Server Load Balancer

FAQs

MORE THAN JUST CLOUD | Alibaba Cloud

FAQs

How to decide the Server Load Balancer protocol

Sever Load Balancer listeners support HTTP, HTTPS, TCP and UDP protocols.

> HTTP is not required for all websites. For most websites with no special HTTP requirements, listening to port 80 in TCP can generally meet business needs.
>
> Server Load Balancer clusters use LVS and Tengine. Layer-4 listening (TCP/UDP) goes through LVS and reaches the backend servers directly, while Layer-7 listening (HTTP/HTTPS) must go through LVS and Tengine to reach the backend servers. As such, Layer-4 listening is recommended for higher performance requirements.

The following table details recommended protocols for different application scenarios, and the features of each protocol.

Protocol	Application Scenarios	Description
TCP	For scenarios with an emphasis on reliability and high requirements on data accuracy, but a lower demand on speed, such as file transmission, e-mail sending or receiving, and remote login. Scenarios involving web applications with no special requirements are also suitable.	- It is a connection-oriented protocol, which requires a reliable connection must be established with the other side. - Session persistence is required based on the source address. - The source address must be visible at the network layer. - The listener must support TCP and HTTP health checks.
HTTP	For applications requiring	- It is an application

	data content identification, such as web applications and small mobile phone games.	layer protocol, which solves how to package the data. - Cookie-based session persistence. - The source address is obtained using X-Forward-For. - The listener only supports HTTP health checks.
HTTPS	For applications requiring encrypted transmission.	- Encrypted data transmission blocks unauthorized access. - Once the certification management service is made uniform, you can upload certificates to the Server Load Balancer. - The decryption can be completed directly on the Server Load Balancer.
UDP	For scenarios that emphasize real-time situations and attach less importance to reliability, such as video chatting and real-time communication of financial market information	- Targeting non-connection-oriented protocols. Data packets are sent directly without the three-way handshake process with the receiver, and there is no error recovery and data retransmission. - Relatively low reliability.

		- Fast data transmission.

Server Load Balancer support for domain name/URL-based forwarding

FAQs

Q: In what scenarios is the Server Load Balancer 's support for domain name/URL-based forwarding functions applicable?

A: The Server Load Balancer supports the forwarding of user requests through an instance to a backend server of the same group. Additionally, Server Load Balancer now supports more refined forwarding control through custom forwarding policies based on domain names and URLs.

This is suitable for scenarios where a single Server Load Balancer instance has multiple different services.

Q: Who can use the Server Load Balancer’ s domain name/URL-based forwarding function?

A: Any user that has a Server Load Balancer instance with layer-7 (HTTP and HTTPS) listeners can use the domain name/URL-based forwarding function.

Q: What is the significance of the Server Load Balancer's support for domain name/URL-based forwarding?

A: The Server Load Balancer supports domain name/URL-based forwarding in two ways:

- The Server Load Balancer allows users to customize backend server groups for all listeners (TCP/UDP/HTTP/HTTPS) based on listener levels and allows forwarding to different server ports.
- For HTTP and HTTPS listeners, the Server Load Balancer allows users to set forwarding rules for listeners based on domain names and URLs and forwards requests to different backend server groups.

Q: How many forwarding rules can be added for each listener?

A: The quantity of forwarding rules that a user can add is unlimited.

Q: Can users set the backend servers in the instance dimension as well as VServer group and forwarding rules dimensions?

A: Server Load Balancer allows users to configure all three of these dimensions and there are no constraints between servers added in the three dimensions.

Q: If a user configures a backend server, a VServer group and a forwarding rule, what is the priority of them?

A: When user traffic passes through a Server Load Balancer port, the system first checks whether it matches the forwarding rule. If it matches, the traffic is forwarded to the backend server group specified by this rule. If it does not match, the traffic is forwarded to the VServer group. If the user do not configure a VServer group for this listener, the traffic is forwarded to the backend servers.

Forwarding rules take highest priority, VServer groups next, and backend servers at the last.

Q: What regions are currently supported? Are console and Open API operations both supported?

A: The forwarding is supported over the Management Console by the Server Load Balancer in South China 1 (Shenzhen) only. It will be available in Open API and in other regions in the future. We will update the list of supported regions at https://intl.aliyun.com/forum.

Console operation changes

The newly released Server Load Balancer 's domain name/URL-based forwarding function includes 3 main changes to the console:

- Adding VServer groups
- Adding service listeners
- Setting forwarding policies

Why is the traffic not balanced?

When session persistence is configured and there are few clients requesting access to the Server Load Balancer instance, it will cause traffic imbalance.

- This is especially common when clients are used to test the Server Load Balancer instance. For example, if session persistence (source IP address-based at Layer-4) is enabled for a TCP listener and a client is used to test the traffic of the Server Load Balancer instance.
- Backend servers with abnormal heath status can also lead to an imbalance. It is easy to ignore the health check status of backend servers, especially during stress testing. If a backend server health check fails or its health status changes frequently, this will cause an imbalance.
- When TCP Keepalive is enabled for some backend servers to maintain persistent connections, these servers will accumulate more connections, causing an imbalance.
- Because of Server Load Balancer’ s underlying architecture, when there are too few connections to allocate properly, an imbalance may occur. In the worst scenario, the connection variance between backend servers may reach 48.

Why are the IP address requests from the backend ECS instances prefixed With 10?

The number of access requests increases greatly when the Server Load Balancer system performs a health check.

The Server Load Balancer system redirects external access requests to backend ECS instances through the intranet IP address of the system server.

The system also performs a health check on ECS (if the health check function is enabled) and monitors the availability of the Server Load Balancer service. The access requests are initiated by the Server Load Balancer system from the following IP address CIDR block:

- 10.158.0.0/16
- 10.159.0.0./16
- 10.49.0.0/16
- 100.64.0.0/10

To ensure the availability of your external services, check that permit rules are configured for the preceding IP address CIDR block.

Server Load Balancer white list FAQs

Restrictions

- Only IP addresses in the list can access the Server Load Balancer Listener once a white list is set.
- If you enable the white list function without setting a white list, no user can access the Server Load Balancer Listener by default.
- User access to the Server Load Balancer Listener may be interrupted in the process of setting the white list.

FAQs

Q: What is a white list and what does it do?

A: A white list is an access control method. It determines which IP addresses can access Server Load

Balancer Listener. It is used to only permit certain IP addresses to access a user’ s applications.

Q: In which regions is the white list function supported?

A: The public cloud in all regions supports white lists.

Q: How do I activate the white list function?

A: You must submit a ticket for white list activation.

Q: Who can apply for the white list function?

A: Users who currently have a Server Load Balancer instance and need access control through white lists can submit a ticket to apply for a trial. This function is in development for all users. Check the Alibaba Cloud website for official announcements.

Q: How to set a white list?

A: Users can configure the white list through the console or API.

Q: Is the white list set in the instance or listener?

A: It is set in the Server Load Balancer Listener.

Q: What is the white list setting format?

A: The white list can be set using IP addresses. These can be single IP address or IP network segments.

Q: What is the maximum supported number of white list records?

A: Up to 300 white list records are supported.

Server Load Balancer limitations

Limitation	Description	Ticket Submission Supports Exception
Number of ECS instances to be added	You must have one ECS instance at least to add to the Server Load Balancer instance, and the region of the ECS instance and the	Not supported

	Server Load Balancer instance must be the same.	
Billing method	Pay-As-You-Go	Not supported
Peak public bandwidth range for a single listener (Pay-As-You-Go)	1 – 1,000 Mbps or unlimited	Not supported
System limit on the peak private bandwidth for a single listener	1 Gbps	Not supported
Default quota of Pay-As-You-Go instances	- Common users: 30 - Ant Financial Cloud users: 30	Supported
Restrictions on Server Load Balancer instance name	Length range is 1–80 characters, including letters, digits, hyphen (-), backslash （/）, period （.） and underscore （_）.	Not supported
Number of Server Load Balancer instance listeners	50 instance listeners	Not supported
Protocol types available for Server Load Balancer monitoring	HTTP/HTTPS/TCP	Not supported
Frontend/Backend port range for Server Load Balancer monitoring	1 – 65535	Not supported
Forwarding rules of Server Load Balancer monitoring	wrr and wlc	Not supported
HTTP protocol-session persistence-cookie processing method	insert and server	Not supported
HTTP-session persistence-cookie timeout time	1 – 86,400 (default is 3,600)	Not supported
HTTP-session persistence-cookie name	No more than 200 characters and cookies must comply with RFC 2965. This means that they can only contain ASCII English letters and digits, and cannot contain commas, semicolons, spaces, or begin with a dollar symbol “$”.	Not supported
HTTP-health check-port	1 – 65,535 (default is the backend server port)	Not supported
HTTP-health check-timeout time	1 – 50 (default is 5. If HCTimeout < Interval, HCTimeout is invalid and the	Not supported

	timeout time will be Interval.)	
HTTP-health check-check interval	1 - 5 (default is 2)	Not supported
HTTP-health check-healthy threshold value	1 - 10 (default is 3)	Not supported
HTTP-health check-unhealthy threshold value	1 - 10 (default is 3)	Not supported
Number of backend ECS instances that can be batch added or deleted	20	Not supported
ECS instance status to be added	Running	Not supported
Weight input range for backend ECS instances	1 - 100 (default is 100)	Not supported
API access frequency limit for a single key	5,000 times/day	No automatic process is available now. Contact customer service or Business Development for help.
Maximum number of certificates uploaded by a single user	100	No automatic process is available now. Contact customer service or Business Development for help.

Server Load Balancer UDP protocol support

FAQ

Q: What is the difference between UDP and TCP protocols?

A: TCP is a connection-oriented protocol, before data is sent or received, a connection must be established with the other side. UDP is a non-connection-oriented protocol; before data can be sent, it directly performs packet transmission instead of performing security checks, such as three handshakes, with the other side.

Q: What scenarios is UDP protocol suitable for?

A: UDP protocol is primarily suited to scenarios that value real-time information over reliability, such as video chats, pushing real-time financial quotations, DNS, and IoT.

Q: What scenarios are suitable to use UDP protocol?

A: UDP protocol is used when real-time information is valued over reliability, such as video, pushing real-time financial quotations, DNS, and IoT.

Q: How can I configure UDP protocol health check?

A: The current 4-layer TCP protocol and UDP protocol configuration items are the same. However, the following parameter configuration is recommended:

- Response timeout: 10s
- Health check interval: 5s
- Unhealthy threshold value: 3
- Healthy threshold value: 3

Restrictions and known issues

- Max connection quantity for each listener is 100,000.
- Fragment packets are currently not supported.
- UDP protocol for Server Load Balancer classic instance does not allow users to view source addresses.
- If the ECS in the VPC needs to access the UDP protocol of the VPC Server Load Balancer instance, IP access to the instance must be allowed in the inbound UDP protocol of these ECS security groups. In this situation, you must create a security group.
- To avoid inconsistencies between the backend server port status and Server Load Balancer health check status, increase the health check interval and response timeout on the Server Load Balancer console:
    - Health check interval: 5s
    - Response timeout: 10s
- If, after performing the preceding adjustments, the problem continues, perform the following operations:
    - Disable sending restrictions for port unreachable type ICMP messages.
    - Execute the following command on the back-end LINUX server:
      sysctl -w net.ipv4.icmp_ratemask=6160
      Note: The default value is 6168.
      If the sending speed for port unreachable ICMP messages is no longer restricted and the RS exposed to the public network experiences a UDP port scan attack, all port unreachable ICMP messages will be returned. This will increase the OS consumption.
- For UDP listeners, configuration may need 5 minutes to take effect in the following scenarios:
    - When a backend ECS instance is removed and switched to a normal ECS instance.
    - When a health check detects an exception in a backend ECS instance with a weight

of 0 and switches to a normal ECS instance.

References

ECS security group configuration instructions

If the ECS in the VPC needs to access the UDP protocol of the VPC Server Load Balancer instance, IP access to the instance must be allowed in the inbound UDP protocol of these ECS security groups.

- Method 1: Authorization through the VPC console's security groups
- Method 2: API Settings

Classic network ECS instances needs to access the UDP protocol of the Server Load Balancer instance. In principle, no additional authorization is needed. If the user has disabled inbound authorization, the UDP protocol must be enabled.

Server Load Balancer UDP health check instructions

For UDP packets sent to the backend server's specified port XX, if the UDP port XX does not have a listener, the backend server protocol stack will return a port XX unreachable ICMP message.

After the Server Load Balancer sends a packet, if it does not receive a corresponding ICMP packet before wait timeout, the RS UDP XX PORT will be considered normal. Otherwise, the UDP PORT XX health check has failed.

Defects: Currently, the Linux backend server protocol stack has a self-defense mechanism. To prevent ICMP attacks, it restricts the backend server’ s ICMP sending speed. In stress testing, it was discovered that the backend server does not respond with a port XX unreachable ICMP message. In this case, although the backend server UDP PORT XX is not available for service, the Server Load Balancer shows a successful health check because it does not receive an ICMP response.

Late-stage complete solution: The check will not be considered successful until the Server Load Balancer sends the user-specified string to the backend server and obtains a user's specified response. This requires the user program to make an appropriate response.

Server Load Balancer VPC support

Server Load Balancer supports VPCs in the following ways:

- Users can apply for using an IP address as the Server Load Balancer private network address and then attach the VPC ECS instances.
- Users can apply for using a public IP address as the Server Load Balancer public network address and then attach the VPC ECS instances.

Note: Currently, Elastic IP addresses cannot be used as the IP addresses of Server Load Balancer instances.

In Alibaba Cloud, there are two network types, classic network and VPC. In combination with the network types, there are three different instance configurations:

- Classic private network instances
- Classic public network instances
- VPC private network instances

What conditions are required for Server Load Balancer to use VPC?

- The ECS instances are of the VPC type and is activated.
- The Server Load Balancer for the target region has been upgraded to support VPC.

How do I activate Server Load Balancer support for VPC?

Server Load Balancer has a built-in support for VPC and there is no needed activation or application process. You only need to activate VPC, create a VPC ECS instance, and apply for a VPC IP address, which can then be used as the Server Load Balancer instance IP address.

Scenarios and restrictions

A Server Load Balancer instance and an ECS instance are created in VPC-supported regions, for example, North China 2 (Beijing).

- Use a VPC IP address as the IP address of the Server Load Balancer private network instance and add VPC ECS instances.
- Use a classic cloud public IP address as the IP address of a Server Load Balancer public network instance and add VPC ECS instances.
- You cannot use a classic cloud private IP address as the IP address of a Server Load Balancer private network instance and add VPC ECS instances.
- You cannot use a VPC IP address as the IP address of a Server Load Balancer private network instance and add non-VPC ECS instances.
- A single instance cannot have both VPC and non-VPC ECS instances at the same time.

At least one ECS instance has been created in a VPC-supported region, but none are created on the Server Load Balancer system in the same region.

- Use a classic public IP address as the IP address of a Server Load Balancer public network instance and add VPC ECS instances.
- By default, you cannot use a classic cloud private IP address as the IP address of a Server Load Balancer private network instance and add VPC ECS instances.
- You cannot use a VPC IP address as the IP address of a Server Load Balancer private

network instance and add non-VPC ECS instances.
- A single instance cannot have both VPC and non-VPC ECS instances at the same time.

Switch private network IP addresses and health check IP addresses to the 100 CIDR block in VPC-supported regions

In regions with support for VPC, the Server Load Balancer private network IP addresses and health check IP addresses will be switched to the 100 CIDR block.

Note the following:

If you have enabled firewalls for your ECS instances, you must permit health check IP addresses within the 100 CIDR block. See Server Load Balancer Health Check IP Address Segment

You must add routes to the 100 CIDR block for earlier ECS instances that lack them.

Once the 100 CIDR block private IP addresses are applied, the ECS instances created earlier will not have routes to them. These ECS instances will not be available because they will fail the health check when being added to a Server Load Balancer instance with an IP address within the 100 CIDR block.

Procedures for adding routes on a Linux OS

Retrieve the gateway IP address.

```
cat /etc/sysconfig/network |grep GATEWAY
```

Note: In the following commands, gateway_ip must be replaced with the gateway address obtained in the preceding action.

Manually add static routing rules to take effect immediately.

```
ip route add 100.64.0.0/10 via gateway_ip dev eth0
```

Add the static route to the configuration file for persistent effect after the system has been restarted.

- On CentOS/Red Hat Linux/Ali OS/SUSE Linux/openSUSE system: echo "100.64.0.0/10 via gateway_ip dev eth0” >> /etc/sysconfig/network-scripts/route-eth0

```
- On Ubuntu/Debian system:
  echo "up route add -net 100.64.0.0 netmask 255.192.0.0 gw gateway_ip dev eth0"
  >> /etc/network/interfaces
- On Gentoo Linux system:
  echo "routes_eth0=(\"100.64.0.0/10 via gateway_ip\")" >> /etc/conf.d/net
```

Confirm the configuration.

```
ip route show | grep '100.64.0.0/10'
100.64.0.0/10 via gateway_ip dev eth0
```

Procedures for adding routes on a Windows OS

Retrieve the gateway IP address.

In the Windows command line, execute route print to view the private network gateway address.

Add a route.

In the Windows command line, execute the following command to make the routing rules effective.

```
route add 100.64.0.0 mask 255.192.0.0 gateway_ip -p
```

Certificate FAQs

Q: How many certificates can each user upload?

A: Each user can upload up to 100 certificates.

Q: How can I upload a certificate?

A: You can upload a certificate through an API or the Server Load Balancer Console.

Q: Do I need to upload the certificate to backend ECS instances?

A: No. Server Load Balancer HTTPS provides a certificate management system that can manage and store user certificates. The certificates do not need to be uploaded to backend ECS instances. The private key uploaded to the certificate management system will be encrypted.

Q: Does the certificate vary between different regions?

A: Yes. For security and performance considerations, if you use a certificate in multiple regions, you need to upload it to all used regions.

Q: What port does the HTTPS listener use?

A: The HTTPS listener can use any port, however, Port 443 is recommended.

Q: How many listeners can one certificate be applied to?

A: A certificate can be applied to multiple listeners.

Q: Can I delete the certificate after it is uploaded?

A: Yes. However, a certificate cannot be deleted if it has been referenced.

HTTPS FAQs

What port does the HTTPS listener use?

The HTTPS can use any port, however, 443 port is recommended.

What certificate format does HTTPS support?

HTTPS supports PEM certificates. If you want to use other formats, you must convert the format.

How can I find the SSL protocol version for the HTTPS listener?

Run the following command to find the SSL protocol version:

```
`ssl_protocols TLSv1 TLSv1.1 TLSv1.2 `
```

What is the difference between the CA certificate and server certificate in HTTPS two-way authentication?

- Server certificate: used to confirm if the certificate sent by the server is signed by a trusted center. The server certificate can be purchased from Alibaba Cloud Security, or from other service providers. The server certificate should be uploaded to the certificate management

system of the Server Load Balancer.
- CA certificate: used to verify the client certificate. The server requires the user browser to send the client certificate and, once the certificate is received, the verification occurs. If the verification fails, the connection will be denied. After two-way authentication is enabled, both the CA certificate and server certificate should be uploaded to the certificate management system of the Server Load Balancer.

Why is the actual traffic generated by the HTTP protocol higher than the billed traffic?

The HTTPS protocol will consume some traffic for protocol handshaking, causing additional traffic to be generated.

What is the HTTPS ticket persistence?

The persistence time is set to 300 seconds. A session ticket is an encrypted data blob that contains the TLS connection information for reuse, such as the session key, and is usually encrypted by the ticket key. Because the server side also determines the ticket key, the server will send a session ticket to the client side in the initial handshake and ticket will be stored on the local client side. When the session is reused, the client side sends the session ticket to the server and the server then decrypts the key and reuses the session.

How do I upload DH PARAMETERS for the HTTPS listener?

The ECDHE algorithm cluster adopted by the online Server Load Balancer Layer-7 HTTPS listener supports forward secrecy technology. The algorithm does not allow users to upload the security enhancement parameter files required by the DHE algorithm cluster. Instead, it refers to the strings containing the BEGIN DH PARAMETERS fields in the PEM certificate file.

Does HTTPS listener support Server Name Indication (SNI)?

SNI is an SSL/TLS extension supporting multiple domain names and certificates by one server. Currently, the Server Load Balancer HTTP listener does not support SNI. If this feature is required, you can switch to the TCP listener and implement the SNI feature on the backend ECS.

How does HTTP protocol access to HTTPS protocol?

It is recommended you enable the Server Load Balancer Listening Protocol with the HTTP header field appended when creating the HTTP listener. After it is enabled, the backend ECS can receSive the HTTP X-Forwarded-Proto header field. If this value indicates HTTP protocol access, return the response for HTTPS domain name access.

Why am I getting insecure access prompts when trying to access an HTTPS page?

Check whether the HTTPS page attempting to be accessed contains the HTTP link reference. If the reference is incorrect, an insecure access prompt is generated.

Server Load Balancer IP addresses FAQs

Is an IP address exclusive to a Server Load Balancer?

A purchased Server Load Balancer IP address is exclusive during its lifecycle.

Will configuration cause a loss of IP addresses?

To ensure configuration of Server Load Balancer does not affect the IP address, perform configurations through the Server Load Balancer Console. Any modification performed in this manner will not result in a Server Load Balancer IP address change.

What is the impact of deleting Server Load Balancer?

If the Server Load Balancer IP address is properly resolved to a domain name and can be used for external service provision, do not delete the Server Load Balancer unless necessary. If the Server Load Balancer is deleted, the corresponding service configuration and IP address will be released, and deleted data cannot be recovered. If you create Server Load Balancer again, it will be assigned a new IP address.

Does Server Load Balancer support CNAMES?

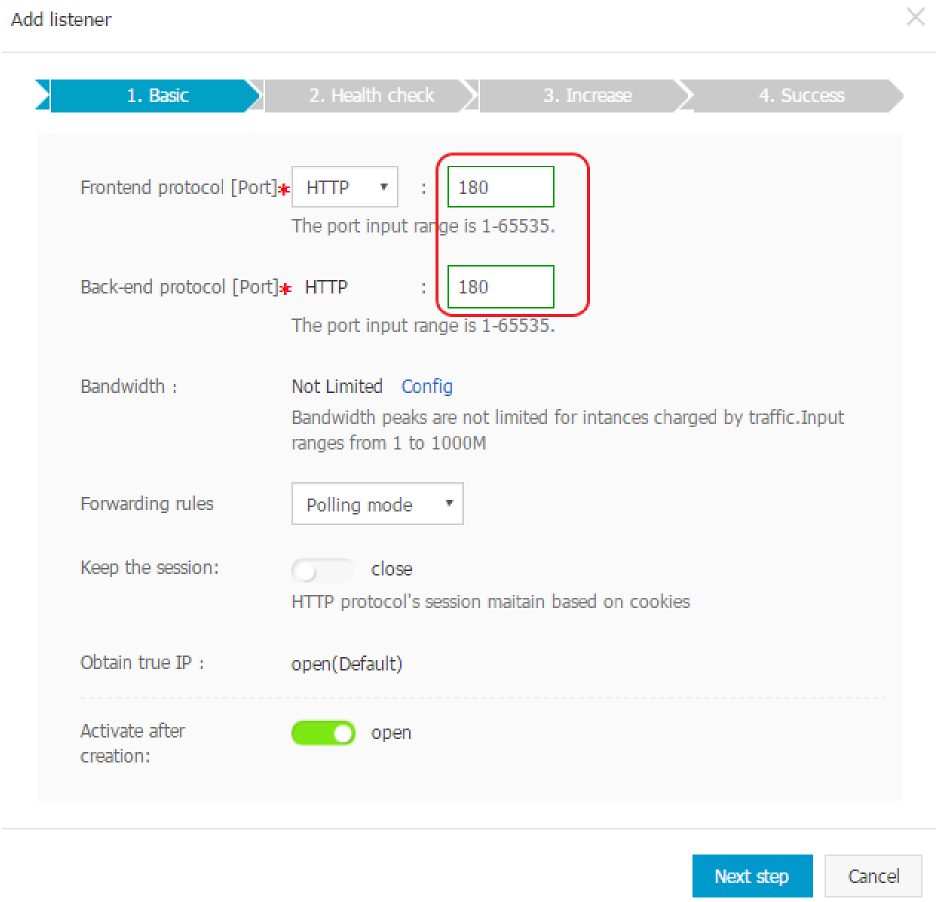
No, Server Load Balancer does not support CNAMEs. The domain names are directly resolved to the service IP address provided by the Server Load Balancer.

How do I configure multiple sites through a single Server Load Balancer instance?

If you have multiple ECS instances and one Server Load Balancer instance. You can configure different host headers for multiple sites to run on the ECS instances at the same time but in different domain names.

Log on to the Server Load Balancer console and then add ECS instances as needed.

Configure the Server Load Balancer listener. In this example, the ports are 80 respectively for the frontend and the backend.



Frontend protocol [Port]* HTTP : 180
The port input range is 1-65535.

Back-end protocol [Port]* HTTP : 180
The port input range is 1-65535.

Bandwidth : Not Limited Config
Bandwidth peaks are not limited for intances charged by traffic.Input ranges from 1 to 1000M

Forwarding rules Polling mode

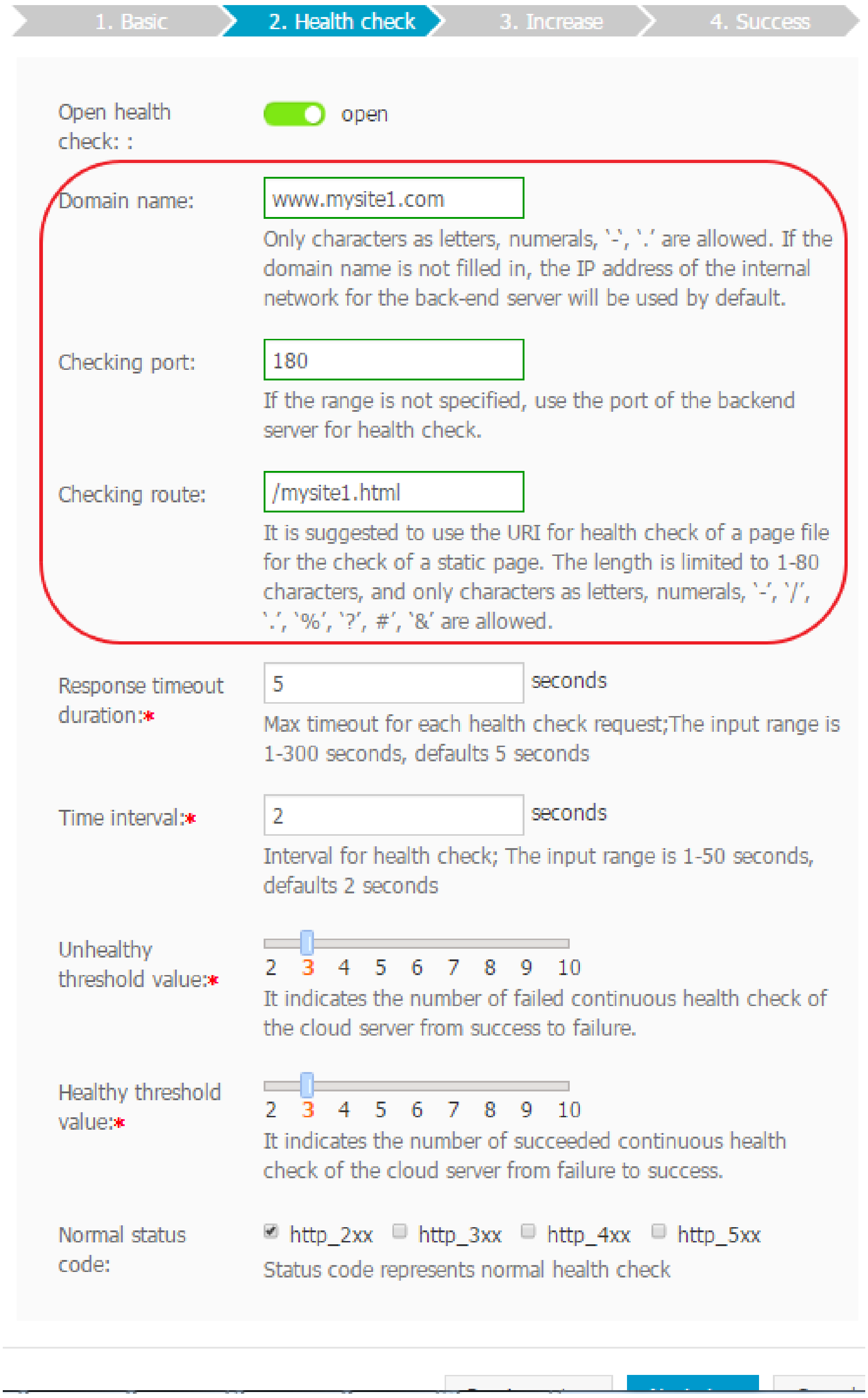
Keep the session: close
HTTP protocol's session maitain based on cookies

Obtain true IP : open(Default)

Activate after creation: open

Configure the Server Load Balancer health check to comply with the following restrictions:

- The domain name is the same as that in the ECS host header.
- The health check port is the same as the port set for the backend ECS. In this example, it is 180.
- The check path is the ECS site’ s file location. In this example, the default page is mysite1.html is used.



Open health check: : open

Domain name: www.mysite1.com

Only characters as letters, numerals, '-', '.' are allowed. If the domain name is not filled in, the IP address of the internal network for the back-end server will be used by default.

Checking port: 180

If the range is not specified, use the port of the backend server for health check.

Checking route: /mysite1.html

It is suggested to use the URI for health check of a page file for the check of a static page. The length is limited to 1-80 characters, and only characters as letters, numerals, '-', '/', '.', '%', '?', #', '&' are allowed.

Response timeout duration:* 5 seconds

Max timeout for each health check request;The input range is 1-300 seconds, defaults 5 seconds

Time interval:* 2 seconds

Interval for health check; The input range is 1-50 seconds, defaults 2 seconds

Unhealthy threshold value:* 2 3 4 5 6 7 8 9 10

It indicates the number of failed continuous health check of the cloud server from success to failure.

Healthy threshold value:* 2 3 4 5 6 7 8 9 10

It indicates the number of succeeded continuous health check of the cloud server from failure to success.

Normal status code: ☑ http_2xx □ http_3xx □ http_4xx □ http_5xx

Status code represents normal health check

Test the configuration.

Can I modify the Server Load Balancer

instance type?

No. The Server Load Balancer system allocates addresses (public/private IP addresses) according to the instance type. To change the type of an instance, you must delete the instance, and then create a new instance and set it to the desired type.

Will Server Load Balancer be affected if the ECS public network adapter is disabled?

Yes. When the public network adapter is enabled, the default route is a public route. When the public network adapter is disabled, packets cannot be returned, which affects Server Load Balancer.

It is recommended that you do not disable the public network adapter. If you need to disable it, change the default route to a private route to avoid affecting Server Load Balancer. You still need to consider service dependence on the public network, such as access to RDS through the public network.

Why can I not add ECS instances in different regions to the same Server Load Balancer instance?

Server Load Balancer does not support cross-region deployment. Only backend ECS servers of the same region and the same account can be added to a Server Load Balancer instance.

Internet bandwidth FAQs

Do backend ECS instances require the Internet bandwidth?

Backend ECS instances do not require Internet bandwidth, as the Server Load Balancer system communicates with backend ECS instances through the intranet. If you need to provide services externally through Server Load Balancer and ECS, you must configure sufficient Internet bandwidth for the corresponding ECS instances or use ECS on a Pay-As-You-Go basis.

Is bandwidth related to the service capability of Server Load Balancer?

The service capability of Server Load Balancer is unrelated to the Internet bandwidth of backend ESC instances.

Forwarding methods supported by Server Load Balancer

Server Load Balancer supports two types of forwarding: weighted round-robin and weighted least-connection.

In Round-Robin Mode, external and internal access requests are distributed to the backend ECS instances in order. In Least Connection Mode, these requests are distributed to the backend ECS instance with the lowest number of connections.

What ports can be enabled in AntCloud Server Load Balancer?

In AntCloud Server Load Balancer, the following ports can be enabled to the Internet:

- Port 80
- Port 443
- Ports 2800–3300
- Ports 6000–10000
- Ports 13000–14000

How can I obtain the real IP address of a visitor?

Use the following methods to obtain the real IP address of a visitor:

- For the Layer-4 Server Load Balancer service, you can obtain the real IP address of a visitor directly on the backend ECS, without additional configuration.
- For the Layer-7 Server Load Balancer service, you have to configure the application server,

and then use the X-Forwarded-For method to obtain the real IP address of a visitor.
This document describes how to obtain the IP address of a visitor for the following Layer-7 Server Load Balancer applications:

- IIS 6
- IIS 7
- Apache
- Nginx

IIS 6 configuration

To obtain the real IP address through the IIS 6 log, you must first install the plug-in F5XForwardedFor.dll.

Download the F5XForwardedFor.dll installation package.

Copy the F5XForwardedFor.dll file from the x86\Release or x64\Release directory into a directory such as C:\ISAPIFilters. The IIS process should be enabled with the right to read the directory.

Open the IIS manager.

Right-click your website and then click **Attribute**.

Switch to **ISAPI filter** and click **Add**.

Enter **F5XForwardedFor** in the **Filter name** field and the full path of F5XForwardedFor.dll in the **Executable file** field, and then click OK.

Reboot the IIS server for the configuration to take effect.

IIS 7 configuration

To obtain the real visitor IP addresses through the F5XForwardedFor module, you must first install the F5XForwardedFor module.

Download the F5XForwardedFor installation package.

Copy the F5XFFHttpModule.dll and F5XFFHttpModule.ini files from the x86\Release or x64\Release directory into a directory such as C:\F5XForwardedFor. The IIS process should have the right to read the directory.

Open the IIS manager.

Double-click **Modules** and then click **Configure Native Modules**.

Click the **Register**.

Add the F5XFFHttpModule.dll file and then click **OK**.

Add the .dll files on **API and CGI restriction** and change into **Allow**.

Reboot the IIS server for the configuration to take effect.

Apache configuration

Install a third-party module mod_rpaf of Apache from the following website.

```
 wget http://stderr.net/apache/rpaf/download/mod_rpaf-0.6.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
```

Modify the apache configuration /alidata/server/httpd/conf/httpd.conf file and add the following at the end.

```
 LoadModule rpaf_module modules/mod_rpaf-2.0.so
RPAFenable On
RPAFsethostname On
RPAFproxy_ips ip address
RPAFheader X-Forwarded-For
```

Note: The IP address for **RPAFproxy_ips** is not the public network IP provided by the Server Load Balancer instance. You can obtain the IP address by checking the Apache log and there are two IP address usually. Both of the IP addresses should be provided as shown in the following example.

```
 LoadModule rpaf_module modules/mod_rpaf-2.0.so
RPAFenable On
RPAFsethostname On
RPAFproxy_ips 10.242.230.65 10.242.230.131
```

```
RPAFheader X-Forwarded-For
```

Reboot Apache.

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

Nginx configuration

The Server Load Balancer uses Nginx to obtain a real IP through http_realip_module. Howver, in the default one-click installation package there is no http_realip_module for Nginx.

Recompile the Nginx and add --with-http_realip_module.

```
 wget http://soft.phpwind.me/top/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`
```

Open the nginx.conf file.

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

Add set_real_ip_from <IP Address>; and real_ip_header X-Forwarded-For; at the end of the following 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;
```

Note: The IP address for **set_real_ip_from** is not the public network IP provided by the Server Load Balancer instance. You can obtain the IP address by checking the Nginx log. If there are multiple IP addresses, all of them should be provided.

Reboot Nginx.

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

Troubleshoot health check exceptions

For information about how to implement health check, click here.

Layer-7 (HTTP protocol)

For a Layer-7 (HTTP protocol) service, if an expectation occurs in the listener health check, do the following:

- Ensure that you can access your application service through ECS directly.
- Ensure the application server port is listening to the intranet address.
- Ensure that the backend server has enabled the corresponding port, and that this port is consistent with the backend port for Server Load Balancer listening.
- Check whether the backend ECS has a firewall or other security protection software enabled, which may blocking the local IP address of the Server Load Balancer system and disabling communication between the Server Load Balancer system and the backend server.
- Check whether the Server Load Balancer health check parameters are correctly set. It is recommended that the default values are used.
- Check that, if the page for health check is not the default page on the backend ECS application server, the URL for the health check configuration page is specified. It is recommended that you use static pages to conduct a health check.
- Check whether there is a high load on the backend ECS resources, which may cause ECS to respond slowly.

Additionally, you can also check the intranet access. Because the Server Load Balancer and backend ECS communicate with each other over the intranet, it is necessary for ECS to monitor intranet ports or all ports. The following details how to check intranet communication:

- If the IDs of the Server Load Balancer frontend port and the ECS backend port are both 80 and the ECS intranet IP address is 10.11.192.1, run the following command:
  - For Windows ECS: netstat -ano | findstr :80
  - For Linux ECS: netstat -anp | grep :80
    If you see 10.11.192.1:80 or 0.0.0.0:80 as being listened to, the port is normal.
- Check whether the server intranet firewall allows port 80. The firewall can be temporarily closed for testing, run the following command to close the firewall:
  - For Windows system: firewall.cpl
  - For Linux system: /etc/init.d/iptables stop

- Check the response status code. The HTTP status code must be 200 or another code that represents normal.
    - For Windows: enter the intranet IP address in the browser in the ECS server directly to check whether it is normal.
    - For Linux: use the curl -I command to check whether the status is HTTP/1.1 200 OK. It is recommended that you specify simple pages in HTML format for health checks. Pages featuring dynamic scripting languages like PHP are not recommended.

Layer-4 (TCP protocol)

For Layer-4 Server Load Balancer, use the telnet command of the backend port to test response. Telnet the IP address 10.11.192.1 80 for testing.

Session persistence FAQs

How do I enable session persistence, and how long does it last?

When you configure Server Load Balancer listeners, you can choose whether or not to enable session persistence. You can configure different session persistence policies for different listeners. The maximum session persistence time is 86,400 seconds (24 hours).

What types of session persistence does Server Load Balancer support?

For Layer-7 services (HTTP), the Server Load Balancer system supports cookie-based session persistence. For Layer-4 services (TCP), the Server Load Balancer system supports IP-address-based session persistence.

Why is session persistence not working?

Session persistence does not work in the following scenarios:

- Layer-4: session persistence is implemented based on the source IP address. If the source IP address is changed after the access request goes through an NAT gateway, the session persistence fails.
- Layer-7: session persistence is implemented based on the cookie. If a cookie becomes invalid, for example, if the backend application returns 302 to redirect the request to a Layer-7 address of another Server Load Balancer, session persistence fails.

Can I set different rules for enabling/disabling session persistence

for different domain names?

Yes. You can select Overwrite Service Cookie in Session Persistence provided by the Server Load Balancer system.

Backend ECS server FAQs

Can I increase or decrease the number of backend ECS instances at any time in Server Load Balancer?

Yes. You can increase or decrease the number of backend ECS instances in Server Load Balancer at any time and switch between different backend ECS instances. To ensure your external services remain available and stable when you perform the preceding operations, enable the health check function of Server Load Balancer and ensure that at least one backend ECS instance in Server Load Balancer runs normally.

How to synchronize data between backend ECS servers in Server Load Balancer?

A wide range of tools are available for synchronizing data, such as Rsync. You can also configure your ECS as a stateless application server, and store data and files on RDS and OSS.

Can Server Load Balancer be applied to ECS instances running different operating systems?

Yes. Server Load Balancer has no requirements for the types of operating systems that backend ECS instances run, but you must ensure that your two ECS instances have consistent data and the same application service deployment. However, for management and maintenance convenience, it is recommended that the two ECS instances run the same operating system.

Why does the backend ECS instance in Layer-4 Server Load Balancer instances fail to access its instance service address?

Access failure is related to the realization mechanism of Server Load Balancer TCP. In Layer-4 (TCP) Server Load Balancer, the ECS instances in the backend ECS pool cannot serve as both real servers as well as clients that can send requests to the corresponding Server Load Balancer instances.

This is because the returned packets will only be forwarded among the ECS instances in the backend ECS pool, but not through Server Load Balancer. It is impossible to access the VIP through the backend ECS instances in Server Load Balancer.

Is additional configuration required for the backend ECS instances in Server Load Balancer?

No special configuration is required for the backend ECS instances added to a Server Load Balancer instance. If a Linux ECS instance cannot be accessed normally after being attached to Layer-4 (TCP) Server Load Balancer, check whether the following three values in the system configuration file /etc/sysctl.conf are **0**:

```
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 in the same intranet segment cannot communicate with each other, check whether the following parameters are set correctly:

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

Run the sysctl -p command to update the parameter settings.

Can I build multiple websites on a set of ECS instances and implement Server Load Balancer?

Yes. A Server Load Balancer instance supports up to 50 listener services. Each listener service corresponds to an application on the backend ECS instances (the frontend port of Server Load Balancer corresponds to the application ports on the backend ECS instances). You can configure different host headers for the applications on the backend ECS instances to meet your requirements.

What is the ECS Weight?

You can specify the forwarding weighting of each ECS instance in the backend ECS pool. An ECS instance with the higher weight ratio will receive more access requests. The forwarding weighting can be set based on the external service capacity and status of the backend ECS instance.

- If the session persistence function is enabled, the number of access requests distributed to backend application servers may vary. It is recommended that you disable the function temporarily and check whether the variation persists.

  If Server Load Balancer distributes requests to backend ECS instances unevenly, perform the following actions:

  i. Record the number of web service access logs that the backend ECS instances generate within a specified period.
     - For Nginx and Apache, run log directory/access.log.
     - For IIS:
       a. Open the IIS manager.
       b. Hover the cursor over the site for which you want to enable access logging.
       c. Right-click the site and select **Properties**.
       d. Click the **Website** tab.
       e. Click **Enable Logging**.

     Check whether the log quantities on multiple backend ECS instances are different based on the configuration of Server Load Balancer.

     Note: If the session persistence function is enabled, remove the access logs with identical IP addresses. If the ECS weight ratio is configured, calculate and check whether the percentage of access traffic recorded in access logs corresponds to the weight ratio.

How many ECS instances does a Server Load Balancer instance support?

There is no limit to the number of backend ECS instances that can be configured for a Server Load Balancer instance.

However, in order to ensure the stability and efficiency of your external services, it is recommended that you configure different Server Load Balancer instances with backend application servers based on business types or application modules and use the backend application servers to provide different services or perform different tasks.

Bandwidth peak value FAQ

Why the connection cannot reach the bandwidth peak?

Because the Server Load Balancer system is deployed in clusters to serve your Server Load Balancer instance, all external access requests will be evenly distributed among the Server Load Balancer system servers for forwarding. The bandwidth peak value you have set will be equally shared by the multiple servers.

The formula for a single connection’ s download traffic is as follows:

Individual connection download peak value = the preset total Server Load Balancer bandwidth / (N-1), where N is the number of packets for traffic forwarding (generally 4).

For example, assume you have set a 10 MB bandwidth limit on the console. In this case, each client’ s maximum download traffic is 10 / (4-1), or 3.33 MB.

It is recommended you configure a suitable bandwidth peak value for individual listeners based on your business needs. This ensures that your normal external services will not be affected or restricted.

What is the bandwidth peak of the Server

Load Balancer?

You can set different bandwidth peaks for different listeners to restrict the capacity of different applications on the backend ECS to offer external services.

The rules for setting bandwidth peaks are as follows:

- Up to 50 listeners can be added to a Server Load Balancer instance. Different rules can be set for each listener.
- The bandwidth peak for a single listener can be set in the range of 5 to 1,000 Mbps.
- The bandwidth peak limit can be lifted for larger bandwidth needs.

Disaster tolerance FAQs

Server Load Balancer disaster tolerance

Server Load Balancer supports the following disaster tolerance features:

- Addition of ECS instances in different zones of the same region to a single Server Load Balancer instance. This allows the Server Load Balancer instance to switch between the two equipment rooms in the same region without user intervention.

  Intra-city disaster tolerance.

  NOTE: Intra-city disaster tolerance cannot be implemented in a region that has only one equipment room.

To implement disaster tolerance for Server Load Balancer:

- The backend ECS instances attached to the same Server Load Balancer instance must be distributed in multiple zones located in the same region.
- Different zones in the same region must be selected as the master and backup zones for the Server Load Balancer instance, which is attached to the backend ECS instances in different zones.
- Multiple Server Load Balancer instances must be created in different regions to provide external services through DNS round robin, achieving cross-region availability.

Server Load Balancer multi-zone FAQs

What is the benefit of multi-zone deployment?

In order to provide more reliable services, Alibaba Cloud Server Load Balancer has already deployed multiple zones in each region to achieve cross-machine-room disaster recovery.

- When the primary zone’ s machine room is faulty and unavailable, Server Load Balancer can rapidly switch to the machine room of the backup zone to restore its service capabilities within 30s.
- When the primary zone is restored, the Server Load Balancer service will automatically switch back to its machine room for service provision.

What is a zone?

Cloud product zones are independent infrastructure sets, also known as Internet datacenters (IDCs). Different zones have independent infrastructures (network, power supply, air-conditioning, etc.). Therefore, an infrastructure fault in one zone will not affect the other zones.

On what dimension is the multi-zone feature based?

Zones belong to specific regions. A single region may have one or more zones. In most regions, the Server Load Balancer service is deployed to two zones.

What are the details for Server Load Balancer zones in each region?

The zone details for each region are as follow:

Region	Zone Type	Primary Zone	Backup Zone
East China 1 (Hangzhou)	Multi-zone	Zone D	Zone B
	Multi-zone	Zone B	Zone D
North China 2 (Beijing)	Single Zone	Zone B	-
South China 1 (Shenzhen)	Multi-zone	Zone A	Zone B
	Multi-zone	Zone B	Zone A
North China 1 (Qingdao)	Multi-zone	Zone A	Zone B
	Multi-zone	Zone B	Zone A
East China 2 (Shanghai)	Multi-zone	Zone A	Zone B

	Multi-zone	Zone B	Zone A
Hong Kong	Single Zone	Zone B	-
Western US (Silicon Valley)	Multi-zone	Zone 1A	Zone 1B
	Multi-zone	Zone 1B	Zone 1A
Eastern US (Virginia)	Single Zone	Zone 1A	-
Asia-Pacific (Singapore)	Single Zone	Zone A	-

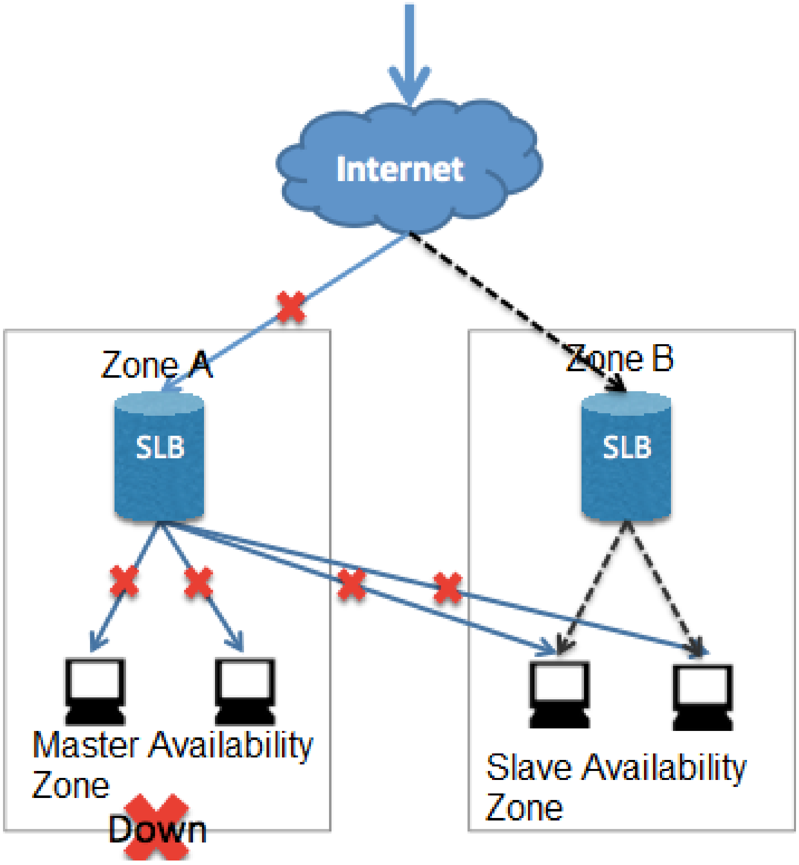
What is the difference between single-zone and multi-zone regions?

In a single-zone region, any instance created by a user can be stored only in a single zone in that region. When a user creates an instance in a multi-zone region, this instance can be stored simultaneously in two zones. By default, the instance is stored in the primary zone. However, when the primary zone is faulty, the instance will automatically switch to the backup zone. This greatly increases local availability.

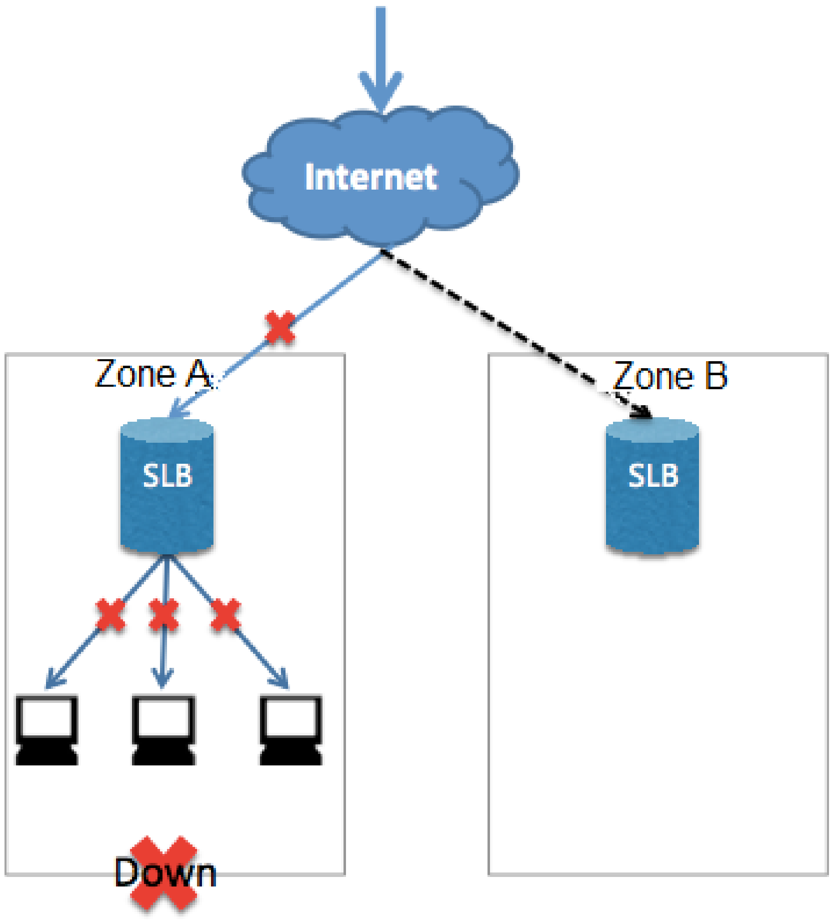
How can the Server Load Balancer multi-zone feature work with other products to support higher-availability or lower-latency solutions?

In addition to selecting a multi-zone region that supports local disaster recovery, it is recommended that users consider deploying backend servers to achieve greater regional reliability and availability.

> As the following figure shows, ECS instances are bound to different zones under a single Server Load Balancer instance. When Zone A works normally, user access traffic will follow the path of the blue line shown in the image; when a fault occurs in Zone A, user access traffic will follow the path of the black line. This prevents a fault in a single zone from causing total service loss.



By selecting the zones for different products, users can have a lower latency as shown in the following figure. ECS instances are bound to the primary zone (Zone A) under a single Server Load Balancer instance. This way, when Zone A works normally, user access traffic will follow the path of the blue line shown in the image above; when a fault occurs in Zone A, user access traffic will follow the path of the black line. This deployment method lowers the latency obviously at the cost of high availability.



Are users charged for the multi-zone feature?

Currently, the multi-zone feature is free of charge for users.

SLB performance FAQ

Why does the Layer-7 Server Load Balancer have a low stress test performance?

Server Load Balancer clusters use LVS and Tengine. Layer-4 listeners go through LVS directly to the backend servers, while Layer-7 listeners must go through Tengine after LVS to finally reach the backend servers. This means that Layer-7 has one more processing stage than Layer-4, resulting in a lower performance.

Some users may also discover that attaching two ECS instances connected to a Layer-7 Server Load Balancer listener does not support a single ECS instance.

The possible causes include:

Case 1: insufficient client ports

An insufficient number of client ports may cause connection establishment to fail, especially during stress testing. By default, the Server Load Balancer will erase the timestamp attributes of TCP connections so that the Linux protocol stack’ s tw_reuse (time_wait status connection reuse) is invalid. As a result, time_wait status connections may build up, causing an insufficiency of client ports.

Solution

Allow client ports to use persistent connection to replace short connection. Use RST packets for disconnection (socket sets the SO_LINGER attribute), rather than sending FIN packets.

Case 2: full backend server accept queue

If the backend server’ s accept queue is full, the backend server will not send syn_ack packets, causing the client to time out.

Solution

The net.core.somaxconn parameter’ s default value is 128. Execute the command sysctl -w net.core.somaxconn=1024 or use a higher value and then restart the backend server application.

Case 3: excessive number of backend server connections

Due to the architectural design, when using a Layer-7 Server Load Balancer, your persistent connections will change to short connections after going through Tengine. This may cause an excessive number of backend server connections, resulting in poor performance during stress testing.

Case 4: applications that backend servers depend on cause bottlenecks

After requests pass through the Server Load Balancer to the backend servers, the backend servers may have all their applications dependent on other applications, such as databases, resulting in bottlenecks.

Case 5: backend server abnormal health check status

If a backend server health check failed, or its health status changes frequently, this may result in poor performance.

Stress Testing Suggestions

A stress test is mainly intended to measure the forwarding capability of a Server Load Balancer, as well as its session persistence and balancing capability. Therefore, it is recommended you use short connections to test Server Load Balancer and backend server processing capabilities. However, short connections for stress testing may cause the problem of insufficient client ports as described in preceding cases.

Recommended parameters

- Persistent connections to stress test Server Load Balancer throughput. This is used to test the upper bandwidth limit or special services.
- A lower stress test tool timeout (5 seconds). If the timeout time is too long, the test result may show an increased average RT, making it difficult to determine whether the proper stress test level has been reached. If the timeout time is kept low, the test result may indicate the success rate, making it easy to determine the stress test level. The backend servers provide a static webpage for stress testing, so as to prevent the loss caused by application logic (I/O and DB).

Recommended stress testing tools

It is recommended you use Alibaba Cloud PTS.

The PTS can select multiple clients as stress test sources and produce clear test results. It also allows you to obtain backend server performance data during a stress test by enabling the monitoring function.

Using apache ab is not recommended, as in highly-concurrent scenarios, ab may experience tiered interruptions of 3s, 6s, and 9s. The ab tool determines whether a request is successful based on the content length. When multiple backend servers are attached to the Server Load Balancer, the returned content length will be inconsistent, affecting the test result.

Recommended stress test configuration requirements

The following settings are only used to perform stress testing on Server Load Balancer capabilities. Your actual production environment does not need to be configured in this way.

- Do not enable session persistence during the monitoring process; otherwise, pressure may be concentrated on a certain backend server.
- Disable health check during the monitoring process to reduce the impact of health check requests on the backend servers.
- Perform stress testing using multiple clients (>5) so that source IP addresses are scattered, to better simulate real conditions.

Declaration of uploading Server Load Balancer instances to performance guaranteed instances

In late April 2017, Alibaba Cloud Server Load Balancer will be upgraded to performance-guaranteed instances, beginning with the Northern China 1 Region. This will allow users to select Server Load Balancer instances of different specifications based on their needs. In addition, super-large performance specifications will be available on Alibaba Cloud in select regions.

Alibaba Cloud will provide a guarantee for the performance of these instances. Users will be able to configure and query performance indicators and view operating data in real-time.

Alibaba Cloud will also increase the maximum bandwidth for Server Load Balancer instances to 5Gbps.

Date of purchase	**Performance type**	**Performance indicators**		
		Max number of connections	Number of new connections per second (CPS)	Number of queries per second (QPS)
Before late April, 2017	Performance sharing			
After late April, 2017	Performance guaranteed			

indicates that the performance is guaranteed. indicates that the performance is not guaranteed.

Besides, from late April, 2017, Alibaba Cloud will adjust the maximum bandwidth for sale for Server Load Balancer instances on the Internet to 5Gbps.

Definition of performance specifications of performance guaranteed instances

According to online data analysis, the performance indicators of Specification 1 can meet needs of most users. To minimize your costs, Alibaba Cloud will not collect performance specification fee on instances of Specification 1.

Performance specification	Max number of connections	Number of new connections per second (CPS)	Number of queries per second (QPS)
Specification 1	5000	3000	1000
Specification 2	50000	5000	5000
Specification 3	100000	10000	10000
Specification 4	200000	20000	20000

Specification 5	500000	50000	50000
Specification 6	1000000	100000	100000

NOTE:

- Server Load Balancer instances of higher performance specifications will be available. If you need such instances, you can submit a Ticket application.
- Definitions of performance specifications of Server Load Balancer instances might be adjusted according to feedback from users during the trial period.

For more information about performance guaranteed instances, you can refer to relevant documentations on Alibaba Cloud official website later.

What are the connection timeout settings for Server Load Balancer listeners?

The Server Load Balancer currently does not allow users to set the timeout time. The connection timeout settings for Server Load Balancer listeners are as follows:

- TCP 900 sec
- UDP 300 sec
- HTTP 60 sec
- HTTPS 60 sec

The following are reasons that may cause access timeout in a VIP connection.

Note: There are many possible reasons. This document focuses on server issues.

Case 1: Security protection activated for VIP connections

This could be due to traffic black holes or cleaning, or WAF protection (WAF is distinguished by sending RTS packets to both the client and LVS after a connection is established.)

Case 2: Insufficient client ports

An insufficient number of client ports may cause connection establishment to fail, especially during stress testing. By default, the Server Load Balancer will erase the timestamp attributes of TCP connections so that the Linux protocol stack’ s tw_reuse (time_wait status connection reuse) is invalid. As a result, time_wait status connections may build up, causing an insufficiency of client ports.

Solution

Allow client ports to use persistent connection to replace short connection. Use RST packets for disconnection (socket sets the SO_LINGER attribute), rather than sending FIN packets.

Case 3: Full backend server accept queue

If the backend server’ s accept queue is full, the backend server will not send syn_ack packets, causing the client to time out.

Solution

The net.core.somaxconn parameter’ s default value is 128. Execute the command sysctl -w net.core.somaxconn=1024 (or use a higher value) and then restart the backend server application.

Case 4: A Layer-4 Server Load Balancer backend server attempts to access the Layer-4 VIP

A Layer-4 Server Load Balancer instance is not permitted to access the VIP Server Load Balancer from a backend server. This causes the connection to fail. This commonly occurs when a user’ s backend application uses URL splicing for jump access.

Case 5: Improper RTS processing for connection timeout

After a TCP connection remains inactive for 900 seconds after being created on the Server Load Balancer, the system will send an RST disconnect request to both the client and RS. Some applications handle RST exceptions improperly and may send data to disabled connections, causing an application to time out.