Server Load Balancer

Quick Start

Quick Start

Overview

Server Load Balancer is a service that distributes traffic among multiple ECS instances. It can extend the service capability of an application system by distributing traffic and enhance availability by eliminating SPOFs.

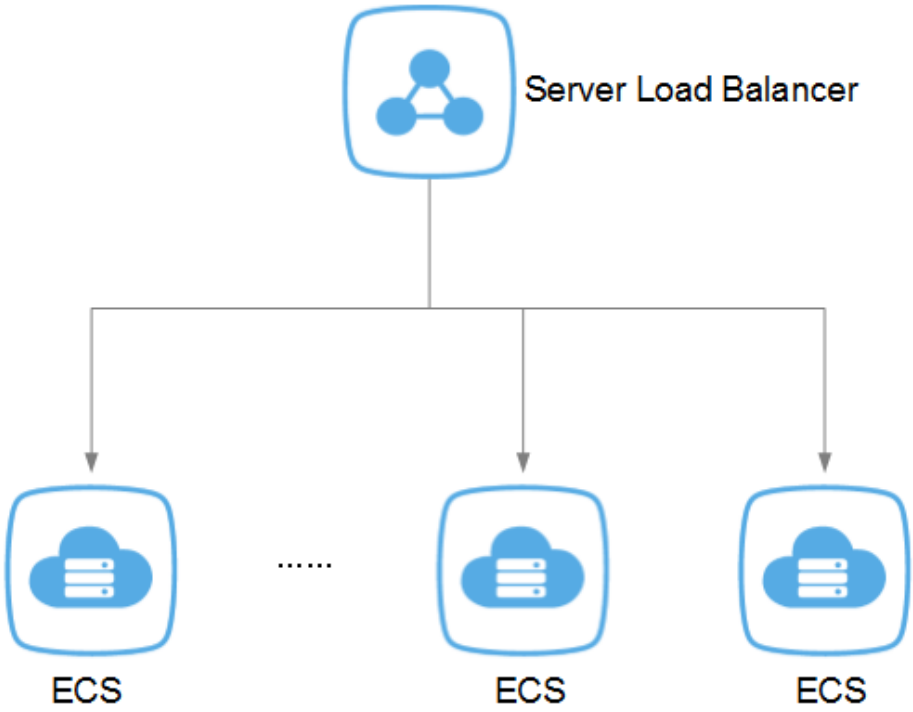
This guide aims to explain how to quickly use the Server Load Balancer service to balance the traffic load of multiple hosts. The main process is shown in the diagram below.

Data planning

Application scenarios

Before constructing the service, the user must have a clear understanding of his or her needs and make appropriate preparations.

For example, a local Beijing BBS forum would need a public VIP, Server Load Balancer and the company’ s ECS instance mounted at the backend.



Data preparation

The user must have an Alibaba Cloud website account.

Unregistered users should refer to Account Registration.

Confirm the service region.

Alibaba Cloud provides the Server Load Balancer service for the following regions: Singapore, China North 1 (Qingdao), US East 1 (Virginia), China North 2 (Beijing), China East 2 (Shanghai), China South 1 (Shenzhen), Hong Kong, US West 1 (Silicon Valley), China East 1 (Hangzhou). To reduce latency time and accelerate download speeds, we suggest choosing the nearest Server Load Balancer node.

Confirm the instance type: public network or private network.

Based on actual business needs, select an appropriate instance type. If you require public network access, select a public network instance; if you only need access to the intranet, select a private network instance.

Confirm the billing method: PayByTraffic or PayByBandwidth.

If you select a Server Load Balancer instance on the public network, select a suitable billing method based on the actual characteristics of your business. For businesses with significant

traffic fluctuations, we suggest PayByTraffic; for businesses with relatively stable bandwidth usage, we suggest PayByBandwidth. Of course, you may select either billing method based on your own preferences.

Create ECS instances

Application scenario

Based on their business needs, users can create ECS instances at any time. For information on specific applications, refer to Typical Application Scenarios.

Considerations

- Because Server Load Balancer does not support cross-regional deployment, you should select an ECS instance in the same region as the Server Load Balancer service.
- Select the CPU and memory options that suit your actual needs. Please note that Windows OS images cannot be used for some CPU and memory combinations.
- Select the public bandwidth appropriate to your actual needs.
- Alibaba Cloud provides a variety of operating systems, including Ubantu, OpenSUSE, CoreOS, Debian, CentOS, Alibaba Cloud Linux, and Windows Server, and FreeBSD. You can select an OS officially provided by Alibaba Cloud based on your actual needs. You can also use an OS image from the image market.
- You can choose to add data disks as needed. Please note that data disks added here cannot be detached.
- The password and ECS instance name set when creating the ECS instance allow you to conveniently log into and manage the purchased ECS instance.

Operation procedure

- For instructions on the creation of Windows instances, refer to Create an instance running Windows.
- For instructions on the creation of Linux instances, refer to Create an instance running Linux.

View instances

After creating an ECS instance, go to the “Management Console” and click “ECS” under “My Products and Services” to go to the ECS management console.

Select “Instances” in the left-side navigation bar to go to the “Instance List”.

Select the relevant region.

View the created instance.

Configure ECS instances

After creating an ECS instance, you need to configure this instance before using it. Below is the basic method of configuring ECS instances.

Configure the Web Server

Suppose we only need to create a simple static page. We will use Apache to configure an Alibaba Cloud Linux ECS instance in a Mac OS environment. The procedure is as follows:

Connect to the ECS instance

First, start the "Terminal" service on the computer. Enter the following command to connect to the ECS instance:

```
ssh root@xxx.xxx.xxx.x05
```

xxx.xxx.xxx.x05 indicates the public IP address of the ECS instance we need to connect to.Next, enter the password for this instance to establish a connection.

Install Apache

After establishing a connection with the ECS instance for the first time, enter the following command to install and configure Apache:

```
yum install httpd*
/etc/init.d/httpd start
```

After that, enter the following command to check the port status:

```
/etc/init.d/httpd status
netstat -an |grep 80
```

If the status is OK, start to edit the static webpage for this ECS instance.

Edit the static webpage

Enter the following command to edit the static webpage for this ECS instance:

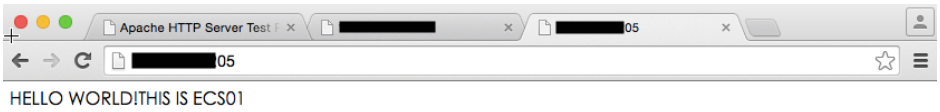
```
cd /var/www/html/
vim index.html
```

After editing, use the following command to view the edited static webpage:

```
cat index.html
```

Access the static webpage

Now, a static webpage has been generated for the ECS instance. You can access this page by entering the public IP address of this instance in your browser.



As shown in the figure above, entering instance ECS01's public IP address (xxx.xxx.xxx.x05) brings you to the static webpage under ECS01.

As shown in the figure above, entering instance ECS02's public IP address (xxx.xxx.xxx.x83) brings you to the static webpage under ECS02.

One-key configuration and installation package

To create a more complex dynamic webpage, you can install more components including Apache. To facilitate your configuration, the Alibaba Cloud website currently provides one-key configuration and installation packages for Windows and Linux operating systems. These are free, open-source installation packages provided by Cloud Technologies in Shanghai. They are safe, stable, and efficient.

Create Server Load Balancer instances

Application scenarios

If users have the following needs, they can solve them by creating a Server Load Balancer instance:

- Flexible traffic distribution, suitable for businesses with high access volumes.
- Horizontally scaling application system service capabilities, suitable for a variety of Web servers and apps with elastic traffic needs.
- Eliminating application system SPOF so that, even if some ECS instances go down, the application system will continue to function normally.
- Improving application systems' disaster tolerance capabilities through multi-zone deployment so that, even if a machine room goes down, the system will continue to function normally.
- Preventing attacks on the application system, suitable for businesses that often suffer WAF and CC troubles.

Considerations

- As Server Load Balancer does not support cross-regional deployment, you should select the same region as the backend ECS instance.
- There are two different types of Server Load Balancer instances: Server Load Balancer instances on the public and private networks.
- The Server Load Balancer service provides two different billing methods: PayByTraffic and PayByBandwidth.

Operation procedure

Log onto the Server Load Balancer Management Console.

Select the same region as the ECS instance and click **Create Server Load Balancer Instances** in the top-right corner.

Select the appropriate region, primary and backup zones, instance type, public bandwidth, and quantity.

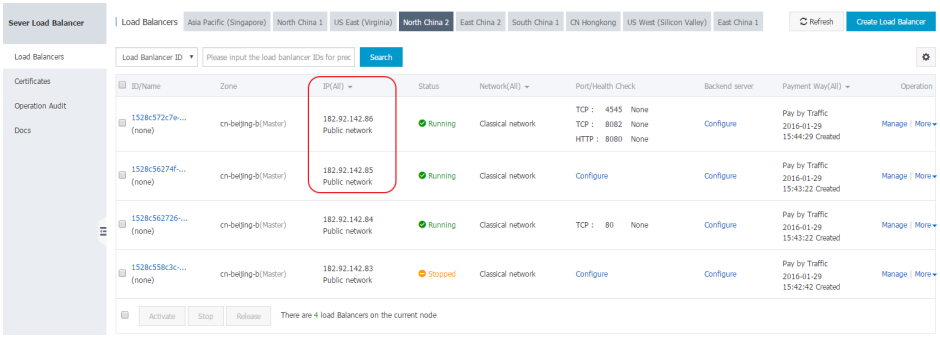
Click **Buy Now** to go to the “Confirm Order” page.

Check the configuration information and then click **Activate**.

View Server Load Balancer instances

Go to the “Management Console” and click “Server Load Balancer” to go to the “Instance Management” page.

Select the relevant region to view the created instances.



As shown in the image above, the user has created a Server Load Balancer instance on the public network in the North China 2 (Beijing) region. You can also see the name, billing method, port/health check setting, backend server setting, status, zone, network type, IP address, and other information for the created Server Load Balancer instance.

Configure Server Load Balancer instances

After creating a Server Load Balancer instance, you must configure it before you can use it. Configuring and managing a Server Load Balancer instance primarily involve three operations: configuring Server Load Balancer instance attributes, configuring Server Load Balancer monitoring, and configuring the backend ECS instance for the Server Load Balancer instance.

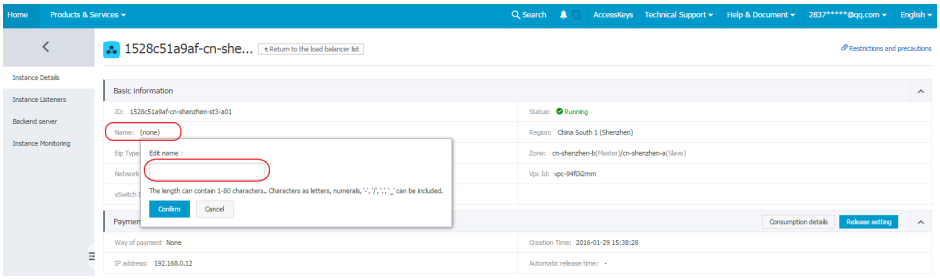
Configure Server Load Balancer instance attributes

Go to the management console and select the “Server Load Balancer” service.

On the “Server Load Balancer” instance list page, select a Server Load Balancer instance.

Click **Manage** to the right to go to the basic information page.

Edit the name of the created Server Load Balancer instance.



As shown in the figure above, we name the created Server Load Balancer instance “sarah.cxj-lb02”.

Configure Server Load Balancer monitoring

Click the “Manage” link of the Server Load Balancer instance to go to the Server Load Balancer instance details page.

Click “Monitoring” in the left-side navigation bar to go to the “Monitoring” page.

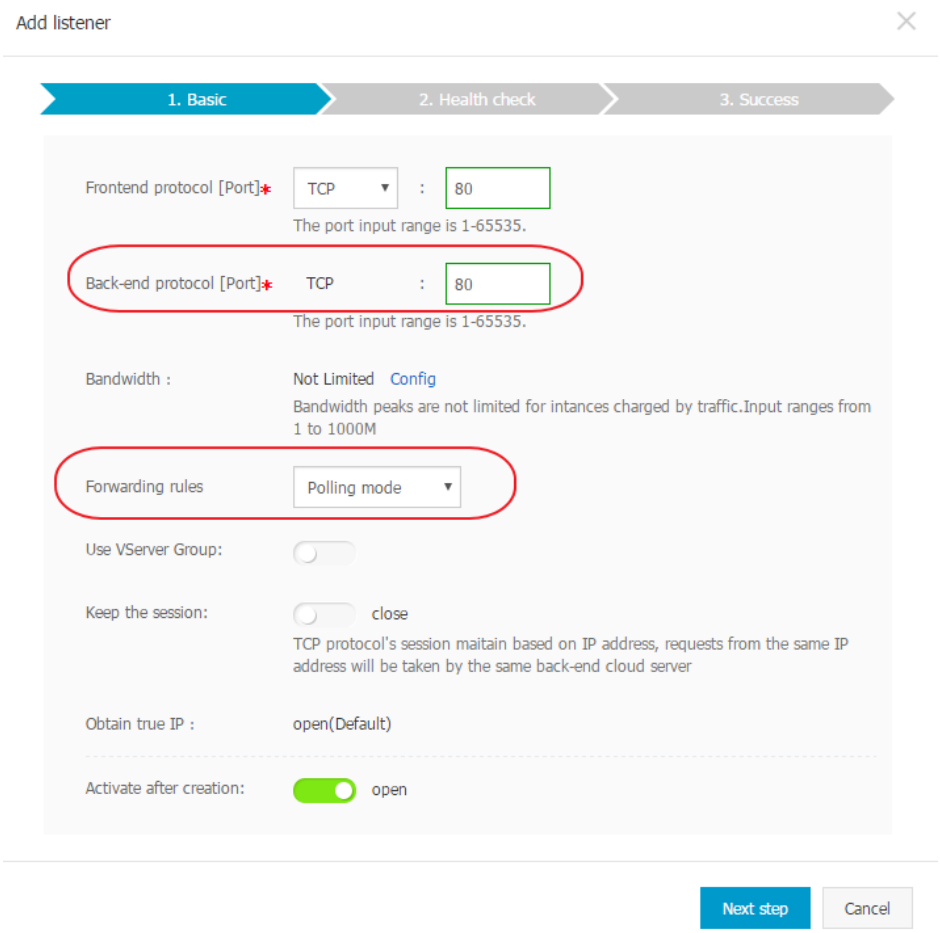
In the “Monitoring Configuration” tab, click **Add Listener** in the top-right corner to bring up the “Add Listener” dialog box.

NOTE:

- One Server Load Balancer instance can have up to 50 listeners.
- Currently, Server Load Balancer is provided for both Layer 4 (TCP, UDP protocol) and Layer 7 (HTTP, HTTPS protocols).

Select TCP protocol and set the listener’s backend ECS port to 80.

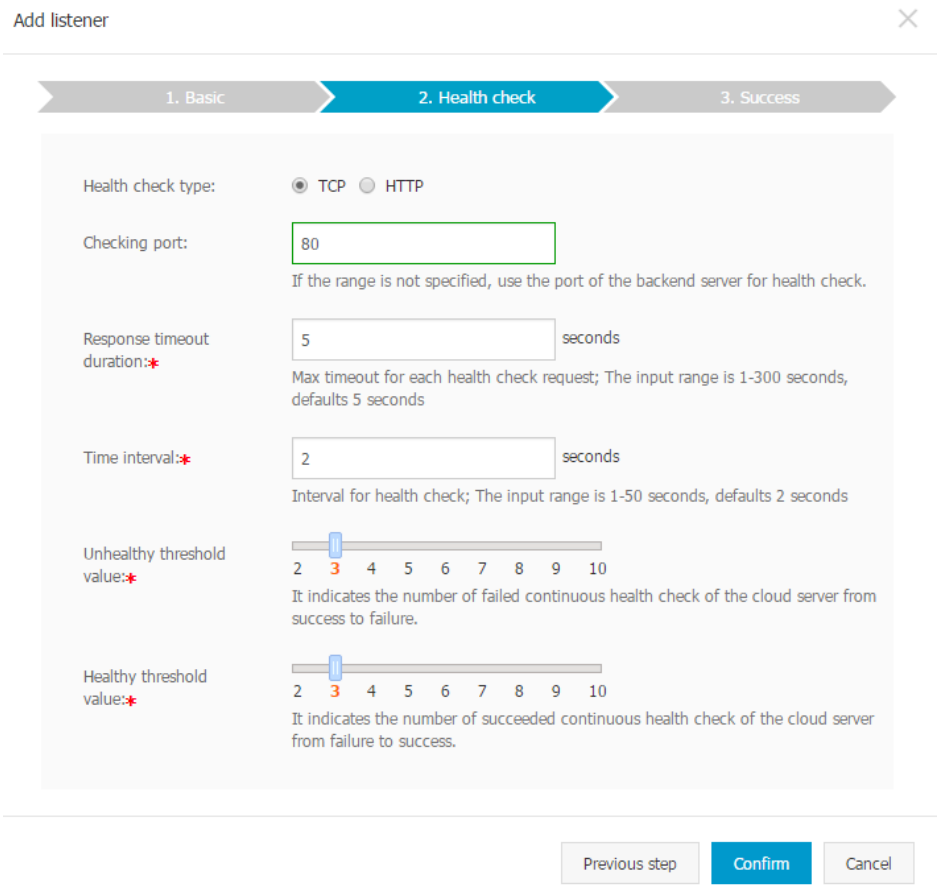
Note: Because only Apache configuration was performed on the previously created ECS instances, we will configure Server Load Balancer service monitoring for Layer 4 (TCP Protocol) in this example.



You may choose from the two forwarding rules “Round Robin” and “Least Connections” based on your needs. External access requests will be forwarded according to your selection.

- In “Round Robin”, external and internal access requests are distributed to the backend ECS instances in sequence for processing.
- In “Least Connections Mode”, these requests will be distributed to the backend ECS instance with the least number of connections for processing.

Click **Next** to go to the health check configuration page.



For health check parameter configuration, we recommend the following for your reference:

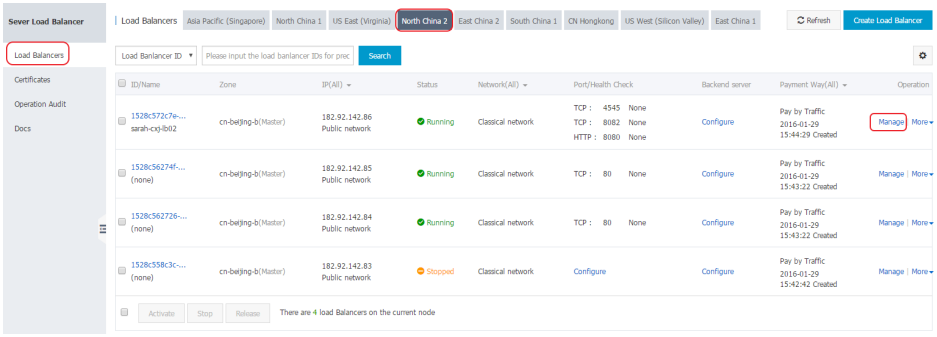
Response timeout: 5 seconds
Health check interval: 2 seconds
Unhealthy threshold value: 3
Healthy threshold value: 3

This configuration facilitates the rapid convergence of user services and application statuses:
ECS health check failure response time: 2×3+5=11 seconds
For faster response, lower the response timeout value. However, ensure that your service is normally processed within a timeframe less than this value.
ECS health check success response time: 2×3=6 seconds

Backend ECS configuration

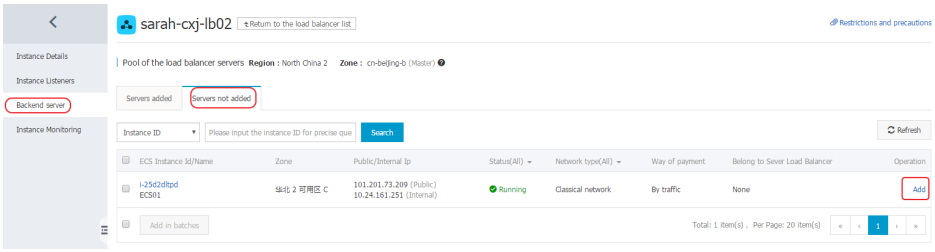
Log onto the Server Load Balancer management console.

Click the “Manage” link of the Server Load Balancer instance to go to the Server Load Balancer instance details page.

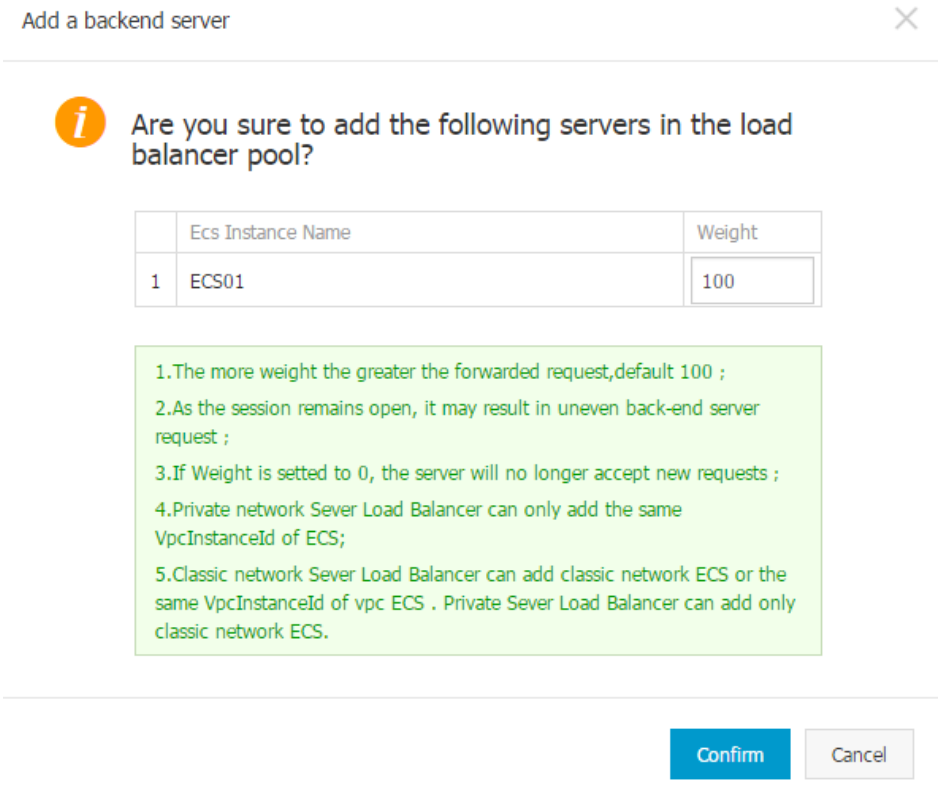


Select “Servers” in the left-side navigation bar to go to the “Server Load Balancer Server Pool”.

Select the “Excluded Servers” tab to go to the relevant ECS list page.

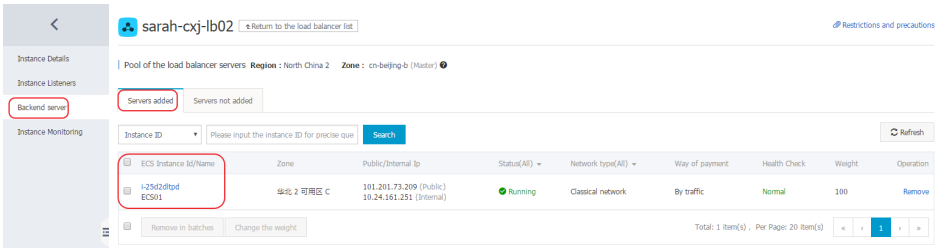


Select an ECS instance and click the **Add** button to its right to bring up the “Add Backend Server” dialog box.



Set the weight. If the weight of each backend ECS instance is set to 100, this means the average forwarding rule is adopted.

As shown in the image below, the two backend ECS instances we have added are now displayed in the “Included Servers” tab.

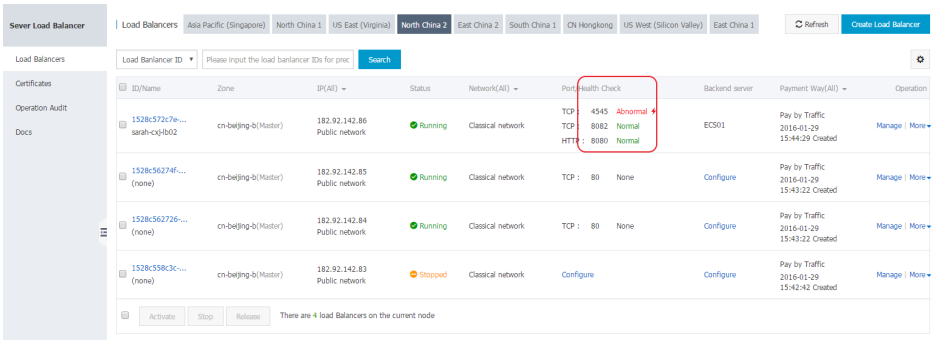


Check health status

Log onto the Server Load Balancer management console.

Click “Instance Management” in the left-side navigation bar to go to the instance management page.

Select the appropriate region and check if the “Port/Health Check” status for the instance in question is “Normal”.



Confirm the status is “Normal” and proceed to the next task.

Verify Server Load Balancer

Prerequisites

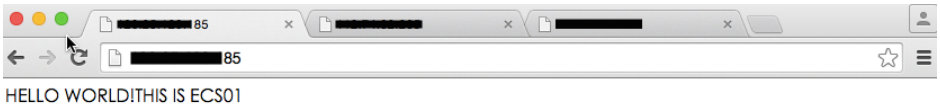
Follow the procedures above to complete basic Server Load Balancer instance configuration.

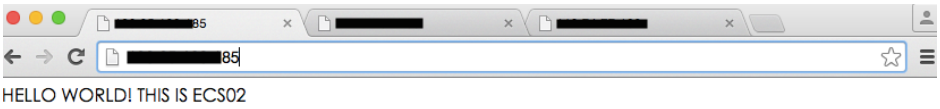
Operation Procedure

Enter the public IP address for this Server Load Balancer instance in your browser.

View the test results to see if this Server Load Balancer instance has been configured successfully.

As shown in the two figures below, by entering the Server Load Balancer instance’s public IP address (xxx.xxx.xxx.x85), you will randomly access the static webpage of one of the two backend ECS instances.





Domain Name resolution

Application scenario

In order to properly work with Server Load Balancer on the Internet for applications that offer external service via DNS, you must resolve the domain name to the IP address of the corresponding Server Load Balancer instance on the website of the domain name registration service provider. You can access the service via DNS only after successful domain name resolution.

For example, assume that an ECS instance has the domain name “taobao.com”, the domain name is resolved to the public IP address 1.1.1.1, and the Server Load Balancer instance’s public IP address is 2.2.2.2. In this case, you must resolve the domain name “taobao.com” to the public IP address 2.2.2.2 to allow users to access it through DNS.

Generally, this is accomplished through A-record resolution.

Operation procedure

You can use Alibaba Cloud DNS to resolve the domain name. For details, refer to Add resolution records.

Delete Server Load Balancer instances

Application scenario

When a user no longer requires a Server Load Balancer instance, it can be deleted.

Note: Deleting a Server Load Balancer instance will not affect the ECS instance at the backend.

Considerations

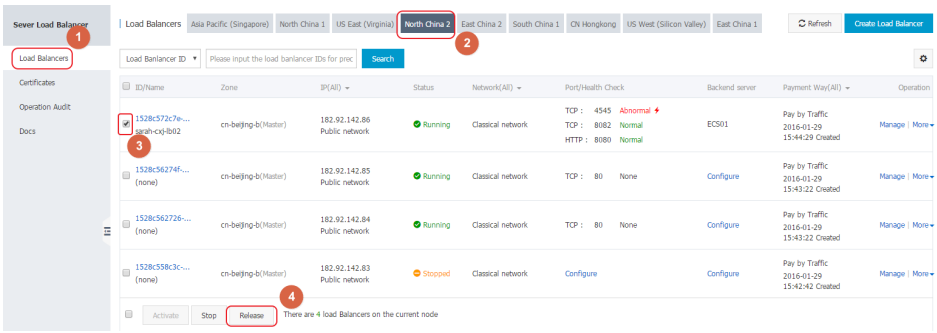
Before deletion, the domain name must be resolved to a new IP address to ensure the business will not be interrupted.

Operation procedure

If you are directing traffic to the Server Load Balancer domain name' s CNAME record, you must resolve the domain name to a new IP address before deleting the Server Load Balancer instance.

Log into the Server Load Balancer console.

Click "Instance Management" in the left-side navigation bar to go to the "Instance Management" page.



Select the region and corresponding Server Load Balancer.

Click the **Release Settings** button to bring up the "Release Settings" dialog box.

Select **Immediate Release** or **Timed Release** as needed.

- If you select **Immediate Release**, just click **Next**.
- If you select **Timed Release**, select the automatic release time and then click **Next**.

On the popped-up release settings confirmation page, click **OK** to go to the cell phone verification stage.

Enter the cell phone verification code and then click **OK** to release the Server Load Balancer.

Subsequent processing

After the Server Load Balancer instance is released, the associated backend ECS instance will continue to run and you will be charged for its operations. This is only applicable to the release of Pay-As-You-Go instances. For specific operations, refer to Release an instance and disabling auto release (for Pay-As-You-Go users).