解决方案实践

基于 Geek-AI 构建 AI 智能助手

文档版本 1.0

发布日期 2024-05-29

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

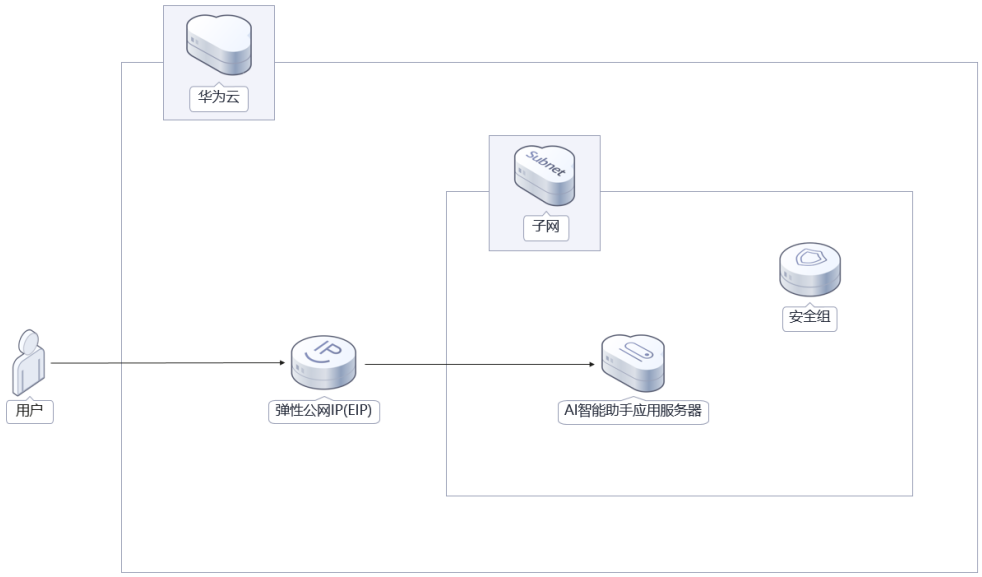
应用场景

该解决方案可以帮助您在华为云弹性云服务器 ECS上基于Geek-AI构建自己的创作系统。Geek-AI是基于 AI 大语言模型 API 实现的 AI 助手全套开源解决方案，自带运营管理后台，开箱即用。集成了 OpenAI、 Azure ChatGLM、讯飞星火、文心一言等多个平台的大语言模型。集成了 MidJourney 和 Stable Diffusion AI绘画、音乐、思维导图生成功能。

方案架构

该解决方案在华为云弹性云服务器 ECS上基于开源Geek-AI构建AI智能助手。该解决方案部署架构如下图所示：

图 1-1 方案架构图



该解决方案会部署如下资源：

- 创建一台Linux 弹性云服务器 ECS，用于搭建Geek-AI创作系统。
- 创建一个弹性公网IP EIP，用于提供访问公网和被公网访问能力。
- 创建安全组，通过配置安全组规则，为弹性云服务器提供安全防护。

方案优势

- 多功能

  支持多个平台大语言模型，集成AI聊天机器人、AI绘画、音乐、思维导图生成功能。
- 开箱即用

  完整前端应用和后台管理系统，装完即用。
- 一键部署

  一键轻松部署，即可实现Geek-AI创作系统应用搭建。

约束与限制

- 该解决方案部署前，需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。如果计费模式选择“包年包月”，请确保账户余额充足以便一键部署资源的时候可以自动支付；或者在一键部署的过程进入**费用中心**，找到“待支付订单”并手动完成支付。
- 请确保你有AI大模型API KEY，具体信息可以参考Geek-AI添加 API KEY文档。

2 资源和成本规划

该解决方案主要部署如下资源，不同产品的花费仅供参考，具体请参考华为云**官网价格**，实际以收费账单为准：

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
弹性云服务器ECS	• 按需计费：0.55元/小时 • 区域：中国-香港 • 计费模式：按需计费 • 规格： 通用计算型 S6 | 2核 | 4GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：高IO | 100GB • 购买量：1	396.00 元
弹性公网IP EIP	• 按需计费：1.05元/GB • 区域：中国-香港 • 计费模式：按需计费 • 线路：动态BGP • 公网带宽：按流量计费 • 购买数量：1	1.05 元/GB
合计	-	**396.00 元 + 流量费用**

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
弹性云服务器ECS	● 区域：中国-香港 ● 计费模式：包月 ● 规格： 通用计算型 S6 | 2核 | 4GB ● 镜像：Ubuntu 22.04 server 64bit ● 系统盘：高IO | 100GB ● 购买量：1	275.50 元
弹性公网IP EIP	● 按需计费：1.05元/GB ● 区域：中国-香港 ● 计费模式：按需计费 ● 线路：动态BGP ● 公网带宽：按流量计费 ● 购买数量：1	1.05 元/GB
合计	-	**275.50 元 + 流量费用**

3 实施步骤

3.1 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。

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该解决方案。

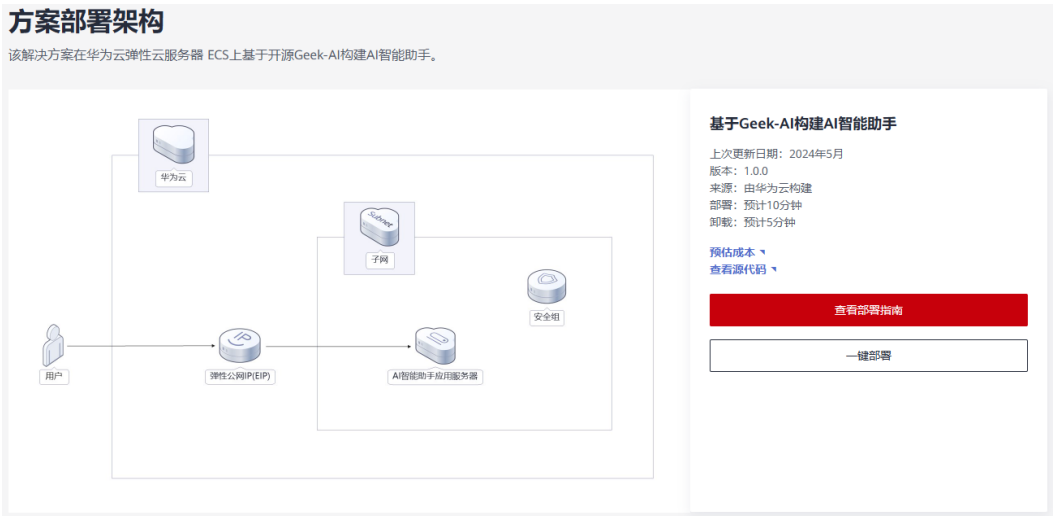
表 3-1 参数填写说明

参数名称	类型	是否必填	参数解释	默认值
vpc_name	String	必填	虚拟私有云 VPC名称，该模板使用新建VPC，不允许重名。取值范围：1-64个字符，仅支持中文、英文字母、数字、_（下划线）、-（中划线）、.（点）。	ai-assistant-with-geek-ai-demo
security_group_name	String	必填	安全组名称，该模板新建安全组，安全组规则请参考**安全组规则修改（可选）**进行修改。取值范围：1-64个字符，支持数字、字母、中文、_（下划线）、-（中划线）、.（点）。	ai-assistant-with-geek-ai-demo
ecs_name	String	必填	弹性云服务器 ECS名称，用于搭建WebUI服务器，不允许重名。取值范围：1-52个字符，仅支持数字、字母、中文、_(下划线)、-（中划线）、.（点）。	ai-assistant-with-geek-ai-demo
ecs_flavor	String	必填	弹性云服务器规格，推荐使用2vCPUs4GB及以上规格，请参考**弹性云服务器规格清单**。	s6.large.2

参数名称	类型	是否必填	参数解释	默认值
ecs_password	String	必填	弹性云服务器初始化密码。取值范围：长度为8-26个字符，密码至少包含大写字母、小写字母、数字和特殊字符（!@$%^-_=+[{()}]:,./?~#*）中的三种，管理员账户默认root。	空
charging_mode	String	必填	计费模式，默认自动扣费。可选值为：postPaid（按需计费）、prePaid（包年包月）。	postPaid
charging_unit	String	必填	订购周期类型。仅当charging_mode为prePaid（包年/包月）生效，此时该参数为必填参数。可选值为：month（月），year（年）。	month
charging_period	Number	必填	订购周期，仅当charging_mode为prePaid（包年/包月）生效，此时该参数为必填参数。当charging_unit=month（周期类型为月）时，取值范围：1-9；当charging_unit=year（周期类型为年）时，取值范围：1-3。	1

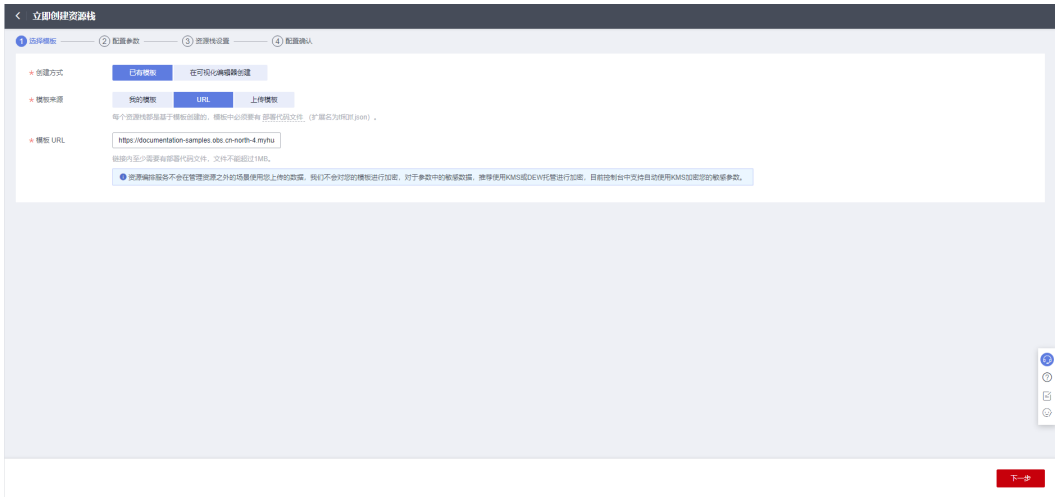
步骤1 登录**华为云解决方案实践**，选择“基于Geek-AI构建AI智能助手”并单击，跳转至该解决方案一键部署界面。

图 3-8 解决方案实施库



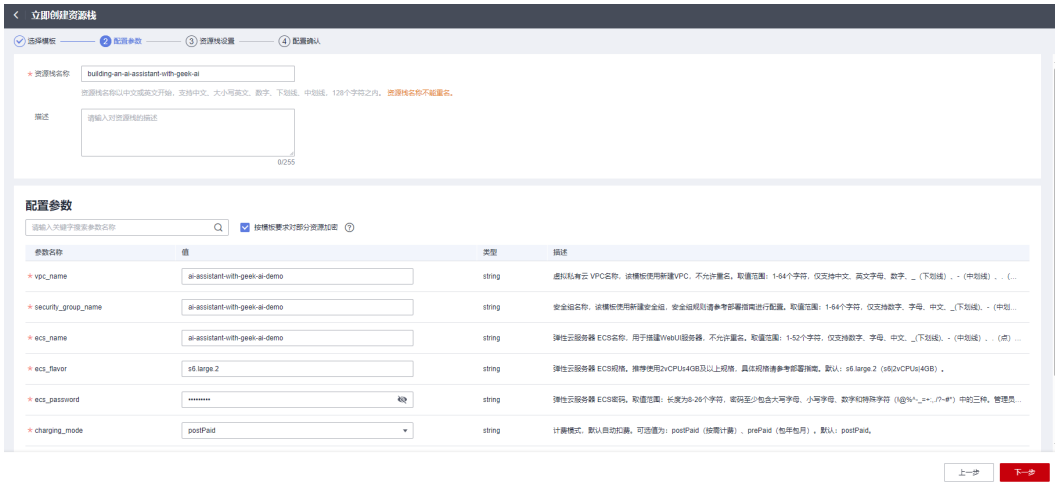
步骤2 单击“一键部署”，跳转至该解决方案创建资源栈部署界面。

图 3-9 创建资源栈



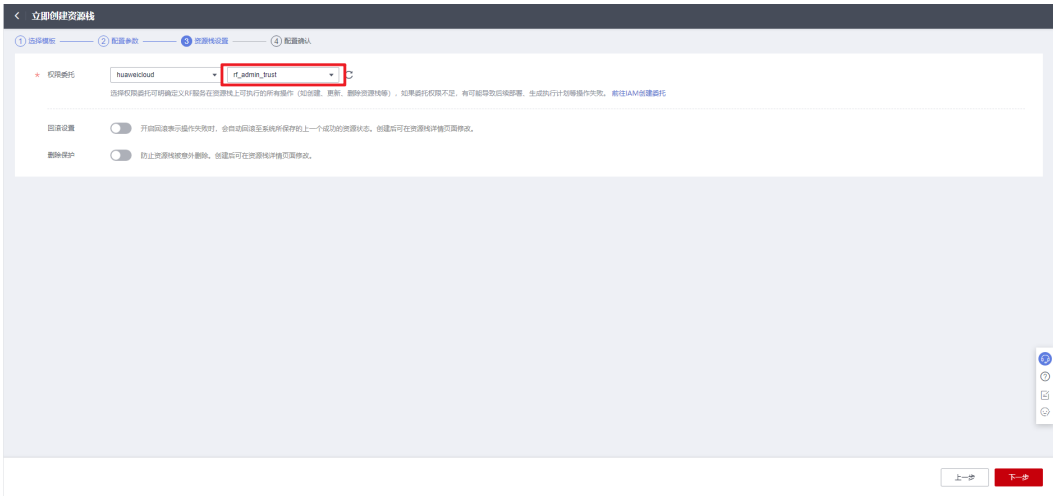
步骤3 单击“下一步”，参考**表3-1**完成自定义参数填写。

图 3-10 参数配置



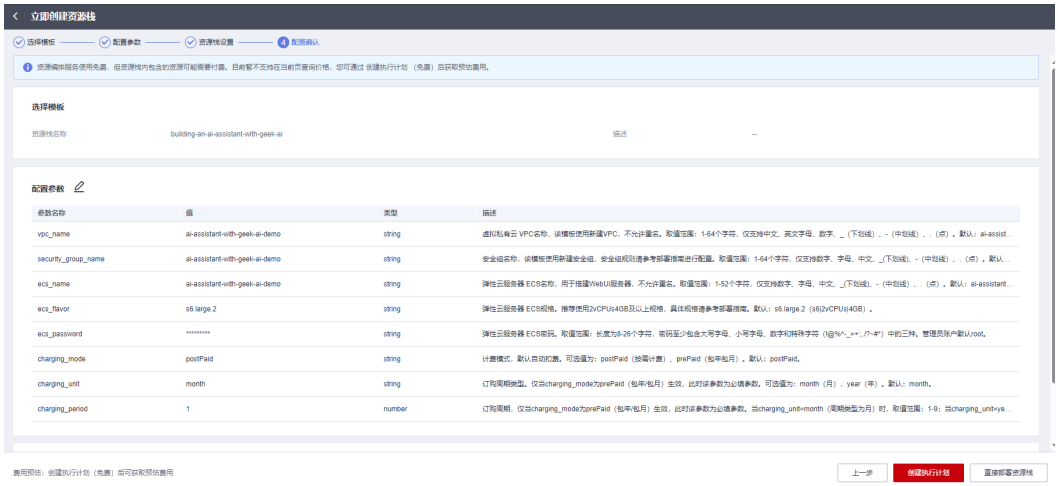
步骤4 （可选）在资源设置界面中，“权限委托”下拉框中选择“rf_admin_trust”委托，单击“下一步”。

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 创建执行计划



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”，弹出执行计划提示信息，单击“执行”确认执行。

图 3-14 执行计划确认

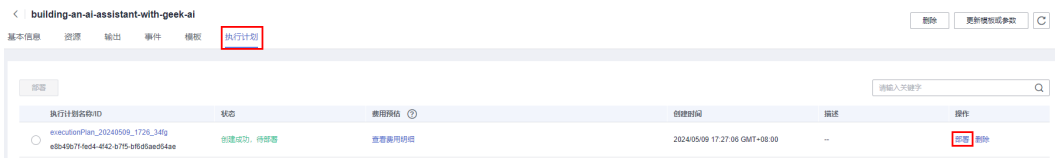
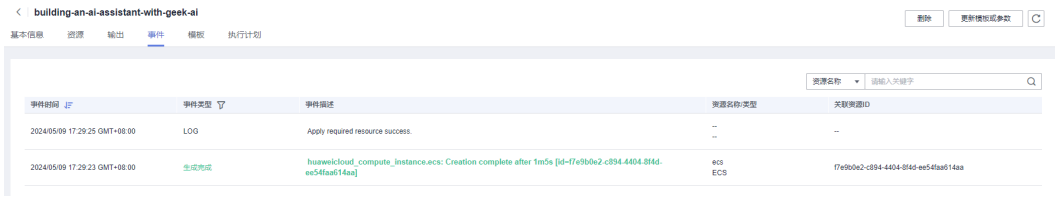


图 3-15 确认执行

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 等待解决方案自动部署。部署成功后，单击“事件”，回显结果如下:

图 3-16 资源创建成功



步骤10 在“输出”中查看WebUI访问说明。堆栈部署成功后，Geek-AI创作系统环境搭建脚本开始执行，耐心等待10分钟左右（受网络波动影响）。

图 3-17 输出

----结束

3.3 开始使用

本方案基于Geek-AI v4.0.6版本部署，详细使用指导请参考**Geek-AI文档说明**。

说明

涉及到AI创作系统的业务端口如下：

- 8080：WebUI访问端口

安全组规则修改（可选）

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 查看**快速部署 步骤3.2-9**访问说明查看公网IP地址。打开浏览器输入：http://EIP:8080，其中EIP为服务器公网IP，访问Geek-AI创作系统页面。

图 3-18 Geek-AI 创作系统页面

步骤2 浏览器输入http://EIP:8080/admin，进入Geek-AI 管理界面，初始用户名：admin ，初始密码：admin123，单击“登录”。

图 3-19 登录 Geek-AI 管理界面

步骤3 初次使用需要添加API KEY，按下图所示单击“API-KEY”新增添加聊天/绘画的API KEY。

图 3-20 添加 API KEY

步骤4 按下图所示，依次选择所属平台，填写名称，选择用途，按照页面提示填写API KEY信息，激活启用状态，单击“提交”。

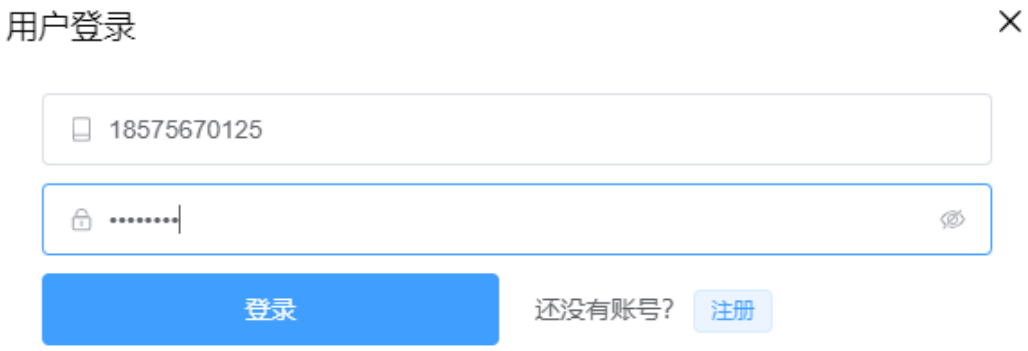
图 3-21 新增 API KEY

步骤5 打开浏览器，输入http://EIP:8080/，进入Geek-AI创作系统界面，单击“AI聊天”，

图 3-22 创作系统界面

步骤6 初始使用需要登录，在对话框中输入任意内容，在弹出的用户登录窗口中输入体验账号：18575670125 密码：12345678。，单击“登录”（移动端登录会自动适配）。

图 3-23 用户登录



步骤7 按下图所示，按照步骤4添加的API-KEY，下拉框选择对应的模型，在下面的对话框中，输入对话内容，单击发送按钮，即可获取对话结果。

图 3-24 AI 对话聊天

----结束

3.4 快速卸载

步骤1 登录**资源编排服务 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，在弹出的“删除资源栈”提示框输入Delete，单击“确定”进行解决方案卸载。

图 3-25 一键卸载

图 3-26 删除资源

----结束

4 附录

名词解释

基本概念、云服务简介、专有名词解释

- **弹性云服务器 ECS**：是一种可随时自助获取、可弹性伸缩的云服务器，可帮助您打造可靠、安全、灵活、高效的应用环境，确保服务持久稳定运行，提升运维效率。
- **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VIP、弹性负载均衡、NAT网关等资源灵活地绑定及解绑。
- **虚拟私有云 VPC**：是用户在云上申请的隔离的、私密的虚拟网络环境。用户可以自由配置VPC内的IP地址段、子网、安全组等子服务，也可以申请弹性带宽和弹性IP搭建业务系统。

5 修订记录

发布日期	修订记录
2024-05-15	第一次正式发布。