应用型负载均衡

产品简介

腾讯云

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或95716。

文档目录

产品简介
产品概述

最近更新时间：2026-07-07 09:56:11

什么是应用型负载均衡?

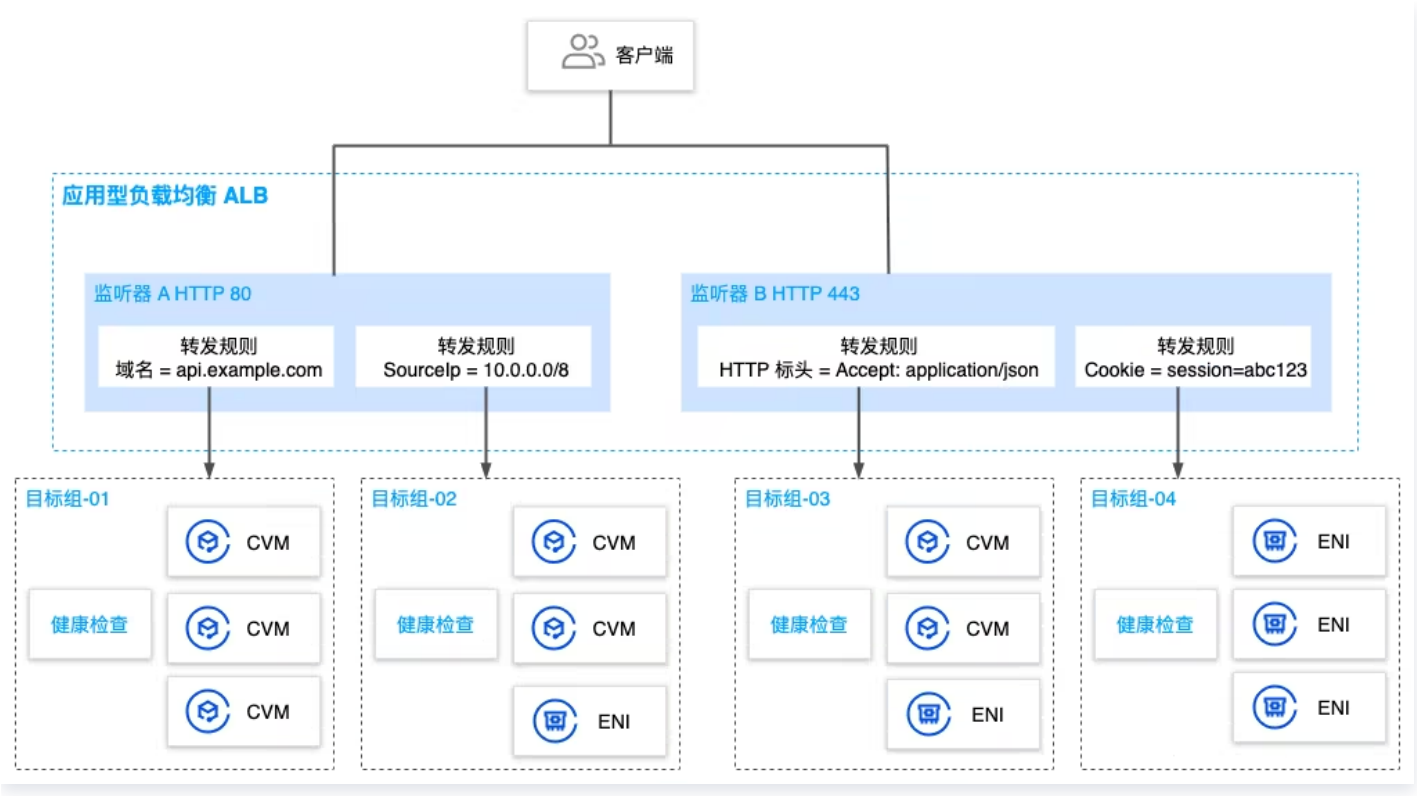
应用型负载均衡 ALB（Application Load Balancer）是面向七层（应用层）的负载均衡服务，专门针对 HTTP、HTTPS 和 QUIC 等应用协议进行流量分发和路由管理。它工作在开放系统互连（OSI）模型的第七层——应用层，能够解析 HTTP/HTTPS/QUIC 等协议内容，并根据请求特征（域名、路径、Header 等）将流量智能分发到不同的后端服务。
ALB 是面向现代应用架构的新一代负载均衡解决方案，特别适用于微服务、容器化、Serverless 等云原生场景，以及需要对 HTTP 流量进行精细化管理的业务系统。

> **说明：**
> 应用型负载均衡（ALB）当前处于公测阶段，暂未全面开放。使用前需申请开通资格。请提交 工单申请 。

核心能力

- **七层高级路由**：基于 HTTP 内容的高级路由规则，支持按域名、URL 路径、HTTP 标头、查询字符串、HTTP 请求方法等条件转发请求。
- **高可用与弹性伸缩**：支持跨可用区部署，消除单点故障；性能随业务流量自动弹性扩容，从容应对突发峰值。
- **安全防护**：支持安全组和 ACL，集成 Web 应用防火墙（WAF），提供 DDoS 防护能力和全链路 HTTPS 加密（支持 TLS 1.3）。
- **健康检查**：自动探测后端服务器运行状态，实时隔离不健康节点，有助于保障业务稳定性。
- **多协议支持**：全面支持 HTTP/1.1、HTTP/2、HTTP/3、QUIC、gRPC、WebSocket、SSE 等主流协议。
- **云原生集成**：深度集成容器服务 TKE、弹性伸缩 AS 等组件，可作为 Kubernetes Ingress Controller 使用。

组成部分



应用型负载均衡由以下核心组件构成：

负载均衡实例

面向七层提供负载均衡服务的实体，是整个系统的入口点和流量汇聚点。

- **网络类型**：分为公网和内网两种，公网实例通过弹性公网 IP（EIP）提供公网服务。
  - 公网实例绑定弹性公网 IP（EIP），直接面向互联网提供服务。
  - 内网实例仅在 VPC 内部提供服务，适合内部系统互联。
- **可用区部署**：建议至少选择两个不同可用区的子网，有助于在单可用区故障时维持服务可用。
- **容量规格**：单 VIP 性能如下，不支持调整。

指标名	最大规格
每秒请求数（QPS）	50万
新建连接数（CPS）	20万/秒
最大并发连接数	500万
出带宽	25Gbps
入带宽	25Gbps

说明：

> 公网 ALB 实例的公网带宽为 ALB 实例下所有 EIP 的带宽总和，默认单个 EIP 的公网带宽为200Mbps。一个 ALB 实例选择2个可用区，即两个 EIP，则该 ALB 实例的带宽上限为400Mbps。

监听器

监听器是 ALB 的最小业务单元，负责检查客户端连接请求并配置转发策略。
主要配置项包括：

配置项	说明
协议类型	根据业务需求选择合适的监听协议，如 HTTP / HTTPS。
监听端口	可选1 - 65535，同一个实例内监听器的端口不能重复。
转发规则	定义请求匹配条件和目标动作。
SSL 设置	证书选择、加密套件、协议版本等。

转发规则

转发规则定义了如何将不同特征的请求路由到不同的后端服务。
条件类型：

- 域名
- 路径
- HTTP 标头
- 查询字符串
- HTTP 请求方法
- Cookie
- SourceIP

动作类型：

- 转发
- 重定向
- 返回固定响应
- 重写
- 写入 Header
- 删除 Header

优先级机制：规则按优先级从高到低依次评估，首条命中的规则生效。未命中任何规则时，按照监听器默认规则进行转发。

目标组

目标组是一组后端服务器的逻辑集合，用于统一管理和调度流量。

- **目标类型**：CVM 实例、ENI 弹性网卡。
- **健康检查**：按目标组维度配置独立的健康检查策略。
- **负载方式**：加权轮询、加权最小连接数。

健康检查

健康检查机制用于实时监控后端可用性，使流量尽量分发至健康的后端

参数	说明	默认值
检查协议	TCP / HTTP / HTTPS	TCP
检查端口	健康检查使用的端口	业务端口
响应超时	等待响应的最大时间	2秒
检查间隔	两次检查的时间间隔	5秒
不健康阈值	判定异常的连续失败次数	3次
健康阈值	判定健康的连续成功次数	3次

名词解释

术语	英文	说明
应用型负载均衡	Application Load Balancer (ALB)	工作在七层（应用层）的负载均衡实体，负责接收并分发应用层流量。
监听器	Listener	配置协议、端口及转发规则的逻辑单元，负责处理客户端连接请求。
转发规则	Rule	定义请求匹配条件和对应动作的策略，实现精细化路由。
目标组	Target Group	一组后端服务器的逻辑集合，作为流量转发的目标。
后端服务器	Backend Server	实际处理业务请求的服务器实例，如 CVM、容器等。
健康检查	Health Check	自动探测后端服务器可用性的机制。
服务域名	ALB Domain Name	ALB 实例提供负载均衡服务的域名，对外暴露的统一入口，解析到多个 VIP。 **警告：**

		此域名不能直接访问，仅支持作为 CNAME 解析的目标地址。
虚拟 IP	Virtual IP (VIP)	负载均衡实例的服务地址，每个可用区一个 VIP。
ALCU	ALB Load Balancer Capacity Unit	性能容量单位，用于衡量 ALB 的资源消耗和计费。

工作原理

基本流程

1. **DNS 解析**：客户端通过域名访问业务系统，平台侧推荐您在 DNS 平台将自定义域名通过 CNAME 的方式解析到平台侧提供的服务域名上。服务域名由平台侧解析至 ALB 的虚拟 IP（VIP）。
2. **请求到达**：ALB 在指定端口接收请求，解析完整的 HTTP 内容（方法、URL、Header 等）。
3. **规则匹配**：按优先级顺序评估转发规则，提取请求特征并逐一比对，首条完全匹配的规则被激活执行。
4. **流量分发**：根据规则动作确定目标组，使用负载均衡算法（如加权轮询）从目标组中选择健康的后端服务器，转发请求。
5. **响应返回**：后端服务器处理请求并返回响应，ALB 将响应转发回客户端。

建议您跨多个可用区配置负载均衡器的后端服务器实例。如果一个可用区变得不可用，负载均衡器会将流量路由到其他可用区正常运行的实例上去，从而避免可用区故障引起的服务中断问题。

请求路由选择

客户端请求通过域名访问服务，在请求发送到负载均衡器之前，DNS 服务器将会解析负载均衡域名，并将收到请求的负载均衡 IP 地址返回到客户端。当负载均衡监听器收到请求时，将会使用不同的负载均衡算法将请求分发到后端服务器中。

健康检查机制

ALB 定期向后端发送探测请求（TCP 连接或 HTTP 请求），根据响应结果判定节点状态，从而尽量只将流量路由到正常运行的实例上去。当负载均衡器检测到运行不正常的实例时，它会停止向该实例路由流量，然后会在它再次检测到实例正常运行之后重新向其路由流量。

跨可用区部署

建议创建 ALB 时选择至少两个不同可用区的子网，实现跨可用区容灾：

- 单个可用区故障时，其他可用区仍可正常提供服务。
- 流量自动分散到多个区域，提升整体吞吐能力。
- 符合企业级业务的容灾要求。

相关服务

> ⚠ **注意：**
> 官网文档里的产品概述、优势、应用场景等文档中的数据描述，均来源于2026年腾讯内部实验测试结果，测试基于特定环境、条件及时间范围，仅反映相应测试场景下的情况，实际效果可能因业务场景、配置及使用情况不同而存在差异。

ALB 作为腾讯云网络生态的核心组件，可与以下产品协同使用：

计算类：

- 云服务器 CVM：作为 ALB 的后端服务器，承载业务应用。
- 容器服务 TKE：通过 Ingress Controller 对接 ALB，管理容器流量。
- 弹性伸缩 AS：AS 创建的实例自动注册到 ALB 目标组，按需扩缩。

网络与安全类：

- 私有网络 VPC：ALB 部署在 VPC 内，有助于保障网络安全。
- 安全组和 ACL：安全组是一种虚拟防火墙，能够通过自定义的访问规则，控制业务流量。ACL 规则是一种子网级别的可选安全层，用于控制进出子网的数据流，可以精确到协议和端口粒度，实现子网粒度流量的精细化控制。
- Web 应用防火墙 WAF：集成到 ALB，防御 SQL 注入、XSS 等攻击。
- DDoS 高防包：为公网 ALB 提供增强型 DDoS 清洗能力。
- SSL 证书：为 HTTPS 监听器申请和管理证书。
- 云解析 DNS：将域名解析至 ALB 的 VIP 或 CNAME。
- 全球网络加速 GA：为 ALB 提供跨国低延迟接入能力。

运维与监控类：

- 腾讯云可观测平台：监控 ALB 的 QPS、延迟、状态码等关键指标。
- 访问管理 CAM：帮助您安全、便捷地管理对腾讯云服务和资源的访问
- 日志服务 CLS：收集和分析 ALB 的访问日志。

产品优势

最近更新时间：2026-07-06 16:57:00

高性能

ALB（Application Load Balancer）采用全托管应用层集群架构，单实例可承载**百万级 QPS 及千万级并发连接数**，可应对日访问量超千万的电商平台、在线教育、互动娱乐等高并发业务场景。ALB 支持 HTTP/2 多路复用与 QUIC 低延迟协议，为您的大流量业务提供稳定的入口。

高可用

ALB 默认采用**多可用区部署**模式，各可用区间互为热备份，任一可用区故障时自动秒级切换，可用性达99.995%。配合 弹性伸缩，当后端服务器组中实例出现异常时，ALB 可自动将流量分发至健康实例，有助于保障业务连续性。跨可用区容灾能力让您的核心服务无需额外投入即可获得企业级高可用能力。

> **说明：**
> 当前有关 SLA 的描述，即可用性达99.995%，为正式商用后的承诺。当前处于公测阶段，不提供 SLA 承诺。

精细化调度

ALB 基于应用层协议特征实现**细粒度流量转发**，支持基于域名、URL 路径、HTTP Header、Query String 及 Cookie 等多维条件的路由匹配，满足微服务架构下复杂的灰度发布、A/B 测试及多租户隔离等场景需求。同时支持请求重定向与 URL 重写功能，您无需修改后端代码即可完成业务逻辑调整，大幅降低迭代成本。

协议丰富

ALB 全面覆盖现代应用的主流通信协议，支持 HTTP、HTTPS、HTTP/2、WebSocket 及 QUIC 协议，并原生支持 gRPC 框架，无缝对接微服务间的高性能通信需求。QUIC 协议特别适用于实时音视频、在线游戏等对延迟敏感的业务场景；SSE（Server-Sent Events）流式传输能力可支撑大模型 AI 应用的实时推理结果推送，一套负载均衡方案即可支撑 Web 应用、移动端 API 与 AI 推理等多元化业务。

安全可靠

ALB 集成多层次安全防护体系，全链路支持 TLS 1.3加密传输，内置 SNI 多域名证书管理能力，单个监听器可关联多张证书以满足大规模 HTTPS 业务需求。您可结合 Web 应用防火墙 和 DDoS 高防包，帮助识别和防护 SQL 注入、XSS 等 Web 攻击及 DDoS 威胁，为后端服务构建从网络层到应用层的纵深防御体系。

云原生集成

ALB 可作为**云原生 Ingress 网关**使用，与 容器服务 TKE 、Serverless 云函数 深度集成，支持通过 Ingress 资源声明式管理路由规则，实现容器化应用的统一流量入口管控。当您的容器服务通过 弹性伸缩 动态扩缩容时，

ALB 可自动感知后端节点变更并实时更新转发策略，您无需手工干预即可实现云原生环境下的弹性流量调度。

弹性计费

ALB 采用 **CU（Capacity Unit）性能容量单位**计费模式，根据实际资源消耗（新建连接数、并发连接数、处理流量、规则评估数）按量计费，相比固定规格实例更贴合弹性业务的实际使用成本。按实际资源消耗计费，业务低谷时相应降低费用，业务高峰时自动扩容无需提前预估容量，帮助您在获得性能的同时有效优化基础设施成本。

应用场景

最近更新时间：2026-07-06 16:57:00

ALB 面向七层（应用层）负载均衡场景，在提供高性能流量分发能力的同时，支持基于内容的高级路由、协议灵活适配及云原生生态深度集成，广泛适用于 Web 服务、微服务、容器化应用、AI 推理等多元化业务场景。

高并发 Web 业务的流量分发与智能调度

对于电商平台的大促活动、在线教育的开课高峰、互动娱乐的新赛季发布等访问量瞬间激增的业务场景，ALB 单实例可承载百万级 QPS 及千万级并发连接数，将客户端请求均匀分发至后端多台服务器，避免单点过载导致响应延迟。当某台后端服务器出现异常时，ALB 通过健康检查自动识别并快速隔离故障节点，将其流量转移至其他健康实例，全程无需人工干预，有助于保障业务连续性。若您的业务部署在多个可用区，建议您将 ALB 实例同时绑定跨可用区的后端服务器组，以便在后端层实现可用区级容灾，进一步提升服务稳定性。

微服务架构的统一流量入口

在微服务架构中，多个服务往往需要对外暴露独立的访问入口，传统方案通常为每个服务配置单独的负载均衡实例，导致资源分散且运维成本高。ALB 基于域名、URL 路径、HTTP Header、Query String 及 Cookie 等多维条件实现细粒度转发规则，单个实例即可承载数十个乃至上百个微服务的流量入口需求。例如，您可以配置 api.example.com/user 转发至用户服务、api.example.com/order 转发至订单服务，或根据请求头中的版本号将流量路由到不同版本的服务实例。相比为每个服务单独配置负载均衡实例，可减少实例数量，帮助您在简化架构的同时降低运维复杂度。

全站 HTTPS 安全改造与协议升级

随着安全合规要求的日益严格，越来越多的 Web 业务需要从 HTTP 协议迁移至 HTTPS 以保障数据传输安全。ALB 支持在监听器上配置重定向规则，将 HTTP 请求自动跳转至对应的 HTTPS 监听器，无需修改后端应用代码即可完成全站加密改造。您可以根据业务需要选择 301（永久重定向）、302（临时重定向）等不同状态码，并支持基于路径、HTTP 标头等条件的差异化跳转策略——例如仅对 /login、/payment 等敏感路径强制 HTTPS，其余路径保持 HTTP 访问。此外，ALB 全链路支持 TLS 1.3 加密协议及 SNI 多域名证书管理，单个监听器可关联数张 SSL 证书，轻松满足大规模多域名 HTTPS 业务的部署需求。

云原生容器化业务的弹性伸缩

对于基于 Kubernetes 编排的容器化应用，ALB 作为 Ingress 网关，可与 容器服务 TKE 深度集成，通过 Ingress 资源声明式定义路由规则，实现容器化应用的统一七层流量入口管控。当您的业务通过 弹性伸缩 或 HPA（水平 Pod 自动扩缩容）动态调整副本数量时，ALB 可自动感知后端 Pod 的变更并实时更新转发策略，新启动的 Pod 自动加入负载均衡池，缩容回收的 Pod 自动移除，整个过程对前端请求基本无感知。

新版本灰度发布与平滑过渡

对于金融系统、支付核心等对发布风险高度敏感的业务，直接全量上线新版本可能带来不可预估的影响。ALB 支持通过服务器组权重分配或基于 Header/Cookie 的转发规则，精确控制新旧版本的流量比例，实现灰度发布。例如，您可以先将 5% 的流量按请求头中的 X-Version: canary 路由到新版服务进行验证观察，确认无误后再逐步提升比例直至全量切换；若发现异常，只需调整规则即可快速回滚至旧版本，整个过程中终端用户基本无感知。这种渐进式的发布机制可显著降低版本迭代风险，适用于更新频繁且需保障业务连续性的生产环境。

使用限制

最近更新时间：2026-06-26 17:22:20

本文为您介绍应用型负载均衡（ALB）的使用限制。

通用配额限制

实例维度

资源项目	默认限制	
	支持数量	是否支持提升
一个账号在单地域可创建的实例数	60	可提升，请 提交工单 咨询。
一个实例可添加的监听器数	50	可提升，请 提交工单 咨询。
一个实例可添加的转发规则数	100，不包含默认规则	不支持提升。
一个实例可添加的扩展证书数	25，不包含默认证书	可提升，请 提交工单 咨询。
一个实例可添加的目标组数	100	不支持提升。
一个实例可添加的后端服务数	10000	不支持提升
一个实例可添加的安全组数	10	可提升，请 提交工单 咨询。

监听器/转发规则维度

资源项目	默认限制	
	支持数量	是否支持提升
一个转发规则可添加的目标组数	5	不支持提升。
一个转发规则可添加的转发条件数	10	不支持提升。
一个转发规则可添加的匹配条件条目数	10	不支持提升。
一个转发规则可添加的转发动作数	5	不支持提升。

目标组

资源项目	默认限制	
	支持数量	是否支持提升
一个账号在单地域可创建的目标组数	1000	可提升，请 提交工单 咨询。
一个目标组中可添加的后端服务数	1000	不支持提升。
一个后端服务（IP）可被添加到目标组的次数	200	不支持提升。
一个目标组可被绑定的转发规则数	50	可提升，请 提交工单 咨询。

其他

资源项目	默认限制	
	支持数量	是否支持提升
一个加密套件模板可关联的监听器数	10	不支持提升。

容量规格

单 VIP 性能指标如下表所示，不支持调整。

指标名	最大规格
每秒请求数（QPS）	50万
新建连接数（CPS）	20万/秒
最大并发连接数	500万
出带宽	25Gbps
入带宽	25Gbps

说明：
公网 ALB 实例的公网带宽为 ALB 实例下所有 EIP 的带宽总和，默认单个 EIP 的公网带宽为200Mbps。一个 ALB 实例选择2个可用区，即2个 EIP，则该 ALB 实例的带宽上限为400Mbps。

支持地域

当前仅支持广州，其他地域即将上线，具体支持的地域请以购买页为准。