

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

参数名称	值	类型	描述
EC2实例规格	ib	字符串	EC2实例的规格名称。规格名称要求：规格名称长度范围为8到12位，规格至少由1个大写字母、1个小写字母、数字和特殊字符（如\$%^&*+=~_!@#~）组成。
实例大小	40	number	设置实例大小（范围400，默认400）。
实例大小	50	number	设置实例的实例大小，如果不填则默认，可以设置为0，可设置实例与实例，默认为0。
版本	选择	字符串	选择版本
vpc ID	102-100-0-0-10	字符串	指定VPC ID。VPC ID格式为：102-100-0-0-10，102-100-0-0-10，102-100-0-0-10，102-100-0-0-10。
子网ID	102-100-10-0-10	字符串	指定子网ID。子网ID格式为：102-100-10-0-10，子网ID至少由1个大写字母、1个小写字母、数字和特殊字符（如\$%^&*+=~_!@#~）组成。
子网名称	102-100-10-1	字符串	子网名称，至少由1个大写字母、1个小写字母、数字和特殊字符（如\$%^&*+=~_!@#~）组成。
计费模式（不指定用户付费）	选择	字符串	指定计费模式，默认为按需计费，也可以选择包年包月。
计费周期（不指定用户付费）	month	字符串	指定计费周期，默认为按月计费，也可以选择按年计费。
计费周期（不指定用户付费）	1	字符串	指定计费周期，默认为按月计费，也可以选择按年计费。

高级设置

安全设置

创建直接计划后，点击 确定

创建直接计划

通过API创建，可以创建直接计划。

执行计划名称

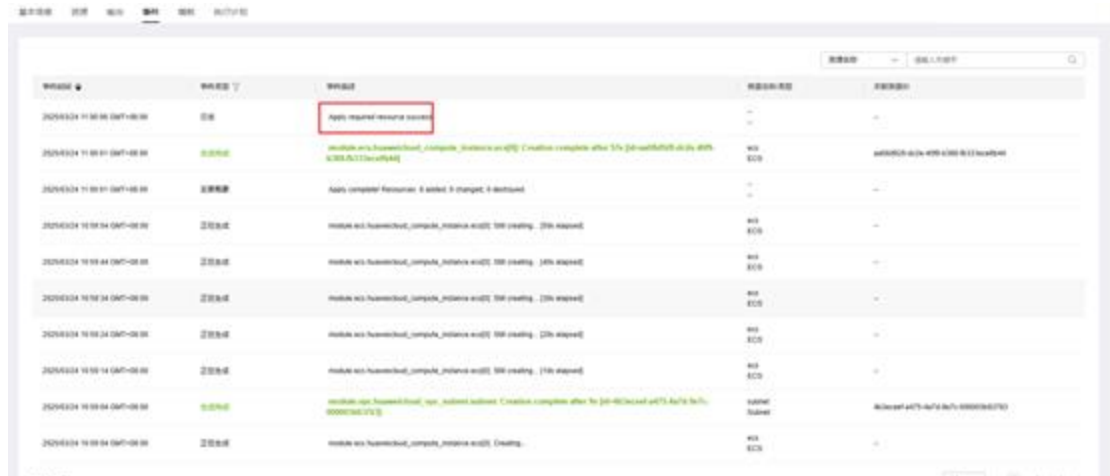
描述

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：

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

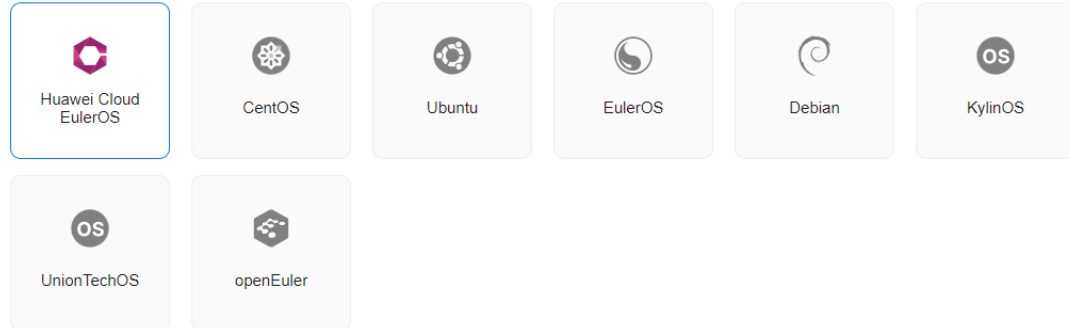
镜像 ?

公共镜像

私有镜像

共享镜像

市场镜像



Huawei Cloud EulerOS 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

云服务器的名称: ecs-kettle ☐ 允许重名

描述:

登录凭证:

密钥对:

云备份:

云备份策略 (可选):

高级选项: ☒ 现在配置

实例自定义数据脚本:

脚本内容:

```
#!/bin/bash
echo 'root:xox' | chpasswd
bash homeint.sh
```

购买: 配置费用 ¥0.3988/小时 + 弹性公网IP费用 ¥0.80/GB

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
- 2025-5-21 华为保密信息, 未经授权禁止扩散 第5页, 共7页

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。

```
Last login: Fri May 16 12:37:51 2025 from 192.168.0.241
cd "/root"
(vllm) [root@localhost ~]# cd "/root"
```

4.1.1 查看 vLLM 帮助

运行命令：vllm -h

```
(vllm) [root@localhost ~]# vllm -h
INFO 05-16 16:05:16 [__init__.py:239] Automatically detected platform cpu.
usage: vllm [-h] [-v] {chat,complete,serve,bench} ...

vLLM CLI

positional arguments:
  {chat,complete,serve,bench}
    chat                Generate chat completions via the running API server
    complete            Generate text completions based on the given prompt via the running API server
    serve               Start the vLLM OpenAI Compatible API server
    bench              vLLM bench subcommand.

options:
  -h, --help            show this help message and exit
  -v, --version         show program's version number and exit
```

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](#)