

Server Load Balancer

Best Practices

Best Practices

What are guaranteed-performance instances?

Performance metrics, such as MaxConnection, CPS, and QPS, are included in the guaranteed-performance instance SLA. In contrast, shared-performance instances do not provide the performance guarantees. The Server Load Balancer resources are shared among the shared-performance instances.

The following are three key metrics of 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)

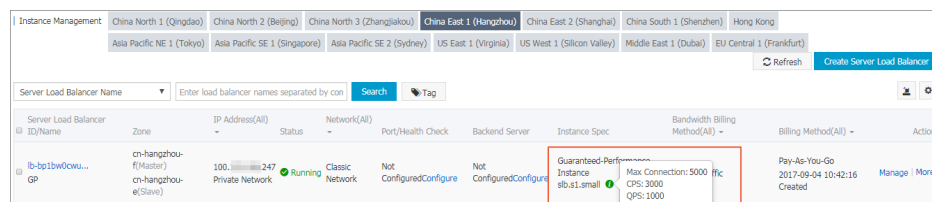
The number of HTTP/HTTPS queries/requests that can be processed per second, which is specific to layer-7 listeners. When the QPS reaches the limits of the specification, the new connection will be dropped.

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

Specification		Max Connection	CPS	QPS
Specification 1	Small I (slb.s1.small)	5000	3000	1000
Specification 2	Standard I (slb.s2.small)	50000	5000	5000
Specification 3	Standard II (slb.s2.medium)	100000	10000	10000
Specification 4	Higher I	200000	20000	20000

	(slb.s3.small)			
Specification 5	Higher II (slb.s3.medium)	500000	50000	30000
Specification 6	Super I (slb.s3.large)	1000000	100000	50000

Before launching guaranteed-performance instances, all the instances that you created previously are shared-performance instances. For the guaranteed-performance instances, you can view the instance type on the console as shown in the following figure.



Release plan of the guaranteed-performance instances

From mid-May 2017, Alibaba Cloud starts upgrading shared-performance instances to guaranteed-performance instances in US East 1 (Virginia). The state quo and release plan for other regions are as follows:

- US East 1 (Virginia): Available
- China North 2 (Beijing): Available
- China East 1 (Hangzhou): Available
- China East 1 (Shanghai): Available
- China North 1 (Qingdao): Available
- China North 3 (Zhangjiakou) and other regions: Upgrading

How to choose specifications for guaranteed-performance instances?

Choose the specification according to your service types, and the overall principle is as follows:

The key factor for layer-4 listeners is the number of concurrent connections of TCP keep-alive connections, then the max connection is considered as the key metric. Depending on the business scenarios, estimate the maximum number of concurrent connections and select the appropriate specification.

The key factor for layer-4 listeners is QPS performance. QPS determines the throughput of a layer-7 application system. Similarly, you also need to estimate the QPS based on experience. After the initial selection of a specification, you can adjust the specification during business

stress test and real test.

Use other monitoring metrics introduced by guaranteed-performance instances to check the actual traffic trends, peak bandwidth, and so on for more accurate selection. For more information, see [Monitoring data](#).

Billing method:

For Pay-AS-You-Go guaranteed-performance instances, you can downgrade or upgrade the configurations, but for the Subscription guaranteed-performance instances, you can only upgrade the configurations.

We recommend that you purchase a Pay-AS-You-Go guaranteed-performance instance for testing. When the specification is decided, purchase a Subscription guaranteed-performance instance.

Additionally, if you also change the billing method (from PayByTraffic to PayByBandwidth, vice versa) when changing the specification of a guaranteed-performance instance, the modification will take effect at 00:00 the next day. If you only change the specification, the modification takes effect immediately. We recommend that you do not change the billing method while changing the specification.

How to handle the original shared-performance instances?

The original shared-performance instances will not be automatically upgraded to guaranteed-performance instances and also will not be charged for the specification fee.

You can manually upgrade them to guaranteed-performance instances. After upgrading, you will be charged for the specification fee accordingly.

Note: Some of the shared-performance instances may be deployed in an old cluster. When upgrading these instances to guaranteed-performance instances, a service interruption of 10-30 seconds may occur during the migration of the instances. We recommend upgrading these instances in a low traffic period. The upgrading of the guaranteed-performance instances has no impact on the services.

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

Guaranteed-performance instances do not guarantee that the three metrics (including the peak

bandwidth) can reach the specification limits at the same time. That is, when one metric first reaches its limit, limitation is triggered.

For example, you have purchased a guaranteed-performance instance of specification higher I (slb.s3.small). 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.

Similarly, if your billing method for the guaranteed-performance instance is PayByBandwidth, when the peak bandwidth is reached, the new connections will be dropped even though the instance does not reach the performance specification limits.

Why sometimes the performance of a guaranteed-performance instance is worse than that of a shared-performance instance?

Shared-performance instances share all the resources. Their performance may be better than guaranteed-performance instances when the traffic is low. However, in the situation of high traffic, shared-performance instances do not guarantee the performance but guaranteed-performance instances do.

When can I use API to create and modify guaranteed-performance instances?

Currently, the creation and modification of guaranteed-performance instances is not supported by Server Load Balancer API. Check your registered email account and Alibaba Cloud website for further notifications.

Can I still buy shared-performance instances?


Yes. However, shared-performance instances will be unavailable in the future. Check the registered email account and Alibaba Cloud website for further notifications.

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 the **Rewrite Cookie** method, you can set the **Cookie Name** as name, and set the key of vip.a.com 's cookie as name on

the backend server.



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

Apache

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

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

Add the following configurations in the VirtualHost file.

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

Nginx

Configure the configuration file 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 configuration file 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" )
}
```

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.

Prerequisites

You have created 2 ECS instances and granted access to their SSH and Web ports.

Procedure

Call `CreateLoadBalancer` interface 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"
}
```

Call `CreateLoadBalancerHttpListener` interface 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`

Call `SetLoadBalancerStatus` interface to active the Server Load Balancer instance.

Request:

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

Call `AddBackendServers` interface to add an ECS instance to backend servers.

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
      }
    ]
  }
}
```

Call `AddBackendServers` interface again to add an ECS instance to backend servers.

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
}
]
}
}
```

Call DescribeLoadBalancerAttribute interface 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.

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.

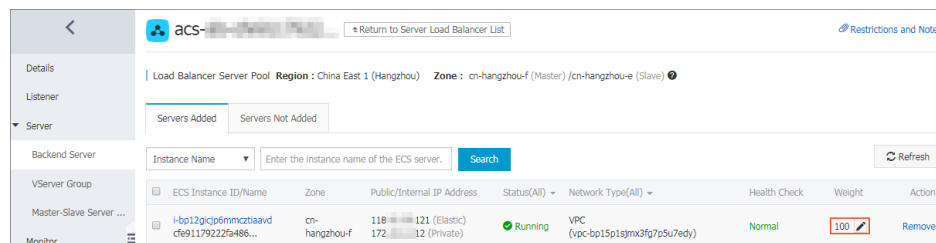
Log on Server Load Balancer console.

Choose a region and then click the ID of the target Server Load Balancer instance.

In the left-side navigation pane, click **Server** > **Backend Server**.

If the ECS instance is added to a server group, click **VServer Group** or **Master-Slave Server Group** accordingly.

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



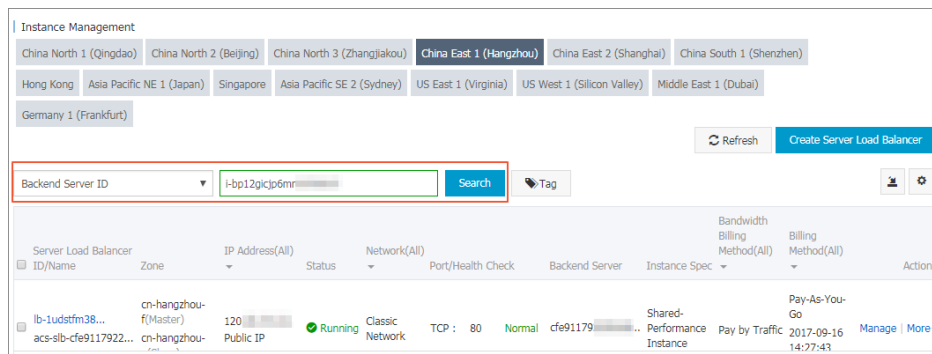
When no traffic is distributed to the ECS instance, click **Remove** to remove it from the backend server pool.

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.

```

~# netstat -ano
Active Internet connections (servers and established)
Proto Recv-Q Send-Q Local Address           Foreign Address         State       Timer
tcp        0      0 0.0.0.0:22              0.0.0.0:*               LISTEN      off (0.00/0/0)
tcp        0      0 0.0.0.0:111             0.0.0.0:*               LISTEN      off (0.00/0/0)
tcp        0      0 172.16.0.0:42285        0.0.0.0:*               ESTABLISHED off (0.00/0/0)
tcp        0      0 172.16.0.0:22          0.0.0.0:*               ESTABLISHED on (0.16/0/0)
tcp6       0      0 :::111                  :::*                     LISTEN      off (0.00/0/0)
udp        0      0 0.0.0.0:42947           0.0.0.0:*               off (0.00/0/0)
udp        0      0 0.0.0.0:68              0.0.0.0:*               off (0.00/0/0)
udp        0      0 0.0.0.0:111             0.0.0.0:*               off (0.00/0/0)
udp        0      0 0.0.0.0:627             0.0.0.0:*               off (0.00/0/0)
udp        0      0 172.16.0.0:123          0.0.0.0:*               off (0.00/0/0)
udp        0      0 127.0.0.1:123           0.0.0.0:*               off (0.00/0/0)
udp        0      0 0.0.0.0:123            0.0.0.0:*               off (0.00/0/0)
udp6       0      0 :::111                  :::*                     off (0.00/0/0)
udp6       0      0 :::627                  :::*                     off (0.00/0/0)
udp6       0      0 :::123                  :::*                     off (0.00/0/0)
udp6       0      0 :::1275                 :::*                     off (0.00/0/0)
Active UNIX domain sockets (servers and established)
Proto RefCnt Flags       Type       State       I-Node  Path
unix   2      [ ]         DGRAM     7689       /run/systemd/shutdown
unix   7      [ ]         DGRAM     7691       /run/systemd/journal/dev-log
unix   2      [ ]         DGRAM     7695       /run/systemd/journal/dev-log

```

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.

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 web applications

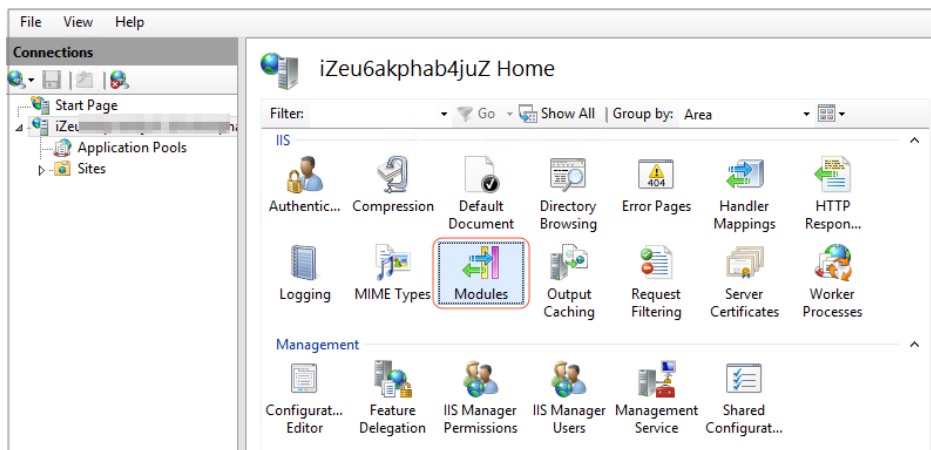
This section introduces some common methods used to configure web applications.

Configure IIS7/IIS8

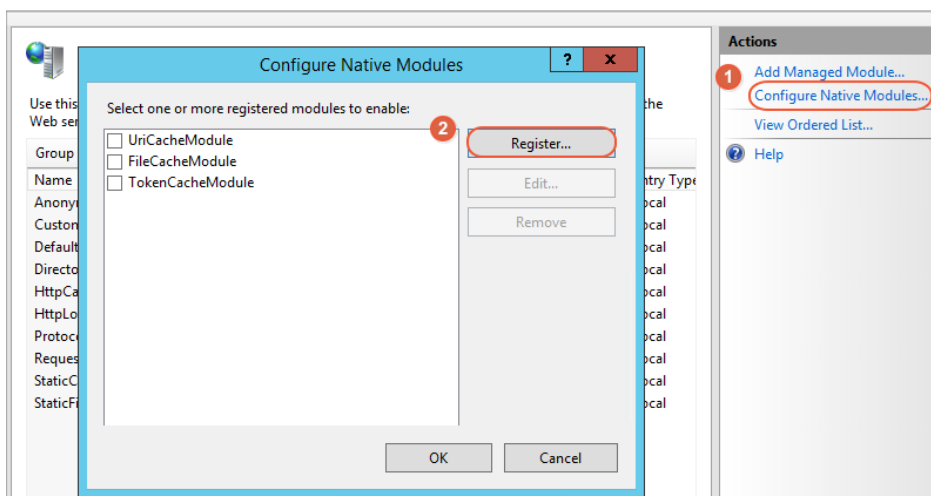
Download and extract the F5XForwardedFor.

Copy the F5XFFHttpModule.dll and F5XFFHttpModule.ini files from the extracted folder to a folder, such as C:\F5XForwardedFor\. Make sure that the IIS process has the write permission to this folder.

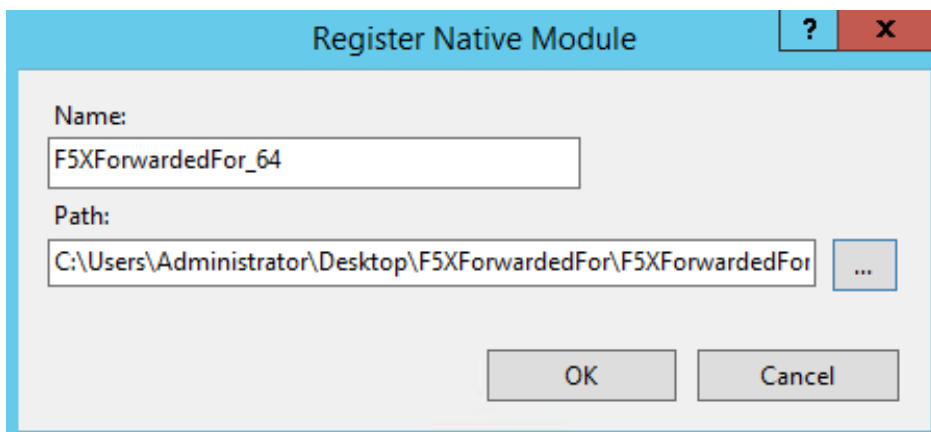
Open the IIS Manager, and then double-click the **Modules** function.



Click **Configure Native Modules**, and then click **Register**.

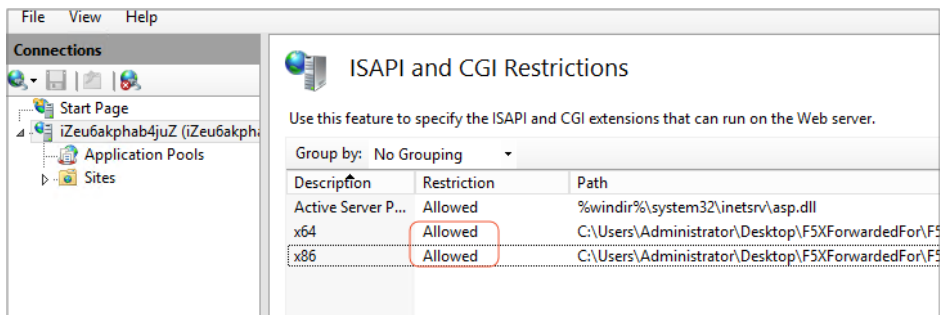


Add the copied the .dll file.



Add the ISAPI and CGI restrictions for the .dll file and set the restriction to **Allowed**.

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



Restart the IIS Manager.

Configure Apache

Run the following command to install the mod_rpaf module.

```
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
```

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

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

`RPAFproxy_ips`: the IP address is not the IP address of the Server Load Balancer instance. Check the Apache log to find the IP address, usually both the two IP addresses are entered.

Run the following command to restart the Apache server.

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

Configure Nginx

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`
```

Run the following command to open the nginx.conf file.

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

Find the following content and add the required information after it.

```
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 information to be added:

```
set_real_ip_from IP_address
real_ip_header X-Forwarded-For;
```

set_real_ip_from IP: the IP address is not the IP address of the Server Load Balancer instance. Check the Nginx log to find the IP address, usually both the two IP addresses are entered.

Run the following command to restart the Nginx server.

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

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 request	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

Instance ID/Name	Zone	IP Address	Status(All)	Network Type(All)
i-bp1huan9mmlu3jvcmk0cmg7 SLB_jerry1	China East 1 Zone F	47.96.179.21(Elastic IP Address) 172.16.19.13(Private IP Address)	Running	VPC
i-bp1k2etkayhmgpku2py SLB_jerry2	China East 1 Zone F	47.96.172.48(Elastic IP Address) 172.16.33.32(Private IP Address)	Running	VPC
i-bp138ue0e0k0k0du7rea SLB_tom1	China East 1 Zone F	115.62.125.14(Elastic IP Address) 172.16.19.20(Private IP Address)	Running	VPC
i-bp132pewrhvmtfhuht SLB_tom2	China East 1 Zone F	47.96.169.125(Elastic IP Address) 172.16.30.36(Private IP Address)	Running	VPC

Procedure

Create an Internet-facing SLB instance.

For details, see [Create a server load balancer](#).

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.

Create two VServer groups.

Locate the newly created instance in the Server Load Balancer console and click the instance ID to go to the **Instance Details** page.

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

Click **Create VServer Group**.

In the dialog box that appears, select the backend servers to be added and set ports and weights for them respectively. The ports for ECSs 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).

Notice: The network type of current server load balancer is VPC, instance type is Intranet IP. This VServer group can only add a VPC ECS.

1

*Group Name: TOM

*Server Network Type: ☐ Classic Network ☒ VPC

Instance Na Enter the name of the ECS instan

Available Servers		
ECS Instance ID/Name	IP Address	Zone
I-bp1han9m6x3ecmk37mqZ SLB_jerry1	47.96.175.121 (EIP) 172.16.33.33 (Private)	cn-hangzhou-f VPC
I-bp162wb1ayjhmp6u2py SLB_jerry2	47.96.172.148 (EIP) 172.16.33.32 (Private)	cn-hangzhou-f VPC
I-bp18us9p89pilgc7yaa SLB_tom1	116.62.128.54 (EIP) 172.16.33.30 (Private)	cn-hangzhou-f VPC
I-bp1e9bjg74lm9ofxxddg SLB_tom2	116.62.158.112 (EIP) 172.16.33.29 (Private)	cn-hangzhou-f VPC

Selected Servers Add(2/20)				
ECS Instance ID/Name	IP Address	Zone	*Port	*Weight
I-bp18us9p89pilgc7yaa SLB_tom1	116.62.128.54 (EIP) 172.16.33.30 (Private)	cn-hangzhou-f VPC	80	100
I-bp1e9bjg74lm9ofxxddg SLB_tom2	116.62.158.112 (EIP) 172.16.33.29 (Private)	cn-hangzhou-f VPC	80	100

Note: Already added an ECS instance in a VPC (ID: vpc-bp1w92wjrgz01fm6pubd8). Only ECS

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

Add a listener.

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

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.

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

training_SLB [Return to Server Load Balancer List](#) [Restrictions and Notes](#)

Listeners [Add Listener](#) [Refresh](#)

Front-end Protocol/Port	Backend Protocol/Port	Status	Forwarding Rules	Session Persistence	Health Check	Peak Bandwidth	Server Group	Actions
HTTP: 80	HTTP: 80	Running	Round Robin	Disable	Enable	No Limits	-	Configure Details Add Forwarding Rules More

Start Stop Delete

[Activate](#)
[Stop](#)
[Delete](#)
[Set Access Control](#)
[Add Forwarding Rules](#)

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

Configure three forwarding rules.

Add Forwarding Rules

Rule Name	Domain Name	URL	VServer Group	Actions
rule1	www.aaa.com	/jerry	JERRY ▼	Delete
rule2	www.aaa.com	/tom	TOM ▼	Delete
rule3	www.aaa.com		TOM ▼	Delete

Add Forwarding Rule

* Domain name rule:
The domain name can contain letters a-z, numbers 0-9, hyphens (-), and periods (.), and wildcard characters. The following two domain name formats are supported:
- Standard domain name: www.test.com
- Wildcard domain name: *.test.com. wildcard (*) must be the first character in the format of (*.*)

* URL rule:
URLs must be 2-80 characters in length. Only letters a-z, numbers 0-9, and characters '-', '/', '?', '%', '#', and '&' are allowed. URLs must be started with the character '/', but cannot be '/' alone.

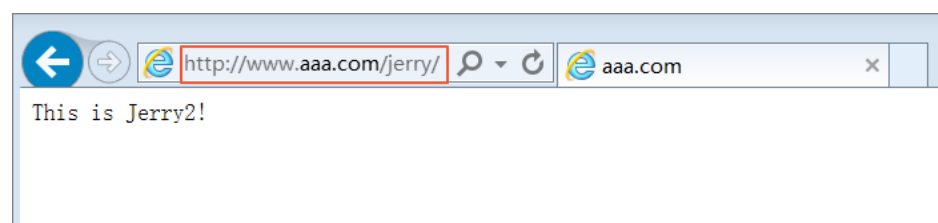
* At least one domain name rule or URL rule is required.

Confirm

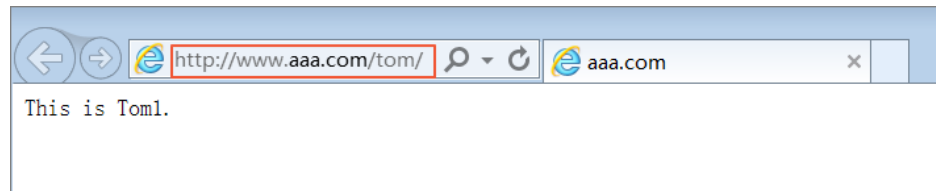
Cancel

Test:

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



Enter www.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.

