解决方案实践

快速搭建 Dify-LLM 应用开发平台

文档版本 1.0

发布日期 2024-11-07

华为技术有限公司

商标声明


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。

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该流程的详细内容请参见如下网址：
https://www.huawei.com/cn/psirt/vul-response-process
如企业客户须获取漏洞信息，请参见如下网址：
https://securitybulletin.huawei.com/enterprise/cn/security-advisory

目 录

1 方案概述

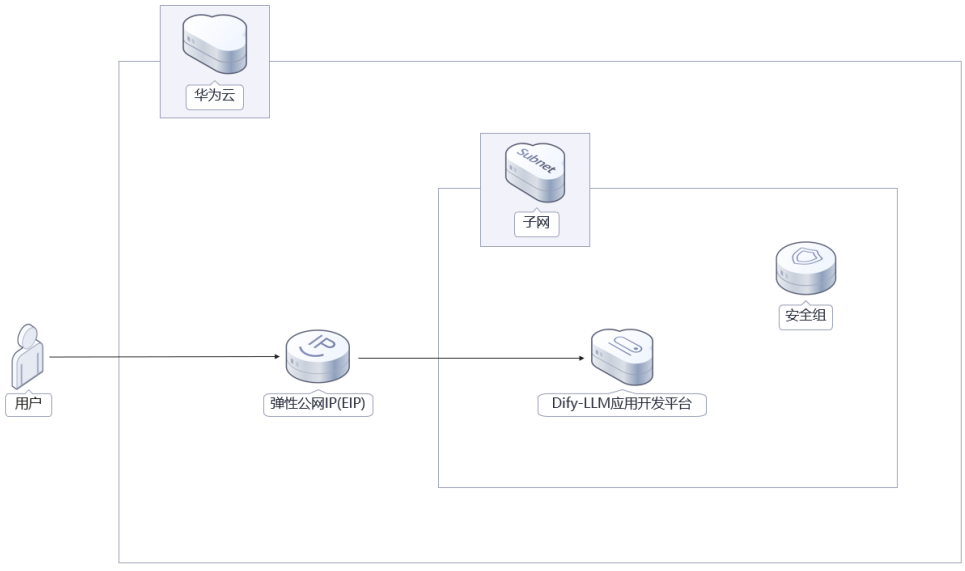
应用场景

该解决方案基于Flexus云服务器X实例帮助您快速部署Dify LLM应用开发平台。Dify是一款开源的大语言模型(LLM)应用开发平台。它融合了后端即服务（Backend as Service）和LLMOps的理念，使开发者可以快速搭建生产级的生成式AI应用。

方案架构

该解决方案基于Flexus云服务器X实例帮助您快速部署Dify LLM应用开发平台。

图 1-1 方案架构图



该解决方案将会部署如下资源：

- 创建1台**华为云Flexus云服务器X实例**，用于搭建Dify-LLM应用开发平台。
- 创建1个**弹性公网IP EIP**并关联华为云Flexus云服务器X实例，提供访问公网和被公网访问能力。

- 创建一个安全组，通过配置安全组规则，为云服务器提供安全防护。

方案优势

- 开箱即用
  后端即服务，帮助开发者快速搭建生产级的生成式AI应用，部署完成即可使用。
- 低成本
  提供高性价比的云服务器，用户可以根据实际需求自定义不同规格的云服务器。
- 一键部署
  一键轻松部署，即可完成云服务器的创建和Dify-LLM应用开发平台的搭建。

约束与限制

- 部署该解决方案之前，您需要注册华为账号并开通华为云，完成实名认证，且账号不能处于欠费或冻结状态，请根据**2 资源和成本规划**中预估价格，确保余额充足。

2 资源和成本规划

该解决方案主要部署如下资源，以下费用仅供参考，具体请参考华为云官网**价格详情**，实际收费以账单为准。

表 2-1 资源和成本规划（按需计费）

华为云服务	配置示例	每月预估花费
华为云Flexus云服务器X实例	• 按需计费 • 区域：华北-北京四 • 规格：Flexus云服务器X实例 | 性能模式（关闭）| x1.2u.4g | 2核 | 4 GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：高IO | 100GB • 购买量：1	197.28元
弹性公网IP EIP	• 区域：华北-北京四 • 计费模式：按需计费 • 线路：动态BGP • 公网带宽：按流量计费 • 带宽大小：300Mbit/s • 购买量：1	0.80元/GB
合计	-	**197.28元 + 弹性公网IP EIP费用**

表 2-2 资源和成本规划（包年包月）

华为云服务	配置示例	每月预估花费
华为云Flexus云服务器X实例	• 包年包月 • 区域：华北-北京四 • 规格：Flexus云服务器X实例 | 性能模式（关闭）| x1.2u.4g | 2核 | 4 GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：高IO | 100GB • 购买量：1	143.00元
弹性公网IP EIP	• 区域：华北-北京四 • 计费模式：按需计费 • 线路：动态BGP • 公网带宽：按流量计费 • 带宽大小：300Mbit/s • 购买量：1	0.80元/GB
合计	-	**143.00元 + 弹性公网IP EIP费用**

3 实施步骤

3.1 准备工作

当您首次使用华为云时注册的账号，则无需执行该准备工作，如果您使用的是IAM用户账户，请确认您是否在admin用户组中，如果您不在admin组中，则需要为您的账号**授予相关权限**，并完成以下准备工作。

创建 rf_admin_trust 委托（可选）

步骤1 进入华为云官网，打开**控制台管理**界面，鼠标移动至个人账号处，打开“统一身份认证”菜单。

图 3-1 控制台管理界面

图 3-2 统一身份认证菜单

步骤2 进入“委托”菜单，搜索“rf_admin_trust”委托。

图 3-3 委托列表

- 如果委托存在，则不用执行接下来的创建委托的步骤
- 如果委托不存在时执行接下来的步骤创建委托

步骤3 单击步骤2界面中的“创建委托”按钮，在委托名称中输入“rf_admin_trust”，委托类型选择“云服务”，选择“RFS”，单击“下一步”。

图 3-4 创建委托

步骤4 在搜索框中输入“Tenant Administrator”权限，并勾选搜索结果，单击“下一步”。

图 3-5 选择策略

步骤5 选择“所有资源”，并单击“下一步”完成配置。

图 3-6 设置授权范围

步骤6 “委托”列表中出现“rf_admin_trust”委托则创建成功。

图 3-7 委托列表

----结束

3.2 快速部署

本章节主要帮助用户快速部署“快速搭建Dify-LLM应用开发平台”解决方案

表 3-1 参数说明

参数名称	类型	是否可选	参数解释	默认值
vpc_name	string	必填	虚拟私有云名称，该模板使用新建VPC，不允许重名。取值范围：1-54个字符，支持中文、英文字母、数字、_（下划线）、-（中划线）、.（点）。	dify-llm-application-development-platform-demo
secgroup_name	string	必填	安全组名称，该模板新建安全组，请参考**安全组规则修改**进行配置。取值范围：1-64个字符，支持字母、数字、中文、下划线（_）、中划线（-）、英文句号（.）。	dify-llm-application-development-platform-demo
ecs_name	string	必填	云服务器实例名称，不支持重名。取值范围：1-60个字符，支持中文、英文字母、数字、_（下划线）、-（中划线）、.（点）。	dify-llm-application-development-platform-demo

参数名称	类型	是否可选	参数解释	默认值
flexus_flavor	string	必填	云服务器实例规格，支持弹性云服务器ECS及华为云Flexus 云服务器X实例。Flexus 云服务器X实例规格ID命名规则为x1.?u.?g，例如2vCPUs4GiB规格ID为x1.2u.4g，具体华为云Flexus 云服务器X实例规格请参考控制台。弹性云服务器规格名称，具体请参考官网**弹性云服务器规格清单**。	x1.2u.4g
ecs_password	string	必填	云服务器密码，长度为8-26位，密码至少必须包含大写字母、小写字母、数字和特殊字符（!@$%^-_=+[{}]:,./?）中的三种，仅支持小写字母、数字、中划线（-）、英文句号（.）。修改密码，请参考**重置云服务器密码**登录ECS控制台修改密码。管理员账户默认root。	空

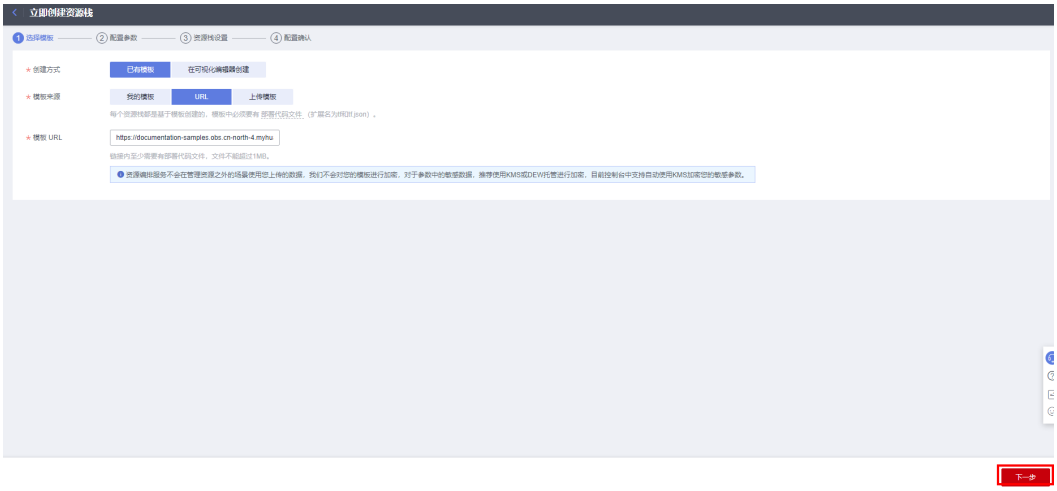
参数名称	类型	是否可选	参数解释	默认值
ecs_volume_size	number	必填	云服务器系统盘大小，磁盘类型默认为高IO，单位：GB，取值范围为40-1,024，不支持缩盘。	100
bandwidth_size	number	必填	弹性公网带宽大小，该模板计费方式为按流量计费。单位：Mbit/s，取值范围：1-300Mbit/s。	300
charging_mode	string	必填	计费模式，默认自动扣费，取值为prePaid（包年包月）或postPaid（按需计费）。	postPaid
charge_period_unit	string	必填	计费周期单位，当计费方式设置为prePaid，此参数是必填项。有效值为：month（包月）和year（包年）。	month
charge_period	number	必填	计费周期，当计费模式设置为prePaid，此参数是必填项。可选值为：1-3（year）、1-9（month）。	1

步骤1 登录**华为云解决方案实践**，选择“快速搭建Dify-LLM应用开发平台”，单击“一键部署”，跳转至解决方案创建资源栈界面。

图 3-8 解决方案实施库

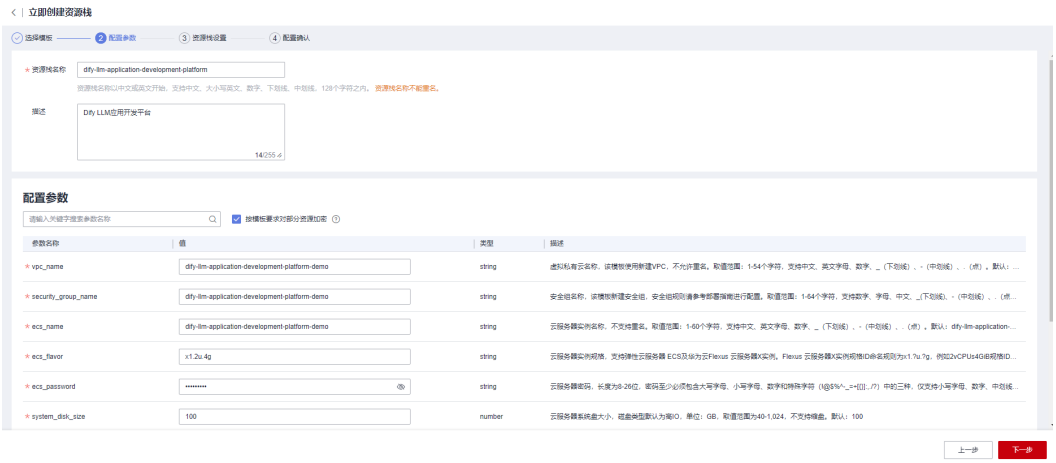
步骤2 在选择模板界面中，单击“下一步”。

图 3-9 选择模板



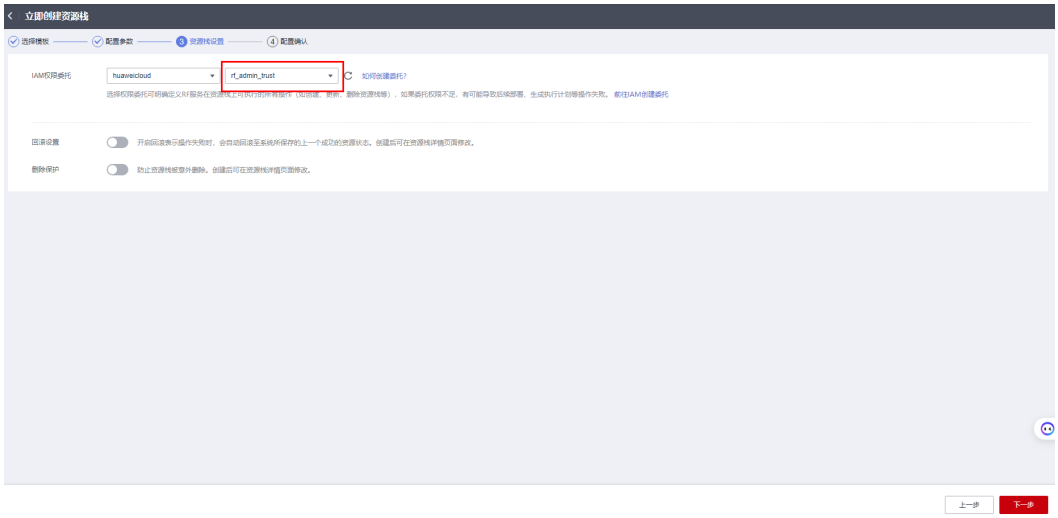
步骤3 在配置参数界面中，参考**表1 参数填写说明**完成自定义参数填写，单击“下一步”。

图 3-10 配置参数



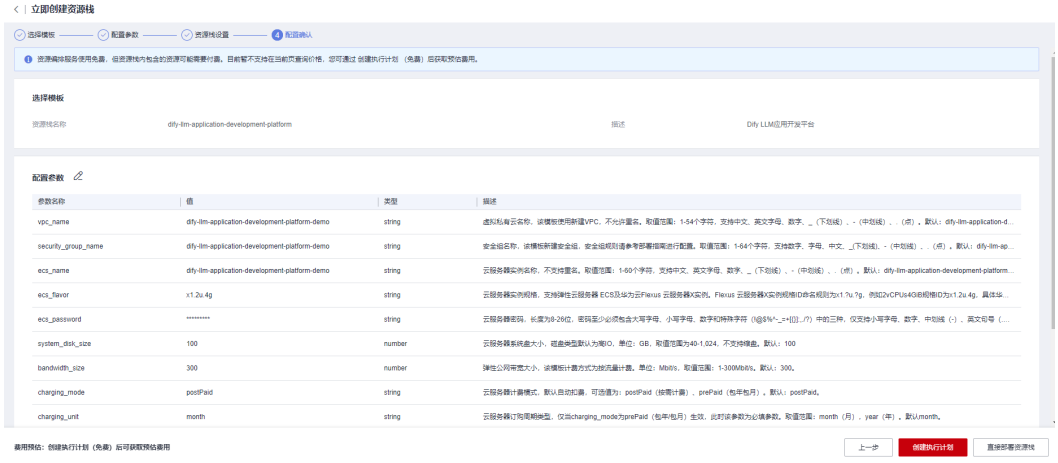
步骤4 在资源设置界面中，在权限委托下拉框中选择“rf_admin_trust”委托（可不选），单击“下一步”。

图 3-11 资源栈设置



步骤5 在配置确认界面中，单击“创建执行计划”。

图 3-12 配置确认



步骤6 在弹出的创建执行计划框中，自定义填写执行计划名称，单击“确定”。

图 3-13 创建执行计划

步骤7 单击“部署”，并且在弹出的执行计划确认框中单击“执行”。

图 3-14 执行计划

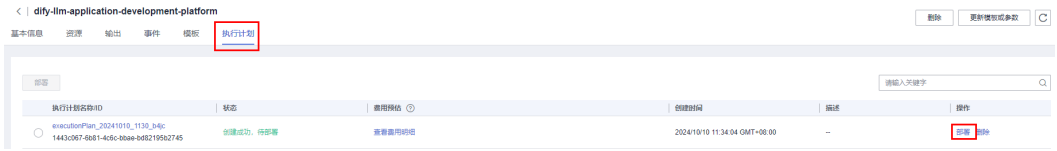


图 3-15 执行计划确认

步骤8 （可选）如果计费模式选择“包年包月”，在余额不充足的情况下（所需总费用请参考**表2-2**）请及时登录费用中心，手动完成待支付订单的费用支付。

步骤9 待“事件”中出现“Apply required resource success”，表示该解决方案已经部署完成。

图 3-16 部署完成

步骤10 刷新页面，在“输出”中查看Dify-LLM应用开发平台访问说明。堆栈部署成功后，Dify应用搭建脚本开始执行，耐心等待5-10分钟左右（受网络波动影响）。

图 3-17 说明

----结束

3.3 开始使用

安全组规则修改（可选）

须知

- 该解决方案使用80端口用来访问Dify，默认对该方案创建的VPC子网网段放开，请参考**修改安全组规则**，配置IP地址白名单，以便能正常访问服务。
- 该解决方案使用22端口用来以SSH方式远程登录云服务器，若需远程登录云服务器，请参考**修改安全组规则**，配置IP地址白名单，以便能正常访问服务。
- 该解决方案部署成功后，环境初始化预计5-10分钟，受网络、带宽影响，部署时间会有波动部署完成之后方可正常访问。

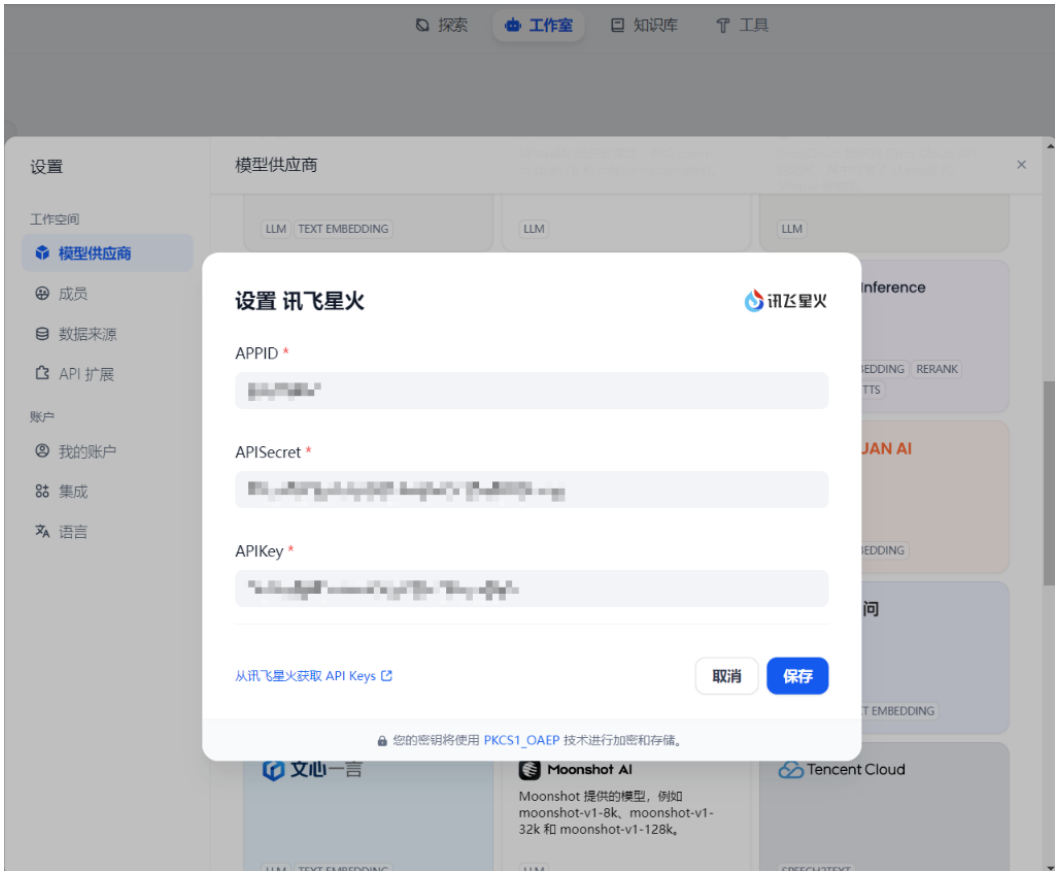
安全组实际是网络流量访问策略，包括网络流量入方向规则和出方向规则，通过这些规则为安全组内具有相同保护需求并且相互信任的云服务器、云容器、云数据库等实例提供安全保护。

如果您的实例关联的安全组策略无法满足使用需求，比如需要添加、修改、删除某个TCP端口，请参考以下内容进行修改。

- 添加安全组规则：根据业务使用需求需要开放某个TCP端口，请参考**添加安全组规则**添加入方向规则，打开指定的TCP端口。
- 修改安全组规则：安全组规则设置不当会造成严重的安全隐患。您可以参考**修改安全组规则**，来修改安全组中不合理的规则，保证云服务器等实例的网络安全。
- 删除安全组规则：当安全组规则入方向、出方向源地址/目的地址有变化时，或者不需要开放某个端口时，您可以参考**删除安全组规则**进行安全组规则删除。

步骤1 登录开发平台：输入**快速部署步骤10**的访问地址，即可浏览Dify的开发平台。首次登录需注册管理员账号，依次填写邮箱、账号、密码。

步骤2 配置大模型API key：进入系统后，首先需要配置大模型的API key，用于Dify调用大模型API的能力。单击系统右上角的设置，选择模型供应商，挑选一家大模型，例如讯飞星火，单击设置，填写对应的API key。



步骤3 创建应用：接着需要创建一个应用，您可以创建空白应用或者从模板中创建。作为开始，选择空白应用。

步骤4 查看API调用：您可以在编排里进行调试与预览，例如编写提示词、导入知识库等。一旦完成，可以单击左侧的访问API，查看API的调用方法。

步骤5 创建密钥：在调用之前，还需要单击右上角“API密钥”创建一个密钥，用作请求鉴权。

----结束

说明

探索应用模板

仅使用空白模板来部署Dify和直接调用大模型的API体验几乎一致，为了充分发挥Dify的能力，您可以单击页面上方的探索，尝试预制的各类应用模板。除此之外，您还可以浏览Dify官方文档获取更多详细的开发指南。

3.4 快速卸载

步骤1 登录资源编排 RFS资源栈，找到该解决方案创建的资源栈，单击资源栈名称右侧“删除”按钮。

图 3-18 一键卸载

步骤2 在弹出的删除资源栈确定框中，删除方式选择删除资源，输入Delete，单击“确定”，即可卸载解决方案。

图 3-19 删除资源栈确认

----结束

4 附录

名词解释

- 华为云Flexus云服务器X实例：Flexus云服务器X实例是新一代面向中小企业和开发者打造的柔性算力云服务器。Flexus云服务器X实例功能接近ECS， 同时还具备独有特点，例如Flexus云服务器X实例具有更灵活的vCPU内存配比、支持热变配不中断业务变更规格、支持性能模式等。
- 弹性云服务器 ECS：是一种云上可随时自助获取、可弹性伸缩的计算服务，可帮助您打造安全、可靠、灵活、高效的应用环境。
- 虚拟私有云 VPC：是用户在华为云上申请的隔离的、私密的虚拟网络环境。用户可以基于VPC构建独立的云上网络空间，配合**弹性公网IP**、**云连接**、**云专线**等服务实现与Internet、云内私网、跨云私网互通，帮您打造可靠、稳定、高效的专属云上网络。
- 弹性公网IP EIP：提供独立的公网IP资源，包括公网IP地址与公网出口带宽服务。可以与弹性云服务器、裸金属服务器、虚拟IP、弹性负载均衡、NAT网关等资源灵活地绑定及解绑，提供访问公网和被公网访问能力。

5 修订记录

表 5-1 修订记录

发布日期	修订记录
2024-11-07	第一次正式发布。