LocalAI平台使用指南

1 商品说明

LocalAI是 OpenAI 的免费开源替代方案。LocalAI 可作为 REST API 的替代，兼容 OpenAI（Elevenlabs、Anthropic 等）API 规范，用于本地 AI 推理。它允许您在本地或使用消费级硬件运行 LLM、生成图像、音频（以及其他内容），并支持多种模型系列。LocalAI 无需 GPU 计算。

本商品在鲲鹏云的AArch64架构上Ubuntu24.04和HCE2.0系统中进行安装后以镜像提供给用户使用。

2 商品购买

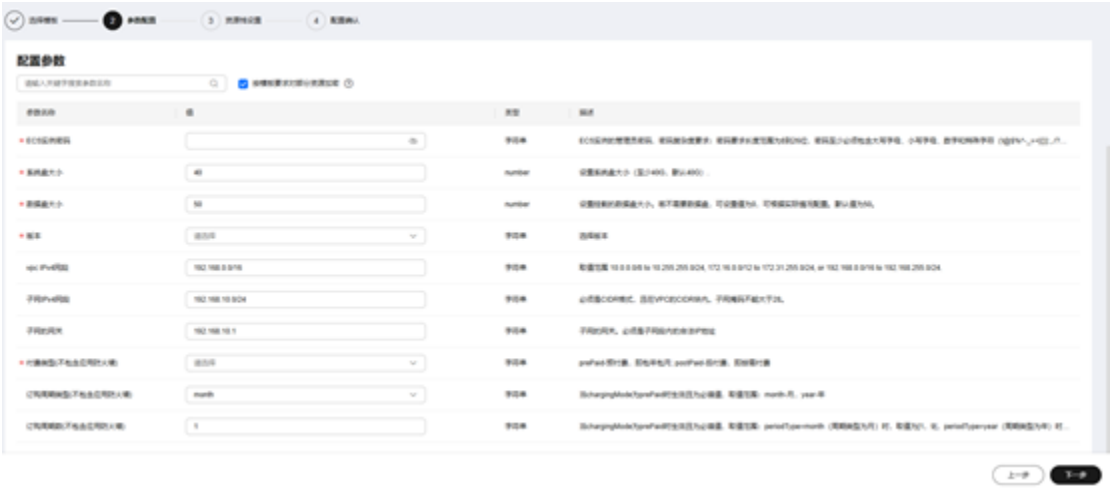
您可以在云商店搜索“LocalAI平台”。

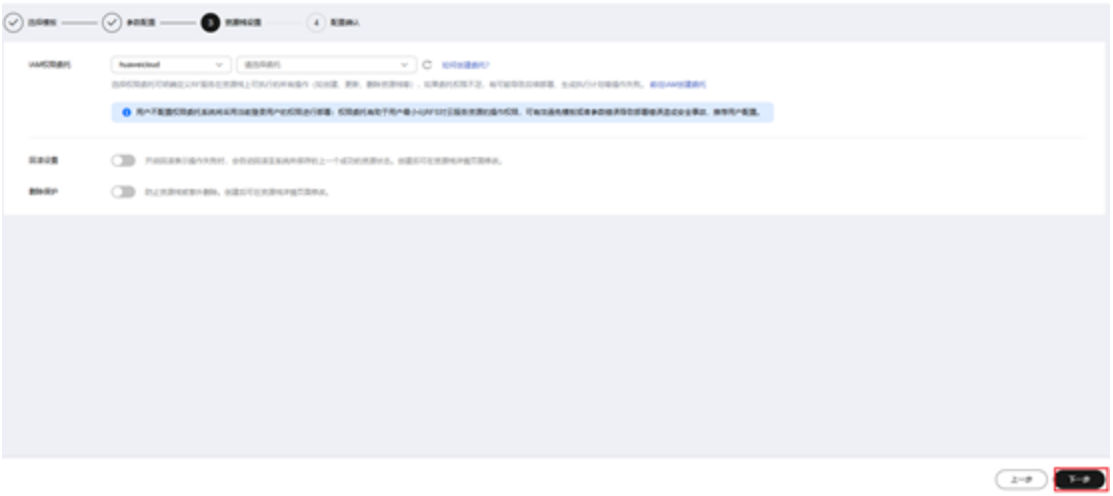
其中，地域、规格、推荐配置使用默认，购买方式根据您的需求选择按需/按月/按年，短期使用推荐按需，长期使用推荐按月/按年，确认配置后点击“立即购买”。

2.1 商品支持自定义 ECS 购买，具体见章节 3.1.1

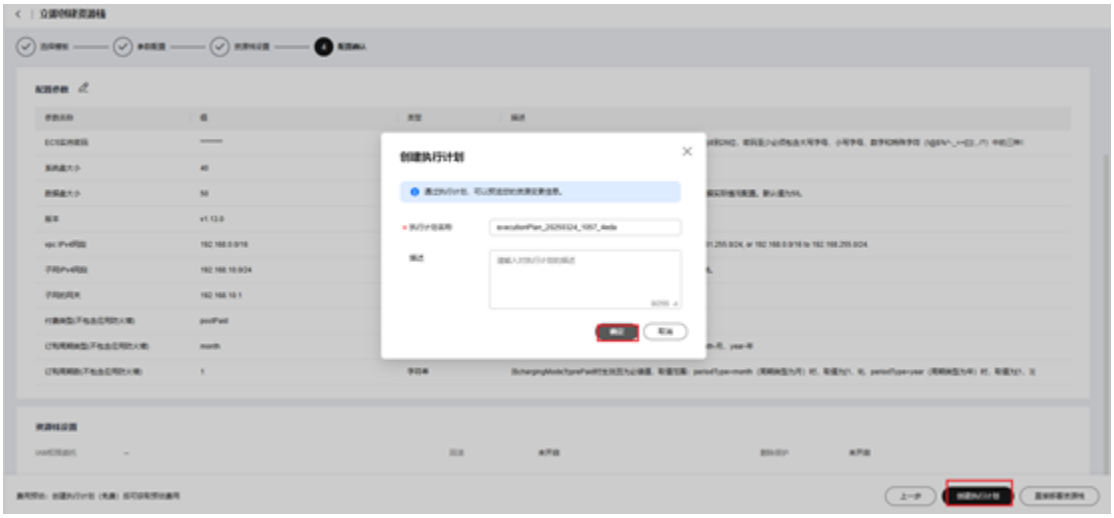
2.2 使用 RFS 模板直接部署

必填项填写后，点击 下一步



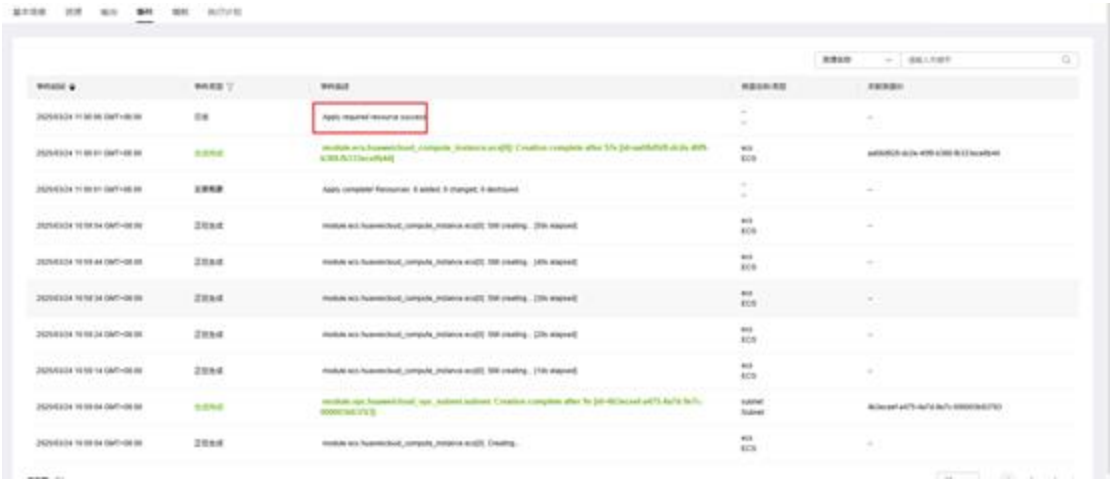


创建直接计划后，点击 确定



点击 部署

如下图“Apply required resource success. ”即为资源创建完成



3 商品资源配置

商品支持**ECS控制台配置**，下面对资源配置的方式进行介绍。

3.1 ECS 控制台配置

3.1.1 准备工作

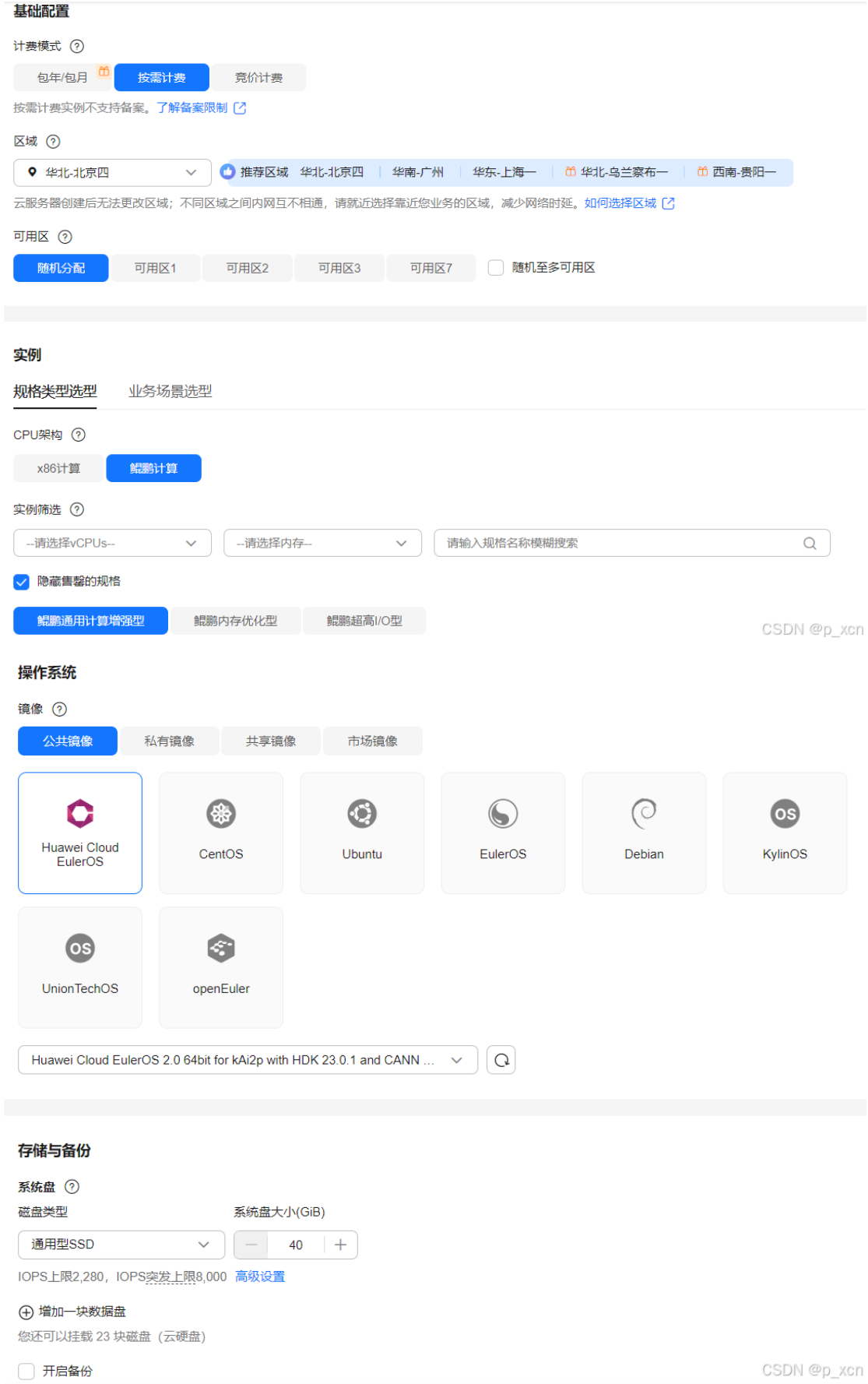
在使用ECS控制台配置前，需要您提前配置好**安全组规则**。

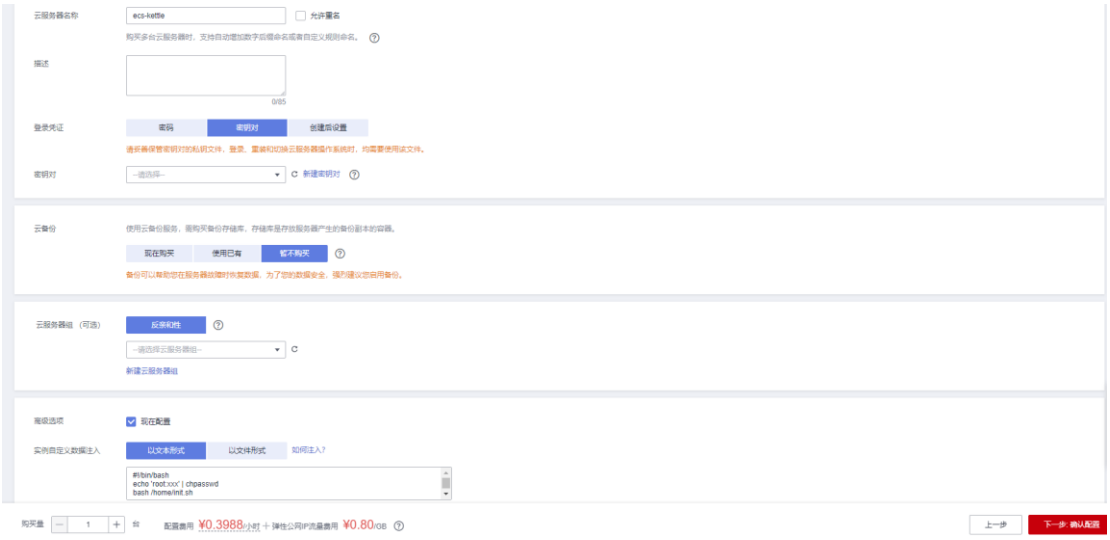
安全组规则的配置如下：

- 入方向规则放通端口8080，源地址内必须包含您的客户端ip，否则无法访问
- 入方向规则放通CloudShell连接实例使用的端口22，以便在控制台登录调试。
- 出方向规则一键放通

3.1.2 创建 ECS

前提工作准备好后，选择ECS控制台配置跳转到购买ECS页面，ECS资源的配置如下图所示：


基础配置
计费模式
包年/包月
按需计费
竞价计费
按需计费实例不支持备案。了解备案限制
区域
华北-北京四
推荐区域 华北-北京四 华南-广州 华东-上海一 华北-乌兰察布一 西南-贵阳一
云服务器创建后无法更改区域；不同区域之间内网互不相通，请就近选择靠近您业务的区域，减少网络时延。如何选择区域
可用区
随机分配
可用区1
可用区2
可用区3
可用区7
随机至多可用区
实例
规格类型选型
业务场景选型
CPU架构
x86计算
鲲鹏计算
实例筛选
--请选择vCPUs--
--请选择内存--
请输入规格名称模糊搜索
隐藏售罄的规格
鲲鹏通用计算增强型
鲲鹏内存优化型
鲲鹏超高I/O型
CSDN @p_xcn
操作系统
镜像
公共镜像
私有镜像
共享镜像
市场镜像
Huawei Cloud EulerOS
CentOS
Ubuntu
EulerOS
Debian
KylinOS
UnionTechOS
openEuler
Huawei Cloud EulerOS 2.0 64bit for kAi2p with HDK 23.0.1 and CANN ...
存储与备份
系统盘
磁盘类型
系统盘大小(GiB)
通用型SSD
40
IOPS上限2,280，IOPS突发上限8,000 高级设置
增加一块数据盘
您还可以挂载 23 块磁盘（云硬盘）
开启备份
CSDN @p_xcn



值得注意的是：

- VPC您可以自行创建
- 安全组选择3.1.1章节中配置的安全组
- 弹性公网IP选择现在购买，推荐选择“按流量计费”，带宽大小可设置为5Mbit/s
- 高级配置需要在高级选项支持注入自定义数据，所以登录凭证不能选择“密码”，选择创建后设置
- 其余默认或按规则填写即可。

4 商品使用

4.1 登录服务器查看 Dify 进程

4.1.1 进入系统后通过命令

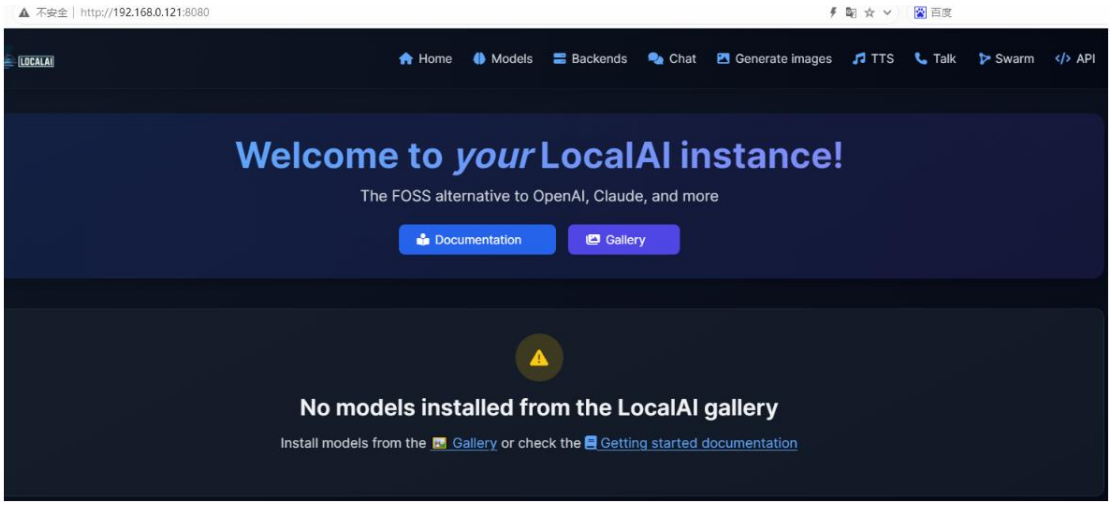
cd ~/localai

docker run -p 8080:8080 -v $PWD/models:/models -ti --rm localai/localai:v3.0.0 --models-path /models

```
root@ecs-kaiyuan-public-0003:~/localai# docker run -p 8080:8080 -v $PWD/models:/models -ti --rm localai/localai:v3.0.0 --models-path /models
@@@@@
Skipping rebuild
@@@@@
If you are experiencing issues with the pre-compiled builds, try setting REBUILD=true
If you are still experiencing issues with the build, try setting CMAKE_ARGS and disable the instructions set as needed:
CMAKE_ARGS="-DGGML_F16C=OFF -DGGML_AVX512=OFF -DGGML_AVX2=OFF -DGGML_FMA=OFF"
see the documentation at: https://localai.io/basics/build/index.html
Note: See also https://github.com/go-skynet/LocalAI/issues/288
@@@@@
CPU info:
CPU: no AVX    found
CPU: no AVX2   found
CPU: no AVX512 found
@@@@@
6:53AM INF env file found, loading environment variables from file envFile=.env
6:53AM INF Setting logging to info
6:53AM INF Starting LocalAI using 2 threads, with models path: /models
6:53AM INF LocalAI version: v3.0.0 (f9b968e19d7cbc556d59dceb2e0e450b828a3fda)
6:53AM INF Preloading models from /models
6:53AM INF core/startup process completed!
6:53AM INF LocalAI API is listening! Please connect to the endpoint for API documentation. endpoint=http://0.0.0.0:8080
6:53AM INF Success ip=192.168.0.250 latency=3.598135ms method=GET status=200 url=/
```

4.1.2 通过浏览器登录 LocalAI 平台

http://yourIp:8080/



4.2 参考文档

完整操作参考 LocalAI手册