Auto Scaling

Service Overview

Issue	01
Date	2022-09-15

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions


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 What Is Auto Scaling?

AS Introduction

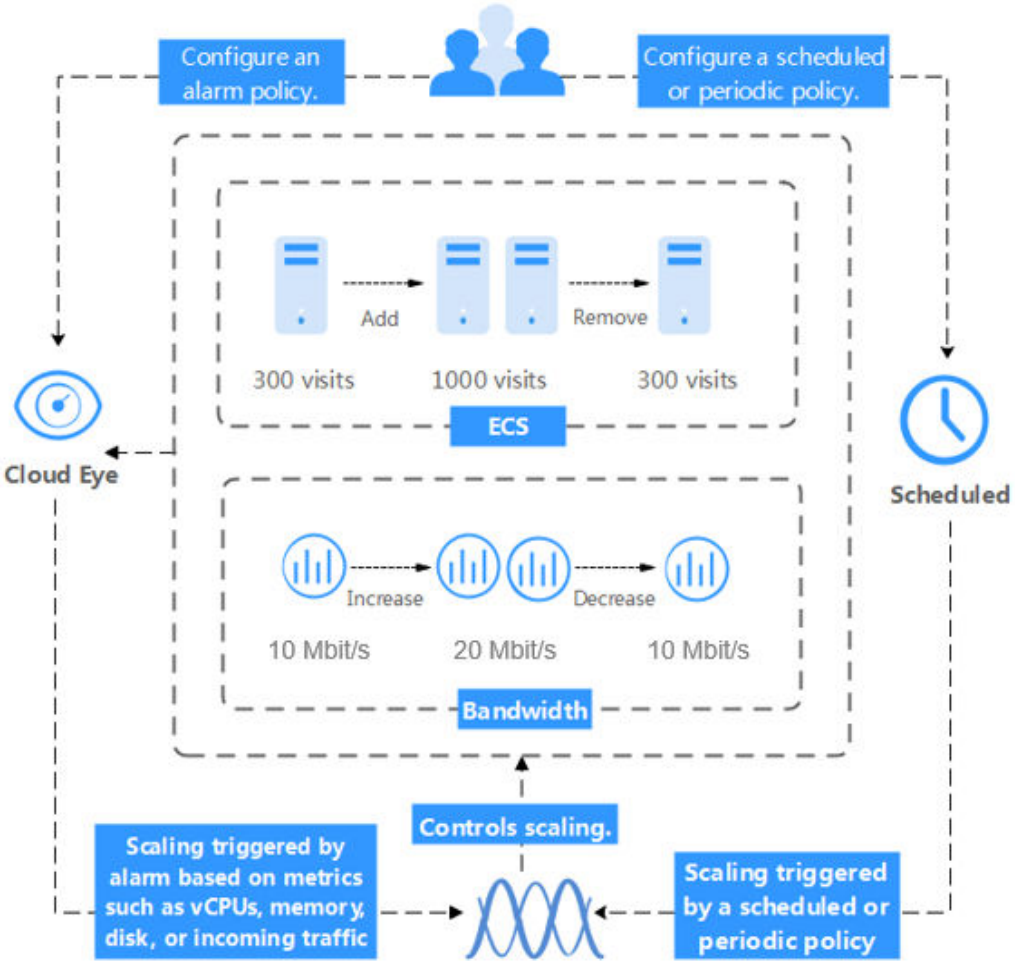
Auto Scaling (AS) helps you automatically scale Elastic Cloud Server (ECS) and bandwidth resources to keep up with changes in demand based on pre-configured AS policies. It allows you to add ECS instances or increase bandwidths to handle increases in load and also save money by removing resources that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

Architecture

AS allows you to scale ECS instances and bandwidths.

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

Figure 1-1 AS architecture



Accessing AS

The public cloudThe cloud service platform 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.

2 AS Advantages

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

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

Automatic Resource Scaling

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

- Scaling ECS Instances on Demand

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

  - Dynamic scaling

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

  - Scheduled scaling

    Scheduled scaling helps you to set up your own 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 provision enough servers so that the application always has enough capacity to meet demand, as shown in **Figure 2-1**. The second option is to provision servers according to the average load of the application, as shown in **Figure 2-2**. 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 2-3**.

Figure 2-1 Over-provisioned capacity

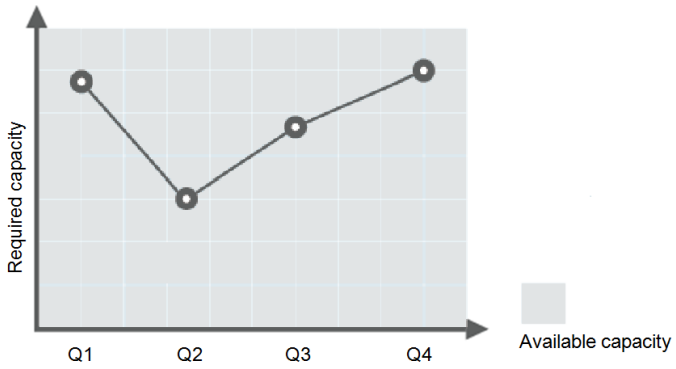


Figure 2-2 Insufficient capacity

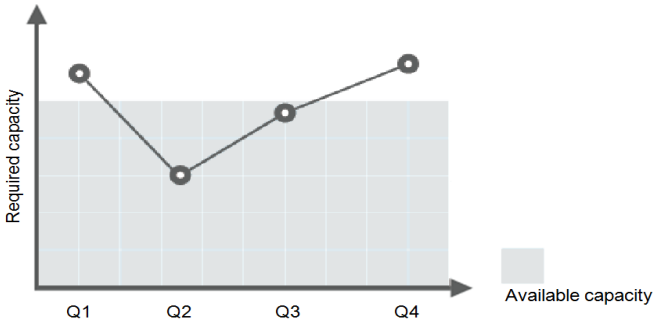
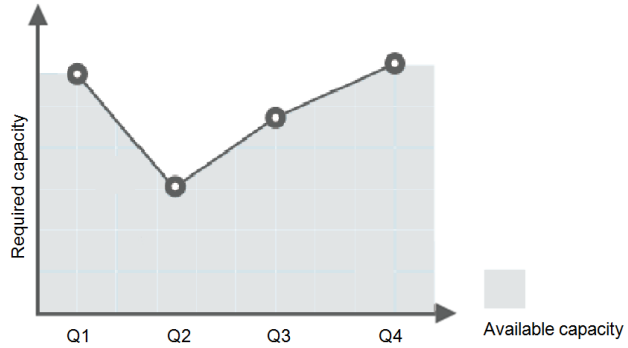


Figure 2-3 Auto-scaled capacity



- 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 live streaming website.

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

- Evenly Distributing Instances by AZ

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

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

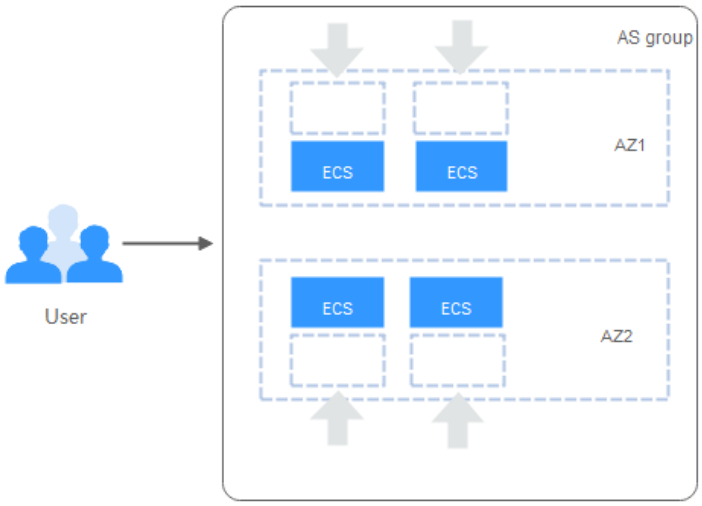
  An AS group can contain ECS instances in one or more AZs within a region. During scale the capacity of an AS group, AS attempts to evenly distribute ECS instances across AZs used by the AS group based on the following rules:

  Evenly distributing new instances to balanced AZs

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

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

Figure 2-4 Evenly distributing instances

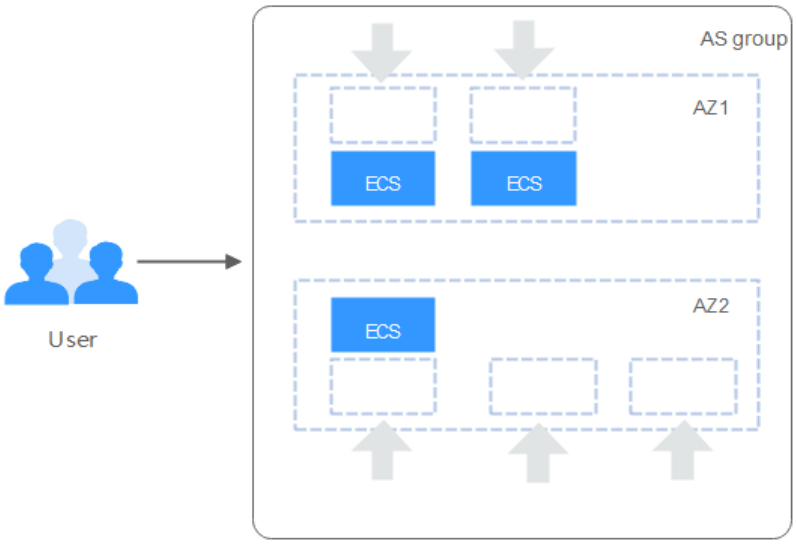


Re-balancing instances across AZs

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 2-5 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

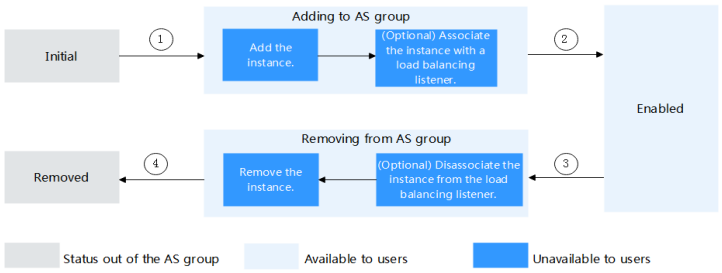
AS monitors instances in an AS group, and replaces any unhealthy instances it detects with new ones.

3 Instance Lifecycle

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

The instance status changes as shown in **Figure 3-1** if you have not added a lifecycle hook to the AS group.

Figure 3-1 Instance lifecycle



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

Table 3-1 Instance statuses

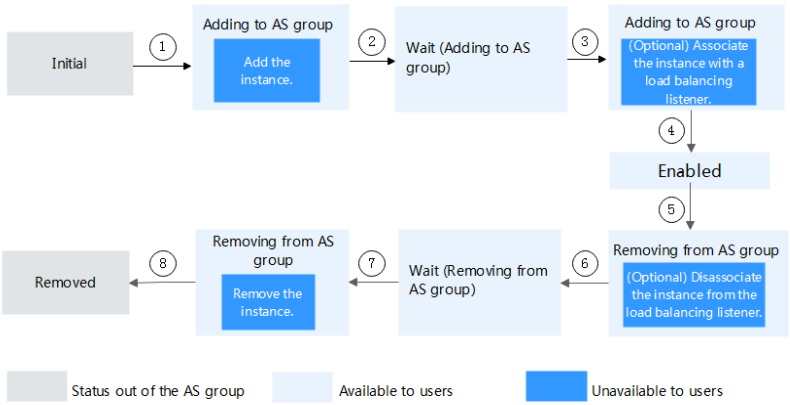
Status	Sub-status	Status Description	Trigger Condition
Initial	None	The instance has not been added to the AS group.	The instance status will be changed to **Adding to AS group** when any of the following conditions is met: • You manually increase the expected number of instances of the AS group.
Adding to AS group	Add the instance.	When trigger condition 1 is met, AS adds the instance to expand the AS group capacity.	

Status	Sub-status	Status Description	Trigger Condition
	(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.	• The system automatically expands the AS group capacity. • You manually add instances to the AS group.
Enabled	None	The instance is added to the AS group and starts to process service traffic.	The instance status is changed from **Enabled** to **Removing from AS group** when any of the following conditions is met: • You manually decrease the expected number of instances of the AS group. • The system automatically removes instances in a scaling action. • A health check shows that an enabled instance is unhealthy, and the system removes it from the AS group. • You manually remove instances from the AS group.
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.	
	Remove the instance.	After the instances are unbound from the load balancing listener, they are removed from the AS group.	
Removed	None	The instance lifecycle in the AS group ends.	None

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

If you have added a lifecycle hook to the AS group, the instance statues change as shown in **Figure 3-2**. 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 3-2 Instance lifecycle



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

Table 3-2 Instance statuses

Status	Sub-status	Status Description	Trigger Description
Initial	None	The instance has not been added to the AS group.	The instance status is changed to **Adding to AS group** when any of the following occurs: • You manually increase the expected number of instances of an AS group. • The system automatically adds instances to the AS group in a scaling action. • You manually add instances to the AS group.
Adding to AS group	Add an instance.	When trigger condition 1 is met, AS adds the instance to expand the AS group capacity.	
Wait (Adding to AS group)	None	The lifecycle hook suspends the instance that is being added to the AS group and puts the instance into waiting state.	The instance status is changed from **Wait (Adding to AS group)** to **Adding to AS group** when either of the following operations is performed: • The default callback action is performed. • You manually perform the callback action.

Status	Sub-status	Status Description	Trigger Description
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.	
Enabled	None	The instance is added to the AS group and starts to process service traffic.	The instance status is changed from **Enabled** to **Removing from AS group** when any of the following occurs: ● You manually decrease the expected number of instances of an AS group. ● The system automatically removes instances in a scaling action. ● A health check shows that the instance is unhealthy after being enabled, and the system removes it from the AS group. ● You manually remove an instance from an AS group.
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.	
Wait (Removing from AS group)	None	The lifecycle hook suspends the instance that is being removed from the AS group and sets the instance to be in waiting state.	The instance status is changed from **Wait (Removing from AS group)** to **Removing from AS group** when either of the following occurs: ● The default callback action is performed. ● You manually perform the callback action.
Removing from AS group	Remove the instance.	When trigger condition 7 is met, AS removes the instance from the AS group.	
Removed	None	The instance lifecycle in the AS group ends.	None

Instances are added to an AS group manually or through a scaling action. Then, they go through the **Adding to AS group**, **Wait (Adding to AS group)**, **Adding to AS group**, **Enabled**, **Removing from AS group**, **Wait (Removing from the AS**

group), and **Removing from AS group** and are finally removed from the AS group.

4 Constraints

AS has the following constraints:

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

  NOTE

  - A stateless process or application can be understood in isolation. There is no stored knowledge of or reference to past transactions. Each transaction is made as if from scratch for the first time.

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

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

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

    Stateful transactions are performed repeatedly, such as online banking or e-mail, which are performed with the context of previous transactions.
- AS can release ECS instances in an AS group automatically, so the instances cannot be used to save application status information (such as session statuses) or related data (such as database data and logs). If the application status or related data must be saved, you can store the information on separate servers.
- AS does not support capacity expansion or deduction of instance vCPUs and memory.
- AS resources must comply with quota requirements listed in **Table 4-1**.

Table 4-1 Quotas

Item	Description	Default
AS group	Maximum number of AS groups per region per account	10

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

5 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. to support high-availability systems.

Selecting a Region

If your target users are in Europe, select the **EU-Dublin** region.

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.

6 Billing

You can use AS for free, but ECS instances automatically created in an AS group are billed on a pay-per-use basis. EIPs used by the instances are also billed. 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 you will be billed for that one hour of use.

7 AS and Other Services

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

Figure 7-1 shows the relationships between AS and other services.

Figure 7-1 Relationships between AS and other services

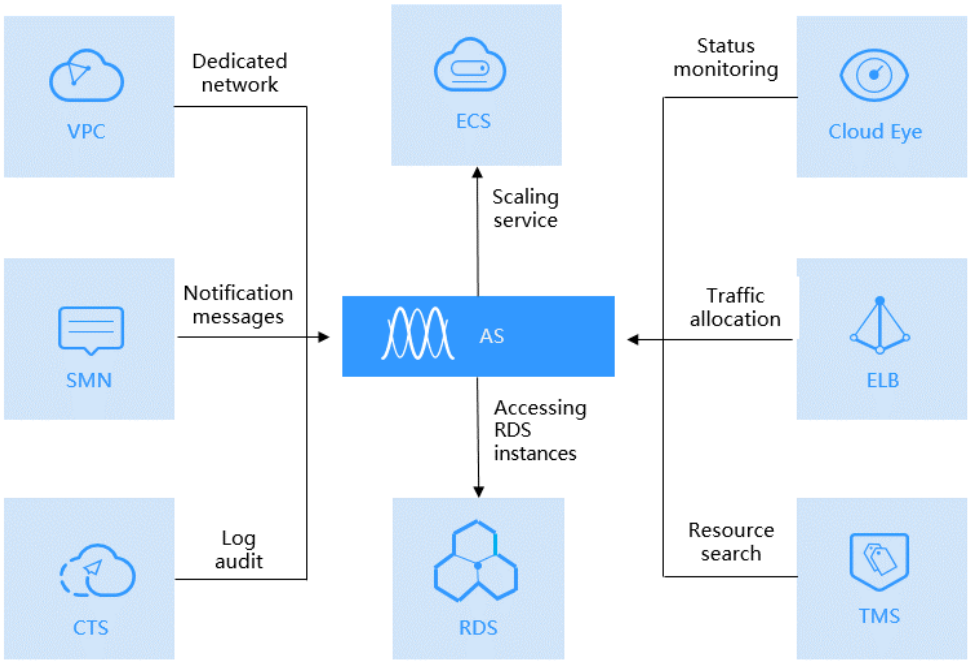


Table 7-1 Related services

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.	AS 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 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**
Simple Message Notification (SMN)	If you enable the SMN service, the system sends you notifications about the status of your AS group in a timely manner.	Message notification	**Configuring Notification for an AS Group**

Service	Description	Interaction	Reference
Cloud Trace Service (CTS)	With CTS, you can record AS operation logs for view, audit, and backtracking.	Log audit	**Recording AS Resource Operations**
Tag Management Service (TMS)	If you have multiple resources of the same type, TMS enables you to manage these resources more easily.	Tags	**Adding Tags to AS Groups and Instances**
Relational Database Service (RDS)	The prerequisites for directly accessing an RDS DB instance from a scaled instance are as follows: • 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 an RDS for MySQL DB Instance Through a Private Network**

8 Permissions Management

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

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

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

You can grant users permissions by using roles and policies.

- Roles: A type of coarse-grained authorization mechanism that defines permissions related to user responsibilities. This mechanism provides only a limited number of service-level roles for authorization. When using roles to grant permissions, you need to also assign other roles on which the permissions depend to take effect. However, roles are not an ideal choice for fine-grained authorization and secure access control.
- Policies: A type of fine-grained authorization mechanism that defines permissions required to perform operations on specific cloud resources under

certain conditions. This mechanism allows for more flexible policy-based authorization, meeting requirements for secure access control. For example, you can grant AS users only the permissions for managing a certain type of ECSs. Most policies define permissions based on APIs. For the API actions supported by AS, see **Permissions Policies and Supported Actions**.

Table 8-1 lists all the system policies supported by AS.

Table 8-1 System-defined permissions supported by AS

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

Table 8-2 lists the common operations supported by each system-defined policy of AS. Select the policies as required.

Table 8-2 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	√
Querying details about an AS group	√	√	√

Operation	AutoScaling FullAccess	AutoScaling ReadOnlyAccess	AutoScaling Administrator
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**

9 Basic Concepts

AS Group

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

AS Configuration

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

AS Policy

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

Scaling Action

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

Cooldown Period

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

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

at 10:32. If the scheduled scaling action ends at 10:36, a new cooldown period starts at 10:36 and ends at 10:41.

Bandwidth Scaling

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

10 Change History

Released On	Description
2022-09-15	This issue is the first official release.

