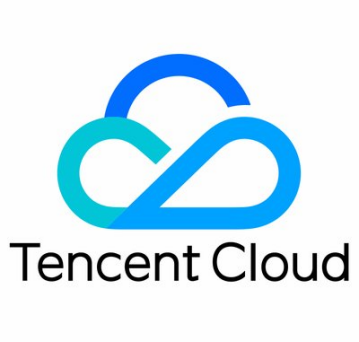
Auto Scaling

Expanding and Reducing Capacity



Contents

Expanding and Reducing Capacity Lifecycle Hook

Last updated: 2024-01-18 11:02:24

Use Cases

Within the scaling group, you can configure elastic expansion and contraction activities. If you wish to perform custom operations before officially launching these instances, the lifecycle hook feature can assist you in accomplishing this.

- After expanding the scaling group, there is a need to delay for a period before mounting the instances to the CLB, and then providing services externally.
- Executing data backup operations when the scaling group releases instances.
- Executing user-defined operations during the elastic expansion or contraction of the scaling group.

> **Note:**
> - The lifecycle hook only takes effect when instances are automatically created or removed; it does not affect other instances within the scaling group.
> - If an expansion mode is set, it will also take effect when manually adding or removing instances, as well as during power on and off.
> - Only ten lifecycle hooks can be created within a single scaling group.

Mode of Operation

After a lifecycle hook is created in a scaling group, the associated scaling activity is suspended when the lifecycle hook event occurs, allowing you to perform custom operations during this suspension period. The suspension will terminate when the lifecycle hook times out.

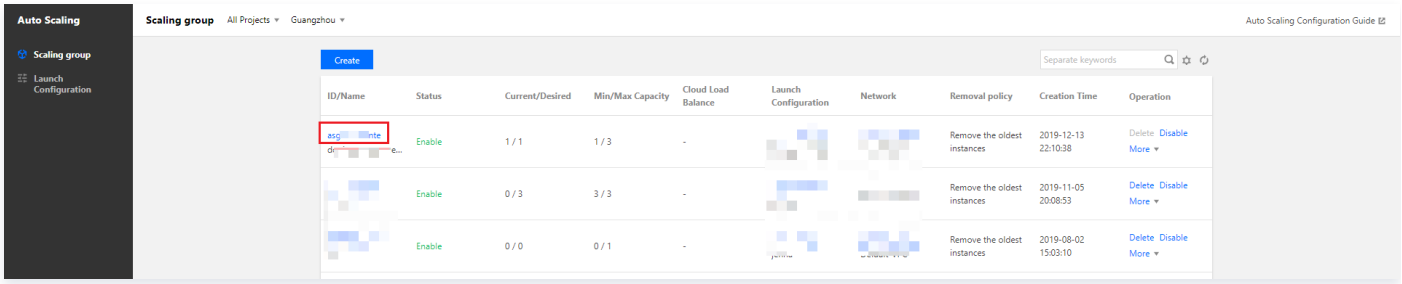
Lifecycle Hook Attributes

Name	Note	Sample
Name	The name of the lifecycle hook, which only supports Chinese, English, numbers, underscores, hyphens, and decimal points.	fehwnl_
Scaling activity type	Elastic contraction activity/Elastic expansion activity	Scale In
Expanded activities	Expansion activities include NORMAL and EXTENSION, with the default value being NORMAL. • NORMAL: Lifecycle hooks are only effective when instances are automatically created or removed. • EXTENSION: Lifecycle hooks will also be effective during manual instance addition, removal, and power cycling.	NORMAL
Timeout period	The default duration for which an instance remains in the waiting state. It must be an integer between 30 and 7200 seconds.	300
Execution policy	The execution policy includes **Proceed** and **Reject**. • Proceed Policy: The suspended scaling activity will continue to execute. • Reject Policy: For elastic expansion activities, the created CVM instances will be directly released. There is no impact on elastic contraction activities. When multiple lifecycle hooks are set for a scaling group, the operations are executed in sequence. However, only the first reject policy is executed, and the reject policies of subsequent lifecycle hooks are not effective.	Deny
Notification method	Notification methods include TDMQ topics, TDMQ queues, and TAT commands. After selecting a notification method, you also need to choose a specific TDMQ topic, TDMQ queue, or TAT command.	CMQ Topic

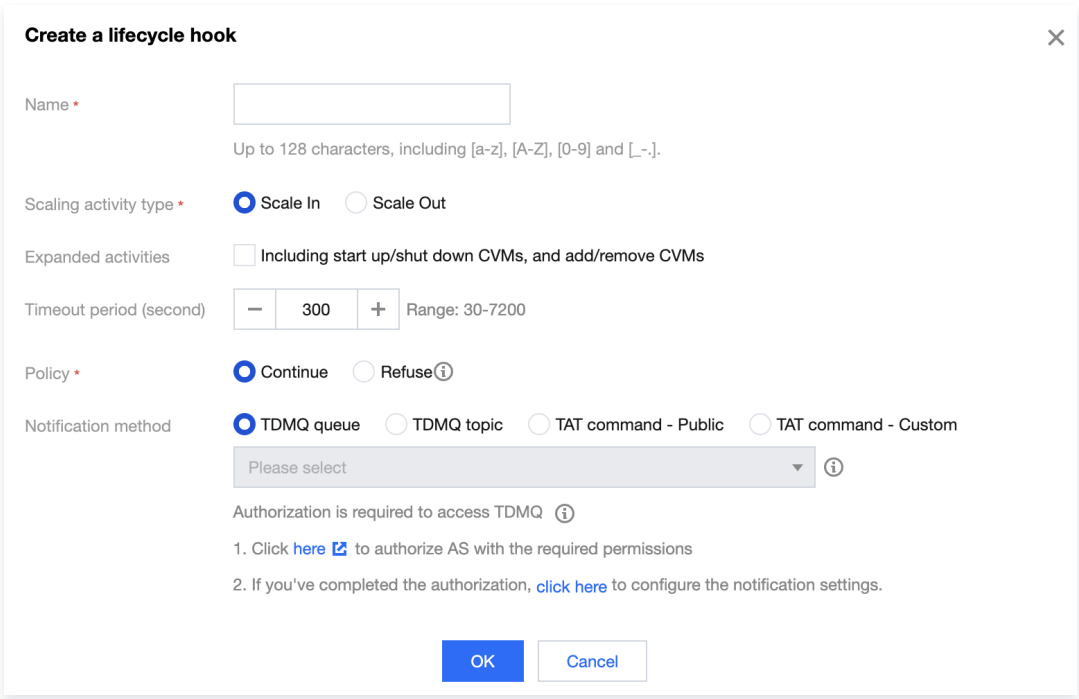
Notification identifier	Each time Auto Scaling pushes a message to the notification object, it concurrently sends the notification identifier you have pre-specified, facilitating the management and tagging of different categories of notification information.	Notification message

Create Lifecycle Hook

1. Log in to the Auto Scaling Console.
2. Select the scaling group to which you need to bind the lifecycle hook, click on the scaling group ID/name, and enter the details page of that scaling group, as shown in the figure below:



3. Select the Lifecycle Hook tab and click on New.
4. In the pop-up New Lifecycle Hook dialog box, fill in the relevant information about the lifecycle hook, as shown in the figure below:



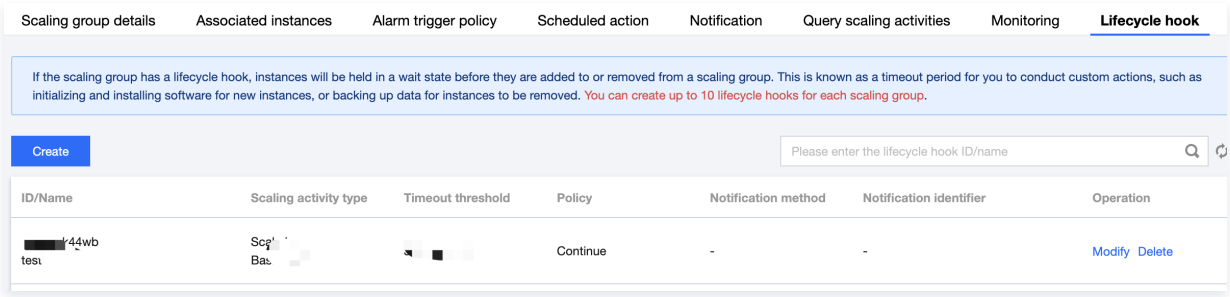
Note:

- When creating a lifecycle, you need to select or create a new TDMQ. This includes creating a TDMQ Topic and a TDMQ Queue.
- In the case of setting multiple lifecycle hooks for a single scaling group, they will wait in sequence, but only the first rejection policy will be executed, subsequent lifecycle hooks will not take effect.
- If you do not specify a notification method, you will not receive any default notifications.
- Within the same scaling group, lifecycle hook names cannot be duplicated.

Modify lifecycle hook

1. Log in to the Auto Scaling Console.
2. Select the scaling group for which you need to modify the lifecycle hook, click on the scaling group ID/name to enter the details page of that scaling group.

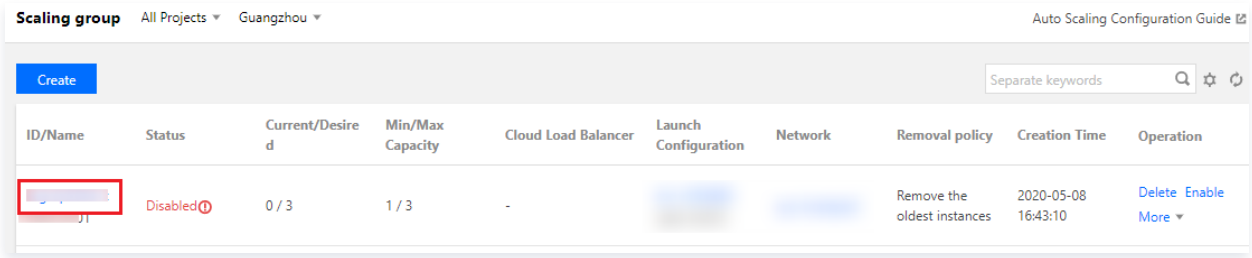
3. Select the **Lifecycle Hook** tab, in the row where you need to modify the lifecycle hook, click on **Modify**, as shown in the figure below:



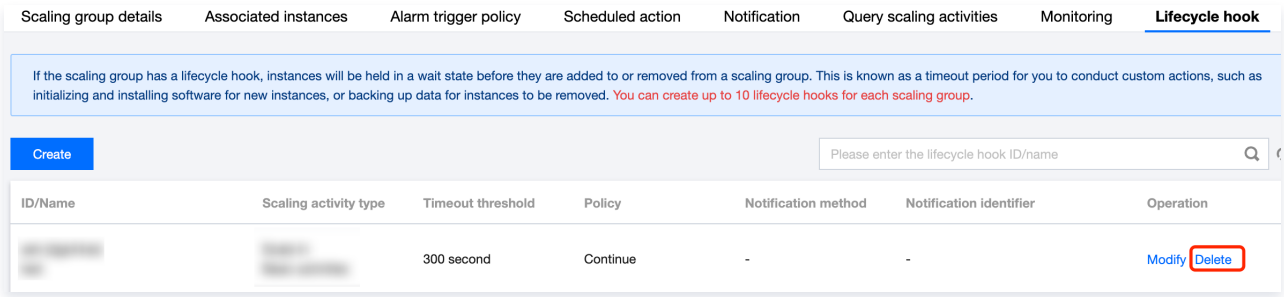
4. In the pop-up **Modify Lifecycle Hook** window, you can modify the information according to your actual needs.

Delete lifecycle hook

1. Log in to the Auto Scaling Console.
2. Select the scaling group from which you need to delete the lifecycle hook, click on the scaling group ID/name to enter the details page of that scaling group, as depicted in the figure below:



3. Select the **Lifecycle Hook** tab, in the row where you need to delete the lifecycle hook, click on **Delete**, as depicted in the figure below:



4. In the pop-up **Delete Lifecycle Hook** window, simply click on **Confirm**.

Managing Scheduled Actions

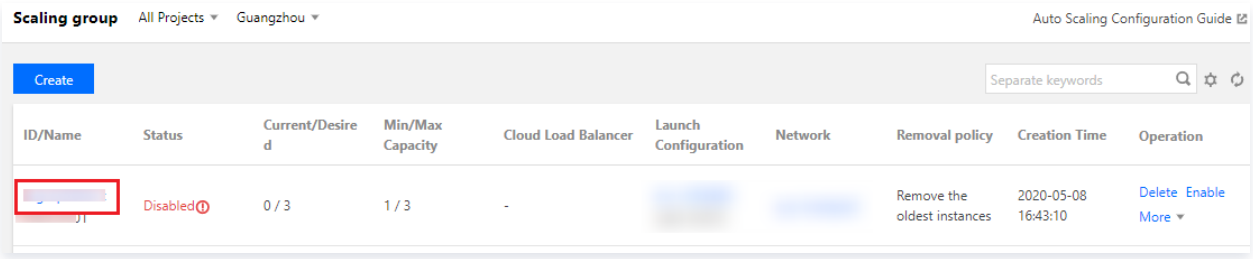
Last updated: 2024-01-17 17:57:07

Introduction to Scheduled Tasks

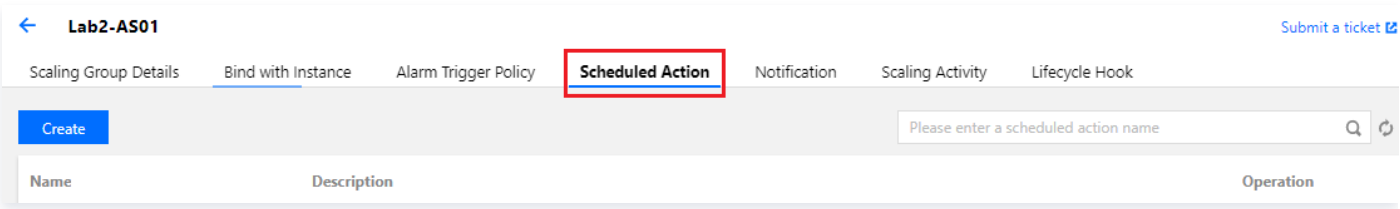
Scheduled tasks, essentially, are time-based plans that allow your business to predictably scale up or down the number of cloud server instances in use, in accordance with load variations.
For instance, if the traffic to your web application begins to increase every Wednesday, maintains a high volume on Thursday, and starts to decrease on Friday, you can plan scaling activities based on this predictable traffic pattern of your web application.
To create a scheduled scaling action, specify the start time when you want the scaling action to take effect, as well as the new minimum size (minimum number of instances), maximum size (maximum number of instances), and desired size (expected number of instances) for the scaling action. At the specified time, Auto Scaling will update the number of instances in the scaling group according to these set values.
You can create pre-planned actions for a one-time scaling, or establish pre-planned actions for scaling on a regular schedule.

Managing Scheduled Tasks

1. Log into the Auto Scaling console and select **Scaling Groups** from the left navigation bar.
2. Select the scaling group you wish to modify, click on the scaling group ID to enter the basic information page of the scaling group, as shown in the figure below:



3. On the details page of the scaling group, select the **Scheduled Action** tab to manage the scheduled tasks associated with the scaling group, as depicted in the figure below:



- Click on **New** to add a new scheduled task.
- Select a specific scheduled task and click on **Edit**. You can modify the task name, adjust the execution time, set whether it is to be executed periodically, and modify the execution activity in the pop-up page.
- Click on **Delete** to remove the selected scheduled task.

> **Note:**
> If you wish to create a task that repeats at scheduled intervals, you can specify a start time. The Auto Scaling (AS) will execute the operation at this time and then proceed according to the repetition schedule. If an end time is specified, AS will cease to execute operations after this time.

Managing an Alarm-triggered Policy

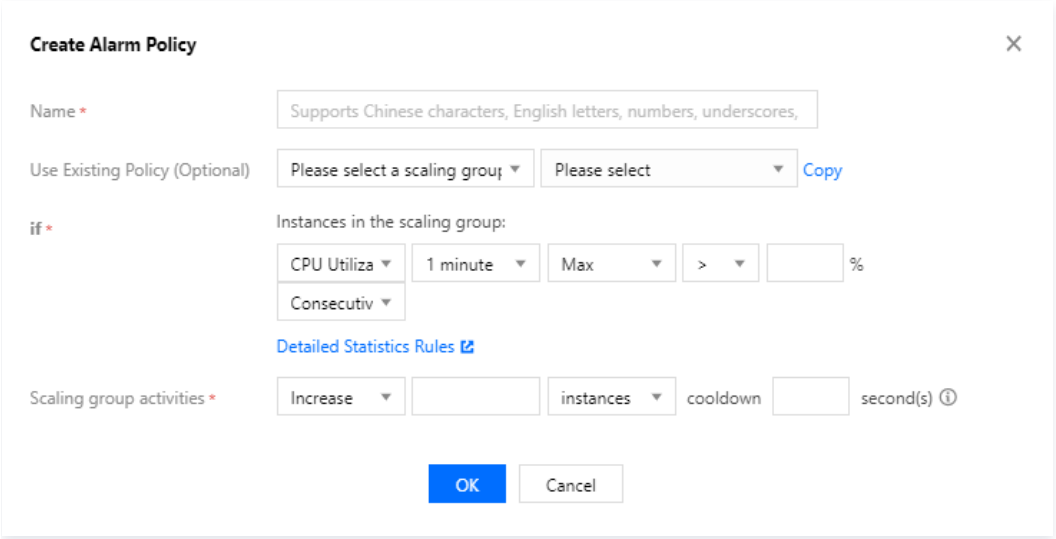
Last updated: 2024-01-17 17:57:16

Feature Overview

Auto Scaling (AS) dynamically adjusts the number of instances in the scaling group based on monitored metrics. You need to define an alarm trigger strategy, which includes the status of the monitored metrics that trigger expansion and how to scale according to demand changes. The alarm trigger strategy encompasses both simple and target tracking strategies.

Basic Strategy

To create an alarm strategy, specific conditions and actions must be designated, as illustrated below:



- The condition format is: a specific metric + threshold + period + number of consecutive periods reaching the threshold. That is, the metric has reached the threshold for N consecutive periods.
- The execution actions are: sending notifications + increasing/decreasing a specified number of cloud servers.

You can create two simple strategies for each scaling group: one for expansion and another for contraction. When the business volume reaches the conditions specified by the alarm strategy, AS will execute the associated strategy to contract (by terminating instances) or expand (by launching instances) the scaling group.

Target Tracking Strategy

Each scaling group supports the creation of a single target tracking strategy. The target tracking strategy will automatically calculate the required number of instances and perform scaling up or down based on the alarm value of the selected monitoring indicator, the set target value, and the number of instances in the scaling group, thereby keeping the monitoring indicator close to the target value. To create a target tracking strategy, it is necessary to specify predefined metrics, target values, warm-up time, and whether to disable scaling down.

- **Select Metric**

  The target tracking strategy has certain limitations on applicable monitoring metrics. The monitoring metrics applicable to the target tracking strategy must be valid usage rate metrics, capable of accurately reflecting the busyness of instances, and the metric values need to increase or decrease proportionally with the number of scaling group instances. Only when these conditions are met, the target tracking strategy can use metric values to proportionally scale up or down the number of instances. When creating a target tracking strategy, the type of monitoring metric must be specified. The supported monitoring metrics include:
  - Average CPU Utilization of Scaling Group
  - Average Outbound Bandwidth of Scaling Group on Internal Network
  - Average Inbound Bandwidth of Scaling Group on Internal Network
  - Average Outbound Bandwidth of Scaling Group on External Network
  - Average Inbound Bandwidth of Scaling Group on External Network
- **Target Value**

A target tracking policy must specify a target value. This value represents the optimal utilization or throughput of the scaling group. Generally, under the conditions that satisfy the Tencent Cloud Observability Platform's alarm, when the monitored metric alarm value exceeds the target value, it triggers the scaling group to expand. When the monitored metric alarm value is less than 80% of the target value, it triggers the scaling group to contract. The number of instances for expansion or contraction is determined by the ratio of the alarm value to the target value (80% of the target value when contracting), and the instances within the scaling group are expanded or reduced proportionally. The calculated number of instances (which may be a decimal) is adjusted by rounding up for expansion and rounding down for contraction. The final number of instances for expansion or contraction is also limited by the minimum and maximum number of instances in the scaling group. For example, if the minimum number of instances in the scaling group is 0 and the maximum is 10, and there are currently 9 instances in the scaling group. At this point, the target tracking policy is triggered, and it is calculated that 1.5 instances need to be expanded proportionally. After rounding up, it becomes 2 instances. However, due to the limitation of the maximum number of instances in the scaling group, only 1 instance is actually expanded.

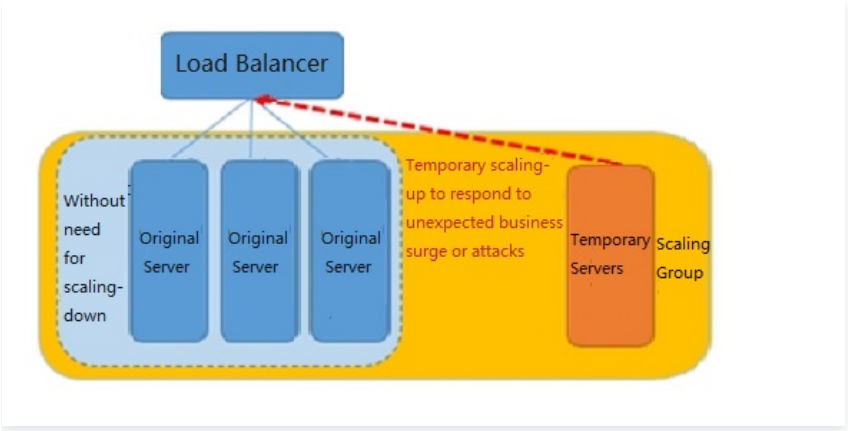
- **Instance Preheating Time**

  When creating a target tracking policy, you can specify the time required for instance preheating. New instances created by the expansion triggered by the target tracking policy will enter a preheating phase. During this specified preheating time, the instance will not affect the monitoring metrics of the scaling group. After a new instance joins the scaling group, it typically needs to go through processes such as business deployment, load balancing health checks, and data collection before it can report stable monitoring data. During this process, it is not suitable to trigger new scaling activities. To limit the frequency of scaling operations, any scaling activities generated by the target tracking policy will be cancelled if there are instances in the scaling group that are currently in the preheating phase.
- **Disable Contraction**

  When the disable contraction feature is enabled, the target tracking policy will only trigger expansion activities and will not trigger contraction activities.
- The conditions for triggering expansion with a target tracking policy are when a specified type of metric exceeds the threshold (target value) for three consecutive periods, each period being one minute. The conditions for triggering contraction are when a specified type of metric falls below the threshold (80% of the target value) for fifteen consecutive periods, each period being one minute.

Scenario Example

For instance, you have an e-commerce website application currently utilizing five instances. If you are conducting a marketing campaign and are concerned that the traffic will far exceed your estimates, you can set up your system to launch two additional instances when the load on the current instances rises to 70%, and then terminate the surplus instances when the load drops to 40%. You can configure two simple policies for the scaling group, one for expansion when the load exceeds 70%, and another for contraction when the load falls below 40%. As illustrated in the following diagram:

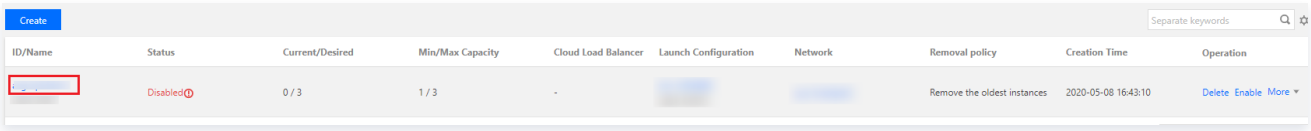


Alternatively, you may wish to maintain the load level of the entire scaling group around 60%, and carry out proportional expansion and contraction based on the actual load level. You can configure a target tracking policy for the scaling group, allowing the number of instances to dynamically change according to the actual load level, in order to achieve a target load level close to the desired value.

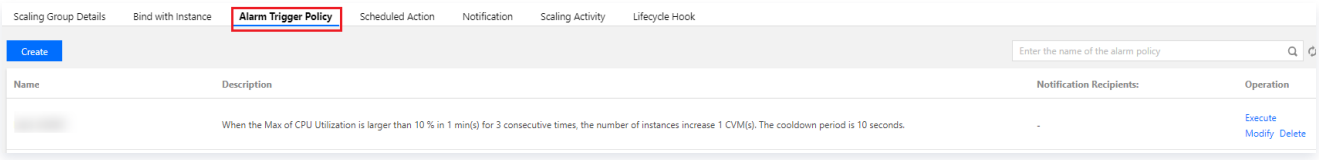
Instructions

1. Log into the Auto Scaling console and select **Scaling Groups** from the left navigation bar.

2. Select the scaling group you wish to modify, click on the scaling group ID to enter the basic information page of the scaling group, as shown in the figure below:



3. On the details page of the scaling group, select the **Alarm Trigger Policy** tab. On this page, manage the alarm trigger policies associated with the scaling group, as depicted in the figure below:



   - Click on **Create** to add a new alarm trigger policy.
   - Click on **Delete** to remove the selected alarm trigger policy.

Designate a specific server to be unaffected by the alarm scaling policy.

Before using auto scaling, your system may already have commonly used servers. You may not want these machines to be removed by the alarm scaling policy for the following reasons:

- **Multifunctional Machine**: A server within the cluster serves multiple purposes in addition to its primary role. For instance, during the initial stages of website development, a particular server might function both as a cache server and a file server. When the cache server cluster is incorporated into the scaling group, you would not want it to be removed by the alarm scaling policy.
- **Data Storage**: The server is stateful or contains data that other servers do not possess. For instance, incremental data generated by other servers in the cluster is uniformly stored on this server.
- **Update Image/Snapshot**: Regularly create images and snapshots using this specific server.

Configuration Method:

1. You can click on the scaling group where the server is located in the Scaling Group List to access the management page.
2. Select the **Associated Instances** tab on the management page and click on "Set Removal Protection" for the instance you wish to configure.

Instance Health Check

Last updated: 2024-01-17 17:57:27

Should you specify a **starting instance count** when creating a new scaling group, the group will generate a number of cloud server instances equivalent to the starting instance count upon the creation of the launch configuration and scaling group. Concurrently, the scaling group will ensure that the number of running instances exceeds the **minimum scaling count** and falls short of the **maximum scaling count**.

> ⚠ **Note:**
> - Minimum Scaling Count: This refers to the smallest permissible quantity of instances within a scaling group. Should the number of CVMs in the scaling group fall below the minimum scaling count, the Auto Scaling (AS) feature will augment the number of instances, aligning the current instance count of the scaling group with the minimum scaling count.
> - Starting Instance Count: This refers to the quantity of cloud servers present when the scaling group is initially created.
> - Maximum Scaling Count: This refers to the largest permissible quantity of instances within a scaling group. Should the number of CVMs in the scaling group exceed the maximum scaling count, the Auto Scaling (AS) feature will remove instances, aligning the current instance count of the scaling group with the maximum scaling count.

To maintain the normal operation of instances within the scaling group, Auto Scaling (AS) conducts regular checks on the operational status of these instances. If an instance is found to be performing poorly, it will be terminated, and a new cloud server instance will be initiated.

Instance health check

The scaling group conducts regular checks on the operational status of each instance to determine their robustness, with the criterion being whether the machine is unreachable for a continuous minute via ping. If an instance is unreachable for more than a minute via ping, Auto Scaling (AS) will mark the instance as performing poorly.

Replace Unhealthy Instances

Once an unhealthy instance has been marked as performing poorly, the scaling group will promptly initiate a new instance to replace it, excluding machines that have been set with "removal protection".

Expanding Capacity Manually

Last updated: 2024-01-17 17:57:35

Auto Scaling (Auto Scaling, AS) not only supports automatic expansion and contraction based on business load, but also allows for your manual intervention, achieving rapid manual scaling effects. You can achieve expansion effects through the following two methods:

- Add existing CVM instances to the scaling group
- By modifying the desired instance count of the scaling group, one-click expansion can be achieved

Incorporate existing CVM instances into the scaling group.

The scaling group provides you with a method to add existing instances to the current scaling group, enabling the ability to observe load and manage in conjunction with the other machines in the scaling group.

Preparations

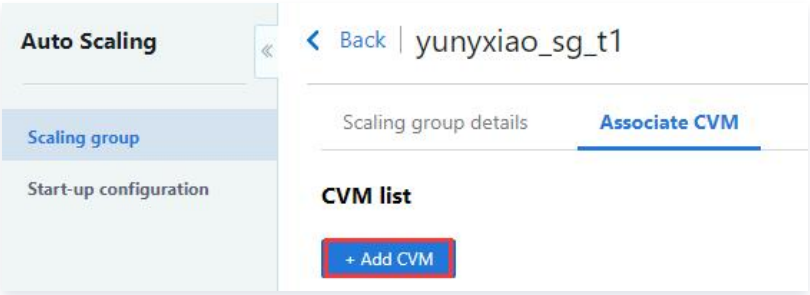
- The instance is in operation.
- The instance and the scaling group are located in the same region.
- The network attributes of the instance must align with the scaling group, meaning they either both belong to the basic network or to the same private network.

Notes

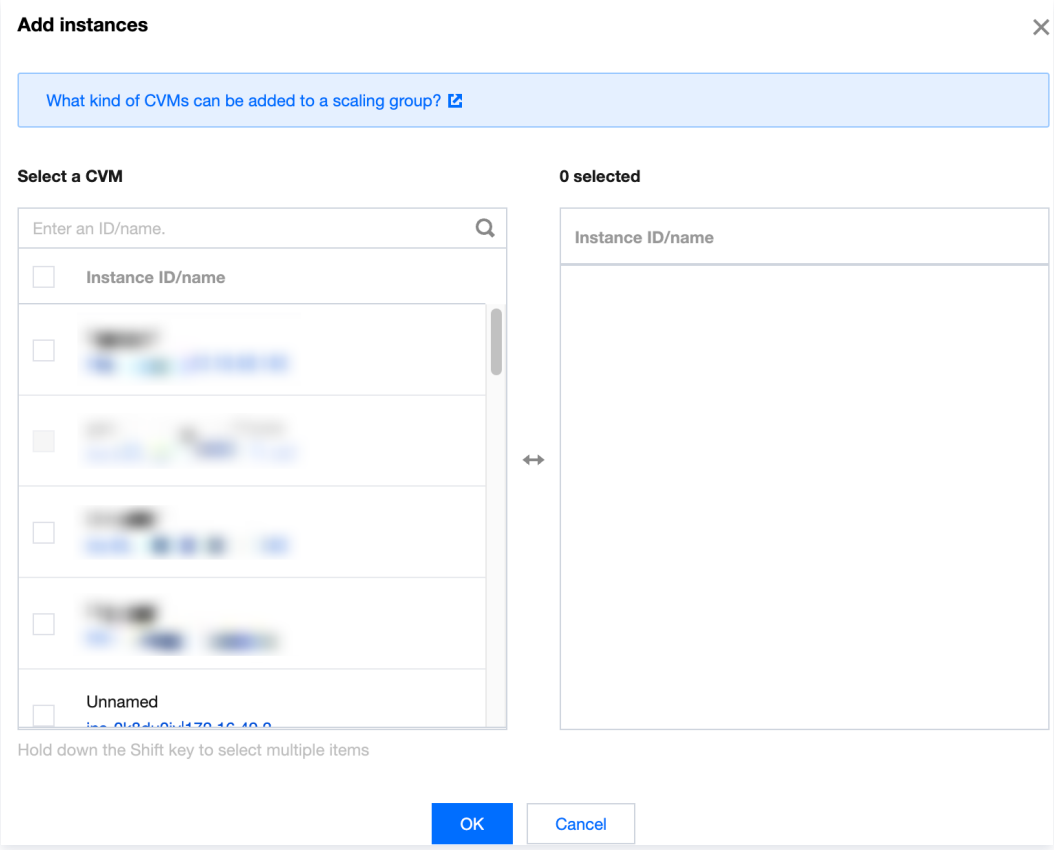
- AS will add the required capacity of the group to the number of instances to be added. For instance, if your current desired instance count is 5, after manually adding 3 instances, your desired instance count will become 5 + 3 = 8. If the number of instances to be added plus the required capacity exceeds the maximum instance count of the scaling group, the request will fail.
- If the scaling group is associated with one or more Cloud Load Balancers (CLB), any manually added instances will automatically register with all the CLBs in the scaling group.
- During the scaling down process, the scaling group will first remove the automatically created machines. Only when there are no automatically created machines left, will it opt to remove the manually added machines.
- When the scaling group removes a manually added instance, it merely disassociates the instance from the scaling group and the CLB. This action ensures that the instance is no longer managed by the scaling group, but it does not terminate your instance.

Manually Adding Instances Using the Console

1. Log in to the Scaling Group Console and click on the ID of the scaling group to which you want to add instances.
2. Navigate to the details page of the scaling group, select **Associated Instances** > **Add Instance**. As shown in the image below:



3. In the dialog box, select the corresponding instances and click **Confirm**. As depicted in the image below:



Modify the desired instance count to achieve one-click scaling.

Scaling Scenarios

If your requirements align with the following scenarios, you can execute one-click scaling via the console. Ensure to complete tasks such as CLB forwarding rules, machine configuration, and business deployment in advance. Even if your business needs to scale up later, you only need to modify the parameters of the scaling group with one click to quickly complete the scaling.

- The peaks and troughs of business are difficult to predict, yet there is a reluctance to leave scaling entirely to the system's discretion. If the business peaks and troughs are predictable, please refer to Managing Scheduled Tasks for more details.
- Your computational requirements are project-based, and the machines used each time are similar. This is applicable to scenarios such as social sentiment collection, gene sequencing, weather forecasting, etc.

Perform one-click scaling via the console.

Execute the following steps to set the CVM template as the launch configuration and configure the corresponding scaling group.

1. Create a custom image, for more details, please refer to Detailed Method for Creating Custom Images.

> **Note:**
> - Subsequent scaling instances will be deployed based on this image environment.
> - Recommended approach for creating a custom image: You may choose an existing CVM or create a new one, deploy your business on it, and set the business to start with the operating system, then export it as a custom image.

2. Create a launch configuration based on this custom image, for more details, please refer to Creating a Launch Configuration.
3. Create a Scaling Group.
   When creating, select the previously created launch configuration. The minimum scaling number, maximum scaling number, and initial instance number should be filled in according to the lower limit, upper limit, and current number of servers you need.
4. Upon completion of the above steps, when your business needs to scale (for instance, initiating a gene sequencing task or activating a request-type machine to collect data), you can modify the scaling group configuration, increase the minimum scaling number, maximum scaling number, and desired instance number. AS will swiftly accomplish the scaling.

Reducing Capacity

Last updated: 2024-01-18 11:18:31

For each scaling group, you have the ability to dictate when instances are added (i.e., scaled out) or removed (i.e., scaled in). You can manually adjust the size of the group by adding or removing instances, or employ scaling policies to allow for automatic execution of this process by the Auto Scaling service.

Note:
- When a scaling group automatically scales in, it is necessary to determine which instances should be terminated first, a decision guided by the removal policy.
- During scale-in operations, you can prevent the termination of specific instances by utilizing instance protection provided by the Auto Scaling service.
- For scaling groups configured with load balancing, instances are automatically unmounted from the associated load balancer when they are scaled in, removed, or deleted from the scaling group.
- For instances with a prepaid billing type, they will not be destroyed during scale-in, removal, or deletion processes, but merely removed from the group.

Removal policy

When a scaling group scales in, the machine to be removed is determined based on the removal policy. You can choose from the following two removal strategies:

- **Remove the oldest instances**: This policy removes the earliest automatically added machines. Once all automatically added machines are removed, it then proceeds to remove the earliest manually added machines.
- **Remove the newest instances**: This policy removes the most recently added automatic machines. Once all automatically added machines are removed, it then proceeds to remove the most recently manually added machines.

> **Note:**
> Regardless of whether the newest or oldest machines are being removed, the Auto Scaling will first remove the automatically created cloud servers, followed by the manually added cloud servers.

Setting and modifying removal policies in the console

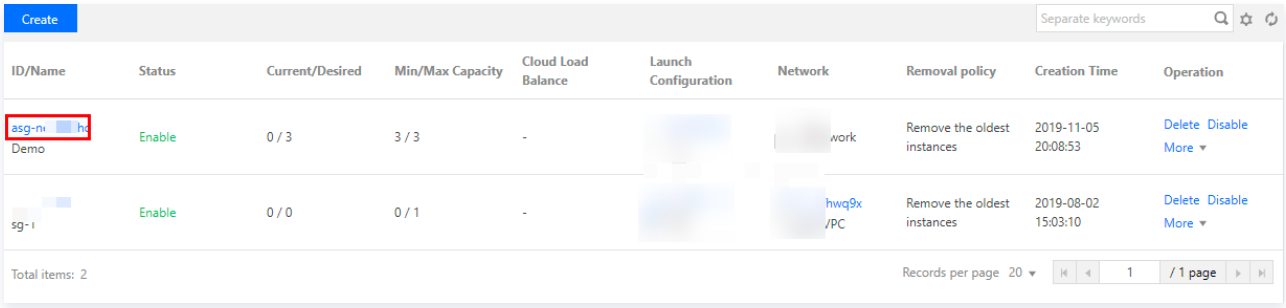
There are two methods for setting:

- When creating a scaling group, select the removal policy that best suits your needs.
- On the details page of the scaling group, click **Edit** to modify the scaling policy.

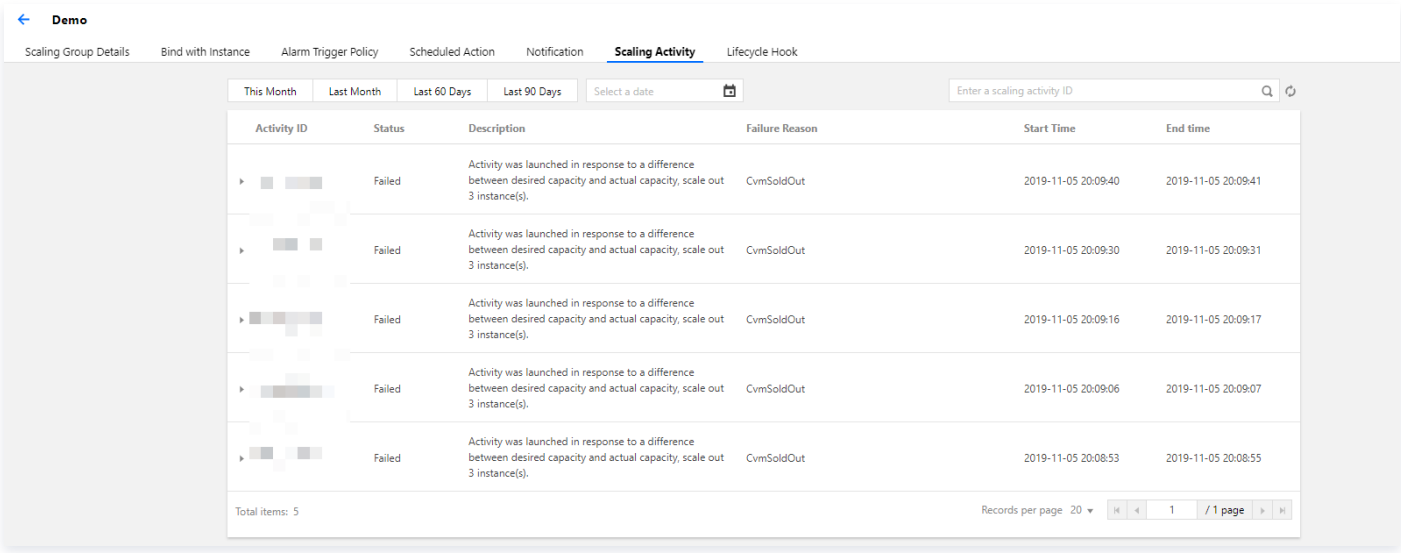
Viewing Scaling Activities

Last updated: 2024-01-17 17:57:49

1. Log into the Auto Scaling console and select **Scaling Groups** from the left navigation bar.
2. Select the desired scaling group and click on the scaling group ID to access the basic information page of the scaling group, as illustrated below:



3. On the details page of this scaling group, select the **Scaling Activities** tab to view the information on scaling activities that have been executed according to the scaling strategy for this group, as depicted below:



Suspending and Resuming Scaling

Last updated: 2024-01-17 17:58:00

Use Cases

Should you need to troubleshoot configuration or other issues related to your web application (such as shutdown password reset, business upgrade, etc.) and wish to make changes to the application without triggering the auto-scaling process, you have the option to pause the scaling group and resume it once completed.

Pausing the Scaling Group

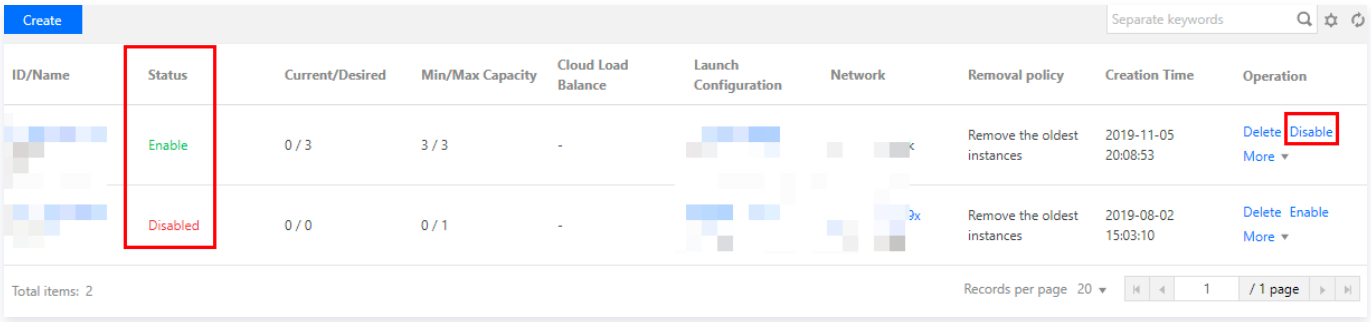
Supports and Limits

Once the scaling group is set to inactive, automatically triggered activities will not proceed.

- Automatically triggered activities include:
  - Alarm Scaling.
  - Scheduled Tasks.
  - Health Check.
  - Manual actions resulting in a mismatch of the expected number of instances.
- After disabling the scaling group:
  - If manually added instances exceed the maximum number of instances, addition will not be permitted.
  - Modifying the minimum or maximum number of instances in the scaling group will not trigger scaling activities, but the changes will take effect.
  - Manually removing instances is not subject to the minimum number of instances restriction.

Instructions

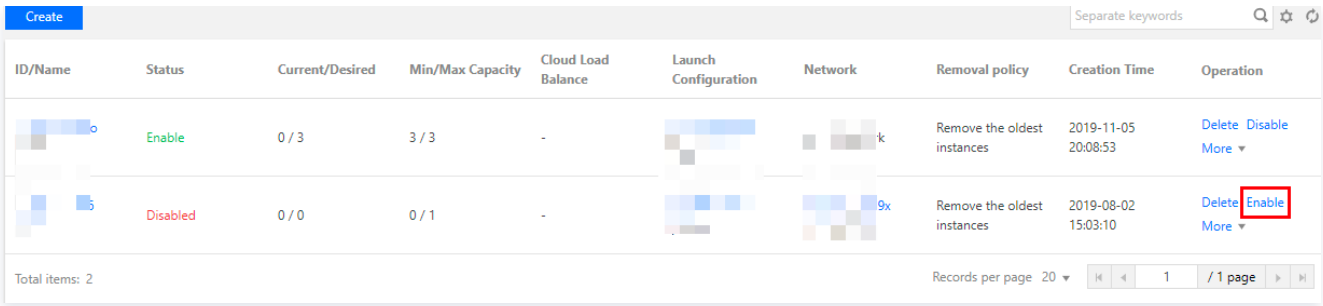
1. Log in to the Auto Scaling console and select **Scaling Groups** from the left navigation bar.
2. On the **Scaling Groups** page, select **Disable** on the right side of the row where the scaling group to be disabled is located, and confirm in the pop-up window.
   You can then see that the scaling group is in the **Disabled** state, as shown in the figure below:



Restore Scaling Group

If you have completed troubleshooting or operations during the pause of the scaling group activity, you can restore the automatic scaling settings for your business.

1. Log in to the Auto Scaling console and select **Scaling Groups** from the left navigation bar.
2. On the **Scaling Groups** page, simply select **Enable** on the right side of the row where the scaling group to be enabled is located, as shown in the figure below:



Scale-in Removal Protection

Last updated: 2024-01-17 17:58:09

Feature Overview

Within the scaling group, you can designate a specific sub-machine to be exempt from reduction during scaling down activities. When such activities occur, the Auto Scaling selects the sub-machines to be reduced from the remaining machines.
You can enable the **Instance Protection** setting for one or more scaling group instances, with the flexibility to modify the scaling group or instance protection settings at any given time.
In the event that all remaining instances within the scaling group are under reduction protection and a scaling down event occurs simultaneously, the Auto Scaling will decrease the required capacity without removing any instances.

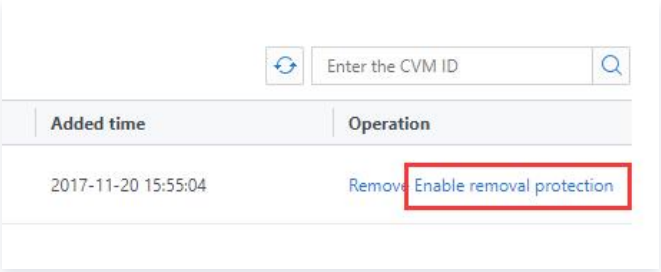
Scenarios

Typically, the machines within a scaling group are stateless, and any machine can be removed at any time. However, in practical application, there are scenarios where it is appropriate to set specific instances to be exempt from reduction:

- **Multipurpose Machine:** For cost considerations, individual machines may serve multiple purposes beyond their roles within the cluster. For instance, a machine might store data generated within the cluster, thus rendering it stateful.
- **Preventing Misoperations:** If there are concerns about policy setting errors impacting operations, "exemption from reduction" can be set for certain machines. This ensures that Auto Scaling will never reduce these machines, maintaining a smooth pathway for "Request-LB-Submachine".

Instructions

1. Log into the Auto Scaling console and select **Scaling Groups** from the left navigation bar.
2. On the **Scaling Groups** page, select the ID of the scaling group that needs to be configured, which will take you to the details page of that scaling group.
3. Select the **Associated Instances** tab, and click on Enable **Removal Protection** on the right side of the row where the instance that needs to be exempted from reduction is located, as shown in the figure below:



4. Click **Confirm** in the pop-up prompt to complete the setup.

Scaling Activity Cancelled

Last updated: 2024-01-17 17:58:22

The cancellation of a scaling activity refers to the scenario where a scheduled task is due or the conditions for alarm scaling are met, triggering the scaling activity. However, due to conflicts, the scaling activity is compelled to be cancelled.

Reasons for Conflict:

- There is an ongoing scaling activity.
- The scaling group is in a cooling-off period.

Will there be a retry after the cancellation of a scaling activity?

- **Alarm scaling** activities, if cancelled, will not be retried. However, if the conditions for alarm scaling persist, it will trigger the next alarm scaling activity.
- **Scheduled tasks** define the desired number of instances, maximum scaling number, and minimum scaling number, hence the scaling group will persistently retry to align the actual number of instances with the desired number.

> **Note:**
> The scaling group is suspended, and it will not attempt any scaling activities. Therefore, no cancelled scaling activities will be recorded in the "Scaling Activities" log.

Scaling Activity Failed

Last updated: 2024-01-17 17:58:32

Scaling activity cancellation is in accordance with expectations, whereas **scaling activity failure** is contrary to expectations.

How can one view the failed scaling activities?

You can view the details of the scaling activity. To be promptly informed about a scaling activity failure, you may configure a notification strategy.

Why would a scaling activity failure occur?

We have categorized the reasons for scaling activities. Please refer to Failure Reason Classification.

Cooldown Period

Last updated: 2024-01-17 17:58:42

What is the Cooling Period?

The Cooling Period in Auto Scaling (AS) is a configurable setting for the scaling group, ensuring that AS will not initiate or terminate any other instances prior to the effectiveness of the previous expansion activity. After dynamically scaling using a simple expansion strategy, AS will wait for the Cooling Period to complete before continuing with the expansion activity.
When manually scaling the group, it defaults to not waiting for the Cooling Period, but you can set a Cooling Period to override the default settings. Please note, if poor instance health is detected, AS will promptly replace the unhealthy instance, without waiting for the Cooling Period to complete.

Why is a Cooling Period necessary?

After a machine joins the scaling group, it requires a certain period to reduce the load. Without a Cooling Period, the system would continuously scale up before the load decreases. Once the newly added machine takes over the business, it finds the load too low, and then scales down.
Before instances are put into use, these instances utilize configuration scripts to install and configure software, thus it takes approximately two to three minutes for an instance to transition from startup to operational status. (Of course, the actual time depends on many factors, such as the size of the instance and whether there are startup scripts to be completed, etc.)
Example Scenario:
A surge in business traffic triggers the alarm policy. When this alarm is activated, AS will launch an instance to help handle the increased demand. However, there is a problem: this instance takes a few minutes to start, and after starting, it needs time to gradually receive requests from the CLB. During this period, the monitoring alarm may continue to trigger, causing AS to launch an additional instance each time the alarm occurs.
However, if you set a Cooling Period, AS will pause all scaling activities caused by simple expansion policies or manual expansion after launching an instance, until the specified amount of time has passed (the default value is 60 seconds). This way, the newly launched instance has time to start handling application traffic.
After the Cooling Period, all paused scaling operations will resume. If the alarm is triggered again, AS will launch another instance, and the Cooling Period will take effect again. However, if the added instance is sufficient to reduce the CPU usage to a normal level, the group will maintain its current size.

Setting the Cooling Period

The default Cooling Period is 60 seconds.
To modify, please follow the steps below:

1. Open the details page of the **Scaling Group**.
2. Click on **Alarm Trigger Policy**, select the alarm scaling policy you wish to set, choose **Modify**, and specify the duration of the Cooling Period at the bottom of the modification box (can be set between 0 – 999999 seconds).