解决方案实践

基于 PyTorch NPU 快速部署开源大模型

文档版本 1.0

发布日期 2024-11-05

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

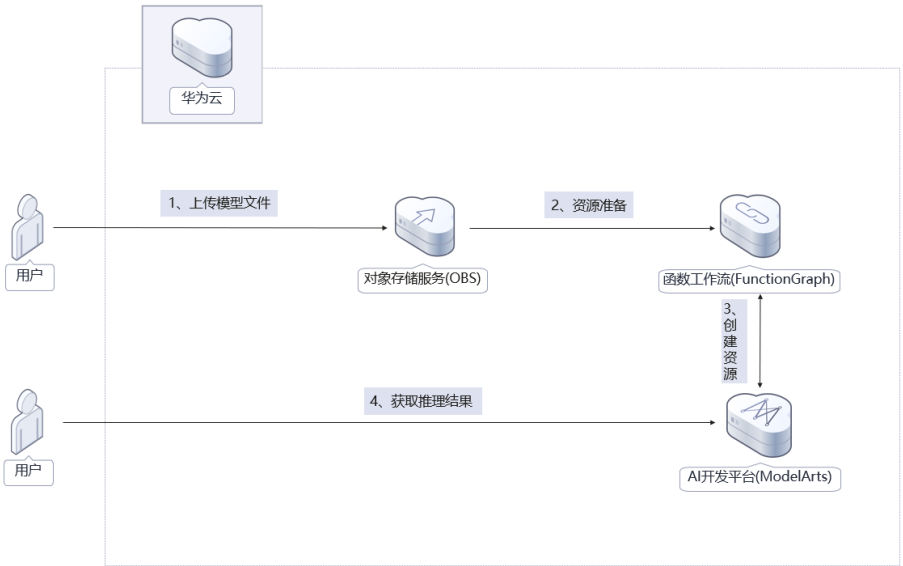
应用场景

该解决方案基于ModelArts Standard资源模式适配PyTorch NPU推理技术，将主流的开源大模型与硬件相结合，实现高速、高效的模型推理。帮助用户快速一键部署AI应用、在线推理，主要适用于自然语言处理 LLM应用场景，为用户提供更加高效、便捷的服务。

方案架构

该解决方案使用主流开源大模型，帮助用户快速搭建基于Standard适配PyTorch NPU的推理系统。

图 1-1 方案架构图



该解决方案会部署如下资源：

- 创建1台**弹性云服务器 ECS**，用于帮助用户制作镜像并上传。

- 创建1个**弹性公网IP EIP**，并关联弹性云服务器 ECS，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。
- 创建一个容器镜像服务组织，用于上传镜像。
- 使用**函数工作流 FunctionGraph**创建一个函数，用于调用AI应用、在线服务接口，实现在**AI开发平台ModelArts**上快速部署推理服务。
- 使用**AI开发平台ModelArts**，创建AI应用，部署在线服务、用于获取推理结果。
- 在**统一身份认证服务 IAM**上创建一个委托，用于授权FunctionGraph，获取IAM用户Token，访问ModelArts在线服务和对象存储服务 OBS桶。

方案优势

- 全栈自主可控

  芯片、芯片使能、AI框架、行业应用国产化，从底层芯片到上层应用实现全栈自主可控。
- 快速推理

  内置开源模型，serverless化调用服务API快速配置模型，自动部署在线服务，实现快速推理。
- 一键部署

  一键轻松部署，即可完成函数工作流、统一身份认证服务 IAM等资源创建，帮助用户快速搭建基于Standard适配PyTorch NPU的推理系统。

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。
- **此方案部署时需先执行“一键部署（制作镜像）”模板，获取镜像地址后方可执行“一键部署（部署模型）”模板。**

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 成本预估

华为云服务	配置示例	每月预估花费
弹性云服务器 ECS	• 区域：西南-贵阳一 • 按需计费：0.31元/小时 • 规格：鲲鹏通用计算增强型 kc1 | 2核 | 4GB • 镜像：EulerOS • 系统盘：高IO | 100GB • 购买量：1	222.48元
弹性公网IP EIP	• 区域：西南-贵阳一 • 按需计费：0.80元/GB • 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
AI开发平台 ModelArts	• 区域：西南-贵阳一 • 按需计费：21.72元/小时 • 计费模式：按需计费 • 业务类型：AI全流程开发 • 资源类型：公共资源池 • 规格：ModelArts昇腾AI加速型(B1)1卡实例 • 购买个数：1	15,635.52元

华为云服务	配置示例	每月预估花费
函数工作流 FunctionGraph	• 区域：西南-贵阳一 • 产品：函数 • 请求次数： 0-100万次： 0元/100万次 100万次以上： 1.33元/100万次 • 计量时间： 0-400,000 GB/秒： 0元/GB-秒 400,000 GB/秒以上： 0.00011108元/GB-秒	费用包括请求次数、计量时间两部分，详细请参考每月账单。
合计	-	**15858元 + 弹性公网IP EIP费用 +OBS费用+函数工作流费用**

3 实施步骤

3.1 准备工作

当您使用租户账号登录华为云时，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

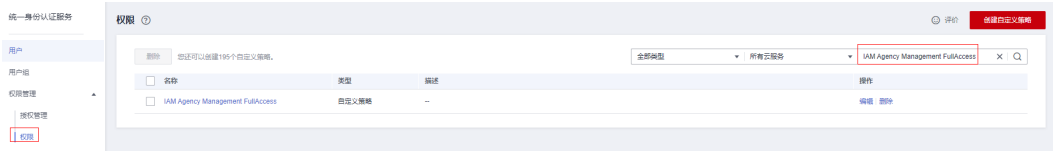
创建 IAM Agency Management FullAccess 权限（可选）

步骤1 打开“统一身份认证”菜单。

图 3-8 统一身份认证菜单

步骤2 进入“权限管理”-> “权限”菜单，在搜索框输入“IAM Agency Management FullAccess”当前账号是否存在IAM委托管理权限。

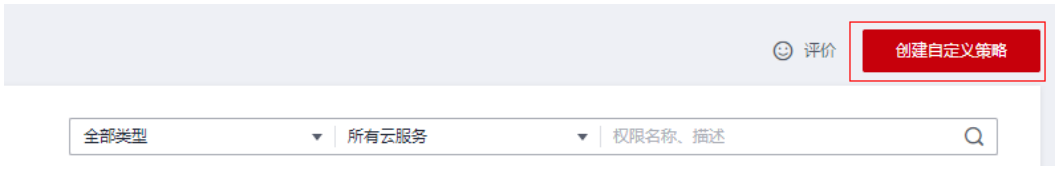
图 3-9 权限列表



- 如果搜索结果不为空，则当前账号已经存在IAM委托管理权限，不需要重复创建
- 如果过搜索结果为空，则继续创建“IAM Agency Management FullAccess”权限

步骤3 单击“创建自定义策略”按钮。

图 3-10 创建自定义策略



步骤4 输入策略名称为“IAM Agency Management FullAccess”，选择“JSON视图”，在策略内容中输入如下JSON代码，单击确认按钮。

图 3-11 创建自定义策略


```
{
   "Version": "1.1",
   "Statement": [
      {
         "Action": [
            "iam:agencies:createAgency",
            "iam:agencies:listAgencies",
```

```
            "iam:agencies:getAgency",
            "iam:agencies:deleteAgency",
            "iam:agencies:updateAgency",
            "iam:permissions:revokeRoleFromAgencyOnProject",
            "iam:permissions:revokeRoleFromAgencyOnDomain",
            "iam:permissions:revokeRoleFromAgency",
            "iam:permissions:grantRoleToAgencyOnDomain",
            "iam:permissions:grantRoleToAgencyOnProject",
            "iam:permissions:grantRoleToAgency",
            "iam:permissions:listRolesForAgencyOnDomain",
            "iam:permissions:listRolesForAgencyOnProject",
            "iam:permissions:checkRoleForAgencyOnDomain",
            "iam:permissions:checkRoleForAgencyOnProject",
            "iam:permissions:listRolesForAgency",
            "iam:permissions:checkRoleForAgency",
            "iam:roles:listRoles"
        ],
        "Effect": "Allow"
    }
  ]
}
```

步骤5 界面无报错，则成功创建IAM Agency Management FullAccess权限。

----结束

给 rf_admin_trust 委托添加 IAM Agency Management FullAccess 权限（可选）

步骤1 打开“统一身份认证”菜单。

图 3-12 统一身份认证菜单

步骤2 进入“委托”菜单，选择rf_admin_trust委托。

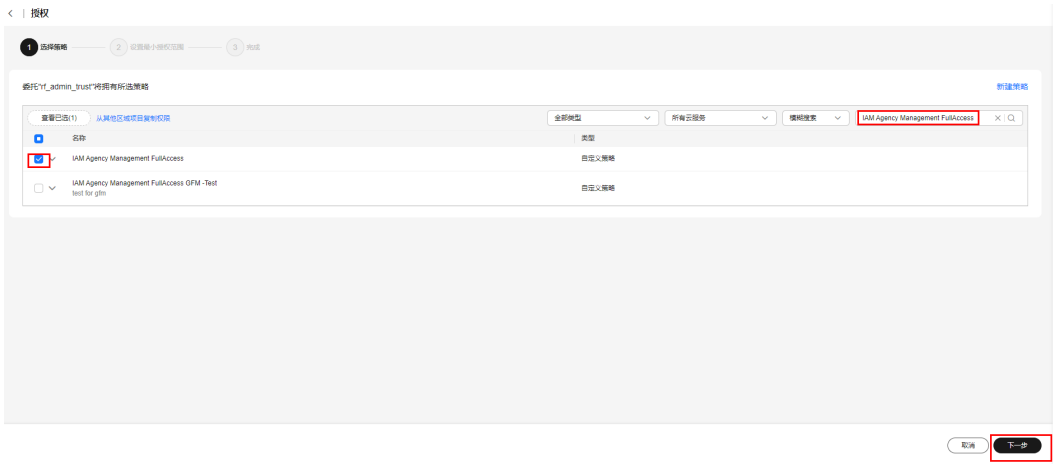
图 3-13 委托列表

步骤3 进入“授权记录”菜单，单击“授权”按钮。

图 3-14 授权记录

步骤4 在搜索框输入IAM Agency Management FullAccess，勾选过滤出来的记录，单击下一步，并确认完成权限的配置。

图 3-15 配置 IAM Agency Management FullAccess 策略



步骤5 配置好后的情况：rf_admin_trust委托拥有Tenant Administrator和IAM Agency Management FullAccess权限。

图 3-16 授权记录列表

----结束

获取 SWR 临时登录指令（镜像制作）

说明

当您使用“一键部署(镜像制作)”模板进行部署时，请按照提供的说明获取登录指令。

步骤1 访问**容器镜像服务控制台**，单击“登录指令按钮”按下图所示获取临时登录指令。

图 3-17 获取临时登录指令

----结束

ModelArts 创建及部署（部署模型）

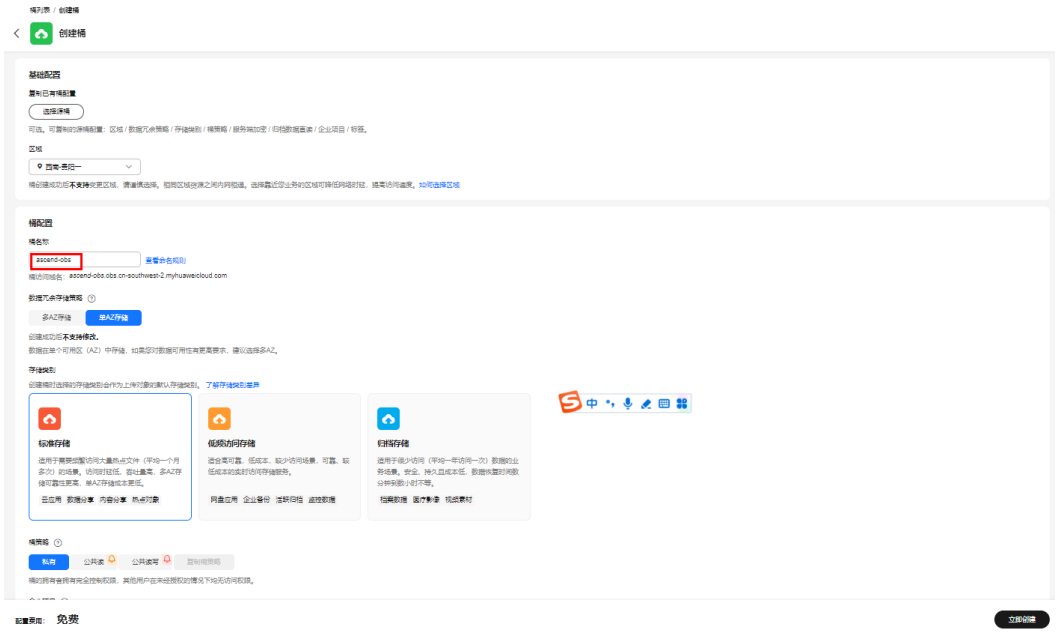
说明

当您使用“一键部署(部署模型)”模板进行部署时，请按以下步骤上传权重文件。

步骤1 准备一个OBS桶：（如果已有，可跳过此步骤）登录华为云**对象存储服务控制台**，单击“创建桶”进入obs桶创建界面，按照提示命名规则输入桶名称，单击“立即创建”。

图 3-18 进入 OBS 桶创建界面

图 3-19 创建 OBS 桶



步骤2 **示例权重文件：**单击此**文件下载地址**，下载权重文件、SSL证书压缩包，解压并通过拖拽文件夹的方式上传至**3.1准备工作步骤1**准备的OBS桶中。

权重文件：若需要部署其他模型，单击此**文件下载地址**，下载权重文件并上传。

文件目录层级如下图所示：

图 3-20 上传权重文件

----结束

3.2 快速部署

本章节主要帮助用户快速部署“基于PyTorch NPU快速部署开源大模型”解决方案。

表 3-1 参数说明（制作镜像）

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）。	model-inference-based-on-npus-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（.）。	model-inference-based-on-npus-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）。	model-inference-based-on-npus-demo
swr_name	string	必填	swr组织名称，不支持重名。取值范围：2-64个字符，小写字母开头，支持小写字母、数字、-（中划线），小写字母或数字结尾。	model-inference-based-on-npus-demo
entry_instructions	string	必填	swr临时登录指令，注意开头和结尾需要加英文双引号（"）,示例："docker login -u cn-southwest-2@xxx -p xxx swr.cn-southwest-2.myhuaweicloud.com"，请参考**获取临时登录指令**。	空
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

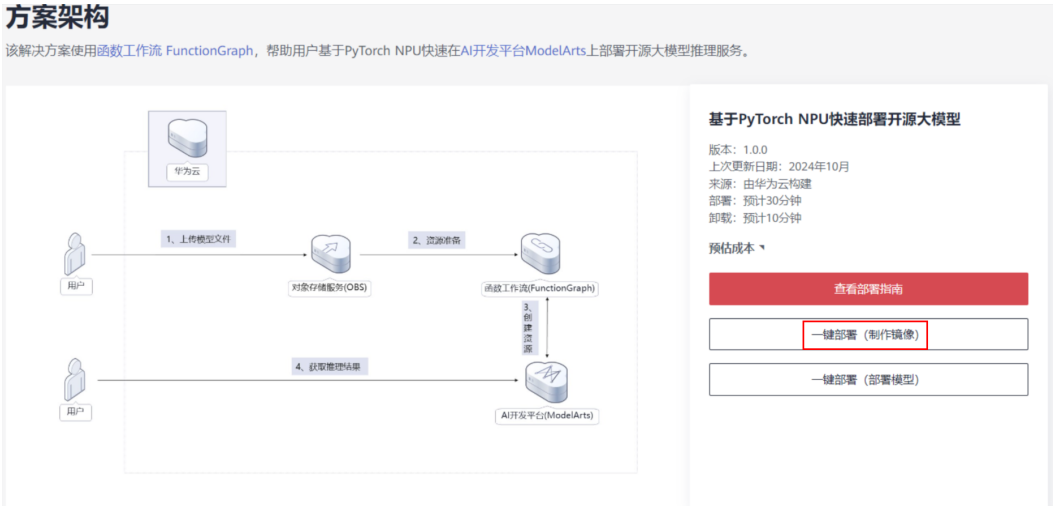
表 3-2 参数说明（部署模型）

参数名称	类型	是否可选	参数解释	默认值
functiongraph_name	string	必填	函数工作流 Functiongraph函数名称，不支持重名。取值范围：长度为2-57个字符，支持字母、数字、_（下划线）和-（中划线），以字母开头，以字母或数字结尾。	model-inference-based-on-npus-demo
domain_username	string	必填	IAM用户所属的华为云账号名称。取值范围：6-30 个字符，以字母开头，支持字母、数字、下划线（_）、中划线（-）。	空
username	string	必填	IAM用户名。取值范围：1-64个字符，支持字母、数字、下划线（_）、中划线（-）、点（.）不能以数字或空格开头，如果您的华为云账号已升级为华为账号，将不支持获取账号Token，建议您为自己创建一个IAM用户，授予该用户必要的权限，获取IAM用户Token。	空
password	string	必填	IAM用户密码。取值范围：8-32个字符，支持字母、数字、特殊字符，不能包含空格，为避免获取Token失败，请务必保证密码输入正确。	空
model_obs_path	string	必填	模型所在的OBS路径。路经格式：https://桶名.obs.cn-southwest-2.myhuaweicloud.com/模型文件路径/。	空
environment_swr_path	string	必填	模型运行的SWR环境路径，请参考**获取镜像地址**，示例：swr.cn-southwest-2.myhuaweicloud.com/xxx/pytorch_2_1_ascend:909。	空
service_name	string	必填	在线服务名称。支持1-64位字符，可包含字母、中文、数字、中划线、下划线。	model-inference-based-on-npus-demo
service_running_time	string	必填	服务运行的时间。单位：小时。取值范围：1-24的正整数。例如：1小时后停止服务，此参数填1。	1

参数名称	类型	是否可选	参数解释	默认值
specification	string	必填	在线服务资源规格。当前版本仅支持公共资源池的规格，可选modelarts.vm.cpu.2u/modelarts.vm.gpu.pnt004(需申请)/modelarts.vm.ai1.snt3(需申请)/custom(仅支持在部署到专属资源池时使用)，需申请的规格请提交工单，由ModelArts运维工程师添加权限。	modelarts.bm.arm.snt9b1
instance_count	string	必填	在线服务模型部署的实例数。取值范围：1-128的正整数，当前限制最大实例数为128，如需使用更多的实例数，需提交工单申请。	1
ascend_rt_visible_devices	string	必填	在线服务NPU卡的数量，单卡设为0，4卡设为0,1,2,3。	0
model_path	string	必填	在线服务模型路径，格式为/home/mind/model/权重文件夹名称，例如：/home/mind/model/Qwen-7B-Chat。	/home/mind/model/Qwen-7B-Chat

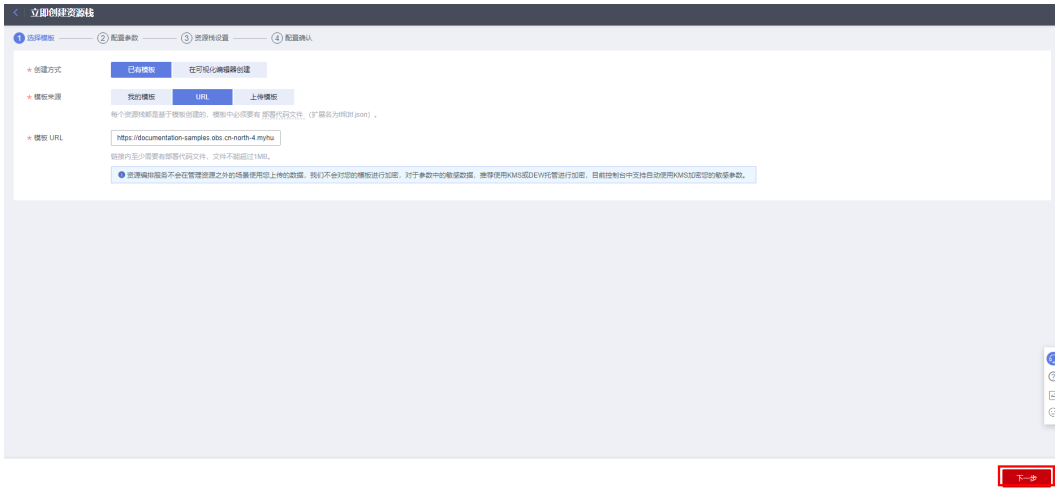
步骤1 登录华为云解决方案实践，选择“基于PyTorch NPU快速部署开源大模型”，单击“一键部署（制作镜像）”，跳转至解决方案创建资源栈界面。

图 3-21 解决方案实践



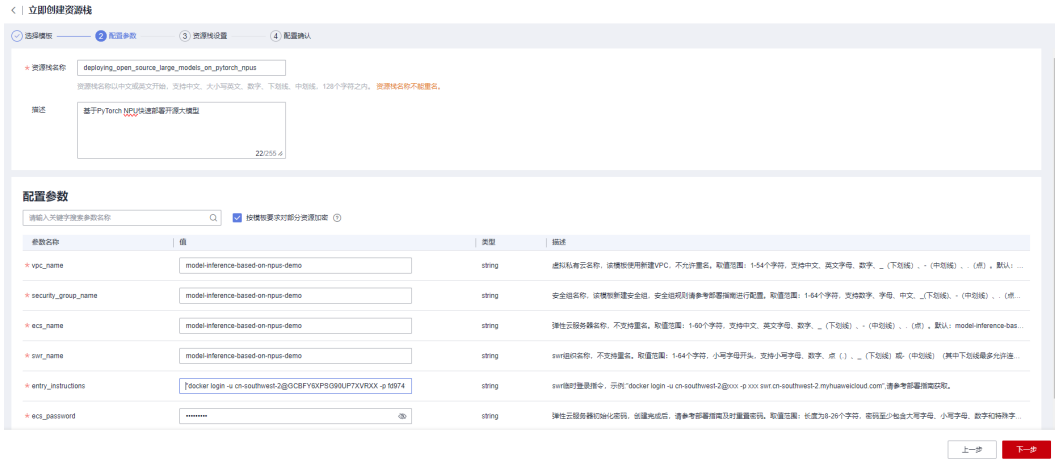
步骤2 在选择模板界面中，单击“下一步”。

图 3-22 选择模板



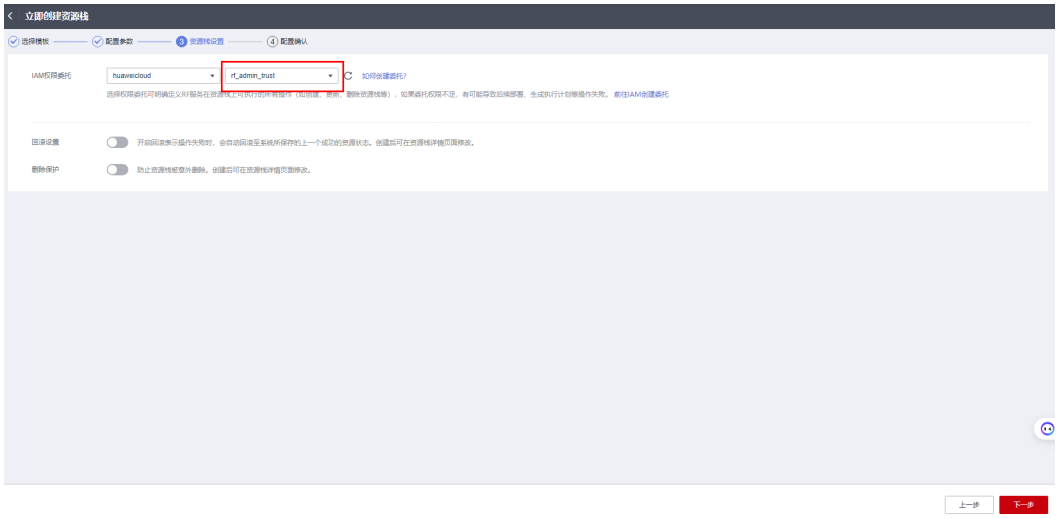
步骤3 在配置参数界面中，请按下面的描述完成对应参数填写。参考**表1 参数说明（制作镜像）**完成自定义参数填写，单击“下一步”。

图 3-23 配置参数



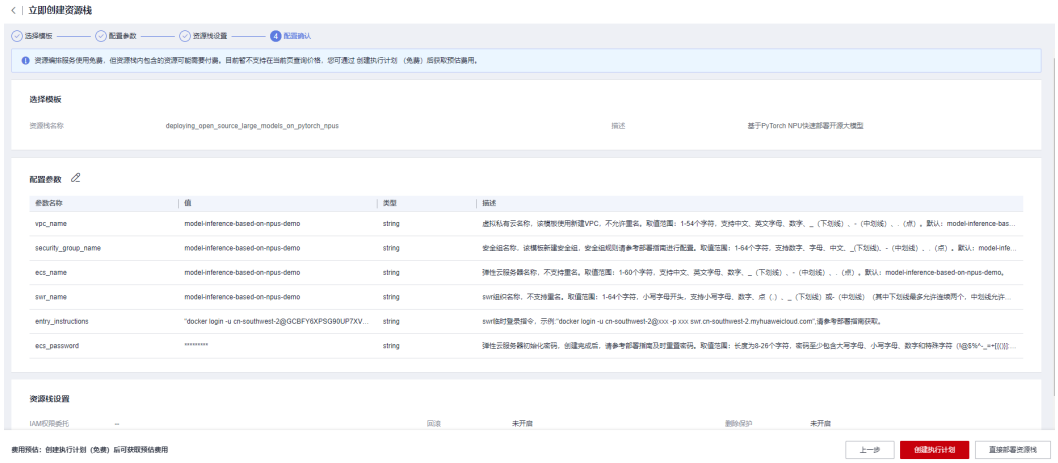
步骤4 在资源设置界面中，在权限委托下拉框中选择“rf_admin_trust”委托（可不选），单击“下一步”。

图 3-24 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 3-25 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 3-26 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 3-27 执行计划

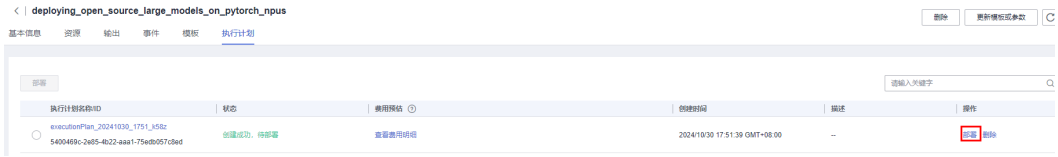
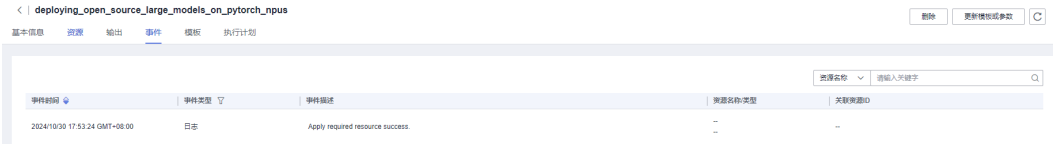


图 3-28 执行计划确认

步骤8 待“事件”中出现“Apply required resource success”，表示该解决方案已经部署完成。

图 3-29 部署完成



步骤9 参考**制作镜像**，获取镜像地址。

步骤10 访问**华为云解决方案实践**，选择“基于PyTorch NPU快速部署开源大模型”，单击“一键部署（部署模型）”，跳转至解决方案创建资源栈界面，其余部署参考以上步骤2-8，（**注：步骤3参考表2 参数说明（部署模型）完成自定义参数填写**）。

----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。

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。

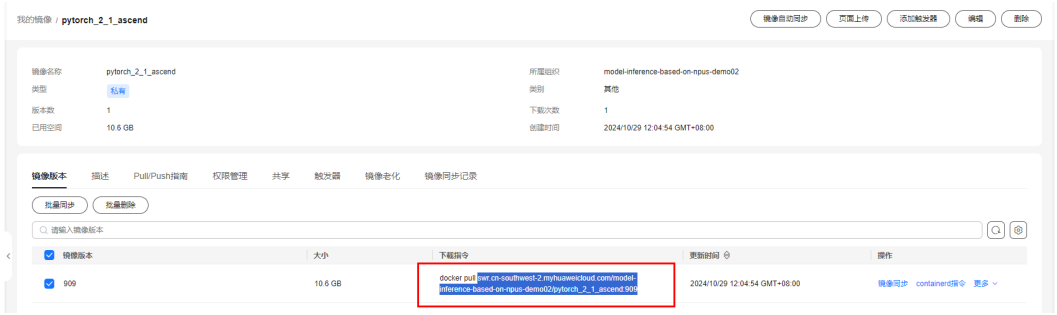
- 添加安全组规则：根据业务使用需求需要开放某个TCP端口，请参考**添加安全组规则**添加入方向规则，打开指定的TCP端口。
- 修改安全组规则：安全组规则设置不当会造成严重的安全隐患。您可以参考**修改安全组规则**，来修改安全组中不合理的规则，保证云服务器等实例的网络安全。
- 删除安全组规则：当安全组规则入方向、出方向源地址/目的地址有变化时，或者不需要开放某个端口时，您可以参考**删除安全组规则**进行安全组规则删除。

制作镜像

步骤1 访问**容器镜像服务控制台**，按下图所示，单击“镜像名称”进入镜像详情页，获取镜像地址（一键部署（制作镜像）模板部署完成后约10分钟，镜像制作完成）**仅复制镜像地址不需要docker pull 命令**。

图 3-30 容器镜像服务

图 3-31 获取镜像地址



----结束

部署模型

步骤1 进入**函数工作流控制台**选择此方案创建的函数，单击函数名称进入函数主页。

图 3-32 进入函数主页

步骤2 单击“测试”在弹出窗口中选择“空白模板”单击“创建”配置测试事件

图 3-33 函数主页

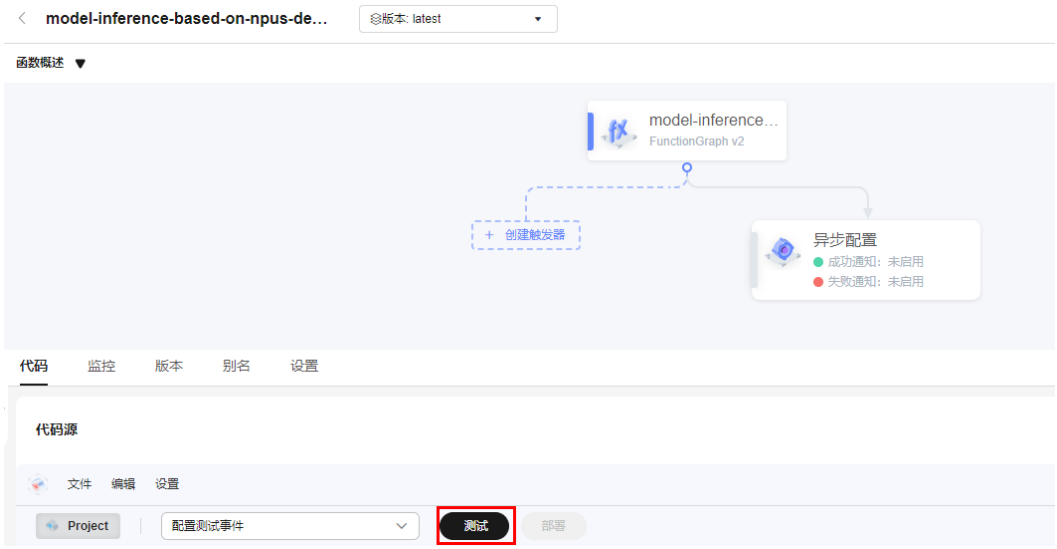
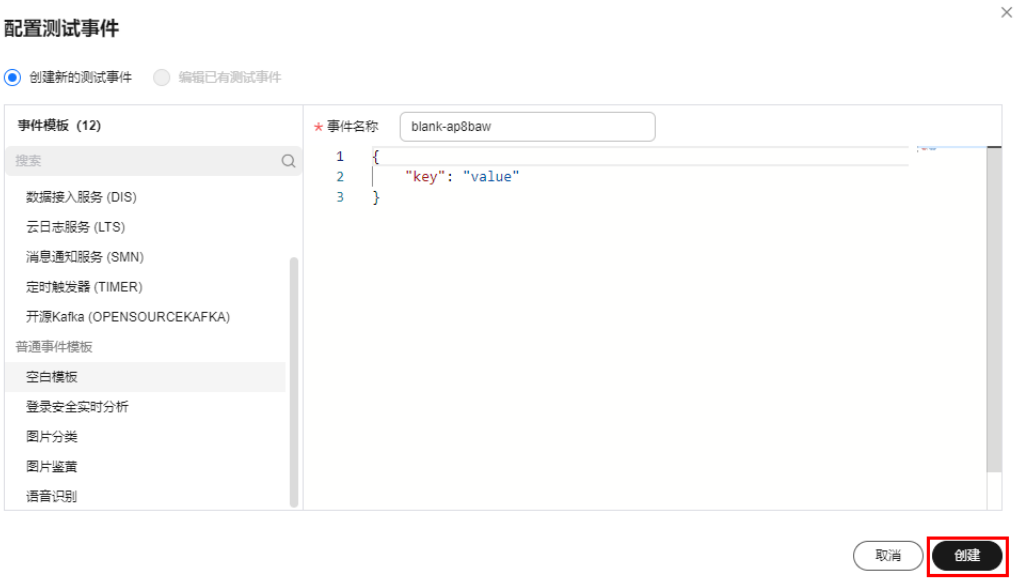
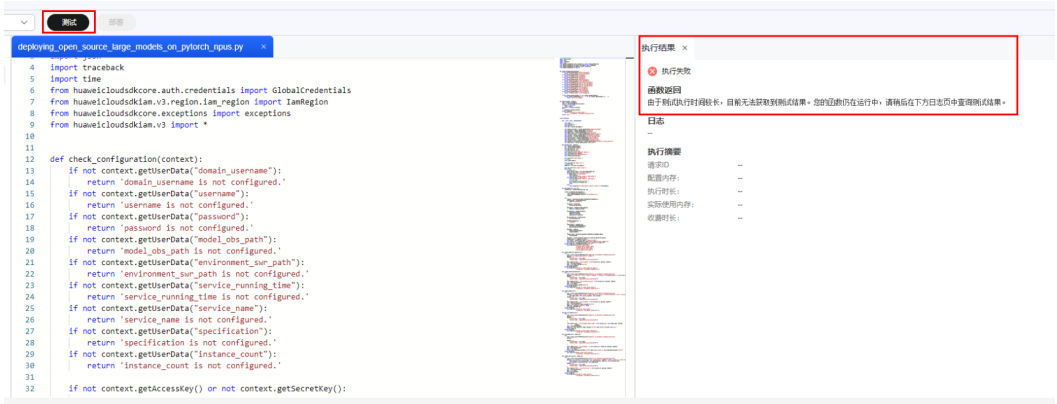


图 3-34 配置测试事件



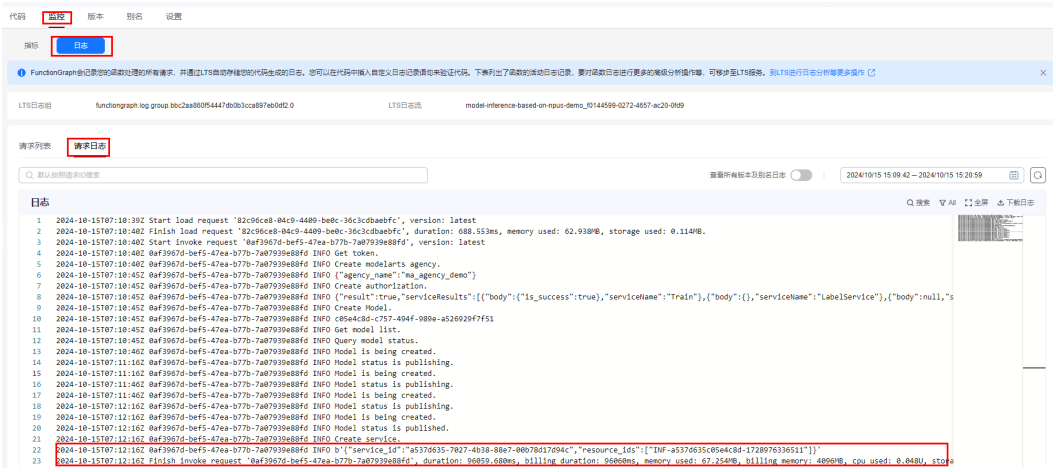
步骤3 在函数主页，单击“测试”调用函数，通过SDK获取token，创建modelarts委托，配置授权、创建AI应用，部署在线服务。（若出现右侧提示，为函数执行时间较长，调用正常，请继续执行下面的步骤。）

图 3-35 调用函数



步骤4 在函数主页，依次单击“监控”，“日志”，请求日志，查看Modelarts资源创建相关日志信息，无红色字体错误信息，显示如下图，则表示资源创建完成。

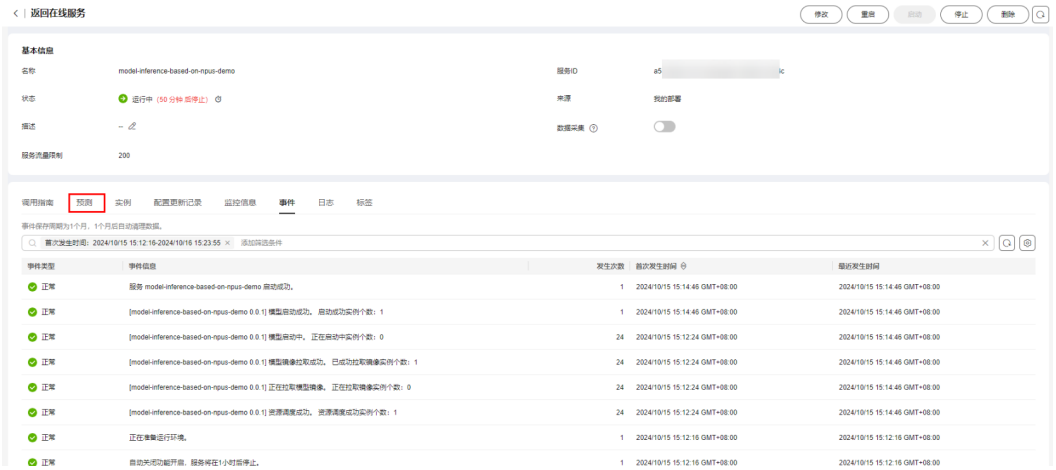
图 3-36 日志信息



步骤5 访问**AI开发平台ModelArts在线服务控制台**单击在线服务名称，进入在线服务管理页面，单击“预测”。

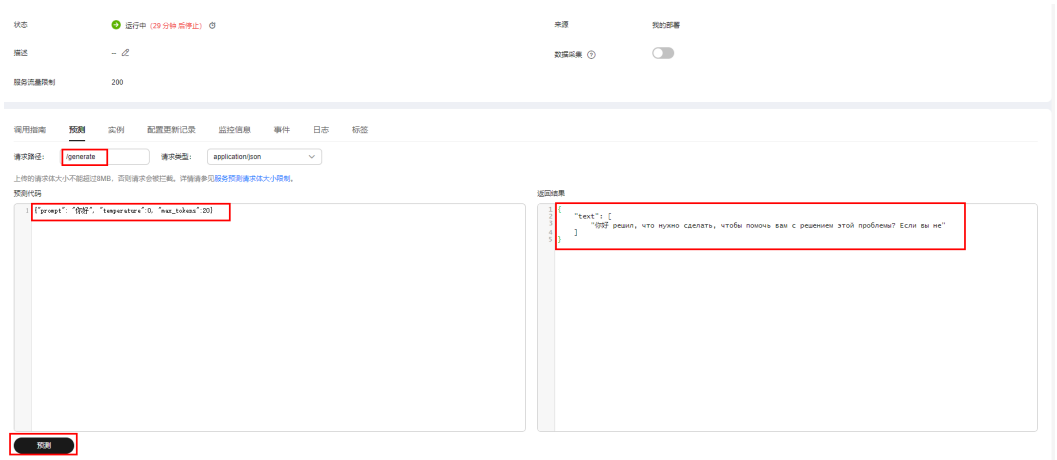
图 3-37 ModelArts 在线服务控制台

图 3-38 在线服务管理页面



步骤6 在在线服务预测界面中，输入请求路径“generate”预测代码中填写 {"prompt": "你好", "temperature":0, "max_tokens":20}，单击"预测"即可获取返回结果。（若以openai接口启动服务，请求路径：“/v1/completions”，输入预测代码“{"prompt": "你是谁","model": "${model_path}","max_tokens": 50,"temperature":0}”，单击“预测”既可看到预测结果。model_path为**表3-2**中model_path的值）

图 3-39 预测



步骤7 按下图所示，在调用指南中获取API接口公网地址，参考右侧调用指南，实现通过API接口调用在线服务。

图 3-40 调用指南

----结束

3.4 快速卸载

手动卸载

删除制作镜像资源栈时需手动删除swr组织下的镜像，请按以下步骤操作，完成后方可执行一键卸载步骤。

步骤1 访问容器镜像服务我的镜像管理页面，按下图所示，选择镜像名称，单击批量删除，在弹出的删除镜像确认框中输入DELETE，单击确定。

图 3-41 SWR 组织管理

----结束

一键卸载

登录资源编排 RFS资源栈，找到该解决方案创建的两个资源栈，参考以下步骤进行资源删除。

步骤1 单击该方案资源栈后的“删除”。

图 3-42 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 3-43 删除资源栈确认

----结束

说明

此方案一键卸载后仍有部分资源残留，需请参考**手动卸载**进行卸载。

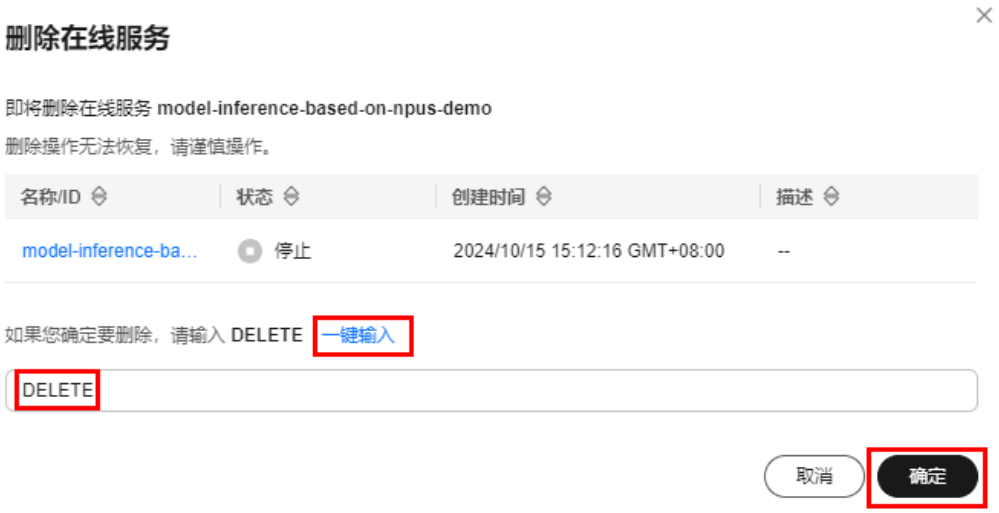
手动卸载的资源： ModelArts在线服务、AI应用、OBS桶

手动卸载

步骤1 删除在线服务：访问**ModelArts 在线服务控制台**，按下图所示，依次单击“更多”“删除”，在弹出的确认窗口中单击“确定”。

图 3-44 删除在线服务

图 3-45 确认删除在线服务



步骤2 删除AI应用：访问**ModelArts AI应用控制台**，如下图所示，单击“删除”，在弹出的确认窗口中单击“确定”。

图 3-46 删除 AI 应用

图 3-47 确认删除 AI 应用

步骤3 （可选）删除手动创建的OBS桶：登录**对象存储服务 OBS控制台**，查找在**3.1准备工作步骤1**创建的OBS桶，单击“删除”，在弹出的确认窗口中单击“确定”（注意：删除桶需桶中无文件）。

----结束

4 附录

名词解释

- 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。
- 对象存储服务 OBS：是一个基于对象的海量存储服务，为客户提供海量、安全、高可靠、低成本的数据存储能力。
- 函数工作流 FunctionGraph：是一项基于事件驱动的函数托管计算服务。使用FunctionGraph函数，只需编写业务函数代码并设置运行的条件，无需配置和管理服务器等基础设施，函数以弹性、免运维、高可靠的方式运行。此外，按函数实际执行资源计费，不执行不产生费用。
- AI开发平台 ModelArts：面向开发者的一站式AI开发平台，可快速创建和部署模型，管理全周期AI工作流，助力千行百业智能升级。
- 统一身份认证服务 IAM：是华为云提供权限管理、访问控制和身份认证的基础服务，您可以使用IAM创建和管理用户、用户组，通过授权来允许或拒绝对云服务和资源的访问，通过设置安全策略提高账号和资源的安全性，同时IAM为您提供多种安全的访问凭证。

5 修订记录

表 5-1 修订记录

发布日期	修订记录
2024-10-30	第一次正式发布。