Server Load Balancer

Quick Start

Quick Start

This section provides a complete tutorial on using Server Load Balancer. An Internet-facing Server Load Balancer instance will be created to distribute the received HTTP requests to the backend servers.

> **Note**: Before creating a Server Load Balancer instance, you need to plan and design your load balancing service, such as the instance type, instance region, and more. For details, see Plan and prepare.

The tutorial includes the following tasks:

Create ECS instances

Server Load Balancer is a complementary service for ECS multi-machine solutions，and must be used in conjunction with ECS. In this tutorial, two ECS instances are created to process the distributed traffic.

Install web pages

Create required applications on the ECS instances. In this tutorial, a static web page is created to test the load balancing service.

Create a Server Load Balancer instance

A Server Load Balancer instance is a running entity of Server Load Balancer. In this tutorial, an Internet-facing Server Load Balancer instance is created.

Configure the Server Load Balancer instance

After creating a Server Load Balancer instance, you have to add at least one listener, and multiple ECS instances as backend servers. In this tutorial, a TCP listener is added, and the ECS instances created in task 1 are used as backend servers.

Delete the Server Load Balancer instance

If you no longer need Server Load Balancer, delete it to avoid additional charges.

Plan the region of the Server Load Balancer instance

Alibaba Cloud provides the Server Load Balancer service in various regions.

To provide more stable and reliable load balancing services, Server Load Balancer has deployed multiple zones in most regions for better disaster tolerance. Additionally, to improve high availability in different regions, you can deploy the Server Load Balancer instances in multiple regions and use DNS to resolve the domain name to the IP addresses of the Server Load Balancer instances.

When selecting the region, note the following considerations:

- To reduce latency and increase the download speed, we recommend choosing a region that is physically closest to where your customers are located.

- Server Load Balancer does not support the cross-region development. Ensure that the region is the same for the Server Load Balancer and the backend ECS instances.

Plan the instance type (Internet or intranet)

Choose the instance type according to your business. After you create a Server Load Balancer instance, a private or public IP is allocated. You can resolve a domain name to the IP to provide services.

- The Internet Server Load Balancer instance only has a public IP, which is accessible from the Internet.

  If you choose the Internet type, you need to consider the billing method:

  - By traffic：Suitable for an application with obvious traffic changes.

  - By bandwidth：Suitable for an application with relatively stable bandwidth.

- The intranet Server Load Balancer instance only has a private IP, which is accessible only from classic network or VPC, and cannot be accessed by the Internet.

Plan the listening protocol

Server Load Balancer supports layer-4 (TCP and UDP) and layer-7 (HTTP and HTTPS) listening.

Layer-4 listener distributes connection requests directly to the backend servers without modifying the HTTP headers. After the request arrives at the listener, the Server Load Balancer server uses the backend protocol port configured in the listener to create a TCP connection with the backend ECS.

Layer-7 listener is an implementation of reverse proxy. After the request arrives at the listener, the Server Load Balancer server uses the TCP connection to transfer the data packets instead of transferring the data packets directly to the backend ECS.

Prepare the backend servers

Before using Server Load Balancer, you have to create ECS instances and build corresponding applications. Then, add the ECS instances to a Server Load Balancer instance as the backend servers to process the distributed requests.

ECS region

Ensure the region is the same for the ECS instance and Server Load Balancer instance. Also, we recommend deploying the ECS instances in different zones to improve availability.

ECS configurations

Additional configuration is not required after deploying the applications. However, if you are going to create a layer-4 listener, and the ECS instances use the Linux operating system, ensure the values of the following parameters in the net.ipv4.conf file are 0:

```
net.ipv4.conf.default.rp_filter = 0
net.ipv4.conf.all.rp_filter = 0
net.ipv4.conf.eth0.rp_filter = 0
```

ECS amount

There is no restriction on the number of ECS instances added to a Server Load Balancer instance. To improve the stability and efficiency of the service, we recommend adding the ECS instances responsible for different tasks or provide different services to different Server Load Balancer instances.

Before using Server Load Balancer, you have to create at least two ECS instances and build corresponding applications. Then, add them to the Server Load Balancer instance to process the distributed client requests.

Follow the instructions in this document to create two ECS instances, ECS01 and ECS02.

Procedure

Log on to the ECS console.

In the left-side navigation pane, click **Instances** and then click **Create Instance**.

On the buy page, configure the ECS instance.

The following are ECS settings used in this tutorial. For more details, refer to Create Linux ECS instances.

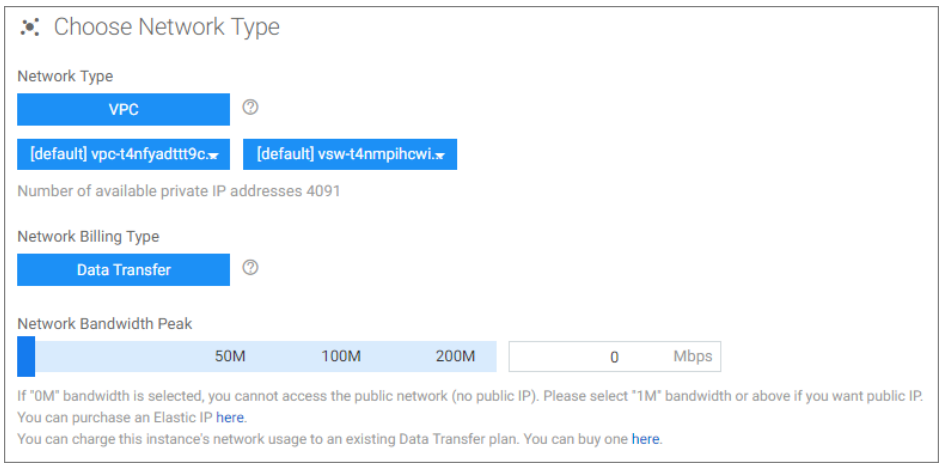
Region: In this tutorial, select **China East 1**.

> **Note**: Server Load Balancer does not support cross-region deployment. The region must be the same for the Server Load Balancer instance and the ECS instance.

Network Type: In this tutorial, select **VPC**. Use the default VPC and VSwitch.

Operating System: In this tutorial, select **Ubuntu 16.04 64 bit**.

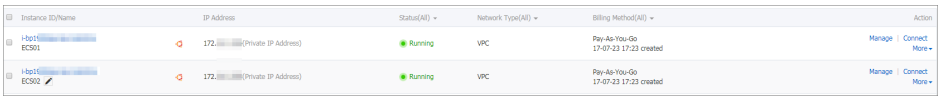
Number of Instances: In this tutorial, select **2**. The system will simultaneously create two ECS instances with identical settings.



Click **Buy Now** and complete payment.

Go back to the **ECS Instance List** page and click **China East 1**.

Hover the mouse pointer over the instance name and click the displayed pencil icon to change the instance name to ECS01 and ECS02, separately.

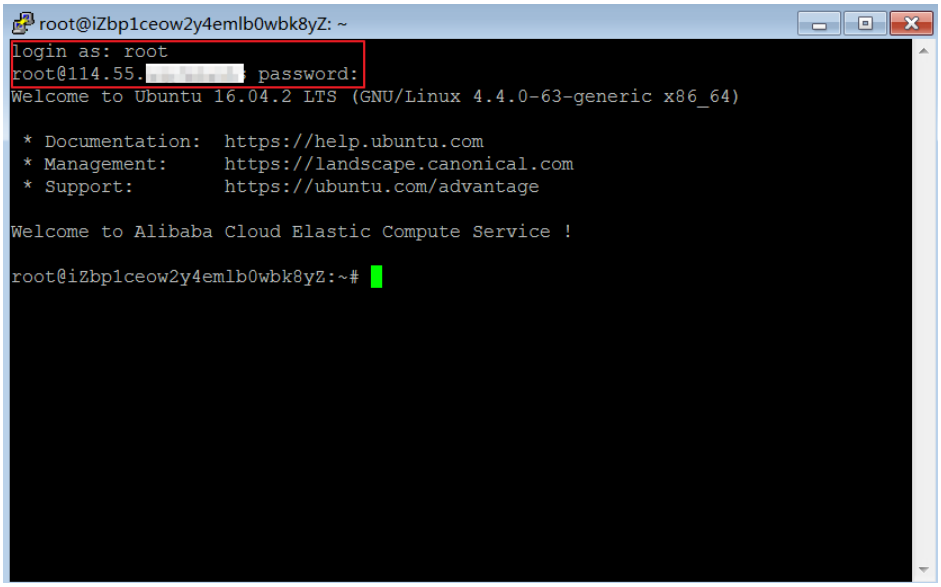


After you create the ECS instances, you need to deploy applications. In this tutorial, two static web pages will be deployed on the ECS instances using Apache.

> **Note**: We use the default settings of Apache and only modify the content of the index file. Additionally, two Elastic IPs are bound to the ECS instances for easy management. For details, refer to Bind an EIP.

Procedure

Log on to the ECS instance.



Enter the following command to install Apache.

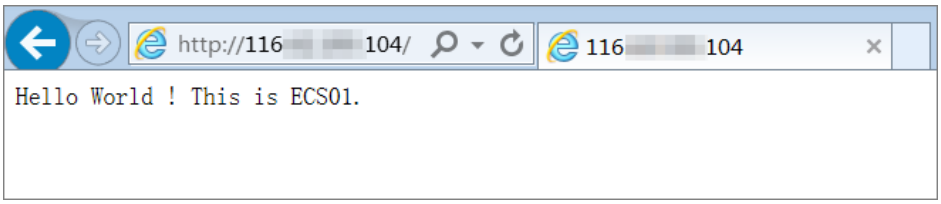
sudo apt-get install apache2

Enter the following command to modify the content of the index.html file.

cd /var/www/html

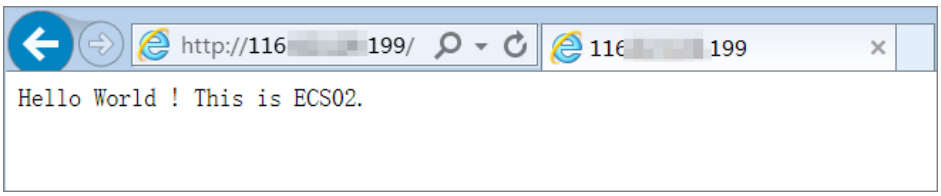
echo "Hello World! This is ECS01." > index.html

After modifying the content, enter the Elastic IP of the ECS instance in the web browser, you will see the following content.



Repeat the previous steps to create a web page on the another ECS instance and change the content to Hello World! This is ECS02..

After modifying the content, enter the Elastic IP of the ECS instance in the web browser, you will see the following content.



Before using Server Load Balancer, you need to create a Server Load Balancer instance. You can add multiple listeners and backend servers to the Server Load Balancer instance.

Follow this document to create an Internet-facing Server Load Balancer instance. After creating the instance, a public IP is allocated to it and resolve a domain name to this IP.

Procedure

Log on to the Server Load Balancer console.

On the **Instance Management** page, click **Create Server Load Balancer**.

Configure the Server Load Balancer instance.

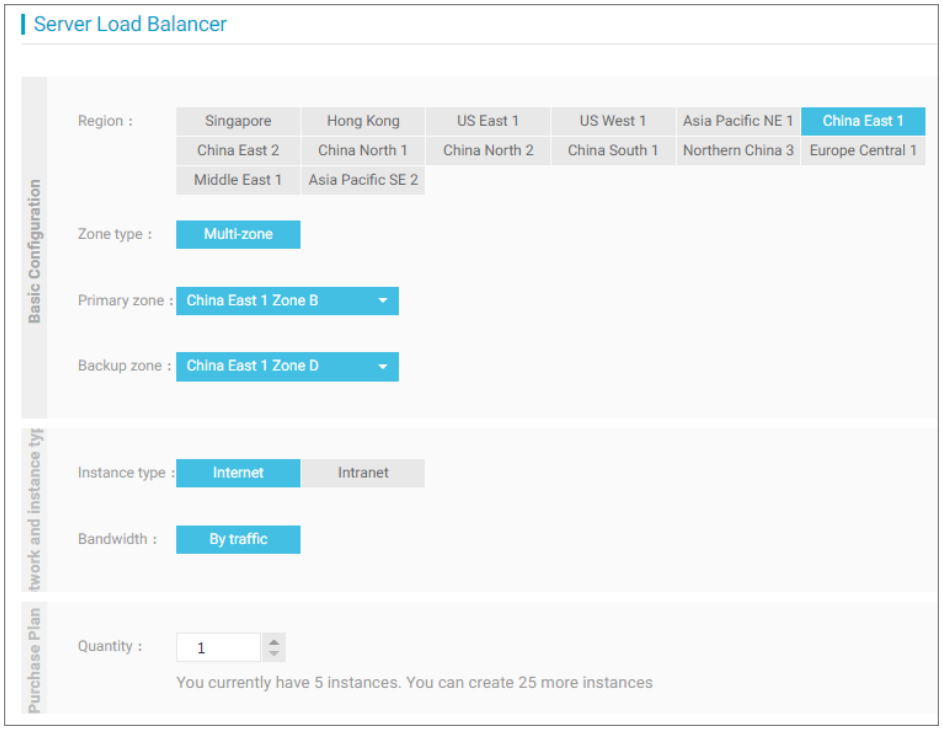
The configurations for the Server Load Balancer instance in this tutorial are as follows. For more details, refer to Server Load Balancer configurations.

Region: Server Load Balancer does not support cross-region deployment. The region must be the same for the Server Load Balancer instance and the ECS instances. In this tutorial, we choose **China East 1**, which is the region of the ECS instance.

Zone type: Server Load Balancer has deployed multiple zones in most regions for better disaster tolerance. If Server Load Balancer service is unavailable in the primary zone, it will switch to a backup zone to restore service (within 30 seconds). Then, it will automatically switch back to the primary zone when the service is restored.

In this tutorial, select **China East 1 Zone B** as the primary zone and **China East 1 Zone D** as the backup zone.

Instance type: Select **Internet**.

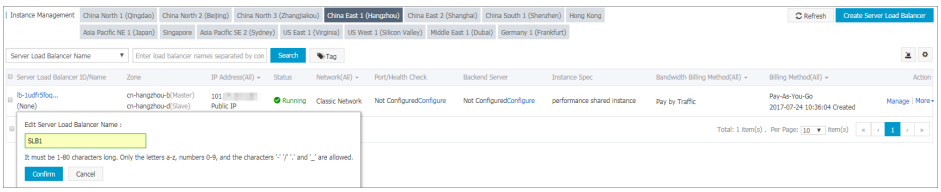


Click **Buy Now**.

Go back to the **Instance Management** page, find the created instance.

Hover the mouse pointer over the instance ID and then click the pencil icon.

Enter the name SLB1 and click **Confirm**.



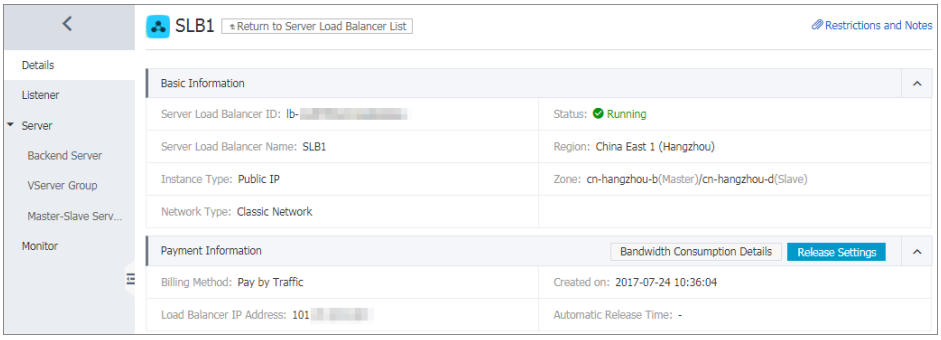
After creating a Server Load Balancer instance, you have to add at least one listener and a group of backend servers to it.

In this tutorial, we will create a TCP listener and add the ECS instances that have deployed web pages as the backend servers.

Procedure

Log on to the Server Load Balancer console.

On the **Instance Management** page, click the ID of the target Server Load Balancer instance.



On the left-side navigation pane, click **Listener** and then click **Add Listener**.

Configure the listener as follows and use the default settings for other options:

Frontend Protocol [Port]: The front-end protocol and port of the Server Load Balancer system that is used to receive and distribute connection requests. The port number cannot be the same in a Server Load Balancer instance.

In this tutorial, select the **TCP** protocol with port number **80**.

Backend Protocol [Port]: The port number that is opened on the ECS instances to receive the distributed requests. The port number can be the same in a Server Load Balancer instance.

In this tutorial, set to **80**.

Peak Bandwidth: You can set a peak bandwidth to limit the service capabilities that the application of the ECS instance can provide.

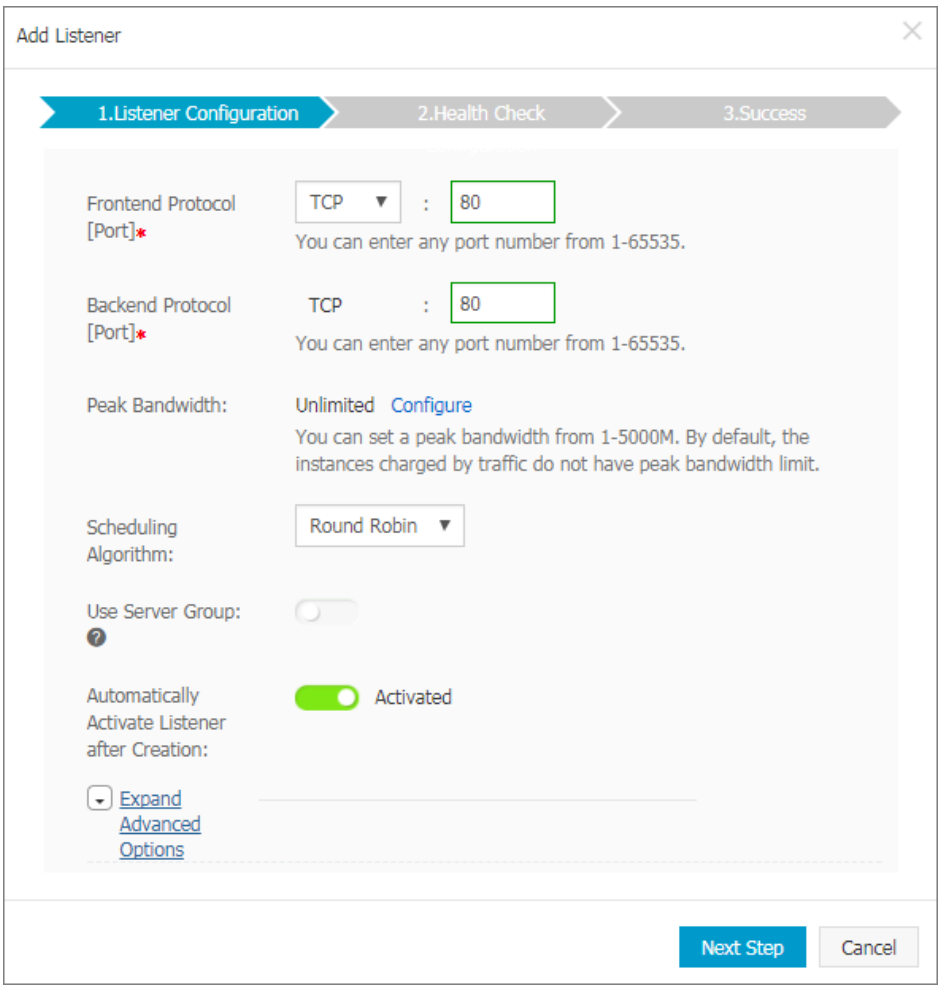
In this tutorial, no need to set the peak bandwidth because the instance is payed by traffic.

Scheduling Algorithm: Server Load Balancer supports the following scheduling algorithms. In this tutorial, the round-robin method is used.

Round robin: Requests are distributed evenly across the group of the backend ECS servers sequentially.

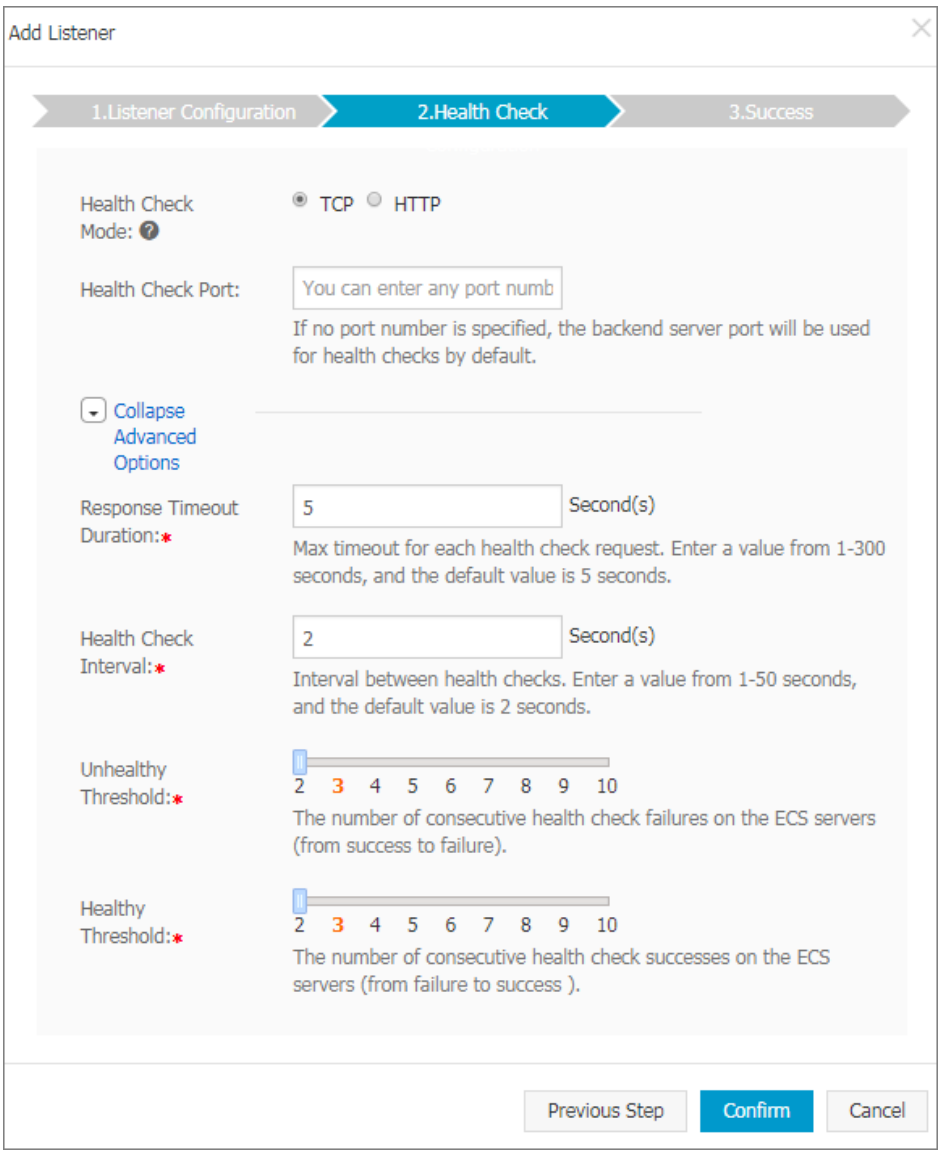
Weighted round robin (WRR): You can set a weight for each backend server. Servers with higher weights receive more requests than those with less weights.

Weighted least connections (WLC): In addition to the weight set to each backend ECS server, the number of connections to the client is also considered. The servers with a higher weight value will receive a larger percentage of live connections at any one time. If the weights are the same, the system directs network connections to the server with the least number of established connections.



Click **Next Step** to configure health check settings. Select the **TCP** mode and keep other settings as default, click **Confirm**.

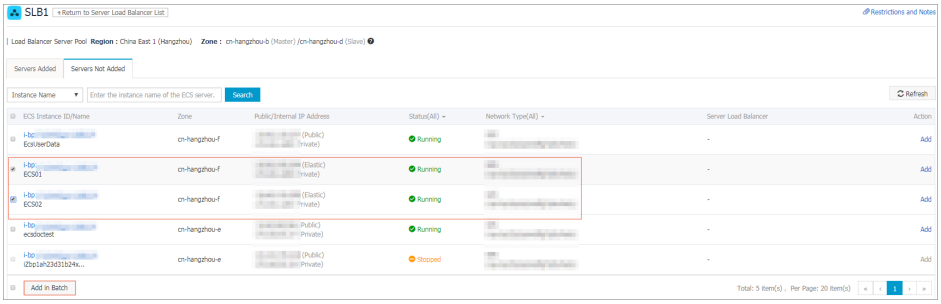
Through health check on backend ECS instances, Server Load Balancer can automatically block abnormal ECS instances and distribute requests to them again when they become normal.



Click **Confirm** to complete the configuration.

On the left-side navigation pane, click **Server** > **Backend Server**.

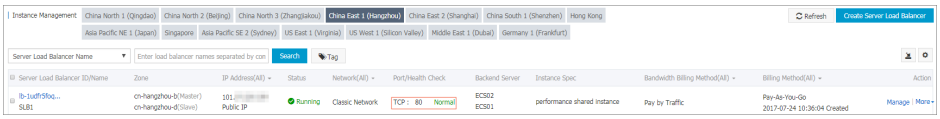
On the **Load Balancer Server Pool** page, click the **Servers Not Added** tab and select the previously created ECS instances, then click **Add in Batch**.



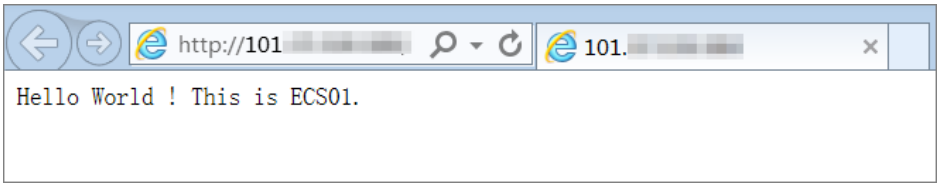
In the **Add a Backend Server** dialog, use the default weight value and click **Confirm**.

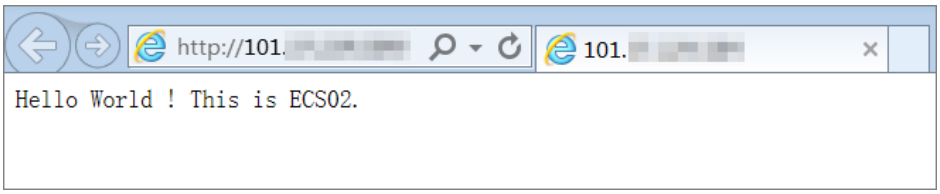
The higher the weight, the more requests are received.

Go back to the **Instance Management** page, click **Refresh**. When the health check is **Normal**, you can send requests to the Server Load Balancer instance.



In the web browser, enter the IP address of the Server Load Balancer instance to test the service.





When you no longer need Server Load Balancer, delete the corresponding instance to avoid additional charges. When deleting the Server Load Balancer instance, the backend ECS will not be deleted or affected.

> **Note**: After the Server Load Balancer instance is released, the backend ECS instances are still running. If you want to release the ECS instances, refer to Release an instance.

Procedure

Log on to the ECS console.

On the **Instance Management** page, select the region where the instance is located.

Select the target instance and click **Release Settings**.

In the **Release Settings** dialog, select **Release Now** or **Timed Release**.

If you select **Timed Release**, select the time to release the instance.

Click **Next Step** and click **Confirm** to finish.