

Alibaba Cloud Server Load Balancer

Instance

Issue: 20190816

Legal disclaimer

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

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.
2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company, or individual in any form or by any means without the prior written consent of Alibaba Cloud.
3. The content of this document may be changed due to product version upgrades, adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.
4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults" and "as available" basis. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, special, or punitive damages, including lost profits arising from the use

or trust in this document, even if Alibaba Cloud has been notified of the possibility of such a loss.

5. By law, all the content of the Alibaba Cloud website, including but not limited to works, products, images, archives, information, materials, website architecture, website graphic layout, and webpage design, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of the Alibaba Cloud website, product programs, or content shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates).
6. Please contact Alibaba Cloud directly if you discover any errors in this document.

Generic conventions

Table -1: Style conventions

Style	Description	Example
	This warning information indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	 Danger: Resetting will result in the loss of user configuration data.
	This warning information indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	 Warning: Restarting will cause business interruption. About 10 minutes are required to restore business.
	This indicates warning information, supplementary instructions, and other content that the user must understand.	 Notice: Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructions, best practices, tips, and other content that is good to know for the user.	 Note: You can use Ctrl + A to select all files.
>	Multi-level menu cascade.	Settings > Network > Set network type
Bold	It is used for buttons, menus, page names, and other UI elements.	Click OK.
Courier font	It is used for commands.	Run the <code>cd / d C :/ windows</code> command to enter the Windows system folder.
<i>Italics</i>	It is used for parameters and variables.	<code>bae log list --instanceid Instance_ID</code>
[] or [a b]	It indicates that it is an optional value, and only one item can be selected.	<code>ipconfig [-all -t]</code>

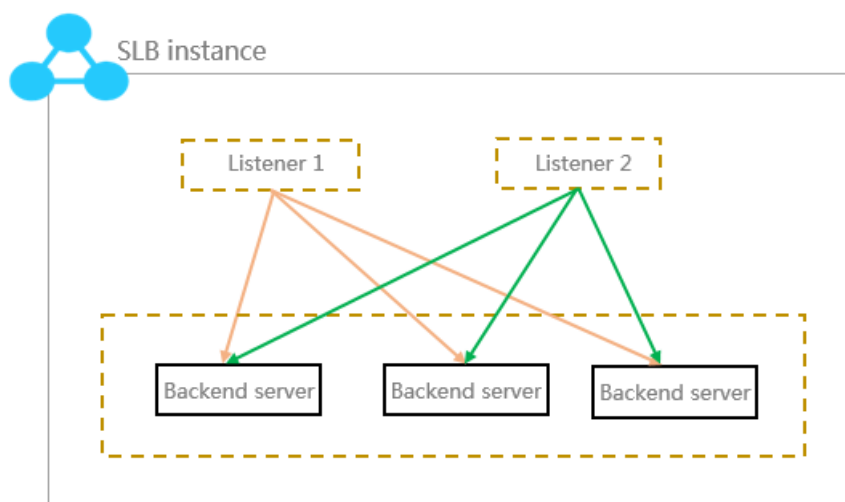
Style	Description	Example
<code>{}</code> or <code>{a b}</code>	It indicates that it is a required value, and only one item can be selected.	<code>swich {stand slave}</code>

Contents

Legal disclaimer.....	I
Generic conventions.....	I
1 SLB instance overview.....	1
2 Network traffic flow.....	4
3 Create an SLB instance.....	7
4 Create an IPv6 instance.....	9
5 Start or stop an SLB instance.....	12
6 Bind an EIP.....	13
7 Release an SLB instance.....	14
8 Tags.....	16
8.1 Overview.....	16
8.2 Add a tag.....	16
8.3 Search for SLB instances by using a tag.....	17
8.4 Delete a tag.....	17
9 Expiring Instances.....	18
10 Change the specification of an SLB instance.....	19
10.1 Overview.....	19
10.2 Change the specification of an SLB instance.....	19
11 Manage idle instances.....	20
12 Health checks of SLB instances.....	22
13 FAQ.....	23
13.1 SLB instance FAQ.....	23
13.2 Guaranteed-performance instance FAQ.....	24

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.

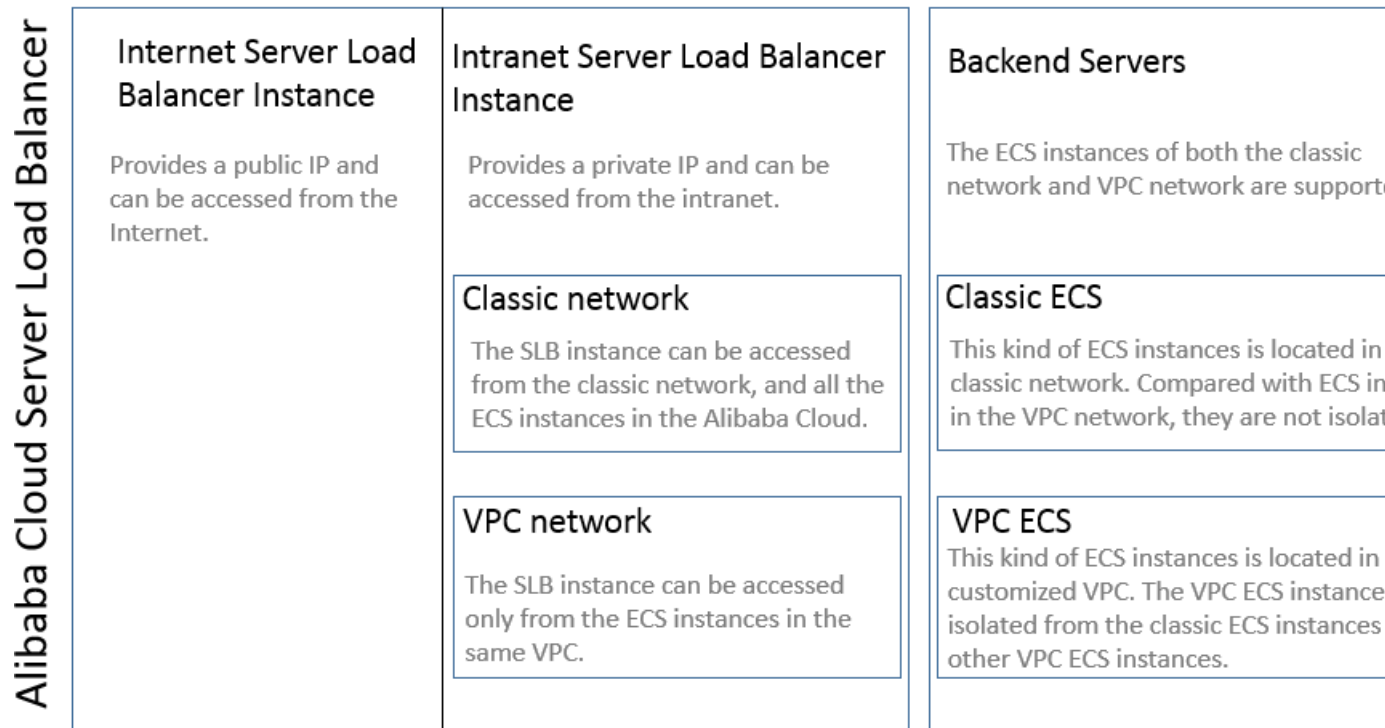


Alibaba Cloud provides Internet SLB service and intranet SLB service. A public or a private IP address is allocated to the SLB instance according to the instance type you select.

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:

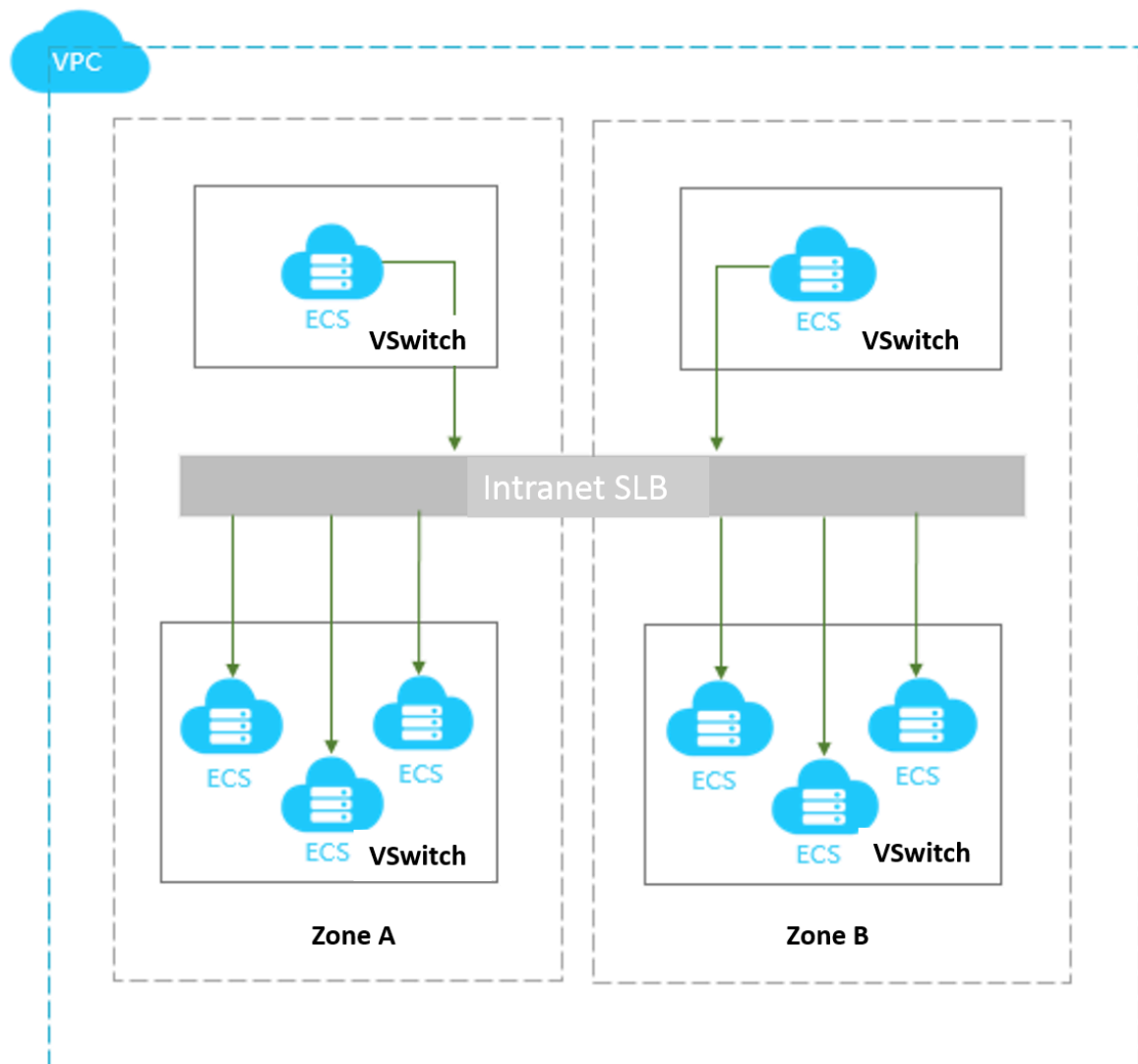
- Classic network

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.



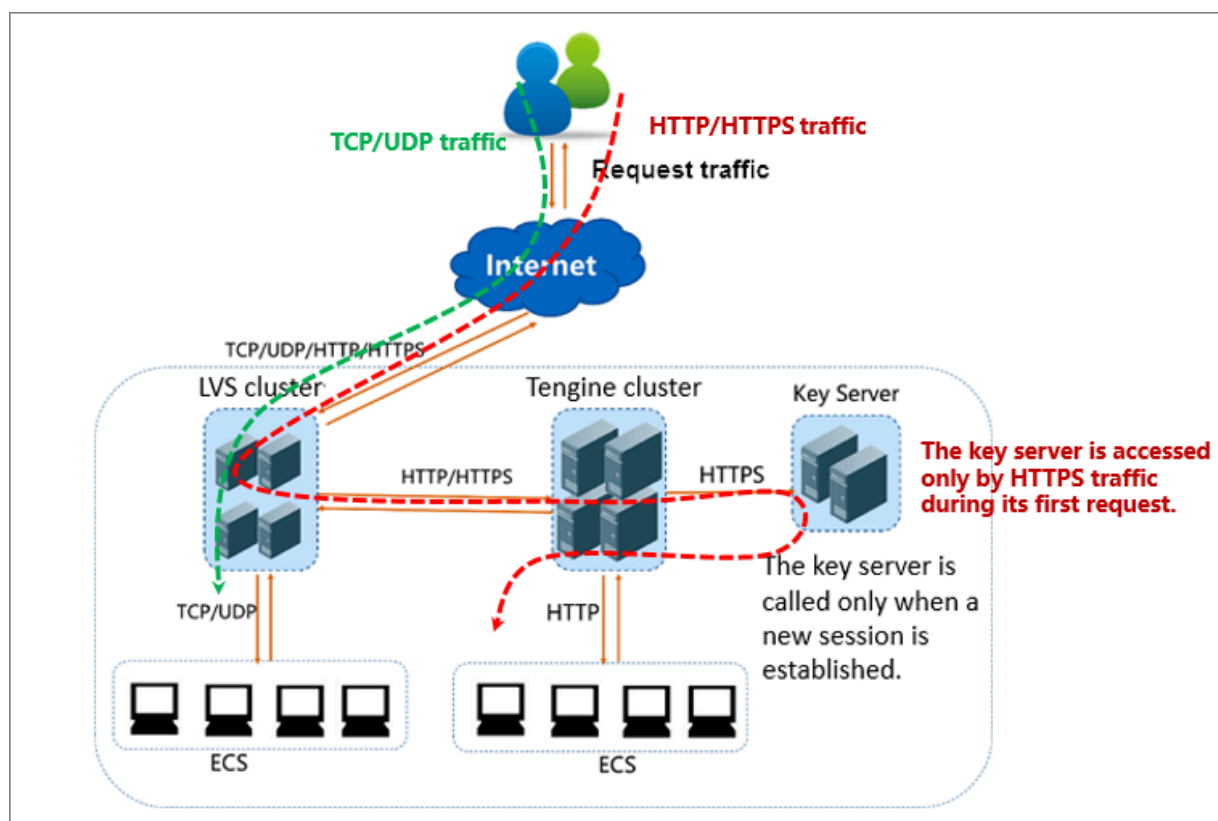
2 Network traffic flow

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

Inbound network traffic flow

SLB distributes incoming traffic according to the forwarding rules configured in the console or by using APIs. The following figure shows the inbound network traffic flow

Figure 2-1: Inbound network traffic flow



1. For TCP, UDP, HTTP, and HTTPS protocols, the incoming traffic must be forwarded through the LVS cluster first.

2. Large amounts of access requests are evenly distributed among all servers in the LVS cluster. Servers synchronize sessions 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 requests to backend ECS instances, the system calls the Key Server to validate certificates and decrypt data packets.

Outbound network traffic flow

SLB communicates with backend ECS instances through the intranet.

- If backend ECS instances only need to handle the traffic distributed from SLB, no public bandwidth (EIP, NAT Gateway, and public IP address) is required, and you do not need to purchase any public bandwidth.



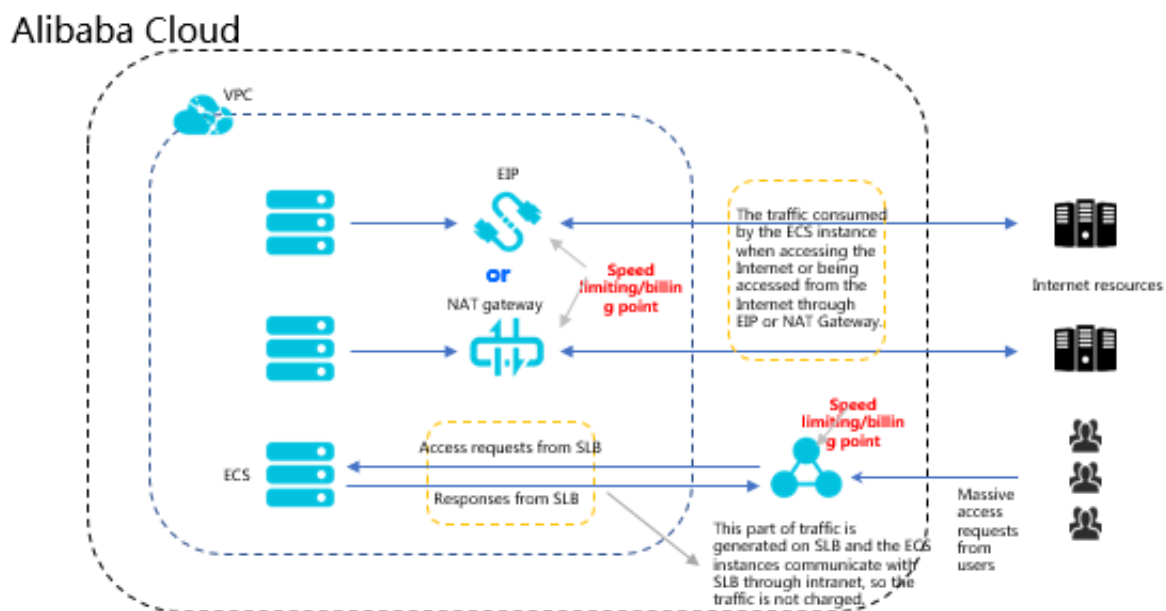
Note:

ECS instances previously created are directly allocated with public IP addresses. You can view the public IP addresses by using the `ifconfig` command. If these ECS instances process requests only through SLB, no traffic fee is incurred for traffic sent through the Internet even traffic statistics are read at the public network interface (NIC).

- If you want to provide external services through backend ECS instances, or backend ECS instances need to access the Internet, you must configure at least one of the following: a public IP address, an EIP, or a NAT Gateway.

The following figure shows the outbound network traffic flow.

Figure 2-2: Outbound network traffic flow



1. For outbound traffic from SLB instances (that is, traffic transferred through the Internet), traffic is sent at speeds dependent on the current network capacity, and is charged. However, you are not charged for intranet communications, such as traffic transferred between SLB instances and backend ECS instances.
2. For outbound traffic from an EIP or from NAT Gateway (that is, traffic transferred through the Internet), traffic is sent at speeds dependent on the current network capacity, and is charged. Additionally, if an ECS instance is configured with a public IP address when it is created, the outbound traffic from this instance is also charged.
3. SLB supports dynamic access to the Internet. Specifically, if a backend ECS instance needs to access the Internet, you must first configure a public IP address for it (by using an EIP or using NAT Gateway).
4. A public IP address (configured when you create an ECS instance), EIP, and NAT gateway all allow mutual Internet access. That is, ECS instances can access the Internet or be accessed from the Internet through any of these. Note, however, that they cannot forward traffic or balance traffic loads.

3 Create an SLB instance


This topic describes how to create a Server Load Balancer (SLB) instance.


Prerequisites

The required instance type, instance region, network type, and listener protocol are known. The environment is prepared. For more information, see [#unique_6](#).

Procedure

1. Log on to the [SLB console](#).
2. In the left-side navigation pane, choose Instances > Server Load Balancer, and click Create SLB Instance in the upper-left corner.
3. Configure the SLB instance according to the following information.

Configuration	Description
Region	<p>Select the region to which the SLB instance belongs.</p> <div> Note: Make sure that the region of the SLB instance is the same as that of backend ECS instances.</div>
Zone Type	<p>The zone type of the selected region is displayed. The zone of a cloud product refers to a set of independent infrastructure and is usually represented by data centers. Different zones have independent infrastructure (network, power supply, air-conditioning and so on). Therefore, an infrastructure fault in one zone does not affect other zones. A zone belongs to a specific region. A single region may have one or more zones. SLB has deployed multiple zones in most regions.</p> <ul style="list-style-type: none">• Single zone: The SLB instance is deployed only in one zone.• Multi-zone: The SLB instance is deployed in two zones. By default, the primary zone is used. If the primary zone is faulty, the secondary zone automatically takes 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 secondary zone for the SLB instance. The secondary zone only takes over traffic when the primary zone is unavailable.

Configuration	Description
Instance name	<p>Enter a name for the SLB instance to be created.</p> <p>The name must be 1 to 80 characters in length and can contain letters, numbers, Chinese characters, hyphens (-), slashes (/), periods (.), and underscores (_).</p>
Resource Group	The resource group to which the SLB instance to be created belongs.
Instance Spec	<p>Select a performance specification for the instance.</p> <p>The performance metrics vary by specification. For more information, see #unique_7.</p>
Instance Type	<p>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 #unique_7.</p> <ul style="list-style-type: none"> • Internet: An Internet SLB instance only provides a public IP address and you can access the SLB service from the Internet. • Intranet: An intranet SLB instance only provides a private IP address and you can access the SLB service only from the intranet.
IP version	<p>Select the IP version of the SLB instance, which can be IPv4 or IPv6.</p> <div>  <p>Note:</p> <p>Currently, IPv6 instances are supported only in the following regions. However, the instances must be guaranteed-performance instances.</p> <ul style="list-style-type: none"> • Zone E and Zone F in the China (Hangzhou) region • Zone F and Zone G in the China (Beijing) region • Zone D and Zone E in the China (Shanghai) region • Zone D and Zone E in the China (Shenzhen) region </div>
Quantity	Select the number of SLB instances to create.

4. Click Buy Now and complete the payment.

4 Create an IPv6 instance

This topic describes how to create an IPv6 Server Load Balancer (SLB) instance. After an IPv6 SLB 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, IPv6 instances are supported in the following zones, but the instances must be guaranteed-performance instances.
 - Zones E and F in the China (Hangzhou) region
 - Zones F and G in the China (Beijing) region
 - Zones D and E in the China (Shanghai) region
 - Zones D and E in the China (Shenzhen) region
- The Internet IPv6 network environment is still in the early stage of construction, and some links may be inaccessible. If such problem occurs, submit a ticket for technical support. SLA is not provided in the open beta test stage.
- IPv6 has a longer IP header than IPv4. Therefore, when you use a UDP listener in an IPv6 SLB instance, you must ensure that the MTU of the NIC communicating with SLB on the backend server (ECS instance) is not greater than 1480 (some applications need to synchronize their configuration files based on this MTU value). Otherwise, the packets may be discarded because they are too large.

If you use a TCP, HTTP, or HTTPS listener, no additional configurations are required because the TCP protocol supports MSS auto-negotiation.
- HTTP listeners can use the `X - Forwarded - For` header field to obtain source IPv6 addresses of clients.

IPv6 instances have the following features:

- Smooth migration and no impact on your service

You can directly associate ECS instances that use IPv4 addresses with an IPv6 SLB instance and smoothly migrate the service to IPv6 without modifying the original system.

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

- IPv6 access control ensures more secure and reliable service deployment

SLB supports IPv6 access control. You can configure access control lists 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. In the left-side navigation pane, choose 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. For more information, see [SLB configurations](#).



Note:

Currently, IPv6 instances are supported in the following zones, but the instances must be guaranteed-performance instances.

- Zones E and F in the China (Hangzhou) region
- Zones F and G in the China (Beijing) region
- Zones D and E in the China (Shanghai) region

- Zones D and E in the China (Shenzhen) region

The screenshot shows the configuration interface for creating a Server Load Balancer instance. The following fields are visible:

- Primary Zone:** A dropdown menu with "China North 2 Zone G" selected.
- Backup Zone:** A dropdown menu with "China North 2 Zone F" selected.
- Instance Name:** A text input field containing "auto_named_slb".
- LoadBalancerSpec:** A dropdown menu with "(slb.s1.small)" selected.
- Instance Type:** A button labeled "Public Network" with a help icon.
- IP Version:** Two radio buttons, "IPv4" and "IPv6". The "IPv6" button is selected and highlighted with a red box.

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

Result

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

The screenshot shows the "Server Load Balancer" console page. The table below lists the instances:

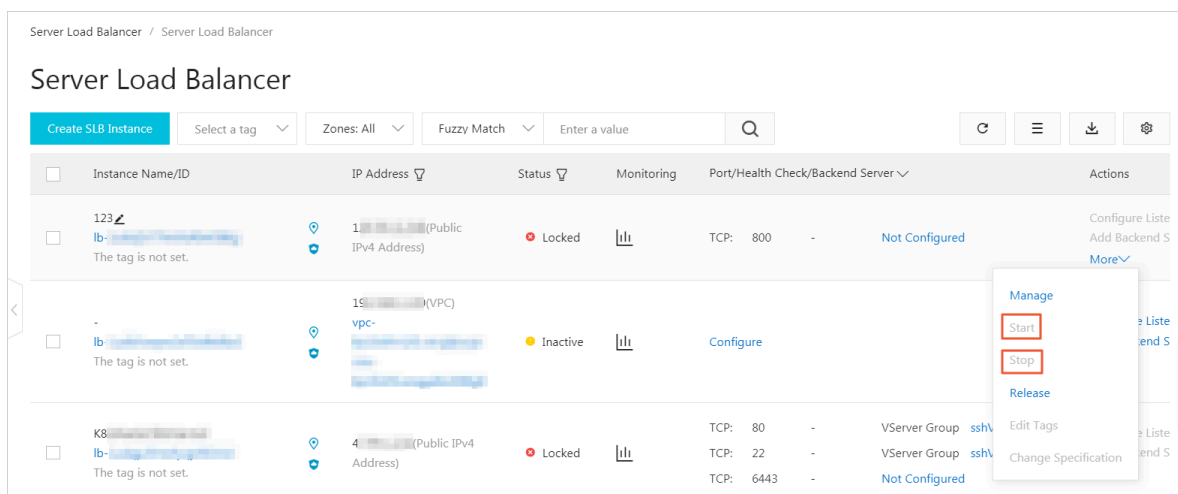
Instance Name/ID	IP Address	Status	Monitoring	Port/Health Check/Backend Server
auto_named_slb	2001:0000:0000:0000:0000:0000:0000:0000(Public IPv6 Address)	Active		Configure

5 Start or stop an SLB instance

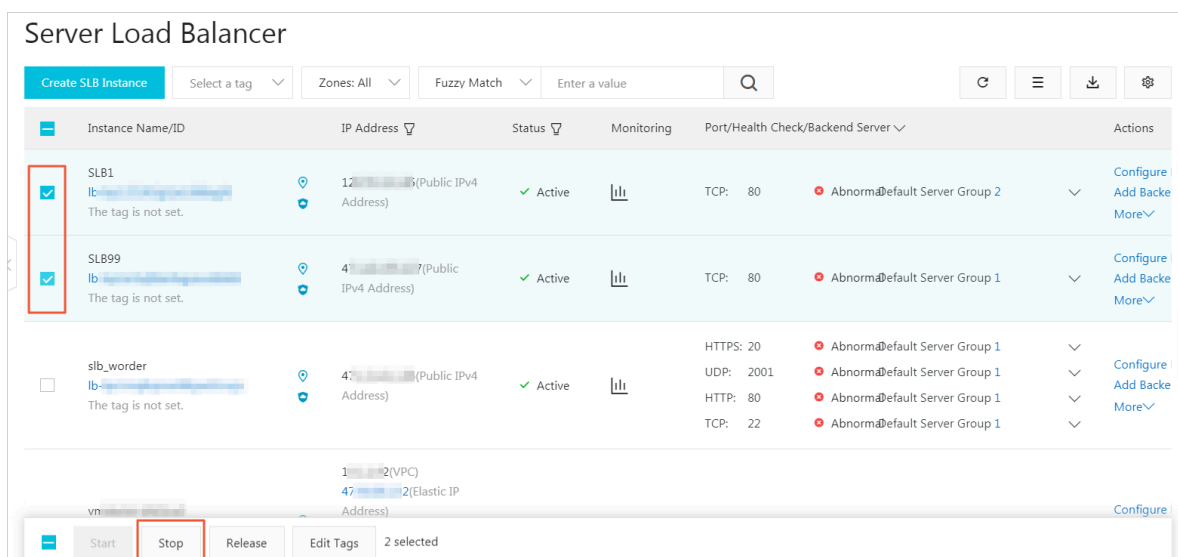
You can start or stop a Server Load Balancer (SLB) instance at any time. After being stopped, an SLB instance does not receive or forward requests any more.

Procedure

1. Log on to the [SLB console](#).
2. In the left-side navigation pane, choose **Instances > Server Load Balancer**.
3. Select the region of the target SLB instance and find the target instance.
4. In the Actions column, choose **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.



6 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

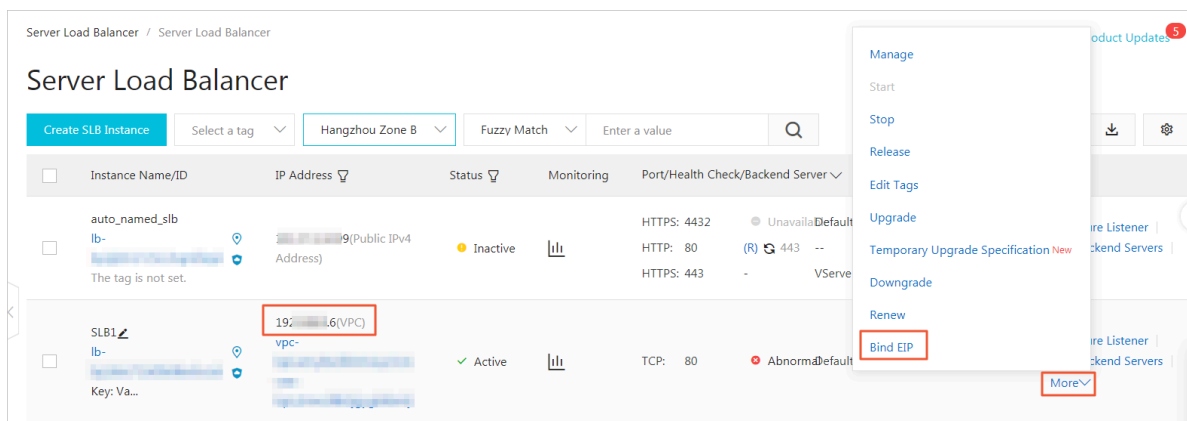
1. Log on to the [SLB console](#).
2. In the left-side navigation pane, click **Instances > Server Load Balancer**.
3. Select a region and find the target instance.



Note:

Ensure that the SLB instance is of the VPC network.

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



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

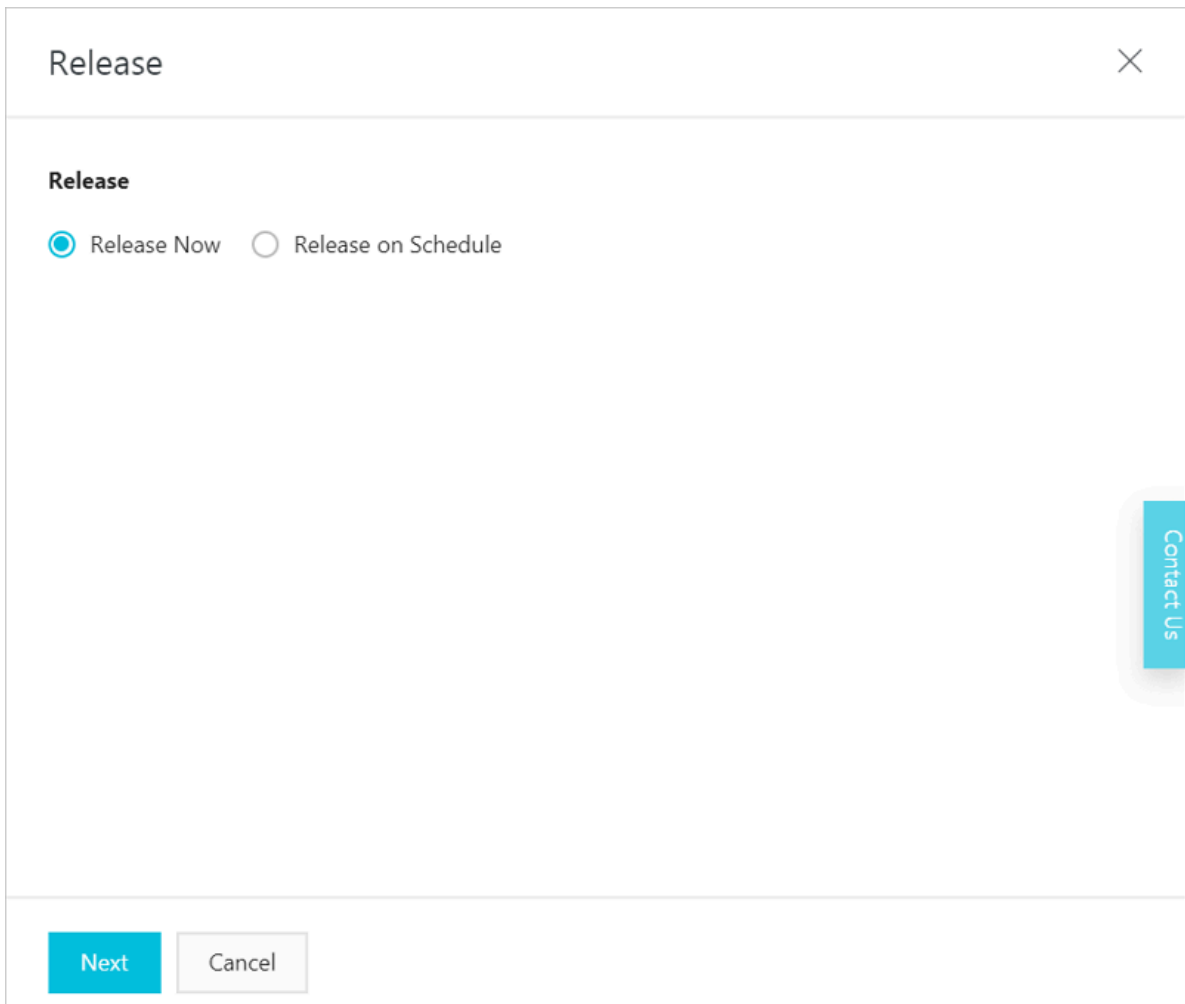
7 Release an SLB instance

This topic describes how to release a Server Load Balancer (SLB) instance. You can release an SLB instance immediately or at a specified time.

Procedure

1. Log on to the [SLB console](#).
2. Find the target instance and click More > Release.

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



The screenshot shows a modal dialog box titled "Release" with a close button (X) in the top right corner. Inside the dialog, under the heading "Release", there are two radio button options: "Release Now" (which is selected) and "Release on Schedule". At the bottom of the dialog, there are two buttons: "Next" (highlighted in blue) and "Cancel". On the right side of the dialog, there is a vertical blue button labeled "Contact Us".

3. On the Release page, select Release Now or Release on Schedule.



Note:

While the system executes the release operation every half hour or one hour cycle, the billing of the instance is stopped immediately at the release time you set.

4. Click Next.
5. Confirm the displayed information and click OK to release the instance.

8 Tags

8.1 Overview

You can classify Server Load Balancer (SLB) instances by using tags.

Each tag consists of a key and a value. Before you use tags, note the following limits:

- A tag cannot exist on its own and must be associated with an SLB instance.
- Up to 10 tags can be associated with an SLB instance.
- The key of each tag associated with an instance must be unique. Tags with the same key will be overwritten.
- Tags cannot be used across regions and are region-specific resources. For example, tags that belong to the China (Hangzhou) region are invisible to the China (Shanghai) region.

8.2 Add a tag

This topic describes how to add a tag for a Server Load Balancer (SLB) instance.

Procedure

1. Log on to the [Server Load Balancer console](#).
2. In the left-side navigation pane, choose Instances > Server Load Balancer.
3. Select the region of the target SLB instance and find the target SLB instance.
4. In the Actions column, choose More > Edit Tags.

5. On the Edit Tags page, add a tag.

To add a tag, follow these steps:

- If existing tags are available, click Saved Tags and then select a tag to add.
- If you want to create a new tag, on the Edit Tags page, click New Tag, enter the key and value of the new tag, and click OK.

6. Click OK.

8.3 Search for SLB instances by using a tag

This topic describes how to search for SLB instances by using a tag.

Procedure

1. Log on to the [Server Load Balancer console](#).
2. In the left-side navigation pane, choose Instances > Server Load Balancer.
3. Select a region.
4. Click Select a tag, and select the tag to be used as the search condition. The SLB instances associated with the selected tag are displayed.
5. To clear the search condition, rest the pointer over the selected tag and click the displayed delete icon.

8.4 Delete a tag

Server Load Balancer (SLB) does not support deleting tags of multiple instances in batches. You can remove the tags of only one instance at a time.

Procedure

1. Log on to the [Server Load Balancer console](#).
2. In the left-side navigation pane, choose Instances > Server Load Balancer.
3. Select the region of the target SLB instance and find the target SLB instance.
4. In the Actions column, choose 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 an SLB instance and is not associated with any other instances, the tag is deleted from the system.

9 Expiring Instances

This topic describes how to renew an expiring Server Load Balancer (SLB) instance.

If an SLB instance has an overdue payment, it is added to the list of expiring instances and, if not handled, released.

Context

If you do not renew an expiring instance, the process by which the instance is released is as follows:

- **Subscription instances:** The SLB instance is stopped and locked, and added to the list of expiring instances after an overdue payment is detected. If after seven days the payment is not settled, the SLB instance is released.
- **Pay-As-You-Go instances:** The SLB instance runs normally for 24 hours after an overdue payment is detected. If after 24 hours the payment is not settled, the SLB instance is stopped and locked, and added to the list of expiring instances, but not released. If after seven days the payment is not settled, the SLB instance is released.
-

Procedure

1. Log on to the [SLB console](#).
2. Choose Instances > Expiring Instances.
3. View detailed information of overdue instances.
4. Find the target SLB instance and click Renew in the Actions column, then the instance will be added back to the Server Load Balancer list.

10 Change the specification of an SLB instance

10.1 Overview

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

Before you modify the instance specification, note the following:

- When 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 instance type in a low-traffic period, or use DNS to schedule services to other SLB instances before you change the instance type.

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

You can use the simple I (slb.s1.small) specification after you change a shared-performance instance to a guaranteed-performance instance. This specification is free of charge.

- Intranet SLB instances only support traffic-based billing and cannot be changed to instances that are billed based on bandwidth.

10.2 Change the specification of an SLB instance

This topic describes how to change the specification of a Server Load Balancer (SLB) instance.

Procedure

1. Log on to the [SLB console](#).
2. Select the region of the target SLB instance.
3. Find the target SLB instance, choose More > Change Specification.
4. In the Configuration upgrade section, select a new specification, and complete the payment.

More information
[#unique_22](#)

11 Manage idle instances

This topic describes how to use the Server Load Balancer (SLB) console to display the Pay-As-You-Go instances that have been idle for more than seven days.

Procedure

1. Log on to the [Server Load Balancer console](#).
2. In the left-side navigation pane, choose SLB Lab > Idle SLB instances.
3. On the Idle SLB instance page, view all the Pay-As-You-Go instances that have not been used for more than seven days. You can click to customize the display of IP Address and Idle Cause.
4. To release an idle instance, click Release from the Actions column to immediately release the instance.



Note:

Given that the data of idle instances has a one-day cache period, make sure that the instance you want to release is not in use to prevent possible exceptions.

Server Load Balancer / Idle SLB instances

Idle SLB instances



Idle SLB instances list all pay-as-you-go SLB instances that have



Instance Name/ID

IP Address



SLB1


lb-3b123456789012345678901234567890

3

12 Health checks of SLB instances

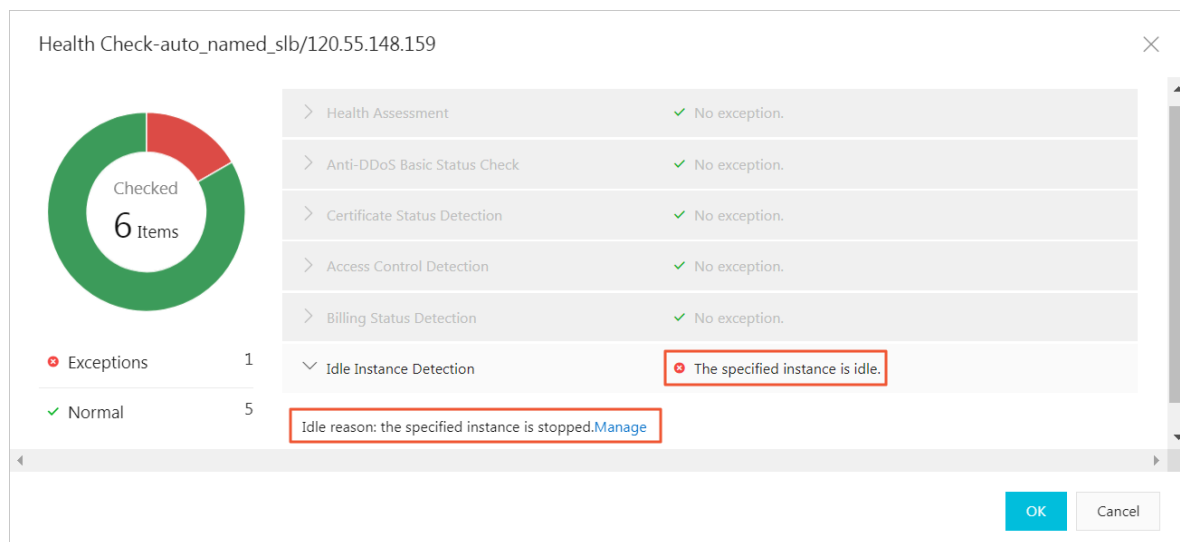
This topic describes how to check the status of Server Load Balancer (SLB) instances, specifically by viewing the results of health assessment, Anti-DDoS Basic status check, certificate status detection, access control detection, billing status detection, and idle instance detection. The health check function also provides causes and standard solutions for detected exceptions.

Procedure

1. Log on to the [Server Load Balancer console](#).
2. In the left-side navigation pane, choose Instances > Server Load Balancer.
3. Find the target SLB instance and click  in the Health Check column.

Optionally, you can click the ID of the target SLB instance and on the instance details page, click Health Check in the upper-right corner.

4. On the health check details page, the health check result of the target SLB instance is displayed. If any exception is detected, the cause and solution are provided.



13 FAQ

13.1 SLB instance FAQ

The following are frequently asked questions about SLB instances:

- [Can I change a shared-performance instance to a guaranteed-performance instance?](#)
- [What are the specifications of a shared-performance instance?](#)
- [How do I select a suitable specification for a guaranteed-performance instance?](#)
- [Can I change the specification of a guaranteed-performance instance?](#)
- [Are shared-performance instances still available for purchase?](#)

Can I change a shared-performance instance to a guaranteed-performance instance?

Yes.

After an SLB instance is changed to the guaranteed-performance type, it cannot be changed back.

What are the specifications of a shared-performance instance?

Shared-performance instances do not guarantee the performance. No specifications are available.

How do I select a suitable specification for a guaranteed-performance instance?

- You can select the largest specification for a Pay-As-You-Go instance, because Pay-As-You-Go instances are charged according to the actual usage and no fees are incurred in idle time.

Can I change the specification of a guaranteed-performance instance?

Yes.

- You can upgrade or downgrade the specification of a guaranteed-performance instance. For more information, see [Change the configuration](#).



Note:

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

- Some previously created instances may exist in older clusters. If you change such a shared-performance instance to a guaranteed-performance instance, a brief disconnection of service may occur for 10 to 30 seconds due to instance migration. Therefore, for such scenarios, we recommend that you make the change when the traffic is low.

Are shared-performance instances still available for purchase?

Yes.

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

13.2 Guaranteed-performance instance FAQ

Guaranteed-performance Server Load Balancer (SLB) instances are instances whose performance is guaranteed in terms of specific indicators, such as the maximum number of connections, Connection Per Second (CPS), and Query Per Second (QPS). Note that Alibaba Cloud now charges specification fees for guaranteed-performance instances.

The following are frequently asked questions about guaranteed-performance instances:

- [#unique_29/unique_29_Connect_42_section_lht_gym_vdb](#)
- [#unique_29/unique_29_Connect_42_section_eth_yzm_vdb](#)
- [#unique_29/unique_29_Connect_42_section_n5z_sl_n_vdb](#)
- [#unique_29/unique_29_Connect_42_section_ifx_kcn_vdb](#)
- [#unique_29/unique_29_Connect_42_p7](#)
- [#unique_29/unique_29_Connect_42_section_gvt_kfn_vdb](#)
- [#unique_29/unique_29_Connect_42_section_hxl_pfn_vdb](#)
- [#unique_29/unique_29_Connect_42_section_ehc_vfn_vdb](#)
- [#unique_29/unique_29_Connect_42_section_flq_wfn_vdb](#)
- [#unique_29/unique_29_Connect_42_section_nfy_xfn_vdb](#)

What is a guaranteed-performance instance?

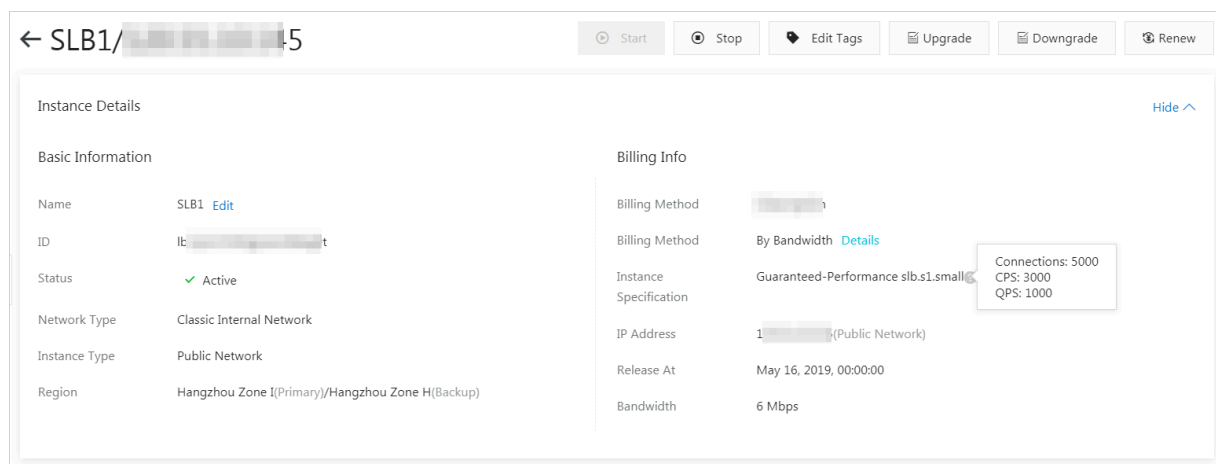
A guaranteed-performance instance provides guaranteed performance metrics (performance SLA) and is opposite to a shared-performance instance. For a shared-

performance instance, performance metrics are not guaranteed and resources are shared by all instances.

All instances were shared-performance instances before Alibaba Cloud launched guaranteed-performance instances. You can view the instance type in the SLB console

.

You can rest the pointer over the question mark icon of the target guaranteed-performance 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 an SLB instance. When the number of connections reaches the limit of the specification, new connection requests will be dropped.

- **Connection Per Second (CPS)**

The rate at which new connections are established per second. When the CPS reaches the limit of the specification, new connection requests will be dropped.

- **Query Per Second (QPS)**

The number of HTTP/HTTPS requests that can be processed per second. This metric is available only for Layer-7 SLB listeners. When the QPS reaches the limit of the specification, new connection requests will be dropped.

Alibaba Cloud SLB provides the following specifications for guaranteed-performance instances:

Type	Specification	Max Connection	CPS	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.

How are guaranteed-performance instances billed?

Guaranteed-performance instances are billed as follows:

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



Note:

The specification fee is charged on intranet guaranteed-performance instances in the same way as Internet guaranteed-performance instances. But no traffic fee or instance fee is charged for intranet guaranteed-performance instances.

The specification fee of a guaranteed-performance instance is charged based on actual usage. No matter what specification you choose, the specification fee will be charged according to the specification you actually use.

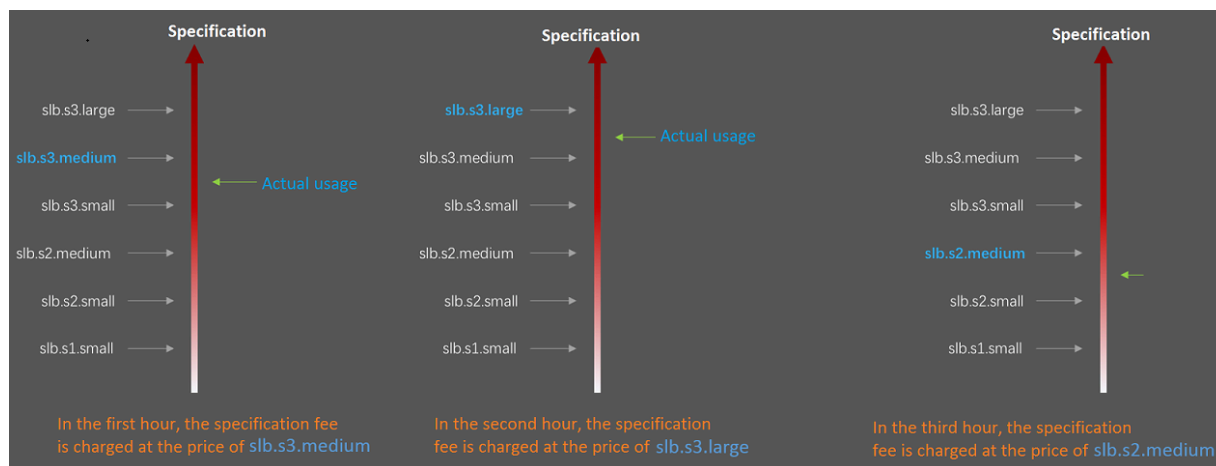
For example, if you purchase the Super I specification (Max Connection: 1,000,000; CPS: 100,000; QPS: 50,000) and the actual usage of your instance in an hour is as follows:

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

- With respect to Max Connection, the actual metric value of 90,000 lies between the limit of 50,000 defined in Standard I (slb.s2.small) and the limit of 100,000 defined in Standard II (slb.s2.medium). Therefore, the specification of the Max Connection metric for this hour is Standard II (slb.s2.medium).
- With respect to CPS, the actual metric value of 4,000 occurs between the limit of 3,000 defined in the Small I (slb.s1.small) specification and the limit of 5,000 defined in the Standard I (slb.s2.small) specification. Therefore, the specification of the CPS metric for this hour is Standard I (slb.s2.small).
- With respect to QPS, the actual metric value of 11,000 occurs between the limit of 10,000 defined in Standard II (slb.s2.medium) and the limit of 20,000 defined in Higher I (slb.s3.small). Therefore, the specification of the QPS metric for this hour is Higher I (slb.s3.small).

Out of the three metrics, QPS has the highest instance specification. Therefore, the specification fee of the instance in this hour is charged according to 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:



The billing is more flexible for guaranteed-performance instances. The specification you select when purchasing an instance is the higher performance limit of the instance. For example, if you select Higher II (slb.s3.medium), new requests will be dropped when requests reach 30,000 in one second.

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 [#unique_30](#).

Region	Specification	Max Connectio	CPS	QPS	Specificat ion fee (USD/ hour)
China (Hangzhou)	Specification 1: Small I (slb.s1. small)	5,000	3,000	1,000	Free of charge
China (Zhangjiakou)	Specification 2: Standard I (slb.s2. small)	50,000	5,000	5,000	0.05
China (Hohhot)	Specification 3: Standard II (slb.s2. medium)	100,000	10,000	10,000	0.10
China (Qingdao)	Specification 4: Higher I (slb.s3. small)	200,000	20,000	20,000	0.20
China (Beijing)	Specification 5: Higher II (slb.s3. medium)	500,000	50,000	30,000	0.31
China (Shanghai)	Specification 6: Super I (slb.s3.large)	1,000,000	100,000	50,000	0.51
China (Shenzhen)					

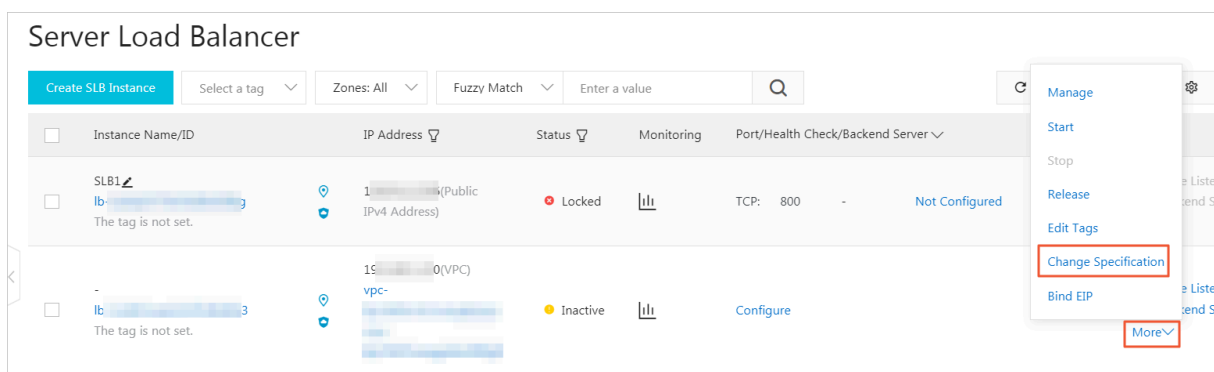
Region	Specification	Max Connectio	CPS	QPS	Specificat ion fee (USD/ hour)
Singapore Malaysia (Kuala Lumpur) Indonesia (Jakarta) India (Mumbai) US (Silicon Valley) US (Virginia) China (Hong Kong)	Specification 1: Small I (slb.s1. small)	5,000	3,000	1,000	Free of charge
	Specification 2: Standard I (slb.s2. 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: Super I (slb.s3.large)	1,000,000	100,000	50,000	0.61

What is the optimal specification for a guaranteed-performance instance?

Because the specification fee is billed based on actual usage, we recommend that you select the largest specification (slb.s3.large). This guarantees your service 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.

Can I change the specification of my SLB instance after it is created?

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

**Note:**

- After a shared-performance instance is changed to a guaranteed-performance instance, it cannot be changed back.
- Some previously created SLB instances 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. Therefore, for such scenarios, we recommend that you make the change when the traffic is low.
- IP addresses of SLB instances will not be affected after you change the instance type or the specification.

**Caution****Warning:**

Your SLB service might be suspended for up to 30 seconds when you switch the SLB instance specification from shared-performance to guaranteed-performance

We recommend that you change the specified SLB instance settings during non-business hours, or after load balancing to another SLB instance is completed.

OK

Cancel

When did Alibaba Cloud begin to charge specification fees on guaranteed-performance instances?

Alibaba Cloud began to charge specification fee on guaranteed-performance instances from April 1, 2018, and continues to sell shared-performance instances.

The charging of specification fee takes effect in batches as follows:

- The first batch:

Start time: From April 1, 2018 to April 10, 2018

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

- The second batch:

Start time: From April 11, 2018 to April 20, 2018

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

- The third batch:

Start time: From April 21, 2018 to April 30, 2018

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

Is an extra fee included for shared-performance instances after Alibaba Cloud starts charging for the specification fee?

No.

Extra fees are not charged for shared-performance instances unless you change them to guaranteed-performance instances.

Why sometimes guaranteed-performance instances cannot reach the performance limit 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 reaches the limit defined in the specification.

When the QPS of the instance reaches 20,000 but the number of maximum connections does not reach 200,000, new connections are still dropped because the QPS has reached the limit.

Are shared-performance instances still available for purchase?

Yes.

Shared-performance instances are still available now, but they will be phased out in the future. Please pay attention to official announcements or emails.

Is a specification fee charged for intranet SLB instances?

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, a specification fee is charged. The calculation method of the specification fee for intranet instances is the same as that for Internet instances. No instance fee or traffic fee is charged for intranet instances.