

Alibaba Cloud Server Load Balancer

Product Introduction

Issue: 20181214

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.	 Danger: 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.	 Warning: 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.	 Note: 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.	 Note: You can use Ctrl + A to select all files.
>	Multi-level menu cascade.	Settings > Network > Set network type
Bold	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 a optional value, and only one item can be selected.	<code>ipconfig [-all -t]</code>
{ } or {a 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.....	6
4 High availability best practice.....	9
5 Scenarios.....	12
6 Terms.....	15

1 What is Server Load Balancer?

Overview

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

By setting a virtual service address, SLB virtualizes the added ECS instances into an application service pool with 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 the added backend servers, and automatically isolates abnormal ECS instances to eliminate single point of failure (SPOF), improving the overall service capability of your application. Additionally, working with Alibaba Anti-DDoS, SLB can defend DDoS attacks.

Server Load Balancer consists of the following components:

SLB consists of the following components:

- SLB instances

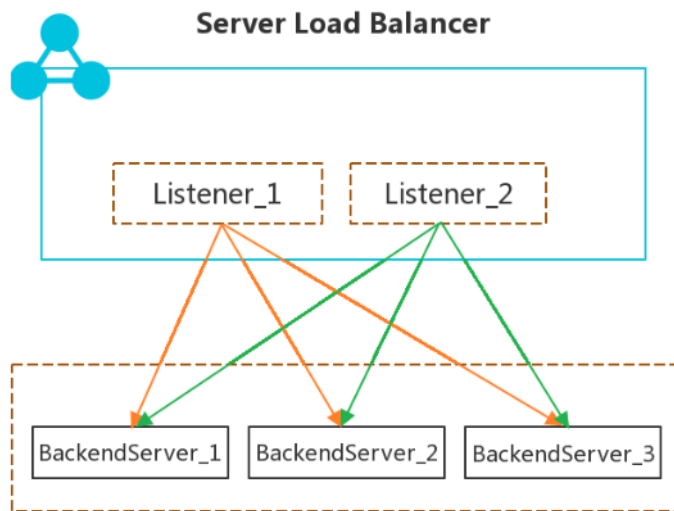
An SLB instance is a running load balancing service that distributes incoming traffic to backend servers. To use the load balancing service, you must create an SLB instance, and then configure the instance with at least one listener and two backend servers.

- Listeners

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

- Backend Servers

Backend servers are the ECS instances added to a 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 to process distributed requests.



Benefits

- High availability

Server Load Balancer is designed to work in the full-redundancy mode without SPOF. Server Load Balancer supports local and cross-region disaster tolerance. When Server Load Balancer is used together with DNS, the service availability is up to 99.95%.

You can scale your service based on the application load, without interrupting services continuity.

- Scalable

You can increase or decrease the number of backend servers as needed to expand the service capabilities of your applications.

- Cost effectiveness

Compared with the traditional hardware load balancing system, Server Load Balancer reduces the cost by 60%.

- Secure

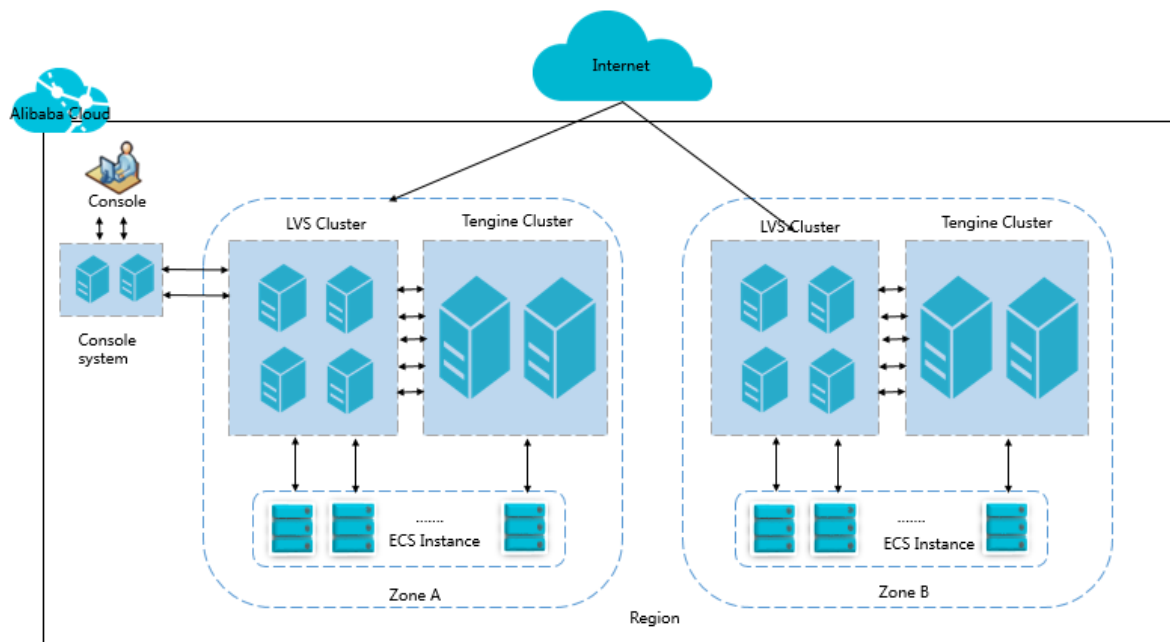
Combined with Alibaba Cloud Security, Server Load Balancer can defend against up to 5 Gbps DDoS attacks, such as HTTP flood and SYN flood attacks.

2 Architecture

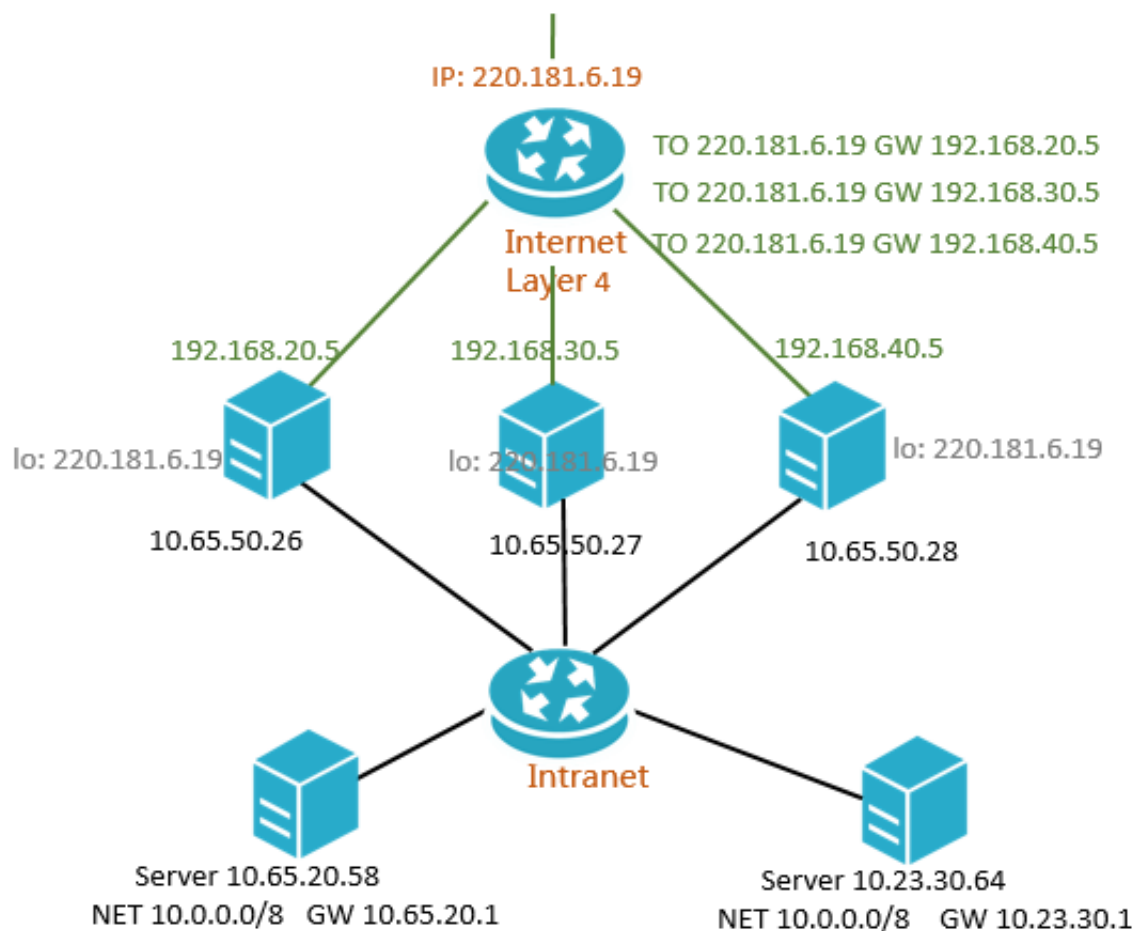
Server Load Balancer is deployed in clusters. The cluster deployment model eliminates Single Point of Failures of backend servers, improves redundancy and increases service stability.

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

- 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 SLB uses Tengine to achieve load balancing. Tengine is a Web server project launched by Taobao. Based on Nginx, it adds a wide range of advanced features dedicated for high-traffic websites.



As shown in the following figure, Layer-4 Server Load Balancer 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.

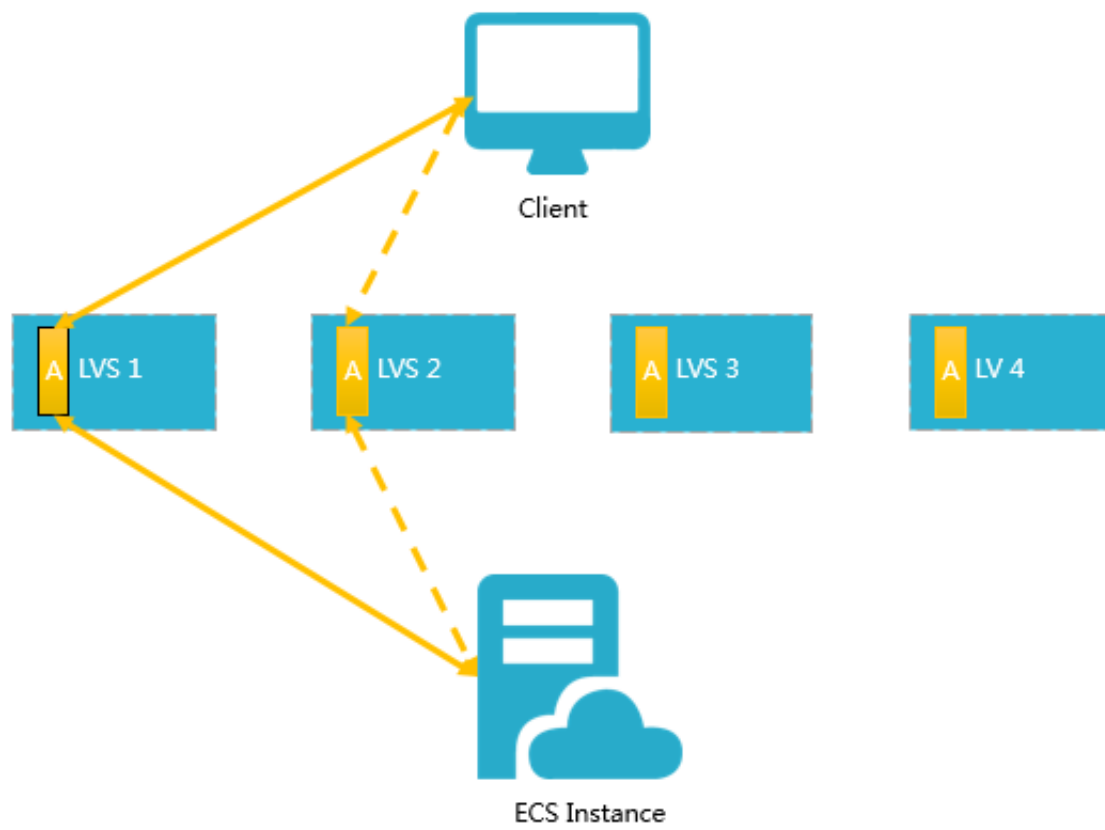


Additionally, the LVS machine in the LVS cluster uses multicast packets to synchronize sessions to other LVS machines. As shown in the following figure, the session A established on LVS1 is synchronized to other LVS machines after three packets are transferred. In normal situations, the session request is sent to LVS1 as the solid line shows. If LVS1 is abnormal or being maintained, the session request will be sent to other normally working machines 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 (three-way handshake is not completed), or a connection has been established but the session synchronization is not triggered, the hot upgrade does not guarantee that the connection is not interrupted and the client needs to re-initiate the connection.



3 Features

Alibaba Cloud provides Layer-4 and Layer-7 load balancing services, and other functions such as health check, session persistence, domain name based forwarding and so on to ensure high availability of your applications.

Functions	Layer-4 Server Load Balancer	Layer-7 Server Load Balancer
Scheduling algorithm Server Load Balancer supports round robin, weighted round robin (WRR), weighted least connections (WLC), and consistent hash.	✓	✓
Health check Server Load Balancer checks the health status of backend servers. If a backend server is declared as unhealthy, Server Load Balancer will stop distributing traffic to it and distribute incoming traffic to other healthy backend servers.	✓	✓
Session persistence Server Load Balancer supports session persistence. In a session, Server Load Balancer can distribute requests from the same client to the same backend server.	✓	✓
Access control Server Load Balancer support adding whitelists and blacklists to control access to your applications.	✓	✓
High availability Server Load Balancer can forward incoming traffic to backend servers in different zones. Additionally, Server Load Balancer is deployed in the active/standby mode in most regions. Server Load Balancer will automatically switch to the standby zone to provide the load balancing service if the primary zone is unavailable.	✓	✓
Security Combined with Alibaba Cloud Security, Server Load Balancer can defend against up to 5 Gbps DDoS attacks.	✓	✓
Internet and intranet load balancing Server Load Balancer provides both Internet and intranet load balancing services. You can create an intranet SLB instance to balance traffic in your VPC network, or create	✓	✓

Functions	Layer-4 Server Load Balancer	Layer-7 Server Load Balancer
an Internet SLB instance to balance traffic coming from the Internet.		
Monitoring With the CloudMonitor service, you can view the number of connections, traffic and more of an SLB instance.	✓	✓
IPv6 support Server Load Balancer supports forwarding requests from IPv6 clients.	✓	✓
Access logs With Log Service, you can analyze access logs of an SLB instance to understand the behavior and geographical distribution of users, troubleshoot problems and more.	—	✓
Health check logs Server Load Balancer stores health check logs of backend servers generated within three days by default. You can store all health check logs in OSS for troubleshooting.	✓	✓
Domain name/URL based forwarding Layer-7 Server Load Balancer supports configuring domain name/URL based forwarding rules to forward requests from different domain names or URLs to different backend servers.	—	✓
Certificate management Server Load Balancer provides centralized certificate management service for applications using HTTPS protocols. You do not need to upload certificates to backend servers. Deciphering is performed on Server Load Balancer to reduce the CPU usage of backend servers.	—	✓
SNI support Server Load Balancer supports configuring multiple certificates in an HTTPS listener to distribute requests with different domain names to different backend servers.	—	✓
Redirection Server Load Balancer supports redirecting HTTP requests to HTTPS requests.	—	✓
WS/WSS support	—	✓

Functions	Layer-4 Server Load Balancer	Layer-7 Server Load Balancer
WebSockets is a new HTML protocol. It provides bi-directional communication channels between a client and a server, saving server resources and bandwidth and achieving real-time communication.		
HTTP/2 support HTTP/2 is the second version of Hypertext Transfer Protocol. It is backward compatible with HTTP1.X and significantly improves performance.	—	✓

4 High availability best practice

Server Load Balancer (SLB) guarantees availability in terms of system design, product configuration and so on. Besides, you can use SLB together with Alibaba Cloud DNS to achieve cross-region disaster recovery according to business needs.

High availability of SLB system

Deployed in clusters, Server Load Balancer (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) with 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 check on the Tengine cluster and removes abnormal machines from the Tengine cluster to ensure the availability of Layer-7 Server Load Balancer.

Best practice:

Session synchronization ensures that persistent connections are not affected by server failure in the cluster. But for short connections or when the session synchronization rule is not triggered by the connection (three-way handshake is not completed), server failures in the cluster may still affect user requests. To prevent session interruption 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 Server Load Balancer are deployed in most regions. If a active zone becomes unavailable, Server Load Balancer rapidly switches to a standby zone to restore its service capabilities within 30 seconds. When the active zone becomes available, Server Load Balancer automatically switches back to the active zone.



Note:

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

Best practice:

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

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

High availability of multiple SLB instances

If your requirements on the availability is extremely high, the availability guaranteeing mechanism of the SLB may meet your demands. For example, when the SLB instance is unavailable due to network attack or configuration error, zone switching is not triggered because no zone-level failure occurs. At this time, you can create multiple SLB instances and schedule requests by using Alibaba Cloud DNS, or achieve cross-region disaster recovery through global Server Load Balancer.

Best practice:

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

High availability of backend ECS instances

Server Load Balancer checks the service availability of backend ECS instances by performing health checks. The health checks improve the overall availability of front-end services and help reduce the impact of service availability when backend servers are abnormal.

When Server Load Balancer 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:

In order for the health check function to work, you need to enable and correctly configure the health check. For more information, see [Configure health check](#).

5 Scenarios

Server Load Balancer is suitable for applications with high traffic, improving the availability and reliability.

Load balance your applications

Server Load Balancer can automatically distribute incoming traffic across multiple backend servers (ECS instances). Additionally, the requests from the same client can be distributed to the same backend server by configuring session persistence.

Scale your applications

To meet the demand of your customers, you can increase the number of the backend servers at any time to scale your applications. SLB is applicable to various web servers and application servers.

Protect your applications from single point of failures

You can add multiple ECS instances to a Server Load Balancer instance. When some ECS instances are faulty, Server Load Balancer automatically shields faulty ECS instances and distributes requests to healthy ECS instances, guaranteeing that the application system can still work normally.

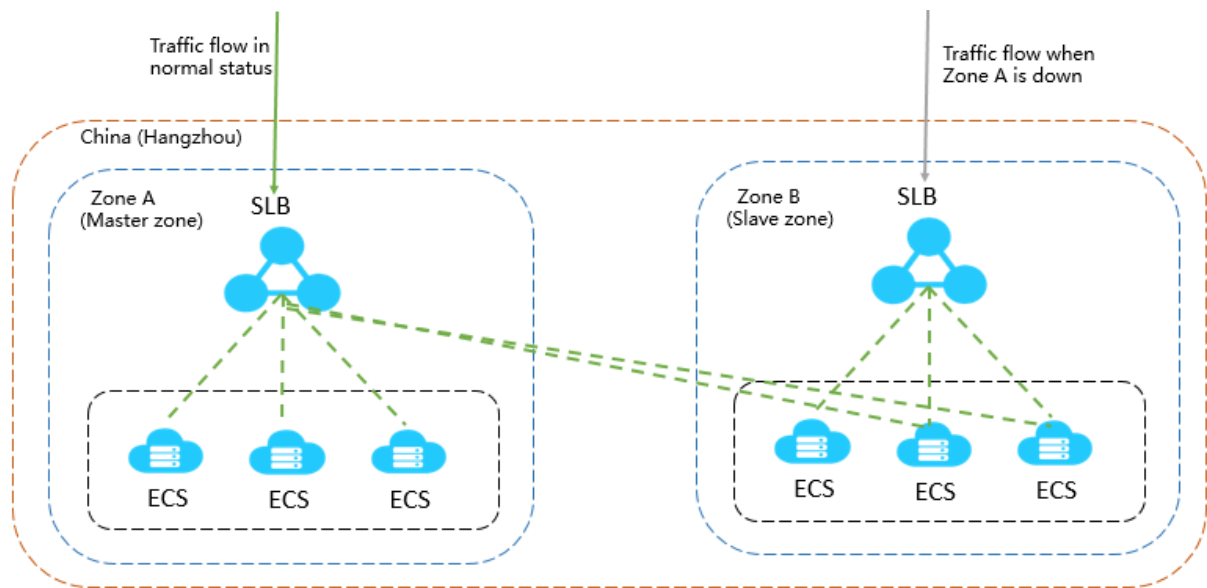
Achieve better disaster tolerance in multiple zones

To provide more stable and reliable services, Server Load Balancer is deployed in multiple zones in most regions to achieve same-region disaster tolerance. If the primary zone becomes unavailable, Server Load Balancer in the standby zone takes over the load balancing service in 30 seconds. Once the primary zone becomes available, Server Load Balancer automatically switches back to the primary zone.

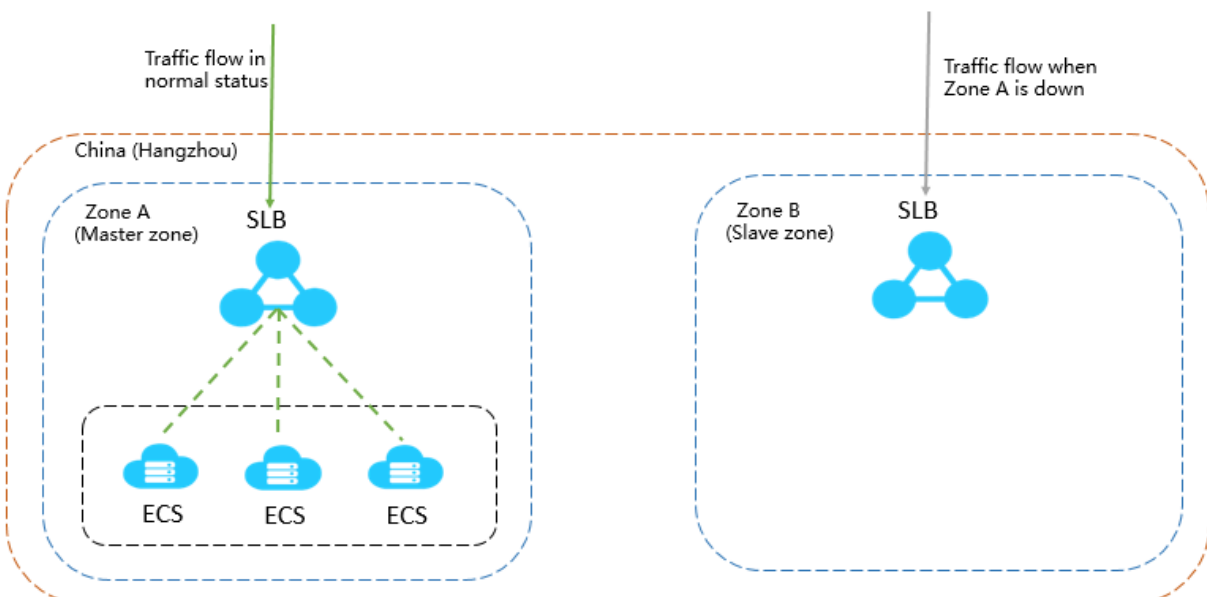
We recommend that you create a Server Load Balancer instance in a region with multiple zones deployed. Additionally, you also need to consider the deployment of backend servers. We recommend that you add at least one backend server in each zone to achieve the highest efficiency.

As shown in the following figure, ECS instances in different zones are added to an SLB instance. In normal situation, Server Load Balancer will distribute incoming traffic to the ECS instances in the primary zone (Zone A). Once the primary zone is unavailable, the incoming traffic will be

distributed to the ECS instances in standby zone. This avoids service interruption caused by the failure of a single zone, and also reduces latency.



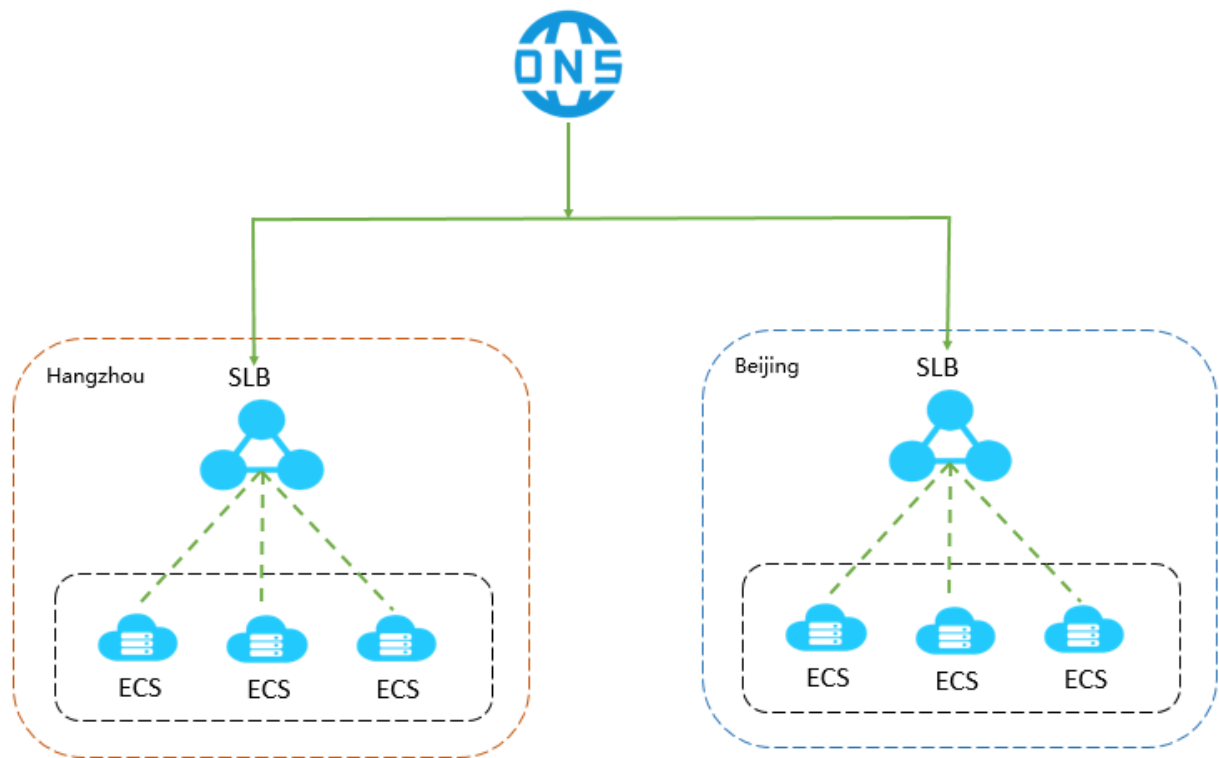
However, if you deploy all ECS instances in the primary zone and have no ECS instances deployed in the standby zone, your service may be interrupted when the primary zone is unavailable, because no ECS instances are available to handle the distributed requests in the standby zone. This deployment mode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



Achieve better disaster tolerance across regions

You can deploy Server Load Balancer instances in different regions, and attach ECS instances of different zones in the corresponding regions. The upper layer uses Alibaba Cloud DNS as

intelligent DNS, and resolves domain names to service addresses of Server Load Balancer instances in different regions, thus implementing global load balancing. When the system becomes unavailable in a region, you can temporarily stop DNS so that no user access is affected



6 Terms

Term	Description
Server Load Balancer	Alibaba Cloud Server Load Balancer (SLB) is a traffic distribution control service that distributes the incoming traffic among multiple Elastic Compute Service (ECS) instances according to the configured forwarding rules.
Server Load Balancer instance	A Server Load Balancer instance is a running entity of the Server Load Balancer service. To use Server Load Balancer, you must first create a Server Load Balancer instance.
Endpoint	An IP address allocated to an SLB 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 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 servers	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, Server Load Balancer will automatically route traffic to the standby server.