Auto Scaling

Expanding and Reducing Capacity

Product Introduction

Copyright Notice

Trademark Notice

Tencent Cloud

Service Statement

This document is intended to provide users with general information about Tencent Cloud's products and services only and does not form part of Tencent Cloud's terms and conditions. Tencent Cloud's products or services are subject to change. Specific products and services and the standards applicable to them are exclusively provided for in Tencent Cloud's applicable terms and conditions.

Contents

Expanding and Reducing Capacity Managing Timed Tasks

Last updated：2018-05-29 14:52:01

Scheduled Task

Scheduled task means to scale based on a schedule. It 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 the scaling activities based on the predictable traffic pattern of your Web application.

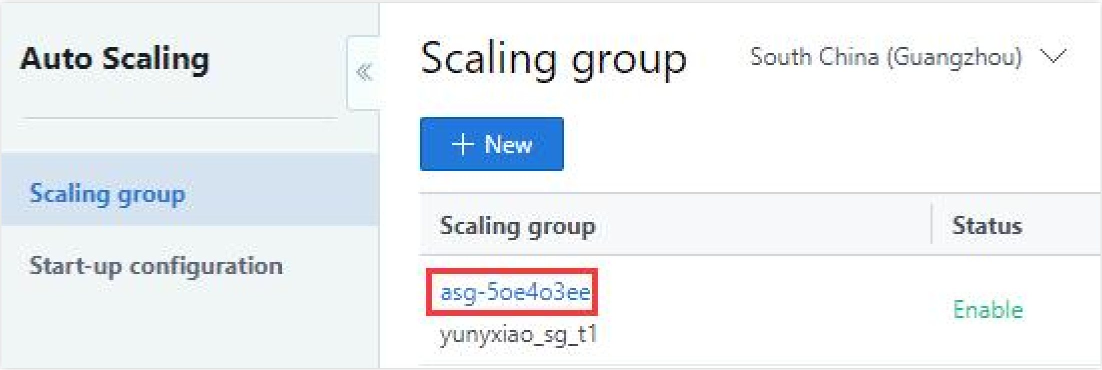
To create a scheduled scaling action, you specify the start time when the scaling action is expected to take effect, as well as the new minimum (minimum number of instances), maximum (maximum number of instances) and required size (expected number of instances) for the scaling action. AS will update the number of instances in the scaling group based on these values at the specified time.

You can create scheduled actions for scaling one time only or for scaling on a recurring schedule.

Scheduled Task Management

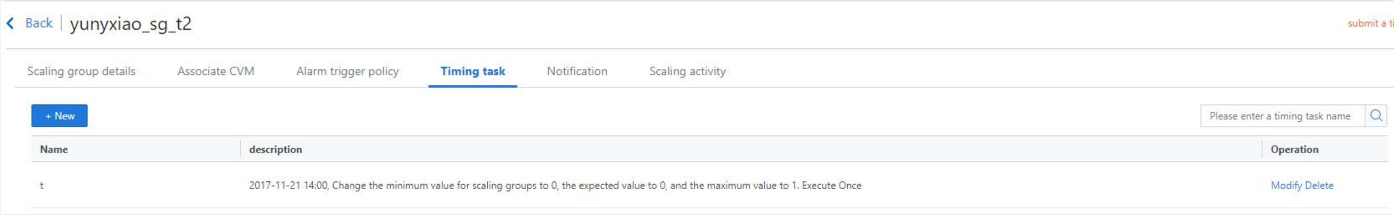
1. Open the Console, and select **Scaling Group** in the navigation bar.

2. Select the scaling group to be modified, and click the scaling group ID to enter the basic information page.



Scaling group	Status
asg-5oe4o3ee yunyxiao_sg_t1	Enable

3. Select **Timing Task** in the top navigation bar, and manage the scheduled task associated with the scaling group on the page:



Name	description	Operation
t	2017-11-21 14:00, Change the minimum value for scaling groups to 0, the expected value to 0, and the maximum value to 1. Execute Once	Modify Delete

- Click **New** to add a scheduled task;
- Select a scheduled task and click **Modify**. On the pop-up page, you can modify the task name, the execution time and the activities to be executed, and choose whether to execute periodically;
- Click **Delete** to delete the scheduled task.

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

Managing Alarm Triggering Policies

Last updated：2018-05-29 14:54:49

Introduction

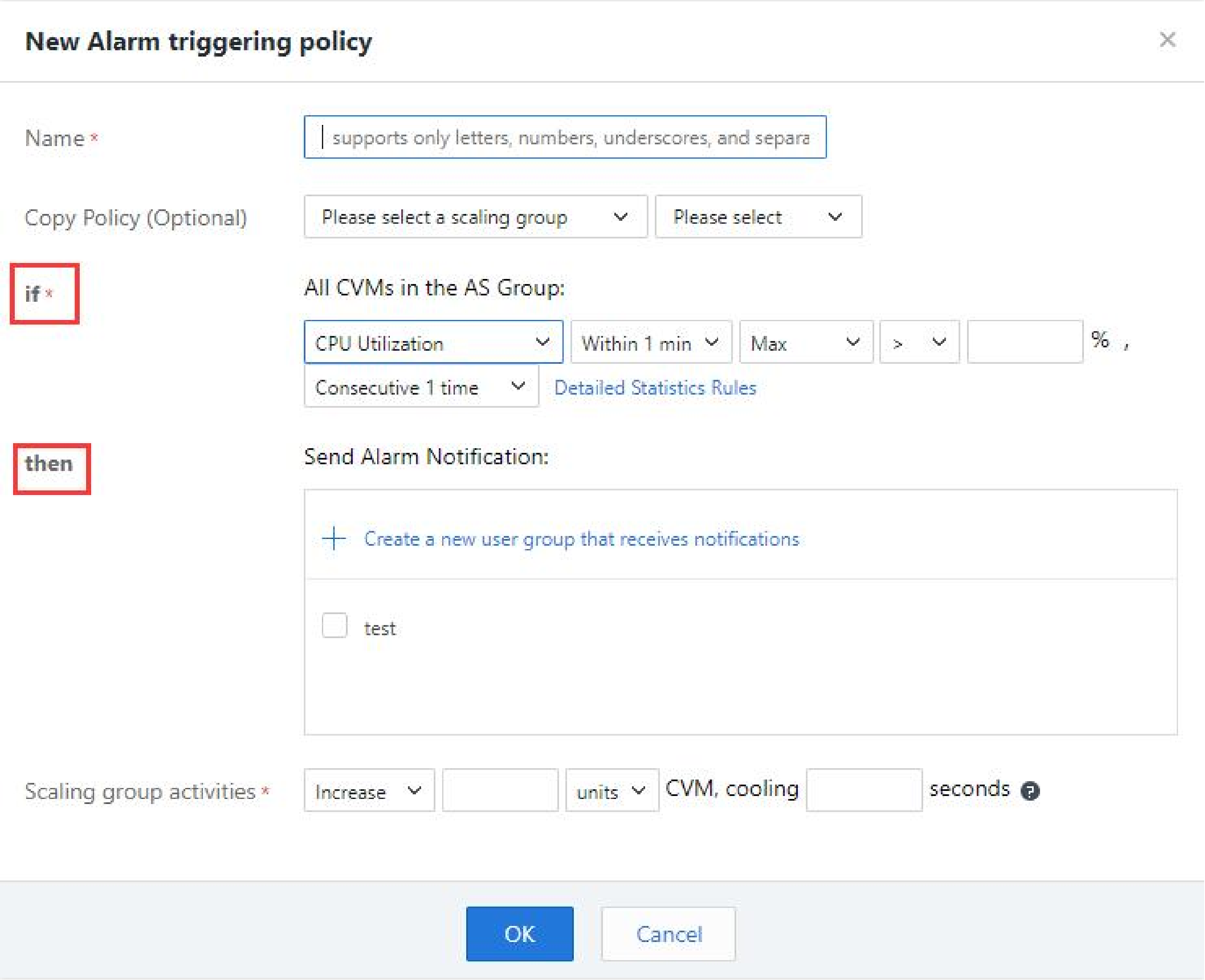
AS supports dynamic scaling of the number of instances in the scaling group based on the monitoring metrics, provided that you define the alarm trigger policy, including the status of the monitoring metrics that trigger the scaling, and how you want to scale in response to changing demand.

You need to specify the conditions and actions when creating an alarm policy:

- Condition format: A metric + threshold + period + number of periods during which the threshold is reached. That is, the value of the metric breaches the threshold that you defined, for the number of periods that you specified.
- Actions: Sending notification + increasing/decreasing a specified number of CVMs.

It is recommended that you create two policies for each scaling change: one policy to scale out and another policy to scale in. Once your business volume breaches the threshold of the alarm policy, AS executes the associated policy to scale your group in (by terminating instances) or out (by launching instances).

As shown in the figure below:



New Alarm triggering policy

Name * supports only letters, numbers, underscores, and separa

Copy Policy (Optional) Please select a scaling group | Please select

if * All CVMs in the AS Group:

CPU Utilization | Within 1 min | Max | > | % ,

Consecutive 1 time | Detailed Statistics Rules

then Send Alarm Notification:

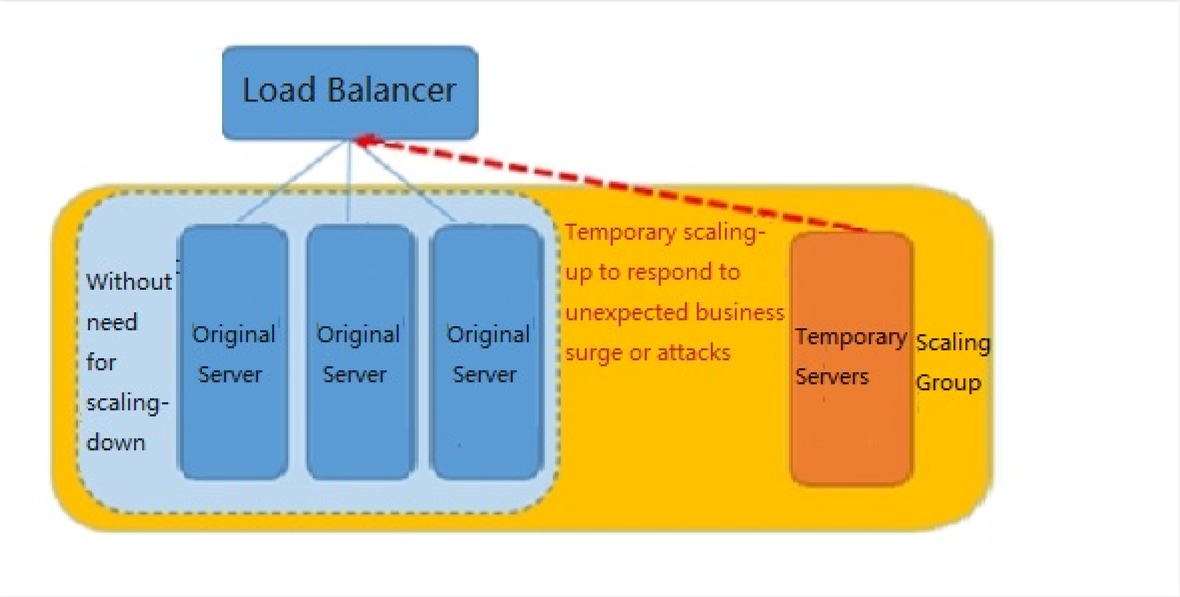
+ Create a new user group that receives notifications

□ test

Scaling group activities * Increase | units CVM, cooling | seconds

Scenario Example

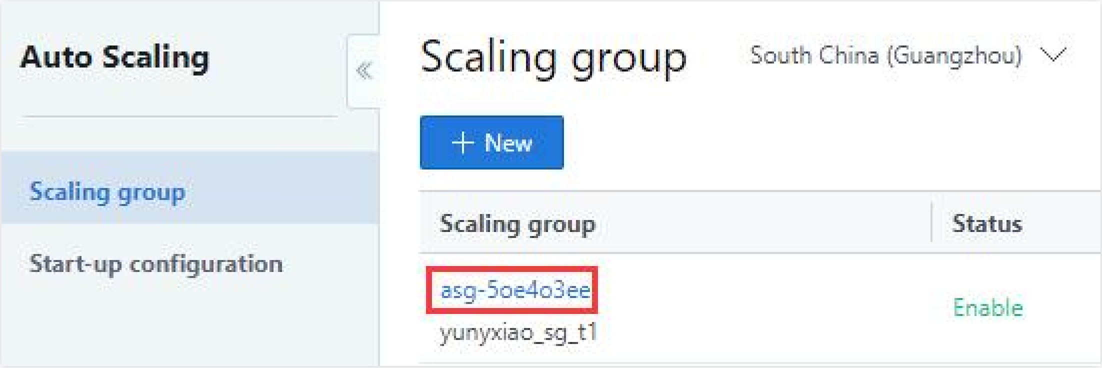
For example, you have an e-commerce web application that currently runs on five instances. You carry out an operating activity, and worry about that the number of visits is much greater than you've expected. You can launch two new instances when the load on the current instance increases to 70%, and terminate these instances when the load decreases to 40%. You can configure the scaling group to scale automatically based on these conditions.


Load Balancer
Without need for scaling-down
Original Server
Original Server
Original Server
Temporary scaling-up to respond to unexpected business surge or attacks
Temporary Servers
Scaling Group

Steps

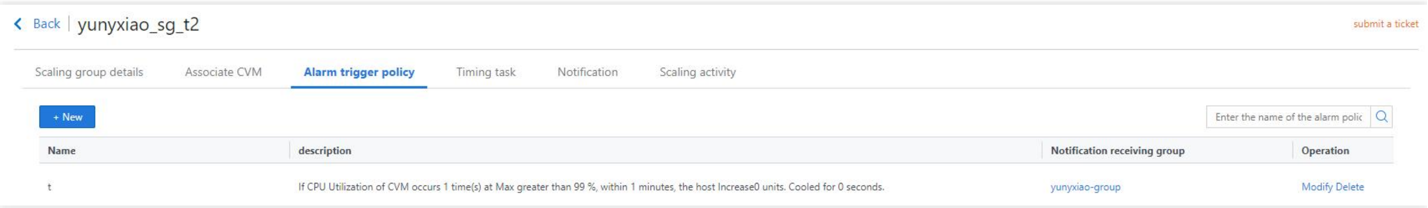
1. Open the Console, and select **Scaling Group** in the navigation bar.

2. Select the scaling group to be modified, and click the scaling group ID to enter the basic information page.



Scaling group	Status
asg-5oe4o3ee yunyxiao_sg_t1	Enable

3. Select **Alarm Trigger Policy** in the top navigation bar, and manage the alarm trigger policy associated with the scaling group on this page.
   - Click **New** to add a new alarm trigger policy;
   - Click **Delete** to delete the alarm trigger policy.



Name	description	Notification receiving group	Operation
t	If CPU Utilization of CVM occurs 1 time(s) at Max greater than 99 %, within 1 minutes, the host Increase0 units. Cooled for 0 seconds.	yunyxiao-group	Modify Delete

Exclude a Server from the Alarm Scaling Policy

Before using auto scaling, your system may have a server that is regularly used. Given the following considerations, you may not want the server to be removed by the alarm scaling policy:

- **One server for multiple uses**: Apart from the tasks specified by the cluster, a server in the cluster is also used for other purposes. For example, in the early stage of website construction, one of your servers is used as both a cache server and a file server. You don't want the server to be removed by the alarm scaling policy when placing the cluster of cache servers into a scaling group.

- **Data storage**: The server is in service or stores data that other servers do not have. For example, the server stores the incremental data of other running servers in a cluster.

- **Image/Snapshot updates**: The server is used to update the image and snapshot regularly.

How to configure:

Step 1: In the Scaling Group List, click the scaling group in which the server resides to enter the management page;

Step 2: In the **CVM List** at the bottom of the management page, click **Set Removal Protection** for the appropriate CVM.

Instance Health Check

Last updated：2017-04-14 15:36:29

If you specify the "initial number of instances" when creating a new scaling group, after the scaling configuration and scaling group are created, the scaling group will create the CVM instances whose number is equal to the initial number of instances. Meanwhile, the scaling group will ensure the instances whose number is larger than "minimum group size" and smaller than "maximum group size" are running.

> Note:
>
> - Minimum group size: Minimum number of instances that can be in a scaling group. If the number of CVMs in the scaling group is smaller than the minimum group size, AS will add the instances till the number of current instances in the scaling group reaches the minimum group size.
> - Initial number of instances: Initial number of CVMs when the scaling group is created.
> - Maximum group size: Maximum number of instances that can be in a scaling group. If the number of CVMs in the scaling group is greater than the maximum group size, AS will remove some instances till the number of current instances in the scaling group is limited to the maximum group size.

AS periodically performs health checks on the instances in your scaling group to ensure normal operation. If an instance is found unhealthy, AS will terminate and replace it with a new CVM.

- **Instance Health check**

AS periodically performs health checks on the instances in your scaling group to determine whether each instance is healthy by checking whether it is able to respond to a ping for one minute. If not, AS will mark it as unhealthy.

- **Replacing unhealthy instances**

If an instance is marked as unhealthy, the scaling group will immediately replace it with a new instance, unless it is under "removal protection".

Expanding Capacity Manually

Last updated：2018-05-29 15:34:57

AS not only supports the capacity scaling based on business load, but also allows you to intervene manually, so as to achieve manual capacity scaling in a fast way. Manual scale-up can be achieved using the following two methods:

- Add existing CVM instances to the scaling group
- Modify the expected number of instances in the scaling group to enable one-click scale-up

Adding Existing CVM Instances to the Scaling Group

The scaling group offers an option: you can add one or more CVM instances to the existing scaling group, and perform load observation and management operations on such instance(s) along with other instances in this scaling group.

- Conditions for Adding Instances

- The instance is in the running status
- The instance is in the same region as the scaling group
- The network attribute of the instance must be exactly the same as that of the scaling group, which means that they both belong to the same basic network or VPC.

- Operations after Adding Instances Manually

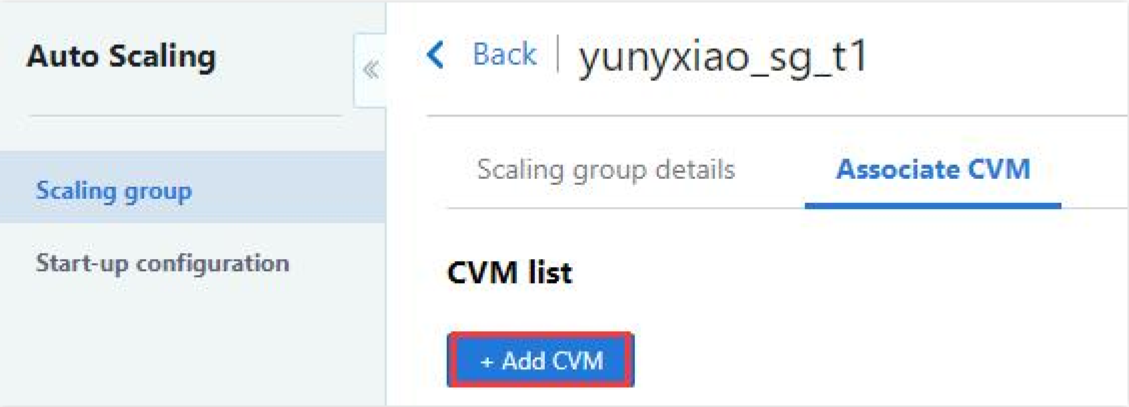
- AS will add the required capacity of the group to the number of instances to be added. For example, if the currently expected number of instance in your scaling group is 5, after 3 instances are added manually, the expected number of instance in your scaling group will be 5 + 3 = 8. (If the sum of the number of instances to be added and required capacity exceeds the maximum capacity of the group, the request will fail)

- If the scaling group is associated with one or more CLBs, the manually-added instances will be automatically registered into all the CLBs of the scaling group.

- The scaling group will remove the automatically created machines first during scale-down. If there is no automatically-created machine, then the manually added machines will be removed. For those manually added instances that are removed by the scaling group, the instances are not destroyed, but only removed from the scaling group and CLB and no longer under the control of the scaling group.

- Example of Adding an Instance in Console

Click the ID of the scaling group to be managed, or click **Management** beside it, to enter the details page of scaling group. In the lower half of the page, click "Add CVM" in CVM list, and check the corresponding CVM in the dialog box, then click "OK".



Modifying the Expected Number of Instances to Enable One-click Scale-up

If the following requirements are met, you are suggested to use AS to achieve one-click scale-up:

- Though it is hard to predict the peaks and troughs of the service, you are not willing to use the system for performing capacity scaling exclusively; (if peaks and troughs are predictable, we suggest that you use scheduled task for capacity scaling)
- Your computing needs are based on projects, and the CVMs to be used every time are similar (for example, collection of social conditions and public opinions, gene sequencing and weather prediction) In such case, you can set the scaling configuration of CVM template, and configure the corresponding scaling group. If you want to scale up later, you can directly modify the required capacity of scaling group.

Performing One-click Scale-up in the Console

1. Create a custom image.
   An instance that is scaled up subsequently will deploy the environment based on this image.
   Suggested steps to create a custom image: You can deploy your services on an existing CVM or a newly-created CVM, set the services to be activated upon the boot of operating system, and export the services as a custom image.
   For any question, refer to Create Custom Images.

2. Create a scaling configuration based on the custom image.
   For more information about creating scaling configuration, refer to Create Scaling Configuration.

3. Create a scaling group.
   During the creation process, select the created scaling configurations. For the minimum group size, maximum group size and initial number of instances, you should fill in these fields based on the upper and lower limits of the required number of servers as well as the current number. After finishing this step, you can scale up at any time.

4. Scale up.
   If the service needs to be scaled up (for example, starting a gene sequencing task or enabling a request-specific server for data collection), you can directly modify the configuration of scaling group to increase the minimum group size, maximum group size and the expected number of instances, and then AS will quickly perform the scale-up operation for you.

Summary: With such operations as CLB forwarding rules, machine configuration and service deployment performed in advance, you can complete scale-up by simply modifying the parameters of scaling group, ensuring the agility of service.

Reducing Capacity

Last updated：2018-05-29 15:33:32

For each scaling group, you can control the time to add instances (scale-up) to it or delete instances (scale-down) from it. You can scale the scaling group manually by adding or removing instances, or allow AS to execute this process automatically by using scaling policy.

When the scaling group is scaled down automatically, you need to know which instances should be terminated in the first place based on the remove policy.

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

Remove Policy

The scaling group will determine which CVM should be removed based on the remove policy during scale-down. You can choose from the following two remove policies:

- **Delete the oldest CVMs**: Delete the oldest CVMs that are added automatically; after this, delete the oldest CVMs that are added manually.
- **Delete the latest CVMs**: Delete the latest CVMs that are added automatically; after this, delete the latest CVMs that are added manually.

> Note: No matter the latest or oldest servers to be deleted, AS will delete the automatically created CVMs in the first place, and then delete manually added CVMs.

Setting and Modifying Remove Policy in the Console

There are two ways to set up:

- Select the remove policy you want when creating the scaling group.
- In the details page of scaling group, click "Edit" to modify the scaling policy.

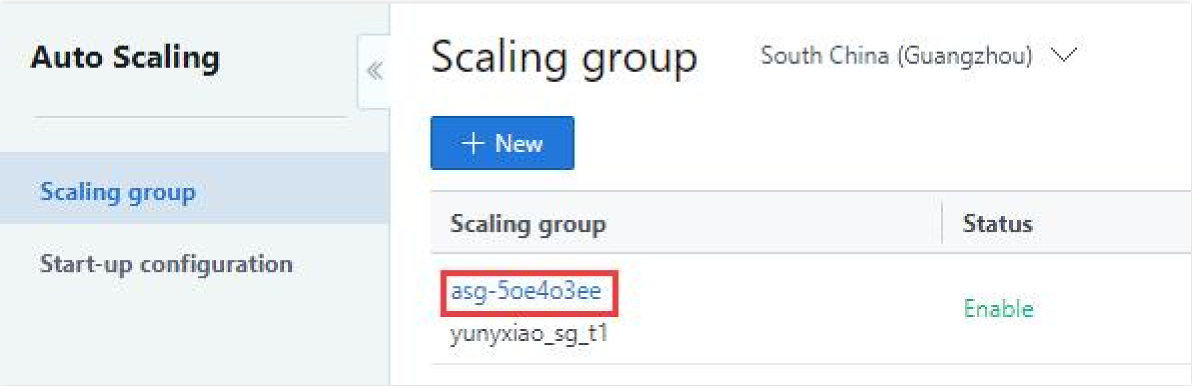
Viewing Scaling Activities

Last updated：2018-05-29 14:55:54

View Scaling Activities

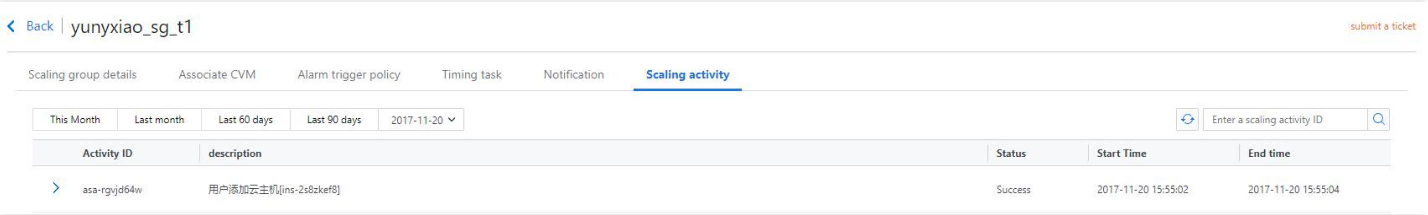
Open the Console, and select **Scaling Group** in the navigation bar.

Select the scaling group to be modified, and click the scaling group ID to enter the basic information page.



Scaling group	Status
asg-5oe4o3ee yunyxiao_sg_t1	Enable

Select **Scaling Activity** in the top navigation bar, and you can view the information of the scaling activity that has been performed by the scaling group based on the scaling policy.



Activity ID	description	Status	Start Time	End time
asa-rgvjd64w	用户添加云主机[ins-2s8zkef8]	Success	2017-11-20 15:55:02	2017-11-20 15:55:04

Suspending and Resuming Scaling

Last updated：2018-05-29 15:35:53

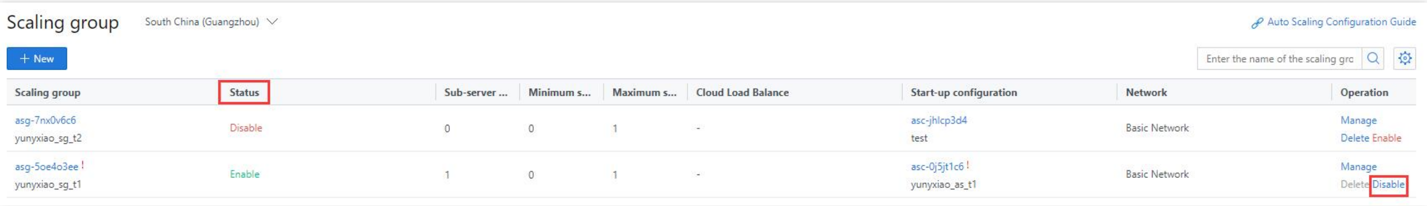
Usage Scenarios

If you need to troubleshoot any problems related to configurations or Web applications (for example, turning off to reset password, upgrading service, etc.), and wish to make modifications to the applications without triggering the auto scaling process, you can suspend the scaling group and resume it after the above operations are performed.

Suspending Scaling Group

Action

Open the Console, select "Scaling Group" in the navigation bar, and click "Disable" at the right side of the scaling group list.



Scaling group South China (Guangzhou)

Auto Scaling Configuration Guide
+ New
Enter the name of the scaling grc

Scaling group	Status	Sub-server ...	Minimum s...	Maximum s...	Cloud Load Balance	Start-up configuration	Network	Operation
asg-7nx0v6c6 yunyxiao_sg_t2	Disable	0	0	1	-	asc-jhlcp3d4 test	Basic Network	Manage Delete Enable
asg-5oe4o3ee ! yunyxiao_sg_t1	Enable	1	0	1	-	asc-0j5jt1c6 ! yunyxiao_as_t1	Basic Network	Manage Delete Disable

When the setting is made, you can see **Disabled** in the **Status** column.

Note

After the scaling group is disabled, the auto capacity scaling of the scaling group will not be triggered, but the restrictions on the scaling group remain in effect.

The activities that are automatically triggered include:

- Alarm Scaling
- Scheduled Task
- Health Check
- Expected number of instances mismatch due to manual operation

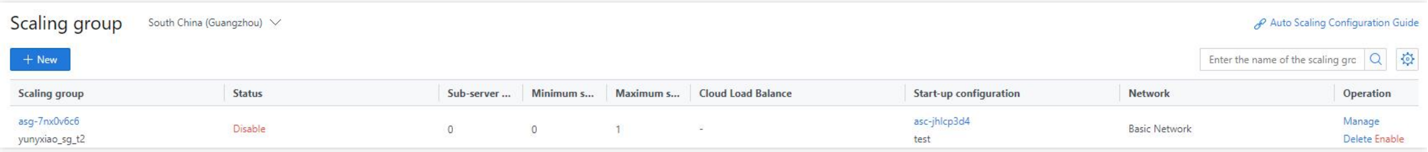
Restrictions on the scaling group include:

- If the number of instances that are removed manually are smaller than the minimum number of instances, the instances are not allowed to be removed;
- If the number of instances that are added manually are larger than the maximum number of instances, the instances are not allowed to be added;
- Increase the minimum or maximum number of instances manually, and do not trigger scaling activity.

Resuming Scaling Group

If you have finished troubleshooting or performed operations when the scaling group activity is suspended, you can resume the auto scaling configuration for your service.

Open the Console, select **Scaling Group** in the navigation bar, and click **Enable** at the right side of the scaling group list.


Scaling group
South China (Guangzhou)
Auto Scaling Configuration Guide
+ New
Enter the name of the scaling grc

Scaling group	Status	Sub-server ...	Minimum s...	Maximum s...	Cloud Load Balance	Start-up configuration	Network	Operation
asg-7nx0v6c6 yunyxiao_sg_t2	Disable	0	0	1	-	asc-jhlcp3d4 test	Basic Network	Manage Delete Enable

Avoiding Reducing Capacity

Last updated：2017-11-22 16:44:09

Introduction

You can specify a submachine in the scaling group to be protected from being scaled down in the scale-down activity. If a scale-down activity occurs, AS will choose a submachine to be scaled down from other CVMs.

You can enable instance protection configuration for one or more scaling group instances. You can modify the scaling group or the instance protection configuration at any time.

If the scaling activity occurs when all the remaining instances in the scaling group are protected from being scaled down, AS will decrease the required capacity instead of removing any instances.

Applicable Scenario

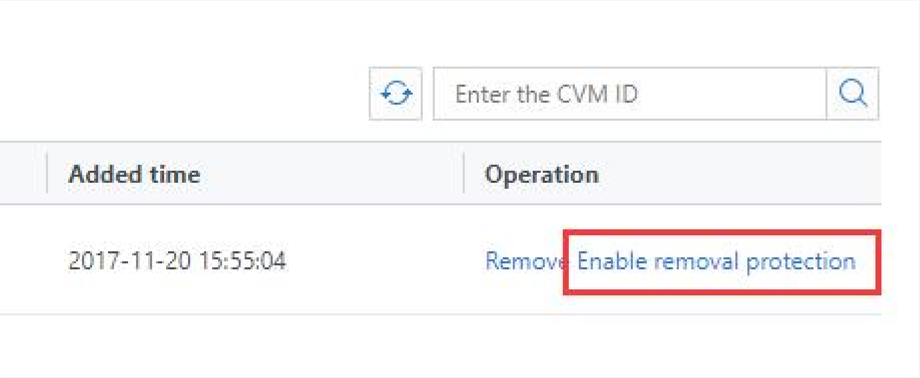
Normally, the CVMs in the scaling group are stateless, and can be removed at any time. In practice, however, the following conditions are applicable to protect specified instances from being scaled down:

- **One server for multiple uses:** In consideration of costs, apart from the tasks specified by the cluster, a server in the cluster is also used for other purposes. For example, the server may be used for storing the data generated in the cluster, so this server is actually stateful.

- **Avoid misoperation:** If you worry that the service will be affected due to policy setup failure, you can set "scale-down exemption" for some servers. In this way, AS will never scale these servers down, and the tunnel "Request-LB-Submachine" will remain unblocked.

Setup

In the list of submachines of the scaling group, you can directly set:



Enter the CVM ID

Added time	Operation
2017-11-20 15:55:04	Remove Enable removal protection

Scaling Activity Failed

Last updated：2017-04-14 14:59:49

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

How to Check the Failed Scaling Activities?

You can check them at Scaling Activity Details.

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 >>