ALIBABA CLOUD

Alibaba Cloud

Server Load Balancer
Product Introduction

Document Version: 20200924

Alibaba Cloud

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 because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should 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 this document based on the "status quo", "being defective", and "existing functions" of its products and services. 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 take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, 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 contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, 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 this document 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 directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
Danger	A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	Danger: Resetting will result in the loss of user configuration data.
Warning	A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	Warning: Restarting will cause business interruption. About 10 minutes are required to restart an instance.
Notice	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	Notice: If the weight is set to 0, the server no longer receives new requests.
Note	A note indicates supplemental instructions, best practices, tips, and other content.	Note: You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Table of Contents

1.What is SLB?

This topic provides an overview of Server Load Balancer (SLB). SLB distributes inbound network traffic across multiple Elastic Compute Service (ECS) instances that act as backend servers based on forwarding rules. You can use SLB to improve the responsiveness and availability of your applications.

Overview

After you add ECS instances that reside in the same region to an SLB instance, SLB uses virtual IP addresses (VIPs) to virtualize these ECS instances into backend servers in a high-performance server pool that ensures high availability. Client requests are distributed to the ECS instances based on forwarding rules.

SLB checks the health status of the ECS instances and automatically removes unhealthy ones from the server pool to eliminate single points of failure (SPOFs). This enhances the resilience of your applications. You can also use SLB to defend your applications against distributed denial of service (DDoS) attacks.

Components

SLB consists of three components:

- SLB instances

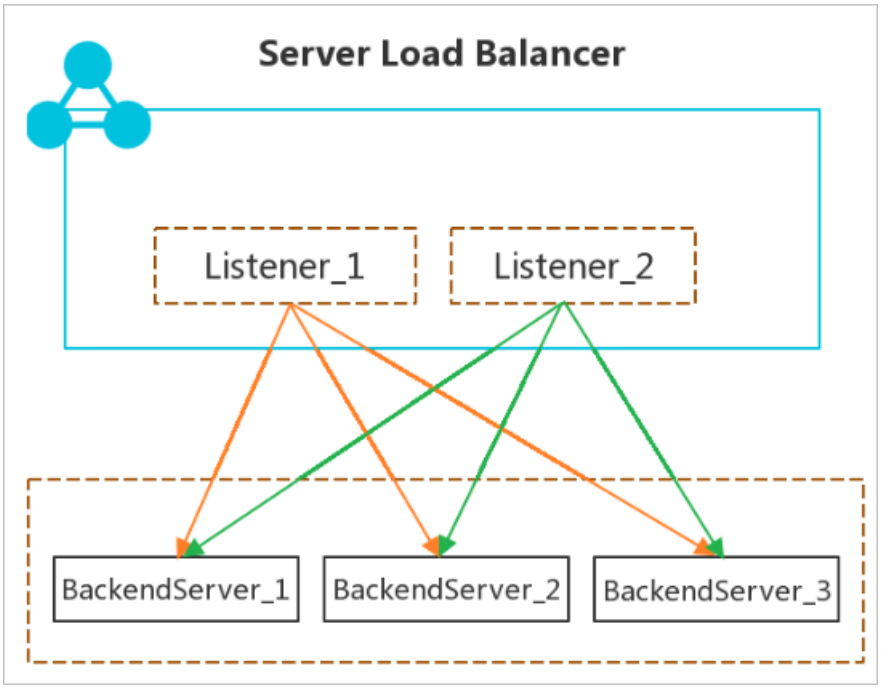
  An SLB instance is a key load-balancing component in SLB. It receives traffic and distributes traffic to backend servers. To get started with SLB, you must create an SLB instance and add at least one listener and two ECS instances to the SLB instance.

- Listeners

  A listener checks for connection requests from clients, forwards requests to backend servers, and performs health checks on backend servers.

- Backend servers

  ECS instances are used as backend servers in SLB to receive and process distributed requests. ECS instances can be added to the default server group of an SLB instance. You can also add multiple ECS servers to VServer groups or primary/secondary server groups after the corresponding groups are created.



Benefits

- High availability

  SLB features full redundancy that avoids SPOFs and supports zone-disaster recovery. You can use SLB with Apsara Stack DNS to achieve geo-disaster recovery with an availability of up to 99.95%.

  SLB can be scaled based on network traffic to protect your services from outages caused by fluctuating traffic flows.

- Strong scalability

  You can increase or decrease the number of backend servers to adjust the load balancing capacity for your applications.

- Low costs

  SLB can save 60% of load balancing costs compared with using traditional hardware solutions.

- Outstanding security

  You can use SLB with Alibaba Cloud Security to defend your applications against 5 Gbit/s distributed denial of service (DDoS) attacks.

- High concurrency

  An SLB cluster supports hundreds of millions of concurrent connections, and a single SLB instance supports tens of millions of concurrent connections.

2.High availability

This topic describes the high-availability architecture of SLB. You can use SLB in concert with DNS to implement geo-disaster recovery. SLB is designed to offer a multi-zone service availability of 99.99% and a single-zone service availability of 99.90%.

High availability of the SLB architecture

SLB instances are deployed in clusters to synchronize sessions and protect backend servers from SPOFs, improving redundancy and ensuring service stability. Layer-4 SLB uses the open-source Linux Virtual Server (LVS) and Keepalived software to balance loads, whereas Layer-7 SLB uses Tengine. Tengine, a web server project launched by Taobao, is based on NGINX and adds advanced features dedicated for high-traffic websites.

Requests from the Internet reach an LVS cluster along Equal-Cost Multi Path (ECMP) routes. In the LVS cluster, each machine uses multicast packets to synchronize sessions with the other machines. At the same time, the LVS cluster performs health checks on the Tengine cluster and removes unhealthy machines from the Tengine cluster to ensure the availability of Layer-7 SLB.

Best practice:

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

The high-availability solution with one SLB instance

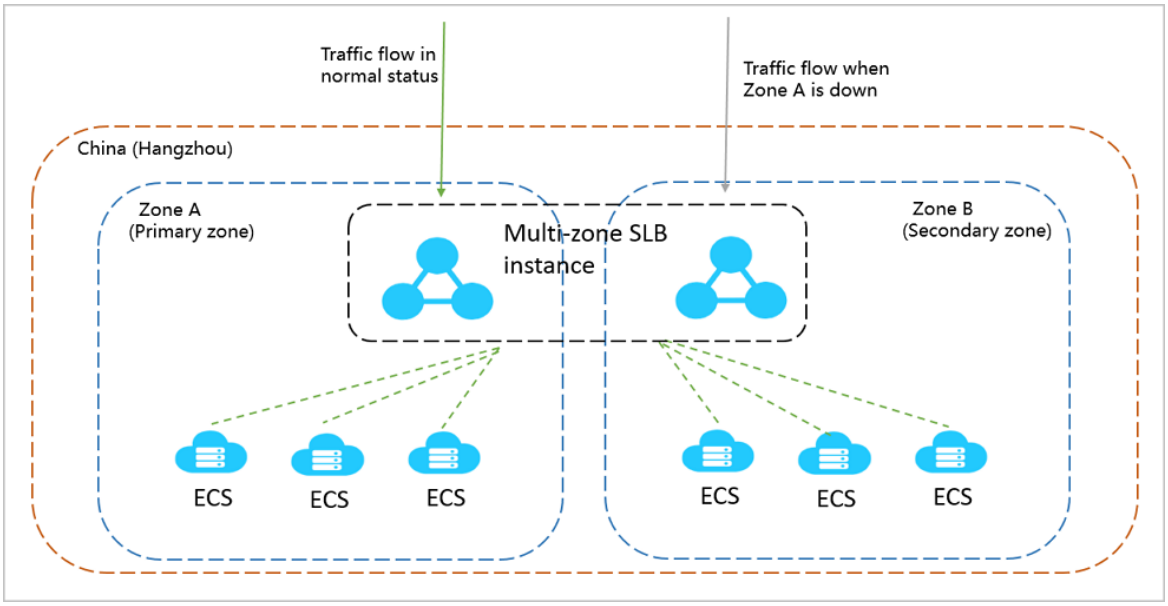
To provide more stable and reliable load balancing services, you can deploy SLB instances across multiple zones in most regions to achieve cross-data-center disaster recovery. Specifically, you can deploy an SLB instance in two zones within the same region whereby one zone acts as the primary zone and the other acts as the secondary zone. If the primary zone suffers an outage, a failover is triggered to redirect requests to the servers in the secondary zone within approximately 30 seconds. After the primary zone is restored, traffic will be automatically switched back to the servers in the primary zone.

> **Note** Zone-disaster recovery is implemented between the primary and secondary zones. SLB implements failovers only when the whole SLB cluster within the primary zone is unavailable or fails, for example, due to power outage or optical cable failures. A failover will not be triggered when a single backend server fails.

Best practice:

1. We recommend that you create SLB instances in regions that support primary/secondary deployment for zone-disaster recovery.
2. You can choose the primary zone for your SLB instance based on the distribution of ECS instances. That is, select the zone where most of the ECS instances are located as the primary zone for minimized latency.

However, we recommend that you do not deploy all ECS instances in the primary zone. When you develop a failover solution, you must deploy several ECS instances in the secondary zone to ensure that requests can still be distributed to backend servers in the secondary zone for processing when the primary zone experiences a downtime.

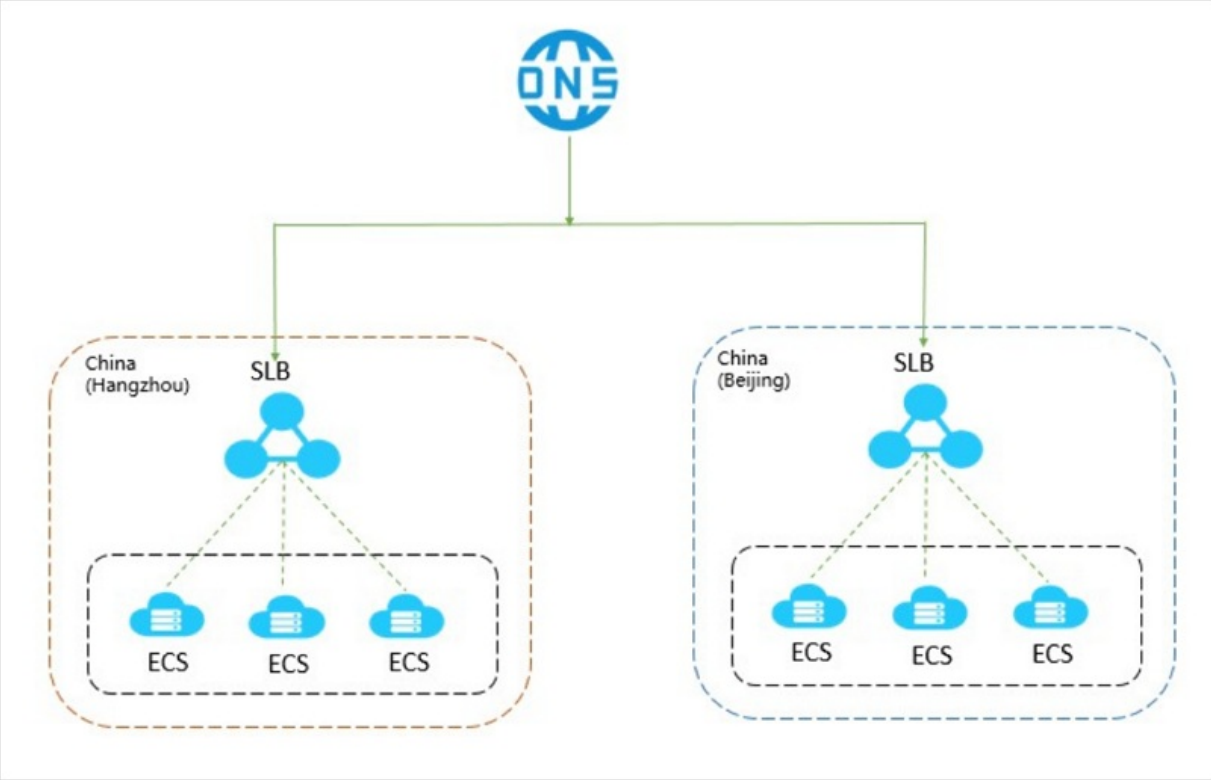


The high-availability solution with multiple SLB instances

In the context of one SLB instance, traffic distribution for your applications can still be compromised by network attacks or invalid SLB configurations, because the failover between the primary zone and the secondary zone is not triggered. As a result, the load-balancing performance is impacted. To avoid this situation, you can create multiple SLB instances to form a global load-balancing solution and achieve cross-region backup and disaster recovery. Also, you can use the instances with DNS to schedule requests so as to ensure service continuity.

Best practice:

You can deploy SLB instances and ECS instances in multiple zones within the same region or across different regions, and then use DNS to schedule requests.



The high-availability solution with backend ECS instances

With health check enabled, SLB verifies the availability of backend ECS instances (or backend servers), and thus improves the availability of frontend services by minimizing downtime that is caused by health issues of ECS instances.

After you enable the health check feature, when an ECS instance is detected unhealthy, SLB distributes new requests to other healthy ECS instances. SLB will only send requests to this backend ECS instance when it is restored and considered healthy. For more information, see Health check overview.

Best practice:

Make sure health check is enabled and properly configured. For more information, see Configure health check.

3.Architecture

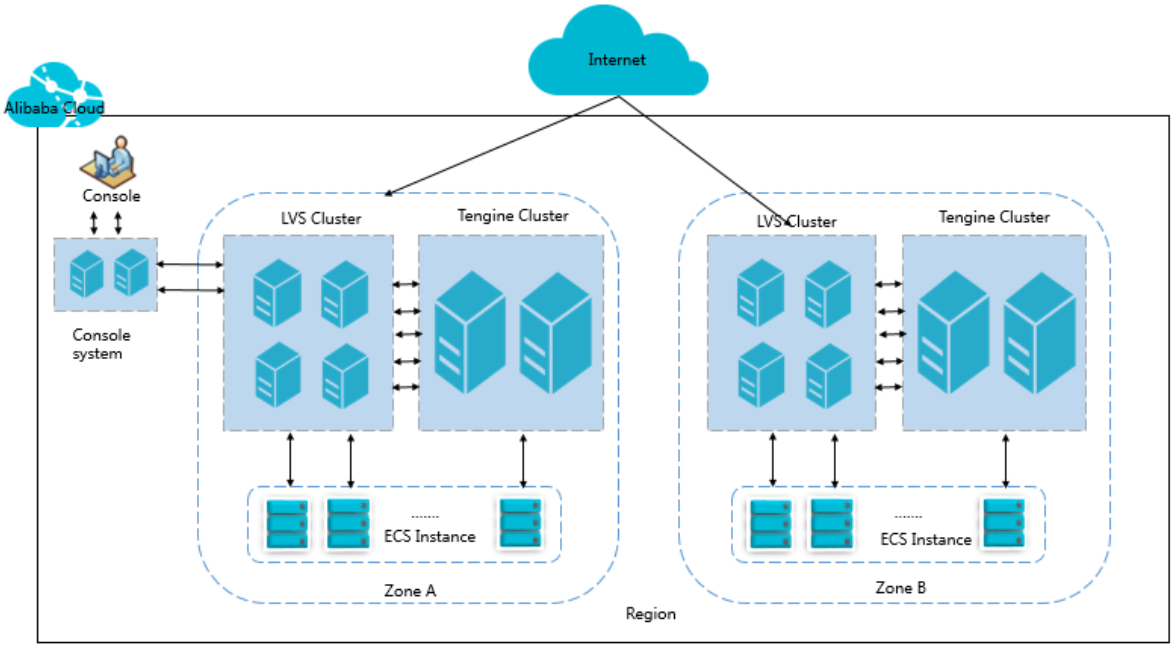
This topic describes the SLB architecture. SLB instances are deployed in clusters to synchronize sessions and protect backend servers from SPOFs, improving redundancy and ensuring service stability. SLB supports Layer-4 load balancing of Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) traffic, and Layer-7 load balancing of HTTP and HTTPS traffic.

SLB forwards client requests to backend servers by using SLB clusters and receives responses from backend servers over internal networks.

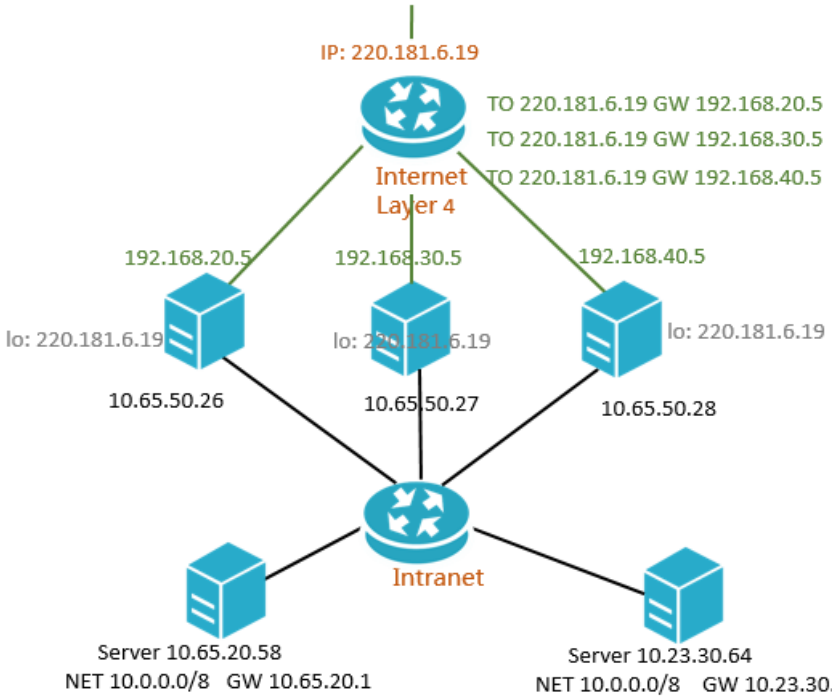
SLB design

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

- Layer-4 SLB combines the open-source Linux Virtual Server (LVS) with Keepalived to balance loads, and implements customized optimizations to meet cloud computing requirements.
- Layer-7 SLB uses Tengine to balance loads. Tengine is a web server project launched by Taobao. Based on NGINX, Tengine has a wide range of advanced features optimized for high-traffic websites.

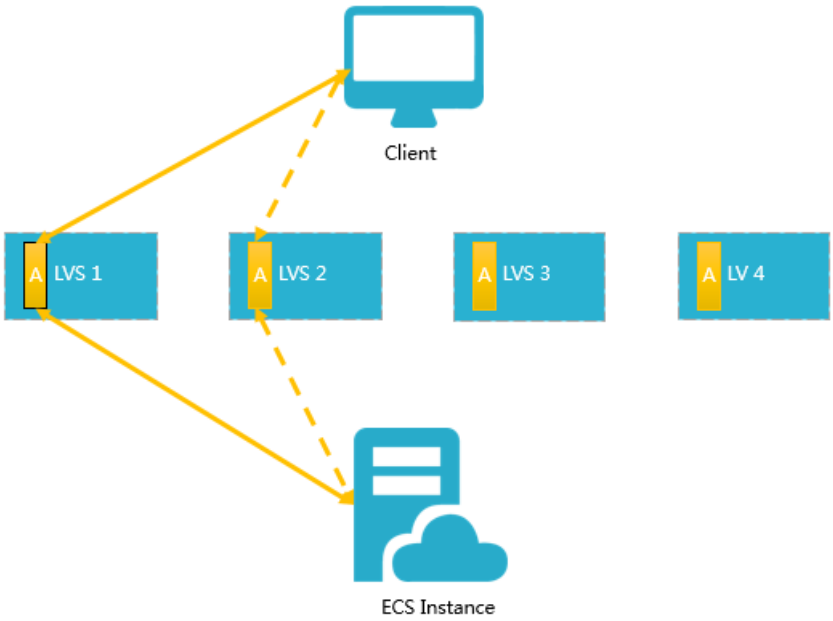


Layer-4 SLB runs in a cluster of LVS machines for higher availability, stability and scalability of load balancing in abnormal cases.



In an LVS cluster, each machine synchronizes sessions with other machines via multicast packets. As shown in the below figure, Session A is established on LVS1 and is synchronized to other LVS machines after the client transfers three data packets to the server. Solid lines indicate the current active connections, while dotted lines indicate that the session requests will be sent to other normally working machines if LVS1 fails or is being maintained. In this way, you can perform hot updates, machine maintenance, and cluster maintenance without affecting business applications.

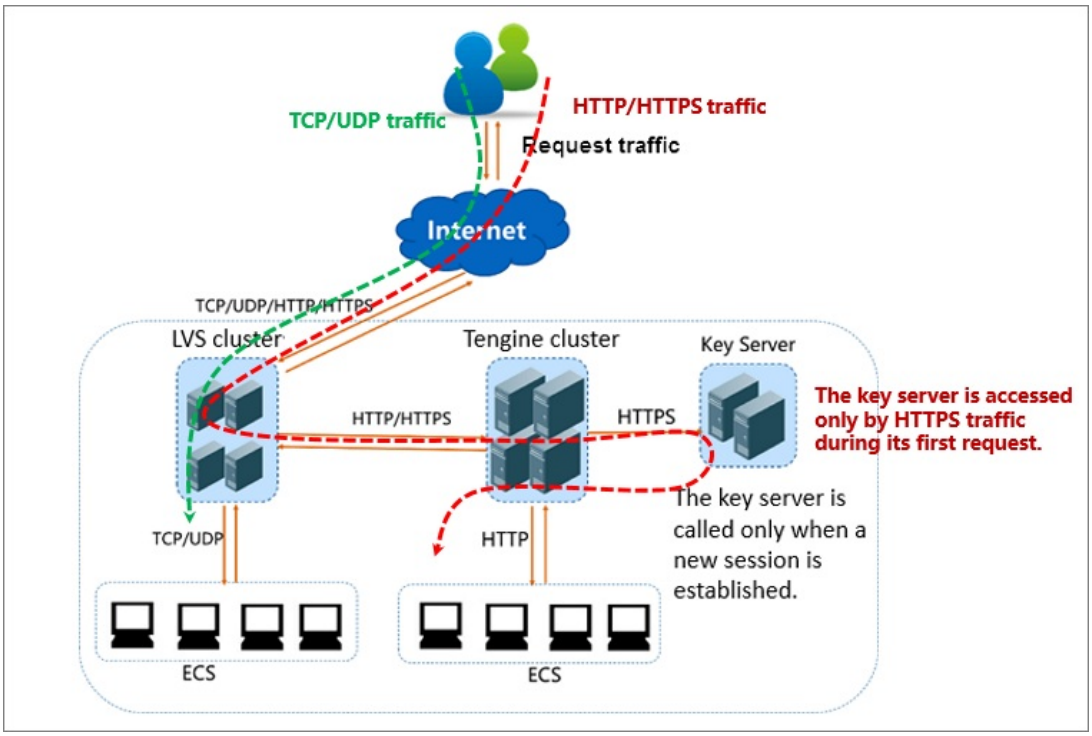
> **Note** If a connection is not established (the three-way handshake is not completed), or if a connection has been established but session synchronization is not triggered during a hot upgrade, your service may be interrupted. In this case, the client needs to re-initiate the connection.



Inbound network traffic flow

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

Inbound network traffic flow



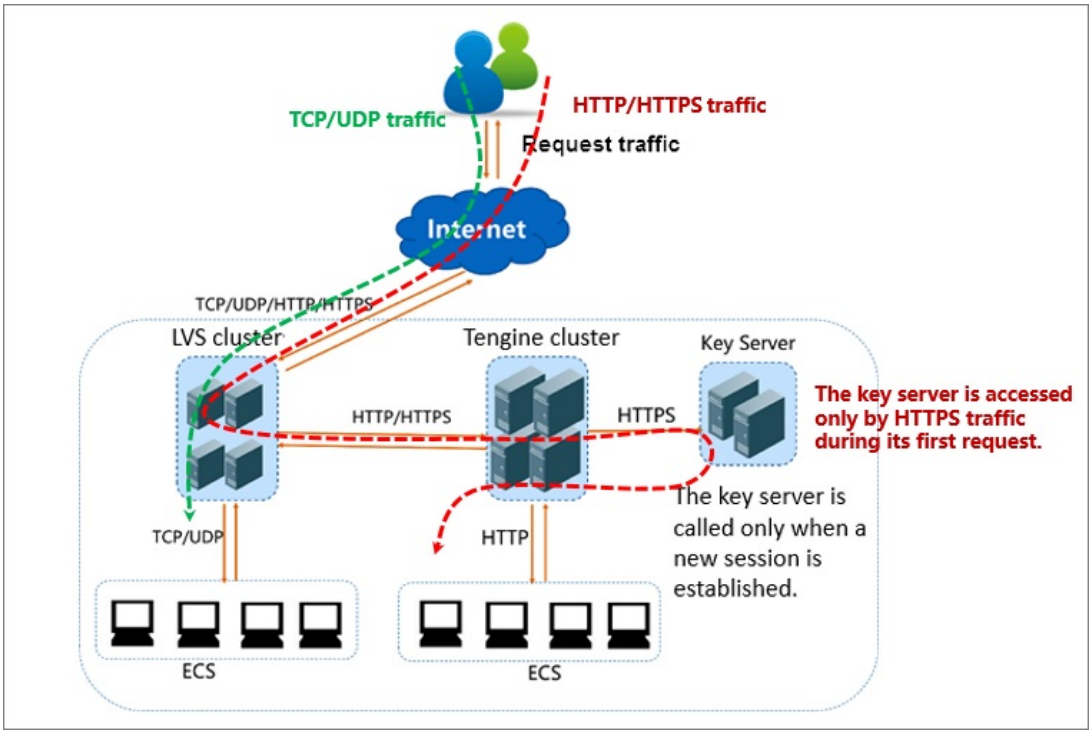
1. For TCP, UDP, HTTP, and HTTPS protocols, the incoming traffic must be forwarded through the LVS cluster first.
2. Large amounts of access requests are evenly distributed among all servers in the LVS cluster. Servers synchronize sessions to guarantee high availability.

- 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 that use the frontend protocol 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 that use the frontend protocol HTTPS, the request distribution is similar to the HTTP protocol. However, before distributing requests to backend ECS instances, the system calls the Key Server to validate certificates and decrypt data packets.

Inbound network traffic flow

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

Inbound network traffic flow



1. For TCP, UDP, HTTP, and HTTPS protocols, the incoming traffic must be forwarded through the LVS cluster first.
2. Large amounts 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 that use the frontend protocol 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 that use the frontend protocol HTTPS, the request distribution is similar to the HTTP protocol. However, before distributing requests to backend ECS instances, the system calls the Key Server to validate certificates and decrypt data packets.

4.Features

This topic describes the key features of Alibaba Cloud SLB, including Layer-4 and Layer-7 load balancing, health check, and session persistence for high availability of backend servers.

5.Scenarios

Server Load Balancer (SLB) can be used to improve the availability and reliability of applications with high access traffic.

Balance the loads of your applications

You can configure listening rules to distribute heavy traffic among ECS instances that are attached as backend servers to SLB instances. You can also use the session persistence feature to forward all of the requests from the same client to the same backend ECS instance to enhance access efficiency.

Scale your applications

You can extend the service capability of your applications by adding or removing backend ECS instances to suit your business needs. SLB can be used for both web servers and application servers.

Eliminate single points of failure (SPOFs)

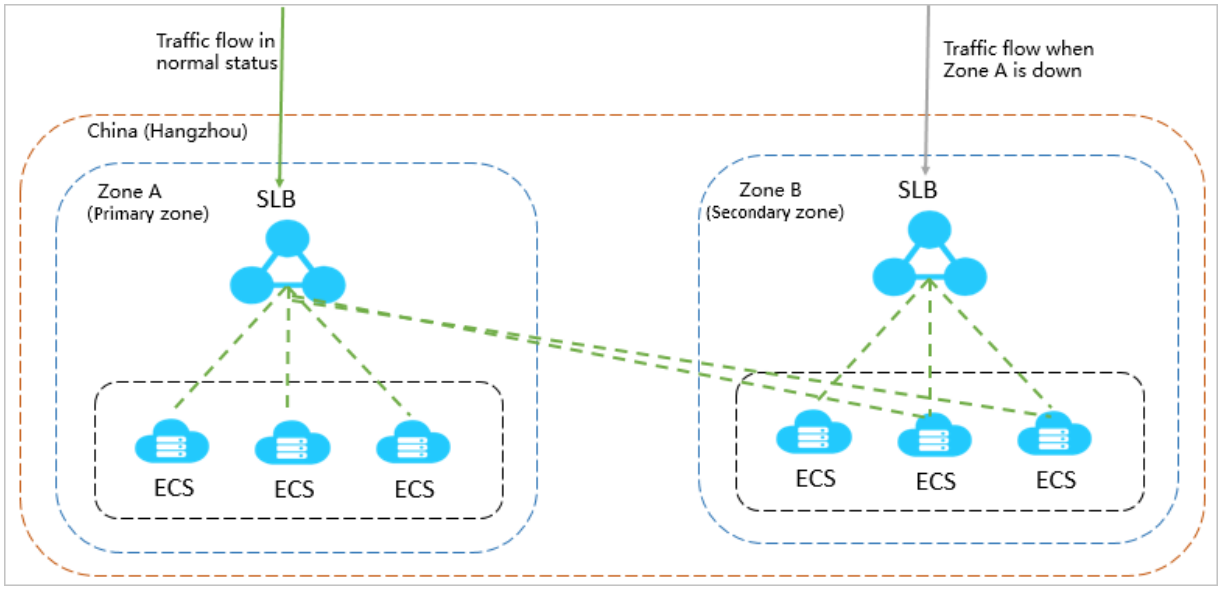
You can attach multiple ECS instances to an SLB instance. When an ECS instance malfunctions, SLB automatically isolates this ECS instance and distributes inbound requests to other healthy ECS instances, ensuring that your applications continue to run properly.

Implement zone-disaster recovery (multi-zone disaster recovery)

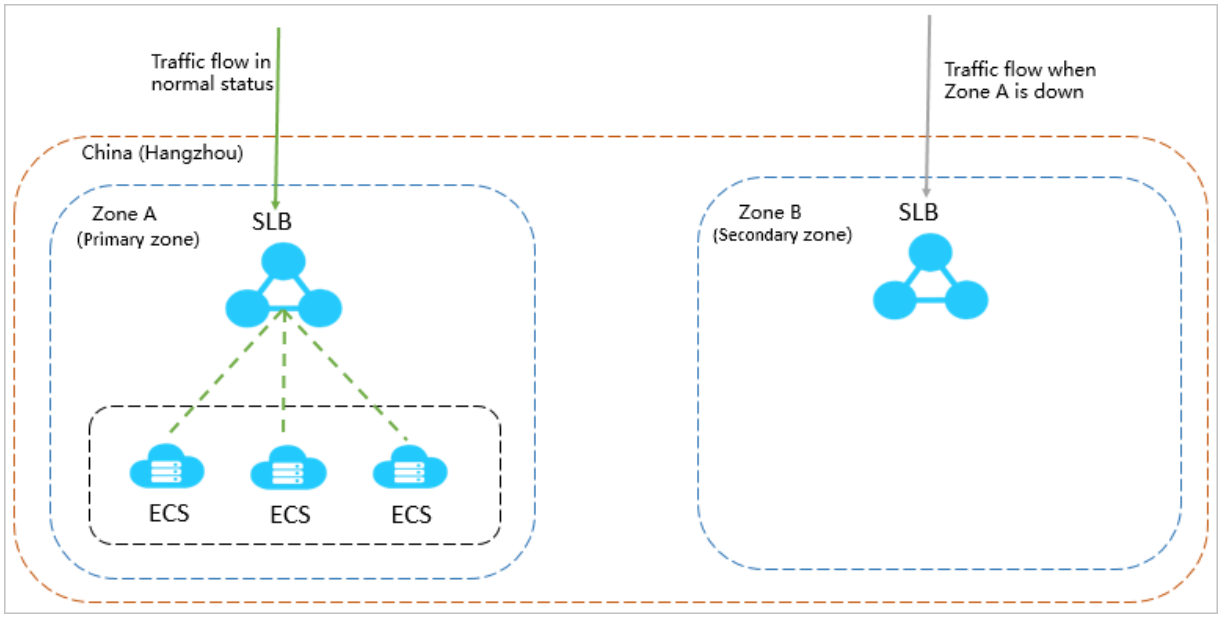
To provide more stable and reliable load balancing services, Alibaba Cloud allows you to deploy SLB instances across multiple zones in most regions for disaster recovery. Specifically, you can deploy an SLB instance in two zones within the same region. One zone is the primary zone, while the other zone is the secondary zone. If the primary zone fails or becomes unavailable, the SLB instance will fail over to the secondary zone in about 30 seconds. When the primary zone recovers, the SLB instance will automatically switch back to the primary zone.

We recommend that you create an SLB instance in a region that has multiple zones for zone-disaster recovery. We recommend that you plan the deployment of backend servers based on your business needs. In addition, we recommend that you add at least one backend server in each zone to achieve the highest load balancing efficiency.

As shown in the following figure, ECS instances in different zones are attached to a single SLB instance. In normal cases, the SLB instance distributes inbound traffic to ECS instances both in the primary zone (Zone A) and in the secondary zone (Zone B). If Zone A fails, the SLB instance distributes inbound traffic only to Zone B. This deployment mode helps avoid service interruptions caused by zone-level failure and reduce latency.

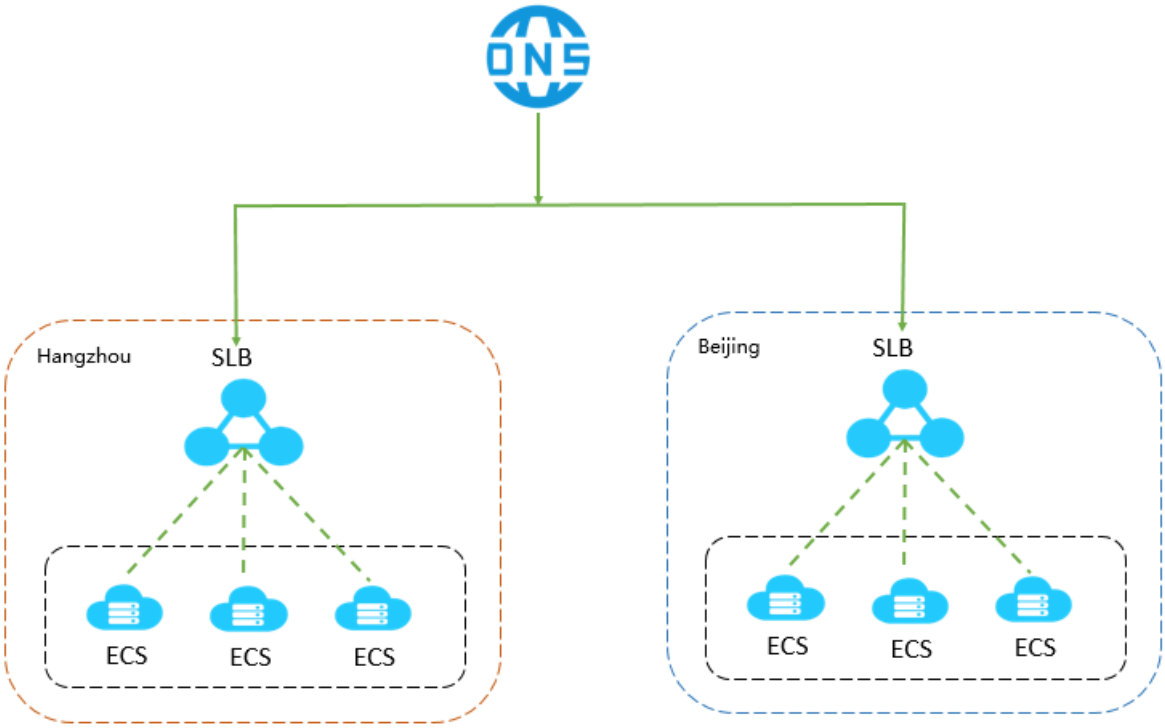


Assume that you deploy all ECS instances in the primary zone (Zone A) and no ECS instances in the secondary zone (Zone B) as shown in the following figure. If Zone A fails, your services will be interrupted because no ECS instances are available in Zone B. This deployment mode achieves low latency at the cost of high availability.



Geo-disaster recovery

You can deploy SLB instances in different regions and attach ECS instances of different zones within the same region to an SLB instance. You can use DNS to resolve domain names to service addresses of SLB instances in different regions for global load balancing purposes. When a region becomes unavailable, you can temporarily stop DNS resolution within that region without affecting user access.


DNS
Hangzhou
SLB
ECS
ECS
ECS
Beijing
SLB
ECS
ECS
ECS

6.Terms

This topic introduces the terms used in SLB.

Term	Description
SLB	SLB distributes traffic across ECS instances. SLB provides Layer-4 and Layer-7 load balancing.
SLB instance	A load-balancing instance in SLB. To get started with SLB, you must create an SLB instance.
Endpoint	An IP address assigned to an SLB instance. The IP address can be either public or private, depending on the type of the SLB instance. You can resolve a domain name to a public IP address of an SLB instance to provide external services.
Listener	A listener distributes requests to backend servers. Each SLB instance must have at least one listener.
Backend server	A backend server is an ECS instance that receives client requests distributed by an SLB instance.
Default server group	A group of ECS instances that process distributed requests. If a listener is not configured with any VServer group or primary/secondary server group, the listener distributes traffic to the backend servers in the default server group.
VServer group	A group of ECS instances that process distributed requests. You can create multiple VServer groups for different listeners of an SLB instance to specify traffic distribution with specific listeners.
Primary/secondary server group	Each primary/secondary server group contains two ECS instances, where one acts as the primary server and the other acts as the secondary server. If the primary server is detected unhealthy, new requests are automatically distributed to the secondary server.