

解决方案实践

OBS 极速同步模型权重文件

文档版本 1.0
发布日期 2025-05-14



版权所有 © 华为技术有限公司 2025。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.....	1
2 资源和成本规划.....	3
3 实施步骤.....	6
3.1 准备工作.....	6
3.2 快速部署.....	9
3.3 开始使用.....	14
3.4 快速卸载.....	18
4 附录.....	20
5 修订记录.....	21

1 方案概述

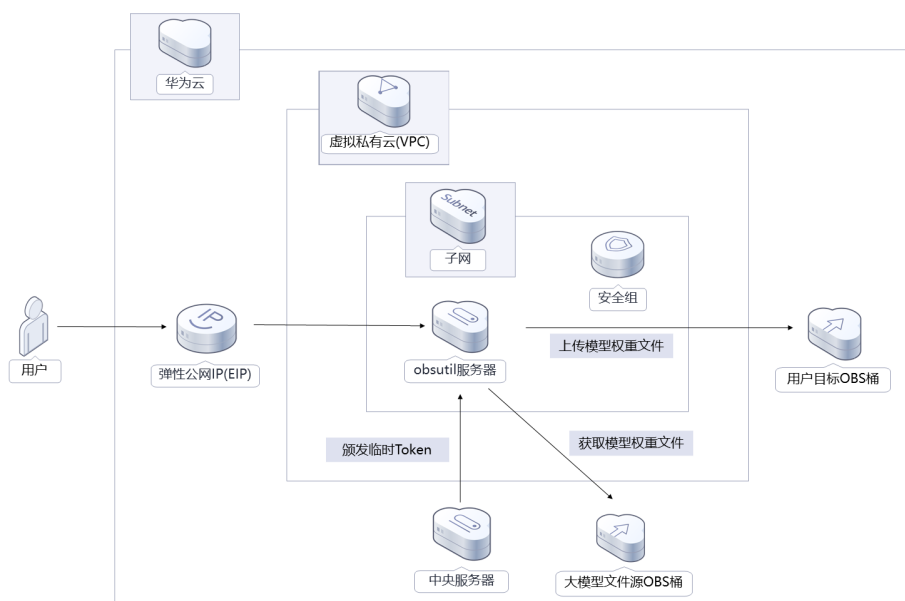
应用场景

大语言模型文件一般是GB到TB级别，从IDC或其它数据中心上传至**对象存储服务 OBS**桶花费时间长，操作繁琐，中转机器磁盘容量要求高。该解决方案通过使用一台轻量级的**弹性云服务器 ECS**作为中转，利用内网带宽将模型权重文件从源端桶流式复制到客户同区域下的目标桶，加快模型权重文件的传输速度，减轻将模型权重文件下载到对象存储时间慢的痛点。

方案架构

云服务器作为中转服务器，将源端桶模型权重文件流式复制到用户同区域下的目标桶，加快模型权重文件的传输速度。

图 1-1 方案架构图



该解决方案将会部署如下资源：

- 创建一个[弹性公网IP EIP](#)，用于ECS访问由华为云解决方案团队维护的中央服务器。
- 创建一台Linux[弹性云服务器 ECS](#)实例作为中转机，并绑定公网IP，安装obsutil工具，用于OBS桶之间的数据复制，并且流量穿透过该机器。
- 使用[对象存储服务OBS](#)桶，用于存储大模型权重文件。

方案优势

- 内网加速
借助华为云OBS内网加速功能，能够显著提升数据传输效率，确保数据快速、稳定地传输。
- 模型文件托管
源端的OBS桶中提供常见的模型权重文件，并由华为云解决方案团队进行维护和管理。
- 低成本
该方案的资源消耗非常低，只需一台轻量级的弹性云服务器 ECS实例就能完全满足需求。

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，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	资源名称	配置示例	数量	每月预估花费
虚拟私有云 VPC	replicating-model-files-across-obs-buckets-demo	- VPC网段：172.16.0.0/16 - 区域：华北-北京四	1	0.00元
子网 Subnet	replicating-model-files-across-obs-buckets-demo-subnet	- 子网网段：172.16.1.0/24 - 区域：华北-北京四	1	0.00元
安全组 SecurityGroup	replicating-model-files-across-obs-buckets-demo	- 允许ping：0.0.0.0/0 - 开放端口22允许Cloud Shell 登录：121.36.59.153/32 - 区域：华北-北京四	1	0.00元
对象存储服务OBS	自定义	- 桶策略：私有 - 数据冗余存储策略：多AZ存储 - 区域：华北-北京四	1	0.00元 + 实际使用费用

华为云服务	资源名称	配置示例	数量	每月预估花费
Flexus云服务器X实例	replicating-model-files-across-obs-buckets-demo	• 按需计费：0.60/小时 • 区域：华北-北京四 • 规格：Flexus云服务器X实例 性能模式（开启） x1e.4u.4g 4核 4GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：通用型 SSD 40GB	1	432.58元
弹性公网IP EIP	replicating-model-files-across-obs-buckets-demo-eip	• 按需计费：0.80元/GB • 区域：华北-北京四 • 线路：动态BGP • 公网带宽：按流量计费 • 带宽大小：300Mbit/s	1	0.80元/GB
合计	-	-		432.58元 + 弹性公网IP EIP费用 + OBS 实际使用费用

表 2-2 资源和成本规划（包年包月）

华为云服务	资源名称	配置示例	数量	每月预估花费
虚拟私有云 VPC	replicating-model-files-across-obs-buckets-demo	• VPC网段：172.16.0.0/16 • 区域：华北-北京四	1	0.00元
子网 Subnet	replicating-model-files-across-obs-buckets-demo-subnet	• 子网网段：172.16.1.0/24 • 区域：华北-北京四	1	0.00元

华为云服务	资源名称	配置示例	数量	每月预估花费
安全组 SecurityGroup	replicating-model-files-across-obs-buckets-demo	• 允许ping：0.0.0.0/0 • 开放端口22允许Cloud Shell 登录：121.36.59.153/32 • 区域：华北-北京四	1	0.00元
对象存储服务OBS	自定义	• 桶策略：私有 • 数据冗余存储策略：多AZ存储 • 区域：华北-北京四	1	0.00元 + 实际使用费用
Flexus云服务器X实例	replicating-model-files-across-obs-buckets-demo	• 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（开启） x1e.4u.4g 4核 4GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：通用型 SSD 40GB	1	298.00元
弹性公网IP EIP	replicating-model-files-across-obs-buckets-demo-eip	• 按需计费：0.80元/GB • 区域：华北-北京四 • 线路：动态BGP • 公网带宽：按流量计费 • 带宽大小：300Mbit/s	1	0.80元/GB
合计	-	-		298.00元 + 弹性公网IP EIP费用+ OBS 实际使用费用

3 实施步骤

[3.1 准备工作](#)

[3.2 快速部署](#)

[3.3 开始使用](#)

[3.4 快速卸载](#)

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

委托 / 创建委托

★ 委托名称

rf_admin_trust

★ 委托类型

☐ 普通帐号

将帐号内资源的操作权限委托给其他华为云帐号。

☒ 云服务

将帐号内资源的操作权限委托给华为云服务。

★ 云服务

RFS

★ 持续时间

永久

描述

请输入委托信息。

0/255

下一步

取消

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

1 选择策略

2 设置委托授权范围

3 完成

委托“rf_admin_trust”将拥有指定策略

策略已选 11

从其他区域策略库复制

全部策略

所有云服务

Tenant Administrator

名称

类型

系统角色

☒

Tenant Administrator

全部云服务管理权限（含IAM管理权限）

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

1 选择策略

2 设置委托授权范围

3 完成

根据当前您选择的策略，系统推荐以下授权范围方案，更便于您最小化授权，可进行编辑。了解如何根据您的应用场景选择合适授权范围方案

选择授权范围方案

☒ 所有资源

授权后，IAM用户可以根据权限使用帐号中所有资源，包括企业项目、区域项目和全局服务资源。

☐ 展开其他方案

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

委托 0						新建委托
删除 您还可以创建49个委托。						全部类型 请输入委托名称进行搜索
☐	委托名称ID 注	委托对象 注	委托时长 注	创建时间 注	描述 注	操作
☐	rt_admin_host	云服务 RFS	永久	2023/05/31 11:07:56 GMT+08:00	--	授权 修改 删除

----结束

3.2 快速部署

操作场景

本章节帮助用户高效地部署“OBS极速同步模型权重文件”解决方案。一键部署该解决方案时，参照本章节中的步骤和说明进行操作，即可完成快速部署。

操作步骤

步骤1 登录[华为云解决方案实践](#)，选择“OBS极速同步模型权重文件”，支持区域下拉选择部署的区域（以北京四为例），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

方案架构

云服务器作为中转服务器，将源端桶模型权重文件流式复制到用户同区域下的目标桶，加快模型权重文件的传输速度。

The diagram illustrates the solution architecture. A user interacts with a solution template. The architecture involves a central server (obsutil服务器) that acts as a中转服务器 (intermediate server). It receives data from a源端桶 (source bucket) and streams it to a用户目标OBS桶 (user target OBS bucket). The process involves uploading model weight files (上传模型权重文件) and retrieving them (获取模型权重文件). The architecture also includes a弹性公网IP (EIP), 虚拟私有云 (VPC), 子网 (subnet), and 安全组 (security group). A 中央服务器 (central server) and 大模型文件源OBS桶 (large model file source OBS bucket) are also shown. The process involves obtaining a临时Token (temporary token) and using it to upload files.

OBS极速同步模型权重文件

版本：1.0.0
上次更新日期：2025年4月
来源：由华为云构建
部署：预计5分钟
卸载：预计5分钟

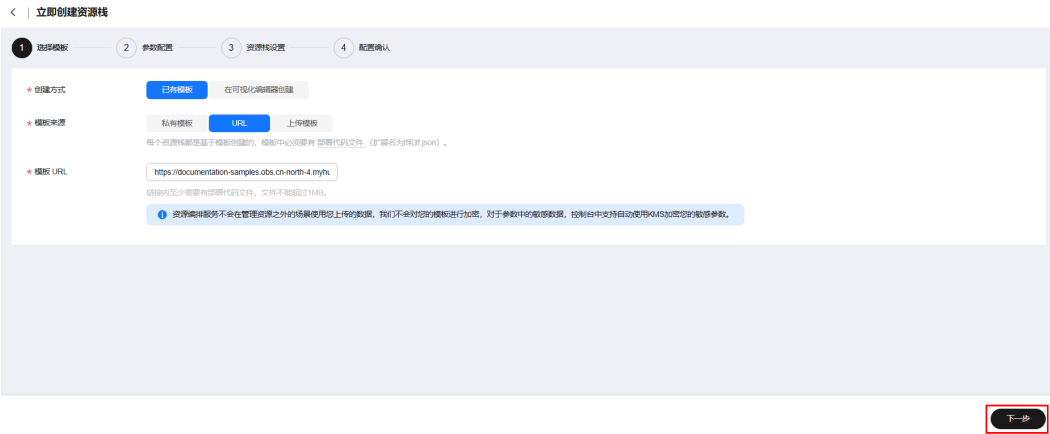
预估成本

支持区域：
西南-贵阳一
西南-贵阳一
查看云 华北-北京四

一键部署

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，部分参数会自动默认填充参数值。如需修改请在参数配置页面删除文本框内的默认值后填写新的参数值，所有参数填写完成后方可单击“下一步”。

图 3-10 配置参数

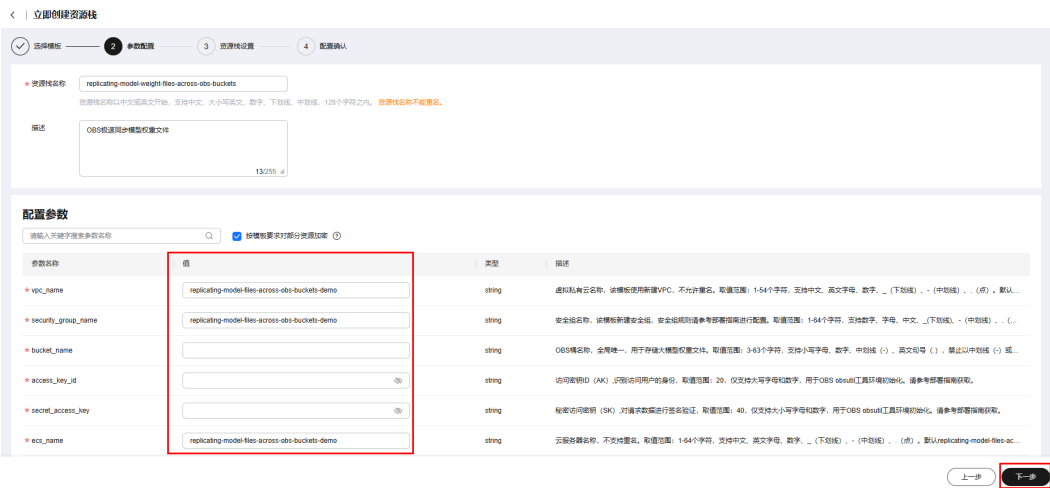


表 3-1 参数填写说明

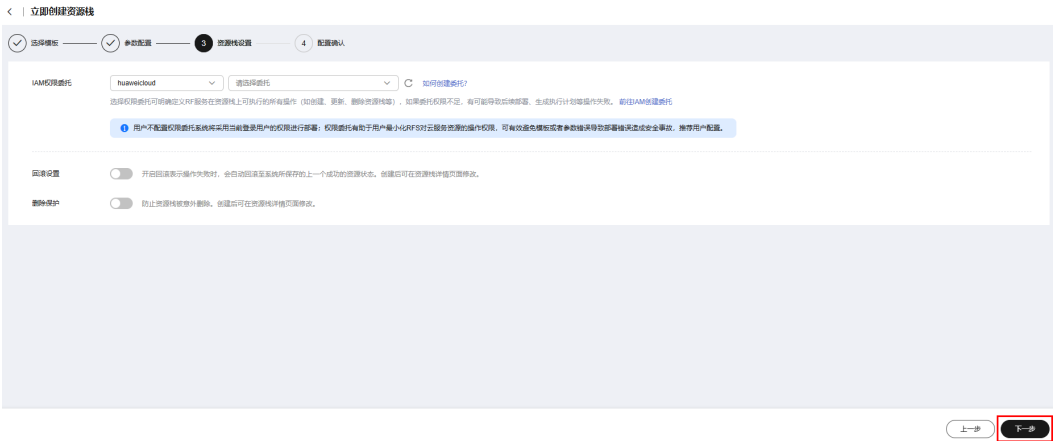
参数名称	类型	是否可选	参数解释	默认值
vpc_name	string	必填	虚拟私有云名称，该模板使用新建VPC，不允许重名。取值范围：1-54个字符，支持中文、英文字母、数字、_（下划线）、-（中划线）、.（点）。	replicating-model-files-across-obs-buckets-demo
security_group_name	string	必填	安全组名称，该模板新建安全组，请参考 安全组规则修改 进行配置。取值范围：1-64个字符，支持字母、数字、中文、下划线（_）、中划线（-）、英文句号（.）。	replicating-model-files-across-obs-buckets-demo

参数名称	类型	是否可选	参数解释	默认值
bucket_name	string	必填	OBS桶名称，全局唯一，用于存储大模型权重文件。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（全局唯一，需要自定义）
access_key_id	string	必填	访问密钥ID（AK），识别访问用户的身份，取值范围：20，仅支持大写字母和数字，用于OBS obsutil工具环境初始化。详见 如何获取访问密钥AK/SK 。	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，取值范围：40，仅支持大小写字母和数字，用于OBS obsutil工具环境初始化。详见 如何获取访问密钥AK/SK 。	空
ecs_name	string	必填	云服务器名称，不支持重名。取值范围：1-64个字符，支持中文、英文字母、数字、_（下划线）、-（中划线）、.（点）。	replicating-model-files-across-obs-buckets-demo
ecs_flavor	string	必填	华为云Flexus 云服务器X实例规格。华为云Flexus 云服务器X实例规格。可选值：x1e.2u.2g（传输速率250MB/s）、x1e.4u.4g（传输速率350MB/s）、x1e.8u.8g（传输速率750MB/s）、x1e.12u.12g（传输速率950MB/s）、x1e.16u.16g（传输速率1.45GB/s）。Flexus 云服务器X实例规格ID命名规则为x1e.?u.?g，例如2vCPUs4GiB规格ID为x1.4u.4g，具体华为云Flexus 云服务器X实例规格请参考控制台。	x1e.4u.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。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
charging_mode	string	必填	云服务器计费模式，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	云服务器订购周期类型，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
charging_period	number	必填	云服务器订购周期，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

步骤4 （可选，如果使用华为主账号或admin用户组下的IAM子账户可不选委托）在资源设置界面中，在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 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[资源和成本规划](#)）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”，堆栈部署成功，表示顺利完成资源的下发和部署。堆栈部署成功后，耐心等待5分钟左右（受网络波动影响）。

图 3-16 部署完成



----结束

3.3 开始使用

说明

该解决方案使用22端口CloudShell方式远程登录云服务器，默认已配置IP地址白名单，若需远程登录云服务器，可直接使用CloudShell远程登录。

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，部署完成方可正常使用。

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。

使用步骤

步骤1 登录[Flexus云服务器X实例控制台](#)，选择部署本方案创建的服务器，单击远程登录。

图 3-17 选择远程登录

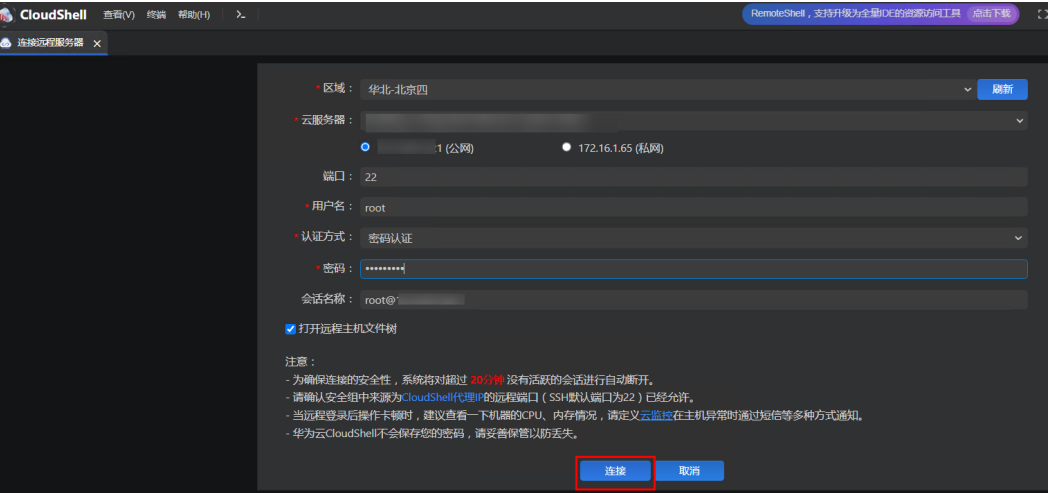


步骤2 选择使用CloudShell登录，在连接配置页面输入云服务器的密码，单击“连接”（部分区域不支持CloudShell，请使用VNC登录）。

图 3-18 选择 CloudShell 登录



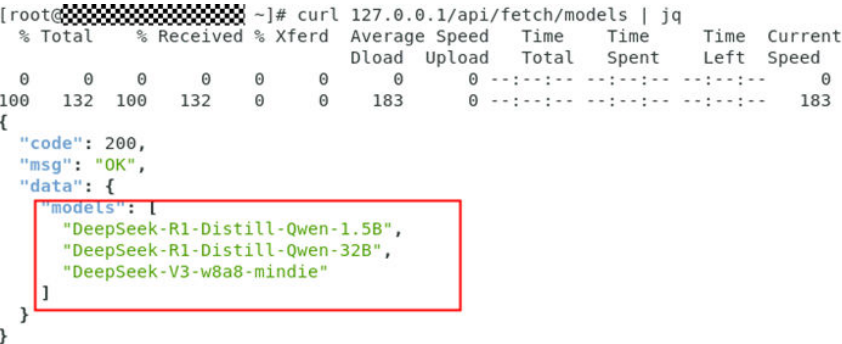
图 3-19 配置连接信息



步骤3 在部署成功的登录Flexus云服务器X实例上执行以下命令查询有哪些模型文件可供分发。

```
curl http://127.0.0.1/api/fetch/models | jq
```

图 3-20 查询可用模型文件



步骤4 执行命令创建模型文件同步任务。

```
curl -XPOST http://127.0.0.1/api/task?modelName=${模型名称}
```

图 3-21 创建任务



步骤5 查看任务状态。

```
curl http://127.0.0.1/api/task | jq
```

图 3-22 任务状态

```
[root@xxxxxxxxxxxx ~]# curl http://127.0.0.1/api/task | jq
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           Dload  Upload   Total     Spent    Left     Speed
100    259    100    259      0      0    433k      0 --:--:-- --:--:-- --:--:--   252k
{
  "code": 200,
  "msg": "OK",
  "data": [
    {
      "taskName": "DeepSeek-R1-Distill-Qwen-32B",
      "status": "RUNNING",
      "progress": "20.70%",
      "sizeProgress": "12.63GB/61.03GB",
      "speed": "1.43GB/s",
      "taskExited": false,
      "destEndpoint": "obs.cn-north-4.myhuaweicloud.com",
      "destBucket": "xxxxxxxxxxxx"
    }
  ]
}
```

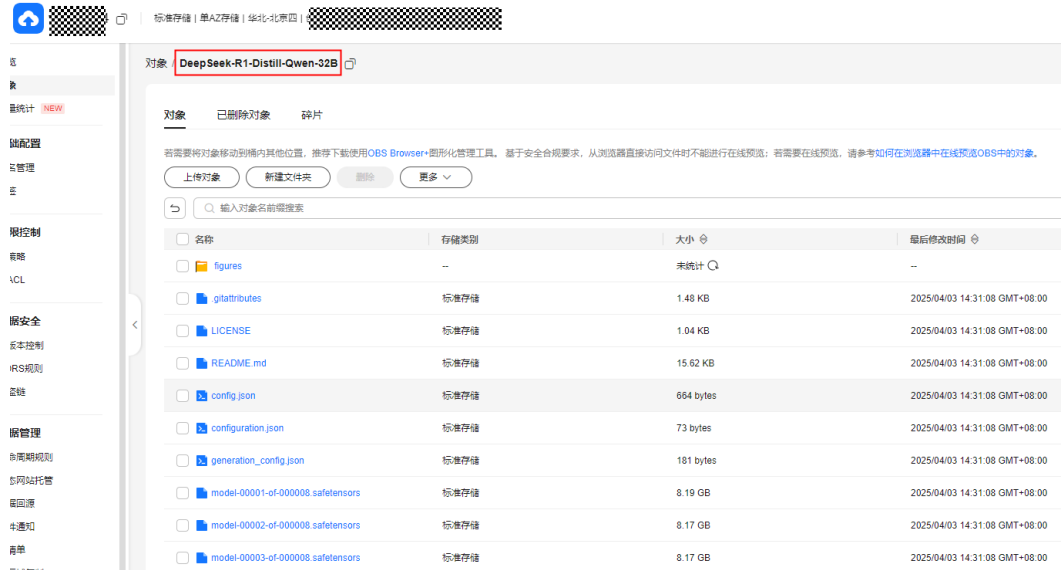
步骤6 待任务状态为“FINISHED”即代表模型权重文件同步成功。

图 3-23 任务完成

```
[root@xxxxxxxxxxxx ~]# curl http://127.0.0.1/api/task | jq
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           Dload  Upload   Total     Spent    Left     Speed
100    241    100    241      0      0    653k      0 --:--:-- --:--:-- --:--:--   235k
{
  "code": 200,
  "msg": "OK",
  "data": [
    {
      "taskName": "DeepSeek-R1-Distill-Qwen-32B",
      "status": "FINISHED",
      "progress": "100.00%",
      "sizeProgress": "61.03GB/61.03GB",
      "taskExited": true,
      "destEndpoint": "obs.cn-north-4.myhuaweicloud.com",
      "destBucket": "xxxxxxxxxxxx"
    }
  ]
}
```

步骤7 登录OBS控制台，查看该解决方案创建的OBS桶，可以看到模型权重文件已同步成功。

图 3-24 同步成功



----结束

3.4 快速卸载

步骤1 登录[OBS控制台](#)，查看该解决方案创建的OBS桶，删除桶中所有对象。

图 3-25 删除所有对象



步骤2 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 3-26 一键卸载



步骤3 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 3-27 删除资源栈确认



----结束

4 附录

名词解释

- 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
2025-04-03	第一次正式发布。