

Auto Scaling

FAQs



Copyright Notice

©2013–2026 Tencent Cloud. All rights reserved.

The complete copyright of this document, including all text, data, images, and other content, is solely and exclusively owned by Tencent Cloud Computing (Beijing) Co., Ltd. ("Tencent Cloud"); Without prior explicit written permission from Tencent Cloud, no entity shall reproduce, modify, use, plagiarize, or disseminate the entire or partial content of this document in any form. Such actions constitute an infringement of Tencent Cloud's copyright, and Tencent Cloud will take legal measures to pursue liability under the applicable laws.

Trademark Notice



This trademark and its related service trademarks are owned by Tencent Cloud Computing (Beijing) Co., Ltd. and its affiliated companies ("Tencent Cloud"). The trademarks of third parties mentioned in this document are the property of their respective owners under the applicable laws. Without the written permission of Tencent Cloud and the relevant trademark rights owners, no entity shall use, reproduce, modify, disseminate, or copy the trademarks as mentioned above in any way. Any such actions will constitute an infringement of Tencent Cloud's and the relevant owners' trademark rights, and Tencent Cloud will take legal measures to pursue liability under the applicable laws.

Service Notice

This document provides an overview of the as-is details of Tencent Cloud's products and services in their entirety or part. The descriptions of certain products and services may be subject to adjustments from time to time.

The commercial contract concluded by you and Tencent Cloud will provide the specific types of Tencent Cloud products and services you purchase and the service standards. Unless otherwise agreed upon by both parties, Tencent Cloud does not make any explicit or implied commitments or warranties regarding the content of this document.

Contact Us

We are committed to providing personalized pre-sales consultation and technical after-sale support. Don't hesitate to contact us at 4009100100 or 95716 for any inquiries or concerns.

FAQs

Last updated: 2026-05-13 11:15:43

What is the Cooling Period?

The Cooling Period refers to a designated lockout duration following the completion of a scaling activity (either addition or removal of a Cloud Virtual Machine, CVM) within the same scaling group. During this interval, the scaling group refrains from executing any scaling activities. The selectable range for the Cooling Period is `0-999999` (seconds).

Does a manually added Cloud Virtual Machine require a Cooling Period?

A manually added Cloud Virtual Machine does not require a Cooling Period.

Is there a charge for Auto Scaling?

The Auto Scaling (Auto Scaling) feature is entirely free of charge, so please feel free to utilize it.

You only need to pay for other services associated with Auto Scaling, such as Cloud Virtual Machine (CVM) instances that are automatically created or manually added through Auto Scaling. For more information on Cloud Virtual Machine billing, please refer to [Billing Modes](#).

How can I increase the maximum number of Cloud Virtual Machines in a scaling group?

Auto Scaling supports up to `2000` Cloud Virtual Machines per scaling group. For the current quota of Pay-As-You-Go Cloud Virtual Machines that a Tencent Cloud user can have in each availability zone, please refer to [Limits on Purchasing Pay-As-You-Go Cloud Virtual Machine Instances](#). If you wish to have more than 2000 Cloud Virtual Machines in a scaling group or want to use more than the fixed quota of Pay-As-You-Go Cloud Virtual Machines within an availability zone, please apply through [Online Support](#).

What type of machine is suitable for using Auto Scaling?

The applications deployed on the Cloud Virtual Machine instances within a scaling group need to be stateless and replaceable. Since instances within the group may be reclaimed during scaling down, the Cloud Virtual Machines used for Auto Scaling should not store application state information (such as sessions) or related data (such as databases, logs, etc.). If state information needs to be stored in the application, consider saving it to a standalone Cloud Virtual Machine outside the scaling group.

What are the requirements for manually adding Cloud Virtual Machines to a scaling group?

The Cloud Virtual Machines manually added to a scaling group must meet the following requirements:

- Resides within the same region as the scaling group;
- The network environment (Private Network VPC or Basic Network) is consistent with the scaling group;
- In an operational state.

What are the requirements for the Load Balancer instance associated with the scaling group?

The Load Balancer instance associated with the scaling group (or the backend instance's Private Network VPC in the case of cross-region Load Balancer instances) must reside within the same network environment (Private Network VPC or Basic Network in the same region) as the scaling group.

Can Auto Scaling automatically adjust the configuration of cloud servers?

Auto Scaling is a management service that automatically adjusts elastic computing resources based on user business requirements and strategies. It can automatically increase cloud servers when business grows and decrease them when business declines. Currently, Auto Scaling does not support "vertical scaling", meaning it temporarily cannot automatically adjust the CPU, memory, and bandwidth of cloud servers.

Is it necessary to use Auto Scaling in conjunction with Load Balancing and Tencent Cloud's Observability Platform?

Auto Scaling can independently expand and contract cloud servers. It can be deployed in conjunction with Load Balancing, or it can be deployed without it.

How should I set up to expand a batch of cloud servers during a specific time period?

Users can set a pair of timed tasks: one timed task defines the expansion operation, setting the desired instance number to the amount you want to expand; the other timed task defines the contraction operation, setting the desired instance number to the amount you want to contract.

For more details, see [Best Practice Examples](#) .

What is the naming rule for cloud server instances that are automatically expanded?

The cloud server instances that are automatically expanded for you by the scaling group will be named in the format of `as-ScalingGroupName+AutomaticallyIncrementingNumber`.

For example, if your scaling group name is `Test`, and two cloud server instances have been automatically expanded, then the names of these two cloud server instances will be `as-Test1` and `as-Test2` respectively.

What are the specific rules for the removal policy of the scaling group?

Tencent Cloud Auto Scaling offers two removal policies:

- Remove the oldest instances: Delete the oldest automatically added machines. Once the automatically added machines are deleted, the earliest manually added machines are removed.
- Remove the newest instances: Delete the most recently added automatic machines. Once the automatically added machines are deleted, the most recently manually added machines are removed.

How does the alarm policy aggregate information from Tencent Cloud's Observability Platform?

Taking the maximum value as an example, the basic strategy for maximum value statistics is to take a value every minute for each cloud server's set monitoring item during each period. If the obtained values meet the set rules for consecutive periods (the number of periods can be customized by the user), an alarm scaling action will be triggered.

For instance, in a scaling group with 5 CVM instances, the defined alarm scaling policy is "If the maximum/minimum/average CPU utilization exceeds 50% within 5 minutes, occurring 3 times", the Auto Scaling collects monitoring data and makes policy judgments, following these steps:

1. The system will take one value from each CVM every minute, resulting in 25 CPU usage values taken within a period (currently set to 5 minutes).
2. The alarm rules are determined based on whether the maximum/minimum/average values align with the configured strategy:
 - Maximum Value: If the highest of these 25 values exceeds the threshold (50%), the period complies with the alarm scaling rules.
 - Minimum Value: If the lowest of these 25 values exceeds the threshold (50%), the period complies with the alarm scaling rules.
 - Average Value: If the average of these 25 values exceeds the threshold (50%), the period complies with the alarm scaling rules.
3. If this rule is met for three consecutive periods (totaling 15 minutes, with each period evaluated every 5 minutes), it will trigger a scaling action.

What is the Desired Instance Count?

The Desired Instance Count refers to the current reasonable number of instances in the scaling group, which falls between the minimum and maximum scaling numbers. You can manually adjust the Desired Instance Count, or trigger adjustments using scheduled tasks and alarm scaling tasks. The scaling group will automatically adjust the actual number of instances to match the Desired Instance Count.

- Upon creation of a scaling group: If the user sets an initial instance count when creating the scaling group, the Desired Instance Count will be the initial instance count.
- Alarm Scaling Task Adjustment: When alarm scaling is triggered, the scaling group will adjust the current instance count to the Desired Instance Count. For instance, if the triggered action is to add two cloud servers, this will be achieved by adding 2 to the current Desired Instance Count. If the system detects a discrepancy between the current instance count of the scaling group and the Desired Instance Count, it will add two cloud servers to equalize the current instance count with the Desired Instance Count.
- Scheduled or Manual Adjustment of Desired Instance Count: When the Desired Instance Count is altered by the user through a scheduled task or direct modification, the system will trigger scaling if it detects a discrepancy between the current instance count and the Desired Instance Count, until they are equal.
- System Adjustment: The Desired Instance Count falls between the maximum and minimum instance counts. If the maximum or minimum instance counts change, it may lead to a change in the Desired Instance Count. For example, if the Desired Instance Count is 3, the minimum scaling number is 2, and the maximum scaling number is 5; if the minimum scaling number is adjusted to 4, then the Desired Instance Count will be adjusted to 4 to comply with the minimum scaling number.

What should be noted when specifying data disk snapshots in the launch configuration?

If a data disk snapshot is specified in the launch configuration, it is necessary to ensure that the data disk can be correctly auto-mounted for the scaling group to successfully auto-scale. You need to perform some operations on the original instance that created the data disk snapshot before setting up Auto Scaling, in order to support the automatic mounting of the data disk when launching new cloud server instances.

For specific methods, please refer to: [Mounting Cloud Disk](#).

What operations will be suspended after the scaling group is deactivated?

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.

What stages does the lifecycle of a cloud server automatically added to the scaling group include?

- Creating: The sub-machine is in the process of being created.
- InService: The sub-machine is currently operational.
- Removing: The sub-machine is currently being removed.
- Attaching: The sub-machine is currently being bound to the scaling group.
- Detaching: The sub-machine is currently being unbound from the scaling group.
- AttachLb: The sub-machine is currently binding to the Load Balancer.
- DetachLb: The sub-machine is currently unbinding from the Load Balancer.
- Preheating: The sub-machine is currently undergoing preheating.

What is the rule for removing a sub-machine?

- Manual addition of sub-machine removal: The removed sub-machine no longer falls within the management scope of the scaling group. Auto Scaling will not delete it, but will unbind this instance. When joining the scaling group, the Load Balancer automatically bound to the instance by Auto Scaling will not be unbound, whereas the Load Balancer manually bound by the user will not be unbound.
- Automatic scaling removal of sub-machine: The sub-machine will be destroyed and the Load Balancer will also be unbound.