Server Load Balancer

Product Introduction

Product Introduction

Alibaba Cloud Server Load Balancer is a traffic distribution control service. It distributes the incoming application traffic among multiple ECS instances according to a scheduling algorithm and listening rules.

By setting a virtual IP address, the Server Load Balancer service virtualizes the ECS instances located in the same region into a high-performing and highly available application service pool. Client requests are distributed to the cloud server pool according to the defined listening rules. This increases the fault tolerance of your applications.

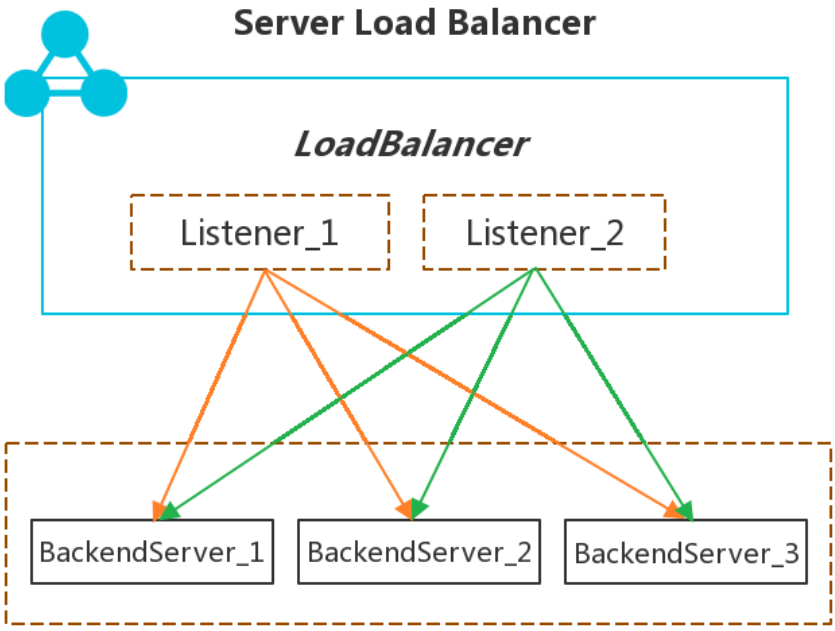
Server Load Balancer checks the health status of the ECS instances in the cloud server pool and automatically isolates any ECS instances with an abnormal status. This resolves the single point of failure (SPOF) problem and improves the overall service capability.

In addition to standard load balancing capabilities, TCP and HTTP listening can defend DDoS attacks, enhancing the protection capability of the application servers.

Components of Server Load Balancer

The Server Load Balancer service consists of three parts: a Server Load Balancer instance, one or more listeners, and multiple backend servers.

As shown in the following figure, after the Server Load Balancer instance receives a client request, the listener forwards the request to the corresponding backend ECS instances according to the configured listening rules.



Server Load Balancer instance: If you want to use the Server Load Balancer service, you must create a Server Load Balancer instance. You can add multiple listeners and backend servers to a Server Load Balancer instance.

Listener: Before using Server Load Balancer, you must add at least one listener to the Server Load Balancer instance, which defines how the client requests are forwarded to the backend servers.

Backend server: The ECS instances added to the Server Load Balancer instances are the backend servers used to process the distributed requests. You can add the ECS instances separately to the backend server pool, or add them in a batch through a VServer group or master-slave server group.

By default, the backend servers are maintained in the Server Load Balancer instance dimension. All listeners can only forward requests to the same ECS instances with the same backend port configured in the listeners. With the VServer group function, you are allowed to maintain the backend servers in the listener dimension. You can create different VServer groups for different listeners, that is, the listeners in a Server Load Balancer instance can forward requests to different backend servers with different ports.

In addition, the layer-7 load balancing service supports the configuration of domain names or URL forwarding rules. The listener can forward the requests from different domain names or URLs to different VServer groups.

Benefits

High availability

Designed to work in full-redundancy mode without SPOF. Server Load Balancer supports local and cross-region disaster tolerance when used together with DNS, delivering service availability of up to 99.95%.

Server Load Balancer can flexibly scale its service based on the application load without interrupting external services during traffic fluctuation.

Low cost

Server Load Balancer is 60% more cost-efficient than traditional hardware load-balancing systems. By providing free access to private network instances without generating any O&M cost, the service completely removes the need to purchase expensive load-balancing equipment.

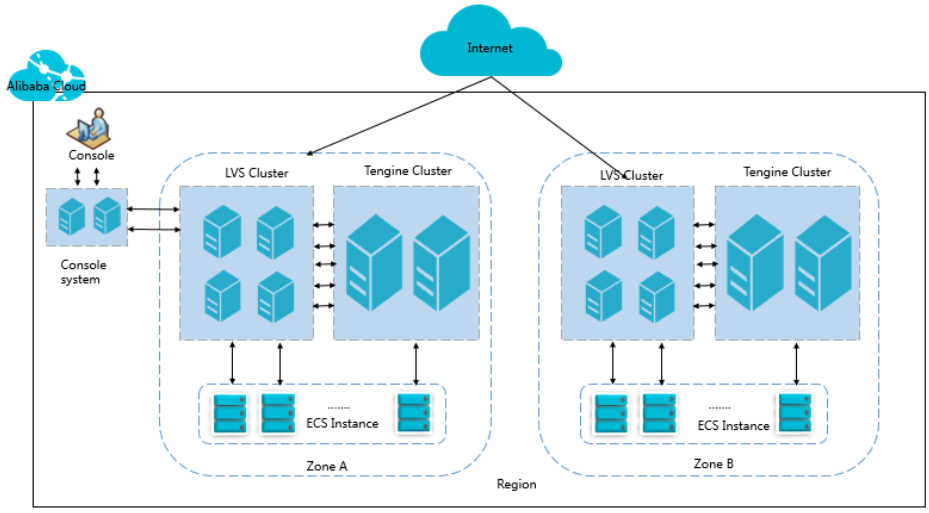
Security

Combined with Alibaba Cloud Security, Server Load Balancer can defend against DDoS attacks, such as HTTP flood and SYN flood attacks.

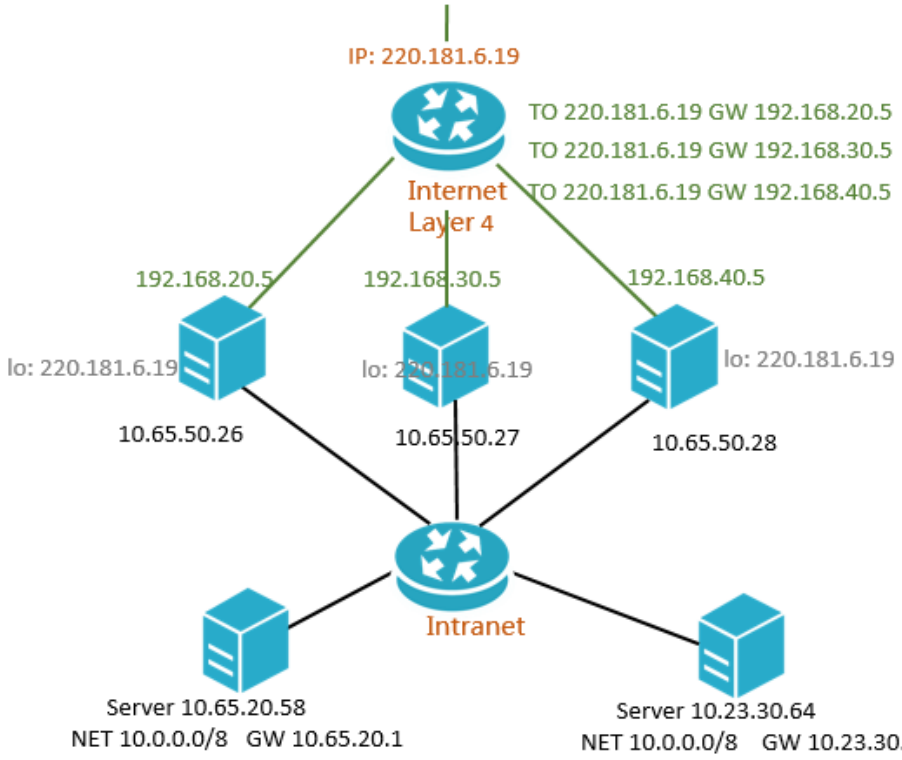
Alibaba Cloud provides the layer-4 (TCP protocol and UDP protocol) and layer-7 (HTTP protocol and HTTPS protocol) load balancing services. Deployed in clusters, Server Load Balancer can synchronize sessions to protect the ECS instances from single points of failure (SPOFs). This improves redundancy and guarantees the service stability.

Layer 4 uses the open source software Linux Virtual Server (LVS) with keepalived to achieve load balancing, and also makes some customization to it according to the cloud computing requirements.

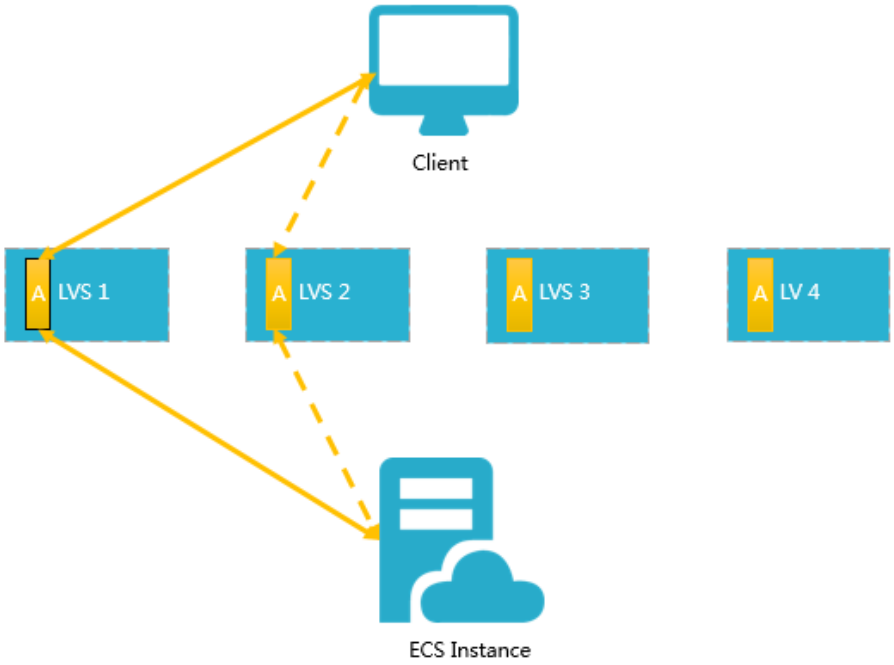
Layer 7 uses Tengine to achieve load balancing. Tengine is a Web server project based on Nginx that adds a wide range of advanced features dedicated for high-traffic websites.



As shown in the following figure, the layer-4 load balancing in each region is actually run in a cluster of multiple LVS machines. The cluster deployment model strengthens the availability, stability, and scalability of the load balancing services in abnormal circumstances.



Additionally, the LVS machine in the LVS cluster uses multicast packets to synchronize sessions to other LVS machines. As shown in the following figure, session A established on LVS1 is synchronized to other LVS machines after three packets are transferred. In normal situations, the session request is sent to LVS1 as the solid line shows. If LVS1 is abnormal or being maintained, the session request will be sent to other machines working normally, as the dotted line shows. In this way, you can perform hot upgrades, machine failure maintenance, and cluster maintenance without affecting business applications.



Note: If a connection is not established (three-way handshake is not completed), or a connection has been established but the session synchronization mechanism is not triggered, the hot upgrade does not guarantee that the connection is not interrupted and the client needs to re-initiate the connection.

Supported protocol

Alibaba Cloud provides both layer-4 (TCP and UDP) and layer-7 (HTTP and HTTPS) load balancing services.

Health check

Through health check on backend ECS instances, Server Load Balancer can automatically block abnormal ECS instances and distribute requests to them when they become normal.

Session persistence

Server Load Balancer supports session persistence. You can set listening rules to forward a session request from a client to the same backend ECS instance during the session lifecycle.

Scheduling algorithm

Server Load Balancer supports the following scheduling algorithms:

Round robin

Requests are distributed across the backend ECS servers sequentially.

Weighted round robin (WRR)

You can set a weight for each backend server. Servers with higher weights receive more requests than those with lower weights.

Weighted least connections (WLC)

In addition to weight set for each backend ECS server, the number of connections to the client is also considered. The servers with a higher weight value will receive a larger percentage of live connections at any one time. If weights are the same, the system directs network connections to the server with the least established connections.

Domain name/URL-based forwarding

For Layer-7 (HTTP and HTTPS) protocols, Server Load Balancer forwards traffic to different VServer groups based on domain names or URLs.

Multiple zones

To provide a more stable and reliable load balancing service, Server Load Balancer has deployed multiple zones in most regions. If the primary zone becomes abnormal, the backup zone automatically takes over the load balancing service from the faulty zone.

Access control

You can set a whitelist to control which IP addresses can access Server Load Balancer.

Security

Server Load Balancer supports application firewalls and HTTP flood protection. Combined with Alibaba Cloud Security, the system can defend against DDoS attacks under 5 Gbps.

Certificate management

Server Load Balancer service provides Certificate Management for the HTTPS protocol listening. With Certificate Management, you do not need to upload certificates to backend ECS instances. Deciphering is performed on Server Load Balancer to reduce the CPU overheads of backend ECS instances.

Bandwidth control

You can set the bandwidth peak for each listener based on the service that the application can provide.

Instance type

You can choose to create an Internet or Intranet Server Load Balancer service. The system will assign a public IP address or private IP address accordingly.

Monitoring

With the Monitor function, you can get the real-time statuses of your Server Load Balancer.

Management methods

You can manage Server Load Balancer instances through various methods, such as the Server Load Balancer console, Open API, and SDK.

Server Load Balancer is applicable to the following scenarios:

Scenario 1: Distribute traffic load for high-traffic applications

If your application traffic is high, you can use Server Load Balancer to distribute the traffic to multiple ECS instances. Additionally, you can use session persistence feature to forward the session requests from a client to the same backend ECS instance to improve access efficiency.

Scenario 2: Expand service capability for applications

You can extend the service capabilities by adding and removing backend ECS instances at any time, depending on your business needs. It is applicable to Web and App applications.

Scenario 3: Eliminate the single point of failure (SPOF)

With the health check feature, Server Load Balancer will automatically block unhealthy ECS instances and distribute requests to healthy ECS instances, eliminating any single point of failure.

Scenario 4: Disaster tolerance in one region

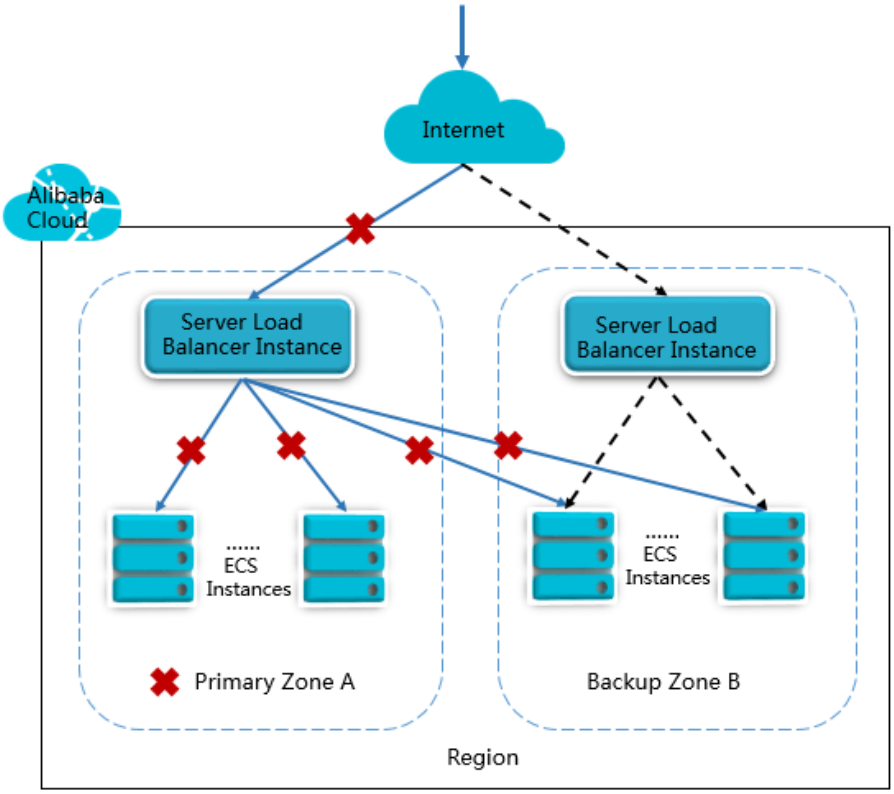
To provide more reliable services, Server Load Balancer has already deployed multiple zones

in most regions.

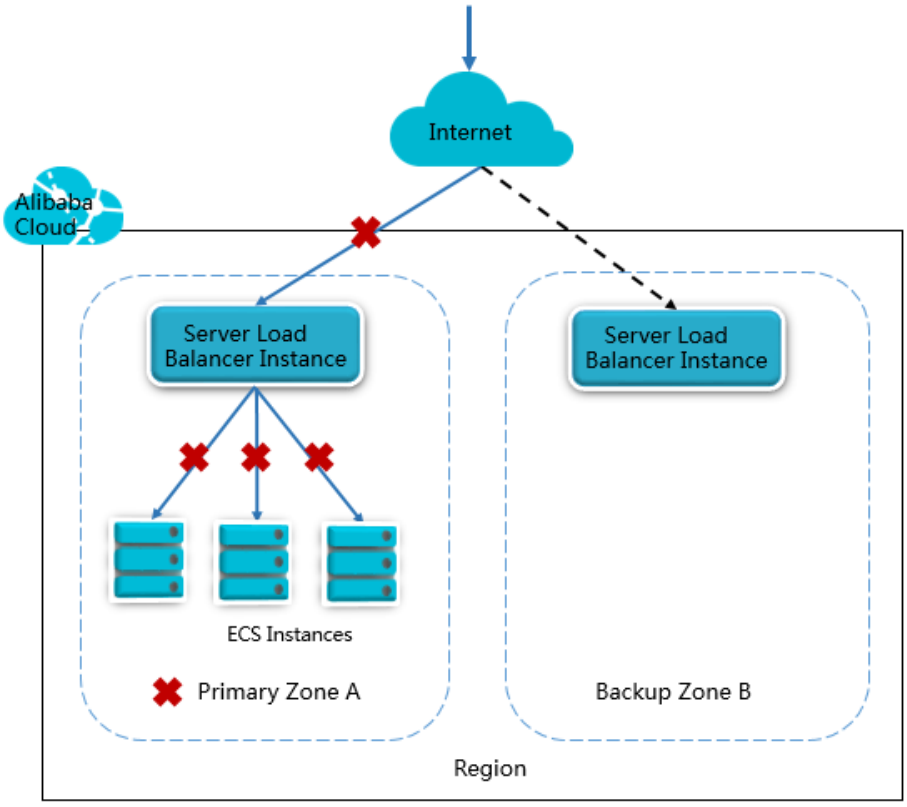
If a primary zone becomes unavailable, Server Load Balancer rapidly switches to a backup zone to restore its service capabilities within 30s. When the primary zone becomes available, Server Load Balancer will automatically switch back to the primary zone.

We recommend creating a Server Load Balancer instance in a region with multiple zones for disaster tolerance. Additionally, you can achieve a higher-availability or lower-latency load balancing by deploying ECS instances through these considerations. It is a best practice to add at last one ECS instance in each zone.

As shown in the following figure, when primary zone A works normally, traffic is distributed to ECS instances in the primary zone A, as the blue line shows. When primary zone A becomes unavailable, traffic is distributed to ECS instances in the backup zone, as the black dotted line shows. This avoids any business interruption because of failure of a single zone, and also reduces the latency between the products in different zones.

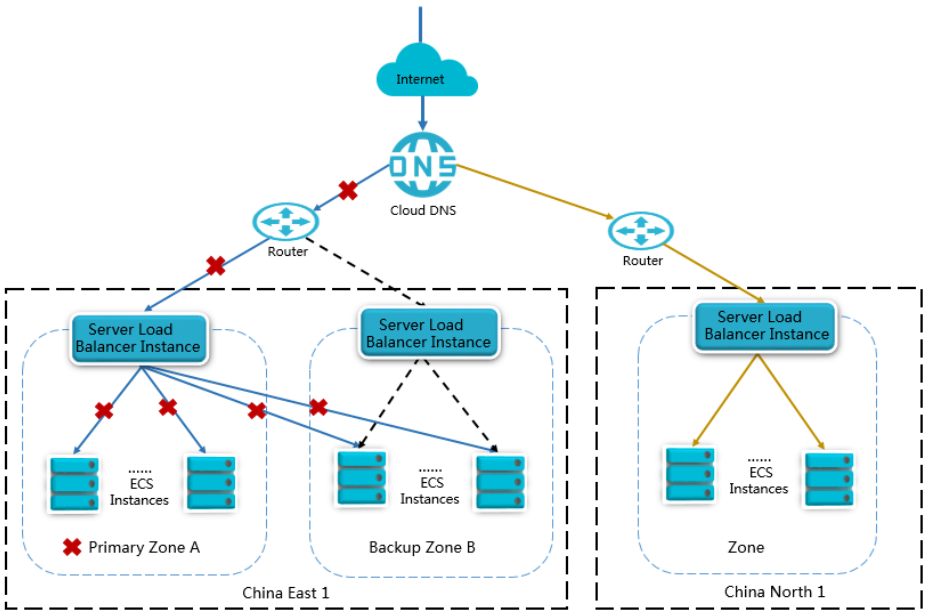


If you add all the ECS instances to the primary zone and have no ECS instances in the backup zone, as shown in the following figure, low latency is achieved at the expense of high availability. In this situation, when the primary zone is unavailable, Server Load Balancer will switch to the backup zone. However, there are no ECS instances to handle the distributed requests.



Scenario 5: Cross-region disaster tolerance

Combined with DNS, you can achieve cross-region disaster tolerance by deploying Server Load Balancer in different regions and using DNS to resolve the domain to the IP addresses of the Server Load Balancer instances. When a region becomes unavailable, you need to stop domain name resolution without affecting user access.



Item	Limits	Exception
ECS instances added to a Server Load Balancer instance	Must have at least one ECS instance, and the region must be the same for the ECS instance and Server Load Balancer instance.	No
Types of a Server Load Balancer instance	Internet/intranet	No
Billing methods	By traffic or bandwidth	No
Public bandwidth range for a Server Load Balancer instance (Pay by bandwidth)	1 to 5000 MB，default is 1 MB	No
Default quota of Pay-As-You-Go instances	60	Submit a ticket to apply for more
Restrictions on Server Load Balancer instance name	The name can be 1 -80 characters in length, and include letters, digits, hyphens (-), backslashes (/), periods (.) and underscores (_).	No
The number of listeners added in a Server Load Balancer instance	50	No
The number of the domain/URL forwarding rules added to a layer-7 listener	20	Yes
Supported protocols	HTTP/HTTPS/TCP/UDP	No
Front-end/backend port range used in a listener	1 to 65535	No
Front-end/backend port range used in a listener (Financial Cloud)	80，443，2800 to 3300，5000 to 10000，13000 to 14000	No
Supported scheduling algorithms	Round robin, weighted round robin (WRR), weighted least connections (WLC)	No
Cookie processing in HTTP session persistence	insert/server，default is insert	No
Cookie timeout for HTTP protocol	1 to 86400, default is 3600	No
Cookie name restrictions	The name can be 1 to 200 characters in length, and must comply with RFC2965.	No

	It can only contain ASCII English letters and digits. It cannot contain commas, semicolons, spaces, or begin with a dollar symbol ($).	
Health check port for HTTP protocol	1 to 65,535 (default is the backend server port)	No
Health check domain for HTTP protocol	The length can be 1 to 80 characters, and include letters, digits, hyphens (-), and periods (.).	No
Health check URI for HTTP protocol	The length can be 1 to 80 characters and include letters, digits, '-', '/', '.', '%', '?', '#', '&'.	No
Health check timeout for HTTP protocol	1 to 300s, default is 5	No
Health check interval for HTTP protocol	1 to 50s, default is 2	No
Healthy threshold for HTTP protocol health check	2 to 10, default is 3	No
Unhealthy threshold for HTTP protocol health check	2 to 10, default is 3	No
Session timeout for TCP protocol	1 to 3600s	No
Health check port for TCP protocol	1 to 65,535 (default is the backend server port)	No
Health check timeout for TCP protocol	1 to 300s, default is 5	No
Health check interval for HTTP protocol	1 to 50s, default is 2	No
Healthy threshold for TCP protocol health check	2 to 10, default is 3	No
Unhealthy threshold for TCP protocol health check	2 to 10, default is 3	No
Health check port for UDP protocol	1 to 65,535 (default is the backend server port)	No
Health check timeout for UDP protocol	1 to 300s, default is 10	No
Health check interval for UDP protocol	1 to 50s, default is 10	No
Healthy threshold for TCP protocol health check	2 to 10, default is 3	No

Unhealthy threshold for UDP protocol health check	2 to 10, default is 3	No
Number of backend ECS instances that can be added or deleted in batch	20	No
API access frequency limit for a single key	5,000 times/day	Currently, an automatic process is not available. Contact customer manager for help.
Maximum number of certificates uploaded by a single user	100	Currently, an automatic process is not available. Contact customer manager for help.

Term	Description
Server Load Balancer	Alibaba Cloud Server Load Balancer is a traffic distribution control service. It distributes incoming application traffic among multiple ECS instances according to the configured scheduling algorithm and listening rules.
Server Load Balancer Instance	A Server Load Balancer instance is a running instance of the Server Load Balancer service. To use Server Load Balancer, you must first create a Server Load Balancer instance. The instance ID is a unique identifier for the Server Load Balancer instance.
Server Load Balancer IP	The IP address allocated to the Server Load Balancer service after creating a Server Load Balancer instance. According to the instance type, the IP address is either a public IP or a private IP. You can resolve a domain name to the public IP address to provide external services.
Listener	A listener defines how the incoming requests are distributed. You must add at least one listener to a Server Load Balancer instance.
Backend Server	The ECS instances that process the distributed requests.
VServer Group	A group of ECS instances that process the distributed requests. Different listeners can use different VServer groups, which allows you to maintain the request distribution in the listener dimension.
Multiple Zones	Server Load Balancer has already deployed multiple zones in most regions for better disaster tolerance.

	By default, an instance in a primary zone is used to provide the load balancing service. However, when the primary zone is unavailable, the instance will automatically switch to the backup zone to continue providing service and then switch back to the primary zone when it becomes available. This increases local availability.