

Cloud Load Balance Product Introduction



Copyright Notice

©2013–2026 Tencent Cloud. All rights reserved.

The complete copyright of this document, including all text, data, images, and other content, is solely and exclusively owned by Tencent Cloud Computing (Beijing) Co., Ltd. ("Tencent Cloud"); Without prior explicit written permission from Tencent Cloud, no entity shall reproduce, modify, use, plagiarize, or disseminate the entire or partial content of this document in any form. Such actions constitute an infringement of Tencent Cloud's copyright, and Tencent Cloud will take legal measures to pursue liability under the applicable laws.

Trademark Notice



This trademark and its related service trademarks are owned by Tencent Cloud Computing (Beijing) Co., Ltd. and its affiliated companies ("Tencent Cloud"). The trademarks of third parties mentioned in this document are the property of their respective owners under the applicable laws. Without the written permission of Tencent Cloud and the relevant trademark rights owners, no entity shall use, reproduce, modify, disseminate, or copy the trademarks as mentioned above in any way. Any such actions will constitute an infringement of Tencent Cloud's and the relevant owners' trademark rights, and Tencent Cloud will take legal measures to pursue liability under the applicable laws.

Service Notice

This document provides an overview of the as-is details of Tencent Cloud's products and services in their entirety or part. The descriptions of certain products and services may be subject to adjustments from time to time.

The commercial contract concluded by you and Tencent Cloud will provide the specific types of Tencent Cloud products and services you purchase and the service standards. Unless otherwise agreed upon by both parties, Tencent Cloud does not make any explicit or implied commitments or warranties regarding the content of this document.

Contact Us

We are committed to providing personalized pre-sales consultation and technical after-sale support. Don't hesitate to contact us at 4009100100 or 95716 for any inquiries or concerns.

Contents

Product Introduction

Overview

Strengths

Use Cases

Technical Principles

Instance Specifications Comparison

Use Limits

Product Introduction

Overview

Last updated: 2024-07-10 09:45:02

You can learn about the concept, components, and working principles of Cloud Load Balance through the following video.

[Watch video](#)

What is Cloud Load Balancer?

Cloud Load Balancer (CLB) is a service that distributes traffic to multiple real servers, enhancing the external service capabilities of application systems and increasing their availability by eliminating single points of failure.

CLB virtualizes multiple real servers **in the same region** into a high-performance and high-availability application service pool by setting a virtual IP address (VIP) and then distributes the network requests from clients to the pool in the manner specified by the applications. CLB checks the health of real servers in the pool and automatically isolates unhealthy ones, thus resolving single points of failure issues and improving the overall service capabilities of the applications.

Tencent CLB, featuring self-service management, self-service fault repair, and defense against network attacks, is applicable for application scenarios such as enterprises, communities, e-commerce, and games.

Components

A working CLB group usually consists of the following components:

- **CLB Instance:** A CLB instance is an operational load balancing service used for traffic distribution.
- **Listener:** A component that inspects client requests and forwards them to the backend servers.
- **Backend Servers:** A set of server instances in the backend, designed to receive requests from the frontend.

The access requests from outside the CLB are distributed by CLB instances to the real servers for processing based on policies and forwarding rules.

Definitions

Term	Full Name	Note
------	-----------	------

CLB	Cloud Load Balancer	Tencent Cloud offers a load balancing service. Users can use CNAME or A records for external access. Only public network load balancing can choose the CNAME method.
CLB listener	Load Balance Listener	Includes listening port, load balancing policy and health check configuration, each of them corresponds to an application service in the backend.
Real Server	Real Server	CLB forwards requests to a group of real server instances for processing based on the rules specified by users.

How does it work

Basic Working Principle

A CLB instance receives traffic from clients and routes requests to real servers in one or more availability zones for processing.

The CLB service is primarily provided by the Cloud Load Balancer listener. The listener monitors requests on the CLB instance and distributes them to real servers based on policies. By configuring the **Client – CLB** and **CLB – Real server** forwarding protocols and protocol ports, CLB can directly forward requests to the real servers.

We recommend that you configure real server instances of CLB across multiple availability zones. This enables CLB to route traffic to healthy instances in other availability zones when one availability zone becomes unavailable, thereby avoiding service interruptions caused by availability zone failures.

Request Routing

The client requests to access the service through the domain name. Before the request is sent to the Cloud Load Balancer, the DNS server resolves the CLB domain name and returns the requested CLB IPs to the client. When the CLB listener receives a request, it uses different load balancing algorithms to distribute the request to the real servers. Tencent Cloud currently supports multiple balancing algorithms including weighted round robin and ip_hash weighted minimum connections.

Backend CVM Monitoring

A CLB instance can also monitor the running status of real server instances to ensure that traffic is only routed to the normally running instances. The CLB instance will stop routing traffic to an abnormal instance and reroute traffic to it when it is running normally again.

Additional Services

CLB works with the following services to improve application availability and scalability:

- **Cloud Virtual Machine** instance: A virtual server where the application runs on the cloud.
- **Auto Scaling**: Controls the instance quantity elastically. If CLB is enabled in auto scaling, the scaled instance is automatically added to the CLB group, and the terminated instance is automatically removed from the CLB group.
- **Tencent Cloud Observability Platform**: Assists in monitoring the operation of CLB instances and all backend instances, and performs necessary actions.
- **Domain Registration** and **DNS Resolution DNSPod**: Convert your custom domain name (such as `www.example.com`) to the IP address used for network communication (such as `192.0.2.1`), thus facilitating the routing of requests to Cloud Load Balancer instances.

Strengths

Last updated: 2026-06-02 17:29:34

This document describes the strengths of CLB.

High Performance

One single CLB cluster (not one instance) can support hundreds of millions of concurrent connections and process millions of data packets per second. This enables you to easily sustain ecommerce websites, social networking platforms, and gaming businesses with over 10 million daily page views.

High Availability

CLB adopts a cluster-based deployment mode to deliver an availability up to 99.95%. In the extreme case where only one CLB physical server is available, it can still support tens of millions of concurrent connections. The cluster system will remove faulty instances in time and keep the healthy ones to ensure that the real server continues to operate properly.

High Elasticity

The CLB cluster scales the service capabilities of the application system elastically according to the business load, and automatically creates and releases CVM instances through the dynamic scaling group of Auto Scaling. These features, in conjunction with a dynamic monitoring system and a billing system that is accurate to the second, eliminate your need to manually intervene or estimate resource requirements, helping you reasonably allocate computing resources and prevent resource waste.

High Security and Stability

It supports Anti-DDoS Basic by default and seamlessly integrates with security products such as [WAF](#) and [Anti-DDoS Pro](#). It also provides end-to-end HTTPS encryption, supports efficient and secure encryption protocols like TLS 1.3, and ensures business security.

Low Costs

With Tencent Cloud CLB, you no longer need to invest in additional Cloud Load Balance hardware or spend excessive time on tedious maintenance tasks, saving you up to 99% in hardware and labor costs. CLB supports various billing options, allowing you to choose the most suitable one based on your needs.

Use Cases

Last updated: 2023-09-04 20:55:05

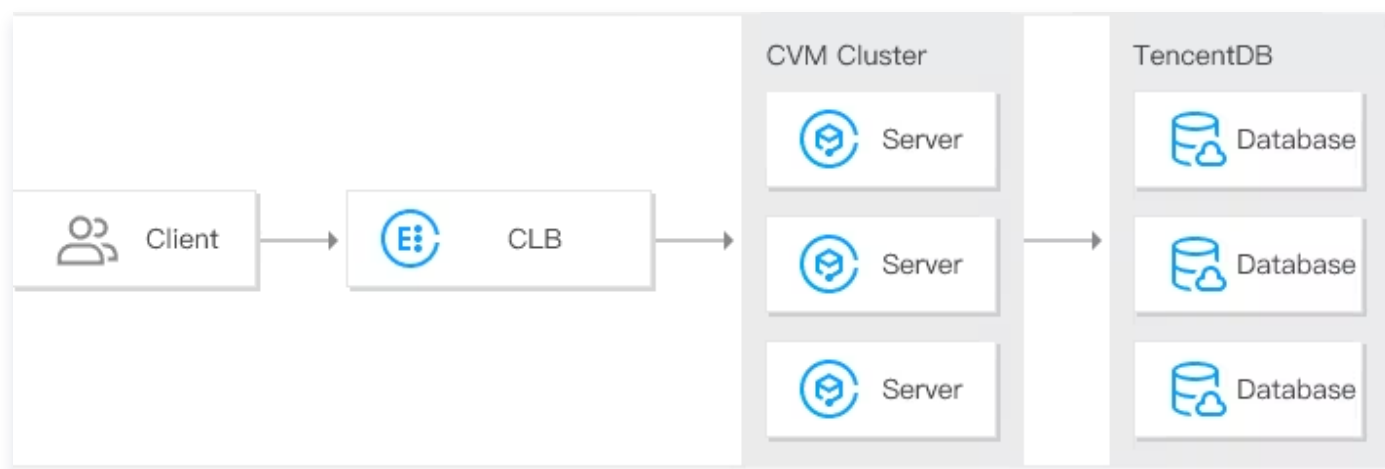
CLB is mainly suitable for the following scenarios:

- Traffic distribution. CLB can distribute the traffic of business with a large number of access requests to multiple real servers.
- Elimination of single point of failure. When some CVM instances become unavailable, CLB automatically blocks them to ensure the normal operation of the application system.
- Horizontal scalability. You can scale out the service capability of web servers and app servers as needed.
- Global CLB. With [DNSPod](#), CLB supports global and multi-regional load balancing for remote disaster recovery.

Traffic Distribution and Elimination of Single Point of Failure

You can use CLB to distribute business traffic to multiple real servers.

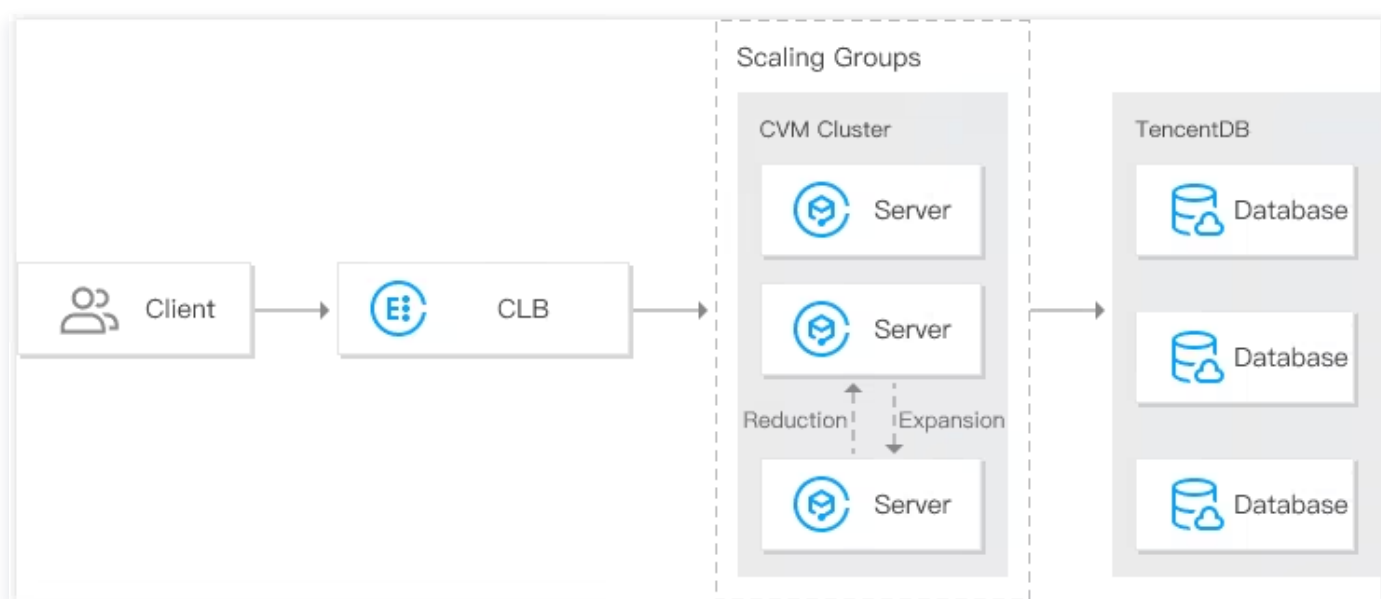
- A business client accesses CLB.
- Multiple real servers form a high-performance and high-availability service pool to which CLB forwards the business traffic.
- When one or more real servers become unavailable, CLB automatically blocks them and distribute the requests to healthy CVM instances, ensuring the operation of the application system.
- If your business is deployed in multiple availability zones, it is recommended that you bind a CLB instance to CVM instances in multiple availability zones at the same time to ensure multi-availability zone disaster recovery at the real server layer.
- The session persistence feature ensures that requests from a client are sent to the same real server, which improves the access efficiency.



Horizontal Scalability

With [Auto Scaling \(AS\)](#), CLB automatically creates and releases CVM instances based on your business needs.

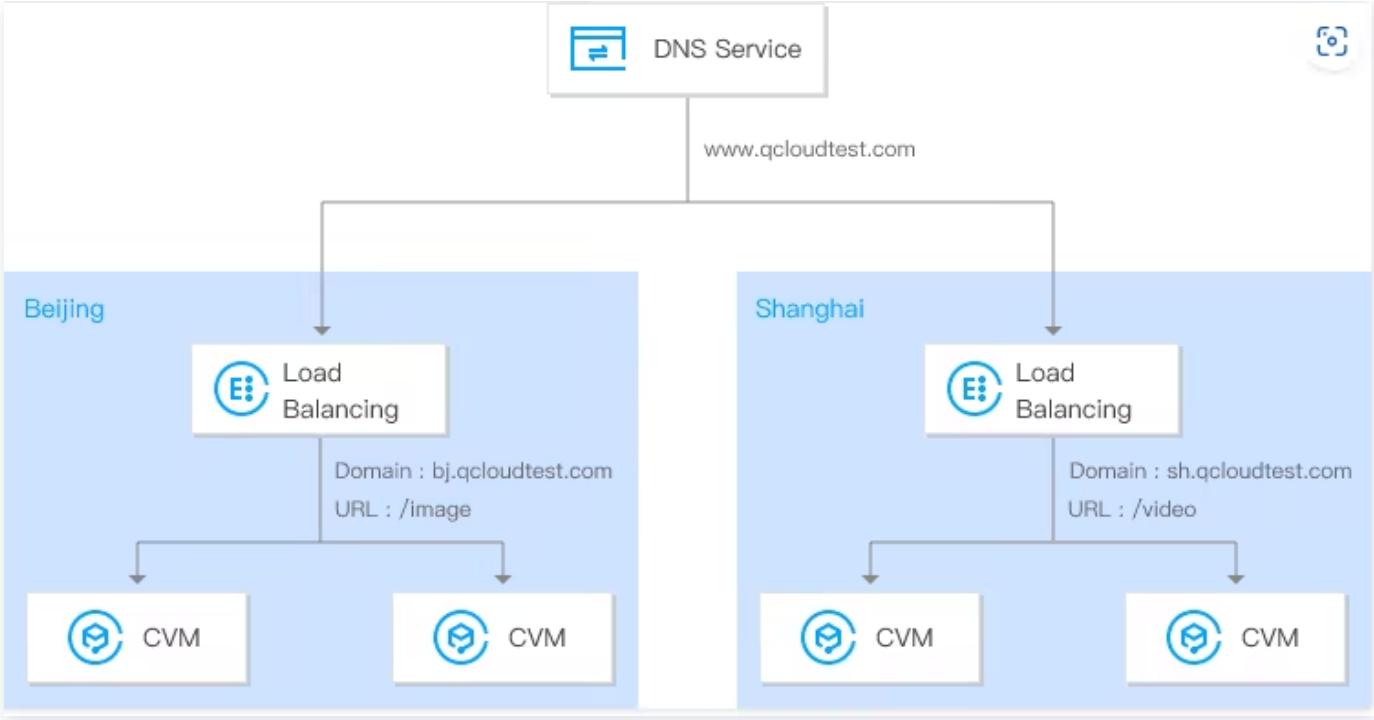
- You can configure auto scaling policies to manage the number of CVM instances, deploy the instance environment, and ensure the smooth operation of your business. During peak demand, CLB automatically increases the number of CVM instances to maintain optimal performance. When demand is low, it reduces the number of CVM instances to minimize costs.
- During major sales events in the ecommerce industry, such as "Singles' Day" and "6.18," web traffic may surge tenfold in an instant and last only for a few hours. Utilizing CLB and Auto Scaling can maximize IT cost savings.



Global Load Balancing

With [DNSPod](#), CLB supports global load balancing for you to resolve business traffic to CLB instances in various regions, ensuring multi-site active-active disaster recovery.

- You can deploy CLB instances in different regions and bind them to real servers in corresponding regions.
- You can use DNSPod to resolve domain names to CLB VIPs in various regions.
- Business traffic will be forwarded to multiple real servers in multiple regions via DNSPod and CLB, achieving global load balancing.
- When a region becomes unavailable, you can suspend resolution of the VIP of the CLB instance in that region to ensure that your business is not affected.



Technical Principles

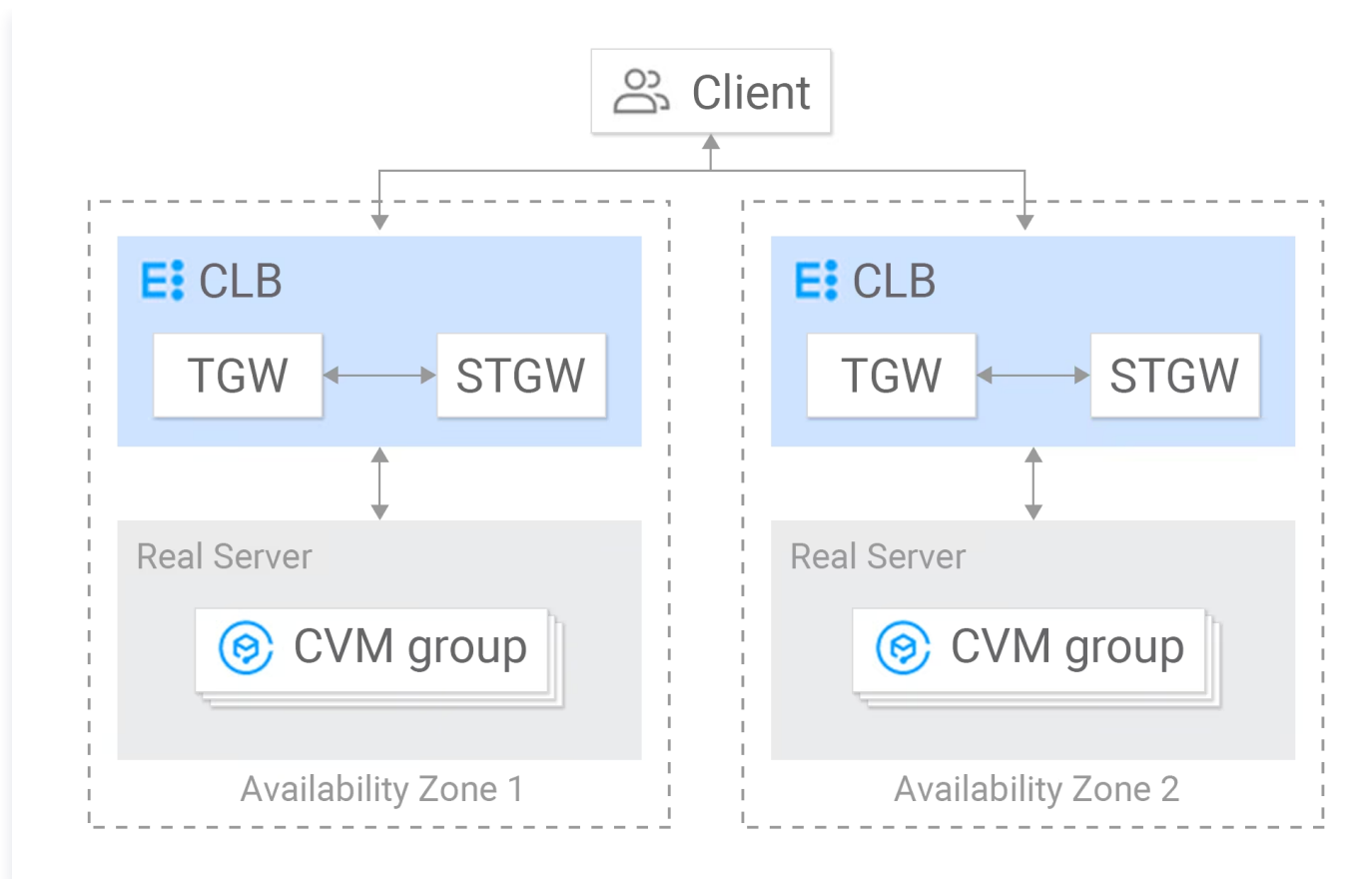
Last updated: 2023-09-04 20:55:18

Tencent Cloud Load Balancer (CLB) offers both Layer 4 (TCP, UDP, TCP SSL protocols) and Layer 7 (HTTP, HTTPS protocols) load balancing. By utilizing CLB, you can distribute business traffic across multiple backend servers, eliminating single points of failure and ensuring service availability. CLB itself employs a clustered deployment, enabling session synchronization, eliminating server bottlenecks, enhancing system redundancy, and guaranteeing service stability. Multiple data centers can be deployed within the same region to achieve disaster recovery within the same city.

Basic Infrastructure

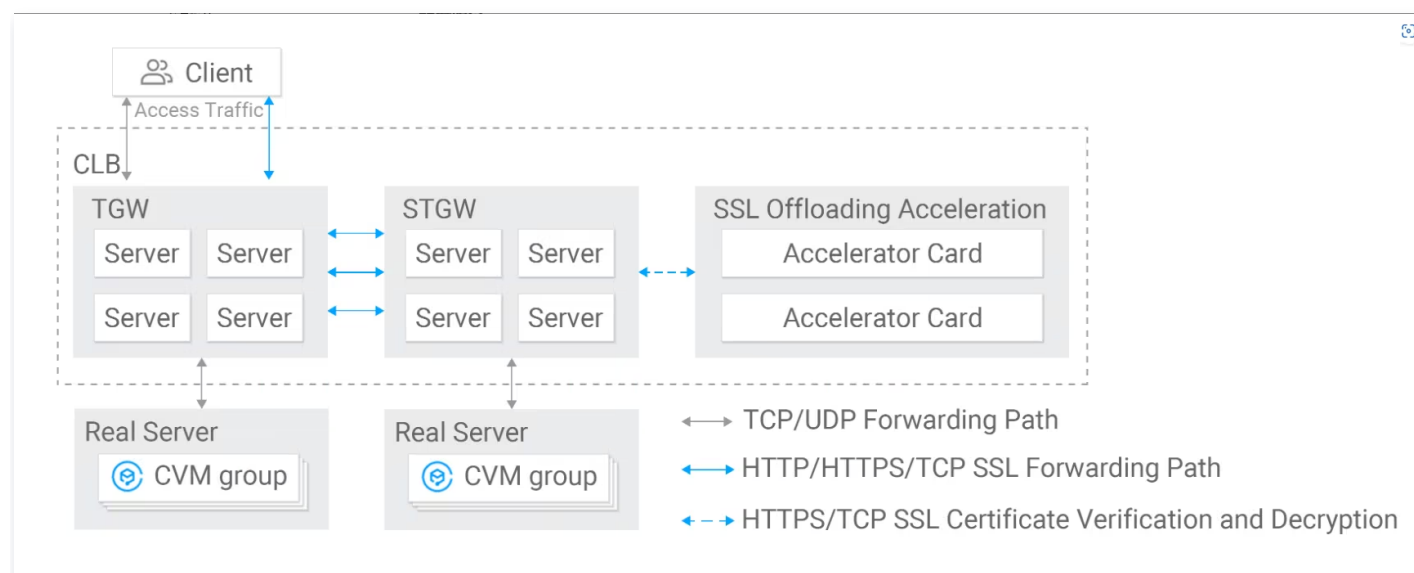
Currently, Tencent Cloud CLB provides Layer-4 and Layer-7 load balancing services:

- Layer 4 load balancing in Tencent Cloud Load Balancer is primarily implemented through Tencent's proprietary Unified Access Gateway (Tencent Gateway, TGW). TGW boasts high reliability, strong scalability, high performance, and robust attack resistance. It supports Data Plane Development Kit (DPDK) for high-performance forwarding, with a single cluster capable of handling billions of concurrent connections and tens of millions of PPS. Numerous Tencent internal services, including Tencent Games, Tencent Video, WeChat, and QQ, utilize TGW for access.
- Layer 7 load balancing in Tencent Cloud Load Balancer is primarily implemented through Secure Tencent Gateway (STGW). STGW is a large-scale, concurrent Layer 7 load balancing service developed by Tencent based on Nginx. It carries a significant amount of Layer 7 business traffic within Tencent, including Tencent News, Licitong, Tencent Games, WeChat, and more.



Forwarding path

Cloud Load Balancer is responsible for forwarding business traffic, while backend services handle the actual processing of requests. CLB communicates with backend CVMs through Tencent Cloud's private network. Both TGW and STGW are deployed on multiple servers, providing load balancing services through clustering. The forwarding path of CLB is illustrated in the following diagram:



1. TCP and UDP protocol:

- The TGW cluster handles the forwarding logic for TCP/UDP protocols.
- Upon arrival of business traffic, TGW forwards it to the backend CVM via Tencent Cloud's private network. The backend CVM's response packets are also sent back to the client through TGW.

2. TCP SSL protocol

- When handling TCP SSL protocol, the business traffic first passes through the TGW cluster and is then forwarded to the backend CVM by the STGW cluster.
- When establishing a new session, it is necessary to go through the accelerator card cluster for certificate verification and encryption/decryption operations as a prerequisite.
- Upon arrival of business traffic, it sequentially passes through TGW, STGW, and backend CVM within Tencent Cloud's internal network. The response packets are then sent back to the client in reverse order.

3. HTTP and HTTPS protocols

- When handling HTTP/HTTPS protocols, business traffic first passes through the TGW cluster and is then forwarded to the backend CVM by STGW, which recognizes the HTTP protocol.
- When establishing a new HTTPS session, it is necessary to go through the accelerator card cluster for certificate validation and encryption/decryption operations. This process converts HTTPS to HTTP protocol before forwarding it to the backend CVM.
- Upon arrival of business traffic, it sequentially passes through TGW, STGW, and backend CVM within Tencent Cloud's internal network. The response packets are then sent back to the client in reverse order.

Instance Specifications Comparison

Last updated: 2023-09-04 21:00:12

Tencent Cloud Load Balancer (CLB) supports shared instances and LCU-supported instances.

Comparison Item	Shared	LCU-supported	
Max throughput	Concurrent connections per minute: 50,000; new connections per second: 5,000; queries per second: 5,000.	Standard	Concurrent connections per minute: 100,000; new connections per second: 10,000; QPS: 10,000; Bandwidth cap: 2 Gbps.
		Advanced I	Concurrent connections per minute: 200,000; new connections per second: 20,000; QPS: 20,000; Bandwidth cap: 4 Gbps.
		Advanced II	Concurrent connections per minute: 500,000; new connections per second: 50,000; QPS: 30,000; Bandwidth cap: 6 Gbps.
		Super Large 1	Concurrent connections per minute: 1,000,000; new connections per second: 100,000; QPS: 50,000; Bandwidth cap: 10 Gbps.
Throttling capability	- Starting from 12:00:00 (UTC+8) July 25, 2023, the shared CLB instances created by new users will be rate-limited according to the instance specifications. - For users who have been using shared instances before 12:00:00 (UTC+8) July 25, 2023, Tencent Cloud will implement rate	Bandwidth limited by specifications	

	limiting based on instance specifications for existing shared CLB instances starting from 12:00:00 (UTC+8) August 25, 2023. Note: Before implementing rate limiting based on specifications, instances exceeding the max throughput limit will share cluster resources with other instances, which may lead to performance preemption.	
Billing Item	CLB instance fee, network fee, cross-region binding fee. In the description: • The network fee is billed only for public network CLB instances. • The cross-region binding fee is billed only after the cross-region binding feature is enabled.	CLB instance fee, network fee, LCU fee, and cross-region binding fee. In the description: • The network fee is billed only for public network CLB instances. • The cross-region binding fee is billed only after the cross-region binding feature is enabled.
Capabilities	Same	Same

Use Limits

Last updated: 2023-09-04 21:00:27

This document describes the use limits of CLB.

General Restrictions

There are some general restrictions when using Tencent Cloud CLB, and different types of CLB instances have specific restrictions for their respective types. For more information on CLB instance types, please refer to [Instance Types](#).

Instance type.	Resource	Default Limitations
General restrictions for all instances	Number of public instances that can be created by an account in a single region.	The number of public instances for regular IPs is limited to 100. For static single-line IP, the limit on the number of public instances is 3 for individual accounts and 15 for enterprise accounts.
	Number of private network instances that can be created by an account in a single region.	100
	The number of listeners that can be added to an instance.	50
	Selectable ports for listeners within an instance	The port number must be an integer between 1 and 65535.
CLB(formerly "application CLB")	In a CLB instance, the number of domain names and URL forwarding rules that can be configured for HTTP/HTTPS listeners.	50
	The number of servers that can be bound to a forwarding rule of a CLB instance.	100
	The number of backend ports that can correspond to a frontend port for a CLB instance.	Multiple ports

Classic CLB	The number of servers that can be bound to a listener in a Classic CLB instance.	100
	The number of backend ports that can be mapped to a frontend port for a classic CLB instance.	One port

CLB **will not actively unbind** from a CVM instance, and when a Cloud Virtual Machine enters an isolated state (such as a monthly/yearly subscription CVM entering the recycle bin or a pay-as-you-go CVM being overdue for more than two hours), it **will not unbind** from the CLB instance either.

Peak Bandwidth

The meaning of peak bandwidth varies by network billing modes as detailed below:

Billing Method	Difference	Description
Billed by fixed bandwidth(including hourly bandwidth subscription and monthly bandwidth)	Bandwidth peak serves as a committed metric. In case of bandwidth resource contention, the peak bandwidth is guaranteed and will not be restricted.	In a single region, the total bandwidth peak for all instances billed by fixed bandwidth (including hourly and monthly bandwidth billing) should not exceed 50 Gbps for both purchased and actual running instances. If you have a higher bandwidth requirement, please contact your account manager to request an adjustment.
Traffic-based billing	The peak bandwidth serves only as the maximum bandwidth cap and is not a committed performance indicator. During bandwidth resource contention, the peak bandwidth may be restricted.	In a single region, the total peak bandwidth of all instances billed by traffic does not exceed 5 Gbps. If your business requires guaranteed bandwidth or a higher peak bandwidth, please purchase public network bandwidth billed by fixed bandwidth.
Bandwidth Packages		In a single region, the total peak bandwidth of all instances billed by shared bandwidth packages should not exceed 50 Gbps. If you have a higher bandwidth requirement, please contact your account manager to apply for an adjustment.

