Auto Scaling

Scaling Groups

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.

Contents

Scaling Groups
Scaling Group Overview

Last updated：2024-01-17 17:54:47

An auto-scaling group is a collection of cloud server instances that adhere to the same rules and cater to the same scenario.
The auto-scaling group delineates the maximum and minimum number of CVM instances within the group, along with attributes such as associated load balancing instances.

Overview

Last updated: 2024-01-17 17:55:01

A hybrid scaling group, combining pay-as-you-go instances and spot instances, can assist you in optimizing the usage costs of Cloud Virtual Machine (CVM), while ensuring the scale and performance you require.

Feature Overview

- Utilize the Spot Instance Reclamation Monitoring to proactively monitor interruption notifications, and gracefully return spot instances before the system triggers reclamation, thereby mitigating the impact of spot instance interruptions on your operations.
- Regulate the quantity and ratio of pay-as-you-go instances to spot instances within the scaling group.
- Specify the spot instance creation strategy:
  - Through the Capacity Optimization Strategy (recommended), you can best utilize spot instance resources, reducing overall resource usage costs while making every effort to minimize the possibility of interruptions.
  - Maximize cost savings through the Cost Optimization Strategy.
- Execute necessary operations before returning instances by integrating lifecycle hooks.

Usage Recommendations

Auto Scaling can assist you in achieving the scale and performance your business actually requires, maintaining service availability with the most suitable cluster capacity. Auto Scaling is designed to support flexible workloads and respond quickly to dynamic capacity changes. When creating a scaling group mixed with pay-as-you-go and spot instances, it is recommended:

- Enhance availability by deploying services across multiple availability zones, using multiple machine types and billing models.
- Configure multiple machine types in the launch configuration, for more details, please refer to Multi-Machine Type Configuration. As the resource amount of each machine type in each availability zone is independent and unrelated, configuring multiple machine types can yield more computing resources.
- Do not restrict yourself to the latest generation of machine types. Opt for other generations to reduce costs and minimize interruptions.
- Opt for a retry strategy, where Auto Scaling will continuously attempt to recreate, ensuring new instances are created even in the event of initial failure.
- To further reduce the possibility of interruptions in spot instances, it is recommended to use in conjunction with spot instance recovery monitoring, multi-machine type configuration, and spot instance allocation strategy (capacity optimization strategy recommended). This is suitable for businesses that can quickly migrate between CVMs, such as containerized workloads, big data and analytics, image and media rendering, batch processing, web applications, etc.
- If you need to perform custom operations before returning instances, you can use the lifecycle hook of the scale-down activity type.

Operations Guide

You may refer to Creating a Scaling Group to create a scaling group that mixes pay-as-you-go and spot instances via the Auto Scaling Console.

Bidding instance recycling monitoring

Last updated: 2024-01-18 10:41:56

Bidding Instances and Interruption Notifications

Bidding Instances are discounted CVM computing resources available for your use, currently sold at 20% of the list price of pay-as-you-go instances. Unlike pay-as-you-go instances, bidding instances will be interrupted and reclaimed when CVM resources are scarce. CVM will issue an interruption notification 2 minutes before the reclaim, and you can check the reclaim status of bidding instances to obtain interruption information for bidding instances.

Introduction to Bidding Instance Reclaim Monitoring

The bidding instance reclaim monitoring feature of Auto Scaling provides an automated experience for managing the lifecycle of bidding instances. It is recommended to enable bidding instance reclaim monitoring when creating a scaling group.

> **Note:**
> Enabling bidding instance reclaim monitoring can help mitigate the risks associated with bidding instance interruptions. However, you will still face risks associated with the interruption characteristics of bidding instances. It is advisable to avoid running services with high stability requirements on bidding instances.

Enabling Bidding Instance Reclaim Monitoring

This feature proactively queries for bidding instance interruption notifications and automatically attempts to replace them with new instances prior to interruption. This helps maintain the ratio of pay-as-you-go instances to bidding instances within the scaling group, thereby ensuring workload availability.
In addition to actively replacing bidding instances to maintain balance in the scaling group capacity, it also supports removing bidding instances from the scaling group through the normal return process. As soon as an interruption notification for a bidding instance is detected, Auto Scaling immediately initiates the return process, returning the instance about to be reclaimed, while simultaneously attempting to proactively replace it with a new instance.

> **Note:**
> - If Auto Scaling detects a bidding instance interruption notification while a scaling activity is in progress, it will first complete the scaling activity before creating a new instance to replace the reclaimed instance.
> - If you need to perform necessary operations before returning an instance, you can utilize the lifecycle hook of the scale-down activity type.

Disabling Bidding Reclaim Monitoring

Auto Scaling will proceed with the replacement only after the bidding instance is interrupted and a health check failure is detected.

Combining Lifecycle Hooks

You can configure the lifecycle hook of the scale-down activity while enabling the bidding instance reclaim monitoring. Once configured, you will have the opportunity to perform custom operations on it before Auto Scaling actively returns the bidding instance. For more information, please refer to Lifecycle Hooks.

> **Note:**
> If the scaling group is already bound to a load balancer, Auto Scaling will invoke the lifecycle hook to perform custom operations after the instance is unbound from the load balancer.

Utilizing the lifecycle hook feature of the scale-down activity provides an opportunity to assist you in completing the following operations before Auto Scaling releases the instance:

- Performing Data Backup Operations
- Uploading System or Business Logs to Object Storage
- Gracefully Shutting Down CMQ Worker Threads
- Deregistering from the Domain Name System

Quantity and ratio of pay-as-you-go and spot instances

Last updated: 2024-01-17 17:55:38

When establishing a scaling group, you can designate the proportion of on-demand instances and spot instances within the group, as well as the base quantity of pay-as-you-go instances.

Designate the base number of pay-as-you-go instances.

The base number of pay-as-you-go instances is the minimum quantity that must be met within the scaling group. If a base number of pay-as-you-go instances is specified, the scaling group will ensure that these instances are created first during expansion.

Specify the percentage of pay-as-you-go instances.

Calculate the portion exceeding the base number of pay-as-you-go instances, then determine the quantity of pay-as-you-go instances and spot instances to be created based on the percentage of pay-as-you-go instances. You can specify any ratio between 0 and 100.

Activate pay-as-you-go instances to supplement spot capacity.

If you activate pay-as-you-go instances to supplement spot capacity, when the spot instance inventory corresponding to the instance type selected in your launch configuration is insufficient, Auto Scaling will attempt to create pay-as-you-go instances corresponding to the instance type.

> **Note:**
> If this feature is enabled, the ratio of pay-as-you-go instances to spot instances within the scaling group cannot be strictly controlled.

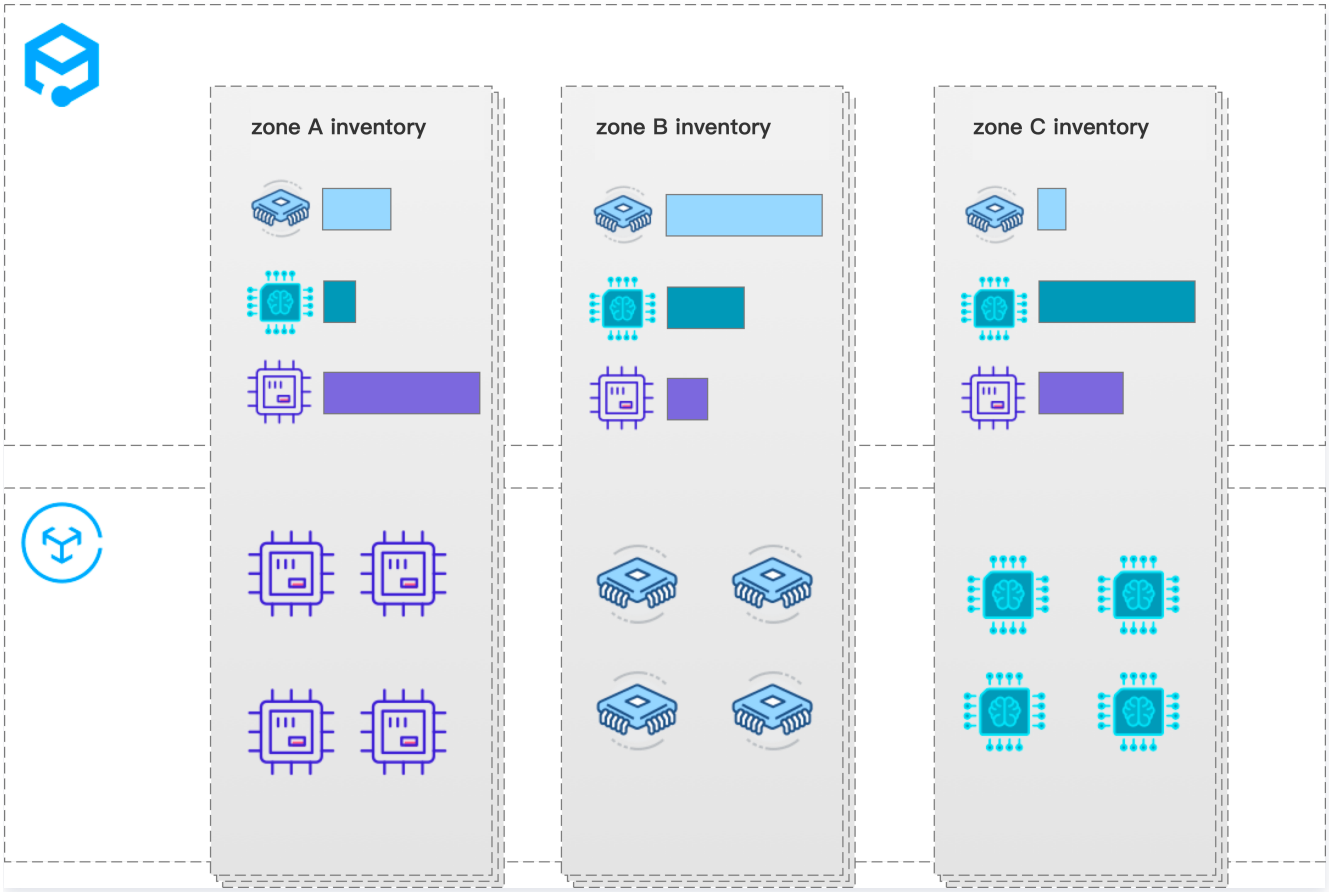
Spot instance creation policy

Last updated: 2024-01-17 17:55:48

The creation strategy for Auto Scaling bid instances can assist you in utilizing bid resources in the most efficient and optimal manner, while still taking advantage of the substantial discounts offered by bid instances. This is a strategy that allows you to optimize the cost of using computing resources. Auto Scaling offers the following strategies for creating bid instances:

Capacity Optimization Strategy (Recommended)

Auto Scaling prioritizes the selection of the most available bid instance types. Expanding in this manner can assist you in making the best use of bid instance resources. The capacity optimization strategy can reduce the overall resource usage cost while minimizing the possibility of interruptions. The schematic diagram is as follows:



This allocation strategy is applicable to:

- Big Data and Analytics
- High performance computing
- Image and Media Rendering
- Business operations with high interruption costs, such as machine learning.

Lowest price

The Auto Scaling strategy prioritizes the selection of the lowest-priced single-core bidding instance model, allocating your instances from the availability zones you specify. This method of expansion can help you save

costs to the greatest extent. The cost optimization strategy is suitable for businesses that require low interruption costs and aim to minimize the cost of computing resource usage. For instance:

- Time-insensitive business workloads
- Extremely brief workloads
- Business operations that are easily supplemented with checkpoints and restarts (utilizing checkpoints can assist in ensuring that your business operates in a reliable and predictable manner).

Creating a Scaling Group

Last updated: 2024-01-19 10:31:22

Scenario

This document provides guidance on how to create a scaling group through the Tencent Cloud Auto Scaling console.

Instructions

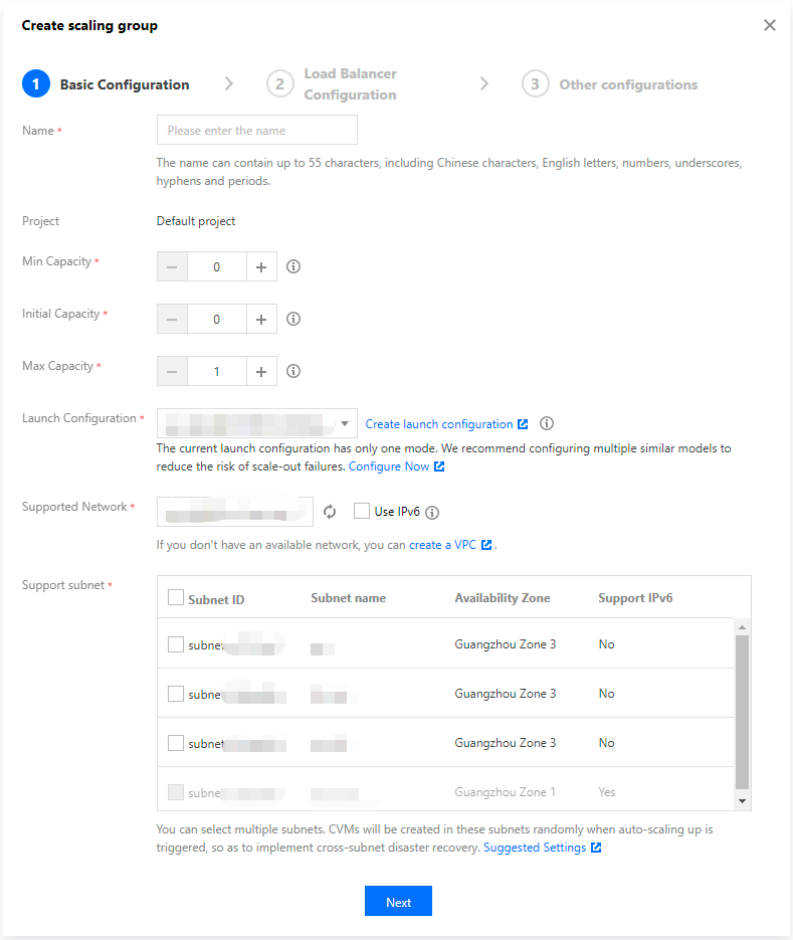
Select a region

1. Log into the Auto Scaling console and select **Scaling Groups** from the left navigation bar.
2. Select the desired region at the top of the "Scaling Groups" page.

   The choice of region restricts the cloud servers that can be manually added and the load balancers that can be bound. For instance, if the launch configuration's region is set to Guangzhou, then the cloud servers automatically added to the scaling group will be from Guangzhou. In a scaling group with the region set to Guangzhou, you cannot manually add cloud servers from other regions such as Shanghai, Beijing, Hong Kong, Toronto, etc., nor can you bind load balancers from these regions.

Configure the Scaling Group

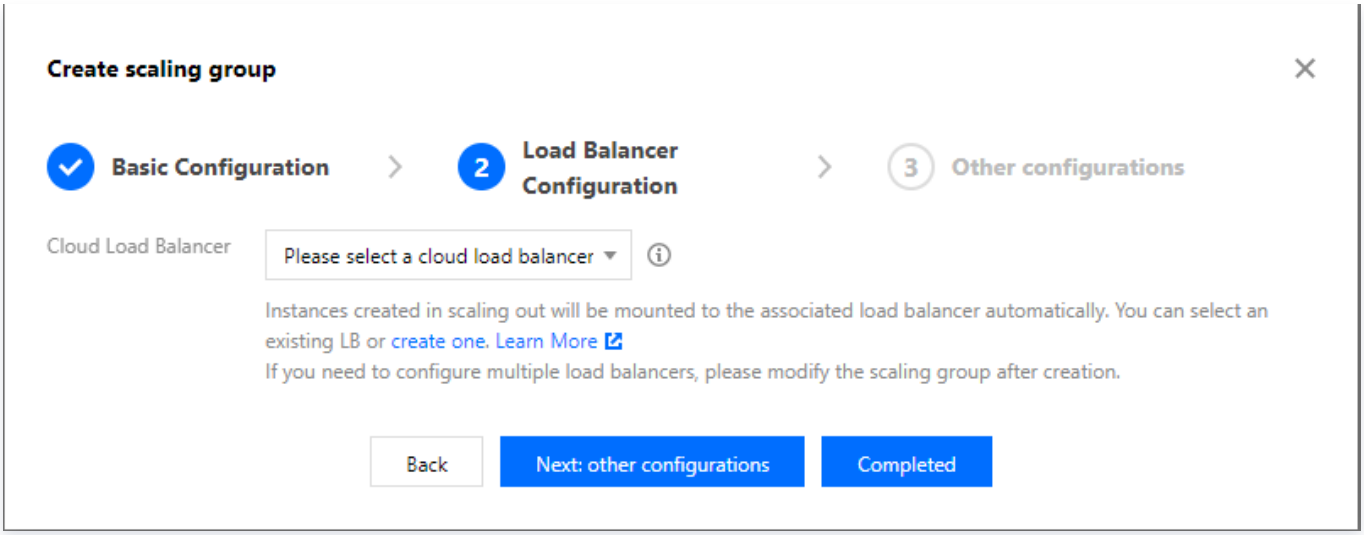
1. On the **Scaling Groups** page, click **New**.



   - **Scaling Group Name**: A custom name used to identify this scaling group.

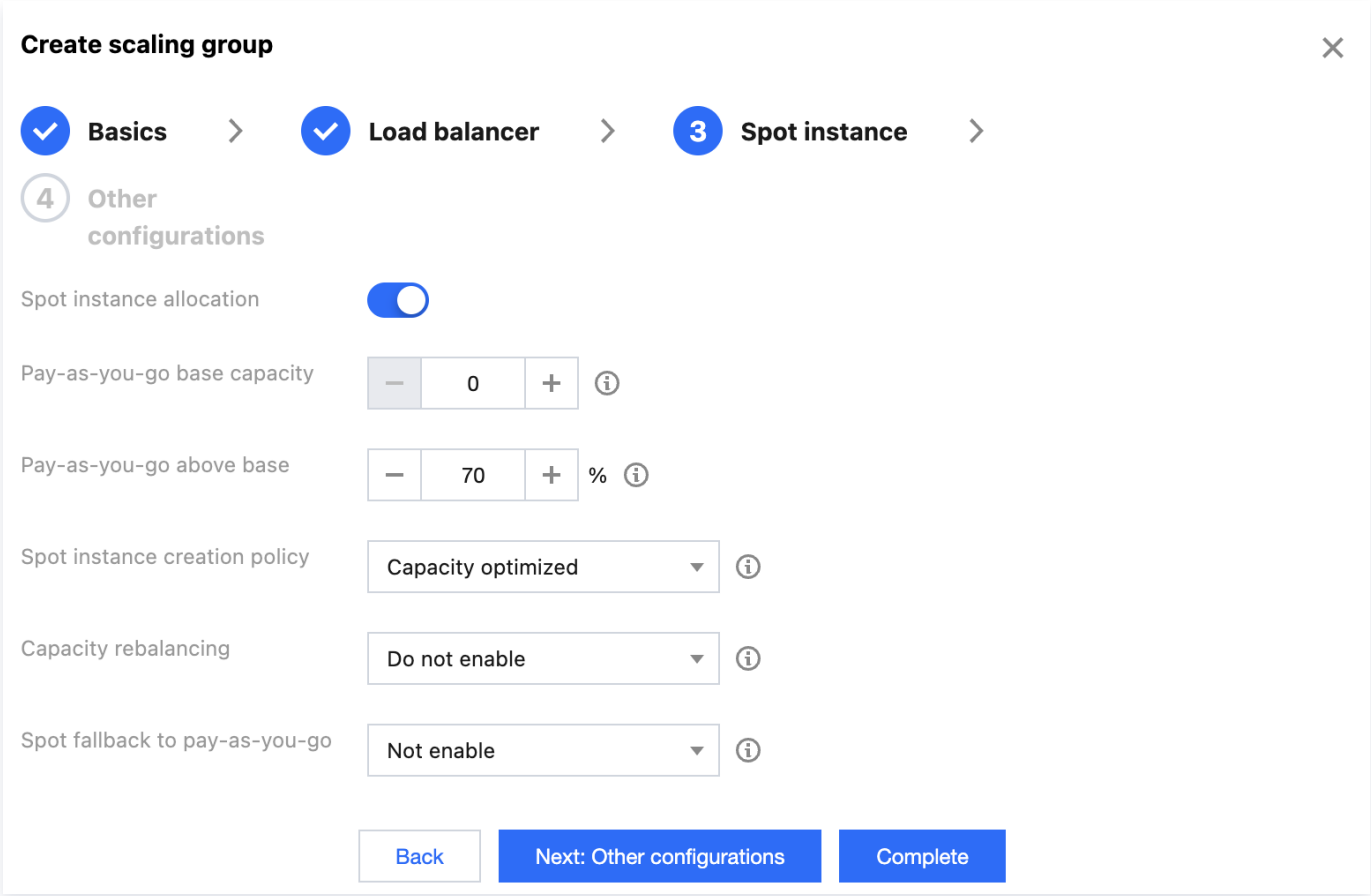
- **Minimum Scaling Number**: Specifies the minimum number of instances in the scaling group.
- **Initial Instance Count**: Specifies the number of instances that the scaling group will **automatically** generate at the start. The corresponding number of instances will be produced after the scaling group is created.
- **Maximum Scaling Number**: Specifies the maximum number of instances in the scaling group.
- **Launch Configuration**: Specifies the created launch configuration. Expansion machines will be created according to this configuration during scaling.
- **Network Support**: Specifies the network attributes of the expanded machines, i.e., whether the expanded machines are in the basic network or in a certain private network (VPC).
- **Subnet Support**: Specifies the subnet where the expanded machines are located. You can choose multiple subnets, and the automatically expanded machines will be randomly created from the subnets you have selected, achieving a cross-subnet disaster recovery effect.

2. Click **Next** to proceed with the load balancing configuration for the scaling group.



You can choose an existing load balancer or create a new one. The expanded machines will automatically mount under the load balancer you have associated. If you need to configure multiple load balancers, please edit the scaling group after creation.

3. (Optional) Click **Next: Spot Instance Allocation** to configure the spot instance allocation strategy. You can also click **Finish** to skip this step.
   Turn on the "Use Spot Instance" switch. Once activated, it appears as shown below:



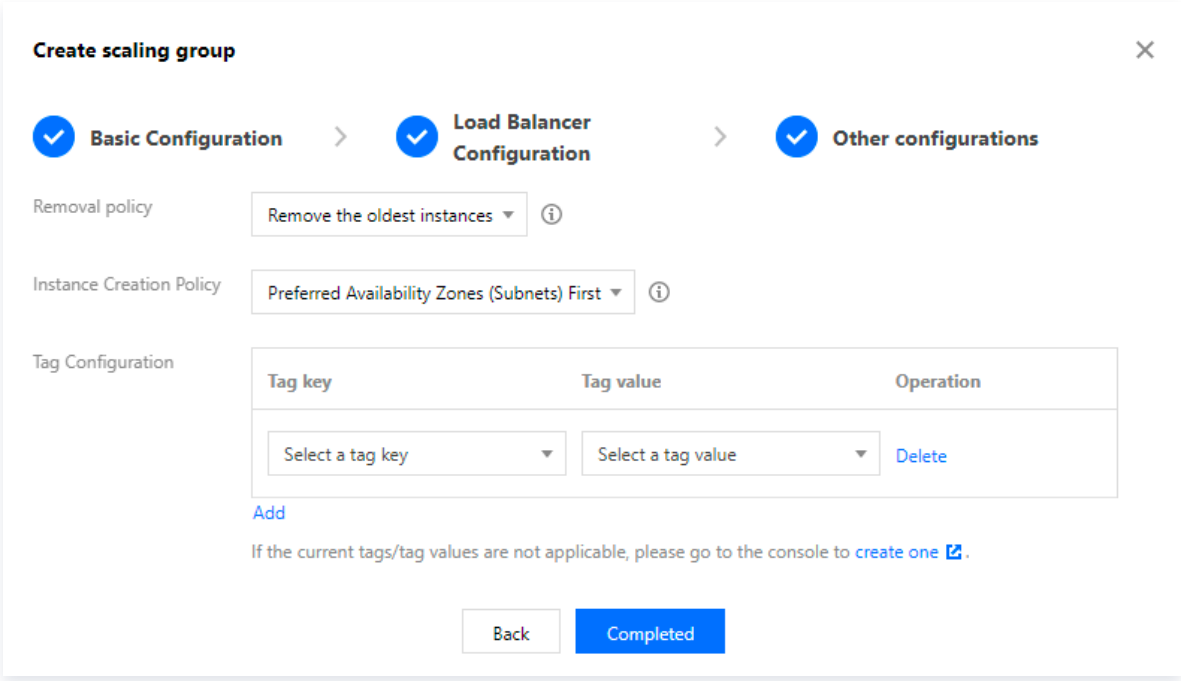
- **Number of On-Demand Base Instances**: The minimum number of on-demand billed instances that must be met within the scaling group. When the scaling group expands, this portion of the instances is expanded first.
- **On-Demand Instance Percentage**: The proportion of on-demand instances, excluding the number of on-demand billed base instances. You can specify any ratio between 0 and 100.
- **Spot Instance Creation Strategy**: The strategy for creating spot instances when the launch configuration configures multiple machine types.
  - **Capacity Optimization Strategy**: Prioritize the selection of the most available spot instance types. Expanding in this manner can help you make the best use of spot instance resources.
  - **Cost Optimization Strategy**: Prioritize the selection of spot instance types with the lowest per-core price. Your instances will be allocated from the availability zones you specify. Expanding in this manner can help you save costs to the greatest extent.
- **Spot Instance Reclamation Monitoring**: When enabled, Auto Scaling will attempt to proactively replace the spot instances in the scaling group that are about to be reclaimed with new instances, thereby helping you maintain the number of instances within the scaling group and the ratio of on-demand instances.
- **On-Demand Instances Supplementing Spot Capacity**: When enabled, it will attempt to create on-demand billed instances for you when the spot instance inventory of your configured machine type is insufficient.

Note:

- For detailed information about scaling groups mixed with on-demand billing and spot instances, please refer to Overview.

- A scaling group mixed with on-demand billing and spot instances can only be created when the specified launch configuration billing mode is on-demand billing.

4. (Optional) Click **Next: Other Configurations** to proceed with other related configurations for the scaling group, or you can skip this step by clicking **Finish**. As shown in the figure below:



- **Removal Policy**: When the scaling group needs to reduce instances and there are multiple options, the instance to be removed will be selected based on the removal policy. You can choose:
  - **Remove the Oldest Instance**: This option removes the oldest automatically added machine. Once all automatically added machines are removed, the earliest manually added machines will be removed. This option is typically selected.
  - **Remove the Newest Instance**: This option removes the most recently added automatic machine. Once all automatically added machines are removed, the most recently manually added machines will be removed.
- **Instance Creation Policy**: When the scaling group needs to add instances and multiple subnets are specified in different availability zones, instances will be added according to this policy. You can choose:
  - **Priority to Preferred Availability Zone (Subnet)**: Based on the order of your configured availability zones (subnets), the system will prioritize the earlier configurations. If it fails, it will automatically retry in order (suitable for architectures that primarily rely on a specific availability zone, with other zones as secondary).
  - **Disperse Across Multiple Availability Zones (Subnets)**: The system will create new instances in the relatively less populated availability zones (subnets) based on the distribution of instances within the scaling group during expansion (suitable for architectures that require evenly distributed instances).
- **Tag Configuration**: You can manage resources categorically through tags. For more details, please refer to Tags.

5. Click **Complete**.

See Also

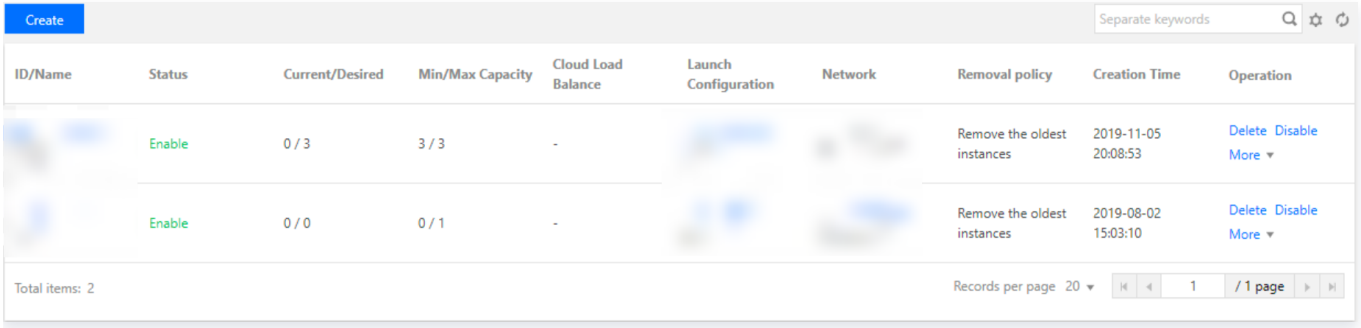
Upon completion of the scaling group creation, the group can accommodate machines, but it does not yet possess the capability for intelligent scaling. We strongly recommend that you proceed with the following three operations:

- Add Existing Cloud Server
- Manage Alarm Trigger Strategy
- Create Notification

Viewing Scaling Group List

Last updated: 2024-01-17 17:55:59

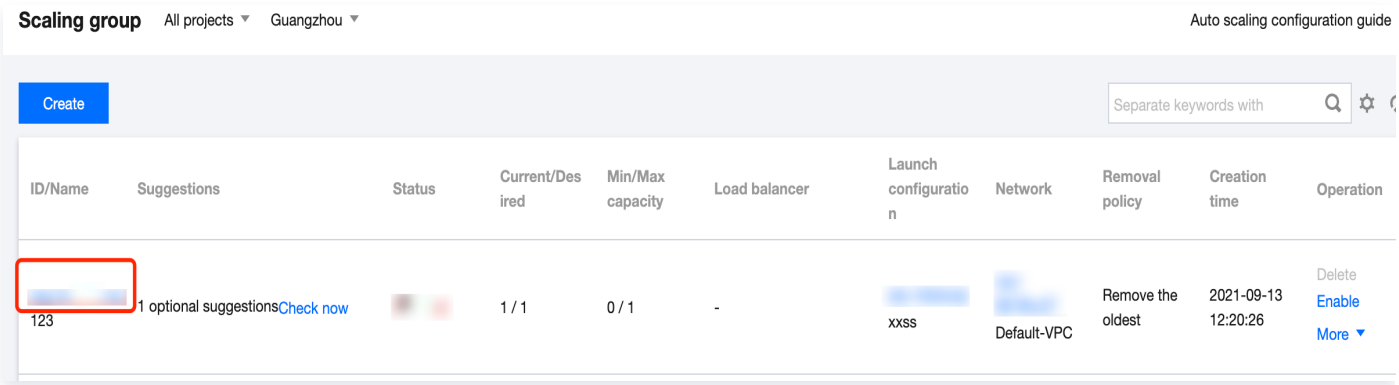
Access the Auto Scaling Console, and select Scaling Group from the navigation bar on the left to view the list, as illustrated below:



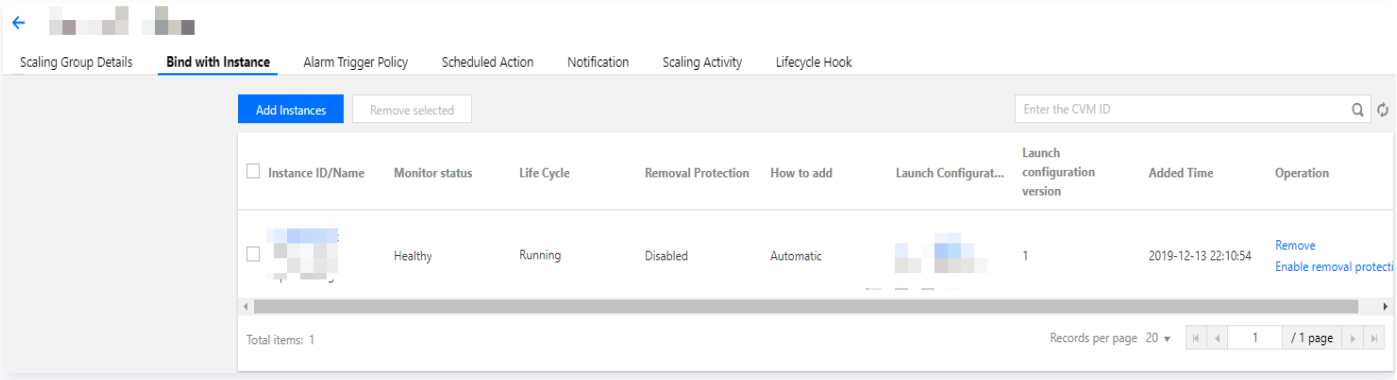
Modifying Bound Instances

Last updated: 2024-01-17 17:56:09

1. Access the Auto Scaling Console, and select **Scaling Group** from the navigation menu on the left.
2. Select the scaling group you wish to modify, and click on the scaling group ID to enter the basic information page for that group, as illustrated below:



3. On the Scaling Group details page, select the **Bind with Instances** tab to view the list of instances associated with this scaling group, as depicted below:



   - To manually add CVM instances to the scaling group, select **Add Instance**, make your selection in the pop-up window (hold Shift for multiple selections), and click **Confirm**.
   - To unbind a particular instance, click on **Remove** located on the right side of the instance's row.

> **Note:**
> - Machines generated automatically will be destroyed upon removal.
> - Machines added manually will not be destroyed upon removal. They will only be removed from the scaling group and unbound from the load balancer.

Modifying Scaling Groups

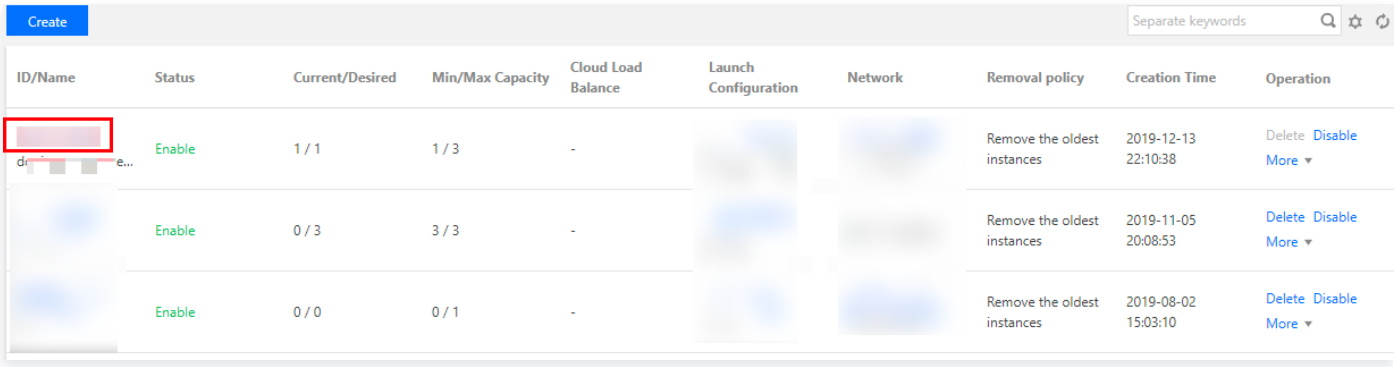
Last updated: 2024-01-17 17:56:20

Scenario

This document provides guidance on how to modify the basic information of the scaling group and the allocation strategy for spot instances via the Auto Scaling console.

Instructions

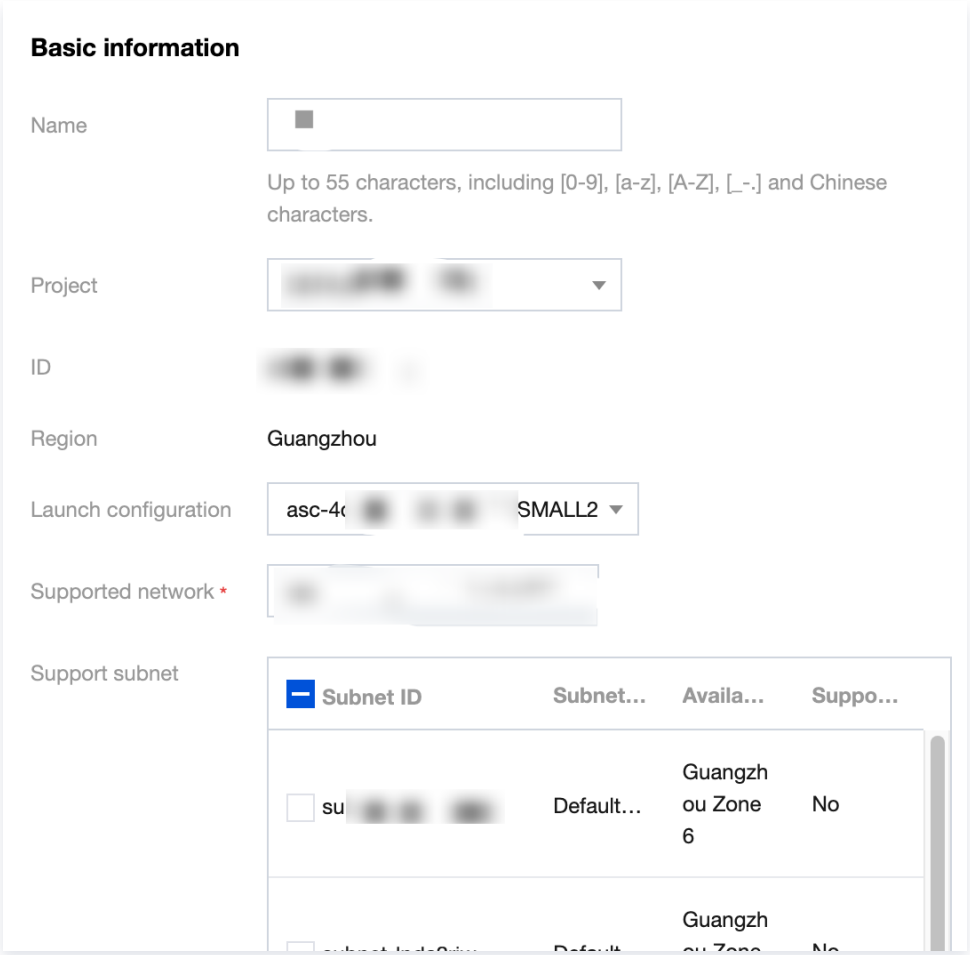
1. Log in to the Auto Scaling console and select **Scaling Group** from the left navigation bar.
2. On the **Scaling Group** page, select the ID of the scaling group you wish to modify, which will take you to the basic information page of that scaling group, as shown in the figure below:



Refer to the following steps to modify the scaling group configuration as needed.

Modifying Basic Information of the Scaling Group

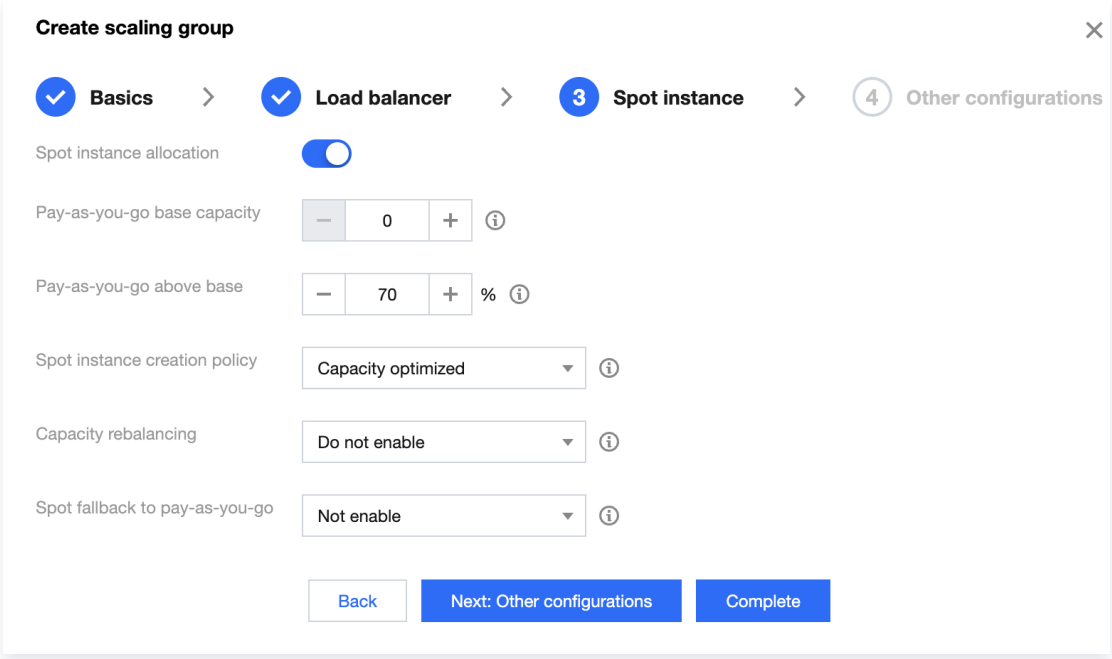
1. After clicking on **Edit** in **Basic Information**, you can modify the scaling group name, adjust the minimum and maximum scaling numbers, and alter the CVM instance removal strategy, as depicted in the figure below:



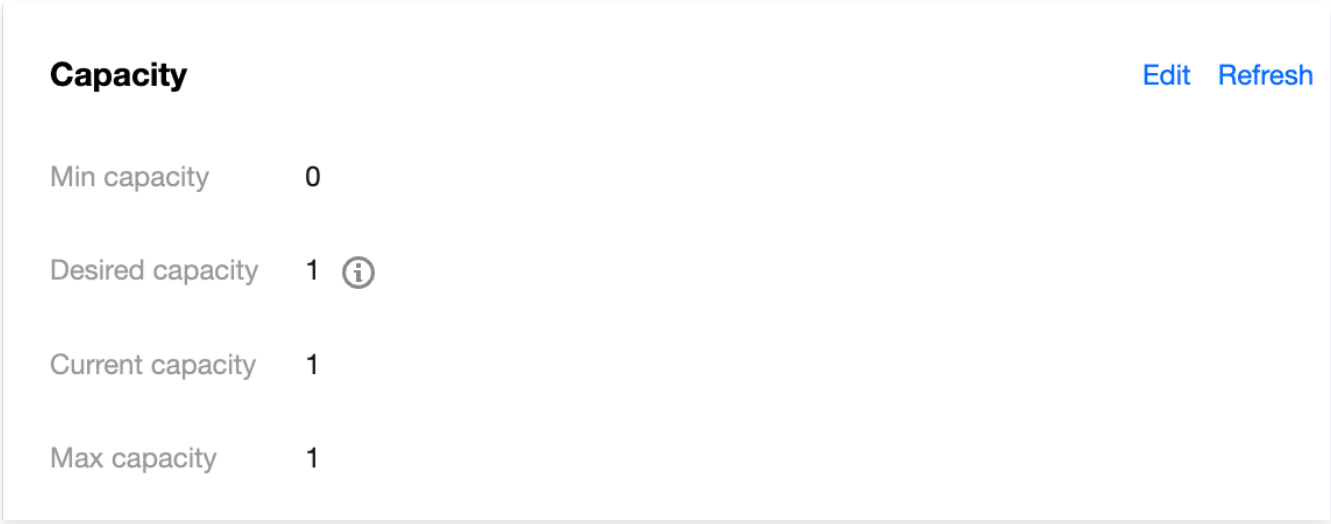
2. Once modifications are complete, click **Save** to apply the changes.

Modifying the Bidding Instance Allocation Strategy of the Scaling Group

1. After clicking on **Edit** in **Bidding Instance Allocation Strategy**, you can enable or disable the use of bidding instances, modify the number of pay-as-you-go base instances, the percentage of pay-as-you-go instances, the creation strategy for bidding instances, the monitoring of bidding instance recovery, and the supplementation of bidding capacity with pay-as-you-go instances, as depicted in the figure below:



2. Once modifications are complete, click **Save** to apply the changes. The number of pay-as-you-go base instances and the percentage of pay-as-you-go instances can be viewed or modified in **Instance Quantity Information**, as depicted in the figure below:



Adding CLBs

Last updated: 2024-01-17 17:56:31

When adding and removing instances from the Auto Scaling, it is crucial to ensure the distribution of application traffic across all cloud server instances. If you desire the expanded machines to be under a certain load balancer and receive the load-balanced forwarded traffic without your intervention, you can assign your machines a specific load balancer. This load balancer will serve as the sole point of contact for all incoming traffic to the instances in your scaling group.

Adding a Load Balancer to the Scaling Group

The integration of the scaling group and load balancing allows you to attach a load balancer to an existing scaling group. Once the load balancer is attached, it automatically registers the instances within the group and distributes incoming traffic amongst them. Subsequently, expanded instances and those manually added to the scaling group will automatically mount under the associated load balancer. Conversely, instances within the scaling group that are scaled down, removed, or deleted will automatically unmount from the load balancer associated with the scaling group.

1. Log into the Auto Scaling console and select **Scaling Groups** from the left navigation bar.
2. On the "Scaling Groups" list page, click on **Create**.
3. During the "Load Balancer Configuration" step of creating a new scaling group, select the load balancer you require. If you have not created one in advance, you can click on **Create** below the option to establish a new load balancer.

> ⚠ **Note:**
> The load balancing instance associated with the scaling group (or the backend instance's private network in the case of cross-regional load balancing) must reside within the same network environment (private network or basic network in the same region) as the scaling group.

Removing the Load Balancer from the Scaling Group

On the "Scaling Groups" list page, click on the ID to enter the details page of the scaling group. In the **Load Balancer Information** module, you can remove the corresponding load balancer.

> ⚠ **Note:**
> Upon deletion, the machines within the scaling group will automatically unbind from the removed load balancer.

Delete Scaling Group

Last updated: 2024-01-18 11:03:05

1. Log into the Auto Scaling console and select **Scaling Groups** from the left navigation bar.
2. Select **Delete** on the right side of the row where the scaling group you wish to remove is located. Confirm in the pop-up window to proceed with the deletion.

> **Note:**
> - The scaling group can only be deleted after the instances within it have been removed.
> - For scaling groups configured with load balancing, when instances within the group are deleted, they are automatically unmounted from the load balancer associated with the scaling group.