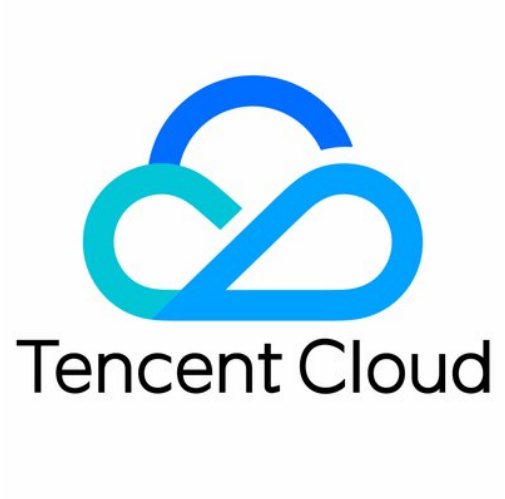
Cloud Load Balancer

Product Introduction

Product Documentation



Contents

Product Introduction
Overview

Last updated：2024-01-04 09:44:16

What is Cloud Load Balancer?

Tencent Cloud Load Balancer (CLB) is a service that distributes traffic to multiple real servers so as to elevate service capabilities of applications and eliminate single points of failure for a higher availability.
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:
Cloud Load Balancer: A CLB instance used for traffic distribution.
VIP (virtual IP): The IP address used for the CLB to provide service to clients.
Backend/Real Server: A set of server instances in the backend used for processing requests.
VPC/Classic Network: Overall network environment.
The access requests from outside the CLB are distributed by CLB instances to the real servers for processing based on policies and forwarding rules.

Glossary

Term	Full Name	Description
CLB	Cloud Load Balancer	A network load balancing service provided by Tencent Cloud, which works with real servers to provide users with TCP/UDP- and HTTP-based load balancing services.

CLB Listener	Cloud Load Balance Listener	CLB listeners allow you to configure listening ports, load balancing policies, and health check settings. Each listening item corresponds to an application service in the backend.
RS	Real Server	CLB forwards requests to a group of real server instances for processing based on the rules specified by users.
VIP	Virtual IP Address	The service address (currently an IP address) assigned by the system. Users can choose whether to disclose the service address to create CLB instances of the public or private network type. **Public network VIP** Regular IP: A general Border Gateway Protocol (BGP) IP address, which balances network quality and costs. Accelerated IP: An IP address accelerated by using Anycast Internet Acceleration (AIA), which makes public network access more stable, reliable, and low-latency. Static single-line IP: An IP address used to access the public network through a single ISP, which is low-cost and convenient for autonomous scheduling. **Private network VIP** VPC: An IP address within a VPC. Classic network: A private IP address on the classic network.

How it Works

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 works primarily by the load balancing listener. The listener monitors the requests on the CLB instance and distributes them to real servers based on policies. The CLB can forward the requests to real servers directly by configuring the **Client - CLB** and **CLB - Real server** forwarding protocols and protocol ports.

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 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.

Real server 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: Helps you monitor CLB instances and the running status of all real server instances and performs required operations.
Domains and 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 CLB instances.

Strengths

Last updated：2024-01-04 09:44:16

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

With the aid of BGP Anti-DDoS system, CLB is capable of defending against most network attacks (such as DDoS and web intrusion attacks) and cleansing attacking traffic in a matter of seconds, which greatly avoids the occurrence of blocked IPs and full occupancy of bandwidth. Its built-in synproxy anti-attack mechanism prevents the backend CVM instances from being crashed by attacks before the Anti-DDoS system takes effect, which makes data more secure and stable.
CLB strictly isolates the traffic of each tenant and provides active protection against DDoS attacks. Public network

CLB supports Anti-DDoS Basic by default. Moreover, CLB further supports Anti-DDoS Pro, Anti-DDoS Advanced, WAF, and other security products to safeguard your businesses.

Note:

If you have higher protection requirements, you can purchase Anti-DDoS Pro. It provides a DDoS protection capability of at least 30 Gbps, and the maximum protection capability is dynamically adjusted according to the actual network conditions in each region.

To protect the application layer, you can purchase WAF. It protects web security at the application layer against web vulnerability attacks, malicious crawlers, and CC attacks.

Low Costs

CLB eliminates your need to invest in additional load balancing hardware and time devoted to tedious Ops work, saving you up to 99% of hardware and labor costs. It supports multiple billing modes for your choice as needed.

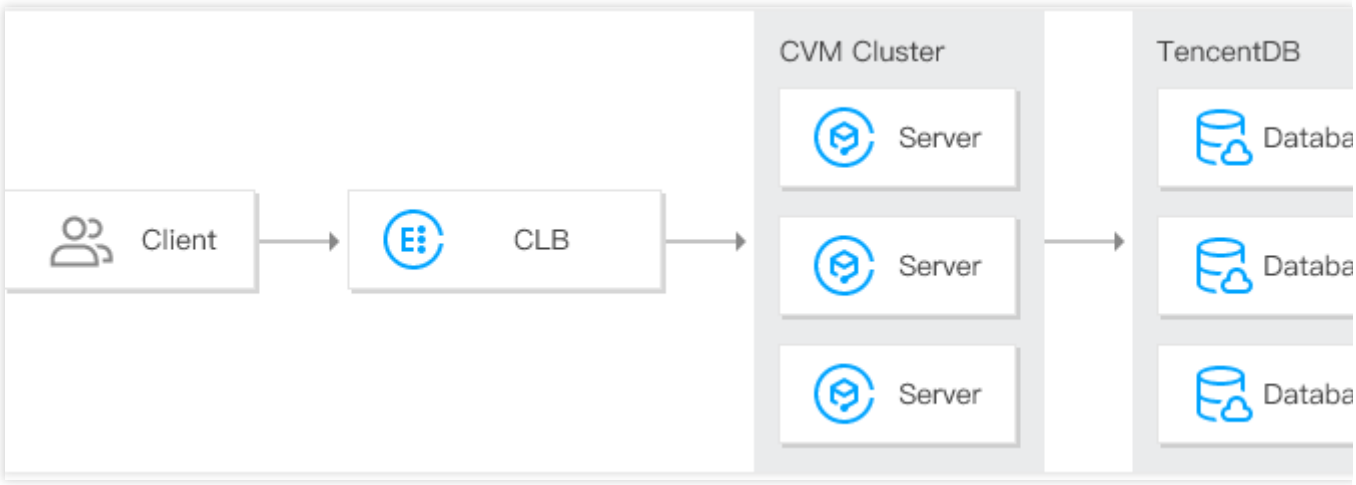
Use Cases

Last updated：2024-01-26 10:21:19

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 load balancing. With DNSPod, CLB supports global and multi-regional load balancing for remote disaster recovery.

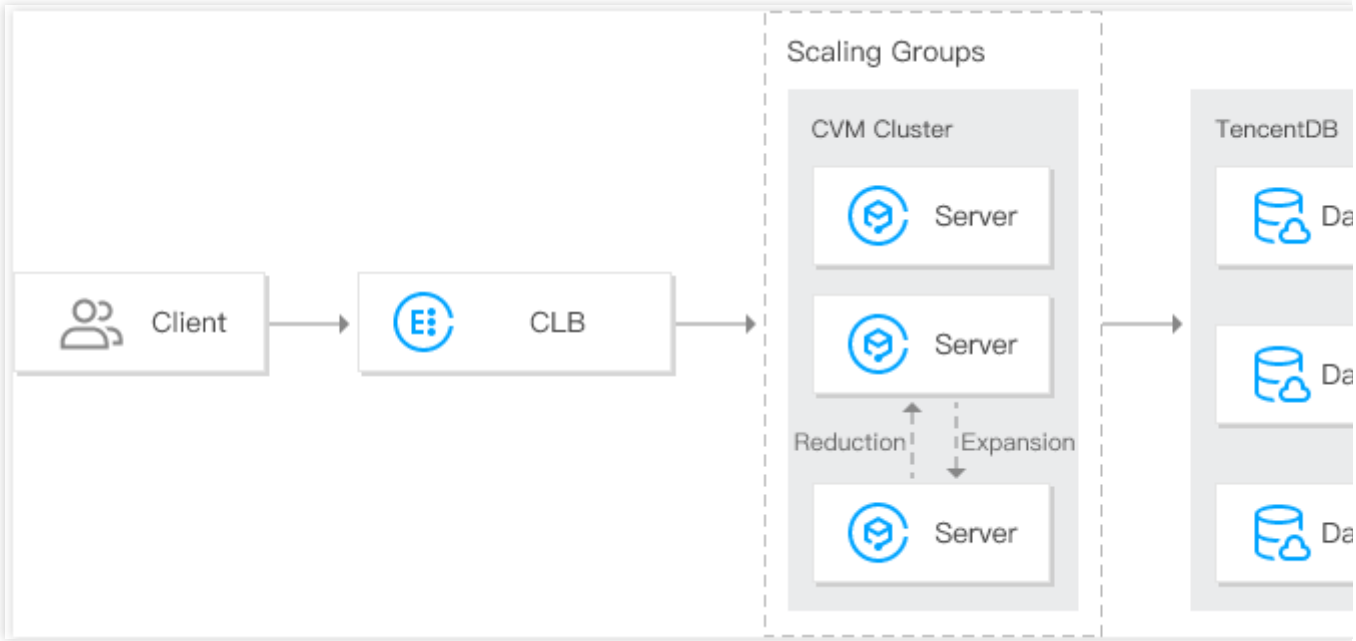
Traffic Distribution and Elimination of Single Point of Failure

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

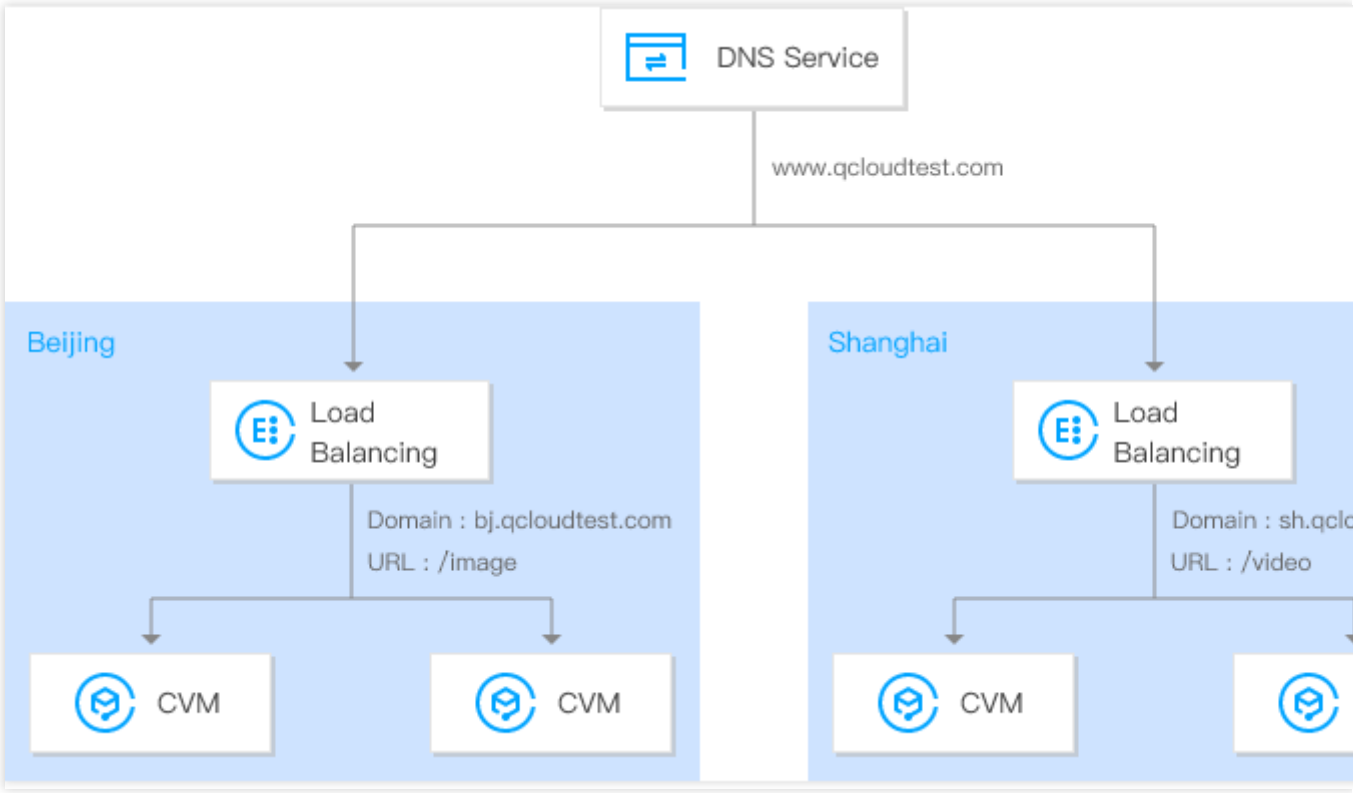
With Auto Scaling (AS), CLB automatically creates and releases CVM instances based on your business needs. You can configure auto scaling policies to manage the number of CVM instances, deploy the instance environment, and ensure the operation of your business. CLB can automatically add CVM instances when demands peak to keep high performance, while removing CVM instances when demands drop to reduce costs.
For example, during major sales campaigns in ecommerce such as Black Friday, web traffic may suddenly increase by 10 times and lasts only for a few hours. In this case, CLB and AS can be used to 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.



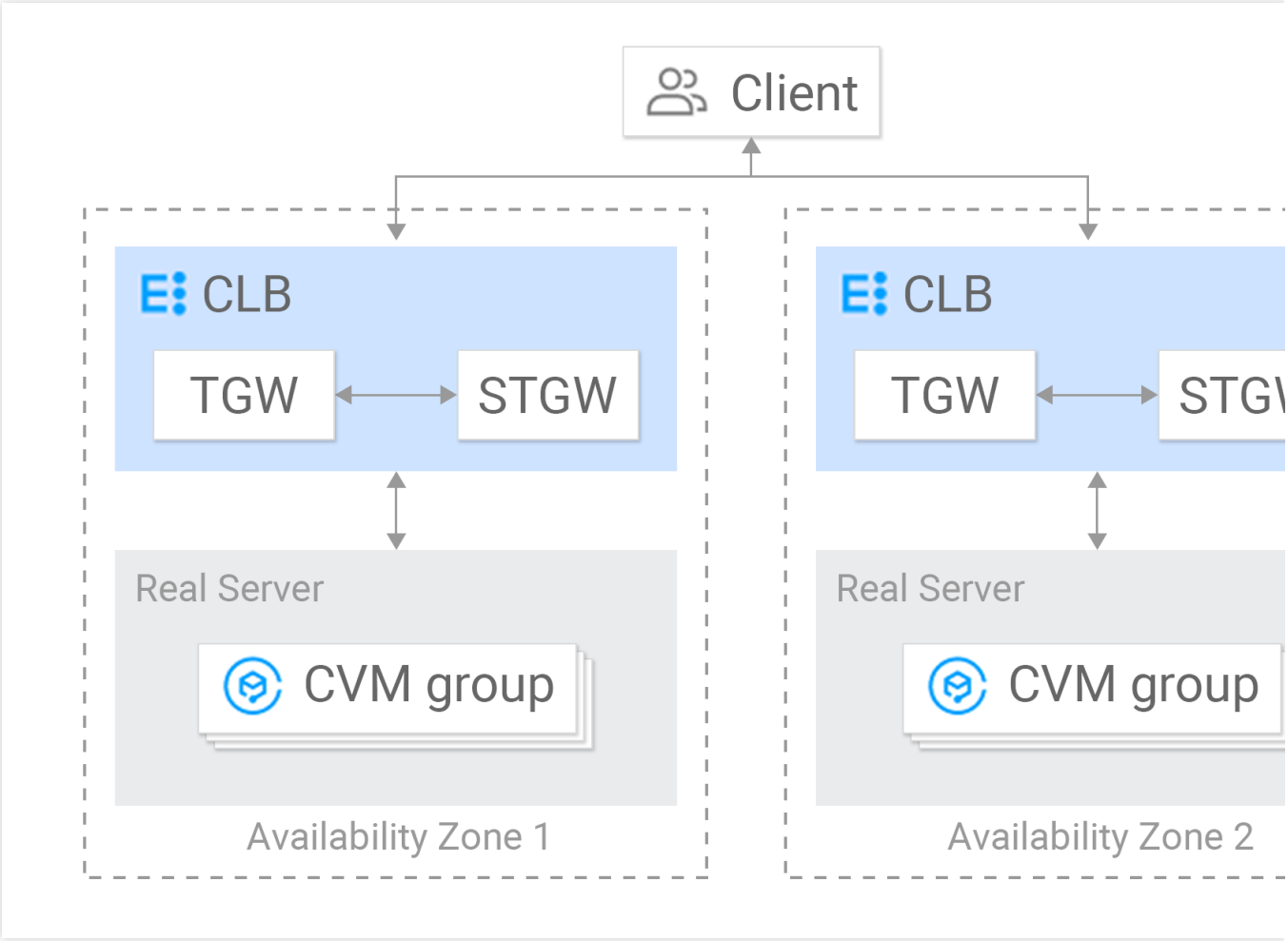
Principles

Last updated：2023-03-23 12:00:14

CLB provides Layer-4 (TCP, UDP, and TCP SSL protocols) and Layer-7 (HTTP and HTTPS protocols) load balancing. You can use CLB to distribute business traffic to multiple real servers to eliminate single point of failure and guarantee business availability. CLB adopts cluster deployment to achieve session synchronization, eliminating server's single point of failure and improving system redundancy to ensure service stability. CLB can be deployed in multiple data centers in the same region to implement intra-city disaster recovery.

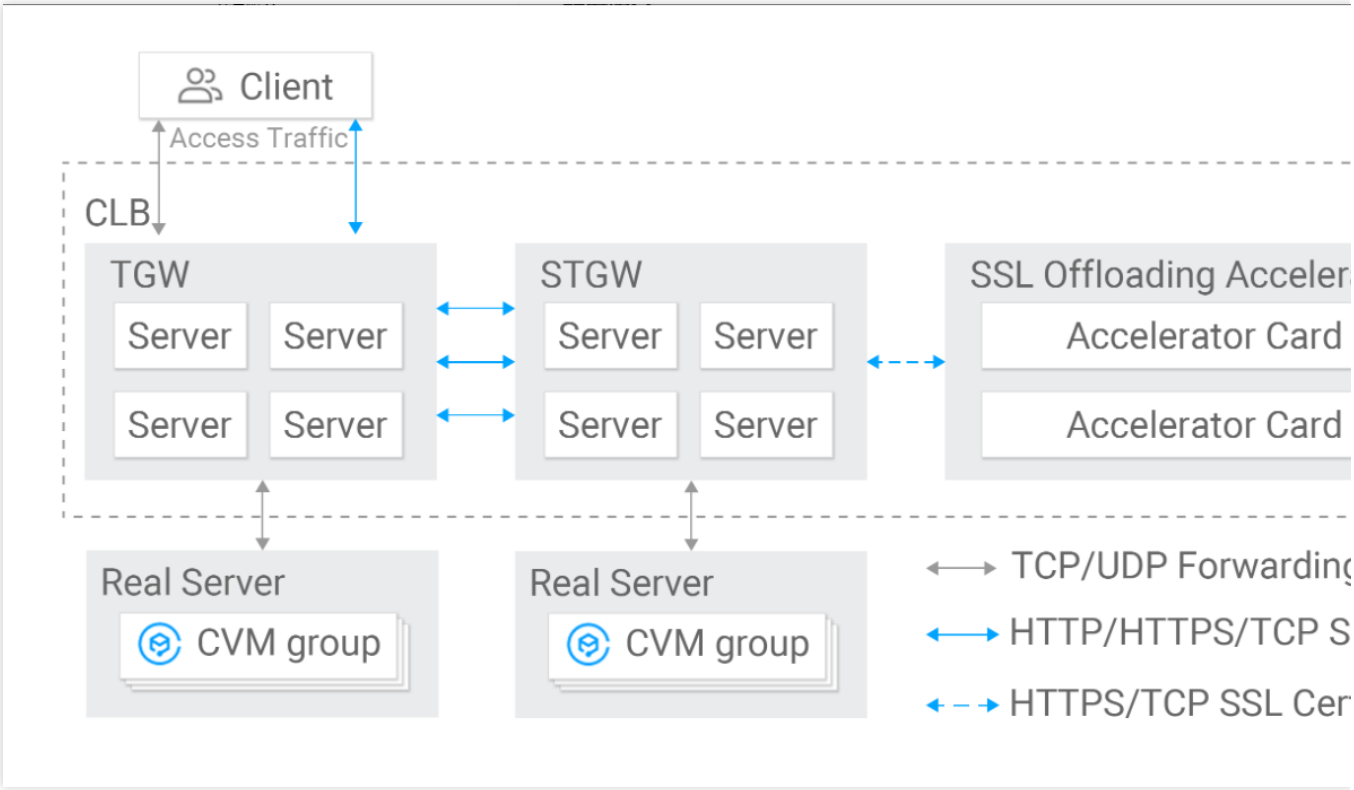
Infrastructure

Currently, Tencent Cloud CLB provides Layer-4 and Layer-7 load balancing services:
At Layer-4, load balancing is implemented based on the unified Tencent Gateway (TGW). TGW has features such as high availability, high scalability, high performance, and strong anti-attack capability. It supports high-performance forwarding based on Data Plane Development Kit (DPDK). With TGW, a single cluster can support hundreds of millions of concurrent requests and tens of millions of packets per second (PPS). Many Tencent businesses, such as Tencent Games, Tencent Video, use TGW for service access.
At Layer-7, load balancing is implemented based on Secure Tencent Gateway (STGW). It is a load balancing service developed by Tencent based on Nginx that supports large-scale concurrence. It carries a large amount of Tencent's Layer-7 business traffic, such as Tencent News, Tencent Games.



Forwarding Path

CLB forwards business traffic and real servers process business requests. CLB communicates with real servers via Tencent Cloud private network. Both TGW and STGW are deployed on multiple servers, and provide load balancing services through clusters. The forwarding path of CLB is as shown below:



1. TCP and UDP protocol:

The forwarding logic of TCP/UDP protocol is processed by TGW cluster.

After receiving the business traffic, TGW forwards it to real servers via Tencent Cloud's private network. The return packets from real servers are also returned to the client via TGW.

2. TCP SSL protocol

When TCP SSL protocol is processed, business traffic passes through the TGW cluster and then STGW cluster, which forwards the traffic to real servers.

Before a new session is established, it must pass through the accelerator card cluster for certificate verification, encryption, decryption and other operations.

When business traffic arrives, it passes through TGW, STGW, and real servers in sequence via Tencent Cloud's private network. The return packets are sent to the client in reverse sequence.

3. HTTP and HTTPS protocols

When HTTP or HTTPS protocol is processed, business traffic passes through the TGW cluster and then STGW cluster, which identifies the HTTP protocol and forwards the traffic to real servers.

Before a new HTTPS session is established, it must pass through the accelerator card cluster for certificate verification, encryption, decryption and other operations. HTTPS will be converted to HTTP protocol and then forwarded to real servers.

When business traffic arrives, it passes through TGW, STGW, and real servers in sequence via Tencent Cloud's private network. The return packets are sent to the client in reverse sequence.

Product Comparison
Instance Types Comparison

Last updated：2024-01-04 09:44:16

CLB offers two types of instances: CLB (formerly "application CLB") and Classic CLB.
CLB supports TCP/UDP/HTTP/HTTPS protocols to provide flexible forwarding capabilities based on domain names and URL paths.
Classic CLB does not support HTTP/HTTPS protocols on the private network but is easy to configure.
CLB has all features of Classic CLB. Given their product features and performance, we recommend that you use CLB.
For a detailed comparison, please see below:

Product Type	CLB		Classic CLB	
	Public network	Private network	Public network	Private network
Layer-7 forwarding (HTTP/HTTPS)	✓	✓	✓	×
Layer-4 forwarding (TCP/UDP)	✓	✓	✓	✓
Encrypted Layer-4 forwarding (TCP SSL)	✓	✓	×	×
HTTP/2 and WebSocket (Secure) support	✓	✓	✓	×
Load balancing policy	IP hash (Layer-7)Weighted round-robinWeighted least-connection scheduling	IP hash (Layer-7)Weighted round-robinWeighted least-connection scheduling	IP hash (Layer-7)Weighted round-robinWeighted least-connection scheduling	Weighted round-robin
Session persistence	✓	✓	✓	✓

Health check	✓	✓	✓	✓
Custom forwarding rule (domain name/URL)	✓	✓	×	×
SNI multi-certificate support	✓	✓	×	×
Forwarding to different real ports	✓	✓	×	×
Custom Layer-7 configuration	✓	✓	×	×
Layer-7 redirect (rewrite)	✓	✓	×	×
Cross-region binding support	✓	✓	×	×
Storing Layer-7 access logs in CLS	✓	✓	✓	×

Note:

CLB instance: a CLB instance supports enabling or disabling the HTTP/2 protocol. For more information, see Configuring an HTTPS Listener.

Classic CLB instance: HTTPS listeners created for Classic CLB before April 2018 do not support the HTTP/2 protocol. HTTPS listeners created after April 2018 support but cannot disable the HTTP/2 protocol.

Instance Specifications Comparison

Last updated：2024-01-04 09:44:16

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

Item	Shared	LCU-supported	
Specification Ceiling	Concurrent connections per minute: 50,000; New connections per second: 5,000; QPS: 5,000.	Standard	Concurrent connections per minute: 100,000; New connections per second: 10,000; QPS: 10,000; Bandwidth cap: 2 Gbps.
		Advanced 1	Concurrent connections per minute: 200,000; New connections per second: 20,000; QPS: 20,000; Bandwidth cap: 4 Gbps.
		Advanced 2	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.
		Super Large 2	Concurrent connections per minute: 2,000,000; New connections per second: 200,000; QPS: 100,000; Bandwidth cap: 20 Gbps.
		Super Large 3	Concurrent connections per minute: 5,000,000; New connections per second: 500,000; QPS: 200,000; Bandwidth cap: 40 Gbps.
		Super Large 4	Concurrent connections per minute: 10,000,000; New connections per second: 1,000,000; QPS: 300,000; Bandwidth cap: 60 Gbps.
Rate limiting	Starting from June 25, 2023, Tencent Cloud added specification-based rate limits on shared CLB instances	Apply rate limits based on the instance specifications	

	created by new users (users who don't have shared CLB instances under their accounts). For users who have shared CLB instances under their accounts before June 25, 2023, the rate limits take effect starting from 12:00:00, August 25, 2023 (UTC+8). **Note:** When the existing shared CLBs require excess performance, they take extra resources from the shared cluster, which may cause performance preemption.	
Billable Item	CLB instance fee, network fee, and cross-region binding fee. Description: The network fee is billed only on 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. 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.

Use Limits

Last updated：2022-05-23 10:08:31

This document describes the use limits of CLB.

General Restrictions

The use of Tencent Cloud CLB has certain restrictions, and different types of CLB instances have their own use limits. For more information on CLB instance types, see Instance Types.

Instance Type	Resource	Default Restriction
General restrictions for all instances	Number of public network CLB instances that can be created in a single region	100 general IP-based instances 3 static IP-based instances under one individual account or 15 under one organizational account.
	Number of private network instances that can be created under one account in a single region	100
	Number of listeners that can be added to an instance	50
	Ports that can be selected by a listener in an instance	An integer between 1 and 65535
CLB(formerly "application CLB")	Number of domain name and URL forwarding rules that can be configured for an HTTP/HTTPS listener in a CLB instance	50
	Number of servers that can be bound to a forwarding rule in a CLB instance	100
	Number of backend ports that can correspond to a frontend port of a CLB instance	Multiple ports
Classic CLB	Number of servers that can be bound to a listener in a classic CLB instance	100
	Number of backend ports that can correspond to a frontend port of a classic CLB instance	1 port

A CLB instance **will not unbind itself** from the CVM instance. After a CVM instance becomes isolated (pay-as-you-go CVM instance has been in arrears for more than 2 hours), it **will not unbind itself** from the CLB instance either.

Peak Bandwidth

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

Billing Mode	Difference	Description
Bill-by-traffic	The peak bandwidth is only regarded as the maximum peak bandwidth, and not as the committed bandwidth. When bandwidth resources are contested, the peak bandwidth may be limited.	The sum of peak bandwidth of all the running bill-by-traffic instances cannot exceed 5 Gbps in one region. If your application requires a guaranteed or higher bandwidth, choose bill-by-bandwidth.
Bandwidth packages		The sum of peak bandwidth of all the running instances that are billed by shared bandwidth package cannot exceed 50 Gbps in one region. If you require a higher bandwidth, contact your sales rep.