Alibaba Cloud
Server Load Balancer

Miscellaneous

Issue: 20181107

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

Contents

1 Best practices

1.1 Use Open API to configure Server Load Balancer

Prerequisites

You have created two ECS instances and granted access to their SSH and Web ports.

Context

In this tutorial, the request parameters are included in the request URL, and the URL does not include common parameters. For more information, see *API overview*.

Note:
To increase readability, the parameter values of the request URL in this example are not URL-encoded.

```
server.modules = ( "mod_setenv" )
$HTTP["host"] == "test.example.com" {
      server.document-root = "/var/www/html/"
      setenv.add-response-header = ( "Set-Cookie" => "name=XXXXXX"
  }
}
```

Procedure

1. Call the **CreateLoadBalancer** API to create a Server Load Balancer instance.

   Request:

   https://slb.aliyuncs.com/?Action=CreateLoadBalancer&RegionId=cn-hangzhou-dg-a01

   Response:

   ```
   {
   "RequestId":"3DE96B24-E2AB-4DFA-9910-1AADD60E13A5",
   "LoadBalancerId":"LoadBalancerId",
   "Address":"SLBIPAddress"
   }
   ```

2. Call the **CreateLoadBalancerHttpListener** API to create a HTTP listener, of which the port is 80, for the Server Load Balancer instance.

   Request:

   https://slb.aliyuncs.com/?Action=CreateLoadBalancerHttpListener&LoadBalancerId=LoadBalancerId&ListenerPort=80&BackendServerPort=80&ListenerStatus=active

3. Call the **SetLoadBalancerStatus** API to activate the Server Load Balancer instance.

Request:

https://slb.aliyuncs.com/?Action=SetLoadBalancerStatus&LoadBalancerId=LoadBalancerId&LoadBalancerStatus=active

4. Call the **AddBackendServers** API to add an ECS instance to the Server Load Balancer instance.

   Request:

   https://slb.aliyuncs.com/?Action=AddBackendServers&LoadBalancerId=LoadBalancerId&BackendServers=[{"ServerId":"ECS1InstanceID"}

   Response:

```
{
     "RequestId" : "FA2F2172-63F2-409D-927C-86BD1D536F13",
     "LoadBalancerId" : "LoadBalancerId",
     "BackendServers" : {
         "BackendServer" : [
             {
                 "ServerId" : "ECS1InstanceId",
                 "Weight" : 100
             }
         ]
     }
 }
```

5. Call the **AddBackendServers** API again to add another ECS instance to the Server Load Balancer instance.

   Request:

   https://slb.aliyuncs.com/?Action=AddBackendServers&LoadBalancerId=LoadBalancerId&BackendServers=[{"ServerId":"ECS2InstanceID"}

   Response:

```
 {
     "RequestId" : "C61FAD0A-2E87-4D0C-80B0-95AB758FCA70",
     "LoadBalancerId" : "LoadBalancerId",
     "BackendServers" : {
     "BackendServer" : [
         {
             "ServerId" : "ECS1InstanceId",
             "Weight" : 100
         },
         {
             "ServerId" : "ECS2InstanceId",
             "Weight" : 100
          }
       ]
     }
```

```
}
```

6. Call the **DescribeLoadBalancerAttribute** API to view the configuration of the Server Load Balancer instance.

   Request:

   https://slb.aliyuncs.com/?Action=DescribeLoadBalancerAttribute&LoadBalancerId=LoadBalancerId

   Response:

```
{
    "RequestId" : "4747E9AE-ADFD-412D-B523-C1CBD45A2154",
    "LoadBalancerId" : "LoadBalancerId",
    "Address" : "SLBIPAddress",
    "IsPublicAddress" : "true",
    "ListenerPorts" : {
        "ListenerPort" : [
            80
        ]
    },
    "BackendServers" : {
        "BackendServer" : [
            {
                "ServerId" : "ECS1InstanceId",
                "Weight" : 100
            },
            {
                "ServerId" : "ECS2InstanceId",
                "Weight" : 100
            }
        ]
    }
}
```

   Use your browser to access the IP address of the Server Load Balancer instance to verify whether the service is working.

1.2 Server Load Balancer + ECS multi-site deployment

You can build multiple sites on an ECS instance by creating virtual sub-directories. Server Load Balancer supports adding this type of ECS instance as a backend server to forward traffic. Both layer-4 and layer-7 Server Load Balancer support adding multi-site ECS instances and the function is not influenced by session persistence.

Context

This tutorial illustrates how to create a TCP listener and add an ECS instance deployed with two sites.

Procedure

1. On the ECS01, create the site www.aaa.com and the site www.bbb.com through creating sub-directories.

   The content displayed on the site www.aaa.com is shown as the following figure.

   The content displayed on the site www.bbb.com is shown as the following figure.

   You can bind the two domain names to the public IP address of the ECS instance in the local host file and use the browser to visit ww.aaa.com and www.bbb.com. If the preceding contents are displayed, the configuration is successful.
2. Create a custom image for the ECS01 and use the image to create ECS02 with the same configuration.
3. Create an Internet Server Load Balancer instance and add a TCP listener.
4. Add the ECS01 and the ECS02 as backend servers.
5. In the local host file, resolve the domain name www.aaa.com and the domain name www.bbb.com to the public IP address of the Server Load Balancer instance.
6. Enter www.aaa.com and www.bbb.com in the browser. If the corresponding backend services can be accessed, the deployment is successful.

1.3 Handle excessive logs caused by health check

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 occurred 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 check uses 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 find the target listener and click **Configure**.
3. On the **Configure Listener** page, click **Next** to configure health check.
4. Adjust the **Health Check Interval** in the range of 1 to 50 seconds. The greater the interval, the lower the health check frequency and the fewer logs generated by the backend server. Modify 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 amount of logs.

Potential risks

After you close HTTP/HTTPS health check, Server Load Balancer does not check backend servers. Once 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 find the target listener and click **Configure**.
3. On the **Configure Listener** page, click **Next** to configure health check.
4. Close **Enable Health Check**.
5. Click **OK**.

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

Layer-4 health check only uses TCP three-way handshakes and generates no application 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 occurred 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 find the target listener and click **Configure**.
3. On the **Configure Listener** page, click **Next** to configure health check.
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 applicatio n 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 tutorial takes ngix 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 find the target listener and click **Configure**.
4. On the **Configure Listener** page, click **Next** to configure health check.
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**.

1.4 Configure cookie in the backend server

Server Load Balancer provides session persistence function. With session persistence enabled, Server Load Balancer can distribute requests from the same client to the same backend server during the session period.

For layer-4 listeners, session persistence is based on the IP address. The listener of Server Load Balancer forwards requests from the same IP address to the same backend server. For layer-7 listeners, session persistence is based on cookies.

If you choose to rewrite the cookie, you must configure the cookie on the backend server. Suppose there are two domain names under your Server Load Balancer service: vip.a.com and img.a.com. If you want to configure session persistence for vip.a.com, you can set the cookie name to name, and set a cookie of which the key is name for vip.a.com on the backend server.

Follow the instructions in this section to set cookies on a backend server.

Apache

1. Open the *httpd.conf* file and make sure that the following line is not commented.

```
LoadModule usertrack_module modules/mod_usertrack.so
```

2. Add the following configurations in the VirtualHost file.

```
CookieName name
CookieExpires "1 days"
CookieStyle Cookie
CookieTracking on
```

Nginx

Configure the cookie as follows.

```
server {
    listen 8080;
    server_name wqwq.example.com;
    location / {
      add_header Set-Cookie name=xxxx;
        root html;
        index index.html index.htm;
    }
}
```

Lighttpd

Configure the cookie as follows.

```
    server.modules  = ( "mod_setenv" )
    $HTTP["host"] == "test.example.com" {
          server.document-root = "/var/www/html/"
          setenv.add-response-header = ( "Set-Cookie" => "name=XXXXXX"
      }
    }
```

1.5 Obtain the real IP address of the client

Introduction to the function of obtaining IP address

Alibaba Cloud Server Load Balancer provides the function of obtaining the real IP address of the client and this function is enabled by default.

- For the Layer-4 load balancing service (TCP protocol), listeners distribute client requests to backend ECS servers without modifying the request headers. Therefore, you can obtain the real IP address from the backend ECS servers without additional configurations.

- For the Layer-7 load balancing service (HTTP/HTTPS protocol), you have to configure the application servers, and then use the X-Forwarded-For header to obtain the real IP addresses of the clients.

  The real client IP is put in the X-Forwareded-For field of the HTTP header in the following format:

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

  When this method is used to obtain the real IP of the client, the first IP obtained is the real IP of the client.

  Note:
  For the HTTPS load balancing service, the SSL certificates are configured in front-end listeners, the backend still uses the HTTP protocol. Therefore, the configurations on application servers 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 of your server to a specified directory, such as *C:\F5XForwardedFor*. Make sure that the IIS process has the write permission to this folder.
3. Open the **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 the 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 &lt;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 (100.64.0.0/10 is reserved by Alibaba Cloud, and will not be used by any user, there is no security risk)) and the address range of Anti-DDoS Pro. Separate multiple CIDR blocks by commas.

3. Restart Apache after the adding.

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

Configure the Nginx server

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_module --without-http-cache --with-
http_ssl_module --with-http_realip_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 behind the following configuration information.

```
fastcgi connect_timeout 300;
fastcgi send_timeout 300;
fastcgi read_timeout 300;
fastcgi buffer_size 64k;
fastcgi buffers 4 64k;
fastcgi busy_buffers_size 128k;
fastcgi temp_file_write_size 128k;
```

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

```
set_real_ip_from IP_address
real_ip_header 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 (100.64.0.0/10 is reserved by Alibaba Cloud, and will not be used by any user, there is no security risk)) and the address range of Anti-DDoS Pro. Separate multiple CIDR blocks by commas.

4. Restart Nginx.

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

1.6 Remove backend ECS instances

Directly removing backend ECS instances from a Server Load Balancer instance may cause service interruption. We recommend setting the weight of an ECS instance to zero first, and then remove it when no traffic is distributed to it.

Remove backend ECS instances

Directly removing backend ECS instances from a Server Load Balancer instance may cause service interruption. We recommend that you follow these steps to remove ECS instances:

1. Log on to the *Server Load Balancer console*.
2. Choose a region and then click the ID of the target Server Load Balancer instance.
3. In the **Details** left-side navigation pane, click **Server** > **Backend Server**.

If the ECS instance is added to a server group, click the VServer group or master-slave server group accordingly.

4. Hover the mouse pointer to the weight of the target ECS instance and then set the value to 0.

5. When no traffic is distributed to the ECS instance, click **Remove** to remove it from the backend server pool.

Troubleshoot

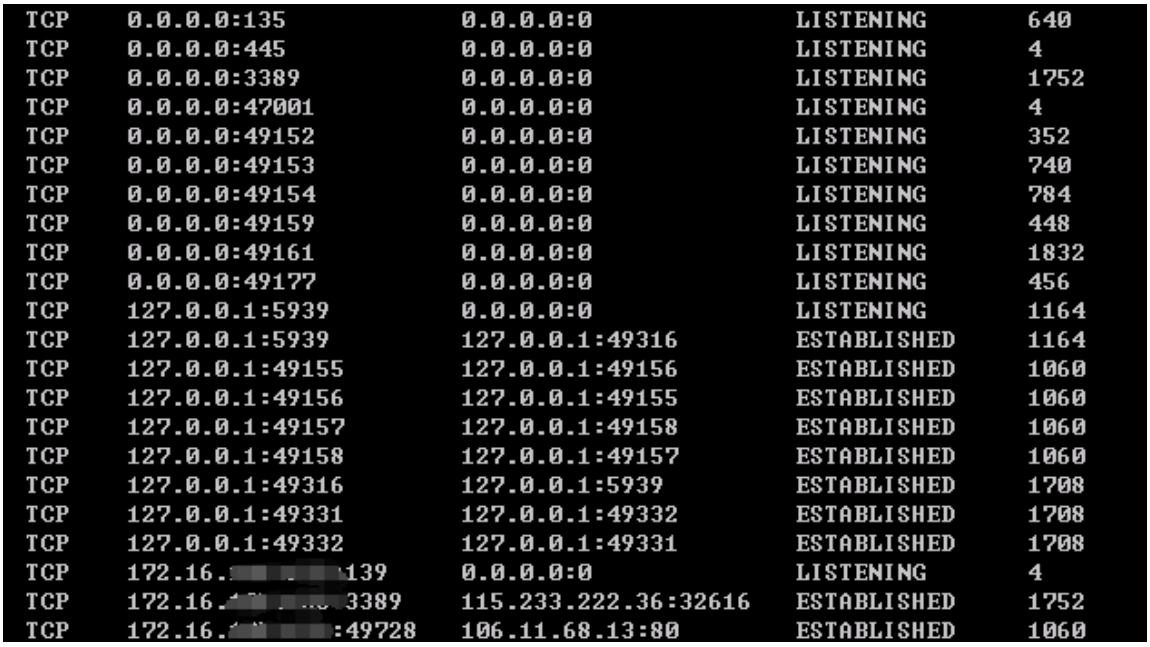
If there are ongoing service requests sent to the ECS instance after removing it from the backend server pool, check the following:

- Whether the ECS instance is added to backend server pools of other Server Load Balancer instances.

  You can use the ECS instance ID to filter Server Load Balancer instances that the ECS instance is added to.

- Log on to the ECS instance, run the `netstat` command to check whether the ECS instance is deployed with public services.

  - Windows: Run `netstat -ano` to view all open ports on the instance.



```
TCP    0.0.0.0:135            0.0.0.0:0              LISTENING       640
TCP    0.0.0.0:445            0.0.0.0:0              LISTENING       4
TCP    0.0.0.0:3389           0.0.0.0:0              LISTENING       1752
TCP    0.0.0.0:47001          0.0.0.0:0              LISTENING       4
TCP    0.0.0.0:49152          0.0.0.0:0              LISTENING       352
TCP    0.0.0.0:49153          0.0.0.0:0              LISTENING       740
TCP    0.0.0.0:49154          0.0.0.0:0              LISTENING       784
TCP    0.0.0.0:49159          0.0.0.0:0              LISTENING       448
TCP    0.0.0.0:49161          0.0.0.0:0              LISTENING       1832
TCP    0.0.0.0:49177          0.0.0.0:0              LISTENING       456
TCP    127.0.0.1:5939         0.0.0.0:0              LISTENING       1164
TCP    127.0.0.1:5939         127.0.0.1:49316        ESTABLISHED     1164
TCP    127.0.0.1:49155        127.0.0.1:49156        ESTABLISHED     1060
TCP    127.0.0.1:49156        127.0.0.1:49155        ESTABLISHED     1060
TCP    127.0.0.1:49157        127.0.0.1:49158        ESTABLISHED     1060
TCP    127.0.0.1:49158        127.0.0.1:49157        ESTABLISHED     1060
TCP    127.0.0.1:49316        127.0.0.1:5939         ESTABLISHED     1708
TCP    127.0.0.1:49331        127.0.0.1:49332        ESTABLISHED     1708
TCP    127.0.0.1:49332        127.0.0.1:49331        ESTABLISHED     1708
TCP    172.16.[redacted]139   0.0.0.0:0              LISTENING       4
TCP    172.16.[redacted]3389  115.233.222.36:32616   ESTABLISHED     1752
TCP    172.16.[redacted]:49728  106.11.68.13:80      ESTABLISHED     1060
```

  - Linux: Run this command to view all open ports on the instance or use other parameters of the `netstat` command.

1.7 High availability best practice

Server Load Balancer (SLB) guarantees availability in terms of system design, product configuration and so on. Besides, you can use SLB together with Alibaba Cloud DNS to achieve cross-region disaster recovery according to business needs.

High availability of SLB system

Deployed in clusters, Server Load Balancer (SLB) can synchronize sessions to protect the ECS instances from single points of failure (SPOFs). This improves redundancy and guarantees the service stability. Layer-4 SLB uses the open source software Linux Virtual Server (LVS) with Keepalived to achieve load balancing. Layer-7 SLB uses Tengine to achieve load balancing. Tengine, a Web server project based on Nginx, adds advanced features dedicated for high-traffic websites.

Requests from the Internet reach the LVS cluster through ECMP routing. Each LVS in the LVS cluster synchronizes the session to other LVS machines in the cluster through multicast packets, thereby implementing session synchronization among machines in the LVS cluster. At the same time, the LVS cluster performs health check on the Tengine cluster and removes abnormal machines from the Tengine cluster to ensure the availability of Layer-7 Server Load Balancer.

Best practice:

Session synchronization ensures that persistent connections are not affected by server failure in the cluster. But for short connections or when the session synchronization rule is not triggered by the connection (three-way handshake is not completed), server failures in the cluster may still affect user requests. To prevent session interruption caused by machine failure in the cluster, you can add a retry mechanism to the service logic to reduce the impact on user access.

High availability of a single SLB instance

To provide more reliable services, multiple zones for Server Load Balancer are deployed in most regions. If a active zone becomes unavailable, Server Load Balancer rapidly switches to a standby zone to restore its service capabilities within 30 seconds. When the active zone becomes available, Server Load Balancer automatically switches back to the active zone.

Note:

The active zone and standby zone form zone-level disaster tolerance. An SLB instance switches to the standby zone only when Alibaba Cloud detects that the current zone is unavailable due to power outage or optical cable failure rather than the failure of an instance.

Best practice:

1. We recommend that you create a Server Load Balancer instance in a region with multiple zones for disaster tolerance.
2. You can deploy ECS instances in the active zone and standby zone respectively as needed. You can set the zone where most ECS instances are located to the active zone to minimize the access latency.

   However, we do not recommended that you deploy all ECS instances in one zone. You also need to deploy a small number of ECS instances in the standby zone, so that the standby zone still can process requests in extreme conditions (the active zone is unavailable).

High availability of multiple SLB instances

If your requirements on the availability is extremely high, the availability guaranteeing mechanism of the SLB may cannot meet your demands. For example, when the SLB instance is unavailable due to network attach or configuration error, zone switching is not triggered because no zone-level failure occurs. At this time, you can create multiple SLB instances and schedule requests by using Alibaba Cloud DNS, or achieve cross-region disaster recovery through global Server Load Balancing.

Best practice:

You can deploy SLB instances and backend ECS instances in multiple zones of a region or in multiple regions, and use Alibaba Cloud DNS to schedule the requests.

High availability of backend ECS instances

Server Load Balancer checks the service availability of backend ECS instances by performing health checks. The health checks improve the overall availability of front-end services and help reduce the impact of service availability when backend servers are abnormal.

When Server Load Balancer discovers that an instance is unhealthy, it distributes requests to other healthy ECS instances, and only resumes distributing requests to the instance when it has restored to a healthy status. For more information, see *Health check overview*.

Best practice:

In order for the health check function to work, you need to enable and correctly configure the health check. For more information, see *Configure health check*.

2 FAQ

2.1 Server Load Balancer FAQ

1. What scheduling algorithms are supported by Server Load Balancer (SLB)?

The following scheduling algorithms are supported:

- Round robin (RR): Requests are distributed across the group of backend ECS servers sequentially.
- Weighted round robin (WRR): You can set a weight for each backend server. Servers with higher weights receive more requests than those with lower weights.
- Weighted least connections (WLC): In addition to the weight set to each backend ECS server, the number of connections to the client is also considered. A server with a higher weight value will receive a larger percentage of live connections at any one time. If the weights are the same , the system directs network connections to the server with the least number of established connections.

2. What is the difference between an Internet Server Load Balancer and an intranet Server Load Balancer?

- The Internet SLB can distribute client requests from the Internet. A public IP is assigned to an Internet SLB instance.
- The intranet SLB can only distribute client requests from the intranet. A private IP is assigned to an intranet SLB instance.

3. Can I modify the Server Load Balancer instance type?

No.

The SLB system allocates an instance IP (a public IP or a private IP) based on the SLB instance type. To switch the SLB instance type, you must first delete the instance and create a new SLB instance of the expected type.

4. Is the allocated IP address exclusive to the SLB Instance?

Yes. The IP address of the SLB instance is exclusive to the load balancing service you purchased during the entire lifecycle. Changing the SLB configurations and listening rules will not affect the IP address.

If the SLB IP has been resolved to a domain name to provide public services, do not delete the corresponding SLB instance unless necessary. The configurations and the IP will be deleted along with the deletion of the instance and cannot be restored. If you recreate a SLB instance, the system will allocate a new IP.

5. Does SLB support HTTP redirection?

Yes.

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

6. What is the difference between pinging the IP of the SLB instance and pinging the IP of a backend ECS instance?

When pinging the IP of an SLB instance, the response is sent from the SLB cluster. The request will not be forwarded to backend ECS instances. When pinging the IP of a backend ECS instance , the response is sent from the backend ECS instance and has no relationship with the SLB.

7. Does SLB rely on the Internet bandwidth?

The communication between the SLB and backend ECS instances goes through the intranet, therefore, no need to configure extra Internet bandwidth for backend ECS instances.

However, if you want to provide public services through both the SLB and ECS instances at the same time, the corresponding ECS instances need to be configured with sufficient Internet bandwidth. The Internet bandwidth of backend ECS instances has no impact on the service capability of the SLB.

8. Is there any impact on the load balancing service if the Internet NIC is disabled?

If the ECS instance has configured a public IP, 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. When the Internet NIC is disabled, the returned data packet cannot be sent. We recommend 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.

9. How to avoid the service failure of the SLB itself?

- Add ECS instances of different zones as backend servers for the Server Load Balancer instance to improve local availability.

- Create multiple SLB instances in the same region and use DNS to provide public services to improve the local availability.
- Create multiple SLB instances in different regions and use DNS to provide public services to improve the cross-region availability.

10. What does the SLB balance?

The SLB distributes network traffic to backend ECS instances according to the specified scheduling algorithm:

- Layer-4 listeners distribute the traffic based on TCP connections. If you create a socket through TCP or UDP to access the Server Load Balancer instance, the source and destination IPs and the ports form a connection.
- Layer-7 listeners distribute the traffic based on HTTP requests, such as an HTTP GET request.

11. Why is the traffic not balanced?

The following are possible reasons:

- Have enabled session persistence

  If session persistence is enabled, it will cause traffic imbalance when fewer clients are accessing the SLB instance. This is especially common when a small number of clients are used to test the SLB 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 enables TCP Keepalive and others do not, the connection s will accumulate on the ECS instances with TCP Keepalive enabled. This scenario will cause an imbalance.
- Troubleshoot

  — Check whether the weights of backend ECS instances are the same;

  — Check if the weight values of the backend ECS instances are the same. Check if the health check of the backend ECS instances fails or 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.

12. Why each connection does not reach the bandwidth peak?

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 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.

13. Why does the SLB compression fail?

View the file's content-type attribute. If the file type is not text/xml, text/plain, text/css, application/javascript, application/x-javascript, application/rss + xml, application/atom + xml or application/xml , the compression fails.

Resolution:

- Modify the file's content-type attribute at the source site to change the file type to a type supported by the Server Load Balancer.
- Modify the Layer-7 SLB instance listener to a Layer-4 SLB instance listener.

14. How to query the load balancing traffic usage?

Consumption details

Log on to the billing management console, in the **Billing Management** navigation bar, click **Resource Packages** > **Usage Records**. Select **Server Load Balancer (SLB)** to view the consumption details of SLB.

Traffic usage

On the SLB details page, click **Monitoring** to view traffic usage.

The monitoring data may be different from the traffic data in the consumption details. For more information, see *Monitoring data and billing data*.

15. Why I cannot modify the bandwidth by calling the API?

If the following error occurs when calling the API to modify the bandwidth, it indicates that the peak bandwidth set on the console conflicts with the value set in the API. Error message: `Code":"InvalidParameter","Message":"The specified parameter bandwidth is not valid."`. You have to change the peak bandwidth on the console.

16. Why is the monitoring data different from the actual billing data?

- Monitoring data is collected every one minute by the Server Load Balancer system, and reported to the cloud monitoring system. Then, the cloud monitoring system calculates the average value of all collected data in each 15 minutes. Billing data is collected at the same granularity and the Server Load Balancer system reports the accumulated value in each hour to the billing system.

  The monitoring data is the calculated average value, but the billing data is the accumulation value. These two data sets are incomparable because they are calculated and generated differently.
- Server Load Balancer provides real-time monitoring data. However, a short delay may inevitably occur in the data collection, calculation, and display process. Although this delay is almost insignificant, it can create a certain degree of discrepancy between the monitoring and billing data. Billing data tolerates a maximum delay of three hours. For example, billing data generated between 01:00-02:00 is normally reported to the billing system at 03:00, but is allowed to be reported to the billing system at 05:00. As a result, there are differences between billing data and monitoring data.
- The product definitions of monitoring and billing data are also different. The purpose of monitoring is to help users observe if the instance is in abnormal conditions. If so, users can resolve the problem as soon as possible. The purpose of billing is to generate bills. Monitoring data cannot be used as the billing data.

17. What are the timeout values of each listener?

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

- HTTPS listener: 60 seconds

18. Why does the SLB connection time out?

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

- The IP of the SLB instance is protected

  Such as the black hole triggering and traffic cleaning, 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.

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

- 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.

2.2 Backend server FAQ

1. How many ECS instances can I add to a Server Load Balancer (SLB) instance at most?

You can add up to 200 ECS instances to an SLB instance. An ECS instance can be associated with up to 50 SLB instances.

Note:

Submit a ticket if you want to add more ECS instances. After the application is approved, you can get more quota.

To guarantee the service stability and efficiency, we recommend that you add ECS instances hosting different applications or performing different tasks on different SLB instances.

2. Can I adjust the number of the backend ECS instances?

Yes.

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

3. Can I use different operating systems for the 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 the backend ECS instances.

4. Can I add the ECS instances in different regions to an SLB instance?

No.

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

5. Why cannot backend ECS instances access SLB?`

It is related to the TCP implementation mechanism of SLB. For the Layer-4 SLB service, a backend ECS instance cannot act both as a real server and a client that sends requests to an SLB instance. The returned packets can be transmitted only inside the ECS instance and cannot pass through SLB, so the SLB instance cannot be accessed from backend ECS instances.

6. Is there any impact on the SLB service if an ECS instance has configured a whitelist before being added to a SLB instance?

If an ECS instance has configured a whitelist before it is added to an SLB instance, add the SLB whitelist to the ECS instance, otherwise a 403 error may occur after it is added to the SLB instance.

The SLB whitelist to be added is 100.64.0.0/10.

7. Why there are IP addresses starting with 100 frequently accessing backend ECS instances?

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

The IP address range used by the SLB system to do health check and availability check is 100.64 .0.0/10.

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

8. Does the backend ECS instance require any special configurations?

No special configuration is required in general.

However, if the ECS instances using Linux systems are added with Layer-4 listeners, make sure the following values in the */etc/sysctl.conf* file are zero:

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

If the ECS instances belonging to the same intranet IP address range cannot communicate with each other, check whether the following parameters are correct. Run **`sysctl -p`** to update the configuration after you make any changes.

```
net.ipv4.conf.default.arp_announce =2
net.ipv4.conf.all.arp_announce =2
```

9. Can I deploy multiple sites on the same group of ECS instances and add them to one SLB instance?

Yes. For more information, see *SLB + ECS multi-site deployment*.

Each SLB instance supports adding up to 50 listeners and each listener corresponds to an application deployed on backend ECS instances. To achieve this, you can configure different host headers for different applications on the backend ECS instances.

10. What is the purpose of setting weights for ECS instances?

An ECS instance with a higher weight value receives more requests. You can configure different weights for different ECS instances based on the external service capabilities and demand.

It is the same if the weights for ECS instance A and ECS instance B are both configured to the same value, for example, 100 or 50. Traffic is evenly distributed to instance A and instance B when the two instances are both in normal status. However, if the weight for the ECS instance A is set to 10 but the weight for the ECS instance B is set to 100, instance A receives 10/(10+100) percent of the traffic and instance B receives 100/(10+100) percent of the traffic.

If you have enabled session persistence, requests may not be evenly distributed to the backend ECS instances. We recommend temporarily disabling session persistence to check if the problem still exists.

When requests are unevenly distributed to backend ECS instances, troubleshoot as follows:

1. Obtain the total number of access logs of web services deployed on the backend ECS instances in a certain time period.

   Use the following methods to get the access logs:

   - Nginx and Apache: open log directory `/access.log`.
   - For IIS: Open the IIS management page, place the cursor over the site for which you want to enable access logging, right-click and choose **Properties**, click the **Website** tab, and click **Enable Logging**.
2. Compare the number of the logs for multiple ECS instances according to the SLB configurations.

   If session persistence is enabled, you must strip the access logs from the same IP. If the SLB configures a weight, calculate whether the proportion of access logs is normal according to the weight.

11. Why are responses returned by SLB compressed while the ECS instance does not configure compression?

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

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

Yes.

2.3 Heath check FAQ

1. How does Server Load Balancer (SLB) health check work?

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

As shown in the following figure, the node servers in the LVS cluster and Tengine cluster perform the data forwarding and health check at the same time. 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 (100.64.0.0/10 is reserved by Alibaba Cloud, and will not be used by any user, there is no security risk) on the intranet NIC.

For more information, see *Health check overview*.

2. What are recommended configurations for health check?

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

The following are recommended health check configurations for UDP listeners.

Configuration	Recommended value
Response timeout	10 seconds

Configuration	Recommended value
Health check interval	5 seconds
Unhealthy threshold	3
Healthy threshold	3

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 lower response timeout value but make sure the processing time in the normal status is less than the specified timeout value.

3. Can I disable the health check?

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

4. How to choose a health check method for TCP listeners?

For TCP listeners, both the TCP health check and HTTP health check are supported:

- Based on network Layer detection, TCP health check uses the traditional three-way handshakes to determine if the backend ECS instance is healthy or not.
- HTTP health check detects the health status by sending head requests. A Tengine node server sends an HTTP head request and compares the returned code to determine if the backend ECS instance is healthy or not.

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

5. Is there any impact on health check if the weight for 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.

6. What method does HTTP listeners use to do health check on backend ECS instances?

HEAD method

If the backend ECS instances disable the HEAD method access, health check 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:

```
echo -e “HEAD /test.html HTTP/1.0\r\n\r\n” | nc -t LAN_IP port
```

7. What are the IP address ranges that HTTP listeners use to perform health check 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.

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

Health check is 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.

9. How to troubleshoot the 502 Bad Gateway error?

Symptoms:

The 502 Bad Gateway error occurred when accessing the SLB service and the health status of the backend servers is indicated abnormal. However, the ECS instance can be accessed from the intranet and the port 80 listening is normal although a 404 error is reported in the web logs.

Cause:

The page used for health check does not exist.

Solution:

Modify the health check configurations accordingly.

10. Does health check consume system resources?

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

11. Why do backend servers of the Server Load Balancer frequently receive requests with a UA of KeepAliveClient?

Symptoms:

The backend ECS instances frequently receive GET requests from an intranet IP even if there is no user access, and User-Agent is KeepAliveClient.

Cause:

The listener uses TCP protocol but the health check uses HTTP protocol. In this situation, SLB uses the GET method instead of the HEAD method to do the health check.

Solution:

Use the same protocol for the listener and corresponding health check.

12. How to handle health check failure caused by a backend database fault?

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.

Solution:

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

13. Why a network connection exception is displayed in the backend service logs but TCP health check is 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.

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 check and the impact on backend service. TCP health check only performs a simple three-way handshake and then directly sends 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.

Solution:

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

14. Why the health check status is abnormal when the service is actually normal?

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.

15. How to troubleshoot health check exceptions?

Follow these steps to troubleshoot health check exceptions:

1. Check if the backend ECS instance can provide services normally. If HTTP health check is configured, make sure that the returned status code is the same with the normal status code configured on the console.
2. Because the health check is performed on backend ECS instances through the intranet, log on to the backend ECS instance and check if the application port is listening on an intranet IP. If not, change the listening port to an intranet IP.
3. Make sure the next hop for 10.0.0.0/8, 100.64.0.0/10, and 11.0.0.0/8 on the ECS instance is set to the intranet gateway. If the ECS instance only has an intranet IP, guarantee the default route (0.0.0.0/0) points to the intranet gateway.

   Configure the route as follows:

   - Windows: Log on to the ECS instance. Download *Windows route adding tool* and double click it to add route configurations.
   - Linux: Log on to the ECS instance, download the *Linux route adding tool* and run the `bash linux_add_routes.sh` command.
   - FreeBSD: Log on to the ECS instance, download the *FreeBSD route adding tool* and run the `bash freebsd_add_routes.sh` command.

     After you run the route adding tool, the intranet route is displayed as normal, as shown in the figure below.

4. Ensure the backend port you configured in the listener is opened on the backend server.
5. Check if the weight value of the backend ECS instance is zero. If so, the health check status is abnormal.

6. Check if there is a firewall or other security software in the backend ECS instance. This kind of software could mask the local IP address of the SLB system so as that the system cannot communication with the backend ECS instance. We recommend closing the firewall or uninstalling the security software to perform a test.
7. Check whether the response time of the backend ECS instance exceeds the response timeout of health check.

   Run the following command to view the response time of the Layer-4 health check.

   ```
   time telnet <SLB IP address> <SLB Port>
   ```

   Run the following command to view the response time of the Layer-7 health check.

   ```
   time echo -e 'HEAD <Check path of health check> HTTP/1.0\r\n\r\n'|nc
    -t <Port>
   ```

   Note:

   Run this command on another ECS instance in the same account and region. For example, the check path of health check is /, the intranet IP of the ECS instance is 192.168.0.1, and the port number is 80, use the following command: `time echo -e 'HEAD / HTTP/1.0\r\n\r\n' | nc -t 192.168.0.1 80.`

   When the value of the **`real`** parameter in the result exceeds the response timeout, the backend ECS instance is declared as unhealthy.

   Solution:

   - Optimize the application or program to reduce the response time.
   - Increase the response timeout value set in the listener check.
8. For the Layer-7 health check, run `echo -e "HEAD /test.html HTTP/1.0\r\n\r\n" | nc -t LAN_IP 80` to check if the returned value is 2XX or 3XX.

   Note:

   /test.html is the check path of health check. Change it as needed.
9. Check if the traffic load on the backend ECS instance is too heavy, which leads to over long response timeout.

10.We recommend using HTML static file to perform health check. The file is used only for checking the returned result. We do not recommend using dynamic scripting languages such as PHP.

2.4 Billing FAQ

1. How is Server Load Balancer billed?

For more information, see *Pay-As-You-Go*.

2. Is the inbound traffic of SLB billed?

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

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. No matter 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 when Alibaba Cloud Security detects an attack to when cleansing traffic begins, there may be a latency of several seconds. Therefore, the response to attack packets during this period will incur data usage and any corresponding fees. Such an attack also consumes the bandwidth resources of SLB.

6. If all the backend ECS instances of a 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:

- Bill 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 mounted. Therefore, we recommend that you stop the SLB instance if it is unnecessary in order to avoid extraneous costs.

- Bill by fixed bandwidth

  In the case of billing by fixed 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.

2.5 Guaranteed-performance instance billing FAQ

1. When will the guaranteed-performance instance start charging the capacity fees?

Alibaba Cloud Server Load Balancer (SLB) launched the guaranteed-performance instances in May, 2017, and started the testing at the same time. For now, the guaranteed-performance instance has been tested for 10 months. SLB will charge the capacity fee on guaranteed-performance instances from April 1, 2018. For more information, see *How to use guaranteed-performance instances?*.

The fee collection for the guaranteed-performance instances take effect in batches by regions:

- The first batch:

  Effective time: From April 1 to April 10

  Effective regions: Asia Pacific SE 1 (Singapore), Asia Pacific SE 3 (Kuala Lumpur), Asia Pacific SE 5 (Jakarta), Asia Pacific SOU 1 (Mumbai), US West 1 (Silicon Valley), US East 1 (Virginia), Asia Pacific SOU 1 (Mumbai)

- The second batch:

  Effective time: From April 11 to April 20

  Effective regions: China East 1 (Hangzhou), China North 3 (Zhangjiakou), China North 5 (Hohhot), Hong Kong

- The third batch:

Effective time: From April 21 to April 30

Effective regions: China North 1 (Qingdao), China North 2 (Beijing), China East 2 (Shanghai), China South 1 (Shenzhen)

2. Will intranet SLB instances be charged for capacity fee?

- If the intranet SLB instance is a shared-performance SLB instance, no capacity fees are collected.
- If the intranet SLB instance is a guaranteed-performance SLB instance, corresponding capacity fees are collected.

  The capacity fees are collected as the same as the Internet guaranteed-performance instances , but intranet guaranteed-performance instances are free from instance fee and traffic fee.

3. Will the billing of shared-performance instances be affected after the capacity fees are charged?

No.

The shared-performance instances will not be charged for the capacity fee.

However, if you change the shared-performance instances to the guaranteed-performance instances, the capacity fees will be collected from April 1, 2018.

4. Can shared-performance instances be changed to guaranteed-performance instances?

Yes.

Once an instance is changed to the guaranteed-performance type, it cannot change back and will be charged from April 1st.

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

Yes.

Guaranteed-performance instances are charged and cannot be changed back to shared-performance instances.

5. What are the capacities for shared-performance instances?

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

6. How to select a performance capacity?

You can select the largest capacity for Pay-As-You-Go instances, because Pay-As-You-Go instance are charged according to the actual usage and are free of charge in idle time.

7. What is the price of the performance capacity?

Six capacities are provided for guaranteed-performance instances. No capacity fee is collected for the Standard I (slb.s1.small) capacity. For more information, see *Billing*.

8. Are the traffic fee and instance fee of guaranteed-performance instances the same as those of shared-performance instances?

Yes.

9. How many guaranteed-performance instances can be created?

Same as shared-performance instances, you can purchase up to 60 guaranteed-performance instances. Open a ticket to apply for more quota.

10. Can I adjust the capacity of a guaranteed-performance instance?

Yes.

You can upgrade or downgrade a Pay-As-You-Go 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 instances may exist in older clusters due to historical stock. These instances need to be migrated when they are changed to guaranteed-performance instances. Therefore a service interruption of 10-30 seconds may occur. We recommend that you change the instance type when the traffic is low or perform load balancing among instances first through *GSLB* and then change the instance type.

11. Can I still buy shared-performance instances?

Yes.

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

2.6 How to use guaranteed-performance SLB instances?

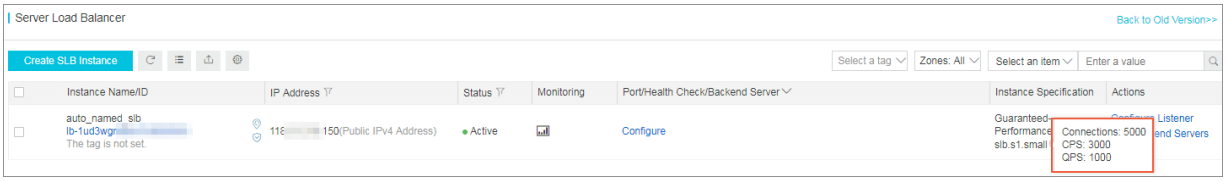
Alibaba Cloud now charges specification fees for guaranteed-performance SLB instances.

1. What are guaranteed-performance instances?

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

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

Hover your mouse pointer to the green icon of the target instance to view the performance metrics , as shown in the following figure.



The following are three key performance metrics for guaranteed-performance instances:

- Max Connection

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

- Connection Per Second (CPS)

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

- Query Per Second (QPS)

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

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

Type	Specification	Max Connection	CPS	Query Per Second (QPS)
Specification 1	Small I (slb.s1.small)	5,000	3,000	1,000

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

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

2. How are guaranteed-performance instances billed?

Guaranteed-performance instances are billed as follows:

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

Note:
The corresponding specification fee is billed for each guaranteed-performance instance regardless of the network type of the instance, and is billed based on the actual usage depending on the specification selected. If the actual performance metrics of an instance occurs between two capacities, the specification fee is charged at the higher specification fee.

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

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

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

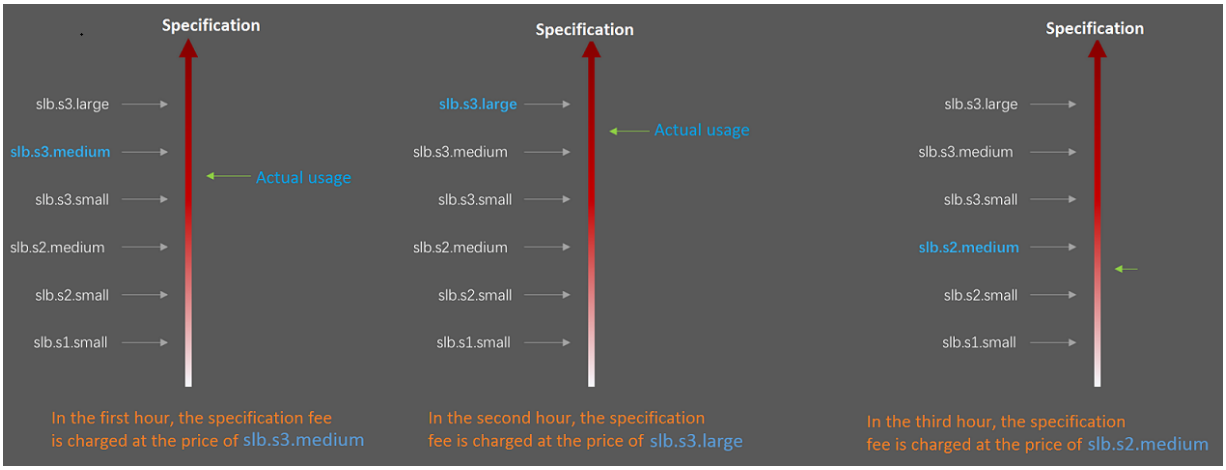
- From the perspective of Max Connection, the actual metrics 90,000 occurs between the limit 50,000 defined in the Standard I (slb.s2.small) specification and the limit 100,000 defined in the

Standard II (slb.s2.medium) specification. Therefore, the specification of the Max Connection metrics in this hour is Standard II (slb.s2.medium).

- From the perspective of CPS, the actual metrics 4,000 occurs between the limit 3,000 defined in the Small I (slb.s1.small) specification and the limit 5,000 defined in the Standard I (slb.s2.small) specification. Therefore, the specification of the CPS metrics in this hour is Standard I (slb.s2.small).
- From the perspective of QPS, the actual metrics 11,000 occurs between the limit 10,000 defined in the Standard II (slb.s2.medium) specification and the limit 20,000 defined in the Higher I (slb.s3.small) specification. Therefore, the specification of the QPS metrics in this hour is Higher I (slb.s3.small).

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

The following figure is an example showing how the specification fee is billed for an SLB instance in the first three hours:



The billing of the guaranteed-performance instances is flexible. The specification you select when purchasing an instance is the performance limitation of the instance. For example, if slb.s3.medium is selected, the new connections are dropped when the HTTP requests in one second reach 30,000.

3. What is the price of each specification?

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

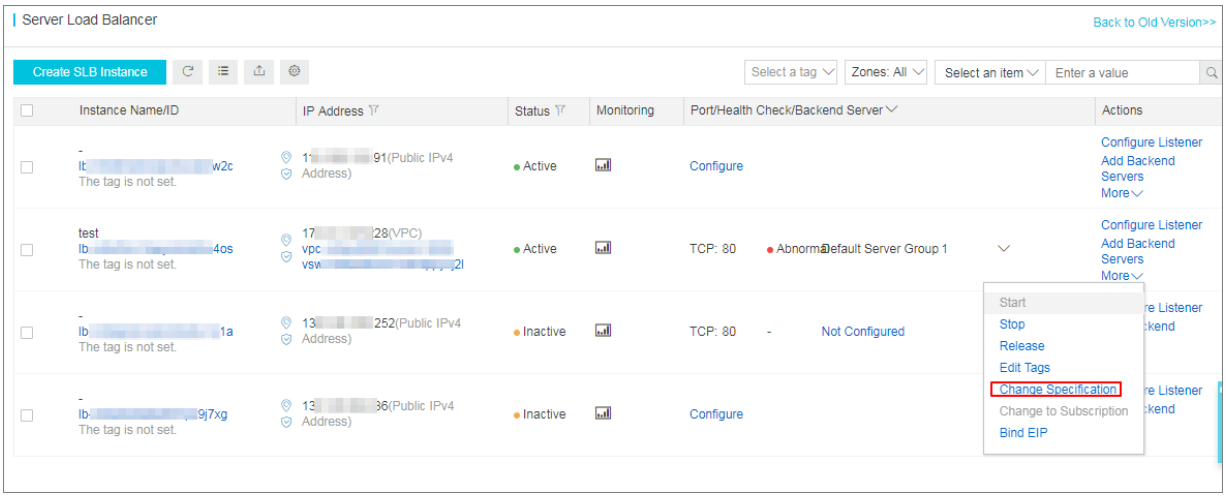
Region	Type	Max Connectio	CPS	QPS	Specification fee (USD/Hour)
China (Hangzhou) China (Zhangjiakou) China (Huhhot) China (Qingdao) China (Beijing) China (Shanghai) China (Shenzhen)	Small I (slb.s1.small)	5,000	3,000	1,000	Free
	Standard I (slb.s2. small)	50,000	5,000	5,000	0.05
	Specification 3: Standard II (slb.s2. medium)	100,000	10,000	10,000	0.10
	Higher I (slb.s3.small)	200,000	20,000	20,000	0.20
	Higher II (slb.s3. medium)	500,000	50,000	30,000	0.31
	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) China (Hong Kong)	Small I (slb.s1.small)	5,000	3,000	1,000	Free
	Standard I (slb.s2. small)	50,000	5,000	5,000	0.06
	Standard II (slb.s2. medium)	100,000	10,000	10,000	0.12
	Higher I (slb.s3.small)	200,000	20,000	20,000	0.24
	Higher II (slb.s3. medium)	500,000	50,000	30,000	0.37
	Super I (slb.s3.large)	1,000,000	100,000	50,000	0.61

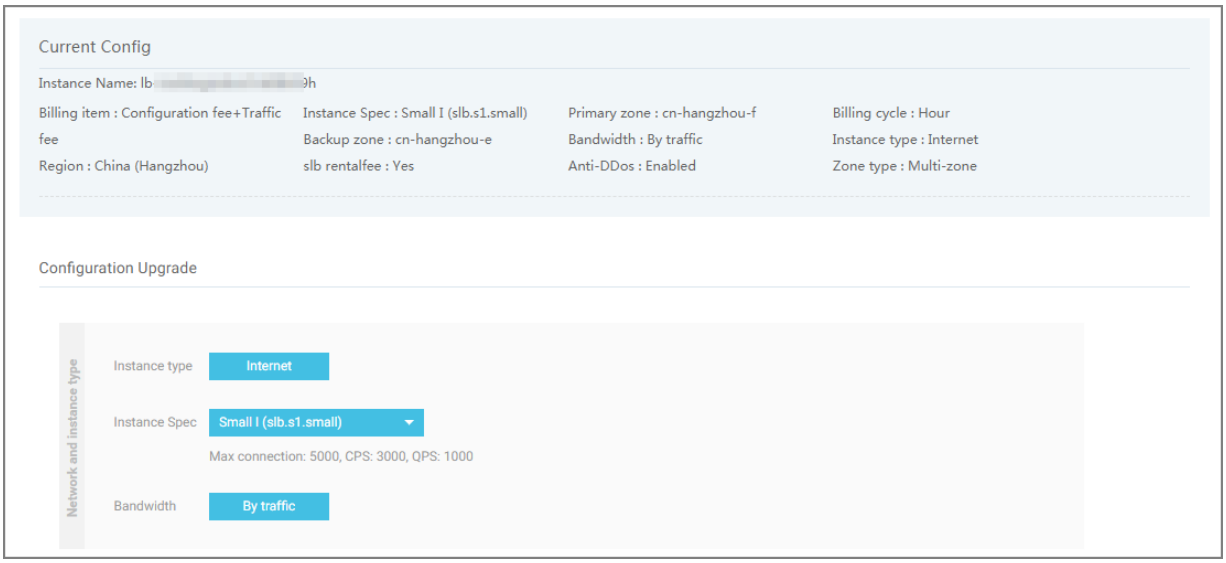
4. How to select a guaranteed-performance instance?

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

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

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





Note:

- Once a shared-performance instance is changed to a guaranteed-performance instance, it cannot be changed back.
- Some instances may exist in older clusters due to historical stock. If you change a shared-performance instance to a guaranteed-performance instance, a brief disconnection of service may occur for 10 to 30 seconds. We recommend that you change the specification when the business is not busy.
- The IP of the SLB instance will not be changed after you changing the instance type or the specification.

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

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

The fee collection for the guaranteed-performance instances take effect in batches by regions:

- The first batch:

  Effective time: From April 1 to April 10

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

  Effective time: From April 11 to April 20

  Effective regions: China (Hangzhou), China (Zhangjiakou), China (Hohhot), China (Hong Kong)
- The third batch:

  Effective time: From April 21 to April 30

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

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

Not at all.

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

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

It applies the cask theory.

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

For example, you have purchased a guaranteed-performance instance of the Higher I (slb.s3. small) specification. When the QPS of the instance reaches 20,000 but the number of maximum

connections does not reach 200,000, the new connections are still dropped because the QPS has reached the limitation.

9. Can I still buy shared-performance instances?

Yes.

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

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

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

2.7 Why is the traffic not balanced?

Reasons

The following are possible reasons:

- 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 persistenc e (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 enables TCP keepalive while some do not, the connection s will accumulate on the ECS instances with TCP keepalive enabled, which will cause the imbalance.

Troubleshoot

- Check if the weights of the backend servers are the same.
- Check if the health check of the backend servers fails or the health status is unstable in a specified period, or the status code is correctly configured in the health check configurations.
- Check if both the WLC scheduling algorithm and session persistence are enabled. If so, change the scheduling algorithm to WRR.

2.8 How to forward traffic with the same domain name but different URLs

Context

In this case, we use four ECSs deployed with Nginx servers as the example to demonstrate how to configure forwarding rules specified by domain name and URL, so as to fulfill traffic forwarding as shown in the following table.

Frontend requests	Forward traffic to
www.aaa.com/tom	Server SLB_tom1 and server SBL_tom2
www.aaa.com/jerry	Server SLB_jerry1 and server SBL_jerry2

Procedure

1. Create an Internet-facing SLB instance.

   For more information, see *Create a Server Load Balancer instance*.

2. Resolve the domain name into the public IP of the SLB instance by using DNS.

   For convenience, the public IP of the SLB instance is bound to domain name www.aaa.com in the host file in this case.

3. Create two VServer groups.

   a) Locate the newly created target instance in the Server Load Balancer console and click the instance ID to enter the details page.

   b) In the left-side navigation pane, click **Server** > **VServer Group**.

   c) Click **Create VServer Group**.

   d) In the displayed dialog box, select the backend servers to be added and set ports and weights for them respectively. The ports for ECS instances in the VServer group can be different.

In this case, enter **TOM** as the server group name, add server SLB_tom1 and server SBL_tom2 into the group, set the port number to 80, and keep the default weight value (100).

e) Repeat the preceding steps to add another VServer group named JERRY, which includes server SLB_jerry1 and server SBL_jerry2.

4. Add a listener.

a) In the left-side navigation pane, click **Listeners**, and click **Add Listener**.

b) Configure the listener. In this case, the listener is configured as follows:

- **Frontend protocol [Port]: HTTP: 80**
- **Backend protocol [Port]: HTTP: 80**
- **Scheduling algorithm: Round-robin.**
- Keep the default values for other configuration items.

c) On the **Listeners** page, click **More** > **Add Forwarding Rules**.

d) On the **Forwarding rules** page, click **Add Forwarding Rules**.

e) Configure three forwarding rules.

5. Test:

- Enter wwww.aaa.com/jerrry in the browser and the following result is returned.

- Enter wwww.aaa.com/tom in the browser and the following result is returned.

Enter www.aaa.com in the browser and the following result is returned.

2.9 Session persistence FAQ

1. What is session persistence?

Session persistence serves to forward session requests from the same client to a specified backend server for processing.

2. How can I enable session persistence?

You can choose whether to enable the session persistence function when configuring listeners . You can configure different session persistence policies for different listeners. The maximum session persistence duration is 86,400 seconds (24 hours).

3. What type of session persistence does SLB support?

- For Layer-4 (TCP protocol) services, session persistence is based on source IP addresses. The maximum duration of Layer-4 session persistence is 3,600 seconds.

- For Layer-7 (HTTP or HTTPS) services, session persistence is based on cookies. The maximum duration of session persistence based on cookie inserting is 86,400 seconds (24 hours).

4. What kind of cookie configurations are supported?

HTTP/HTTPS listeners can use cookie inserting and rewriting methods to achieve session persistence.

- Cookie inserting: When this method is used, you only need to specify the cookie timeout. For the first access by a client, SLB inserts a cookie (inserts a SERVERID string in the HTTP/ HTTPS response message) in the response. The next request from the client will contain this cookie and SLB will forward the request to the same ECS instance.
- Cookie rewriting: When this method is used, you can specify the cookie to be inserted in the HTTPS/HTTP response as needed. You must maintain the timeout and TTL of the cookie in the backend ECS instance. SLB rewrites the original cookie when it discovers a customized cookie. The next request from the client will contain this rewritten cookie and SLB will forward the request to the same ECS instance. For more information, see *Configure cookie in the backend server*.

5. Can I configure different session persistence rules for different domain names?

Yes.

You can configure different session persistence rules by using the cookie rewriting method.

6. What timeout value should I set for a cookie?

- For cookie inserting, you can set a timeout value from 1 to 86400 seconds on the SLB console.

- For cookie rewriting, you must maintain the timeout value on the backend ECS instance.

7. How to check the session persistence string?

You can use developer tools in the browser to view whether the response message contains the SERVERID string or a user-specified keyword, or 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 initiate the access.

8. Why the session persistence does not work sometimes?

- Check whether the session persistence function has been enabled in the listener configuration.
- HTTP/HTTPS listeners cannot insert session persistence cookies to the response messages containing 4xx response codes that are returned by backend servers.

  Resolution: Use TCP listeners instead. TCP listeners achieve session persistence based on the source IP address of the client. Additionally, you can configure cookies for the backend ECS instances and add cookie detection logic to guarantee that session persistence works.
- 302 redirection changes the SERVERID string.

  If a 302 redirection packet is returned by the backend servers when SLB is inserting a cookie, the SERVERID string in session persistence will be changed. Therefore, session persistence fails to work.

  Troubleshoot: Capture the request and returned response on the browser or use a tool to capture packets to check whether a 302 response message is returned. Then, compare the SERVERID strings in the packets to see if they are different.

  Resolution: Use TCP listeners instead. TCP listeners achieve session persistence based on the source IP address of the client. Additionally, you can configure cookies for the backend ECS instances and add cookie detection logic to guarantee that session persistence works.
- Too short session persistence duration will also cause session persistence failure.

9. How can I use Linux curl to test the session persistence?

1. Create test pages.

   Create test pages on all the backend ECS instances. The local intranet IP address is displayed , as shown in the following figure. The intranet IP address is used to verify the backend server

to which client requests are distributed. Observe the consistency of this IP address to check whether session persistence works.

2. Perform curl test in a Linux environment.

   Assume that the IP address of an SLB instance is 1.1.1.1, and the URL of the created test page is *http://1.1.1.1/check.jsp*.

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

   b. Run the following command to obtain the cookie.

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

   Note:
   The default session persistence method of SLB is cookie inserting, and the curl test does not save or send a cookie. Therefore you must save the cookie for test first. Otherwise, the curl test result is random. As a result, you will consider that session persistence does not work by mistake.

   c. Run the following command to test session persistence.

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

   Note:
   a<=30 is the number of tests to do, you can change the number as needed. grep ‘10.170.*' is the IP address to display, you can change it ccording to the intranet IP address of the backend ECS instance.

   d. Observe the IP addresses returned in the preceding tests. If they are the intranet IP address of the same ECS instance, then session persistence works; otherwise, there is something wrong with the SLB session persistence.

2.10 HTTP/2 support FAQ

What is HTTP/2?

HTTP/2 (Hypertext Transfer Protocol Version 2) is the second version of Hypertext Transfer Protocol (HTTP). It is compatible with HTTP/1.X and has significant performance improvements. Comparing with HTTP/1.X, HTTP/2 has the following advantages:

- Multiplexing: Allows multiple request-response messages to be initiated simultaneously over a single HTTP/2 connection.
- Binary framing and header compression: Improves the efficiency of data transmission in the network.
- Server push: The server actively sends data to the client to reduce the number of requests, which improves efficiency.
- Additionally, features such as flow control, active request resetting, and request priority greatly improve the performance of web services, as shown in the following figure.
-

How to enable HTTP/2 on Server Load Balancer of Alibaba Cloud?

No configuration is required. HTTP/2 is supported by HTTPS listeners by default.

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

The Server Load Balancer HTTPS listener detects the ALPN field in the handshake message ClientHello sent from the client to negotiate the protocol version. If the ALPN field is not included in the ClientHello message, HTTP/1.x is used to handle the request. If the ALPN field is included, HTTP/2 is used.

> **Note:**
> HTTP/2 only supports HTTPS listeners and does not support HTTP/2 Cleartext.

Supported regions

The HTTP/2 support is available in all regions.

Limits

The limits for HTTP/2 support are as follows:

- HTTP/2 only supports HTTP listeners and does not support HTTP/2 Cleartext.
- Currently HTTP/2 is enabled only on the link between the client and Server Load Balancer. The connection between Server Load Balancer and the backend servers still uses HTTP/1. X.
- The requests of HTTP/2 are counted into the QPS of the listener/instance together with the QPS of HTTP/1.X requests.

- For HTTP/2, the head field in the response sent from the backend server to SLB is changed to lowercase, such as Content-Type is changed to content-type.
- A single connection can support up to 128 concurrent streams.
- The connection timeout value of HTTP/2 is 180 seconds.

Billing

HTTP/2 support is free of charge.

2.11 HTTPS/HTTP listener FAQ

1. Why are some response header parameters deleted after the 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.

2. Why an additional header Transfer-Encoding: chunked is added to the 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.

Cause: 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.

3. Why style sheets are not loaded when opening a website through an HTTPS listener?

Symptoms:

An HTTP and HTTPS listeners 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.

4. Which port does HTTPS listeners use?

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

5. 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.

6. Does the 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*.

7. 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*.

8. How many certificates can I upload with one account?

A maximum of 100 certificates per account are allowed, including CA certificates and server certificates.

9. Why does the KeyEncryption error occur when uploading certificates?

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

10. How many certificates can be added to an HTTPS listener?

If you use HTTPS one-way authentication, you can only bind one server certificate to a listener; if you use HTTPS mutual authentication, you must bind a server certificate and a CA certificate to a listener.

The HTTPS listener of a guaranteed-performance SLB instance supports attaching multiple certificates to forward requests with different domain names to different backend server groups. For more information, see *Tutorial: Configure a domain name extension*.

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

TLSv1, TLSv1.1, and TLSv1.2.

12. Why is the actual traffic generated by HTTPS listeners more than the billed traffic of HTTPS listeners?

HTTPS listeners consume some traffic for three-way handshake, so the actual traffic generated is more than the billed traffic.

13. What is the lifetime of an HTTPS session ticket?

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

14. 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.

15. Does 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*.

16. Which HTTP version is used by HTTP/HTTPS listeners to access the backend servers?

HTTP/1.0.

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

Yes.

18. After SLB forwards a request to a backend server, if the client disconnects from SLB before it receives the response from the backend server, will SLB close the connection to the backend server at the same time?

No. SLB will not close the connection to the backend servers during the reading and writing process.

19. Do HTTP/HTTPS listeners support the WebSocket/SSL WebSocket?

Yes, WebSocket/SSL WebSocket protocol is supported in all regions. For more information, see *WS/WSS protocol FAQ*.

20. What are timeout values specified for HTTP/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/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.

21. Does SLB support configuring domain and URL based forwarding rules?

Yes. For more information, see *Configure domain and URL based forwarding rules*.

22. How many forwarding rules can be configured for each listener?

You can add a maximum of 20 forwarding rules to each listener.

2.12 WS/WSS support FAQ

What is WS/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.

Its biggest difference from HTTP is that WebSocket is a two-way communication protocol. Once the connection is established, both the WebSocket server and the client can send or receive data to or 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/WSS?

With the increasing popularity and accessibility of the Internet, a multitude of varied web applications are emerging. Many applications require real-time push capabilities 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 to enable WS/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 WSS/WS support is available in all regions.

Limitations

The limitations for WS/WSS protocol are as follows:

- Server Load Balancer is connected to backend ECS instances by using HTTP/1.1. We recommend 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/WSS protocol support is free of charge.

3 Troubleshooting

3.1 500/502/504 troubleshooting

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 carrier, 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 document lists the causes, resolutions 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.*
   - *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. *Submit 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.

   Resolution: Put the domain name or 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 carrier.

Resolution: Submit a ticket to Alibaba Cloud after-sales personnel who decides if there is blockage through packet capture. If so, contact the carrier 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) The IP address belongs to the IP address range of the SLB server and is mainly used for health check 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.

   Resolution: Configure a whitelist for antivirus or firewall software, or unload the software to perform 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.

Resolution: 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, new PHP requests can be processed successively.
- If the existing PHP requests are not processed timely, new PHP requests will remain in a waiting mode. If the value of fastcgi_read_timeout of Nginx is exceeded, a 504 Gateway Time-out error occurs.
- If the existing PHP requests are not processed in a timely manner, new PHP requests will remain in a waiting mode. If 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 /timeoutmore.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 Quick network, or backend ECS instance configurat ions.
- If Anti-DDoS 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 carrier.
- 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.

Submit a ticket

Perform the troubleshooting procedures step by step and record the test results in detail. Provide the test results when you submit 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.

3.2 Troubleshoot ECS instance exceptions

After you enable health check 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 you can directly access your service through the ECS instance.
- Ensure 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 check. If the page you use for health check 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 simple html page for health check 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 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 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 tutorial, 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 tutorial, 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 tutorial, use `curl -I 10.11.192.1`.

3.3 Cannot access Server Load Balancer

When the client cannot access Server Load Balancer, the possible reasons and resolutions are as follows:

1. If you haven't added any listener to the SLB instance, you can't successfully ping the endpoint of the SLB instance.

   Resolution: Configure listeners. For more information, see *Configure listeners*.
2. The backend Linux ECS kernel of Layer-4 SLB isn't correctly configured.

   You must disable the rp_filter feature of the Linux kernel of the Linux ECS instance added to Layer-4 SLB. Otherwise, the client may not be able to access the endpoint of SLB by using telnet, but the health check is normal.

Linux's rp_filter feature is used to implement reverse filtering (URPF). It verifies the flow of reverse packets to avoid attacks using a forged IP. However, this feature may conflict with the bottom-layer LVS policy route of SLB and causes access exceptions.

Resolution: Ensure that the values of the following three parameters in the system configuration files of the backend ECS instance are zero. Edit `/etc/sysctl.conf` and run the `sysctl -p` command the make the configurations take effect.

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

3. The backend Windows ECS parameters of Layer-4 SLB are not correctly configured.

   For Layer-4 SLB service, a backend ECS instance cannot directly provide service for clients and act as the backend server of the SLB service at the same time.

   If the ECS instance act the two roles at the same time, access requests are forwarded to the same ECS instance, which may cause a data access loop and the failure of access from the ECS instance to the SLB service.

   Resolution:

   a. Install Windows loopback adapter: Right-click**Computer** > **Property** . On the Control Panel home page, click**Device Manager** > **Add hardware** > **Install the hardware that I manually select from the list** > **Show all devices** , Select the device shown in the following figure and install it.

   b. Enable weak host Model, and perform the following command to view idx of all network interfaces.

      ```
      netsh interface ipv4 show interface
      ```

   c. Configure `weakhostsend=enabled`, `weakhostsend=enabled` for all network interfaces. For example, configure the interface of which the idx is 12 as follows:

      ```
      netsh interface ipv4 set interface 12 weakhostsend=enabled
      netsh interface ipv4 set interface 12 weakhostreceive=enabled
      ```

4. The local network of the client or the intermediate link of the service provider is abnormal.

   For Internet Server Load Balancer service, client network exceptions or exceptions in the service provider's network between the client and Server Load Balancer may also lead to the failure of access from the client to Server Load Balancer.

Troubleshoot: Perform access test 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, it is determined that the problem is caused by a network exception. Then you can do further troubleshooting and analysis through ongoing Ping tests or MTR route tracing.

5. Client IP address is intercepted by Alibaba Cloud Security.

   For Internet Server Load Balancer service, if the client network is a shared network (all LAN servers share a public IP to access the Internet). Meanwhile, certain servers in the local network launch malicious attacks such as continuous scanning against IPs of related Alibaba Cloud services due to factors such as being infected by viruses.

   Resolution: You can see the following steps to add the public IP of the local network to the whitelist of SLB:

   a. Visit http://ip.taobao.com in the network environment of the client to obtain the public IP of the client network.
   b. Add the IP to the SLB whitelist to allow all requests from the IP to SLB.

      Note:
      This operation may pose security risk. Ensure that the IP in the whitelist does not launch malicious attacks on the SLB.

   If the problem persists, you can provide the following information when opening a ticket 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 long-time ping and MTR route tracing tests performed by the client on the IP of the SLB.