Alibaba Cloud
Server Load Balancer

User Guide

MORE THAN JUST CLOUD | C-J 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 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 intellectu

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

Contents

1 Server Load Balancer instance

1.1 SLB instance overview

An SLB instance is a running entity of the Server Load Balancer service. To use the load balancing service, you must create an SLB instance first, and then add listeners and backend servers to it.
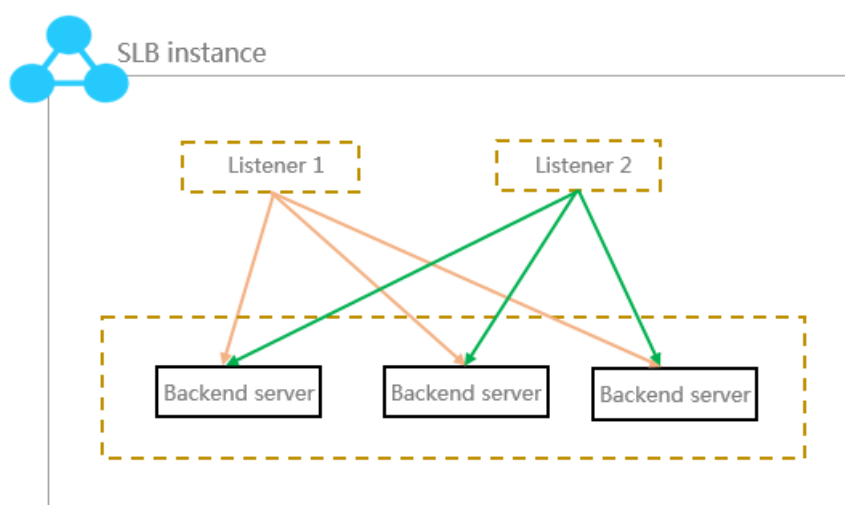
Alibaba Cloud provides Internet SLB service and intranet SLB service. A public or a private IP address is allocated to the SLB instance according to the instance type you select.

Internet SLB instances

An Internet SLB instance distributes client requests over the Internet to backend ECS servers according to configured forwarding rules.

After you create an Internet Server Load Balancer instance, the system will allocate a public IP to the instance. You can resolve a domain name to the public IP to provide public services.

Intranet SLB instances

Intranet SLB instances can only be used inside Alibaba Cloud and can only forward requests from clients that can access the intranet of SLB.

For an intranet SLB instance, you can further select the network type:

- Classic network

  If you choose classic network for the intranet SLB instance, the IP of the SLB instance is allocated and maintained by Alibaba Cloud. The classic SLB instance can only be accessed by the classic ECS instances.

- VPC network

  If you choose VPC network for the intranet SLB instance, the IP of the SLB instance is allocated from the CIDR of the VSwitch that the instance belongs to. SLB instances of the VPC network can only be accessed by ECS instances in the same VPC.
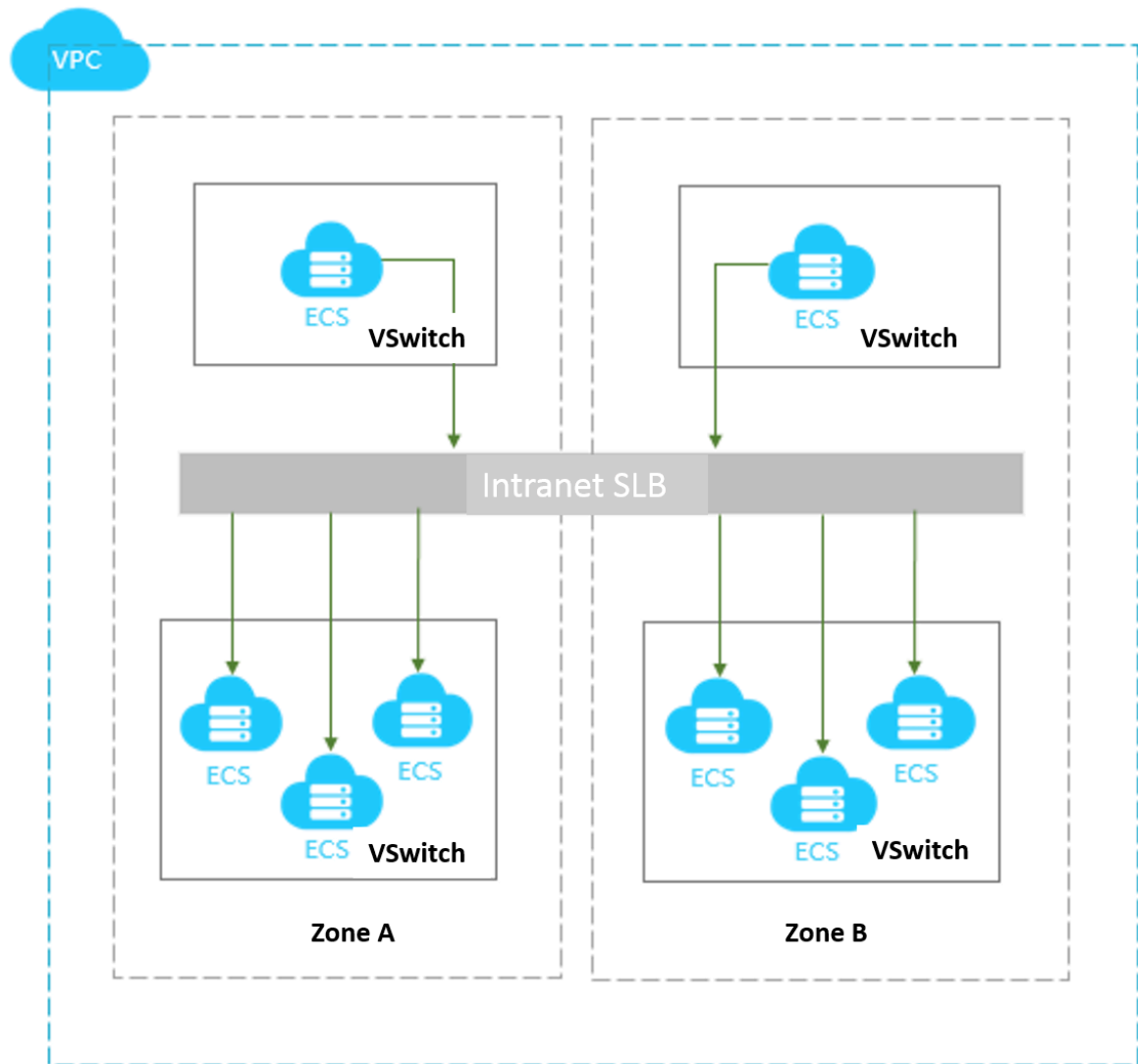
1.2 Guaranteed-performance instances

Alibaba Cloud plans to charge specification fee on guaranteed-performance Server Load Balancer instances from April 1th, 2018, and continue to sell shared-performance Server Load Balancer instances.

1. What are guaranteed-performance instances?

A guaranteed-performance instance provides guaranteed performance metrics (performance SLA) and is opposite to a shared-performance instance. For a shared-performance instance, the performance metrics are not guaranteed and the resources are shared by all instances.

All instances are shared-performance instances before Alibaba launches guaranteed-performance instances. You can view the instance type on the console.

Hover your mouse pointer to the green icon of the target instance to view the performance metrics , as shown in the following figure.
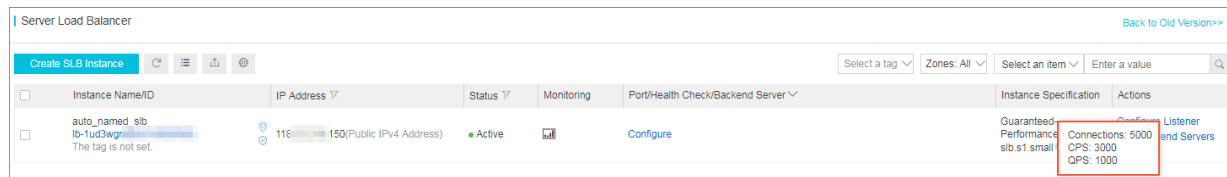
The following are three key performance metrics for guaranteed-performance instances:

• Max Connection

The maximum number of connections to a SLB instance. When the maximum number of connections reaches the limits of the specification, the new connection will be dropped.

• Connection Per Second (CPS)

The rate at which a new connection is established per second. When the CPS reaches the limits of the specification, the new connection will be dropped.

• Query Per Second (QPS)

The number of HTTP/HTTPS requests that can be processed per second. When the QPS reaches the limits of the specification, the new connection will be dropped.

Alibaba Cloud Server Load Balancer provides the following capacities for guaranteed-performance instances:

Type	Specification	Max Connection	Connection Per Second (CPS)	Query Per Second (QPS)
Specification 1	Small I (slb.s1. small)	5,000	3,000	1,000
Specification 2	Standard I (slb.s2 .small)	50,000	5,000	5,000
Specification 3	Standard II (slb. s2.medium)	100,000	10,000	10,000
Specification 4	Higher I (slb.s3. small)	200,000	20,000	20,000
Specification 5	Higher II (slb.s3. medium)	500,000	50,000	30,000
Specification 6	Super I (slb.s3. large)	1,000,000	100,000	50,000

If you want to use a larger specification, contact your customer manager.

2. How are guaranteed-performance instances billed?

Guaranteed-performance instances are billed as follows:

Total fee (per instance) = instance fee + traffic fee + specification fee

> **Note:**
>
> You can select a shared-performance intranet SLB instance or a guaranteed-performance intranet SLB instance. Specification fee is charged on intranet guaranteed-performance instances in the same way as Internet guaranteed-performance instances. But intranet guaranteed-performance instances are free from traffic fee and instance fee.

The specification fee of a performance-guarantee instance is charged by usage. No matter what kind of specification you choose, the instance specification fee will be charged according to the specification you actually use.

For example, if you purchase the slb.s3.large specification (1,000,000; CPS 500,000; QPS 50,000) and the actual usage of your instance in an hour is as follow:

Max Connection	CPS	QPS
90,000	4,000	11,000

- From the perspective of Max Connection, the actual metrics 90,000 occurs between the limit 50 ,000 defined in the Standard I (slb.s2.small) specification and the limit 100,000 defined in the Standard II (slb.s2.medium) specification. Therefore, the specification of the Max Connection metrics in this hour is Standard II (slb.s2.medium).

- From the perspective of CPS, the actual metrics 4,000 occurs between the limit 3,000 defined in the Small I (slb.s1.small) specification and the limit 5,000 defined in the Standard I (slb.s2. small) specification. Therefore, the specification of the CPS metrics in this hour is Standard I (slb.s2.small).

- From the perspective of QPS, the actual metrics 11,000 occurs between the limit 10,000 defined in the Standard II (slb.s2.medium) specification and the limit 20,000 defined in the Higher I (slb.s3.small) specification. Therefore, the specification of the QPS metrics in this hour is Higher I (slb.s3.small).

  Comparing these three metrics, the specification of the QPS metrics is highest, therefore, the specification fee of the instance in this hour is charged at the price of the Higher I (slb.s3.small) specification.

The following figure is an example showing how the specification fee is billed for an SLB instance in the first three hours:
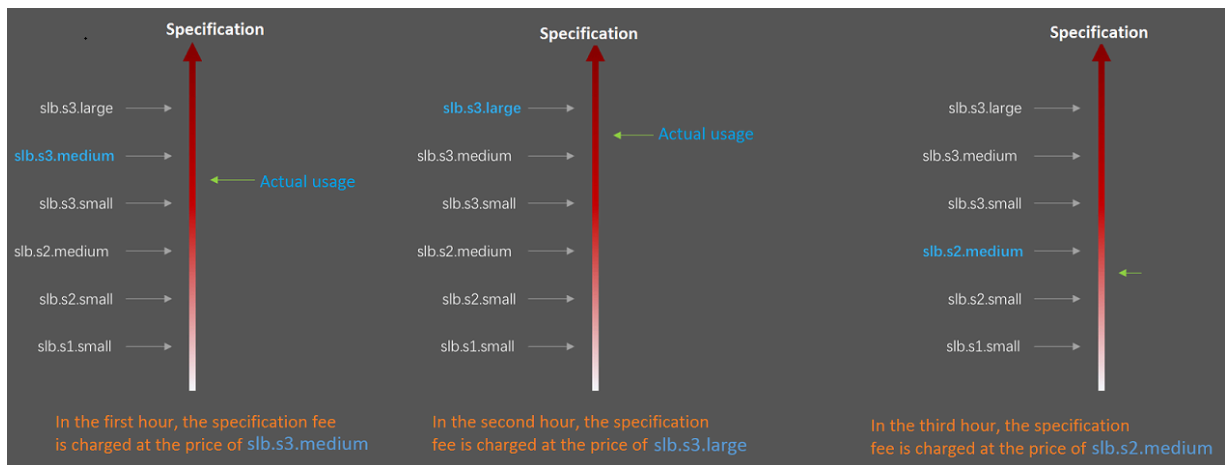
The billing of the guaranteed-performance instances is flexible. The specification you select when purchasing an instance is the performance limitation of the instance. For example, if slb.s3 .medium is selected, the new connections are dropped when the HTTP requests in one second reach 30,000.

3. What is the price of each specification?

The following table lists the price of each specification. In addition to the specification fee, you are also charged for instance fee and traffic fee. For more information, see *Pay-As-You-Go*.

Region	Specification	Max Connection	CPS	QPS	Specification fee (USD/Hour)
China (Hangzhou) China (Zhangjiakou) China (Hohhot) China (Qingdao) China (Beijing) China (Shanghai) China (Shenzhen)	Specification 1: Small I (slb.s1.small)	5,000	3,000	1,000	Free
	Specification 2: Standard I (slb.s2. small)	50,000	5,000	5,000	0.05
	Specification 3: Standard II (slb.s2. medium)	100,000	10,000	10,000	0.10
	Specification 4: Higher I (slb.s3.small)	200,000	20,000	20,000	0.20
	Specification 5: Higher II (slb.s3.medium)	500,000	50,000	30,000	0.31

Region	Specification	Max Connection	CPS	QPS	Specification fee (USD/Hour)
	Specification 6: Super I (slb.s3.large)	1,000,000	100,000	50,000	0.51
Singapore Malaysia (Kuala Lumpur) Indonesia (Jakarta) India (Mumbai) US (Silicon Valley) US (Virginia) Hong Kong	Specification 1: Small I (slb.s1.small)	5,000	3,000	1,000	Free
	Specification 2: Standard I (slb.s2. small)	50,000	5,000	5,000	0.06
	Specification 3: Standard II (slb.s2. medium)	100,000	10,000	10,000	0.12
	Specification 4: Higher I (slb.s3.small)	200,000	20,000	20,000	0.24
	Specification 5: Higher II (slb.s3.medium)	500,000	50,000	30,000	0.37
	Specification 6: Extra I (slb.s3.large)	1,000,000	100,000	50,000	0.61

4. How to select a guaranteed-performance instance?

Because the specification fee is billed based on the actual usage, we recommend that you select the largest specification (slb.s3.large). This guarantees the business flexibility (flexibility) and will not cause extra costs. If your traffic does not reach the largest specification, you can select a more reasonable specification, such as slb.s3.medium.

5. Can I modify the specification after the instance is created?

Yes. You can change the specification at any time and the change takes effect immediately.

Note:

- Once a shared-performance instance is changed to a guaranteed-performance instance, it cannot be changed back.

- Some SLB servers are deployed in old clusters. If you change a shared-performance instance to a guaranteed-performance instance, a brief disconnection of service may occur for 10 to 30 seconds. We recommend that you change the specification when the business is not busy.

- The IP of the SLB instance will not be changed after you changing the instance type or the specification.

Caution

When you change the configuration of an SLB instance or change a shared-performance instance to a guaranteed-performance instance, a brief disconnection of service may occur for 10 to 30 seconds. We recommend that you perform this operation when the service is not busy or after the service migrates to another SLB instance by using Global Server Load Balancer. (Changes made to the billing method and network bandwidth of the SLB instance will not affect the service.)

[I Agree]   [No, Not Now]

6. When will the guaranteed-performance instances be charged?

Alibaba Cloud plans to charge specification fee on guaranteed-performance Server Load Balancer instances from April 1st, 2018, and continue to sell shared-performance Server Load Balancer instances.

The charging of specification fee will take effect in batches as follows:

• The first batch:

Effective time: From April 1 to April 10

Regions: Singapore, Malaysia (Kuala Lumpur), Indonesia (Jakarta), India (Mumbai), US (Silicon Valley), US (Virginia)

• The second batch:

Effective time: From April 11 to April 20

Effective regions: China (Hangzhou), China (Zhangjiakou), China (Hohhot), China (Hong Kong)

• The third batch:

Effective time: From April 21 to April 30

Effective regions: China (Qingdao), China (Beijing), China (Shanghai), China (Shenzhen)

7. After Alibaba Cloud starts to charge specification fee on guaranteed-performance instances, will extra fees be charged on shared-performance instances?

No.

The billing of the original shared-performance instances is the same if you do not change it to a performance-guaranteed instance. However, if you change the shared-performance instance to the guaranteed-performance one, the specification fee will be charged.

8. Why sometimes guaranteed-performance instances cannot reach the performance limit as defined in the specification?

It applies the cask theory.

Guaranteed-performance instances do not guarantee that the three metrics can reach the specification limits at the same time. The limitation is triggered as long as a metric first reaches the limitation defined in the specification.

For example, you have purchased a guaranteed-performance instance of the Higher I (slb.s3. small) specification. When the QPS of the instance reaches 20,000 but the number of maximum connections does not reach 200,000, the new connections are still dropped because the QPS has reached the limitation.

9. Can I still buy shared-performance instances?

Yes, you can.

However, shared-performance instances will be phased out in the future. Please pay attention to the official announcement.

10. Will intranet SLB instances be charged for specification fee?

If the intranet SLB instance is a shared-performance instance, no specification fee is charged. If the intranet SLB instance is a guaranteed-performance instance, corresponding specification fee is charged, and no other fees are charged. The specification fees are collected as the same as the Internet guaranteed-performance instances, but intranet guaranteed-performance instances are free from instance fee and traffic fee.

1.3 Network traffic flow

As a traffic forwarding service, SLB forwards requests from clients to backend servers through the SLB cluster and then the backend servers return the responses to SLB through the intranet.

Inbound network traffic

SLB distributes incoming traffic according to forwarding rules configured on the console or API. The inbound network traffic flow is shown as *Figure 1-1: Inbound network traffic*.

Figure 1-1: Inbound network traffic
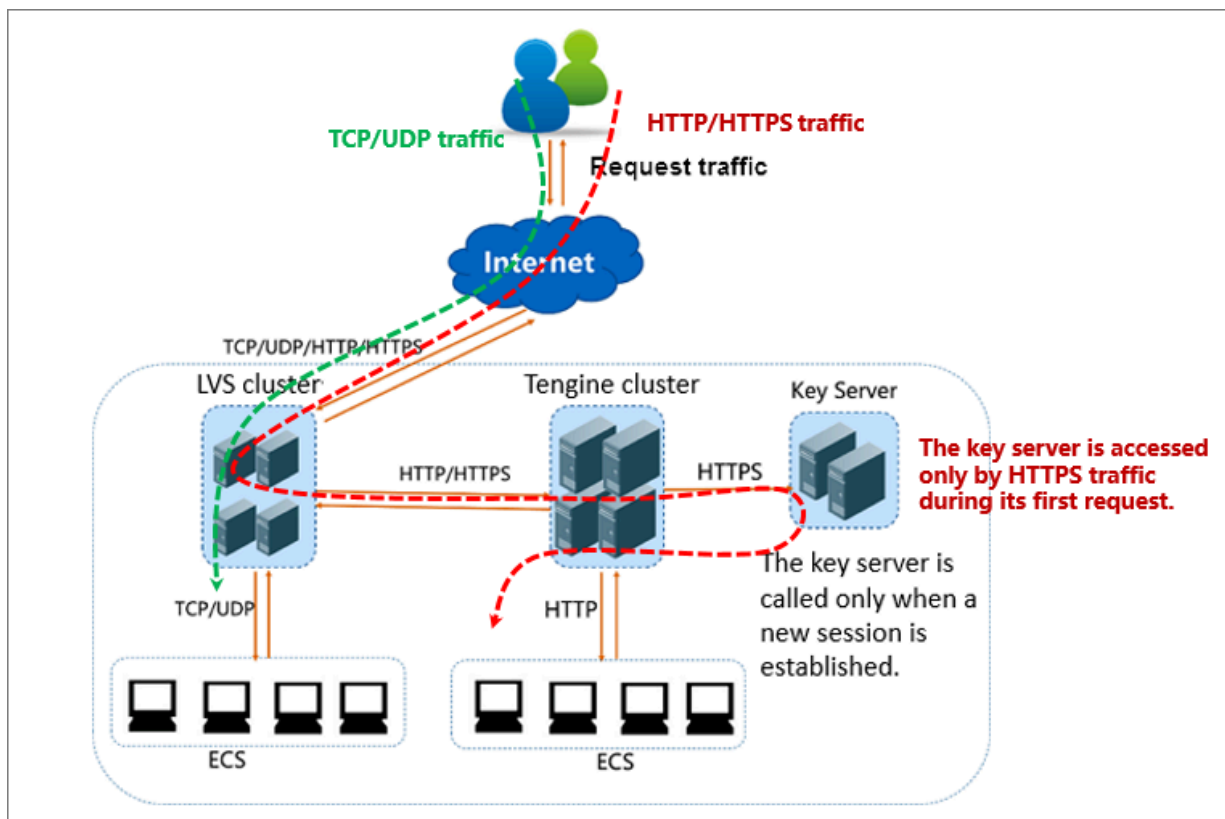
1. For TCP/UDP protocol and HTTP/HTTPS protocol, the incoming traffic must be forwarded through the LVS cluster first.

2. Numerous inbound traffic is distributed evenly among all node servers in the LVS cluster, and the node servers synchronizes session to guarantee high availability.

   • For Layer-4 listeners (the frontend protocol is UDP or TCP), the node servers in the LVS cluster distribute requests directly to backend ECS instances according to the configured forwarding rules.

   • For Layer-7 listeners (the frontend protocol is HTTP), the node servers in the LVS cluster first distribute requests to the Tengine cluster. Then, the node servers in the Tengine cluster

distribute the requests to backend ECS instances according to the configured forwarding
rules.

- For Layer-7 listeners (the frontend protocol is HTTPS), the request distribution is similar to
  the HTTP protocol. However, before distributing the requests to backend ECS instances,
  the system will call the Key Server to validate certificates and decrypt data packets.

Outbound network traffic

SLB communicates with backend ECS instances through the intranet.

- If the backend ECS instances only need to handle the traffic distributed from SLB, no public
  bandwidth (EIP, NAT Gateway and public IP) is required.

  If the ECS instance already has a public IP, but only processes requests from SLB, no fee is
  generated.

- However, if you want to provide external services from a backend ECS instance, or the
  backend ECS instance needs to access the Internet, you must configure a public IP, an EIP or
  a NAT Gateway.

The outbound network traffic flow is shown as *Figure 1-2: Outbound network traffic*.

Figure 1-2: Outbound network traffic
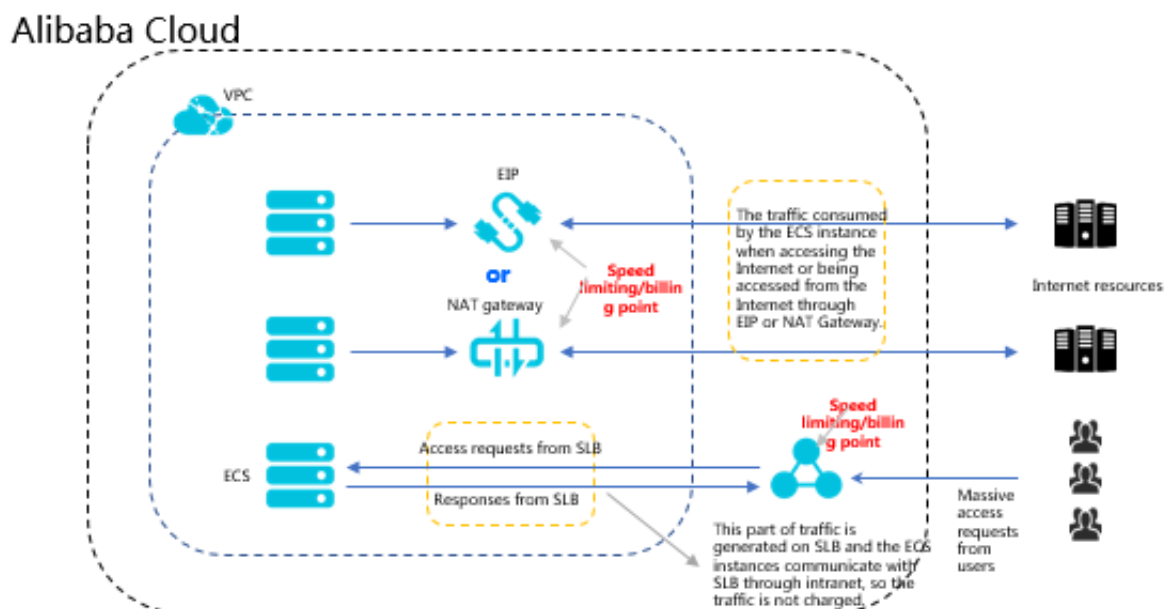
In general, the traffic goes out from where it comes in:

1. For the traffic coming from SLB, billing and speed limitation are done on SLB. You are charged
   by the outbound traffic and not the inbound traffic (the rule may change in the future). SLB

communicates with the backend ECS instances through the intranet and no traffic fee is
charged for the internal communication.

2. For the traffic coming from the EIP or NAT Gateway, billing and speed limitation are done on
   EIP or NAT Gateway. If the ECS instance has configured a public IP when it is created, the
   billing and speed limitation are done on the ECS instance.

3. SLB only provides the function of being accessed from the Internet. That is, a backend ECS
   instance can only access the Internet when it responds to the request forwarded by SLB. If you
   want to actively access the Internet from a backend ECS instance, you must configure a public
   IP (configure EIP or NAT gateway) for the ECS instance.

4. A public IP (configured when you create an ECS instance), EIP, and NAT gateway can all
   achieve mutual Internet access (access or accessed), but they cannot forward traffic or balance
   traffic loads.

1.4 Create an SLB instance

Prerequisites

Before creating an SLB instance, make sure that you have properly prepared the environment. For
more information, see *Plan and prepare*.

Procedure

1. Log on to the *SLB console*.

2. In the left-side navigation pane, click **Instances** > **Server Load Balancer**, and click **Create
   SLB Instance** in the upper-left corner.

3. Configure the SLB instance according to the following information.

Configurat ion	Description
Region	Select the region where the SLB instance is located. 📋 **Note:** Make sure that the region of the SLB instance is the same as that of backend ECS instances.

Configuration	Description
Zone Type	Displays the zone type of the selected region. The zone of a cloud product refers to a set of independent infrastructure and is usually represented by Internet data centers (IDCs). Different zones have independent infrastructure (network, power supply, air-conditioning and so on). Therefore, an infrastructure fault in one zone will not affect other zones. A zone belongs to a specific region, however, a single region may have one or more zones. SLB has deployed multi-zone in most regions. • Single zone: The SLB instance is deployed only in one zone. • Multi-zone: The SLB instance is deployed in two zones. By default, the instance in the primary zone is used to distribute traffic. If the primary zone is faulty, the instance in the backup zone will automatically take over the load balancing service.
Primary Zone	Select the primary zone for the SLB instance. The primary zone carries traffic in normal conditions.
Backup Zone	Select the backup zone for the SLB instance. The backup zone only takes over traffic when the primary zone is unavailable.
Instance Spec	Select a performance specification for the instance. The performance metrics vary by specification. For more information, see *Guaranteed-performance instances*.
Instance Type	Select the instance type based on your business needs. A public or a private IP address is allocated to the SLB instance based on the instance type. For more information, see *SLB instance overview*. • Internet: An Internet SLB instance only provides a public IP and you can access the SLB service from the Internet. • Intranet: An intranet SLB instance only provides a private IP and you can only access the SLB service from the intranet.
Network type	If the selected instance type is intranet, you have to select a network type for the instance. • Classic network: The IP of the instance is allocated and managed by Alibaba Cloud in a unified manner. • VPC: The IP of the instance is allocated from the VSwitch CIDR block specified by you.
Quantity	Select the number of instances to create.

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

1.5 Create an IPv6 instance

Server Load Balancer supports creating IPv6 instances. After an IPv6 instance is created, the system allocates a public IPv6 address to the instance to forward requests from IPv6 clients.

Context

IPv6 is the abbreviation of Internet Protocol Version 6. IPv6 is the next-generation IP protocol designed by IETF (Internet Engineering Task Force) to replace the current version of IP protocol (IPv4). By extending the length of IPv4 address from 32 bits to 128 bits, it expands the address space by 79,228,162,514,264,337,593,543,950,336 times. After IPv6 is used, each grain of sand on the world can be allocated with an IP address.

> **Notice:**
>
> - Currently, only Zone E and Zone F in the China (Hangzhou) region as well as Zone F and Zone G in the China (Beijing) region support creating IPv6 instances and the instances must be guaranteed-performance instances.
>
> - The Internet IPv6 network environment is still in the early stages of construction, and some links may cannot be accessed. If such problem occurs, submit a ticket. Besides, SLA is not provided in the pre-release stage.
>
> - Because IPv6 has a longer IP head than IPv4, when you use a UDP listener on an IPv6 SLB instance, you must ensure that the MTU of the NIC communicating with the SLB on the backend server (ECS instance) is not greater than 1480 (some applications require synchronizing its configuration files based on this MTU value), otherwise the packets may be discarded because they are too large.
>
>   If you use a TCP/HTTP/HTTPS listener, no additional configurations are required because the TCP protocol supports MSS auto-negotiation.

SLB IPv6 has the following features:

- Smooth migration, which is not sensed by the service

  You can directly bind ECS instances using IPv4 addresses to an IPv6 SLB instance and smoothly migrate the service to IPv6 without transforming the original system.

  Adding the IPv6 entry has no impact on the original IPv4 service. If the traffic volume increases, you only need to increase the backend ECS instances.

- IPv6 access control ensures more security and reliable service deployment

Alibaba Cloud SLB supports IPv6 access control. You can configure the access control list according to your business needs.

- A blacklist can effectively block the access of malicious addresses to the SLB service.

- If a whitelist is configured, only addresses in the whitelist can access the SLB service.

Procedure

1. Log on to the *SLB console*.

2. Select **Instances** > **Server Load Balancer**.

3. On the **Server Load Balancer** page, click **Create SLB Instance** in the upper-left corner.

4. Configure the SLB instance. For the IP version, select **IPv6**.

   Other configurations are the same as configurations of common instances. See *SLB configurations*.

   > **Note:**
   > Currently, only Zone E and Zone F in the China (Hangzhou) region as well as Zone F and Zone G in the China (Beijing) region support creating IPv6 instances and the instances must be guaranteed-performance instances.

5. Go back to the Server Load Balancer page to view the created IPv6 instance.

Result

Once the IPv6 instance is created, the system allocates an IPv6 address to it.


1.6 Start or stop an SLB instance

You can start or stop an SLB instance at any time. After being stopped, an SLB instance does not receive or forward requests any more.

Procedure

1. Log on to the *SLB console*.

2. In the left-side navigation pane, click **Instances** > **Server Load Balancer**.

3. Select a region and find the target instance.

4. In the **Actions** column, click **More** > **Start** or **More** > **Stop**.
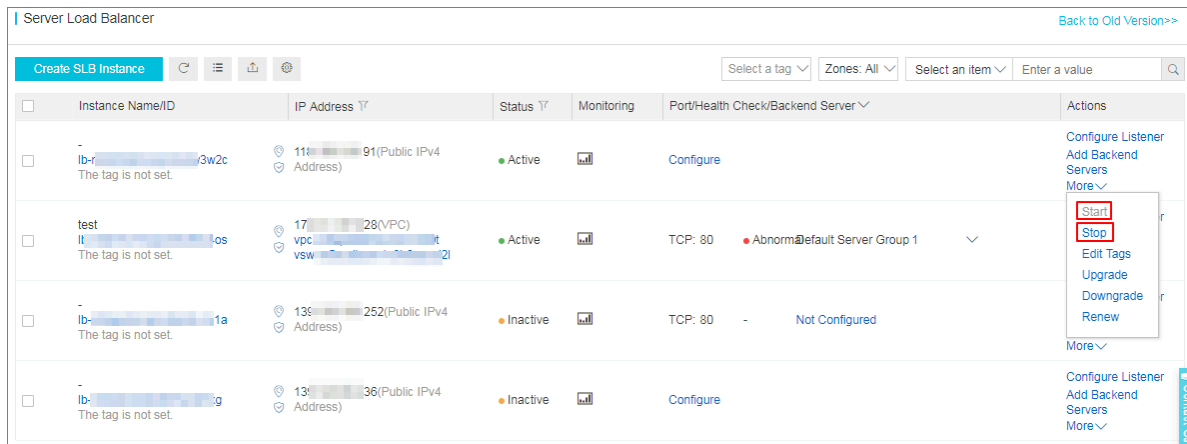
5. If you want to start or stop multiple instances at a time, select the target instances and click **Start** or **Stop** at the lower part of the page.
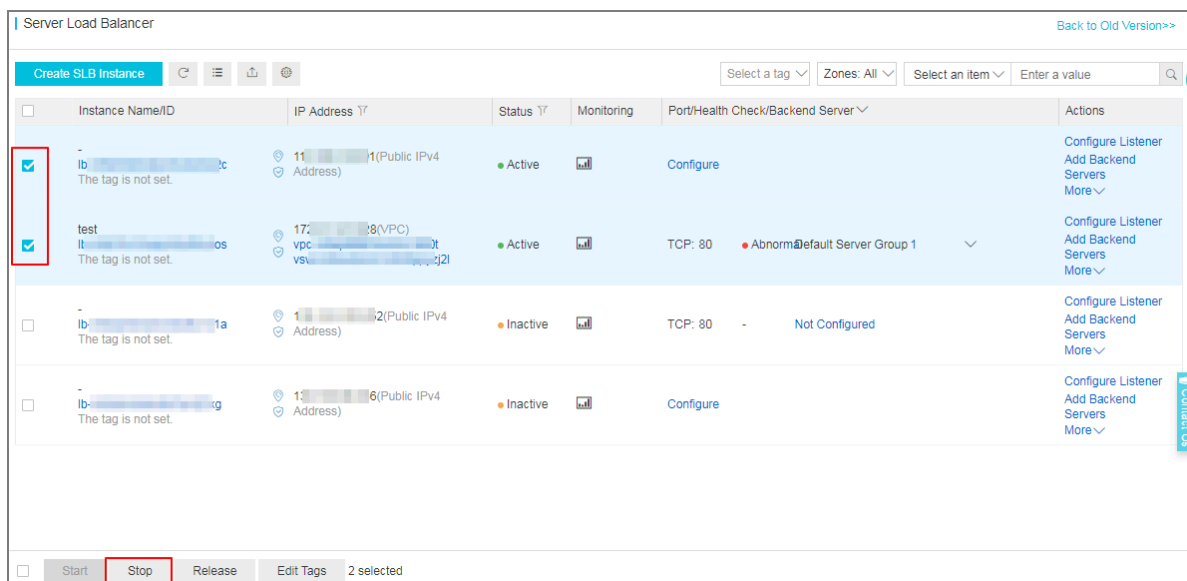
1.7 Bind an EIP

You can bind an EIP to an SLB instance of the VPC network. After being bound to an EIP, the SLB instance can forward requests from the Internet.

Procedure

1. Log on to the *SLB console*.

2. In the left-side navigation pane, click **Instances** > **Server Load Balancer**.

3. Select a region and find the target instance.

> **Note:**
>
> Ensure that the SLB instance is of the VPC network.

4. Click **More** > **Bind EIP**.

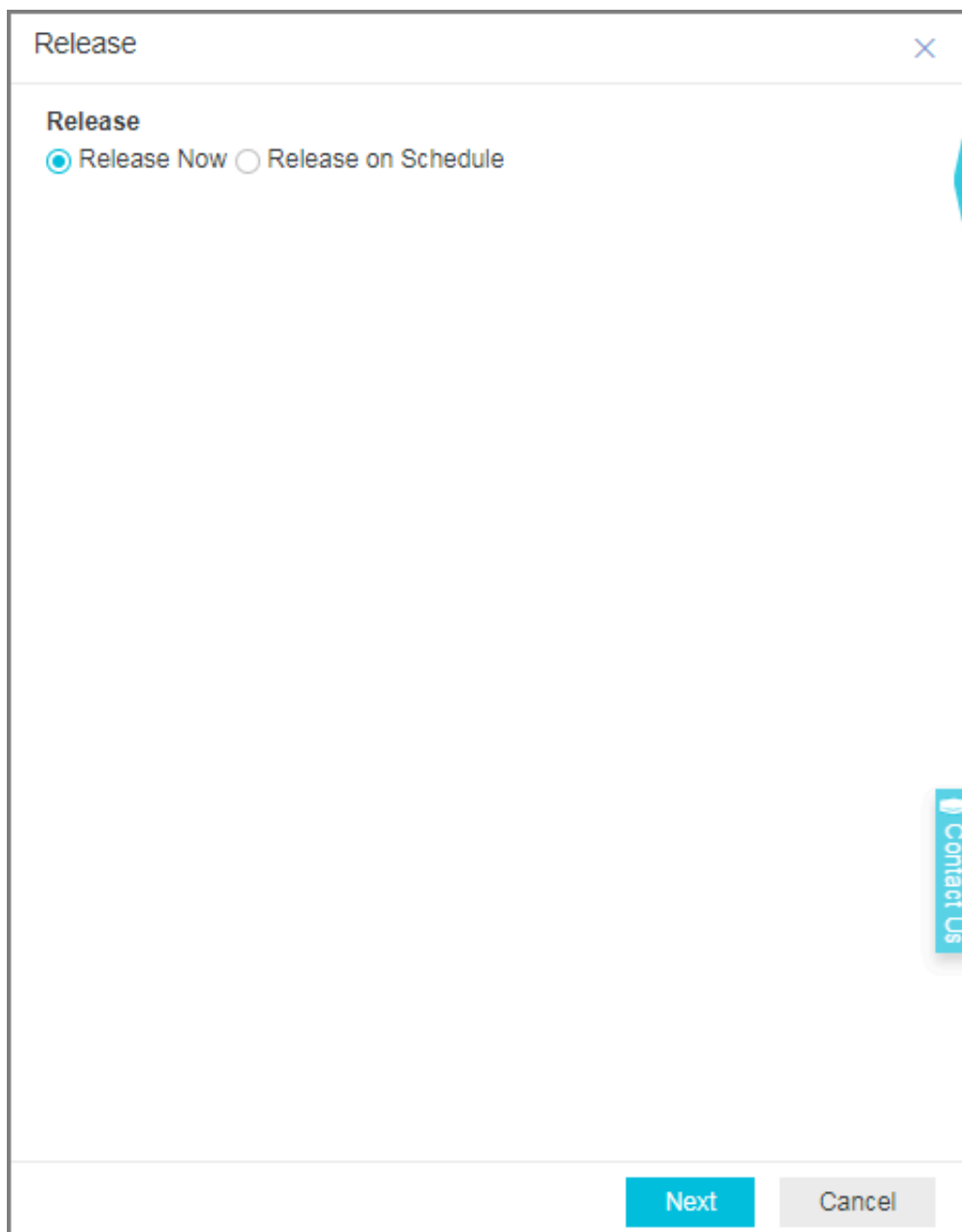5. Select an EIP and click **OK**.

1.8 Release an SLB instance

Context

You can release an SLB instance immediately or on schedule.

Procedure

1. Log on to the *SLB console*.

2. Find the target instance and then click **More** > **Release**.

   You can select multiple SLB instances at a time and click **Release** in the lower part of the page to release SLB instance in batches.

3. On the **Release** page, select to release now or release on schedule.

> 📋 **Note:**
>
> The system executes the release operation every one hour or every 30 minutes, but will stop
>
> billing at the release time you set.

4. Click **Next**.

5. Confirm the displayed information and click **OK** to release the instance.

1.9 Manage tags

With tagging, you can classify Server Load Balancer instances by tags.

Each tag consists of a key and a value. Note the following limits when using tags:

- A tag cannot exist on its own and must be bound to an SLB instance.

- Up to 10 tags can be bound to an SLB instance.

- The key of each tag added to an instance must be unique. Otherwise, tags of the same key will be overwritten.

- Tags cannot be used across regions and are region-specific resources. For example, tags created in China (Hangzhou) are invisible in China (Shanghai).

Add a tag

To add a tag, complete these steps:

1. Log on to the *SLB console*.

2. In the left-hand navigation pane, select **Instances** > **Server Load Balancer**.

3. Select a region and find the target instance.

4. In the **Actions** column, select **More** > **Edit Tags**.
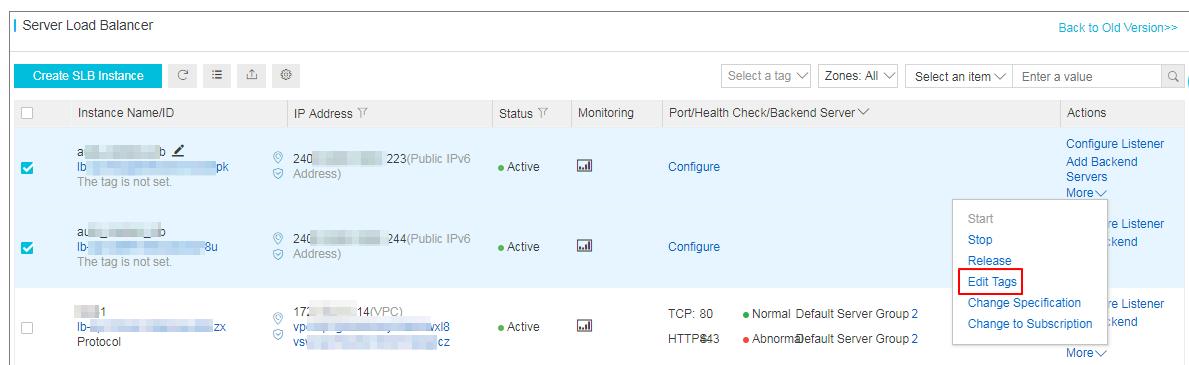
5. On the **Edit Tags** page, complete these steps:

   a. If there are available tags, click **Saved Tags** and the select the tag to add.

   b. If you want to create a new tag, on the **Edit Tags** page, click **New Tag**, then enter the key and value of the new tag and click **OK**.

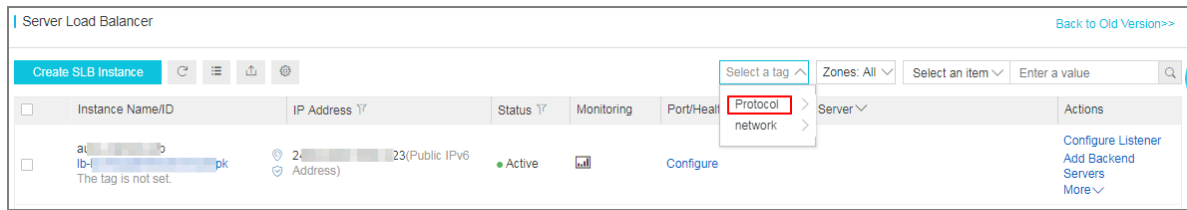   **c.** Click **OK**.

Search instances using a tag

    To search instances using a tag, complete these steps:

    1. Log on to the *SLB console*.

    2. In the left-hand navigation pane, click **Instances** > **Server Load Balancer**.

    3. Select a region and find the target instance.

    4. Click **Select a tag**, and then select the tag to be used as the search criteria.

5. You can click the delete icon next to the selected tag to clear the filter.

Delete a tag

SLB does not support deleting tags of multiple instances in batches. You can only remove the tags
of an instance at a time.

To delete a tag, complete these steps:

1. Log on to the *SLB console*.

2. In the left-hand navigation pane, click **Instances** > **Server Load Balancer**.

3. Select a region and find the target instance.

4. In the **Actions** column, select **More** > **Edit Tags**.

5. On the **Edit Tags** page, click the delete icon next to the tag to be removed, and then click **OK**.

> **Note:**
>
> If a tag is removed from one instance and is not bound to any other instances, the tag is
> removed from the system.
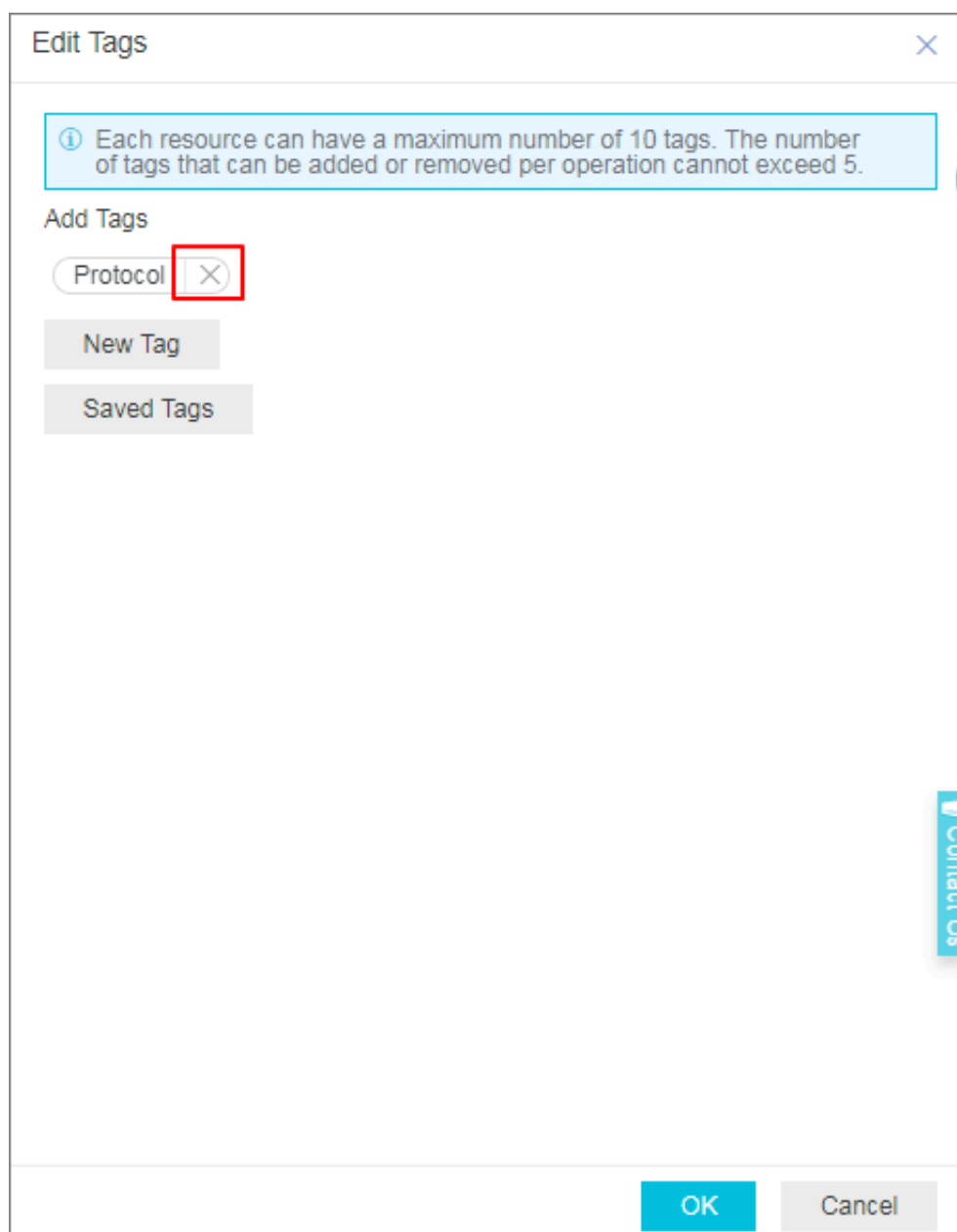
1.10 Expiring Instances

You can manage overdue instances.

Context

If you do not renew an expiring instance, the instance will be released automatically.

- After an instance is added to expiring instances, it will only be reserved for one day.

Procedure

1. Log on to the *SLB console*.

2. Select **Instances** > **Expiring Instances**.

3. View detailed information of overdue instances.

4. Click **Renew** in the **Actions** column of the target SLB instance, then the instance will added back to the Server Load Balancer list.

1.11 Change the instance specification

You can change a shared-performance instance to a guaranteed-performance instance, or change the specification of a guaranteed-performance instance.

Context

Before modifying the instance configuration, note the following:

- If you change a shared-performance instance to a guaranteed-performance instance, a brief disconnection of service may occur for 10 to 30 seconds.

    We recommend that you change the configuration in the low traffic period, or use DNS to schedule services to other SLB instances first before changing the configuration.

- After you change a shared-performance instance to a guaranteed-performance instance, you cannot change it back.

    You can use the (slb.s1.small) specification after changing the instance to a guaranteed-performance instance.

Procedure

1. Log on to the *SLB console*.

2. Select a region.

3. Find the target instance, select **More** > **Change Specification**.

4. In the **Configuration Upgrade** area, select a new specification, and complete the payment.

2 Listeners

2.1 Listener overview

After creating a Server Load Balancer instance, you need to configure a listener for it. The listener checks connection requests and then distributes them to backend servers according to the configured rules.

Alibaba Cloud provides Layer-4 (TCP and UDP protocols) and Layer-7 (HTTP and HTTPS protocols) load balancing services. Select the protocol based on your business needs:

Protocol	Description	Scenarios
TCP	- A connection-oriented protocol. A reliable connection must be established with the peer end before data can be sent and received. - Source address-based session persistence. - The source address is visible at the network layer. - Fast data transmission.	- Applicable to scenarios with high requirements on reliability and data accuracy but with tolerance for low speeds, such as file transmission, sending or receiving e-mails, and remote logon. - Web applications without special requirements. For more information, see *Add a TCP listener*.
UDP	- A non-connection-oriented protocol. Before sending data, UDP directly performs data packet transmission instead of making three handshakes with the other party. It does not provide error recovery and data retransmission. - Fast data transmission, however, the reliability is relatively low.	Applicable to scenarios with preference for real-time content over reliability, such as video chats and pushes of real-time financial quotations. For more information, see *Add a UDP listener*.
HTTP	- An application layer protocol mainly used to package data. - Cookie-based session persistence. - Use X-Forward-For to obtain the source IP address.	Applicable to applications that need to recognize data content, such as web applications and small-sized mobile games. For more information, see *Add an HTTP listener*.

Protocol	Description	Scenarios
HTTPS	• Similar to HTTP, but with an encrypted connection that prevents unauthorized access. • Unified certificate management service. Users can upload certificates to the Server Load Balancer and the decryption operations are completed directly on the Server Load Balancer	Applications requiring encrypted transmission For more information, see *Add an HTTPS listener*.

📋 **Note:**

Server Load Balancer supports HTTP/2 and WSS/WS protocols in all regions now. For more information, see *HTTP/2 FAQ* and *WS/WSS FAQ*.

2.2 Add a TCP listener

TCP listeners are applicable to scenarios with high requirements on reliability and data accuracy but with tolerance for low speed, such as file transmission, sending or receiving emails and remote logon. You can add a TCP listener to forward requests from the TCP protocol.
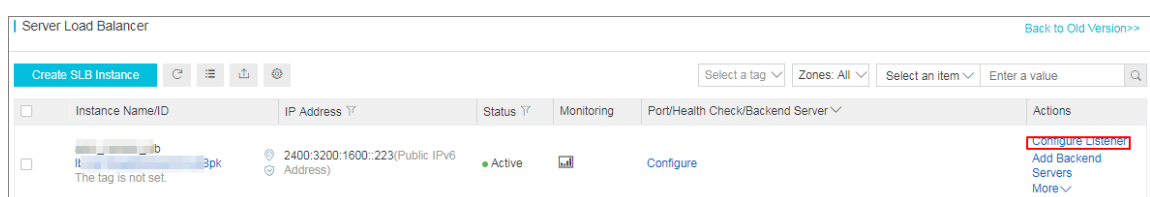
Prerequisites

Create an SLB instance.

Step 1 Open the listener configuration wizard

To open the listener configuration wizard, complete these steps:

1. Log on to the *SLB console*.

2. In the left-side navigation pane, select **Instances** > **Server Load Balancer**.

3. Select the region of the instance.

4. Select one of the following methods to open the listener configuration wizard:

   • On the **Server Load Balancer** page, find the target instance and then click **Configure Listener**.

- On the **Server Load Balancer** page, click the ID of the target SLB instance. On the

  Listeners page, click **Add Listener**.
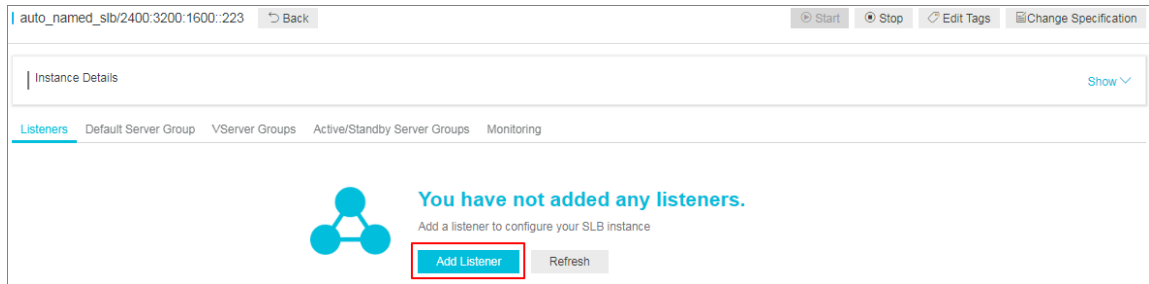
Step 2 Configure a TCP listener

To configure a UDP listener, complete these steps:

1. On the **Protocol and Listener** page, configure the TCP listener according to the following
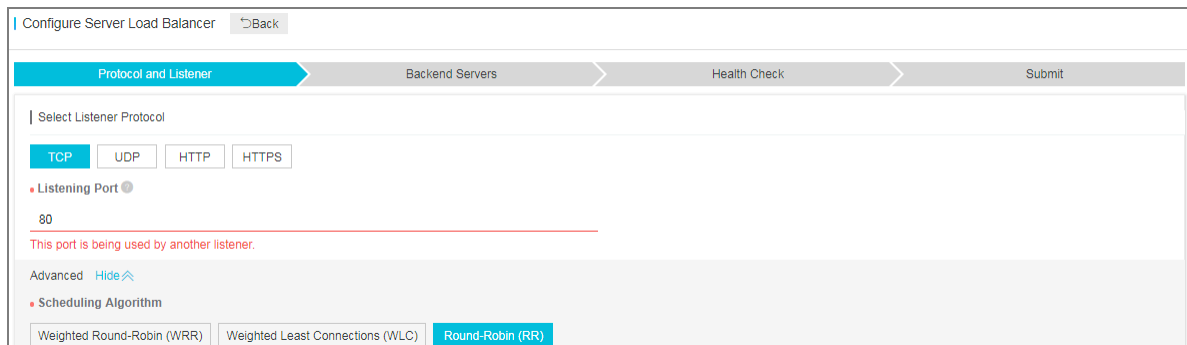
   information.

Configuration	Description
Select Listener Protocol	Select the protocol type of the listener. In this tutorial, select **TCP**.
Listening Port	The listening port used to receive requests and forward the requests to backend servers. The port number is in the range of 1-65535. **Note:** The listening ports must be unique in a Server Load Balancer instance.
Advanced configurations	

Configuration	Description

Scheduling Algorithm	Server Load Balancer supports three scheduling algorithms: round robin, weighted round robin (WRR), and weighted least connections (WLC).
	
	 • **Weighted Round-Robin (WRR)**: Backend servers with higher weights receive more requests than those with smaller weights.
	
	 • **Round-Robin (RR)**: Requests are evenly and sequentially distributed to the backend servers.
	
	 • **Weighted Least Connections (WLC)**: A server with a higher weight will receive a larger percentage of live connections at any one time. When the weight value is the same, a backend server with a smaller number of connections is more frequently (and probably) accessed.
	
	 • **Consistency Hash (CH)**:
	
	    - **Source IP**: The consistent hash based on the source IP address. The same source IP addresses are scheduled to the same backend server.
	
	    - **Tuple**: The consistent hash based on the quaternion (source IP + destination IP + source port + destination port). The same steams are scheduled to the same backend server.

Inside the Description cell (Note box):

> 📋 **Note:**
>
> The consistency hash (CH) algorithm is only supported in the following areas currently:
>
> - Japan (Tokyo)
> - Australia (Sydney)
> - Malaysia (Kuala Lumpur)
> - Indonesia (Jakarta)
> - Germany (Frankfurt)
> - US (Silicon Valley)
> - US (Virginia)
> - UAE (Dubai)
> - China (Hohhot)

Configuration	Description
Enable Access Control	Select whether to enable session persistence. If session persistence is enabled, all session requests from the same client are sent to the same backend server. For TCP listeners, session persistence is based on IP addresses. Requests from the same IP address are forwarded to the same backend server.
Enable Access Control	Select whether to enable the access control function.
Access Control Method	Select an access control method after enabling the access control function: • **Whitelist**: Only requests from IP addresses or CIDR blocks in the selected access control lists are forwarded. It applies to scenarios where the application only allows access from some specific IP addresses. Enabling whitelist poses some business risks. After a whitelist is configured, only the IP addresses in the list can access the listener. If you enable the whitelist without adding any IP entry in the corresponding access control list, all requests are forwarded. • **Blacklist**: Requests from IP addresses or CIDR blocks in the selected access control lists are not forwarded. It applies to scenarios where the application only denies access from some specific IP addresses. If you enable a blacklist without adding any IP entry in the corresponding access control list, all requests are forwarded.
Access Control List	Select an access control list as the whitelist or the blacklist. **Note:** An IPv6 instance can only bind IPv6 access control lists and an IPv4 instance can only bind IPv4 access control lists. For more information, see *Configure an access control list*.

Configuration	Description
Enable Peak Bandwidth Limit	Select whether to configure the listening bandwidth. If the SLB instance is billed by bandwidth, you can set different peak bandwidths for different listeners to limit the traffic passing through the listeners. The sum of the peak bandwidths of all listeners under an instance cannot exceed the bandwidth of that instance. By default, all listeners share the bandwidth of the SLB instance. **Note:** Instances billed by traffic have no peak bandwidth limit by default.
Idle Timeout	Specify the idle connection timeout in seconds. Valid value: 10-900.
Listener Name	Configure the name of the listener.
Get Client Source IP Address	The backend server of a Layer-4 listener can directly obtain the real IP of the client.
Automatically Enable Listener after Creation	Choose whether to enable listener after the listener is configured. The listener is enabled by default.

2. Click **Next**.

| Configure Server Load Balancer   ⤺Back |

| Protocol and Listener | Backend Servers | Health Check | Submit |

| Select Listener Protocol

| TCP | UDP | HTTP | HTTPS |

▪ Listening Port ⓘ

80

This port is being used by another listener.

Advanced   Hide⌃

▪ Scheduling Algorithm

| Weighted Round-Robin (WRR) | Weighted Least Connections (WLC) | Round-Robin (RR) |

Step 3 Add backend servers

Add backend servers to process requests. You can use the default server group configured for the instance, or configure a VServer group or an active/standby server group for the listener. For more information, see *Backend server overview*.

In this tutorial, the default server group is used:

1. Select **Default Server Group** and then click **Add**.

2. Select the ECS instances to add and then click **Add to Selected Server List**. Click **OK**.

3. Configure the ports and weights of the added backend servers.

   • Port

   The port opened on the backend server (ECS instance) to receive requests. The port number is in the range of 1-65535. Ports of backend servers can be the same in an SLB instance.

- Weight

  The weight of the backend server (ECS instance). An ECS instance with a higher weight will receive a larger number of connection requests.

  > **Note:**
  >
  > If the weight is set to 0, no requests will be sent to the ECS instance.

4. Click **Next**.

Step 4 Configure health check

Server Load Balancer checks the service availability of the backend servers (ECS instances) by performing health checks. The health check function improves the overall availability of your services and avoids the impact of backend server failures. Click **Modify** to change health check configurations. For more information, see *Configure health check*.

Step 5 Submit the configurations

To confirm the listener configurations, complete these steps:

1. On the **Submit** page, check listener configurations. You can click **Modify** the change the configurations.

2. Click **Submit**.

3. On the **Submit** page, click **OK** after the configurations are successful.

| Configure Server Load Balancer   ↰Back |

| Protocol and Listener | Backend Servers | Health Check | Submit |

| Submit |

Default Server Group	Success
Layer-4 listener	Success
Start Listener	Success

| OK | Cancel |

After the configurations are successful, you can view the created listener on the Listeners page.

| Instance Details | Show ⌄ |

Listeners   Default Server Group   VServer Groups   Active/Standby Server Groups   Monitoring

| Add Listener | ↻ |

	Frontend Protocol/Port	Backend Protocol/Port	Name	Health Status	Monitoring	Forwarding Rule	Session Persistence	Peak Bandwidth	Server Group	Access Control List	Actions
☐	UDP:143	UDP:80	udp_143	● Abnormal	📊	Weighted Round-Robin	Disabled	No Limit	Default Server Group	Disabled	Configure Details More ⌄
☐	TCP:80	TCP:80	tcp_80	● Normal	📊	Weighted Round-Robin	Disabled	No Limit	Default Server Group	Disabled	Configure Details More ⌄
☐	HTTPS:443	HTTP:80	-	● Abnormal	📊	Weighted Round-Robin	Disabled	No Limit	Default Server Group	Disabled	Configure Details Add Forwarding Rules More ⌄

Related operations

- *Configure health check*

- *Manage a default server group*

- *Manage a VServer group*

- *Manage an active/standby server group*

- *Configure access control*

2.3 Add a UDP listener

Applicable to scenarios with preference for real-time content over reliability, such as video chats and pushes of real-time financial quotations. You can add a UDP listener to forward requests from the UDP protocol.

Limits

Note the following before adding a UDP listener:

- The maximum number of connections per listener: 100,000.

- Currently, fragmented packets are not supported.

- UDP listeners of an SLB instance of the classic network do not support viewing the source IP address.

- In the following two scenarios, UDP listener configurations take effect after five minutes:

  - Remove the backend ECS instances.

  - Set the weight of a backend ECS instance to zero after the instance is declared as unhealthy.

- Because IPv6 has a longer IP head than IPv4, when you use a UDP listener on an IPv6 SLB instance, you must ensure that the MTU of the NIC communicating with the SLB on the backend server (ECS instance) is not greater than 1480 (some applications require synchronizing its configuration files based on this MTU value), otherwise the packets may be discarded because they are too large.

  If you use a TCP/HTTP/HTTPS listener, no additional configurations are required because the TCP protocol supports MSS auto-negotiation.

Prerequisites

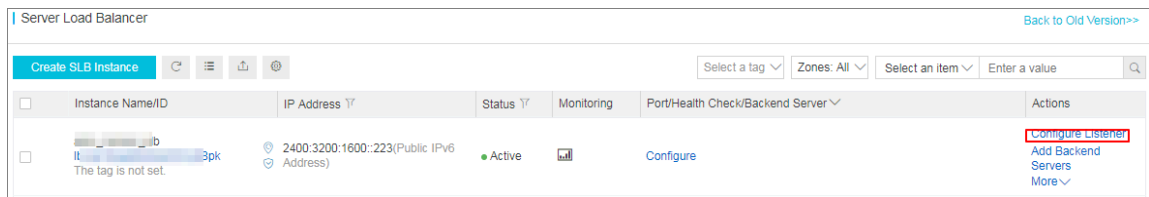*Create an SLB instance*.

Step 1 Open the listener configuration wizard

To open the listener configuration wizard, complete these steps:

1. Log on to the *SLB console*.

2. In the left-side navigation pane, select **Instances** > **Server Load Balancer**.

3. Select the region of the instance.

4. Select one of the following methods to open the listener configuration wizard:

- On the **Server Load Balancer** page, find the target instance and then click **Configure Listener**.

- On the **Server Load Balancer** page, click the ID of the target SLB instance. On the **Listeners** page, click **Add Listener**.
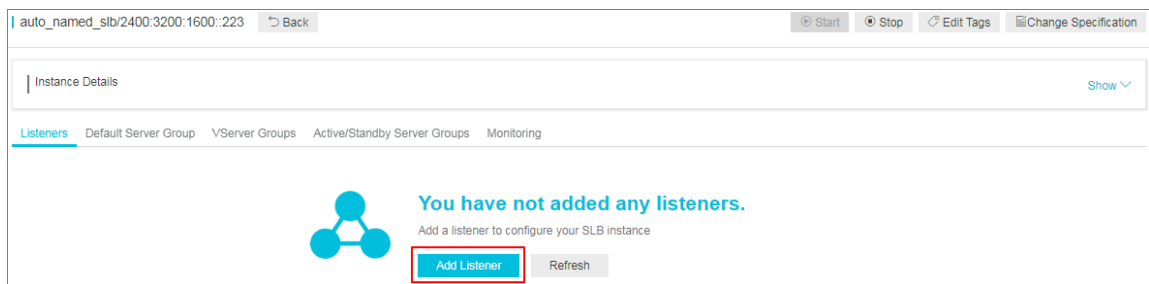
Step 2 Configure a UDP listener

To configure a UDP listener, complete these steps:

1. On the **Protocol and Listener** page, configure the UDP listener according to the following information.
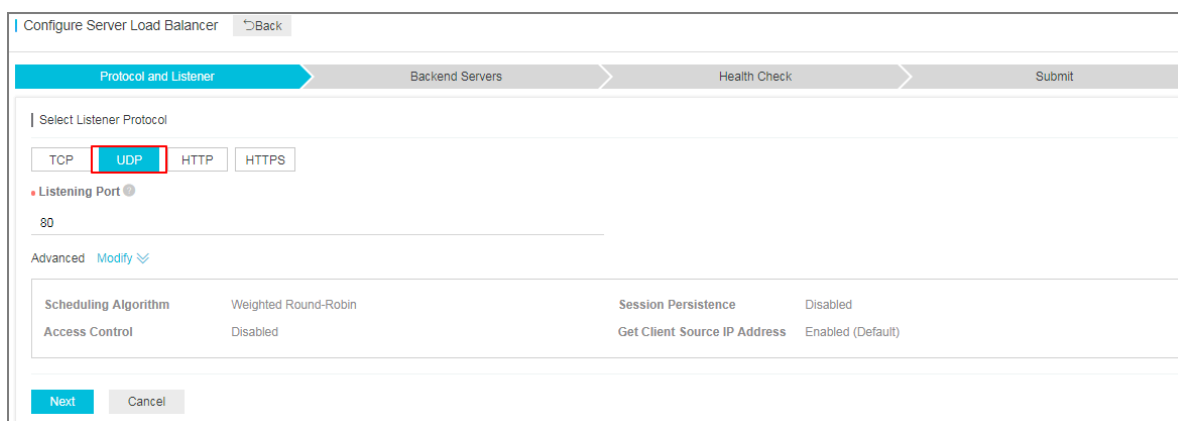
Configuration	Description
Select Listener Protocol	Select the protocol type of the listener. In this tutorial, select **UDP**.
Listening Port	The listening port used to receive requests and forward the requests to backend servers. The port number is in the range of 1-65535. 📋 **Note:** The listening ports must be unique in a Server Load Balancer instance.
Advanced configurations	

Configuration	Description

Scheduling Algorithm	Server Load Balancer supports three scheduling algorithms: round robin, weighted round robin (WRR), and weighted least connections (WLC).
	
	 • **Weighted Round-Robin (WRR)**: Backend servers with higher weights receive more requests than those with smaller weights.
	
	 • **Round-Robin (RR)**: Requests are evenly and sequentially distributed to the backend servers.
	
	 • **Weighted Least Connections (WLC)**: A server with a higher weight will receive a larger percentage of live connections at any one time. When the weight value is the same, a backend server with a smaller number of connections is more frequently (and probably) accessed.
	
	 • **Consistency Hash (CH)**:
	
	   - **Source IP**: The consistent hash based on the source IP address. The same source IP addresses are scheduled to the same backend server.
	
	   - **Tuple**: The consistent hash based on the quaternion (source IP + destination IP + source port + destination port). The same steams are scheduled to the same backend server.
	
	   - **QUIC ID**: Consistent hash based on the QUIC Connection ID. The same QUIC Connection IDs are scheduled to the same backend server.

Notice:

The QuIC protocol is rapidly evolving, and the algorithm is based on *draft-ietf-quic-transport-10* and does not guarantee the compatibility of all QUIC versions. We recommend that you do enough tests before using it for the production environment.

Note:

The consistency Hash (CH) algorithm is only supported in the following areas currently:

- Japan (Tokyo)

- Australia (Sydney)

- Malaysia (Kuala Lumpur)

- Indonesia (Jakarta)

- Germany (Frankfurt)

- US (Silicon Valley)

- US (Virginia)

Configuration	Description
Enable Access Control	Select whether to enable the access control function.
Access Control Method	Select an access control method after enabling the access control function: • **Whitelist**: Only requests from IP addresses or CIDR blocks in the selected access control lists are forwarded. It applies to scenarios where the application only allows access from some specific IP addresses. Enabling whitelist poses some business risks. After a whitelist is configured, only the IP addresses in the list can access the listener. If you enable the whitelist without adding any IP entry in the corresponding access control list, all requests are forwarded. • **Blacklist**: Requests from IP addresses or CIDR blocks in the selected access control lists are not forwarded. It applies to scenarios where the application only denies access from some specific IP addresses. If you enable a blacklist without adding any IP entry in the corresponding access control list, all requests are forwarded.
Access Control List	Select an access control list as the whitelist or the blacklist. 📋 **Note:** An IPv6 instance can only bind IPv6 access control lists and an IPv4 instance can only bind IPv4 access control lists. For more information, see *Configure an access control list*.
Enable Peak Bandwidth Limit	Select whether to configure the listener bandwidth. If the SLB instance is billed by bandwidth, you can set different peak bandwidths for different listeners to limit the traffic passing through the listeners. The sum of the peak bandwidths of all listeners under an instance cannot exceed the bandwidth of that instance. By default, all listeners share the bandwidth of the SLB instance. 📋 **Note:** Instances billed by traffic have no bandwidth peak limit by default.

Configuration	Description
Get Client Source IP Address	The backend server of a UDP listener can directly obtain the real IP of the client. 📋 **Note:** UDP listeners of an SLB instance of the classic network do not support viewing the source IP address.
Automatically Enable Listener After Creation	Choose whether to enable listener after the listener is configured. The listener is enabled by default.

2. Click **Next**.

Step 3 Add backend servers

Add backend servers to process requests. You can use the default server group configured for the instance, or configure a VServer group or an active/standby server group for the listener. For more information, see *Backend server overview*.

In this tutorial, the default server group is used:

1. Select **Default Server Group** and then click **Add**.

2. Select the ECS instances to add and then click **Add to Selected Server List**. Click **OK**.

3. Configure the ports and weights of the added backend servers.

  • Port

    The port opened on the backend server (ECS instance) to receive requests. The port number is in the range of 1-65535. Ports of backend servers can be the same in an SLB instance.

  • Weight

    The weight of the backend server (ECS instance). An ECS instance with a higher weight will receive a larger number of connection requests.

    📋 **Note:**
    If the weight is set to 0, no requests will be sent to the ECS instance.

4. Click **Next**.

Step 4 Configure health check

Server Load Balancer checks the service availability of the backend servers (ECS instances) by performing health checks. The health check function improves the overall availability of your services and avoids the impact of backend server failures. Click **Modify** to change health check configurations. For more information, see *Configure health check*.

Step 5 Submit the configurations

To confirm the listener configurations, complete these steps:

1. On the **Submit** page, check listener configurations. You can click **Modify** the change the configurations.

2. Click **Submit**.

3. On the **Submit** page, click **OK** after the configurations are successful.

After the configurations are successful, you can view the created listener on the Listeners page.

Related operations

- *Configure health check*

- *Manage a default server group*

- *Manage a VServer group*

- *Manage an active/standby server group*

- *Configure access control*

2.4 Add an HTTP listener

It is applicable to applications that need to recognize data contents, such as web applications and small-sized mobile games. You can add an HTTP listener to forward requests from the HTTP protocol.

Prerequisites

Create an SLB instance.

Step 1 Open the listener configuration wizard

To open the listener configuration wizard, complete these steps:

1. Log on to the *SLB console*.

2. In the left-side navigation pane, select **Instances** > **Server Load Balancer**.

3. Select the region of the instance.

4. Select one of the following methods to open the listener configuration wizard:

    - On the **Server Load Balancer** page, find the target instance and then click **Configure Listener**.

    - On the **Server Load Balancer** page, click the ID of the target SLB instance. On the **Listeners** page, click **Add Listener**.
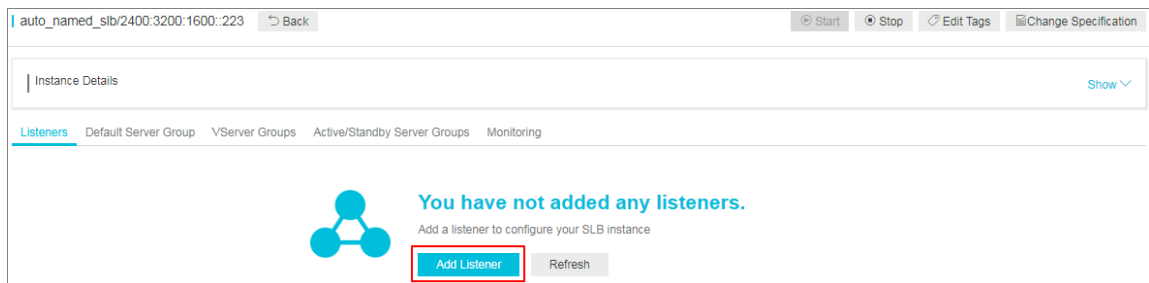
Step 2 Configure an HTTP listener

To configure an HTTP listener, complete these steps:

1. On the **Protocol and Listener** page, configure the HTTP listener according to the following information.

Configuration	Description
Select Listener Protocol	Select the protocol type of the listener. In this tutorial, select **HTTP**.
Listening Port	The listening port used to receive requests and forward the requests to backend servers. The port number is in the range of 1-65535. 📋 **Note:** The listening ports must be unique in a Server Load Balancer instance.
Advanced configurations	

Configuration	Description
Scheduling Algorithm	Server Load Balancer supports three scheduling algorithms: round robin, weighted round robin (WRR), and weighted least connections (WLC). • **Weighted Round-Robin (WRR)**: Backend servers with higher weights receive more requests than those with smaller weights. • **Round-Robin (RR)**: Requests are evenly and sequentially distributed to the backend servers. • **Weighted Least Connections (WLC)**: A server with a higher weight will receive a larger percentage of live connections at any one time. When the weight value is the same, a backend server with a smaller number of connections is more frequently (and probably) accessed.
Redirection	Select whether to forward traffic of the HTTP listener to an HTTPS listener. **Note:** If you enables listener forwarding, make sure that you have created an HTTPS listener.

Configuration	Description

Session Persistence	Select whether to enable session persistence.
	If session persistence is enabled, all session requests from the same client are sent to the same backend server.
	HTTP session persistence is based on cookies. The following two methods are supported:
	
	• **Insert cookie**: You only need to specify the cookie timeout period.
	
	   SLB adds a cookie to the first response from the backend server (insert SERVERID in the HTTP/HTTPS response packet). The next request will contain the cookie and the listener will distribute the request to the same backend server.
	
	• **Rewrite cookie**: You can set the cookie to insert to the HTTP/HTTPS response according to your needs. You must maintain the timeout period and lifecycle of the cookie on the backend server.
	
	   SLB will overwrite the original cookie when it discovers that a new cookie is set. The next time the client carries the new cookie to access SLB, the listener will distribute the request to the recorded backend server. For more information, see *Session persistence*.

Enable Access Control	Select whether to enable the access control function.

Configuration	Description

Access Control Method	Select an access control method after enabling the access control function:
	
	• **Whitelist**: Only requests from IP addresses or CIDR blocks in the selected access control lists are forwarded. It applies to scenarios where the application only allows access from some specific IP addresses.
	
	  Enabling whitelist poses some business risks. After a whitelist is configured, only the IP addresses in the list can access the listener. If you enable the whitelist without adding any IP entry in the corresponding access control list, all requests are forwarded.
	
	• **Blacklist**: Requests from IP addresses or CIDR blocks in the selected access control lists are not forwarded. It applies to scenarios where the application only denies access from some specific IP addresses.
	
	  If you enable a blacklist without adding any IP entry in the corresponding access control list, all requests are forwarded.

Access Control List	Select an access control list as the whitelist or the blacklist. **Note:** An IPv6 instance can only bind IPv6 access control lists and an IPv4 instance can only bind IPv4 access control lists. For more information, see *Configure an access control list*.
Enable Peak Bandwidth Limit	Select whether to configure the listener bandwidth. If the SLB instance is billed by bandwidth, you can set different peak bandwidths for different listeners to limit the traffic passing through the listeners. The sum of the peak bandwidths of all listeners under an instance cannot exceed the bandwidth of that instance. By default, all listeners share the bandwidth of the SLB instance. **Note:** Instances billed by traffic have no peak bandwidth limit by default.

Configuration	Description
Idle Timeout	Specify the idle connection timeout in seconds. Valid value: 1-60 If no request is received during the specified timeout period, Server Load Balancer will close the connection and restart the connection when the next request comes. This function is available in all regions.
Request Timeout	Specify the request timeout in seconds. Valid value: 1-180 If no response is received from the backend server during the specified timeout period, Server Load Balancer will stop waiting and send an HTTP 504 error code to the client. Currently, this function is available in all regions.
Enable Gzip Compression	Choose whether to enable Gzip compression to compress files of specific formats. Now Gzip supports the following file types: text/xml, text/plain, text/css, application/javascript, application/x-javascript application/rss+xml, application/atom+xml and application/xml.
Add HTTP Header Fields	Select the custom HTTP headers that you want to add: • Use the `X-Forwarded-For` field to retrieve the client source IP address. • Use the `X-Forwarded-Proto` field to retrieve the listener protocol used by the SLB instance. • Use the `SLB-IP` field to retrieve the public IP address of the SLB instance. • Use the `SLB-ID` field to retrieve the ID of the SLB instance.
Get Client Source IP Address	HTTP listener uses X-Forwarded-For to obtain the real IP of the client.
Automatically Enable Listener After Creation	Choose whether to enable listener after the listener is configured. The listener is enabled by default.

2. Click **Next**.

Step 3 Add backend servers

Add backend servers to process requests. You can use the default server group configured for the instance, or configure a VServer group or an active/standby server group for the listener. For more information, see *Backend server overview*.

In this tutorial, the default server group is used:

1. Select **Default Server Group** and then click **Add**.

2. Select the ECS instances to add and then click **Add to Selected Server List**. Click **OK**.

3. Configure the ports and weights of the added backend servers.

- Port

  The port opened on the backend server (ECS instance) to receive requests. The port
  number is in the range of 1-65535. Ports of backend servers can be the same in an SLB
  instance.

- Weight

  The weight of the backend server (ECS instance). An ECS instance with a higher weight will
  receive a larger number of connection requests.

  > **Note:**
  > If the weight is set to 0, no requests will be sent to the ECS instance.

4. Click **Next**.

Step 4 Configure health check

Server Load Balancer checks the service availability of the backend servers (ECS instances) by performing health checks. The health check function improves the overall availability of your services and avoids the impact of backend server failures. Click **Modify** to change health check configurations. For more information, see *Configure health check*.

Step 5 Submit the configurations

To confirm the listener configurations, complete these steps:

1. On the **Submit** page, check listener configurations. You can click **Modify** the change the configurations.

2. Click **Submit**.

3. On the **Submit** page, click **OK** after the configurations are successful.

After the configurations are successful, you can view the created listener on the Listeners page.

Related operations

- *Configure health check*

- *Manage a default server group*

- *Manage a VServer group*

- *Manage an active/standby server group*

- *Configure access control*

- *Add domain-name based or URL-based forwarding rules*

- *Manage a domain name extension*

2.5 Add an HTTPS listener

It is applicable to applications requiring encrypted transmission. You can add an HTTPS listener to forward requests from the HTTPS protocol.

Prerequisites

Create an SLB instance.

Step 1 Open the listener configuration wizard

To open the listener configuration wizard, complete these steps:

1. Log on to the *SLB console*.

2. In the left-side navigation pane, select **Instances** > **Server Load Balancer**.

3. Select the region of the instance.

4. Select one of the following methods to open the listener configuration wizard:

- On the **Server Load Balancer** page, find the target instance and then click **Configure Listener**.

- On the **Server Load Balancer** page, click the ID of the target SLB instance. On the **Listeners** page, click **Add Listener**.
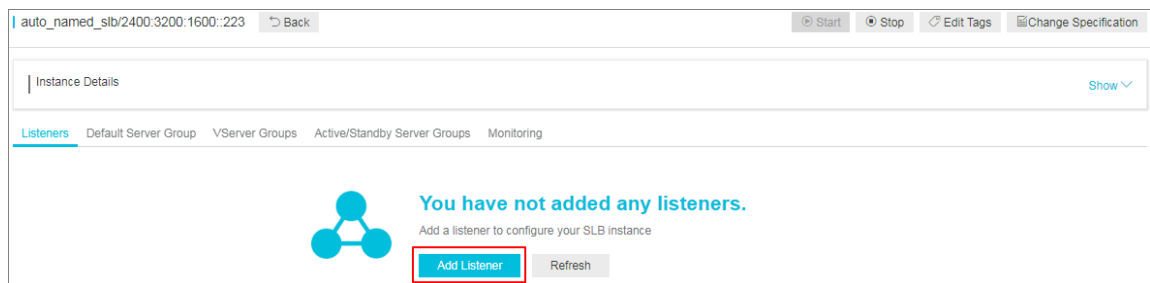
Step 2 Configure an HTTPS listener

To configure an HTTPS listener, complete these steps:

1. On the **Protocol and Listener** page, configure the HTTPS listener according to the following information.

Configuration	Description
Select Listener Protocol	Select the protocol type of the listener. In this tutorial, select **HTTPS**.
Listening Port	The listening port used to receive requests and forward the requests to backend servers. The port number is in the range of 1-65535. 📋 **Note:** The listening ports must be unique in a Server Load Balancer instance.
Advanced configurations	

Configuration	Description
Scheduling Algorithm	Server Load Balancer supports three scheduling algorithms: round robin, weighted round robin (WRR), and weighted least connections (WLC). • **Weighted Round-Robin (WRR)**: Backend servers with higher weights receive more requests than those with smaller weights. • **Round-Robin (RR)**: Requests are evenly and sequentially distributed to the backend servers. • **Weighted Least Connections (WLC)**: A server with a higher weight will receive a larger percentage of live connections at any one time. When the weight value is the same, a backend server with a smaller number of connections is more frequently (and probably) accessed.
Enable Session Persistence	Select whether to enable session persistence. If session persistence is enabled, all session requests from the same client are sent to the same backend server. HTTP session persistence is based on cookies. The following two methods are supported: • **Insert cookie**: You only need to specify the cookie timeout period. SLB adds a cookie to the first response from the backend server (insert SERVERID in the HTTP/HTTPS response packet). The next request will contain the cookie and the listener will distribute the request to the same backend server. • **Rewrite cookie**: You can set the cookie to insert to the HTTP/HTTPS response according to your needs. You must maintain the timeout period and lifecycle of the cookie on the backend server. SLB will overwrite the original cookie when it discovers that a new cookie is set. The next time the client carries the new cookie to access SLB, the listener will distribute the request to the recorded backend server. For more information, see *Session persistence*.
Enable HTTP2.0	Select whether to enable HTTP 2.0.
Enable Access Control	Select whether to enable the access control function.

Configuration	Description
Access Control Method	Select an access control method after enabling the access control function: • **Whitelist**: Only requests from IP addresses or CIDR blocks in the selected access control lists are forwarded. It applies to scenarios where the application only allows access from some specific IP addresses. Enabling whitelist poses some business risks. After a whitelist is configured, only the IP addresses in the list can access the listener. If you enable the whitelist without adding any IP entry in the corresponding access control list, all requests are forwarded. • **Blacklist**: Requests from IP addresses or CIDR blocks in the selected access control lists are not forwarded. It applies to scenarios where the application only denies access from some specific IP addresses. If you enable a blacklist without adding any IP entry in the corresponding access control list, all requests are forwarded.
Access Control List	Select an access control list as the whitelist or the blacklist. 📋 **Note:** An IPv6 instance can only bind IPv6 access control lists and an IPv4 instance can only bind IPv4 access control lists. For more information, see *Configure an access control list*.
Enable Peak Bandwidth Limit	Select whether to configure the listening bandwidth. If the SLB instance is billed by bandwidth, you can set different peak bandwidths for different listeners to limit the traffic passing through the listeners. The sum of the peak bandwidths of all listeners under an instance cannot exceed the bandwidth of that instance. By default, all listeners share the bandwidth of the SLB instance. 📋 **Note:** Instances billed by traffic have no peak bandwidth limit by default.

Configuration	Description
Idle Timeout	Specify the idle connection timeout in seconds. Valid value: 1-60 If no request is received during the specified timeout period, Server Load Balancer will close the connection and restart the connection when the next request comes. This function is available in all regions.
Request Timeout	Specify the request timeout in seconds. Valid value: 1-180 If no response is received from the backend server during the specified timeout period, Server Load Balancer will stop waiting and send an HTTP 504 error code to the client. Currently, this function is available in all regions.
TLS Security Policy	Only guaranteed-performance instances support selecting the TLS security policy to use. The TLS security policy contains available TLS protocol versions and supporting encryption algorithm suites. For more information, see *Manage TLS security policies*.
Enable Gzip Compression	Choose whether to enable Gzip compression to compress files of specific formats. Now Gzip supports the following file types: text/xml, text/plain, text/css, application/javascript, application/x-javascript application/rss+xml, application/atom+xml and application/xml .
Add HTTP Header Fields	Select the custom HTTP headers that you want to add: • Use the `X-Forwarded-For` field to retrieve the client source IP address. • Use the `X-Forwarded-Proto` field to retrieve the listener protocol used by the SLB instance. • Use the `SLB-IP` field to retrieve the public IP address of the SLB instance. • Use the `SLB-ID` field to retrieve the ID of the SLB instance.
Get Client Source IP Address	HTTP listener uses X-Forwarded-For to obtain the real IP of the client.
Automatically Enable Listener After Creation	Choose whether to enable listener after the listener is configured. The listener is enabled by default.

2. Click **Next**.

Step 3 Configure the SSL certificate

To add an HTTPS listener, you must upload a server certificate or CA certificate, as shown in the following table.

Certificate	Description	Required for one-way authentication	Required for mutual authentication
Server certificate	Used to identify a server. The client uses it to check whether the certificate sent by the server is issued by a trusted center.	Yes The server certificate must be uploaded to the certificate management system of the Server Load Balancer.	Yes Upload the server certificate to SLB.
Client certificate	Used to identify a client. The client user can prove its true identity when communicating with the server. You can sign a client certificate with a self-signed CA certificate.	No	Yes Install the client certificate on the client.
CA certificate	The server uses the CA certificate to authenticate the signature on the client certificate, as part of the authorization before launching a secure connection. If the authentication fails, the connection will be rejected.	No	Yes Upload the server certificate to SLB.

Note the following before uploading certificates:

- The uploaded certificate must be in the PEM format. For more information, see *Certificate requirements*.

- After the certificate is uploaded to SLB, SLB can manage the certificate and you do not need to bind the certificate on backend ECS instances.

- It usually takes one to three minutes to activate the HTTPS listener because the uploading, loading, and validation of certificates take some time. Normally it takes effect in one minute and it will definitely take effect in three minutes.

- The ECDHE algorithm cluster used by HTTPS listeners supports forward secrecy, but does not support uploading security enhancement parameter files required by the DHE algorithm cluster, such as strings containing the `BEGIN DH PARAMETERS` field in the PEM certificate file. For more information, see *Certificate formats*.

- Currently, Server Load Balancer HTTPS listeners do not support SNI (Server Name Indication). You can use TCP listeners instead, and then configure SNI on the backend ECS instances.

- The session ticket retention time of HTTPS listeners is 300 seconds.

- The actual amount of traffic is larger than the billed traffic amount because some traffic is used for the protocol handshaking.

- In the case of a large number of new connections, HTTPS listeners consume more traffic.

To configure the SSL certificate, complete these steps:

1. Select the server certificate that has been uploaded, or click **Create Server Certificate** to upload a server certificate.

    For more information, see *Create a certificate*.

2. If you want to enable HTTPS mutual authentication, click **Modify** and enable mutual authentication.

3. Select an uploaded CA certificate, or click **Create CA Certificate** to upload a CA certificate.

    You can use a self-signed CA certificate. For more information, see *Generate a CA certificate*.

Step 4 Add backend servers

Add backend servers to process requests. You can use the default server group configured for the instance, or configure a VServer group or an active/standby server group for the listener. For more information, see *Backend server overview*.

In this tutorial, the default server group is used:

1. Select **Default Server Group** and then click **Add**.

2. Select the ECS instances to add and then click **Add to Selected Server List**. Click **OK**.

3. Configure the ports and weights of the added backend servers.

- Port

The port opened on the backend server (ECS instance) to receive requests. The port
number is in the range of 1-65535. Ports of backend servers can be the same in an SLB
instance.

- Weight

  The weight of the backend server (ECS instance). An ECS instance with a higher weight will receive a larger number of connection requests.

  > **Note:**
  >
  > If the weight is set to 0, no requests will be sent to the ECS instance.

4. Click **Next**.

Step 5 Configure health check

Server Load Balancer checks the service availability of the backend servers (ECS instances) by performing health checks. The health check function improves the overall availability of your services and avoids the impact of backend server failures. Click **Modify** to change health check configurations. For more information, see *Configure health check*.

Step 6 Submit the configurations

To confirm the listener configurations, complete these steps:

1. On the **Submit** page, check listener configurations. You can click **Modify** to change the configurations. Click **Submit**.

2. On the **Submit** page, click **OK** after the configurations are successful.

After the configurations are successful, you can view the created listener on the listeners page.

Related operations

- *Configure health check*

- *Manage a default server group*

- *Manage a VServer group*

- *Manage an active/standby server group*

- *Generate a CA certificate*

- *Create a certificate*

- *Configure access control*

- *Add domain-name based or URL-based forwarding rules*

- *Manage a domain name extension*

2.6 Support TLS security policies

Guaranteed-performance instances support selecting the TLS security policy to use when you
create or configure an HTTPS listener.

You can select the TLS security policy when you set advanced configurations of **Protocol and
Listener** during adding or configuring an HTTPS listener. For more information, see *Add an
HTTPS listener*.

The TLS security policy contains available TLS protocol versions and supporting encryption
algorithm suites.

TLS security policy

Security policy	Features	Supported TLS versions	Supported encryption algorithm suites
tls_cipher _policy_1_0	Best compatibil ity and low security	TLSv1.0, TLSv1.1, and TLSv1.2	Supported encryption algorithm suites: ECDHE -RSA-AES128-GCM-SHA256, ECDHE-RSA-AES256-GCM-SHA384, ECDHE-RSA-AES128 -SHA256, ECDHE-RSA-AES256-SHA384 , AES128-GCM-SHA256, AES256-GCM-SHA384, AES128-SHA256, AES256-SHA256 , ECDHE-RSA-AES128-SHA, ECDHE-RSA-AES256-SHA, AES128-SHA, AES256-SHA and DES-CBC3-SHA.
tls_cipher _policy_1_1	Good compatibility and security	TLSv1.1 and TLSv1.2	ECDHE-RSA-AES128-GCM-SHA256, ECDHE -RSA-AES256-GCM-SHA384, ECDHE-RSA -AES128-SHA256, ECDHE-RSA-AES256-SHA384, AES128-GCM-SHA256, AES256-GCM-SHA384, AES128-SHA256, AES256-SHA256, ECDHE-RSA-AES128-SHA, ECDHE -RSA-AES256-SHA, AES128-SHA, AES256-SHA and DES-CBC3-SHA.
tls_cipher _policy_1_2	Good compatibil ity and high security	Tlsv1.2	ECDHE-RSA-AES128-GCM-SHA256, ECDHE -RSA-AES256-GCM-SHA384, ECDHE-RSA -AES128-SHA256, ECDHE-RSA-AES256-SHA384, AES128-GCM-SHA256, AES256-GCM-SHA384, AES128-SHA256, AES256-SHA256, ECDHE-RSA-AES128-SHA, ECDHE -RSA-AES256-SHA, AES128-SHA, AES256-SHA and DES-CBC3-SHA.
tls_cipher _policy_1_ 2_strict	Only support encryption suites of forward security and have extremely high security	TLSv1.2	ECDHE-RSA-AES128-GCM-SHA256, ECDHE -RSA-AES256-GCM-SHA384, ECDHE-RSA -AES128-SHA256, ECDHE-RSA-AES256-SHA384, ECDHE-RSA-AES128-SHA and ECDHE-RSA-AES256-SHA.

Differences among TLS security policies

Security policy		tls_cipher _policy_1_ 0	tls_cipher _policy_1_ 1	tls_cipher _policy_1_ 2	tls_cipher _policy_1_ 2_strict
TLS		1.2/1.1/1.0	1.2/1.1	1.2	1.2
CIPHER	ECDHE-RSA-AES128-GCM-SHA256	✓	✓	✓	✓
	ECDHE-RSA-AES256-GCM-SHA384	✓	✓	✓	✓
	ECDHE-RSA-AES128-SHA256	✓	✓	✓	✓
	ECDHE-RSA-AES256-SHA384	✓	✓	✓	✓
	AES128-GCM-SHA256	✓	✓	✓	
	AES256-GCM-SHA384	✓	✓	✓	
	AES128-SHA256	✓	✓	✓	
	AES256-SHA256	✓	✓	✓	
	ECDHE-RSA-AES128-SHA	✓	✓	✓	✓
	ECDHE-RSA-AES256-SHA	✓	✓	✓	✓
	AES128-SHA	✓	✓	✓	
	AES256-SHA	✓	✓	✓	
	DES-CBC3-SHA	✓	✓	✓	

2.7 Manage a domain name extension

HTTPS listeners of guaranteed-performance SLB support configuring multiple certificates, allowing you to forward requests with different domain names to different backend servers.

Introduction to SNI

Server Name Indication (SNI) is an extension to the SSL/TLS protocol, allowing a server to install multiple SSL certificates on the same IP address. When a client accesses SLB, the certificate configured for the domain name is used by default. If no certificate is configured for the domain name, the certificate configured for the HTTPS listener is used.

> **Note:**
>
> Only guaranteed-performance SLB instances support SNI.

If you want to resolve multiple domain names to the IP address of an SLB instance, and distribute requests with different domain names to different backend servers, use the domain name extension function.

The domain name extension function is available in all regions.

Add a domain name extension

1. Log on to the *SLB console*.

2. Select a region and all SLB instance in this region are displayed.

3. Click the ID of the Server Load Balancer instance.

4. In the left-side navigation pane, click **Listeners**.

5. On the **Listeners** page, find the created HTTPS listener, and then click **More** > **Additional Domains**.

6. Click **Add Additional Domain** and configure the domain name:

    a. Enter a domain name. The domain name can only contain letters, numbers, dashes, or dots.

    Domain name based forwarding rules support exact match and wildcard match.

    • Exact domain name: www.aliyun.com

    • Wildcard domain name (generic domain name): *.aliyun.com, *.market.aliyun.com

    When a request matches multiple forwarding rules, exact match takes precedence over small-scale wildcard match and small-scale wildcard match takes precedence over large-scale wildcard match, as shown in the following table.

Type	Request URL	Domain name based forwarding rule		
		www. aliyun. com	***.aliyun. com**	***.market .aliyun. com**
Exact match	www.aliyun.com	✓	×	×
Wildcard match	market.aliyun.com	×	✓	×
Wildcard match	info.market.aliyun.com	×	×	✓

   b. Select the certificate associated with the domain name.

> 📋 **Note:**
>
> The domain name in the certificate must be the same as the added domain name
>
> extension.

   c. Click **OK**.

7. On the **Listeners** page, find the created HTTPS listener and click **Add Forwarding Rules**.

8. On the **Forwarding Rules** page, click **Add Forwarding Rules**.

9. For more information, see *Traffic forwarding based on domain name or URL*.

> 📋 **Note:**
>
> Make sure that the domain name configured in the forwarding rule is the same as the added
>
> domain name extension.

Edit a domain name extension

You can replace the certificate used by an added domain name extension.

To edit a domain name extension, complete these steps:

1. Log on to the *SLB console*.

2. Select a region and all SLB instance in this region are displayed.

3. Click the ID of the SLB instance.

4. In the left-side navigation pane, click **Listeners**.

5. On the **Listeners** page, find the created HTTPS listener, and then click **More** > **Additional Domains**.

6. Find the target domain name extension and then click **Edit**.

7. In the **Edit Additional Domain** dialog box, select a new certificate and then click **OK**.
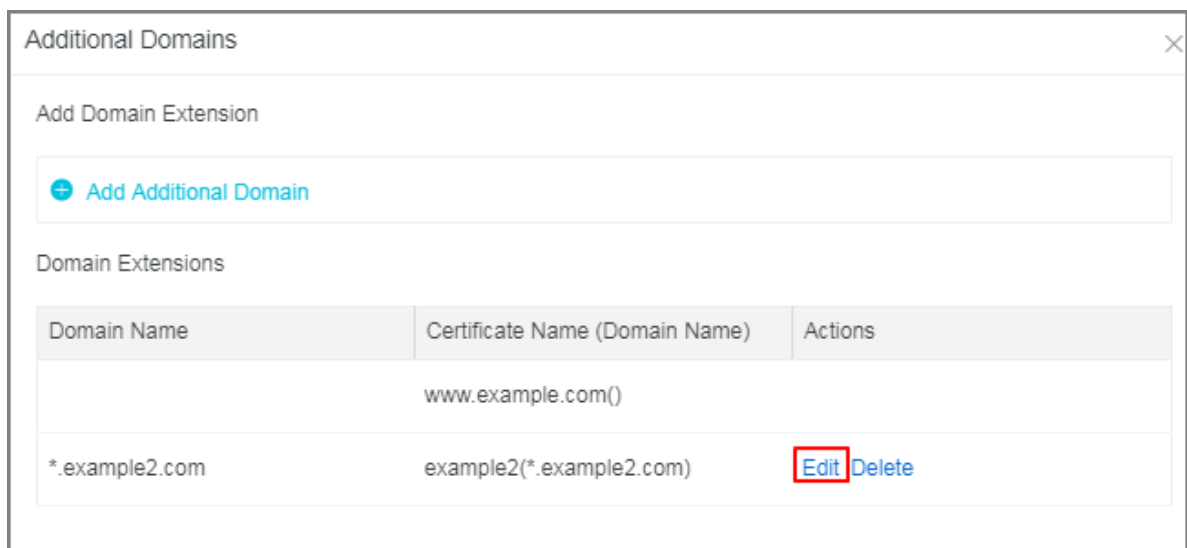
Delete a domain name extension

To delete a domain name extension, complete these steps:

1. Log on to the *SLB console*.

2. Select a region and all SLB instance in this region are displayed.

3. Click the ID of the SLB instance.

4. In the left-side navigation pane, click **Listeners**.

5. On the **Listeners** page, find the created HTTPS listener, and then click **More** > **Additional Domains**.

6. Find the target domain name extension and then click **Delete**.

7. In the displayed dialog box, click **OK**.

3 Health check

3.1 Health check overview

Server Load Balancer checks the service availability of the backend servers (ECS instances) by performing health checks. Health check improves the overall availability of the front-end service, and avoids impact on the entire service caused by exceptions of the backend ECS instances.

After enabling the health check function, SLB stops distributing requests to the instance that is discovered as unhealthy and restarts forwarding requests to the instance only when it is declared healthy.

If your business is highly sensitive to traffic load, frequent health checks may impact normal service. You can reduce this impact by reducing the frequency of health checks, increasing the health check interval, or changing the HTTP health check to TCP health check. To guarantee the service availability, we do not recommend disabling all health checks.

Health check process

Server Load Balancer is deployed in clusters. Data forwarding and health checks are handled at the same time by the node servers in the LVS cluster and Tengine cluster.

The node servers in the cluster independently perform health checks in parallel, according to the health check configuration. If a node server discovers an ECS instance is unhealthy, the node server will stop distributing requests to the ECS instance. This operation is synchronized through all node servers.

The IP address range used to perform the health check is 100.64.0.0/10. The backend servers cannot block this CIDR block. You do not need to additionally configure a security group rule to allow access from this CIDR block. However, if you have configured security rules such as iptables, allow access from this CIDR block (100.64.0.0/10 is reserved by Alibaba Cloud, and other users cannot use any IP address in this CIDR block, so there is no security risk).
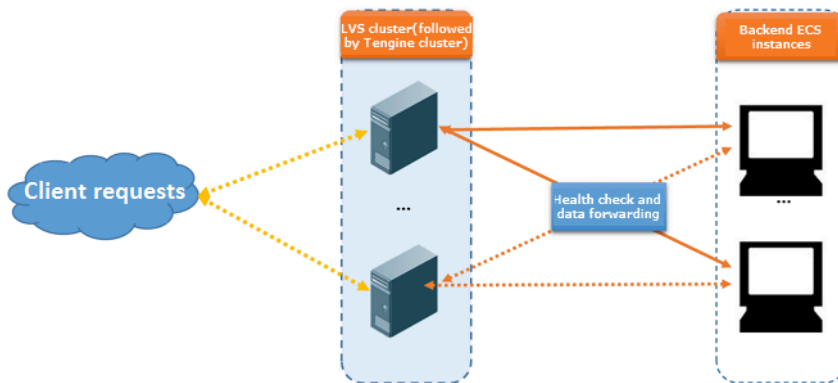
Health check of HTTP/HTTPS listeners

For Layer-7 (HTTP or HTTPS) listeners, SLB detects the status of backend servers by sending HTTP HEAD requests, as shown in the following figure.

For HTTPS listeners, certificates are managed in SLB. Data exchange (including health check data and service interaction data) between SLB and backend ECS instances is not transmitted over HTTPS to improve the system performance.
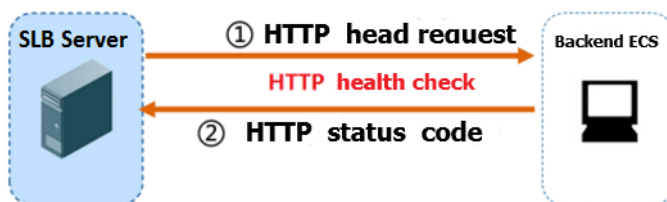
The health check process of a Layer-7 listener is as follows:

1. The Tengine node server sends an HTTP HEAD request to the intranet IP + 【Health Check Port】 + 【Health Check Path】 of the ECS instance according to the health check settings.

2. After receiving the request, the backend server returns an HTTP status code based on the running status.

3. If the Tengine node server does not receive the response from the backend server within the 【Response Timeout】 period, the ECS instance is declared unhealthy.

4. If the Tengine node server receives the response from the backend ECS instance within the 【Response Timeout】 period, it compares the returned status code with the status code specified in the listener configuration. If the status code is the same, the backend server is declared healthy. Otherwise, the backend server is declared unhealthy.

Health check of TCP listeners

For TCP listeners, SLB detects the status of backend servers by sending TCP detections, as the following figure shows.
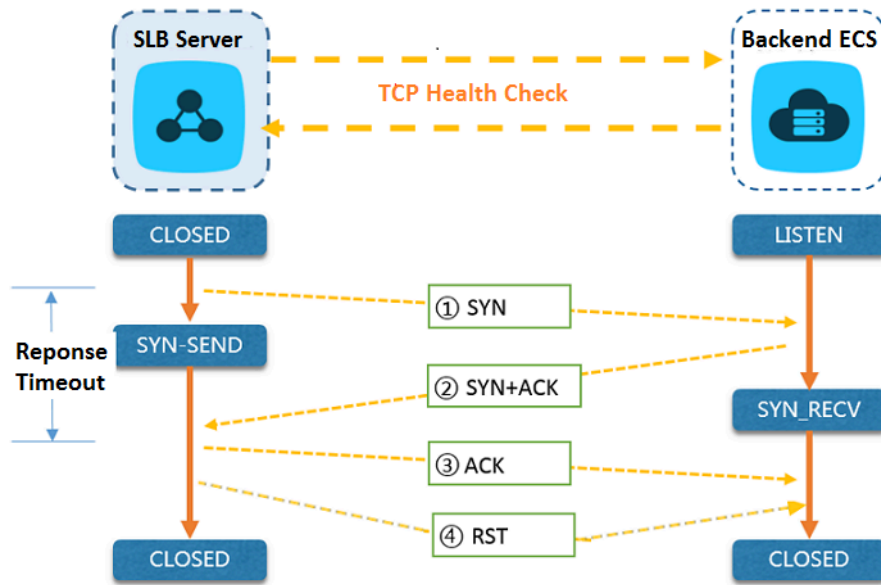
The health check process of a TCP listener is as follows:

1. The LVS node server sends a TCP SYN packet to the intranet IP + 【Health Check Port】of the backend ECS instance.

2. After receiving the request, the backend server returns a TCP SYN and ACK packet if the corresponding port is listening normally.

3. If the LVS node server does not receive the required data packet from the backend server within the【Response Timeout】period, the ECS instance is declared unhealthy. Then, the LVS node server sends an RST data packet to the backend server to terminate the TCP connection.

4. If the LVS node server receives the data packet from the backend ECS instance within the【Response Timeout】period, the ECS instance is declared healthy. Then, the LVS node server sends an RST data packet to the backend server to terminate the TCP connection.

> 📋 **Note:**
>
> In general, TCP three-way handshakes are conducted to establish a TCP connection. After the LVS node server receives an SYN + ACK data packet from the backend ECS instance, the LVS node server sends an ACK data packet, and then immediately sends an RST data packet to terminate the TCP connection.
>
> This process may make backend server think an error (such as an abnormal exit) occurred in the TCP connection and then throw a corresponding error message, such as `Connection reset by peer`.
>
> Resolution:

- Use the HTTP health check.

- If obtaining real IP is enabled, ignore the connection errors caused by access of the SLB IP address.

Health check of UDP listeners

For UDP listeners, Server Load Balancer detects the status of the backend servers through UDP packet detection, as shown in the following figure.
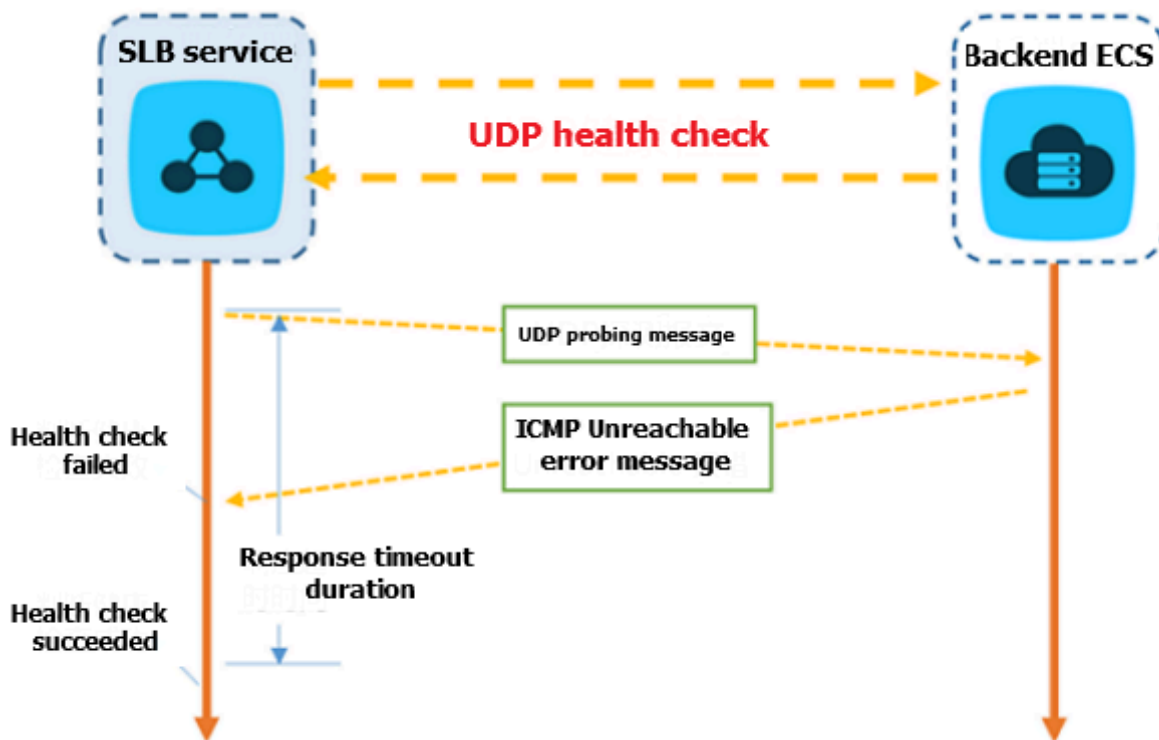
The health check process of a UDP listener is as follows:

1. The LVS node server sends a UDP packet to the intranet IP + 【Health Check Port】 of the ECS instance according to health check configurations.

2. If the corresponding port of the ECS instance is not listening normally, the system will return an ICMP error message, such as `port XX unreachable`. Otherwise, no message is sent.

3. If the LVS node server receives the ICMP error message within the 【Response Timeout】 period, the ECS instance is declared unhealthy.

4. If the LVS node server does not receive any messages within the 【Response Timeout】 period , the ECS instance is declared healthy.

> 📋 **Note:**

For UDP health checks, the real status of the backend server and the health check result may not be the same in the following situation:

If the ECS instance uses a Linux operating system, the speed of sending ICMP messages in high-concurrency scenarios is limited due to the anti-ICMP attack protection in Linux. In this case, even if an exception occurs in the ECS instance, SLB may declare the backend server healthy because the error message `port XX unreachable` is not returned. As a result, the actual service status is different from the health check result.

Resolution:

Set a pair of custom request and response for the UDP health check. If the custom response is returned, the ECS instance is considered healthy. Otherwise, the ECS instance is considered unhealthy. To achieve this, you must add corresponding configurations for the client.

Health check time window

The health check function has effectively improved the availability of your business services . However, to reduce impact on the system availability caused by frequent system switches because of health check failure, SLB declares an ECS instance healthy or unhealthy only after successive successes or failures within a specified timeframe. The health check time window is determined by the following three factors:

- Health check interval (How often the health check is performed.)
- Response timeout (The amount of time to wait for the response.)
- Health check threshold (The number of consecutive successful or failed health checks.)

The health check time window is calculated as follows:

- Health check failure time window = Response Timeout x Unhealthy Threshold + Health Check Interval X (Unhealthy Threshold) -1)
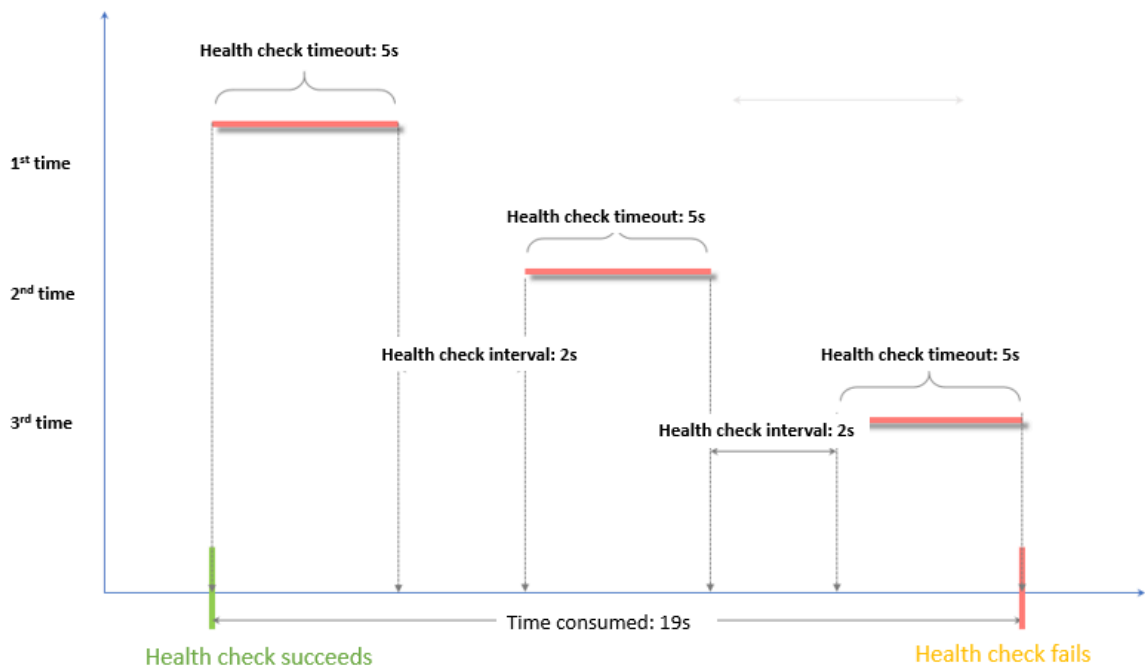
- Health check success time window = (Response Time of a Successful Health Check X Healthy Threshold) + Health Check Interval X (Healthy Threshold-1)

> **Note:**
>
> The success response time of a health check is the duration from the time when the health check request is sent to the time when the response is received. For TCP health check, the time is very short and almost negligible because TCP health check only detects whether the port is alive. For HTTP health check, the time depends on the performance and load of the application server and is generally within seconds.
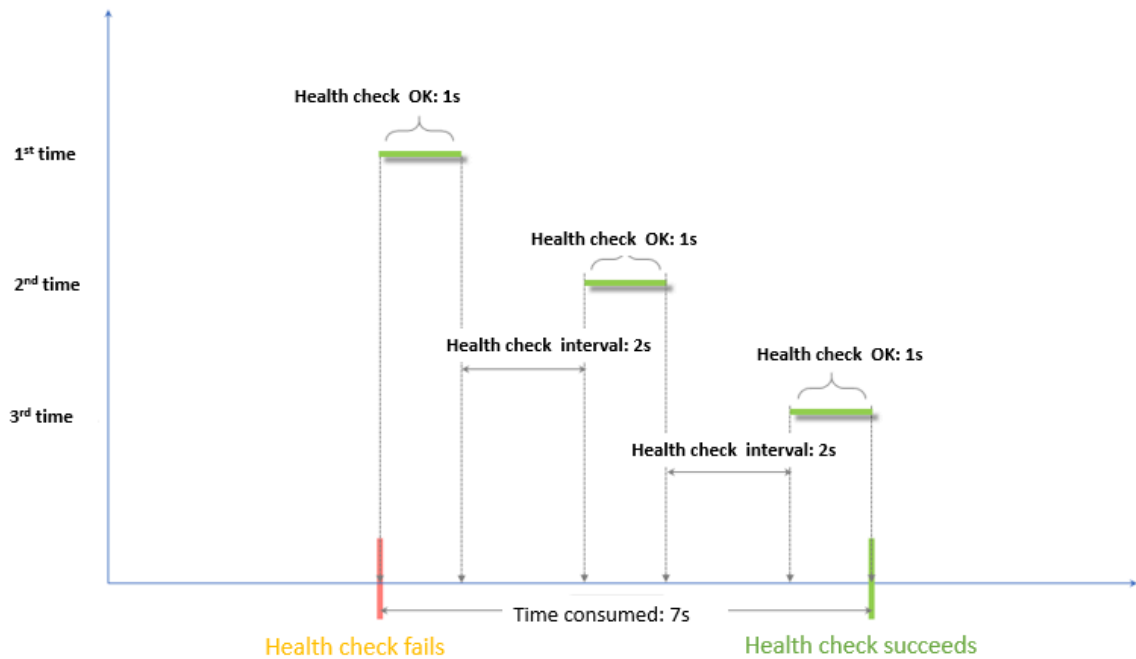
The health check result has the following impact on the requests forwarding:

- If the health check of the target ECS instance fails, new requests will not be distributed to the ECS instance. Therefore, there is no impact on the client access.

- If the health check of the target ECS instance succeeds, new requests will be distributed to it. The client access is normal.

- If a request arrives during a health check failure window, the request is still sent to the ECS instance because the ECS instance is being checked and has not been declared unhealthy. As a result, the client access fails.
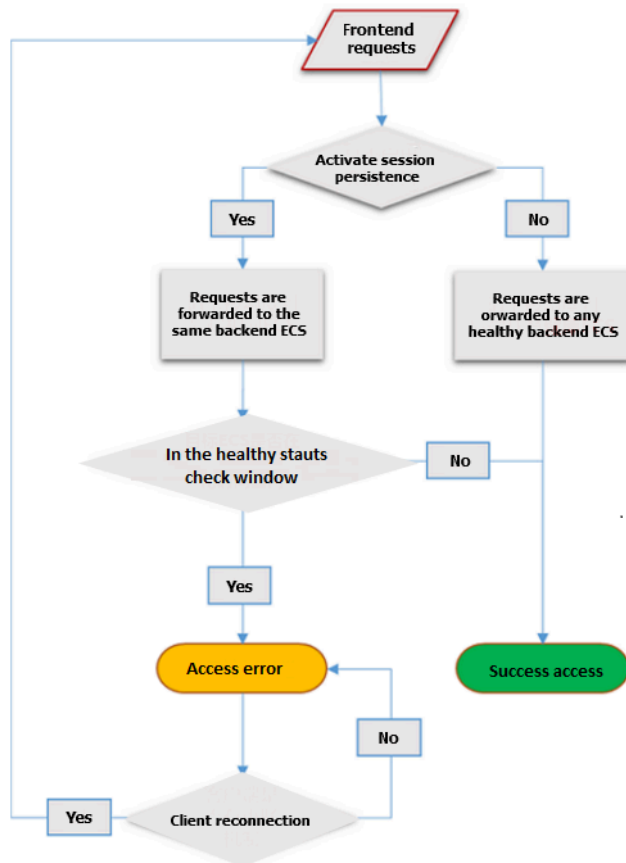
3.2 Configure health check

You can configure health check settings when adding a listener. In general, the default settings can meet your requirements.

Configure health check

You can configure the health check of a listener on the console or through API. For more information, see *Health check overview* and *Health check FAQ*.

To configure the health check, complete these steps:

1. Log on to the *SLB console*.

2. Select a region and all SLB instances in this region are displayed.

3. Click the ID of the target SLB instance.

4. On the **Instance Details** page, click the **Listeners** tab.

5. Click **Add Listener** or the **Configure** option of the target listener.

6. On the **Health Check** page, configure the health check.

   We recommend that you use the default values when configuring the health check.

Table 3-1: Health check configurations

Configuration	Description
Health Check Protocol	For TCP listeners, both the TCP health check and HTTP health check are supported. • TCP health check is based on network layer detection. • HTTP health check is performed by sending head requests.
Health Check Path and Domain Name (HTTP health check only)	By default, SLB sends an HTTP head request to the default homepage configured on the application server through the intranet IP address of the backend ECS instance to do the health check. If you do not use the default homepage of the application server to do health check, you must specify the URL for health check. Some application servers verify the host field in the request, therefore, the request header must contain the host field. If a domain name is configured in the health check, Server Load Balancer adds the domain name to the host field when forwarding the request to the backend server. If not, the health check request will be denied by the server and the health check may fail. Therefore, if your application server verifies the host field in the request, you must configure a domain name to make sure the health check works.
Normal Status Code (HTTP health check only)	Select the HTTP status code indicating that the health check is normal. The default values are http_2xx and http_2xx.
Health Check Port	The detection port used by the health check to access the backend ECS instances. By default, the backend port configured in the listener is used. 📋 **Note:** If a VServer group or an active/standby server group is configured for the listener, and the ECS instances in the group use different ports, leave this option empty. SLB uses the backend port of each ECS instance to do health check.
Response Timeout	The amount of time in seconds to wait for the response from a health check. If the backend ECS instance does not respond correctly within the specified time, the health check fails. The timeout range is 1-300 seconds. The default value is 10 seconds for UDP listeners and 5 seconds for HTTP/HTTPS/TCP listeners.

Configuration	Description
Health Check Interval	The time interval between two consecutive health checks. All node servers in the LVS cluster independently and concurrently perform health check on backend ECS instances according to the interval. The statistics from a health check request on a single ECS instance cannot reflect the health check interval because the health check time of each node server is not synchronized. The time range is 1-50 seconds. The default value is 5 seconds for UDP listeners, and 2 seconds for HTTP/HTTPS/TCP listeners.
Unhealthy Threshold	The number of consecutive failures of health check performed by the same LVS node server on the same ECS instance (from success to failure). Valid value: 2-10. The default value is 3.
Healthy Threshold	The number of consecutive successes of health check performed by the same LVS mode server on the same ECS instance (from failure to success). Valid value: 2-10. The default value is 3.
Health Check Requests and Results	When configuring health check for UDP listeners, you can enter the request contents (such as **youraccountID**) in **Health Check Request** and the expected response (such as **slb123**) in **Health Check Response**. Add the corresponding health check response logic to the application logic of the backend server. For example, return slb123 when youraccountID is received. If SLB receives the expected response from the backend server, the health check succeeds. Otherwise, the health check fails. This method can maximally guarantee the reliability of health check.

| Configure Server Load Balancer   ⟲ Back |

| Protocol and Listener | Backend Servers | **Health Check** | Submit |

Configure Health Check

ⓘ Health checks enable an SLB instance to automatically exclude unhealthy backend servers.

Enable Health Check

Advanced   Modify ⌄

Health Check Protocol	TCP		Health Check Port	Backend Server Port
Response Timeout	5 Seconds		**Health Check Interval**	2 Seconds
Healthy Threshold	3 Times		**Unhealthy Threshold**	3 Times

Previous   Next   Cancel

Example of health check response timeout and health check interval

Take the following health check configurations as the example:

- **Response Timeout: 5 seconds**

- **Health Check Interval: 2 seconds**

- **Healthy Threshold: 3 times**

- **Unhealthy Threshold: 3 times**

Health check failure time window = Response Timeout × Unhealthy Threshold + Health Check Interval × (Unhealthy Threshold - 1). That is, 5x3+2x(3-1)=19s.

The following figure shows the process to declare an unhealthy backend server:
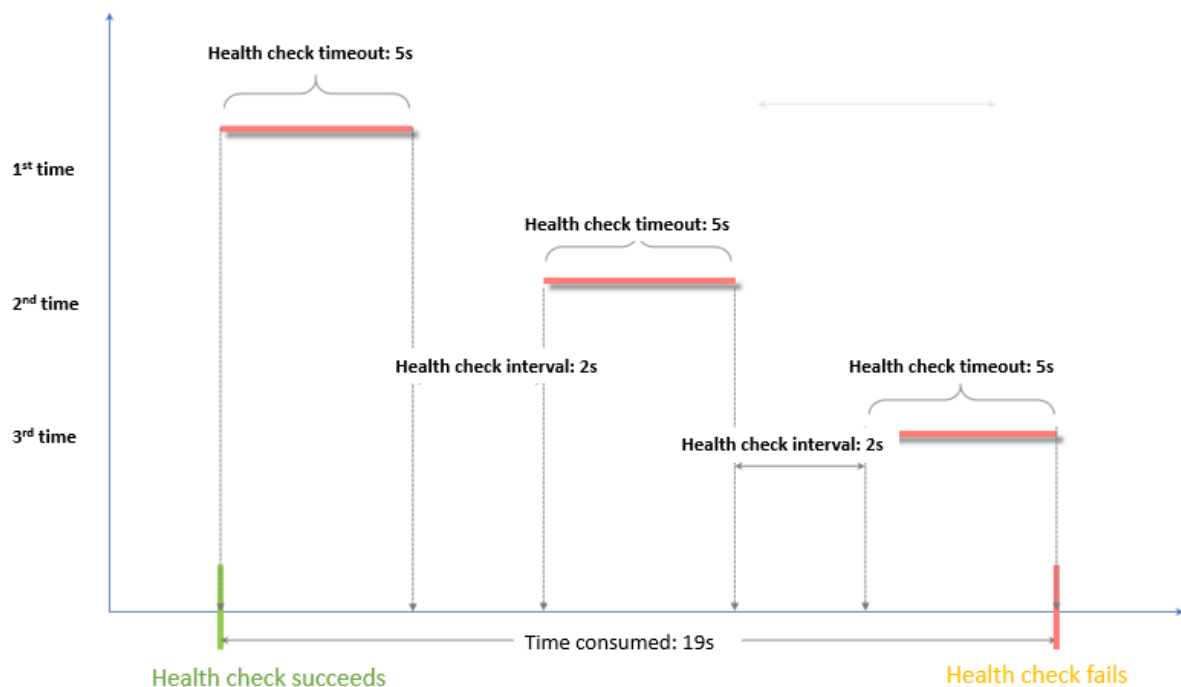
Health check success time window = Health check response time × Healthy Threshold + Health Check Interval × (Healthy Threshold - 1). That is, (1x3)+2x(3-1)=7s.

> **Note:**
>
> Health check response time is the duration from the time when the health check request is sent to the time when the response is received. When the TCP health check is used, the time is very short and almost negligible because the health check only detects whether the port is alive. When the HTTP health check is used, the response time depends on the performance and load of the application server and is usually within seconds.

The following figure shows the process to declare a healthy backend server (Assume that it takes 1 second for the backend server to respond to the health check request):
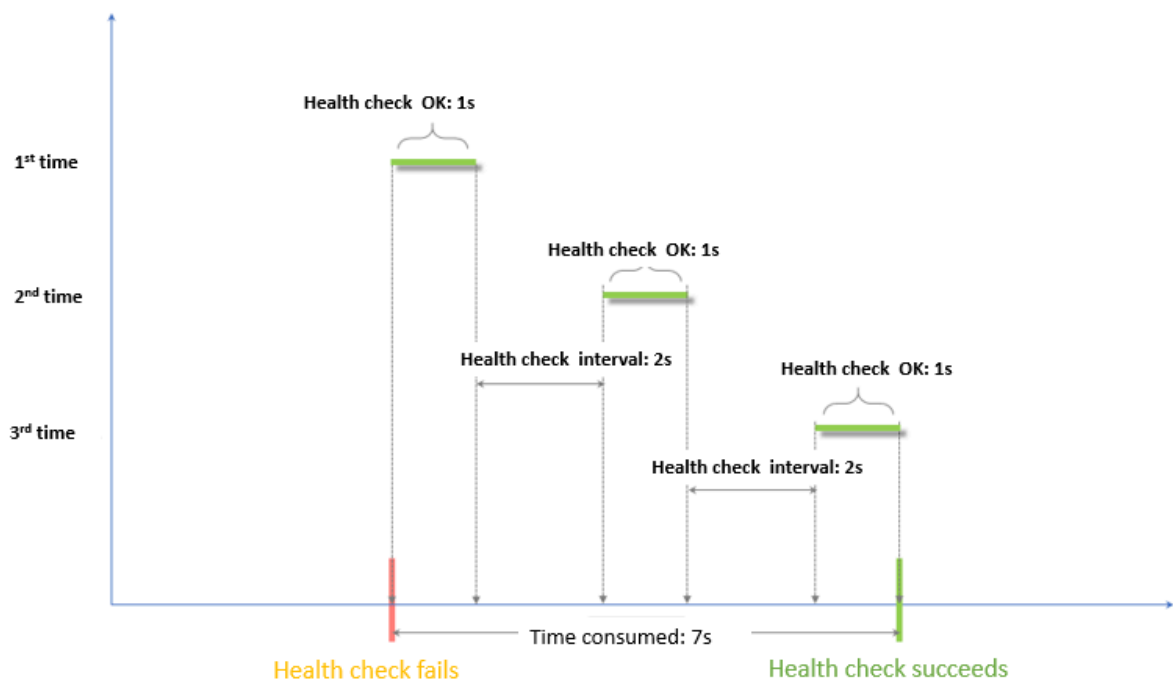
Configure a domain name in HTTP health check

When the HTTP health check is used, you can set a domain name for the health check, but it is not a required option. Some application servers verify the host field in the request, therefore, the request header must contain the host field. If a domain name is configured in the health check , Server Load Balancer adds the domain name to the host field when forwarding the request to the backend server. If not, the health check request will be denied by the server and the health check may fail. Therefore, if your application server verifies the host field in the request, you must configure a domain name to make sure the health check works.

3.3 Close health check

If health check is closed, requests may be distributed to unhealthy ECS instances, which can lead to service interruption. In general, we do not recommend closing health check.

Context

> 📋 **Note:**
>
> You can only close health check for HTTP and HTTPS listeners. The health check of UDP and TCP listeners cannot be closed.

Procedure

1. Log on to the *SLB console*.

2. On the **Instances** page, click the ID of the target instance.

3. In the **Listeners** tab, click **Configure** in the **Actions** column of the target listener.

4. On the **Configure Listener** page, click **Next** until the **Health Check** tab is displayed.

5. Close the health check switch, click **Next**, click **Submit**, and then click **OK**.

4 Backend servers

4.1 Backend server overview

Before using the load balancing service, you must add one or more ECS instances as the backend servers to an SLB instance to process the distributed client requests.

SLB service virtualizes the added ECS instances in the same region into an application pool featured with high performance and high availability. You can also manage backend servers through a VServer group. Different listeners can be associated with different server groups so that different listeners of an SLB instance can forward requests to the backend servers with different ports.

> **Note:**
>
> After a VServer group is configured for a listener, the listener will forward requests to the ECS instances in the associated VServer group instead of the ECS instances in the default server group.

You can increase or decrease the number of the backend ECS instances at any time and specify the ECS instances that receive requests. However, we recommend that you enable the health check function, and there must be at least one normal ECS to maintain service stability.

When adding ECS instances to an SLB instance, note the following:

- SLB does not support cross-region deployment. Make sure that the ECS instances and the SLB instance are in the same region.

- SLB does not limit the operating system used in the ECS instances as long as the applications deployed in the ECS instances are the same, and the data is consistent. However, we recommend that you use the same operating system for better management and maintenance.

- Up to 50 listeners can be added to an SLB instance. Each listener corresponds to an application deployed on backend ECS instances. The listening port of an SLB instance corresponds to the application port opened on the ECS instance.

- You can specify a weight for each ECS instance in the backend server pool. An ECS instance with a higher weight will receive a larger number of connection requests.

- If you have enabled the session persistence function, the requests distributed to the backend ECS instances may be imbalanced. If so, we recommend that you disable the session persistence function to check if the problem persists.

When the traffic is not distributed evenly, troubleshoot as follows:

1. Collect the access logs of the web service within a period of time.

2. Check if the number of logs of multiple ECS instances are different according to SLB configurations. If session persistence is enabled, you need to strip the access logs for the same IP address. If weights are configured for SLB, you need to calculate whether the percentage of access logs matches the weight ratio.)

- When an ECS instance is undergoing live migration, the persistent connections of the SLB may be interrupted and can be restored by reconnecting them. Be prepared for the reconnection.

Default server group

A default server group contains ECS instances that receive requests. If a listener is not associated with a VServer group or an active/standby server group, requests are forwarded to ECS instances in the default server group by default.

See *Manage a default server group* to create a default server group.

Active/standby server group

An active/standby server group only contains two ECS instances. One acts as the active server and the other acts as the standby server. No health check is performed on the standby server. When the active server is declared as unhealthy, the system forwards traffic to the standby server. When the active server is declared as healthy and restores service, the traffic is forwarded to the active server again.

See *Manage an active/standby server group* to create an active/standby server group.

> **Note:**
> Only Layer-4 listeners (TCP and UDP protocols) support configuring active/standby server groups.

VServer group

When you need to distribute different requests to different backend servers, or you want to configure domain name or URL based forwarding rules, you can use VServer groups.

See *Manage a VServer group* to create a VServer group.

4.2 Manage a default server group

Before using the SLB service, you must add at least one default server to receive client requests forwarded by SLB.

Add default servers

Before adding ECS instances to the default server group, make sure the following conditions are met:

- You have *created an SLB instance*.

- You have created ECS instances and deployed applications to process distributed requests.

To add ECS instances, complete these steps:

1. Log on to the *SLB console*.

2. On the **Server Load Balancer** page, select a region.

3. Click the ID of the target instance.

4. Click the **Default Server Group** tab.

5. Click **Add**.
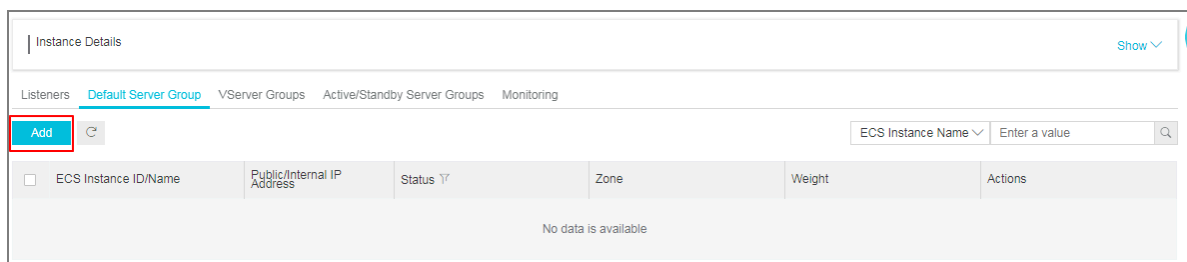
6. On the **Available Servers** page, find the target ECS instance and click **Add**, or click multiple target ECS instances and click **Add to Selected Server List** at the bottom of the page.
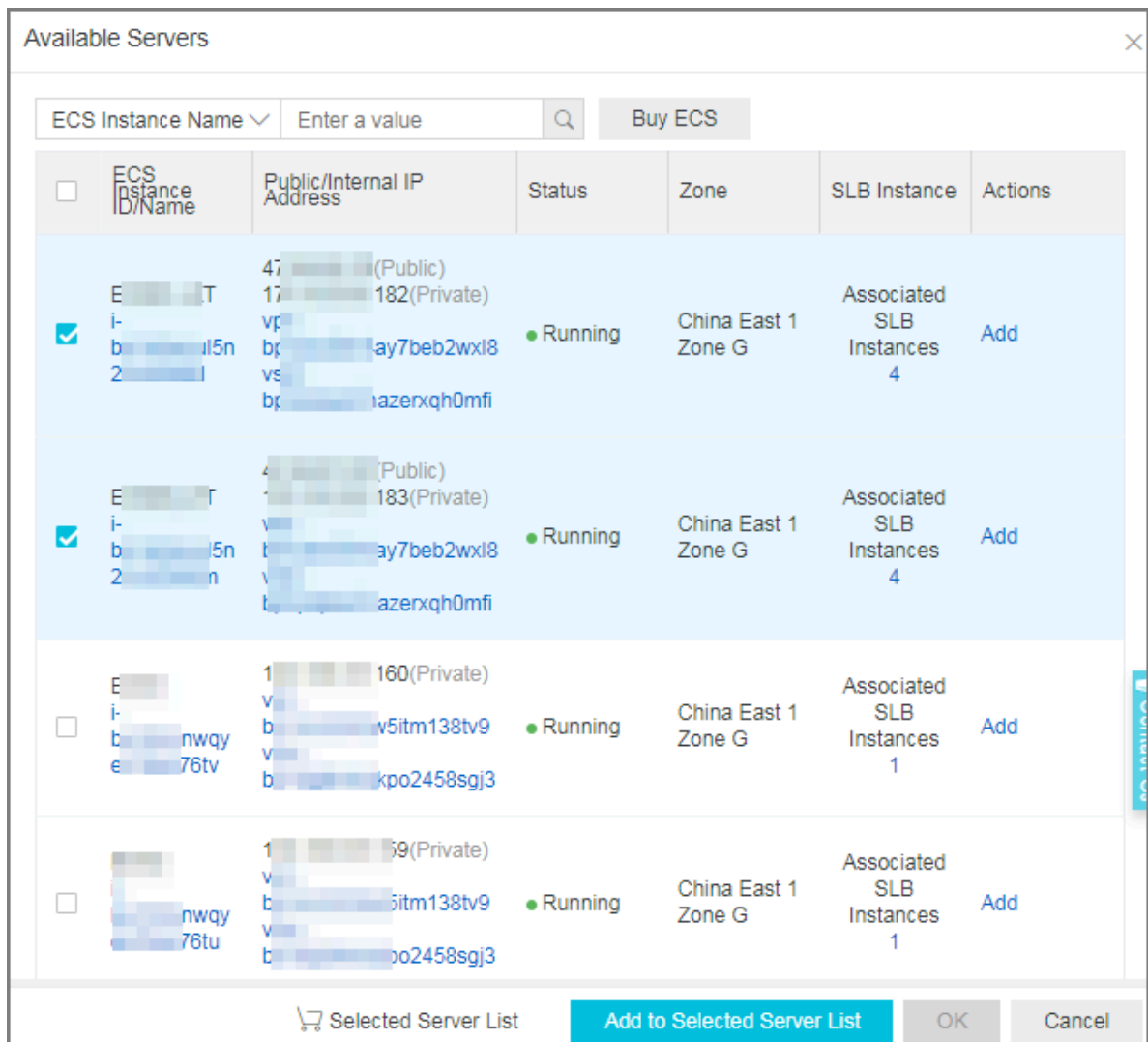
7. Click **OK**.

8. In the **Available Servers** dialog box, specify the weights of ECS instances and click **OK**.

   Weight: An ECS instance with a higher weight receivers more requests.

   > ⚠ **Notice:**
   >
   > If the weight is set to 0, no requests will be sent to the ECS instance.

   You can modify the ports and weights of added servers in batches.

   - Click ⤓ : Duplicate to below. If you modify the weight of the current server, the weights of

     all servers blow are also changed.

   - Click ⤒ : Duplicate to above. If you modify the weight of the current server, the weights of

     all servers above are also changed.

- Click ⬚ : Duplicate to all. If you modify the weight of the current server, the weights of all

    servers in the default server group are also changed.

- Click ⬚ : Clear all. If you clear the weight of the current server, the weights of all servers

    in the default server group are also cleared.

9. Click **OK**.

Edit the weight of a backend server

To edit the weight of a backend server, complete these steps:

1. Log on to the *SLB console*.

2. On the **Server Load Balancer** page, select a region.

3. Click the ID of the target SLB instance.

4. Click the **Default Server Group** tab.

5. Hover the mouse to the weight area of the target backend server, and then click the displayed

    pencil icon.

6. Modify the weight and then click **OK**.

    An ECS instance with a higher weight will receive a larger number of connection requests.

    > ⊘ **Notice:**
    >
    > If the weight is set to 0, no requests will be sent to the ECS instance.

Remove a backend server

To remove a backend server, complete these steps:

1. Log on to the *SLB console*.

2. On the **Server Load Balancer** page, select a region.

3. Click the ID of the target SLB instance.

4. Click the **Default Server Group** tab.

5. Click **Remove** in the **Actions** column to remove the backend server.

4.3 Manage a VServer group

A virtual server group (VServer group) is a group of ECS instances. If you associate a VServer group with a listener, the listener distributes requests to the associated VServer group instead of other backend servers.

If you add default backend servers, VServer groups and forwarding rules to the same Layer-7 listener, the sequence of request forwarding is as follows:

*   If the requests match a forwarding rule, the requests are distributed to the VServer group associated with the rule.

*   If no forwarding rule is matched and a VServer group is configured on the listener, the requests are distributed to the VServer group associated with the listener.

*   If no VServer group is configured on the listener, the requests are forwarded to ECS instances in the default server group.

Create a VServer group

Before creating a VServer group, make sure the following conditions are met:

- You have *created an SLB instance*.

- You have created ECS instances and deployed applications to process distributed requests.

Note the following when creating a VServer group:

- The ECS instances added to the VServer group and the SLB instance must be located in the same region.

- One ECS instance can be added to multiple VServer groups.

- One VServer group can be associated with multiple listeners.

- A VServer group consists of ECS instances and application ports.

To add ECS instances, complete these steps:

1. Log on to the *SLB console*.

2. On the **Server Load Balancer** page, select a region.

3. Click the ID of the target instance.

4. Click the **VServer Groups** tab.

5. On the **VServer Group** page, click **Create VServer Group**.

6. On the **Create VServer Group** page, complete these steps:

    a. In the **VServer Group Name** text box, enter the name of the VServer group.

    b. Click **Add** to select the server to add in the **Available Servers** list.

    c. Click **Add to Selected Server List** and click **OK**.

    d. Under the **Servers Added** tab, complete the following configuration and click **OK**.

       - **Port**: The backend port opened on the ECS instance to receive requests.

         The backend ports in a Server Load Balancer instance can be the same.

       - **Weight**: An ECS instance with a higher weight receivers more requests.

         ⓘ **Notice:**

         If the weight is set to 0, no requests will be sent to the ECS instance.

You can modify the ports and weights of added servers in batches.

- Click [ ]: Duplicate to below. If you modify the port or weight of the current server, the

  ports or weights of all servers blow are also changed.

- Click [ ]: Duplicate to above. If you modify the port or weight of the current server, the

  ports or weights of all servers above are also changed.

- Click [ ]: Duplicate to all. If you modify the port or weight of the current server, the

  ports or weights of all servers in the VServer group are also changed.

- Click [ ]: Clear all. If you clear the port or weight of the current server, the ports or

  weights of all servers in the VServer group are also cleared.

Edit a VServer group

To modify the ECS instance configuration in a VServer group, complete these steps:

1. Log on to the *SLB console*.

2. On the **Server Load Balancer** page, select a region.

3. Click the ID of the target instance.

4. Click the **VServer Groups** tab.

5. Find the target VServer group, and then click the **Edit** option.

6. Modify the port and weight of the ECS instance or click **Delete** to remove the ECS instance from the VServer group, and then click **OK**.

Delete a VServer group

To delete a VServer group, complete these steps:

1. Log on to the *SLB console*.

2. On the **Server Load Balancer** page, select a region.

3. Click the ID of the target instance.

4. Click the **VServer Groups** tab.

5. Find the target VServer group, and then click the **Delete** option.

Listeners	Default Server Group	VServer Groups	Active/Standby Server Groups	Monitoring	
Create VServer Group ○ C					
Group Name	**Group ID**	**Listener**	**Forwarding Rule**		**Actions**
test1	rsp-bp1d2e3qel4wb	--	--		Edit Delete
test2	rsp-bp1h6b45s4y5c	--	--		Edit Delete

6. In the displayed dialog box, click **OK**.

4.4 Manage an active/standby server group

If you have traditional active/standby requirement, where one backend server is used as the active server and the other as the standby server, create an active/standby server group. When the active server works normally, requests are distributed to it; when the active server is down, the requests will be distributed to the standby server to avoid service interruption.

An active/standby server group only contains two ECS instances. One acts as the active server and the other acts as the standby server. No health check is performed on the standby server. When the active server is declared as unhealthy, the system forwards traffic to the standby server. When the active server is declared as healthy and restores service, the traffic is forwarded to the active server again.

> (!) **Notice:**
> Only Layer-4 listeners (TCP and UDP protocols) support configuring active/standby server groups.

Create an active/standby server group

Before creating an active/standby server group, make sure the following conditions are met:

• You have *created an SLB instance*.

• You have created ECS instances and deploy applications to process distributed requests.

To add ECS instances, complete these steps:

1. Log on to the *SLB console*.

2. On the **Server Load Balancer** page, select a region.

3. Click the ID of the target SLB instance.

4. Click the **Active/Standby Server Groups** tab.

5. On the **Active/Standby Server Groups** page, click **Create Active/Standby Server Group**.

6. On the **Create Active/Standby Server Group** page, complete these steps:

    a. In the **Name** text box, enter the name of the active/standby server group.

b. Click **Add** to select the server to add in the **Available Servers** list.

You can add up to two ECS instances to an active/standby server group.

c. Click **Add to Selected Server List** and click **OK**.

d. In the **Servers Added** tab, complete the following configuration and click **OK**.

- **Port**: The backend port opened on the ECS instance to receive requests.

    The back-end ports in a Server Load Balancer instance can be the same.

- **Server**: Select a server to act as the active server.

Delete an active/standby server group

To delete an active/standby server group, complete these steps:

1. Log on to the *SLB console*.

2. On the **Server Load Balancer** page, select a region.

3. Click the ID of the target SLB instance.

4. Click the **Active/Standby Server Groups** tab.

5. Click **Delete** next to the target active/standby server group.

6. In the displayed dialog box, click **OK**.

5 Certificate management

5.1 Certificate requirements

Server Load Balancer only supports certificates in the PEM format. Before uploading a certificate, make sure that the certificate, certificate chain, and private key conform to the rules described in this section.

Certificates issued by a root CA

If the certificate is issued by a root CA, the received certificate is the only one required to be uploaded to Server Load Balancer. The website that is configured with the certificate will be trusted by the web browser without configuring additional certificates.

The certificate format must meet the following requirements:

- The certificate content is placed between `-----BEGIN CERTIFICATE-----`, `-----END CERTIFICATE-----`. Include the header and footer when uploading the certificate.
- Each line except the last must contain exactly 64 characters. The last line can contain 64 or fewer characters.
- Space is not allowed in the content.

The following is a sample certificate issued by a root CA.

```
-----BEGIN CERTIFICATE-----
MIIE+TCCA+GgAwIBAgIQU306HIX4KsioTW1s2A2krTANBgkqhkiG9w0BAQUFADCB
tTELMAkGA1UEBhMCVVMxFzAVBgNVBAoTDlZlcmlTaWduLCBJbmMuMR8wHQYDVQQL
ExZWZXJpU2lnbiBUcnVzdCBOZXR3b3JrMTsw0QYDVQQLEzJUZXJtcyBvZiB1c2Ug
YXQgaHR0cHM6Ly93d3cudmVyaXNpZ24uY29tL3JwYSoAYykwOTEvMC0GA1UEAxMm
VmVyaVNpZ24gQ2xhc3MgMyBTZWN1cmUgU2VydmVyIENBIC0gRzIwHhcNMTAxMDA4
MDAwMDAwWhcNMTMxMDA3MjM1OTU5WjBqMQswCQYDVQQGEwJVUzETMBEGA1UECBMK
V2FzaGluZ3RvbjEQMA4GA1UEBxQHU2VhdHRsZTEYMBYGA1UEChQPQW1hem9uLmNv
bSBJbmMuMRowGAYDVQQDFBFpYW0uYW1hem9uYXdzLmNvbTCBnzANBgkqhkiG9w0B
AQEFAAOBjQAwgYkCgYEA3Xb0EGea2dB8QGEUwLcEpwvGawEkUdLZmGL1rQJZdeeN
3vaF+ZTm8Qw5Adk2Gr/RwYXtpx04xvQXmNm+9YmksHmCZdruCrW1eN/P9wBfqMMZ
X964CjVov3NrF5AuxU8jgtw0yu//C3hWn0uIVGdg76626gg0oJSaj48R2n0MnVcC
AwEAAa0CAdEwggHNMAkGA1UdEwQCMAAwCwYDVR0PBAQDAgWgMEUGA1UdHwQ+MDww
0qA4oDaGNGh0dHA6Ly9TVlJJTZWN1cmUtRzItY3JsLnZlcmlzaWduLmNvbS9TVlJJT
ZWN1cmVHMi5jcmwwRAYDVR0gBD0w0zA5BgtghkgBhvhFAQcXAzAqMCgGCCsGAQUF
BwIBFhxodHRwczovL3d3dy52ZXJpc2lnbi5jb20vcnBhMB0GA1UdJQQWMBQGCCsG
AQUFBwMBBggrBgEFBQcDAjAfBgNVHSMEGDAWgBSl7wsRzsBBA6NKZZBIshzgVy19
RzB2BggrBgEFBQcBAQRqMGgwJAYIKwYBBQUHMAGGGGh0dHA6Ly9vY3NwLnZlcmlz
aWduLmNvbTBABggrBgEFBQcwAoY0aHR0cDovL1NWUlNlY3VyZS1HMi1haWEudmVy
aXNpZ24uY29tL1NVUlNlY3VyZUcyLmNlcjBuBggrBgEFBQcBDARiMGChXqBcMFow
WDBWFglpbWFnZS9naWYwITAfMAcGBSsOAwIaBBRLa7kolgYMu9BSOJsprEsHiyEF
GDAmFiRodHRwOi8vbG9nby52ZXJpc2lnbi5jb20vdnNsb2dvMS5naWYwDQYJKoZI
hvcNAQEFBQADggEBALpFBXeG782QsTtGwEE9zBcVCuKjrsl3dWK1dFiq3OP4y/Bi
ZBYEywBt8zNuYFUE25Ub/zmvmpe7p0G76tmQ8bRp/4qkJoiSesHJvFgJ1mksr3IQ
3gaE1aN2BSUIHxGLn9N4F09hYwwbeEZaCxfgBiLdEIodNwzcvGJ+2LlDWGJOGrNI
NM856xjqhJCPxYzk9buuCl1B4Kzu0CTbexz/iEgYV+DiuTxcfA4uhwMDSe0nynbn
1qiwRk450mCOnqH4ly4P4lXo02t4A/DI1I8ZNct/Qfl69a2Lf6vc9rF7BELT0e5Y
R7CKx7fc5xRaeQdyGj/dJevm9BF/mSdnclS5vas=
-----END CERTIFICATE-----
```

Certificates issued by an intermediate CA

If a certificate is issued by an intermediate CA, you will obtain multiple intermediate certificates. You must upload both the server certificate and the immediate certificates to Server Load Balancer .

The format of the certificate chain must meet the following requirements:

- Put the server certificate in the first place and the intermediate certificates in the second place without any blank line in between.

- Space is not allowed in the content.

- Blank lines are not allowed in the content. Each line must contain exactly 64 characters. For more information, see *RFC1421*.

- Conform to the certificate requirements as described in the certificate description. In general , the intermediate CA will provide an instruction about the certificate format when issuing the certificate, the certificate chain must conform to the format requirements.

The following is a sample certificate chain.

```
-----BEGIN CERTIFICATE-----
-----END CERTIFICATE-----
-----BEGIN CERTIFICATE-----
-----END CERTIFICATE-----
```

```
      -----BEGIN CERTIFICATE-----
      -----END CERTIFICATE-----
```

RSA private key

When uploading a server certificate, you also need to upload the private key of the certificate.

The RSA private key format must meet the following requirements:

- The key is placed between `-----BEGIN RSA PRIVATE KEY-----`, `-----END RSA PRIVATE KEY-----`. Include the header and footer when uploading the key.

- Blank line is not allowed in the content. Each line except the last must contain exactly 64 characters. The last line can contain 64 or fewer characters. For more information, see *RFC1421*.

If your private key is encrypted, for example, the header and footer are `-----BEGIN PRIVATE KEY-----`, `-----END PRIVATE KEY-----` or `-----BEGIN ENCRYPTED PRIVATE KEY-----`, `-----END ENCRYPTED PRIVATE KEY-----`, or the private key contains `Proc-Type : 4,ENCRYPTED`, run the following command to convert the private key before uploading it to Server Load Balancer:

```
openssl rsa -in old_server_key.pem -out new_server_key.pem
```

The following is a sample RSA private key.

5.2 Create a certificate

To configure an HTTPS listener, you can directly use a certificate in SSL Certificate Service or upload the required third-party server certificate and CA certificate to the SLB. Then you no longer need to configure certificates on the backend servers.

SLB supports certificates from the following two resources:

- Certificates issued or hosted by Alibaba Cloud SSL Certificate Service: You can select the certificate from Alibaba Cloud SSL Certificate Service. You will receive alerts when the certificate is about to expire, and can renew the certificate easily.

  Currently client CA certificates are not supported.

- Third-party certificates: To upload a third-party certificate, you must have the public key/private key files of the certificate.

  HTTPS server certificates and client CA certificates are supported.

Note the following before creating a certificate:

- If you want to use a certificate in multiple regions, you must select multiple regions when creating the certificate.

- You can create up to 100 certificates per account.

Select a certificate from SSL Certificate Service

Alibaba Cloud SSL Certificate Service issues digital certificates of different brands to provide HTTPS services for websites, so that the websites are trustworthy and are prevented from hijacking, tampering and interception. It also uniformly manages the life cycles of certificates to simplify certificate deployment. For more information, see *SSL certificate service*.

To use certificates in SSL Certificate Service, you must log on to the *SSL Certificate console* to buy a certificate or upload a third-party certificate to the SSL Certificate Service.

To select a certificate from SSL Certificate Service, complete these steps:

1. Log on to the *SLB console*.

2. In the left-side navigation pane, click **Certificates**.

3. Click **Create Certificate**. On the **Create Certificate** page, select **Select Certificate From SSL Certificate Service**.
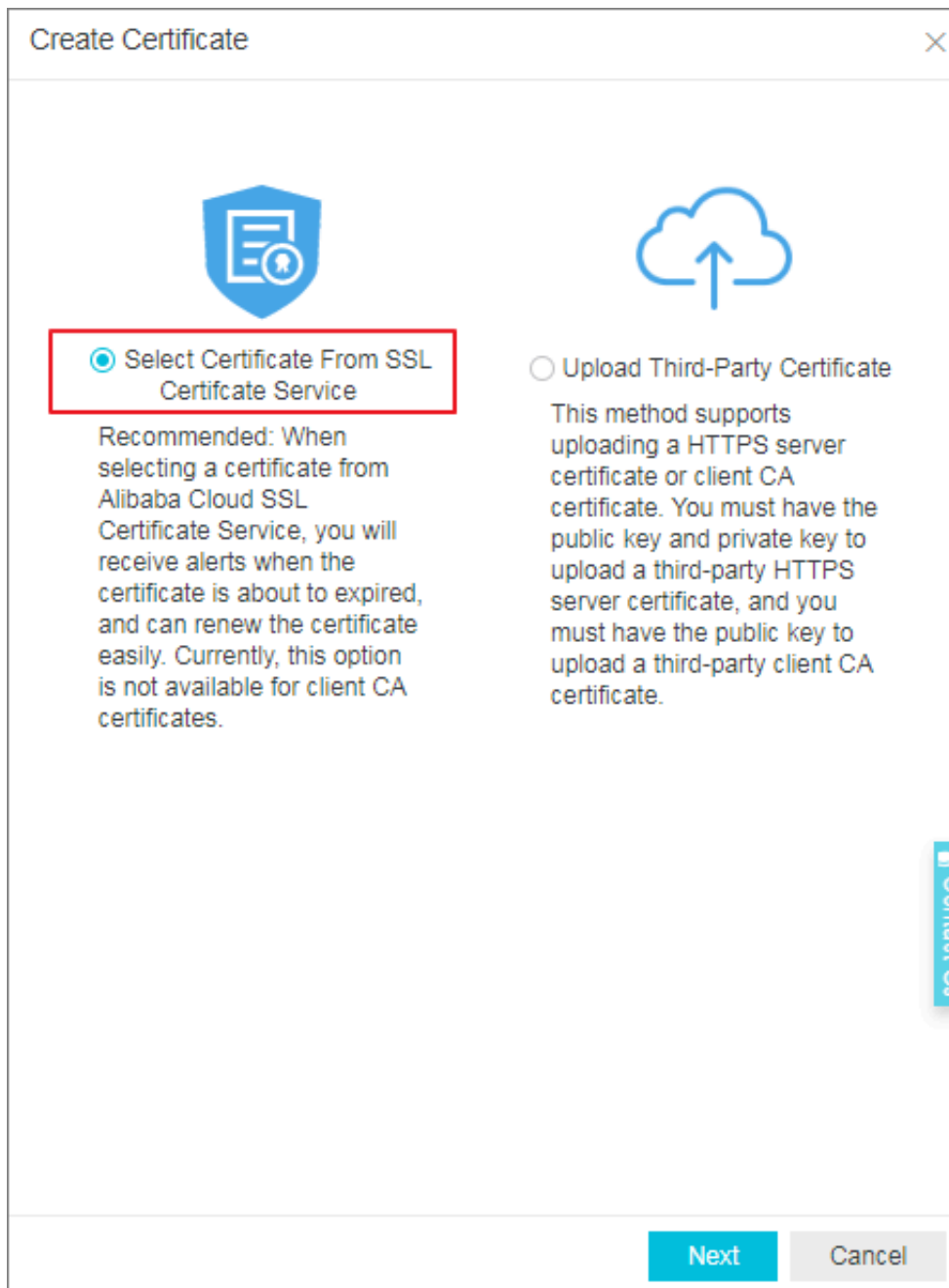
4. Click **Next**. On the **Select Certificate From SSL Certificate Service** page, select the region to deploy the certificate and then select the SSL certificate to use from the certificate list.

   A certificate cannot be used across regions. If a certificate is to be used in multiple regions, select all these regions.

5. Click **OK**.

Upload a third-party certificate

Before uploading a third-party certificate, make sure the following conditions are met:

- You have purchased a server certificate.

- You have generated a CA certificate and a client certificate. For more information, see

    Generate a CA certificate.

To upload a third-party certificate to the SLB instance, complete these steps:

1. Log on to the *SLB console*.

2. In the left-hand navigation pane, click **Certificates**.

3. Click **Create Certificate**.

4. On the **Create Certificate** page, click **Upload Third-party Certificate**.
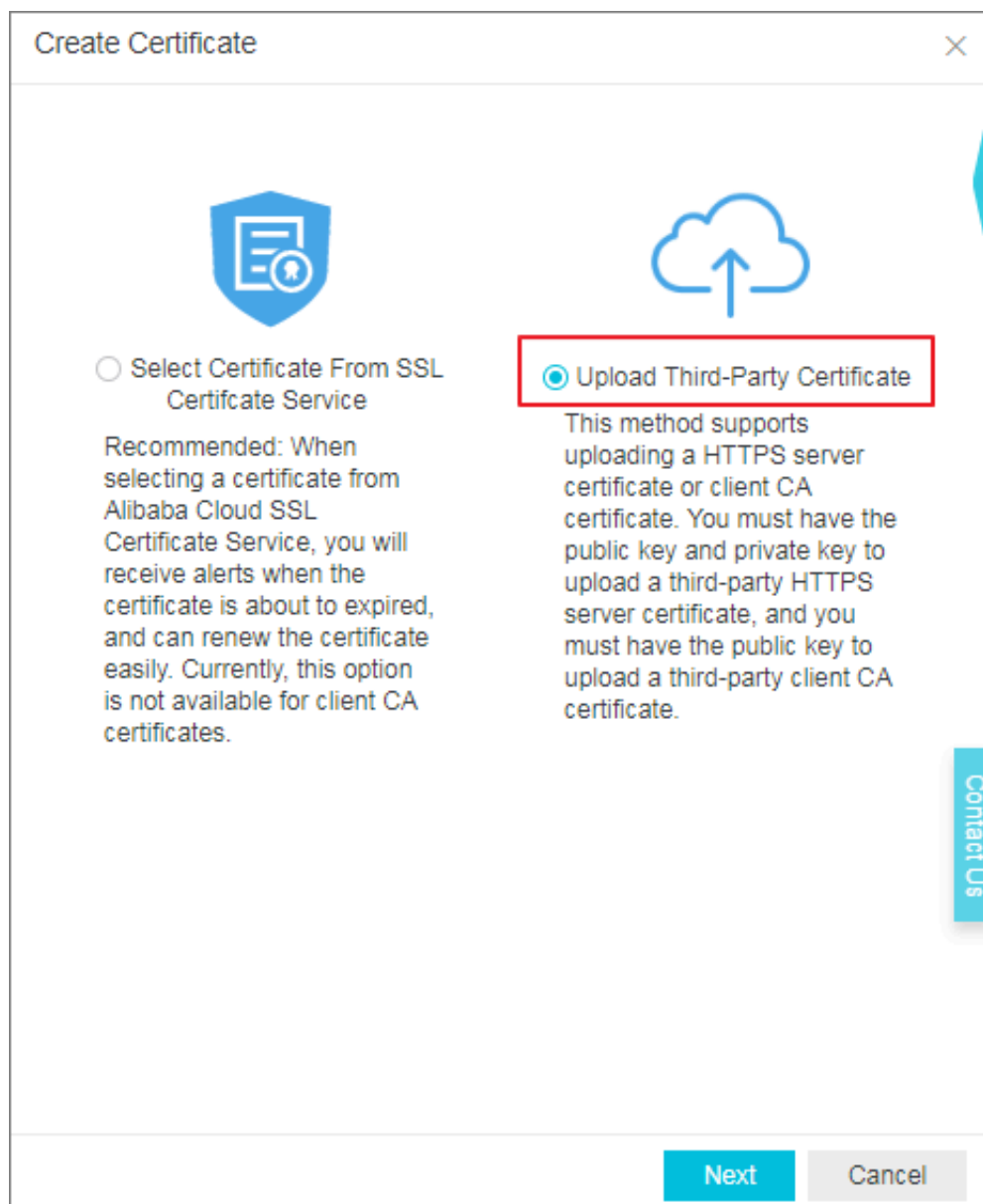
5. Click **Next**. On the **Upload Third-Party Certificate** page, upload the certificate content.

Configuration	Description
Certificate Name	Enter a certificate name. The name must be 1-80 characters long, and can only contain letters , numbers and the following special characters: _/. -
Regions	Select one or more regions where the certificate is uploaded. A certificate cannot be used across regions. If a certificate is to be used in multiple regions, select all these regions.
Certificate Type	Select the type of the certificate to be uploaded: • **Server Certificate**: For HTTPS one-way authentication, only the server certificate and the private key are required. • **CA Certificate**: For HTTPS mutual authentication, both the server certificate and the CA certificate are required.
Certificate Content	Paste the certificate content in the editor. Click **View Sample Certificate** to view the valid certificate formats. For more information, see *Certificate requirements*.
Private Key	Paste the private key of the server certificate in the editor. Click **View Sample Certificate** to view the valid certificate formats. For more information, see *Certificate requirements*. (!) **Notice:** A private key is only required when uploading a server certificate.

6. Click **OK**.

5.3 Generate a CA certificate

When configuring HTTPS listeners, you can use self-signed CA certificates. Follow the instructions in this document to generate a CA certificate and use the CA certificate to sign a client certificate.

Generate a CA certificate by using Open SSL

1. Run the following commands to create a `ca` folder in the `/root` directory and then create four sub folders under the `ca` folder.

```
$ sudo mkdir ca
$ cd ca
$ sudo mkdir newcerts private conf server
```

• `newcerts` is used to store the digit certificate signed by a CA certificate.

- *private* is used to store the private key of the CA certificate.

- *conf* is used to store the configuration files.

- *server* is used to store the server certificate.

2. Create an *openssl.conf* file that contains the following information in the *conf* directory.

```
[ ca ]
default_ca = foo
[ foo ]
dir = /root/ca
database = /root/ca/index.txt
new_certs_dir = /root/ca/newcerts
certificate = /root/ca/private/ca.crt
serial = /root/ca/serial
private_key = /root/ca/private/ca.key
RANDFILE = /root/ca/private/.rand
default_days = 365
default_crl_days= 30
default_md = md5
Unique_subject = No
Policy = policy_any
[ policy_any ]
countryName = match
stateOrProvinceName = match
organizationName = match
organizationalUnitName = match
localityName = optional
commonName = supplied
emailAddress = optional
```

3. Run the following command to generate a private key.

```
$ cd /root/ca
 $ sudo openssl genrsa -out private/ca.key
```

The following figure is an example of key generation.

4. Run the following command and input the required information according to the prompts. Press

Enter to generate a *csr* file.

```
$ sudo openssl req -new -key private/ca.key -out private/ca.csr
```

> 📋 **Note:**
>
> Common Name is the domain name of the SLB instance.

```
root@iZbp1hfvivcqx1jbwap31iZ:~/ca# sudo openssl req -new -key private/ca.key -ou
t private/ca.csr
You are about to be asked to enter information that will be incorporated
into your certificate request.
What you are about to enter is what is called a Distinguished Name or a DN.
There are quite a few fields but you can leave some blank
For some fields there will be a default value,
If you enter '.', the field will be left blank.
-----
Country Name (2 letter code) [AU]:CN
State or Province Name (full name) [Some-State]:ZheJiang
Locality Name (eg, city) []:HangZhou
Organization Name (eg, company) [Internet Widgits Pty Ltd]:Alibaba
Organizational Unit Name (eg, section) []:Test
Common Name (e.g. server FQDN or YOUR name) []:mydomain
Email Address []:a@alibaba.com

Please enter the following 'extra' attributes
to be sent with your certificate request
A challenge password []:
An optional company name []:
root@iZbp1hfvivcqx1jbwap31iZ:~/ca#
```

5. Run the following command to generate a `crt` file.

```
$ sudo openssl x509 -req -days 365 -in private/ca.csr -signkey
private/ca.key -out private/ca.crt
```

6. Run the following command to set the start sequence number for the private key, which can be any four characters.

```
$ sudo echo FACE > serial
```

7. Run the following command to create a CA key library.

```
$ sudo touch index.txt
```

8. Run the following command to create a certificate revocation list for removing the client certificate.

```
$ sudo openssl ca -gencrl -out /root/ca/private/ca.crl -crldays 7 -
config "/root/ca/conf/openssl.conf"
```

The response is as follows:

```
Using configuration from /root/ca/conf/openssl.conf
```

Sign the client certificate

1. Run the following command to generate a `users` folder under the `ca` directory to store the client key.

```
$ sudo mkdir users
```

2. Run the following command to create a key for the client certificate.

```
$ sudo openssl genrsa -des3 -out /root/ca/users/client.key 1024
```

📋 **Note:**

Enter a pass phrase when creating the key. It is the password to protect the private key from unauthorized access. The pass phrase entered is the password for this key.

3. Run the following command to create a `csr` file for requesting certificate sign.

```
$ sudo openssl req -new -key /root/ca/users/client.key -out /root/ca
/users/client.csr
```

Enter the pass phrase set in the previous step when prompted.
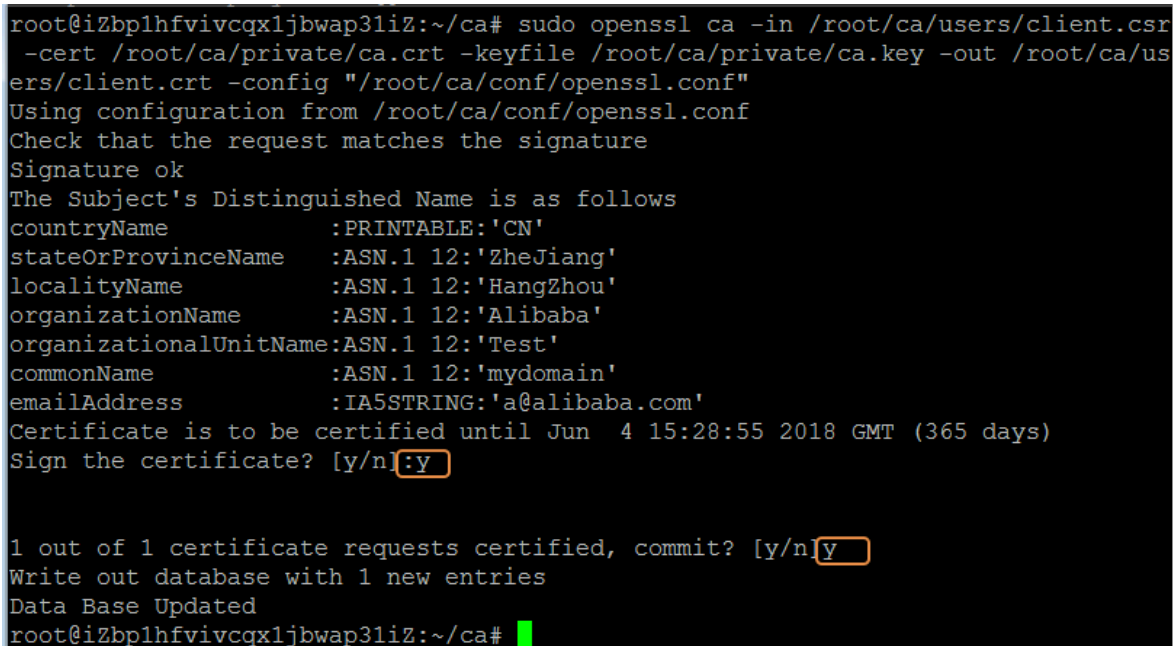
📋 **Note:**

> A challenge password is the password of the client certificate. Note that it is not the password
>
> of the client key.

4. Run the following command to sign the client key.

```
$ sudo openssl ca -in /root/ca/users/client.csr -cert /root/ca/
private/ca.crt -keyfile /root/ca/private/ca.key -out /root/ca/users/
client.crt -config "/root/ca/conf/openssl.conf"
```

Enter `y` twice when prompted.

```
root@iZbp1hfvivcqx1jbwap31iZ:~/ca# sudo openssl ca -in /root/ca/users/client.csr
 -cert /root/ca/private/ca.crt -keyfile /root/ca/private/ca.key -out /root/ca/us
ers/client.crt -config "/root/ca/conf/openssl.conf"
Using configuration from /root/ca/conf/openssl.conf
Check that the request matches the signature
Signature ok
The Subject's Distinguished Name is as follows
countryName           :PRINTABLE:'CN'
stateOrProvinceName   :ASN.1 12:'ZheJiang'
localityName          :ASN.1 12:'HangZhou'
organizationName      :ASN.1 12:'Alibaba'
organizationalUnitName:ASN.1 12:'Test'
commonName            :ASN.1 12:'mydomain'
emailAddress          :IA5STRING:'a@alibaba.com'
Certificate is to be certified until Jun  4 15:28:55 2018 GMT (365 days)
Sign the certificate? [y/n]:y


1 out of 1 certificate requests certified, commit? [y/n]y
Write out database with 1 new entries
Data Base Updated
root@iZbp1hfvivcqx1jbwap31iZ:~/ca#
```

5. Run the following command to convert the certificate to a `PKCS12` file.

```
$ sudo openssl pkcs12 -export -clcerts -in /root/ca/users/client.crt
 -inkey /root/ca/users/client.key -out /root/ca/users/client.p12
```

Enter the password of the client key when prompted. Then, enter the password used for

exporting the client certificate. This password is used to protect the client certificate, which is

required when installing the client certificate.

6. Run the following command to view the generated client certificate.

```
cd users
```

```
ls
```

5.4 Convert the certificate format

Server Load Balancer supports PEM certificates only. Certificates in other formats must be converted to the PEM format before they can be uploaded to Server Load Balancer. We recommend that you use Open SSL for conversion.

Convert DER to PEM

DER: This format is usually used on a Java platform. The certificate file suffix is generally *. der*, *. cer*, or *. crt*.

- Run the following command to convert the certificate format:

```
openssl x509 -inform der -in certificate.cer -out certificate.pem
```

-

- Run the following command to convert the private key:

```
openssl rsa -inform DER -outform PEM -in privatekey.der -out
privatekey.pem
```

Convert P7B to PEM

P7B: This format is usually used in a Windows server and Tomcat.

Run the following command to convert the certificate format:

```
openssl pkcs7 -print_certs -in incertificate.p7b -out outcertificate.
cer
```

Convert PFX to PEM

PFX: This format is usually used in a Windows server.

- Run the following command to extract the certificate:

```
openssl pkcs12 -in certname.pfx -nokeys -out cert.pem
```

- Run the following command to extract the private key:

```
openssl pkcs12 -in certname.pfx -nocerts -out key.pem -nodes
```

5.5 Replace a certificate

To avoid the impact of certificate expiration on your service, replace the certificate before the certificate expires.

Procedure

1. Create and upload a new certificate.

   For more information, see *Upload certificates* and *Generate certificates*.

2. Configure the new certificate in HTTPS listener configuration.

   For more information, see *Add an HTTPS listener*.

3. On the **Certificates** page, find the target certificate, and then click **Delete**.

4. In the displayed dialog box, click **OK**.

6 Log management

6.1 View operation logs

You can view the logs of operations performed on SLB instances, HTTP listeners and server certificates in the past one month.

Context

The operation logs are recorded in ActionTrail. ActionTrail records the operations acted upon your Alibaba Cloud resources. You can query operation records and store the records to OSS.

Procedure

1. Log on to the *SLB console*.

2. In the left-side navigation pane, click **Logs** > **Operation Log**.

3. Click **View Operation Logs**.

4. On the **History Search** page, complete these steps to view operation logs:

   a) Select **Resource Type** as a filter.

   b) Select the SLB resource of which operation logs you want to view.

   In this tutorial, **LoadBalancer** is selected.

   c) Select an event type.

   In this tutorial, **All** is selected.
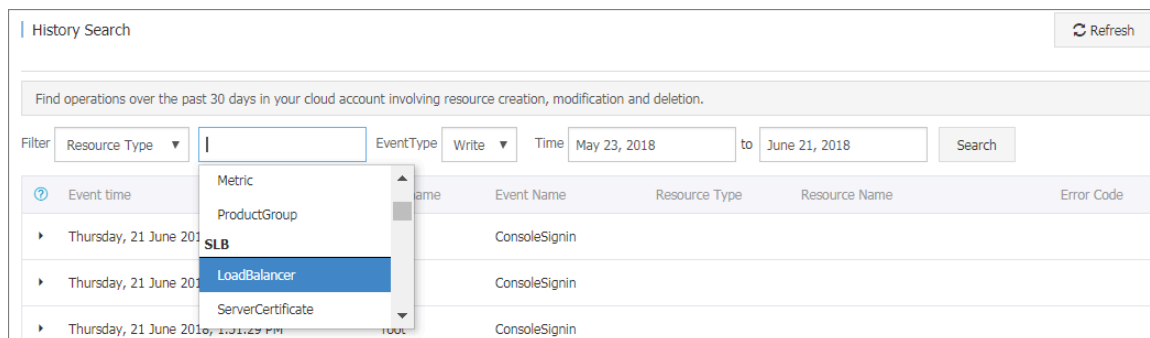
   d) Select the time range to search.

   e) Click **Search** to view logs of operations performed on the selected resource.

   Expand a record to view more detailed information.

6.2 Manage health check logs

You can view the health logs within three days on **Health Check Logs** page. If you want to get health check logs generated three days or longer before, you can store the health check logs to OSS. In this way, you can download complete health check logs.

Store health check logs

You can view the health check logs of the backend servers by using the health check log function. Currently, logs in the past three days are provided. If you want to view more logs, store the health check logs to OSS.

You can enable and disable the storage function at any time. After the storage function is enabled, SLB will create a folder named `AliyunSLBHealthCheckLogs` in the selected bucket to store the health check logs. The health logs are generated hourly and the system will create a subfolder named after the date to store the log files generated in that day, for example `20170707`.

The log files in a day are named after the time when they are generated. For example, the file name of a log file generated between 00:00-01:00 is `01.txt` and the file name of a log file generated between 01:00-02:00 is `02.txt`.

> **Note:**
>
> The health check logs are generated only when the backend server is abnormal. Health check logs are generated only when the backend server is abnormal. If no failures occur for all the backend servers in an hour, no health check logs are generated in that hour.

To store health check logs, complete these steps:

1. *Create a bucket*

2. *Authorize SLB to access OSS*

3. *Configure log storage*

Step 1 Create a bucket

1. Open the *OSS product page* and click **Buy Now** to activate the OSS service.

2. Log on to the OSS console.

3. Click **Create Bucket**.
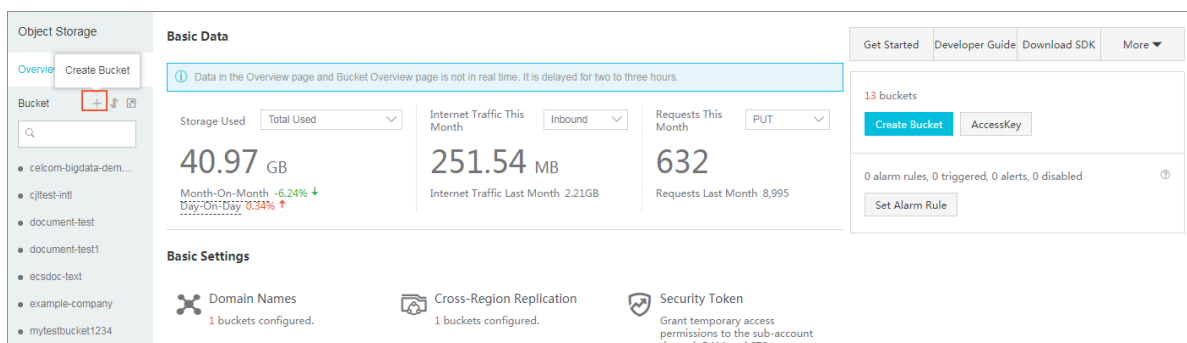
4. In the **Create Bucket** dialog box, configure the bucket and click **OK**.

> **Note:**
>
> Make sure that the bucket and the SLB instance are in the same region.

Step 2 Authorize SLB to access OSS

After creating a bucket, you have to authorize the log role (`SLBLogDefaultRole`) to access OSS resources.

> **Notice:**
>
> The authorization is required only for the first time.

1. On the SLB console, click **Logs** > **Health Check Logs**.

2. Click **1. Activate OSS.** if OSS has not been activated yet.

3. On the **Health Check Logs** page, click **Authorize Now** in the **2. Authorize the required RAM role.** section.

4. Read the authorization description, and then click **Confirm Authorization Policy**.

5. Log on to the RAM console.

6. In the left-side navigation pane, click **Roles** and find the role named SLBLogDefaultRole, and then click **Authorize**.

7. In the **Edit Role Authorization Policy** dialog box, find the **AliyunOSSFullAccess** policy, and

then click **OK**.
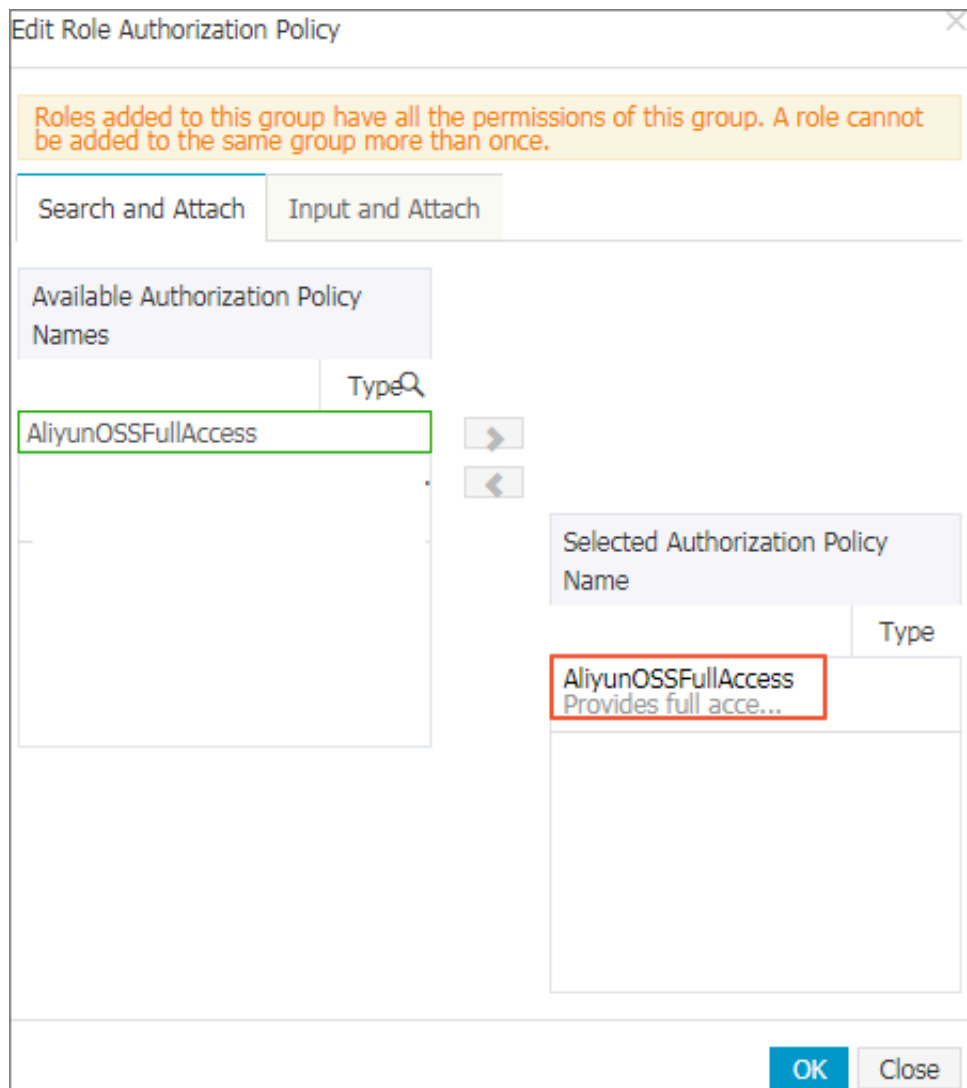
After the authorization, click **SLBLogDefaultRole**, and then click **Role Authorization Policies**

to view the attached policy.

Step 3 Configure log storage

1. Log on to the *SLB console*.

2. In the left-side navigation pane, click **Logs** > **Health Check Logs**.

3. On the **Health Check Logs** page, click the **Log Storage** tab.

4. Click **Configure Log Storage** link of the target region.

Health Check Logs			
Logs Log Storage			
Region	Status	Details	Actions
China East 1 (Hangzhou)		Bucket: slbyh	Configure Log Storage
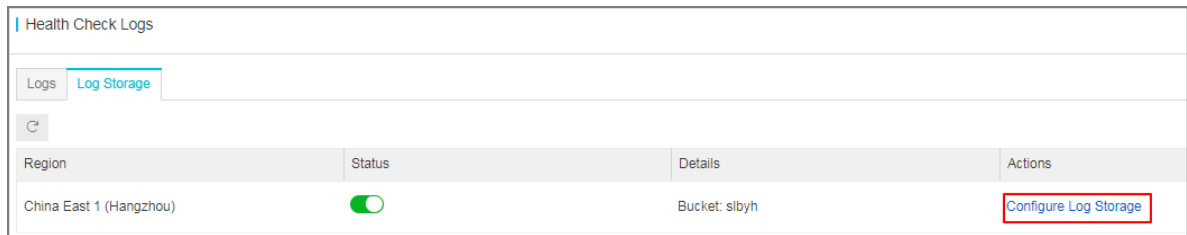
5. In the **Configure Log Storage** dialog box, select a bucket to store health check logs, and then click **OK**.

6. Click the switch in the status column to enable log storage.

View health check logs

To view the health check logs generated in the past three days, complete these steps:

1. Log on to the *SLB console*.

2. In the left-side navigation pane, click **Logs** > **Health Check Logs**.

3. On the **Health Check Logs** page, click the **Logs** tab.

> 📋 **Note:**
>
> Health check logs are generated only when the health status of a backend server is abnormal. Health check logs are generated every one hour. If no failure occurs to all the backend servers in an hour, no health check logs are generated in that hour.

- The `SLB_instance_IP:port to Added_ECS_instance_IP:port abnormal; cause:XXX` log message indicates that the backend server is abnormal. Troubleshoot according to the detailed error message.

- The `SLB_instance_IP:port to Added_ECS_instance_IP:port normal` log message indicates that the backend server becomes normal again.

Health Check Logs			
Logs Log Storage			
ⓘ			
Instance ID	Time	Details	
lb▓▓▓▓▓...zx	08/13/2018, 23:17:55	[172.16.213.14]:443 to 172.16.32.254:80 abnormal; cause: check time out	
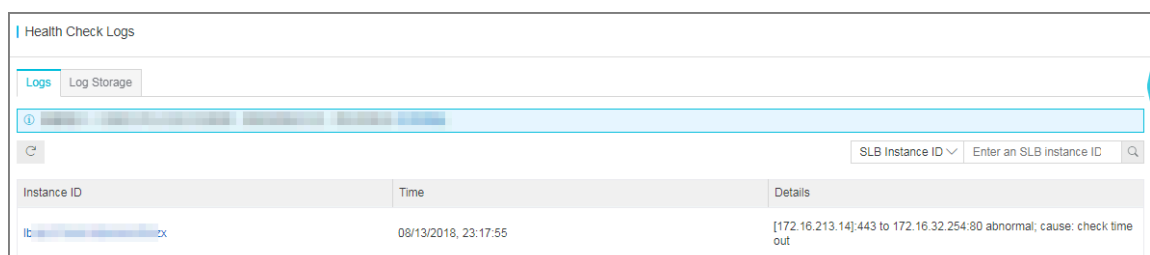
Download health check logs

You can download the completed health check logs stored in OSS.

1. Log on to the OSS console.

2. On the **Overview** page, click the target bucket and then click **Files**.

3. On the **Files** page, click *AliyunSLBHealthCheckLogs/*.
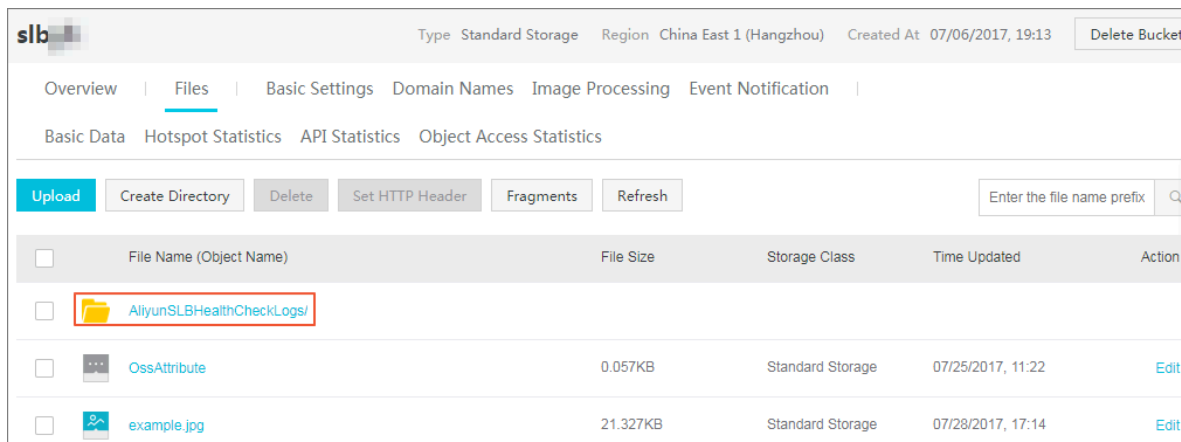
4. Click the folder of the heath logs to download.

5. Click **Edit** of the target folder. Then, click **Copy File URL** in the displayed page.

6. Enter the copied URL in the web browser to download the logs.

6.3 Authorize a RAM user to use access logs

Before a RAM user starts to use the access log function, the RAM user must be authorized by the primary account.

Prerequisites

The primary account has enabled the access log function

1. Log on to the RAM console with your primary account.

2. Click **Roles** to see whether the primary account has the **AliyunLogArchiveRole**.

If the primary account does not have this role, log on to the SLB console with the primary account, select **Logs** > **Access Logs**, click **Authorize**. In the displayed dialog box, click **Confirm Authorization Policy**.

> 📋 **Note:**
>
> This operation is required only at the first time.

Procedure

1. Create an authorization policy:

   a) Use the primary account to log on to the RAM console.

   b) In the left-side navigation pane, click **Policies**, and then click **Create Authorization Policy**.
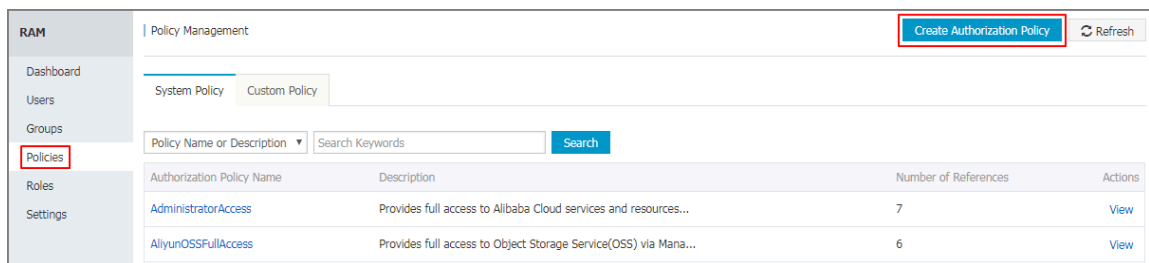
   c) Click **Blank Template**.
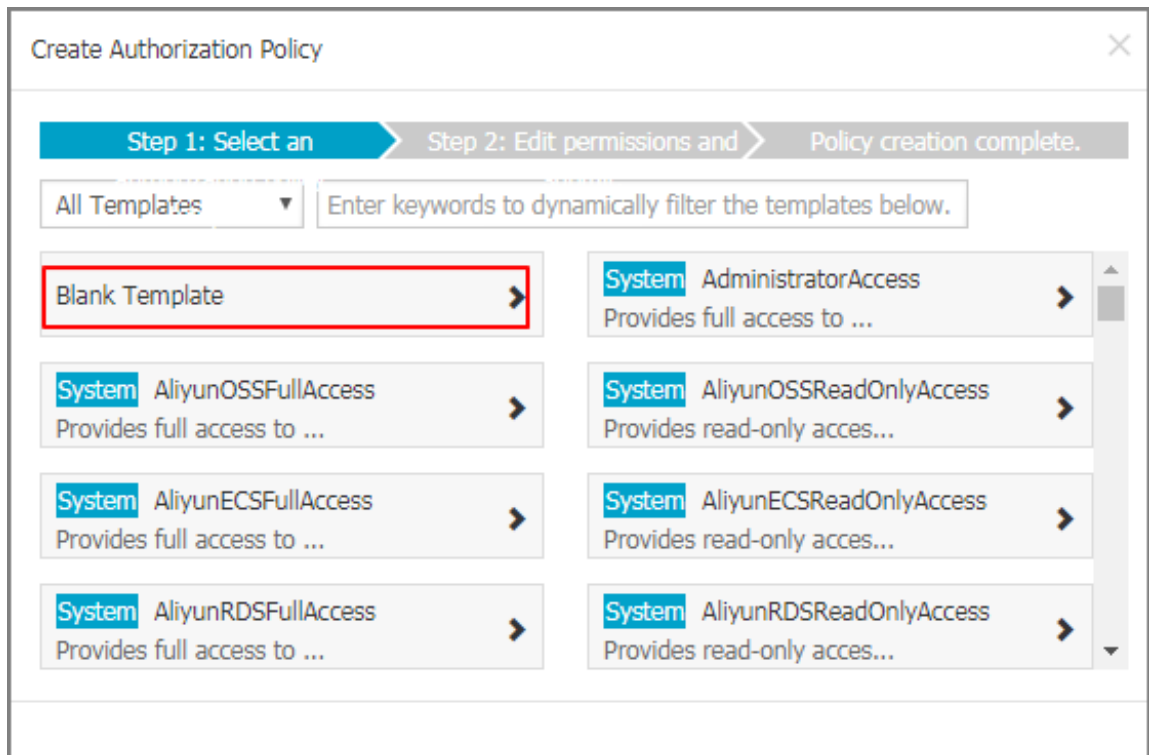
d) Enter a policy name, such as **SlbAccessLogPolicySet**, and then enter the following policy.

Click **Create Authorization Policy**.

```
{
"Statement ":[
 {
   "Action ":[
     "slb:Create*",
     "slb:List*"
   ],
   "Effect": "Allow",
   "Resource": "acs:log:*:*:project/*"
 },
 {
   "Action": [
     "log:Create*",
     "log:List*"
   ],
   "Effect": "Allow",
   "Resource": "acs:log:*:*:project/*"
 },
 {
   "Action": [
     "log:Create*",
     "log:List*",
     "log:Get*",
     "log:Update*"
   ],
   "Effect": "Allow",
   "Resource": "acs:log:*:*:project/*/logstore/*"
 },
 {
   "Action": [
     "log:Create*",
     "log:List*",
     "log:Get*",
     "log:Update*"
   ],
   "Effect": "Allow",
   "Resource": "acs:log:*:*:project/*/dashboard/*"
 },
 {
   "Action": "cms:QueryMetric*",
   "Resource": "*",
   "Effect": "Allow"
 },
 {
   "Action": [
     "slb:Describe*",
     "slb:DeleteAccessLogsDownloadAttribute",
     "slb:SetAccessLogsDownloadAttribute",
     "slb:DescribeAccessLogsDownloadAttribute"
   ],
   "Resource": "*",
   "Effect": "Allow"
 },
 {
   "Action": [
     "ram:Get*",
     "ram:ListRoles"
   ],
   "Effect": "Allow",
```

```
      "Resource": "*"
    }
  ],
  "Version": "1"
}
```

a) Click **Close**.

2. Attach the created policy to a RAM user:

   a) In the left-side navigation pane, click **Users**.

   b) Find the target user (the user who uses the SLB Access Log function) and click **Authorize**.
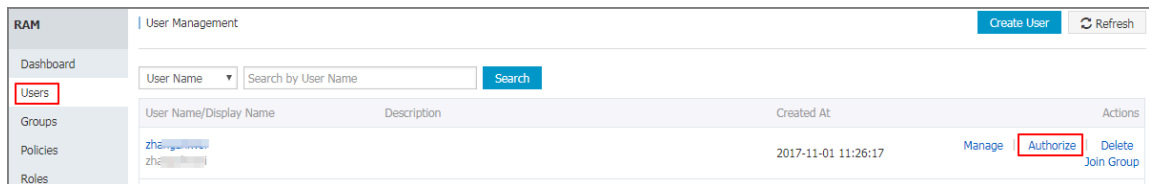
   c) Search the created authorization policy and attach the policy to the RAM user.

d) Click **OK**.

e) Go back to the User Details page to check if the policy has been attached to target RAM user.

6.4 Configure access logs

Integrated with Alibaba Cloud Log Service, you can analyze the access logs of a Server Load Balancer instance to understand the behavior and geographical distribution of client users and troubleshoot problems.

What are access logs?

The access logs collect detailed information of all requests sent to a Server Load Balancer instance, including the request time, client IP address, latency, request URL, server response, and so on. As the entry of Internet access, Server Load Balancer receives massive client requests. You can use access logs to analyze user behavior and geographical distribution, and troubleshoot.

After you enable the SLB access logs, you can store the access logs in the Logstore of SLS to collect and analyze the access logs. You can also disable access logs at any time.

There is no extra fee for Server Load Balancer access logs. You only need to pay for the Log Service.

> (!) **Notice:**
>
> - Only Layer-7 Server Load Balancer supports configuring access logs and this function is available in all regions now.
> - Make sure that the HTTP header value does not contain ||, otherwise, the exported logs may be misplaced.

Benefits

The following are benefits of Server Load Balancer access logs:

- Easy to use

  Free developers and maintenance staff from tedious and time-consuming log processing so that they can concentrate on business development and technical research.

- Cost-effective

  Access logs are typically massive. Processing access logs takes a lot of time and consumes a lot of resources. With Log Service, the access log processing is faster and cost-effective than self-build open-source solutions. Log Service can analyze one hundred million logs in one second.

- Real-time

Scenarios such as DevOps, monitoring, and alerting require real-time log data. Traditional data storage and analysis tools cannot meet this requirement. For example, it takes long time to ETL data to Hive where a lot of time is spent on data integration. Powered by its powerful big data computing capability, Log Service can process and analyze access logs in seconds.

- Flexible

You can enable or disable Server Load Balancer access logs according to the instance specification. Additionally, you can set the storage period (1 to 365 days) as needed and the Logstore's capacity is scalable to meet increasing service demands.

Configure access logs

Before configuring access logs, make sure:

1. A Layer-7 listener is added.
2. Log Service is activated.

To configure access logs, complete these steps.

1. Log on to the *SLB console*.
2. In the left-side navigation pane, click **Logs** > **Access Logs**.
3. Select a region.
4. Click **Authorize**, and then click **Confirm Authorization Policy** to authorize Server Load Balancer to write logs to Log Service.

   If you are using a RAM account, the primary account must perform authorization. For more information, see *Authorize a RAM user to use access logging*.

   > 📋 **Note:**
   > This operation is required only at the first time.
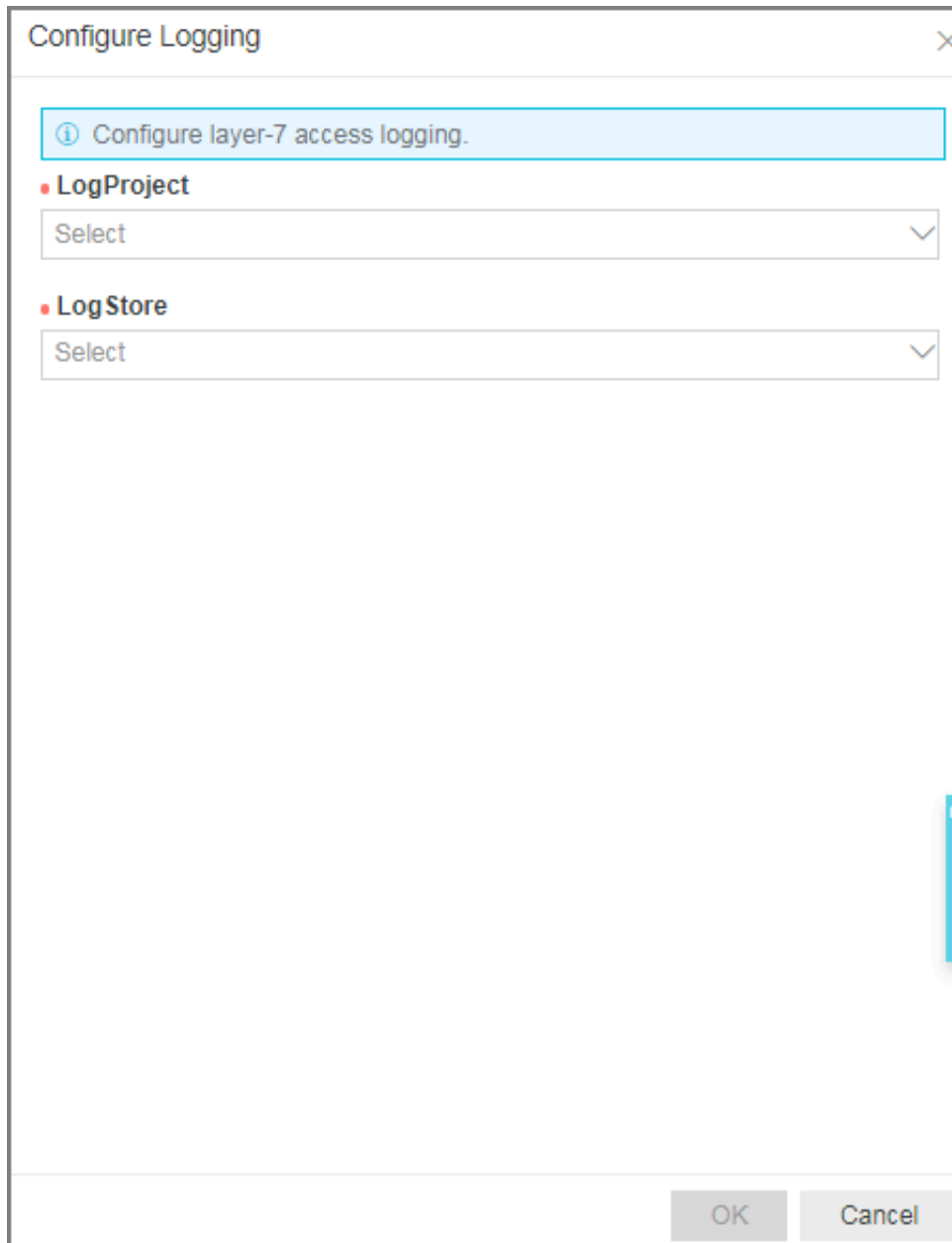
5. On the **Access Logs** page, find the target Server Load Balancer instance and click **Configure Logging**.
6. Select the LogProject and LogStore of Log Service and then click **OK**.

   If there is no available LogStore, click **Log Service console** to create log projects.

   > 📋 **Note:**
   > Make sure that the name of the LogProject is globally unique and the region of the LogProject is the same as that of the Server Load Balancer instance.

Configure Logging   ✕

ⓘ Configure layer-7 access logging.

● LogProject

| Select   ⌄ |

● LogStore

| Select   ⌄ |

OK   Cancel

Search and analyze access logs

After configuring Server Load Balancer access logs, you can search and view logs using the following indexing fields.

Field	Description
body_bytes_sent	The size of HTTP body (in byte) sent to the client.
client_ip	The client IP.
host	The host header in the request.
http_user_agent	The received http_user_agent header in the request.

Field	Description
request_length	The length of the request including startline, HTTP header and HTTP body.
request_method	The request method.
request_time	The interval between the time the Server Load Balancer receives the first request and the time the Server Load Balancer returns a response.
request_uri	The URI of the received request.
Slbid	The ID of the Server Load Balancer instance.
status	The status of the SLB response.
Upstream_addr	The IP address and port number of the backend server.
upstream_response_time	The interval between the time Server Load Balancer sends a request to the backend server and the time Server Load Balancer sends a response to the client.
upstream_status	The response status code of the backend server received by SLB.

Search access logs

To search access logs, complete these steps:

1. Go to the log search page. You can navigate to the search page from the Server Load Balancer console or the Log Service Console:

   • From the Server Load Balancer console:

     On the **Access Logs** page, click **View Logs**.

   • From the Log Service Console:

On the **Logstores** page, click **Search** of the target Logstore.
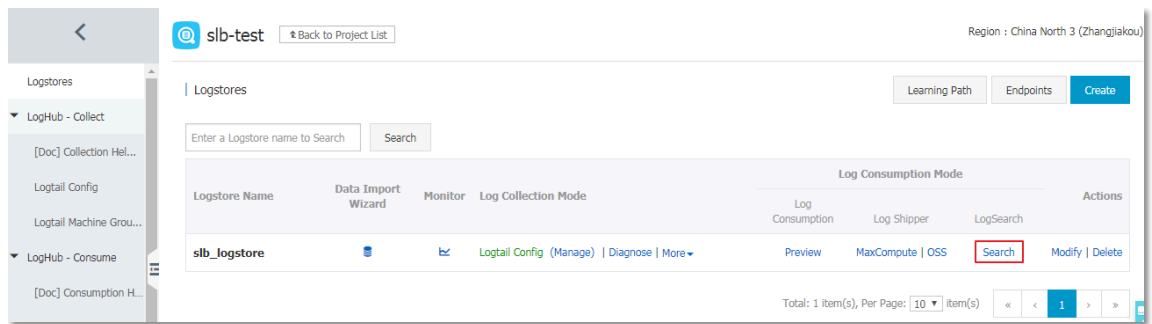
2. Click the target log field to view detailed information.

3. Enter an SQL statement to query access logs.

For example, enter the following SQL statement to query the Top20 clients, which is used for analyzing the request resource to assist business decision-making.

```
* | select ip_to_province(client_ip) as client_ip_province, count
(*) as pv group by
     client_ip_province order by pv desc limit 50
```
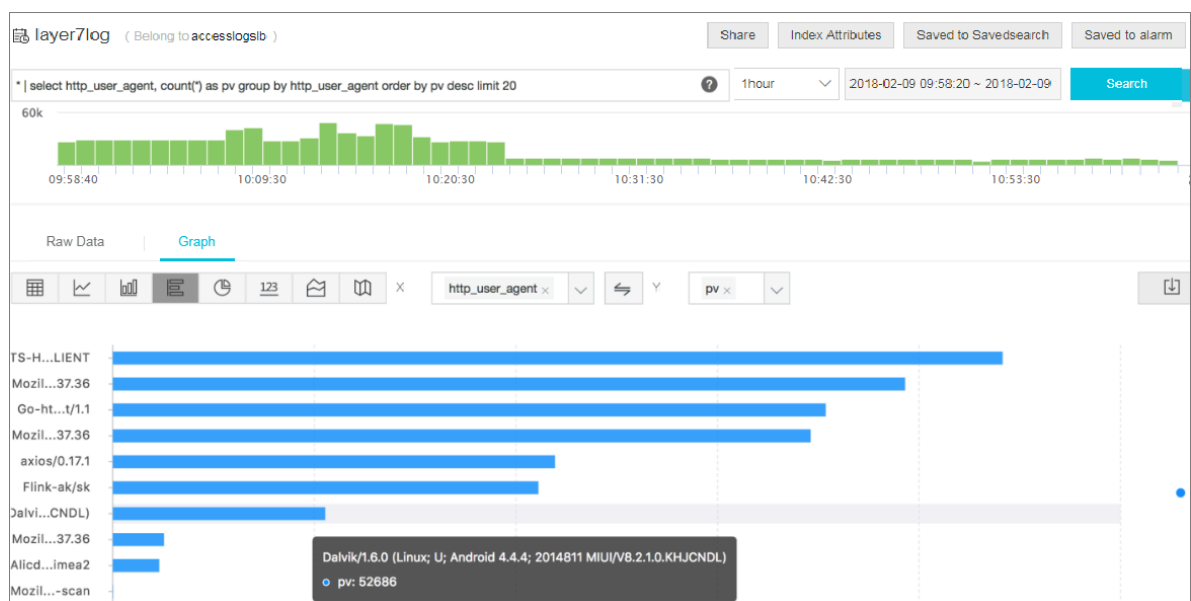

Analyze access logs

You can analyze access logs through the dashboard, which provides rich graphic information.

To analyze access logs, complete these steps:

1. On the Log Service console, click the project link of the SLB instance.

2. In the left-side navigation pane, click **Search/Analytics - Query** > **Dashboard**, and then click the name of the access log.
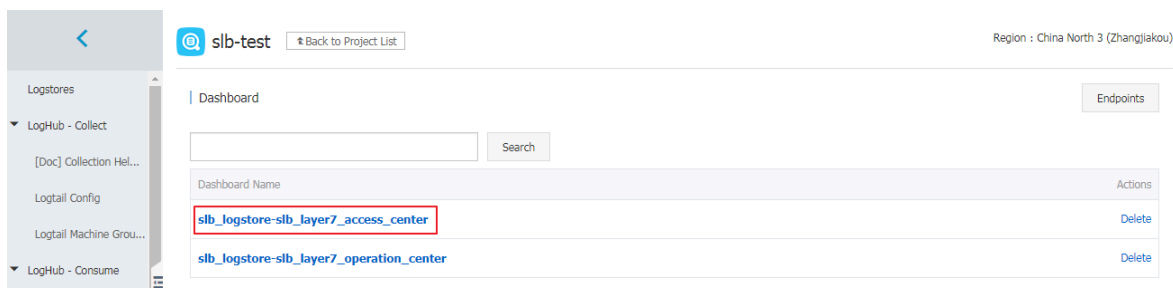
Disable access logging

To disable access logging, complete these steps:

1. Log on to the *SLB console*.

2. In the left-side navigation pane, click **Logs** > **Access Logs**.

3. Select a region.

4. On the **Access Logs** page, find the target instance and click **Delete** to disable access logging.

5. In the displayed dialog box, click **OK**.

7 Access control

7.1 Configure an access control list

Server Load Balancer (SLB) provides you with the access control function. You can configure different access control rules (access whitelist or blacklist) for different listeners. Before configuring access control for listeners, you must first configure an access control list.

You can create multiple access control lists. Each list contains multiple IP addresses or CIDR blocks. Limits on access control lists are as follows:

Resource	Limit
The maximum number of access control lists per region.	50
The maximum number of IP addresses added each time.	50
The maximum number of entries per access control list.	300
The maximum number of listeners that an access control list can be added to	50

Create an access control list

To create an access control list, complete these steps:

1. Log on to the *SLB console* .

2. Select a region.

3. In the left-side navigation pane, click the **Access Control** tab.

4. Click **Create Access Control List**, enter the access control list name, and click **OK**.

Add IP entries

To add IP entries, complete these steps:

1. Log on to the *SLB console*.

2. Select a region.

3. In the left-side navigation pane, click the **Access Control** tab.

4. Find the target access control list and click **Manage**.

5. Add IP entries:

- Click **Add Multiple Entries**. In the displayed dialog box, add IP addresses or CIDR blocks and click **OK**.

  Note the following when adding entries:

  - Each line is one entry. Use the Enter key to break lines.

  - Use "|" to separate an IP address or CIDR block with the description. For example, "192. 168.1.0/24|description".

- Click **Add Entry**. In the displayed dialog box, add an IP address or CIDR block and the description, and click **OK**.

Delete IP entries

To delete IP entries, complete these steps:

1. Log on to the *SLB console*.

2. Select a region.

3. In the left-side navigation pane, click the **Access Control** tab.

4. Find the target access control list and click **Manage**.

5. Click **Delete** in the **Actions** column of the target IP entry, or select multiple IP entries and click

   Delete at the bottom of the entry table.

6. In the displayed dialog box, click **OK**.

7.2 Configure access control

Server Load Balancer allows you to configure access control for listeners. You can configure different whitelists or blacklists for different listeners.

You can configure access control when you create a listener or change access control configuration after a listener is created.

This document introduces how to configure access control after a listener is created.

Enable access control

Before enabling access control, make sure:

- You have created an access control list. For more information, see *Configure an access control list*.

- You have created a listener.

To enable access control, complete these steps:

1. Log on to the *SLB console*.

2. Select a region.

3. Click the ID of the target SLB instance.

4. On the **Instance Details** page, click the **Listeners** tab.

5. Find the target listener, and then click **More** > **Set Access Control**.
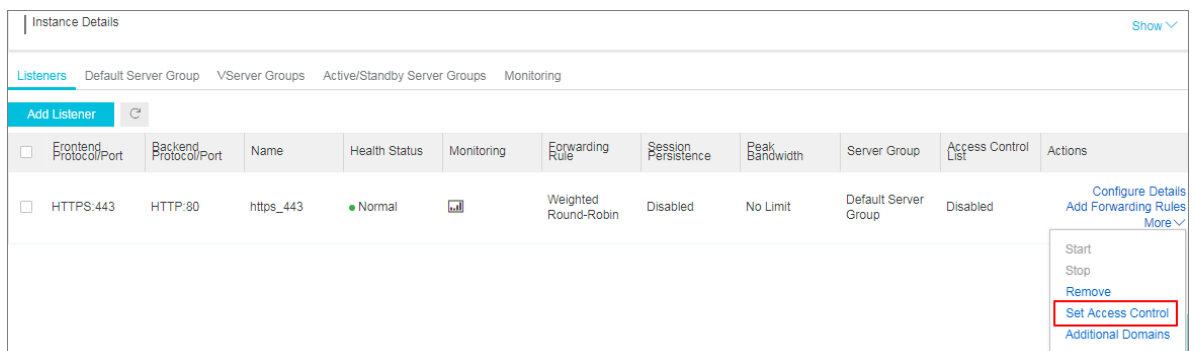
6. On the **Access Control Settings** page, enable access control, select the access control method and access control list, and click **OK**.

- **Whitelist**: Only requests from IP addresses or CIDR blocks in the selected access control lists are forwarded. It applies to scenarios where the application only allows access from some specific IP addresses.

Enabling whitelist poses some business risks. After a whitelist is configured, only the IP addresses in the list can access the listener. If you enable the whitelist without adding any IP entry in the corresponding access control list, all requests are forwarded.

- **Blacklist**: Requests from IP addresses or CIDR blocks in the selected access control lists are not forwarded. It applies to scenarios where the application only denies access from some specific IP addresses.

  If you enable a blacklist without adding any IP entry in the corresponding access control list, all requests are forwarded.

Disable access control

To disable access control, complete these steps:

1. Log on to the *SLB console*.

2. Select a region.

3. Click the ID of the target SLB instance.

4. On the **Instance Details** page, click the **Listeners** tab.

5. Find the target listener, and then click **More** > **Set Access Control**.

6. On the **Access Control Settings** page, disable access control and click **OK**.

7.3 Migrate to the new access control

If you have already configured a whitelist for a listener, Server Load Balancer can automatically add the IP addresses or CIDR blocks in the whitelist to an access control list and apply the list to the listener.

Migrate a whitelist to an access control list

To migrate a previously configured whitelist to an access control list, complete these steps:

1. Log on to the *SLB console*.

2. Select the region where the SLB instance is located, and then click the ID of the target SLB instance.

3. Click the **Listeners** tab.

4. Find the target listener, select **More** > **Set Access Control**.

5. Click **Use New Access Control Features**.

6. Enter a name of the access control list and click **Create Access Control List**.

7. Click **Apply** to apply the list to the listener as a whitelist.

> **Note:**
>
> If you do not apply the list to a listener, the whitelist does not take effect.

View the migrated access control list

To view the migrated access control list, complete these steps:

1. Log on to the *SLB console*.

2. Select a region.

3. In the left-side navigation pane, click **Access Control**.

4. Find the created access control list and view the associated listener. You can click **Manage** to
   manage IP entries.

7.4 Configure a whitelist

Whitelist is a method used to control the access of Server Load Balancer. It applies to scenarios
where an application only allows access from some specific IP addresses.

Context

> **Note:**
>
> SLB has released a new version of the access control function, allowing you to configure both
> whitelists and blacklists. You can migrate the previously configured whitelist to the new version.
> For more information, see *Migrate to the new access control*.

Note the following when configuring whitelists:

· Enabling whitelist poses some business risks. After a whitelist is configured, only the IP
   addresses in the list can access the listener.

· If you enable the whitelist function without adding any IP entry in the corresponding access
   control list, no requests are forwarded.

· When you configure a whitelist, the access to Server Load Balancer may be interrupted for a
   short time.

Procedure

1. Log on to the *SLB console*.

2. Select the region where the target SLB instance is located.

3. Click the ID of the target SLB instance.

4. In the **Listeners** tab, select **More** > **Set Access Control**.

5. In the displayed dialog box, configure as follows:

a) Click the **Enable Access Control** switch.

b) Enter the IP addresses that are allowed to access the listener.

Separate multiple IP addresses using comma. Up to 300 IP addresses are allowed. You can also enter CIDR blocks.

c) Click **OK**.

8 Monitoring

8.1 View monitoring data

With the CloudMonitor service, you can view the number of connections, and other traffic information of SLB listeners.

Procedure

1. Log on to the *SLB console*.

2. Select the region where the SLB instance is located.

3. Click the monitoring icon [icon] next to the target SLB instance.

4. Select the monitor metrics that you want to view.
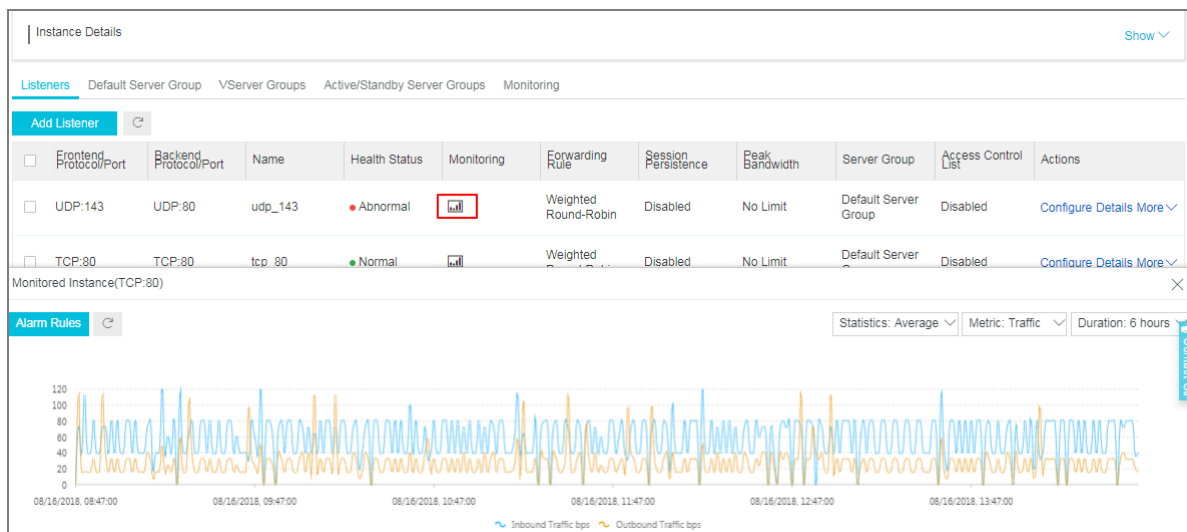
SLB supports viewing the following monitor metrics.

Monitor metrics	Description
Traffic	• Inbound Traffic: The traffic consumed by external access. • Outbound Traffic: The traffic consumed by Server Load Balancer.
Packets	• RX Packets Count: The number of request packets received per second. • TX Packets Count: The number of response packets sent per second.

Monitor metrics	Description
Concurrent Connections	• Active Connections Count: The number of established TCP connections. If persistent connections are used, a connection can transfer multiple file requests at one time. • Inactive Connections Count: The number of TCP connections that are not in the established status. You can use `netstat -an` command to view the active connections. • Max Concurrent Connections Count: The total number of TCP connections.
Average Connection Requests Count	The average number of new TCP connections established between clients and the Server Load Balancer in the statistical period.
Dropped Traffic	• Dropped Inbound Traffic: The amount of inbound traffic dropped per second. • Dropped Outbound Traffic: The amount of outbound traffic dropped per second.
Dropped Packets	• Dropped RX Packets: The number of inbound packets dropped per second. • Dropped TX Packets: The number of outbound packets dropped per second.
Dropped Connections	The number of TCP connections dropped per second.
The following metrics are specific to Layer-7 listeners.	
Layer-7 Protocol QPS	The number of HTTP/HTTPS requests that can be handled per second.
Response Time (Listener)	The average response time of Server Load Balancer.
HTTP Status Code 2XX/3XX/4XX/5XX/Others (Listener)	The average number of HTTP response codes generated by the listener.
Response Code 4xx/5xx (Server)	The average number of HTTP response codes generated by the backend server.
Response Time (Server)	The average response time of the backend server.

8.2 Configure alarm rules

After activating the CloudMonitor service, you can configure alarm rules for SLB instances on the CloudMonitor console.

Context

Note:

If a listener or an SLB instance is deleted, its alarm settings are deleted correspondingly.

Procedure
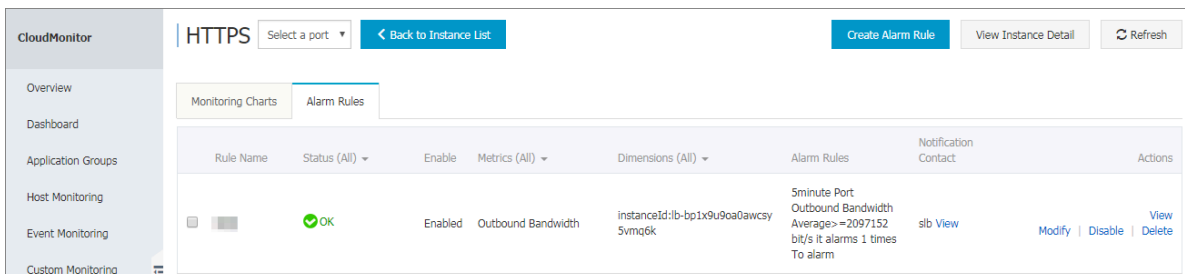
1. Log on to the *SLB console*.

2. Select the region where the SLB instance is located.

3. Find the target instance and click .

Notice:

Make sure the instance has configured with listeners and enabled health check.

4. Click **Alarm Rules**. You are then directed to CloudMonitor console.

5. Click **Create Alarm Rule**.

6. Configure the alarm rule.

1 **Related Resource**

Products : [Server Load Balancer ▾]

Resource
Range : [Instances ▾] ⃝ When selecting an application group, you can use an alarm template. Click View alarm template best practices.

Region : [China East 1 (Hangzhou) ▾]

Instances : [lb-bp1x9u9oa0awcsy5v... ▾]

2 **Set Alarm Rules**

Alarm
Rule : [   ]

Rule
Describe : [Number of Active Port ▾] [5mins ▾] [Average ▾] [>= ▾] [Thresho] unit

Port : AnyPort☑ [All   ]

+Add Alarm Rule

Mute for : [24h   ▾] ⃝

Triggered
when
threshold
is
exceeded
for : [1   ▾] ⃝
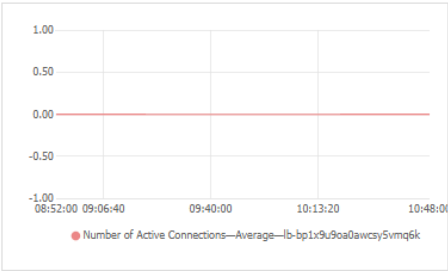
Effective
Period : [00:00   ▾] To: [23:59   ▾]

```
1.00

0.50

0.00

-0.50

-1.00
08:52:00 09:06:40      09:40:00      10:13:20      10:48:00
● Number of Active Connections—Average—lb-bp1x9u9oa0awcsy5vmq6k
```

9 API Inspector

API Inspector is an experimental feature. It enables you to view the API calls behind each step on the console, and automatically generates API codes of different languages. You can debug online through Cloud Shell or API Explorer.

Features

API Inspector, API Explorer, and Cloud Shell form an integrated solution for you to learn and debug API. It has the following features:

- Automatic recording: To obtain related API calls, you only need to perform operations on the console. For more information, see *Automatically record API calling*.

- Code generation with one click: API code scripts in different languages with pre-filled parameters are generated and can be run directly. For more information, see *Generate API codes with one click*.

- Online debugging: When API Inspector is used together with API Explorer and Cloud Shell, one-click online debugging can be implemented and you do not need to build the development environment. What you get is what you see. For more information, see *Debug online through API Explorer* and *Debug online through Cloud Shell*.

Enable API Inspector

To enable API Inspector, complete these steps:

1. Log on to the *SLB console*.

2. In the left-side navigation pane, select **SLB Lab** > **API Inspector**.

3. On the **API Inspector** page, enable API Inspector. Then the API suspended pendant is displayed on the right side of the page.
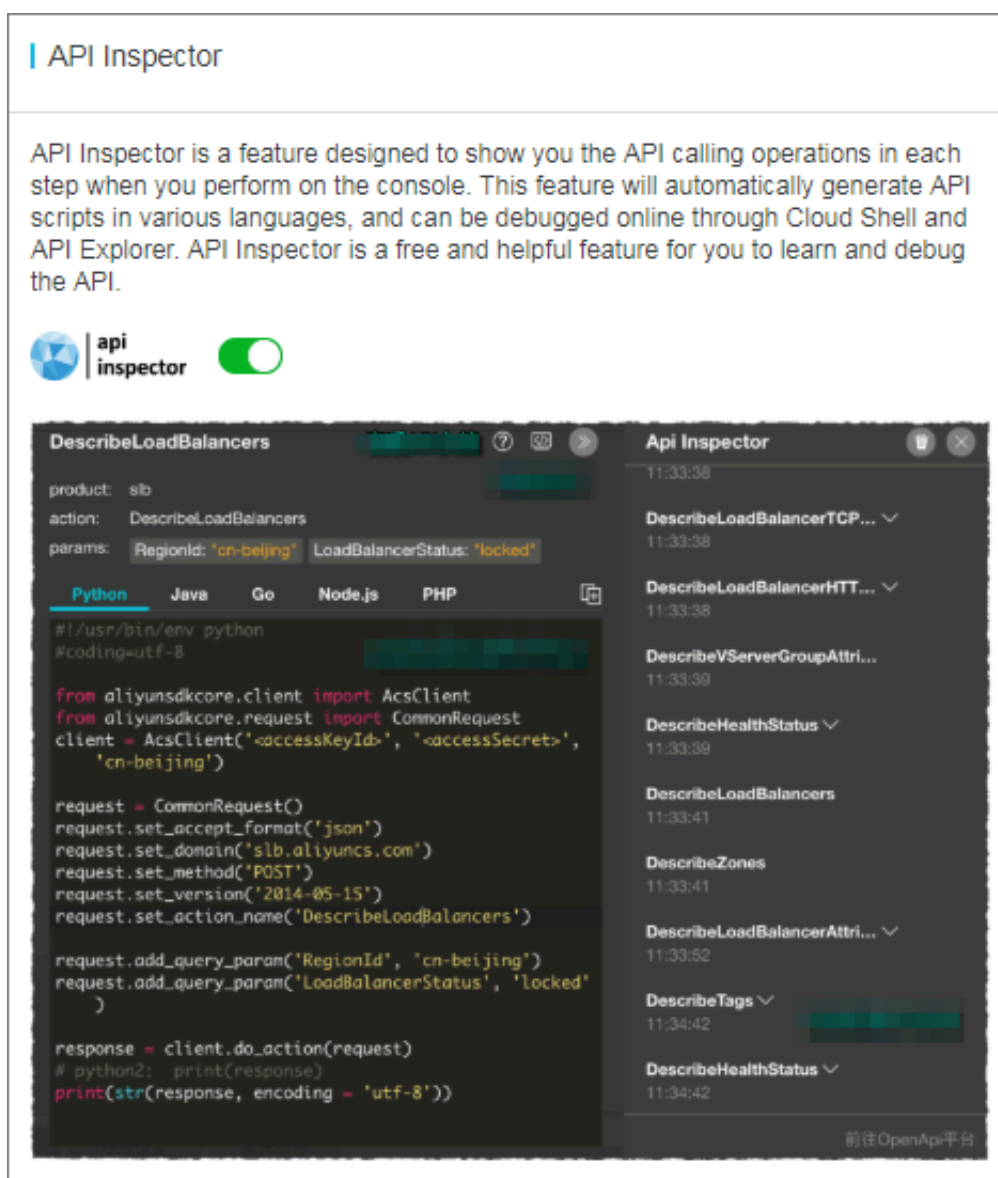
Automatically record API calling

Modifying the name of an SLB instance is taken as an example to demonstrate the automatic recording function of API Inspector.

1. Select **Instances** > **Server Load Balancer**.

2. Modify the name of an SLB instance to **SLB1**.

3. Click **OK**.

4. Click on the right side of the page. Then you can see all API calls related to the

    preceding operation.
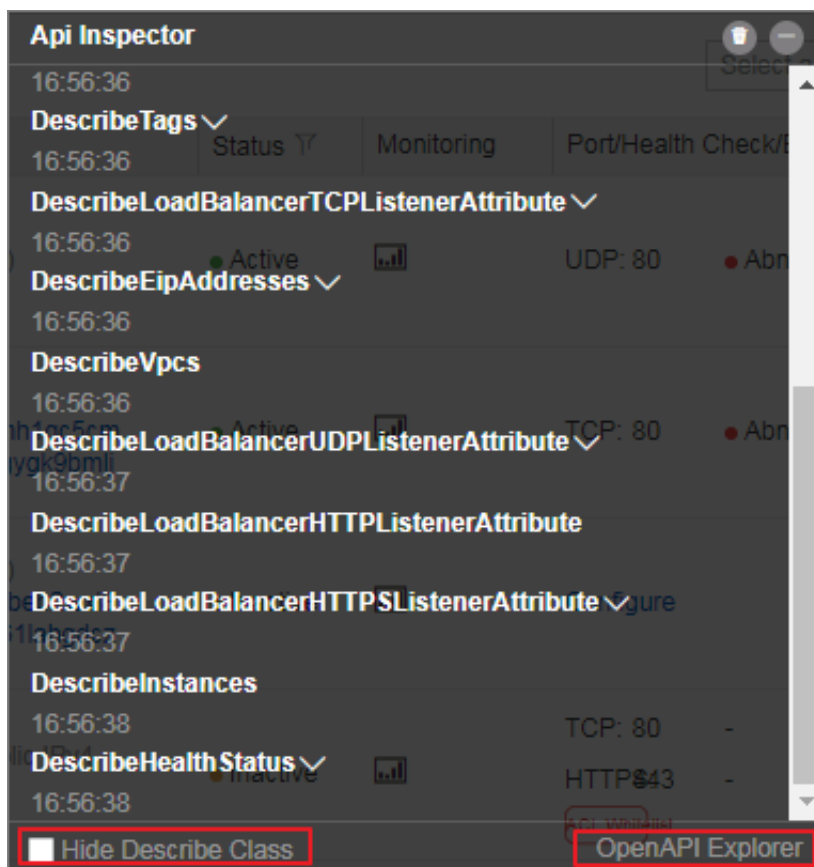
5. You can click **Hide Describe Class OpenAPI Explorer** to view core APIs. In the example, the
   core API is `SetLoadBalancerName`.
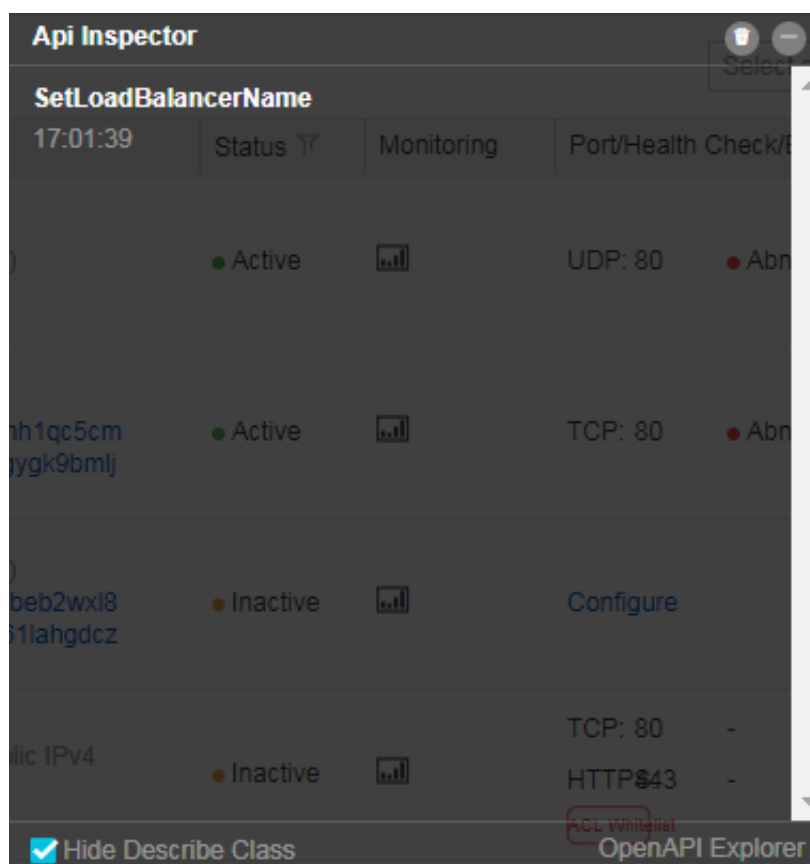
Generate API codes with one click

After API recording is completed, click the API name to generate API code scripts in Python, Java

, Go, Node.js, and PHP, with pre-filled parameters.

> **Note:**
>
> Click [icon] to copy the code scripts of the corresponding format, which can be run directly.
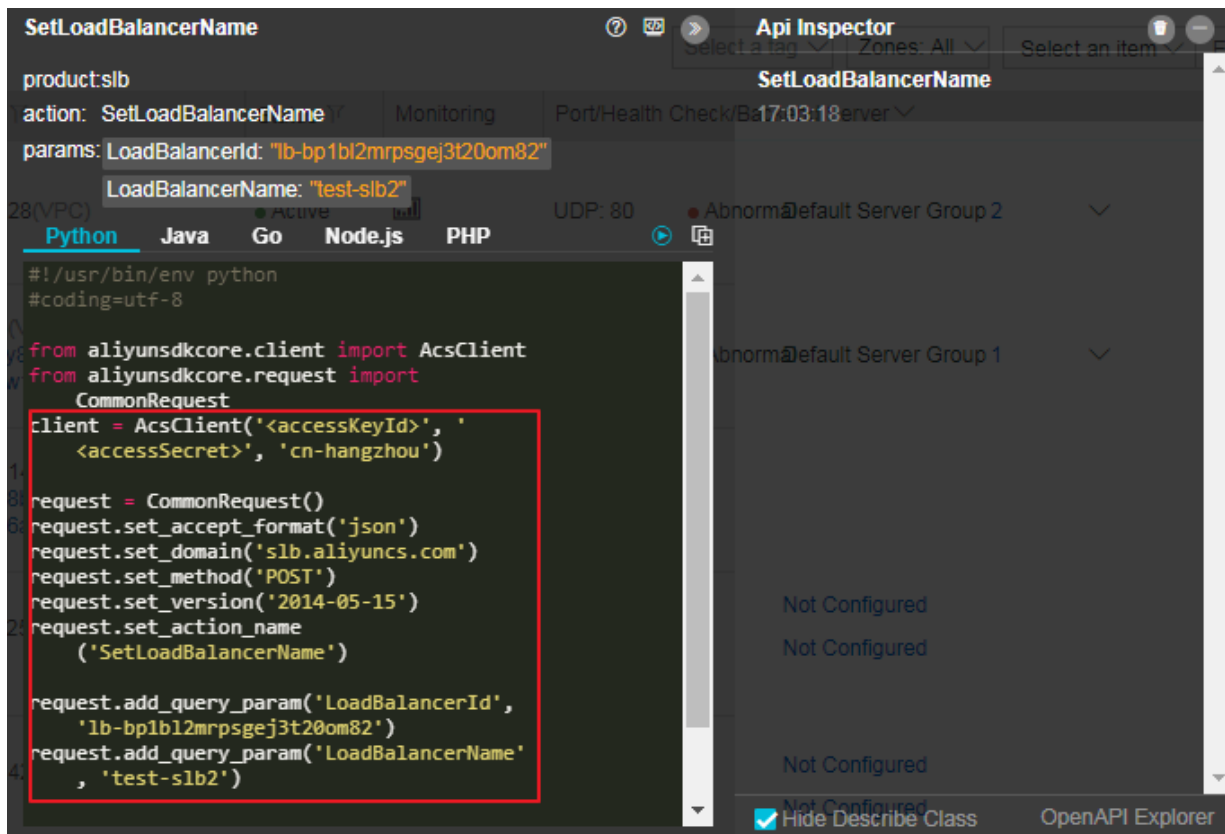
Debug online through API Explorer

After the API recording is completed, click **OpenAPI Explorer** or [icon] to go to the *OpenAPI*

Explorer console to debug the corresponding function. The API parameter values have been

automatically generated according to operations on the console.

Note:

Click to view the document describing the parameter details of the called API.

Debug online through Cloud Shell

After API recording, unfold the API calling details and click to use the online one-click

debugging of Cloud Shell.

Note:

If you use the one-click debugging of Cloud Shell, we recommend that you associate and create an OSS Bucket to store your frequently used scripts and files. However, a small amount of OSS costs will be generated. You can also choose to do not create an OSS Bucket.

The cloud command line format for the Cloud Shell debugging of SLB is as follows:
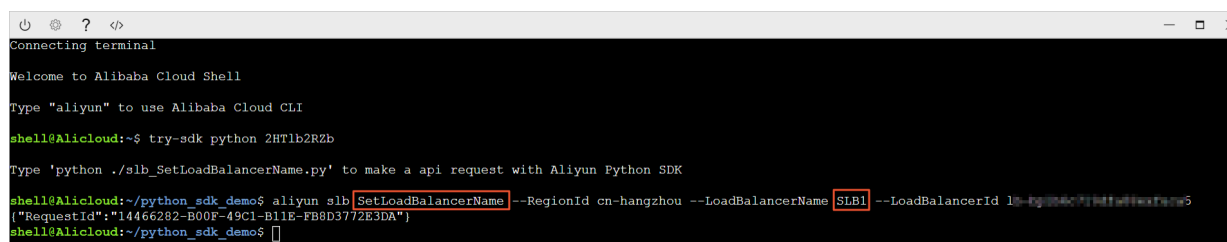
```
aliyun slb actionName --parameter1 value1 --paramter2 value2...
```

In this example, the called **SetLoadBalancerName** API modifies the name of the SLB instance to SLB1, so run the following cloud command line:

```
aliyun slb SetLoadBalancerName --RegionId cn-hangzhou --LoadBalanc
erName SLB1 --LoadBalancerId lb-bp1b6c719dfa08exfuca5
```

The returned value is:

```
{"RequestId":"14466282-B00F-49C1-B11E-FB8D3772E3DA"}
```

10 Multiple-zone deployment

You can create SLB instances in a region with multiple zones to improve the availability.

What is multiple-zone deployment?

A cloud product zone refers to a set of independent infrastructures. Different zones have independent infrastructures (such as network, power supply and air-conditioning), thus an infrastructure fault in one zone does not affect other zones.

To provide more reliable services, SLB has deployed multiple zones in most regions to achieve disaster recovery across data centers. When the data center in the primary zone is faulty and unavailable, SLB is able to switch to the data center in the backup zone to restore its service capabilities within 30 seconds.

Note the following about SLB primary/backup zones:

- SLB supports attaching ECS instances in different zones as long as the ECS instances and the SLB instance are in the same region. SLB can distribute traffic to the ECS instances in different zones.

- Normally, the SLB instance in the backup zone is in the standby state. You cannot manually switch the primary/backup state of an SLB instance. SLB switches to the backup zone only when the whole primary zone is unavailable due to data center outage or exit cable failure and more. SLB will not switch to the backup zone if only one instance in the primary zone is faulty.

- SLB and ECS instances are deployed in different clusters. When an SLB instance in Zone A is unavailable, the ECS instances in Zone A are not necessarily unavailable. Therefore, after SLB switches to the backup zone, the SLB instance in the backup zone still can distribute traffic to the added ECS instances. However, if all clusters in a zone is unavailable or the optical cable is faulty, all services (including but not limited to SLB instances and ECS instances) in the zone cannot work anymore.

For more information, see *SLB high availability*.
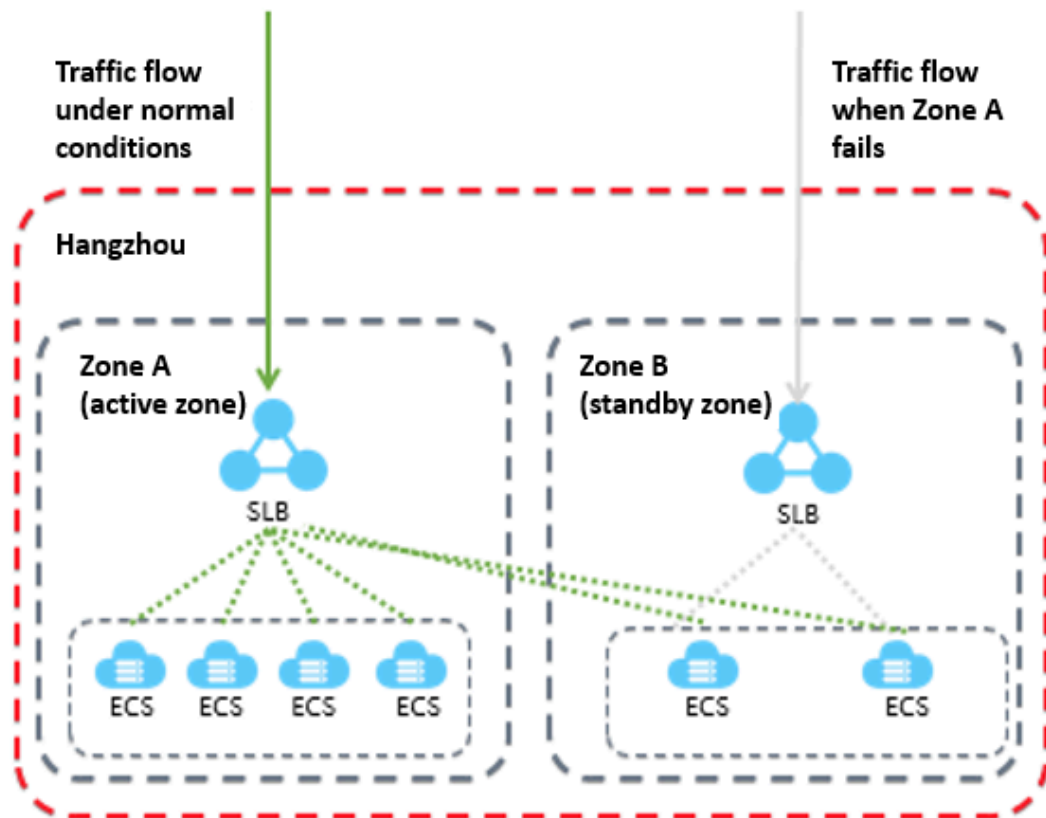
Primary/backup zone list

The following table lists the primary/backup zones in different regions. You can call the

`DescribeZones` API to obtain available primary/backup zones in a region.

Region	Zone type	Zones	
China (Hangzhou)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone B	Zone D Zone G
		Zone D	Zone E
		Zone E	Zone D Zone F
		Zone F	Zone E
		Zone G	Zone B Zone H
		Zone H	Zone G
China (Shanghai)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone B

Region	Zone type	Zones	
		Zone B	Zone A Zone C Zone D
		Zone C	Zone B
		Zone D	Zone B Zone E
		Zone E	Zone D Zone F
		Zone F	Zone E
China (Shenzhen)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone B
		Zone B	Zone A Zone C
		Zone C	Zone B Zone D
		Zone D	Zone C Zone E
		Zone E	Zone D
China (Qingdao)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone B	Zone C
		Zone C	Zone B
China (Beijing)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone B Zone D Zone E
		Zone B	Zone C
		Zone C	Zone E
		Zone D	Zone A
		Zone E	Zone C Zone F
		Zone F	Zone E Zone G
		Zone G	Zone F

Region	Zone type	Zones	
China (Zhangjiakou)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone B
		Zone B	Zone A
China (Hohhot)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone B
		Zone B	Zone A
Germany (Frankfurt)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone B
		Zone B	Zone A
UK (London)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone A
		Zone B	Zone B
UAE (Dubai)	Single-zone	Zone A	
Singapore	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone B
		Zone B	Zone A
		Zone C	Zone B
Australia (Sydney)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone B
		Zone B	Zone A
Malaysia (Kuala Lumpur)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone B
		Zone B	Zone A
Indonesia (Jakarta)	Single-zone	Zone A	
India (Mumbai)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone B
		Zone B	Zone A
Japan (Tokyo)	Single-zone	Zone A	

Region	Zone type	Zones	
Hong Kong	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone B	Zone C
		Zone C	Zone B
US (Virginia)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone B
		Zone B	Zone A
US (Silicon Valley)	Multiple-zone	**Primary zone**	**Available backup zones**
		Zone A	Zone B
		Zone B	Zone A

11 Achieve cross-region load balancing through Global Traffic Manager

Global traffic management

Load balancing is divided into local load balancing and global load balancing according to the geographical structure of its application. Local load balancing balances loads of server groups in the same region. Global load balancing balances server groups that are in different regions and have different network structures.

By using Global Traffic Manager, you can deploy global traffic management above the local traffic balancing to achieve cross-region disaster tolerance, accelerate access from different regions and achieve intelligent resolution.

- Multi-line intelligent resolution service

    Using DNS intelligent resolution and health check on the running status of the application, global traffic management directs user accesses to the most appropriate IP addresses so that the users can obtain the fastest and smoothest experience.

- Cross-region disaster tolerance

    Global traffic management supports adding IP addresses of different regions to different address pools and configuring health check. In access policy configurations, set the **address pool A** as the default IP address pool and **address pool B** as the failover IP address pool. Then active-standby IP disaster tolerance of the application service can be achieved.

- Accelerate accesses from different regions

    By using Global Traffic Manager, you can direct user accesses from different regions to different IP address pools, thus achieving grouped user management and grouped access and helping the application improve user experience.

Deploy Global Traffic Manager

This tutorial takes aliyuntest.club as an example (most users of the website are in Singapore and China) to show how to achieve global load balancing through global traffic management and load balancing.

Step 1 Purchase and configure ECS instances

Purchase and configure at least two ECS instances in each region where the users of the application service are located.

In this tutorial, two ECS instances are purchased in Beijing, Shenzhen and Singapore separately, and a simple static web page is built on each ECS instance.

Step 2 Purchase and configure Server Load Balancing instances

1. See *Create an SLB instance* to create three Internet Server Load Balancing instances in Beijing, Shenzhen and Singapore separately.

2. See *Configure an SLB instance* to add listeners and add configured ECS instances to backend server pools.

- Example of the SLB instance in the Beijing region

- Example of the SLB instance in the Shenzhen region

- Example of the SLB instance in the Singapore region

Step 3 Configure Global Traffic Manager

1. Purchase a Global Traffic Manager instance.

   a. Log on to the *Alibaba Cloud DNS console*.

   b. In the left-side navigation pane, click **Global Traffic Manager**.

   c. On the **Global Traffic Manager** page, click **Create Instance**.

   d. Select the version, quantity and service time.

   e. Click **Buy Now**.

   After the instance is successfully purchased, the system automatically allocates a CNAME access domain name.
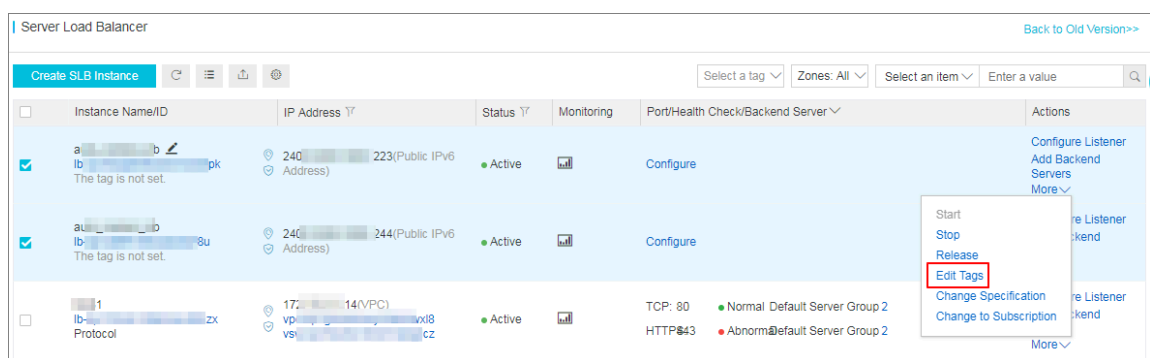
2. Configure the Global Traffic Manager instance.

   a. On the **Global Traffic Manager** page, click the ID of the Global Traffic Manager instance or click **Configure** in the **Actions** column.

   b. In the left-side navigation pane, click **Configurations**.

   c. In the **Global Settings** tab, click **Edit** to configure the parameters of the Global Traffic Manager instance.

Configure the following parameters and use the default values for the remaining options.

- Instance Name: Used for identifying the instance used for a certain application and can be customized.

- Primary Domain: The primary domain name is used by you to access the application. In this tutorial, enter aliyuntest.club.

- Alert Group: Select a contact group you configured in CloudMonitor. When an error occurs, the contact group is notified.

    d. Click **Confirm**.

3. Configure the IP address pool.

    a. In the **Address Pool Configurations** tab, click **Create Address Pool**.

    b. On the **Create Address Pool** page, configure the IP address pool.

In this tutorial, three IP address pools are to be added and each IP address pool accommodates one of the three SLB addresses in different regions.

- Address Pool Name: Custom. For example, China North_Beijing, China East_Hangzhou , Singapore.

- Address: The SLB public IP address to be added to this region.

c. Click **Confirm**.

4. Configure health check.

In this operation, you need to configure health check for the three address pools separately.

a. In the **Address Pool** tab, click **Add** next to health check.

b. Configure health check parameters.

Monitoring Node shows the locations of monitoring nodes. Select the monitoring node according to the region of the address pool.

5. Configure the access policy.

In this tutorial, add different access policies for the three different regions.

    a. In the **Access Policy** tab, click **Add Access Policy**.

    b. On the **Add Access Policy** tab, configure the access policy.

- Configure the corresponding default address pools for different access regions, and set an address pool of another region as the failover address pool.

- Select the access region. When users in this region access the application, the address pool configured in the access policy is matched.

  There must be an access policy with **Global** selected. Otherwise some areas cannot access the application.

6. Configure CNAME access.

    a. Log on to the Alibaba Cloud DNS console.

    b. Find the domain name aliyuntest.club and click **Configure** in the **Actions** column.

    c. On the **DNS Settings** page, click **Add Record**.

    d. On the **Add Record** page, direct the domain name aliyuntest.club accessed by end users to the alias record of the Global Traffic Manager instance in the form of CNAME.

    e. Click **Confirm**.

Step 4 Test

Remove the ECS instances of the SLB instance in the Beijing region so as that the SLB service is
 unavailable.

Visit the website to see if the access is normal.

> 📋 **Note:**
>
> It takes one to two minutes for Global Traffic Manager to make judgment after it detects that your
>
> IP is down. If you set the monitoring frequency to 1 minute, it takes two to three minutes for the
>
> link switching caused by exceptions to take effect.

12 Anti-DDoS Basic

You can view Alibaba Cloud Security thresholds of an Internet SLB instance on the SLB console.

Introduction to Anti-DDoS Basic

Alibaba Cloud provides up to 5 Gbps Anti-DDoS Basic for SLB. As shown in the following figure, all traffic from the Internet must first go through Alibaba Cloud Security before arriving at SLB. Anti-DDoS Basic scrubs and filters common DDoS attacks and protects your services against attacks such as SYN flood, UDP flood, ACK flood, ICMP flood, and DNS Query flood.
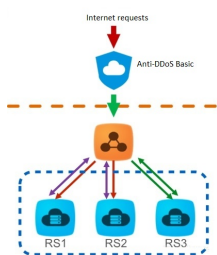
Anti-DDoS Basic sets the scrubbing threshold and blackholing threshold according to the bandwidth of the Internet SLB instance. When the inbound traffic reaches the threshold, scrubbing or blackholing is triggered:

- Scrubbing: When the attack traffic from the Internet exceeds the scrubbing threshold or matches certain attack traffic model, Alibaba Cloud Security starts scrubbing the attack traffic. The scrubbing includes packet filtration, traffic speed limitation, packet speed limitation and more.

- Blackholing: When the attack traffic from the Internet exceeds the blackholing threshold, blackholing is triggered and all inbound traffic is dropped.

The thresholds are calculated based on the following principles:

- The thresholds are determined by the bandwidth of the SLB instance, that is, the outbound bandwidth of the SLB instance. The thresholds are high when the bandwidth of the instance is high and vise versa.

- The blackholing threshold is determined by the security credit score of the user.

> **Note:**
> The security credit score only influences the blackholing threshold and does not influence the scrubbing threshold.

Complete these steps to calculate the threshold:

1. The SLB backstage provides the recommended threshold value that can ensure normal running of the instance according to the purchased bandwidth.

> **Note:**
> The outbound bandwidth of a Pay-As-You-Go instance is the peak bandwidth in the region. Currently the peak bandwidth in Mainland China is 5G. For more information, see *Peak bandwidths in different regions*.

- The relationship between SLB bandwidth and traffic scrubbing threshold (bits/s)

    - When the SLB bandwidth < 100 Mbps, the default traffic scrubbing threshold (bits/s) = 120 Mbps

    - When the SLB bandwidth > 100 Mbps, the default traffic scrubbing threshold (bits/s) = bandwidth*1.2

- The relationship between SLB bandwidth and traffic scrubbing threshold (packets/s)

    Traffic scrubbing threshold (packets/s) = (SLB bandwidth/500) * 150000

    The bandwidth is in Mbps.

- The relationship between SLB bandwidth and blackholing threshold (bits/s)

    - When the SLB bandwidth < 1 Gbps, the default blackholing threshold (bits/s) = 2 Gbps

    - When the SLB bandwidth> 1 Gbps, the default blackholing threshold (bits/s) = MAX (SLB bandwidth*1.5, 2 G)

2. Alibaba Cloud Security calculates the threshold according to the recommended value, the security credit score and the resource conditions in different regions.

- Rules for determining the traffic scrubbing threshold (bits/s) and the traffic scrubbing threshold (packets/s)

    The minimum traffic scrubbing threshold (bits/s) is 1,000 M and the minimum traffic scrubbing threshold (packets/s) is 300,000.

    - If the threshold recommended by SLB is lower than the minimum cleaning threshold, the minimum threshold is used.

    - If the threshold recommended by SLB is higher than the minimum cleaning threshold, the recommended threshold is used.

- Alibaba Cloud Security determines the blackholing threshold according to the security credit score of the user.

View thresholds

You can view the thresholds of an instance on the SLB console as a RAM user. If not, you must authorize the RAM account first. For more information, see *Allow read-only access to Anti-DDoS Basic*.

To view thresholds, complete these steps:

1. Log on to the *SLB console*.

2. Select a region.

3. Hover the mouse pointer to the DDoS icon next to the target instance to view the following thresholds. You can click the link to go to the DDoS console to view more information.

   • Traffic Scrubbing Threshold (bits/s): When the inbound traffic exceeds this value, scrubbing is triggered.

   • Traffic Scrubbing Threshold (packets/s): When the inbound packets exceed this value, scrubbing is triggered.

   • Blackholing Threshold: When the inbound traffic exceeds this value, blackholing is triggered.
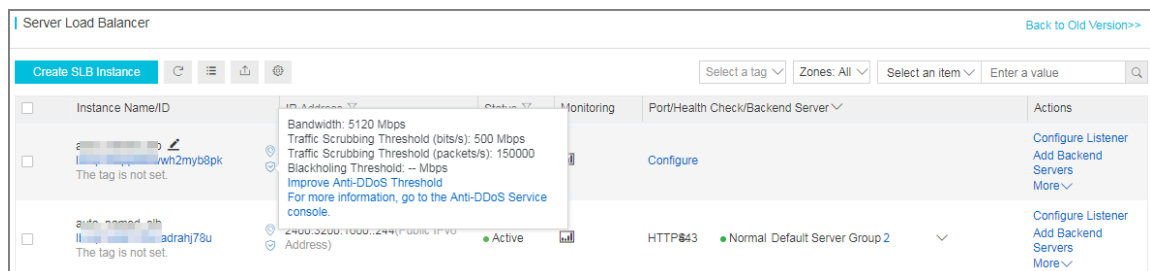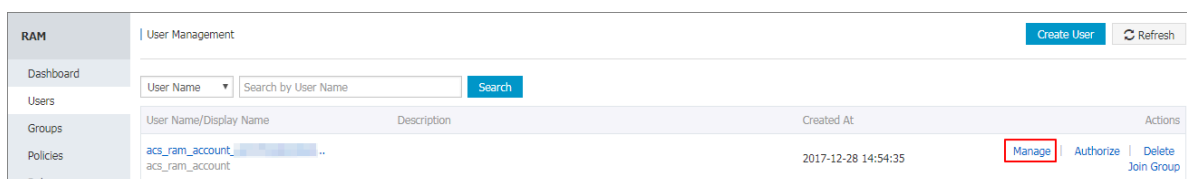

Allow read-only access to Anti-DDoS Basic

To allow read-only access to Anti-DDoS Basic, complete these steps:

> **Note:**
> Use the primary account to complete the authorization.

1. Use the primary account to log on to the RAM console.

2. In the left-side navigation pane, click **Users**, find the target RAM user and click **Manage**.

3. Click **User Authorization Policies**, and then click **Edit Authorization Policy**.

4. In the displayed dialog box, search **AliyunYundunDDosReadOnlyAccess**, and then add it to the Selected Authorization Policy Name list. Click **OK**.
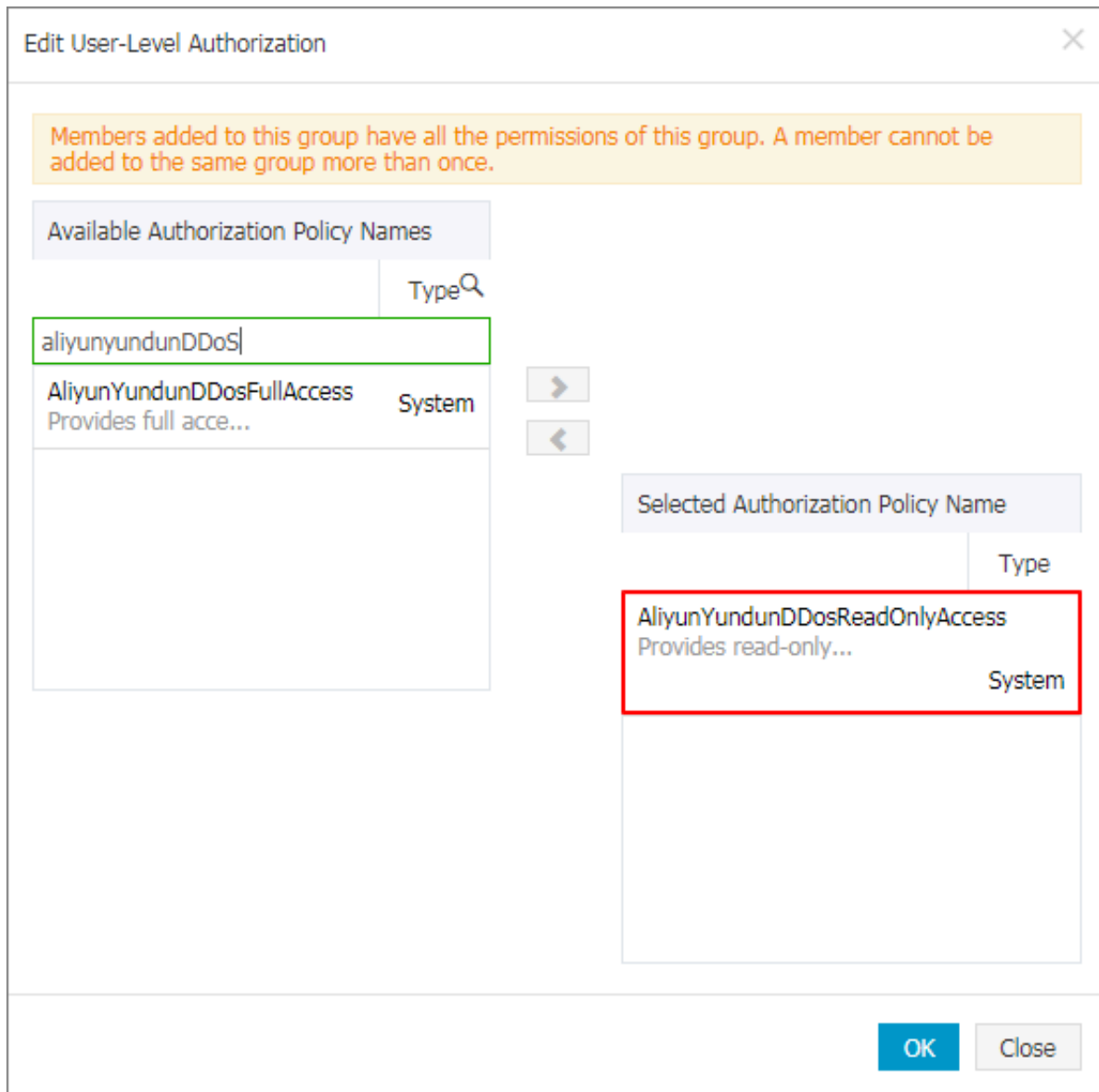
View the security credit score

The security credit score is provided by Alibaba Cloud based on your attack history, purchase history, account activity, security level, expectation and more. With a higher security credit score, you can have a higher free blackholing threshold and a shorter blackholing duration (how long the blackholing status lasts).

Complete these steps to view the security credit score:

1. Log on to the *Anti-DDoS Basic console*.

2. Select **Anti-DDoS Basic** > **Instances**.

3. Click the **Security Credibility** link to view the security credit score of the account.

Note:

Security credit scores are region-based.

Security Credit Details   ✕

Inspect the data and improve your security credit.

The system updates the following statistics daily, but only statistics updated by the end of the previous day are displayed.

| Attack History | Purchase History | Account Activity | Service Compliance | Security Levels |

Your DDoS attack history contributes to your security credit.

Attack Duration of Last 30 Days:-Hour(s)

Blackholing Events of Last 30 Days:-Times

See Alibaba Cloud Anti-DDoS Service Best Practices

Security Credit Score Trend for the Latest 30 Days