解决方案实践

基于开源模型构建高可用 AIGC 应用

文档版本 1.0

发布日期 2023-07-18

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

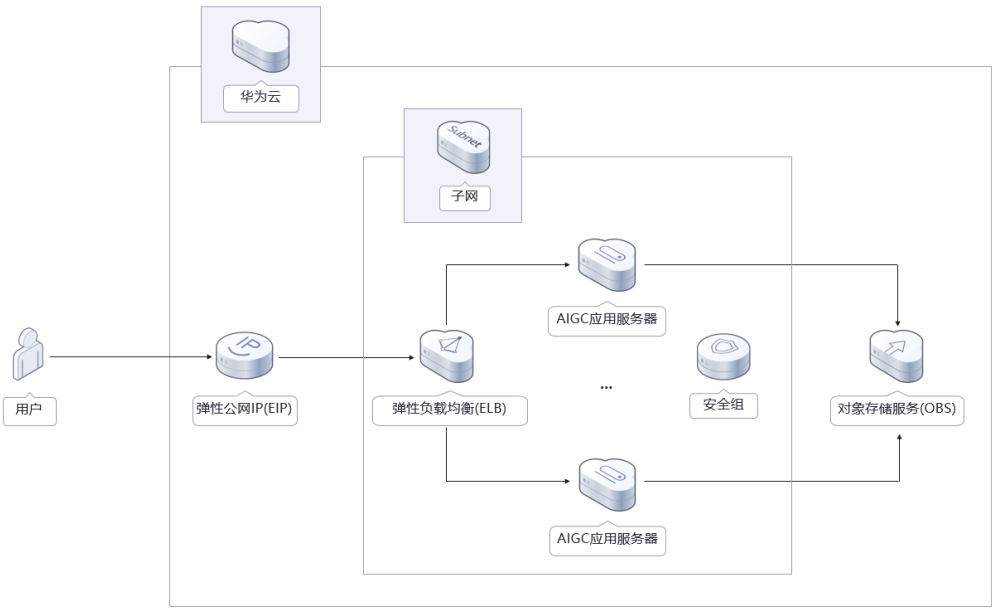
应用场景

该解决方案可以帮助您在华为云弹性云服务器 ECS上基于Stable Diffusion构建高可用AIGC Web应用。Stable diffusion是一种基于潜在扩散模型（Latent Diffusion Models）的文本到图像生成模型，能够根据输入文本和图像生成高质量图像。

方案架构

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2台Linux GPU加速型弹性云服务器 ECS，用于搭建AIGC应用系统。
- 创建三个弹性公网IP EIP，分别绑定到两个到服务器及**弹性负载均衡 ELB**，用于提供访问公网和被公网访问能力。

- 部署一个弹性负载均衡 ELB，用于业务流量跨可用区进行分发。
- 创建一个对象存储服务 OBS桶，用于保存生成的图片文件。
- 在两台Linux弹性云服务器 ECS上分别完成Stable Diffusion WebUI应用、inotify-tools工具安装，以及对象存储服务 OBS obsutil工具安装，用于自动上传备份在页面上保存的图片。

方案优势

- 高可用

  弹性云服务器 ECS跨可用区部署，提供多可用区容灾能力，够快速自动完成故障切换。

- 开源和定制化

  该解决方案是开源的，用户可以免费用于商业用途，并且还可以在源码基础上进行定制化开发。

- 一键部署

  一键轻松部署，即可实现基于Stable Diffusion的高可用AIGC应用系统搭建。

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。

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：7.65元/小时 • 区域：亚太-新加坡 • 计费模式：按需计费 • 规格： GPU加速型 Pi2 | 8核 | 32GB | 加速卡：1 * NVIDIA T4 / 1 * 16G • 镜像： Ubuntu 20.04 server 64bit with Tesla Driver 460.73.01 and CUDA 11.2 Ubuntu 20.04 server 64bit with Tesla Driver 470.182.03 and CUDA 11.4 • 系统盘：高IO | 100GB • 购买量：2	11016.00 元
弹性公网IP EIP	• 按需计费：0.82元/GB • 区域：亚太-新加坡 • 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：2	84.15元

华为云服务	配置示例	每月预估花费
弹性公网IP EIP	按需计费（按带宽计费）：0.34元/小时 区域：亚太-新加坡 ● 计费模式：按带宽计费 ● 线路：动态BGP ● 公网带宽：按带宽计费 ● 带宽大小：5Mbit/s ● 购买量：1	644.98元
弹性负载均衡 ELB	共享型负载均衡(性能保障模式) ● 按需计费：0.32元/小时 ● 区域：亚太-新加坡 ● 计费模式：按需计费 ● 购买量：1	230.40元
对象存储服务 OBS	● 区域：亚太-新加坡 ● 存储空间：数据存储（多AZ存储） ● 默认存储类别：标准存储 ● 桶策略：私有 ● 请求费用：0.0100元/万次 ● 存储空间：0.1390元/GB/月 ● 流量费用： 内/公网流入流量（数据上传到OBS） 0元 内网流出流量 0元 公网流出流量 / 00:00-08:00（闲时）0.2500元/G 公网流出流量 / 08:00-24:00（忙时）0.5000元/GB 费用包括存储空间、请求费用、流量费用两部分，具体请参考**OBS计费详情**。	费用包括存储空间、请求费用、流量费用两部分，详细请参考每月账单。
合计	-	**11975.53元 + OBS服务产生费用**

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：亚太-新加坡 • 计费模式：包月 • 规格：GPU加速型 Pi2 | 8核 | 32GB | 加速卡：1 * NVIDIA T4 / 1 * 16G • 镜像： Ubuntu 20.04 server 64bit with Tesla Driver 460.73.01 and CUDA 11.2 Ubuntu 20.04 server 64bit with Tesla Driver 470.182.03 and CUDA 11.4 • 系统盘：高IO | 100GB • 购买量：2	8,283.00元
弹性公网IP EIP	• 按需计费：0.82元/GB • 区域：亚太-新加坡 • 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：20GB	84.15元
弹性公网IP EIP	• 区域：亚太-新加坡 • 计费模式：按带宽计费 • 线路：动态BGP • 公网带宽：按带宽计费 • 带宽大小：5Mbit/s • 购买量：1	407.75元
弹性负载均衡 ELB	共享型负载均衡(性能保障模式) • 按需计费：0.32元/小时 • 区域：亚太-新加坡 • 计费模式：按需计费 • 购买量：1	230.40元

华为云服务	配置示例	每月预估花费
对象存储服务 OBS	• 区域：亚太-新加坡 • 存储空间：数据存储（多AZ存储） • 默认存储类别：标准存储 • 桶策略：私有 • 请求费用：0.0100元/万次 • 存储空间：0.1390元/GB/月 • 流量费用： 内/公网流入流量（数据上传到OBS） 0元 内网流出流量 0元 公网流出流量 / 00:00-08:00（闲时）0.2500元/G 公网流出流量 / 08:00-24:00（忙时）0.5000元/GB 费用包括存储空间、请求费用、流量费用两部分，具体请参考**OBS计费详情**。	费用包括存储空间、请求费用、流量费用两部分，详细请参考每月账单。
合计	-	**9005.30 元 + OBS服务产生费用**

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

获取 AK、SK 密钥

部署该方案之前，需要您在华为云控制台获取AK、SK密钥，您将会在**3.2-快速部署**中填写参数以完成高可用AIGC应用系统的搭建。

根据**官方文档**指引，在控制台--我的凭证--访问密钥中配置访问密钥AK并下载秘密访问密钥SK。

图 3-8 创建 AK，SK

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）。	high-availability-aigc-applications-demo
security_group_name	String	必填	安全组名称，该模板新建安全组。取值范围：1-64个字符，支持数字、字母、中文、_（下划线）、-（中划线）、.（点）。	high-availability-aigc-applications-demo

参数名称	类型	是否必填	参数解释	默认值
ecs_name	String	必填	弹性云服务器名称，不支持重名。命名方式为{ecs_name}-数字，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）。	high-availability-aigc-applications-demo
image_bucket_name	String	必填	OBS桶名称，全局唯一，用于自动上传WebUI生成的图片。取值范围：3-63个字符，支持小写字母、数字、中划线（-）、英文句号（.），禁止以中划线（-）或英文句号（.）开头及结尾。	空
ecs_count	String	必填	弹性云服务器数量，取值范围：大于等于1，上限由用户配额决定。具体请登录华为云官网**我的配额**查看。	2
ecs_flavor	String	必填	弹性云服务器规格，需选取GPU加速型（仅支持p2v、pi2、g6、g5系列服务器），请参考**弹性云服务器规格清单**。（使用前请到**华为云服务器控制台**查询，需选择**表2-1**中支持的镜像规格，不然会导致方案创建失败。）	pi2.2xlarge.4（根据站点适配， 具体以一键部署模板代码展示的默认值为准。）
ecs_password	String	必填	弹性云服务器初始化密码，创建完成后请参考**3.3开始使用步骤1**重置密码。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，Windows系统密码不能包含用户名或用户名的逆序，不能包含用户名中超过两个连续字符的部分。管理员账户默认root。	空

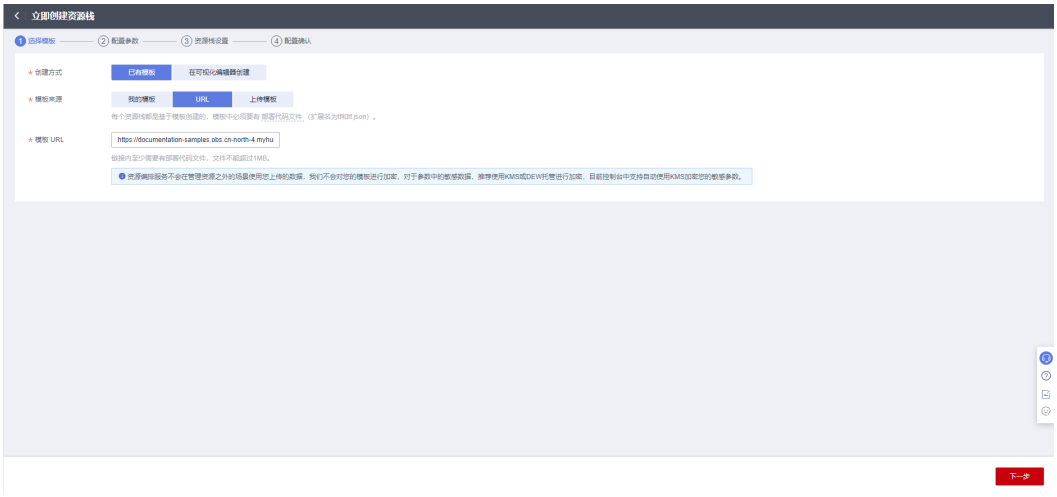
参数名称	类型	是否必填	参数解释	默认值
elb_name	String	必填	弹性负载均衡 ELB名称，取值范围：1-64个字符组成，支持中文、英文字母、数字、_（下划线）、-（中划线）、.（英文句号）。	high-availability-aigc-applications-demo
eip_bandwidth_size	Number	必填	弹性公网带宽大小，该模板计费方式为按流量计费。取值范围：1-300Mbit/s。	300
charging_mode	String	必填	计费模式，默认自动扣费，取值为prePaid（包年包月）或postPaid（按需计费）。	postPaid
charge_period_unit	String	必填	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。取值范围：month（月），year（年）。	month
charge_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。取值范围：charging_unit=month（周期类型为月）时，取值为1-9；charging_unit=year（周期类型为年）时，取值为1-3。	1
access_key_id	String	必填	访问密钥ID（AK）,识别访问用户的身份，用于将生成的图像上传至OBS桶。参考**获取AK、SK密钥**。	空
secret_access_key	String	必填	秘密访问密钥（SK）,对请求数据进行签名验证，用于将生成的图像上传至OBS桶。参考**获取AK、SK密钥**。	空

步骤1 登录**华为云解决方案实践**，选择“基于开源模型构建高可用AIGC应用”并单击，跳转至该解决方案一键部署界面。

图 3-9 解决方案实施库

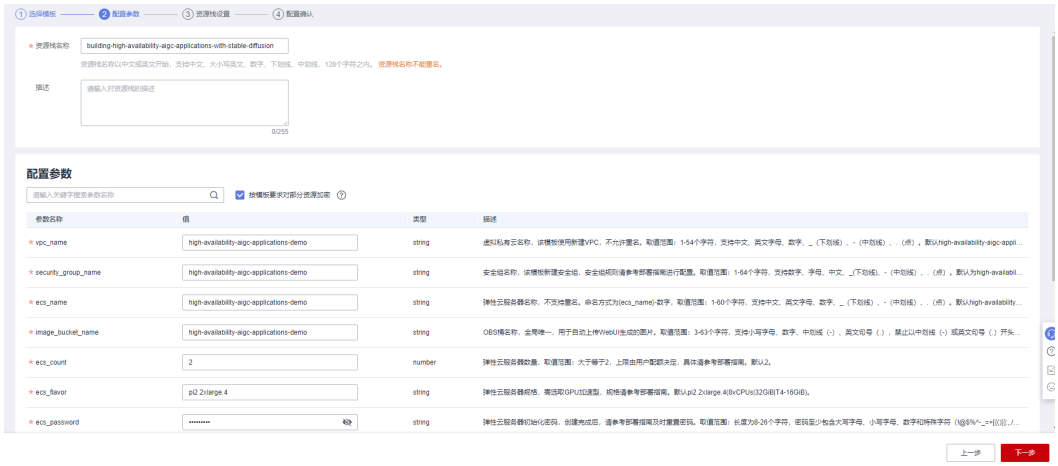
步骤2 单击“一键部署”，跳转至该解决方案创建资源栈部署界面。

图 3-10 创建资源栈



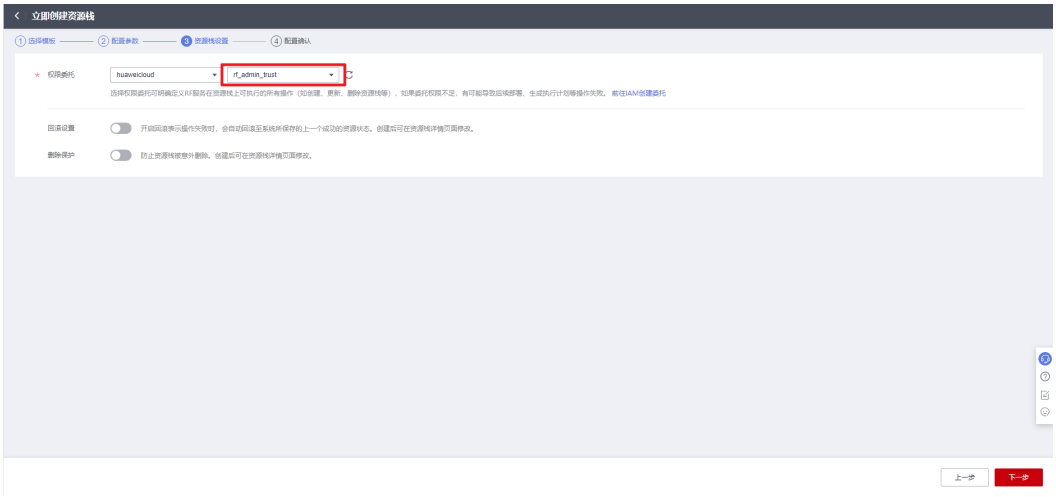
步骤3 单击“下一步”，参考**表3-1**完成自定义参数填写。

图 3-11 参数配置



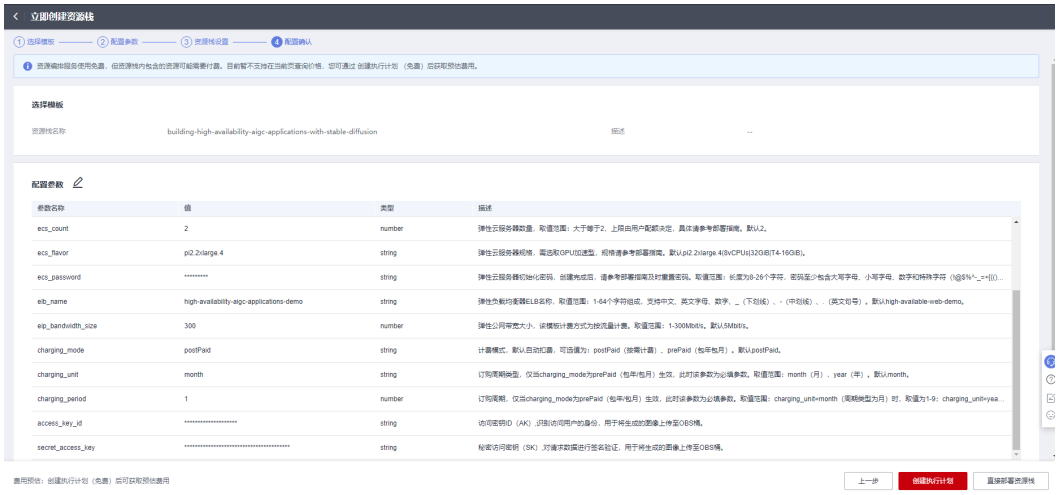
步骤4 （可选）在资源设置界面中，“权限委托”下拉框中选择“rf_admin_trust”委托，单击“下一步”。

图 3-12 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 3-13 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 3-14 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 3-15 执行计划确认

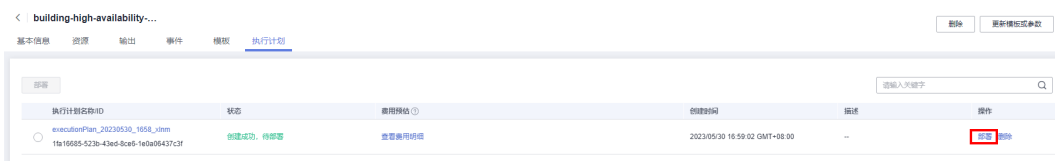
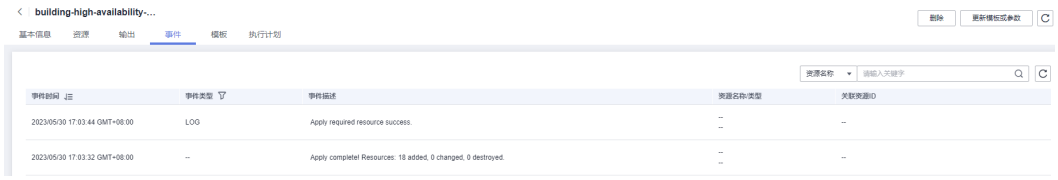


图 3-16 确认执行

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 3-17 资源创建成功



步骤10 刷新页面，在“输出”中查看WebUI访问说明。

图 3-18 输出

----结束

3.3 开始使用

安全组规则修改（可选）

须知

该解决方案默认只创建ping安全组规则，用户需在登录弹性云服务器前添加入方向规则。比如登录Windows弹性云服务器，指定登录端口为3389，并添加白名单IP。

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。

为应用配置域名（可选）

配置域名解析。网站解析将域名与**3.2快速部署步骤9**中网址IP地址相关联，实现通过在浏览器中直接输入域名访问网站。具体解析流程参考**快速添加域名解析**。

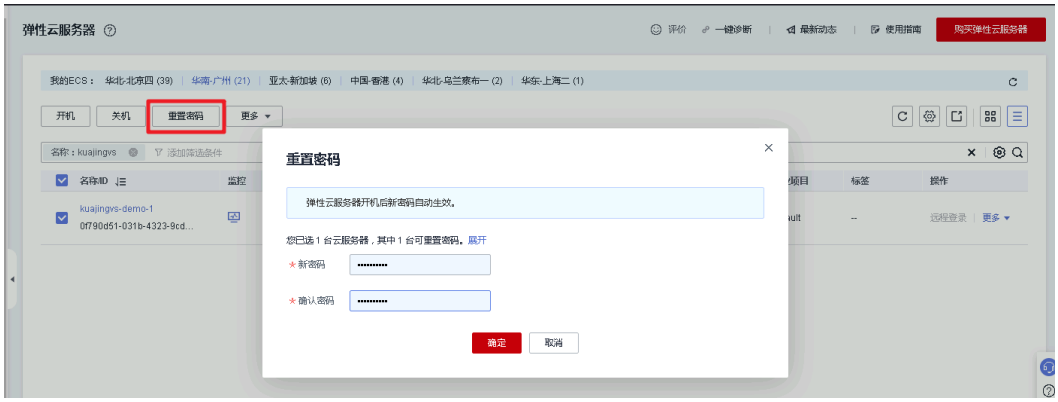
使用 AIGC WebUI 应用

步骤1 （可选）登录**华为云服务器控制台**，按下图所示，单击“立即开启防护”，开启防护。

图 3-19 开始防护

步骤2 登录**华为云服务器控制台**，修改初始化密码。参考**在控制台重置弹性云服务器密码**，进行密码重置。

图 3-20 重置密码



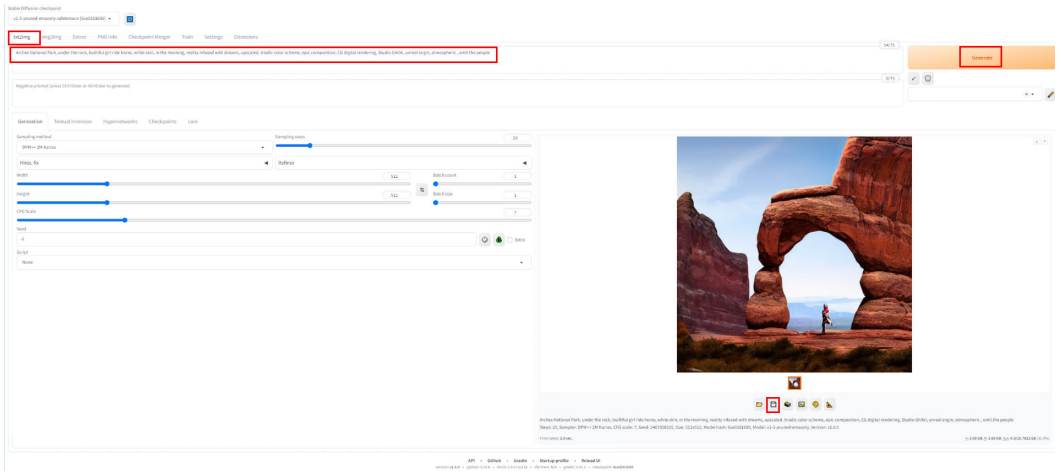
步骤3 进入**弹性负载均衡控制台**，在左侧导航栏单击“后端服务器组”单击“名称”选择“后端服务器”查看服务器状态是否正常。（说明：按照默认参数资源部署完成，20分钟后业务初始化完成，后端服务器7860端口状态正常。）

图 3-21 后端服务器组

图 3-22 查看服务器状态

步骤4 查看**快速部署 步骤3.2-10**访问说明，访问AIGC web UI界面。单击“txt2img”在提示词框中填写提示词，单击“Generate”生成图像，待图像生成完成后，单击保存。Stable diffusion web UI 界面的详细使用说明，请访问开源项目**stable diffusion webui**或者查询网络相关教程获取。**本方案新建aigc用户，默认密码为aigc@123 。**

图 3-23 AIGC WebUI 应用界面



提示词示例：
Arches National Park,under the rock,in the evening,nightsky,reality infused with dreams,upscaled,triadic color scheme,epic composition,CG digital rendering,Studio Ghibli,unreal engine,atmospheric,omit the people,

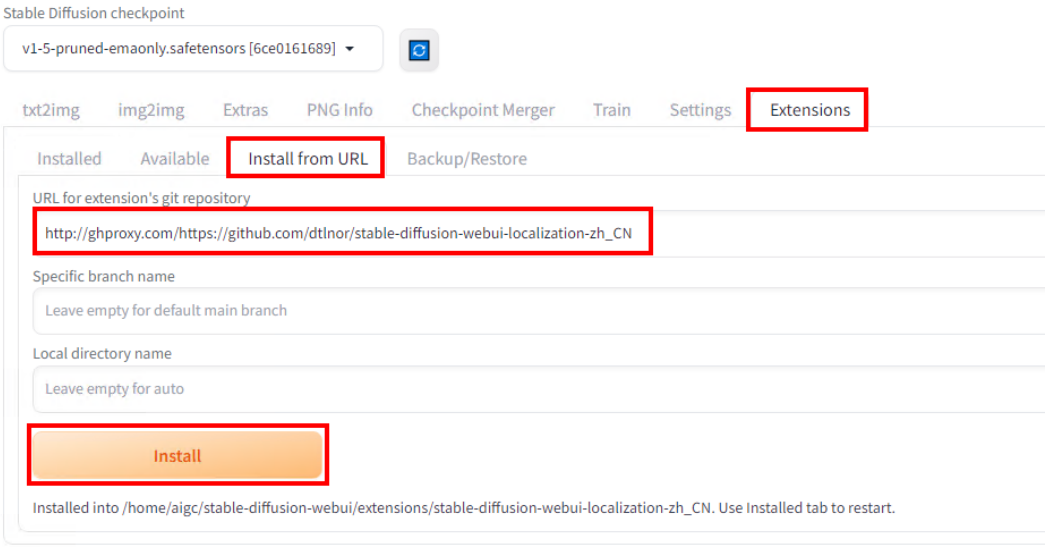
步骤5 进入**对象存储服务控制台**单击**快速部署 步骤3.2-3**创建的OBS桶进入，即可查看步骤3中保存的图片，还可以通过"分享"按钮，分享图片。更多OBS功能请查看**对象存储服务 OBS用户指南对象管理文档**。

图 3-24 对象存储服务桶列表

图 3-25 查看保存的图片

步骤6 （可选）界面汉化插件安装，进入界面后依次单击 extension 选项卡，Install from URL 子选项卡，复制 git 仓库网址：https://github.com/dtlnor/stable-diffusion-webui-localization-zh_CN （国内region使用此网址: http://ghproxy.com/https://github.com/dtlnor/stable-diffusion-webui-localization-zh_CN ），单击 install 即可安装完成。

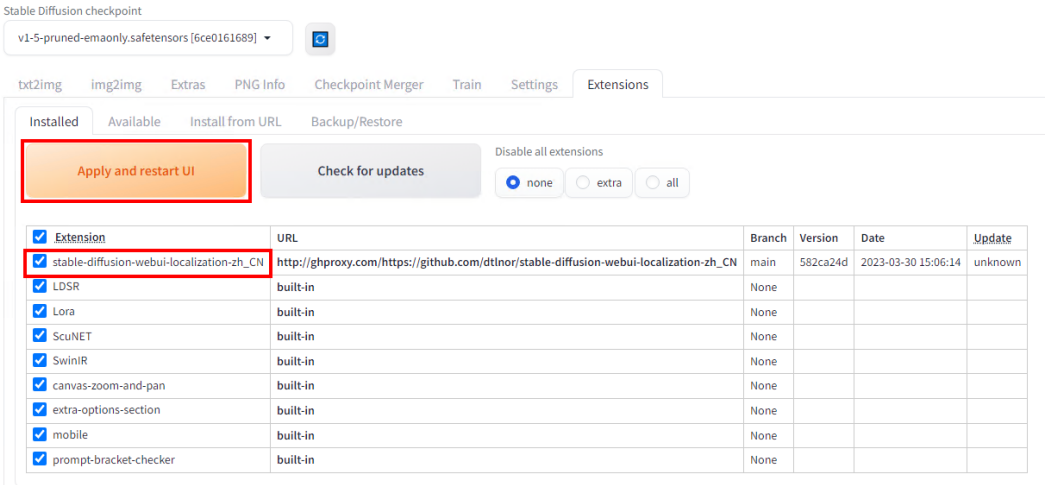
图 3-26 安装汉化插件



步骤7 （可选）界面汉化插件配置，在 Settings 选项卡，单击 页面右上角的 橙色 Reload UI 按钮 刷新扩展列表，在 Extensions 选项卡，确定已勾选汉化扩展插件☑，如未勾选，勾选后单击橙色按钮启用汉化扩展。

图 3-27 重载 UI 界面

图 3-28 启用汉化扩展



步骤8 （可选）界面汉化插件使用，在 Settings 选项卡中，选择 User interface 子选项 ，选择 Localization (requires restart UI) ，在下拉框选中 zh_CN （如果没有单击□按钮），依次单击 Apply settings 按钮 保存设置，Reload UI 按钮 重启webUI。即可完成汉化。

图 3-29 使用汉化拓展

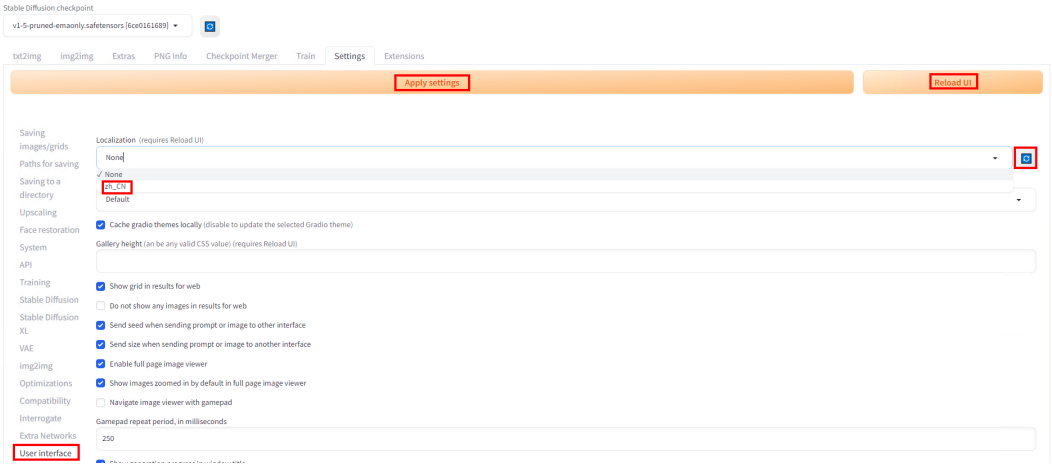
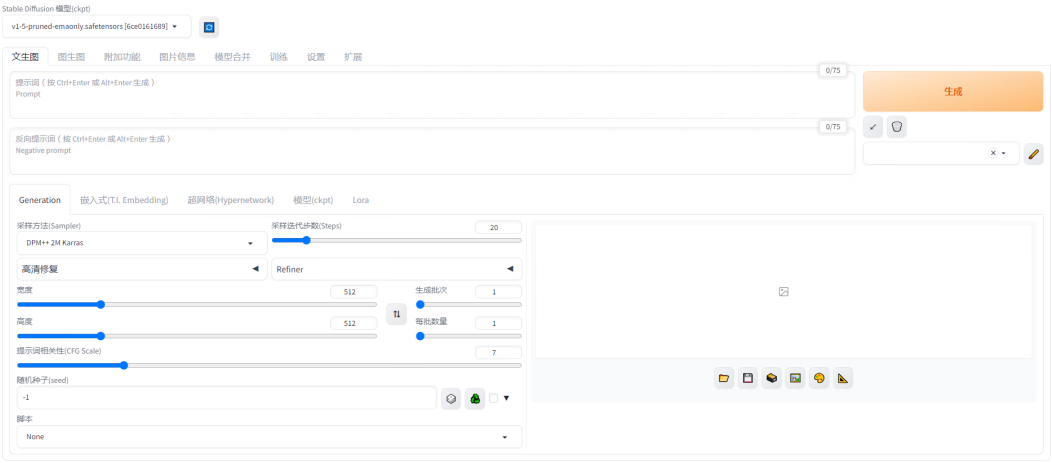


图 3-30 汉化完成



说明

本方案配置使用inotify-tools和对象存储服务 OBS obsutil工具，并设置了开机自启动，实现将您在web UI界面保存的图片自动上传到对象存储服务OBS桶。您也可以在浏览器单击鼠标右键选择另存为保存图片。服务开机自启动，重启机器后无需任何操作即可使用此AIGC应用。

服务启动命令示例：

前台启动

```
cd /home/aigc && sudo -u aigc bash -c "source /home/aigc/webui.sh --listen --port 7860 --api --enable-insecure-extension-access"
```

后台启动

```
cd /home/aigc && sudo -u aigc bash -c "source /home/aigc/webui.sh --listen --port 7860 --api --enable-insecure-extension-access &" >> /home/aigc/aigc-applications.log （日志保存路径可根据需要自行修改）
```

----结束

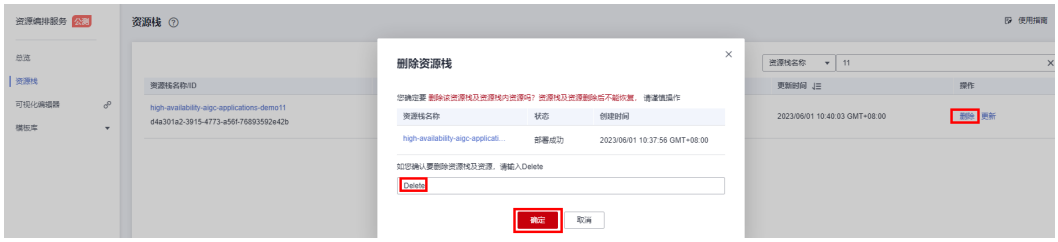
3.4 快速卸载

须知

该解决方案涉及到对象存储服务 OBS桶，如果OBS桶中有数据的话会导致资源栈删除失败。请确保数据以及迁移备份后清空OBS桶中的数据，再卸载该解决方案。

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 3-31 一键卸载



----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。
- 弹性负载均衡 ELB：将访问流量自动分发到多台云服务器，扩展应用系统对外的服务能力，实现更高水平的应用容错。
- 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。
- 对象存储服务 OBS：对象存储服务（Object Storage Service，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。
- 安全组：安全组是一个逻辑上的分组，为同一个VPC内具有相同安全保护需求并相互信任的弹性云服务器提供访问策略。安全组创建后，用户可以在安全组中定义各种访问规则，当弹性云服务器加入该安全组后，即受到这些访问规则的保护。
- inotify-tools：inotify-tools是一个Linux下的命令行工具，用于监控文件系统的变化并触发相应的操作。

5 修订记录

发布日期	修订记录
2023-5-30	第一次正式发布。
2023-9-15	修订开始使用，WebUI界面增加汉化步骤。