Cloud Load Balancer

Product Introduction

Product Documentation
Copyright Notice

©2013-2019 Tencent Cloud. All rights reserved.

Copyright in this document is exclusively owned by Tencent Cloud. You must not reproduce, modify, copy or distribute in any way, in whole or in part, the contents of this document without Tencent Cloud's the prior written consent.

Trademark Notice

All trademarks associated with Tencent Cloud and its services are owned by Tencent Cloud Computing (Beijing) Company Limited and its affiliated companies. Trademarks of third parties referred to in this document are owned by their respective proprietors.

Service Statement

This document is intended to provide users with general information about Tencent Cloud's products and services only and does not form part of Tencent Cloud's terms and conditions. Tencent Cloud's products or services are subject to change. Specific products and services and the standards applicable to them are exclusively provided for in Tencent Cloud's applicable terms and conditions.
Contents

Product Introduction
  Overview
  Strengths
  Use Cases
  Principles
  Use Limits
  Instance Types
What is Cloud Load Balancer?

Cloud Load Balancer (CLB) is a service that distributes traffic to multiple CVMs. CLB expands the external service capabilities of the application system through traffic distribution and improves the availability of the application system by eliminating single-point failures. For information on CVM, see CVM Overview.

CLB virtualizes multiple CVM instances 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 application.

CLB checks the health of the instances 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 CVM instances in the backend used for processing requests.
- **VPC/Classic Network**: Overall network environment.

The access requests from outside the CLB are distributed to the backend CVMs based on policies and forwarding rules for processing through the CLB instance.

Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Full Name</th>
<th>Description</th>
</tr>
</thead>
</table>

©2013-2022 Tencent Cloud. All rights reserved.
A network load balancing service provided by Tencent Cloud, which can be combined with CVM to provide users with TCP/UDP- and HTTP-based load balancing services.

Includes listening port, load balancing policy and health check configuration, each of them corresponds to an application service in the backend.

A group of CVM instances that receive the requests forwarded by the CLB according to the rules set by the users.

The service address (currently the IP address) assigned by the system. Users can choose whether to disclose the service address to create public/private network-based load balancing services.

- Public network VIP
  - Regular IP: Common BGP IP, used to balance network quality and cost.
  - Accelerated IP: Accelerated by using Anycast to make public network access more stable, reliable and low-latency.
  - Static single-line IP: Accesses the public network through a single ISP, which is low-cost and convenient for autonomous scheduling.

- Private network VIP
  - VPC: IP within the VPC.
  - Classic network: Private IP of the classic network.

How it Works

Basic Working Principle

A CLB instance receives inbound traffic from clients and routes requests to backend CVMs 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 the backend CVMs by executing policies. The CLB can forward the requests to the backend CVMs directly by configuring the forwarding protocols and protocol ports on the two dimensions: Client - CLB and CLB - Backend CVM.

It is recommended that you configure the backend CVM instances across multiple availability zones for the CLB. This enables CLB to route traffic to CVM instances running normally in other availability zones when one availability zone is 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 backend CVMs. Tencent Cloud currently supports multiple balancing algorithms including weighted round robin and weighted minimum connections.

**Backend CVM Monitoring**

A load balancer can also monitor the running status of backend CVMs to ensure that traffic is only routed to the normally running CVMs. The load balancer will stop routing traffic to an abnormal CVM 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 number elastically. If a 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.
- **Cloud Monitor:** Helps you monitor the CLB and the running status of all backend CVM instances and perform required operations.
Strengths

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 instance 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, CC, 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

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 CVM instances.
- Elimination of single point of failure. When some CVM instances become unavailable, CLB can automatically block them to ensure the normal operation of the application system.
- Horizontal scalability. You can scale out the service capability of application systems as needed, which is suitable for web and app servers.
- Global load balancing. CLB supports global multi-region load balancing to ensure remote disaster recovery.

Traffic Distribution and Elimination of Single Point of Failure

You can use CLB to distribute business traffic to multiple CVM instances.

- Business client accesses CLB.
- Multiple CVM instances form a high-performance and high-availability service pool, to which CLB forwards the business traffic.
- When one or more CVM instances become unavailable, CLB can automatically block them and distribute the requests to normal 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 can forward requests from the same client to the same real server, improving access efficiency.
Horizontal Scalability

With Auto Scaling, CLB can automatically create and release 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
Combined with DNSPod, CLB enables you to resolve business traffic to the global load balancing in various regions, ensuring remote multi-site active-active and disaster recovery.

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

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, WeChat, and QQ, 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, Licaitong, Tencent Games, and WeChat.

### 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.
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.

<table>
<thead>
<tr>
<th>Instance Type</th>
<th>Resource</th>
<th>Default Restriction</th>
</tr>
</thead>
</table>
| 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:

<table>
<thead>
<tr>
<th>Billing Mode</th>
<th>Difference</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill-by-traffic</td>
<td>The peak bandwidth is only regarded as the <strong>maximum peak bandwidth</strong>, and not as the committed bandwidth. When bandwidth resources are contested, the peak bandwidth may be limited.</td>
<td>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.</td>
</tr>
<tr>
<td>Bandwidth packages</td>
<td></td>
<td>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.</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>Product Type</th>
<th>CLB</th>
<th>Classic CLB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Public network</td>
<td>Private network</td>
</tr>
<tr>
<td>Layer-7 forwarding (HTTP/HTTPS)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Layer-4 forwarding (TCP/UDP)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Encrypted Layer-4 forwarding (TCP SSL)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HTTP/2 and WebSocket (Secure) support</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Load balancing policy</th>
<th>IP hash (Layer-7) Weighted round-robin</th>
<th>IP hash (Layer-7) Weighted round-robin</th>
<th>IP hash (Layer-7) Weighted round-robin</th>
<th>Weighted round-robin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Weighted least-connection scheduling</td>
<td>Weighted least-connection scheduling</td>
<td>Weighted least-connection scheduling</td>
<td></td>
</tr>
<tr>
<td>Session persistence</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Feature</td>
<td>Classic CLB</td>
<td>Express CLB</td>
<td>Standard CLB</td>
<td>Premium CLB</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Health check</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Custom forwarding rule (domain name/URL)</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>SNI multi-certificate support</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Forwarding to different real ports</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Custom Layer-7 configuration</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Layer-7 redirect (rewrite)</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Cross-region binding support</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Storing Layer-7 access logs in CLS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
</tbody>
</table>

**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.