

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 is created to distribute received HTTP requests to 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 more information, 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, multiple zones for Server Load Balancer are deployed in most regions for better disaster tolerance. Additionally, to improve cross-region availability, you can deploy 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.

Note:

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

Server Load Balancer does not support cross-region deployment. 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 as needed. 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.

An Internet Server Load Balancer instance only has a public IP and is accessible from the Internet.

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

Billing by traffic: Suitable for an application with obvious traffic changes.

Billing by bandwidth: Suitable for an application with relatively stable bandwidth.

An intranet Server Load Balancer instance only has a private IP and is accessible only from a classic network or VPC.

Plan the listening protocol

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

A layer-4 listener distributes connection requests directly to backend servers without modifying HTTP headers. After a request arrives at a layer-4 listener, the Server Load

Balancer server uses the backend port configured in the listener to create a TCP connection with backend ECS instances.

A layer-7 listener is an implementation of reverse proxy. After a request arrives at a layer-7 listener, the Server Load Balancer server uses a TCP connection to transmit the data packets to backend ECS instances instead of transmitting the data packets directly.

Prepare the backend servers

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

ECS regions and zones

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

ECS configurations

Additional configuration is not required after applications are deployed on the ECS instances. However, if you 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 deployment

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

Before using Server Load Balancer, you have to create at least two ECS instances and deploy corresponding applications, and add the instances to the Server Load Balancer instance to process 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 **Elastic Compute Services (ECS)** page, configure the ECS instance.

The following are ECS settings used in this tutorial. For more information, see [Create Linux ECS instances](#).

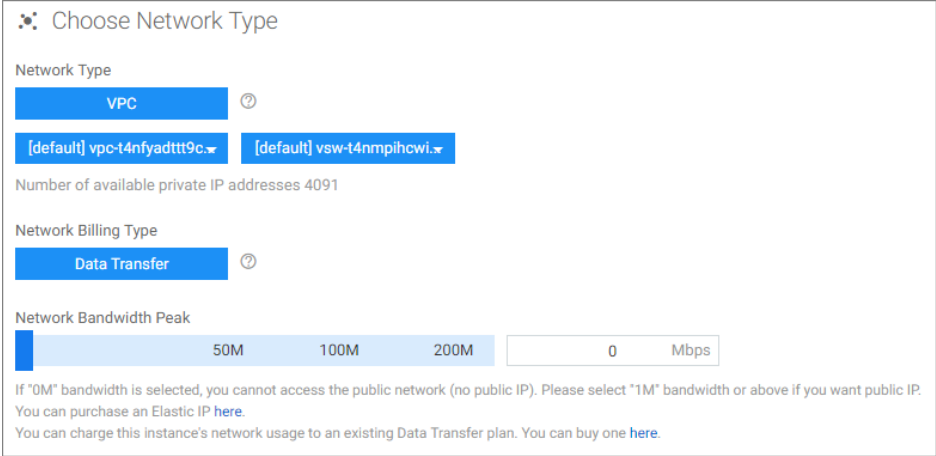
Region: In this tutorial, select **China East 1**.

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 instances.

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 simultaneously creates two ECS instances with identical settings.



The screenshot shows the 'Choose Network Type' configuration window. It includes the following settings:

- Network Type:** VPC (selected). Below it, the default VPC is '[default] vpc-t4nfyadt9c' and the default VSwitch is '[default] vsw-t4nmpihcwi'.
- Number of available private IP addresses:** 4091.
- Network Billing Type:** Data Transfer (selected).
- Network Bandwidth Peak:** A slider bar with options 50M, 100M, 200M, and 0. The '0' option is currently selected.

Below the slider, there is a note: 'If "0M" bandwidth is selected, you cannot access the public network (no public IP). Please select "1M" bandwidth or above if you want public IP. You can purchase an Elastic IP [here](#). You can charge this instance's network usage to an existing Data Transfer plan. You can buy one [here](#).'

Click **Buy Now** and complete the payment.

Go back to the **Instance List** page and click **China East 1**. The two newly created ECS instances are displayed.

Hover the mouse pointer over one instance name and click the displayed pencil icon to change the instance name to ECS01. Then change the other instance name to ECS02.

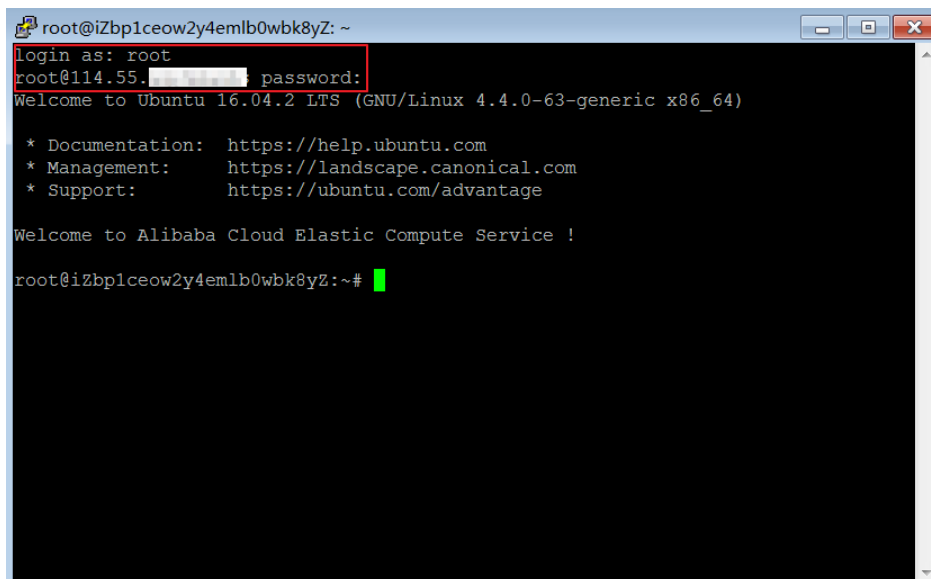
Instance ID/Name	IP Address	Status(All)	Network Type(All)	Billing Method(All)	Action
 i-b011... ECS01	172.17.0.1 (Private IP Address)	Running	VPC	Pay-As-You-Go 17-07-23 17:23 created	Manage Connect More
 i-b011... ECS02	172.17.0.2 (Private IP Address)	Running	VPC	Pay-As-You-Go 17-07-23 17:23 created	Manage Connect More

After you create the ECS instances, you need to deploy applications. In this tutorial, two static web pages are 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 more information, see [Bind an EIP](#).

Procedure

Log on to the ECS instance.



```
root@iZbp1ceow2y4emlb0wbk8yZ: ~  
login as: root  
root@114.55.114.55: password:  
Welcome to Ubuntu 16.04.2 LTS (GNU/Linux 4.4.0-63-generic x86_64)  
  
* Documentation:  https://help.ubuntu.com  
* Management:    https://landscape.canonical.com  
* Support:        https://ubuntu.com/advantage  
  
Welcome to Alibaba Cloud Elastic Compute Service !  
root@iZbp1ceow2y4emlb0wbk8yZ:~#
```

Enter the following command to install Apache.

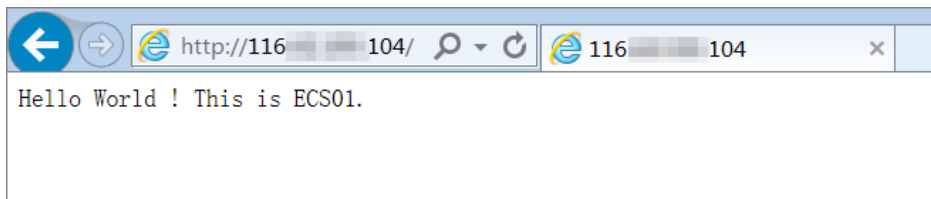
```
sudo apt-get install apache2
```

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

```
cd /var/www/html
```

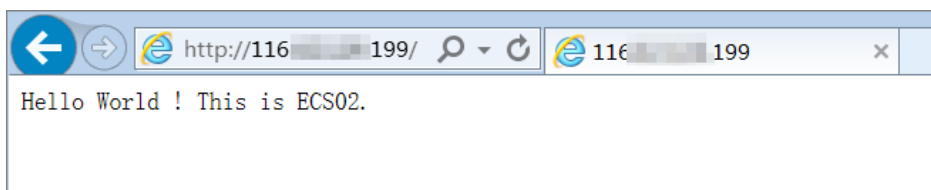
```
echo "Hello World! This is ECS01." > index.html
```

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



Repeat the preceding steps to create a web page on the other 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 the instance is created, a public IP is allocated to it. You can resolve a domain name to this IP.

Procedure

Log on to the Server Load Balancer console.

On the **Instances** 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 information, see [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 ECS instances. In this tutorial, we select **China East 1**, which is the region of the ECS instances.

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

The screenshot shows the 'Server Load Balancer' configuration interface. It is divided into three main sections: 'Basic Configuration', 'Instance type', and 'Purchase Plan'.

- Basic Configuration:**
 - Region:** A grid of region buttons. 'China East 1' is highlighted in blue.
 - Zone type:** A button labeled 'Multi-zone' is selected.
 - Primary zone:** A dropdown menu showing 'China East 1 Zone B'.
 - Backup zone:** A dropdown menu showing 'China East 1 Zone D'.
- Instance type:**
 - Instance type:** Two buttons, 'Internet' (selected) and 'Intranet'.
 - Bandwidth:** A button labeled 'By traffic' is selected.
- Purchase Plan:**
 - Quantity:** A numeric input field set to '1'.
 - A message below the input: 'You currently have 5 instances. You can create 25 more instances'.

Click **Buy Now** and complete the payment.

Go back to the **Instances** 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**.

The screenshot shows the 'Instances' page of the Server Load Balancer console. A table lists instances, with the first instance selected. An 'Edit' modal is open for this instance.

Server Load Balancer ID/Name	Zone	IP Address(A)	Status	Network(A)	Port/Health Check	Backend Server	Instance Spec	Bandwidth Billing Method(A)	Billing Method(A)	Action
slb-1uff0f0e...	cn-hangzhou-b(Native)	101.101.101.101	Running	Classic network	Not Configured/Configure	Not Configured/Configure	performance shared instance	Pay by Traffic	Pay-As-You-Go	2017-07-24 10:36:04 Created

Edit Server Load Balancer Name:

SLB1

It must be 1-60 characters long. Only the letters a-z, numbers 0-9, and the characters '-' and '_' are allowed.

Confirm **Cancel**

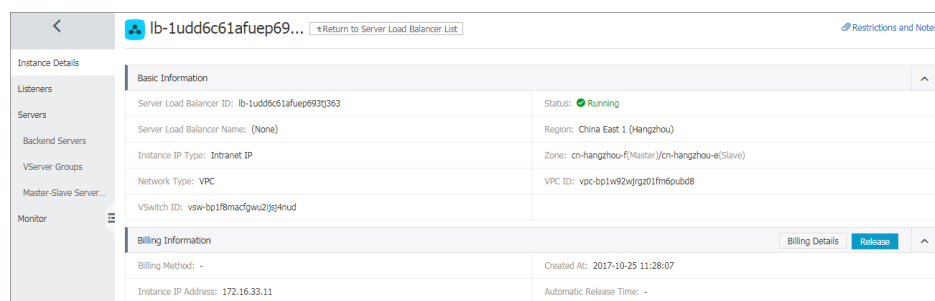
After creating a Server Load Balancer instance, you have to add at least one listener and a group of backend servers to it.

In this tutorial, we will add one TCP listener and add two ECS instances deployed with web pages as the backend servers.

Procedure

Log on to the Server Load Balancer console.

On the **Instances** page, click the ID of the target Server Load Balancer instance and enter the **Details** page.



In the left-side navigation pane, click **Listeners** 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. Two port numbers 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 that is opened on ECS instances to receive distributed requests. It can be the same in a Server Load Balancer instance.

In this tutorial, set the backend port number to **80**.

Peak Bandwidth: You can set a peak bandwidth to limit the service capabilities that applications on the ECS instances can provide.

In this tutorial, you do not need to set the peak bandwidth because the instance is paid 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 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 lower 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. A server 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.

Add Listener

1.Listener Configuration

2.Health Check

3.Success

Frontend Protocol
[Port]*

TCP : 80

You can enter any port number from 1-65535.

Backend Protocol
[Port]*

TCP : 80

You can enter any port number from 1-65535.

Peak Bandwidth:

Unlimited [Configure](#)

You can set a peak bandwidth from 1-5000M. By default, the instances charged by traffic do not have peak bandwidth limit.

Scheduling Algorithm:

Round Robin

Use Server Group:

?

Automatically
Activate Listener
after Creation:

Activated

Expand
Advanced
Options

Next Step

Cancel

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

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

The screenshot shows the 'Add Listener' dialog box with the '2. Health Check' step selected. The 'Health Check Mode' is set to 'TCP'. The 'Health Check Port' field is empty, with a hint that the backend server port will be used by default. The 'Response Timeout Duration' is set to 5 seconds. The 'Health Check Interval' is set to 2 seconds. The 'Unhealthy Threshold' is set to 3, and the 'Healthy Threshold' is set to 2. The 'Previous Step', 'Confirm', and 'Cancel' buttons are at the bottom.

Add Listener

1.Listener Configuration **2.Health Check** 3.Success

Health Check Mode: ☒ TCP ☐ HTTP

Health Check Port:
If no port number is specified, the backend server port will be used for health checks by default.

☐ Collapse Advanced Options

Response Timeout Duration: Second(s)
Max timeout for each health check request. Enter a value from 1-300 seconds, and the default value is 5 seconds.

Health Check Interval: Second(s)
Interval between health checks. Enter a value from 1-50 seconds, and the default value is 2 seconds.

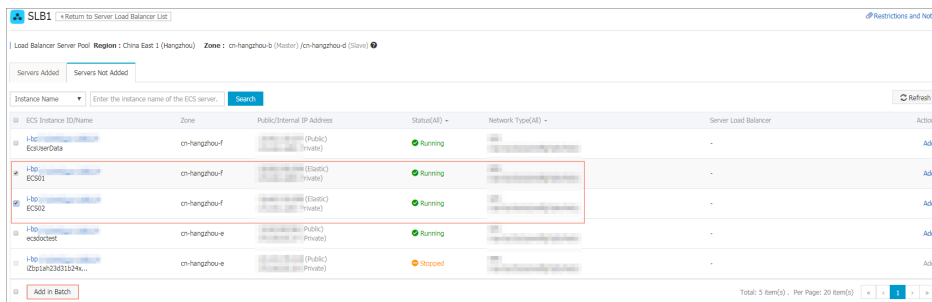
Unhealthy Threshold: 2 3 4 5 6 7 8 9 10
The number of consecutive health check failures on the ECS servers (from success to failure).

Healthy Threshold: 2 3 4 5 6 7 8 9 10
The number of consecutive health check successes on the ECS servers (from failure to success).

Click **Confirm** to complete the configuration.

In the left-side navigation pane, click **Servers** > **Backend Servers**.

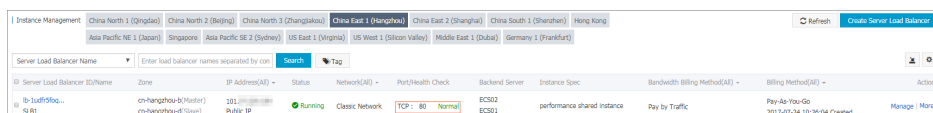
On the **Load Balancer Server Pool** page, click the **Servers Not Added** tab, select the previously created ECS instances, and then click **Add in Batch**.



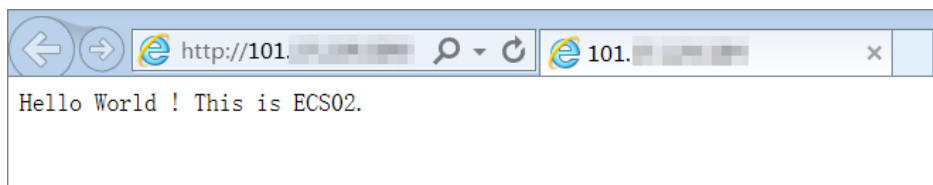
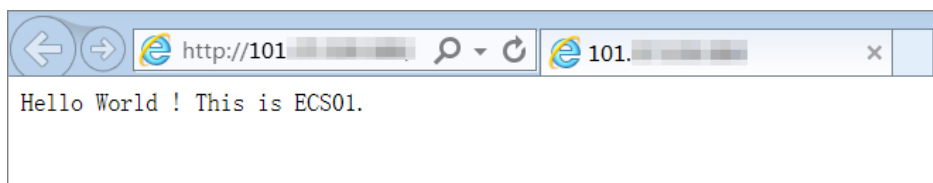
In the **Add a Backend Server** dialog box, use the default weight value and click **Confirm**.

A server with a higher weight value receives more requests. The default weight value is 100.

Go back to the **Instances** page and click **Refresh**. When health check is **Normal**, the corresponding ECS instance can process requests forwarded by the Server Load Balancer instance normally.



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



You can resolve a domain name to the public address of the Server Load Balancer instance. For example, the domain name of your website is *www.abc.com* and the website is running on an ECS instance with the public IP 1.1.1.1. After creating a Server Load Balancer instance, a public IP 2.2.2.2 is allocated to the instance. You have to add the ECS instance hosting the website to the backend server pool and resolve the domain name *www.abc.com* to 2.2.2.2. We recommend that you add an A record resolution (resolve a domain to an IP address).

For more information, see [Add and manage records](#).

When you no longer need Server Load Balancer, delete the corresponding instance to avoid additional charges. Deleting the Server Load Balancer instance does not delete or affect backend ECS instances.

Note: After the Server Load Balancer instance is released, the backend ECS instances are still running. If you want to release the ECS instances, see [Release an instance](#).

Procedure

Log on to the ECS console.

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

Select the target instance and click **More > Release**.

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

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

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