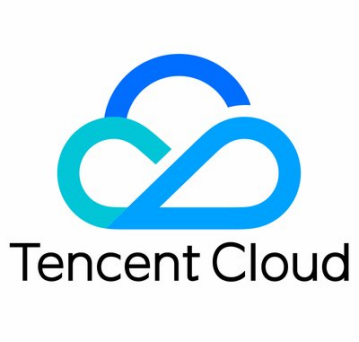
Auto Scaling

Getting Started



Contents

Getting Started
Creating a Scaling Plan in 5 Minutes

Last updated: 2024-01-18 11:07:39

Feature Overview

The Quick Start Guide outlines how to create a comprehensive auto-scaling solution, which can be accomplished in the following three steps:

- Step 1:Creating a Launch Configuration
- Step 2:Creating a Scaling Group
- Step 3:Creating a Scaling Policy

> **Note:**
> The operations in the Quick Start Guide are demonstrated using the Auto Scaling console as an example. If you prefer using APIs, please refer to API Usage Examples.

Step 1:Creating a Launch Configuration

Last updated: 2024-01-17 17:52:34

Scenario

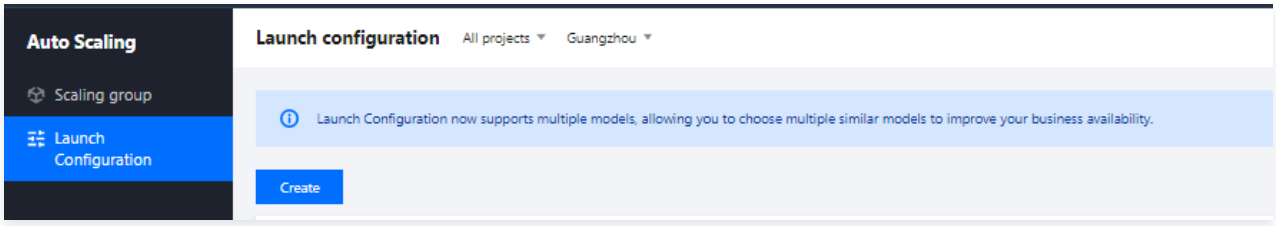
The launch configuration delineates the configuration information for the Cloud Virtual Machine (CVM) instances used for Auto Scaling, encompassing the CVM's image, storage, network, security group, login method, and other configuration details.

> **Note:**
> The creation of the launch configuration service **is entirely complimentary**; please feel free to create one.

Instructions

Select a region

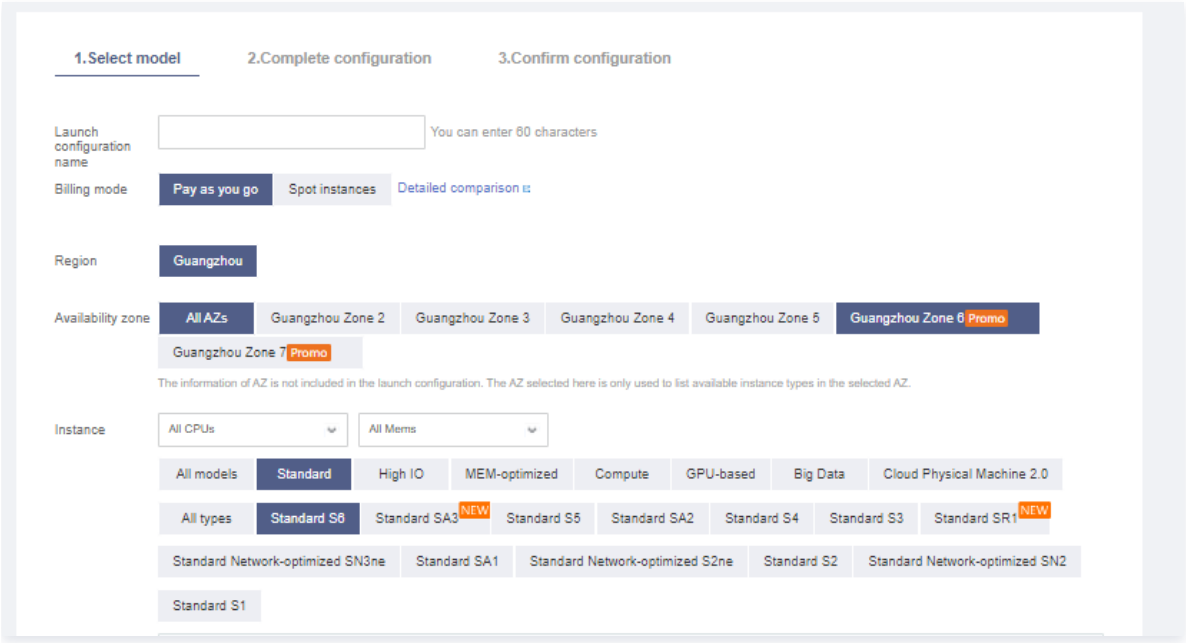
1. Log in to the Auto Scaling console and select Launch Configuration from the left navigation bar.
2. At the top of the **Launch Configuration** page, select the project and region for the launch configuration, as shown in the figure below:



   The selection of the region restricts the instances that can be manually added and the load balancers that can be bound. For example, if the Guangzhou region is selected for the launch configuration, the instances automatically added to the scaling group will be from Guangzhou. A scaling group in the Guangzhou region cannot manually add instances from other regions such as Shanghai, Beijing, Hong Kong, Toronto, etc., nor can it bind load balancers from these regions.
3. Click on **New** to navigate to the "Create Launch Configuration" page.

Select a model

In the **Select Model** step, refer to the following information to set the launch configuration name, availability zone, and model, as shown in the figure below:

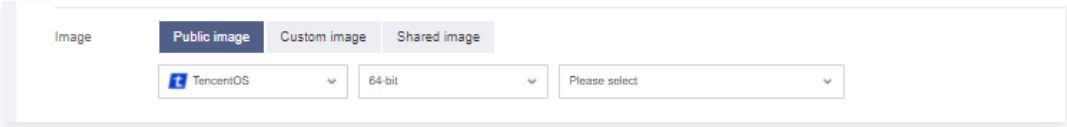


- **Launch Configuration Name**: Customize the name of the launch configuration.

- **Billing Mode**: Supports both Pay-as-you-go and Bidding Instance modes.
- **Availability Zone, Model**: Select the model of the instance you wish to bind with the scaling group.

Select Image, Storage, and Bandwidth

1. When creating a launch configuration, you can use public images, custom images, shared images, or images from the marketplace. For more details, please refer to Image Overview. As shown in the figure below:

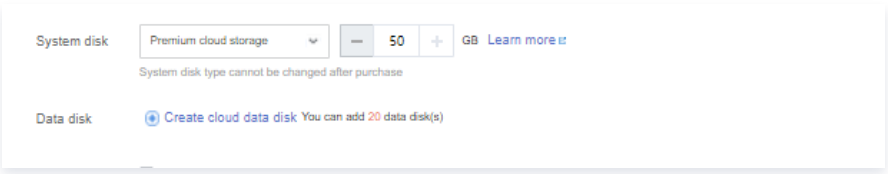


It is recommended to use custom images with pre-deployed environments for the following reasons:

- If you choose a **public image**, the instances scaled out will be a pure OS, and the application environment will still need to be manually deployed.
- If you opt for a **custom image**, by creating an image from a CVM instance with a pre-deployed environment, and then using this image to create CVM instances in bulk, the newly created instances will possess a software environment consistent with the previous CVM instance, thereby achieving the purpose of bulk deployment.

> **Note:**
> For instructions on how to create an image of the "instance expected to be bound to the scaling group", please refer to Creating Custom Images.

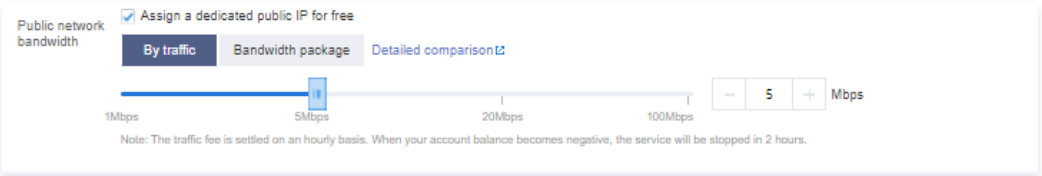
2. Refer to the information below to set up the disk in the launch configuration, as shown in the figure:



If a cloud disk is selected for the system disk, then a data disk snapshot can be chosen for the data disk:

- For users with a large amount of data, data disks are often used for storage. When a snapshot file is created from data disk A, users can utilize this snapshot file to rapidly clone multiple disks, thereby achieving the purpose of swift server deployment.
- When Auto Scaling automatically adds new CVM instances, if a data disk snapshot is specified in the launch configuration data disk, in conjunction with the cloud disk, it can support the ability to automatically mount a data disk containing set data after the CVM instance is launched, thereby fulfilling the requirement for automatic data copying.
- 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. Prior to setting up Auto Scaling, you need to operate on the original instance that created the data disk snapshot, enabling it to support auto-mounting of the data disk when launching new CVM instances. For more details, please refer to Auto-Mounting.

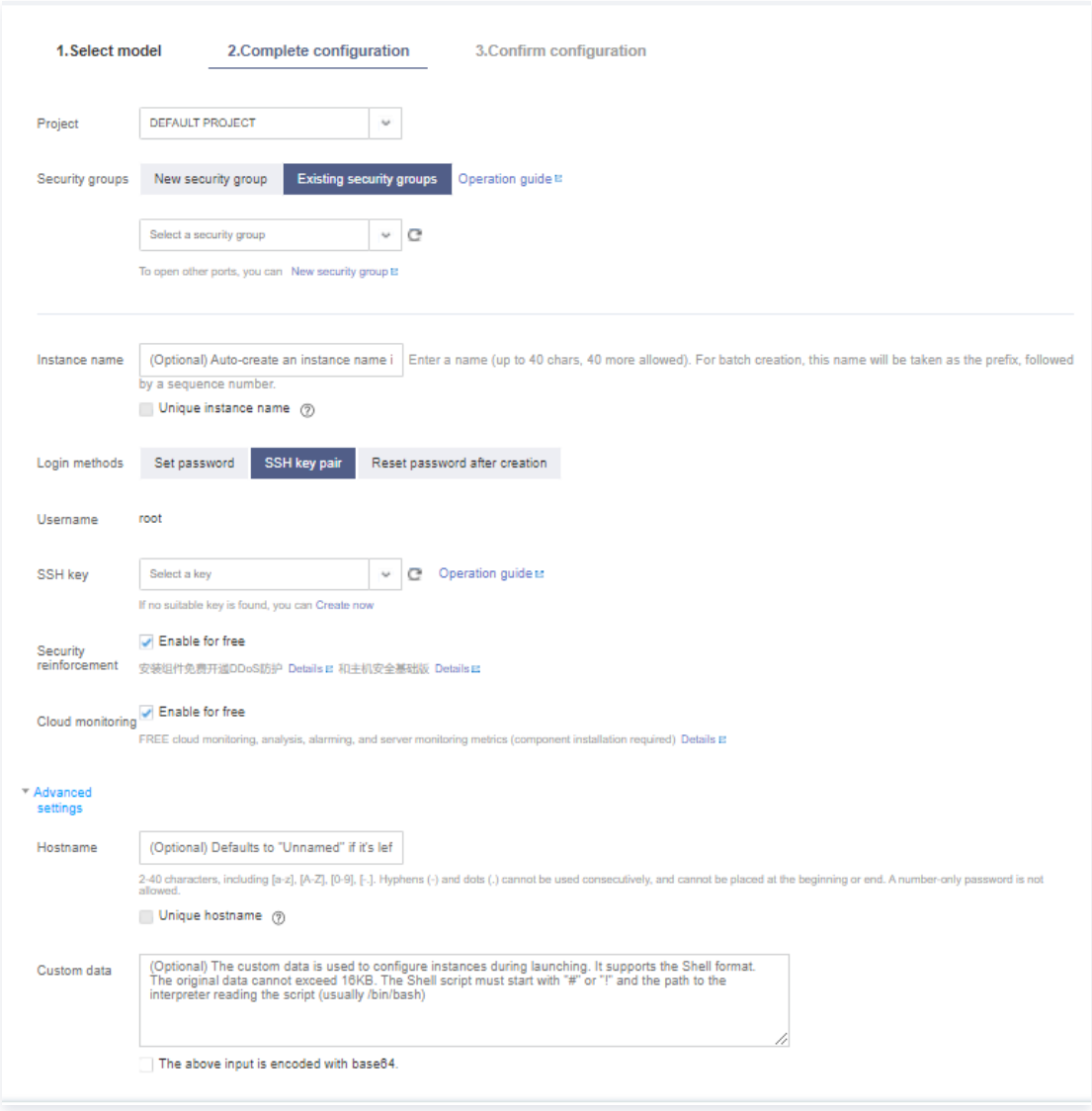
3. By default, an independent public IP is allocated free of charge. Please choose the network billing mode according to the actual situation, as shown in the figure:



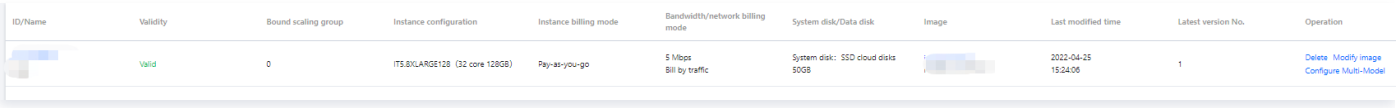
> **Note:**
> The Auto Scaling service is free of charge. Additional CVMs, disks, and networks will be billed on a pay-as-you-go basis for the CVM instances, disks, and networks. The pricing will be displayed on this page according to your settings.

Configuration Information

1. In the **Set Up Host** step, select the login method and security group. The CVM instances added through the Auto Scaling service enjoy cloud security and cloud monitoring services by default, free of charge. As shown in the figure:



2. Once the configuration is confirmed and successfully created, you can view the created launch configuration on the **Launch Configuration** page, as depicted in the figure:



Step 2:Creating a Scaling Group

Last updated: 2024-01-17 17:52:44

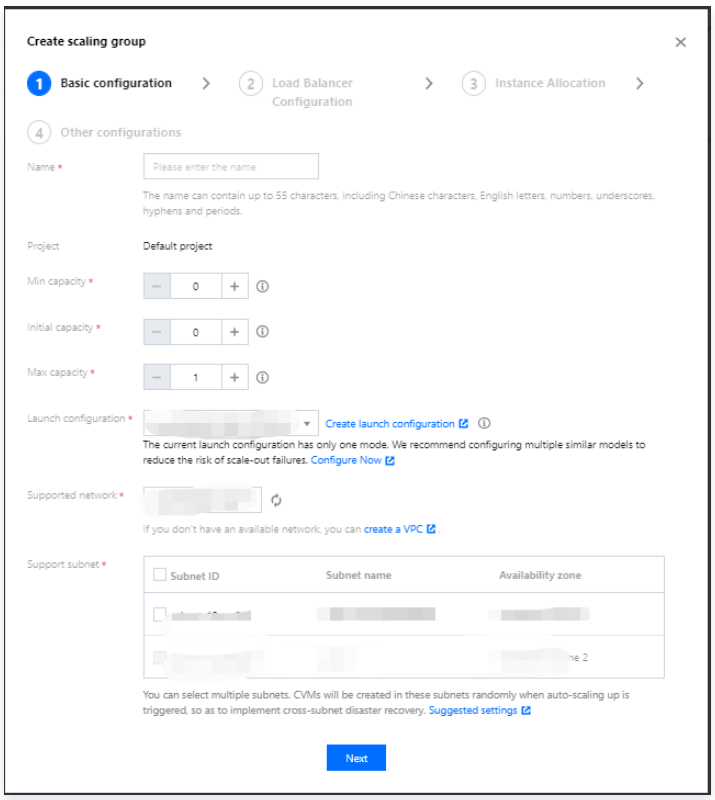
Scenario

An auto scaling group is a collection of cloud server instances that adhere to the same rules and cater to the same scenario. This document outlines the process of creating an auto scaling group via the Auto Scaling Console.

Instructions

Create scaling group

1. Log into the Elastic Services Console and select **Auto Scaling Group** from the left navigation bar.
2. On the **Auto Scaling Group** management page, click **Create**.
3. On the pop-up **Create Scaling Group** page, refer to the following information to fill in the basic information of the auto scaling group. Fields marked with * are mandatory. As shown in the figure below:

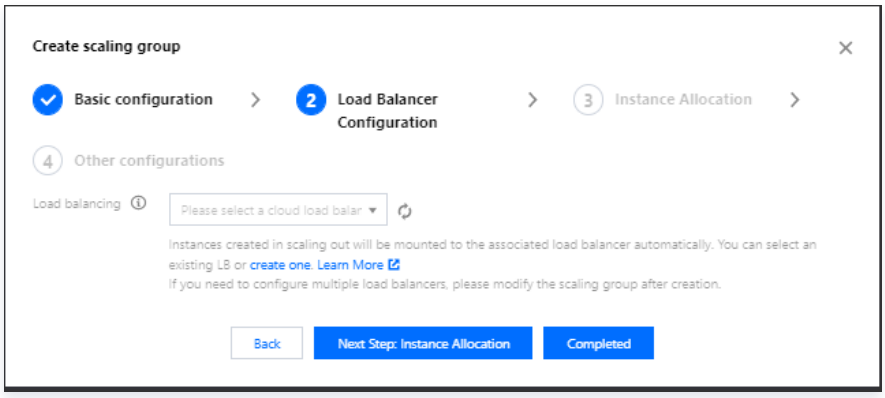


   - **Name**: Customize the name of the auto scaling group.
   - **Minimum Scaling**: The minimum number of instances allowed in the scaling group.
   - **Initial Instance Count**: The number of instances when the auto scaling group is first created. The auto scaling group will automatically create the corresponding number of instances for you.
   - **Maximum Scaling**: The maximum number of instances allowed in the scaling group.

     > **Note:**
     > The current number of CVM instances in the auto scaling group will be maintained between the minimum and maximum scaling numbers.

   - **Launch Configuration**: Specifies the created launch configuration. Expansion machines will be created according to this configuration during scaling.
   - **Supported Networks and Availability Zones**: Select networks and availability zones as needed.
4. Click **Next**.

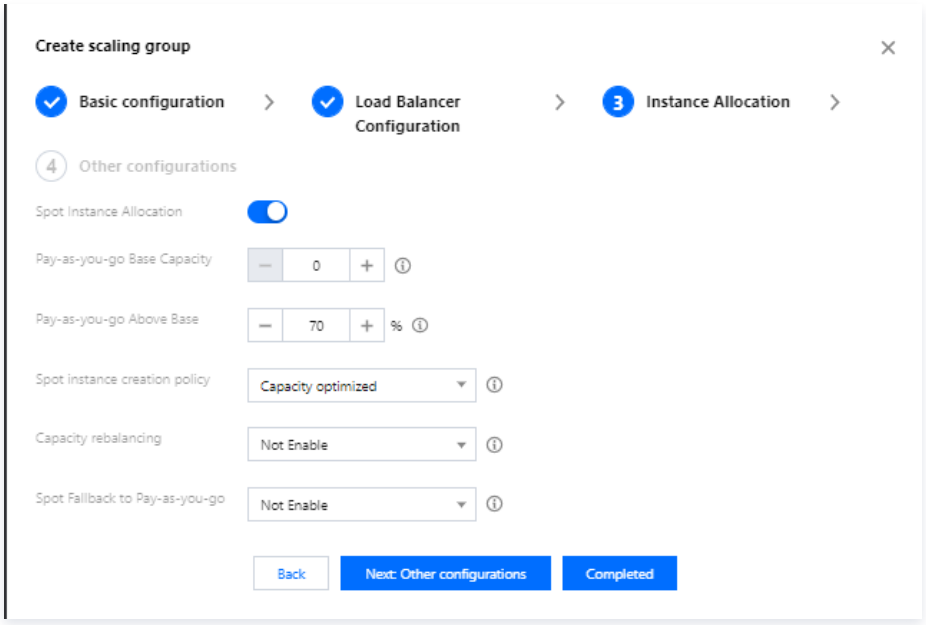
5. (Optional) In the **Load Balancing Configuration** step, choose to associate an existing load balancing policy or create a new load balancer, then click **Next: Spot Instance Allocation**. As shown in the figure below:



6. (Optional) In the **Spot Instance Allocation** step, configure the spot instance allocation strategy. You may also click **Next: Additional Configurations** to skip this step.

> **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.

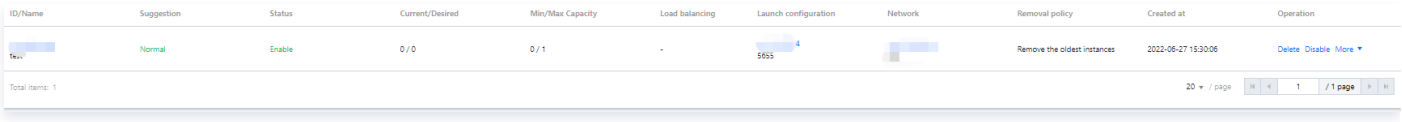
Enable the "Use Spot Instances" switch. Once enabled, it appears as shown in the figure below:



- **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.

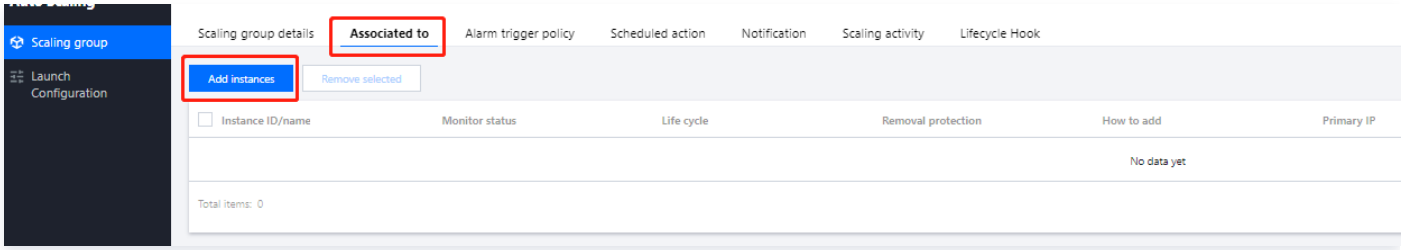
- **Spot Instance Reclamation Monitoring**: When enabled, Auto Scaling will attempt to proactively replace the spot instances in the scaling group that are about to be reclaimed with new instances, thereby helping you maintain the number of instances within the scaling group and the ratio of on-demand 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.

7. In the "Additional Configuration" step, refer to the following information to set the removal strategy and instance creation strategy.
   - **Removal Strategy**: When the scaling group needs to reduce instances and there are multiple choices, the instance to be removed will be selected based on the removal strategy. Supports "Remove the Oldest Instance" and "Remove the Newest Instance".
   - **Instance Creation Strategy**:
     - **Preferred Availability Zone (Subnet) Priority**: Based on the configured order of availability zones (subnets), it prioritizes the earlier configuration items, and automatically retries in order upon failure. This is suitable for architectures that primarily rely on a particular availability zone, with other zones serving as auxiliary.
     - **Multi-Availability Zone (Subnet) Dispersion**: The system will select the relatively fewer availability zones (subnets) to create new instances based on the distribution of instances in different availability zones (subnets) within the scaling group during expansion. This is suitable for architectures that require evenly distributed instances.
8. Click **Complete** to finish creation. The created scaling group can be viewed on the "Scaling Group" page, as shown in the image below:



Add Instance (Optional)

1. On the Scaling Group page, select the scaling group ID to enter the details page of that scaling group.
2. Select the **Associate Instance** tab, and click on **Add Instance**, as depicted in the image below:



3. In the pop-up **Add Instance** window, select the instance that needs to be bound, and click **Confirm**.

> **Note:**
> If you encounter situations where you cannot add or remove instances, please check the maximum and minimum scaling numbers set in the scaling group.

Step 3:Creating a Scaling Policy

Last updated：2024-01-17 17:52:54

Scenario

The Auto Scaling Group adjusts the quantity of cloud servers based on the scaling strategy:

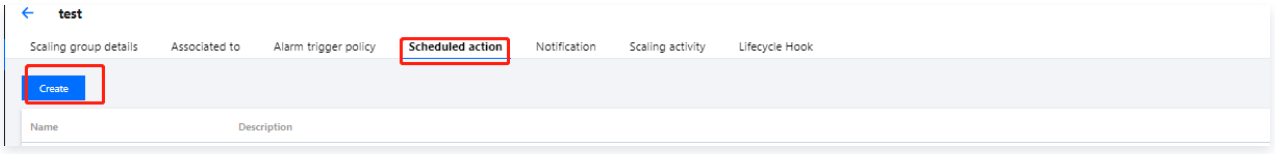
- Create a **Scheduled Task** for the timely execution of scaling activities, with the option to set whether it is to be performed periodically.
- Establish a **Alarm Trigger Strategy**, executing scaling activities based on metrics from Tencent Cloud's Observable Platform (such as CPU, memory usage, etc.).

Instructions

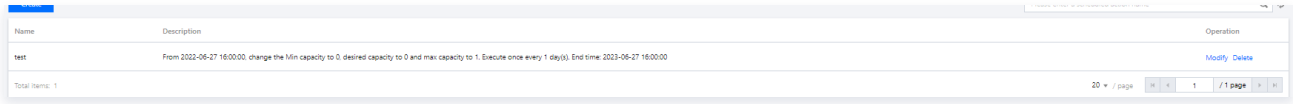
Creating a Scheduled Task

If your load variations are predictable, you can plan your device expansion activities by setting up scheduled tasks. This feature allows for the timely and periodic automatic addition or reduction of CVM instances, thereby flexibly responding to business load changes, enhancing device utilization, and saving on deployment and instance costs.

1. On the Scaling Group page, select the scaling group ID to enter the details page of that scaling group.
2. Select the **Scheduled action** tab and click on **Greate**, as shown in the image below:



3. In the "Create New Scheduled Task" pop-up window, specify information such as the scheduled task name, scaling group activity, and repetition cycle.
4. After completing the settings, click on **Confirm** to view the scheduled task, as depicted in the image below:

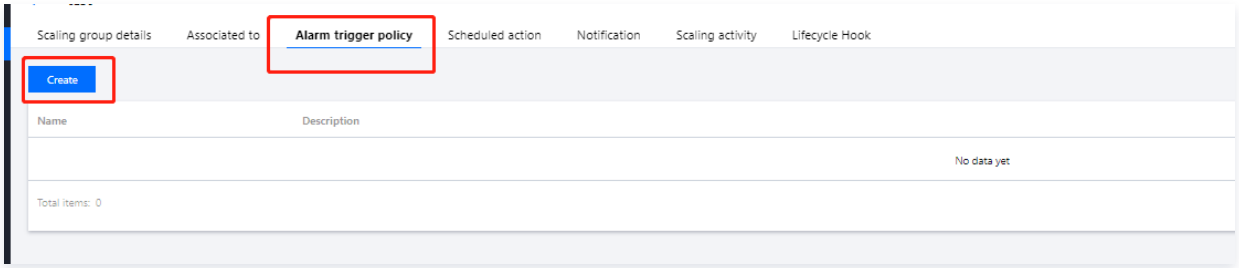


Creating an Alarm Trigger Strategy

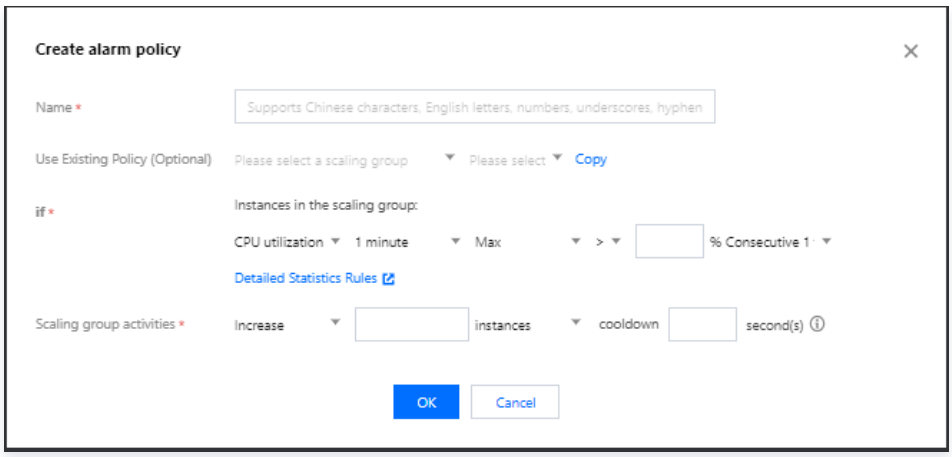
If you wish to adjust your business deployment based on CVM metric conditions, you can plan your device expansion activities through a custom alarm trigger strategy. When the business load causes the metrics to reach a threshold, this strategy will assist you in automatically increasing or decreasing the number of CVM instances. This allows for flexible response to changes in business load, enhances device utilization, and saves on deployment and instance costs.

> **Note:**
> - Upon the creation of a scaling group, a default ping unreachable alarm trigger strategy is established to replace unhealthy sub-machines.
> - Before utilizing the alarm trigger strategy, it is necessary to install the new version of Tencent Cloud's Observable Platform Agent in the CVM image. For more details, please refer to Installing Monitoring Components.

1. On the Scaling Group page, select the scaling group ID to enter the details page of that scaling group.
2. Select the **Alarm Trigger Strategy** tab and click on **Greate** As shown in the figure below:



3. In the pop-up **New Alarm Trigger Strategy** window, set up to automatically add or subtract a specified number or percentage of CVM instances for the scaling group based on Tencent Cloud's Observable Platform performance metrics (such as CPU, memory, bandwidth, etc.).
   You can also directly copy existing strategies from an existing scaling group to the current scaling group through **Copy Strategy (optional)**. As shown in the figure below:



4. After completing the settings, click **Confirm** to view the alarm trigger strategy. As shown in the figure below: