负载均衡

产品简介

产品简介

什么是负载均衡

负载均衡（Server Load Balancer）是将访问流量根据转发策略分发到后端多台云服务器（ECS实例）的流量分发控制服务。负载均衡扩展了应用的服务能力，增强了应用的可用性。

负载均衡通过设置虚拟服务地址，将添加的ECS实例虚拟成一个高性能、高可用的应用服务池，并根据转发规则，将来自客户端的请求分发给云服务器池中的ECS实例。

负载均衡默认检查云服务器池中的ECS实例的健康状态，自动隔离异常状态的ECS实例，消除了单台ECS实例的单点故障，提高了应用的整体服务能力。此外，负载均衡还具备抗DDoS攻击的能力，增强了应用服务的防护能力。

组成部分

负载均衡由以下三个部分组成：

负载均衡实例 （Server Load Balancer instances）

一个负载均衡实例是一个运行的负载均衡服务，用来接收流量并将其分配给后端服务器。要使用负载均衡服务，您必须创建一个负载均衡实例，并至少添加一个监听和两台ECS实例。

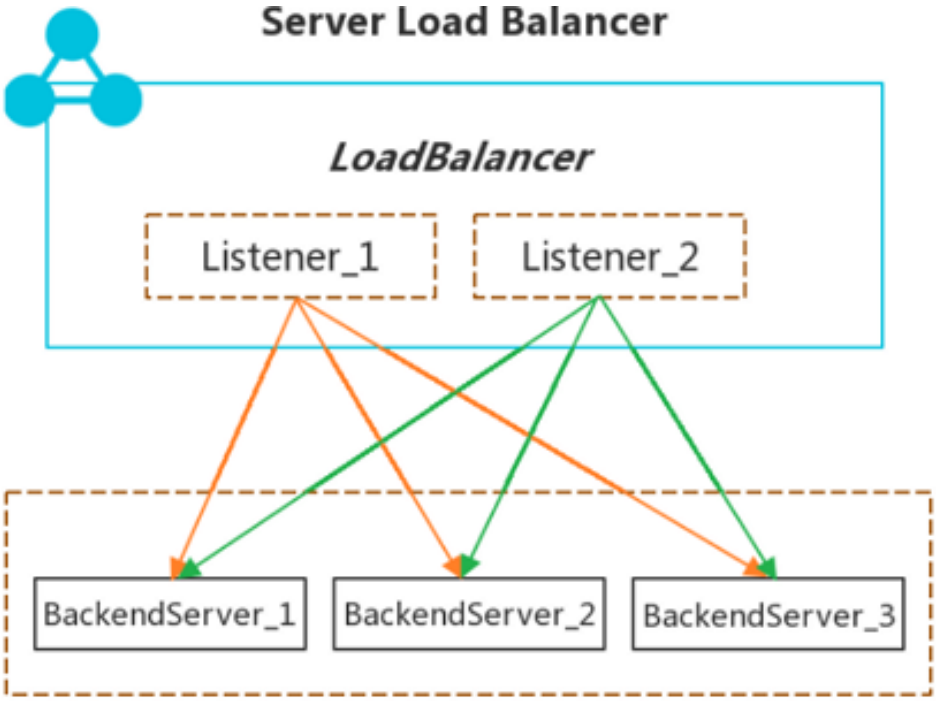
监听 （Listeners）

监听用来检查客户端请求并将请求转发给后端服务器。监听也会对后端服务器进行健康检查。

后端服务器（Backend Servers）

一组接收前端请求的ECS实例。您可以单独添加ECS实例到服务器池，也可以通过虚拟服务器组或主备服务器组来批量添加和管理。

如下图所示，来自客户端的请求经过负载均衡实例后，监听会将请求根据配置的监听规则分发给后端添加的ECS实例处理。



产品优势

高可用

采用全冗余设计，无单点，支持同城容灾。搭配DNS可实现跨地域容灾，可用性高达99.95%。

根据应用负载进行弹性扩容，在流量波动情况下不中断对外服务。

低成本

与传统硬件负载均衡系统高投入相比，成本可下降60%。

安全

阿里云对开源四层负载均衡LVS的管理软件Keepalived进行了全面优化，使得基于LVS的四层负载均衡具备接近于实时防御的能力。结合云盾，可提供5G以下的防DDOS攻击能力。

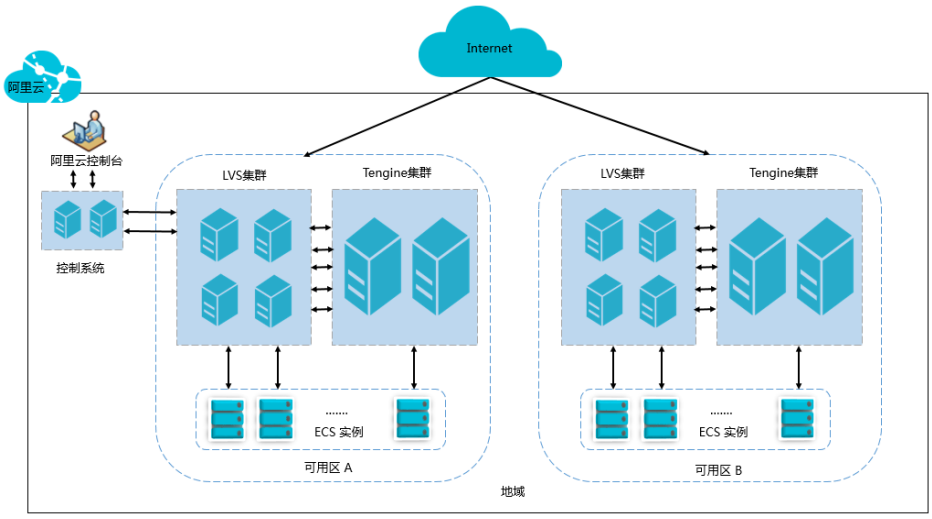
采用Tengine作为负载均衡基础模块的七层负载均衡具备多维度的CC攻击防御能力。

基础架构

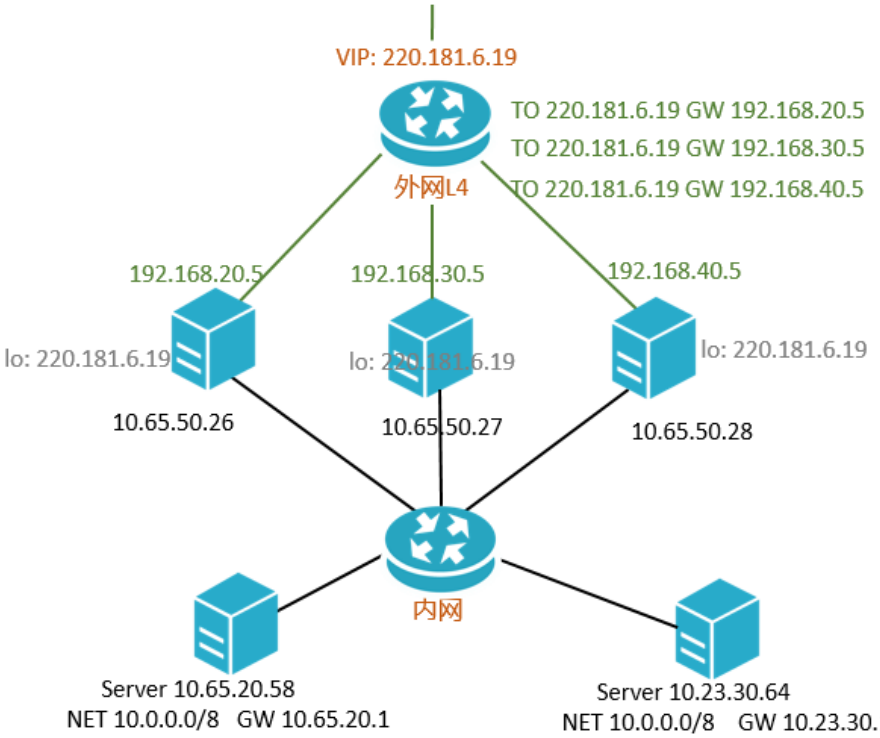
负载均衡采用集群部署，可实现会话同步，以消除服务器单点故障，提升冗余，保证服务的稳定性。阿里云当前提供四层（TCP协议和UDP协议）和七层（HTTP和HTTPS协议）的负载均衡服务。

四层采用开源软件LVS（Linux Virtual Server）+ keepalived的方式实现负载均衡，并根据云计算需求对其进行了个性化定制。

七层采用Tengine实现负载均衡。Tengine是由淘宝网发起的Web服务器项目，它在Nginx的基础上，针对有大访问量的网站需求，添加了很多高级功能和特性。

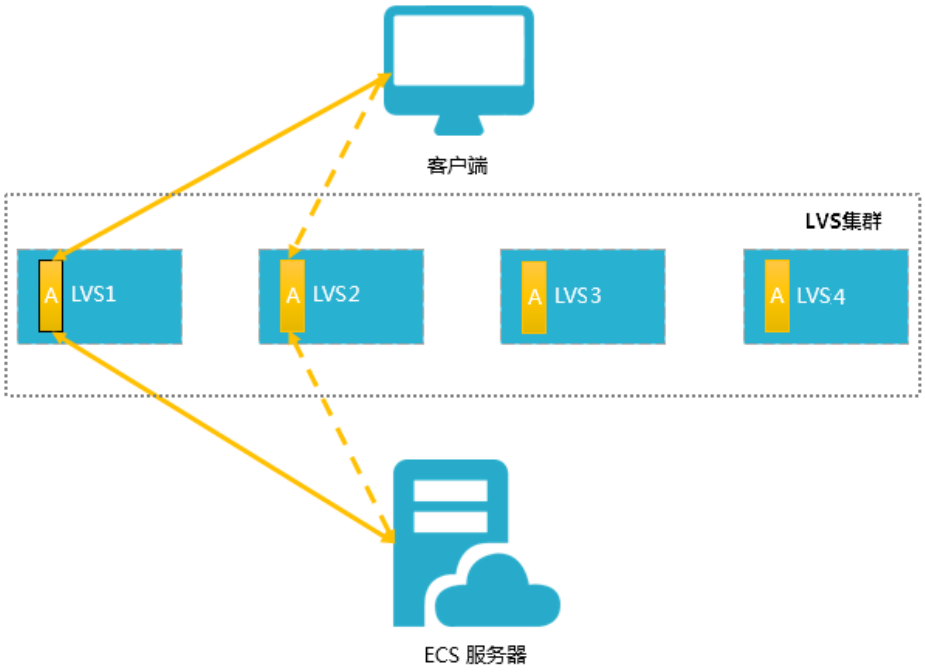


如下图所示，各个地域的四层负载均衡实际上是由多台LVS机器部署成一个LVS集群来运行的。采用集群部署模式极大地保证了异常情况下负载均衡服务的可用性、稳定性与可扩展性。



LVS集群内的每台LVS都会进行会话，通过组播报文同步到该集群内的其它LVS机器上，从而实现LVS集群内各台机器间的会话同步。如下图所示，当客户端向服务端传输三个数据包后，在LVS1上建立的会话A开始同步到其它LVS机器上。图中实线表示现有的连接，图中虚线表示当LVS1出现故障或进行维护时，这部分流量会走到一台可以正常运行的机器LVS2上。因而负载均衡集群支持热升级，并且在机器故障和集群维护时最大程度对用户透明，不影响用户业务。

注意：对于连接未建立（三次握手未完成），或者已建立连接但未触发会话同步机制，热升级不保证连接不中断，需要依靠客户端重新发起连接。



功能概述

阿里云负载均衡服务提供以下功能：

协议支持

当前提供四层（TCP协议和UDP协议）和七层（HTTP和HTTPS协议）的负载均衡服务。

健康检查

支持对后端ECS实例进行健康检查。负载均衡服务会自动屏蔽异常状态的ECS实例，待该ECS实例恢复正常后自动解除屏蔽。

会话保持

提供会话保持功能。在会话的生命周期内，可以将同一客户端的会话请求转发到同一台后端ECS实例上。

调度算法

支持轮询、加权轮询（WRR）、加权最小连接数（WLC）三种调度算法。

轮询：按照访问次数依次将外部请求依序分发到后端ECS实例上。

加权轮询：用户可以对每台后端服务器设置权重值。权重值越高的服务器，被轮询到的次数（概率）也越高。

加权最小连接数：除了根据对每台后端服务器设定的权重值来进行轮询，同时还考虑后端服务器的实际负载（即连接数）。当权重值相同时，当前连接数越小的后端服务器被轮询到的次数（概率）也越高。

域名URL转发

针对七层协议（HTTP协议和HTTPS协议），支持按设定的访问域名和URL将请求转发到不同的虚拟服务器组。

多可用区

支持在指定可用区创建负载均衡实例。在多可用区部署的地域还支持主备可用区，当主可用区出现故障时，负载均衡可自动切换到备可用区上提供服务。

访问控制

通过添加负载均衡监听的访问白名单，仅允许特定IP访问负载均衡服务。

安全防护

结合云盾，可提供5G以下的防DDoS攻击能力。

证书管理

针对HTTPS协议，提供统一的证书管理服务。证书无需上传到后端ECS实例，解密处理在负载均衡上进行，降低后端ECS实例的CPU开销。

带宽控制

支持根据监听设置其对应服务所能达到的带宽峰值。

实例类型

提供公网和私网类型的负载均衡服务。您可以根据业务场景来选择配置对外公开或对内私有的负载均衡服务，系统会根据您的选择分配公网或私网服务地址。

公网类型的负载均衡默认使用经典网络；私网类型的负载均衡服务可以选择使用经典网络或专有网络。

监控

提供丰富的监控数据，实时了解负载均衡运行状态。

管理方式

提供控制台、API、SDK多种管理方式。

应用场景

负载均衡主要应用于以下场景中：

场景一：应用于高访问量的业务

如果您的应用访问量很高，您可以通过配置监听规则将流量分发到不同的ECS实例上。此外，您可以使用会话保持功能将同一客户端的请求转发到同一台后端ECS，提高访问效率。

场景二：横向扩张系统

您可以根据业务发展的需要，通过随时添加和移除ECS实例来扩展应用系统的服务能力，适用于各种Web服务器和App服务器。

场景三：消除单点故障

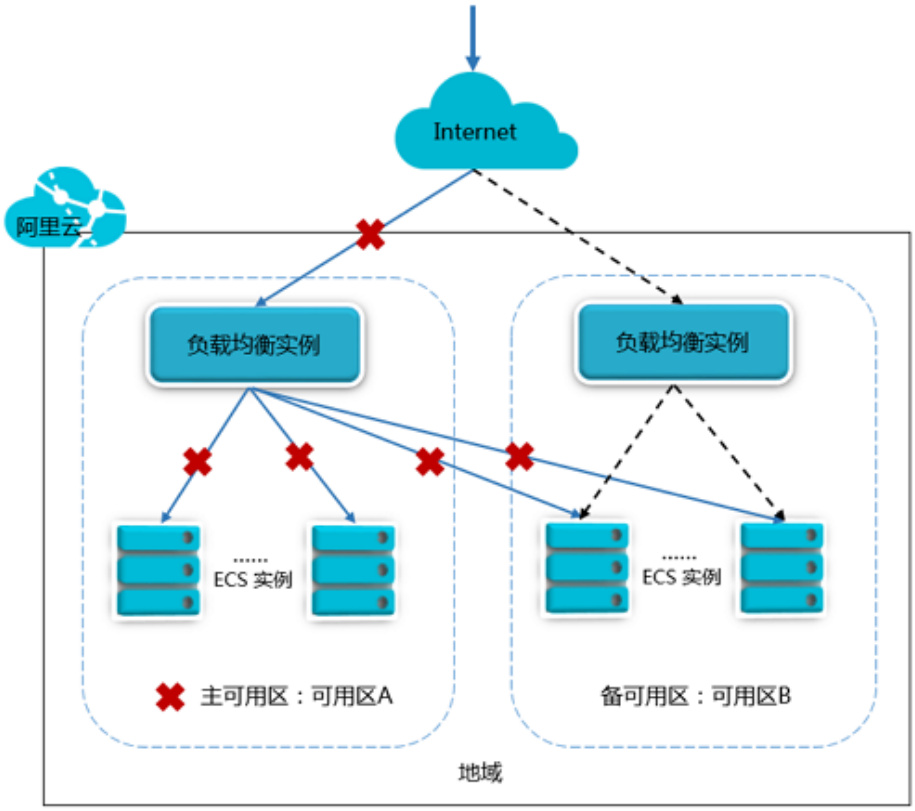
您可以在负载均衡实例下添加多台ECS实例。当其中一部分ECS实例发生故障后，负载均衡会自动屏蔽故障的ECS实例，将请求分发给正常运行的ECS实例，保证应用系统仍能正常工作。

场景四：同城容灾（多可用区容灾）

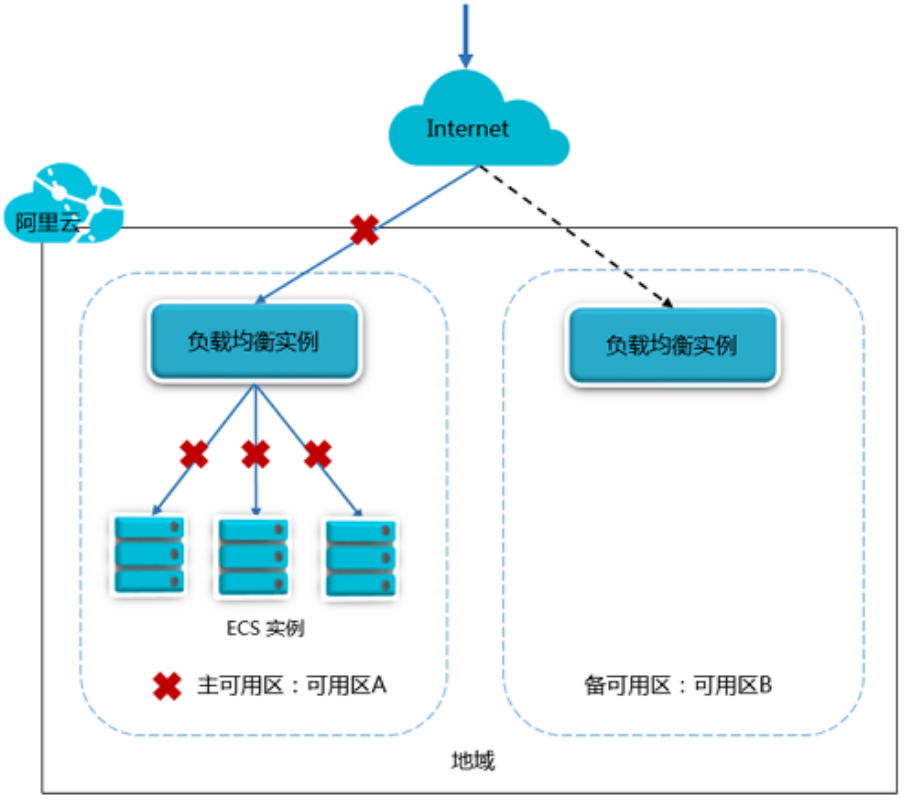
为了提供更加稳定可靠的负载均衡服务，阿里云负载均衡已在各地域部署了多可用区以实现同地域容灾。当主可用区出现机房故障或不可用时，负载均衡仍然有能力在非常短的时间内（大约30s中断）切换到另外一个备可用区恢复服务能力；当主可用区恢复时，负载均衡同样会自动切换到主可用区提供服务。

使用负载均衡时，您可以将负载均衡实例部署在支持多可用区的地域以实现同城容灾。此外，建议您结合自身的应用需要，综合考虑后端服务器的部署。如果您的每个可用区均至少添加了一台ECS实例，那么此种部署模式下的负载均衡服务的效率是最高的。

如下图所示，在负载均衡实例下绑定不同可用区的ECS实例。正常情况下，用户访问流量将转发至主可用区内的ECS实例；当可用区A发生故障时，用户访问流量将转发至备可用区内的ECS实例。此种部署既可以避免因为单个可用区的故障而导致对外服务的不可用，也可以通过不同产品间可用区的选择来降低延迟。

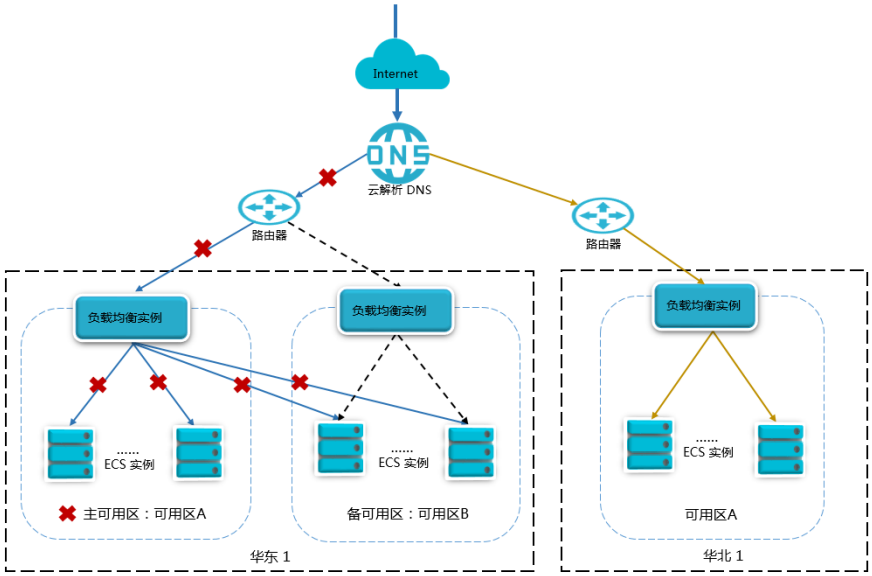


如果您采取如下图所示的部署方案，即在负载均衡实例的主可用区下绑定多台ECS实例，而在备可用区没有任何ECS实例。当主可用区发生故障时会造成业务中断，因为备可用区没有ECS实例来接收请求。这样的部署方式很明显是以牺牲高可用性为代价来获取低延时。



场景五：跨地域容灾

您可以在不同地域下部署负载均衡实例，并分别挂载相应地域内不同可用区的ECS。上层利用云解析做智能DNS，将域名解析到不同地域的负载均衡实例服务地址下，可实现全局负载均衡。当某个地域出现不可用时，暂停对应解析即可实现所有用户访问不受影响。配置详情参考结合云解析实现跨地域负载均衡。



使用限制

限制项	普通用户限制描述	例外申请方式（例外上限）
创建负载均衡实例的财务限制	账户余额大于等于100元现金	工单
创建负载均衡实例的用户限制	在控制台上创建，需要实名认证 使用Open API创建，无限制	没有例外
单台SLB实例后端挂载ECS服务器的数量	200	提交工单
创建负载均衡实例的ECS限制	至少拥有一个ECS实例，且只能选择有ECS实例的地域创建负载均衡实例	没有例外
创建负载均衡实例可用的类型	公网/私网	没有例外
负载均衡实例的计费方式	按使用流量或按固定带宽	没有例外
负载均衡实例公网带宽可选范围（按固定带宽）	1-5000 MB，缺省 1 MB	没有例外
用户默认按量付费的实例配额	60	工单
负载均衡实例名称输入范围	长度限制为1-80个字符，允许包含字母、数字、‘-’、‘/’、‘.’、‘_’这些字符	没有例外
负载均衡实例监听数量	50	没有例外
一个七层监听最多可添加的域名、URL转发策略数量	20	工单申请
负载均衡实例监听可选择的协议类型	HTTP/HTTPS/TCP/UDP	没有例外
负载均衡实例监听可选择的前端/后端端口范围（公共云）	1-65535	没有例外
负载均衡实例监听可选择的前端端口范围（金融云）	80，443，2800-3300，5000-10000，13000-14000	没有例外
支持的调度算法	轮询、加权轮询、加权最小连接数	没有例外
HTTP协议会话保持Cookie处理方式	insert/server，缺省insert	没有例外
HTTP协议会话保持Cookie超时时间	1-86400，缺省3600	没有例外
HTTP协议会话保持Cookie名称	长度限制为1-200，cookie必须遵守RFC2965。这意味着它只能包含ASCII英文字母和数字	没有例外

	，不能包含逗号、分号或空格，也不能以$字符开头	
HTTP健康检查端口	输入范围为1-65535，缺省使用后端服务端口	没有例外
HTTP健康检查域名	长度限制为1-80，只能使用字母、数字、‘-’、‘.’这些字符	没有例外
HTTP健康检查URI	长度限制为1-80，只能使用字母、数字、‘-’、‘/’、‘.’、‘%’、‘?’、‘#’、‘&’这些字符	没有例外
HTTP健康检查超时时间	输入范围1-300秒，缺省5 注意：如果健康检查Timeout < Interval，则健康检查Timeout无效，超时时间为Interval	没有例外
HTTP健康检查间隔	输入范围1-50秒，缺省2	没有例外
HTTP健康检查健康阈值	输入范围2-10，缺省3	没有例外
HTTP健康检查不健康阈值	输入范围2-10，缺省3	没有例外
TCP会话保持超时时间	1-3600秒	没有例外
TCP健康检查端口	输入范围为1-65535，缺省使用后端服务端口	没有例外
TCP健康检查超时时间	输入范围1-300秒，缺省5	没有例外
TCP健康检查间隔	输入范围1-50秒，缺省2	没有例外
TCP健康检查健康阈值	输入范围2-10，缺省3	没有例外
TCP健康检查不健康阈值	输入范围2-10，缺省3	没有例外
UDP健康检查端口	输入范围为1-65535，缺省使用后端服务端口	没有例外
UDP健康检查超时时间	输入范围1-300秒，缺省10	没有例外
UDP健康检查间隔	输入范围1-50秒，缺省5	没有例外
UDP健康检查健康阈值	输入范围2-10，缺省3	没有例外
UDP健康检查不健康阈值	输入范围2-10，缺省3	没有例外
后端ECS批量添加/删除数量	20	没有例外
单个key的API访问频率限制	5000次/天	联系客服或BD提高限制
每个用户上传证书数量限制	100	联系客服或BD提高限制

名词解释

名词	英文	说明
负载均衡	Server Load Balancer	阿里云提供的一种网络负载均衡服务，结合阿里云提供的ECS服务，提供四层和七层负载均衡服务。
负载均衡实例	Server Load Balancer Instance	负载均衡实例是一个运行的负载均衡服务。要使用负载均衡服务，必须先创建一个负载均衡实例。LoadBalancerId是识别负载均衡实例的唯一标识。
服务地址	IP Address	系统为创建的负载均衡实例分配的服务IP地址。根据创建的负载均衡实例的类型，服务地址可能是公网IP也可能是私网IP。您可以将域名解析到公网IP地址提供对外服务。
监听	Listener	负载均衡服务监听规定了如何将请求转发给后端服务器。一个负载均衡实例至少添加一个监听。
后端服务器	Backend Server	处理负载均衡分发的前端请求的ECS实例。
虚拟服务器组	VServer Group	一组处理负载均衡分发的前端请求的ECS实例。不同的监听可以关联不同的虚拟服务器组，实现监听维度的请求转发。
多可用区	Mutlipe Zones	负载均衡在大部分地域部署了多个可用区以实现同地域容灾。如果您在多可用区创建了一个负载均衡实例，该实例将默认工作在主可用区。当主可用区发生故障时，负载均衡实例可切换到备可用区工作并且会在主可用区恢复时自动切换回主可用区。