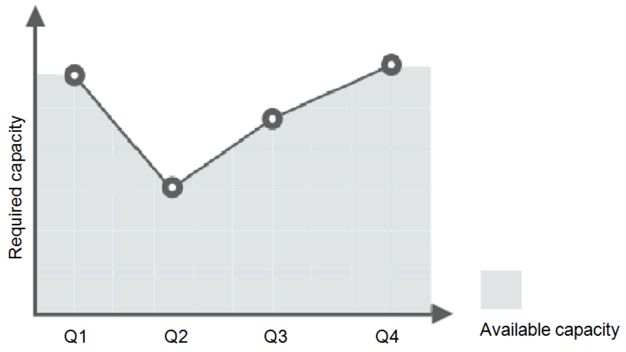


Figure 3-3 Auto-scaled capacity



- Scaling Bandwidth on Demand

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

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

- Alarm-based policies

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

- Scheduled policies

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

- Periodic policies

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

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

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

- Evenly Distributing Instances by AZ

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

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

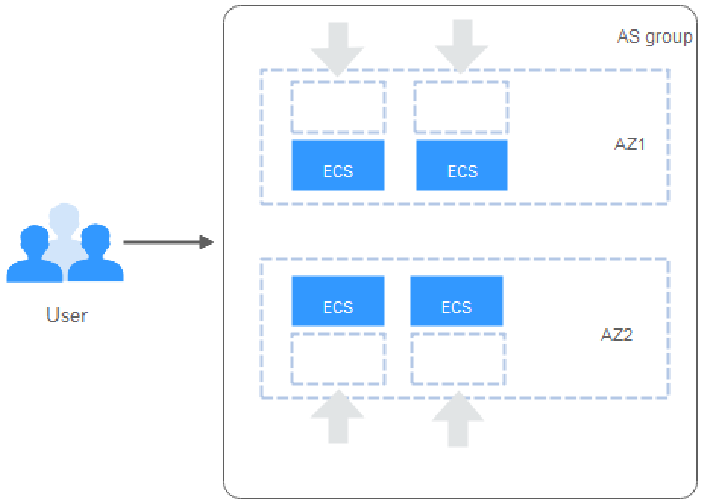
  An AS group can contain ECS instances in one or more AZs within a region. When scaling the capacity of an AS group, AS attempts to evenly distribute ECS instances across AZs used by the AS group based on the following rules:

  Evenly distributing new instances to balanced AZs

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

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

Figure 3-4 Evenly distributing instances

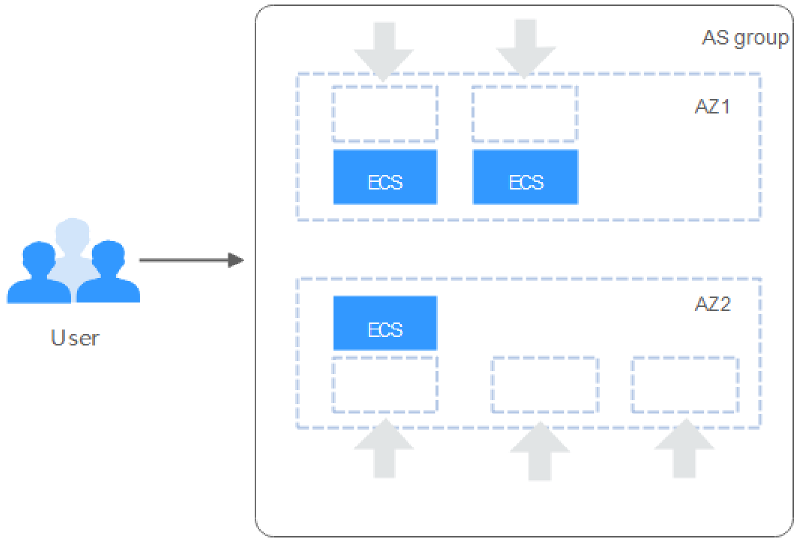


Re-balancing instances across AZs

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

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

Figure 3-5 Re-balancing instances



Enhanced Cost Management

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

Higher Availability

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

Using ELB with AS

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

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

High Fault Tolerance

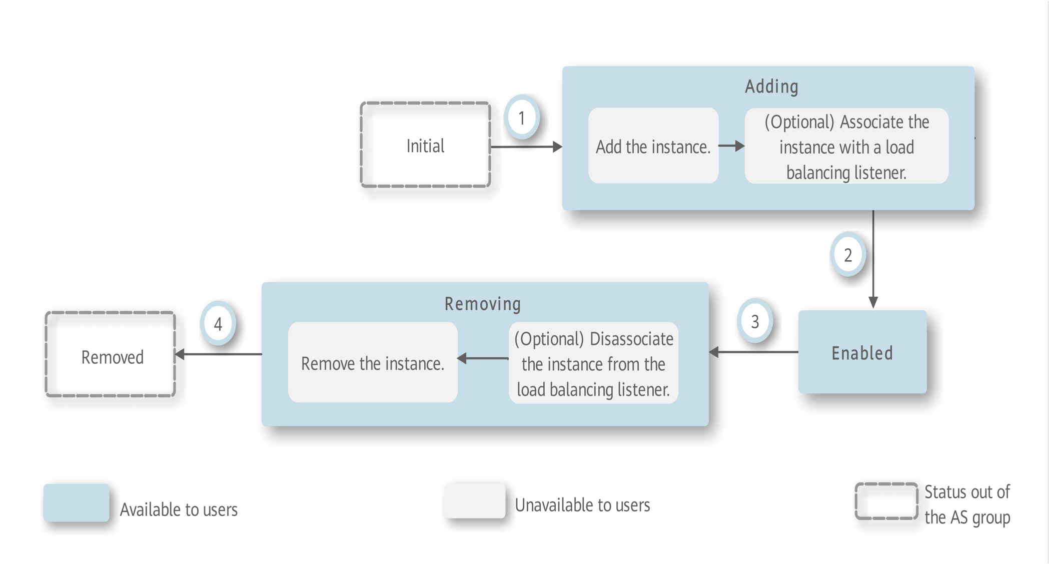
AS monitors instances in an AS group, and replaces any unhealthy instances it detects with new ones.

4 Instance Lifecycle

An ECS instance in an AS group goes through different statuses from its creation to its removal.

The instance status changes as shown in **Figure 4-1** if you have not added a lifecycle hook to the AS group.

Figure 4-1 Instance lifecycle


Initial
1
Adding
Add the instance.
(Optional) Associate the instance with a load balancing listener.
2
Enabled
3
Removing
(Optional) Disassociate the instance from the load balancing listener.
Remove the instance.
4
Removed
Available to users
Unavailable to users
Status out of the AS group

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

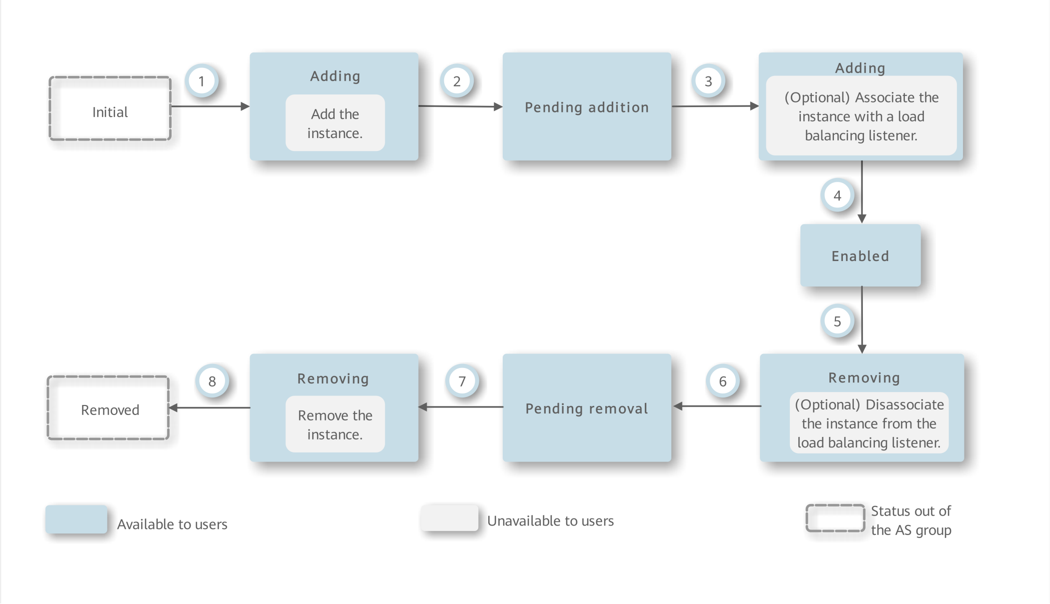
Table 4-1 Instance statuses

Status	Action	Description	Trigger Condition
Initial	-	The instance has not been added to the AS group.	The instance status changes to **Adding** when any of the following conditions occurs: ● You manually increase the expected number of instances of the AS group. ● The system automatically expands the AS group capacity. ● You manually add instances to the AS group.
Adding	Add the 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.	
Enabled	-	The instance is added to the AS group and starts to process service traffic.	The instance status changes from **Enabled** to **Removing** when any of the following conditions is met: ● You manually decrease the expected number of instances of the AS group. ● The system automatically reduces the AS group capacity. ● A health check shows that an enabled instance is unhealthy, and the system removes it from the AS group. ● You manually remove instances from the AS group.
Removing	(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.	
	Remove the instance.	After the instance is unbound from the load balancing listener, it is removed from the AS group.	
Removed	-	The instance lifecycle in the AS group ends.	-

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

If you have added a lifecycle hook to the AS group, the instance statuses change as shown in **Figure 4-2**. When a scale-out or scale-in event occurs in the AS group, the required instances are suspended by the lifecycle hook and remain in the wait status until the timeout period ends or you manually call back the instances. You can perform custom operations on the instances when they are in the wait status. For example, you can install or configure software on an instance before it is added to the AS group or download log files from an instance before it is removed.

Figure 4-2 Instance lifecycle



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

Table 4-2 Instance statuses

Status	Action	Description	Trigger Condition
Initial	-	The instance has not been added to the AS group.	The instance status changes to **Adding** when any of the following conditions occurs: ● You manually change the expected number of instances of the AS group. ● The system automatically expands the AS group capacity. ● You manually add instances to the AS group.
Adding	Add the instance.	When trigger condition 1 is met, AS adds the instance to expand the AS group capacity.	

Status	Action	Description	Trigger Condition
Pending addition	-	The lifecycle hook suspends the instance that is being added to the AS group and puts the instance into a wait status.	The instance status changes from **Pending addition** to **Adding** when either of the following conditions occurs: ● The default callback action is performed. ● You manually perform the callback action.
Adding	(Optional) Associate the instance with a load balancing listener.	When trigger condition 3 is met, AS associates the instance with the load balancing listener.	
Enabled	-	The instance is added to the AS group and starts to process service traffic.	The instance status changes from **Enabled** to **Removing** when any of the following conditions occurs: ● You manually change the expected number of instances of the AS group. ● The system automatically reduces the AS group capacity. ● A health check shows that the instance is unhealthy after being enabled, and the system removes it from the AS group. ● You manually remove an instance from an AS group.
Removing	(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.	
Pending removal	-	The lifecycle hook suspends the instance that is being removed from the AS group and puts the instance into a wait status.	The instance status changes from **Pending removal** to **Removing** when either of the following conditions occurs: ● The default callback action is performed. ● You manually perform the callback action.
Removing	Remove the instance.	When trigger condition 7 is met, AS removes the instance from the AS group.	
Removed	-	The instance lifecycle in the AS group ends.	-

Instances are added to an AS group manually or automatically. Then, they go through statuses **Adding**, **Pending addition**, **Adding**, **Enabled**, **Removing**, **Pending removal**, and **Removing** and are finally removed from the AS group.

5 Constraints

Function 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.

    ECS instances where stateless applications are running do not store data that needs to be persisted locally.

    Think of stateless transactions as a vending machine: a single request and a response.
  - Stateful applications and processes, however, are those that can be returned to again and again. They are performed in the context of previous transactions and the current transaction may be affected by what happened during previous transactions.

    ECS instances where stateful applications are running store data that needs to be persisted locally.

    Stateful transactions are performed repeatedly, such as online banking or e-mail, which are performed in the context of previous transactions.
- AS can release ECS instances in an AS group automatically, so the instances cannot be used to save application status information (such as session statuses) or related data (such as database data and logs). If the application status or related data must be saved, you can store the information on separate servers.
- AS does not support capacity expansion or deduction of instance vCPUs and memory.

Quotas

AS resources must comply with quota requirements listed in **Table 5-1**.

Table 5-1 Quotas

Item	Description	Default
AS group	Maximum number of AS groups per region per account	10
AS configuration	Maximum number of AS configurations per region per account	100
AS policy	Maximum number of AS policies per AS group	10
Instance	Maximum number of instances per AS group	300
Bandwidth scaling policy	Maximum number of bandwidth scaling policies per region per account	10
Lifecycle hook	Maximum number of lifecycle hooks per AS group	5
Notification	Maximum number of notifications per AS group	5
Tag	Maximum number of tags per AS group	10

6 Region and AZ

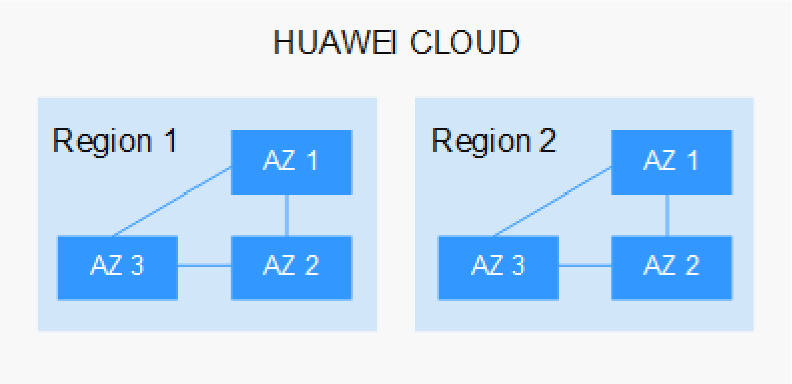
Concept

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

- Regions are divided based on 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 into universal regions and dedicated regions. A universal region provides universal cloud services for common tenants. A dedicated region provides specific services for specific tenants.
- An AZ contains one or more 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 support cross-AZ high-availability systems.

Figure 6-1 shows the relationship between regions and AZs.

Figure 6-1 Regions and AZs



Huawei Cloud provides services in many regions around the world. You can select a region and an AZ based on requirements. For more information, see **Huawei Cloud Global Regions**.

Selecting a Region

When selecting a region, consider the following factors:

- Location

  It is recommended that you select the closest region for lower network latency and quick access.

  - If your target users are in Asia Pacific (excluding the Chinese mainland), select the **CN-Hong Kong**, **AP-Bangkok**, or **AP-Singapore** region.
  - If your target users are in Africa, select the **AF-Johannesburg** region.
  - If your target users are in Latin America, select the **LA-Santiago** region.

  NOTE

  The **LA-Santiago** region is located in Chile.

- Resource price

  Resource prices may vary in different regions. For details, see **Product Pricing Details**.

Selecting an AZ

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

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

Regions and Endpoints

Before you use an API to call resources, specify its region and endpoint. For more details, see **Regions and Endpoints**.

7 Billing

You can use AS for free, but ECS instances automatically created in an AS group are billed on a pay-per-use basis. For pricing details, see **ECS Billing**. EIPs used by the instances are also billed. For pricing details, see **EIP Billing**. When the AS group scales in, the automatically created instances will be removed from the AS group and be deleted. After the deletion, these instances are no longer billed. Instances manually added are still billed after being removed from the AS group. If you do not need these instances, unsubscribe from them on the ECS console.

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

8 Security

8.1 Shared Responsibilities

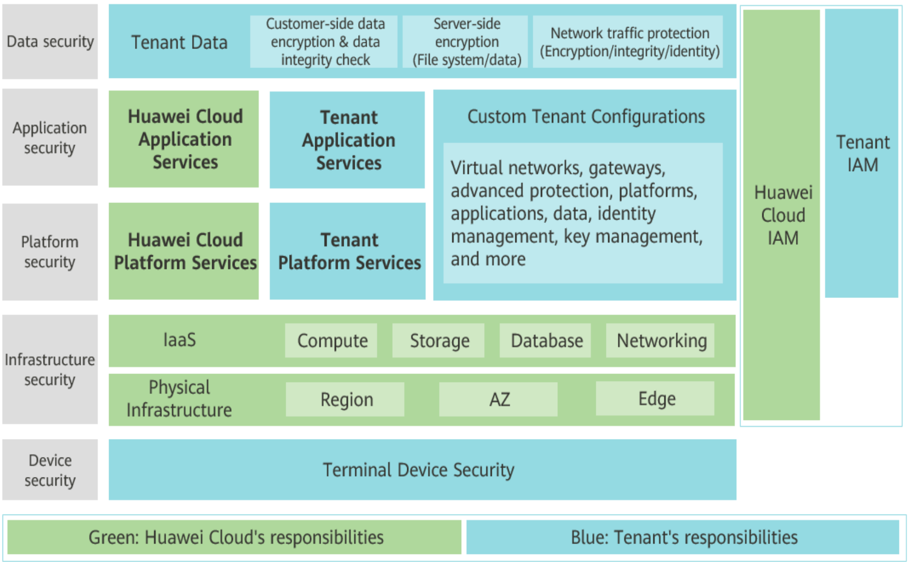
Huawei guarantees that its commitment to cyber security will never be outweighed by the consideration of commercial interests. To cope with emerging cloud security challenges and pervasive cloud security threats and attacks, Huawei Cloud builds a comprehensive cloud service security assurance system for different regions and industries based on Huawei's unique software and hardware advantages, laws, regulations, industry standards, and security ecosystem.

Figure 8-1 illustrates the responsibilities shared by Huawei Cloud and users.

- **Huawei Cloud**: Ensure the security of cloud services and provide secure clouds. Huawei Cloud's security responsibilities include ensuring the security of our IaaS, PaaS, and SaaS services, as well as the physical environments of the Huawei Cloud data centers where our IaaS, PaaS, and SaaS services operate. Huawei Cloud is responsible for not only the security functions and performance of our infrastructure, cloud services, and technologies, but also for the overall cloud O&M security and, in the broader sense, the security and compliance of our infrastructure and services.
- **Tenant**: Use the cloud securely. Tenants of Huawei Cloud are responsible for the secure and effective management of the tenant-customized configurations of cloud services including IaaS, PaaS, and SaaS. This includes but is not limited to virtual networks, the OS of virtual machine hosts and guests, virtual firewalls, API Gateway, advanced security services, all types of cloud services, tenant data, identity accounts, and key management.

Huawei Cloud Security White Paper elaborates on the ideas and measures for building Huawei Cloud security, including cloud security strategies, the shared responsibility model, compliance and privacy, security organizations and personnel, infrastructure security, tenant service and security, engineering security, O&M security, and ecosystem security.

Figure 8-1 Huawei Cloud shared security responsibility model


Data security
Tenant Data
Customer-side data encryption & data integrity check
Server-side encryption (File system/data)
Network traffic protection (Encryption/integrity/identity)
Application security
Huawei Cloud Application Services
Tenant Application Services
Custom Tenant Configurations
Virtual networks, gateways, advanced protection, platforms, applications, data, identity management, key management, and more
Platform security
Huawei Cloud Platform Services
Tenant Platform Services
Infrastructure security
IaaS
Compute
Storage
Database
Networking
Physical Infrastructure
Region
AZ
Edge
Device security
Terminal Device Security
Huawei Cloud IAM
Tenant IAM
Green: Huawei Cloud's responsibilities
Blue: Tenant's responsibilities

8.2 Identity and Access Management

8.2.1 Access Control for AS

Identity Authentication

Identity and Access Management (IAM) provides identity authentication, permissions management, and access control, helping you securely manage access to your Huawei Cloud resources.

With IAM, you can use your account to create IAM users, and assign permissions to the users to control their access to specific resources. For example, you can assign permissions to allow some software developers to use AS resources but disallow them to delete or perform any high-risk operations on the resources.

Access Control

AS supports access control by using IAM permissions, IAM projects, enterprise projects, critical operation protection, and security groups.

Table 8-1 AS access control

Method	Description	Reference
Permissions control through IAM	By default, new IAM users do not have permissions assigned. You need to add a user to one or more groups, and attach permissions policies or roles to these groups. Users inherit permissions from the groups to which they are added and can perform specified operations on cloud services based on the permissions.	**Permission Management**
IAM projects and enterprise projects	Both IAM projects and enterprise projects can be managed by one or more user groups. You can authorize a user group by applying policies to it. Then users inherit permissions defined by the policies.	**IAM Protects and Enterprise Projects**
Critical operation protection	After critical operation protection is enabled, identity authentication is required when you delete an AS group.	**Critical Operation Protection**
Security groups	A security group is a collection of access control rules for ECSs that have the same security requirements and are mutually trusted. After a security group is created, you can add different access rules to the security group, and these rules will apply to all ECSs added to this security group. Your account automatically comes with a default security group that allows all outbound traffic and denies all inbound traffic. Your ECSs in the security group can communicate with each other without the need to add rules.	**Configuring Security Group Rules**

8.3 Data Protection

User encryption allows you to use the encryption feature provided on the cloud platform to encrypt ECS resources, improving data security. If you use an encrypted ECS to create an AS configuration, the AS configuration is encrypted as the ECS. For more information, see **User Encryption**.

8.4 Auditing and Logging

Cloud Trace Service (CTS) is a log audit service for Huawei Cloud security. It allows you to collect, store, and query cloud resource operation records. You can use these records to perform security analysis, audit compliance, track resource changes, and locate faults.

AS works with 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.

After you enable CTS, whenever an AS API is called, the operation is recorded in a log file, which is then dumped to a specified OBS bucket for storage based on time and data changes. On the CTS console, you can query operation records of the last seven days. To store operation records for a longer period, you can transfer them to Object Storage Service (OBS) buckets.

- For details about how to enable and configure CTS, see **Enabling CTS**.
- For details about AS resource operations that can be recorded by CTS, see **Recording AS Resource Operations**.
- For details about how to view traces, see **Viewing Traces**.

8.5 Security Risk Monitoring

Monitoring Metrics

AS reports monitoring metrics to Cloud Eye. You can use Cloud Eye to query metrics and alarms generated by AS. For details about monitoring metrics reported by AS, see **Monitoring Metrics**.

Health Check

Cloud Eye provides you with insights into the running statuses of your ECSs. You can view monitoring metrics of an AS group to better understand performance of ECSs in the AS group. For details, see **Viewing Monitoring Metrics**.

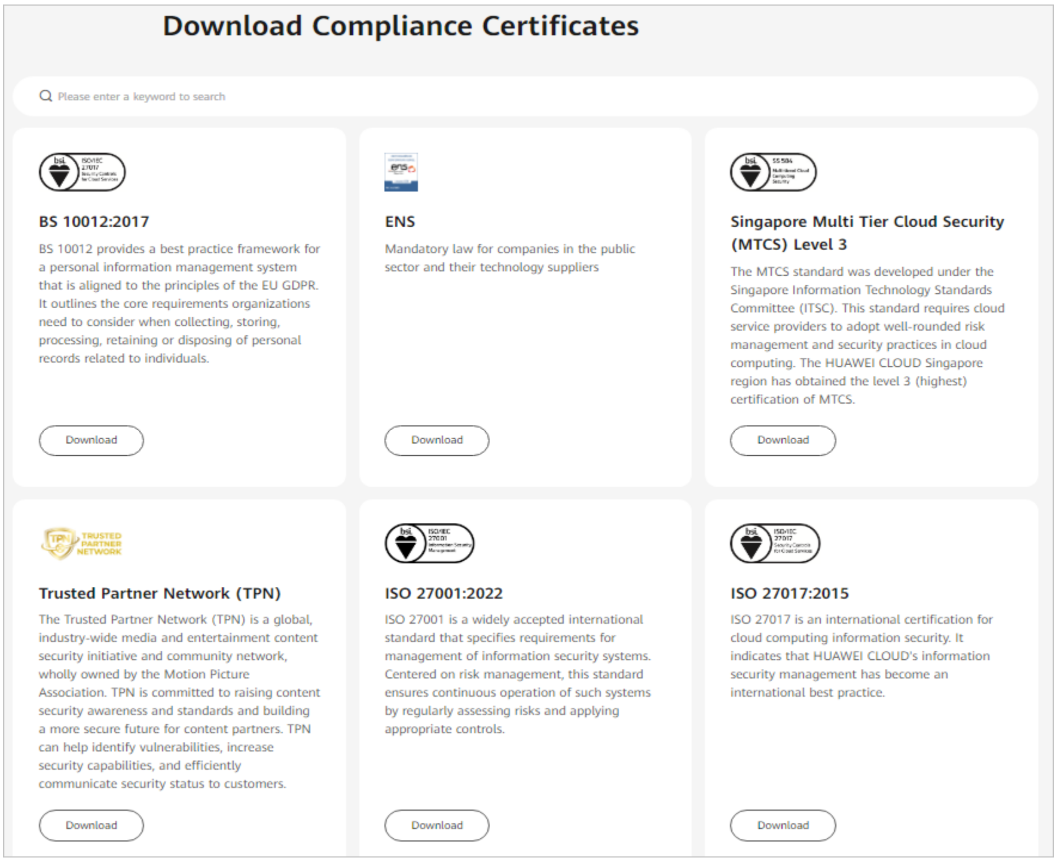
8.6 Certificates

Compliance Certificates

Huawei Cloud services and platforms have obtained various security and compliance certifications from authoritative organizations, such as International

Organization for Standardization (ISO). You can **download** them from the console.

Figure 8-2 Downloading compliance certificates



Download Compliance Certificates

Please enter a keyword to search

BS 10012:2017

BS 10012 provides a best practice framework for a personal information management system that is aligned to the principles of the EU GDPR. It outlines the core requirements organizations need to consider when collecting, storing, processing, retaining or disposing of personal records related to individuals.

Download

ENS

Mandatory law for companies in the public sector and their technology suppliers

Download

Singapore Multi Tier Cloud Security (MTCS) Level 3

The MTCS standard was developed under the Singapore Information Technology Standards Committee (ITSC). This standard requires cloud service providers to adopt well-rounded risk management and security practices in cloud computing. The HUAWEI CLOUD Singapore region has obtained the level 3 (highest) certification of MTCS.

Download

Trusted Partner Network (TPN)

The Trusted Partner Network (TPN) is a global, industry-wide media and entertainment content security initiative and community network, wholly owned by the Motion Picture Association. TPN is committed to raising content security awareness and standards and building a more secure future for content partners. TPN can help identify vulnerabilities, increase security capabilities, and efficiently communicate security status to customers.

Download

ISO 27001:2022

ISO 27001 is a widely accepted international standard that specifies requirements for management of information security systems. Centered on risk management, this standard ensures continuous operation of such systems by regularly assessing risks and applying appropriate controls.

Download

ISO 27017:2015

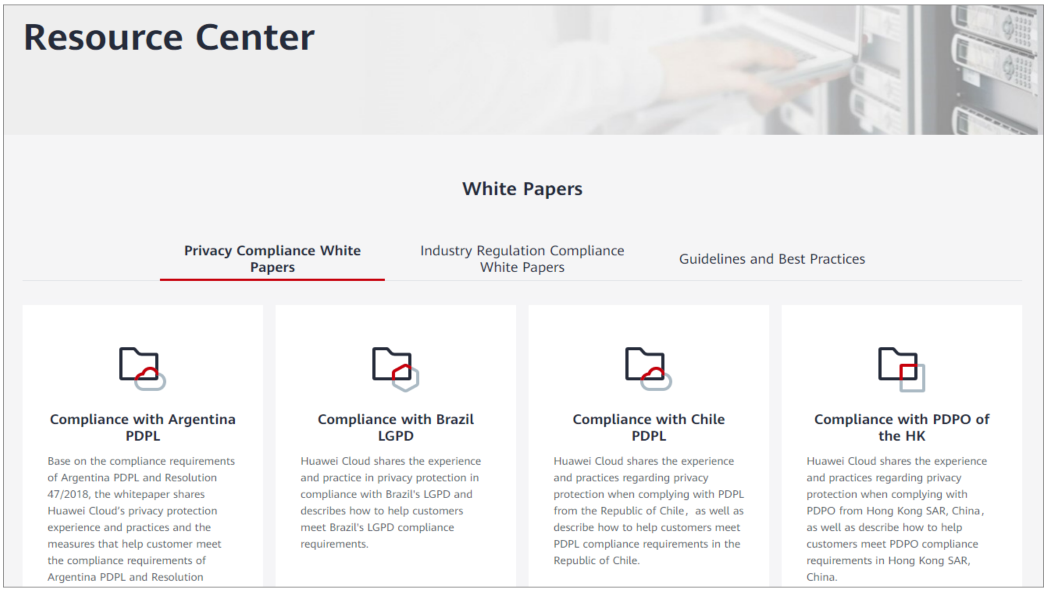
ISO 27017 is an international certification for cloud computing information security. It indicates that HUAWEI CLOUD's information security management has become an international best practice.

Download

Resource Center

Huawei Cloud also provides the following resources to help users meet compliance requirements. For details, see **Resource Center**.

Figure 8-3 Resource center



Resource Center

White Papers

Privacy Compliance White Papers

Industry Regulation Compliance White Papers

Guidelines and Best Practices

Compliance with Argentina PDPL

Base on the compliance requirements of Argentina PDPL and Resolution 47/2018, the whitepaper shares Huawei Cloud's privacy protection experience and practices and the measures that help customer meet the compliance requirements of Argentina PDPL and Resolution

Compliance with Brazil LGPD

Huawei Cloud shares the experience and practice in privacy protection in compliance with Brazil's LGPD and describes how to help customers meet Brazil's LGPD compliance requirements.

Compliance with Chile PDPL

Huawei Cloud shares the experience and practices regarding privacy protection when complying with PDPL from the Republic of Chile， as well as describe how to help customers meet PDPL compliance requirements in the Republic of Chile.

Compliance with PDPO of the HK

Huawei Cloud shares the experience and practices regarding privacy protection when complying with PDPO from Hong Kong SAR, China， as well as describe how to help customers meet PDPO compliance requirements in Hong Kong SAR, China.

9 AS and Other Services

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

Figure 9-1 shows the relationships between AS and other services.

Figure 9-1 Relationships between AS and other services

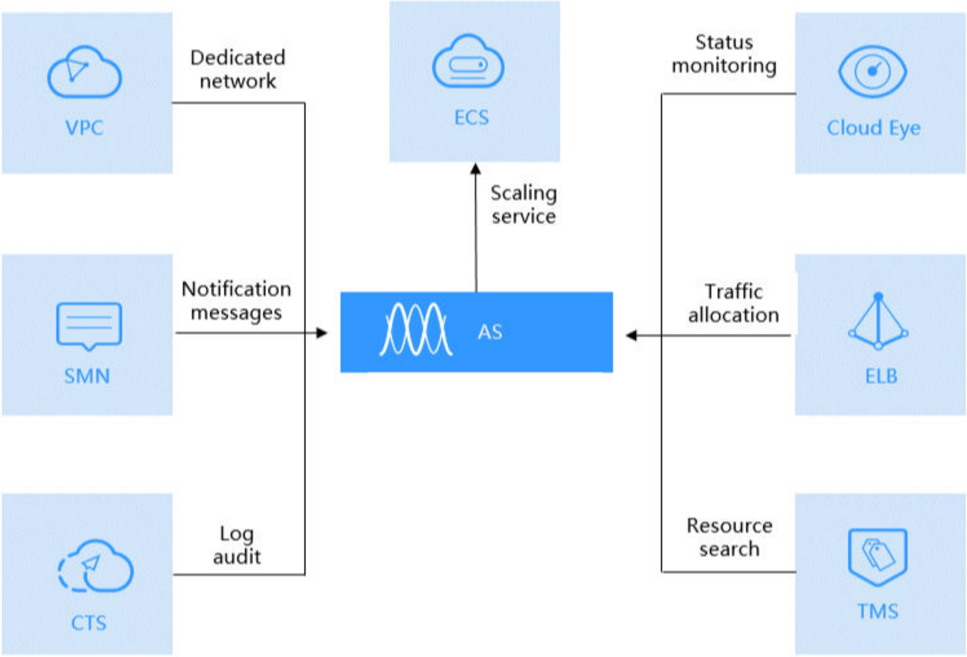


Table 9-1 Related services

Service	Description	Interaction	Reference
Elastic Load Balance (ELB)	After ELB is configured, AS automatically associates ECS instances to a load balancer listener when adding ECSs, and unbinds them when removing the instances. For AS to work with ELB, the AS group and load balancer must be in the same VPC.	ELB distributes traffic to all ECSs in an AS group.	**Adding a Load Balancer to an AS Group**
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 instance status monitored by Cloud Eye.	**AS Metrics**
ECS	ECS instances added in a scaling action can be managed and maintained on the ECS console.	AS automatically adjusts the number of ECS instances.	**Dynamically Expanding Resources**
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**

Service	Description	Interaction	Reference
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 more easily.	Tags	**Adding Tags to AS Groups and Instances**

10 Permissions Management

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

With IAM, you can create IAM users and assign permissions to the users to control their access to specific resources. For example, you can assign permissions to allow some software developers to use AS resources but disallow them to delete or perform any high-risk operations on the resources.

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

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

AS Permissions

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

When you grant AS permissions to a user group, set **Scope** to **Region-specific projects** and then select projects (for example, **ap-southeast-2**) in region **AP-Bangkok** for the permissions to take effect. If you select **All projects**, the permissions will take effect for the user group in all region-specific projects. When accessing AS, the users need to switch to a region where they have been authorized to use this service.

You can grant users permissions by using roles and policies.

- Roles: A type of coarse-grained authorization mechanism that defines permissions related to user responsibilities. This mechanism provides only a limited number of service-level roles for authorization. When using roles to grant permissions, you also also need to attach any existing role dependencies. However, roles are not an ideal choice for fine-grained authorization and secure access control.

- Policies: A type of fine-grained authorization mechanism that defines permissions required to perform operations on specific cloud resources under certain conditions. This mechanism allows for more flexible policy-based authorization, meeting requirements for secure access control. For example, you can grant AS users only the permissions for managing a certain type of ECSs. Most policies define permissions based on APIs. For the API actions supported by AS, see **Permissions Policies and Supported Actions**.

  Table 10-1 lists all the system policies supported by AS.

Table 10-1 System-defined permissions supported by AS

Policy Name	Description	Category	Dependency
AutoScaling FullAccess	Full permissions for all AS resources	System-defined policy	None
AutoScaling ReadOnlyAccess	Read-only permissions for all AS resources	System-defined policy	None
AutoScaling Administrator	Full permissions for all AS resources	System role	The **ELB Administrator**, **CES Administrator**, **Server Administrator**, and **Tenant Administrator** roles need to be assigned in the same project.

Table 10-2 lists the common operations supported by each system-defined policy of AS. Select the policies as required.

Table 10-2 Common operations supported by each system-defined policy of AS

Operation	AutoScaling FullAccess	AutoScaling ReadOnlyAccess	AutoScaling Administrator
Creating an AS group	√	x	√
Modifying an AS group	√	x	√

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

Helpful Links

- What Is IAM?
- Creating a User and Granting AS Permissions
- Permissions Policies and Supported Actions

11 Basic Concepts

AS Group

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

AS Configuration

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

AS Policy

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

Scaling Action

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

Cooldown Period

To prevent an alarm-based policy from being repeatedly triggered by the same event, you can set a cooldown period. A cooldown period (in seconds) is the period of time between two scaling actions. AS recounts the cooldown period after a scaling action is complete. During the cooldown period, AS denies all scaling requests triggered by alarm-based policies. Scaling requests triggered manually or by scheduled or periodic policies are not affected.

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

Bandwidth Scaling

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

12 Change History

Released On	What's New
2022-11-15	This issue is the fifth official release. Modified the following content: Added section "Security."
2021-10-30	This issue is the fourth official release. Modified the following content: Added section "Permissions Management."
2020-10-19	This issue is the third official release. Modified the following content: Added section "Access Methods."
2019-09-30	This issue is the second official release.
2018-11-19	This issue is the first official release.