Auto Scaling

User Guide

User Guide

Working principles

Operation procedure

Procedure

To create a complete Auto Scaling solution, complete the following steps:


1. Create a scaling group. Configure the minimum and maximum number of ECS instances in the scaling group, and select the associated Server Load Balancer and RDS instances.
2. Create scaling configuration. Configure the ECS instances attributes for Auto Scaling, such as Image ID and Instance Type.
3. Enable the scaling group with the scaling configuration created in Step 2.
4. Create a scaling rule. For example, add N ECS instances.
5. Create a scheduled task. For example, to trigger the scaling rule created in Step 4 at 12:00 AM.
6. Create an alarm task (CloudMonitor API PutAlarmRule). For example, to add 1 ECS instance when the average (it can also be max or min) CPU usage is greater than or equal to 80%.

Workflow

Workflow

The following diagram shows the Auto Scaling workflow.
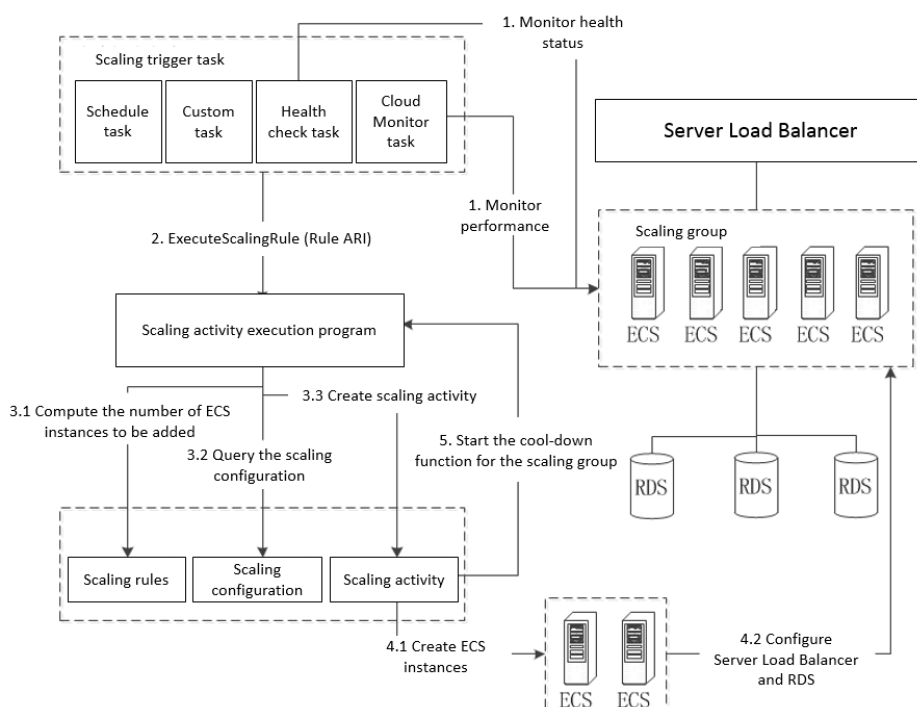
After a scaling group, scaling configuration, scaling rule, and scaling trigger task are created, the system executes the following process (in this example, ECS instances are added):

The task triggers a scaling activity according to the trigger condition.

- The CloudMonitor task monitors the performance of ECS instances in the scaling group in real time and triggers the request for executing a scaling rule based on the configured alarm rules. For example, when the average CPU usage of all ECS instances in the scaling group exceeds 60%.
- The scheduled task triggers the request for executing a scaling rule at the specified time.
- The custom task triggers the request for executing a scaling rule based on the monitoring system and alarm rules. For example, the number of online users or the job queue.
- Health check tasks regularly check the health status of the scaling group and its ECS instances. If an ECS instance is found to be unhealthy (not in running status), the health check task triggers a request to remove the ECS instance from the group.

The system triggers a scaling activity through the ExecuteScalingRule interface and specifies the scaling rule to be executed by its unique Alibaba Cloud resource identifier (ARI) in this interface.

If a custom task needs to be executed, you must have the ExecuteScalingRule interface called in your program.

The system obtains information about the scaling rule, scaling group, and scaling configuration based on the scaling rule ARI entered in Step 2 and creates a scaling activity.

i. The system uses the scaling rule ARI to query the scaling rule and scaling group, computes the number of ECS instances to be added, and configures the Server Load Balancer and RDS instances.

ii. According to the scaling group, the system queries the scaling configuration to determine the correct parameters (CPU, memory, bandwidth) to use when creating new ECS instances.

iii. The system creates scaling activity based on the number of ECS instances to be added, the ECS instance configuration, and the Server Load Balancer and RDS instance configurations.

During the scaling activity, the system creates ECS instances and configures Server Load Balancer and RDS instances.

i. The system creates the specified number of ECS instances based on the instance configuration.

ii. The system adds the intranet IP addresses of the created ECS instances to the whitelist of the specified RDS instance and adds the created ECS instances to the specified Server Load Balancer instance.

After the scaling activity is completed, the system starts the cool-down function for the scaling group. The cool-down time must elapse before the scaling group can execute any new scaling activity.

Considerations

Scaling rules

When you run and compute a scaling rule, the system can automatically adjust the number of ECS instances according to the MaxSize value and the MinSize value of the scaling group. For example, if the number of ECS instances is set to 50 in the scaling rule, but the MaxSize value of the scaling group is set to 45, we compute and run the scaling rule with 45 ECS instances.

Scaling activity

Only one scaling activity can be executed at a time in a scaling group.

A scaling activity cannot be interrupted. For example, if a scaling activity to add 20 ECS instances is being executed, it cannot be forced to terminate when only five instances have been created.

When a scaling activity fails to add or remove ECS instances to or from a scaling group, the system maintains the integrity of ECS instances rather than the scaling activity. That is, the system rolls back ECS instances, not the scaling activity. For example, if the system has created 20 ECS instances for the scaling group, but only 19 ECS instances are added to the Server Load Balancer instance, the system only releases the failed ECS instance.

Since Auto Scaling uses Alibaba Cloud's Resource Access Management (RAM) service to replace ECS instances through ECS API, the rollback ECS instance is still charged.

Cool-down time

During the cool-down time, only scaling activity requests from CloudMonitor alarm tasks are rejected by the scaling group. Other tasks, such as manually executed scaling rules and scheduled tasks, can immediately trigger scaling activities without waiting for the cool-down time to expire.

The cool-down time starts after the last ECS instance is added to or removed from the scaling group by a scaling activity.

Usage notes

Removal policies

Removal policies

There are two types of removal policies: default policy and custom policy.

Default removal policy

This policy first performs level-1 instance screening on the ECS instances created according to the oldest scaling configuration (OldestScalingConfiguration), and then performs level-2 screening on the oldest ECS instances (OldInstances).

> This policy first selects the ECS instances created according to the oldest scaling configuration (OldestScalingConfiguration) of the scaling group, and then selects the oldest ECS instance (OldestInstance) from these ECS instances. If more than one oldest ECS instance is found, one of them is selected at random and removed from the scaling group.

> Manually added ECS instances are not first selected for removal because they are not associated with any scaling configuration.

> If all ECS instances associated with the scaling configuration have been removed, but more instances still need to be removed from the scaling group, this policy selects the instance that was manually added earliest.

Custom release policy

You can set multiple policies to select and remove ECS instances successively from the scaling group.

Release policy types

> OldestInstance: This policy selects the ECS instance that was created earliest. As level-1 screening, the policy selects the earliest ECS instance, either created manually or automatically.

> NewestInstance: This policy selects the ECS instance that was created most recently. As level-1 screening, the policy selects the newest ECS instance, either created manually or automatically.

> OldestScalingConfiguration: This policy selects the instance created according to the oldest scaling configuration and skips over manually added instances. However, if all ECS instances

associated with scaling configurations have been removed, but more instances still need to be removed from the scaling group, this policy randomly selects a manually added ECS instance (an instance not associated with any scaling configuration).

ECS instance lifecycle

ECS instance lifecycle management

There are two types of ECS instances: automatically created and manually added instances.

Automatically created ECS instances

Automatically created ECS instances are automatically created according to scaling configuration and rules.

Auto Scaling manages the lifecycle of this ECS instance type. It creates ECS instances during scale-up and stops and releases them during scale-down.

Manually added ECS instances

Manually added ECS instances are manually attached to a scaling group.

Auto Scaling does not manage the lifecycle of this ECS instance type. When this ECS instance is removed from a scaling group, either manually or as the result of a scaling-down activity, Auto Scaling does not stop or release the instance.

Instance statuses

During its lifecycle, an ECS instance is in one of the following states:

- Pending: The ECS instance is being added to a scaling group. For example, Auto Scaling creates the ECS instance and adds it to the Server Load Balancer instance, or to the RDS access whitelist.
- InService: The ECS instance has been added to a scaling group and is functioning correctly.
- Removing: The ECS instance is being removed from a scaling group.

Instance health statuses

An ECS instance may be in the following health conditions:

> - Healthy
> - Unhealthy

ECS instances are regarded as unhealthy when they are not running. Auto Scaling automatically removes unhealthy ECS instances from scaling groups.

Note: Auto Scaling only stops and releases automatically created unhealthy instances. It does not stop and release manually added ones.

Cool-down time

Cool-down time

Cool-down time refers to a period during which Auto Scaling cannot execute any new scaling activity after another scaling activity is executed successfully in a scaling group.

> During cool-down time, the scaling activity requests from CloudMonitor alarm tasks are rejected. Other tasks, such as manually executed scaling rules and scheduled tasks, can immediately trigger scaling activities without waiting for the cool-down time to expire.

> During cool-down time, only the corresponding scaling group is locked. Scaling activities set for other scaling groups can be executed.

>> When the cool-down time specified by the scaling rule and the default cool-down time for the scaling group overlap, the former takes precedence over the latter.

>> When more than one instance is added to or removed from a scaling group, the cool-down time starts after the last instance is added to or removed from the scaling group.

>> As long as a scaling activity adds or removes one or more ECS instances, the cool-down time begins when the last ECS instance is added or removed.

If a scaling activity does not successfully add or remove any ECS instances, the cool-down time will not start.

After a scaling group is re-enabled after being disabled, the cool-down time is no longer in effect. For example, if a scaling activity is completed at 12:00 PM and the cool-down time is 15 minutes, the scaling group is then disabled and re-enabled, and the cool-down time is no longer in effect. If a request for triggering a scaling activity is sent at 12:03 PM from the CloudMonitor, the requested scaling activity is executed immediately.

Scaling group statuses

A scaling group has three statuses: Active, Inactive, and Deleting

Status	OpenAPI indicator
Creating	Inactive
Created	Inactive
Enabling	Inactive
Running	Active
Disabling	Inactive
Stopped	Inactive
Deleting	Deleting

Scaling activity process

A scaling activity's lifecycle starts with determining the scaling group's health status and boundary

conditions and ends with enabling the cool-down time.

Automatic scaling

Scaling up

1. Determine the scaling group's health status and boundary conditions.
2. Allocate the activity ID and execute the scaling activity.
3. Create ECS instances.
4. Modify Total Capacity.
5. Allocate IP addresses to the created ECS instances.
6. Add the ECS instances to the RDS access whitelist.
7. Launch the ECS instances.
8. Attach the ECS instances to the Server Load Balancer and set the weight to 0.
9. Wait 60s and then set the weight to 50.
10. Complete the scaling activity, and enable the cool-down time.

Scaling down

1. Determine the scaling group's health status and boundary conditions.
2. Allocate the activity ID and execute the scaling activity.
3. The Server Load Balancer stops forwarding traffic to the ECS instances.
4. Wait 60s and then remove the ECS instances from the Server Load Balancer.
5. Disable the ECS instances.
6. Remove the ECS instances from the RDS access whitelist.
7. Release the ECS instances.
8. Modify Total Capacity.
9. Complete the scaling activity, and enable the cool-down time.

Manually add or remove an existing ECS instance

Manually add an existing ECS instance

1. Determine the scaling group's health status and boundary conditions, and check the ECS instance's status and type.
2. Allocate the activity ID and execute the scaling activity.
3. Add the ECS instance.
4. Modify Total Capacity.
5. Add the ECS instance to the RDS access whitelist.
6. Attach the ECS instance to the Server Load Balancer and set the weight to 0.
7. Wait 60ss and then set the weight to 50.
8. Complete the scaling activity, and enable the cool-down time.

Manually remove an existing instance

1. Determine the scaling group's health status and boundary conditions.
2. Allocate the activity ID and execute the scaling activity.
3. The Server Load Balancer stops forwarding traffic to the ECS instance.
4. Wait 60s and then remove the ECS instance from the Server Load Balancer.
5. Remove the ECS instance from the RDS access whitelist.
6. Modify Total Capacity.
7. Remove the ECS instance from the scaling group.
8. Complete the scaling activity, and enable the cool-down time.

Scaling activity statuses

Scaling activity statuses

A scaling activity is in the **Rejected** state if the request for execution is rejected.

A scaling activity is in the **In Progress** state if it is being executed.

After a scaling activity is completed, there are three possible states:

Successful: The scaling activity has successfully added or removed the ECS instances to or from the scaling group as specified by the MaxSize value or the MinSize value adjusted by the scaling rule.

Note: When an ECS instance is successfully added to a scaling group, it has been created and added to the Server Load Balancer instance and the RDS access whitelist. If any of the above steps fail, the ECS instance is considered "failed".

Warning: The scaling activity fails to add or remove at least one ECS instance to or from the scaling group as specified by the MaxSize value or the MinSize value adjusted by the scaling rule.

Failed: The scaling activity fails to add or remove any ECS instance to or from the scaling group as specified by the MaxSize value or the MinSize value adjusted by the scaling rule.

Example

A scaling rule is defined to be added 5 ECS instances. The existing Total Capacity of the scaling group is 3 ECS instances, and the MaxSize value is 5 ECS instances. When the scaling rule is executed, Auto Scaling adds only 2 ECS instances as specified by the MaxSize value. After the scaling activity is completed, there are three possible states:

- Successful: 2 ECS instances are created successfully and correctly added to the Sever Load Balancer instance and the RDS access whitelist.
- Warning: 2 ECS instances are created successfully, but only one is correctly added to the Sever Load Balancer instance and the RDS access whitelist. The other one failed, and is rolled back and released.
- Failed: No ECS instances are created. Or 2 ECS instances are created successfully, but neither are added to the Server Load Balancer instance or the RDS access whitelist. Both are rolled back and released.

Instance rollback resulting from a scaling activity failure

Instance rollback resulting from a scaling activity failure

When a scaling activity fails to add one or more ECS instances to a scaling group, the failed ECS instances are rolled back. The scaling activity is not rolled back.

For example, if a scaling group has 20 ECS instances, out of which 19 instances are added to the Server Load Balancer instance, only the one ECS instance that fails to be added is automatically released.

Auto Scaling uses Alibaba Cloud's Resource Access Management (RAM) service to adjust the resources of ECS instances through ECS Open APIs. Therefore, API usage fees apply.

Remove an unhealthy ECS instance

Remove an unhealthy ECS instance

After an ECS instance has been successfully added to a scaling group, the Auto Scaling service regularly scans its status. If the ECS instance is not in the running state, Auto Scaling removes the ECS instance from the scaling group.

- If the ECS instance was created automatically, Auto Scaling immediately removes and releases it.
- If the ECS instance was added manually, Auto Scaling immediately removes it, but does not stop or release it.

The removal of unhealthy ECS instances is not restricted by the MinSize value. If, due to removal, the number of ECS instances (Total Capacity) in the scaling group is smaller than the MinSize value, Auto Scaling automatically adds ECS instances to the group until the number of instances reaches the MinSize value.

Notification

Notification

A text message or email is sent when a scaling activity meets either of the following conditions:

- The scaling activity is triggered by a scheduled task, CloudMonitor alarm task, or health check.
- An ECS instance has been created or released.

Forced intervention

Forced intervention

Auto Scaling does not prevent users from performing forced interventions, such as deleting

automatically created ECS instances from the ECS console. Auto Scaling handles forced interventions in the following ways:

Resource	Forced intervention types	Solutions
ECS	An ECS instance is deleted from a scaling group through the ECS console or OpenAPI.	Auto Scaling determines if the ECS instance is in an unhealthy state through Health check, and if so, removes the instance from the scaling group. The ECS instance's intranet IP address is not automatically deleted from the RDS access whitelist. When the number of ECS instances (Total Capacity) in the scaling group is smaller than the MinSize value, Auto Scaling automatically adds ECS instances to the group until the number of instances reaches the MinSize value.
ECS	The ECS OpenAPI permissions are revoked from Auto Scaling.	All scaling activity requests are rejected.
Server Load Balancer	An ECS instance is removed from a Server Load Balancer instance by force through the Server Load Balancer console or OpenAPI.	Auto Scaling does not automatically detect this action or handle such an exception. The ECS instance remains in the scaling group, but is released if it was selected according to the removal policy of a scale-down activity.
Sever Load Balancer	A Server Load Balancer instance is deleted (or its health check function is disabled) by force through the Server Load Balancer console or OpenAPI.	No ECS instance is added to the scaling group that has been added to the Server Load Balancer instance. Scaling tasks can trigger scaling rules to remove ECS instances from the scaling group. ECS instances deemed unhealthy by the health check function can also be removed.
Server Load Balancer	A Server Load Balancer instance becomes unavailable (due to overdue payment or a fault).	All scaling activities fail, except for activities that are manually triggered to remove ECS instances.
Server Load Balancer	The Server Load Balancer OpenAPI permissions are revoked from Auto Scaling.	Auto Scaling rejects all scaling activity requests for the scaling groups added to the Server Load Balancer

		instance.
RDS	The IP address of an ECS instance is removed from an RDS whitelist through the RDS console or OpenAPI.	Auto Scaling does not detect this action automatically or handle such an exception. The ECS instance remains in the scaling group. If this instance is selected according to the removal policy of a scale-down activity, it is released.
RDS	An RDS instance is deleted by force through the RDS console or OpenAPI.	The scaling group that configured the RDS instance will no longer add ECS instances. No ECS instance is added to the scaling group that has been added to this RDS instance. Scaling tasks can trigger scaling rules to remove ECS instances from the scaling group. ECS instances determined to be unhealthy by the health check function can also be removed.
RDS	An RDS instance becomes unavailable (due to overdue payment or a fault).	All scaling activities fail except for those manually triggered to remove ECS instances.
RDS	The RDS OpenAPI permissions are revoked from Auto Scaling.	Auto Scaling rejects all scaling activity requests for the scaling groups added to the RDS instance.

Quantity restrictions

At present, the quantity limits of Auto Scaling are as follows:

You can create up to 20 scaling groups.

- Up to 10 scaling configurations can be created for a scaling group.
- Up to 50 scaling rules can be created for a scaling group.
- Up to 6 event notifications can be created for a scaling group.
- Up to 6 lifecycle hooks can be created for a scaling group.

You can scale up to 1,000 ECS instances for all scaling groups in all regions. This restriction

applies to the ECS instances automatically created, but does not apply to those manually added.

You can create up to 20 scheduled tasks.

Scaling configurations

Create a scaling configuration

Auto Scaling automatically adds ECS instances into Auto Scaling groups according to the conditions you set when demands for your instances increase. In **Create Scaling Configuration**, you can specify the options of the ECS instances listed.

Note:

- You cannot modify a scaling configuration once it is created. You can replace it with a new configuration if needed, you must update it with the new scaling configuration.
- ECS instances created by the replaced scaling configuration work as usual.
- You can create up to 10 scaling configurations for one Auto Scaling group, and only one scaling configuration can be at the **Activated** status. You can **use** certain scaling configuration according to your needs.

Create a scaling configuration by using the console

Log on to the **Auto Scaling console**.

Select a **Scaling Group**.

On the navigation pane, select **Scaling Configuration**, then click **Create Scaling**

Configuration.
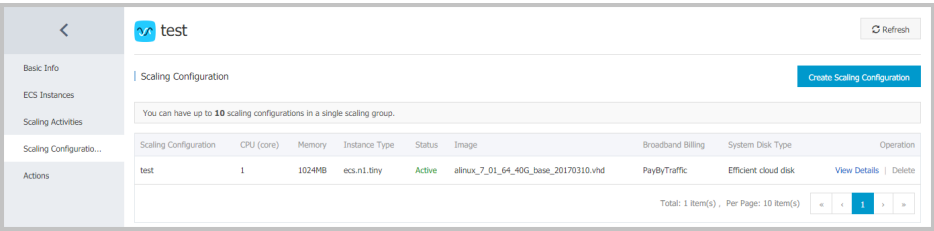
On the **Create Scaling Configuration** page, enter a name for this scaling configuration.

Select an option for each of the listed parameters on the page.

Note:

- For **Image Type**, if you want functions like the automatic start of your Web server, or the automatic download of code and scripts, select **Custom Image**.
- If you want to include CloudMonitor triggered alarms in your overall solutions of Auto Scaling, You can install CloudMonitor Agent on the ECS instance.

View a scaling configuration

You can view a scaling configuration.

Scaling configurations have the following life cycle status:

- Active: The scaling group uses the scaling configuration in active status to create ECS instances.
- Inactive: Inactive scaling configurations are still in a scaling group, but are not used to create ECS instances.

Procedure

1. On the **Scaling Group List** page, click **Manage** next to the group to be managed.

On the **Scaling Configuration** page, click **View Details**.

Delete scaling configuration

Delete scaling configuration

You can delete scaling configuration.

Note:

- Active scaling configuration cannot be deleted.
- If any ECS instances still used for a scaling group and are created according to the scaling
  configuration, the scaling configuration cannot be deleted.

Scaling groups

Create a scaling group

Create a scaling group

A scaling group is a collection of ECS instances with similar configuration deployed in an application
scenario. It defines the maximum and minimum number of ECS instances in the group, associated

Server Load Balancer and RDS instances, and other attributes.

See Removal policies to remove an ECS instance from a scaling group.

Parameter description

This operation creates a scaling group according to input parameters.

MaxSize and MinSize define the maximum and minimum number of ECS instances in the scaling group.

When the number of ECS instances (Total Capacity) in the scaling group is smaller than the MinSize value, Auto Scaling adds ECS instances to the group until the MinSize value is reached.

When the number of ECS instances (Total Capacity) in the scaling group is greater than the MaxSize value, Auto Scaling removes ECS instances from the group until the MaxSize value is reached.

DefaultCooldown specifies the default cool-down time for the scaling group.

During the cool-down time after a scaling activity (adding or removing ECS instances) is executed, the scaling group cannot execute any other scaling activity.

Currently, this only applies to the scaling activities triggered by the alarm tasks of CloudMonitor.

RemovalPolicy determines how ECS instances should be removed from the scaling group when multiple candidates for removal exist.

Considerations for working with Server Load Balancer

If a Server Load Balancer instance is specified for a scaling group, Auto Scaling automatically adds the ECS instances to the Server Load Balancer instance while adding them to the scaling group.

You must enable the specified Server Load Balancer instance.

You must enable the health chech for all listener ports of the Server Load Balancer instance.

Otherwise, the scaling group creation will fail.

The default weight of ECS instances added to a Server Load Balancer instance is 50.

Considerations for working with RDS

If an RDS instance is specified for a scaling group, the scaling group adds the intranet IP addresses of the ECS instances in the group to the specified RDS instance's access whitelist.

The specified RDS instance must be in the running state.

The specified RDS instance's access whitelist must have room for more IP addresses.

The scaling group does not take effect immediately after creation. It must be enabled to trigger scaling rules and execute scaling activities.

Restrictions

The scaling group, Server Load Balancer instance, and RDS instance must be in the same region.

You can create 20 scaling groups at most.

Procedure

On the Scaling Group List page, click Create Scaling Group.
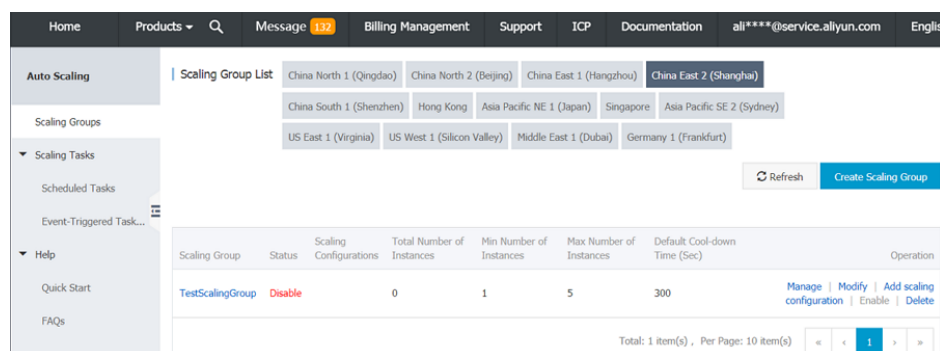
On the Create Scaling Group page, enter the Scaling Group Name.

Enter the **Maximum Number of Instances Allowed for Scaling (Unit)** and **Minimum Number of Instances Allowed for Scaling (Unit)**. If they are set to **1**, an ECS instance is created automatically upon the creation of the scaling solution.

Enter the **Default Cool-down Time(Sec)**, **Removal Policy** and **Network Type**.

Select the **Sever Load Balancer** and **RDS** database instances as needed.

Click **Submit**.

Query a scaling group

Query a scaling group

You can query a scaling group's information.

Scaling groups have the following life cycle states:

- Active: The scaling group can receive requests to execute scaling rules and trigger scaling activities.

- Inactive: The scaling group cannot receive any requests to execute scaling rules.
- Deleting: The scaling group is being deleted and cannot receive any requests to execute scaling rules.

Procedure

On the **Scaling Groups** page, click **Manage** next to the scaling group to query a scaling group.

Modify a scaling group

You can modify the attributes of a scaling group after it is created.

Note the following attributes cannot be modified:

- Region
- Server Load Balancer
- RDS database instance

This operation can be executed only for scaling groups in active or inactive status.

If the number of ECS instances (Total Capacity) in the scaling group does not meet the new MaxSize or MinSize settings, Auto Scaling adds or removes ECS instances to or from the group until the MaxSize or MinSize value is reached.

Procedure

On the **Scaling Groups List** page, click **Modify** next to the scaling group whose settings are to be modified.

Scaling Group Name/ID	Status	Scaling Configurations	Total Number of Instances	Min Number of Instances	Max Number of Instances	Default Cool-down Time (Sec)	Operation
classic asg-uf6f3xewn3dvz4bsy7r1	Enable	classic	1	1	1	300	Manage │ Modify │ Disable │ Delete

On the **Modify Scaling Group** page, modify the group's settings as needed.

Enable a scaling group

You can enable a scaling group so that, if the existing number of ECS instances is less than the MinSize value of an enabled (Active) scaling group, Auto Scaling creates Pay-As-You-Go ECS instances until the MinSize value is reached.

For example, if MinSize is set to 5 during the creation of a scaling group, and the group has only 2 ECS instances when enabled, Auto Scaling creates 3 ECS instances to add to the group.

You can only enable inactive scaling groups with active scaling configurations.

A single scaling group can have one active scaling configuration at a time.

Procedure

On the **Scaling Group List** page, click **Enable** next to the scaling group to enable a scaling group.

Disable a scaling group

Disable a scaling group

You can click **Disable** to disable a scaling group.

   - A scaling group can be disabled only when it is not executing any scaling activity.
   - This operation can only be executed for active scaling groups.

Delete a scaling group

Delete a scaling group

You can delete a scaling group.

The ForceDelete attribute indicates whether to forcibly delete a scaling group and remove and release ECS instances if it has ECS instances, or if scaling activities are in progress.

   - This attribute can only be viewed using OpenAPI.
   - By default, a scaling group is deleted in ForceDelete mode from the console.

If ForceDelete is set to False, a scaling group must meet the following two conditions before being deleted:

   - Condition 1: The scaling group is not executing any scaling activity.
   - Condition 2: The existing number of ECS instances (Total Capacity) in the scaling group is 0.

When the two conditions are met, the scaling group is stopped and deleted.

When ForceDelete is set to True:

   1. Stop the scaling group so it rejects new scaling activity requests.
   2. Wait until the ongoing scaling activities are completed.
   3. Remove all ECS instances from the scaling group and delete the group. Manually added ECS instances are removed from the scaling group, while ECS instances created by Auto Scaling are automatically deleted.

Deleting a scaling group also deletes its scaling configurations, scaling rules, scaling activities, and scaling requests.

However, deleting a scaling group does not delete scheduled tasks, CloudMonitor alarm tasks, Server

Load Balancer instances, or RDS instances.

Procedure

1. On the **Scaling Groups** page, click **Delete** next to the scaling group to be deleted.
2. On the **Delete Scaling Group** page, click **Confirm**.

Scaling rules

Create a scaling rule

A scaling rule defines specific scaling actions; for example, adding or removing ECS instances. If, as a result of executing a scaling rule, the number of ECS instances in a scaling group is less than the MinSize value or greater than the MaxSize value, Auto Scaling automatically adjusts the number of ECS instances to be added or removed. This occurs by executing the adjust scaling group instance quantity to MinSize or MaxSize rules.

For example, a scaling group is created with MaxSize 3, Total Capacity is 2, and the scaling rule is Add 3 ECS instances. In this case, Auto Scaling only **adds 1** ECS instance (the scaling rule will still be set to Add 3 ECS instances, but the Adjust scaling group instance quantity to MaxSize rule will also be run).

Imagine a second scaling group, in which MinSize is 2, Total Capacity is 3, and the scaling rule is Remove 5 ECS instances. Here, Auto scaling only **removes 1** ECS instance (since the Adjust scaling group instance quantity to MinSize rule will also be run).

Description

You can create a scaling rule according to input parameters.

If Adjusted Type is set to Adjust to Specified Total Capacity, the corresponding Adjusted Value must be greater than or equal to 0.

If Adjusted Type is set to Percent Change In Capacity, Auto Scaling computes the number of ECS instances to be added or removed by using the following formula: current number of ECS instances (Total Capacity) * Adjusted Value/100, rounding the result to the nearest

integer.

After a scaling activity is executed for a scaling group, the group is cooled down for the time specified in the scaling rule. If not specified in the scaling rule, the Default cool-down time of the scaling group is applied.

You can create up to 50 scaling rules for a single scaling group.

After a scaling rule is created, a unique scaling rule identifier (scaling rule ARI) is generated and used in these OpenAPIs:

Set it in the ScalingRuleAri parameter of the ExecuteScalingRule interface, and you can manually execute the scaling rule.

Set it in the ScheduledAction parameter of the CreateScheduledTask interface, and you can schedule the execution of the scaling rule.

Set it in the AlarmActions parameter of the CloudMonitor PutAlarmRule interface, and you can dynamically execute the scaling rule with the performance indicators.

Procedure

On the **Scaling rule** page, click **Create scaling rule**.
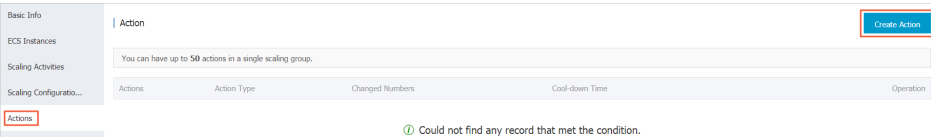
Enter the rule name and set the parameters.

In the **Create Scaling rule** dialog box, click **Create Scaling rule**. The new scaling rule will be displayed on the **Scaling rule** page.

Modify, query, or delete a scaling rule

Modify, query, or delete a scaling rule

You can modify, delete, or query a scaling rule.

Triggered tasks

Execute a scaling rule

Execute a scaling rule

You can execute a scaling rule for a scaling group given the following conditions:

- The scaling group is active.
- The scaling group is not executing any scaling activity.

If no scaling activity is being executed for the scaling group, the scaling rule is executed directly without waiting for the cool-down time.

A successful return indicates that Auto Scaling will shortly execute the scaling activity, but does not mean the scaling activity will be successfully executed. Use the returned ScalingActivityID to check the status of the scaling activity.

Execution rules:

- If the number of ECS instances to be added by the scaling rule plus the number of existing ECS instances in the group (Total Capacity) exceeds the MaxSize value, the Total Capacity is adjusted to the MaxSize value.
- If the number of existing ECS instances in the scaling group (Total Capacity) minus the number of ECS instances to be removed by the scaling rule is less than the MinSize value, the Total Capacity is adjusted to the MinSize value.

Up to 1,000 ECS instances are supported for auto scaling for all scaling groups in all regions. This restriction applies only to ECS instances that were automatically created, not those manually added.

Procedure

On the **Scaling group management** page, click **Manage** next to the group to be managed.

Click **Scaling rule** on the left side, and then click **Execute** next to the scaling rule to be executed.

Click **Confirm**.
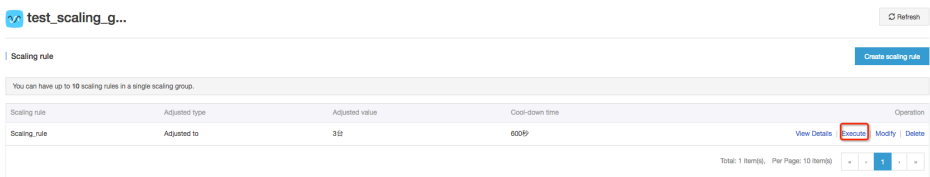
Add an ECS instance

Add an ECS instance

You can add an ECS instance to a scaling group. There are two types of ECS instances: Subscription and Pay-As-You-Go. The ECS instance to be added must meet the following conditions:

- The ECS instance is in the same region as the scaling group.
- The ECS instance is the type as specified in the active scaling configuration.
- The ECS instance is running.
- The ECS instance is not in any other scaling group.
- The ECS instance can be the classic type or VPC, but has the following restrictions:
    - If the scaling group is the classic type, only classic type instances can be added.
    - If the scaling group is the VPC type, only instances belonging to the same VPC can be added.

To add an ECS instance, the scaling group must meet the following conditions:

- The scaling group is active.
- The scaling group is not executing any scaling activity.

When no scaling activity is being executed for the scaling group, adding an ECS instance is executed directly without waiting for the cool-down time.

A successful return indicates that the Auto Scaling service will shortly execute the scaling activity, but does not mean that the scaling activity will be successfully executed. Use the returned ScalingActivityID to check the scaling activity status.

If the number of ECS instances to be added by the scaling rule plus the number of existing ECS instances in the scaling group (Total Capacity) exceeds the MaxSize value, the operation fails.
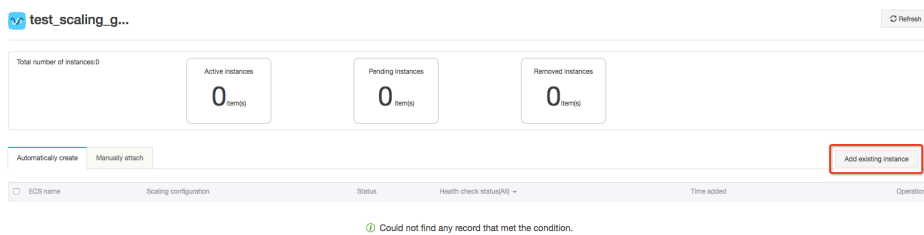
Manually added ECS instances are not associated with the active scaling configuration in the scaling group.

Procedure

On the **Scaling group management** page, click **Manage** next to the scaling group.

Click **ECS instance list** on the left side.

Click **Add existing instance**.

Remove an ECS instance

Remove an ECS instance

You can remove an ECS instance from a specified scaling group.

- When an automatically created ECS instance is removed from a scaling group, the instance is
  stopped and released.
- When a manually added ECS instance is removed from a scaling group, the instance is not
  stopped or released.

The operation will succeed under the following conditions:

- The scaling group is active.
- The scaling group is not executing any scaling activity.

When no scaling activity is being executed for the scaling group, removing an ECS instance is
executed directly without waiting for the cool-down time.

A successful return indicates that the Auto Scaling service will shortly execute the scaling activity, but
it does not mean that the scaling activity will be successfully executed. Use the returned
ScalingActivityID to check the scaling activity status.

If the number of existing ECS instances in the scaling group (Total Capacity) minus the number of ECS
instances to be removed is less than the MinSize value, the operation fails.

Procedure

On the **Scaling group management** page, click **Manage** next to the scaling group.

Click **ECS instance list** on the left side.

Click **Remove from scaling group and release** next to the ECS instance to be removed.
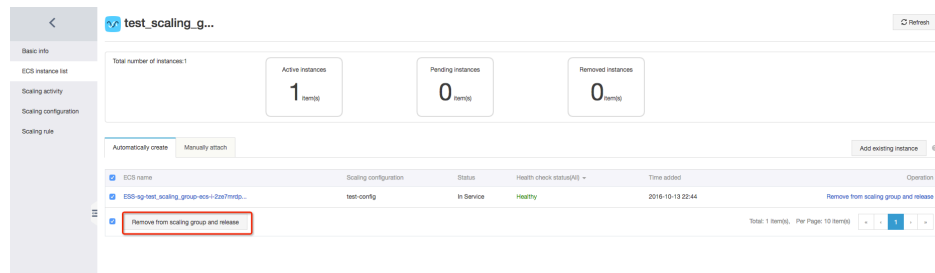
Scheduled tasks

Create a scheduled task

Create a scheduled task

You can create up to 20 scheduled tasks according to input parameters.

If a scheduled task fails to trigger the execution of a scaling rule because the scaling group is executing a scaling activity or the scaling group is disabled, the scheduled task is automatically retried within the Launch Expiration Time. After the Launch Expiration Time expires, the scheduled task is abandoned.

If multiple tasks are scheduled at similar times to execute the scaling rule of a scaling group, the earliest task triggers the scaling activity first. Other tasks will attempt to execute the rule within their Launch Expiration Time, but a scaling group executes only one scaling activity at a time. If another scheduled task is still triggering attempts within its Launch Expiration Time after the scaling activity is finished, the scaling rule is executed and the corresponding scaling activity is triggered.

If multiple tasks are scheduled at the same time, the latest scheduled task is executed.

Procedure

Click **Scheduled task** under **Auto-trigger task management** to display the Scheduled Task page.

Click **Create scheduled task** to display the **Create Scheduled task** dialog box.

Enter the task name.

Enter the execution time. If recurrence is not set, the task is executed once on the designated date and time. Otherwise, this task is executed periodically at the specified time.

Enter the recurrence.

Select a scaling group and scaling rule to be triggered by the scheduled task.

Click **Submit**. The scheduled task is displayed on the **Scheduled Task** page.

Modify, disable or delete a scheduled task

Modify, disable or delete a scheduled task

You can modify, disable, or delete a scheduled task.

Procedure

Click **Scheduled task** under **Auto-trigger task management** to display the Scheduled task page.
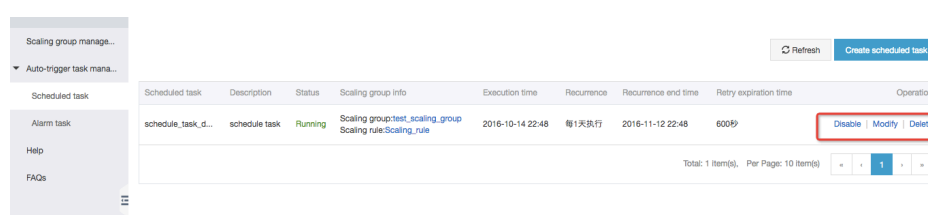
Click **Modify**, **Disable**, or **Delete** next to the scheduled task and change it as desired.

Alarm tasks

Create an alarm task

Create an alarm task

You can create an alarm task according to input parameters. The alarm task names must be unique within a user account. If an alarm task with the same name already exists, it will be overwritten with the new values.

Procedure

Click **Alarm task** under **Auto-trigger task management** to display the alarm task list page.

Click **Create alarm task** to display the alarm task creation dialog box.

3. Enter the task name.
4. Select the scaling group to be monitored.
5. Select the item to be monitored.
6. Enter the statistical period. The finer the granularity of the statistical cycle, the more sensitive the alarm trigger mechanism.
7. Enter the statistical method.
8. Enter the number of recurrences before an alarm is triggered.
9. Select the scaling rule triggered by the alarm.

Click **Submit**. The alarm task will be displayed on the **Alarm Task** page.
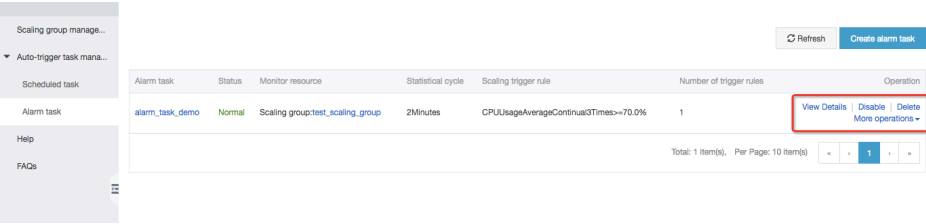

Modify, query, or delete an alarm task

Modify, query, or delete an alarm task

You can modify, query, or delete an alarm task.

Procedure

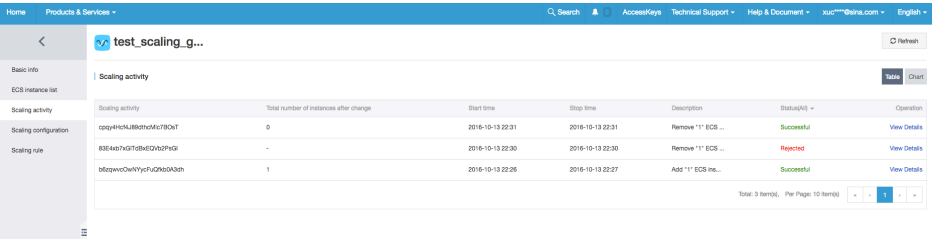Click **Alarm task** under **Auto-trigger task management** to access the Alarm task page.

Click **Modify**, **Disable**, or **Delete** next to the alarm task to be changed.

View scaling activities

View scaling activities

This operation queries the information of scaling activities performed in the last 30 days.

Example

Move ECS instance to Standby

Auto Scaling allows you to set the Standby status for one or more ECS instances. After an ECS instance is in the Standby status, you can upgrade or maintain the ECS instance. Meanwhile, we do not either perform health check for the specified instance or release it.

If an ECS instance is set to the Standby status:

- It is not in service until you resume the ECS instance.
- Its lifecycle is controlled by you rather than Auto Scaling service.
- The weight of the ECS instance is deregistered to zero if the scaling group has Server Load Balancer instances attached.
- You can **stop instance**, **restart instance**, or do other maintenance operations, such **upgrade the instance configurations**, **change the operating system**, **reinitialize the cloud disk**, or **migrate from the classic network to a VPC**.
- It is not removed from the scaling group whenever a scaling event happens.
- The health status is not updated even the specified instance is stopped or restarted.
- It must be removed from the scaling group before you release the instance.
- It is resumed for a short while when you delete the related scaling group and then it is release along with the scaling group.

If an ECS instance is back to the in service status:

- It handles application traffic actively again.
- The weight of the ECS instance is set to a predefined value if the scaling group has Server Load Balancer instances attached.
- The health status is updated if the specified instance is stopped or restarted.
- Its lifecycle is controlled by Auto Scaling service rather than you.

Console-based operation

Move to Standby

Log on to the **Auto Scaling management console**.

In the left-side navigation pane, click **Scaling Groups**.

Choose a **region**, such as China East 2 (Shanghai).

Find and click the target scaling group.

In the left-side navigation pane, click **ECS instances**.

Find and click the target ECS instance, click **Move to Standby**.

Remove from Standby

Log on to the **Auto Scaling management console**.

In the left-side navigation pane, click **Scaling Groups**.

Choose a **region**, such as China East 2 (Shanghai).

Find and click the target scaling group.

In the left-side navigation pane, click **ECS instances**.

Find and click the target ECS instance, click **Remove from Standby**.

API-based operation

Move to Standby: **EnterStandby**

Remove from Standby: **ExitStandby**

References

- What is Server Load Balancer
- Remove an unhealthy ECS instance

Query the ECS instance list

Query the ECS instance list

Query the ECS instance list of a scaling group

There are two types of ECS instances: automatically created and manually added.

- Automatically created ECS instances are created by the Auto Scaling service based on scaling
  configuration and rules.
- Manually added ECS instances are manually added to a scaling group, not created by the
  Auto Scaling service.

Lifecycle of ECS instances in a scaling group

An ECS instance in a scaling group may be in the following states during its lifecycle:

- Pending: The ECS instance is being added to the scaling group. The instance may be being
  created, added to the Server Load Balancer instance, or added to the RDS access whitelist.
- In Service: The ECS instance has been added to the scaling group and is providing services.
- Removing: The ECS instance is being removed from the scaling group.

ECS instance health status

An ECS instance in a scaling group may be in the following health states:

- Healthy
- Unhealthy

ECS instances are unhealthy when they are not running. Auto Scaling automatically removes
unhealthy ECS instances from scaling groups. Auto Scaling stops and releases unhealthy ECS
instances that were created automatically. Auto Scaling does not stop and release unhealthy ECS
instances that were created manually.

Procedure

On the **Scaling group management** page, click **Manage** next to the scaling group.

Click **ECS instance list** on the left side.

Lifecycle Hook