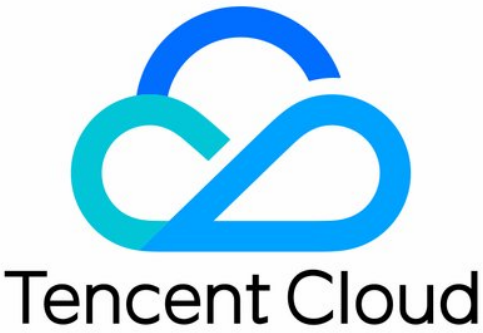
Auto Scaling

Best Practices

Product Introduction



©2013-2018 Tencent Cloud. All rights reserved.

Copyright Notice

©2013-2018 Tencent Cloud. All rights reserved.

Copyright in this document is exclusively owned by Tencent Cloud. You must not reproduce, modify, copy or distribute in any way, in whole or in part, the contents of this document without Tencent Cloud's the prior written consent.

Trademark Notice

Tencent Cloud

All trademarks associated with Tencent Cloud and its services are owned by Tencent Cloud Computing (Beijing) Company Limited and its affiliated companies. Trademarks of third parties referred to in this document are owned by their respective proprietors.

Service Statement

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

Contents

Best Practices
- Setting Static Outbound IP for Scaling Group
- WeChat Mini Program
- Cluster Solution
- Auto-enable Services on Startup

©2013-2018 Tencent Cloud. All rights reserved.

Best Practices
Setting Static Outbound IP for Scaling Group

Last updated：2018-06-21 10:13:04

In this section, we will introduce to you about how to set up a fixed IP address for external access if the cluster requires to actively make external access.

Required Conditions

If three conditions are satisfied in the cluster of your scaling group at the same time as follows:

- Requests are received from CLB;
- Cluster machines need to actively make external access;
- When making external access, it is desired to use a **fixed** public IP.

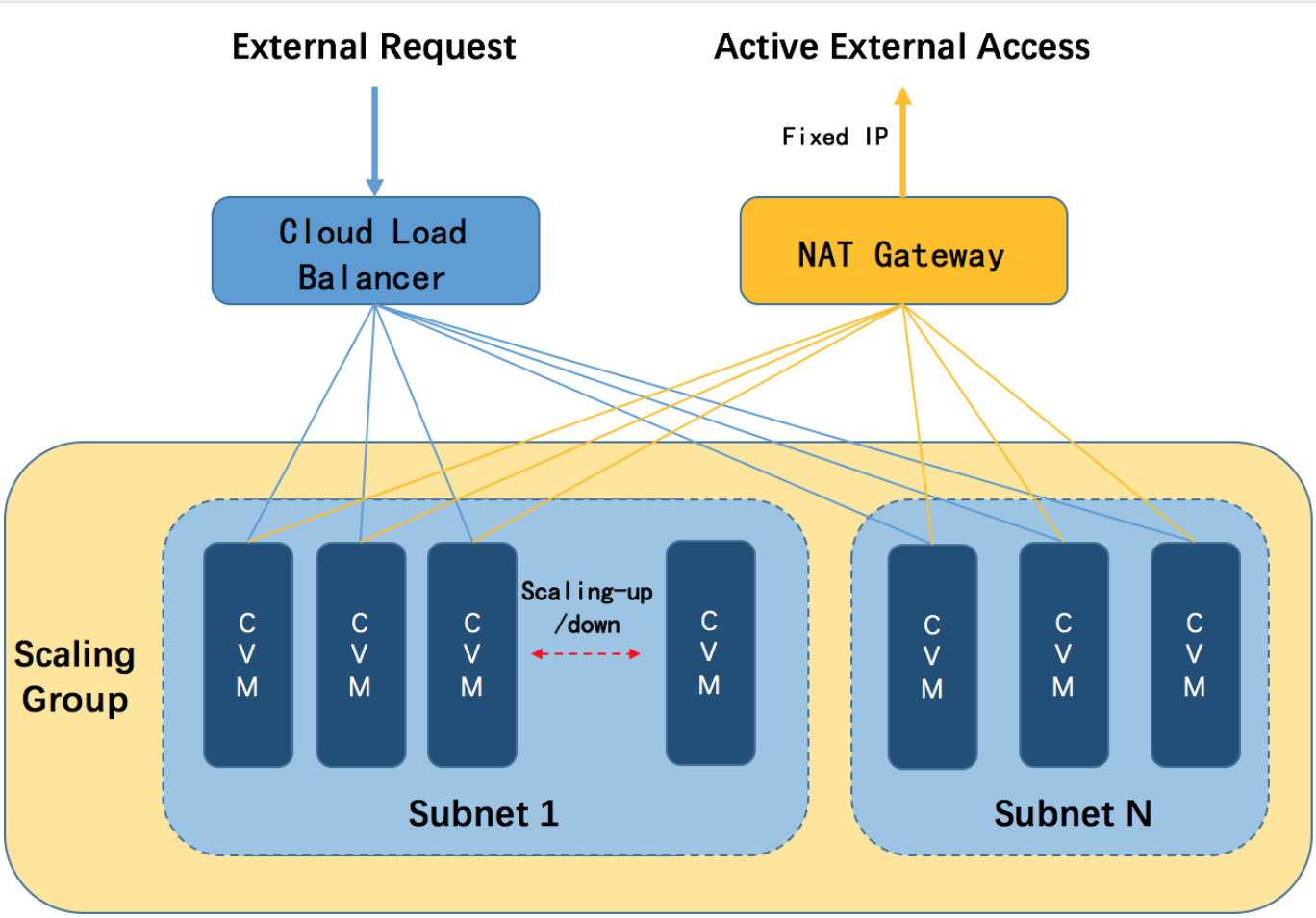
the IP address can be set up according to the following solution.

Solution Overview

1. Receive and respond to external requests via CLB.
2. Place the machine in the subnet of VPC, direct the routing table to the NAT gateway and all active external request are delivered via the public IP of the NAT gateway.
3. The network attribute of the scaling group is set as this subnet, so that all the expanded machines will actively make external access via the NAT gateway.

The setting procedure is shown as follows:

©2013-2018 Tencent Cloud. All rights reserved.



Setting Method

Step One: Create a VPN and a Subnet

1. Create a VPC

1. Log in to Tencent Cloud Console, click "Virtual Private Cloud" in the navigation bar, or click "Experience" in Tencent Cloud's VPC Overview Page to enter the VPC Console.

2. Select a region in the drop-down box above the list and click "New" to create a VPC. For example, select the region of "North China (Beijing)".

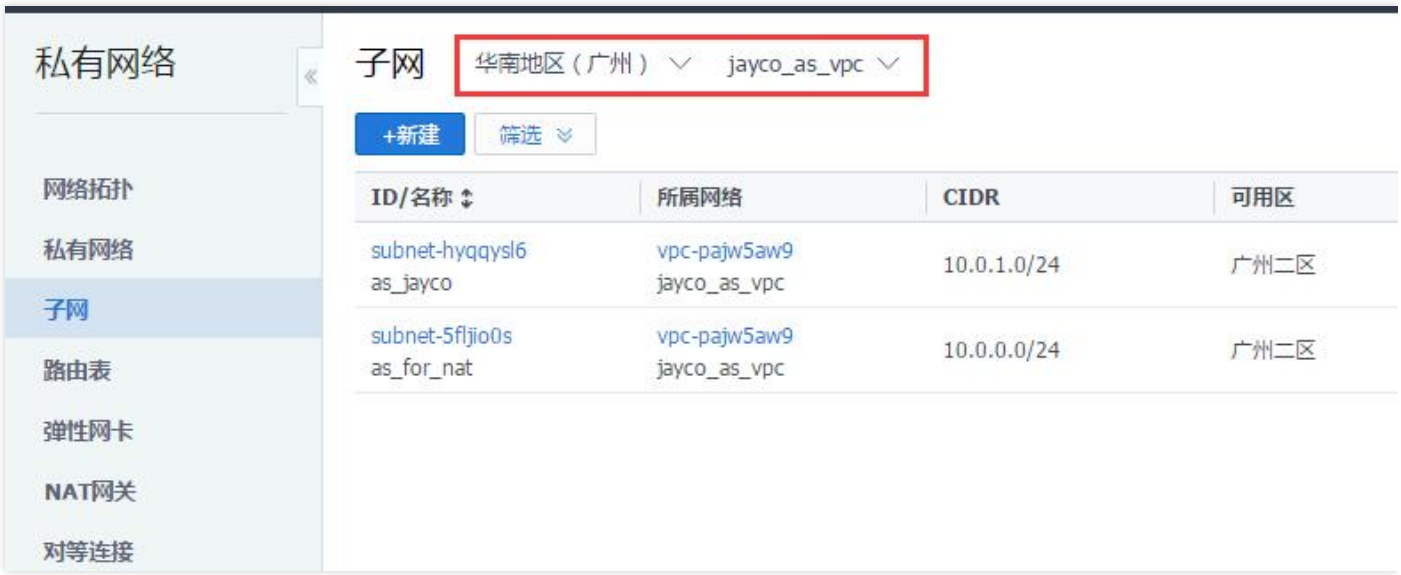
3. Enter names of the VPC and subnet as well as CIDR, and select the availability zone for the subnet.

4. Click "Create".

©2013-2018 Tencent Cloud. All rights reserved.

2. Create a subnet

1. Log in to Tencent Cloud Console, click "Virtual Private Cloud" in the navigation bar, and then click "Subnet" in the left navigation bar. Choose a region and VPC in the drop-down box.



2. Click "New", and then enter the subnet name, CIDR, availability zone and associated routing table. After that, click "Create" to confirm.

3. After the creation is completed, you can purchase machines for this subnet.

Step 2: Create a NAT gateway

1. Purchase desired products

1. Log in to Tencent Cloud Console, select "Virtual Private Cloud" tab, and then select "NAT Gateway".

2. Click the "New" button at the upper left corner, and enter or specify the following parameters in the pop-up box:
   - Gateway name
   - Gateway type (which can be changed after creation)
   - VPC of the NAT Gateway service (which was just created by you)
   - Assign the elastic IP for the NAT gateway. This IP is the fixed IP via which your machine makes external access.

3. After selection, click "OK" to complete the creation of NAT gateway.

©2013-2018 Tencent Cloud. All rights reserved.

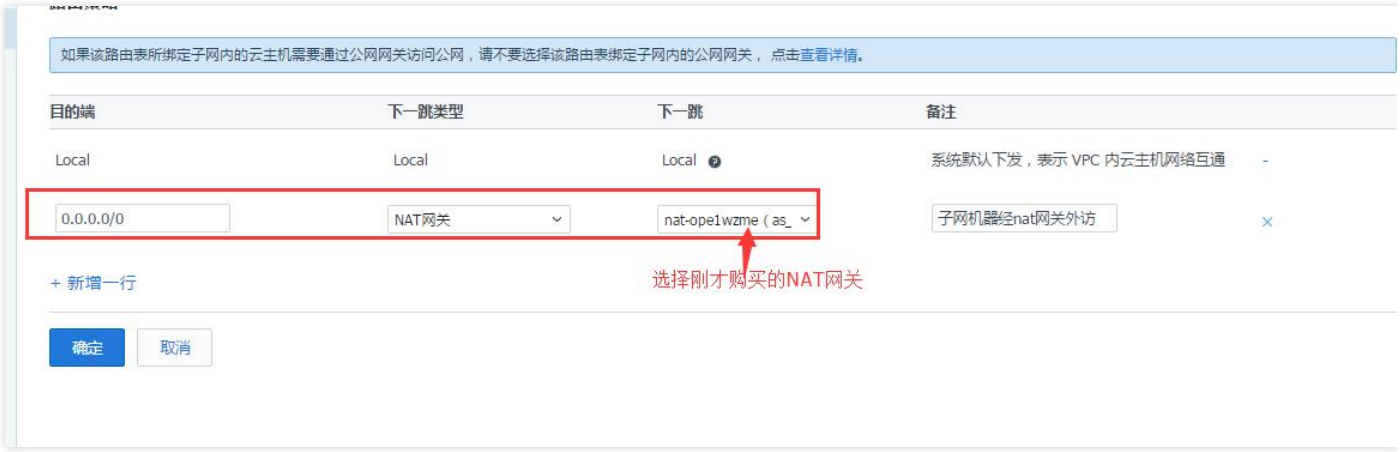
4. After the creation of the NAT gateway, you need to configure the routing rules on the Routing Table page of the VPC Console to direct the subnet traffic to the NAT gateway.

2. Set the routing table (highlights)

1. Log in to Tencent Cloud Console, click "Virtual Private Cloud" in the navigation bar to enter the Virtual Private Cloud Console, and then select "Routing Table".

2. In the routing table list, click the ID of the routing table associated to the subnet that needs to access the Internet to go to the details page of the routing table, and then click "Edit" button in the "Routing Policies".

3. Click "New Line" to fill in the destination (for example, "0.0.0.0/0" can be entered in this scenario), select "NAT Gateway" as the next hop type, select the created NAT gateway ID, and then click "OK".



Now, even without public IP, your machine in this subnet can actively get external access via the fixed IP of the NAT gateway, during which this IP is shown to be fixed externally.

Even if I purchase a CVM without any public IP and with the bandwidth of zero, I still can actively make external access, which is shown as follows:

```
[root@VM_1_162_centos ~]# ping www.baidu.com
PING www.a.shifen.com (220.181.111.188) 56(84) bytes of data.
64 bytes from 220.181.111.188: icmp_seq=1 ttl=54 time=45.7 ms
64 bytes from 220.181.111.188: icmp_seq=2 ttl=54 time=45.8 ms
```

However, the scaling group needs to identify this subnet and ensure that all the machines are created on this subnet.

Step 3: Set up the scaling group

©2013-2018 Tencent Cloud. All rights reserved.

It is used to direct the subnet information to the scaling group so that the newly expanded machine can be installed in this subnet by the scaling group.
In this way, the expanded machine will use the fixed exit IP address of the NAT gateway to access external websites.

In the Auto Scaling Console, click "New":

- Fill in the name of the scaling group, launch configuration (to be set up well in advance), maximum group size, minimum group size, initial number of instances, and other information.
- Select "Network" and "Subnet", and then direct them to the selected VPC and subnet **(important)**.

The setting procedure is shown as follows:

The setting has been completed.

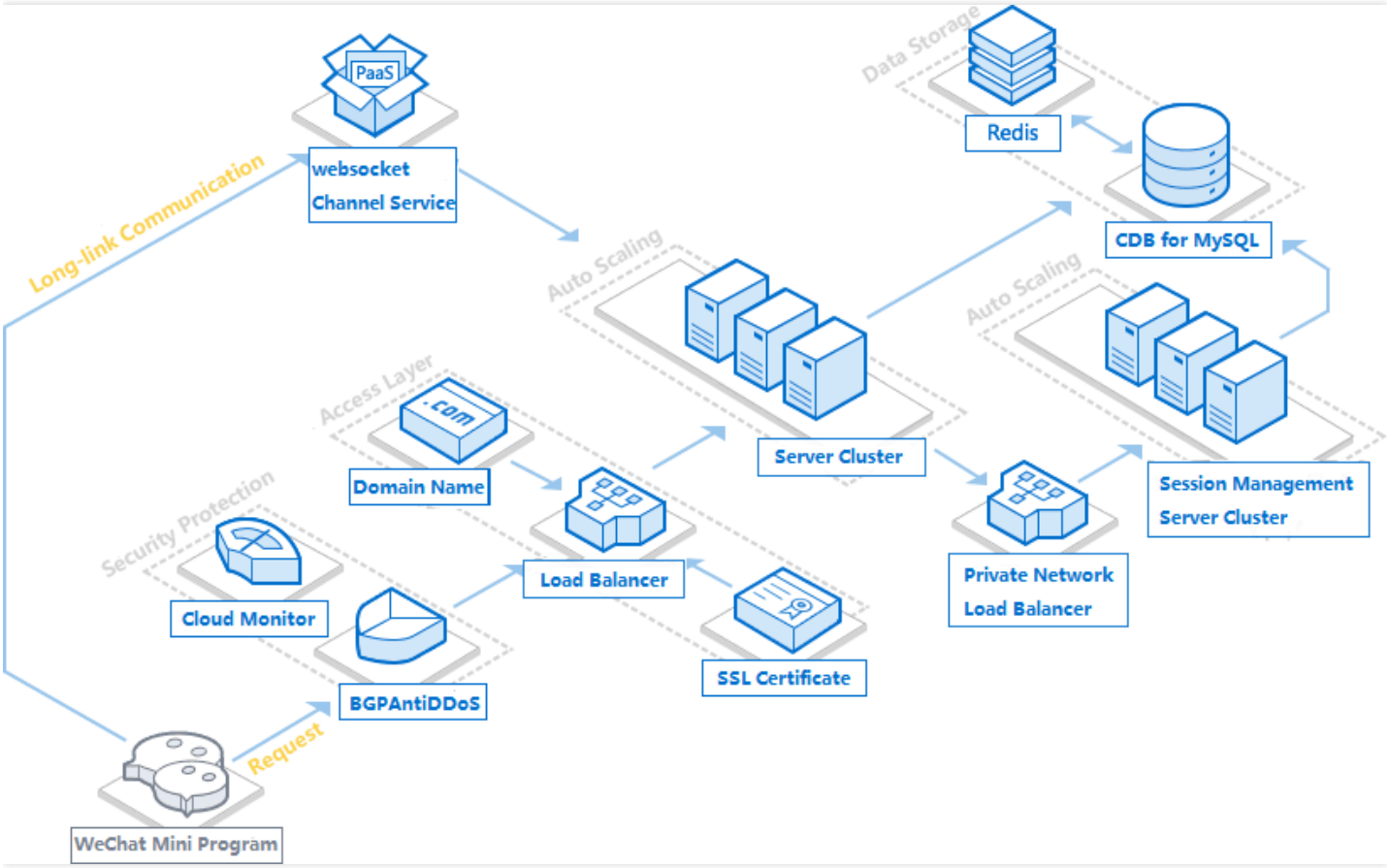
©2013-2018 Tencent Cloud. All rights reserved.

WeChat Mini Program

Last updated：2018-10-08 18:07:59

Deployment Architecture

We can create a WeChat Lite App with the following architecture. Auto Scaling is required for business server cluster and session server cluster.



Why Is the Auto Scaling Configuration Required?

Auto Scaling in the following two scenarios may reduce costs and improve business continuity:

1. Lite Apps visits with obvious peak and idle hours: According to the prediction, if more than 1 set of CVM for business server cluster and session server cluster with less than 8 peak hours are required, 30% of the cost shall be saved if the method of **fixed servers only for idle hours + temporary servers**

©2013-2018 Tencent Cloud. All rights reserved.

added for peak hours is used. With AS's scheduled scaling capability, Tencent Cloud allows you to add temporary servers in peak hours and reclaim and terminate the surplus servers in idle hours.

2. Constant lite Apps visits: You can configure monitoring and alarm based scaling policy to deal with unexpected high load, maintain continuous service and earn time for troubleshooting. Abnormally high load includes CC attack and unexpected traffic (for example, superfast spreading speed of "MYOTee" upon its launch, or burst visits brought by a certain event).

Note: Auto Scaling is free, but the scaled CVM will be charged by seconds.

What Will Auto Scaling Do?

It will:

1. Increase or decrease CVMs on a scheduled basis;
2. Automatically increase or decrease CVMs based on the load of the cluster server;
3. Automatically register the additional CVMs into the cloud load balancer to achieve fully automatic scaling.

Preconditions of Configuration

Your WeChat Lite App should support auto scaling.

Configuring Auto Scaling Policy for Session Servers

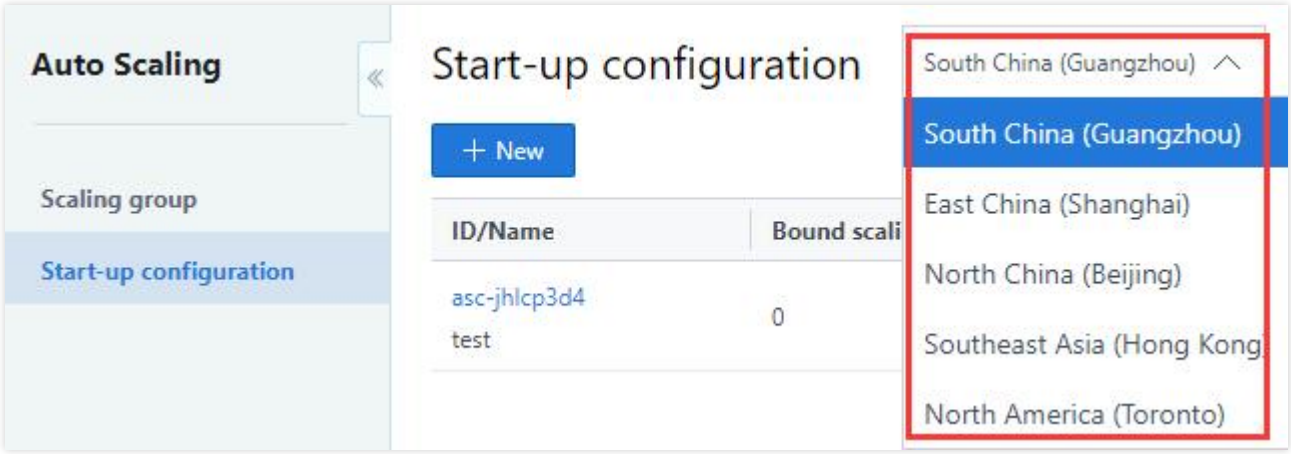
1. Create Scaling Configuration

As **scaling configuration** is the template for CVMs creation when scaling, we specify the region, model, image through scaling configuration in advance.

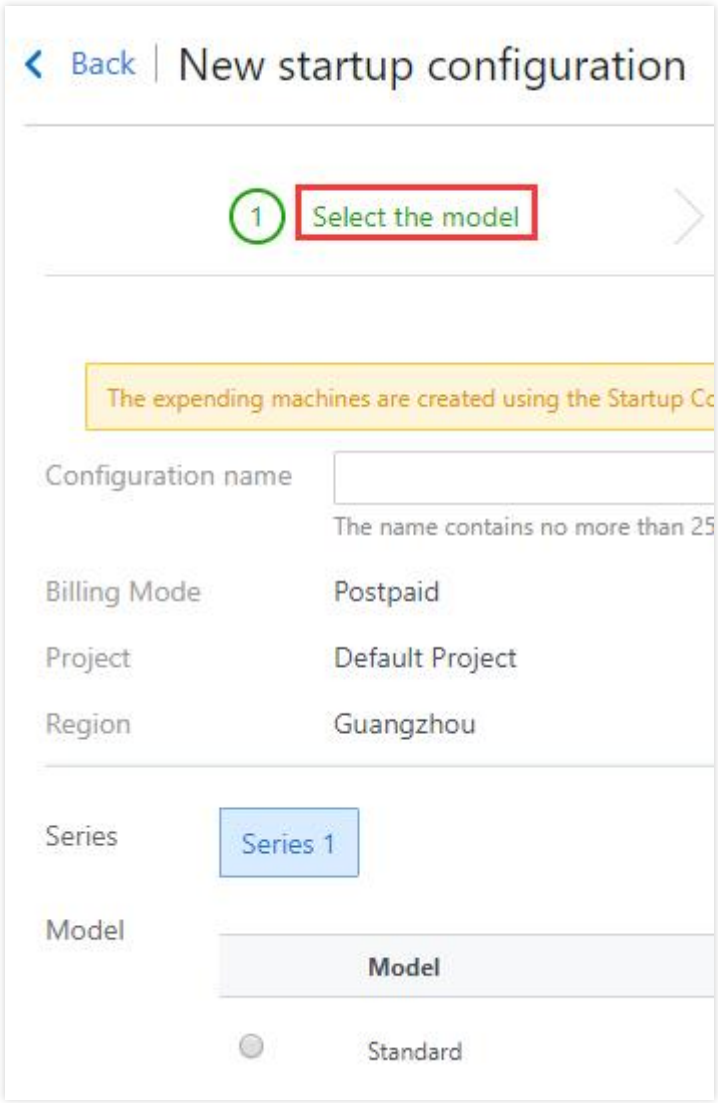
1. Log in to Auto Scaling Console, and click "Scaling Configuration" in the navigation bar.

2. Choose the project and region. Note that you must select the project and region of your WeChat Lite App.

©2013-2018 Tencent Cloud. All rights reserved.

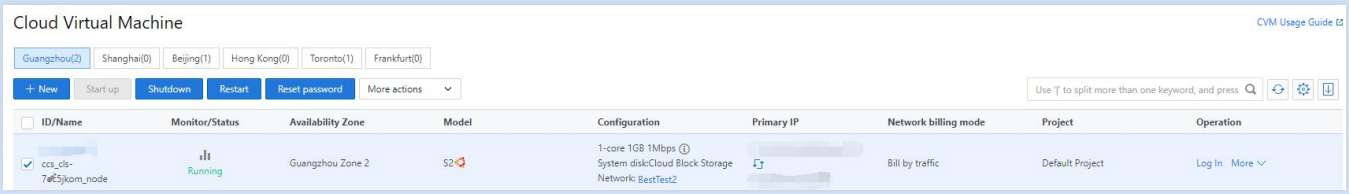


3. The following procedure is similar to purchasing CVMs. You can follow the guidelines to complete the creation of scaling configuration.



©2013-2018 Tencent Cloud. All rights reserved.

Note:
You need to create an image in advance, and ensure that the application of the image can be activated with the operating system. In such case, the scaled CVM can work directly without manual intervention.

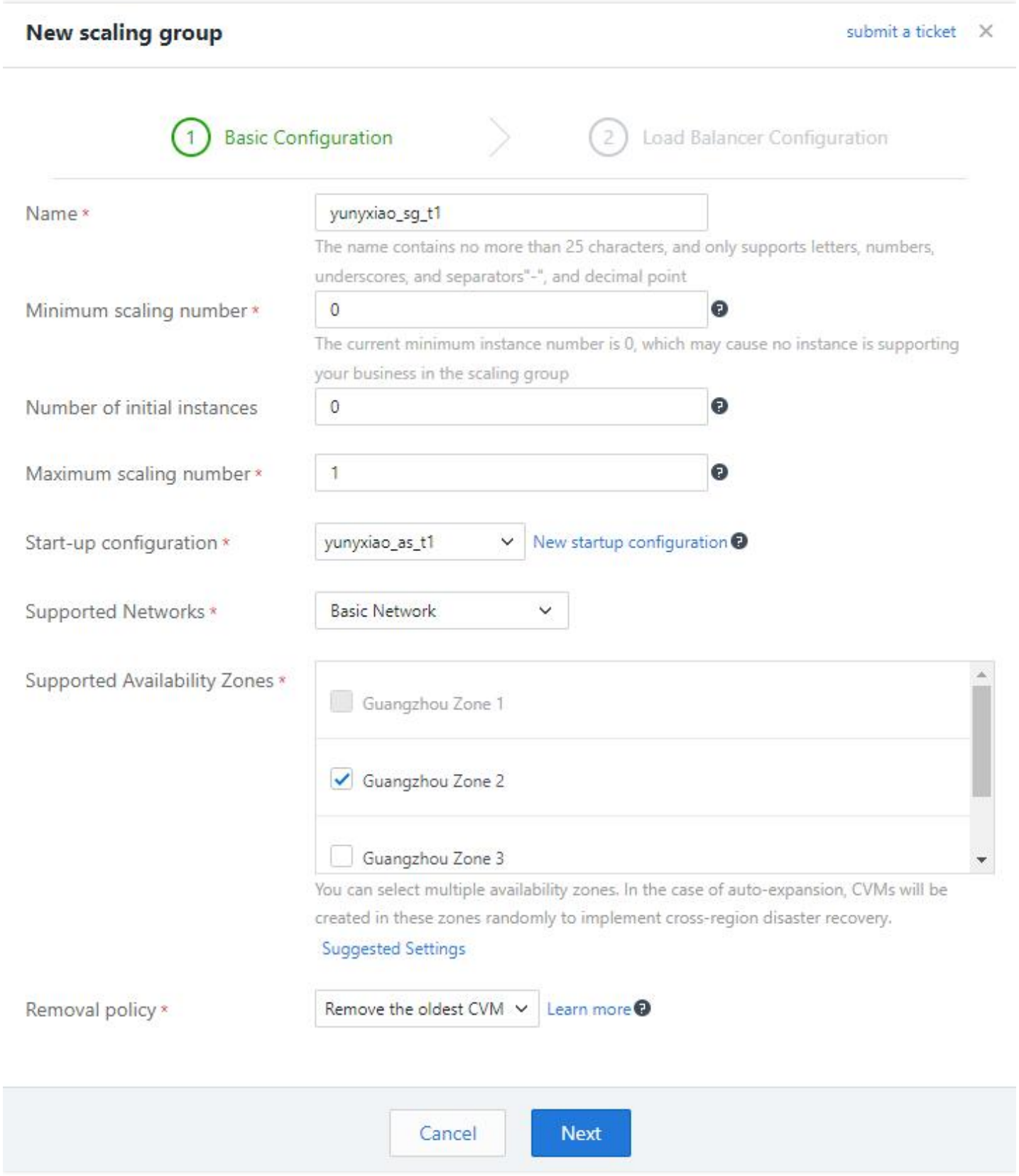


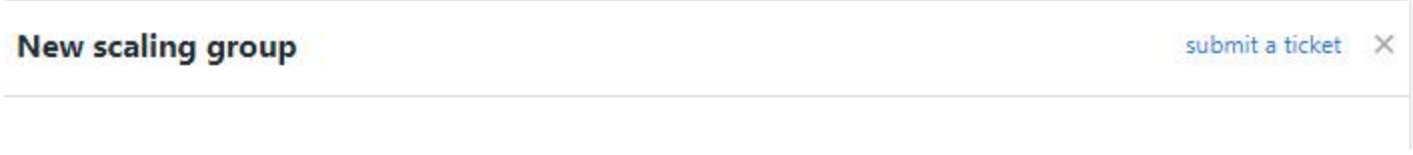
2. Create a Scaling Group

In the Auto Scaling Console, click "New" and fill in the cluster management information as follows:

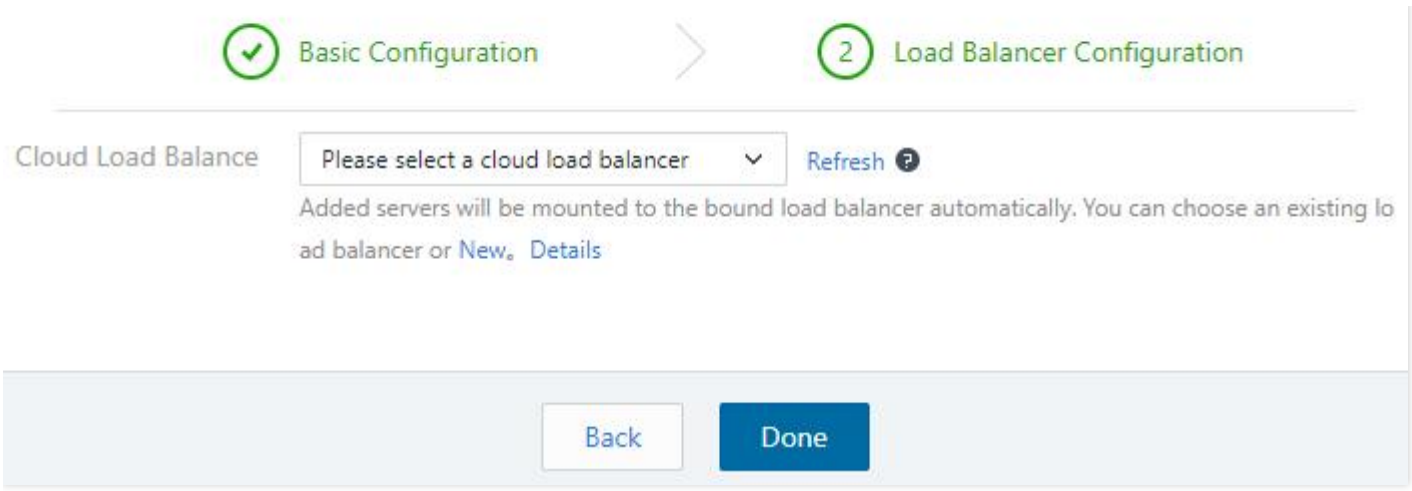
- **Name**: Fill in an appropriate name, such as "Session Server Cluster".
- **Minimum group size**: Lower limit of the number of cluster servers, such as "0".
- **Initial number of instances**: Number of CVMs automatically created upon the creation of the scaling group, such as "0".
- **Maximum group size**: Upper limit of the number of cluster servers. Fill as needed.
- **Scaling Configuration**: Select the scaling configuration you have just created.
- **Supported network**: The network environment of the session server. Select "Basic Network" generally.
- **Supported availability zone**: Select an availability zone for the CVM to be scaled. Check the availability zone of the session server.
- **Removal policy**: Select the default.
- **Cloud load balancer**: Select the cloud load balancer of the session server.

©2013-2018 Tencent Cloud. All rights reserved.





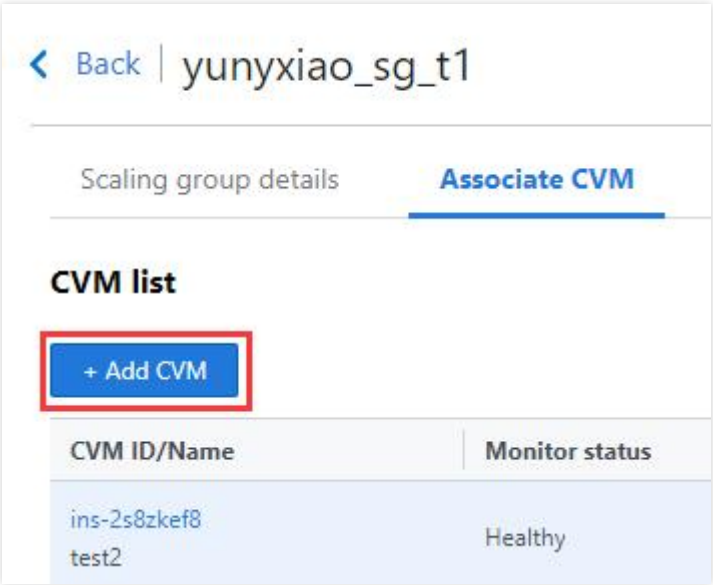
©2013-2018 Tencent Cloud. All rights reserved.



Click **Done** to finish the creation.

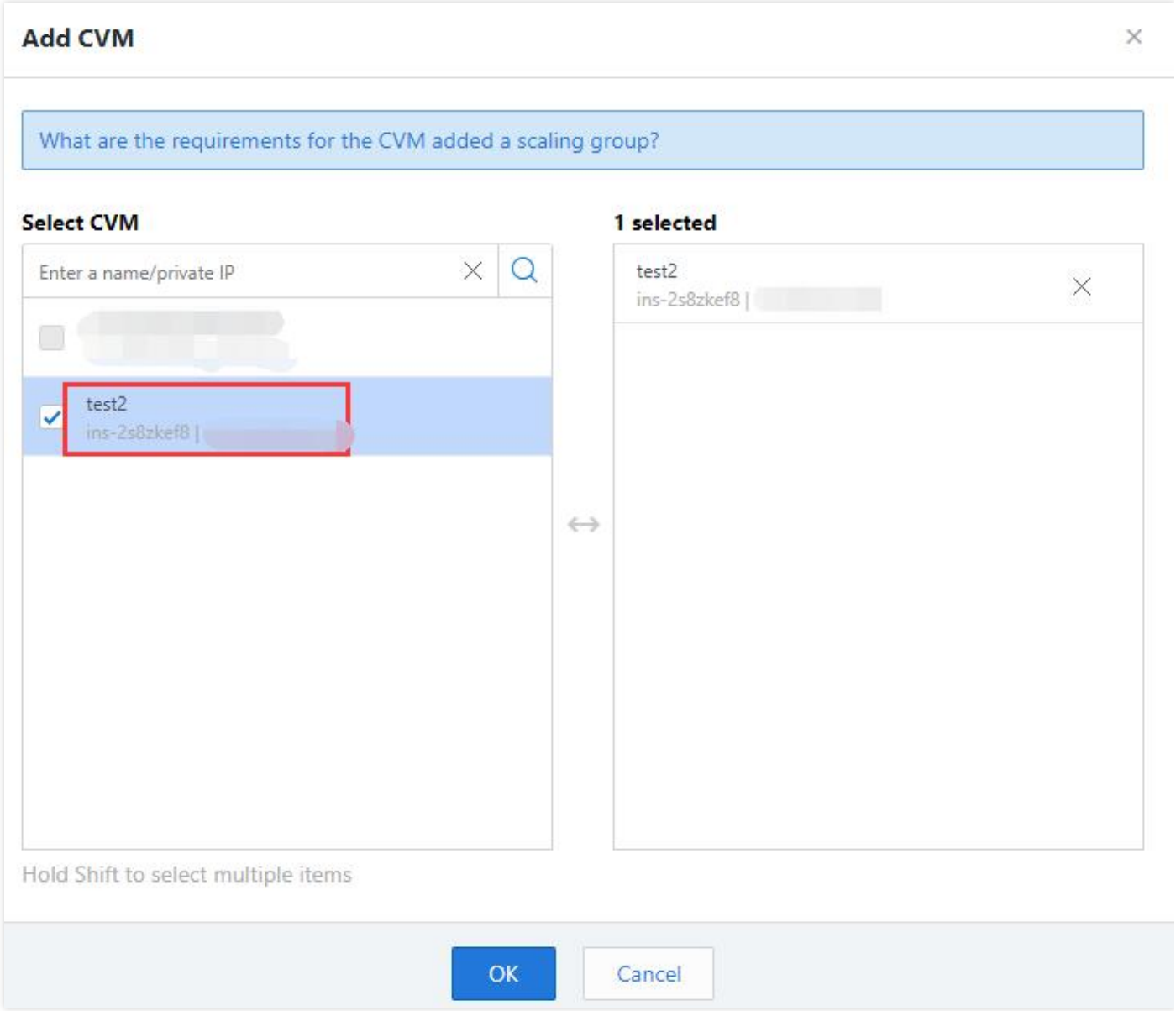
3. Add Existing CVMs to the Scaling Group

1. In the Auto Scaling Console, click on the scaling group name to enter the management page, and click "Add CVM" at the bottom of the page.

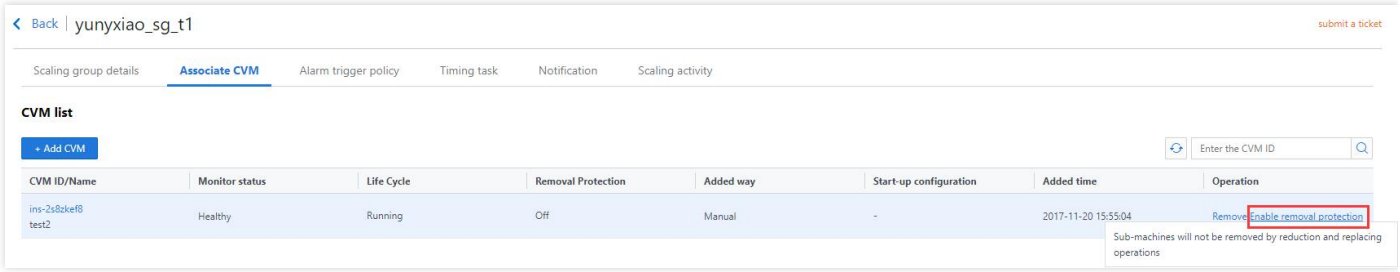


©2013-2018 Tencent Cloud. All rights reserved.

2. Select the existing session server to add to the scaling group in the pop-up dialog box.



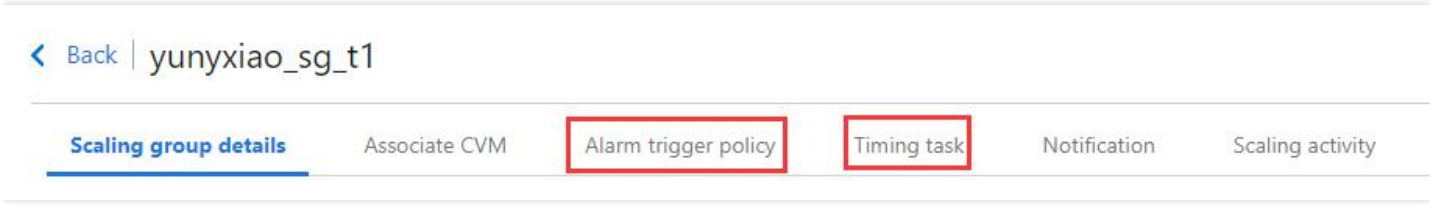
3. Set the added server as "Scale-down exemption", so as to protect it from being removed for scale-down by the scaling group.



©2013-2018 Tencent Cloud. All rights reserved.

4. Set the Scaling Policy

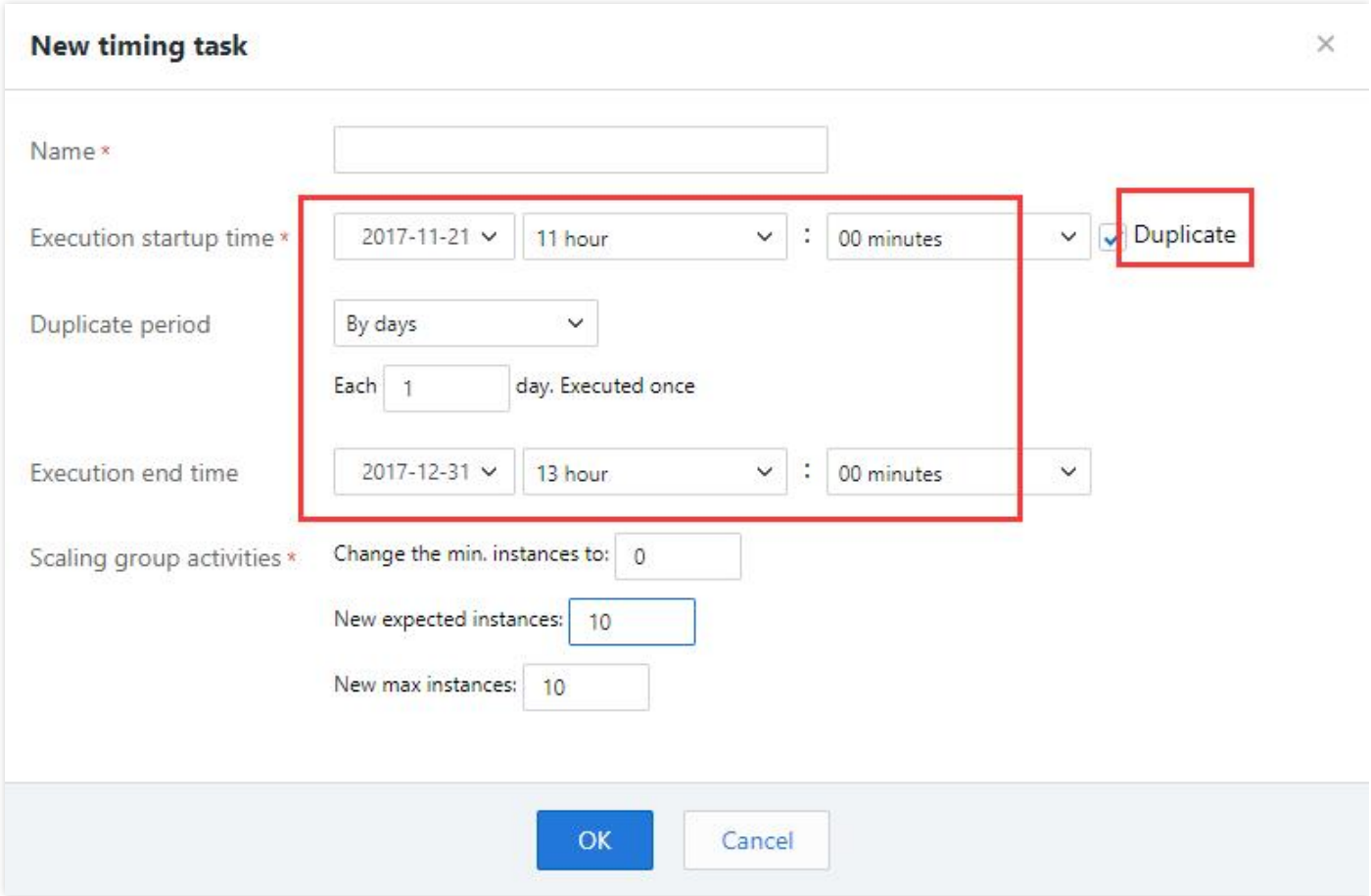
You can select scheduled scaling, or dynamic scaling based on the alarm (usually scaling up and scaling down tasks are provided in pairs).



a. Scheduled scaling:

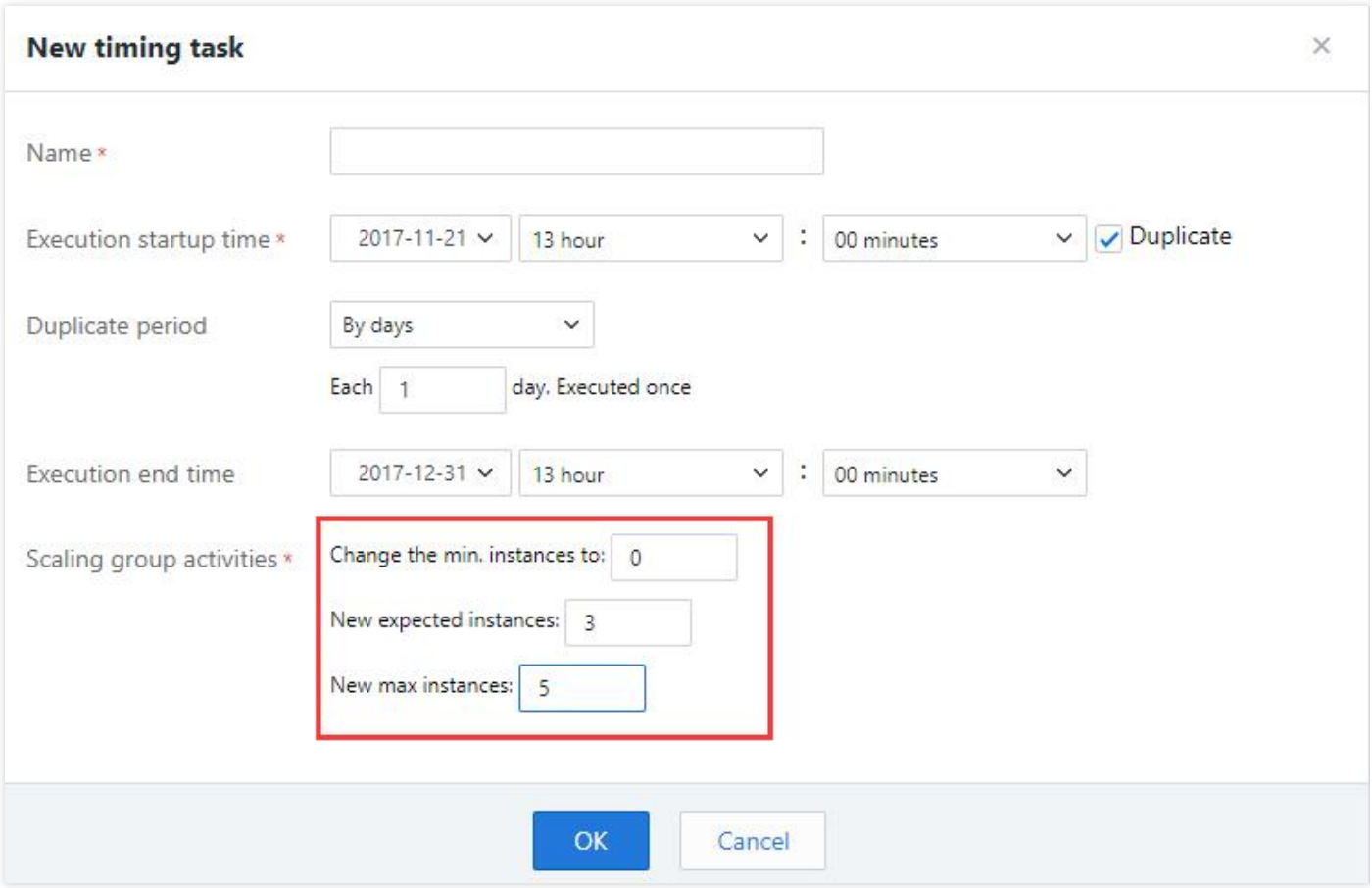
For example, a WeChat Lite App that orders lunches is expected to see higher load during the lunch time. So you can add two extra servers during 11: 00-13: 00 for such load.

- First, set a scheduled scale-up task:



©2013-2018 Tencent Cloud. All rights reserved.

- Then set a scheduled scale-down task:

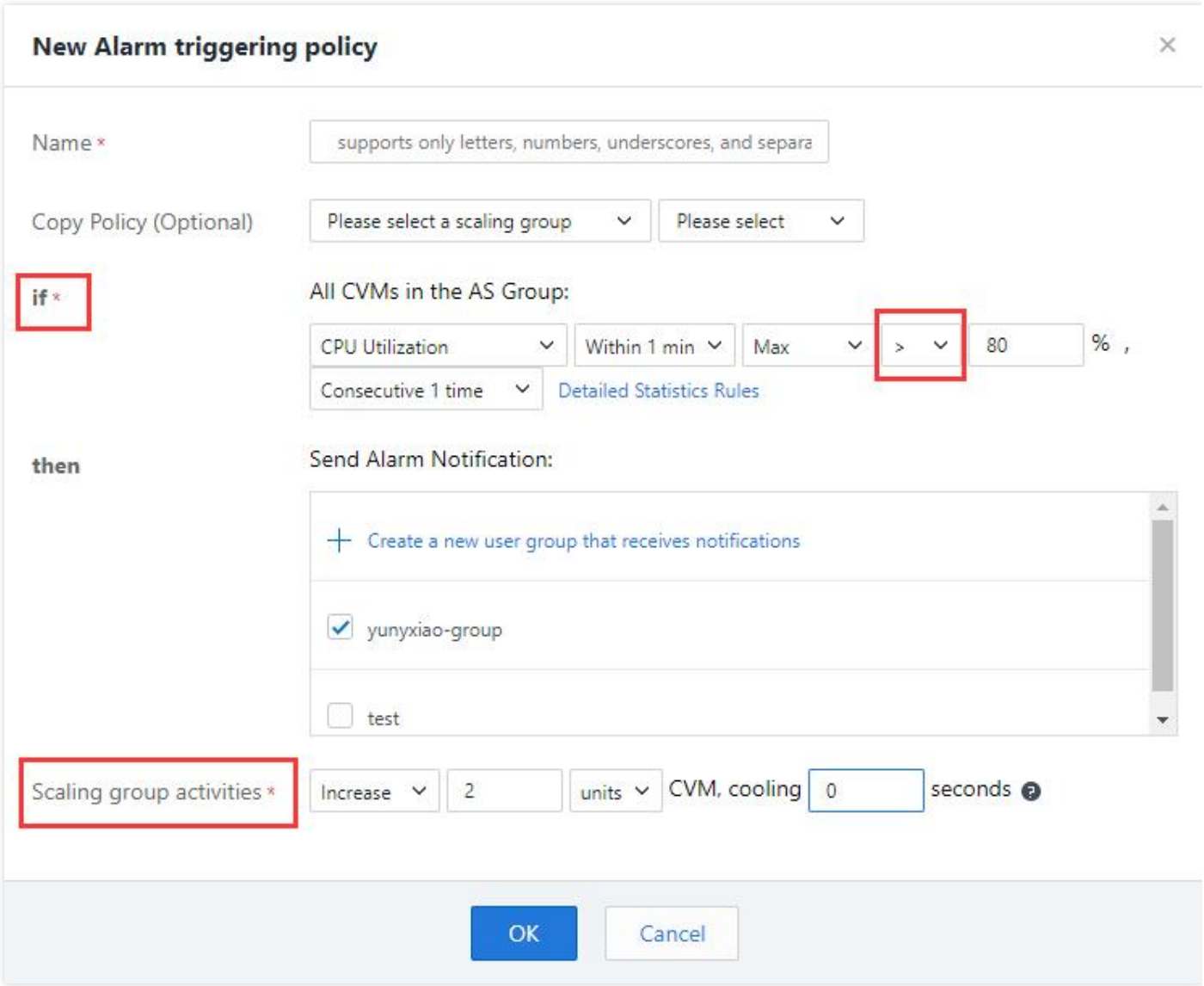


b. Scale-up based on the alarm:

Predict the uncertain scale-up while keep watch on the unexpected traffic/attacks:

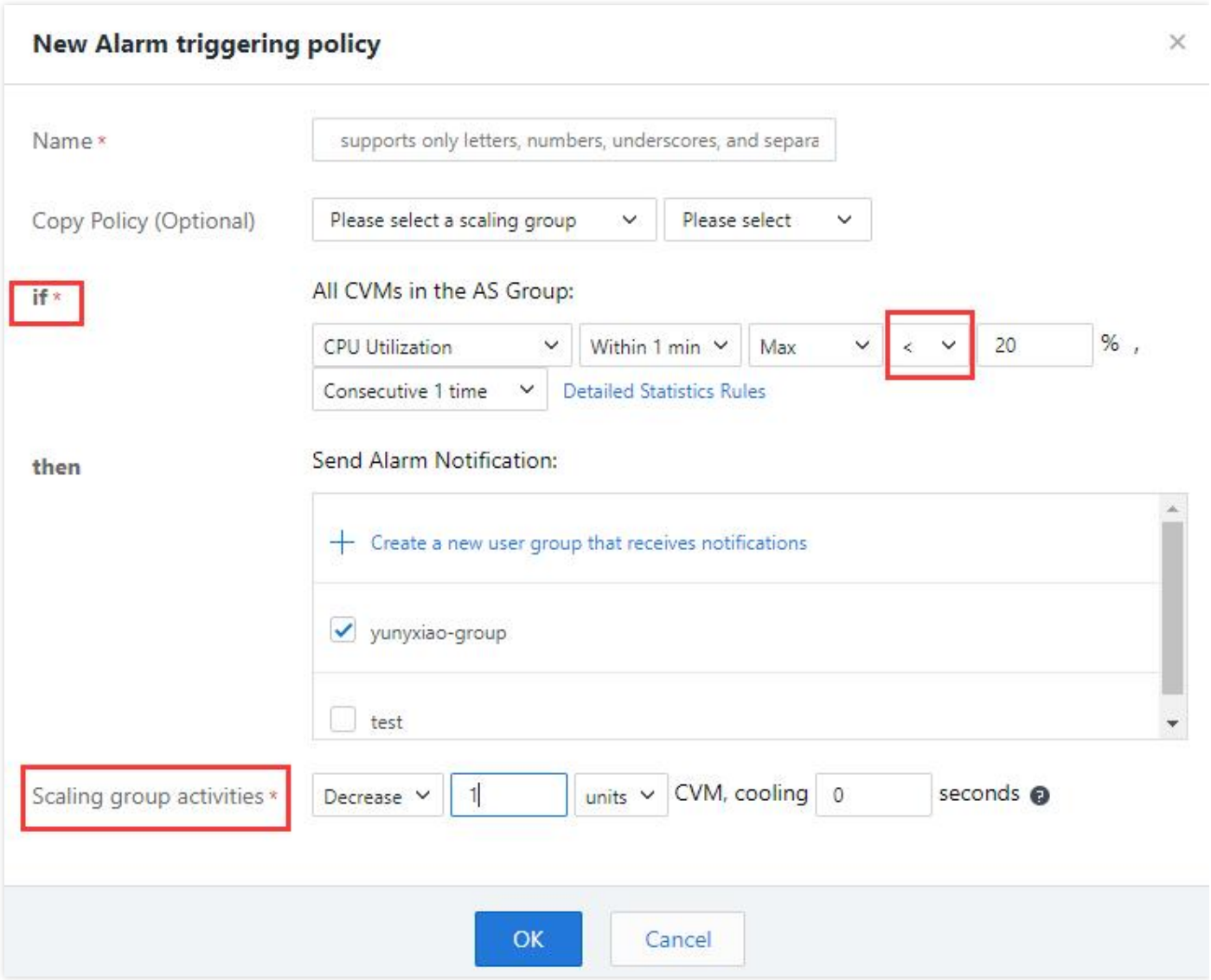
©2013-2018 Tencent Cloud. All rights reserved.

- First, set a scale-up policy based on the alarm which deals with abnormal traffic:



©2013-2018 Tencent Cloud. All rights reserved.

- Then set a scale-down policy based on the alarm which removes surplus servers:



Configuring the Auto Scaling Policy for Business Servers

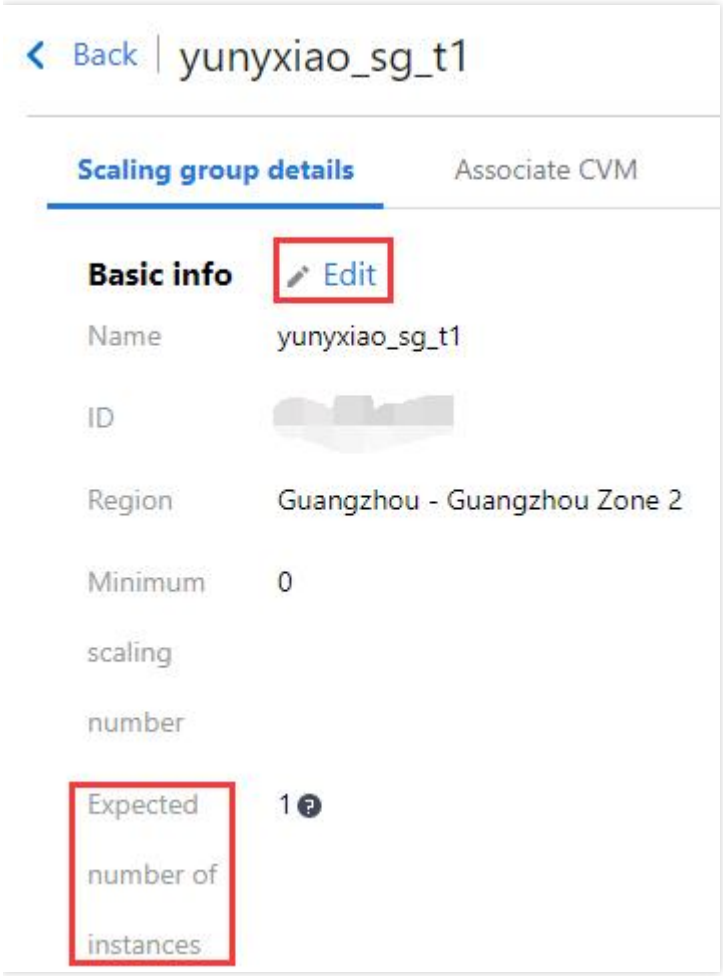
This process is similar to that of configuring the session servers:

- Create a scaling configuration for the business server;
- Create a scaling group "business server cluster" for the business server, directed to the load balance of the business server;
- Configure the auto scaling policy for the business server.

Verifying the Scalability and Viewing the Scaling Activities

©2013-2018 Tencent Cloud. All rights reserved.

Add 1 to the expected instance number of the scaling group and the group shall automatically scale up one server to the cluster. If the newly scale-up machine can handle requests normally, it indicates that the scaling group is working properly.



The scaling group also supports View Scaling Activity History, ensuring your complete grasp of the scaling activity.

Now, your WeChat Lite App is capable of intelligent capacity scaling. You no longer need to worry about the scale-up and scale-down. Just pay attention to the scaling group notification or check from time to time the history of scaling activities.

©2013-2018 Tencent Cloud. All rights reserved.

Cluster Solution

Last updated：2018-10-08 16:53:53

This solution is suitable for websites/APPs deployed based on clusters.

If your business meets the following conditions, just spend 5 minutes deploying this solution and you can save 30% in costs:

- Your website is running based on a cluster, and the cluster has more than 1 server;
- Your website has long idle time. According to the statistics of Tencent Cloud, for 90% of clusters, the load from 00:00 to 09:00 is less than 30%.

Most of the websites have a peak period of not more than 8 hours. So, for the remaining 16 hours, you can absolutely scale down idle servers to reduce costs.

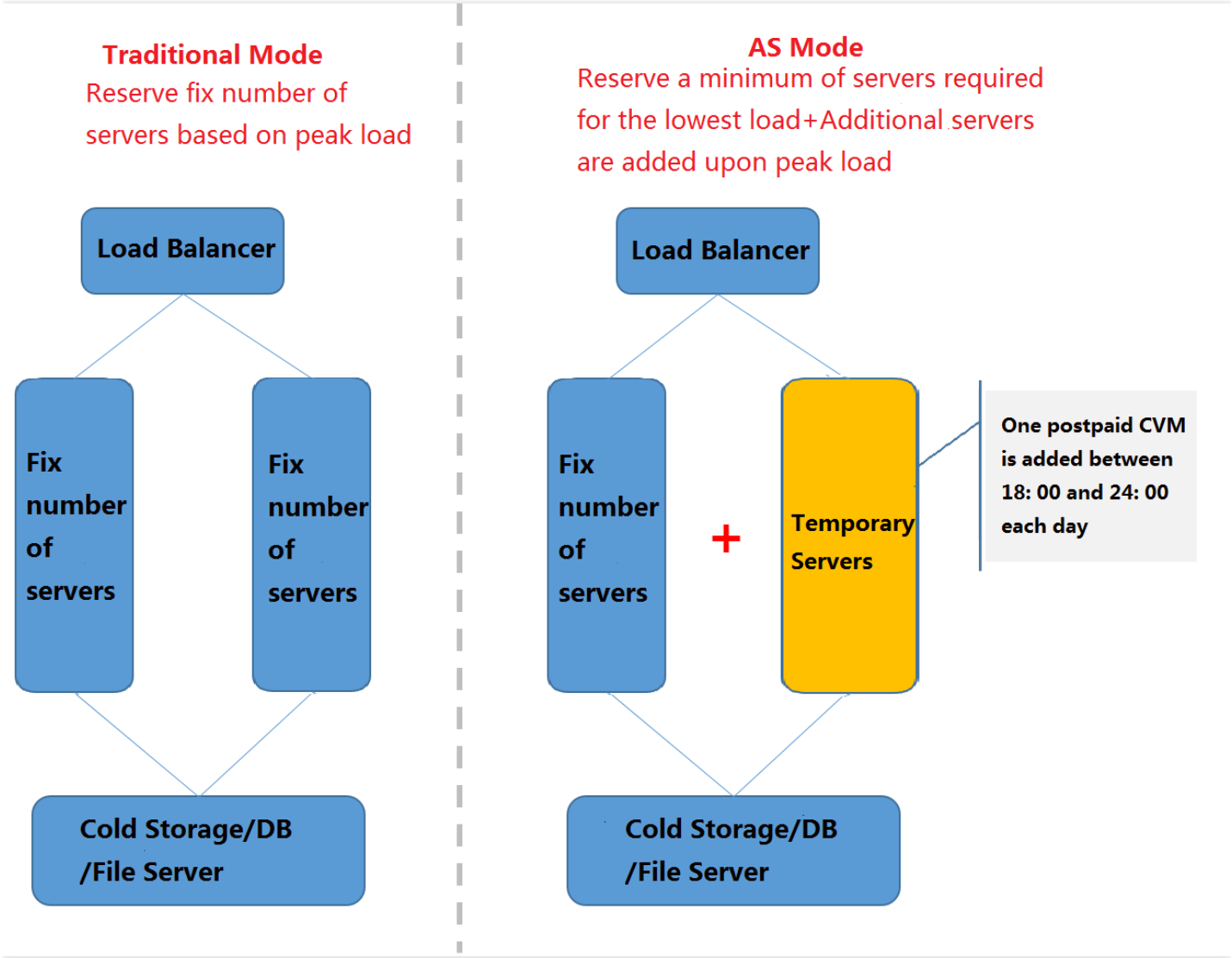
This document takes a leisure website as an example, for which the peak period is from 20:00 to 24:00.

Solution Overview

- CVMs with an annual or monthly plan can be used for deploying fixed resources based on the load in off-peak hours;
- Pay-by-usage CVMs can be used for insufficient part in peak hours. With the scheduled task, the server increases one CVM at 20:00 and scales down at 24:00.

©2013-2018 Tencent Cloud. All rights reserved.

Comparison of old and new solutions:



Revenue

Assume that the original solution needs two 4-core 8GB CVMs working for 24 hours every day. The new solution changes it to one 4-core 4GB CVM working for 24 hours every day and one temporary CVM working for 4 hours every day, which can save you around 41% in expenses.

The small website in Guangzhou in this example thus can save 1440 USD per year:

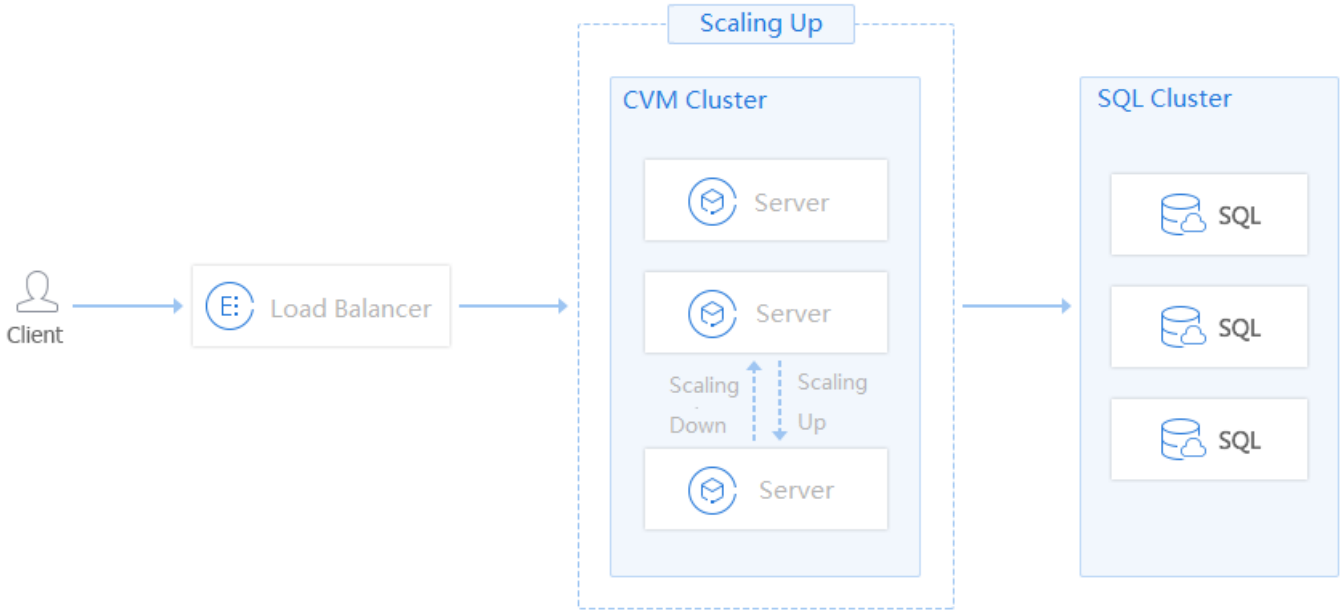
	Original Solution			Auto Scaling Solution		
Resources Required	Resources	Quantity	Fee	Resources	Quantity	Fee
	4-core 8GB CVM	2	288	4-core 8GB CVM per	1	144

©2013-2018 Tencent Cloud. All rights reserved.

	per month			month		
				4-core 4GB temporary CVM for 120 hours	1	24
	load balance	1	2.1	load balance	1	2.1
Total Cost			290.1			170.1
Save	120 USD/month					

Specific Procedure

The website for an instance is simply structured - there is only the application server cluster. For complex websites, there will be application server cluster, frontend server cluster, and cache server cluster, etc. Each cluster can carry out similar operations and each cluster corresponds to one scaling group.



Step 1. Create a custom image for the cluster CVM

This step is very simple. The custom image can be created based on an existing cluster CVM. If you have any questions, refer to Create Custom Image

©2013-2018 Tencent Cloud. All rights reserved.

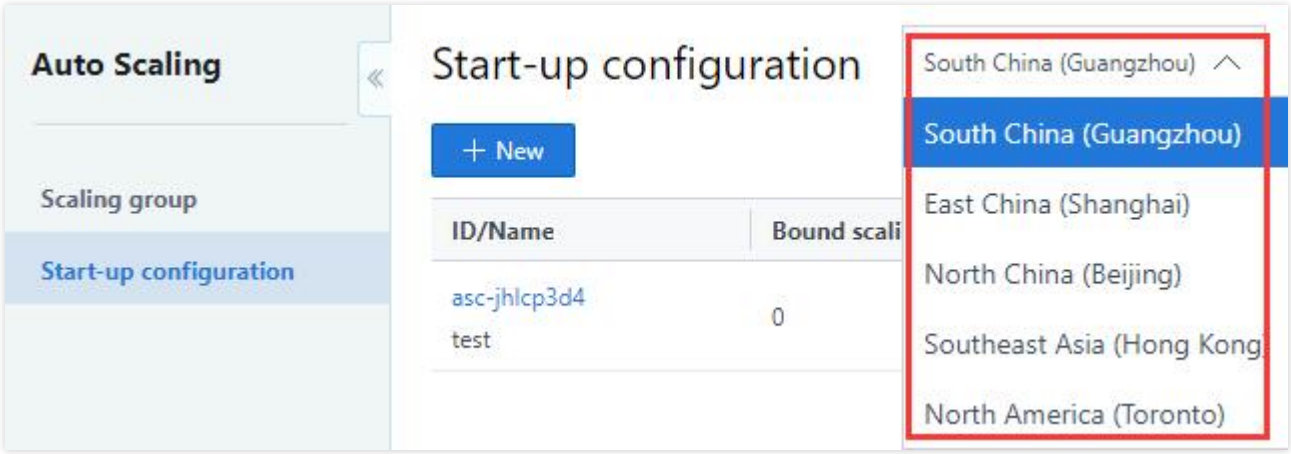
Note:
You need to deploy the environment in the image in advance, ensuring that the application of the image can be activated with the operating system. In such case, the scaled CVM can work directly without manual intervention.

Step 2. Create scaling configuration

AS uses the scaling configuration as a template to create a CVM for scale-up. Therefore, we can specify the region, model, and image through scaling configuration in advance.

1. Log in to Auto Scaling Console, and click "Scaling Configuration" in the navigation bar.

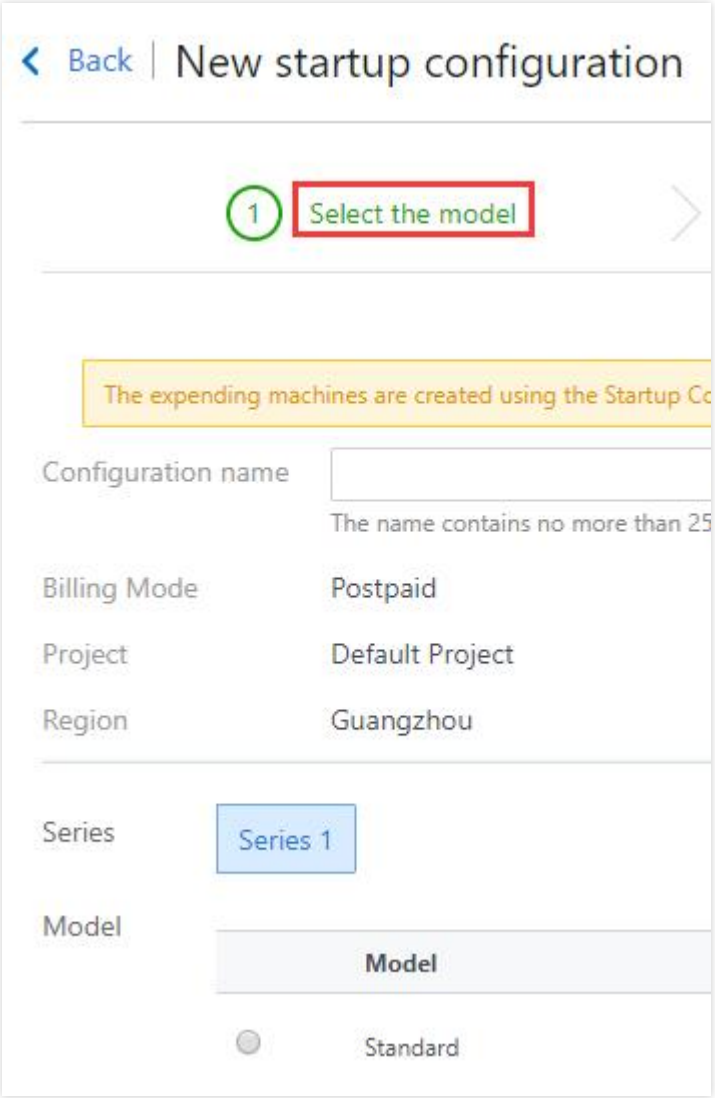
2. Select a project and a region. Note that you must select the project and region of your Web application.



3. The following procedure is similar to purchasing CVMs. You can follow the guidelines to complete the creation of scaling configuration. Please note that you should specify the image you have just created

©2013-2018 Tencent Cloud. All rights reserved.

in the custom image.



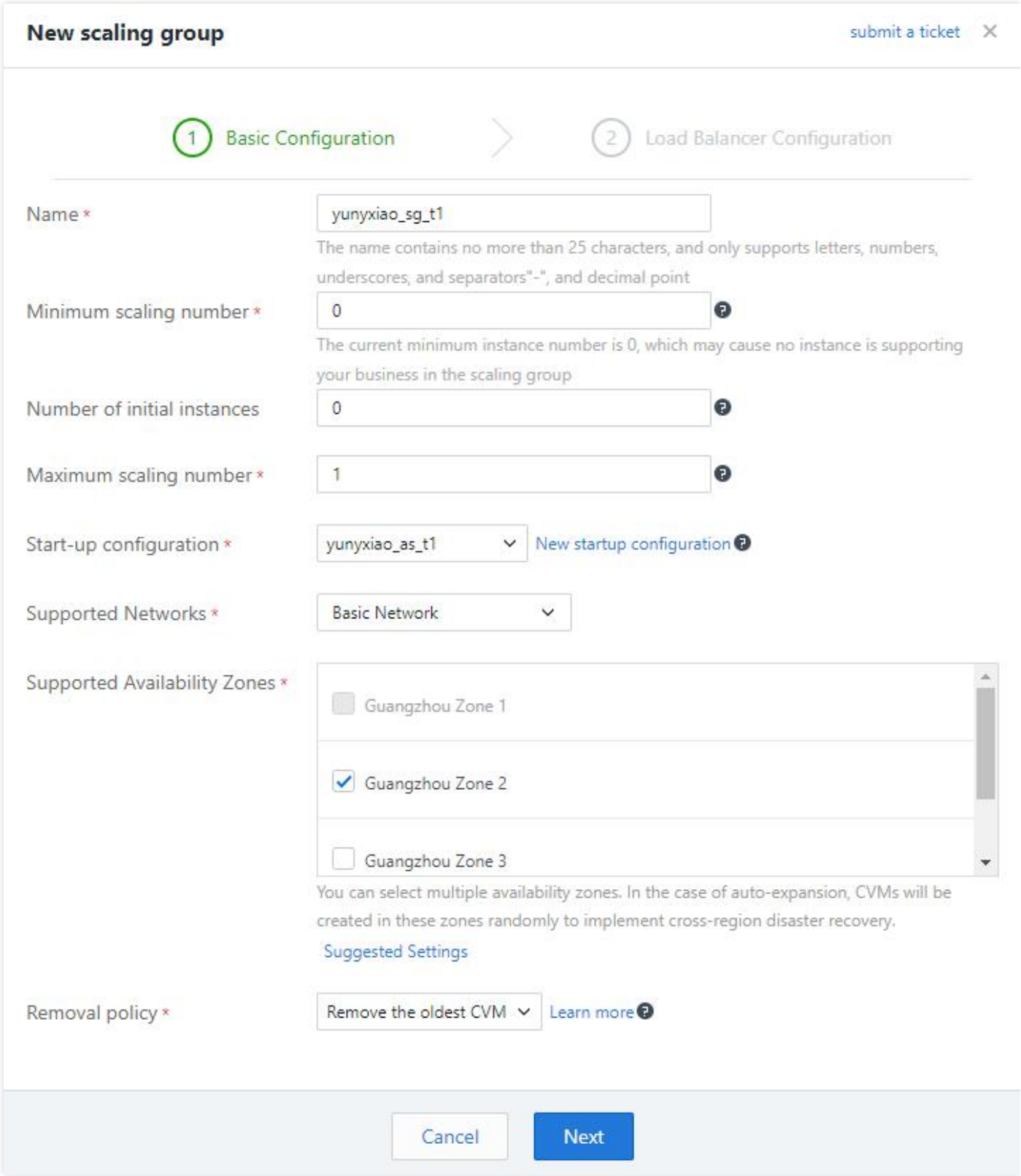
Step 3. Create a scaling group for the CVM

In the Auto Scaling Console, click "New" and fill in the cluster management information as follows:

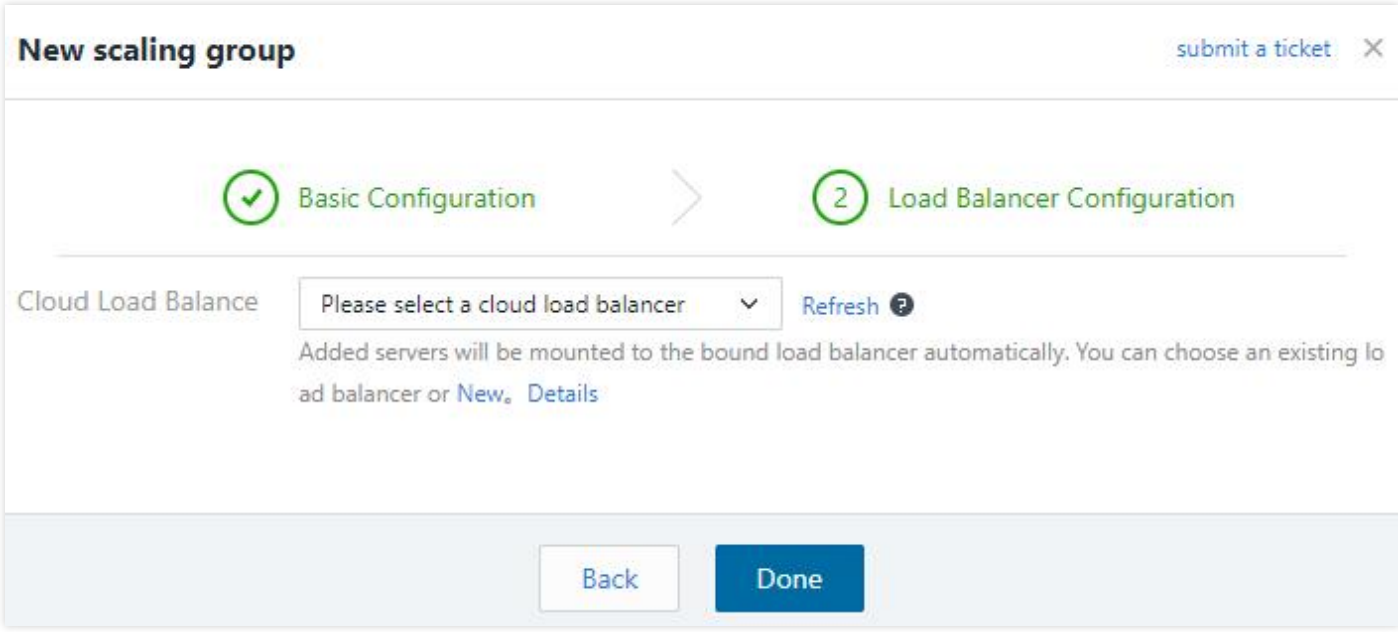
- **Name**: Fill in an appropriate name, such as "Application Server Cluster" in here.
- **Minimum group size**: Lower limit of the number of cluster servers. Fill in 0 in this example.
- **Initial number of instances**: Number of CVMs **automatically created** upon the creation of the scaling group. Generally, no CVM will be created automatically when a scaling group is created. It is recommended to enter 0.
- **Maximum group size**: Upper limit of the number of cluster servers. Fill as needed. We enter 5 here, which means the scaling group has a maximum of 5 CMVs.
- **Scaling Configuration**: Select the scaling configuration you have just created.
- **Supported network**: The network environment of the session server. Select "Basic Network" generally.

©2013-2018 Tencent Cloud. All rights reserved.

- **Supported availability zone**: Select an availability zone for the CVM to be scaled. Check the availability zone of the session server.
- **Remove policy**: Select the default value.
- **Cloud load balancer**: Select the cloud load balancer of the cluster.



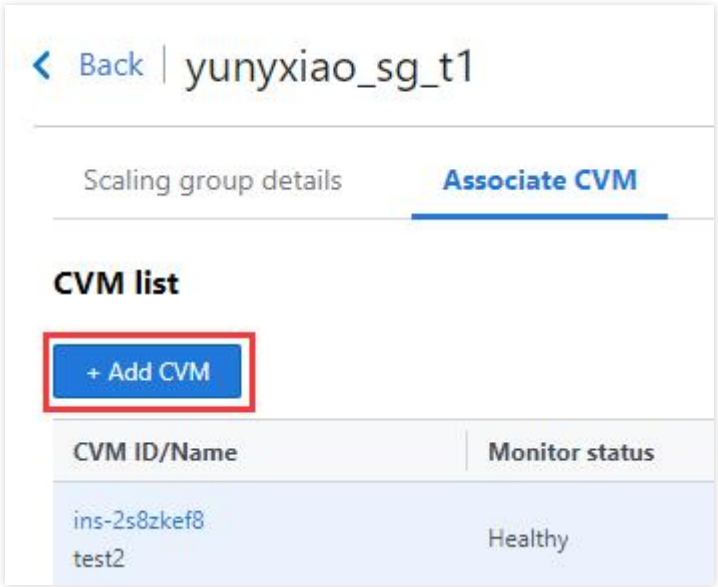
©2013-2018 Tencent Cloud. All rights reserved.



Click **OK** to finish the creation.

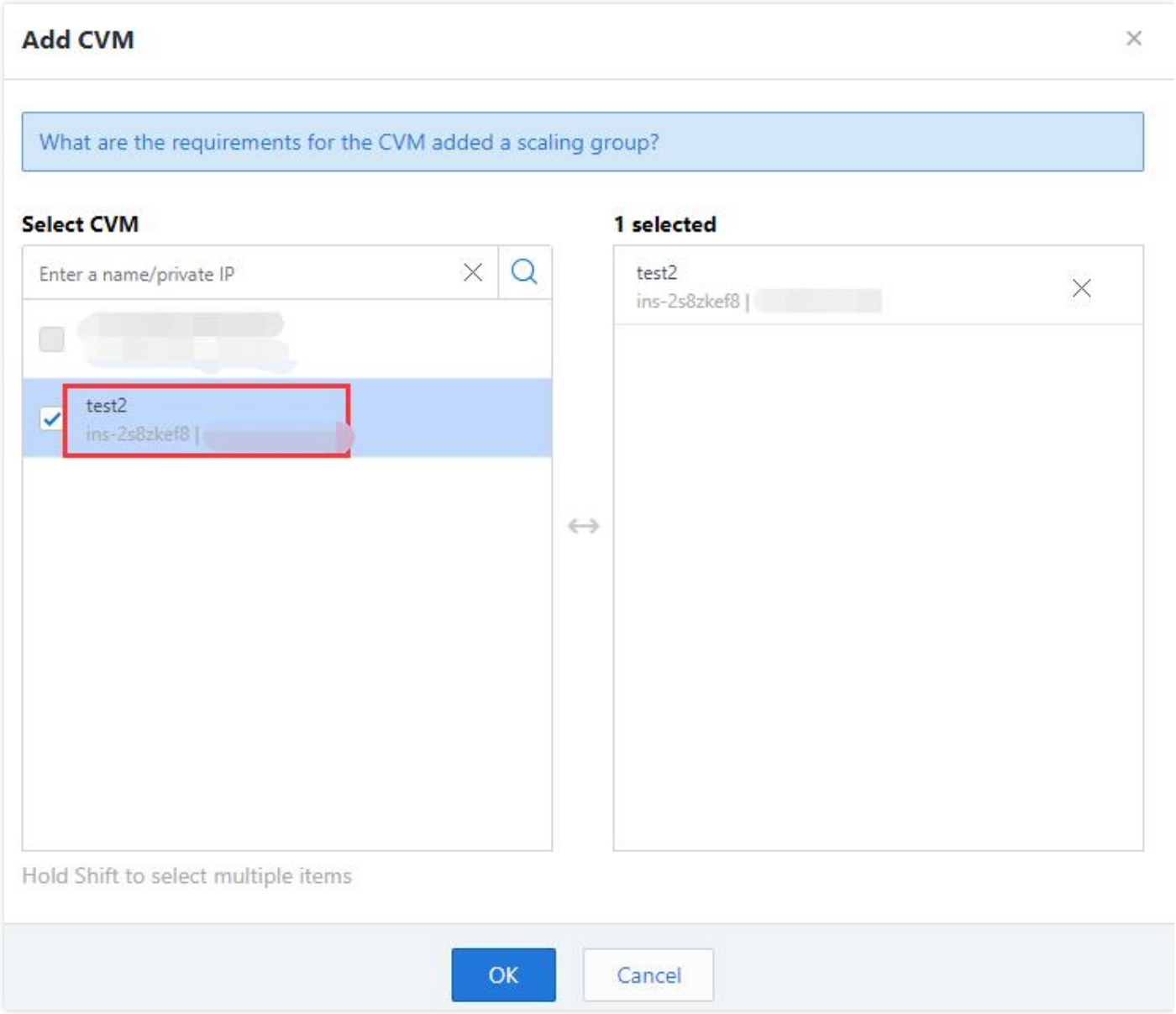
Step 4. Add an existing CVM to the scaling group

1. In the Auto Scaling Console, click on the scaling group name to enter the management page, and click "Add CVM" at the bottom of the page.

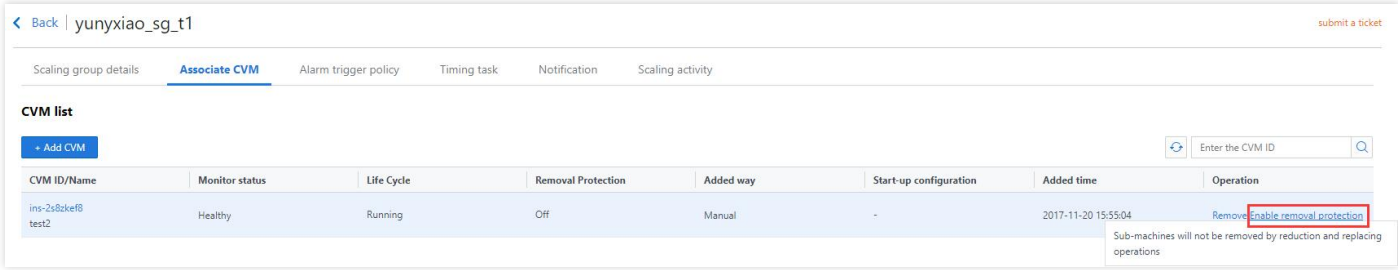


2. In the pop-up dialog box, select an existing server in the cluster to add it to the scaling group. During the off-peak hours, the server that is not fully utilized in the cluster can be returned to save costs.

©2013-2018 Tencent Cloud. All rights reserved.



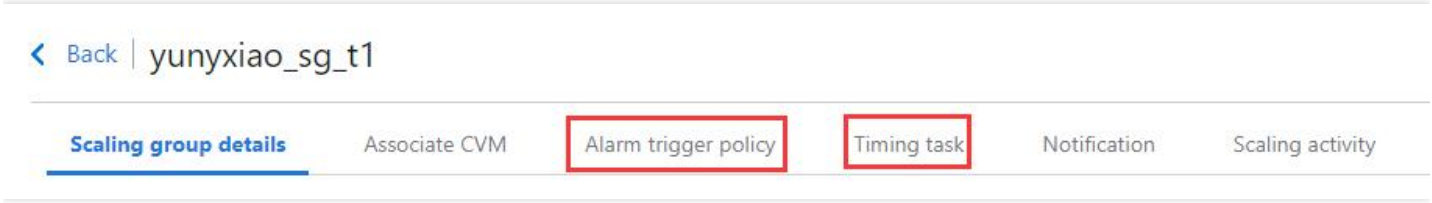
3. Set the added server as "Scale-down exemption", so as to protect it from being removed for scale-down by the scaling group. In this way, this CVM is always in service in the cluster and AS will not change it.



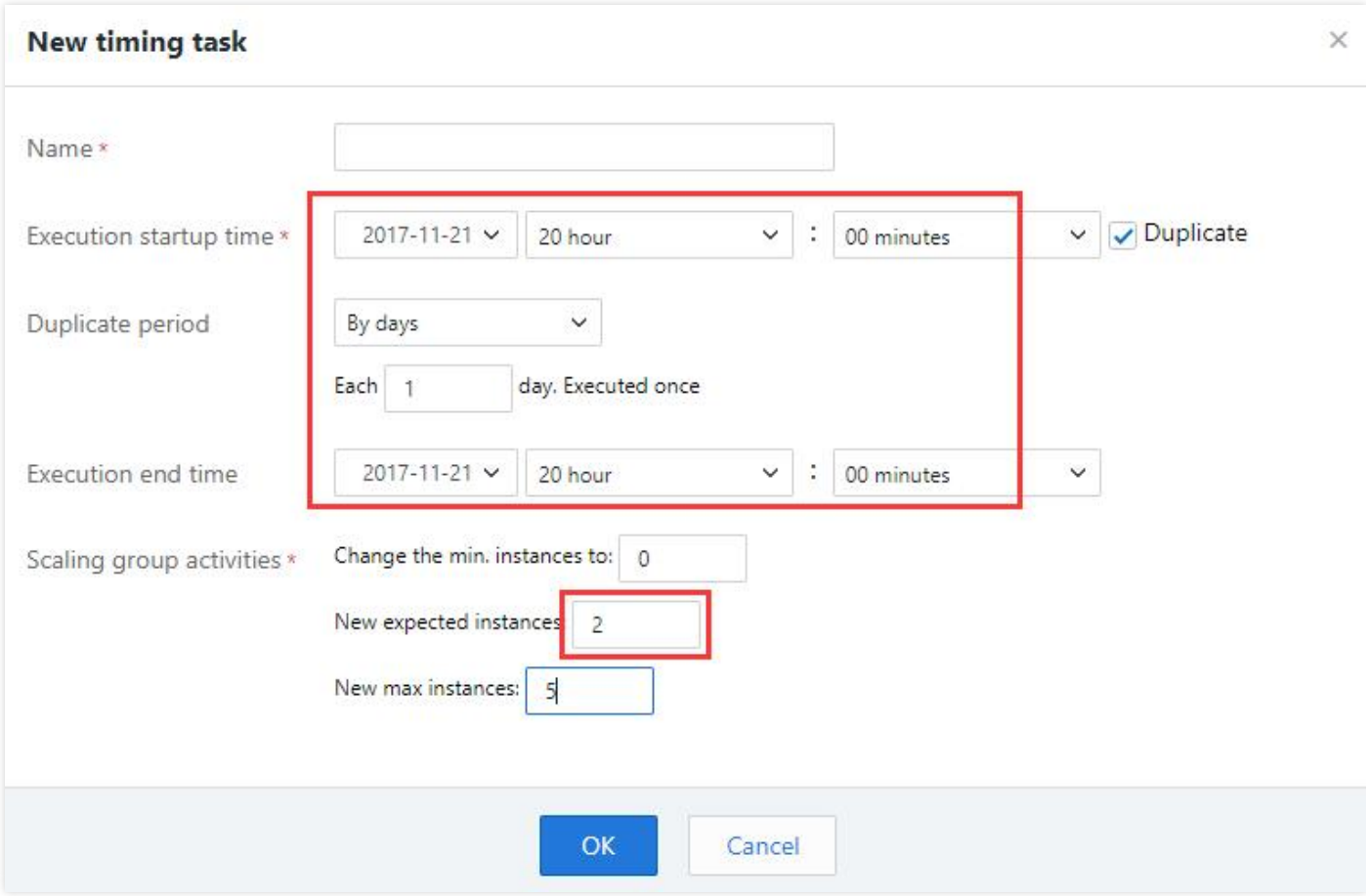
Step 5. Set the scaling policy (important!)

©2013-2018 Tencent Cloud. All rights reserved.

AS supports scheduled scale-up or dynamic scale-up based on the alarm, and can allow you to receive scaling notifications and check the scaling history. Everything is under your control.



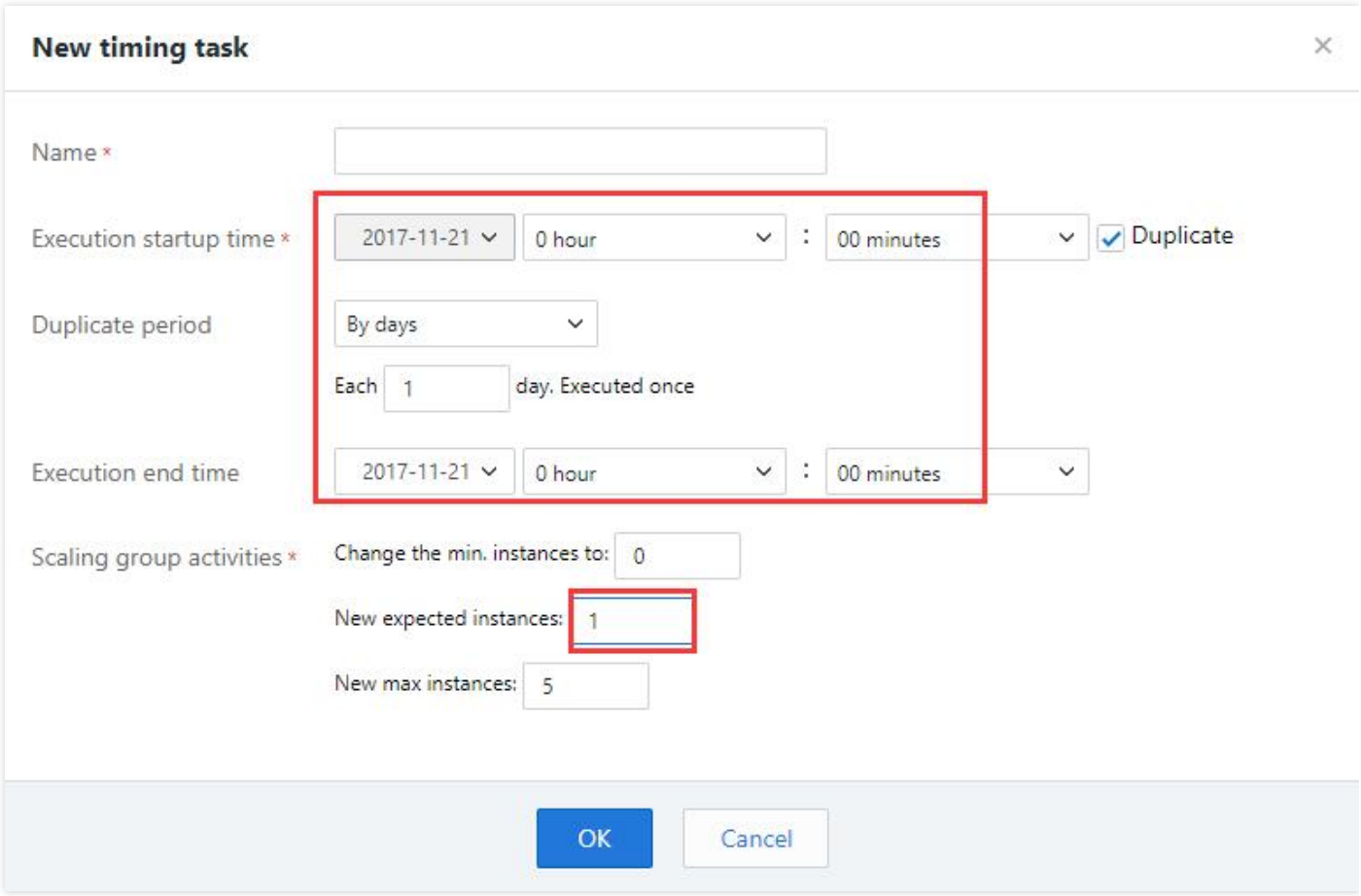
- **First, set a scheduled scale-up task at 20:00**



Note:
It takes about 1 minute to create a Tencent Cloud CVM, and may take more time if the custom image is large. You can set the execution start time 5 minutes earlier.

©2013-2018 Tencent Cloud. All rights reserved.

- **Then, set a scheduled scale-down task at 24:00**



Now you're done!

The backend cluster of the website is changed to " 1 fixed application server + 1 application server that will be created as scheduled during peak hours".
Other cluster CVMs that are not added to the scaling group are not fully utilized most of the time. You can return them for cost saving.

©2013-2018 Tencent Cloud. All rights reserved.

Auto-enable Services on Startup

Last updated：2017-12-06 10:34:36

Set the Service to Be Activated upon Boot

Usage Scenarios

We hope there is no manual intervention in the whole process of scale-up via auto scaling. Thus, we strongly recommend that you set the service to be activated on boot for the machine for auto scaling. The services include:

-**httpd** service
-**mysqld** service
-**php-fpm** service
-**tomcat** service

- etc.

The setting can be done within one minute - Modify the file /etc/rc.d/rc.local!

Setting Method (taking centos as an example):

Step 1: Open file "rc.local"

Input

```
vim /etc/rc.d/rc.local
```

Do not change the existing contents and add contents at the end of the file.

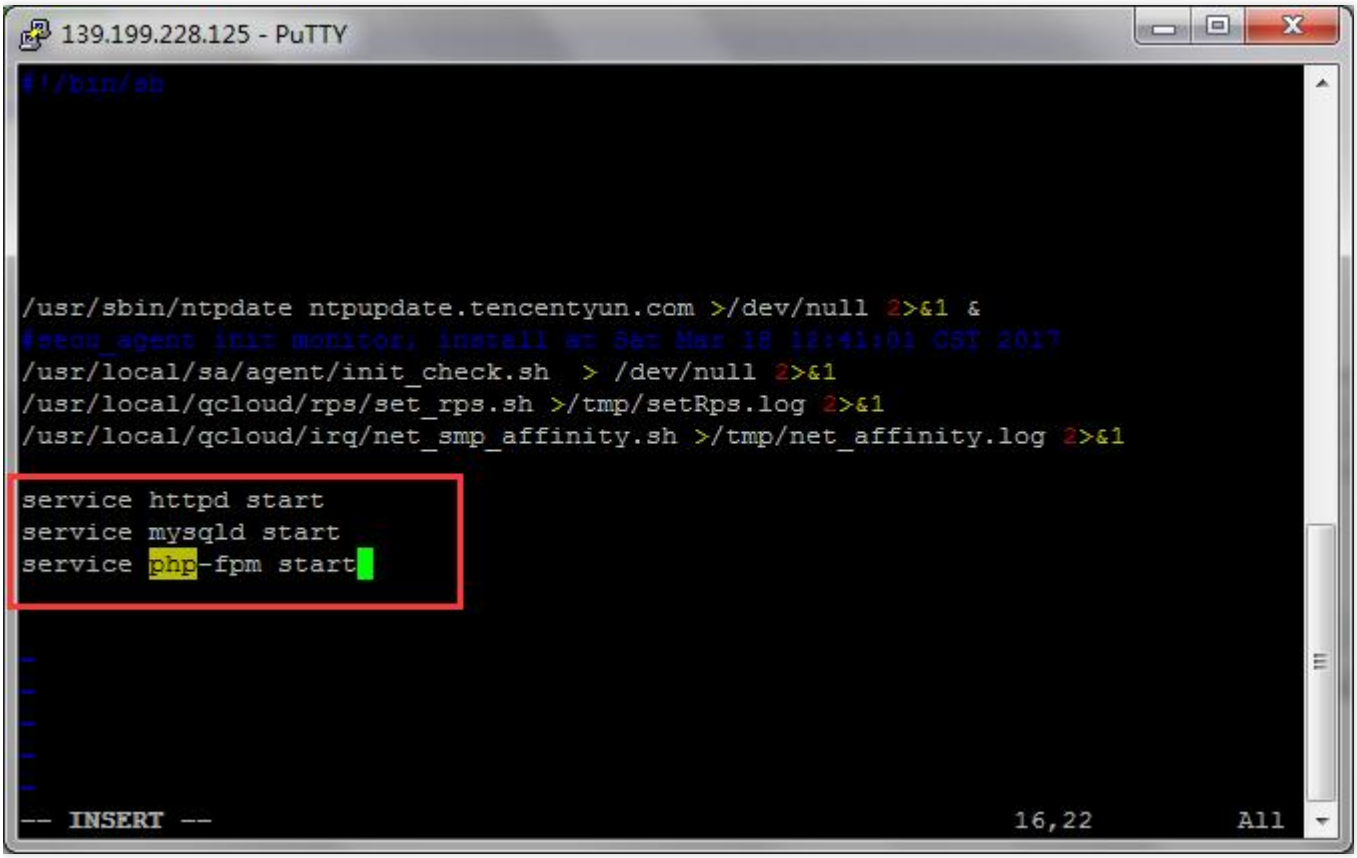
> Here is the detailed operation for starters **(experienced users may skip this)**:
>
> Input "i" to enter the insert mode of vim, and then you can press the direction key "↓" to reach the end of the file and enter the content.

Step 2: Input the service to be activated

This example demonstrates how to set the services httpd, mysqld and php-fpm to be activated upon boot on the built website. So the sentences below are added behind rc.local:

©2013-2018 Tencent Cloud. All rights reserved.

service httpd **start**
service mysqld **start**
service php-fpm **start**



Save and exit. Then, these services will be accessed automatically by the website after the server is booted. Please note that different websites require different services. Set services in this step according to your needs.

Operation TIPS **(experienced users may skip this):**

After entering the content, press the Esc key, **hold down the Shift key and press "Z" twice** to exit, that is, inputting ZZ.

Step 3: Verification (optional)

Reboot the server (Input reboot, or reboot the server in the console). After the server is rebooted, refresh the webpage on the website first without entering the server to see if there is a response. If there is a

©2013-2018 Tencent Cloud. All rights reserved.

response, it means the setting is successful.

Step 4: Create an image based on this machine and use this image when creating scaling configurations

This step is simple. Refer to the guides below if you have problems with the operation:

Create Custom Image

Create Scaling Configurations

©2013-2018 Tencent Cloud. All rights reserved.