Alibaba Cloud Server Load Balancer

Quick Start (New Console)

Issue: 20190806

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 due to product version upgrades , adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall 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 the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults " and "as available" basis. 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 bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, 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 content of the Alibaba Cloud website, including but not limited to works, products, images, archives, information, materials, website architecture, website graphic layout, and webpage design, 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 the Alibaba Cloud website, product programs, or content 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 contact Alibaba Cloud directly if you discover any errors in this document.

Generic conventions

Table -1: Style conventions

Style	Description	Example
	This warning information 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.
	This warning information 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 restore business.
	This indicates warning information, supplementary instructions, and other content that the user must understand.	 Notice: Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructions, best practices, tips, and other content that is good to know for the user.	 Note: You can use Ctrl + A to select all files.
>	Multi-level menu cascade.	Settings > Network > Set network type
Bold	It is used for buttons, menus , page names, and other UI elements.	Click OK.
`Courier font`	It is used for commands.	Run the `cd / d C :/ windows` command to enter the Windows system folder.
Italics	It is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	It indicates that it is a optional value, and only one item can be selected.	`ipconfig` *[-all|-t]*

Style	Description	Example
{} or {a|b}	It indicates that it is a required value, and only one item can be selected.	`swich {stand | slave}`

Contents

1 Tutorial overview

This tutorial guides you to create an Internet SLB instance to forward requests to backend ECS instances.

> 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.

The tutorial includes the following tasks:

1. Create an SLB instance

   Creates an SLB instance. An SLB instance is a running entity of Server Load Balancer.
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. Delete an SLB instance

   If you no longer need the SLB instance, release it to avoid additional charges.

2 Before you begin

Before you use the Server Load Balancer (SLB) service, you need to determine the instance type and region, network type, listener protocol, and backend servers according to your business needs.

Instance region

Consider the following scenarios when you select the region to which the SLB instance belongs:

- To reduce latency and increase the download speed, select a region closest to your customers.
- To provide more stable and reliable load balancing services, deploy primary and secondary zones for zone-level disaster tolerance. To do this, make sure that you select a region in which primary and backup zones are available.
- SLB does not support cross-region deployment. Therefore, make sure that the region selected for the SLB instance is the same as the region for your backend ECS instances.

Network type

SLB provides Internet and intranet load balancing services. Consider the following scenarios when you select the network type of an SLB instance:

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

  An Internet SLB instance is provided with a public IP address to receive requests from the Internet.
- If you want to use SLB to distribute requests from the intranet, create an intranet SLB instance.

  An intranet SLB instance only has a private IP address and is accessible only from a classic network or VPC.

Instance type

When you create an SLB instance, you need to choose either a guaranteed-performance or a shared-performance instance type. The guaranteed-performance instance type provides greater flexibility in resource utilization to guarantee service

availability. For guaranteed-performance instances, Alibaba Cloud SLB provides six specifications for these instances to better meet your specific requirements.

- We recommend that you select the highest specification, Super I (slb.s3.large). This guarantees the running of your services and will not incur extra costs. However, if you do not require the highest specification available, you can choose a lower specification instance such as Higher II (slb.s3.medium).

Listener protocol

SLB supports Layer-4 (TCP and UDP) and Layer-7 (HTTP and HTTPS) load balancing.

- A Layer-4 listener distributes requests directly to backend servers without modifying HTTP headers. After a request arrives at a Layer-4 listener, SLB uses the backend port configured in the listener to establish 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, SLB uses a TCP connection to transmit the data packets to backend ECS instances instead of transmitting the data packets directly.

  The Layer-7 listener has one more procedure than the Layer-4 listener when forwarding incoming requests. Due to this additional procedure, the performance of the Layer-7 listener is inferior to that of the Layer-4 listener. Moreover, scenarios involving insufficient client ports or excessive connections to the backend servers also affect the performance of Layer-7 listeners. Therefore, if you require high performance, we recommend that you use Layer-4 listeners.

  For more information, see Protocols.

Backend servers

Before you use the SLB service, you must create an ECS instance, deploy applications on it, and add it to an SLB instance.

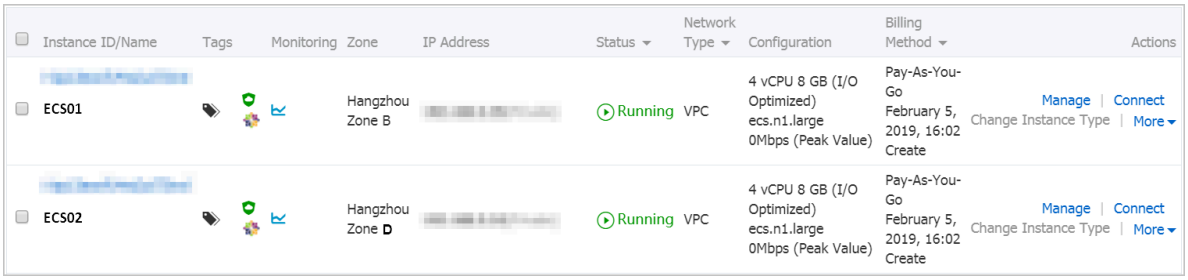
When you create and configure an ECS instance, note the following:

- The region and zone of the ECS instance

  Make sure that the region of the ECS instance is the same as that of the SLB instance. Additionally, we recommend that you deploy each ECS instance in

different zones to improve availability. For more information, see Create an instance by using the wizard.

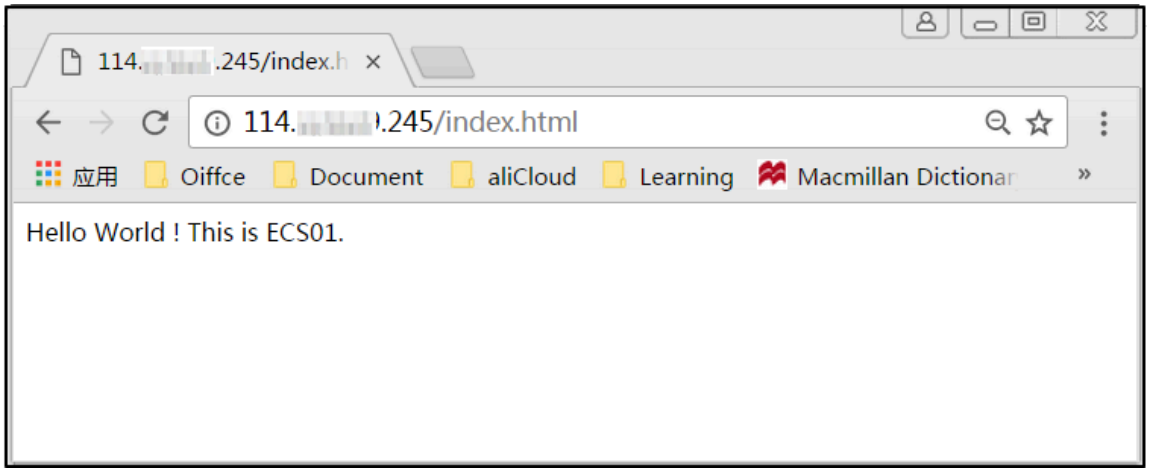
In this example, two ECS instances are created in the China (Hangzhou) region. They are named as ECS01 and ECS02 as shown in the following figure.



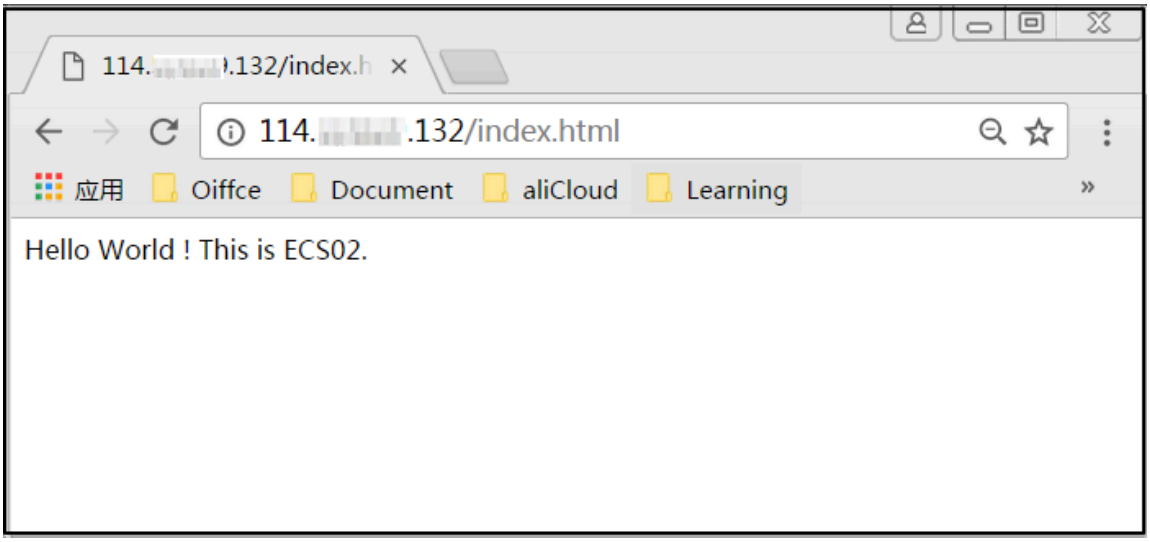
- Application configuration

  In this example, a static website is deployed on ECS01 and another on ECS 02 by using Apache, as shown in the following figure.

  - Enter the EIP address attached to ECS01 in the browser:



  - Enter the EIP address attached to ECS02 in the browser:



  No additional configuration is required after you deploy applications on the ECS instances. However, if you want to use Layer-4 listeners, and the ECS instances use a Linux operating system, make sure that the values of the following parameters in the `net.ipv4.conf` file in `/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 an Internet Server Load Balancer (SLB) instance. After an Internet 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 SLB instance.
3. Configure the SLB instance. For more information, see Create an SLB instance.

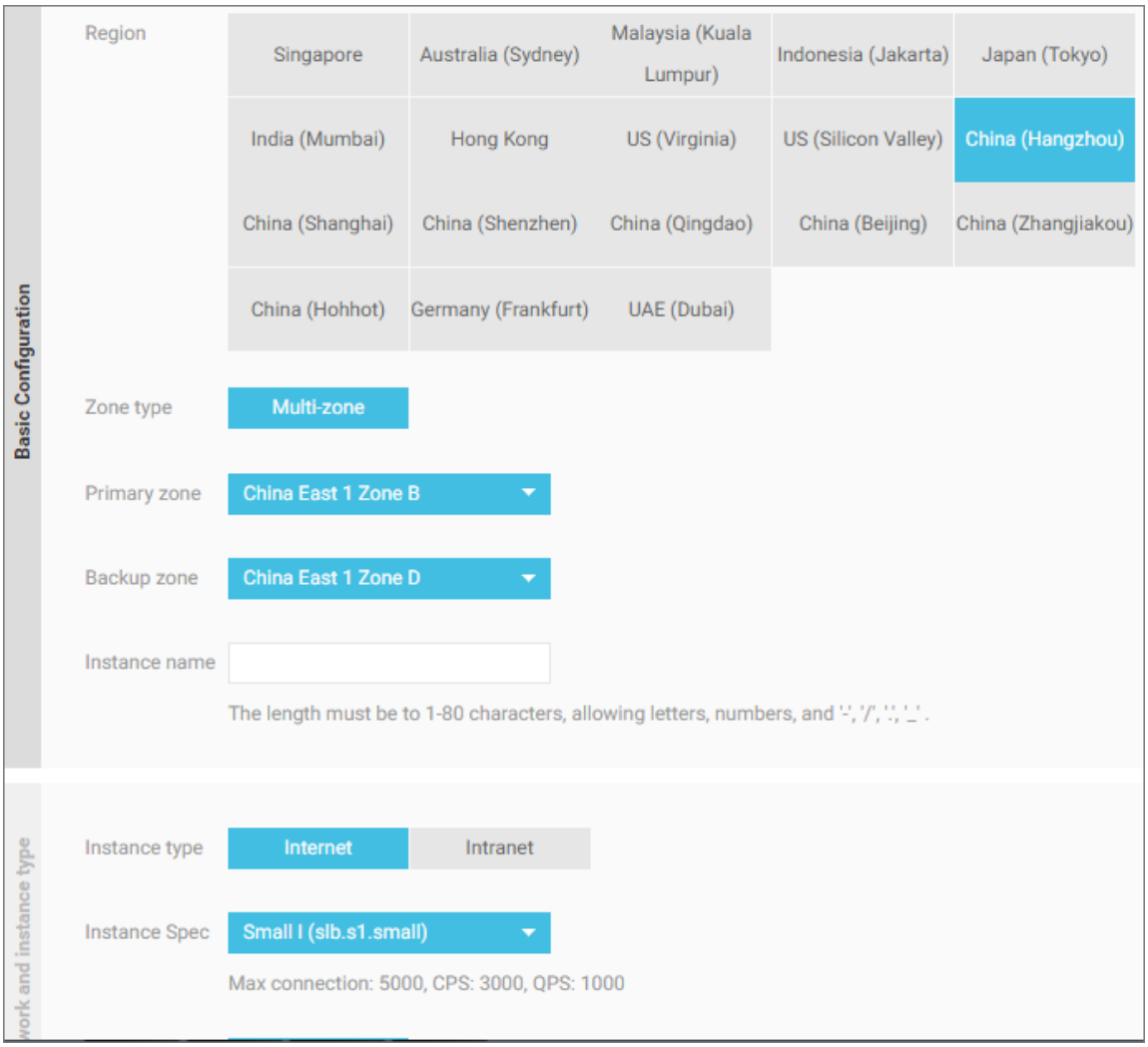
   In this example, configure the SLB instance as follows:

   - 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, and will automatically switch back to the primary zone when the primary zone is recovered.

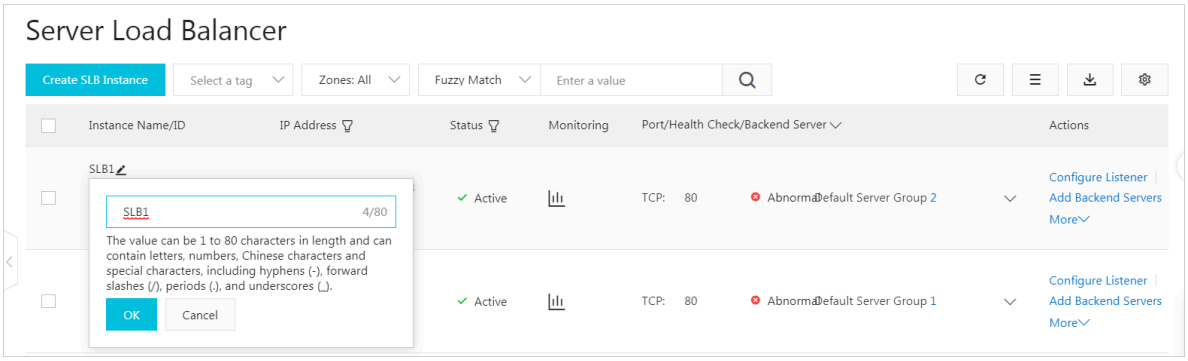
In this example, select China East 1 Zone B as the primary zone, and China East 1 Zone D as the secondary zone.

- 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. Rest the pointer near the name of the created instance and then click the pencil icon. Enter a name, for example, `SLB1` and click OK.



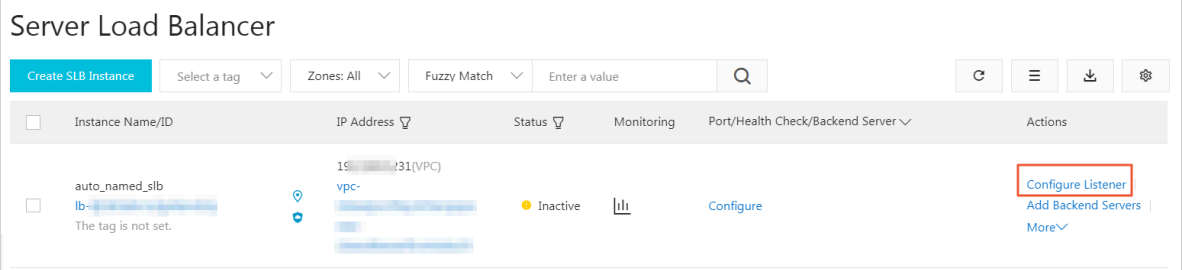
7. Optional: You can resolve a domain name to the public IP address of the SLB instance. For more information, see Resolve a domain name.

4 Configure an SLB instance

This topic describes how to add listeners and backend servers to a Server Load Balancer (SLB) instance. After you create an SLB instance, you must configure it before it can distribute traffic. Specifically, you need to add at least one listener and one group of backend servers to the instance.

Procedure

1. Log on to the SLB console.
2. On the Server Load Balancer page, find the target SLB instance and click Configure Listener.



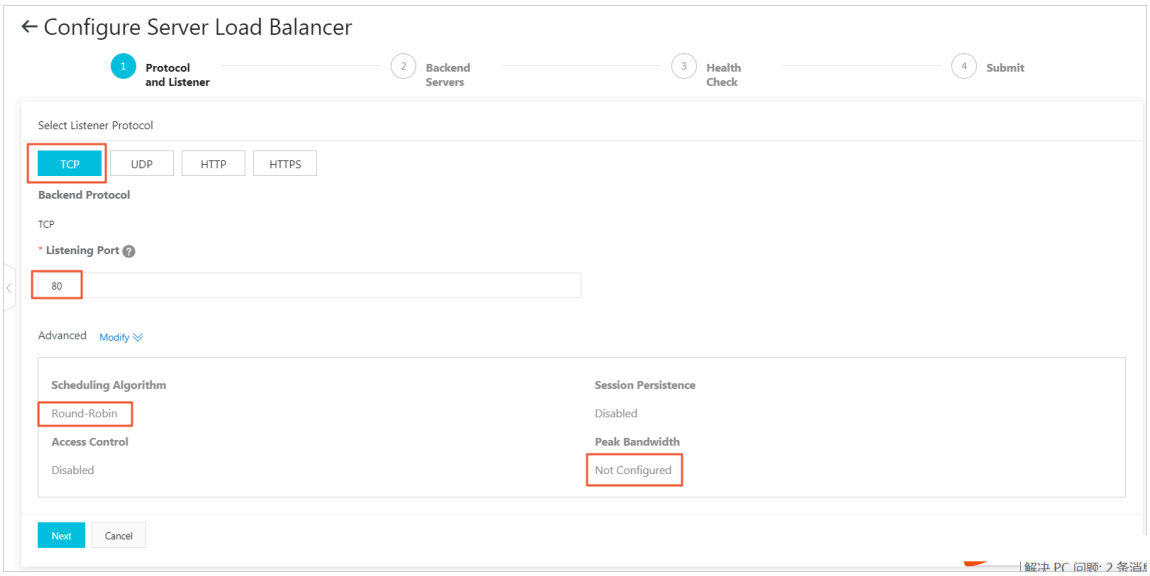
3. On the Protocol and Listener tab, configure the listening rule according to the following information and use the default values for the remaining configurations.
   - Select Listener Protocol: In this example, select TCP.
   - Listening Port: the frontend protocol and port used to receive requests and forward the requests to backend servers.

     In this example, enter the port number 80.
   - Enable Peak Bandwidth Limit: You can set a peak bandwidth to limit the service capabilities that applications on the ECS instances can provide.

     In this example, you do not need to set the peak bandwidth because the target SLB instance is billed by traffic.
   - Scheduling Algorithm: SLB supports the following scheduling algorithms. In this example, select Round-Robin (RR).
     - Weighted Round-Robin (WRR): Distributes requests according to the weights of backend servers. Servers with higher weights receive more requests.
     - Weighted Least Connections (WLC): In addition to the weights of backend ECS servers, the number of connections also contributes to scheduling changes.

A server with a higher weight value receives more connections. If the weights of two backend servers are the same, the server with the least number of connections receives more requests.

- Round-Robin (RR): Requests are distributed evenly and sequentially to backend ECS servers.

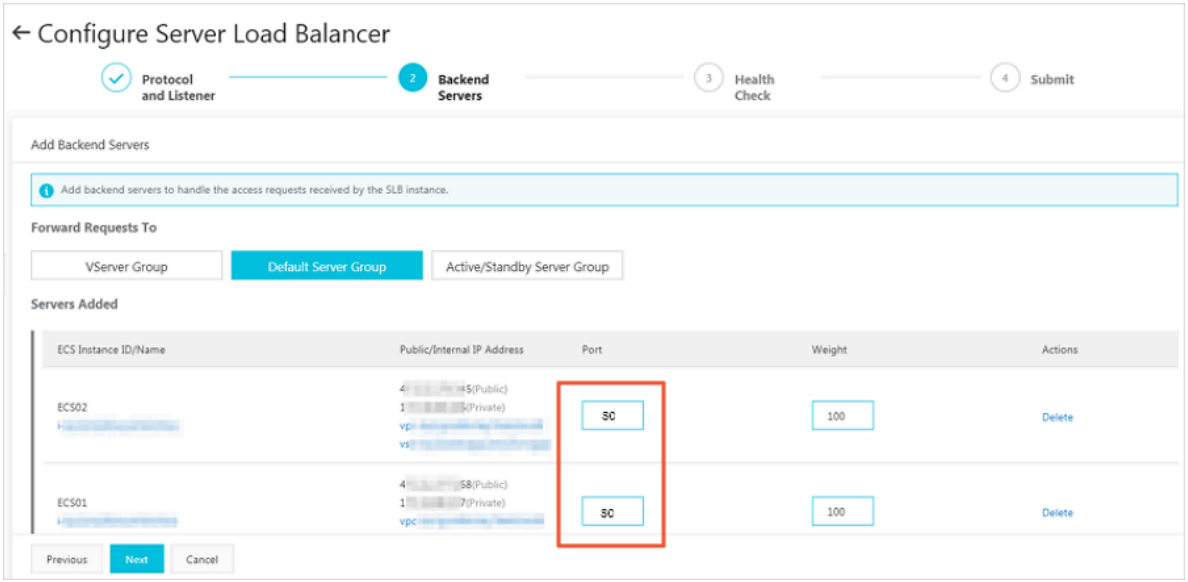


4. Click Next. On the Backend Servers tab, click Default Server Group, and then click Add More.

   a) On the Available Servers page, select the ECS instances ECS1 and ECS2 and click Next: Set Weight and Port.

   b) Configure ports and weights for the added backend servers.

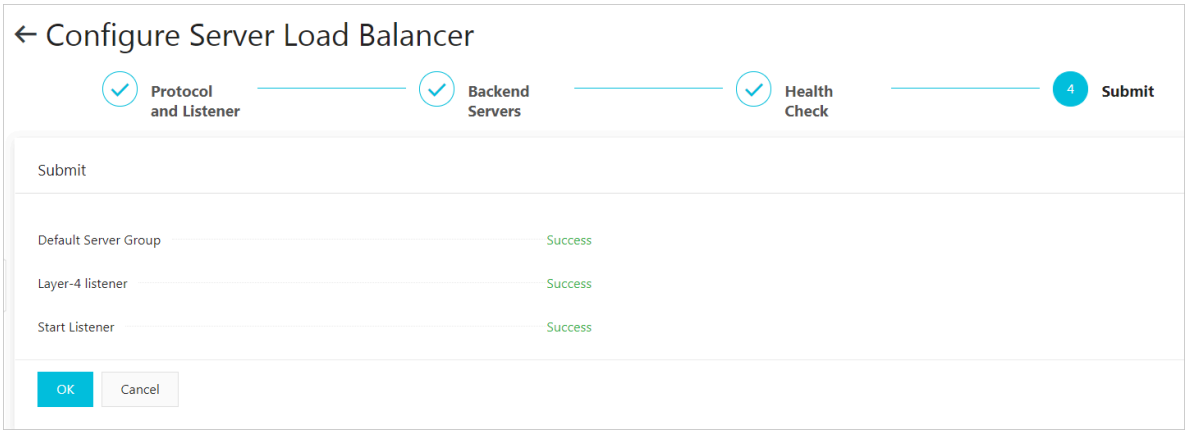
   - Port: the ports opened on backend ECS instances to receive requests. They can be the same in an SLB instance. In this example, set the backend port numbers to `80`.
   - Weight: An ECS instance with a higher weight receives a larger number of requests. The default value is 100 and we recommend that you use the default value.



5. Click Next to configure health checks. In this example, use default health check configurations.

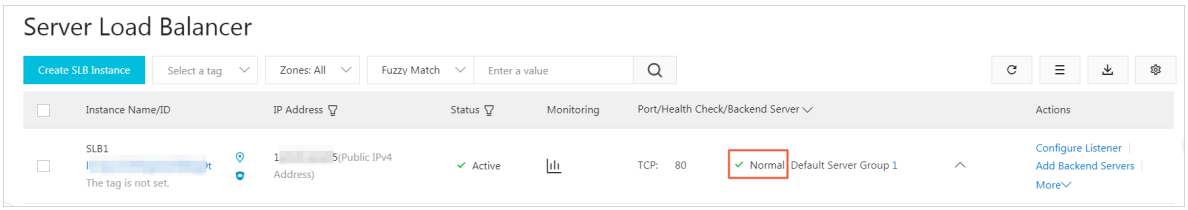
   With health check enabled, when an ECS instance is declared as unhealthy, SLB distributes requests to other healthy ECS instances and restores service to the original instance when it is healthy again.

6. Click Next. On the Submit page, click Submit.



7. Click OK. Go back to the Server Load Balancer page and click  .

   When the health check status of the backend server is Normal, it indicates that the backend server can process forwarded client requests.



8. In the web browser, enter the IP address of the SLB instance to test the service.

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 Delete an SLB instance

Delete the SLB instance when you no longer need the load balancing service to avoid additional charges. Deleting the Server Load Balancer instance does not delete or affect backend ECS instances.

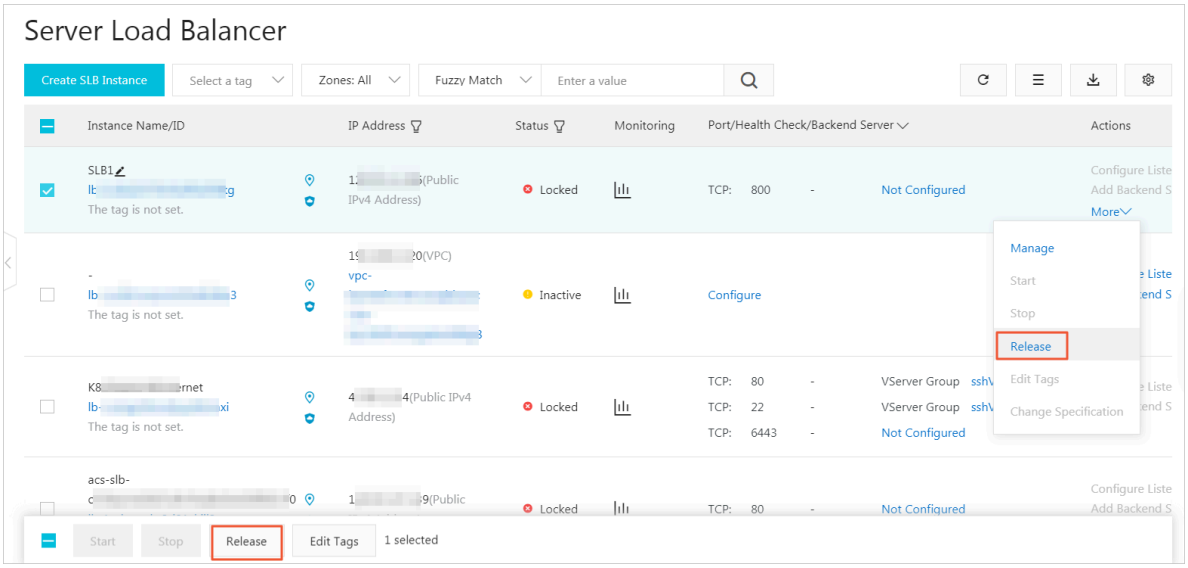
Context

Note:

- If you have resolved a domain name to the SLB endpoint, resolve it to another IP address first to avoid service interruption.
- Only Pay-As-You-Go SLB instances can be released. Subscription SLB instances are automatically released if they are not renewed timely.
- The backend ECS instances are still running after the SLB instance is released. You can release the backend ECS instances if you do not need them anymore.

Procedure

1. Log on to the SLB console.
2. On the Instances page, select the region where the instance is located.
3. Locate the target instance, click Release at the bottom of the list or click More > Release in the actions column.



4. In the Release dialog box, select Release Now or Release on Schedule.

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

5. Click Next.
6. Click OK to release the SLB instance.