vLLM大模型服务使用指南

1 商品说明

vLLM 是一个快速且易于使用的 LLM 推理和服务库。

vLLM 最初由加州大学伯克利分校的天空计算实验室开发，现已发展成为一个由学术界和工业界共同贡献的社区驱动项目。

目前官方没有提供针对鲲鹏云的AArch64架构的安装包，需要手工进行安装。

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

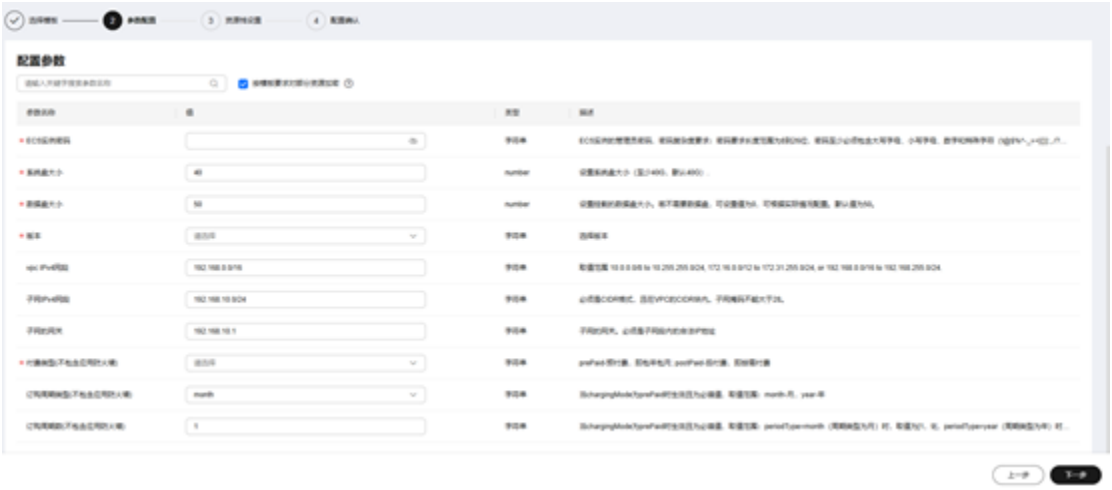
您可以在云商店搜索“vLLM大模型服务”。

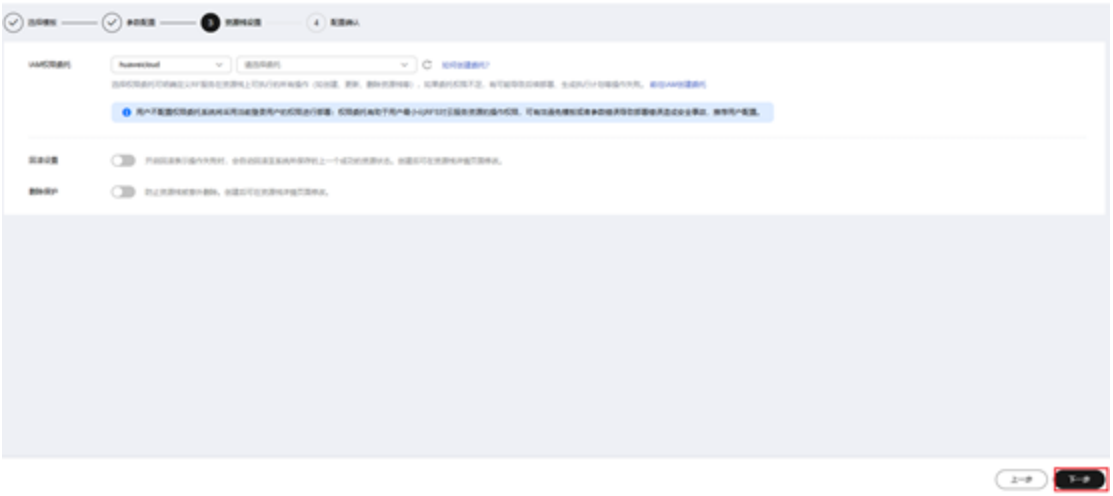
其中，地域、规格、推荐配置使用默认，购买方式根据您的需求选择按需/按月/按年，短期使用推荐按需，长期使用推荐按月/按年，确认配置后点击“立即购买”。

2.1 商品支持自定义 ECS 购买，具体见章节 3.1.1

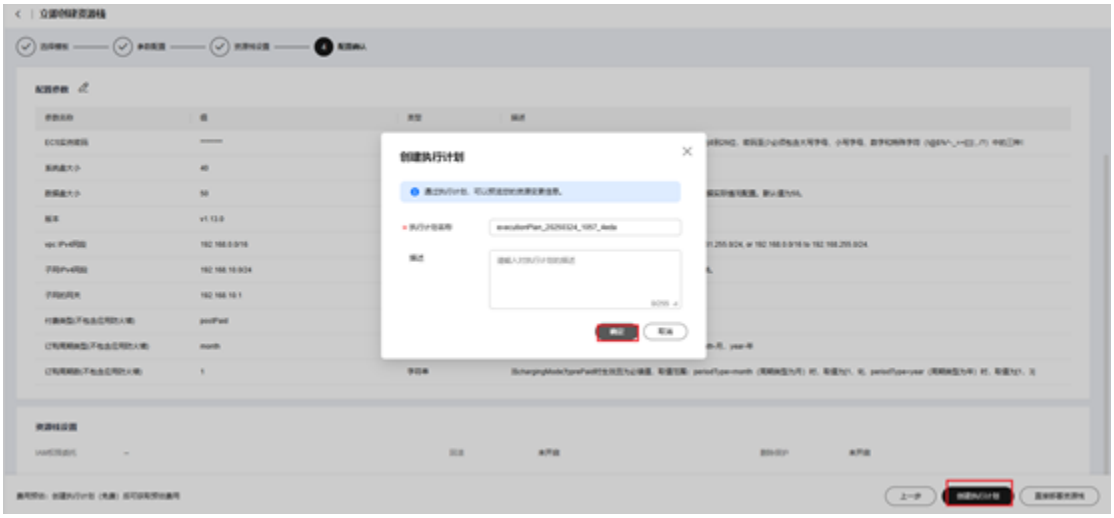
2.2 使用 RFS 模板直接部署

必填项填写后，点击 下一步



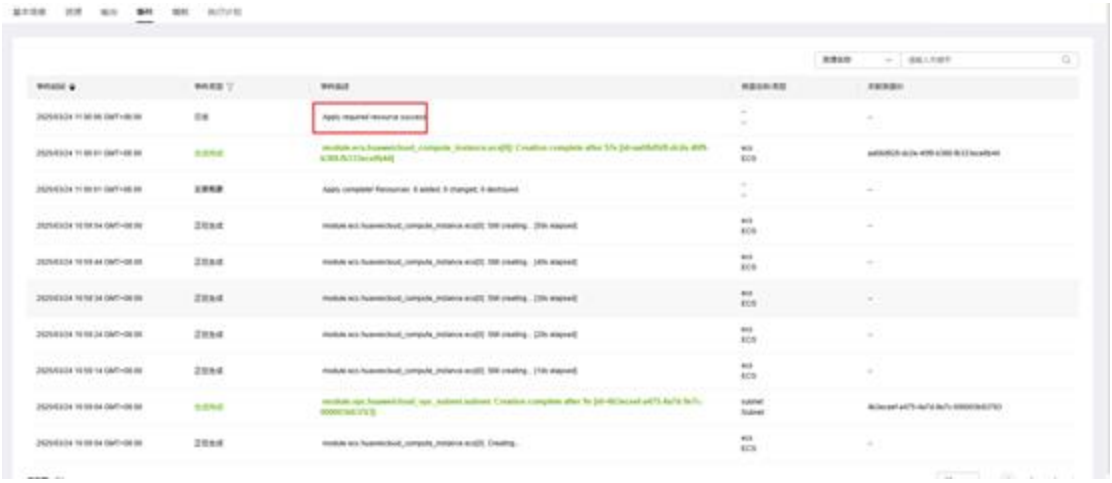


创建直接计划后，点击 确定



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8000，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

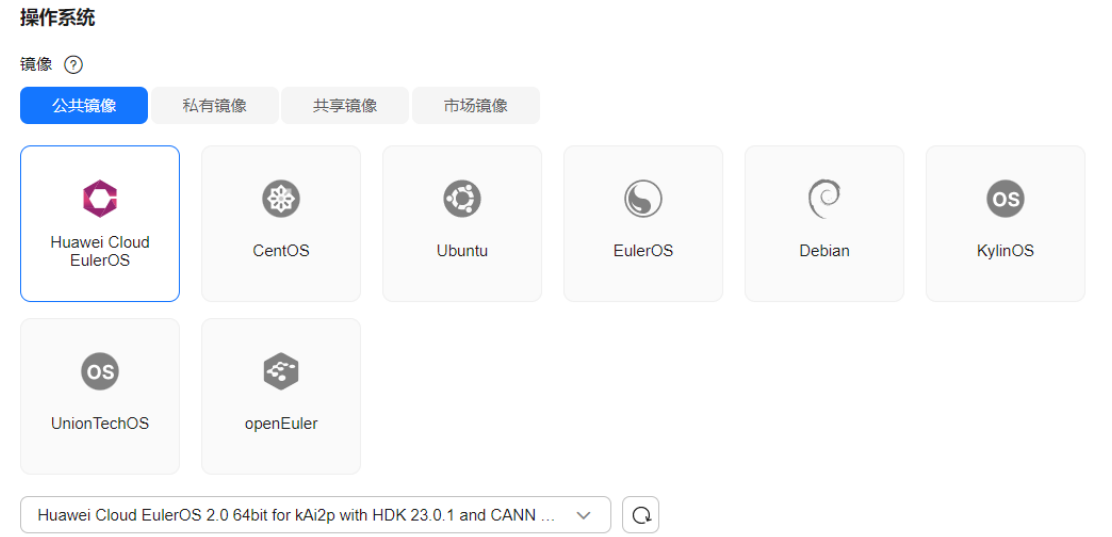
3.1.2 创建 ECS

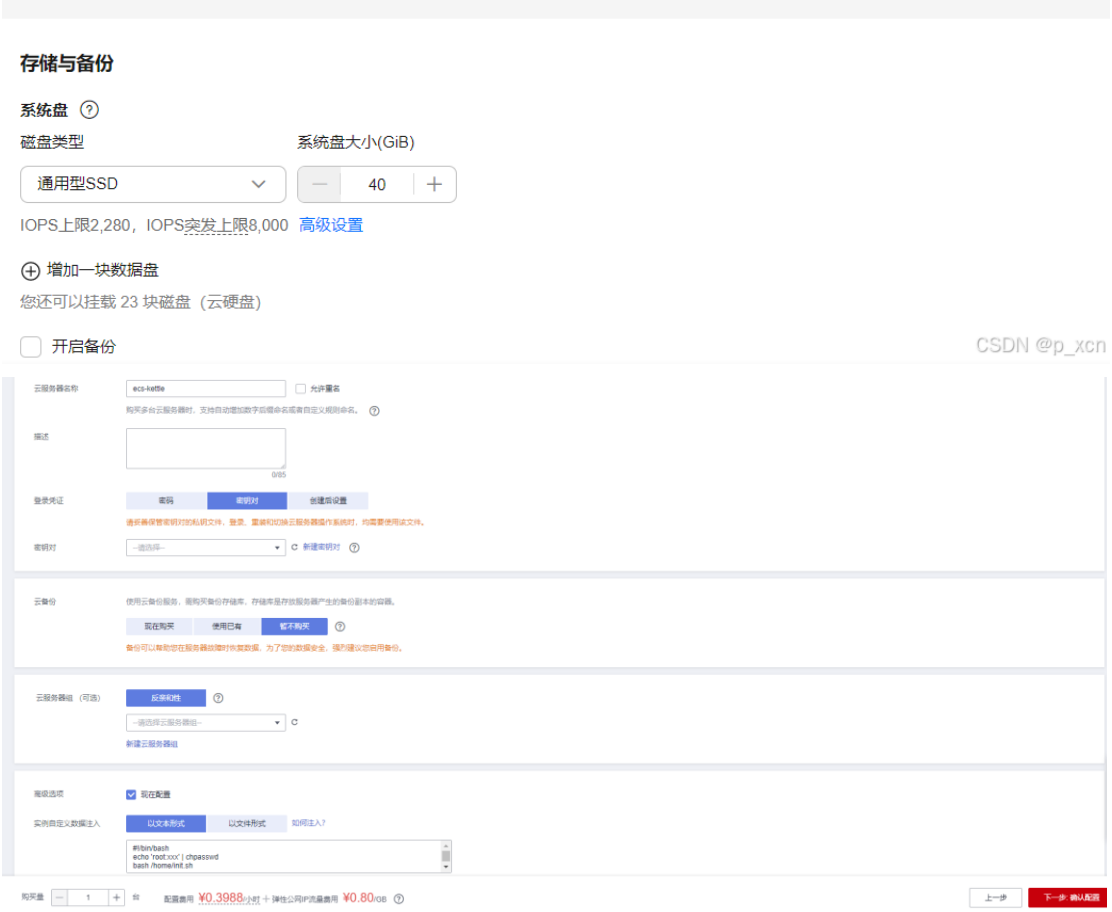
注意：**要根据想部署的模型规格和精度选择ECS的规格**，例如：

- 1.5B模型Float16 至少 4G内存

- 7/8B模型Float16 建议24G内存

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：





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

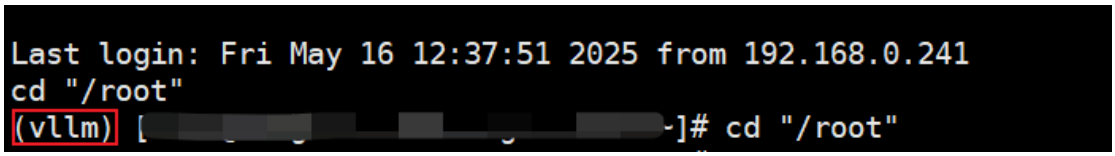
择创建后设置

- 其余默认或按规则填写即可。

4 商品使用

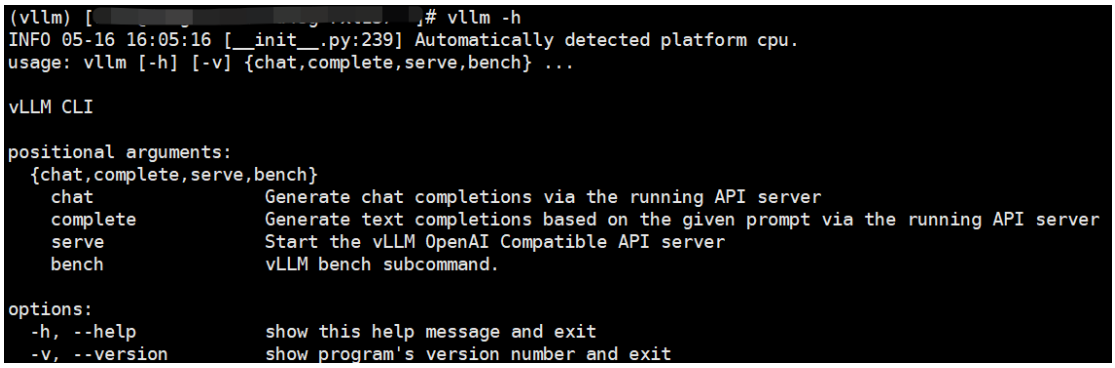
4.1 vLLM 使用

进入系统后默认会切换到Python的虚拟环境vllm。



4.1.1 查看 vLLM 帮助

运行命令：vllm -h



4.1.2 下载模型运行服务

- 使用魔塔社区的模型，运行 set_modelscope.sh 进行配置。
- 默认使用Hugging Face 模型

下面以deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B 为例 启动服务：

vllm serve deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

启动成功后访问接口查看模型信息：

http://your_ip:8000/v1/models

4.2 参考文档

vLLM使用手册