

弹性负载均衡

产品介绍

文档版本 01
发布日期 2026-01-28



版权所有 © 华为云计算技术有限公司 2026。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目录

1 网络型+应用型	1
1.1 什么是弹性负载均衡.....	1
1.2 产品优势.....	3
1.3 弹性负载均衡是如何工作的.....	5
1.4 应用场景.....	10
1.5 产品功能.....	13
1.5.1 弹性负载均衡产品类型简介.....	13
1.5.2 弹性负载均衡功能对比.....	16
1.6 公网和私网负载均衡器.....	23
1.7 ELB 网络流量路径说明.....	25
1.8 独享型负载均衡实例规格.....	27
1.9 约束与限制.....	32
1.10 安全.....	36
1.10.1 责任共担.....	36
1.10.2 ELB 服务的访问控制.....	37
1.10.3 审计与日志.....	38
1.10.4 监控安全风险.....	38
1.10.5 认证证书.....	38
1.11 权限管理.....	40
1.12 基本概念.....	45
1.12.1 产品基本概念.....	45
1.12.2 区域和可用区.....	46
1.13 与其他服务的关系.....	48
2 网关型	50
2.1 什么是网关型负载均衡.....	50
2.2 网关型负载均衡是如何工作的.....	52
2.3 应用场景.....	53
2.4 产品功能.....	53
2.5 约束与限制.....	55

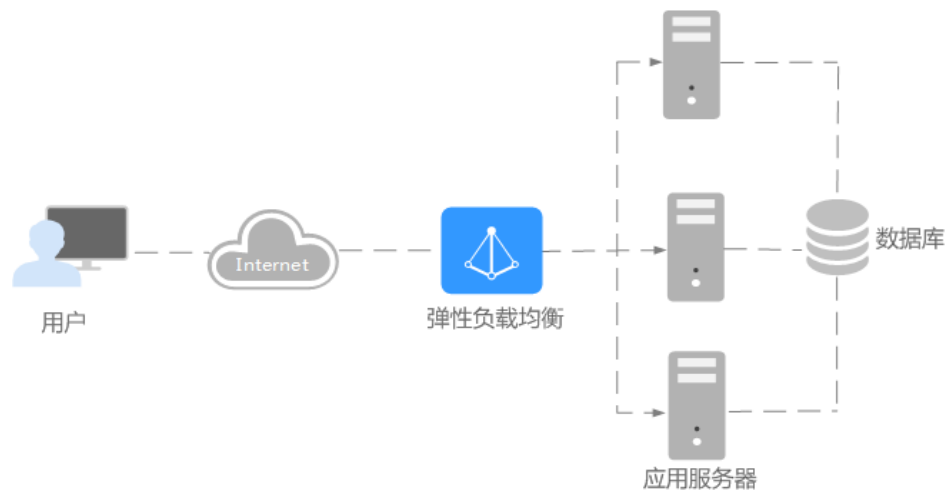
1 网络型+应用型

1.1 什么是弹性负载均衡

弹性负载均衡（Elastic Load Balance，简称ELB）是将访问流量根据分配策略分发到后端多台服务器的流量分发控制服务。弹性负载均衡可以通过流量分发扩展应用系统对外的服务能力，同时通过消除单点故障提升应用系统的可用性。

如下图所示，弹性负载均衡将访问流量分发到后端三台应用服务器，每个应用服务器只需分担三分之一的访问请求。同时，结合健康检查功能，流量只分发到后端正常工作的服务器，从而提升了应用系统的可用性。

图 1-1 使用弹性负载均衡实例



弹性负载均衡的组件

弹性负载均衡由以下部分组成：

图 1-2 弹性负载均衡组件图

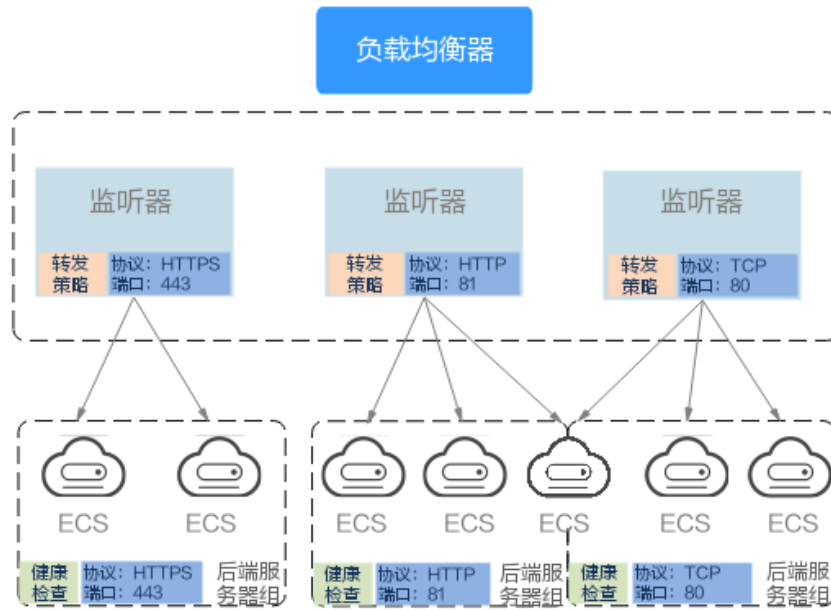


表 1-1 弹性负载均衡的组件

负载均衡器	接受来自客户端的传入流量并将请求转发到一个或多个可用区中的后端服务器。
监听器	您可以向您的弹性负载均衡器添加一个或多个监听器。监听器使用您配置的协议和端口检查来自客户端的连接请求，并根据您定义的分配策略和转发策略将请求转发到一个后端服务器组里的后端服务器。
后端服务器组	后端服务器组是一个或多个后端服务器的逻辑集合，用于将客户端的流量转发到一个或多个后端服务器，满足用户同时处理海量并发业务的需求。后端服务器可以是云服务器实例、辅助弹性网卡或IP地址。
后端服务器	负载均衡器会将客户端的请求转发给后端服务器处理。例如，您可以添加ECS实例作为负载均衡器的后端服务器，监听器使用您配置的协议和端口检查来自客户端的连接请求，并根据您定义的分配策略将请求转发到后端服务器组里的后端服务器。

弹性负载均衡的类型

弹性负载均衡支持独享型负载均衡、共享型负载均衡。

- 独享型负载均衡：独享型负载均衡实例资源独享，实例的性能不受其它实例的影响，您可根据业务需要选择不同规格的实例。
- 共享型负载均衡：属于集群部署，实例资源共享，实例的性能会受其它实例的影响，不支持选择实例规格。

独享型负载均衡和共享型负载均衡的详细区别请参见[弹性负载均衡产品类型简介](#)。

如何访问弹性负载均衡

可以使用以下方式访问和管理弹性负载均衡：

- 管理控制台
请使用管理控制台方式访问弹性负载均衡。可直接登录管理控制台，从主页选择“弹性负载均衡”。
- 查询API
通过调用API的方式访问弹性负载均衡，具体操作请参见《[弹性负载均衡API参考](#)》。

1.2 产品优势

弹性负载均衡和 LVS 和 Nginx 等开源软件对比的优势

表 1-2 弹性负载均衡的优势

对比项	弹性负载均衡	LVS/Nginx 负载均衡
运维方式	全托管、免运维。	用户自行安装、升级和维护。
计费模式	<ul style="list-style-type: none">● 弹性规格：按照实际使用量付费。● 固定规格：提供多种规格，支持差异化的性能指标。	按照业务峰值为预留资源付费。
部署方式	<ul style="list-style-type: none">● 集群模式部署。● 多可用区部署。	虚拟机或容器模式部署。
可靠性	<ul style="list-style-type: none">● 具备弹性能力，流量突发时，支持扩展计算资源。● 集群模式部署具备节点/可用区级容灾，服务等级定义SLA（Service Level Agreement）承诺99.99%。	<ul style="list-style-type: none">● 无弹性能力，需要按照业务峰值预留计算资源。● 七层性能依赖底层计算资源配置，无云厂商服务等级定义SLA（Service Level Agreement）承诺。
性能	支持千万级并发连接，百万级新建连接。	四层仅支持主备部署模式，性能受制于资源规格限制。
配置变更	支持配置动态加载。	<ul style="list-style-type: none">● 配置更新需要reload进程，对长连接有损。● Lua插件变更需要Reload进程。
SSL卸载	证书在ELB上进行卸载，不占用底层计算资源。	使用计算资源卸载证书，性能更低。

对比项	弹性负载均衡	LVS/Nginx 负载均衡
周边服务集成	<ul style="list-style-type: none"> 支持对接WAF、LTS、CES等周边服务。 敏捷部署监控与告警。 按需开启访问日志。 	用户自行手动部署。

独享型负载均衡的优势

表 1-3 独享型负载均衡的优势说明

超高性能	<p>可实现性能独享，资源隔离，单实例单AZ最高支持2千万并发连接，满足用户的海量业务访问需求。</p> <p>选择多个可用区之后，对应的最高性能规格（新建连接数/并发连接数等）会加倍。例如：单实例单AZ最高支持2千万并发连接，那么单实例双AZ最高支持4千万并发连接。</p>
高可用	支持多可用区的同城多活容灾，无缝实时切换。完善的健康检查机制，保障业务实时在线。
超安全	支持TLS 1.3，提供全链路HTTPS数据传输，支持多种安全策略并支持创建自定义安全策略，根据业务不同安全要求灵活选择安全策略。
多协议	支持TCP/UDP/TLS/HTTP/HTTPS/QUIC/GRPC协议，满足不同协议接入需求。
更灵活	支持请求方法、HEADER、URL、PATH、源IP等不同应用特征，并可对流量进行转发、重定向、固定返回码等操作。
无边界	提供混合负载均衡能力，可以将云上的资源和云下、多云之间的资源进行统一负载。
简单易用	快速部署ELB，实时生效，支持多种协议、多种调度算法可选，用户可以高效地管理和调整分发策略。

共享型负载均衡的优势

表 1-4 共享型负载均衡的优势说明

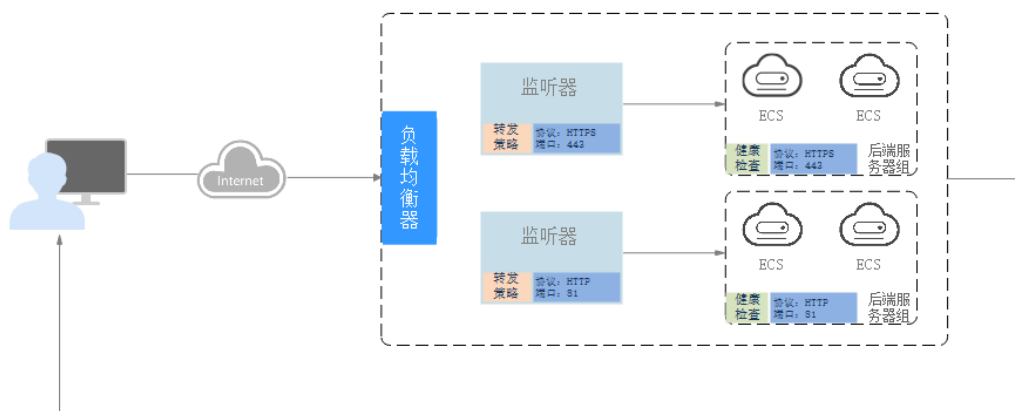
高性能	共享型负载均衡支持性能保障模式后，提供并发连接数5万、每秒新建连接数5000、每秒查询速率5000 的保障能力。
高可用	采用集群化部署，支持多可用区的同城双活容灾，无缝实时切换。完善的健康检查机制，保障业务实时在线。
多协议	支持TCP/UDP/HTTP/HTTPS，满足不同协议接入需求。

简单易用	快速部署ELB，实时生效，支持多种协议、多种调度算法可选，用户可以高效地管理和调整分发策略。
可靠性	支持跨可用区双活容灾，流量分发更均衡。

1.3 弹性负载均衡是如何工作的

工作原理

图 1-3 ELB 工作原理图



弹性负载均衡的工作原理如下：

1. 客户端发起请求：客户端向您的应用程序发起请求。
2. 监听器接收请求：负载均衡器中的监听器接收与您配置的协议和端口匹配的请求。
3. 负载均衡转发请求：
 - a. 监听器根据您的配置将请求转发至相应的后端服务器组。
 - b. 如果您配置了转发策略，监听器会评估传入的请求是否匹配转发策略。如果匹配上某条转发策略，请求将按照相应的转发动作进行转发。
4. 后端服务器处理请求：后端服务器组中健康检查正常的后端服务器将根据分配策略和您在监听器中配置的转发策略的转发接收流量，处理流量并返回客户端。

请求的流量分发与负载均衡器所绑定的监听器配置的**转发策略**和后端服务器组配置的**分配策略**类型相关。

分配策略类型

独享型负载均衡支持加权轮询算法、加权最少连接、源IP算法、连接ID算法。

共享型负载均衡支持加权轮询算法、加权最少连接、源IP算法。

加权轮询算法

图1-4展示弹性负载均衡器使用加权轮询算法的流量分发流程。假设可用区内有2台权重相同的后端服务器，负载均衡器节点会将50%的客户端流量分发到其可用区中的每一台后端服务器。

图 1-4 加权轮询算法流量分发

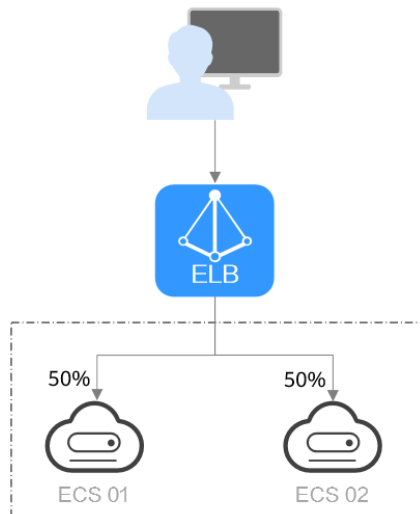


表 1-5 加权轮询算法说明

概述	加权轮询算法根据组内后端服务器设置的权重，依次将请求分发给不同的服务器。权重大的后端服务器被分配的概率高，相同权重的服务器处理相同数目的连接数。
推荐场景	加权轮询算法常用于短连接服务，例如HTTP等服务。 <ul style="list-style-type: none">● 灵活负载：当对后端服务器的负载分配有更精细的要求时，可以通过设置不同的权重来实现对服务器的灵活调度，使得性能较好的服务器能够处理更多的请求。● 动态负载：当后端服务器的性能和负载情况经常发生变化时，可以通过动态调整权重来适应不同的场景，实现负载均衡。
缺点	<ul style="list-style-type: none">● 加权轮询算法需要配置每个后端服务器的权重，对于有大量后端服务器或频繁变动的场景，运维工作量较大。● 权重设置不准确可能会导致负载不均衡的情况，需要根据后端服务器的实际性能进行调整。

加权最少连接

图1-5展示弹性负载均衡器使用加权最少连接算法的流量分发流程。假设可用区内有2台权重相同的后端服务器，ECS 01已有100个连接，ECS 02已有50个连接，则新的连接会优先分配到ECS 02上。

图 1-5 加权最少连接算法流量分发

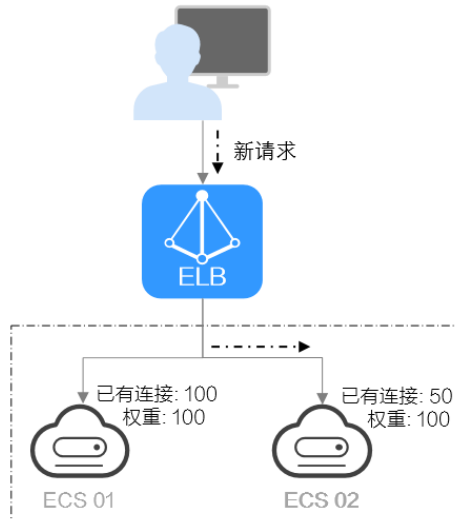


表 1-6 加权最少连接说明

<p>概述</p>	<p>最少连接是通过当前活跃的连接数来评估服务器负载情况的一种动态负载均衡算法。加权最少连接就是在最少连接数的基础上，根据服务器的不同处理能力，给每个服务器分配不同的权重，使其能够接受相应权值数的服务请求。</p>
<p>推荐场景</p>	<p>加权最少连接常用于长连接服务，例如数据库连接等服务。</p> <ul style="list-style-type: none"> ● 灵活负载：当后端服务器的性能差异较大时，同时考虑后端服务器的连接数和权重来进行负载，可以更精确地将请求分配到后端服务器上，避免出现过载或空闲的情况。 ● 动态负载：当后端服务器的连接数和负载情况经常发生变化时，可以通过实时监控连接数变化进行动态的负载调整。 ● 更高稳定负载：对于需要高稳定性的业务场景，加权最少连接算法可以降低后端服务器的峰值负载，提高业务的稳定性和可靠性。
<p>缺点</p>	<ul style="list-style-type: none"> ● 加权最小连接算法的实现更复杂：需要实时监控负载均衡器与后端服务器之间的连接数变化。 ● 对后端服务器的连接数存在依赖：算法依赖于准确获取负载均衡服务和后端服务器的连接数，如果获取不准确或监控不及时，可能导致负载分配不均衡。同时由于算法只能统计到负载均衡器与后端服务器之间的连接，后端服务器整体连接数无法获取，因此对于后端服务器挂载到多个弹性负载均衡的场景，也可能导致负载分配不均衡。 ● 新增后端服务器时可能导致过载：如果已有的连接数过大，大量的新建连接会被分配到新加入的后端服务器上，可能会导致新加入的后端服务器瞬间过载影响系统稳定性。

源 IP 算法

图1-6展示弹性负载均衡器使用源IP算法的流量分发流程。假设可用区内有2台权重相同的后端服务器，ECS 01已经处理了一个IP-A的请求，则IP-A新发起的请求会自动分配到ECS 01上。

图 1-6 源 IP 算法流量分发

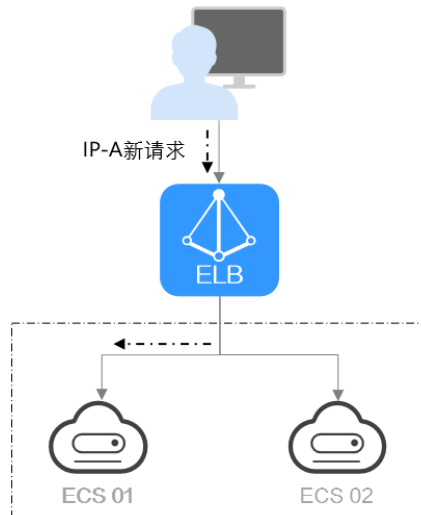


表 1-7 源 IP 算法说明

概述	根据请求的源IP地址进行一致性哈希计算，源IP地址相同的请求会被分配到同一台后端服务器。
推荐场景	<p>源IP算法常用于需要保持用户状态或会话的应用。</p> <ul style="list-style-type: none"> ● 基于源IP的会话保持：源IP算法可以确保源IP相同的请求具有相同的哈希值并被分配到同一台后端服务器上，从而实现会话保持。 ● 保持数据一致：一致性哈希算法将相同哈希值的请求调度到相同后端服务器上，保证多次请求数据的一致性。 ● 均衡性要求较高：一致性哈希算法能够提供相对均衡的负载分配效果，减少后端服务器的负载差异。
缺点	<ul style="list-style-type: none"> ● 后端服务器数量变动可能导致不均衡：一致性哈希算法在后端服务器数量变动时会尽力保障请求的一致性，部分请求会重新分配。当后端服务器数量较少时，重新分配过程中有可能导致负载不均衡的情况发生。 ● 扩展复杂性增加：由于一致性哈希算法将请求根据哈希因子进行哈希计算，当后端服务器数量变化时，会导致一部分请求需要重新分配，这会引入一定的复杂性。

连接 ID 算法

图1-7展示弹性负载均衡器使用连接ID算法的流量分发流程。假设可用区内有2台权重相同的后端服务器，ECS 01已经处理了一个客户端A的请求，则客户端A上新发起的请求会自动分配到ECS 01。

图 1-7 连接 ID 算法流量分发

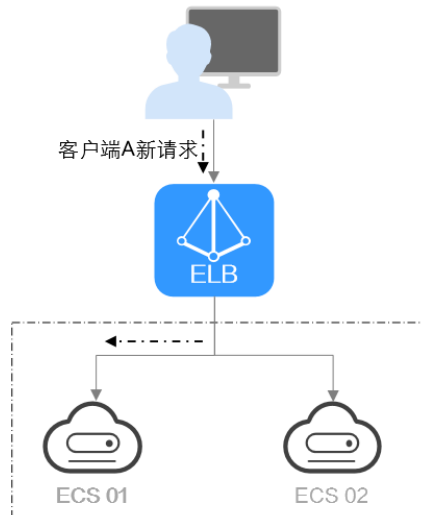


表 1-8 连接 ID 算法说明

概述	<p>根据QUIC 协议请求的QUIC ID进行哈希计算，相同QUIC连接上的请求会被分配到同一台后端服务器。QUIC ID是QUIC连接的唯一标识符，连接ID算法可以实现基于连接级别的负载均衡。</p> <p>仅QUIC协议的后端服务器组支持连接ID算法。</p>
推荐场景	<p>连接ID算法常用于实现连接级别负载均衡的应用。</p> <ul style="list-style-type: none">● 基于QUIC连接的会话保持：连接ID算法可以确保源相同QUIC连接上的请求具有相同的哈希值并被分配到同一台后端服务器上，从而实现会话保持。● 保持数据一致：一致性哈希算法将相同哈希值的请求调度到相同后端服务器上，保证多次请求数据的一致性。● 均衡性要求较高：一致性哈希算法能够提供相对均衡的负载分配效果，减少后端服务器的负载差异。
缺点	<ul style="list-style-type: none">● 后端服务器数量变动可能导致不均衡：一致性哈希算法在后端服务器数量变动时会尽力保障请求的一致性，部分请求会重新分配。当后端服务器数量较少时，重新分配过程中有可能导致负载不均衡的情况发生。● 扩展复杂性增加：由于一致性哈希算法将请求根据哈希因子进行哈希计算，当后端服务器数量变化时，会导致一部分请求需要重新分配，这会引入一定的复杂性。

影响负载均衡的因素

一般情况下，影响负载均衡分配的因素包括分配策略、会话保持、长连接、权重等。换言之，最终是否均匀分配不仅与分配策略相关，还与使用的长短连接、后端的性能负载等相关。

假设可用区内有2台权重相同且不为0的后端服务器，流量分配策略选择“加权最少连接”，未开启会话保持，ECS 01已有100个连接，ECS 02已有50个连接。

如果有客户端A使用长连接访问了ECS 01，长连接未断开期间，客户端A的业务流量将持续转发到ECS 01，其他客户端的业务流量则根据分配策略优先分配到ECS 02。

说明

后端服务器健康检查异常或权重设置为0时，ELB不会转发业务流量到该后端服务器。

配置流量分配策略请参考[流量分配策略](#)。

检查请求不均衡请参考[如何检查请求不均衡?](#)。

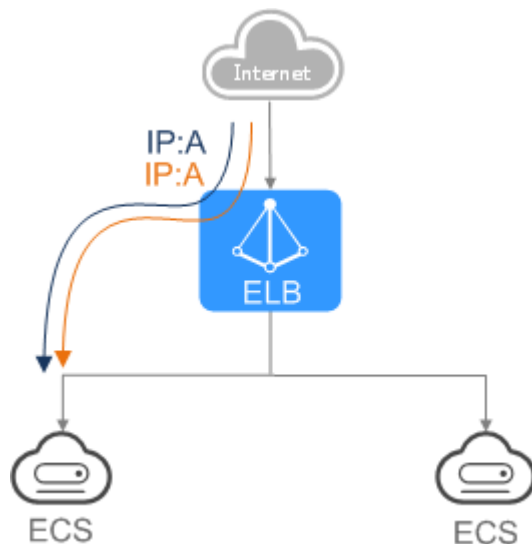
1.4 应用场景

使用 ELB 为高访问量业务进行流量分发

对于业务量访问较大的业务，可以通过ELB设置相应的分配策略，将访问量均匀地分到多个后端服务器处理。例如大型门户网站，移动应用市场等。

同时您还可以开启会话保持功能，保证同一个客户请求转发到同一个后端服务器。从而提升访问效率，如[图1-8](#)所示。

图 1-8 会话保持流量分发

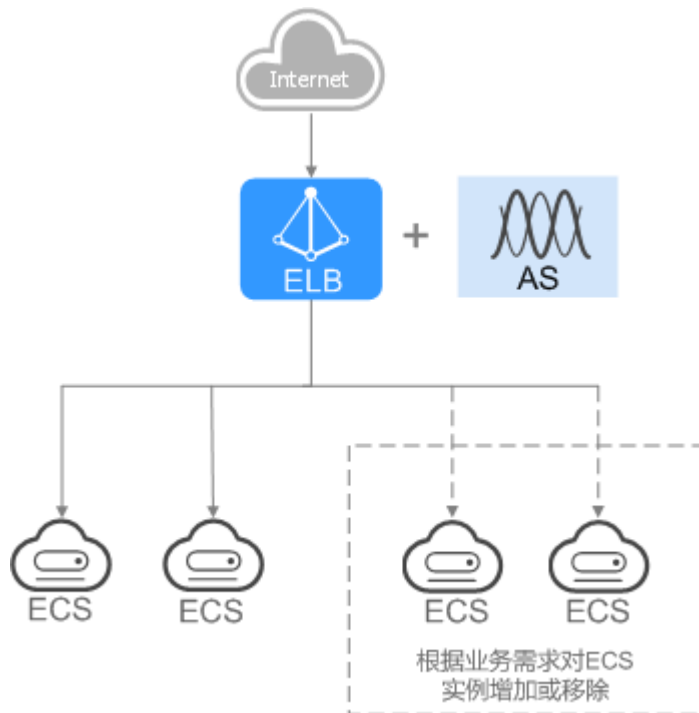


使用 ELB 和 AS 为潮汐业务弹性分发流量

对于存在潮汐效应的业务，结合弹性伸缩服务，随着业务量的增长和收缩，弹性伸缩服务自动增加或者减少的ECS实例，可以自动添加到ELB的后端服务器组或者从ELB的后端服务器组移除。负载均衡实例会根据流量分发、健康检查等策略灵活使用ECS实例

资源，在资源弹性的基础上大大提高资源可用性，如图1-9所示。例如电商的“双11”、“双12”、“618”等大型促销活动，业务的访问量短时间迅速增长，且只持续短暂的几天甚至几小时。使用负载均衡及弹性伸缩能最大限度地节省IT成本。

图 1-9 灵活扩展

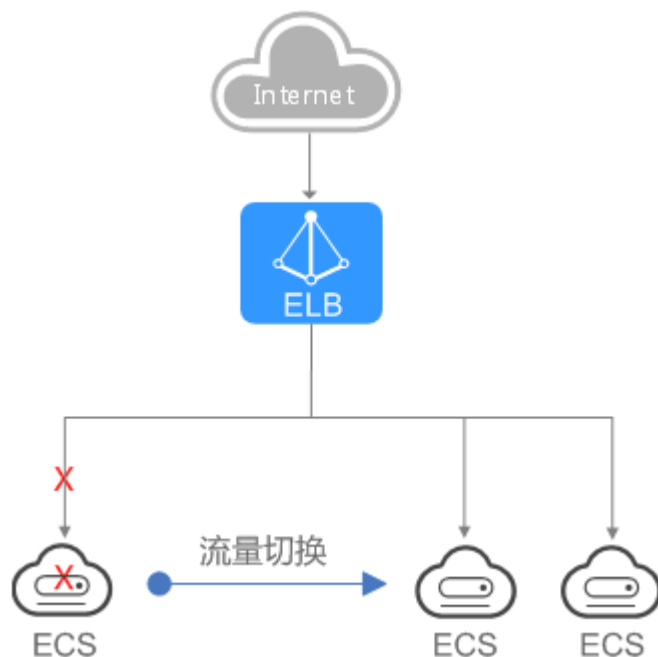


使用 ELB 消除单点故障

对可靠性有较高要求的业务，可以在负载均衡器上添加多个后端服务器。负载均衡器会通过健康检查及时发现并屏蔽有故障的服务器，并将流量转发到其他正常运行的后端服务器，确保业务不中断，如图1-10所示。

例如官网，计费业务，Web业务等。

图 1-10 消除单点故障

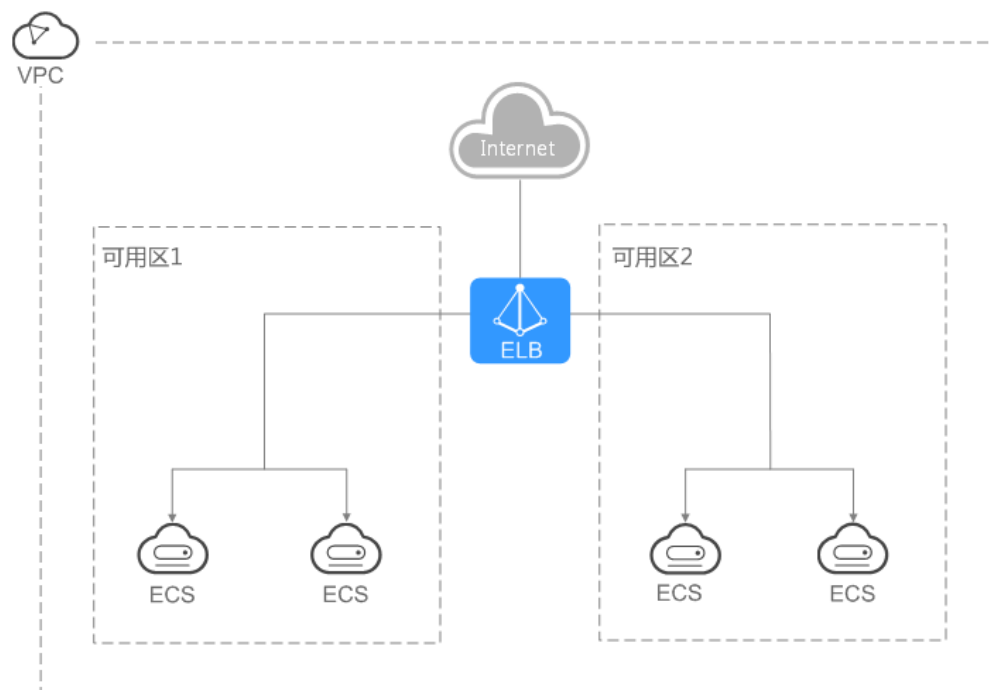


使用 ELB 跨可用区特性实现业务容灾部署

对可靠性和容灾有很高要求的业务，弹性负载均衡可将流量跨可用区进行分发，建立实时的业务容灾部署。即使出现某个可用区网络故障，负载均衡器仍可将流量转发到其他可用区的后端服务器进行处理，如图1-11所示。

例如银行业务，警务业务，大型应用系统等。

图 1-11 多可用区部署



1.5 产品功能

1.5.1 弹性负载均衡产品类型简介

产品简介

弹性负载均衡（Elastic Load Balance，简称ELB）是将访问流量根据分配策略分发到后端多台服务器的流量分发控制服务。弹性负载均衡可以通过流量分发扩展应用系统对外的服务能力，同时通过消除单点故障提升应用系统的可用性。

产品类型

弹性负载均衡服务提供**独享型负载均衡**和**共享型负载均衡**两种类型的实例供用户选择。

表 1-9 产品类型对比

产品类型	独享型负载均衡	共享型负载均衡
部署模式	负载均衡实例资源独享，性能不受其它实例的影响，您可根据业务需要选择不同规格的实例。	集群部署，实例资源共享，支持性能保障模式。
实例规格	<ul style="list-style-type: none">弹性规格：按弹性规格的实际使用量计费。固定规格：提供多种规格供选择，不同固定规格的实例提供差异化的性能指标。 详情请参见 独享型负载均衡实例规格 。	不涉及。
性能上限	单实例单可用区最高支持2千万并发连接，选择多个可用区后，对应的最高性能规格（新建连接数/并发连接数等）加倍。 例如：单实例单AZ最高支持2千万并发连接，那么单实例双AZ最高支持4千万并发连接。	<ul style="list-style-type: none">未开启性能保障模式的实例，不提供性能保障。开启性能保障模式后，提供并发连接数5万、每秒新建连接数5000、每秒查询速率5000的保障能力，超出部分无法保障性能。

产品类型	独享型负载均衡	共享型负载均衡
可用区	<p>支持自定义可用区。</p> <ul style="list-style-type: none"> 对于公网访问，会根据源IP的不同将流量分配到创建的多个AZ中的ELB上，多个AZ的ELB性能加倍。 对于内网访问： <ul style="list-style-type: none"> 当从创建ELB的AZ发起访问时，流量将被分配到本AZ中的ELB上，当本AZ的ELB不可用时，容灾到创建的其他AZ的ELB上。如果本AZ的ELB正常，但是本AZ的流量超过规格，业务也会受影响，内网场景要考虑客户端访问的均衡性。内网流量使用率建议通过AZ粒度监控观察是否超限。 当从未创建ELB的AZ访问时，会根据源IP的不同将流量分配到创建的多个AZ中的ELB上。 对于云专线访问，流量优先分配到专线对接的AZ下部署的ELB，否则分配到其他AZ下的ELB。 对于客户端跨VPC访问，流量优先分配至源VPC子网所在AZ部署的ELB，否则分配到其他AZ下的ELB。 	不涉及。
计费项	<ul style="list-style-type: none"> 固定规格：实例规格费（按固定规格折算LCU数量收取）。 弹性规格：实例费和实际使用的LCU费用。 	性能保障模式下收取实例费用。

产品关键功能对比项

表 1-10 产品类型对比

对比项	独享型负载均衡	共享型负载均衡
产品定位	具备强大的四层和七层处理能力，支持多种应用协议和高级转发策略。	基础的四层与七层能力。

对比项	独享型负载均衡	共享型负载均衡
推荐应用场景	大流量高并发的业务场景，如大型网站、云原生应用、车联网、多可用区容灾应用。	流量负载较低的业务场景，如小型网站和普通高可用应用。
前端协议	TCP、UDP、HTTP、HTTPS、QUIC、TLS。	TCP、UDP、HTTP、HTTPS。
后端协议	TCP、UDP、HTTP、HTTPS、QUIC、TLS、GRPC。	TCP、UDP、HTTP。
转发能力对比	<p>具备强大的四层和七层处理能力，详情参见高级转发策略。</p> <ul style="list-style-type: none"> 支持基于域名、路径、HTTP请求方法、HTTP请求头、查询字符串、网段的转发规则。 支持转发至后端服务器组、重定向至监听器、重定向至URL、重写、返回固定响应。 	<p>支持基础的四层与七层处理能力，详情参见转发策略（共享型）。</p> <ul style="list-style-type: none"> 仅支持基于域名或路径的转发规则。 仅支持转发至后端服务器组、重定向至监听器。
后端服务器组关键功能	<ul style="list-style-type: none"> 健康检查 会话保持 慢启动 支持被多个ELB实例/监听器重复使用 	<ul style="list-style-type: none"> 健康检查 会话保持 仅支持被一个监听器使用
后端服务器组流量分配策略	<ul style="list-style-type: none"> 加权轮询算法 加权最少连接 源IP算法 连接ID算法 	<ul style="list-style-type: none"> 加权轮询算法 加权最少连接 源IP算法
后端服务器组转发模式	<ul style="list-style-type: none"> 负载均衡 主备转发 	负载均衡
后端服务器类型对比	<ul style="list-style-type: none"> 弹性云服务器 IP类型后端 辅助弹性网卡 裸金属服务器 CCE 集群 	<ul style="list-style-type: none"> 弹性云服务器 裸金属服务器 CCE 集群

1.5.2 弹性负载均衡功能对比

协议对比

表 1-11 负载均衡器支持的协议对比

协议类型	描述	独享型负载均衡	共享型负载均衡
四层（TCP/UDP）协议	网络型负载均衡支持TCP和UDP协议，监听器收到访问请求后，将请求直接转发给后端服务器。 转发效率高，数据传输快。	√	√
七层（HTTP/HTTPS）协议	应用型负载均衡支持HTTP和HTTPS协议，监听器收到访问请求后，需要识别并通过HTTP/HTTPS协议报文头中的相关字段，进行数据的转发。 支持加密传输、基于Cookie的会话保持等高级功能。	√	√
全链路HTTPS协议	监听器的前端协议选择“HTTPS”，后端服务器组的后端协议也支持选择“HTTPS”。	√	×
TLS协议	网络型负载均衡：支持TLS协议，适用于需要超高性能和大规模TLS卸载的场景。	√	×
QUIC协议	前端协议为QUIC和UDP的监听器，支持QUIC（Quick UDP Internet Connection）作为后端协议。配合连接ID算法，将同一个连接ID的请求转发到后端服务器。 使用QUIC协议的监听器具有低延迟、高可靠和无队头阻塞的优点，非常适合移动互联网应用、支持在WIFI和运营商网络中无缝切换，而不用重新建立连接。	√	×
HTTP/2协议	HTTP/2，即超文本传输协议HTTP2.0，向下兼容HTTP1.X协议版本，同时性能和安全性都得到了提升。 此功能目前仅支持协议类型为HTTPS的监听器。	√	√

协议类型	描述	独享型负载均衡	共享型负载均衡
GRPC协议	GRPC是一种高性能、通用的RPC开源软件框架，支持弹性负载均衡实现全链路HTTP/2通信。 仅支持前端协议为HTTPS的监听器，开启HTTP/2功能后，选择GRPC协议作为后端协议。	√	×
WebSocket协议	WebSocket (WS)是HTML5一种新的协议，实现了浏览器与服务器全双工通信，能更好地节省服务器资源和带宽并达到实时通讯。	√	√

网络设置对比

表 1-12 负载均衡器支持的网络设置对比

网络设置	描述	独享型负载均衡	共享型负载均衡
IPv4公网	支持通过公网IPv4地址对外提供服务，将来自公网的客户端请求按照指定的负载均衡策略分发到后端服务器进行处理。	√	√
IPv4私网	支持通过私网IPv4地址对外提供服务，将来自同一个VPC的客户端请求按照指定的负载均衡策略分发到后端服务器进行处理。	√	√
IPv6网络	支持负载均衡转发来自IPv6客户端的请求。	√	×
修改IPv4私有地址	可以将负载均衡当前使用IPv4私有IP修改为当前子网或者其它子网的目标IP地址。	√	×
绑定/解绑EIP	根据业务需要为负载均衡实例绑定EIP地址，或者将负载均衡实例已经绑定的IP地址进行解绑。	√	√
修改公网带宽	当通过公网带宽提供负载均衡器和公网之间的访问流量时，可以按照实际需求更改ELB实例关联弹性公网IP的公网带宽。	√	√

监听器的关键功能对比

表 1-13 监听器的关键功能对比

监听器功能	描述	独享型负载均衡	共享型负载均衡
全端口监听	全端口监听器对负载均衡IP地址上的所有端口（1-65535）进行监听，并将监听端口上接收到的请求转发到后端服务器的后端端口。 仅前端协议为TCP和UDP协议的监听器支持全端口监听。	√	×
访问控制	通过添加白名单和黑名单的方式控制访问负载均衡监听器的IP。 <ul style="list-style-type: none"> 通过白名单能够设置允许特定IP访问，而其它IP不许访问。 通过黑名单能够设置允许特定的IP不能访问，而其它IP允许访问。 	√	√
HTTPS双向认证	负载均衡实例与访问用户互相提供身份认证，从而允许通过认证的客户端访问负载均衡实例。 仅前端协议为HTTPS协议的监听器支持双向认证。	√	√
SNI多域名证书	SNI（Server Name Indication）是为解决一个服务器使用多个域名和证书的TLS扩展。开启SNI之后，用户需要添加域名对应的证书。 仅前端协议为HTTPS协议的监听器支持开启SNI。	√	√
获取客户端IP	后端服务器可以获取客户端真实IP地址。 独享型负载均衡默认开启，不支持关闭。	√	√
HTTP/HTTPS监听器的高级配置			
默认安全策略	使用安全策略提高业务的安全性，安全策略包含TLS协议版本和配套的加密算法套件。 仅前端协议为HTTPS协议的监听器支持设置安全策略。	√	√
自定义安全策略	支持选择TLS协议版本和加密算法套件，创建用户自定义的安全策略。 仅前端协议为HTTPS协议的监听器支持设置安全策略。	√	×

监听器功能	描述	独享型负载均衡	共享型负载均衡
获取弹性公网IP	通过X-Forwarded-ELB-IP头字段获取负载均衡实例的公网IP地址。	√	√
获取负载均衡实例ID	通过X-Forwarded-ELB-ID头字段获取负载均衡实例的ID。	√	×
获取监听器端口号	通过X-Forwarded-Port头字段获取ELB实例监听器端口号。	√	×
获取客户端请求端口号	通过X-Forwarded-For-Port头字段获取客户端请求端口号。	√	×
重写X-Forwarded-Host	ELB以客户端请求头的Host重写X-Forwarded-Host传递到后端服务器。	√	√
重写X-Forwarded-Proto	ELB以监听器的前端协议重写X-Forwarded-Proto头字段传递到后端服务器。	√	×
重写X-Real-IP	ELB以客户端的源IP地址重写X-Real-IP传递到后端服务器。	√	×

转发能力对比

HTTP和HTTPS监听器支持设置转发策略，共享型负载均衡支持基础的转发策略，独享型负载均衡支持开启[高级转发策略](#)。

转发策略支持分别设置转发规则和转发动作，详细对比见[表1-14](#)和[表1-15](#)。

表 1-14 支持的转发规则对比

转发规则	描述	独享型负载均衡	共享型负载均衡
域名	触发转发的域名，仅支持精确域名。	√	√
路径	触发转发的路径。 路径的匹配规则有：精确匹配、前缀匹配、正则匹配。	√	√
HTTP请求方法	触发转发的HTTP请求方法。 主要有：GET、POST、PUT、DELETE、PATCH、HEAD、OPTIONS。	√	×
HTTP请求头	触发转发的HTTP请求头。 请求头是键值对的形式，需要分别设置值。	√	×

转发规则	描述	独享型负载均衡	共享型负载均衡
查询字符串	当请求中的字符串与设置好的转发策略中的字符串相匹配时，触发转发。	√	×
网段	触发转发的请求网段。	√	×

表 1-15 支持的转发动作对比

转发动作	描述	独享型负载均衡	共享型负载均衡
转发至后端服务器组	如果满足转发规则，则将请求转发至配置好的后端服务器组。	√	√
重定向至监听器	如果满足转发规则，则将请求转发至配置好的HTTPS监听器上。	√	√
添加重定向至URL	如果满足转发规则，则将请求重定向至配置好的URL。 客户端访问ELB网址A后，ELB返回302或者其他3xx返回码和目的网址B，客户端自动跳转到网址B，网址B可自定义。	√	×
返回固定响应	如果满足转发规则，则返回固定响应。 用户访问ELB实例后，ELB直接返回响应，不向后端服务器继续转发，返回响应的状态码和内容可以自定义。	√	×
转发动作（可选）			
重写	如果满足转发规则的条件，则将请求重写为配置好的URL后再访问后端服务器组。	√	×
写入Header	如果满足转发规则的条件，则将在请求中写入配置的Header后再访问后端服务器组。 输入头字段名称和头字段内容，将覆盖请求中的头变量。	√	×
删除Header	如果满足转发规则的条件，则将在请求中删除配置的Header后再访问后端服务器组。 输入Header头字段名称，将删除请求Header中对应的键值对内容。	√	×

转发动作	描述	独享型负载均衡	共享型负载均衡
限速	转发动作转发至后端服务器组和返回固定响应支持设置限速。 请求速率超过设置的限速后，新建连接请求将被丢弃，并会返回给客户端503状态码。	√	×
跨域	支持添加跨域资源共享CORS（Cross-Origin Resource Sharing）标头以允许浏览器跨域访问 Web应用程序。	√	×
流量镜像至后端服务器组	将转发至后端服务器组的流量请求镜像到选择的后端服务器组，适用于网络流量检查、审计分析以及问题定位等场景。 添加的其他可选转发动作也会对镜像到后端服务器组中的流量请求生效。	√	×

后端服务器组关键功能对比

表 1-16 后端服务器组的关键功能对比

后端服务器组功能	描述	独享型负载均衡	共享型负载均衡
后端服务器组重复使用	同一企业项目下，一个后端服务器组可关联至多个负载均衡实例和监听器使用。	√	×
健康检查	负载均衡器会定期向后端服务器发送请求以测试其运行状态，这些测试称为健康检查。通过健康检查来判断后端服务器是否可用。	√	√
会话保持	会话保持功能可以识别客户与服务器之间交互过程的关联性，在做负载均衡的同时，还保证一系列相关联的访问请求会保持分配到同一台服务器上。	√	√
慢启动	负载均衡器在慢启动时间内向组内新增的后端服务器线性增加请求分配权重。 慢启动能够实现业务的平滑启动，避免业务抖动问题。	√	×

后端服务器组功能	描述	独享型负载均衡	共享型负载均衡
主备转发	当主机健康检查结果正常时，负载均衡将流量转发至主机；当主机健康检查结果异常时，流量将被切换至备机。 必须向后端服务器组中添加两个后端服务器，一个为主服务器，一个为备服务器。	√	×
全端口转发	后端服务器组添加后端服务器时无需指定后端端口，监听器将按照前端请求端口转发流量至后端服务器对应的端口。 仅TCP、UDP和QUIC类型的后端服务器组支持全端口转发功能。	√	×
延迟注销	负载均衡器停止向移除的后端云服务器或者健康检查失败的后端云服务器发送新的请求，保持现有连接在延迟注销时间内正常传输。	√	×

后端服务器组的流量分配策略对比

表 1-17 后端服务器组的流量分配策略对比

分配策略类型	描述	独享型负载均衡	共享型负载均衡
加权轮询算法	当后端服务器的权重相同情况下，将按照简单的轮询策略分发请求。	√	√
加权最少连接	将请求分发给（当前连接/权重）比值最小的后端服务器进行处理。	√	√
源IP算法	后端服务器的权重属性不再生效，在一段时间内，同一个客户端的IP地址的请求会被分发至同一个后端服务器上。	√	√
连接ID算法	利用报文里的连接ID字段进行一致性hash算法，按照运算结果将请求分发到后端服务器上。	√	×

后端服务器类型对比

表 1-18 支持的后端服务器类型对比

后端类型	描述	独享型负载均衡	共享型负载均衡
IP类型后端	后端服务器组不仅支持添加云上VPC内的服务器，还支持添加其他VPC、其他Region、云下数据中心的服务器的IP地址，帮助用户根据业务诉求灵活配置后端服务器。	√	×
辅助弹性网卡 (SubENI)	后端服务器支持绑定辅助弹性网卡。	√	×
弹性云服务器 (ECS)	后端服务器支持弹性云服务器云主机。	√	√
裸金属服务器 (BMS)	后端服务器支持裸金属服务器物理机。	√	√
CCE 集群	支持通过弹性负载均衡访问CCE集群，详情请参见《云容器引擎用户指南》。	√	√

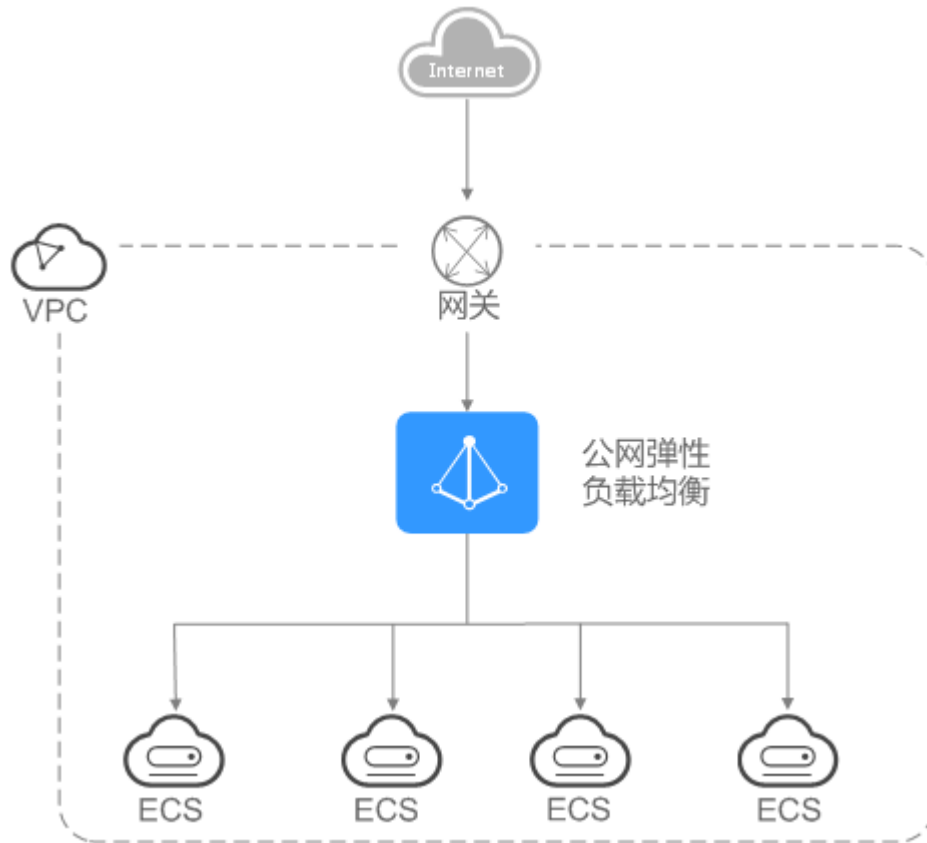
1.6 公网和私网负载均衡器

负载均衡按照支持的网络类型的不同分为公网负载均衡器和私网负载均衡器。

公网负载均衡器

通过给负载均衡器绑定弹性公网IP，使其支持转发公网流量请求，称为公网负载均衡器。通过公网IP对外提供服务，将来自公网的客户端请求按照指定的负载均衡策略分发到后端服务器进行处理。

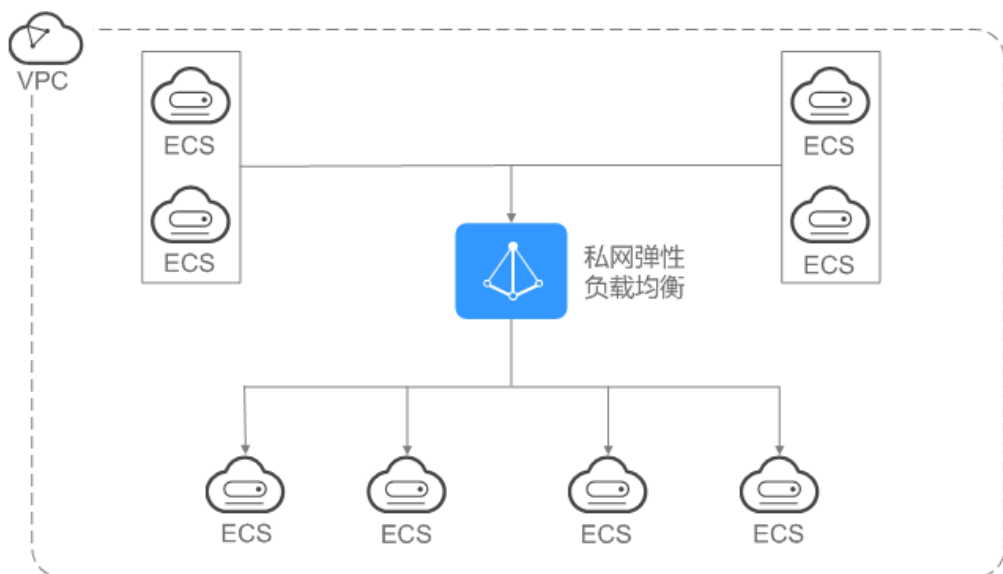
图 1-12 公网负载均衡器



私网负载均衡器

通过给负载均衡器绑定私网IP，使其支持转发私网流量请求，称为私网负载均衡器。通过私网IP对外提供服务，将来自同一个VPC的客户端请求按照指定的负载均衡策略分发到后端服务器进行处理。

图 1-13 私网负载均衡器



实例规格类型与公网/私网负载均衡器的对应关系

表 1-19 独享型负载均衡与公网/私网负载均衡器的对应关系

实例规格类型	网络类型	对应关系
独享型负载均衡	IPv4公网	ELB绑定弹性公网IP，支持IPv4公网流量请求，称为 公网负载均衡器 。
	IPv4私网	ELB绑定私网IP，支持IPv4私网流量请求，称为 私网负载均衡器 。
	IPv6网络	既支持 IPv6公网请求 又支持 IPv6私网请求 。 <ul style="list-style-type: none">ELB绑定弹性公网IP，支持IPv6公网流量请求，称为公网负载均衡器。ELB绑定私网IP，支持IPv6私网流量请求，称为私网负载均衡器。

表 1-20 共享型负载均衡与公网/私网负载均衡器的对应关系

实例规格类型	网络类型	对应关系
共享型负载均衡	IPv4公网	ELB绑定弹性公网IP，支持公网流量请求，称为 公网负载均衡器 。
	IPv4私网	ELB绑定私网IP，支持私网流量请求，称为 私网负载均衡器 。 说明 共享型负载均衡默认支持IPv4私网，且不支持修改。

1.7 ELB 网络流量路径说明

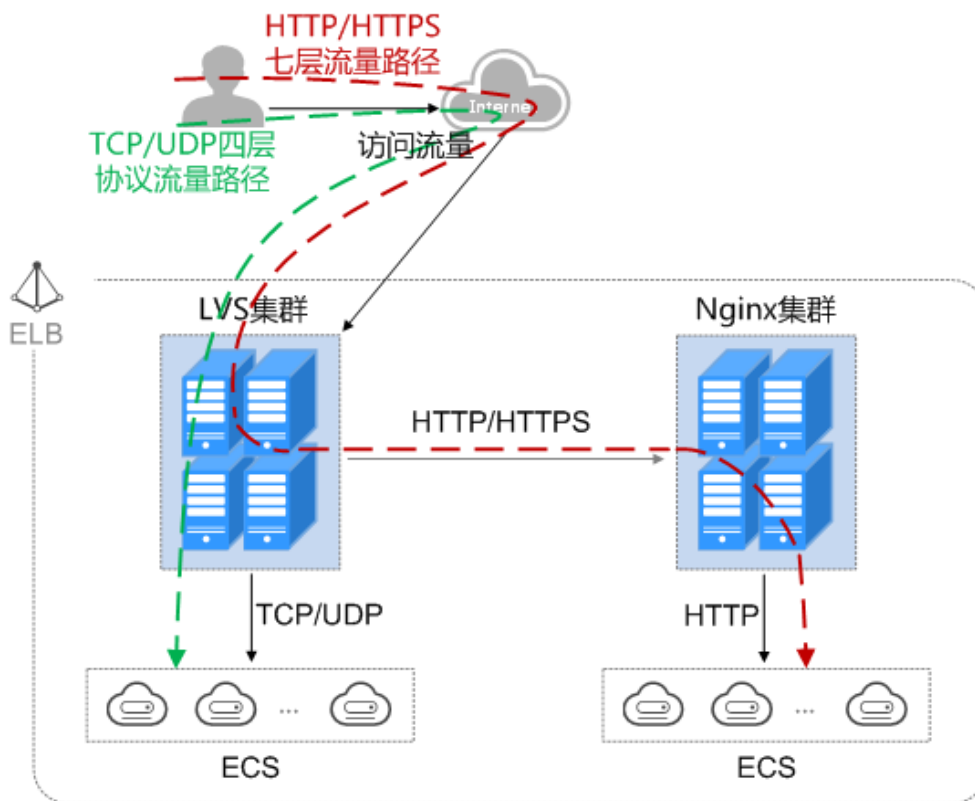
负载均衡将来自客户端的请求通过负载均衡器分发至后端服务器，后端服务器再将响应通过内网返回给负载均衡。负载均衡器和后端服务器之间是通过内网进行通信的。

- 如果负载均衡器后端服务器仅处理来自负载均衡的访问请求，服务器可以不购买EIP或者NAT网关等服务，仅有私网IP即可。
- 如果负载均衡器后端服务器还需要直接对公网提供服务，或者需要访问公网资源，则服务器需要购买EIP或者NAT网关等服务。

入网流量路径

对于入网流量，负载均衡会根据用户配置的流量分配策略，对来自公网或者私网的访问请求进行转发和处理。如图1-14所示。

图 1-14 入网流量路径



当负载均衡器使用四层协议TCP/UDP时:

- 四层协议TCP/UDP的流量只经过LVS集群进行转发。
- LVS集群的所有节点会根据负载均衡器的流量分配策略，将接收到的访问请求直接分发到后端服务器。

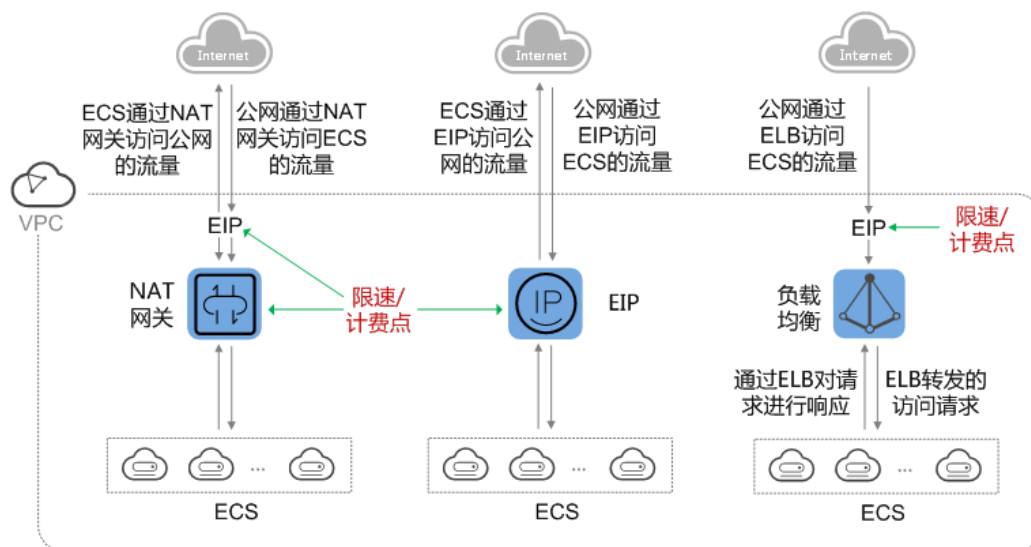
当负载均衡器使用七层协议HTTP/HTTPS时:

- 七层协议HTTP/HTTPS的流量，需要经过LVS集群先将访问请求平均分发到Nginx集群的所有节点，然后Nginx集群的节点再根据负载均衡器的转发策略，将接收到的请求最终分发到后端服务器。
- 七层协议HTTPS的流量，在最终分发到服务器前，还需要在Nginx集群内进行证书验证以及数据包的解密操作。然后通过HTTP协议将请求分发到后端服务器。

出网流量路径

出网流量遵循请求从哪进来，响应从哪出去的原则。如出网流量路径所示。

图 1-15 出网流量路径



- 通过负载均衡器进入的访问流量，对应的响应流量通过负载均衡器返回。
由于负载均衡器实际是通过绑定的EIP接收来自公网的流量和响应请求，所以负载均衡器的限制实际是在负载均衡器绑定的EIP上，并在EIP上进行计费。从负载均衡器到后端云服务器之间通过VPC内网进行通信，不收取费用。
- 通过NAT网关进入的访问流量，对应的响应流量通过NAT网关返回。在NAT网关上限速和计费。
由于NAT网关实际是通过绑定的EIP接收来自公网的流量和访问公网，所以NAT网关上进行的是连接数的限制，带宽或者流量的限制是在NAT网关绑定的EIP上，并分别在NAT网关和弹性公网IP上进行计费。
- 通过EIP进入的访问流量，对应的响应流量通过EIP返回，在EIP上限速和计费。

1.8 独享型负载均衡实例规格

独享型负载均衡支持按弹性规格和固定规格两种规格进行购买，两种规格对比项如表 1-21，请您根据自身业务规划选择实例规格。

表 1-21 独享型负载均衡规格对比

规格对比项	弹性规格	固定规格
适用场景	<ul style="list-style-type: none"> • 业务用量波动较大的场景 • 资源使用具有临时性和突发性 	<ul style="list-style-type: none"> • 业务用量较为稳定的场景 • 资源长期稳定使用
网络型（TCP/UDP/TLS）性能上限	实例性能随可用区数量叠加，单可用区实例性能上限见表1-23。	实例性能随可用区数量叠加，单可用区实例有对应固定规格性能上限，详情见表1-26。

规格对比项	弹性规格	固定规格
应用型（HTTP/HTTPS/QUIC）性能上限	实例性能随可用区数量叠加，单可用区实例性能上限见 表1-23 。	实例性能随可用区数量叠加，单可用区实例有对应固定规格性能上限，详情见 表1-27 。
支持计费模式	按需计费	<ul style="list-style-type: none"> • 按需计费 • 包年/包月
计费项	<ul style="list-style-type: none"> • LCU费用 • 实例费用 	LCU费用
产品能力	无差异	

📖 说明

在私网访问场景，当从创建ELB的AZ发起访问时，流量将被分配至本AZ中的ELB上，当本AZ的ELB不可用时，容灾切换到创建的其他AZ的ELB上。

如果本AZ的ELB正常，但是本AZ的流量超过规格，此时业务也会受影响，因此私网场景要考虑客户端访问的均衡性。

建议通过[可用区的监控](#)来观察私网流量是否超过性能上限。

弹性规格

弹性规格适用于具有周期性或波动较大的业务，例如游戏、视频等行业。弹性规格提供了网络型（TCP/UDP/TLS）和应用型（HTTP/HTTPS/QUIC）两种规格类型，请根据自身业务规划选择实例。

📖 说明

选定对应的规格类型的负载均衡实例才可以创建该规格类型的监听器。

弹性规格支持的性能指标如[表1-22](#)，实际业务超过弹性规格性能上限如[表1-23](#)时，会导致新建连接请求受限以及丢包问题。

表 1-22 弹性规格的性能指标

最大并发连接数-Max Connection	最大并发连接数是指一个负载均衡实例每分钟平均能够承载的最大连接数量。当实例上的连接数超过弹性规格上限的最大连接数时，为了保障已有的连接业务性能，新建连接请求将被丢弃。
每秒新建连接数-Connection Per Second (CPS)	每秒新建连接数是指新建连接的速率。当新建连接的速率超过弹性规格上限的每秒新建连接数时，为了保障已有的连接业务性能，新建连接请求将被丢弃。
每秒查询速率-Query Per Second (QPS)	每秒查询速率是指仅在七层监听时，每秒可以处理的HTTP/HTTPS的查询请求的数量。当请求速率超过弹性规格上限的每秒查询速率时，为了保障已有的连接业务性能，新建连接请求将被丢弃。

带宽 (Mbit/S)	每秒带宽可以保障带宽的性能。
---------------	----------------

表 1-23 弹性规格性能指标上限

协议类型	最大并发连接数	新建连接数 (CPS)	每秒查询速率 (QPS)	带宽 (Mbit/S)
网络型 (TCP/UDP)	20,000,000	400,000	-	10,000
网络型 (TLS)	20,000,000	20,000	-	10,000
应用型 (HTTP)	8,000,000	80,000	160,000	10,000
应用型 (HTTPS)	8,000,000	80,000	160,000	10,000

⚠ 注意

弹性规格因各地域资源情况不同，开放的弹性规格上限可能略有差异，请以控制台创建页为准。

固定规格

固定规格实例的关键指标如下，请根据自身业务规划选择实例规格。实际业务超过固定规格限定时，会导致新建连接请求受限以及丢包问题。

表 1-24 固定规格的性能指标

最大并发连接数-Max Connection	最大并发连接数是指一个负载均衡实例每分钟平均能够承载的最大连接数量。当实例上的连接数超过规格定义的最大并发连接数时，为了保障已有的连接业务性能，新建连接请求将被丢弃。
每秒新建连接数-Connection Per Second (CPS)	每秒新建连接数是指新建连接的速率。当新建连接的速率超过规格定义的每秒新建连接数时，为了保障已有的连接业务性能，新建连接请求将被丢弃。 对于七层监听器，HTTPS监听器在建立连接时，使用SSL握手会占用更多系统资源。如小型I-应用型 (HTTP/HTTPS) 的HTTP每秒新建连接数为2000，HTTPS每秒新建连接数为200，计算示例请参见 表1-25 。
每秒查询速率-Query Per Second (QPS)	每秒查询速率是指仅在七层监听时，每秒可以处理的HTTP/HTTPS的查询请求的数量。当请求速率超过规格所定义的每秒查询速率时，为了保障已有的连接业务性能，新建连接请求将被丢弃。

带宽 (Mbit/S)	每秒带宽可以保障带宽的性能。
--------------------	----------------

以小型I-应用型 (HTTP/HTTPS) 规格为例进行说明。

- 假如该规格实例仅使用HTTP监听器业务，HTTP新建连接数的上限为2000。
- 假如该规格实例仅使用HTTPS监听器业务，HTTPS新建连接数的上限为200。
- 假如实例同时使用HTTP与HTTPS监听器业务，需按叠加公式计算得出新建连接数，其上限为规格定义的HTTP新建连接数。

叠加公式：新建连接数 = 实际HTTP新建连接数 + 实际HTTPS新建连接数 * 规格中HTTP与HTTPS的比值

小型I-应用型 (HTTP/HTTPS) 规格下，HTTP与HTTPS的比值为10，计算详情参见表1-25。

表 1-25 小型 I-应用型 (HTTP/HTTPS) 规格新建连接数计算示例

参数	场景一	场景二
实际HTTP新建连接数	1000	1000
实际HTTPS新建连接数	50	150
计算所得新建连接数	$1000+50*10=1500$	$1000+150*10=2500$
新建连接请求说明	<ul style="list-style-type: none"> • 计算所得新建连接数不超过规格定义的HTTP新建连接数 • 新建连接请求完整 	<ul style="list-style-type: none"> • 计算所得新建连接数超过规格定义的HTTP新建连接数 • 产生新建连接请求丢弃

说明

表1-25中的场景参数仅为示例参考。

独享型负载均衡开放的实例规格，如表1-26和表1-27所示。

注意

- 各地域因资源情况不同，开放的规格可能略有差异，请以控制台创建页为准。
- 选定对应的规格类型的负载均衡实例才可以创建该规格类型的监听器。

表 1-26 固定规格-网络型(TCP/UDP/TLS)

规格类型	最大并发连接数 (TCP/UDP)	最大并发连接数 (TLS)	新建连接数(CPS) (TCP/UDP)	新建连接数(CPS) (TLS)	带宽 (Mbit/S)	折算LCU数 (个/AZ)
小型 I	500,000	30,000	10,000	500	50	10
小型 II	1,000,000	60,000	20,000	1,000	100	20
中型 I	2,000,000	120,000	40,000	2,000	200	40
中型 II	4,000,000	240,000	80,000	4,000	400	80
大型 I	10,000,000	600,000	200,000	10,000	1,000	200
大型 II	20,000,000	1,200,000	400,000	20,000	2,000	400

表 1-27 固定规格-应用型(HTTP/HTTPS)

规格类型	最大并发连接数	新建连接数(CPS) (HTTP)	新建连接数(CPS) (HTTPS)	每秒查询速率(QPS) (HTTP)	每秒查询速率(QPS) (HTTPS)	带宽 (Mbit/S)	折算LCU数 (个/AZ)
小型 I	200,000	2,000	200	4,000	2,000	50	10
小型 II	400,000	4,000	400	8,000	4,000	100	20
中型 I	800,000	8,000	800	16,000	8,000	200	40
中型 II	2,000,000	20,000	2,000	40,000	20,000	400	100
大型 I	4,000,000	40,000	4,000	80,000	40,000	1,000	200
大型 II	8,000,000	80,000	8,000	160,000	80,000	2,000	400

说明

- 如果一个负载均衡实例下创建了多个监听器，则上述表格中的每秒查询速率（QPS）是指该负载均衡实例下的所有监听器的QPS之和不超过规格所定义的QPS值。
- 带宽规格是指入流量或出流量不超过表中的数值。如：对于小型规格，入流量≤50Mbit/S，出流量≤50Mbit/S。
- 带宽规格是负载均衡实例所能提供的带宽保障范围，保障范围内资源可用；超出保障范围的，无法保障带宽性能。

1.9 约束与限制

本文为您介绍弹性负载均衡资源的使用限制。

弹性负载均衡的服务配额

为防止资源滥用，平台限制了各服务资源的配额，对用户的资源数量和容量做了限制。如您最多可以创建多少台弹性云服务器、多少块云硬盘。

默认资源配额如表1-28，不同用户拥有的实际资源配额略有差异，请参考[怎样查看我的配额？](#)，登录控制台查询您的配额详情。

如果当前资源配额限制无法满足使用需要，您可以参考[如何申请扩大配额？](#)，提升资源配额。

表 1-28 弹性负载均衡的服务配额

资源名称	资源说明	默认配额
弹性负载均衡	一个用户创建弹性负载均衡的数量	50个
弹性负载均衡监听器	一个用户创建监听器的数量	100个
弹性负载均衡转发策略	一个用户创建转发策略的数量	500条
弹性负载均衡后端主机组	一个用户创建转发后端服务器组的数量	500个
弹性负载均衡证书	一个用户拥有弹性负载均衡证书的数量	120个
弹性负载均衡后端服务器	一个用户拥有后端服务器的数量	500个
单负载均衡器可添加监听器数量	一个负载均衡器支持添加监听器的数量	50个
单监听器可添加转发策略数量	一个监听器支持添加转发策略的数量	100个

说明

以上配额说明针对单租户情况。

弹性负载均衡的资源配额

除[弹性负载均衡的服务配额](#)外，弹性负载均衡的使用中还存在部分资源配额限制。

您可以参考[查询配额详情](#)，调用API接口查询负载均衡相关的各类资源配额，如[表 1-29](#)。

表 1-29 弹性负载均衡的资源配额

资源名称	资源说明	默认配额
单转发策略可添加的转发条件数量	一条转发策略支持的转发条件数量	10个
单转发策略可添加的后端服务器组数量	一条转发策略支持的转发至后端服务服务器组数量	5个
单后端服务器组可添加的后端服务器数量	一个后端服务器组支持添加后端服务器的数量	500个
单后端服务器组可关联的监听器数量	一个后端服务器组可关联监听器的数量	50个
IP地址组		
弹性负载均衡IP地址组	一个用户创建IP地址组的数量	50个
单IP地址组可关联监听器	一个IP地址组可关联监听器的数量	50个
单IP地址组可添加IP地址数量	一个IP地址组支持添加IP地址的数量	300个

负载均衡器

- 创建资源前请参考[规划和准备](#)，根据业务需求对负载均衡器的区域、类型、协议以及后端服务器等进行合理规划。
- 负载均衡器对转发数据的限制：
 - 四层监听器：无限制。
 - 七层监听器：
 - 上传文件大小限制为10GB。
 - HTTP请求行加HTTP请求头之和限制为32KB。

监听器

- 为了保持ELB的性能和减少管理复杂度，请您根据业务规模合理评估单ELB实例下配置的监听器数量。如果监听器的默认配额无法满足需求，建议您增加ELB实例的数量。
- 独享型负载均衡的监听器最多与50个后端服务器组关联使用。
- 一个证书最多支持关联到600个监听器。

- SNI证书：
 - 共享型：
 - 一个HTTPS监听器默认支持配置30个SNI证书。
 - 一个证书的域名个数不超过30个，监听器关联的所有SNI证书默认支持的域名总数为30个。
 - SNI证书中单个域名长度不超过100个字符，域名总长度不超过1024个字符。
 - 独享型：
 - 一个HTTPS监听器默认支持配置30个SNI证书，最多支持调整为50个SNI证书。如您有需求，可提交[工单](#)进行处理。
 - 一个证书的域名个数不超过100个，监听器关联的所有SNI证书默认支持的域名总数为200个。
 - SNI证书中单个域名长度不超过100个字符，域名总长度不超过10000个字符。
- 监听器的前端协议和端口设置后不允许修改。

转发策略

- 仅HTTP和HTTPS协议的监听器支持配置转发策略。
- 不支持创建相同的转发策略。
- 单个监听器最多支持配置100条转发策略，超过配额的转发策略不生效。
- 转发条件数量：
 - 未开启高级转发策略：一种转发规则仅支持一个转发条件
 - 高级转发策略：一种转发规则支持多个转发条件，一条转发策略最多支持10个转发条件。

表 1-30 弹性负载均衡的转发策略限制

实例类型	高级转发策略	转发规则	转发动作	更多详情
共享型负载均衡	不支持高级转发策略	域名、路径	转发至后端服务器组、重定向至监听器	转发策略（共享型）
独享型负载均衡	未开启高级转发策略	域名、路径	转发至后端服务器组、重定向至监听器	转发策略（独享型）
	开启高级转发策略	域名、路径、HTTP请求方法、HTTP请求头、查询字符串、网段、Cookie	转发至后端服务器组、重定向至监听器、重定向至URL、返回固定响应、重写、写入Header、删除Header、限速	高级转发策略（独享型）

后端服务器组

- 仅前端协议与后端协议匹配的监听器和后端服务器组才可关联使用，协议匹配关系详见表4 前端/后端协议匹配关系。

表 1-31 前端/后端协议匹配关系

ELB的规格类型	监听器的前端协议	后端服务器组的后端协议
网络型	TCP	TCP
网络型	UDP	<ul style="list-style-type: none">UDPQUIC
网络型	TLS	<ul style="list-style-type: none">TLSTCP
应用型	HTTP	HTTP
应用型	HTTPS	<ul style="list-style-type: none">HTTPHTTPSGRPC
应用型	QUIC	<ul style="list-style-type: none">HTTPHTTPS

后端服务器

- 开启获取客户端IP功能后，不支持同一台服务器既作为后端服务器又作为客户端的场景。
- 同一台ECS可以重复作为后端服务器的次数为800次。

IP 地址组

单监听器的访问控制策略，最多支持选择5个IP地址组。IP地址组内合计最多可添加300个IP地址或网段。

TLS 安全策略

一个用户可以创建50个自定义TLS安全策略。

关联服务

当前独享型ELB支持接入云模式的WAF实例，为ELB实例添加WAF策略配置时有如下限制：

表 1-32 WAF 策略配置

资源名称	数量
防护域名	域名：支持单域名和泛域名。 <ul style="list-style-type: none">● 单域名：<ul style="list-style-type: none">- 输入防护的单域名，例如：www.example.com。- 如果各子域名对应的服务器IP地址不相同：请将子域名按“单域名”方式逐条添加，默认支持30个单域名。● 泛域名：如果各子域名对应的服务器IP地址相同：输入防护的泛域名。例如：子域名a.example.com，b.example.com和c.example.com对应的服务器IP地址相同，可以直接添加泛域名*.example.com，仅支持1个泛域名。
WAF策略绑定的监听器	对弹性负载均衡下的应用型（HTTP/HTTPS/QUIC）监听器进行防护。 <ul style="list-style-type: none">● 所有监听器：防护对应负载均衡实例下的所有监听器。● 指定监听器：最多支持选择5个监听器。

说明

独享型ELB支持接入云模式的WAF实例陆续上线中，已发布区域请以控制台实际为准。

1.10 安全

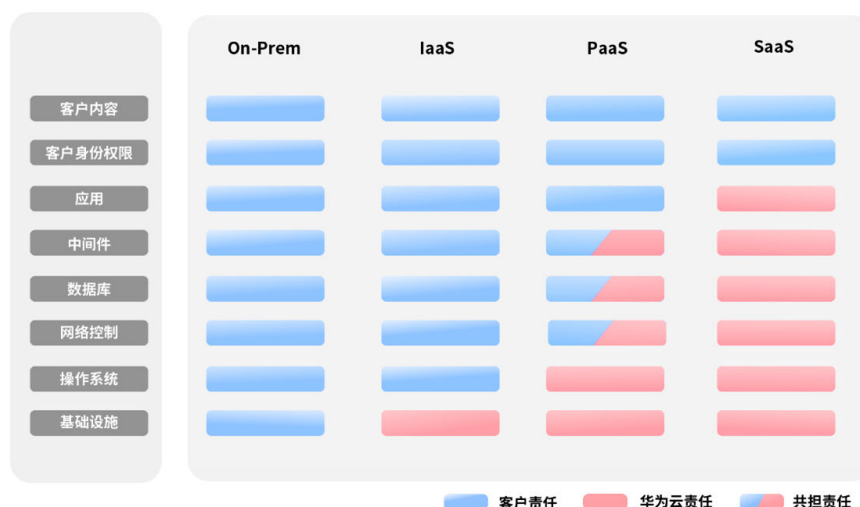
1.10.1 责任共担

华为云秉承“将对网络和业务安全性保障的责任置于公司的商业利益之上”。针对层出不穷的云安全挑战和无孔不入的云安全威胁与攻击，华为云在遵从法律法规业界标准的基础上，以安全生态圈为护城河，依托华为独有的软硬件优势，构建面向不同区域和行业的完善云服务安全保障体系。

与传统的本地数据中心相比，云计算的运营方和使用方分离，提供了更好的灵活性和控制力，有效降低了客户的运营负担。正因如此，云的安全性无法由一方完全承担，云安全工作需要华为云与您共同努力，如图1-16所示。

- **华为云**：无论在任何云服务类别下，华为云都会承担基础设施的安全责任，包括安全性、合规性。该基础设施由华为云提供的物理数据中心（计算、存储、网络等）、虚拟化平台及云服务组成。在PaaS、SaaS场景下，华为云也会基于控制原则承担所提供服务或组件的安全配置、漏洞修复、安全防护和入侵检测等职责。
- **客户**：无论在任何云服务类别下，客户数据资产的所有权和控制权都不会转移。在未经授权的情况下，华为云承诺不触碰客户数据，客户的内容数据、身份和权限都需要客户自身看护，这包括确保云上内容的合法合规，使用安全的凭证（如强口令、多因子认证）并妥善管理，同时监控内容安全事件和账号异常行为并及时响应。

图 1-16 华为云安全责任共担模型



云安全责任基于控制权，以可见、可用作为前提。在客户上云的过程中，资产（例如设备、硬件、软件、介质、虚拟机、操作系统、数据等）由客户完全控制向客户与华为云共同控制转变，这也就意味着客户需要承担的责任取决于客户所选取的云服务。如图1-16所示，客户可以基于自身的业务需求选择不同的云服务类别（例如IaaS、PaaS、SaaS）。不同的云服务类别中，每个组件的控制权不同，这也导致了华为云与客户的责任关系不同。

- 在On-prem场景下，由于客户享有对硬件、软件和数据等资产的全部控制权，因此客户应当对所有组件的安全性负责。
- 在IaaS场景下，客户控制着除基础设施外的所有组件，因此客户需要做好除基础设施外的所有组件的安全工作，例如应用自身的合法合规性、开发设计安全，以及相关组件（如中间件、数据库和操作系统）的漏洞修复、配置安全、安全防护方案等。
- 在PaaS场景下，客户除了对自身部署的应用负责，也要做好PaaS服务中间件、数据库、网络控制的安全配置和策略工作。
- 在SaaS场景下，客户对客户内容、账号和权限具有控制权，客户需要做好自身内容的保护以及合法合规、账号和权限的配置和保护等。

传统本地部署(On-Prem)：由客户在自有数据中心内部署和管理软件及IT基础设施，而非依赖于远程的云服务提供商；

基础设施即服务(IaaS)：由云服务提供商提供计算、网络、存储等基础设施服务，如[弹性云服务器 ECS](#)、[虚拟专用网络 VPN](#)、[对象存储服务 OBS](#)；

平台即服务(PaaS)：由云服务提供商提供应用程序开发和部署所需要的平台，客户无需维护底层基础设施，如[AI开发平台 ModelArts](#)、[云数据库 GaussDB](#)；

软件即服务(SaaS)：由云服务提供商提供完整应用软件，客户直接应用软件而无需安装、维护应用软件及底层平台和基础设施，如[华为云会议 Meeting](#)。

1.10.2 ELB 服务的访问控制

身份认证

弹性负载均衡支持通过IAM权限策略进行访问控制。IAM权限是作用于云资源的，IAM权限定义了允许和拒绝的访问操作，以此实现云资源权限访问控制。管理员创建IAM

用户后，需要将用户加入到一个用户组中，IAM可以对这个组授予ELB所需的权限，组内用户自动继承用户组的所有权限。

详情请参见[权限管理](#)。

访问控制策略

访问控制策略：用户可以通过添加白名单和黑名单的方式控制访问负载均衡监听器的IP。通过白名单能够设置允许特定IP访问，而其它IP不许访问。通过黑名单能够设置允许特定的IP不能访问，而其它IP允许访问，详情参见[访问控制策略](#)。

1.10.3 审计与日志

云审计服务（Cloud Trace Service, CTS），是华为云安全解决方案中专业的日志审计服务，提供对各种云资源操作记录的收集、存储和查询功能，可用于支撑安全分析、合规审计、资源跟踪和问题定位等常见应用场景。

用户开通云审计服务后，ELB可记录ELB的操作事件用于审计。

- CTS的详细介绍和开通配置方法，请参见[CTS入门指引](#)。
- ELB支持审计的操作事件请参见[支持审计的关键操作](#)。
- 查看审计日志请参见[查看审计日志](#)。

1.10.4 监控安全风险

云监控（Cloud Eye）服务是面向华为云资源的监控平台，提供了实时监控、及时告警、资源分组、站点监控等能力，使您全面了解云上资源的使用情况和业务的运行状况。

通过云监控，可以按时间轴查看ELB的网络流量，错误日志相关情况，动态告警分析潜在风险。

关于弹性负载均衡服务支持的监控指标，以及如何创建监控告警规则等内容，请参见[监控弹性负载均衡](#)

1.10.5 认证证书

合规证书

华为云服务及平台通过了多项国内外权威机构（ISO/SOC/PCI等）的安全合规认证，用户可自行[申请下载](#)合规资质证书。

图 1-17 合规证书下载

资源中心

华为云还提供以下资源来帮助用户满足合规性要求，具体请查看[资源中心](#)。

图 1-18 资源中心

合规资质证书

华为云安全服务提供了网络安全专用产品安全检测证书、软件著作权等证书，供用户下载和参考。具体请查看[合规资质证书](#)。

图 1-19 网络安全专用产品安全检测证书&软件著作权证书



1.11 权限管理

如果您需要对华为云上购买的ELB资源，员工设置不同的访问权限，以达到不同员工之间的权限隔离，您可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制华为云资源的访问。如果华为云账号已经能满足您的要求，不需要通过IAM对用户进行权限管理，您可以跳过本章节，不影响您使用ELB服务的其它功能。

IAM是华为云提供权限管理的基础服务，无需付费即可使用，您只需要为您账号中的资源进行付费。

通过IAM，您可以通过授权控制他们对华为云资源的访问范围。例如您的员工中有负责软件开发的人员，您希望他们拥有ELB的使用权限，但是不希望他们拥有删除ELB等高危操作的权限，那么您可以使用IAM进行权限分配，通过授予用户仅能使用ELB，但是不允许删除ELB的权限，控制他们对ELB资源的使用范围。

目前IAM支持两类授权，一类是角色与策略授权，另一类为身份策略授权。

两者有如下的区别和关系：

表 1-33 两类授权的区别

名称	核心关系	涉及的权限	授权方式	适用场景
角色与策略授权	用户-权限-授权范围	<ul style="list-style-type: none"> 系统角色 系统策略 自定义策略 	为主体授予角色或策略	核心关系为“用户-权限-授权范围”，每个用户根据所需权限和所需授权范围进行授权，无法直接给用户授权，需要维护更多的用户组，且支持的条件键较少，难以满足细粒度精确权限控制需求，更适用于对细粒度权限管控要求较低的中小企业用户。

名称	核心关系	涉及的权限	授权方式	适用场景
身份策略授权	用户-策略	<ul style="list-style-type: none"> 系统身份策略 自定义身份策略 	<ul style="list-style-type: none"> 为主体授予身份策略 身份策略附加至主体 	核心关系为“用户-策略”，管理员可根据业务需求定制不同的访问控制策略，能够做到更细粒度更灵活的权限控制，新增资源时，对比角色与策略授权，基于身份策略的授权模型可以更快地直接给用户授权，灵活性更强，更方便，但相对应的，整体权限管控模型构建更加复杂，对相关人员专业能力要求更高，因此更适用于中大型企业。

例如：如果需要对IAM用户授予可以创建华北-北京四区域的ECS和华南-广州区域的OBS的权限，基于角色与策略授权的场景中，管理员需要创建两个自定义策略，并且为IAM用户同时授予这两个自定义策略才可以实现权限控制。在基于身份策略授权的场景中，管理员仅需要创建一个自定义身份策略，在策略中通过条件键“g:RequestedRegion”的配置即可达到策略对于授权区域的控制。将身份策略附加主体或为主体授予该身份策略即可获得相应权限，权限配置方式更细粒度更灵活。

两种授权场景下的策略/身份策略、授权项等并不互通，推荐使用身份策略进行授权。[角色与策略权限管理](#)和[身份策略权限管理](#)分别介绍两种模型的系统权限。

关于IAM的详细介绍，请参见[IAM产品介绍](#)。

角色与策略权限管理

ELB服务支持角色与策略授权。默认情况下，管理员创建的IAM用户没有任何权限，需要将其加入用户组，并给用户组授予策略或角色，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于被授予的权限对云服务进行操作。

ELB部署时通过物理区域划分，为项目级服务。授权时，“授权范围”需要选择“指定区域项目资源”，然后在指定区域（如华北-北京1）对应的项目（cn-north-1）中设置相关权限，并且该权限仅对此项目生效；如果“授权范围”选择“所有资源”，则该权限在所有区域项目中都生效。访问ELB时，需要先切换至授权区域。

如表1-34所示，包括了ELB的所有系统权限。角色与策略授权场景的系统策略和身份策略授权场景的并不互通。

表 1-34 ELB 系统权限

系统角色/策略名称	描述	类别	依赖关系
ELB FullAccess	弹性负载均衡管理员权限，拥有该权限的用户可以操作并使用所有弹性负载均衡。	系统策略	无

系统角色/策略名称	描述	类别	依赖关系
ELB ReadOnlyAccess	弹性负载均衡只读权限，拥有该权限的用户仅能查看弹性负载均衡。	系统策略	无
ELB Administrator	对弹性负载均衡服务的所有执行权限。	系统角色	依赖Tenant Administrator、VPC Administrator、CES Administrator、Server Administrator、Tenant Guest策略，在同项目中勾选依赖的策略。

表1-35列出了ELB常用操作与系统权限的授权关系，您可以参照该表选择合适的系统权限。

表 1-35 常用操作与系统策略的关系

操作	ELB FullAccess	ELB ReadOnlyAccess	ELB Administrator
创建负载均衡器	√	×	√
查询负载均衡器	√	√	√
查询负载均衡器状态树	√	√	√
查询负载均衡器列表	√	√	√
更新负载均衡器	√	×	√
删除负载均衡器	√	×	√
创建监听器	√	×	√
查询监听器	√	√	√
修改监听器	√	×	√
删除监听器	√	×	√
创建后端服务器组	√	×	√
查询后端服务器组	√	√	√
修改后端服务器组	√	×	√

操作	ELB FullAccess	ELB ReadOnlyAccess	ELB Administrator
删除后端服务器组	√	×	√
创建后端服务器	√	×	√
查询后端服务器	√	√	√
修改后端服务器	√	×	√
删除后端服务器	√	×	√
创建健康检查	√	×	√
查询健康检查	√	√	√
修改健康检查	√	×	√
关闭健康检查	√	×	√
创建弹性公网IP	×	×	√
绑定弹性公网IP	×	×	√
查询弹性公网IP	√	√	√
解绑弹性公网IP	×	×	√
查看监控指标	×	×	√
查看访问日志	×	×	√

说明

创建负载均衡器时需注意以下事项：

- 创建包年/包月负载均衡器，还需要配置费用中心的BSS Administrator权限，具体详见《费用中心用户指南》。
- 创建公网弹性负载均衡器，还需要配置VPC服务的vpc:publicips:create和vpc:publicips:update细粒度权限，具体详见《弹性公网IP服务API参考》
- 将按需计费的负载均衡器转为包年/包月计费，还需要配置费用中心的BSS Administrator权限，具体详见《费用中心用户指南》。

身份策略权限管理

ELB服务支持身份策略授权。如表1-36所示，包括了ELB身份策略中的所有系统身份策略。身份策略授权场景的系统身份策略和角色与策略授权场景的并不互通。

表 1-36 ELB 系统身份策略

系统身份策略名称	描述	策略类别
ELBFullAccessPolicy	弹性负载均衡服务所有权限	系统身份策略

系统身份策略名称	描述	策略类别
ELBReadOnlyAccessPolicy	弹性负载均衡服务只读权限	系统身份策略

表1-37列出了ELB常用操作与系统身份策略的授权关系，您可以参照该表选择合适的系统身份策略。

表 1-37 常用操作与系统身份策略的关系

操作	ELBFullAccessPolicy	ELBReadOnlyAccessPolicy
创建负载均衡器	√	×
查询负载均衡器	√	√
查询负载均衡器状态树	√	√
查询负载均衡器列表	√	√
更新负载均衡器	√	×
删除负载均衡器	√	×
创建监听器	√	×
查询监听器	√	√
修改监听器	√	×
删除监听器	√	×
创建后端服务器组	√	×
查询后端服务器组	√	√
修改后端服务器组	√	×
删除后端服务器组	√	×
创建后端服务器	√	×
查询后端服务器	√	√
修改后端服务器	√	×
删除后端服务器	√	×
创建健康检查	√	×
查询健康检查	√	√
修改健康检查	√	×
关闭健康检查	√	×

操作	ELBFullAccessPolicy	ELBReadOnlyAccessPolicy
创建弹性公网IP	×	×
绑定弹性公网IP	×	×
查询弹性公网IP	√	√
解绑弹性公网IP	×	×
查看监控指标	×	×
查看访问日志	×	×
查询企业项目列表	√	√

相关链接

- [IAM产品介绍](#)
- [通过IAM进行授权通过角色或策略授予使用ELB的权限](#)
- [权限与授权项说明](#)

1.12 基本概念

1.12.1 产品基本概念

表 1-38 弹性负载均衡基本概念

名词	说明
负载均衡器	负载均衡器是指您创建的承载业务的负载均衡服务实体。
监听器	监听器负责监听负载均衡器上的请求，根据配置的流量分配策略，分发流量到后端云服务器处理。
后端服务器	负载均衡器会将客户端的请求转发给后端服务器处理。例如，您可以添加ECS实例作为负载均衡器的后端服务器，监听器使用您配置的协议和端口检查来自客户端的连接请求，并根据您定义的分配策略将请求转发到后端服务器组里的后端服务器。
后端服务器组	把具有相同特性的后端服务器放在一个组，负载均衡实例进行流量分发时，流量分配策略以后端服务器组为单位生效。
健康检查	负载均衡器会定期向后端服务器发送请求以测试其运行状态，这些测试称为健康检查。通过健康检查来判断后端服务器是否可用。负载均衡器如果判断后端服务器健康检查异常，就不会将流量分发到异常后端服务器，而是分发到健康检查正常的后端服务器，从而提高了业务的可靠性。当异常的后端服务器恢复正常运行后，负载均衡器会将其自动恢复到负载均衡服务中，承载业务流量。

名词	说明
重定向	HTTPS是加密数据传输协议，安全性高，如果您需要保证业务建立安全连接，可以通过负载均衡的HTTP重定向功能，将HTTP访问重定向至HTTPS。
会话保持	会话保持，指负载均衡器可以识别客户与服务器之间交互过程的关联性，在实现负载均衡的同时，保持将其他相关联的访问请求分配到同一台后端服务器上。
WebSocket	WebSocket (WS)是HTML5一种新的协议。它实现了浏览器与服务器全双工通信，能更好地节省服务器资源和带宽并达到实时通讯。WebSocket建立在TCP之上，同HTTP一样通过TCP来传输数据，但是它与HTTP最大不同在于，WebSocket是一种双向通信协议，在建立连接后，WebSocket服务器和Browser/Client Agent都能主动地向对方发送或接收数据，就像Socket一样；WebSocket需要类似TCP的客户端和服务器端通过握手连接，连接成功后才能相互通信。
SNI	SNI (Server Name Indication)是为了解决一个服务器使用域名证书的TLS扩展，开启SNI之后，用户需要添加域名对应的证书。开启SNI后，允许客户端在发起SSL握手请求时就提交请求的域名信息，负载均衡收到SSL请求后，会根据域名去查找证书，如果找到域名对应的证书，则返回该证书；如果没有找到域名对应的证书，则返回缺省证书。
长连接	长连接是指在一个连接上可以连续发送多个数据包，在连接保持期间，如果没有数据包发送，需要双方发链路检测包。
短连接	短连接是指通讯双方有数据交互时，就建立一个连接，数据发送完成后，则断开此连接，即每次连接只完成一项业务的发送。
并发连接	并发连接指客户端向服务器发起请求并建立了TCP连接的总和，负载均衡的并发连接是指每秒钟所能接收并处理的TCP连接总和。

1.12.2 区域和可用区

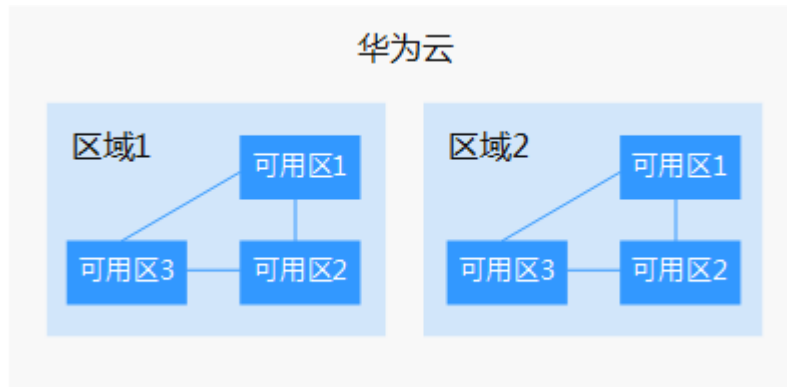
什么是区域、可用区？

区域和可用区用来描述数据中心的位置，您可以在特定的区域、可用区创建资源。

- 区域 (Region)：从地理位置和网络时延维度划分，同一个Region内共享弹性计算、块存储、对象存储、VPC网络、弹性公网IP、镜像等公共服务。Region分为通用Region和专属Region，通用Region指面向公共租户提供通用云服务的Region；专属Region指只承载同一类业务或只面向特定租户提供业务服务的专用Region。
- 可用区 (AZ, Availability Zone)：一个AZ是一个或多个物理数据中心的集合，有独立的风火水电，AZ内逻辑上再将计算、网络、存储等资源划分成多个集群。一个Region中的多个AZ间通过高速光纤相连，以满足用户跨AZ构建高可用性系统的需求。

图1-20阐明了区域和可用区之间的关系。

图 1-20 区域和可用区



目前，华为云已在全球多个地域开放云服务，您可以根据需求选择适合自己的区域和可用区。更多信息请参见华为云全球站点。

如何选择区域？

选择区域时，您需要考虑以下几个因素：

- 地理位置

一般情况下，建议就近选择靠近您或者您的目标用户的区域，这样可以减少网络时延，提高访问速度。

- 在除中国大陆以外的亚太地区有业务的用户，可以选择“中国-香港”、“亚太-曼谷”或“亚太-新加坡”区域。
- 在非洲地区有业务的用户，可以选择“非洲-约翰内斯堡”区域。
- 在拉丁美洲地区有业务的用户，可以选择“拉美-圣地亚哥”区域。

📖 说明

“拉美-圣地亚哥”区域位于智利。

- 资源的价格

不同区域的资源价格可能有差异，请参见华为云服务价格详情。

如何选择可用区？

是否将资源放在同一可用区内，主要取决于您对容灾能力和网络时延的要求。

- 如果您的应用需要较高的容灾能力，建议您将资源部署在同一区域的不同可用区内。
- 如果您的应用要求实例之间的网络延时较低，则建议您将资源创建在同一可用区内。

区域和终端节点

当您通过API使用资源时，您必须指定其区域终端节点。有关华为云的区域和终端节点的更多信息，请参见[地区和终端节点](#)。

1.13 与其他服务的关系

弹性负载均衡服务与其它服务的依赖关系如图1所示：

图 1-21 弹性负载均衡服务其它服务的关系示例图

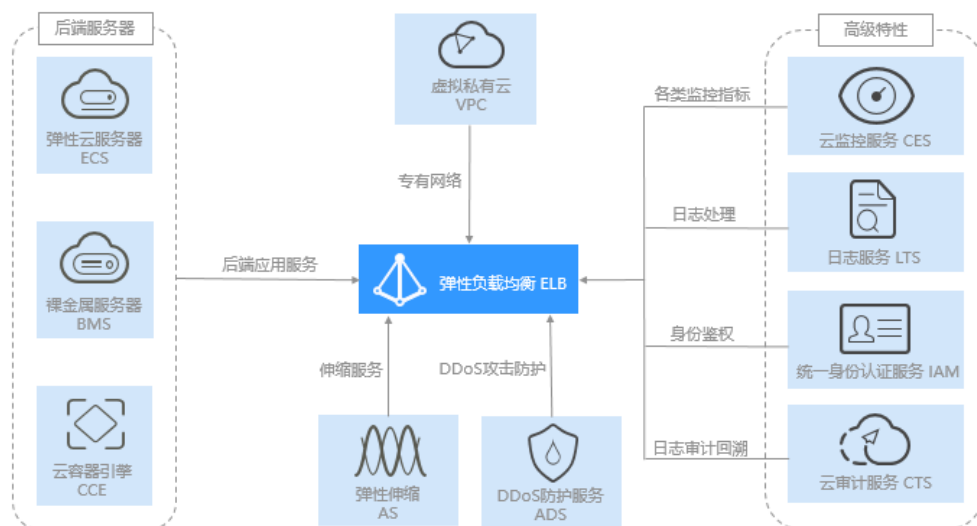


表 1-39 与其他服务之间关系

服务名称	交互功能	相关内容
弹性云服务器（Elastic Cloud Server, ECS）	通过相关服务部署用户业务，并接收ELB分发的访问流量。	搭建后端服务
裸金属服务器（Bare Metal Server, BMS）		添加后端服务器
云容器引擎（Cloud Container Engine, 简称CCE）		创建负载均衡类型的服务
弹性公网IP（Elastic IP）	为ELB配置EIP和公网带宽，使得ELB可以处理公网的访问流量。	新建负载均衡器并配置EIP
弹性伸缩（Auto Scaling, AS）	当配置了负载均衡服务后，弹性伸缩在添加和移除云服务器时，自动在负载均衡服务中添加和移除云服务器。	创建伸缩组
统一身份认证服务（Identity and Access Management, IAM）	需要统一身份认证提供鉴权。	创建用户组并授权
云审计服务（Cloud Trace Service, CTS）	使用云审计服务记录弹性负载均衡服务的资源操作。	查看审计日志

服务名称	交互功能	相关内容
云监控服务（Cloud Eye Service）	当用户开通了弹性负载均衡服务后，无需额外安装其他插件，即可在云监控查看对应服务的实例状态。	监控弹性负载均衡
DDoS防护服务（Anti-DDoS Service, AAD）	当用户购买了DDoS防护服务后，配置了负载均衡器的公网IP，确保了弹性负载均衡服务免受外部攻击，提高安全可靠性的。	配置Anti-DDoS防护策略
云日志服务（Log Tank Service, LTS）	配置访问日志时需要您对接云日志服务，查看和分析对七层负载均衡HTTP和HTTPS进行请求的详细访问日志记录。	访问日志

2 网关型

2.1 什么是网关型负载均衡

网关型负载均衡 GWLB (Gateway Load Balancer) 是运行在网络层 (OSI七层模型) 的负载均衡, 支持通过IP监听将所有端口的流量分发给后端服务器组中的网络虚拟设备NVA (Network Virtual Appliance), 帮助您高效实现各类网络虚拟设备的高可用部署, 例如: 防火墙、入侵检测和预防、流量镜像、深度报文检测等。

说明

网关型负载均衡目前在公测阶段, 如需使用, 请提交[工单](#)申请公测。

GWLB 组成

图 2-1 GWLB 组成

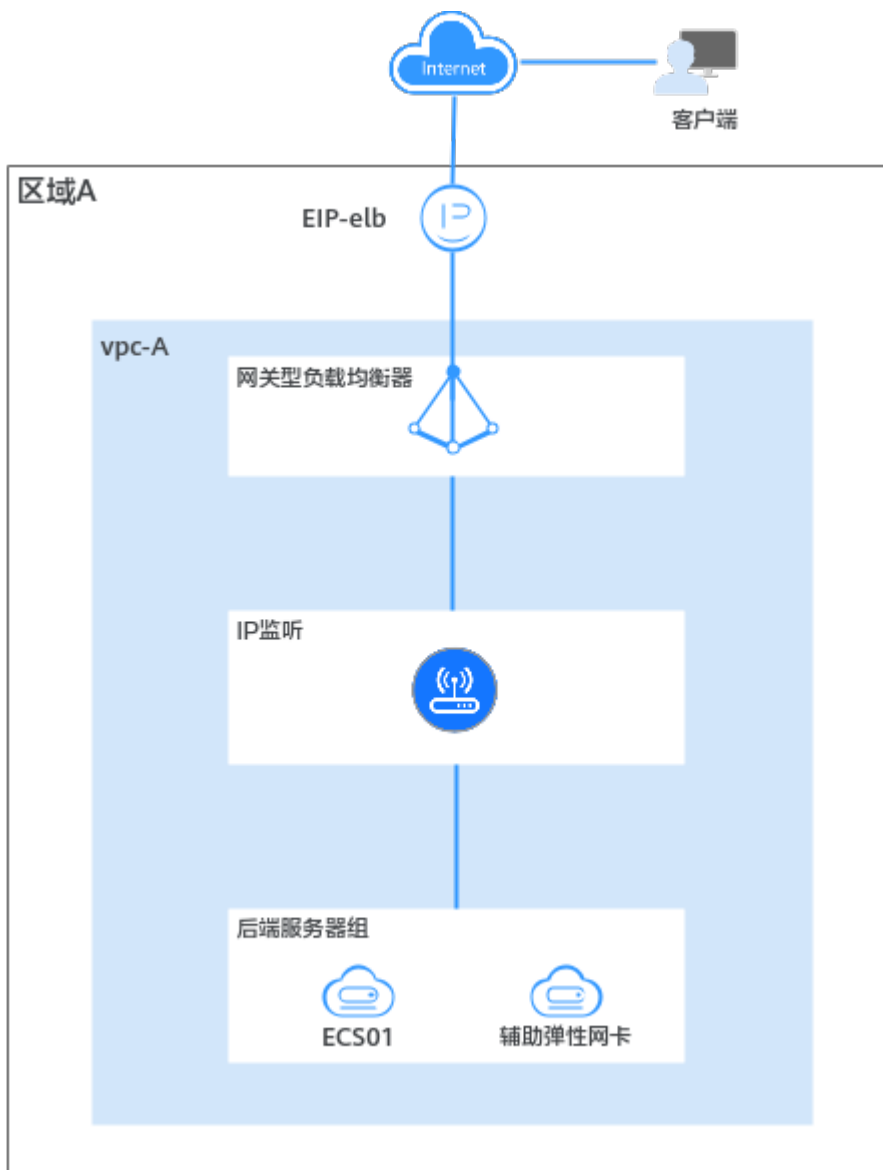


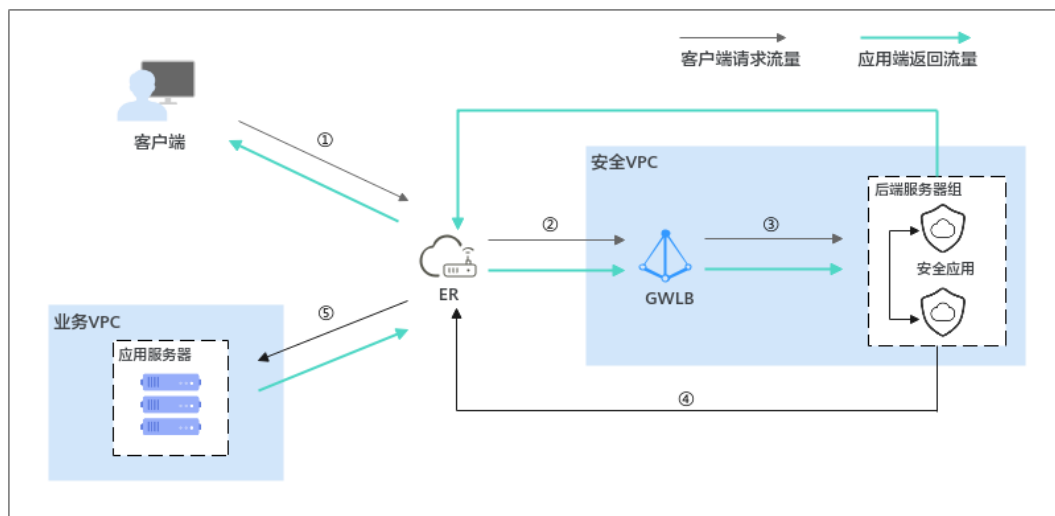
表 2-1 GWLB 组成

概念	说明
GWLB实例	网关型负载均衡 GWLB (Gateway Load Balancer) 是一种工作在 OSI参考模型第三层 (即网络层) 的负载均衡, 通过将流量透明地分发到不同的后端服务器来提高系统的安全性和可用性。
监听器	GWLB使用IP监听, 将所有端口中的IP数据包流量转发至后端服务器组。 一个GWLB实例仅支持一个IP监听器。

概念	说明
后端服务器组	<p>后端服务器组是一个逻辑分组，包含用于接收GWLB分发请求的一组后端服务器。</p> <p>GWLB中的后端服务器组独立于GWLB存在，可以将同一后端服务器组挂载在不同GWLB内，但同一个GWLB实例仅支持挂载一个后端服务器组。</p> <p>GWLB的服务器组支持添加云服务器和辅助弹性网卡作为后端服务器，一个服务器组中支持添加多个后端服务器。</p>
健康检查	<p>GWLB通过健康检查来判断后端服务器的可用性，支持TCP/UDP/HTTP/HTTPS协议的健康检查探测能力。GWLB探测服务器组中不健康的服务器，并避免将新的请求分发给不健康的服务器。GWLB支持丰富灵活的健康检查配置，如协议、端口以及各种健康检查阈值。</p>

2.2 网关型负载均衡是如何工作的

图 2-2 网关型负载均衡工作原理图



GWLB需要与企业路由器（Enterprise Router,ER）搭配使用，您可以通过在ER中配置路由表连通GWLB所在的VPC和业务VPC。

GWLB基于IP进行监听，可以将客户端请求的流量透明地转发到后端服务器组中。GWLB支持基于五元组、三元组和二元组的一致性哈希算法来进行流量分配，GWLB会将哈希计算值相同的流量转发到相同后端服务器。

客户端入方向请求路径

客户端的请求访问GWLB由企业路由器的路由表和VPC子网的路由表控制。客户端请求流量通过ER与GWLB连通，经过后端服务器组内的安全应用检测后返回到ER，最终转发到应用服务器。

1. 客户端发起请求：客户端向您的应用程序发起请求。

2. 企业路由器中转请求流量：GWLB实例与企业路由器（ER）服务的实例搭配使用，通过在ER中配置路由表将客户端VPC与安全VPC之间建立路由表信息，从而将客户端请求流量路由到GWLB实例所在子网。
3. GWLB转发请求：实例配置的IP协议监听器监听所有端口流量，并根据您的配置将客户端请求流量转发到后端服务器组中的虚拟设备检测过滤。
4. 安全应用转发检测后请求：后端虚拟设备将过滤后的请求流量送回到ER中。
5. 应用端接收请求：ER根据路由表信息将流量转发到应用服务器进行业务处理。

服务端出方向请求路径

1. 应用端返回响应：应用端处理完请求后向客户端返回响应。
2. 企业路由器中转响应流量：GWLB实例与企业路由器（ER）服务的实例搭配使用，通过在ER中配置路由表将业务VPC与安全VPC之间建立路由表信息，从而将应用端响应流量路由到GWLB实例所在子网。
3. GWLB转发请求：实例绑定的IP协议监听器监听所有端口流量，并根据您的配置将应用端响应流量转发到后端服务器组中的虚拟设备检测过滤。
4. 安全应用返回检测后的响应：后端虚拟设备将过滤后的响应流量送回到ER中。
5. 客户端接收响应：ER根据路由表信息将流量转发到相应的客户端。

2.3 应用场景

网关型负载均衡作为第三层（即网络层）的负载均衡，应用场景适用于在网络边界处理大规模、多类型的流量并确保高可用和高效分发。

通过 GWLB 部署 NAT 边界防火墙

NAT网关通常是云上资源访问互联网的网络出口。为应用复杂的网络安全挑战，企业可通过GWLB将所有经过NAT网关的流量集中至统一的管理层，从而确保所有出入互联网的请求都经过防火墙进行安全审核，以实现对流量的全面控制。

通过 GWLB 部署 VPC 边界防火墙

在同一区域内，多个VPC之间的云资源需要互通时，可以通过企业路由器（ER）将访问流量引导至GWLB，GWLB将访问流量分发至后端安全设备进行流量过滤，实现仅允许经过安全过滤的流量相互通信，从而提升网络安全性。

2.4 产品功能

使用网关型负载均衡 GWLB（Gateway Load Balancer）作为统一业务流量的入口，可将流量分发至健康的虚拟网络后端服务器，帮助您实现安全防御、流量检测过滤、网络分析等核心诉求。

本文为您介绍网络型负载均衡的主要功能。

协议版本

- **IPv4**：GWLB支持IPv4流量接入。
- **IPv6**：GWLB支持IPv6流量接入。

基于 IP 的监听

GWLB支持IP监听，可以监听多协议、所有端口的IP数据包并进行转发。在GWLB层面的配置上无需设置端口。

GWLB 的流量分配策略

GWLB的流量分配策略支持基于一致性哈希的调度算法。

- **五元组**：根据请求的五元组（包括源IP、源端口、目标IP、目标端口和传输协议）进行哈希计算。相同五元组的请求会分发到同一台后端服务器。
- **三元组**：根据请求的三元组（包括源IP、目标IP和传输协议）进行哈希计算。相同三元组的请求会分发到同一台后端服务器。
- **二元组**：根据请求的二元组（包括源IP和目标IP）进行哈希计算。相同二元组的请求会分发到同一台后端服务器。

📖 说明

如果后端服务器组的后端服务器发生了变更或后端权重进行了修改，流量分配策略可能会将新的请求重新分配到新的后端服务器。

可靠性

- **多可用区部署**：GWLB实例支持多可用区部署，当某个可用区出现故障时，其他可用区的仍然可以正常运行，保证业务的持续性，降低单点故障风险。
- **跨可用区转发**：GWLB实例支持将流量跨可用区转发至网络虚拟设备中，**不要求**GWLB实例与网络虚拟设备的可用区保持一致，提高了部署灵活性。当某个可用区的网络虚拟设备出现故障时，GWLB可将新的请求转发至其他可用区的网络虚拟设备上，保证业务的持续性，降低单点故障风险。

后端服务器

后端服务器支持添加云服务器和辅助弹性网卡。

高级特性

- **延迟注销**：开启延迟注销功能后，负载均衡器停止向移除的后端云服务器或者健康检查失败的后端云服务器发送新的请求，保持现有连接在延迟注销时间内正常传输。
- **后端异常转发模式**：可以避免因健康检查配置错误导致检查结果异常，造成的业务不可使用情况，提高业务的可用性。
 - 后端全异常转发模式默认关闭时，如果后端服务器健康检查结果全部异常，将会无法将请求转发到该后端服务器组。
 - 开启后端全异常转发模式后，如果后端服务器健康检查结果全部异常，将忽略健康检查结果，转发请求到组内的所有后端服务器。

该功能可以避免因健康检查配置错误导致检查结果异常，造成的业务不可使用情况，提高业务的可用性。
- **关闭权重为0后端的健康检查**：开启该功能后，将停止对权重设置为0的后端服务器进行健康检查探测，这些后端的健康检查状态将显示为未知。在业务变更等场景，您可以将暂时不需要的后端服务器权重设置为0，关闭权重为0后端的健康检查后，将避免向这些后端发送健康检查探测报文，这些后端服务器也不会触发异常主机告警。

监控运维

实例监控：GWLB可结合云监控服务，从实例维度、监听器维度、可用区维度提供连接数、带宽、健康检查等指标信息，方便您了解业务的运行状况。网关型实例与网络型实例支持的监控指标一致。

2.5 约束与限制

本文介绍网关型负载均衡 GWLB（Gateway Load Balancer）资源配额的名称、默认值、是否支持调整，以及GWLB相关限制等信息。

GWLB 的配额

为防止资源滥用，平台限定了各服务资源的配额，对用户的资源数量和容量做了限制。如您最多可以创建多少台弹性云服务器、多少块云硬盘。

如果当前资源配额限制无法满足使用需要，您可以参考[如何申请扩大配额?](#)《用户指南》中《关于配额》章节，提升资源配额。

网关型负载均衡与所有独享型ELB实例共享配额，详情请见[弹性负载均衡的服务配额](#)，仅在以下资源存在差异。

表 2-2 网关型弹性负载均衡的服务配额

资源名称	资源说明	默认配额	是否支持申请
单负载均衡器可添加监听器数量	一个网关型负载均衡器支持添加监听器的数量	1个	否
单监听器可添加后端服务器组数量	一个监听器支持添加后端服务器组的数量	1个	否

GWLB 使用限制

- 创建GWLB实例时，**GWLB的前端子网和后端服务器所在的子网必须使用同一VPC下不同的子网**，否则将会导致访问不通。
- GWLB基于IP进行监听转发，仅基于IP进行监听，将所有端口所有IP协议的数据包转发到后端服务器，因此在GWLB的配置中**不需要配置端口**。
 - 创建监听器时，**不支持设置监听端口**，默认配置为0端口且不支持修改。
 - 为后端服务器组添加后端服务器时，**不支持设置业务端口**，默认配置为0端口且不支持修改。
- 为后端服务器组配置健康检查时，**不支持使用后端服务器的默认业务端口（0端口）**，需要您支持指定特定端口，支持指定端口范围为1~65535端口。