ALIBABA CLOUD

Alibaba Cloud

Server Load Balancer
Quick Start (New Console)

Document Version: 20200901

Alibaba Cloud

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.

2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company or individual in any form or by any means without the prior written consent of Alibaba Cloud.

3. The content of this document may be changed because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.

4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides this document based on the "status quo", "being defective", and "existing functions" of its products and services. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, special, or punitive damages, including lost profits arising from the use or trust in this document (even if Alibaba Cloud has been notified of the possibility of such a loss).

5. By law, all the contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of this document shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates.

6. Please directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
Danger	A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	Danger: Resetting will result in the loss of user configuration data.
Warning	A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	Warning: Restarting will cause business interruption. About 10 minutes are required to restart an instance.
Notice	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	Notice: If the weight is set to 0, the server no longer receives new requests.
Note	A note indicates supplemental instructions, best practices, tips, and other content.	Note: You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Table of Contents

1.Overview

This topic describes the process of creating a Server Load Balancer (SLB) instance to forward requests to backend servers.

> **Note** Before creating an SLB instance, you must plan your SLB service, such as the instance type, region, and more. For more information, see Before you begin.

To create an SLB instance and forward requests to backend servers, follow these steps:

1. Create an SLB instance.

   Create an SLB instance. An SLB instance is a running entity of the SLB service.

2. Configure listeners and add backend servers.

   After creating an SLB instance, you must add at least one listener, and add ECS instances as backend servers.

3. Resolve a domain name (optional).

   Use Alibaba Cloud DNS to resolve a domain name to the IP address of the SLB instance to provide external services.

4. Release an SLB instance.

   If you no longer need the SLB instance, delete it to avoid extra fees.

2.Before you begin

This article presents the essential considerations for configuring an SLB instance. Before you create an SLB instance, you must determine the types of listeners and the network traffic you want to balance.

Instance region

When you select a region, note the following points:

- To reduce latency and increase the download speed, we recommend that you select a region closest to your end-users.
- SLB offers stable and reliable load balancing services by providing support for primary/secondary failovers in most regions. This implements disaster recovery across different zones within the same region. We recommend that you select a region that supports the primary/secondary SLB deployment.
- SLB instances cannot span across regions. Therefore, you must make sure that the SLB instance and its backend Elastic Compute Service (ECS) instances are located in the same region.

Network traffic

SLB provides load balancing services for both Internet and internal network traffic:

- If you need to use SLB to distribute requests from the Internet, you can create an Internet-facing SLB instance.

  An Internet-facing SLB instance comes with a public IP address to receive requests from the Internet.

- If you need to use SLB to distribute requests from the internal network, you can create an internal SLB instance.

  Internal SLB instances only have private IP addresses and are only accessible from the internal network and not from the Internet.

Instance type

When you create an SLB instance, you can choose a guaranteed-performance instance or a shared-performance instance. The guaranteed-performance SLB instance provides greater flexibility in resource utilization to guarantee service availability. SLB provides six types of guaranteed-performance instances.

- For a pay-as-you-go SLB instance, we recommend that you select the instance type that provides the highest level of performance. This guarantees a flexible load balancing service without incurring any additional costs. However, if the capacity of Super I (slb.s3.large), the highest-performance instance type, far exceeds the demand of your business, you can select a more appropriate type based on the workload of your business, for example, Higher II (slb.s3.medium).

Listener protocol

SLB supports Layer-4 load balancing of Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) traffic, and Layer-7 load balancing of HTTP and HTTPS traffic.

- A Layer-4 listener directly distributes requests to backend servers without modifying packet headers. After a client request reaches a Layer-4 listener, SLB uses the backend port configured for the listener to establish a TCP connection with an Elastic Compute Service (ECS) instance (backend server).
- A Layer-7 listener is implemented as a reverse proxy. After a client request reaches a Layer-7 listener, SLB establishes a new TCP connection over HTTP to a backend server, instead of directly forwarding the request to the backend server (ECS instance).

  Compared with Layer-4 listeners, Layer-7 listeners require an additional step of Tengine processing. Therefore, Layer-4 listeners provide better performance than Layer-7 listeners. In addition, the performance of Layer-7 listeners can also be affected by factors such as insufficient client ports or excessive backend server connections. Therefore, we recommend that you use Layer-4 listeners for high-performance load-balancing services.

  For more information, see Protocols.

Backend servers

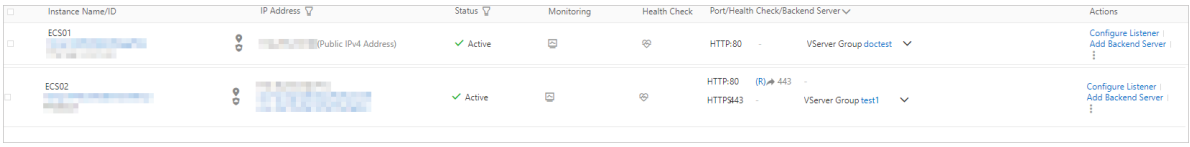
Before you use the SLB service, you must create ECS instances, deploy applications on them, and add the ECS instances to your SLB instance to process client requests.

When you create and configure an ECS instance, note the following points:

- Select a region and zone for the ECS instance

  Make sure that the ECS instance resides in the same region and Virtual Private Cloud (VPC) as the SLB instance. We recommend that you deploy ECS instances in different zones to improve availability. For more information about how to create an ECS instance, see Create an instance by using the provided wizard.

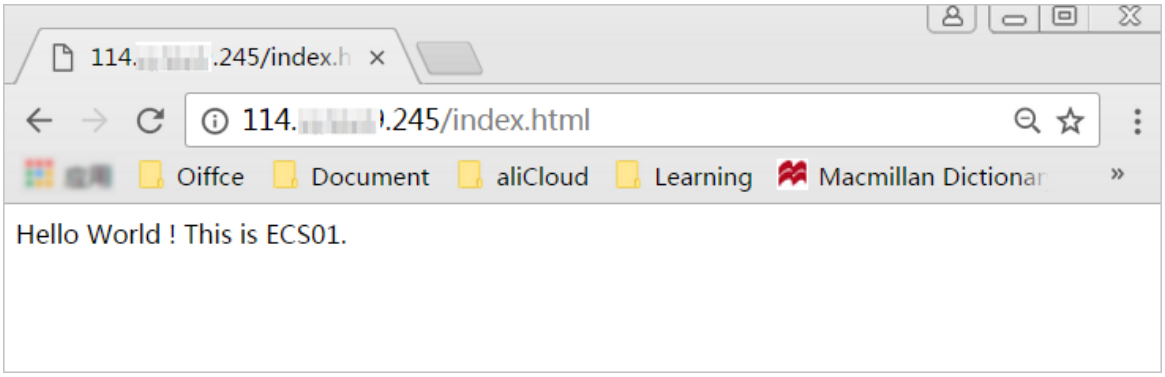
  In this example, two ECS instances named ECS01 and ECS02 are created in the **China (Hangzhou)** region. The following figure shows their basic configurations.



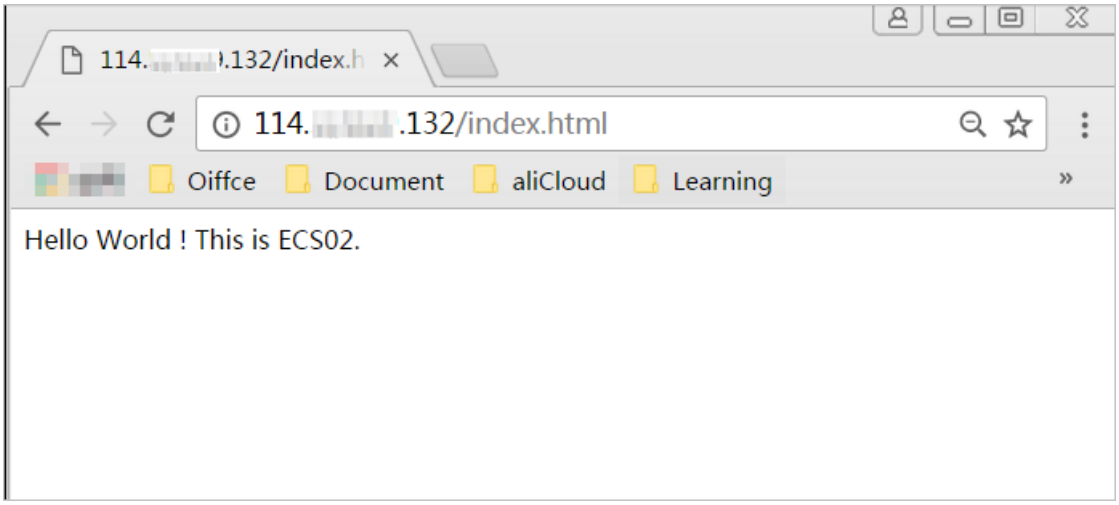
- Configure applications

  In this example, two static web pages are built on ECS01 and ECS02 by using Apache.

- Enter the Elastic IP address (EIP) associated with ECS01 in the address box of your browser.



- Enter the EIP associated with ECS02 in the address box.



No additional configuration is required after you deploy applications on the ECS instances. However, if you need to use a Layer-4 (TCP or UDP) listener and the ECS instances run on Linux, make sure that the following parameters in the *net.ipv4.conf* file under */etc/sysctl.conf* are set to 0:

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

3.Create an SLB instance

This topic describes how to create a public-facing Server Load Balancer (SLB) instance. After a public-facing SLB instance is created, a public IP address is allocated to the instance and you can resolve a domain name to this address.

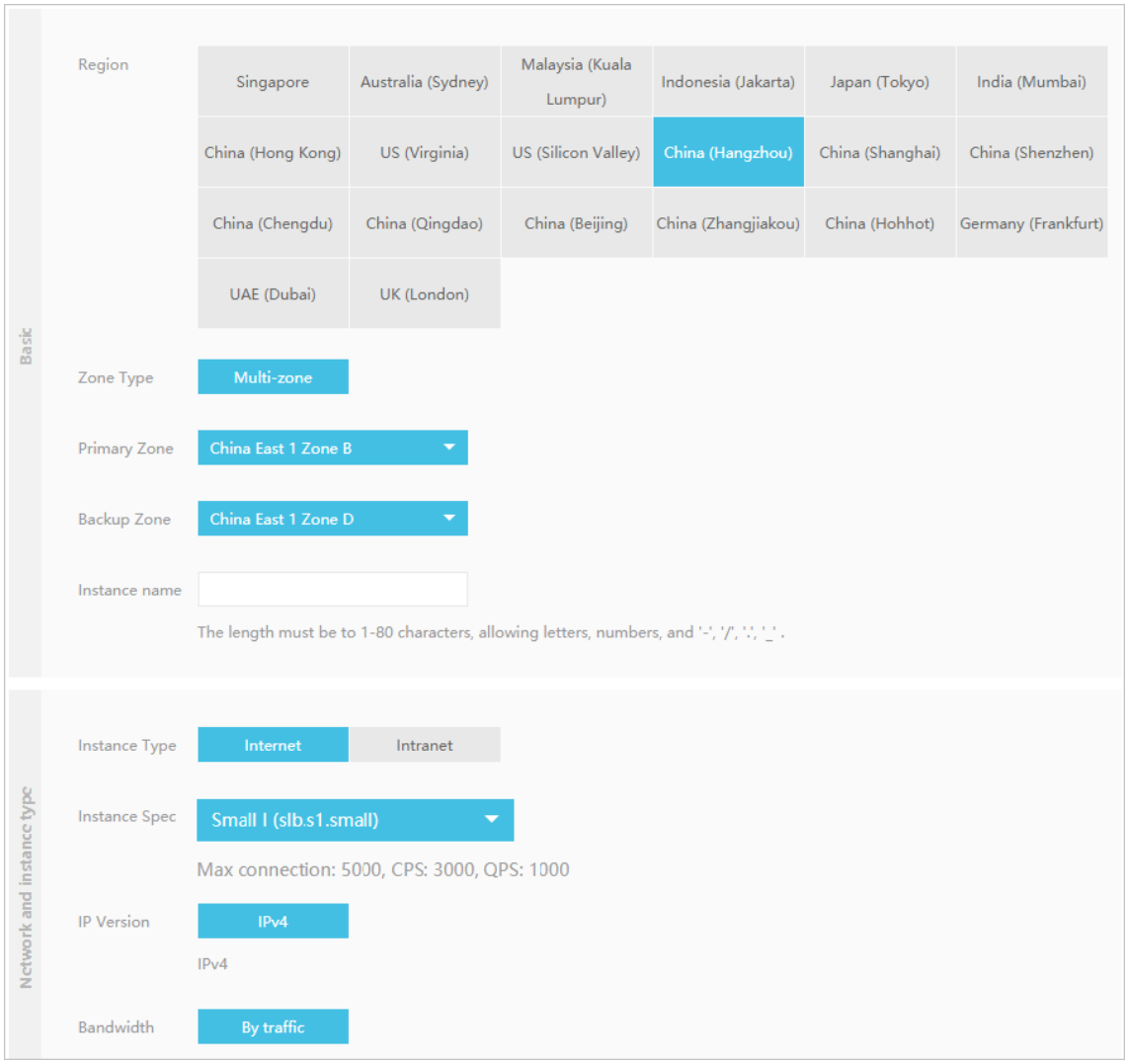
Context

You can add multiple listeners and backend servers to an SLB instance.

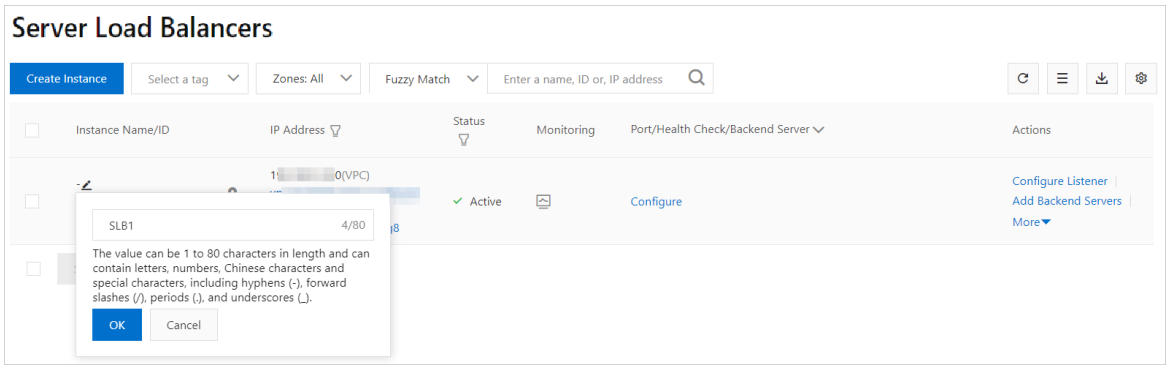
Procedure

1. Log on to the SLB console.
2. On the **Server Load Balancer** page, click **Create Instance**.
3. Configure the SLB instance. For more information, see Create an SLB instance. In this example, configure the SLB instance as follows:
   - **Region**: SLB does not support cross-region deployment. The region of the SLB instance must be the same as that of ECS instances. In this example, select **China (Hangzhou)**.
   - **Zone Type**: Multiple zones have been deployed in most regions for better disaster tolerance. SLB can switch to the secondary zone to provide the load balancing service when the primary zone is unavailable. When the primary zone is recovered, SLB automatically switches back to the primary zone.

     In this example, select **China East 1 Zone B** as the primary zone, and **China East 1 Zone D** as the secondary zone.
   - **Instance Type**: Select **Internet**.



4. Click **Buy Now** and complete the payment.
5. Go back to the SLB console.
6. To edit the instance name, select the **China (Hangzhou)** region on the **Server Load Balancer** page. Move the pointer near the name of the created instance and then click the pencil icon. Enter a name, for example, *SLB1* and click **OK**.



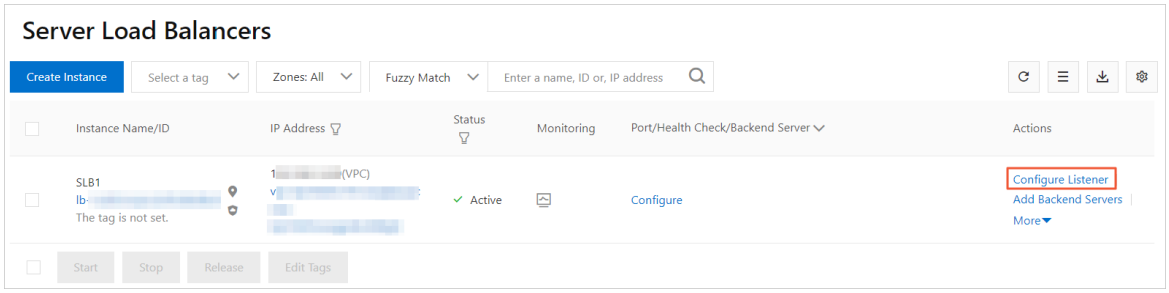
7. (Optional)You can resolve a domain name to the public IP address of the SLB instance. For more information, see Resolve a domain name.

4.Configure an SLB instance

This topic describes how to configure an SLB instance. After you create an SLB instance, you must add at least one listener and a group of backend servers to this SLB instance so that it can forward traffic. The following example sets up a TCP listener and adds two ECS instances (ECS01 and ECS02) that host static web pages as backend servers to an SLB instance.

Procedure

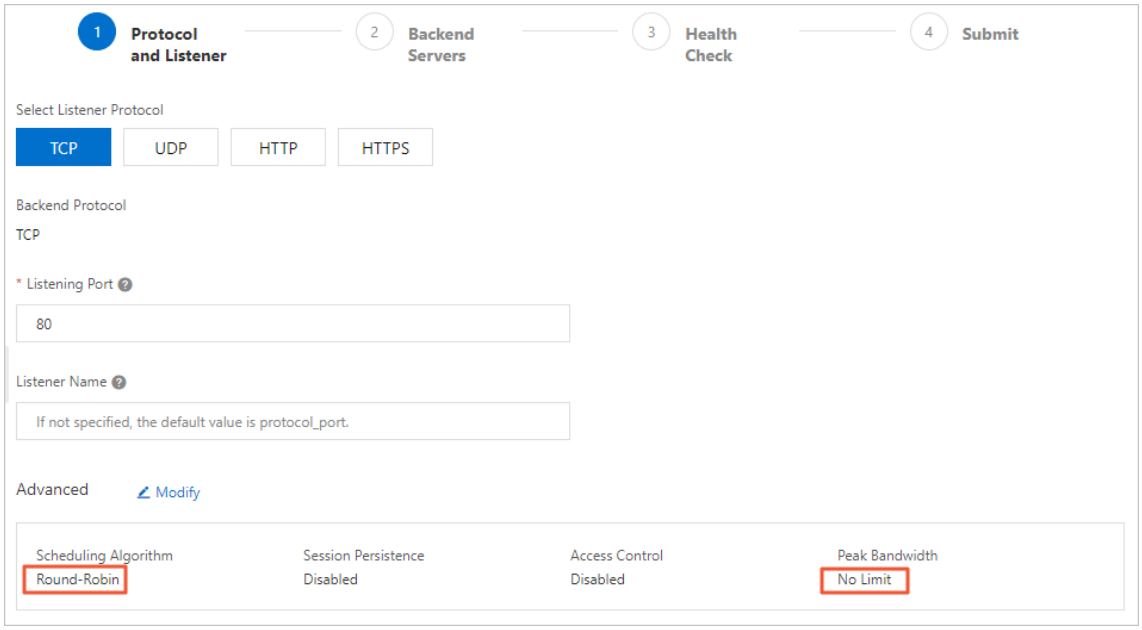
1. Log on to the SLB console.
2. On the **Server Load Balancers** page, find the target SLB instance and click **Configure Listener** in the Actions column.



3. In the **Protocol and Listener** step, complete the following information and use the default values for other fields to configure the listening rule.
   - **Select Listener Protocol**: Select a listener protocol. In this example, select **TCP**.
   - **Listening Port**: Set a frontend port to receive and forward requests to backend servers.

     In this example, set the port number to **80**.
   - **Enable Peak Bandwidth Limit**: You can switch on this option and then set a bandwidth limit to control the bandwidth that is used by the applications running on backend servers to provide external services.

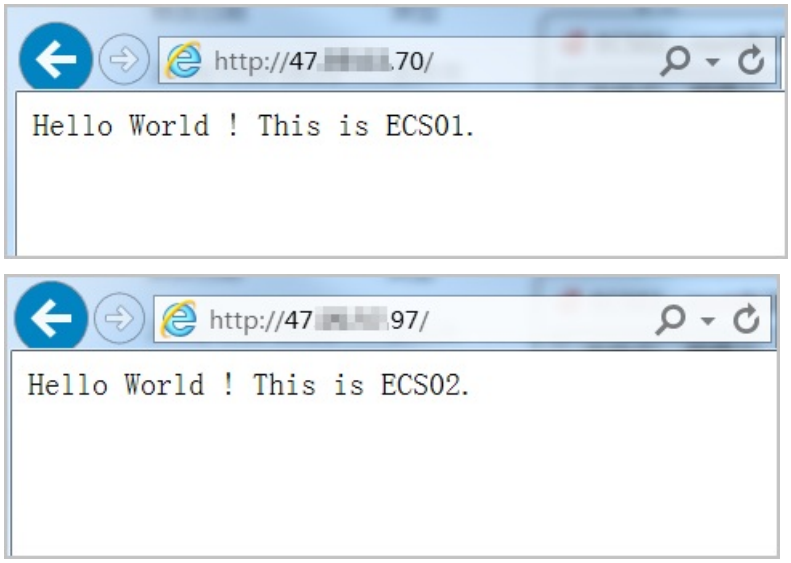
     In this example, the SLB instance incurs fees based on the amount of transmitted data, not on the bandwidth. Therefore, the bandwidth limit is not set.
   - **Scheduling Algorithm**: Select a scheduling algorithm for distributing requests to backend servers. SLB supports the following scheduling algorithms. In this example, select **Round-Robin (RR)**.
     - **Weighted Round-Robin (WRR)**: Requests are distributed proportionally based on the assigned weights of backend servers. Backend servers with higher weights receive more requests.
     - **Weighted Least Connections (WLC)**: Requests are distributed based on the combination of the weights and active connections of backend servers. If multiple backend servers have the same weight, requests are routed to the backend server with the lowest number of active connections.
     - **Round-Robin (RR)**: Requests are evenly and sequentially distributed to backend servers.



4. Click **Next**. In the **Backend Servers** step, select **Default Server Group** and click **Add More** to add backend servers.
   i. In the **Select Servers** dialog box, select the previously created ECS01 and ECS02 instances and click **Next**.
   ii. Configure the weights of the ECS instances.The higher the weight, the more requests a backend server receives. The default weight is 100. We recommend that you use the default value.
   iii. Click **Add**.
   iv. In the **Default Server Group** section, specify backend server ports.A backend server uses this port to receive requests. You can specify the same port for multiple backend servers of an SLB instance. In this example, set the port number to 80.
5. Click **Next** to configure the health check. The default health check settings are used in this example.

   After you enable the health check feature, when a backend server is detected unhealthy, SLB bypasses requests from this backend server to other healthy backend servers. SLB will only send requests to this backend server when it has been restored and is considered healthy.
6. Click **Next**. In the **Submit** step, check the configuration and click **Submit**.
7. Click **OK** to go back to the **Server Load Balancers** page, and click  to refresh the page.

   If the health check state of a backend ECS instance is **Active**, the backend server is working properly and is able to process requests.
8. In the web browser, enter the service IP address of the SLB instance in the address box to test network load balancing.


http://47 .70/
Hello World ! This is ECS01.
http://47 97/
Hello World ! This is ECS02.

5.Resolve a domain name

This topic describes Alibaba Cloud DNS. Alibaba Cloud DNS is a distributed database that maps domain names to IP addresses. You can use the DNS service to resolve a domain name to the public IP address of a Server Load Balancer (SLB) instance.

Context

In this example, the domain name of your website is www.abc.com and the website runs on an ECS instance of which the public IP address is 1.1.1.1. An SLB instance is created and a public IP address 2.2.2.2 is allocated to the instance. You want 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 name to an IP address).

Procedure

1. Log on to the Alibaba Cloud DNS console.
2. Click **Add Domain Name** to add a domain name.
3. On the **Manage DNS** page, find the added domain name, click **Configure** in the **Actions** column, and complete the DNS configuration.

6.Release an SLB instance

This topic describes how to release an SLB instance. You can release an SLB instance when you no longer need it to avoid accumulating unnecessary charges. Releasing an SLB instance will not affect the operation of the backend ECS instances.

Context

Note

- If you have resolved a domain name to the IP address of the SLB instance that you want to delete, resolve the domain name to the IP address of another SLB instance in advance to avoid service interruption.
- Only pay-as-you-go SLB instances can be released. Subscription SLB instances are automatically released if their subscriptions are not renewed upon expiration.
- After an SLB instance is released, the backend ECS instances that are added to the SLB instance continue running. You can release backend ECS instances that are no longer needed.

Procedure

1. Log on to the SLB console.
2. On the **Server Load Balancers** page, select the region of the target SLB instance.
3. Find and select the target SLB instance and click **Release** at the bottom of the list. You can also choose  > **Release** in the Actions column.
4. In the **Release** dialog box, select **Release Now** or **Release on Schedule**.

   If you select **Release on Schedule**, you must set a release time.

   Note The system performs release operations at 30-minute and hour marks. However, billing for the SLB instance is stopped at the specified release time.
5. Click **Next**.
6. Click **OK**.

   Note Pay-as-you-go SLB instances cannot be restored once deleted. We recommend that you exercise caution when you release SLB instances.