Server Load Balancer

Product Introduction

Product Introduction

Product Overview

Overview

Server Load Balancer is a service that distributes traffic among multiple ECS instances. It can extend the service capability of an application system by distributing traffic and enhancing availability by eliminating any Single Point Of Failure (SPOF).

Major capabilities are highlighted below:

- Server Load Balancer sets a virtual IP address to virtualize multiple Elastic Compute Service (ECS) instances in the same region into an application pool featuring high performance and high availability. Based on the method specified by the application, Server Load Balancer distributes network requests from clients to the ECS pool.
- Server Load Balancer checks the health status of ECS instances in the ECS pool and automatically isolates abnormal ECS instances, so as to solve the SPOF possibility of a single ECS instance and improves the overall service capabilities of the application.

In addition to the standard load balancing function, Server Load Balancer can also defend against DDoS attacks utilizing the TCP and HTTP methods, and thereby improves the overall defense capability of the application server.

- Server Load Balancer is a service supporting multiple ECS instances, and is to be used together with ECS.

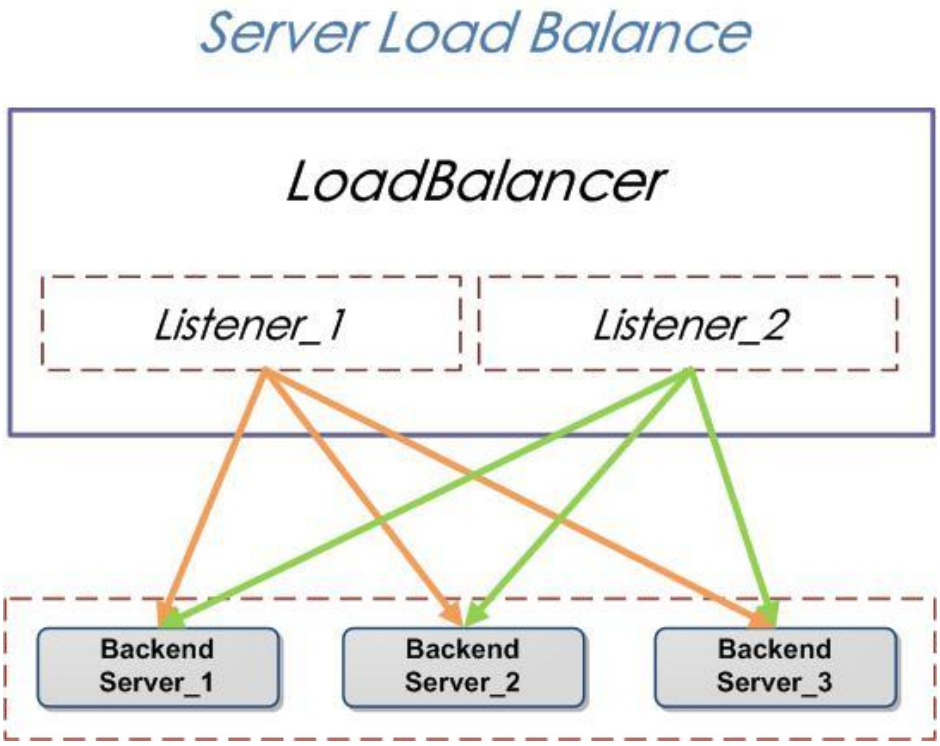
Core Concepts

Server Load Balancer has the following three core concepts:

- LoadBalancer represents a Server Load Balancer instance.
- Listener represents custom load balancing policies and forwarding rules.
- BackendServer represents a group of backend ECS instances.

External access requests are forwarded to the backend ECS instances for processing through the Server Load Balancer instance in accordance with relevant policies and forwarding rules.

The core concepts of Server Load Balancer are shown in the figure below



Restrictions

- Server Load Balancer does not support cross-region deployment. Only backend ECS instances of the same region can be added to a Server Load Balancer instance.
- In Layer-4 (TCP) Server Load Balancer, the ECS instances in the backend ECS pool cannot serve both as real servers and as clients which can send requests to the corresponding Server Load Balancer instances. The reason is that 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.
- Server Load Balancer does not restrict the IP addresses through which ECS instances are pinged to the Server Load Balancer instance. However, some newly purchased ECS instances in North China 1 (Qingdao) node, North China 2 (Beijing) node and East China 1 (Hangzhou) node cannot be pinged to the Server Load Balancer instance through private IP addresses. The restriction does not affect the communication between the Server Load Balancer instance and backend ECS instances.
- For security compliance considerations of AntCloud users, Server Load Balancer instances on the public network only enable the following ports for external accesses: 80, 443, 2800-3300, 5000-10000 and 3000-14000.

Usage Notes

- Before using Server Load Balancer to offer external services, it is required to install and

correctly configure the application services on all backend ECS instances, and ensure that the ECS service IP address can access Server Load Balancer.
- Server Load Balancer does not support data synchronization among backend ECS instances. If application services deployed on the backend ECS instances are stateless, independent ECS or RDS can be used to store data. If the application services deployed on the backend ECS instances are stateful, data on these ECSs must be synchronized.
- When the IP address of the Server Load Balancer instance is resolved to a normal domain name and external services are offered, do not delete the Server Load Balancer instance. When the Server Load Balancer instance is deleted, its IP address is released and the external services are interrupted. When a new Server Load Balancer instance is created, the system will assign an IP address for the instance.

Function Description

Protocols Supported

Currently, Server Load Balancer is available for both Layer-4 (TCP and UDP) and Layer-7 (HTTP and HTTPS) protocols.

Health Check

Through health check on backend ECS instances, Server Load Balancer can automatically block abnormal ECS instances and restore them when they become functional again。

Session Persistence

Server Load Balancer provides session persistence feature that forwards requests of a client to the same backend ECS instance during the session lifecycle.

Scheduling Algorithm

Server Load Balancer supports weighted round robin (WRR) and weighted least connections (WLC). WRR distributes external requests sequentially to backend ECS instances. WLC distributes external requests to the backend ECS instance that currently has the minimum number of connections. A backend ECS instance with greater weight is more likely to receive the requests.

Domain Name/URL-based Forwarding

For Layer-7 (HTTP and HTTPS) protocols, Server Load Balancer forwards traffic to different virtual

server groups based on users' domain names or URLs.

Multi-zone

Supports the creation of a Server Load Balancer instance for a specified zone. It is possible to deploy master and slave zones in a multi-zone region. When the master zone is at fault, the slave zone automatically takes over services from the faulty zone.

Resource Access Management (RAM)

Supports the white list feature. A white list is an access control method used to determine which IP addresses can access Server Load Balancer Monitoring. It applies when a user application only permits access by certain IP addresses.

Security

Supports application firewalls and CC protection. The cluster has a WAF module, which allows you to enable WAF protection without modifying CNAME. When used in combination with Yundun, the system can defend against DDoS attacks no more than 5 Gbps.

Certificate Management

This is a unified certificate management service over the HTTPS protocol, eliminating the need to upload certificates to backend ECS instances. Deciphering is performed on Server Load Balancer, so as to reduce the CPU overheads of backend ECS instances.

Bandwidth Control

A bandwidth peak can be configured for each service based on the listener results.

Type of Instances/Networks Supported

- Server Load Balancer is supported on public or private networks.
- Server Load Balancer is also supported on classic networks or VPCs.

Monitoring

Provides a rich set of monitoring data to keep you informed about the real-time running status of Server Load Balancer.

Management Methods

Provides multiple management methods, including the Server Load Balancer Console, APIs, and SDKs.

Product Terminology

Glossary

Term	Full Name	Description
Server Load Balancer	Server Load Balancer	A network load balancing service provided by Alibaba Cloud. In combination with ECS, it provides TCP and HTTP load balancing services based on ECS instances.
LoadBalancer	Load Balancer	A Server Load Balancer instance can be understood as a running instance of Server Load Balancer. To use Server Load Balancer, the user must first create a Server Load Balancer instance. The LoadBalancerId is the unique identifier of the Server Load Balancer instance.
Listener	Listener	A listener includes a monitoring port, Server Load Balancer policy and health check configuration. Each listener corresponds to a backend application service.
BackendServer	Backend Server	A set of ECS instances that accepts the requests distributed by Server Load Balancer, and forwards external access requests to the backend ECS instances for processing according to user-defined rules.
Address	Address	Service address assigned by the system. It is currently the IP address. THe user can determine whether the service address is publicly available or not, so as to create Server Load Balancer

		instances for public and private networks respectively.
Certificate	Certificate	Certificates are used in HTTPS. After uploading a certificate to Server Load Balancer, the user bind the certificate during HTTPS listener creation to provide HTTPS service.
Master Availability Zone	Master Availability Zone	Server Load Balancer can be deployed in multiple zones in a region. The user specify primary and standby zones for a Server Load Balancer instance, which runs in the primary zone by default.
Slave Availability Zone	Slave Availability Zone	Server Load Balancer can be deployed in multiple zones in a region. The user can specify primary and standby zones for the Server Load Balancer Instance. The Server Load Balancer instance runs in the primary zone by default.

Configuration Instructions

Configuring and managing a Server Load Balancer instance primarily involves three operations:

1.) Configuring Server Load Balancer instance attributes.

2.) Configuring Server Load Balancer Monitoring.

3.) Configuring backend ECS instances for the Server Load Balancer instance.

The user can configure instance attributes to define the type of Server Load Balancer instance, configure service listeners to define policies and forwarding rules for the Server Load Balancer instance, and configure the Server Load Balancer instance with backend ECS instances used to process user requests.

Configuring Server Load Balancer Instance Attributes

Server Load Balancer Instance Name

The user can specify an easily recognizable name for the Server Load Balancer instance created. If the

name of the instance is not specified, the LoadBalancerID of the Server Load Balancer instance is used as the instance name. LoadBalancerID is the unique identifier of the Server Load Balancer instance and cannot be modified.

Types of Server Load Balancer

There are two variants of Server Load Balancer available:

1) For the public network.

2) For the private network.

The user may choose public or private Server Load Balancer service based on business scenarios. The system will assign a public IP address or private IP address according to choice.

IP Addresses of Server Load Balancer

Different service IP addresses are distributed based on the selected type of Server Load Balancer. For applications that offer external services via a domain name, you must resolve the domain name to the public IP address of the corresponding Server Load Balancer instance. You can access the service via the domain name only after successful domain name resolution.

Configuring Server Load Balancer Monitoring

Server Load Balancer Protocols/Ports

- Protocols: Currently, Server Load Balancer is available for both Layer-4 (TCP and UDP) and Layer-7 (HTTP and HTTPS) protocols.
- Ports: In a Server Load Balancer instance offering external or internal services, the frontend ports which receive requests and forward the requests to backend servers shall not be repeated within the same Server Load Balancer instance.

Backend Protocols/Ports

- Protocols: Currently, Server Load Balancer is available for both Layer-4 (TCP and UDP) and Layer-7 (HTTP and HTTPS) protocols.
- Ports: The group of backend ports added after the Server Load Balancer instance for processing external or internal requests and receiving requests from ECS can be repeated within the same Server Load Balancer instance. When multiple application services are deployed in the same group of backend ECS instances, a Server Load Balancer instance supports up to 50 monitoring services at the moment.

Forwarding Rule

Currently, Server Load Balancer supports two types of forwarding: round-robin and 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 least connections.

Obtaining Visitors’ Real IP Addresses

- Layer-7 (HTTP) Server Load Balancer replaces the IP address of the HTTP header file to forward requests. Therefore, the access IP address displayed on the backend ECS instances is the local IP address of the Server Load Balancer system instead of the real IP address of the visitor. Therefore, the system allows the use of X-Forwarded-For to obtain the visitor’s real IP address. The “Obtain the Visitor’s Real IP Address” function is enabled for Layer-7 (HTTP) Server Load Balancer Monitoring by default, and cannot be disabled. For how to configure common application servers, click here.
- For Layer-4 (TCP) Server Load Balancer, the backend ECS instances directly obtain the visitors real IP address.

Session Persistence

After session persistence is enabled, Server Load Balancer will distribute the access requests from a client to the same backend ECS instance for processing.For Layer-7 (HTTP and HTTPS) Server Load Balancer, the system supports cookie-based session persistence. The Server Load Balancer system provides two cookie-based processing methods:

- Cookie seeding means the cookie on the client side is directly allocated and managed by the Server Load Balancer system. The user needs to specify the timeout time of session persistence during configuration.

NOTE: When the “Cookie seeding” Mode is configured for session persistence and the HTTP status code returned from backend RS is 4xx, Server Load Balancer does not seed the Set-Cookie header, this may result in session persistence failure. Server Load Balancer seeds the header necessary for session persistence when the HTTP status code returned from the backend RS is 200, 201, 204, 206 301, 302, 303, 304, or 307.

- Cookie rewriting means the Server Load Balancer system will allocate and manage the cookie seeding operation on the client side according to the custom cookie name. This allows you to identify and distinguish the custom cookie name, and therefore set session persistence rules for different applications on backend servers. You need to specify a name for the corresponding cookie during configuration. For how to configure different persistence rules under multiple domain names, click here.

For Layer-4 (TCP and UDP) Server Load Balancer, the system supports IP address-based session persistence. Server Load Balancer forwards access requests from an IP address to the same backend ECS instance for processing.

Health Check

- After health check is enabled, if a backend ECS instance is abnormal, Server Load Balancer forwards requests to other normal ECS instances, and automatically restores the ECS instance when it becomes normal again and can provide external or internal services.
- For Layer-7 (HTTP and HTTPS) Server Load Balancer, the system performs health check as follows: By default, the Server Load Balancer system uses the intranet IP address of backend ECS to initiate a HTTP head request to the default page configured by the application server (the default page is accessed through the backend ECS port specified in the server listener configuration). When a 200 OK message is returned, the backend ECS runs normally; otherwise, the backend ECS runs abnormally. If the page for heath check is not the default page of the application server, you need to specify a corresponding URI. If you define parameters of the host field for the HTTP head request, you need to specify a corresponding URL. You can set the health check frequency, healthy threshold value and unhealthy threshold value to better control the health check function.
- For Layer-4 (TCP and UDP) Server Load Balancer, the system performs health check as follows: By default, the Server Load Balancer system uses the backend ECS port specified in the service listener configuration to initiate an access request. If the port can access the system, the backend ECS instances run normally; otherwise, they run abnormally.
- For how to troubleshoot health check exceptions, click here.

The following TCP/HTTP/HTTPS health check parameter settings are recommended:

Response timeout: 5 secondsHealth check interval: 2 secondsUnhealthy threshold value: 3Healthy threshold value: 3

This configuration facilitates the rapid convergence of user services and application statuses:ECS health check failure response time (in case of a network exception): (2 + 5) × 3 = 21 secondsFor faster response, lower the response timeout value. However, ensure that your service is normally processed within a timeframe less than this value.ECS health check success response time: 2×3=6 seconds

The following UDP health check parameter settings are recommended:

Response timeout time: 10 secondsHealth check interval: 5 secondsUnhealthy threshold value: 6Healthy threshold value: 6

This configuration facilitates the rapid convergence of user services and application statuses:ECS health check failure response time (in case of a network exception): (5 + 10) × 6 = 90 secondsFor faster response, lower the response timeout value. However, ensure that your service is normally processed within a timeframe less than this value.ECS health check success response time: 5 × 6 = 30 seconds

Bandwidth Peak

You can set different bandwidth peaks for different listeners to restrict the capacity of different applications on the backend ECS instances 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 higher bandwidth needs.

Backend ECS Configuration for Server Load Balancer

In principle, 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 and UDP) 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.

Backend ECS Weight

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

Product Strengths

High availability

Designed to work in full redundancy mode without single point of failure (SPOF), Server Load Balancer supports local and cross-region disaster tolerance when used together with DNS, and delivers availability of upto 99.99%.

Server Load Balancer achieves elastic scaling based on the applied load and does not interrupt external services during traffic fluctuation.

Low cost

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

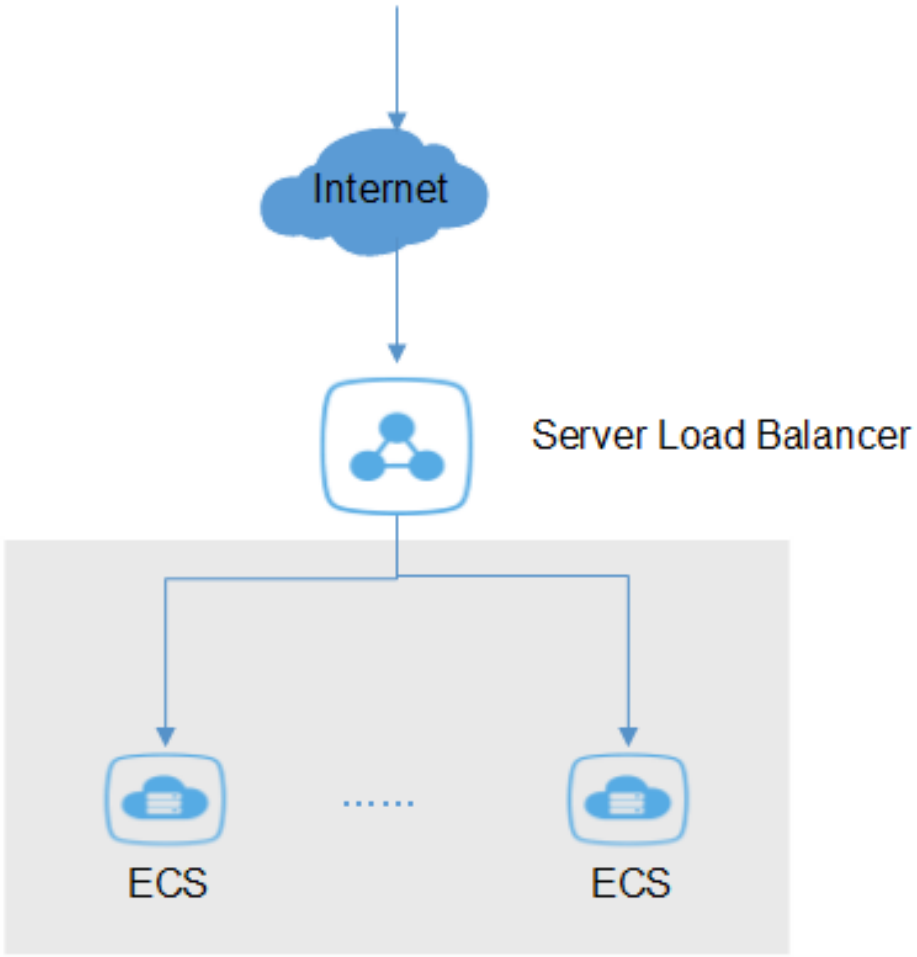
Security

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

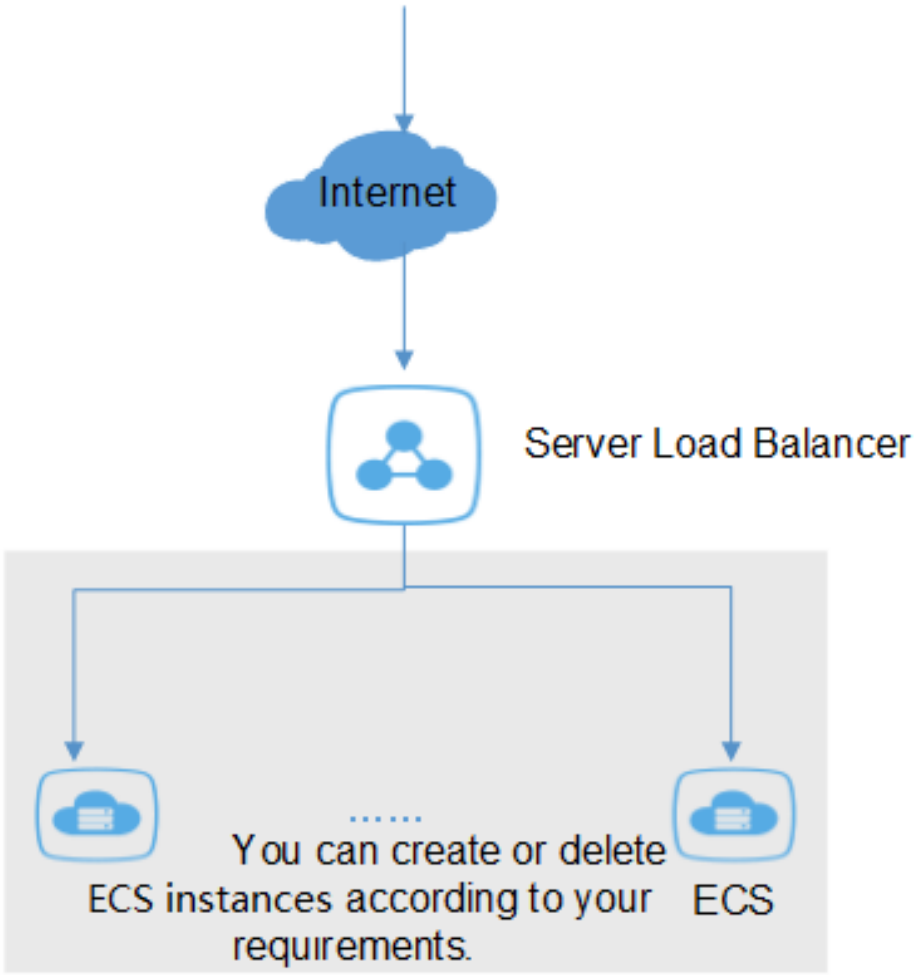
Application Scenarios

Server Load Balancer is mainly applied in the following scenarios:

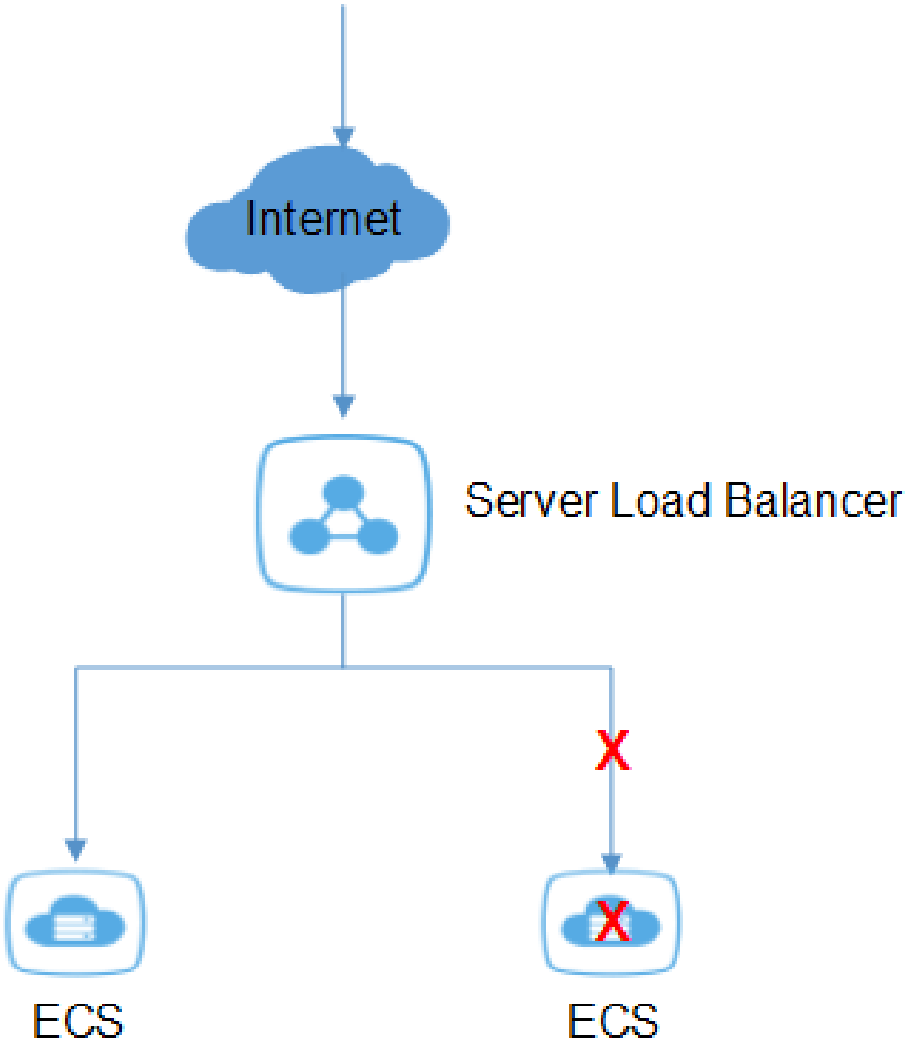
- Flexible traffic distribution, suitable for businesses with high access volumes.



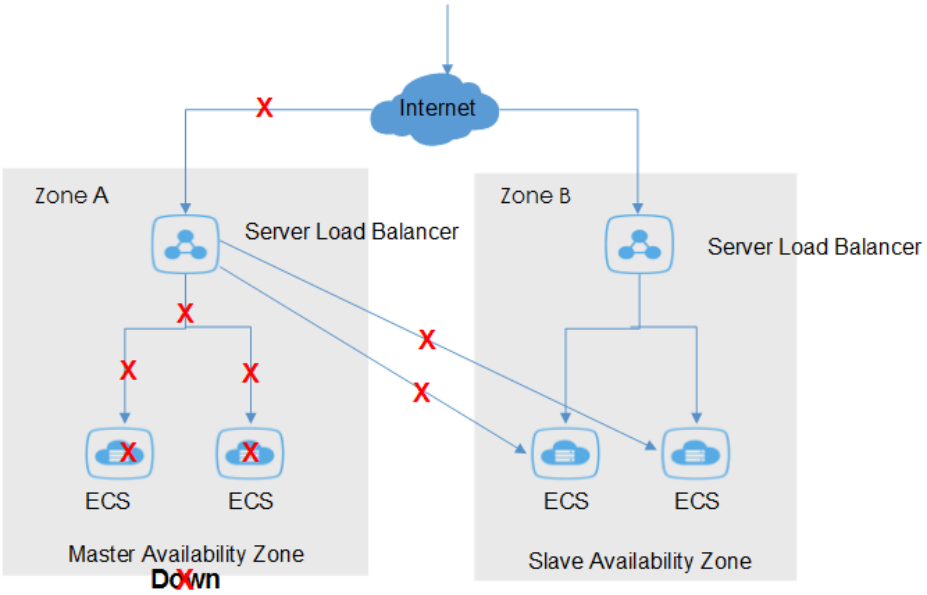
- Service capabilities of horizontally scaling application systems, suitable for a variety of Web servers and app servers.



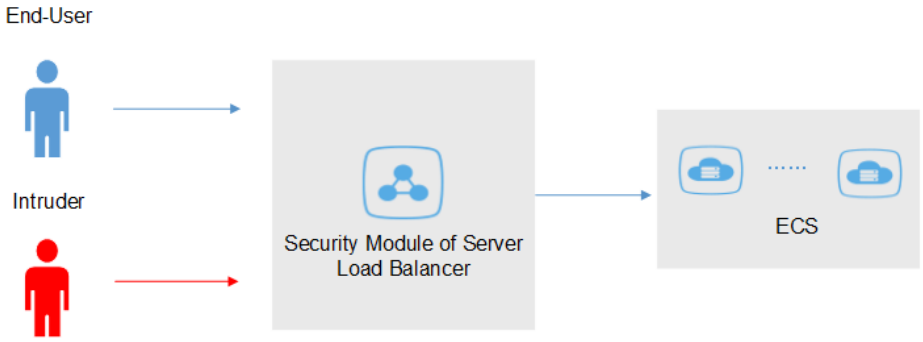
- Eliminating application system SPOF so that, even if some ECS instances go down, the application system will continue to function normally.



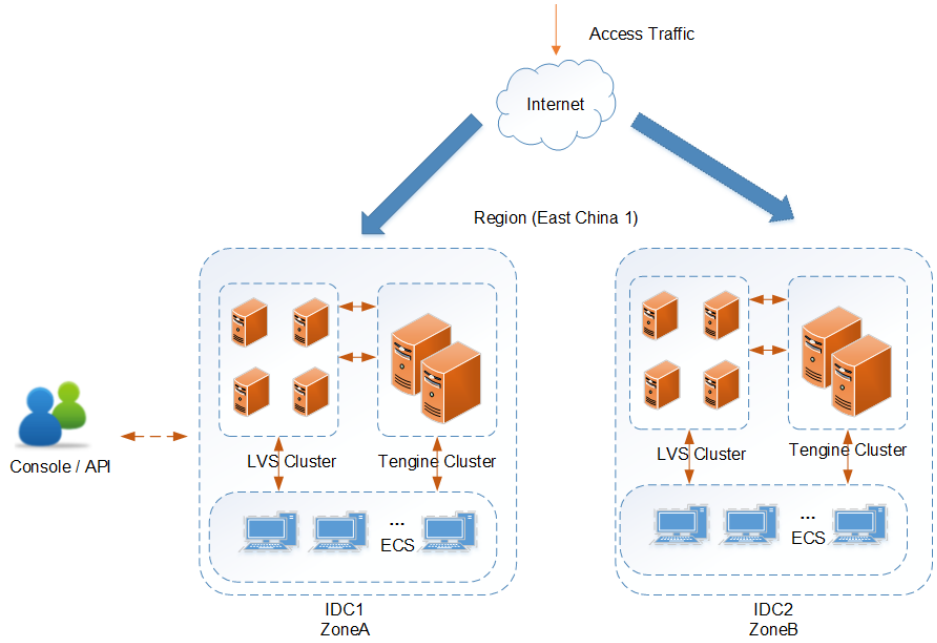
- Improving application systems' disaster tolerance capabilities through multi-zone deployment so that, even if a machine room goes down, the system will continue to function normally.



- Preventing attacks on the application system, suitable for businesses that often suffer WAF and CC troubles.



Architecture



The architecture of Server Load Balancer is shown in the figure above. Details are described below:

- Currently, Server Load Balancer is available for both Layer-4 and Layer-7 protocols.
- Layer-4 Server Load Balancer is based on open source software LVS + keeplived.
- Layer-7 Server Load Balancer is powered by Tengine, a Web server project launched by Taobao. Based on Nginx, Tengine includes a wide range of additional advanced functions and features geared to the needs of high-traffic websites.
- Deployed in clusters, Server Load Balancer can synchronize sessions to protect servers from single point of failures (SPOFs), so as to improve redundancy and ensure service stability.
- Two machine rooms are deployed in certain regions for local disaster tolerance.