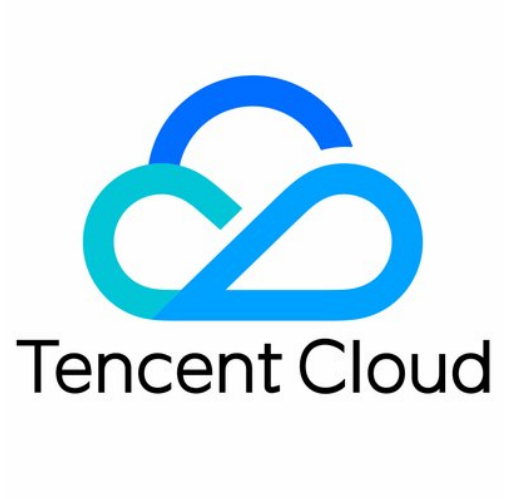
Auto Scaling

Expanding and Reducing Capacity

Product Documentation



Contents

Expanding and Reducing Capacity
Managing Scheduled Actions

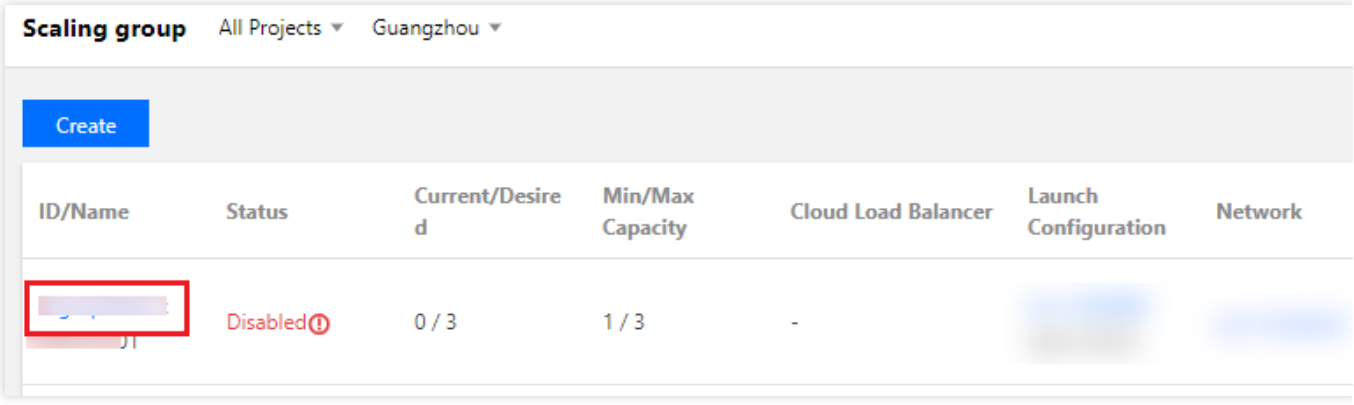
Last updated：2024-01-08 17:53:29

Scheduled Actions

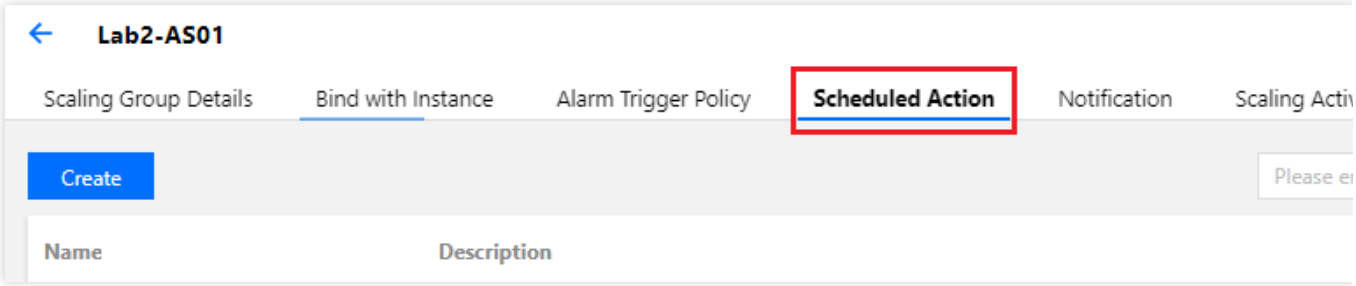
Scheduled actions refer to scheduling scaling activities. This allows you to scale the number of CVM instances in response to predictable load changes.
For example, every week the traffic to your Web application starts to increase on Wednesday, remains high on Thursday, and starts to decrease on Friday. You can schedule scaling activities based on the predictable traffic pattern of your Web application.
To create a scheduled scaling action, specify the start time of the scaling action, the new minimum (min capacity), maximum (max capacity), and required size (desired capacity) for the scaling action. At the specified time, AS will update the number of instances in the scaling group based on these values.
You can create scheduled actions for one-time scaling or for a recurring schedule.

Scheduled Action Management

1. Log in to the Auto Scaling Console and click **Scaling Groups** in the left sidebar.
2. Select the scaling group to be modified and click the **Scaling Group ID** to go to the basic information page, as shown in the figure below:



3. On the details page of the scaling group, select the **Scheduled Action** tab and configure the scheduled action associated with the scaling group on the page, as shown in the figure below:



Click **Create** to add a new scheduled action.

Select a scheduled action and click **Modify**. On the pop-up page, modify the action name, the execution time, and the activities to be executed and choose whether to execute the action periodically.

Click **Delete** to delete the scheduled action.

Note:

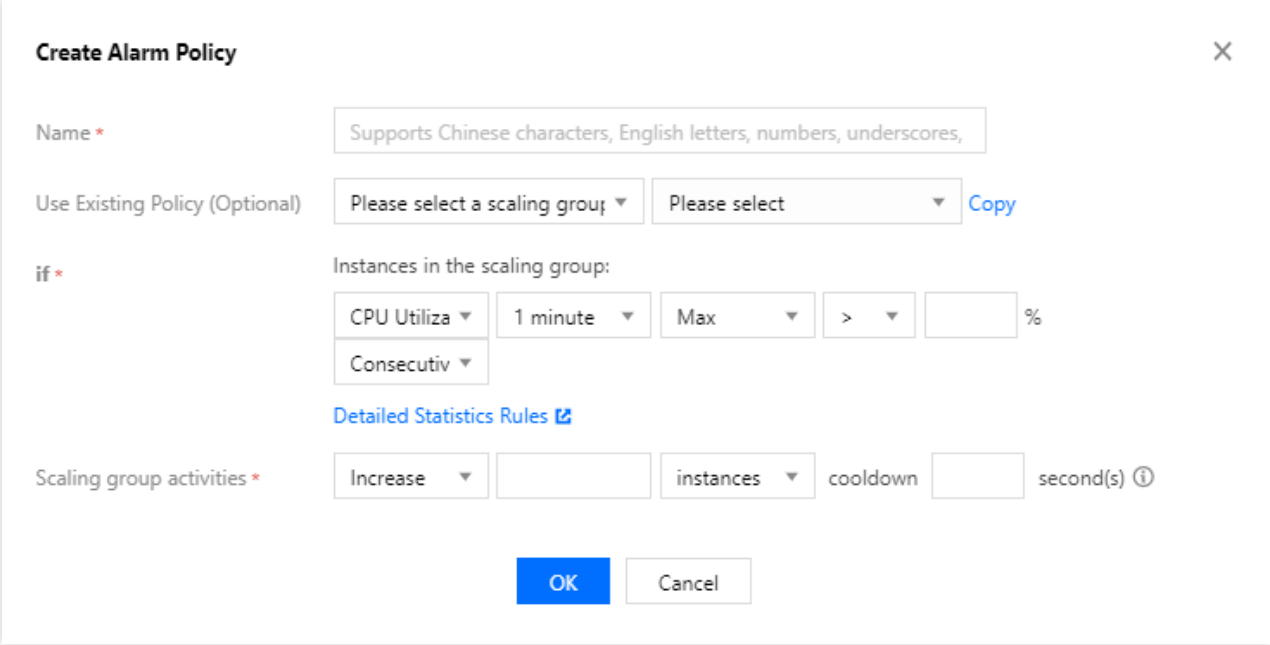
If you want to create a scheduled action on a recurring schedule, you can specify the start time. AS will perform the action at this time and then performs the action based on the recurring schedule. If you specify an end time, AS will not perform the action after the set time.

Managing an Alarm-triggered Policy

Last updated：2024-01-08 17:53:30

Overview

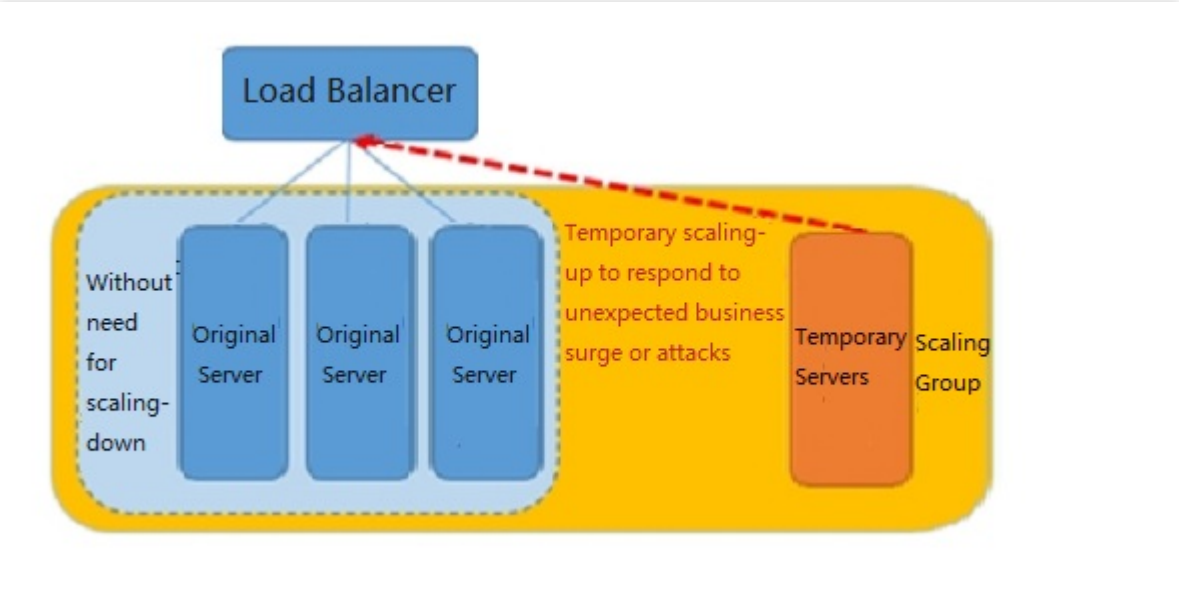
With Auto Scaling (AS), you can add and remove CVMs to or from a scaling group based on monitoring metrics. You only need to define an alarm-triggered policy, specifying the status of the monitoring metrics that trigger scaling and the related scaling activity.
You need to specify the conditions and actions when creating an alarm policy, as shown in the figure below:



Condition format: a metric + threshold + period + number of consecutive periods during which the threshold is reached. This indicates an alarm is triggered when the value of the metric breaches the threshold that you defined for the number of periods that you specified.
Execution actions: sending notification(s) + adding/removing the specified number of CVMs
We recommend you create two policies for each scaling group, one for scale-out and one for scale-in. Once the traffic to your web application reaches the threshold of the alarm policy, AS executes the associated policy to scale your group in (by terminating instances) or out (by launching instances).

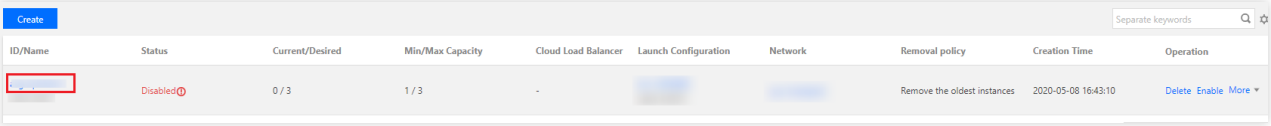
Scenarios

For example, assume you have an e-commerce web application that currently runs on five instances. You plan to carry out a promotional activity and are concerned that the access traffic might be much greater than you expect. In this case, you can configure a scaling group to add two new instances when the load on the current instances reaches 70%, and terminate the extra instances when the load decreases to 40%. This is shown in the figure below:

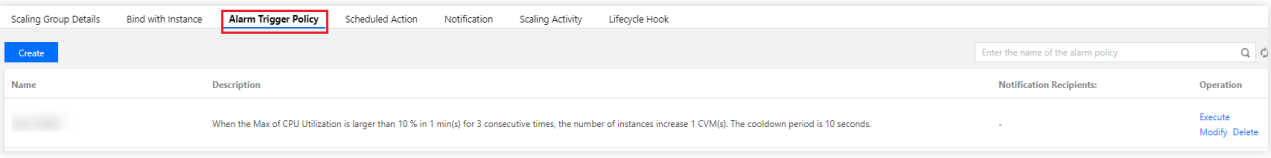


Directions

1. Log in to the Auto Scaling console and click **Scaling group** in the left sidebar.
2. Click the **ID/Name** of the scaling group you want to modify to enter its details page, as shown in the figure below:



3. Select the **Alarm Trigger Policy** tab and configure the alarm-triggered policy associated with the scaling group on this page, as shown in the figure below:



Click **Create** to add a new alarm-triggered policy.
Click **Delete** to delete the alarm-triggered policy.

Preventing Specified CVMs from Being Removed by the Scaling Policy

For the proper running of your existing business, if the CVMs in the cluster are used for the following purposes, you need to prevent them from being removed by the scale-in policy:
Multiple purposes: apart from the tasks specified by the cluster, a CVM in the cluster is also used for other purposes, for example the CVM is used as both a cache server and a file server.
Data storage: the CVM is stateful or stores data that other CVMs do not have. For example, the CVM stores the incremental data of other running CVMs in a cluster.
Image/Snapshot updates: the CVM is used to regularly update images and snapshots.
Configuration:
1. In the scaling group list, click the scaling group to which the CVM belongs to go to the management page.
2. Select the **Bind with Instance** tab and click **Enable Removal Protection** for the instance.

Instance Health Check

Last updated：2024-01-08 17:53:30

If you specify the **Initial Capacity** when creating a new scaling group, after the launch configuration and scaling group have been created, the scaling group will create the same number of CVM instances as the initial number of instances. Meanwhile, the scaling group will ensure that the number of running instances is larger than the **Min Capacity** and smaller than the **Max Capacity**.

Note：

Min Capacity: The minimum number of instances allowed in the auto scaling group. When the number of instances in the group is smaller than this value, scale-out is triggered and AS adds instances, making the number of instances equal to the mini capacity.

Initial Capacity: The initial number of CVMs when the scaling group is created.

Max Capacity: The maximum number of instances that is allowed in a scaling group. When the number of instances in the group is larger than this value, scale-in is triggered and AS removes instances, making the number of instances equal to the maximum scaling group size.

To ensure the normal operations of the instances in the scaling group, AS periodically performs health checks on the instances. If it is found that the running status of a CVM instance is unhealthy, AS will stop this CVM and launch a new CVM.

Instance health check

The scaling group periodically checks the running status of instances to confirm whether each instance is robust (whether it responses to the ping command within 1 minute). If the pinged instance is unreachable for more than one minute, AS marks this instance as unhealthy.

Replacing unhealthy instances

If an instance is marked as unhealthy, the scaling group will immediately replace it with a newly launched instance, unless it is under "Removal Protection".

Expanding Capacity Manually

Last updated：2024-01-08 17:53:29

Beside adding and removing instances automatically, AS also allows you to add and remove instances manually.

Adding existing CVM instances to a scaling group

Modifying the desired capacity of a scaling group to implement one-click scale out

Adding existing CVM instances to a scaling group

You can add existing CVM instances to a scaling group manually.

Prerequisites

The instance is in the running status.

The instance is in the same region as the scaling group.

The instance and the scaling group must be in the same basic network or VPC.

Notes

AS will add the number of instances to be added to the desired capacity of the group.

For example, if the current desired capacity of your scaling group is 5, and you add 3 instances manually, the desired capacity of your scaling group will become 5 + 3 = 8. If the sum of the number of instances to be added and the desired capacity exceeds the maximum capacity of the group, the request will fail.

If the scaling group is associated with one or more load balancers, manually added instances will be automatically associated with these load balancers.

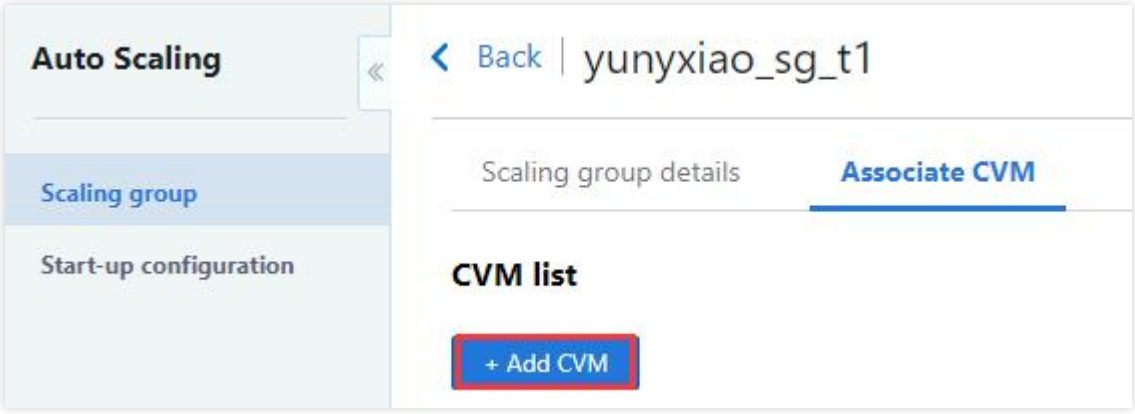
The scaling group will remove automatically created CVMs first during scale-in. If there are no more automatically-created CVMs, then the manually added CVMs will be removed.

For manually added instances, when they are removed from the scaling group, they will be unbound with the CLB and will not be terminated.

Manually adding instances in console

1. Log in to the Scaling Group Console, and click the ID of the scaling group to which you want to add an instance.

2. Go to the scaling group details page. Select **Associate Instances** -> **Add an Instance**. This is shown in the following figure:



3. In the dialog box, check the corresponding instance, and click **OK**. This is shown in the following figure:

Modifying the desired capacity to scale out quickly

Scenarios

For the following scenario, we recommended to use launch configuration as instructed in Scale-out with Launch Configuration. You can complete CLB forwarding rules, CVM configurations and business deployment in advance, and then you can modify the parameters of the scaling group to rapidly complete scale out.
It's hard to predict your business traffic, but you do not want to leave the scaling decisions entirely to the system. If the business fluctuations are predictable, see Managing scheduled actions.
Your computing needs are based on projects, and the CVMs to be used every time are similar. For example, this may involve collection of social conditions and public opinions, sequencing of genes, or prediction of weather.

Scale-out with Launch Configuration

Execute the following steps to configure a CVM template as the launch configuration, and to configure the corresponding scaling group.
1. Create a custom image. For more information, see Creating a custom image.
Note:
Instances added later for scale-out will be configured using this image.
Suggested steps to create a custom image: You can deploy your services on an existing CVM or a new CVM, set the services to be launched on the boot of the operating system, and create a custom image.
2. For information about creating a launch configuration based on this custom image, see Creating a launch configuration.
3. Create a scaling group.
During the creation process, select the created launch configuration. Enter the minimum capacity, maximum capacity, and initial capacity of the cluster as required by your business.

4. After you complete the preceding steps, when the business needs to scale out, you can modify the scaling group configurations to raise the minimum capacity, the maximum capacity, and the desired capacity, and the AS will rapidly perform scaling.

Scale-in Process

Last updated：2024-01-08 17:53:30

For each scaling group, you can set the time for adding (scale-out) or removing (scale-in) instances. You can scale the scaling group manually by adding or removing instances, or you can enable AS to execute this process automatically by using a scaling policy.

Note:

When a scaling group is scaled in automatically, it needs to know which instances should be terminated first, and the selection is based on the removal policy.

During scale-in, you can prevent specified instances from being terminated by AS by using instance protection.

For a scaling group configured with a CLB instance, when the instances are automatically or manually removed or deleted, they will be automatically disassociated from the CLB instance.

Removal policy

AS will determine which CVM should be removed based on the removal policy during scale-in. You can choose from the following two removal policies:

Remove the oldest instances: Remove the earliest automatically added instances from the scaling group. Instances added automatically are removed first, followed by the earliest manually added instances.

Remove the latest instances: Remove the latest automatically added CVM instances from the scaling group. Instances added automatically are removed first, followed by the latest manually added instances.

Note:

Under both of the removal policies, AS will remove automatically created CVM instances before those manually added.

Setting and modifying removal policy in the console

There are two ways to set this up:

Select the removal policy you want when creating the scaling group.

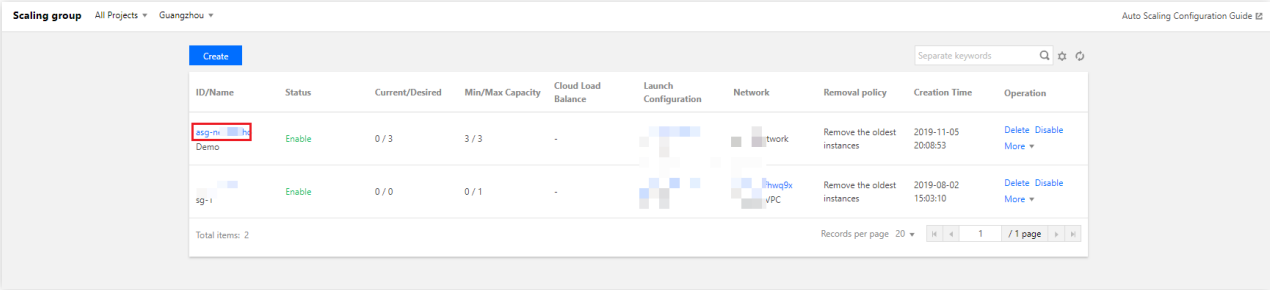
On the **Scaling Group Details** page, click **Edit** to modify the scaling policy.

Viewing Scaling Activities

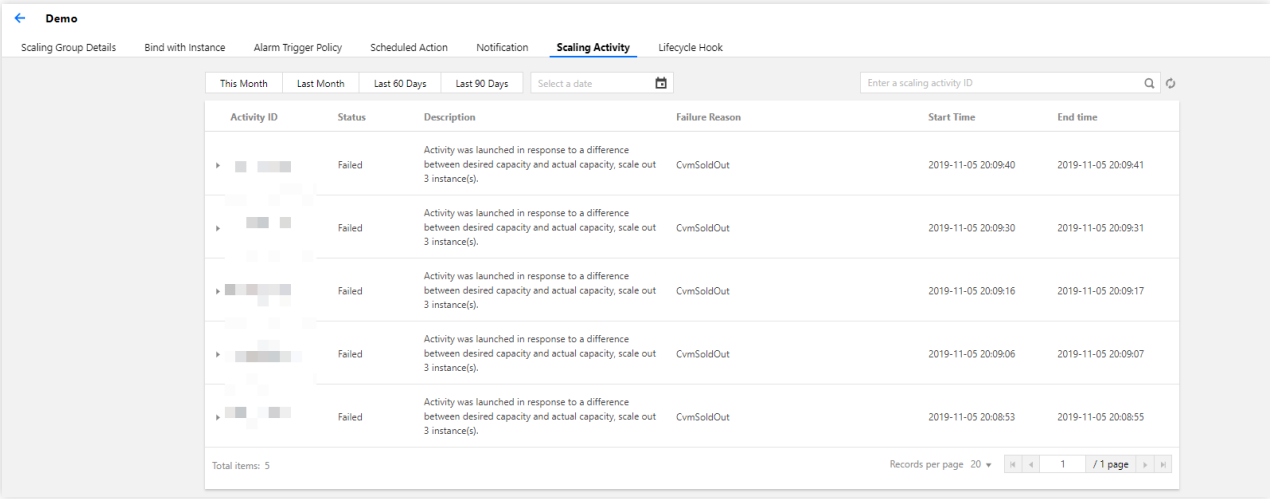
Last updated：2024-01-08 17:53:30

Viewing scaling activities

1. Log in to the Auto Scaling Console, and click Scaling group in the left sidebar.
2. Locate the scaling group you want to view details, and click its ID/Name to enter the basic information page, as shown below:



3. Click the **Scaling Activity** tab, and you can view the information of the scaling activities that have been performed by the scaling group based on the scaling policy, as shown below:



Suspending and Resuming Scaling

Last updated：2024-01-08 17:53:30

Overview

If you need to troubleshoot any problems related to configurations or Web applications (for example, shutdown to reset a password and upgrade services), and you want to modify the applications without triggering the auto scaling process, you can disable the scaling group and recover it after troubleshooting.

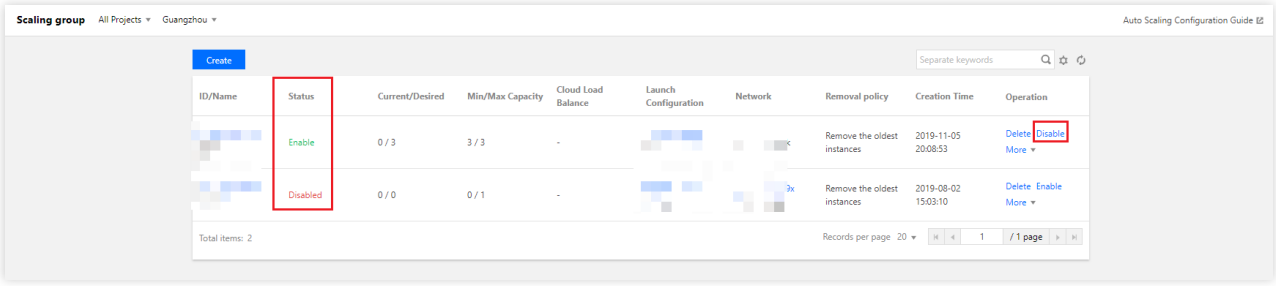
Suspending a Scaling Group

Notes

After a scaling group is disabled, the automatically triggered activities will not proceed.
Automatically triggered activities include:
Alarm-based scaling
Scheduled actions
Health checks
Matching the current capacity due to manual operations with the desired capacity
Scaling group restrictions include:
If manual addition causes the current capacity to be more than the max capacity, the addition will not be permitted.
Modifying the min or max capacity of the scaling group does not trigger a scaling action, but the modification takes effect.
Manual removal of instances is not limited by the min capacity.

Directions

1. Log in to the Auto Scaling console and select **Scaling group** on the left sidebar.
2. On the **Scaling group** page, locate the scaling group to be disabled, click **Disable** under the **Operation** column, and click **OK** in the pop-up window.
Then, you can see that the scaling group is in “Disabled” status, as shown in the following figure:

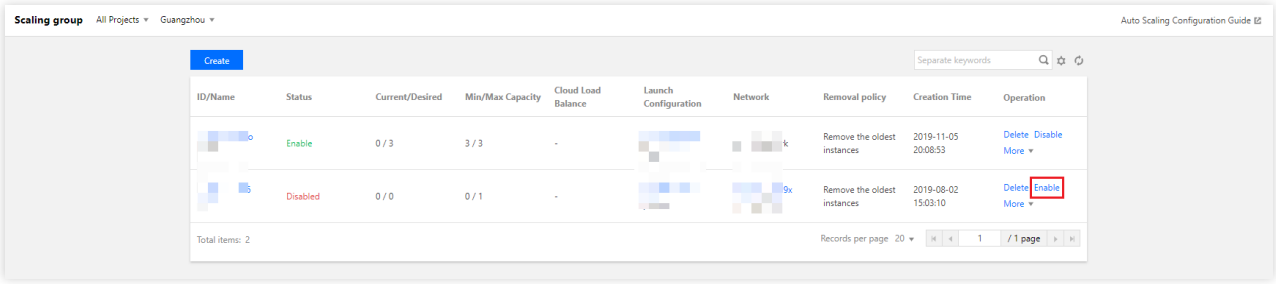


Recovering a Scaling Group

If you have finished troubleshooting, you can recover the auto scaling configuration.

1. Log in to the Auto Scaling console and select **Scaling group** on the left sidebar.

2. On the **Scaling group** page, locate the scaling group to be enabled, and click **Enable** under the **Operation** column, as shown in the following figure:



Scale-in Removal Protection

Last updated：2024-01-08 17:53:30

Introduction

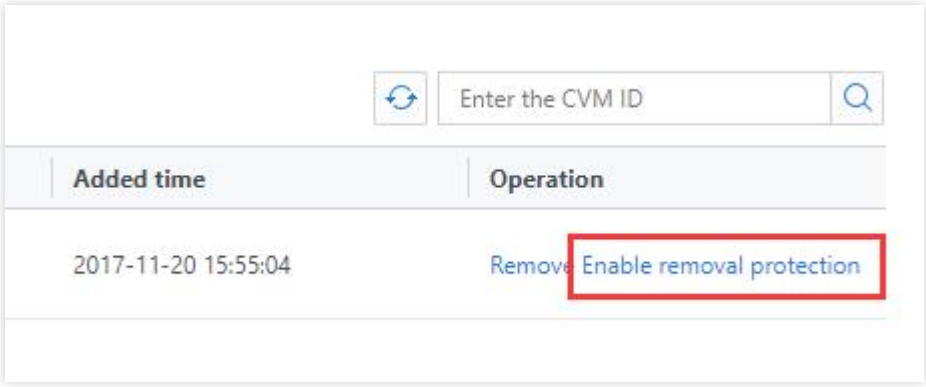
Removal protection allows you to specify CVMs in the scaling group that will not be removed during scale-in. When scale-in occurs, AS will remove other CVMs.
You can enable **Removal Protection** for one or more instances in a scaling group. You can modify the scaling group or the instance protection configuration at any time.
If a scale-in activity occurs when all the remaining instances in the scaling group are marked with “Removal Protection”, AS will decrease the capacity instead of removing any instances.

Application Scenarios

Usually, the CVMs in a scaling group are stateless and can be removed at any time. However, in the following circumstances, you may need to prevent certain instances from being removed during scale-in:
One CVM for multiple uses: if the CVM in the cluster is also used for other purposes, for example, if it is also used for storing the data generated in the cluster, then this CVM is actually stateful.
Avoid misoperation: if you worry that your business will be affected due to a policy settings error, you can enable **Removal Protection** for some CVMs. In this way, AS will never remove these CVMs in scale-in, and the tunnel "Request-LB-CVM" will remain unblocked.

Directions

1. Log in to the Auto Scaling console and select **Scaling Groups** in the left sidebar.
2. On the **Scaling Groups** page, select the target scaling group ID to go to the details page of the scaling group.
3. Select the **Associated Instances** tab and click **Enable Removal Protection** on the right side of the row of the target instance, as shown in the following figure:



4. Click **OK** in the pop-up box.

Scaling Activity Cancelled

Last updated：2020-08-17 11:12:41

A scaling activity will be canceled under such conditions that a conflict occurs when the scaling activity is triggered as the scheduled task starts or the requirement of the alarm scaling policy is met.

Causes of Conflict:

- Other scaling activities are in progress.
- The scaling group is under cooldown period.

Will the alarm scaling activity be rebooted if canceled?

- No, the **alarm scaling** activity cannot be rebooted if canceled. If the requirements for alarm scaling are met, another alarm scaling activity will be triggered.
- **Scheduled task** defines the expected, maximum and minimum group size, so the scaling group will keep trying until the actual number of instances equals the expected number of instances.

Note: A suspended scaling group does not try any scaling activity. That's why such activity is not recorded as canceled scaling activity in the "Scaling Activity".

Scaling Activity Failed

Last updated：2024-01-08 17:53:30

Unlike a **canceled scaling activity**, a **failed scaling activity** is not acceptable.

How to Check the Failed Scaling Activities?

You can check them at Viewing Scaling Activities.

You can configure notification policy to be informed of the failure of scaling activities at the earliest possible time.

Why Does a Scaling Activity Fail?

We have categorized the causes for failure of scaling activities. For details, refer to Failure Causes >>

Cooldown Period

Last updated：2024-01-08 17:53:30

Cooldown Period

The AS cooldown period is a configurable setting for your scaling group that helps to ensure that AS doesn't launch or terminate other instances before the previous scaling activity takes effect. After the scaling group dynamically scales using a simple scaling policy, AS waits for the cooldown period to complete before resuming scaling activities.
When you manually scale your scaling group, the default is not to wait for the cooldown period, but you can overwrite the default by setting a new cooldown period. Note that if an instance becomes unhealthy, AS does not wait for the cooldown period to complete before replacing the unhealthy instance.

Importance of Cooldown Period

After an instance is added to the scaling group, it will take some time to decrease the load. If there is no cooldown period, the system will keep scaling in before the load decreases. After the newly added instance takes over the service, it has to scale out due to low load.
These instances use a configuration script to install and configure software before the instance is put into service. As a result, it takes around two or three minutes for the instances to be put into service after they are enabled. (The actual time, of course, depends on several factors, such as the size of the instance and whether there are startup scripts to complete.)
Scenario Example:
A spike in traffic occurs, which triggers the alarm policy. When it does, AS enables an instance to help with the increase in demand. However, there's a problem: the instance takes a couple of minutes to enable, and it will also take some time for the enabled instance to receive requests from CLB. During that time, the monitor alarm could continue to be triggered, causing AS to enable another instance each time the alarm is triggered.
However, with a cooldown period in place, AS enables an instance and then suspends scaling activities due to simple scaling policies or manual scaling until the specified time elapses (the default is 60 seconds). This gives newly-enabled instances time to start handling application traffic.
After the cooldown period expires, any suspended scaling actions resume. If the alarm is triggered again, AS enables another instance, and the cooldown period takes effect again. If, however, the additional instance was enough to bring the CPU utilization back down, then the group remains at its current size.

How to Configure the Cooldown Period

Cooldown period is 60 seconds by default.
Use the following steps if you need to modify the cooldown period:
Open the details page of the scaling group;

Click **Alarm Trigger Policy**, select an appropriate alarm scaling policy, and then select **Modify** to specify the cooldown duration below the modification box (value range: 0-999,999 seconds)