Server Load Balancer

Best Practices

MORE THAN JUST CLOUD | Alibaba Cloud

Best Practices

What are guaranteed-performance instances?

The performance metrics, such as MaxConnection, CPS, and QPS, are included in the guaranteed-performance instance SLA. In contrast, the 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)

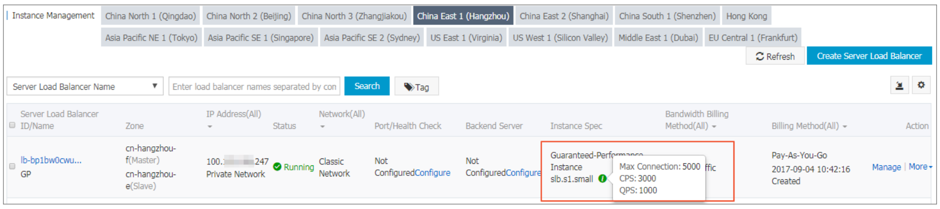
The number of HTTP/HTTPS queries/requests that can be processed per second, which is specific to the layer-7 listener. When the QPS reaches the limits of the specification, the new connection will be dropped.

Alibaba Cloud Server Load Balancer provides the following specifications of the guaranteed-performance instances for you to choose:

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 specification on the console as shown in the following figure.


Instance Management
China North 1 (Qingdao)
China North 2 (Beijing)
China North 3 (Zhangjiakou)
China East 1 (Hangzhou)
China East 2 (Shanghai)
China South 1 (Shenzhen)
Hong Kong
Asia Pacific NE 1 (Tokyo)
Asia Pacific SE 1 (Singapore)
Asia Pacific SE 2 (Sydney)
US East 1 (Virginia)
US West 1 (Silicon Valley)
Middle East 1 (Dubai)
EU Central 1 (Frankfurt)
Refresh
Create Server Load Balancer
Server Load Balancer Name
Enter load balancer names separated by con
Search
Tag

Server Load Balancer ID/Name	Zone	IP Address(All)	Status	Network(All)	Port/Health Check	Backend Server	Instance Spec	Bandwidth Billing Method(All)	Billing Method(All)	Action

lb-bp1bw0cwu...
GP
	cn-hangzhou-f(Master)
	cn-hangzhou-e(Slave)
		100.[redacted]247
		Private Network
			Running	Classic Network	Not ConfiguredConfigure	Not ConfiguredConfigure	Guaranteed-Performance Instance
							slb.s1.small
							Max Connection: 5000
							CPS: 3000
							QPS: 1000
								ffic	Pay-As-You-Go
									2017-09-04 10:42:16 Created
										Manage | More

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), following with China South 1 (Shenzhen) and China East 2 (Shanghai). The release plan for other regions are as follows:

- China North 2 (Beijing) and China East 1 (Hangzhou): Mid-August
- China North 1 (Qingdao): Late-August
- China North 3 (Zhangjiakou) and other regions: End of August

How to choose the specification of the guaranteed-performance instances?

You need to choose the specification according to your service types, the overall principle is as follows:

> The key factor of the layer-4 listeners is the number of the concurrent connections of the TCP keep-alive connections, then the max connection is considered as the key metrics. Depending on the business scenarios, you need to estimate the maximum number of concurrent connections and select the appropriate specification.
>
> The key factor of the layer-4 listeners is the performance of the QPS. QPS determines the throughput of a layer-7 application system. Similarly, you also need to estimate the QPS based on the experience. After the initial selection of a specification, you can adjust the specification during the business stress test and real test.

Combined with other monitoring metrics introduced with the guaranteed-performance instances to check the actual business trends, peak bandwidth, and so on for more accurate selection. For details, see Monitoring data.

Billing method:

For the Pay-AS-You-Go guaranteed-performance instances, you can downgrade or upgrade the configurations, but for the Subscription guaranteed-performance instances, you can just upgrade the configurations, but cannot downgrade.

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 in next day at 00：00. 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 the guaranteed-performance instances. After upgrading, you will be charged for the specification fee accordingly.

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

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

The 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, the metrics that the first reaches the limitation, on which the limitation is triggered.

For example, you purchase a guaranteed-performance instance of the specification higher I (slb.s3.small). When the QPS of the instance reaches 20,000 but the number of the maximum connections does not reach 200,000, the new connections are still dropped because the QPS reaches the limitation.

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

Why sometimes the performance of the guaranteed-performance instance are worse than the shared-performance instance?

The shared-performance instances share all the resources. The performance may be better than the guaranteed-performance instances when the traffic is low. However, in the situation of high traffic, the shared-performance instances does not guarantee the performance while the guaranteed-performance instance does.

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

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

Can I still buy the shared-performance instances?

Yes. However, the shared-performance instances will be unavailable in the future. Please check the registered email account and Alibaba Cloud website for 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-7 listeners, session persistence is based on cookies. If you choose the **Rewrite Cookie** method, you have to configure the cookie in the backend server.

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

Apache

Open the httpd.conf file and ensure 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, an ECS instance deployed with a static web page using Nginx is used as an example, and a security rule for allowing access through SSH and web ports is added for the ECS instance.

Task 1 Clone an ECS instance

Create a snapshot for the system disk.

Query the system disk ID of the instance.

Request:

https://ecs.aliyuncs.com/?Action=DescribeInstanceDisks&InstanceId=id5ab1760-3498-4d95-9687-a91545ef90b3

Response:

```
{
"RequestId" : "9F2188AC-AFAC-4F43-B452-C88463B9F069",
"Disks" : {
"Disk" : [
{
"DiskId" : "1008-27930",
"Size" : 20,
"Type" : "system"}]
}
}
```

Create a snapshot for the system disk.

Request:

https://ecs.aliyuncs.com/?Action=CreateSnapshot&InstanceId=id5ab1760-3498-4d95-9687-a91545ef90b3&DiskId=1008-27930&SnapshotName=mytesthost1-init

Response:

```
 {
"RequestId" : "5CA4F9E6-81D2-42E1-A317-4C25284C6939",
"SnapshotId" : "1008-27930-1097358"
}
```

Query the snapshot creation process. When the progress is 100, it indicates that the snapshot has been created.

Request:

https://ecs.aliyuncs.com/?Action=DescribeSnapshotAttribute&RegionId=cn-hangzhou-dg-a01&SnapshotId=1008-27930-1097358

Response:

```
 {
"RequestId" : "8307863A-1415-40EF-9520-8974871E651C",
"SnapshotId" : "1008-27930-1097358",
"SnapshotName" : "mytesthost1-snp-init",
"Progress" : "100",
"CreationTime" : "2013-05-19T03:19Z"
}
```

Create a custom image with the newly created snapshot.

Request:

https://ecs.aliyuncs.com/?Action=CreateImage&RegionId=cn-hangzhou-dg-a01&SnapshotId=1008-27930-1097358&Description=for creating test instances

Response:

```
 {
"RequestId" : "38C930E9-5CE9-4E24-A392-8538FC20D503",
"ImageId" : "m8a1f80fe-ed9d-4156-a7a8-432f66305c36"
}
```

Clone the ECS instance.

With the custom image, an ECS instance with the same configuration can be cloned and the second ECS instance will be created with this ImageID: ImageId=m8a1f80fe- ed9d-4156-a7a8-432f66305c36.

In this example, the ECS instance configuration is as follows.

```
 {
"RequestId" : "850ED7ED-A4D5-40A1-A7EF-C33B74B1296B",
"InstanceId" : "i6b47cd72-843f-4558-b911-2776acae06fb",
"ImageId" : "m8a1f80fe-ed9d-4156-a7a8-432f66305c36",
"RegionId" : "cn-hangzhou-dg-a01",
"ZoneId" : "cn-hangzhou-gy002-a",
"InstanceType" : "ecs.t1.small",
"HostName" : "mytesthost2",
"Status" : "Stopped",
"SecurityGroupIds" : {
"SecurityGroupId" : [
"g1f91e6e8-3c4b-4923-98dd-78aacbd09d17"
]
},
"PublicIpAddress" : {
"IpAddress" : [
"10.10.10.173"
]
},
"InnerIpAddress" : {
"IpAddress" : [
"10.32.148.152"
]
},
"InternetMaxBandwidthIn" : 2,
"InternetMaxBandwidthOut" : 2,
"SerialNumber" : "1fec6c01-7186-2c3e-fa10-a672b8c300ec"
}
```

To distinguish this new ECS instance, change the sample sentence in the Body of the file /usr/share/nginx/www/default/index.html. For example: Welcome to nginx on mytesthost2!.

Task 2 Create a Server Load Balancer instance

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" : "13ebb82ceaa-cn-hangzhou-dg-a01",
"Address" : "10.10.10.77"
}
```

A Server Load Balancer with the ID 13ebb82ceaa-cn-hangzhou-dg-a01 is created. You can use the same method to create a Layer-4 instance as follows.

https://slb.aliyuncs.com/?Action=CreateLoadBalancerHttpListener&LoadBalancerId=13ebb82ceaa-cn-hangzhou-dg-a01&ListenerPort=80&BackendServerPort=80&ListenerStatus=active

Activate the Server Load Balancer instance.

Request:

https://slb.aliyuncs.com/?Action=SetLoadBalancerStatus&LoadBalancerId=13ebb82ceaa-cn-hangzhou-dg-a01&LoadBalancerStatus=active

Task 3 Add backend servers

Add a backend server through the AddBackendServers interface.

Request:

https://slb.aliyuncs.com/?Action=AddBackendServers&LoadBalancerId=13ebb82ceaa-cn-hangzhou-dg-a01&BackendServers=[{"ServerId":"id5ab1760-3498-4d95-9687-a91545ef90b3"}]

Response:

```
 {
"RequestId" : "FA2F2172-63F2-409D-927C-86BD1D536F13",
"LoadBalancerId" : "13ebb82ceaa-cn-hangzhou-dg-a01",
"BackendServers" : {
"BackendServer" : [
{
"ServerId" : "id5ab1760-3498-4d95-9687-a91545ef90b3",
"Weight" : 100
}
]
}
}
```

Add another backend server.

Request:

https://slb.aliyuncs.com/?Action=AddBackendServers&LoadBalancerId=13ebb82ceaa-cn-hangzhou-dg-a01&BackendServers=[{"ServerId":"i6b47cd72-843f-4558-b911-2776acae06fb"}]

Response:

```
 {
"RequestId" : "C61FAD0A-2E87-4D0C-80B0-95AB758FCA70",
"LoadBalancerId" : "13ebb82ceaa-cn-hangzhou-dg-a01",
"BackendServers" : {
"BackendServer" : [
{
"ServerId" : "id5ab1760-3498-4d95-9687-a91545ef90b3",
"Weight" : 100
},
{
"ServerId" : "i6b47cd72-843f-4558-b911-2776acae06fb",
"Weight" : 100
}
]
}
}
```

View the configuration details of the Server Load Balancer instance.

Request:

https://slb.aliyuncs.com/?Action=DescribeLoadBalancerAttribute&LoadBalancerId=13ebb82ceaa-cn-hangzhou-dg-a01

Response:

```
 {
"RequestId" : "4747E9AE-ADFD-412D-B523-C1CBD45A2154",
"LoadBalancerId" : "13ebb82ceaa-cn-hangzhou-dg-a01",
"Address" : "10.10.10.77",
"IsPublicAddress" : "true",
"ListenerPorts" : {
"ListenerPort" : [
80
]
},
"BackendServers" : {
"BackendServer" : [
{
"ServerId" : "id5ab1760-3498-4d95-9687-a91545ef90b3",
"Weight" : 100
},
{
"ServerId" : "i6b47cd72-843f-4558-b911-2776acae06fb",
"Weight" : 100
}
]
}
}
```

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

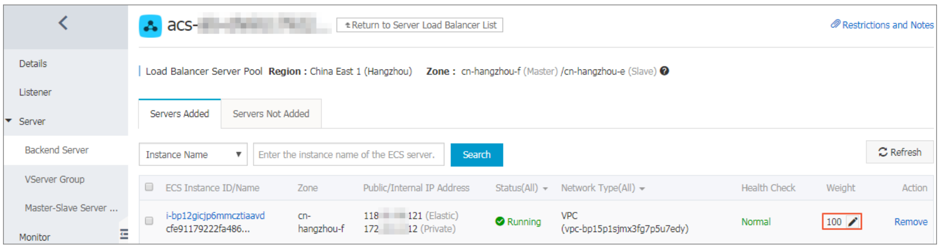
Log on Server Load Balancer console.

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 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 value to **0**.



Load Balancer Server Pool **Region :** China East 1 (Hangzhou) **Zone :** cn-hangzhou-f (Master) /cn-hangzhou-e (Slave)

□	ECS Instance ID/Name	Zone	Public/Internal IP Address	Status(All)	Network Type(All)	Health Check	Weight	Action
□	i-bp12gicjp6mmcztiaavd cfe91179222fa486...	cn-hangzhou-f	118[redacted]121 (Elastic) 172[redacted]12 (Private)	Running	VPC (vpc-bp15p1sjmx3fg7p5u7edy)	Normal	100	Remove

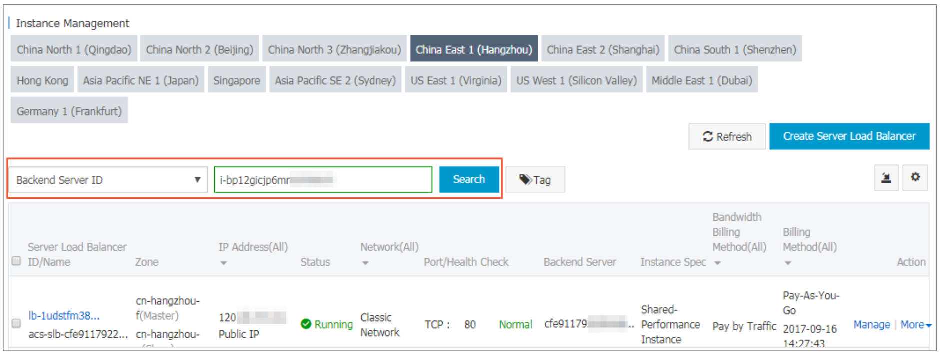
When no traffic is distributed to the ECS instance, click **Remove** to remove it from the backend server pool.

Troubleshoot

If there are ongoing business 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.

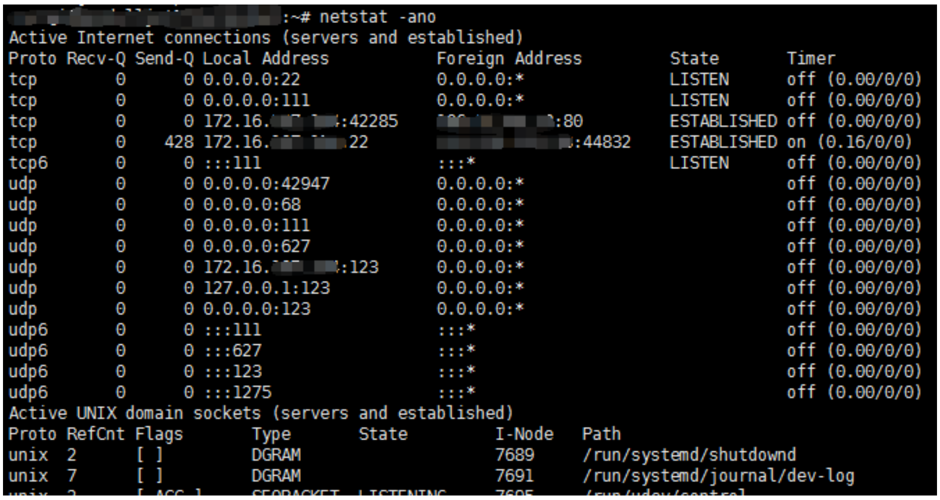
You can use the ECS instance ID to filter the Server Load Balancer that the ECS instance is added to.



Backend Server ID: i-bp12gicjp6mr[redacted]

□	Server Load Balancer ID/Name	Zone	IP Address(All)	Status	Network(All)	Port/Health Check		Backend Server	Instance Spec	Bandwidth Billing Method(All)	Billing Method(All)	Action
□	lb-1udstfm38... acs-slb-cfe9117922...	cn-hangzhou-f(Master) cn-hangzhou-	120[redacted] Public IP	Running	Classic Network	TCP : 80	Normal	cfe91179[redacted]..	Shared-Performance Instance	Pay by Traffic	Pay-As-You-Go 2017-09-16 14:27:43	Manage | More

Log on to the ECS instance, run the netstat command to check whether the ECS instance has deployed public services.



```
[redacted]:~# 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.[redacted]:42285 [redacted]:80           ESTABLISHED off (0.00/0/0)
tcp        0    428 172.16.[redacted]:22    [redacted]:44832        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.[redacted]: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/shutdownd
unix  7      [ ]         DGRAM                    7691     /run/systemd/journal/dev-log
```

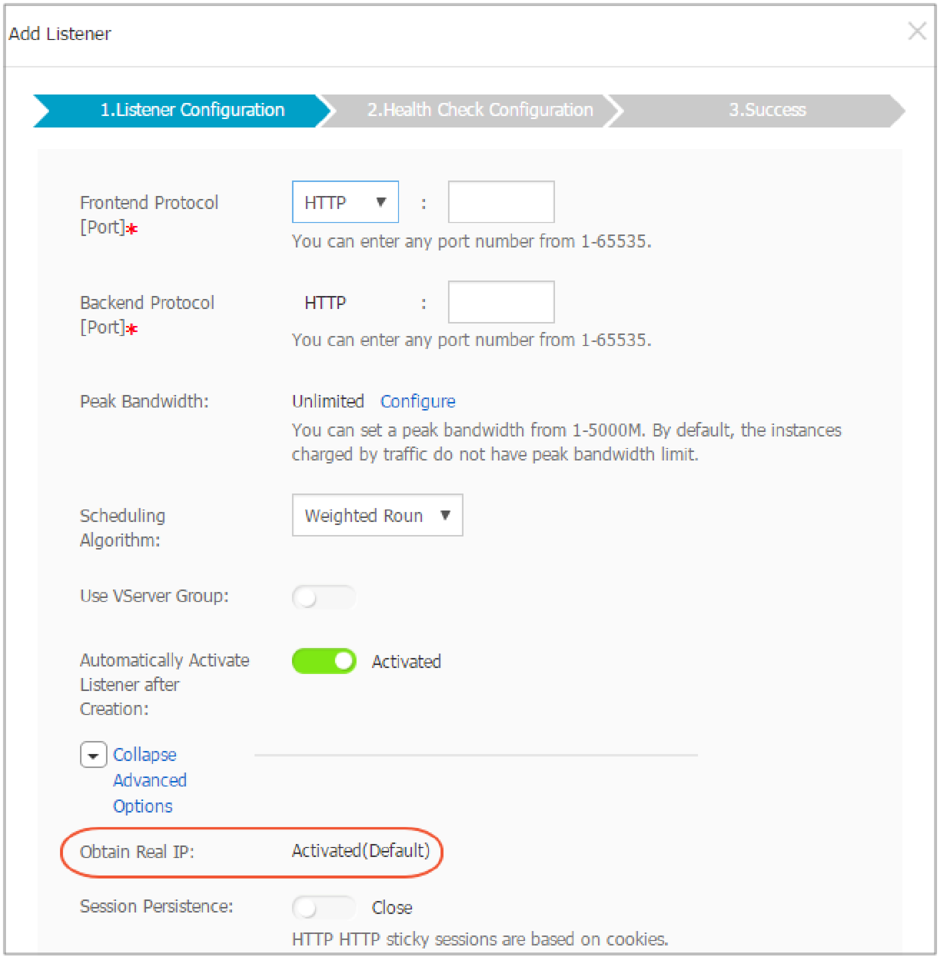
Introduction to the obtaining IP address function

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), the listener distributes the client requests to the 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 server, and then use the X-Forwarded-For header to obtain the real IP address of the client.

Note: For the HTTPS load balancing service, the SSL certificates are configured in the front-end listener, the backend still uses the HTTP protocol. Therefore, the configurations on the application server are the same for HTTP and HTTPS protocols.



Frontend Protocol [Port]*: HTTP :
You can enter any port number from 1-65535.

Backend Protocol [Port]*: HTTP :
You can enter any port number from 1-65535.

Peak Bandwidth: Unlimited Configure
You can set a peak bandwidth from 1-5000M. By default, the instances charged by traffic do not have peak bandwidth limit.

Scheduling Algorithm: Weighted Roun

Use VServer Group:

Automatically Activate Listener after Creation: Activated

Collapse Advanced Options

Obtain Real IP: Activated(Default)

Session Persistence: Close
HTTP HTTP sticky sessions are based on cookies.

Configure web applications

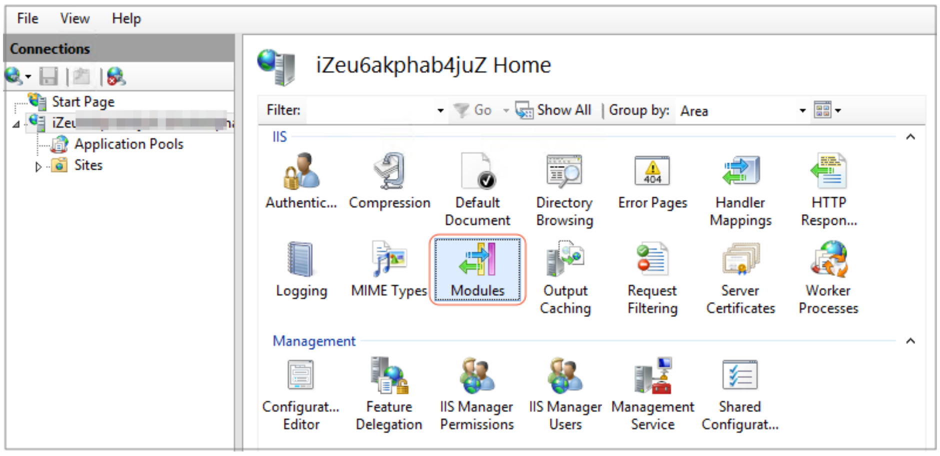
This section introduces some common methods used to configure web applications.

Configure IIS7/IIS8

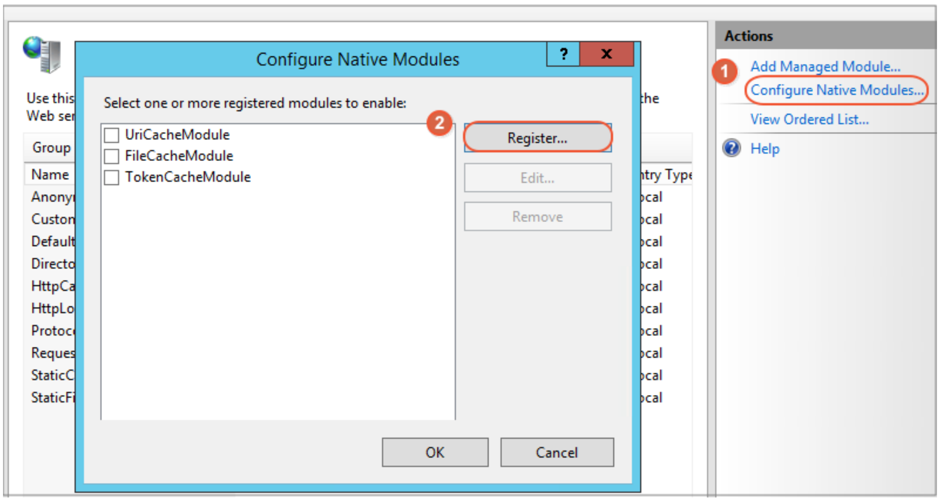
Download and extract the F5XForwardedFor.

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

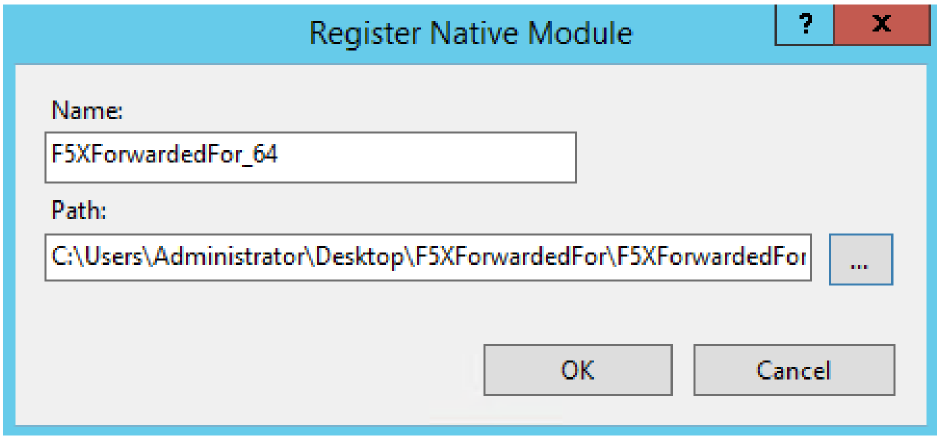
Open the IIS Manager, and then double-click the **Modules** function.



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



Add the copied the .dll file.

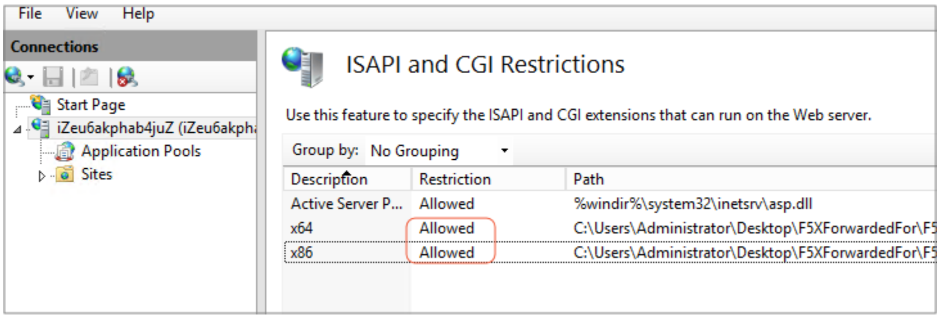


Name: F5XForwardedFor_64

Path: C:\Users\Administrator\Desktop\F5XForwardedFor\F5XForwardedFo

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

Ensure that you have installed the ISAPI and CGI applications.



ISAPI and CGI Restrictions

Use this feature to specify the ISAPI and CGI extensions that can run on the Web server.

Group by: No Grouping

Description	Restriction	Path
Active Server P...	Allowed	%windir%\system32\inetsrv\asp.dll
x64	Allowed	C:\Users\Administrator\Desktop\F5XForwardedFor\F5
x86	Allowed	C:\Users\Administrator\Desktop\F5XForwardedFor\F5

Restart the IIS Manager.

Configure Apache

Run the following command to the 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 start the Nginx server.

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