Alibaba Cloud Server Load Balancer

FAQ

Issue: 20190523

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

Generic conventions

Table -1: Style conventions

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

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

Contents

Legal disclaimer..... I
Generic conventions..... I
1 Why am I unable to access the SLB instance?.....1
2 Why is the traffic between my ECS instances unevenly distributed?.....5
3 Obtain real client IP addresses.....7
4 What can I do if health checks generate an excessive number of logs?.....11
5 What can I do if my ECS instance is declared unhealthy after I enable health checks for Server Load Balancer?..... 14
6 Troubleshoot the health check exception of a Layer-4 listener (TCP/UDP)..... 16
7 Troubleshoot a health check exception of a Layer-7 listener (HTTP/HTTPS).....19
8 How do I perform a stress test?..... 22
9 How do I troubleshoot HTTP 5xx errors?..... 25
10 Billing FAQ.....30
11 SLB instance FAQs.....32
12 Guaranteed-performance instances..... 34
13 Server Load Balancer FAQs.....42
14 Backend server FAQs.....47
15 Heath check FAQs..... 49
16 HTTPS and HTTP listener FAQs..... 55
17 WS and WSS support FAQs..... 59

1 Why am I unable to access the SLB instance?

This tutorial describes the causes for why the client cannot access an SLB instance and the troubleshooting methods.

Note:
In this example, the frontend port of the SLB instance is 80, the backend port of the ECS instance is 80, and the intranet IP address of the ECS instance is 10.11.192.1. You must configure the port and intranet IP address according to your actual situation.

No.	Causes	Resolutions
1	For Layer-4 SLB service, a backend ECS instance cannot directly provide service for clients and function as the backend server of the SLB service at the same time.	-
2	Health check exception.	For more information, see *Troubleshoot the health check exception of a Layer-4 listener (TCP/UDP)* and *Troubleshoot a health check exception of a Layer-7 listener (HTTP/HTTPS)*.
3	Using FTP, tftp, h323 , sip and related protocols through SLB is not supported.	· For a Linux system, you can configure the forwarding of Port 22 and use sftp to connect and transmit data. · You can attach an EIP to an FTP server to provide external FTP service through the EIP cut-through mode.

4	The intranet firewall of the server does not allow Port 80.	You can run the following commands to temporarily disable the firewall.
		· For a Windows server, run:
		`firewall . cpl`
		· For a Linux server, run:
		`/ etc / init . d / iptables stop`

No.	Causes	Resolutions
5	Backend port exception.	· For Layer-4 SLB service, the backend port is normal if you receive a response after performing the telnet test. Example: Use `telnet 10.11.192.1 80` to test. · For Layer-7 SLB service, the HTTP status code must be a status code that indicates a normal condition, such as 200. The test methods are as follows: - Windows: Access the intranet IP address of the ECS instance directly from the ECS instance to check if access is normal. Example: `http://10.11.192.1` - Linux: Use the `curl -I` command to check if the status is `HTTP/1.1 200 OK`. Example: `curl -I 10.11.192.1`

6	The rp_filter feature conflicts with the policy route of the bottom-layer LVS of SLB.	1. Log on to the ECS instance of the Linux system added to the SLB.
		2. Edit the */etc/sysctl.conf* file and set the following three parameters in the system configuration file to `0`.
		`net.ipv4.conf.default.rp_filter = 0`
		`net.ipv4.conf.all.rp_filter = 0`
		`net.ipv4.conf.eth0.rp_filter = 0`
		3. Run the `sysctl -p` command to make the configurations take effect.

No.	Causes	Resolutions

7	Listener exception	Run the following command on the server.
		If you can see the monitoring information of 10.1.1.192.1: 80, or the monitoring information of 0.0.0.0: 80, the listening function of the ports is normal.
		· For a Windows server, run:
		`netstat -ano | findstr :80`
		· For a Linux server, run:
		`netstat -anp | grep :80`

8	No listener is added to the SLB instance.	Configure a listener. For more information, see *Configure a listener*.
9	The SLB cannot be accessed through the domain name. Therefore, an error may occur in domain name resolution.	-
10	Exception of the local network of the client or exception of the intermediate link of the service provider.	Perform access tests on the service port of SLB in different regions and network environments. If the exception only occurs when it is accessed from the local network, the problem is caused by a network exception. In this case, you can do further troubleshooting and analysis through ongoing ping tests or MTR route tracing.

No.	Causes	Resolutions
11	The client IP address is blocked by Alibaba Cloud Security.	1. Visit `http :// ip . taobao . com` in the network environment of the client to obtain the public IP of the client network. 2. Add the IP address to the SLB whitelist to allow access from the IP address. Note: This operation may pose security risks. Ensure that the IP does not launch malicious attacks on the SLB.
12	When the SLB returns to Anti-DDoS Basic after using Anti-DDoS Pro, the whitelist is not disabled.	Disable the ACL whitelist.
If the problem persists, open a ticket and submit the following details so that we can help you solve the problem more efficiently: · The ID of the SLB instance or the IP address of the SLB service. · The public IP of the client obtained when you visit `ip . taobao . com` . · Screenshots of ping and MTR route tracing tests performed by the client using the IP address of the SLB service.		

2 Why is the traffic between my ECS instances unevenly distributed?

Reasons

Traffic may be unevenly distributed due to the following reasons:

- Only a small number of requests are being received by ECS instances.
- The target ECS instances have different network capacities.

  > Note:
  > The memory usage of ECS instances does not indicate whether requests are evenly distributed.

- Session persistence is enabled

  If session persistence is enabled, it will cause traffic imbalance when fewer clients are accessing the Server Load Balancer instance. This is especially common when a small number of clients are used to test the Server Load Balancer instance. For example, session persistence (source-IP-based) is enabled for a TCP listener and a client is used to test the load balancing service.

- Abnormal ECS status

  Backend servers with abnormal heath status can also lead to an imbalance especially during stress test. If the health check for a backend ECS instance fails or its health status changes frequently, this will cause an imbalance.

- TCP keepalive

  When some backend ECS instances enable TCP keepalive while some do not, the connections will accumulate on the ECS instances with TCP keepalive enabled, which will cause the imbalance.

Troubleshooting

- Check whether the weights of backend ECS instances are the same.
- Check if the health check of the backend ECS instances fails or if the health status is unstable in a specified period. Check if the health check is correctly configured with the status code.

- **Check if both the WLC scheduling algorithm and session persistence are enabled. If so, change the scheduling algorithm to WRR and session persistence.**

3 Obtain real client IP addresses

SLB supports obtaining real client IP addresses.

Overview

Support for obtaining real IP addresses in SLB is enabled by default.

- For Layer-4 SLB service (TCP protocol), listeners distribute client requests to backend ECS servers without modifying the request headers. Therefore, you can obtain real client IP addresses directly.
- For Layer-7 SLB service (HTTP and HTTPS protocols), you need to configure application servers, and then use the `X-Forwarded-For` header to obtain real client IP addresses.

  Real client IP addresses are put in the `X-Forwarded-For` fields of HTTP headers in the following format:

  ```
  X-Forwarded-For: the real IP address of the user, the proxy server 1-IP, the proxy server 2-IP, ...
  ```

  When you use the `X-Forwarded-For` header to obtain the real client IP address, the first IP address obtained is the real IP address.

  > **Note:**
  > For the HTTPS SLB service, SSL certificates are configured in frontend listeners, and the backend still uses the HTTP protocol. Therefore, the configurations on application servers for obtaining real client IP addresses are the same for HTTP and HTTPS protocols.

Configure IIS7/IIS8

1. *Download* and extract *F5XForwardedFor*.
2. Copy the *F5XFFHttpModule.dll* and *F5XFFHttpModule.ini* files from the *x86\Release* or *x64\Release* directory (depending on the operating system version) of your server to a directory, such as *C:\F5XForwardedFor*. Make sure that the IIS process has write access to this directory.
3. Open IIS Manager and double-click the Modules function.

4. Click Configure Native Modules, and then click Register in the displayed dialog box.

5. Add the downloaded *.dll* file.

6. Add the ISAPI and CGI restrictions for the added files and set the restrictions to Allowed.

Note:
Make sure that you have installed the ISAPI and CGI applications.

7. Restart IIS Manager.

Configure Apache

1. Run the following command to install the mod_rpaf module:

```
wget https://github.com/gnif/mod_rpaf/archive/v0.6.0.tar.gz
tar zxvf mod_rpaf-0.6.tar.gz
cd mod_rpaf-0.6
/alidata/server/httpd/bin/apxs -i -c -n mod_rpaf-2.0.so mod_rpaf-2.0.c
```

2. Open the */alidata/server/httpd/conf/httpd.conf* file and add the following information at the end of the content:

```
LoadModule rpaf_module modules/mod_rpaf-2.0.so
RPAFenable On
RPAFsethostname On
RPAFproxy_ips <IP_address>
RPAFheader X-Forwarded-For
```

Note:
To obtain the IP address of the proxy server, add the CIDR block of the proxy server to `RPAFproxy_ips <IP_address>`, such as the IP address range of SLB 100.64.0.0/10 and the address range of Anti-DDoS Pro. Separate multiple CIDR blocks by using commas (,). 100.64.0.0/10 is reserved by Alibaba Cloud. Therefore, it is not used by any user and will not pose any security risk.

3. Restart Apache.

```
/ alidata / server / httpd / bin / apachectl   restart
```

Configure Nginx

1. Run the following command to install http_realip_module:

```
 wget   http :// nginx . org / download / nginx - 1 . 0 . 12 . tar
. gz
 tar   zxvf   nginx - 1 . 0 . 12 . tar . gz
 cd   nginx - 1 . 0 . 12
./ configure  -- user = www  -- group = www  -- prefix =/ alidata /
server / nginx  -- with - http_stub_  status_mod  ule  -- without
- http - cache  -- with - http_ssl_m  odule  -- with - http_reali
p_module
 make
 make   install
 kill  - USR2  ` cat  / alidata / server / nginx / logs / nginx .
pid `
 kill  - QUIT  ` cat  / alidata / server / nginx / logs /  nginx .
pid . oldbin `
```

2. Open the *nginx . conf* file.

```
vi  / alidata / server / nginx / conf / nginx . conf
```

3. Add new configuration fields and information at the end of the following configuration information:

```
 fastcgi   connect_ti  meout   300 ;
 fastcgi   send_timeo  ut   300 ;
 fastcgi   read_timeo  ut   300 ;
 fastcgi   buffer_siz  e   64k ;
 fastcgi   buffers   4   64k ;
 fastcgi   busy_buffe  rs_size   128k ;
 fastcgi   temp_file_  write_size   128k ;
```

The configuration fields and information that need to be added are:

```
 set_real_i  p_from   IP_address
 real_ip_he  ader   X - Forwarded - For ;
```

Note:

To obtain the IP address of the proxy server, add the CIDR block of the proxy server to `set_real_i  p_from  < IP_address >`, such as the IP address range of SLB 100.64.0.0/10 and the address arrange of Anti-DDos Pro. Separate multiple CIDR blocks by using commas (,). 100.64.0.0/10 is reserved by Alibaba Cloud. Therefore, it is not used by any user and will not pose any security risk.

4. **Restart Nginx.**

```
/ alidata / server / nginx / sbin / nginx  - s   reload
```

4 What can I do if health checks generate an excessive number of logs?

The log management of Server Load Balancer can automatically save health check logs generated within three days. If there are too many health check logs that may affect your maintenance, you can configure to reduce health check logs or prevent the logs from being generated in certain scenarios through the following methods.

> Note:
> After reducing health check logs, you may be unable to identify problems that occur during the running of Server Load Balancer instance in a timely manner. So configure it according to the actual situation.

- *Get access logs*
- *Adjust health check frequency*
- *Close Layer-7 health check*
- *Change Layer-7 Server Load Balancer to Layer-4 Server Load Balancer*
- *Disable application logs on the health check page*

Get access logs

HTTP health checks use the head request method by default (the get method will be used later). Therefore, you can obtain access logs by filtering out head requests.

Adjust health check frequency

You can increase the interval between two health checks to reduce the health check frequency and generated logs.

Potential risks

After you increase the interval, if the backend ECS instance fails, the time needed for Server Load Balancer to detect the faulty ECS instance is increased accordingly.

Procedure

1. Log on to the *SLB console*.
2. On the Server Load Balancer page, click the ID of the target SLB instance. Then, on the Listeners tab page, find the target listener and click Configure.

3. On the Configure Listener page, click Next and then click Next again to configure health checks.
4. Adjust the Health Check Interval in the range of 1 to 50 seconds. The greater the interval is, the lower the health check frequency is, and the fewer logs are generated by the backend server. Modify the interval according to your actual situation.
5. Click OK.

Close Layer-7 health check

When Layer-7 (HTTP/HTTPS) Server Load Balancer is used, the health check is implemented by HTTP head requests. Application logs of the backend ECS instance record the health check requests, leading to a large number of logs.

Potential risks

After you close HTTP/HTTPS health checks, Server Load Balancer does not check backend servers. If a backend server fails, the traffic cannot be automatically forwarded to other normal backend servers.

Procedure

1. Log on to the *SLB console*.
2. On the Server Load Balancer page, click the ID of the target SLB instance. Then, on the Listeners tab page, find the target listener and click Configure.
3. On the Configure Listener page, click Next and then click Next again to configure health checks.
4. Close Enable Health Check.
5. Click OK.

Change Layer-7 Server Load Balancer to Layer-4 Server Load Balancer

Layer-4 health checks only use TCP three-way handshakes and generate no applicatio n logs. If your service can be changed to Layer-4 Server Load Balancer, this method can reduce generated application logs.

Potential risks

After you change the HTTP/HTTPS Server Load Balancer to the TCP Server Load Balancer, Server Load Balancer checks only the status of the listener port and does not check the HTTP status. In this way, Server Load Balancer cannot detect the exceptions that occur to HTTP applications in real time.

Procedure

1. Log on to the *SLB console*.
2. On the Server Load Balancer page, click the ID of the target SLB instance. Then, on the Listeners tab page, find the target listener and click Configure.
3. On the Configure Listener page, click Next and then click Next again to configure health checks.
4. Change the Health Check Protocol to TCP.
5. Click OK.

Disable application logs on the health check page

Configure the health check site that is independent from the service site, and disable the application logs of the health check page, then the number of health check logs can be reduced. For example, the service site is abc. 123.com, use test. 123.com as the health check site and disable logging of test. 123.com.

Potential risks

If the health check site is running normally, but the service site encounters an exception, the health check cannot detect the exceptions of the service site.

Procedure

1. Create a new health check site and health check page on the backend server and disable logging. This topic takes nginx as an example.

2. Log on to the *SLB console*.
3. On the Server Load Balancer page, click the ID of the target SLB instance. Then, on the Listeners tab page, find the target listener and click Configure.
4. On the Configure Listener page, click Next and then click Next again to configure health checks.
5. Enter the domain name of the health check site in Health Check Domain Name and enter the path of the health check page in Health Check Path.
6. Click OK.

5 What can I do if my ECS instance is declared unhealthy after I enable health checks for Server Load Balancer?

After you enable health checks of Server Load Balancer, when one backend ECS instance is declared as unhealthy, requests are forwarded to other normal ECS instances. When the faulty ECS instance becomes normal, Server Load Balancer forwards requests to the ECS instance again.

For Layer-7 SLB service, when an ECS instance is declared as unhealthy, you can troubleshoot the ECS instance from the following aspects:

- Make sure that you can directly access your service through the ECS instance.
- Make sure that the backend port you configured in the listener is opened on the backend server.
- Check whether the backend ECS instance has installed a firewall or other security protection software. This type of software may easily block the local IP address of the Server Load Balancer service, and thus disable the communication between the Server Load Balancer service and the backend server.
- Check whether the Server Load Balancer health check parameters are correctly set . We recommend that you use default health check parameters.
- We recommend that you use a static page for health checks. If the page you use for health checks isn't the default homepage of the backend ECS instance, you must enter the URL of the health check page in health check configurations. We recommend that you use a simple html page for health checks and the page is only used for checking the response. We do not recommend that you use dynamic scripting languages such as php.
- Check whether the backend ECS instance has high loads, which slow the ECS instance's response in offering services.

Besides, because the Layer-7 SLB service communicates with the backend ECS instance through intranet, the ECS instance must listen to intranet or all-network ports. You can check the ECS instance using the following methods:

1. Check whether the listening function is normal.

   If the frontend port is 80 and the backend port is 80, the intranet IP address of the ECS instance is 10.11.192.1. Run the following command on the server. If you can see the monitoring information of 10.1.1.192.1: 80, or the monitoring information of 0.5.0.0: 80, the listening function of the ports is normal.

   - Run the following command on the Windows server: `netstat -ano | findstr :80`
   - Run the following command on the Linux server: `netstat -anp | grep :80`

2. Check whether the intranet firewall of the server allows port 80. You can disable the firewall temporarily to do the test. Enter the following command to disable the firewall.

   - Windows: `firewall.cpl`
   - Linux: `/etc/init.d/iptables stop`

3. Check whether the backend port is normal.

   - For Layer-4 SLB service, the backend port is normal if you receive response after performing the telnet test. In this topic, use `telnet 10.11.192.1 80` to do the test.
   - For Layer-7 SLB service, the HTTP status code must be a status code that indicates a normal condition, such as 200. The test methods are as follows:
     - Windows: Directly access the intranet IP of the ECS instance. In this topic, access *http://10.11.192.1*.
     - Linux: Use the `curl -I` command to check if the status is HTTP/1.1 200 OK. In this topic, use `curl -I 10.11.192.1`.

6 Troubleshoot the health check exception of a Layer-4 listener (TCP/UDP)

The health check function is used to determining whether your backend servers are normal. When a health check exception occurs, it generally means that your backend server is abnormal. The exception may also be caused by incorrect health check configurations. This topic describes how to troubleshoot a health check exception of a Layer-7 (TCP/UDP) listener.

Procedure

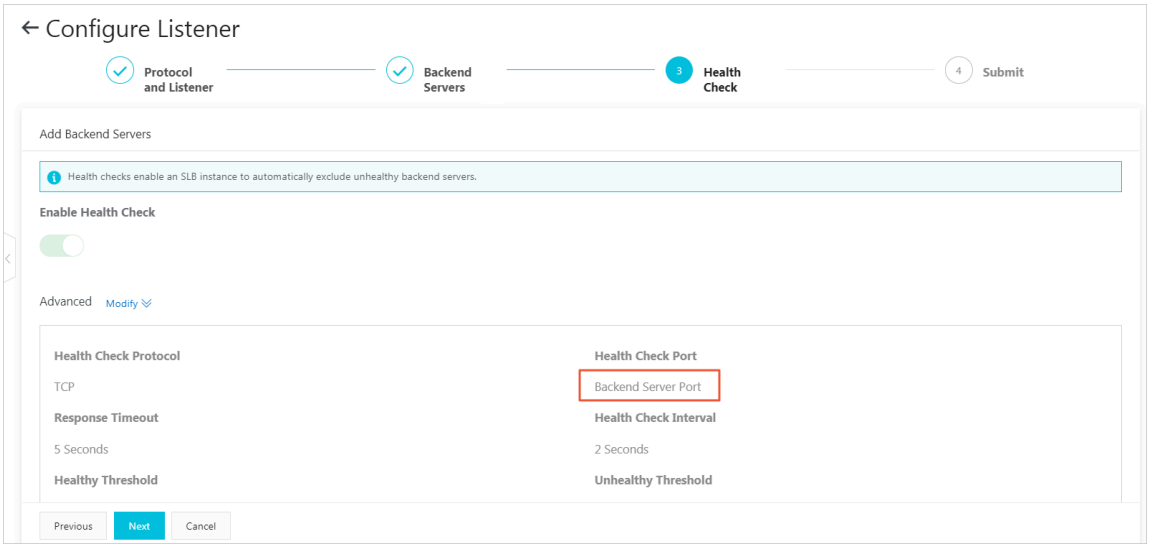
1. **Ensure that the backend server does not block the CIDR block 100.64.0.0/10 through iptables or other third-party firewalls/security software.**

   The SLB instance communicates with backend servers by using IP addresses in the reserved CIDR block 100.64.0.0/10. If the CIDR block is blocked, a health check exception occurs and the SLB instance cannot work normally.

2. Run the `telnet` command to test the backend server.

   a) Log on to the SLB console and click the ID of the target SLB instance. On the Listeners tab page, click Configure in the Actions column of the target listener. Then you can view the health check configurations.

   By default, the port of the backend server is used as the Health Check Port. You can also set the port manually. In this example, the port of the backend server, namely port 80, is used.



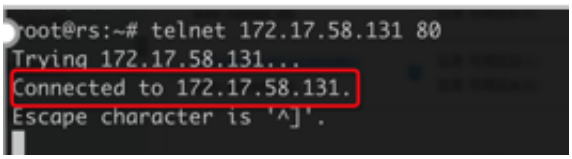
   b) Run the following command to connect to the health check port. The health check port configured on the SLB instance must be the same as the listening port on the backend server.

```
telnet   172 . 17 . 58 . 131   80
```

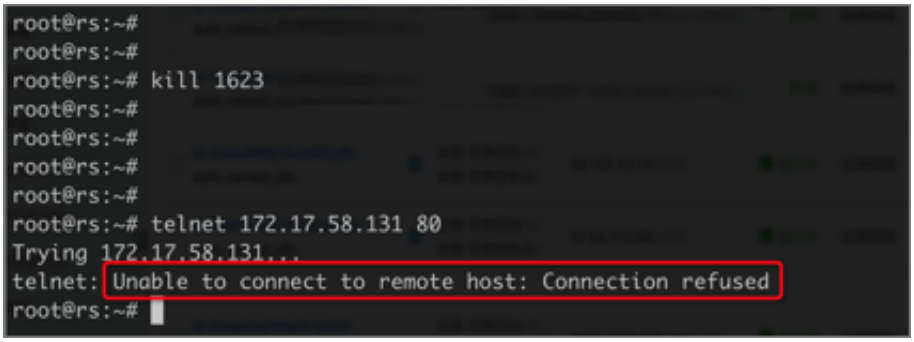
   In this example, *172.17.58.131* is the intranet IP address of the backend server, and *80* is the health check port. By default, the port of the backend server is used as the health check port. You can configure the health check port according to your actual situation.

   - In normal conditions, `Connected to xxx . xxx . xxx . xxx` is returned. This indicates that the port on the backend server is working

(listening) normally and the health check is normal, as shown in the following figure.



- Exception example: Assume you do not change the listener configurations of the SLB instance but stop the listening process of port 80 on the backend server. Then, if you run the `telnet` command, the system prompts that the host cannot be connected. This means that the health check exception occurs if the listening process of port 80 stops, as shown in the following figure.



3. Optional: Layer-4 listeners also support HTTP health check. To use HTTP health check, see *Troubleshoot a health check exception of a Layer-7 listener (HTTP/HTTPS)*. The method for troubleshooting HTTP health check exceptions is the same for Layer-4 listeners and Layer-7 listeners.

7 Troubleshoot a health check exception of a Layer-7 listener (HTTP/HTTPS)

The health check function is used to determine whether your backend servers are normal. If a health check exception occurs, it generally means that your backend server is abnormal. The exception may also be caused by incorrect health check configurations. This topic describes how to troubleshoot a health check exception of a Layer-7 (HTTP/HTTPS) listener.

Procedure

1. **Ensure that the backend server does not block the CIDR block 100.64.0.0/10 through iptables or other third-party firewalls/security software.**

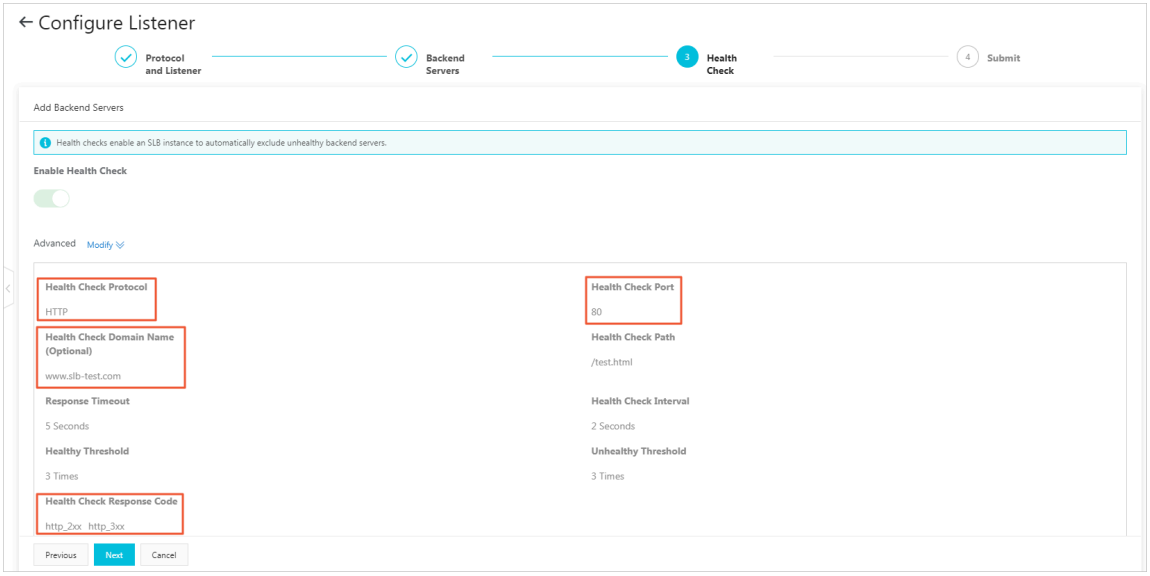
   The SLB instance communicates with backend servers by using IP addresses in the reserved CIDR block 100.64.0.0/10. If the CIDR block is blocked, a health check exception occurs and the SLB instance cannot work normally.

2. Access the HTTP service on the backend server from the backend server to ensure that the HTTP service works normally.

   a) Log on to the SLB console and click the ID of the target SLB instance. On the Listeners tab page, click Configure in the Actions column of the target listener. Then you can view the health check configurations.

   In this example, an HTTP listener is used and the intranet IP of the backend server with the health check exception is 10.0.0.2. Other health check configurations are as follows:

   - Health Check Port: 80
   - Health Check Domain Name: `www . slb - test . com`
   - Health Check Path: `/ test . html`



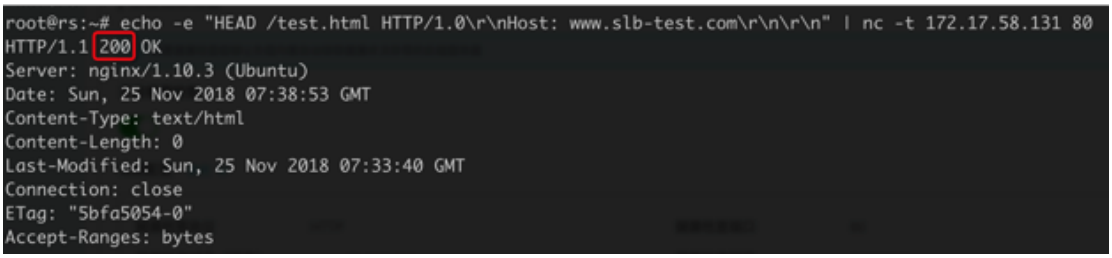
   b) The following example uses a Linux environment. Run the `nc` or `curl` command to test the HTTP service on the backend server. Ensure that the configurations of health check path, health check port, and health check

domain name are the same for the HTTP service and the backend server. Otherwise, a health check exception occurs.

In this example, the `nc` command is used. Configure the health check path, health check domain name, internet IP address, and health check port according to your actual situation.

```
echo -e "HEAD /test.html HTTP/1.0\r\nHost: www.slb-test.com\r\n\r
\n" | nc -t 172.17.58.131 80
```

- In normal conditions, `200` or `2xx / 3xx` is returned, as shown in the following figure.



- Exception example: Assume you do not change the listener configurations of the SLB instance but delete the /test.html page on the backend server. Then, when you run the `nc` command, the error code 404, instead of 2xx or 3xx, is returned, indicating a health check exception has occurred, as shown in the following figure.

8 How do I perform a stress test?

Preparations

Before you perform a stress test, we recommend the following:

- Use short-lived connections to test the forwarding performance of an SLB instance.
- Enable persistent connections to test the throughput of an SLB instance.
- Set a small timeout value, for example, five seconds, for the test tool.
- Build a static web page on backend servers.

We also recommend the following listener configurations:

- Do not enable session persistence.
- Disable the health check function.
- Use at least five clients.

Alibaba Cloud PTS tool

We recommend that you use *Alibaba Cloud PTS* as the test tool. We do not recommend Apache Bench (ab) because SLB returns inconsistent content lengths when multiple backend servers are tested, and Apache ab uses the content length to determine whether requests are successful. This will result in inaccurate stress test results.

Example stress test using Alibaba Cloud PTS

In this example, an SLB instance has been created and added with two ECS instances as backend servers. Additionally, a TCP listener and an HTTP listener have been created, and the backend port has been set to 80. The ECS instances are configured with a single-core CPU, 512 MB memory, and CentOS 6.3 (64-bit). To perform a stress test, run the following commands as follows:

1. Install Apache Web Server to provide web services.

```
yum   install  - y   httpd
```

2. Initialize the default home page index.html.

```
echo  " testvm " > / var / www / html / index . html
```

3. Start the HTTP service.

```
service   httpd   start
```

4. Visit the local port 80 to confirm that the web service is available.

```
curl   localhost
```

5. Create a test script in PTS and confirm the following parameters are set as recommended in Preparations. Then, start the stress test.
   - Set the timeout period to five seconds: PTS.HttpUtilities.setTimeout(5000)
   - Disable persistent connections: PTS.HttpUtilities.setKeepAlive(False)

Why is my Layer-7 listener performing poorly during a stress test?

Layer-4 Server Load Balancer (SLB) uses LVS (Linux Virtual Server) and Keepalived to provide the load balancing service, whereas Layer-7 SLB uses Tengine. In a Layer-4 listener, requests are directly sent to backend servers after being passed through LVS . However, in a Layer-7 listener, requests are sent to Tengine before they are sent to backend servers. Due to this additional step, the performance of a Layer-7 listener is inadequate when compared with a Layer-4 listener.

It may happen that a Layer-7 listener shows poor performance during a stress test. A Layer-7 listener with two ECS instances is surprisingly inferior to a Layer-4 listener with one ECS instance. Except the preceding reason regarding the process, the following situations may also cause low performance of a Layer-7 listener during a stress test:

- Insufficient client ports

  During a stress test, an insufficient number of client ports causes connection failures. In detail, the timestamp attribute of TCP connections is erased by SLB by

default. As a result, tw_reuse of Linux protocol stack (reuse of ports in time_wait state) does not work, and connections in time_wait state accumulate.

Solution: We recommend that you enable persistent connections on clients and use RST (set SO_LINGER attribute for sockets) to close connections, instead of using FIN packets.

- The backend server accept queue is full

  If the backend server accept queue is full, the backend server does not respond with the syn_ack packet and the client times out.

  Solution: Run the command `sysctl -w net.core.somaxconn=1024` to change the value of net.core.somaxconn and restart the application on the backend server. The default value of net.core.somaxconn is 128.

- Excessive connections to backend servers

  When you use a Layer-7 SLB instance, persistent connections are changed to short-lived connections after the connections pass through Tengine due to the design of the network architecture. As a result, there are too many connections sent to the backend server, which leads to poor performance of a Layer-7 SLB instance during a stress test.

- Limitations from dependencies of backend servers

  If backend servers work normally after requests are sent to them, but the system still exhibits poor performance, it may be caused by the dependencies of the backend servers (such as inadequate system support from the databases).

- Backend servers are unhealthy

  If a backend server is declared as unhealthy after health check results are received, or the health status of the server is unstable, such effects can also lead to poor performance of a Layer-7 SLB instance during a stress test.

9 How do I troubleshoot HTTP 5xx errors?

After an SLB instance is configured, errors such as 500 Internal Server Error, 502 Bad Gateway, and 504 Gateway Timeout may occur. They can be caused by the blockage of the service provider, Alibaba Cloud blockage caused by abnormal client activities , wrong configurations of the SLB instance, health check failure, failure in accessing web applications on the backend ECS instances, and more.

This topic lists the causes, solutions, and troubleshooting steps of these problems.

1. *Potential causes and resolutions*
   - *The source site domain name is not put on record or it is not configured with any Layer-7 forwarding rule in Anti-DDoS Pro or security network.*
   - *The source IP address of the client is blocked by Alibaba Cloud Security*
   - *The source IP address is blocked by the security protection software of the backend ECS instance*
   - *Parameter error of the Linux kernel of the backend ECS instance*
   - *Performance bottleneck of the backend ECS instance*
   - *SLB reports 502 error due to health check failure*
   - *The health check is normal but the web application reports 502 error*
   - *The HTTP header is too long*
2. *Troubleshooting*
3. *Open a ticket*

Potential causes and resolutions

1. The source site domain name is not put on record or it is not configured with any Layer-7 forwarding rule in Anti-DDoS Pro.

   Solution: Put the domain name on record. If the SLB instance is in Anti-DDoS Pro or security network, configure corresponding domain name rules.

2. The source IP address of the client is intercepted by Alibaba Cloud Security.

   Test if the problem occurs to clients of other carriers. If not, the problem is generally caused by the blockage of the service provider.

   Solution: Open a ticket to Alibaba Cloud after-sales personnel who decides if there is blockage through packet capture. If so, contact the service provider to solve the problem.
3. Blocked by the security protection software of the backend ECS instance

   100.64.0.0/10（100.64.0.0/10 is reserved by Alibaba Cloud, and will not be used by any user, there is no security risk) is the IP address range of the SLB server and is mainly used for health checks and request forwarding. If security software or a firewall inside the system is installed, add the IP address to the whitelist to avoid 500 or 502 error.

   Solution: Configure a whitelist for antivirus or firewall software, or unload the software to perform a quick test.
4. Parameter error of the Linux kernel of the backend ECS instance

   If the backend ECS instance is using the Linux system, disable the rp_filter feature in system kernel parameters when changing the Layer-7 listener to a Layer-4 listener.

   Set the values of the following parameters in the system configuration file `/ etc / sysctl . conf` to zero, and then run `sysctl - P` .

   ```
   net . ipv4 . conf . default . rp_filter = 0
   net . ipv4 . conf . all . rp_filter = 0
   net . ipv4 . conf . eth0 . rp_filter = 0
   ```
5. Performance bottleneck of the backend ECS instance

   High CPU utilization or no extra bandwidth may cause access exceptions.

   Resolution: Check the performance of the backend ECS instance to solve performance bottlenecks. If the overall system capacity is insufficient, you can increase the number of backend ECS instances.
6. SLB reports 502 error due to health check failure

   For more information, see *Resolve health check failures*.

   Besides, 502 error also occurs if the health check function of SLB is disabled and the web service in the backend server cannot process HTTP requests.

7. The health check is normal but the web application reports 502 errors

   The 502 Bad Gateway error message indicates that SLB can forward requests from the client to the backend servers, but the web application in the backend ECS instance cannot process the requests. Therefore, you must check the configurations and running status of the web application in the backend server. For example, the time used by the web application to process HTTP requests exceeds the timeout value of SLB. For Layer-7 listeners, if the time used by the backend server to process PHP requests exceeds the proxy_read_timeout of 60 seconds, SLB reports 504 Gateway Time-out. For Layer-4 listeners, the timeout value is 900 seconds.

   Solution: Make sure that the web service and related services run normally. Check if PHP requests are processed properly, and optimize the processing of PHP requests by the backend server. Take Nginx+php-fpm as an example:

   a. The number of PHP requests being processed has reached the limit.

      If the total number of PHP requests being processed in the server has reached the limit set by max_children in php-fpm, and more PHP requests are being sent to the server, then 502 or 504 errors may occur:

      - If existing PHP requests in the backend server are processed timely and new PHP requests can be processed successively, no error occurs.
      - If the existing PHP requests are not processed timely and new PHP requests remain in a waiting mode, when the value of fastcgi_read_timeout of Nginx is exceeded, a 504 Gateway timeout error occurs.
      - If the existing PHP requests are not processed in a timely manner and new PHP requests remain in a waiting mode, when the value of request_terminate_timeout in Nginx is exceeded, a 502 Bad Gateway error occurs.

   b. If the PHP script execution time exceeds the limit, namely, the time used by php-fpm to process PHP scripts exceeds the value of request_terminate_timeout in Nginx, a 502 error occurs and the following error log is shown in Nginx logs:

```
[ error ]  1760 # 0 : * 251777   recv ()  failed  ( 104 :
 Connection   reset   by   peer )  while   reading   response
 header   from   upstream ,  client :  xxx . xxx . xxx . xxx ,
 server :  localhost ,  request : “ GET  / timeoutmor  e . php
```

```
HTTP / 1 . 1 ”,  upstream : “ fastcgi :// 127 . 0 . 0 . 1 : 9000
”
```

c. The health check is performed on static pages. Errors occur when exceptions are detected in the process handling dynamic requests. For example, php-fpm is not running.

8. The HTTP header is too long

   An HTTP header that is too long may make SLB unable to process relevant data, resulting in 502 errors.

   Resolution: Decrease the amount of data transmitted by the header or change the Layer-7 listener to a Layer-4 listener.

9. Problem of service access logic

   Make sure that no backend ECS instance in SLB accesses the public IP of SLB. When the backend server accesses its own port through the IP address of SLB, the requests may be scheduled to the server itself based on the scheduling rules of SLB . This will lead to an infinite loop, thus resulting in 500 or 502 error for the requests .

   Resolution: Make sure SLB is correctly used and that no backend ECS instance is accessing the public IP of SLB.

Troubleshooting

- Check the screenshot of 500/502/504 error to determine the cause of the error. The cause of the error could be with SLB, Anti-DDoS or security network, or backend ECS instance configurations.
- If Anti-DDoS or security network is used, make sure that the Layer-7 forwarding rules are correctly configured.
- Check whether the problem occurs in all clients. If not, check whether the client indicating an error has been blocked by Alibaba Cloud Security. Also, check whether the domain name or IP of SLB is intercepted by the service provider.
- Check the status of SLB and whether there are any health check failures in any backend ECS instances. If so, resolve the detected health check failure.
- Bind the service address of SLB to the IP address of the backend server by using the hosts file on the client. If a 5xx error occurs at intervals, it is possible that a backend ECS server is not correctly configured.

- Change the Layer-7 SLB instance to a Layer-4 SLB instance to see whether the problem occurs again.
- Check the performance of backend ECS servers and whether there is performance bottleneck of the CPU, memory, disk, or bandwidth.
- If it is determined that the error is due to the backend server, check whether there are any related errors in web server logs of the backend ECS instance. Check whether the web service is running normally and whether the web access logic is correct. Test by uninstalling anti-virus software on the server and restarting the server.
- Check whether the TCP kernel parameters of the Linux system on the backend ECS instance are correctly configured.

Open a ticket

Perform the troubleshooting procedures step by step and record the test results in detail. Provide the test results when you open the ticket so that our after-sales technical support can help you solve the problem as soon as possible.

If the problem persists, contact Alibaba Cloud after-sales technical support.

10 Billing FAQ

1. How is Server Load Balancer billed?

   See *Billing method*.

2. Is the inbound traffic of SLB billed?

   No. Only the outbound traffic of SLB is billed. For more information, see *Network traffic flow*.

3. Is the traffic generated by health check billed?

   No. The traffic generated by the health check of SLB is not included in the billed traffic.

4. Will the billing of an ECS instance be affected after it is added to the backend server pool of SLB?

   No. Regardless of the billing method used for the backend ECS instance, the billing method remains the same after you associate the instance with SLB. SLB and the backend ECS instance are billed separately based on usage.

5. Is attack traffic billed?

   Integrating with Alibaba Cloud Security, SLB can defend against up to 5 Gbps DDoS attacks. From the time when the attack reaches the scrubbing or blackholing threshold to the time when the Alibaba Cloud Security starts scrubbing or blackholing, there may be a latency of several seconds. Therefore, the response to attack packets during this period will incur fees. Such an attack also consumes the bandwidth resources of SLB.

6. If all the backend ECS instances of an SLB instance are stopped, or all the backend ECS instances are removed, is the SLB billed?

   Yes. Fees may be incurred and billable based on the following:

   - Billing by traffic

     In the case of billing by traffic, no traffic fee is generated when an instance is stopped, released, or not accessed.

     SLB is a traffic distribution control service in front of backend ECS instances and provides services through its service address. If all backend ECS instances are

stopped, but the SLB instance is not stopped, inbound traffic can still reach the service address of the SLB instance. In this case, the SLB instance will respond if it discovers that there are no available backend ECS instances by performing health check.

For Layer-4 SLB, only three-way handshake packets are returned. For Layer-7 SLB, a Tengine 503 error page is prompted because the service is provided by Tengine. If there are ongoing requests, SLB will respond continuously. Such response traffic is billed.

This also applies to SLB instances with no ECS instances added. Therefore, we recommend that you stop the SLB instance if it is unnecessary in order to avoid extraneous costs.

- Billing by bandwidth

  In the case of billing by bandwidth, fees are independent of instance status and traffic usage. You are charged if you activate the service. Billing ends only after the instance is released.

11 SLB instance FAQs

- *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 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:

- Once 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.

12 Guaranteed-performance instances

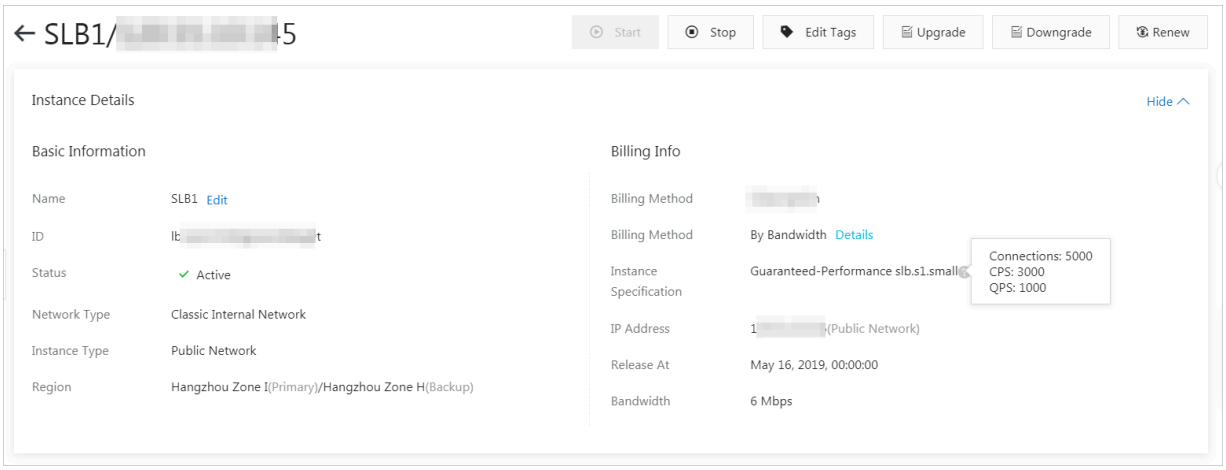
Guaranteed-performance Server Load Balancer (SLB) instances are instance types of which performance is guaranteed in terms of specific indicators, such as the maximum number of connections, Connection Per Second (CPS), and Query Per Second (QPS). Note that Alibaba Cloud now charges specification fees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

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 launched guaranteed-performance instances. You can view the instance type in the console.

You can rest the pointer over 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 an SLB instance. When the number of connections reaches the limit of the specification, new connections 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 connections 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 connections will be dropped.

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

Note:
The specifications that are available may vary according to the region. The specification displayed in the console is the actual specification.

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.

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 acutal usage. No matter what specification you choose, the specification fee will be charged according to the specification you actually use.

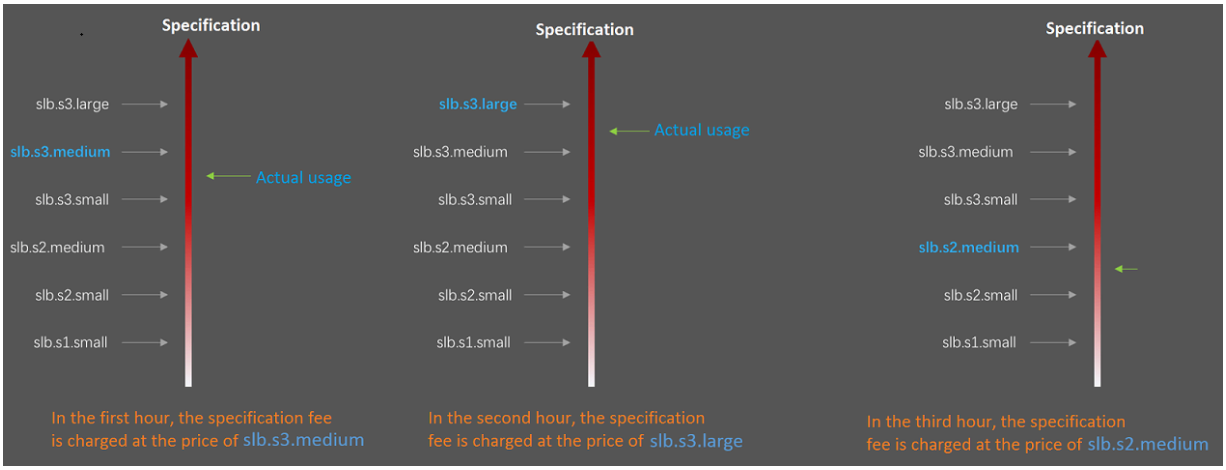
For example, if you purchase the Super I specification (Max Connection: 1,000,000; CPS: 500,000; QPS: 50,000) and the actual usage of your instance in an hour is as follows:

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 in the first three hours.



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 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 *Billing method*.

Region	Specification	Max Connectio	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

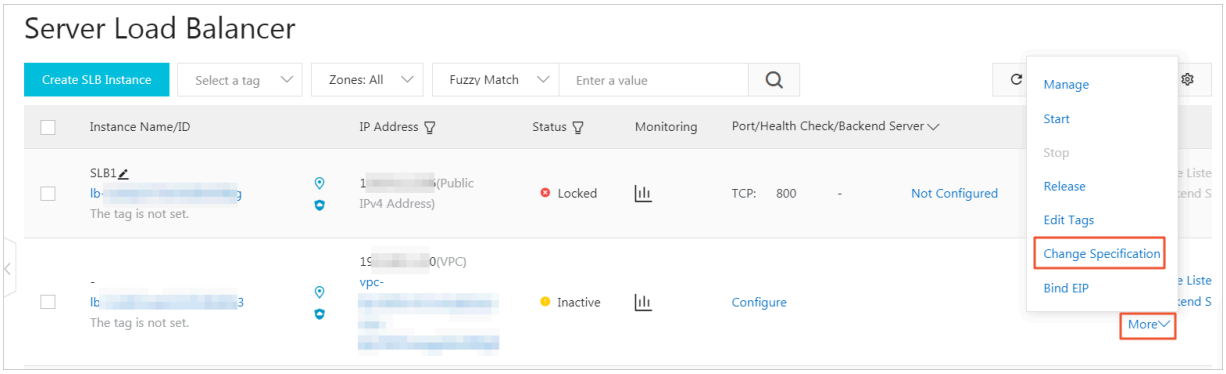
Region	Specification	Max Connectio	CPS	QPS	Specification fee (USD/ hour)
	Specification 5: Higher II (slb.s3. medium)	500,000	50,000	30,000	0.31
	Specification 6: Super I (slb.s3.large)	1,000,000	100,000	50,000	0.51
Singapore Malaysia (Kuala Lumpur) Indonesia (Jakarta) India (Mumbai) US (Silicon Valley) 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: 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.

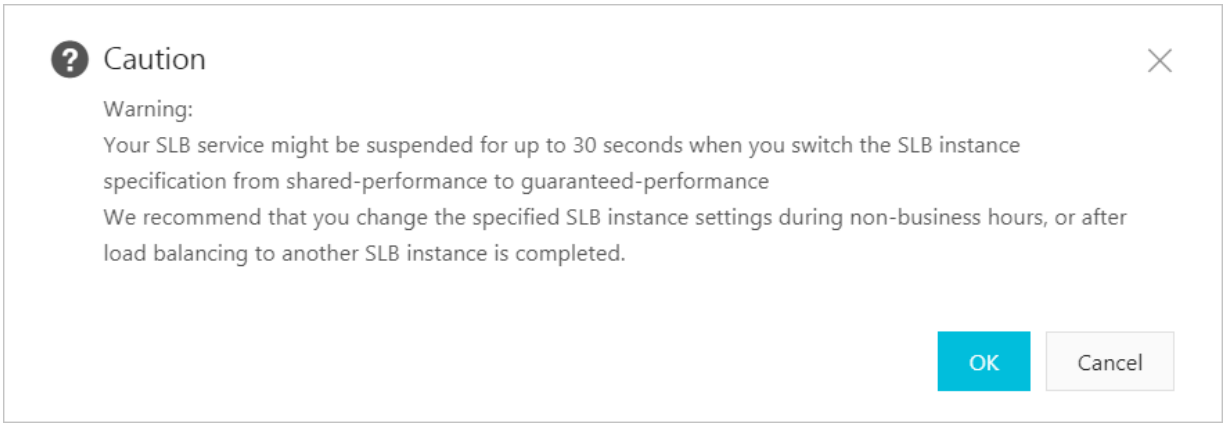
Can I change the specification of my SLB instance after it is created?

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



Note:

- Once 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 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 transfer. Therefore, for such scenarios, we recommend that you make the change when the traffic is low. Other changs of instance type or specifications do no affect your service.
- IP addresses of SLB instances will not be affected after you change the instance type or the specification.



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?

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.

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, 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, a specification fee is charged.
- If the intranet SLB instance is a guaranteed-performance instance, no specification fee is charged.

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

13 Server Load Balancer FAQs

- *Does SLB support HTTP redirection?*
- *If I disable NIC, will my SLB service be affected?*
- *Why are my connections unable to reach the peak bandwidth value?*
- *What are the timeout values of each listener?*
- *Why does my SLB connection time out?*
- *Why does session persistence sometimes fail?*
- *How can I view the session persistence string?*
- *How can I test session persistence by using Linux curl?*
- *If I disconnect my client from SLB before I receive a response from the backend servers, will SLB disconnect from the backend servers?*

Does SLB support HTTP redirection?

Yes.

SLB supports redirecting HTTP to HTTPS. For more information, see *Redirect HTTP to HTTPS*.

If I disable NIC, will my SLB service be affected?

If the ECS instance has configured a public IP address, disabling the Internet NIC will impact the load balancing service.

The traffic goes through the Internet NIC if the backend ECS is configured with a public IP address. When the Internet NIC is disabled, the returned data packet cannot be sent. We recommend that you do not disable the Internet NIC. But if you have to, you can modify the default route to intranet to avoid the impact on the service. However, you need to consider whether the business is Internet-dependent, such as accessing RDS through the Internet.

Why are my connections unable to reach the peak bandwidth value?

Because the SLB is deployed in cluster to provide the load balancing service, all requests are distributed evenly on the SLB system servers. Similarly, the specified bandwidth is also evenly distributed to these servers.

The calculation method of the traffic ceiling for a single connection download is: single connection download peak = the configured total bandwidth of Server Load Balancer / (N-1). N represents the number of traffic forwarding groups, and the current value is 4. For example, if you have set the bandwidth ceiling to 10 MB in the console, the maximum traffic for downloading of each client is 10/(4-1), or 3.33 MB.

Considering the implementation principles of Server Load Balancer, we recommend that you set a reasonable bandwidth peak value for a single listener based on your business conditions and implementation modes to eliminate negative impact and limitations on your external services.

What are the timeout values of each listener?

- TCP listener: 900 seconds
- UDP listener: 90 seconds
- HTTP listener: 60 seconds
- HTTPS listener: 60 seconds

Why does my SLB connection time out?

From the server side, the following situations may cause the connection timeout:

- The IP address of the SLB instance is protected

  Such as the blackholing and scrubbing, as well as WAF protection.

- Insufficient client ports

  Lack of client ports may lead to connection failure especially in the stress test. The SLB erases the timestamp attribute of the TCP connection. Therefore, the tw_reuse parameter does not work and the time_wait state connection heap causes the lack of the client ports.

  Solution: Do not enable TCP Keepalive for the clients and use the RST packet instead of FIN packet to terminate the connection.

- The accept queue of the backend server is full

  If the accept queue of the backend server is full, the backend server cannot sent the SYN_ACK packet. Therefore, the connection times out.

  Resolution: The default value of `net . core . somaxconn` is 128. Run the `sysctl -w net . core . somaxconn = 1024` command to change its value and restart applications on the backend servers.

- Access the Layer-4 load balancing service from the backend servers

  For the Layer-4 load balancing service, the connection fails if you access the service from a backend server.
- Improper RST configuration

  If no data is transferred within 900 seconds after the TCP connection is established on the SLB, the system will send the RST packet to the client and the backend server to terminate the connection. If the RST configuration is not correct on the backend server, the backend server may send data to a closed connection, which leads to connection timeout.

Why does session persistence sometimes fail?

Possible causes for session persistence failure, and corresponding solutions, are described as follows:

- Make sure that you have enabled session persistence.
- HTTP/HTTPS listeners cannot insert the cookies needed for session persistence into the 4xx code messages returned by backend servers.

  Solution: Change HTTP/HTTPS listeners to TCP listeners. Session persistence in TCP listeners is based on source client IP addresses, which means cookies can be inserted in backend servers.
- An HTTP 302 redirect changes the SERVERID string in the session persistence.

  When SLB inserts cookies to backend ECS instances, if the HTTP status code 302 redirect is returned by ECS instances, the SERVERID string in session persistence will be changed, resulting in session persistence failure.

  To verify the cause, check the requests and responses in your browser or by using packet checking software. Then, check whether a 302 code is included in the packets and whether the SERVERID string in the cookie is changed.

  Solution: Use TCP listeners. Session persistence in TCP listeners is based on source client IP addresses, which means cookies can be inserted into backend servers.
- The timeout period is too short. You need to modify the timeout period to a greater value.

How can I view the session persistence string?

You can press F12 in your browser to check whether the SERVERID string or any keywords that you specified are included in the response message. Alternatively, you

can run `curl www.xxx.com -c /tmp/cookie123` to save the cookie and then run `curl www.xxx.com -b /tmp/cookie123` to visit the cookie.

How can I test session persistence by using Linux curl?

1. Create a test page.

   Create a test page on each of the backend ECS instances and make sure that the intranet IP address of the server is displayed on the test page. The following figure shows an example of a test page.

2. Test using curl.

   In this example, the IP address of the SLB instance running Linux is 1.1.1.1 and the URL of the created page is *http://1.1.1.1/check.jsp*.

   a. Log on to the Linux server used for the test.

   b. Run the following command to check the value of the cookie inserted in the server.

   ```
   curl -c test.cookie http://1.1.1.1/check.jsp
   ```

   > **Note:**
   > By default, SLB maintains session persistence by inserting cookies. However, curl does not save or send any cookies. Therefore, you must save the corresponding cookie first. Otherwise, the curl test result may mistakenly determine that session persistence has failed.

   c. Run the following command to continue the test.

   ```
   for ((a=1;a<=30;a++));2="" do="" curl="" -b
   ="" 1.cookie=""
   check.jsp="">/dev/null | grep '10.170
   .*';sleep 1;done`
   ```

   > **Note:**
   > In `a<=30`, `30` is the number of repeated tests and can be changed to your required testing number. In `grep '10.170.*'`, `10.170.*` is the IP

address to be searched and you can change it to the intranet IP address of your server.

d. Check the IP addresses returned by the preceding tests. If they are the same address, then session persistence is working. If the addresses are different, session persistence has failed.

If I disconnect my client from SLB before I receive a response from the backend servers, will SLB disconnect from the backend servers?

No. SLB does not disconnect from backend servers during read/write operations.

14 Backend server FAQs

- *Can I adjust the number of backend ECS instances while my SLB instance is running?*
- *Can I use different operating systems for different backend ECS instances?*
- *Can I add ECS instances from different regions to the same SLB instance?*
- *Why do my records show frequent access to my backend ECS instances from IP addresses that start with 100?*
- *Why are responses returned by SLB compressed even though my ECS instance is not configured for compression?*
- *Is chunked transfer encoding supported if my backend ECS instances use HTTP1.0?*
- *Why do my backend ECS instances frequently receive requests where the value of the UA string is KeepAliveClient?*

Can I adjust the number of backend ECS instances while my SLB instance is running?

Yes.

You can increase or decrease the number of backend ECS instances in an SLB instance at any time and switch between different ECS instances. Before you perform these operations, make sure that health check is enabled and that there is at least one normally running backend ECS instance to avoid service interruption.

Can I use different operating systems for different backend ECS instances?

Yes.

There is no limitation on the operating system used on backend ECS instances as long as applications deployed on the ECS instances are the same and the data is consistent. To facilitate daily management and maintenance, we recommend that you use the same operating system for backend ECS instances.

Can I add ECS instances from different regions to the same SLB instance?

No.

Server Load Balancer does not support cross-region deployment. The ECS instances to be added must belong to the same region as the SLB instance.

Why do my records show frequent access to my backend ECS instances from IP addresses that start with 100?

In addition to forwarding external requests to backend ECS instances by using the intranet IP address of the system server, the SLB system also accesses the ECS instances to perform health checks and monitor service availability.

The IP address range of the SLB system is 100.64.0.0/10（100.64.0.0/10 is reserved by Alibaba Cloud, and will not be used by any user, there is no security risk) , so there are many IP addresses beginning with 100 accessing ECS instances.

To guarantee the service availability, you have to configure appropriate access rules for these IP address ranges.

Why are responses returned by SLB compressed even though my ECS instance is not configured for compression?

The possible reason is that the client web browser supports compression. You can disable Gzip function when creating listeners in the console or use TCP listeners instead.

Is chunked transfer encoding supported if my backend ECS instances use HTTP1.0?

Yes.

Why do my backend ECS instances frequently receive requests where the value of the UA string is KeepAliveClient?

Issue

Backend ECS instances frequently receive GET requests, but there are no visitor IP addresses. Instead, source IP addresses of these requests are intranet IP addresses of Alibaba Cloud, and the value of the User-Agent string is KeepAliveClient.

Cause

TCP listeners are being used, which use the HTTP protocol for health checks. Specifically, when health checks that use HTTP protocol are performed in TCP listeners, GET requests are used by default.

Solution

We recommend that you use the same protocol for both listeners and health checks.

15 Heath check FAQs

- *How does the health check function of Server Load Balancer (SLB) work?*
- *What are the recommended configurations for health checks in SLB?*
- *Can I disable the health checks?*
- *What is the recommended health check method for TCP listeners?*
- *Is there any impact to health checks if the weight of an ECS instance is zero?*
- *What health check method is used for HTTP listeners on backend ECS instances?*
- *7. What are the ranges of IP address that HTTP listeners use to perform health checks on backend ECS instances?*
- *Why is the health check frequency that is displayed on the console different from that recorded in the web logs?*
- *Do health checks use system resources?*
- *How can I handle a health check failure caused by a faulty backend database?*
- *Why is a network connection exception recorded in the backend service logs, but the TCP health check is displayed as successful?*
- *Why is the health check result returned as abnormal when the service is running normally?*

How does the health check function of Server Load Balancer (SLB) work?

SLB checks the service availability of backend servers (ECS instances) by performing health checks on backend servers. When SLB determines that an instance is unhealthy, it stops distributing requests to that instance. Distributions to that instance will resume when it becomes healthy again.

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).

For more information, see *Health check overview*.

What are the recommended configurations for health checks in SLB?

However, to avoid the impact of switching caused by frequent health check failures on the system availability, status is switched (health check succeeded or failed) only when the health check continuously succeeds or fails for multiple times in the time window. For more information, see *Health check configurations*.

The following are recommended health check configurations for TCP/HTTP/HTTPS listeners.

Configuration	Recommended value
Response timeout	5 seconds
Health check Interval	2 seconds
Unhealthy threshold	3 times

The following are recommended health check configurations for UDP listeners.

Configuration	Recommended value
Response timeout	10 seconds
Health check Interval	5 seconds
Unhealthy threshold	3 times
Healthy threshold	3 times

Note:

These configurations are conducive to restoring the service when the health check of a backend server fails. If you have higher requirements, you can specify a smaller response timeout value, but you need to make sure that the processing time in the normal status is less than the specified timeout value.

Can I disable the health checks?

You can only disable health check for HTTP and HTTPS listeners. The health check of UDP and TCP listeners cannot be disabled. For more information, see *Close health check*.

Note:

When the health check function is disabled, requests may be distributed to unhealthy ECS instances, which can lead to service interruption. We do not recommend disabling health checks.

What is the recommended health check method for TCP listeners?

For TCP listeners, both TCP health checks and HTTP health checks are supported:

- TCP health checks send SYN handshake packets to backend servers to check whether the ports of backend servers are normal.
- HTTP health checks detect the health status of applications on backend servers by sending HEAD and GET requests to simulate visits from the browser of a user.

The TCP health check minimally impacts performance and consumes less resources on the backend ECS instances. Choose TCP health checks if the traffic load on the backend ECS instance is large, and choose HTTP health checks if not.

Is there any impact to health checks if the weight of an ECS instance is zero?

In this situation, SLB will no longer forward traffic to this ECS instance and the health check will indicate abnormal for the Layer-4 listeners, while the health check is normal for Layer-7 listeners.

Setting the weight value to zero is equal to manually removing the ECS instance from Server Load Balancer. The weight is set to zero only when restarting, adjusting, or maintaining the ECS instance.

What health check method is used for HTTP listeners on backend ECS instances?

HEAD method

If the backend ECS instances disable the HEAD method access, health checks on the backend ECS instances will fail. We recommend you access your own IP address using the HEAD method on the ECS instance for testing:

```
curl -v -0 -I -H " Host :" -X HEAD http :// IP : port
```

7. What are the ranges of IP address that HTTP listeners use to perform health checks on backend ECS instances?

The IP address range used by the health check of SLB is 100.64.0.0/10（100.64.0.0/10 is reserved by Alibaba Cloud, and will not be used by any user, there is no security risk) . If the backend ECS instance enables access control such as iptables, you need to allow the access of 100.64.0.0/10 on the intranet NIC.

Why is the health check frequency that is displayed on the console different from that recorded in the web logs?

Health checks are performed in the cluster to avoid single points of failure. Therefore , the health check frequency recorded in the logs is different from the frequency configured in the console.

Do health checks use system resources?

HTTP health checks consume few resources of the backend ECS instances.

How can I handle a health check failure caused by a faulty backend database?

Symptoms:

Two web sites are configured on an ECS instance. The website www.test.com is a static website, and the website app.test.com is a dynamic website. A 502 error occurred when accessing www.test.com due to a backend database fault.

Cause:

Domain name app.test.com is configured for health check. RDS or self-built database failure causes the access error to app.test.com, so the health check fails.

Resolution:

Configure the domain name used for health checks to www.test.com.

Why is a network connection exception recorded in the backend service logs, but the TCP health check is displayed as successful?

Symptoms:

After configuring the backend TCP port in a SLB listener, a network connection exception is frequently shown in the backend service logs. The requests are sent from the SLB instance and the SLB instance also sends RST packets to the backend server at the same time.

```
java.io.IOException: Connection reset by peer
        at sun.nio.ch.FileDispatcherImpl.read0(Native Method)
        at sun.nio.ch.SocketDispatcher.read(SocketDispatcher.java:39)
        at sun.nio.ch.IOUtil.readIntoNativeBuffer(IOUtil.java:223)
        at sun.nio.ch.IOUtil.read(IOUtil.java:192)
        at sun.nio.ch.SocketChannelImpl.read(SocketChannelImpl.java:379)
        at io.netty.buffer.UnpooledUnsafeDirectByteBuf.setBytes(UnpooledUnsafeDirectByteBuf.java:446)
        at io.netty.buffer.AbstractByteBuf.writeBytes(AbstractByteBuf.java:871)
        at io.netty.channel.socket.nio.NioSocketChannel.doReadBytes(NioSocketChannel.java:225)
        at io.netty.channel.nio.AbstractNioByteChannel$NioByteUnsafe.read(AbstractNioByteChannel.java:115)
        at io.netty.channel.nio.NioEventLoop.processSelectedKey(NioEventLoop.java:507)
        at io.netty.channel.nio.NioEventLoop.processSelectedKeysOptimized(NioEventLoop.java:464)
        at io.netty.channel.nio.NioEventLoop.processSelectedKeys(NioEventLoop.java:378)
        at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:350)
        at io.netty.util.concurrent.SingleThreadEventExecutor$2.run(SingleThreadEventExecutor.java:116)
        at java.lang.Thread.run(Thread.java:745)
```

Cause:

The problem is related to the health check mechanism.

TCP is transparent to the upper-Layer applications and is utilized to reduce the cost of health checks and the impact on backend service. TCP health checks only perform a simple three-way handshake and then directly send RST packets to terminate the TCP connection. The data exchange process is as follows:

1. The SLB instance sends a SYN packet to the backend port.
2. The backend server replies with a SYN-ACK if the backend port is in normal status.
3. After successfully receiving the response from the backend port, the SLB instance considers that the port is in normal status and the status of the backend server is normal.
4. The SLB instance sends a RST packet to the backend port to actively terminate the connection. For now, health check is completed.

After the health check succeeds, the SLB instance directly sends RST packets to terminate the connection and no data is sent afterwards. Therefore, upper-Layer services (such as Java connection pool) deem that the connection is abnormal and errors such as `Connection reset by peer` are displayed.

Resolution:

- Use HTTP protocol instead.
- In the service layer, filter the logs from the SLB IP address range and ignore related error messages.

Why is the health check result returned as abnormal when the service is running normally?

Symptoms:

The HTTP health check always fails, but the status code obtained by performing the `curl -I` test is normal as follows:

```
echo -e 'HEAD /test.html HTTP/1.0\r\n\r\n' |
nc -t 192.168.0.1 80
```

Cause:

If the returned status code is different from the normal status code configured on the console, the backend ECS instance is declared as unhealthy. For example, if

the configured normal status code is http_2xx, all other status codes returned not matching this status code will be considered as health check failure.

No error occurred when a curl test is performed on the Tengine/Nginx cluster, but a 404 error occurred in the `test.html` test file because the default site is used in the echo test.

Resolution:

- **Modify the main configuration file and annotate the default site.**
- **Add the domain name used for health check in the health check configurations.**

16 HTTPS and HTTP listener FAQs

- *Why are some response header parameters deleted after requests are forwarded by Layer-7 listeners?*
- *Why is an additional header, namely the Transfer-Encoding: chunked, added to an HTTP request?*
- *Why do the style sheets fail to load when I open a website through an HTTPS listener?*
- *Which port number do HTTPS listeners use?*
- *What types of certificates does SLB support?*
- *Does SLB support keytool-created certificates?*
- *Can I use certificates in the PKCS#12(PFX) format?*
- *Why does a KeyEncryption error occur when uploading certificates?*
- *What SSL protocol versions are supported by the HTTPS Server Load Balancer service?*
- *What is the lifetime of an HTTPS session ticket?*
- *Can I upload a certificate containing DH PARAMETERS?*
- *Do HTTPS listeners support SNI?*
- *Which HTTP version is used by HTTP and HTTPS listeners to access the backend servers?*
- *Can backend ECS instances obtain the protocol version used by the client to access the HTTP or HTTPS listener?*
- *What are the timeout values specified for HTTP and HTTPS listeners?*

Why are some response header parameters deleted after requests are forwarded by Layer-7 listeners?

Symptoms: SLB modifies the values of the Date, Server, X-Pad, X-Accel-Redirect and other parameters in the response headers to achieve session persistence.

Solution:

- Add a prefix to the custom header, such as xl-server or xl-date.
- Change the Layer-7 listener to a Layer-4 listener.

Why is an additional header, namely the Transfer-Encoding: chunked, added to an HTTP request?

Symptoms: After a domain name is resolved into the IP address of a Layer-7 SLB instance, a Transfer-Encoding: chunked field is added in the HTTP request header

when accessing the domain name from a local host. However, no such field is found in the request when accessing backend servers directly from the local host.

Layer-7 SLB is based on the Tengine reverse proxy. The Transfer-Encoding field indicates how the Web server encodes the response message body. For example, Transfer-Encoding: chunked indicates the chunked transfer encoding is used.

> Note:
> This header is not added in the requests forwarded by Layer-4 listeners, because Layer-4 listeners only distribute traffic.

Why do the style sheets fail to load when I open a website through an HTTPS listener?

Symptoms:

An HTTP listener and an HTTPS listener are created respectively, and they use the same backend servers. When accessing the website over the HTTP listener with the specified port number, the website is displayed normally. However, the website layout is messy when accessing the website through the HTTPS listener.

Cause:

By default, SLB does not block loading and transferring JavaScript files. The possible reasons are as follows:

- The certificate is not compatible with the security level of the web browser.
- The certificate is an unqualified third-party certificate. In this case, contact the certificate issuer to check the certificate.

Solution:

1. When you open the website, click the prompt in the browser's address bar to load the script.
2. Add the required certificate to the browser.

Which port number do HTTPS listeners use?

There are no special requirements on ports. However, we recommend that you use 443 as the port number for HTTPS listeners.

What types of certificates does SLB support?

SLB supports uploading server certificates and CA certificates in the PEM format.

For the server certificates, you must upload both the certificate content and the private key. For the CA certificates, you only need to upload the certificate content.

Does SLB support keytool-created certificates?

Yes.

However, you must convert the certificate format to PEM before uploading the certificate to SLB. For more information, see *Convert certificate format*.

Can I use certificates in the PKCS#12(PFX) format?

Yes.

However, you must convert the certificate format to PEM before uploading the certificate to SLB. For more information, see *Convert certificate format*.

Why does a KeyEncryption error occur when uploading certificates?

The private key contains incorrect contents. For more information on private key format, see *Certificate formats*.

What SSL protocol versions are supported by the HTTPS Server Load Balancer service?

TLSv1, TLSv1.1, and TLSv1.2.

What is the lifetime of an HTTPS session ticket?

The lifetime of an HTTPS session ticket is set to 300 seconds.

Can I upload a certificate containing DH PARAMETERS?

No. The ECDHE method used by HTTPS listeners supports forward secrecy, but does not support uploading the PEM files that contain the security enhancement parameters, such as BEGIN DH PARAMETERS.

Do HTTPS listeners support SNI?

Yes. SNI (Server Name Indication) is an extension to SSL/TLS protocol so that a server can use multiple domain names and certificates. SLB HTTPS supports the SNI function. For more information, see *Configuration tutorial*.

Which HTTP version is used by HTTP and HTTPS listeners to access the backend servers?

- When the protocol used by client requests is HTTP/1.1 or HTTP2/0, Layer-7 listeners use HTTP/1.1 to access backend servers.

- When the protocol used by client requests is neither HTTP/1.1 or HTTP2/0, Layer-7 listeners use HTTP/1.0 to access backend servers.

Can backend ECS instances obtain the protocol version used by the client to access the HTTP or HTTPS listener?

Yes.

What are the timeout values specified for HTTP and HTTPS listeners?

- A maximum of 100 requests can be sent continuously in an HTTP persistent connection. The connection is closed when the limit is reached.
- The timeout between two HTTP or HTTPS requests in an HTTP persistent connection is 15 seconds. The TCP connection is closed when the timeout exceeds 15 seconds. If you want to use the HTTP persistent connection, try to send heartbeat requests within 13 seconds.
- The timeout for the TCP three-way handshake between SLB and a backend ECS instance is 5 seconds. After the handshake times out, SLB selects the next ECS instance. You can find the timeout by checking the upstream response time in the access logs.
- The time that SLB waits for the response from an ECS instance is 60 seconds. If the wait time exceeds 60 seconds, a 504 or 408 status code is sent to the client. You can find the timeout by checking the upstream response time in the access logs.
- The HTTPS session reuse times out after 300 seconds. After the timeout, the client needs to perform the complete SSL handshake process again.

17 WS and WSS support FAQs

What are WS and WSS?

WebSocket is a new HTML5 protocol, which provides full-duplex communication between the browser and the server. This protocol conserves server resources and bandwidth, enabling real-time communication. WebSocket is built on top of TCP and transmits data over TCP like HTTP.

One major difference between WebSocket and HTTP is that WebSocket is a two-way communication protocol. Once the connection is established, both the WebSocket server and the client can send data to or receive data from each other actively like Socket. The WebSocket server and client have to complete a handshake to establish a WebSocket connection.

WebSocket Secure (WSS) is the encrypted version of WebSocket.

Why use WS and WSS?

With the increasing popularity and accessibility of the Internet, a multitude of varied web applications are emerging. Many applications require real-time push capabiliti es of the server (such as broadcast rooms and chat rooms). In the past, many websites used the round robin technique to achieve real-time push. With the round robin technique, the browser sends HTTP requests to the server at specific intervals (for example, per second) and the server returns the most recent data to the browser of the client. However, this method has an obvious disadvantage of inefficiency. The browser must send requests constantly to the server. The headers of HTTP requests may be very long with only a few effective messages. Therefore, many bandwidth resources are potentially wasted.

In this situation, HTML5 defines the WebSocket protocol, which can help conserve server resources and bandwidth and facilitate real-time communication. WebSocket provides the full-duplex communication between the browser and the server. This allows the server to send data to the client actively without being solicited by the client.

The communication process of the WebSocket protocol is shown in the following figure:

How can I enable WS and WSS on Server Load Balancer?

No configuration is required. The HTTP listener supports the WS protocol and the HTTPS listener supports WSS protocol by default.

Note:

You must upgrade the instance to a guaranteed-performance instance. For more information, see *How to use guaranteed-performance instances*.

Supported regions

The WS and WSS support is available in all regions.

Limits

The limitations for WS and WSS protocol are as follows:

- Server Load Balancer is connected to backend ECS instances by using HTTP/1.1. We recommend that backend servers use a web server that supports HTTP/1.1.
- If there is no message interaction between Server Load Balancer and a backend ECS instance within 60 seconds, the connection is terminated. If you need to maintain the connection, enable Keepalive to ensure message interaction at the frequency of once every 60 seconds.

Billing

WS and WSS protocol support is free of charge.