

A decorative graphic on the left side of the page, consisting of a network of interconnected nodes and lines, resembling a web or a complex data structure. The nodes are represented by small circles of varying sizes, and the lines are thin, light blue lines connecting them. The overall shape is roughly triangular, pointing towards the top right.

# Auto-Scaling Tutorial

Version: ZStack 3.10.0

Issue: V3.10.0

# Copyright Statement

---

Copyright © 2020 Shanghai Yunzhou Information and Technology Ltd. All rights reserved.

Without its written consent, any organization and any individual do not have the right to extract, copy any part or all of, and are prohibited to disseminate the contents of this documentation in any manner.

## Trademark

Shanghai Yunzhou Information and Technology Ltd. reserves all rights to its trademarks, including , but not limited to ZStack and other trademarks in connection with Shanghai Yunzhou Information and Technology Ltd.

Other trademarks or registered trademarks presented in this documentation are owned or controlled solely by its proprietaries.

## Notice

The products, services, or features that you purchased are all subject to the commercial contract and terms of Shanghai Yunzhou Information and Technology Ltd., but any part or all of the foregoing displayed in this documentation may not be in the scope of your purchase or use. Unless there are additional conventions, Shanghai Yunzhou Information and Technology Ltd. will not claim any implicit or explicit statement or warranty on the contents of this documentation.

In an event of product version upgrades or other reasons, the contents of this documentation will be irregularly updated and released. Unless there are additional conventions, this documentation, considered solely as a using manual, will not make any implicit or explicit warranty on all the statements, information, or suggestions.

# Contents

---

<b>Copyright Statement.....</b>	<b>I</b>
<b>1 Overview.....</b>	<b>1</b>
<b>2 Preparation.....</b>	<b>5</b>
<b>3 Quick Start.....</b>	<b>6</b>
<b>4 Auto Scaling Group.....</b>	<b>7</b>
<b>5 Typical Scenario Practice.....</b>	<b>28</b>
<b>Glossary.....</b>	<b>37</b>

# 1 Overview

---

ZStack offers auto-scaling capabilities that let you automatically add or remove VM instances from an auto scaling group (ASG) in response to load balancing of VM instances, your business load changes, and predefined scaling policies. With the auto scaling service, you can better leverage the Cloud resources, reduce the O&M costs, and ensure smooth business operations. Currently, the auto scaling service is applicable to KVM VM instances.

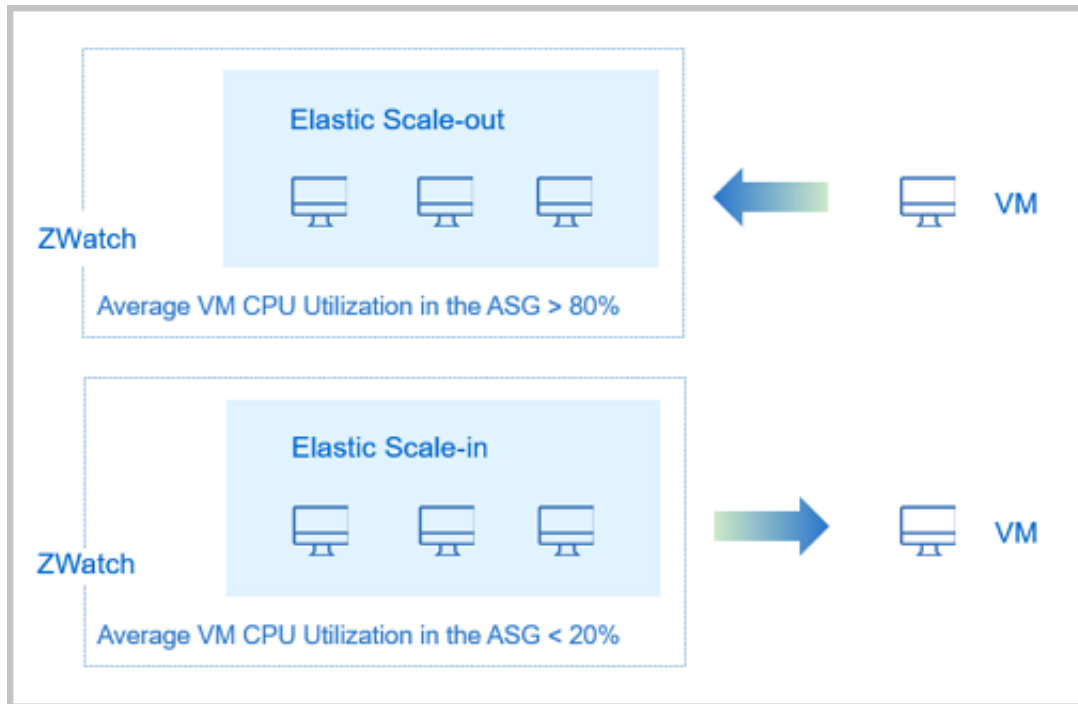
## Scaling Mode

Currently, the Cloud supports the following two types of scaling mode:

### 1. Auto Scaling

- Auto scaling includes elastic scale-out and elastic scale-in. For the elastic scale-out, when your businesses are growing, VM instances will be automatically added to ensure your business continuity. For the elastic scale-in, if your businesses decrease, VM instances will be automatically reduced.
- With the ZWatch monitoring alarm, the auto scaling mode can be triggered. You can select an endpoint type as needed, including email, DingTalk, HTTP application, short message service (SMS), Microsoft Teams, and system endpoint.

The auto scaling mode is shown in [Auto Scaling](#).

**Figure 1-1: Auto Scaling**

## 2. Elastic Self-Health

- In the elastic self-health mode, an auto scaling group monitors the health status of the VM instances within the group, automatically removes the unhealthy VM instances, and automatically adds additional VM instances. In this regard, healthy VM instances within the group will be ensured to be automatically adjusted not lower than the minimum specified number of VM instances.
- Two types of health check are provided to trigger the elastic self-health, including load balancer health check and VM health check. If an auto scaling group configures the load balancing service, we recommend that you select the health check mechanism native to a load balancer.

The elastic self-health is shown in [Elastic Self-Health](#).

**Figure 1-2: Elastic Self-Health**

### Trigger Mechanism for Auto Scaling Group

The following illustrations show the scale-out trigger conditions:

1. When the workload of VM instances in an auto scaling group exceeds the specified thresholds, the scale-out policy is triggered.
  - If the auto scaling group starts to calculate the cooldown time, a scale-out activity is suspended.
  - If the number of the VM instances in the auto scaling group has reached the specified maximum capacity limit, additional VM instances are no longer added.
  - The number of the VM instances in the auto scaling group must not be greater than the specified maximum number of the VM instances.
2. When the number of the VM instances in an auto scaling group is less than the specified minimum number of the VM instances, the scale-out policy is triggered.
  - If the number of the VM instances in the auto scaling group has reached the specified maximum capacity limit, additional VM instances are no longer added.
  - The number of the VM instances in the auto scaling group must not be greater than the specified maximum number of the VM instances.

The following illustrations show the scale-in trigger conditions:

1. When the workload of the VM instances in an auto scaling group is less than the specified thresholds, the scale-in policy is triggered.
  - If the auto scaling group starts to calculate the cooldown time, a scale-in activity is suspended.

- If the number of the VM instances in the auto scaling group has reached the specified minimum capacity limit, VM instances are no longer removed.
  - The number of the VM instances in the auto scaling group must not be lower than the specified minimum number of the VM instances.
2. When the number of the VM instances in an auto scaling group exceeds the specified minimum number of the VM instances, the scale-in policy is triggered.
- If the number of the VM instances in the auto scaling group has reached the specified minimum capacity limit, VM instances are no longer removed.
  - The number of the VM instances in the auto scaling group must not be lower than the specified minimum number of VM instances.

Trigger conditions for elastic self-health: unhealthy VM instance

- Notice that the Cloud will delete VM instances detected as unhealthy. After these VM instances are deleted, if the number of the VM instances in the auto scaling group is less than the specified minimum number of the VM instances, the scale-out policy is triggered. Hence, VM instances will be automatically added.

### Typical Usage Scenario

Three types of typical usage scenario about the auto-scaling feature are introduced as follows:

- Elastic scale-out:

An online retailer has initiated a sales promotion such as red envelop fever and limited-time offer during Double Eleven, the Spring Festival, and other festivals and shopping sprees. As the business workload is skyrocketing, VM instances need to be automatically scaled out in time to avoid access delay and excessive resource loads.

- Elastic scale-in:

After the festivals and shopping sprees, the business workload of the retailer obviously declines. In that case, VM instances need to be automatically removed in time to avoid a waste of resources.

- Elastic self-health:

To ensure that the core businesses of the retailer are running normally, the number of the healthy VM instances must not be lower than some threshold.

## 2 Preparation

---

Make sure that admins install the latest ZStack in advance, and deploy the necessary resources used for creating VM instances. For more information, see related installation and deployment topics in [User Guide](#).

This Tutorial describes the usage of the autoscaling service in detail.



## 3 Quick Start

---

Create an auto-scaling service.

### Procedure

1. Create an auto scaling group. For more information, see [Auto Scaling Group](#).
2. Manage the auto scaling group. For more information, see [Auto Scaling Group](#).

## 4 Auto Scaling Group

---

An auto scaling group contains a logical collection of VM instances that share the same usage scenarios for the purposes of automatic scaling and management. With an auto scaling group, auto scaling or elastic self-health services can be automatically achieved based on business changes.

An auto scaling group supports the following operations:

- Create the auto scaling group.
- Check the details of the auto scaling group.
- Enable the auto scaling group.
- Disable the auto scaling group.
- Delete the auto scaling group.

### Create Auto Scaling Group

In the navigation pane of the ZStack Private Cloud UI, choose **Resource Pool > Auto Scaling Group**. On the **Auto Scaling Group** page, click **Create Auto Scaling Group**. On the displayed **Create Auto Scaling Group** page, create an auto scaling group.

To create an auto scaling group, follow these steps:

1. Set the basic information. To set the basic information, set the following parameters:
  - **Group Name:** Enter a name for the auto scaling group.
  - **Description:** Optional. Enter a description for the auto scaling group.
  - **Minimum Group Size:** Set the minimum number of VM instances in the auto scaling group.
    - When an elastic scale-in policy is triggered, the number of the VM instances in the auto scaling group must not be smaller than that of the specified minimum number of the VM instances.
    - Enter an integer. Value range: 1-1000, inclusive. Make your configurations as needed.
  - **Maximum Group Size:** Set the maximum number of VM instances in the auto scaling group.
    - When an elastic scale-out policy is triggered, the number of the VM instances in the auto scaling group must not be greater than that of the specified maximum number of the VM instances.
    - Enter an integer. Value range: 1-1000, inclusive. Make your configurations as needed.

- **Desired Group Size:** Set the number of VM instances that the auto scaling group attempts to maintain.
  - If you create an auto scaling group for the first time, the number of VM instances in the auto scaling group is equal to the desired capacity of VM instances.
  - Enter an integer. Value range: 1-1000, inclusive. The desired capacity of VM instances must be set between the minimum number of the VM instances and the maximum number of VM instances. Make your configurations as needed.
- If you choose to configure the load balancing service, we recommend that you set the following parameters:
  - **Load Balancer:** Select a load balancer.
    - Ensure that you create a load balancer in advance, and bind one or more listeners to it.
    - For more information about how to use the load balancing service, see [Load Balancing](#) in the [User Guide](#).
  - **Listener:** Select a listener after you choose the load balancer.
    - The listener list displays all listeners bound by the load balancer.
    - If you select multiple listeners, the same collection of the VM instances within the corresponding listeners will be listened separately via different ports.
  - **L3 Network:** Select an L3 network used for creating VM instances.
    - If you do not use the load balancing service, you can select a VPC network, vRouter network, or flat network. Specifically, the VPC network must be attached to a VPC vRouter.
    - If you use the load balancing service, you can select networks available to the service.
      - If the load balancer that you selected provides the load balancing service by using the VIP created from a public network, the following three scenarios are supported:
        - Scenario 1: Assume that the L3 network is a VPC network attached to the same VPC vRouter. In that case, make sure that the L3 network that is created from the VPC vRouter offering and the public network where the VIP resides are the same.
        - Scenario 2: Assume that the L3 network is a vRouter network. In that case, make sure that this vRouter network that attaches a vRouter offering and the public network where the VIP resides are the same.

- Scenario 3: Assume that the L3 network is a flat network. In that case, make sure that the flat network that attaches a vRouter offering and the public network where the VIP resides are the same.
- If the load balancer that you selected provides the load balancing service by using the VIP created from a VPC network, you can select the VPC network attached by the same VPC vRouter. In that case, the VPC network that you selected and the VPC network that provides the VIP must be attached to the same VPC vRouter.
- If the load balancer that you selected provides the load balancing service by using the VIP created from a flat network, the following two scenarios are supported:
  - Scenario 1: The L3 network is a flat network that creates a VIP. In that case, the flat network with a vRouter offering is supported.
  - Scenario 2: The L3 network is other types of flat network. In that case, make sure that the flat network that attaches a vRouter offering and the flat network where the VIP resides are the same.

**Note:**

- If the selected listener has bound a VM NIC, this L3 network and the network attached by the listener must belong to the same router.

— **Health Check:** We recommend that you select health check for the load balancer.

- Load balancer health check is the health check mechanism native to the load balancer.
- For more information about the details of the health check mechanism on the load balancer, see [Load Balancing](#) in the [User Guide](#).

— **Health Check Grace Period:** Set the health check grace period after you select the load balancer health check.

- Health check grace period is a period of time after the VM instances in the auto scaling group are created and booted. During this period of time, application services related to the VM instances are probably still booting, and the auto scaling group will not perform health checks for the load balancer. If this period of time is exceeded, the health status of the VM instances are listened based on the health check mechanism of the load balancer.

- Enter an integer that is greater than 10. Unit: second | minute | hour. Make your configurations as needed, as shown in the [\(Recommended\) Auto Scaling Group Configures Load Balancing](#).

**Figure 4-1: (Recommended) Auto Scaling Group Configures Load Balancing**

The screenshot shows a configuration interface for an Auto Scaling Group. It includes the following sections:

- Load Balancer:** A text input field that is currently empty, with a minus button to its right.
- Listener \*:** A list of listeners. It contains 'Listener-2' and 'Listener-1', each with a minus button. Below them is an empty input field with a plus button to add a new listener.
- L3 Network \*:** A dropdown menu showing 'L3-Private Network-vRouter' with a minus button to its right.
- Health Check \*:** A dropdown menu showing 'Load Balancer' with a downward arrow to its right.
- Health Check Grace Period \*:** A text input field containing '300' and a unit dropdown menu set to 'second'.

- If you do not configure the load balancing feature, we recommend that you set the following parameters:
  - **Load Balancer:** Leave this field blank.
  - **Listener:** Leave this field blank.
  - **L3 Network:** Select a private L3 network.



**Note:**

Currently, our auto scaling service offers autoscaling capabilities that let you automatically add or delete VM instances in the scenarios such as vRouter network and VPC network.

— **Health Check:** Default to display the health check of the VM instances.

- VM health check enables you to check the health status of the VM instances in real time. If VM instances are detected as unhealthy (including stopped, unknown, and deleted), the unhealthy VM instances will be automatically removed, and new VM instances will be created. Hence, the number of the healthy VM instances in the auto scaling group are ensured to be not lower than that of the specified minimum number of the VM instances.

The following figure shows the scenario where load balancing is not configured for the auto scaling group, as shown in [Auto Scaling Group Do not Configure Load Balancing](#).

**Figure 4-2: No Load Balancing Configured for Auto Scaling Group**

The screenshot shows a configuration panel for an auto-scaling group. It contains four sections, each with a text input field and a circular icon with a plus or minus sign:

- Load Balancer:** An empty text field with a plus icon.
- Listener:** An empty text field with a plus icon.
- L3 Network \*:** A text field containing 'L3-Private Network-vRouter' with a minus icon.
- Health Check \*:** A dropdown menu showing 'VM Instance' with a downward arrow icon.

- **Enable alarm notification:** Select whether to enable the alarm notifications. If selected, the ZWatch monitoring alarm mechanism can be configured to trigger the auto scaling service for the auto scaling group.
  - By default, this checkbox is not selected. Associated alarm messages can be viewed at the Notification Center.
  - If selected, you must specify one or more endpoints.

**Endpoint:** Specify one or more endpoints.

- You can either select the default system endpoint or choose custom endpoints such as email, DingTalk, HTTP application, short message service, and Microsoft Teams.

- For more information about how to create endpoints, see [Endpoint](#) in the [User Guide](#).
- **Apply immediately after creation:** Select whether to enable the auto scaling group immediately after you create the auto scaling group. By default, this checkbox is not selected.

Step 1 is about how to set the basic information, as shown in [Step 1 Set Basic Information](#).

**Figure 4-3: Step 1 Set Basic Information**

Next(1/3) Cancel

Create Auto Scaling Group: Basic Configuration

Zone

ZONE-1

Group Name \*

Auto Scaling Group-Business A

Description

Minimum Group Size \*

5

Maximum Group Size \*

10

Desired Group Size \*

5

Load Balancer

Load Balancer

Listener \*

Listener-2

Listener-1

L3 Network \*

L3-Private Network-vRouter

Health Check \*

Load Balancer

Health Check Grace Period \*

300

second

Endpoint \*

System Endpoint

☒ Apply immediately after creation

2. Configure the auto-scaling VM instance. Auto-scaling configurations define the template configuration information of VM instances in the auto scaling group. To configure the information, set the following parameters:

- **VM Name:** Enter a name for the VM instance.
  - The unified naming convention for the VM instances in an auto scaling group is **asg -Auto Scaling Group name-VM name-the first 5-digit VM UUID**. Specifically, asg stands for auto scaling group.
- **VM Description:** Optional. Enter a description for the VM instance.



- **Instance Offering** : Select an instance offering for the VM instance.
- **Image**: Select an image for the VM instance.

**Note:**

- Under this scenario, you can add two types of VM image: qcow2 and raw.
  - If you want to use the internal monitoring metrics, select either an image that has installed an agent, or install the agent manually by using the User Data script.
  - If you change the image after you create the VM instance, the new image will only be effective for the subsequently created VM instances, and the original image remains unchanged for the previously created VM instances.
- **L3 Network**: Default to display the private L3 network that you have set in the previous step.
  - **Advanced**: Make advanced settings for the VM instance as needed.
    - **Data Disk Offering**: Select a data disk offering. The data disk offering can be created directly and then attached to the VM instance.
    - **Security Group**: Select a security group. VM instances in the auto scaling group will share the same security group rules.
    - **Console Password**: Enter the password (VNC password) for the VM console. Password length: 6-18.
    - **SSH Public Key**: Inject an SSH public key into the VM instance. By doing so, you can log in to the VM instance without passwords via SSH.
      - To inject an SSH public key into the VM instance, install cloud-init for the image in advance.
      - For more information about the details of the SSH public key, see [SSH Key Management](#) in the [User Guide](#).
    - **User Data**: Import User Data to the VM instance. By uploading custom parameters or a script, customize configurations or complete specific tasks for the VM instance.
      - To import User Data to a Linux VM instance, this Linux VM instance must install cloud-init in advance.

The following is a sample script of importing User Data to a Linux VM instance:

```
#cloud-config
users:
  - name: test
```

```

    shell: /bin/bash
    groups: users
    sudo: ['ALL=(ALL) NOPASSWD:ALL']
    ssh-authorized-keys:
      - ssh-rsa AAAAB3NzaC1lXCJfjroD1lT root@10-0-0-18
  bootcmd:
    - mkdir /tmp/temp
  write_files:
    - path: /tmp/ZStack_config
      content: |
        Hello,world!
      permissions: '0755'
  hostname: Perf-test
  disable_root: false
  ssh_pwauth: yes
  chpasswd:
    list: |
      root:word
    expire: False
  runcmd:
    - echo ls -l / >/root/list.sh

```

The preceding sample script can achieve the following functionalities:

1. When you create a VM instance, create a user with the name test by using ssh-key .
  2. Start the VM instance, and write the `/etc/hosts` file. Also, create the `/tmp/temp` directory, and create a file under the directory and write contents into the file.
  3. Set hostname, enable the user root, log in to the VM instance by using the SSH password, and change the password for the root.
  4. Run the `echo ls -l /` command.
- To import User Data to a Windows VM instance, this Windows VM instance must install Cloudbase-Init in advance. For more information about the detailed installation method, see [Cloudbase-Init Documentation](#).

The following is a sample script of importing User Data to a Windows VM instance:

```

#cloud-config
write_files:
  - encoding: b64
    content: NDI=
    path: C:\b64
    permissions: '0644'
  - encoding: base64
    content: NDI=
    path: C:\b64_1
    permissions: '0644'
  - encoding: gzip
    content: !!binary |
      H4sIAGUfoFQC/zMxAgCIsCQyAgAAAA==
    path: C:\gzip

```

```
permissions: '0644'
```

The preceding sample script can achieve the following functionalities: When the VM instance boots, three files: **b64**, **b64\_1**, and **gzip** are created in the C drive.

**Note:**

When you use User Data, note that you can configure only one L3 network for an L2 network.

Step 2 is about how to configure the autoscaling VM instance, as shown in [Step 2 Configure Autoscaling VM Instance](#).

**Figure 4-4: Step 2 Configure Autoscaling VM Instance**

Advanced ^

Data Disk Offering

Data Disk Offering

Security Group

Security Group

Console Password

.....

SSH Public Key

ssh-rsa AAAAB3Nza1C1yc2EAAAABIwAAAQEkIOUpk

User Data

#cloud-config
users:
- name: test
 shell: /bin/bash


**Note:**

Exercise caution. If you delete resources (such as instance offering, image, and network) in the template configurations, you will fail to create the auto scaling group.

### 3. Configure the scaling policy. A scaling policy includes scale-out and scale-in.

- Scale-out policy:
  - When businesses are growing, the auto scaling group automatically adds additional VM instances to avoid access latency and excessive resource load.
  - An elastic scale-out policy can be triggered by the ZWatch monitoring mechanism that you set.

For example, when the average memory utilization of all VM instances in the auto scaling group is detected to continuously exceed 80% within a period of time, an appropriate number of VM instances will be automatically created. Hence, the auto scaling group will regain the reasonable load balancing.

To configure the scale-out policy, set the following parameters:

- **Trigger Metric:** Select a trigger metric, including average VM CPU utilization and average VM memory utilization.
  - Average VM CPU utilization: the sum of the utilization for a single VM CPU in an auto scaling group/the total number of VM instances in the auto scaling group
  - Average VM memory utilization: the sum of the memory utilization for a single VM instance in an auto scaling group/the total number of VM instances in the auto scaling group
  - Average VM CPU utilization (install agent): the sum of the utilization for a single VM CPU in an auto scaling group/the total number of VM instances in the auto scaling group
  - Average VM memory utilization (install agent): the sum of the memory utilization for a single VM instance in an auto scaling group/the total number of VM instances in the auto scaling group

**Note:**

- We recommend that you use an agent, internal monitoring tool, to monitor the average memory utilization of VM instances. By doing this, the monitoring data is more accurate.
  - If you must select trigger metrics that require you to install an agent, make sure that you select an image with an installed agent when you create VM instances.
  - Linux VM instances enable you to install agents by using the User Data script. For more information, see the **User Data** part in this topic.
  - If you did not install an agent for internal monitoring and still wanted to select trigger metrics that require you to install the agent, the auto scaling group would fail to take effect.
- **Trigger Condition:** Set a trigger condition. Options: > | ≥.
    - Enter an integer between 1-100, inclusive. Unit: %. Make your configurations as needed.
  - **Duration:** Set the duration.
    - Enter an integer that is greater than 0. Unit: second | minute | hour. Make your configurations as needed.
  - **Cooldown Time:** Set the cooldown time.

- Cooldown time refers to a period of time during which an auto scaling group that is in the locked status rejects any new scaling activity after one scaling activity is launched successfully in the auto scaling group.
- Enter an integer that is greater than 0. Unit: second | minute. Make your configurations as needed.
- **VMs To Be Added Per Time:** Add more VM instances when the auto scaling group performs a scale-out activity.

**Note:**

Each time the minimum allowed scale-out number of VM instances is 1. If the value is too large in this field, the scale-out activity will fail.

The scale-out policy can be configured, as shown in [Configure Scale-out Policy](#).

**Figure 4-5: Configure Scale-out Policy**

Scale-out Policy

Trigger Metric \*

VM Instance Memory Average Used In Percent

Trigger Condition \*

>
80
%

Duration \*

180
second

Cooldown Time \*

300
second

VMs To Be Added Per Time \*

2

- Scale-in policy:
  - When businesses are declining, the auto scaling group automatically deletes VM instances to avoid a waste of resources.

- An elastic scale-in policy can be triggered by the ZWatch monitoring mechanism that you set.

For example, when the average memory utilization of all VM instances in the auto scaling group is detected to continuously drop below 20% within a period of time, an appropriate number of VM instances will be automatically removed. Hence, the auto scaling group will regain the reasonable load balancing.

To configure the scale-in policy, set the following parameters:

- **Trigger Metric:** When you set the scale-in policy, the trigger metric is unselected in which this trigger metric is consistent with that of the scale-out policy.
- **Trigger Condition:** Set a trigger condition. Options: < | ≤.
  - Enter an integer between 1-100, inclusive. Unit: %. This trigger condition cannot be in conflict with that of the scale-out policy. Make your configurations as needed.
- **Duration:** Set the duration.
  - Enter an integer that is greater than 0. Unit: second | minute | hour. Make your configurations as needed.
- **Cooldown Time:** Set the cooldown time.
  - Cooldown time refers to a period of time during which an auto scaling group that is in the locked status rejects any new scaling activity after one scaling activity is launched successfully in the auto scaling group.
  - Enter an integer that is greater than 0. Unit: second | minute. Make your configurations as needed.
- **Removal Policy:** Select a removal policy, including:
  - Most recent created VM instance (default): When the auto scaling group starts performing scale-in activities, the latest created VM instances will be removed successively at first.
  - Earliest created VM instance: When the auto scaling group starts performing scale-in activities, the earlier created VM instances will be removed successively at first.
  - VM instance with the minimum CPU utilization: When the auto scaling group starts performing scale-in activities, the VM instances with the minimum CPU utilization will be removed successively at first.

- VM instance with the minimum memory utilization: When the auto scaling group starts performing scale-in activities, the VM instance with the minimum memory utilization will be removed successively at first.
- **VMs To Be Removed Per Time:** Remove VM instances when the auto scaling group performs a scale-in activity.

**Note:**

Each time the minimum allowed scale-in number of VM instances is 1. If the value is too large in this field, the scale-in activity will fail.

The scale-in policy can be configured, as shown in [Configure Scale-in Policy](#).

**Figure 4-6: Configure Scale-in Policy**

### Scale-in Policy

Trigger Metric \*

VM Instance Memory Average Used In Percent

Trigger Condition \*

<

20

%

Duration \*

180

second

Cooldown Time \*

300

second

Removal Policy \*

Most Recent Created VM Instance

VMs To Be Removed Per Time \*

2

Step 3 is about how to configure the auto scaling policies, as shown in [Step 3 Configure Auto Scaling Policy](#).



**Figure 4-7: Step 3 Configure Auto Scaling Policy**

PreviousOKCancel

Create Auto Scaling Group: Configure Scaling Policy

Scale-out Policy

Trigger Metric \*

VM Instance Memory Average Used In Percent ▾

Trigger Condition \*

> ▾80%

Duration \*

180second ▾

Cooldown Time \*

300second ▾

VMs To Be Added Per Time \*

2

### Scale-in Policy

Trigger Metric \*

VM Instance Memory Average Used In Percent

Trigger Condition \*

< ▾

20

%

Duration \*

180

second ▾

Cooldown Time \*

300

second ▾

Removal Policy \*

Most Recent Created VM Instance

VMs To Be Removed Per Time \*

2

## Check Auto Scaling Group Details

On the **Auto Scaling Group** page, select an auto scaling group, expand its details page, and then check the state and information of the currently created auto scaling group, including basic attributes, VM instance, scaling record, and audit.

- Basic information:
  - Display the current state, name, description, basic information (such as the minimum number of VM instances, the maximum number of VM instances, the number of the current VM instances, the desired number of the VM instances, health check mechanism, load balancing, and endpoints that push alarm notifications), scaling configurations (the template configuration information of VM instances in the auto scaling group), and scaling policy (such as scale-out and scale-in policies).
  - Specifically, you can change the name, description, scale-out policy, and scale-in policy for the auto scaling group.
- VM instance:

- Display a list of the currently healthy VM instances in the auto scaling group.
- The auto scaling group monitors the health state of VM instances based on the health check mechanism. We recommend that you use load balancer health check. If the unhealthy VM instances are detected, these unhealthy VM instances will be automatically deleted, and new VM instances will be created. Hence, the number of the healthy VM instances in the auto scaling group are ensured to be not lower than that of the specified minimum number of the VM instances.
- Display the monitoring status of the auto scaling group. The monitoring status has the following two types:
  - Collect normal: The auto scaling group normally collect the monitoring data of VM instances.
  - Insufficient data: The auto scaling group cannot normally collect the monitoring data of VM instances. The reasons may be as follows:
    1. For newly-created VM instances, collecting the monitoring data takes a while.
    2. For VM instances without an agent installed, the monitoring data cannot be collected.  
To collect the internal monitoring data of VM instances, install GuestTools by going to the details page of the VM instances.
    3. For abnormal VM instances, check the state of the VM instances.
- Scaling record: Display the scaling record of the scaling activities performed by the auto scaling group. You can search the scaling record by adjusting an appropriate time range.
- Audit: Check related operations of the auto scaling group.

### Enable/Disable Auto Scaling Group

- Enable auto scaling group: Enable the auto scaling group that is stopped.
- Disable auto scaling group: Disable the auto scaling group.



#### Note:

- Assume that the auto scaling group has triggered the scaling activity. If you disable the auto scaling group, the ongoing scaling activity will not be affected. After the scaling activity is completed, a new scaling activity will be rejected.
- Assume that the auto scaling group is being inspected. If you disable the auto scaling group, ZWatch or the health check mechanism will stop inspecting the auto scaling group immediately, and will reject a new scaling activity.

## Delete Auto Scaling Group

Exercise Caution. Deleting an auto scaling group will delete all VM instance in the auto scaling group as well.

### More Information

- Business applications running on the VM instances of an auto scaling group must be stateless and can be extensible horizontally.
- The auto scaling service will automatically release VM instances. We recommend that you do not manually attach volumes, NICs, or security groups to the VM instances in the auto scaling group. If the VM instances in the auto scaling group contain stateful information, associated data will be lost.
- The auto scaling group cannot be extended vertically. That is, instance offerings, network bandwidth, and other resources cannot be automatically scaled in or out.
- If you want to change an external monitoring metric to an internal monitoring metric, delete the corresponding auto scaling group and create a new one.
- The auto scaling service can be set in the global settings as follows:

- When the auto scaling group uses the health check mechanism of a load balancer, you can configure the health check interval of the VM instances in the load balancer.

Method: Go to **Settings > Global Settings > Advanced**, locate **Health checking interval of loadBalancing VM instance**, and set a value as needed. Default value: 10. Unit: Second. Minimum value: 10 seconds. Maximum value: 1000 seconds.

- When the auto scaling group uses the health check mechanism of a load balancer, you can configure the thread count of the VM health check in the load balancer.

Method: Go to **Settings > Global Settings > Advanced**, locate **Health checking threads of loadBalancing VM instance**, and set a value as needed. Default value: 10. Minimum thread count: 10. Maximum thread count: 1000.

- When the auto scaling group uses the health check mechanism of a VM instance, you can make configurations to delete the health check interval of unhealthy VM instances in the auto scaling group.

Method: Go to **Settings > Global Settings > Advanced**, locate **Interval for remove unhealthy instance in group**, and set a value as needed. Default value: 30. Unit: Second. Minimum value: 10 seconds. Maximum value: 1000 seconds.

- When the auto scaling group uses the health check mechanism of a VM instance, you can make configurations to delete the thread count of unhealthy VM instances in the auto scaling group.

Method: Go to **Settings > Global Settings > Advanced**, locate **Threads limitation for Unhealthy thread instance**, and set a value as needed. Default value: 10. Minimum thread count: 10. Maximum thread count: 1000.

- The health check interval of the VM count in the auto scaling group can be set.

Method: Go to **Settings > Global Settings > Advanced**, locate **Checking interval for instance count in group**, and set a value as needed. Default value: 20. Unit: Second. Minimum value: 10 seconds. Maximum value: 1000 seconds.

### Notice

- If an auto scaling group repeatedly performs the scaling policies, such as continuously creating and removing VM instances, the reasons may be as follows:
  - The newly-created VM instances cannot enter health status within the tolerance time. Then , the Cloud triggers the elastic self-health policy which deletes the unhealthy VM instances and add additional VM instances. Hence, an elastic self-health loop is formed. In that case , you need to check the health status or modify the health check mechanism of the VM instances.
  - The scale-in or scale-out thresholds are set inappropriately. For example, set a trigger condition where a scale-out activity is launched as the CPU of VM instances is below 40 %. If an auto scaling group has only one VM instance, the average CPU load of the VM instance in the auto scaling group is 60%. Then, two VM instances are added after the scale-out activity is triggered, and the average CPU workload of the VM instances in the auto scaling group reduces to 60%. Hence, a loop is formed. At this time, you need to set an appropriate threshold for the scaling policies.
- If an auto scaling does not perform the scaling policies but continuously triggers alarms, the reason may be as follows:
  - The maximum number of the VM instances and the scale-out trigger condition are set inappropriately. When the number of the VM instance in the auto scaling group has reached the specified maximum limit, and the average workload of the VM instances are still higher than the scale-out threshold, alarms will continuously be triggered. In that case, you need

to configure the appropriate thresholds for the maximum number of VM instances and the scale-out trigger condition.

## 5 Typical Scenario Practice

---

### Context

Scenario: Assume that an online retail has deployed a latest ZStack Private Cloud environment. To meet business needs, the retailer needs to deploy a business VM instance, and wants to use the auto scaling service where an auto scaling group provides scaling activities for the business VM instance based on the load balancing feature.

The auto scaling group enables you to trigger the auto scaling service by using an agent, internal monitoring tool. This scenario takes external monitoring as an example. The following is the detailed procedure to introduce the external monitoring.

### Procedure

1. Create an auto scaling group.
2. Verify functionalities: elastic self-health, elastic scale-out, and elastic scale-in.

### Procedure

1. Create an auto scaling group.

In the navigation pane of the ZStack Private Cloud UI, choose **Resource Pool > Auto Scaling Group**. On the **Auto Scaling Group** page, click **Create Auto Scaling Group**. On the displayed **Create Auto Scaling Group** page, create an auto scaling group.

Create an auto scaling group.

- a) Set the basic information.

To set the basic information, set the following parameters:

- **Group Name:** Enter a name for the auto scaling group, such as Auto Scaling Group-Business A.
- **Description:** Optional. Enter a description for the auto scaling group.
- **Minimum Group Size:** Enter a number, such as 5.
- **Maximum Group Size:** Enter a number, such as 10.
- **Desired Group Size:** Enter a number, such as 5.
- **Load Balancer:** Select a load balancer that you created.
- **Listener:** Select a listener.
- **L3 Network:** Select an available network.
- **Health Check:** Select the load balancer health check.

- **Health Check Grace Period:** Enter a value, such as 300. This value defaults to 300 seconds.
- **Enable alarm notification:** Select the checkbox. By doing so, the alarm notification will be enabled.
- **Endpoint:** Specify one or more endpoints that you created.
- **Apply immediately after creation:** Select the checkbox. By doing so, the auto scaling group will be enabled after you create the auto scaling group.

**Figure 5-1: Step 1 Set Basic Information**

Next(1/3) Cancel

Create Auto Scaling Group: Basic Configuration

Zone

ZONE-1

Group Name \*

Auto Scaling Group-Business A

Description

Minimum Group Size \*

5

Maximum Group Size \*

10

Desired Group Size \*

5



Load Balancer

Load Balancer

Listener \*

Listener-2

Listener-1

L3 Network \*

L3-Private Network-vRouter

Health Check \*

Load Balancer

Health Check Grace Period \*

300

second

Endpoint \*

System Endpoint

☒ Apply immediately after creation

b) Configure the autoscaling VM instance.

To configure the autoscaling VM instance, set the following parameters:

- **VM Name:** Enter a name for the VM instance, such as VM.
- **VM Description:** Optional. Enter a description for the VM instance.
- **Instance Offering:** Select an instance offering for the VM instance.
- **Image:** Select an image for the VM instance.
- **L3 Network:** Default to display the vRouter network that you set in the previous step.
- **Advanced:** Make advanced settings for the VM instance as needed.

**Figure 5-2: Step 2 Configure Autoscaling VM Instance**

Previous

Next(2/3)

Cancel

Create Auto Scaling Group: Configure VM Instance

VM Name \*

VM

VM Description

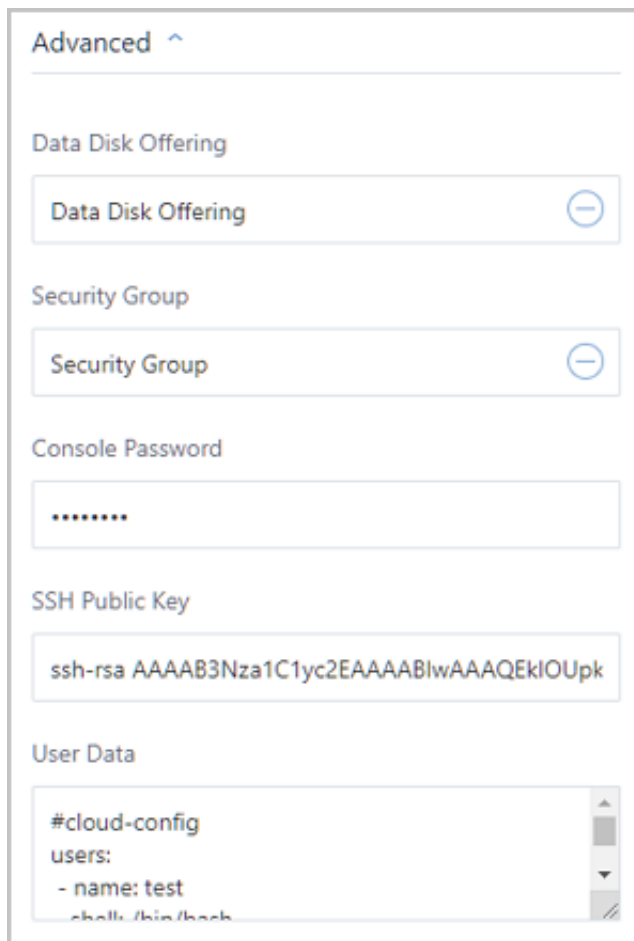
Instance Offering \*

InstanceOffering-1

Image \*

Image-1

To use internal monitoring items, select the image with the agent installed, or use User Data to install the agent.



Advanced ^

Data Disk Offering

Data Disk Offering

Security Group

Security Group

Console Password

.....

SSH Public Key

ssh-rsa AAAAB3Nza1C1yc2EAAAABlwAAAQEkIOUpk

User Data

```
#cloud-config
users:
  - name: test
    shell: /bin/bash
```

c) Configure the auto scaling policies.

To configure a scale-out policy, set the following parameters:

- **Trigger Metric:** Select a trigger metric, such as the average VM memory utilization.
- **Trigger Condition:** Set a trigger condition that is greater than 80%.
- **Duration:** Set a duration, such as 180 seconds.
- **Cooldown Time:** Set a cooldown time, such as 300 seconds.
- **VMs To Be Added Per Time:** Enter a value, such as 2.

To configure a scale-in policy, set the following parameters:

- **Trigger Metric:** Default to display the average VM memory utilization.
- **Trigger Condition:** Set a trigger condition that is lower than 20%.
- **Duration:** Set a duration, such as 180 seconds.
- **Cooldown Time:** Set a cooldown time, such as 300 seconds.
- **Removal Policy:** Select a removal policy, such as the earliest created VM instance.
- **VMs To Be Removed Per Time:** Set a value, such as 2.

Step 3 is about how to configure the auto scaling policies, as shown in [Step 3 Configure Auto Scaling Policy](#).

**Figure 5-3: Step 3 Configure Auto Scaling Policy**

Previous OK Cancel

Create Auto Scaling Group: Configure Scaling Policy

Scale-out Policy

Trigger Metric \*

VM Instance Memory Average Used In Percent

Trigger Condition \*

>	80	%
---	----	---

Duration \*

180	second
-----	--------

Cooldown Time \*

300	second
-----	--------

VMs To Be Added Per Time \*

2
---

### Scale-in Policy

Trigger Metric \*

VM Instance Memory Average Used In Percent ▾

Trigger Condition \*

< ▾

20

%

Duration \*

180

second ▾

Cooldown Time \*

300

second ▾

Removal Policy \*

Most Recent Created VM Instance ▾

VMs To Be Removed Per Time \*

2

**2. Verify functionalities: elastic self-health, elastic scale-out, and elastic scale-in.**

- Elastic self-health: The number of healthy VM instances in the auto scaling group continues to stay at 5 or above to ensure the normal functions of your businesses.

Unhealthy VM instances are automatically deleted in the auto scaling group, as shown in [Delete Unhealthy VM Instance](#).

**Figure 5-4: Delete Unhealthy VM Instance**

VM Instance:

Name	Default IP	Health Status	State
asg-Auto Scaling Group-Busin...	172.30.90.197	● Healthy	● Running
asg-Auto Scaling Group-Busin...	172.30.105.27	● Healthy	● Running
asg-Auto Scaling Group-Busin...	172.30.105.20	● Healthy	● Running
asg-Auto Scaling Group-Busin...	172.30.118.5	● Healthy	● Running
asg-Auto Scaling Group-Busin...	172.30.108.171	● Healthy	● Running
asg-Auto Scaling Group-Busin...	172.30.126.215	● Healthy	● Running

- Elastic scale-out: During Double Eleven, the Spring Festival, and other great events, businesses are skyrocketing, VM instances in the auto scaling group will be automatically added. Specifically, during peak periods, up to 10 VM instances are automatically added, whereby effectively avoiding access latency and excessive resource load, as shown in [Elastic Scale-out](#).

**Figure 5-5: Elastic Scale-out**

VM Instance Available(9) Deleted(0)

Name	Tag(Admin)	CPU	Memory	Default IP	Host IP	Cluster(All)	State(All)	Owner	HA Level
asg-Auto Scaling Gr...	None	1	1 GB	172.30.84.206	10.0.31.198	Ceph	● Running	admin	None
asg-Auto Scaling Gr...	None	1	1 GB	172.30.82.187	10.0.193.29	Cluster-1	● Running	admin	None
asg-Auto Scaling Gr...	None	1	1 GB	172.30.66.48	10.0.165.46	Ceph	● Running	admin	None
asg-Auto Scaling Gr...	None	1	1 GB	172.30.107.240	10.0.165.46	Ceph	● Running	admin	None
asg-Auto Scaling Gr...	None	1	1 GB	172.30.78.249	10.0.108.114	Shared Cluster	● Running	admin	None
asg-Auto Scaling Gr...	None	1	1 GB	172.30.118.58	10.0.193.29	Cluster-1	● Running	admin	None
asg-Auto Scaling Gr...	None	1	1 GB	172.30.116.5	10.0.193.29	Cluster-1	● Running	admin	None
asg-Auto Scaling Gr...	None	1	1 GB	172.30.121.86	10.0.108.114	Shared Cluster	● Running	admin	None
asg-Auto Scaling Gr...	None	1	1 GB	172.30.96.3	10.0.31.198	Ceph	● Running	admin	None

- Elastic scale-in: After the festivals and shopping sprees, the business workload obviously declines. In that case, VM instances need to be automatically removed in time to avoid a waste of resources, as shown in [Elastic Scale-in](#).

**Figure 5-6: Elastic Scale-in**

VM Instance

Available(9)

Deleted(0)

Create VM Instance

Start

Stop

Actions

Tag

<input type="checkbox"/>	Name	Tag(Admin) ▾	CPU	Memory	Default IP	Host IP	Cluster(All) ▾	State(All) ▾	Owner	HA Level
<input type="checkbox"/>	asg-Auto Scaling Gr...	None	1	1 GB	172.30.83.211	10.0.108.114	Shared Cluster	<div><div></div>Running</div>	admin	None
<input type="checkbox"/>	asg-Auto Scaling Gr...	None	1	1 GB	172.30.76.73	10.0.193.29	Cluster-1	<div><div></div>Running</div>	admin	None
<input type="checkbox"/>	asg-Auto Scaling Gr...	None	1	1 GB	172.30.83.4	10.0.165.46	Ceph	<div><div></div>Running</div>	admin	None
<input type="checkbox"/>	asg-Auto Scaling Gr...	None	1	1 GB	172.30.109.27	10.0.165.46	Ceph	<div><div></div>Running</div>	admin	None
<input type="checkbox"/>	asg-Auto Scaling Gr...	None	1	1 GB	172.30.118.69	10.0.31.198	Ceph	<div><div></div>Running</div>	admin	None
<input type="checkbox"/>	asg-Auto Scaling Gr...	None	1	1 GB	172.30.100.39	10.0.193.29	Cluster-1	<div><div></div>Running</div>	admin	None
<input type="checkbox"/>	asg-Auto Scaling Gr...	None	1	1 GB	172.30.116.176	10.0.31.198	Ceph	<div><div></div>Running</div>	admin	None
<input type="checkbox"/>	asg-Auto Scaling Gr...	None	1	1 GB	172.30.120.145	10.0.31.198	Ceph	<div><div></div>Running</div>	admin	None
<input type="checkbox"/>	asg-Auto Scaling Gr...	None	1	1 GB	172.30.116.184	10.0.108.114	Shared Cluster	<div><div></div>Running</div>	admin	None

## What's next

So far, we have introduced the usage of the auto scaling group.

# Glossary

---

## Zone

A zone is a logical group of resources such as clusters, L2 networks, and primary storages. Zone is the largest resource scope defined in ZStack.

## Cluster

A cluster is a logical group of analogy hosts (compute nodes). Hosts in the same cluster must be installed with the same operating system, have the same network configuration, and be able to access the same primary storage. In a real data center, a cluster usually maps to a rack.

## Management Node

A management node is a host with operating system installed to provide UI management and Cloud deployment.

## Compute Node

A compute node is a physical server (also known as a host) that provides VM instances with compute, network, and storage resources.

## Primary Storage

A primary storage is a storage server used to store disk files in VM instances. Local storage, NFS, Ceph, Shared Mount Point, and Shared Block are supported.

## Backup Storage

A backup storage is a storage server used to store image template files. ImageStore, SFTP (Community Edition), and Ceph are supported. We recommend that you deploy backup storage separately.

## ImageStore

ImageStore is a type of backup storage. You can use ImageStore to create images for VM instances that are in the running state and manage image version updates and release.

ImageStore allows you quickly upload, download, export images, and create image snapshots as needed.



## VM Instance

A VM instance is a virtual machine instance running on a host. A VM instance has its own IP address to access public network and run application services.

## Image

An image is an image template used by a VM instance or volume. Image templates include system volume images and data volume images.

## Volume

A volume can either be a data volume or a root volume. A volume provides storage to a VM instance. A shared volume can be attached to one or more VM instances.

## Instance Offering

An instance offering is a specification of the VM instance CPU and memory, and defines the host allocator strategy, disk bandwidth, and network bandwidth.

## Disk Offering

A disk offering is a specification of a volume, which defines the size of a volume and how the volume will be created.

## L2 Network

An L2 network is a layer 2 broadcast domain used for layer 2 isolation. Generally, L2 networks are identified by names of devices on the physical network.

## L3 Network

An L3 network is a collection of network configurations for VM instances, including the IP range, gateway, and DNS.

## Public Network

A public network is generally allocated with a public IP address by Network Information Center (NIC) and can be connected to IP addresses on the Internet.

## Private Network

A private network is the internal network that can be connected and accessed by VM instances.

## **L2NoVlanNetwork**

L2NoVlanNetwork is a network type for creating an L2 network. If L2NoVlanNetwork is selected, VLAN settings are not used for host connection.

## **L2VlanNetwork**

L2VlanNetwork is a network type for creating an L2 network. If L2VlanNetwork is selected, VLAN settings are used for host connection and need to be configured on the corresponding switches in advance.

## **VXLAN Pool**

A VXLAN pool is an underlay network in VXLAN. You can create multiple VXLAN overlay networks (VXLAN) in a VXLAN pool. The overlay networks can operate on the same underlay network device.

## **VXLAN**

A VXLAN network is a L2 network encapsulated by using the VXLAN protocol. A VXLAN network belongs to a VXLAN pool. Different VXLAN networks are isolated from each other on the L2 network.

## **vRouter**

A vRouter is a custom Linux VM instance that provides various network services.

## **Security Group**

A security group provides L3 network firewall control over the VM instances. It can be used to set different security rules to filter IP addresses, network packet types, and the traffic flow of network packets.

## **EIP**

An elastic IP address (EIP) is a method to access a private network through a public network.

## **Snapshot**

A snapshot is a point-in-time capture of data status in a disk. A snapshot can be either an automatic snapshot or a manual snapshot.