

# Alibaba Cloud Server Load Balancer

## Product Introduction

Issue: 20190614

## Legal disclaimer

---

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.
2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company, or individual in any form or by any means without the prior written consent of Alibaba Cloud.
3. The content of this document may be changed due to product version upgrades, adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.
4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults" and "as available" basis. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, special, or punitive damages, including lost profits arising from the use

or trust in this document, even if Alibaba Cloud has been notified of the possibility of such a loss.

5. By law, all the content of the Alibaba Cloud website, including but not limited to works, products, images, archives, information, materials, website architecture, website graphic layout, and webpage design, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of the Alibaba Cloud website, product programs, or content shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates).
6. Please contact Alibaba Cloud directly if you discover any errors in this document.



# Generic conventions

Table -1: Style conventions

Style	Description	Example
	This warning information indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	 <b>Danger:</b> Resetting will result in the loss of user configuration data.
	This warning information indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	 <b>Warning:</b> Restarting will cause business interruption. About 10 minutes are required to restore business.
	This indicates warning information, supplementary instructions, and other content that the user must understand.	 <b>Notice:</b> Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructions, best practices, tips, and other content that is good to know for the user.	 <b>Note:</b> You can use Ctrl + A to select all files.
>	Multi-level menu cascade.	Settings > Network > Set network type
<b>Bold</b>	It is used for buttons, menus, page names, and other UI elements.	Click OK.
Courier font	It is used for commands.	Run the <code>cd / d C :/ windows</code> command to enter the Windows system folder.
<i>Italics</i>	It is used for parameters and variables.	<code>bae log list --instanceid Instance_ID</code>
[ ] or [a b]	It indicates that it is an optional value, and only one item can be selected.	<code>ipconfig [-all -t]</code>

Style	Description	Example
<b><code>{}</code> or <code>{a b}</code></b>	It indicates that it is a required value, and only one item can be selected.	<code>swich {stand   slave}</code>



# Contents

---

Legal disclaimer.....	I
Generic conventions.....	I
1 What is Server Load Balancer?.....	1
2 Architecture.....	3
3 Features.....	9
4 High availability of SLB.....	13
5 Scenarios.....	16
6 Terms.....	19



# 1 What is Server Load Balancer?

---

Server Load Balancer (SLB) is a traffic distribution and control service that distributes inbound traffic among multiple ECS instances according to configured forwarding rules. SLB expands service capabilities of applications and enhances their availability.

## Overview

By setting a virtual service address, SLB virtualizes added ECS instances into an application service pool that has high performance and high availability, and distributes client requests to ECS instances in the server pool based on forwarding rules.

SLB also checks the health status of added backend servers, and automatically isolates abnormal ECS instances to eliminate Single Point of Failures (SPOFs), improving the overall service capability of your application. Additionally, working with Alibaba Anti-DDoS, SLB can defend DDoS attacks.

## Components

SLB consists of the following components:

- SLB instances

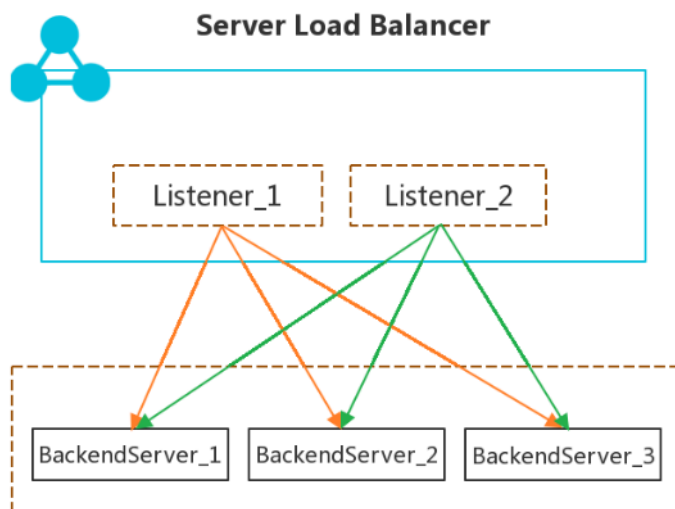
An SLB instance is a running load balancing service that distributes incoming traffic to backend servers. To use the SLB service, you must create an SLB instance, and then configure the instance with at least one listener and two backend servers.

- Listeners

A listener checks client requests and forwards the requests to backend servers according to the configured rules. It also performs health checks on backend servers.

- Backend servers

Backend servers are the ECS instances added to an SLB instance to process the distributed requests. You can add ECS instances to the default server group, a VServer group, or an active/standby server group for better management.



## Benefits

- High availability

SLB is designed to work in full-redundancy mode and avoids SPOFs. It supports local and cross-region disaster tolerance. When SLB is used together with DNS, the service availability is up to 99.95%.

You can scale your service based on application loads, without interrupting service continuity.

- Scalable

You can increase or decrease the number of backend servers to adapt to the service needs of your applications.

- Cost effectiveness

Compared with traditional load balancing hardware, SLB reduces the cost by 60%.

- Secure

Combined with Alibaba Cloud Security, SLB can defend against up to 5 Gbit/s DDoS attacks, such as HTTP flood and SYN flood attacks.

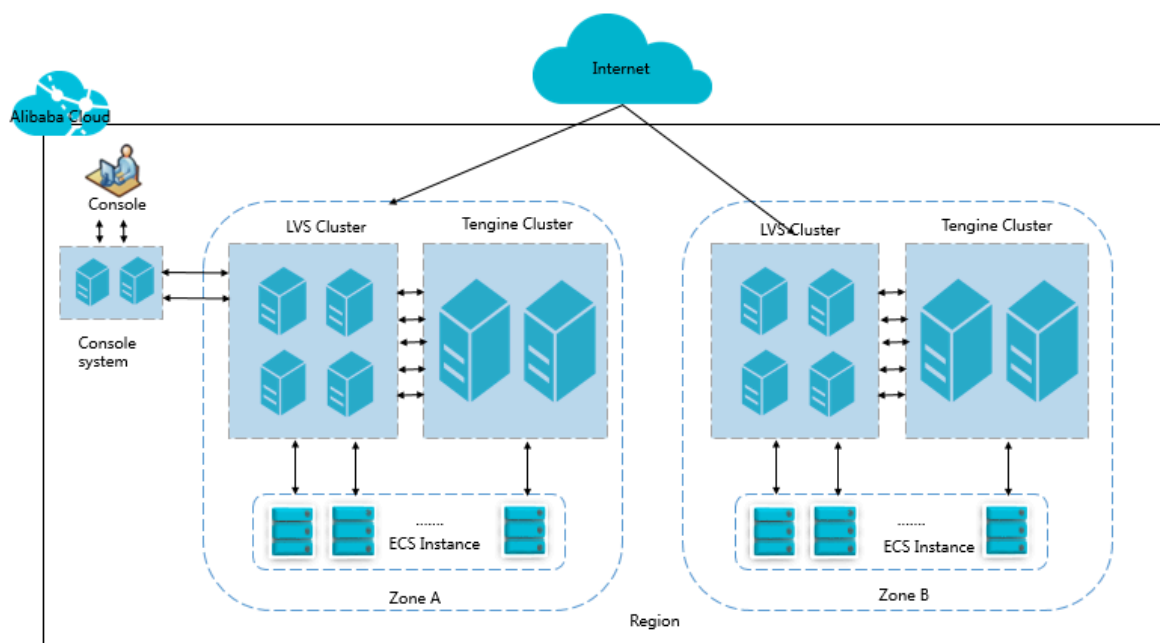
## 2 Architecture

This topic describes the architecture of Server Load Balancer (SLB). SLB is deployed in clusters, which enables session synchronization and mitigates the creation of SPOFs in backend servers. As a traffic forwarding service, requests from clients are forwarded to backend servers through SLB clusters, and backend servers return the responses to SLB through the intranet.

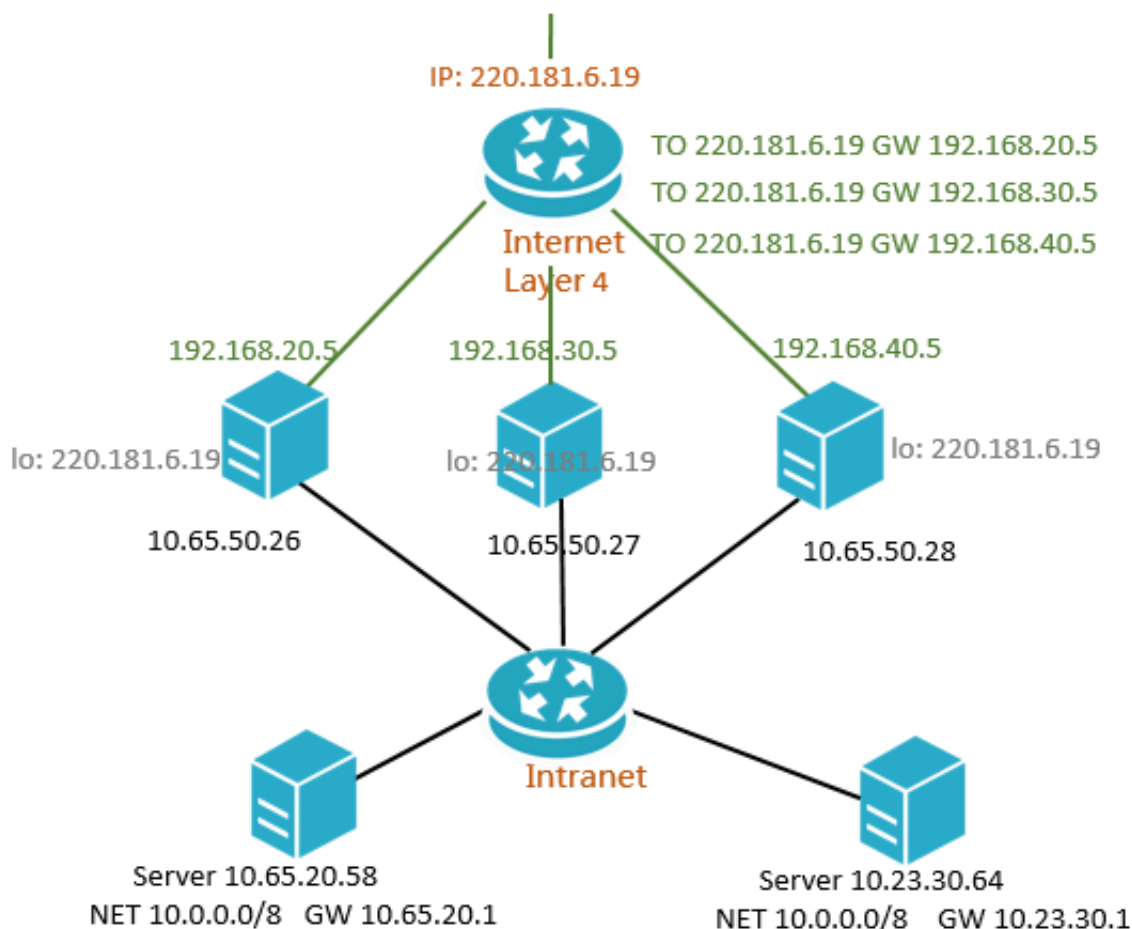
### Architecture

Currently, Alibaba Cloud provides Layer-4 (TCP protocol and UDP protocol) and Layer-7 (HTTP protocol and HTTPS protocol) load balancing services.

- Layer-4 SLB uses the open source software Linux Virtual Server (LVS) and Keepalived to achieve load balancing. It also customizes the software to adapt to cloud computing requirements.
- Layer-7 SLB uses Tengine to achieve load balancing. Tengine is a Web server project launched by another branch of the Alibaba Cloud Group (Taobao). Based on Nginx, Tengine provides a wide range of advanced features to support high-traffic websites.



As shown in the following figure, Layer-4 SLB in each region actually runs 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.

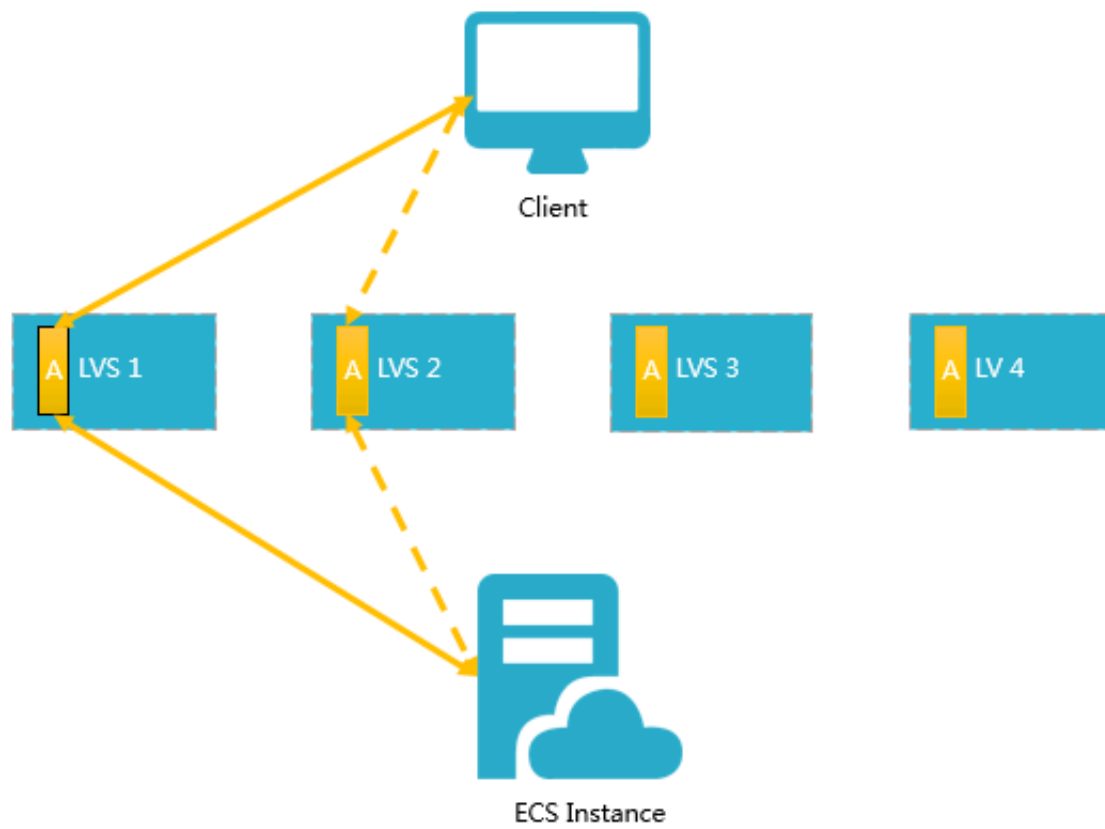


Additionally, each LVS machine in an LVS cluster uses multicast packets to synchronize sessions to other LVS machines. As shown in the following figure, after the client sends three packets to the server, session A is established on LVS1 and this session is synchronized to other LVS machines. In normal situations, the session request is sent to LVS1 as the solid line shows. If LVS1 fails or is being maintained, the session request will be sent to other normally working machines, such as LVS2, as the dotted line shows. Therefore, SLB clusters support hot upgrade, and machine failure or system maintenance will not affect your business.



#### Note:

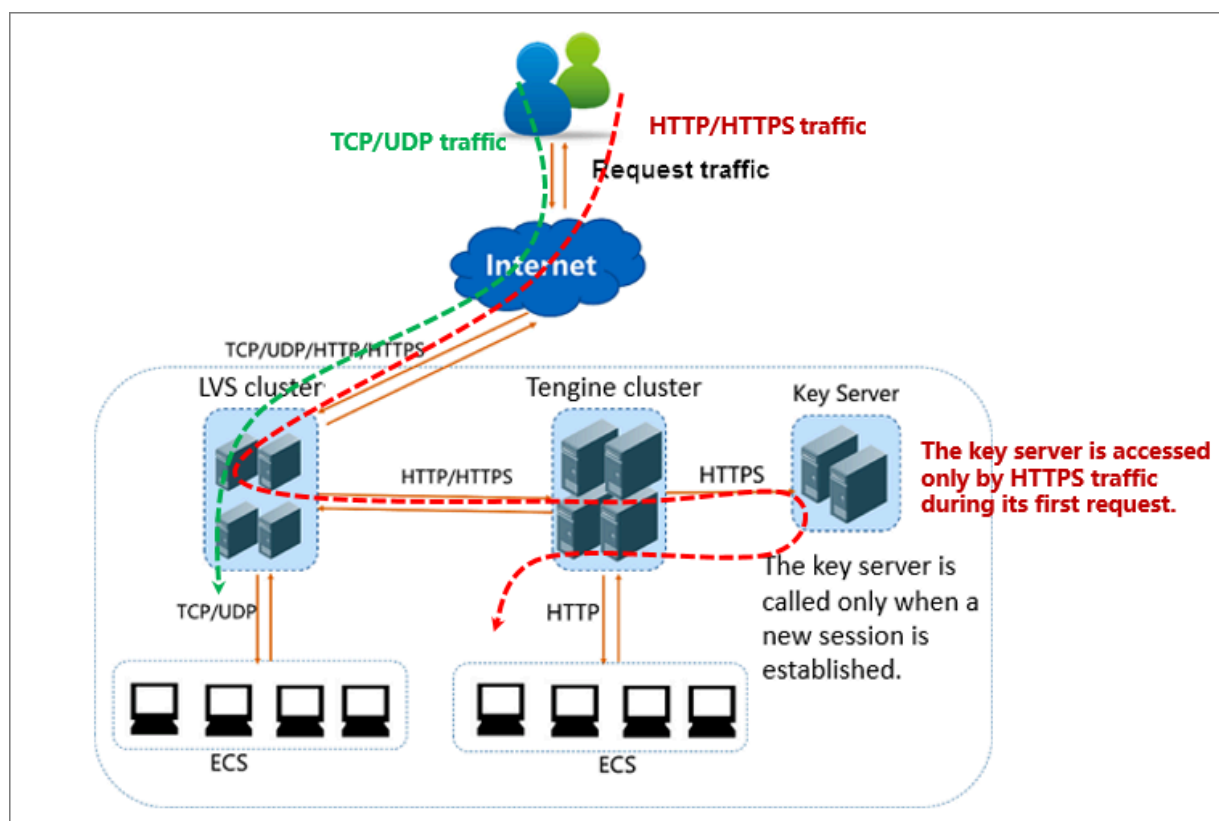
If a connection is not established (that is, a three-way handshake is not completed), or a connection has been established but the session synchronization is not triggered during a hot upgrade, your service may be interrupted and the client needs to re-initiate the connection.



## Inbound network traffic flow

SLB distributes incoming traffic according to the forwarding rules configured on the console or by using APIs. The following figure shows the inbound network traffic flow

Figure 2-1: Inbound network traffic flow



1. For TCP/UDP protocols and HTTP/HTTPS protocols, the incoming traffic must be forwarded through the LVS cluster first.
2. A massive number of access requests are evenly distributed among all servers in the LVS cluster. Servers synchronize sessions to guarantee high availability.
  - For Layer-4 listeners (the frontend protocol is UDP or TCP), the node servers in the LVS cluster distribute requests directly to backend ECS instances according to the configured forwarding rules.
  - For Layer-7 listeners (the frontend protocol is HTTP), the node servers in the LVS cluster first distribute requests to the Tengine cluster. Then, the node

servers in the Tengine cluster distribute the requests to backend ECS instances according to the configured forwarding rules.

- For Layer-7 listeners (the frontend protocol is HTTPS), the request distribution is similar to the HTTP protocol. However, before distributing requests to backend ECS instances, the system will call the Key Server to validate certificates and decrypt data packets.

#### Outbound network traffic

SLB communicates with backend ECS instances through the intranet.

- If backend ECS instances only need to handle the traffic distributed from SLB, no public bandwidth (EIP, NAT Gateway, and public IP) is required, and you do not need to purchase any public bandwidth.



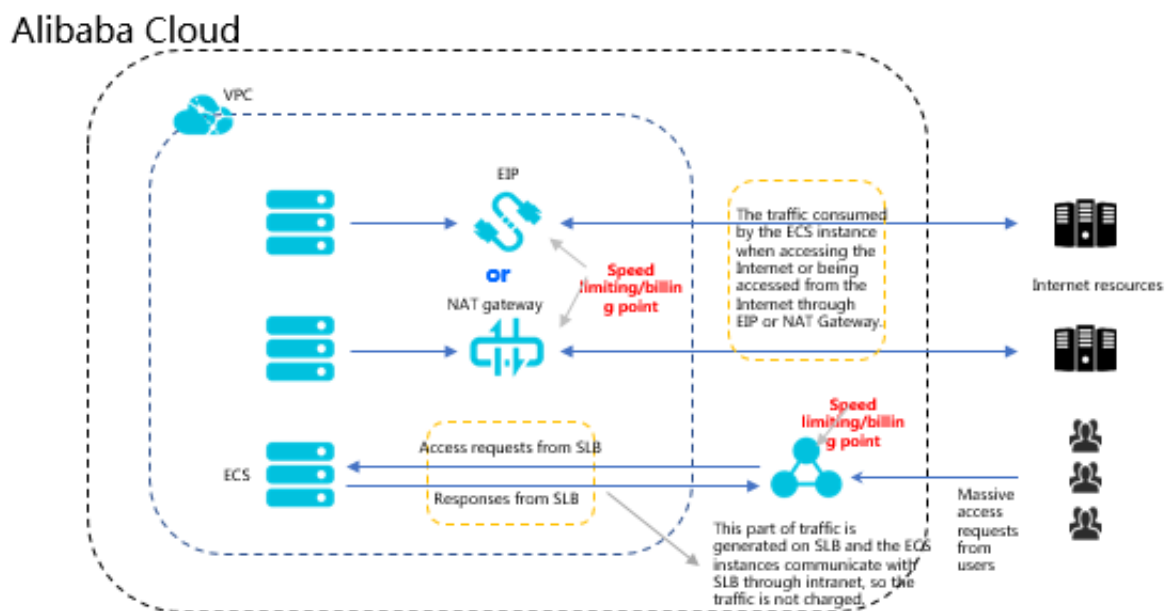
#### Note:

ECS instances of the previous generation are directly allocated with public IP addresses. You can view the public IP addresses by using the `ifconfig` command. If these ECS instances process requests only through SLB, no traffic fee is incurred for traffic sent through the Internet even traffic statistics are read at the public network interface (NIC).

- If you want to provide external services through backend ECS instances, or backend ECS instances need to access the Internet, you must configure at least one of the following: a public IP address, an EIP, or a NAT Gateway.

The following figure shows the outbound network traffic flow.

Figure 2-2: Outbound network traffic flow




- For outbound traffic from SLB instances (that is, traffic transferred through the Internet), traffic is sent at speeds dependent on the current network capacity, and is charged. However, you are not charged for intranet communications, such as traffic transferred between SLB instances and backend ECS instances.
- For outbound traffic from an EIP or from NAT Gateway (that is, traffic transferred through the Internet), traffic is sent at speeds dependent on the current network capacity, and is charged. Additionally, if an ECS instance is configured with a public IP address when it is created, the outbound traffic from this instance is also charged.
- SLB supports dynamic access to the Internet. Specifically, if a backend ECS instance needs to access the Internet, you must first configure a public IP address for it (by using an EIP or using NAT Gateway) and add it to the instance.
- A public IP address (configured when you create an ECS instance), EIP, and NAT gateway all allow mutual Internet access. That is, ECS instances can access the Internet or be accessed from the Internet through any of these. Note, however, that they cannot forward traffic or balance traffic loads.



## 3 Features

This topic lists key functions of Server Load Balancer (SLB) and indicates whether these functions are supported by Layer-4 and Layer-7 SLB instances.

Function	Layer-4 SLB	Layer-7 SLB
<p><b>Scheduling algorithms</b></p> <p>SLB supports round robin, weighted round robin, weighted least connections, and consistent hash.</p>	#	#  <b>Note:</b> Currently, Layer-7 SLB does not support the scheduling algorithm of consistent hash.
<p><b>Health checks</b></p> <p>SLB checks health status of backend servers. If a backend server is declared as unhealthy, SLB will stop distributing traffic to it and distribute incoming traffic to other healthy backend servers.</p>	#	#
<p><b>Session persistence</b></p> <p>SLB supports session persistence. In a session, SLB can distribute requests from the same client to the same backend server.</p>	#	#
<p><b>Access control</b></p> <p>SLB supports whitelists and blacklists to control access to your applications.</p>	#	#

Function	Layer-4 SLB	Layer-7 SLB
<b>High availability</b>  SLB can forward incoming traffic to backend servers in different zones. Additionally, SLB is deployed in active/standby mode in most regions. It will automatically switch to the secondary zone to provide load balancing service if the primary zone is unavailable.	#	#
<b>Security</b>  Combined with Alibaba Cloud Security, SLB can defend against up to 5 Gbit/s DDoS attacks.	#	#
<b>Internet and intranet load balancing</b>  SLB provides both Internet and intranet load balancing services. You can create an intranet SLB instance to balance traffic in your VPC network , or create an Internet SLB instance to balance traffic coming from the Internet.	#	#
<b>Monitoring</b>  With the CloudMonitor service, you can view the monitoring data of SLB, such as the number of connections and inbound and outbound traffic.	#	#
<b>IPv6 support</b>  SLB is able to forward requests from IPv6 clients.	#	#
<b>Access logs</b>  With Log Service, you can analyze access logs of an SLB instance to understand the behavior and geographical distribution of users, and troubleshoot problems.	—	#

Function	Layer-4 SLB	Layer-7 SLB
<b>Health check logs</b>  SLB stores health check logs of backend servers generated within three days by default. You can store all health check logs in OSS for troubleshooting.	#	#
<b>Domain name-based and URL-based forwarding</b>  Layer-7 SLB supports domain name-based and URL-based forwarding rules to forward requests from different domain names or URLs to different backend servers.	—	#
<b>Certificate management</b>  SLB provides a centralized certificate management service for applications using HTTPS protocols. You do not need to upload certificates to backend servers. Instead, deciphering is performed on SLB to reduce the CPU usage of backend servers.	—	#
<b>SNI support</b>  SLB supports configuring multiple certificates in an HTTPS listener to distribute requests with different domain names to different backend servers.	—	#
<b>Redirection</b>  SLB supports redirecting HTTP requests to HTTPS requests.	—	#

Function	Layer-4 SLB	Layer-7 SLB
<b>WS/WSS support</b>  WebSocket is a new HTML protocol. It provides bi-directional communication channels between clients and servers, saving server resources and bandwidth, and achieving real-time communication.	—	#
<b>HTTP/2 support</b>  HTTP/2 is the second version of Hypertext Transfer Protocol. It is backward compatible with HTTP1.X and significantly improves performance.	—	#

## 4 High availability of SLB

---

This topic describes the high-availability architecture of Server Load Balancer (SLB) in terms of different system designs and product configurations to meet different business needs. You can also use SLB together with Alibaba Cloud DNS to achieve cross-region disaster recovery.

### High availability of the SLB system

Deployed in clusters, SLB can synchronize sessions to protect the ECS instances from single points of failure (SPOFs). This improves redundancy and guarantees the service stability. Layer-4 SLB uses the open source software Linux Virtual Server (LVS) and Keepalived to achieve load balancing. Layer-7 SLB uses Tengine to achieve load balancing. Tengine, a Web server project based on Nginx, adds advanced features dedicated for high-traffic websites.

Requests from the Internet reach the LVS cluster through ECMP routing. Each LVS in the LVS cluster synchronizes the session to other LVS machines in the cluster through multicast packets, thereby implementing session synchronization among machines in the LVS cluster. At the same time, the LVS cluster performs health checks on the Tengine cluster and removes abnormal machines from the Tengine cluster to ensure the availability of Layer-7 SLB.

#### Best practice:

Session synchronization protects persistent connections from being affected by server failure in the cluster. However, for short connections or when the session synchronization rule is not triggered by the connection (the three-way handshake is not completed), server failure in the cluster may still affect user requests. To prevent session interruptions caused by machine failure in the cluster, you can add a retry mechanism to the service logic to reduce the impact on user access.

### High availability of a single SLB instance

To provide more reliable services, multiple zones for SLB are deployed in most regions. If a primary zone becomes unavailable, SLB rapidly switches to a secondary zone to restore its service capabilities within 30 seconds. When the primary zone becomes available, SLB automatically switches back to the primary zone.

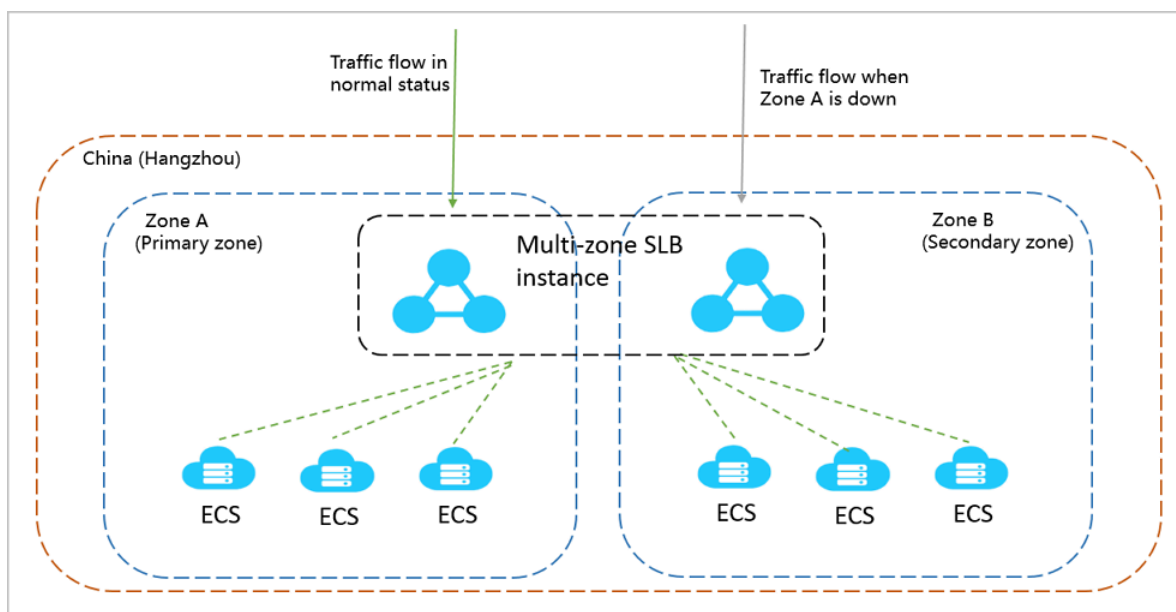
**Note:**

The primary zone and secondary zone form zone-level disaster tolerance. An SLB instance switches to the secondary zone only when Alibaba Cloud detects that the current zone is unavailable due to power outage or optical cable failures rather than the failure of an instance.

**Best practice:**

1. We recommend that you create an SLB instance in a region with multiple zones for disaster tolerance.
2. You can deploy ECS instances in the primary zone and secondary zone respectively as needed. You can set the zone where most ECS instances are located to the primary zone to minimize access latency.

However, we recommend that you do not deploy all ECS instances in one zone. You also need to deploy a small number of ECS instances in the secondary zone, so that the secondary zone can still process requests in extreme conditions (the primary zone is unavailable).

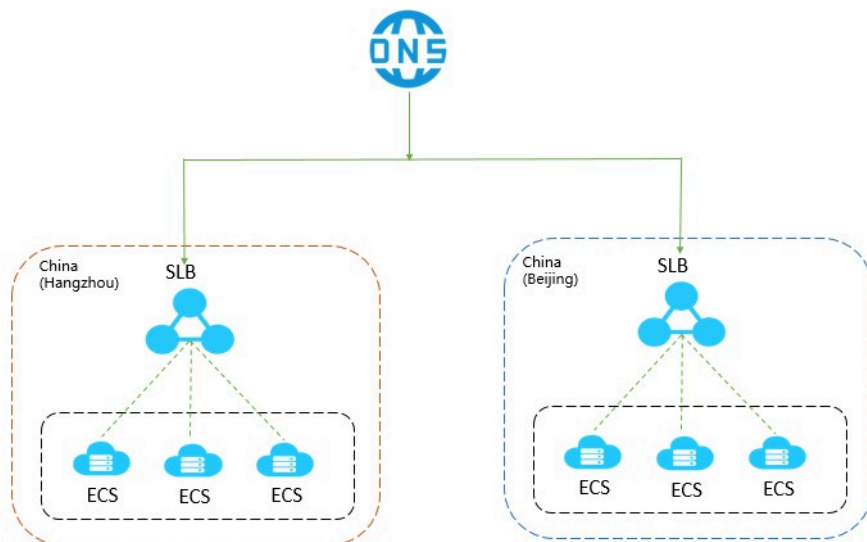
**High availability of multiple SLB instances**

If your availability requirements are extremely high, the availability guaranteeing mechanism of a single SLB may fail to meet your demands. For example, when the SLB instance is unavailable due to network attacks or configuration errors, zone switching is not triggered because no zone-level failure occurs. To solve that problem

, you can create multiple SLB instances and schedule requests by using Alibaba Cloud DNS, or achieve cross-region disaster recovery through global SLB.

**Best practice:**

You can deploy SLB instances and ECS instances in multiple zones of a region or in multiple regions and schedule access requests by using Alibaba Cloud DNS.



### High availability of backend ECS instances

SLB checks the service availability of backend ECS instances by performing health checks. Health checks improve the overall availability of frontend services and help reduce the impact of service availability when backend servers are abnormal.

When SLB discovers that an instance is unhealthy, it distributes requests to other healthy ECS instances, and only resumes distributing requests to the instance when it has restored to a healthy status. For more information, see [Health check overview](#).

**Best practice:**

You need to enable and correctly configure the health check function. For more information, see [Configure health checks](#).

## 5 Scenarios

---

Server Load Balancer (SLB) is suitable for applications with high access traffic to improve availability and reliability.

### Balance the loads of your applications

If your applications experience high access traffic, SLB can help you distribute inbound traffic among multiple backend servers (ECS instances) according to the listening rules you set. Additionally, you can configure session persistence to distribute the requests from the same client to the same backend server.

### Scale your applications

You can add or remove backend servers at any time to scale the serving capacity of your applications. SLB supports various web servers and application servers.

### Protect your applications from single points of failure

You can add multiple ECS instances to an SLB instance. If some ECS instances are faulty, SLB automatically shields faulty ECS instances and distributes requests to healthy ECS instances, guaranteeing that your applications can still work normally.

### Realize disaster tolerance across zones

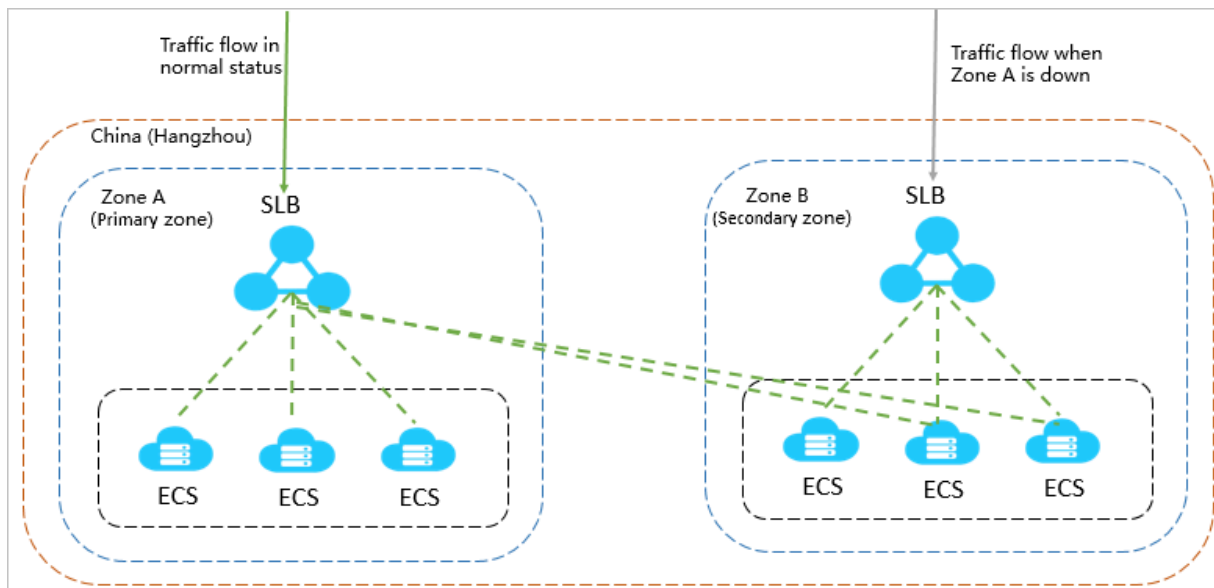
To provide more stable and reliable services, SLB provides multiple zones in most regions to achieve same-region disaster tolerance. If the primary zone becomes unavailable, SLB switches to the secondary zone in about 30 seconds and the load balancing service will hardly be affected. Once the primary zone becomes available, SLB automatically switches back to the primary zone.

We recommend that you create an SLB instance in a region that has deployed multiple zones. You also need to consider the deployment of backend servers. We recommend that you add at least one backend server in each zone to achieve the highest efficiency

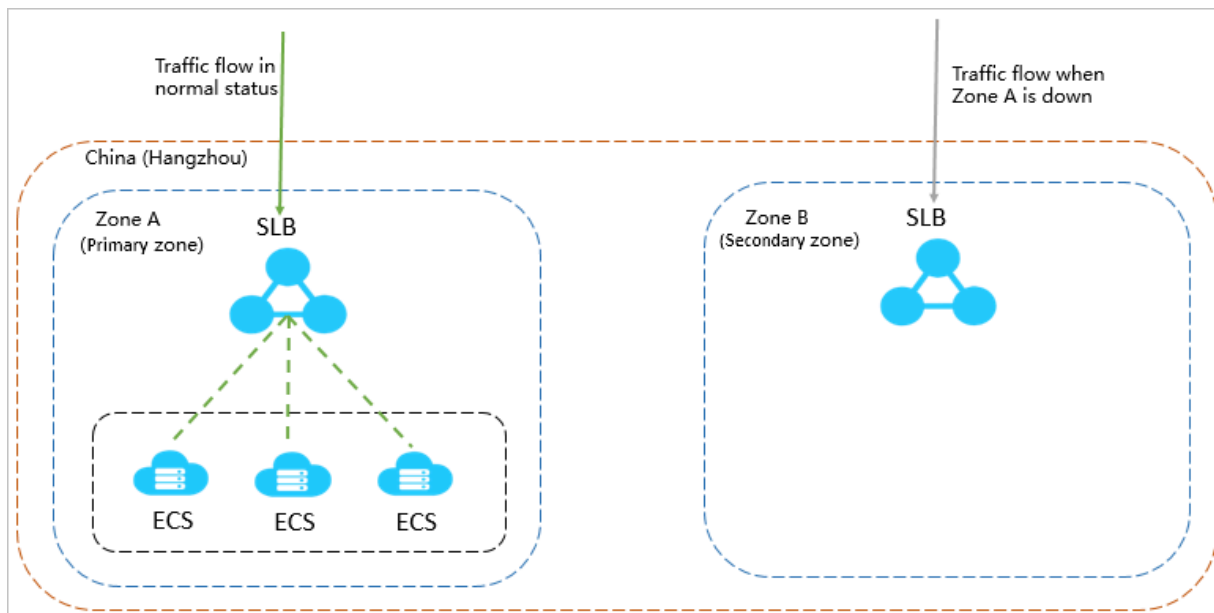
As shown in the following figure, ECS instances in different zones are added to an SLB instance. In normal situation, SLB will distribute traffic to ECS instances both in the primary zone (Zone A) and in the secondary zone (Zone B). Once the primary zone is unavailable, the traffic will be distributed to the ECS instances in the secondary



zone. This avoids service interruption caused by the failure of a single zone, and also reduces latency.



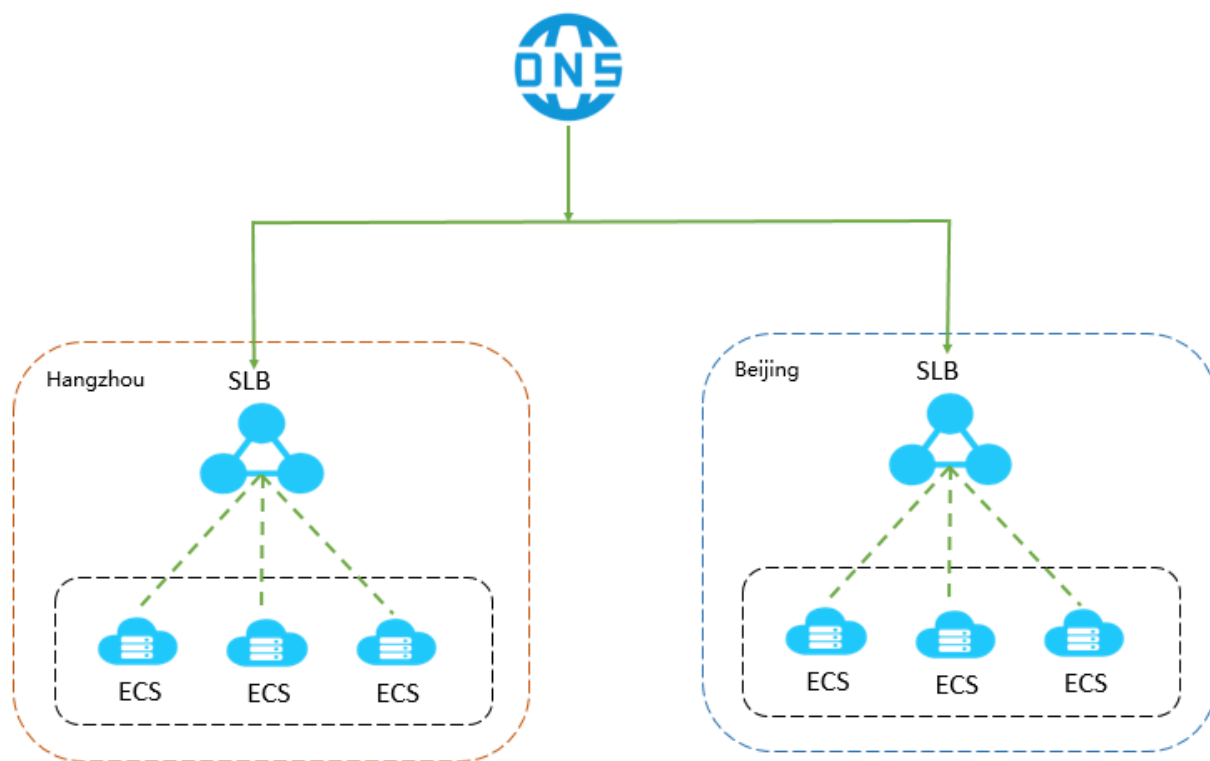
However, if you deploy all ECS instances in the primary zone and have no ECS instances deployed in the secondary zone, your service will be interrupted when the primary zone is unavailable, because no ECS instances are available in the secondary zone to handle the distributed requests. This deployment mode achieves low latency at the expense of high availability.



### Achieve disaster tolerance across regions

You can deploy SLB instances in different regions, and add ECS instances of different zones in the same region to an instance. Then, you can use Alibaba Cloud DNS to

resolve domain names to service addresses of SLB instances in different regions, thus implementing global load balancing. When a region becomes unavailable, you can temporarily stop DNS in that region, and no user access will be affected.



## 6 Terms

---

This topic describes the terms commonly used in Alibaba Cloud Server Load Balancer.

Term	Description
Server Load Balancer	Alibaba Cloud Server Load Balancer (SLB) is a traffic distribution and control service that distributes the incoming traffic among multiple Elastic Compute Service (ECS) instances according to configured forwarding rules.
Server Load Balancer instance	A Server Load Balancer instance is a running entity of the SLB service. To use SLB, you must first create an SLB instance.
Endpoint	An IP address allocated to an SLB instance. According to the instance type, the IP address is either a public IP address or a private IP address. You can resolve a domain name to a public IP address to provide external services.
Listener	A listener defines how incoming requests are distributed. An SLB instance must contain at least one listener.
Backend server	The ECS instances that are added to an SLB instance to process distributed requests.
Default server group	A group of ECS instances that process the distributed requests.  If a listener does not configure a VServer group or an active/standby server group, the default server group is used.  Incoming traffic is distributed to ECS instances in the default server group.
VServer group	A group of ECS instances that process the distributed requests.  Different listeners can be associated with different VServer groups to distribute different requests to different backend servers.
Active/standby server group	An active/standby server group contains only two ECS instances. One is the active server and the other one is the standby server. When the health check of the active server fails, SLB will automatically route traffic to the standby server.