llama.cpp推理框架使用指南

1 商品说明

llama.cpp采用纯 C/C++ 编写，无需任何外部依赖， 提供最佳的性能和跨平台兼容性，从嵌入式设备到高性能服务器都能流畅运行。专为大语言模型推理优化，支持量化技术和硬件加速， 在保持模型质量的同时显著降低内存占用和计算成本。支持 CPU、GPU（CUDA、OpenCL、Metal）等多种硬件平台， 兼容 Windows、macOS、Linux 等操作系统，满足各种部署需求。

本商品在鲲鹏云的上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

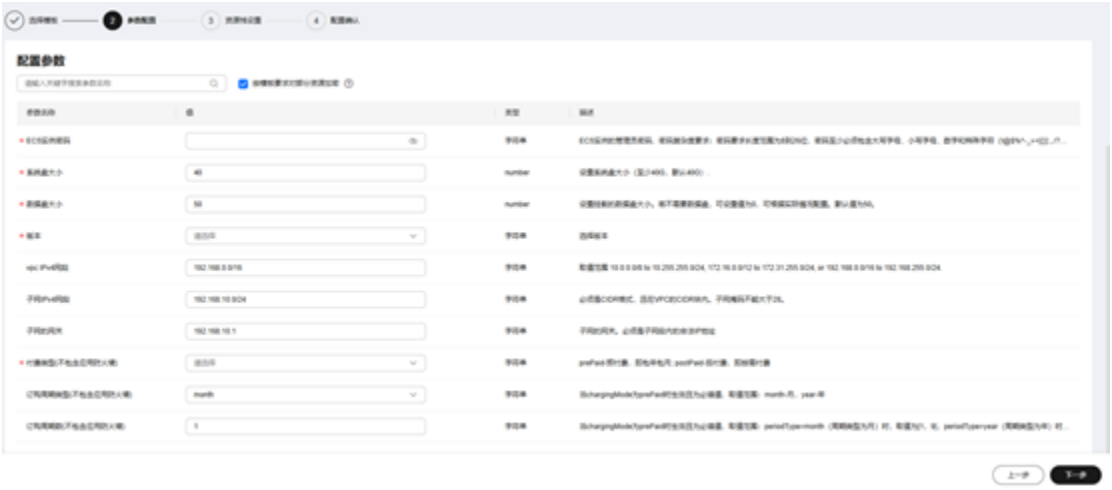
您可以在云商店搜索“llama.cpp推理框架”。

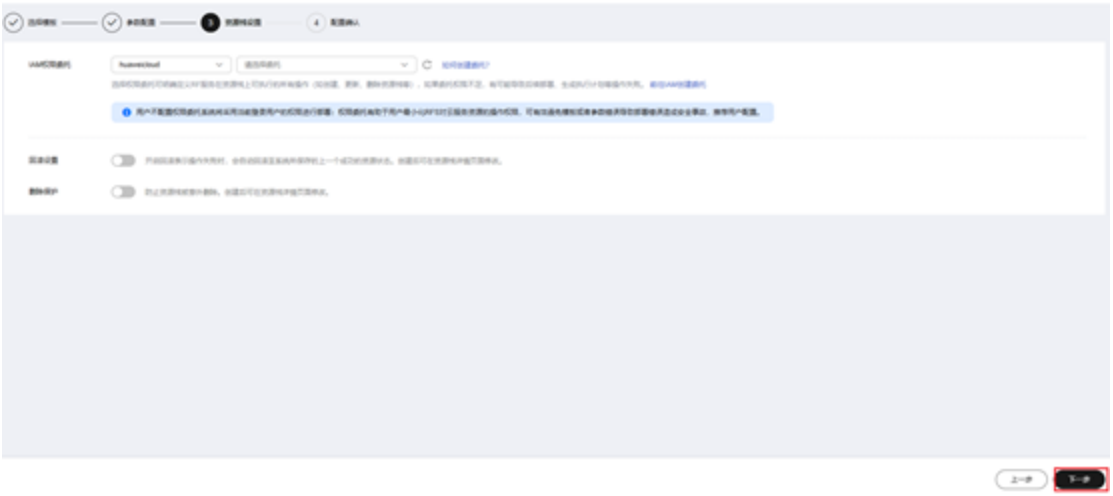
其中，地域、规格、推荐配置使用默认，购买方式根据您的需求选择按需/按月/按年，短期使用推荐按需，长期使用推荐按月/按年，确认配置后点击“立即购买”。

2.1 商品支持自定义 ECS 购买，具体见章节 3.1.1

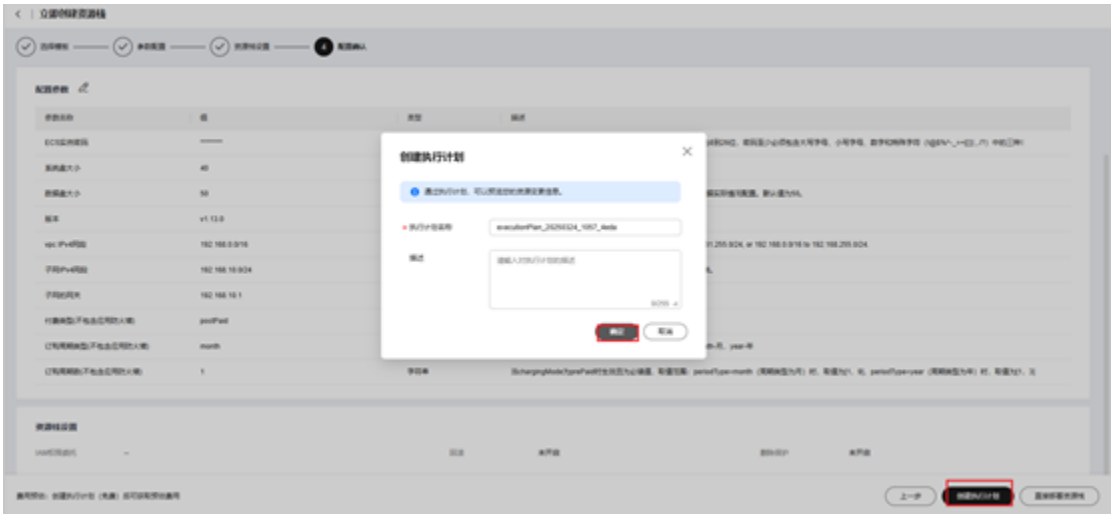
2.2 使用 RFS 模板直接部署

必填项填写后，点击 下一步



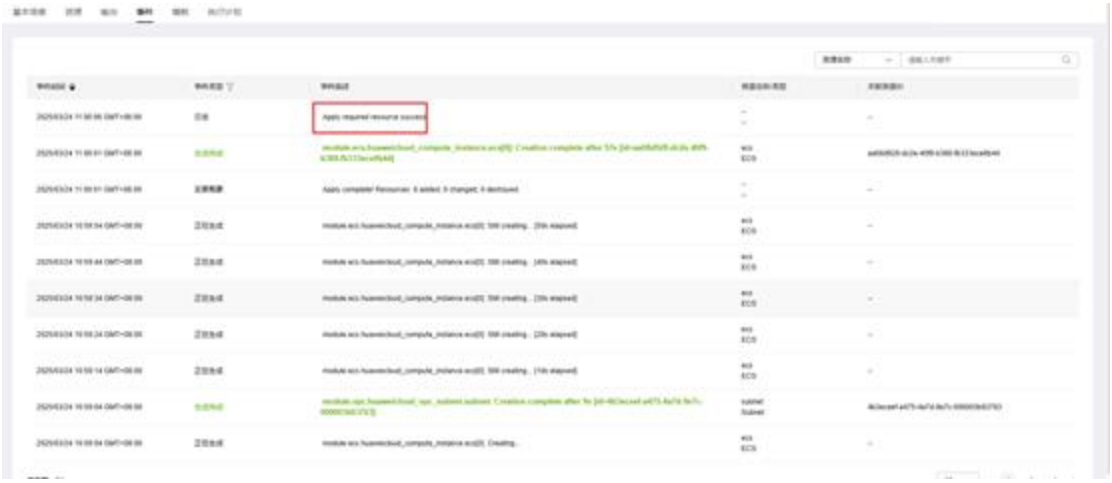


创建直接计划后，点击 确定



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

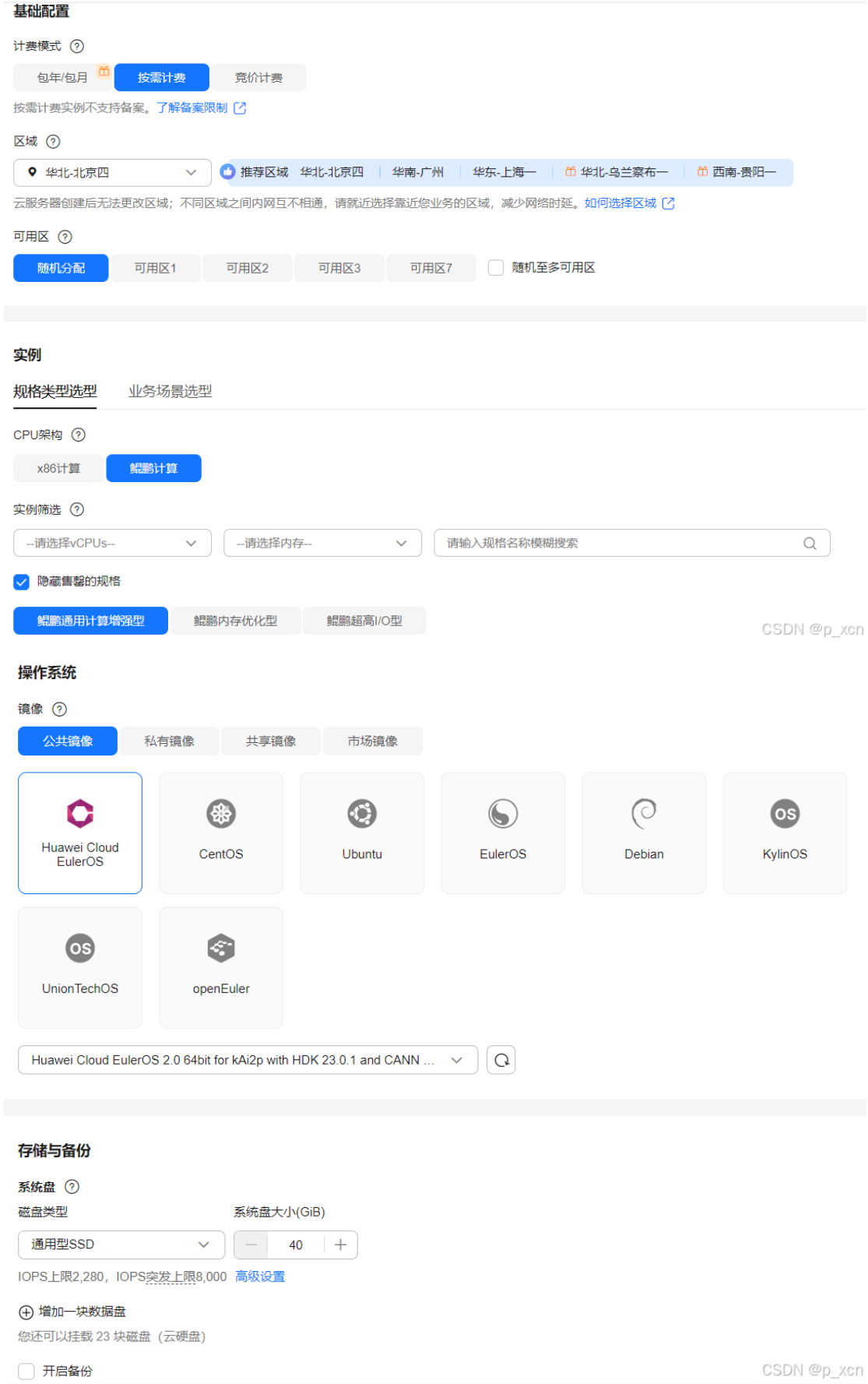
在使用ECS控制台配置前，需要您提前配置好**安全组规则**。

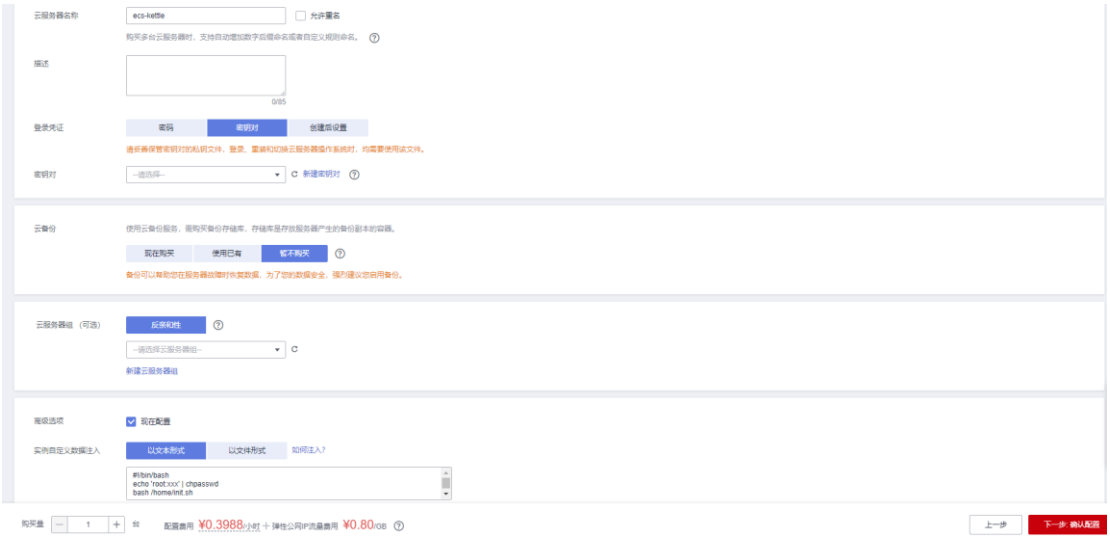
安全组规则的配置如下：

- 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：

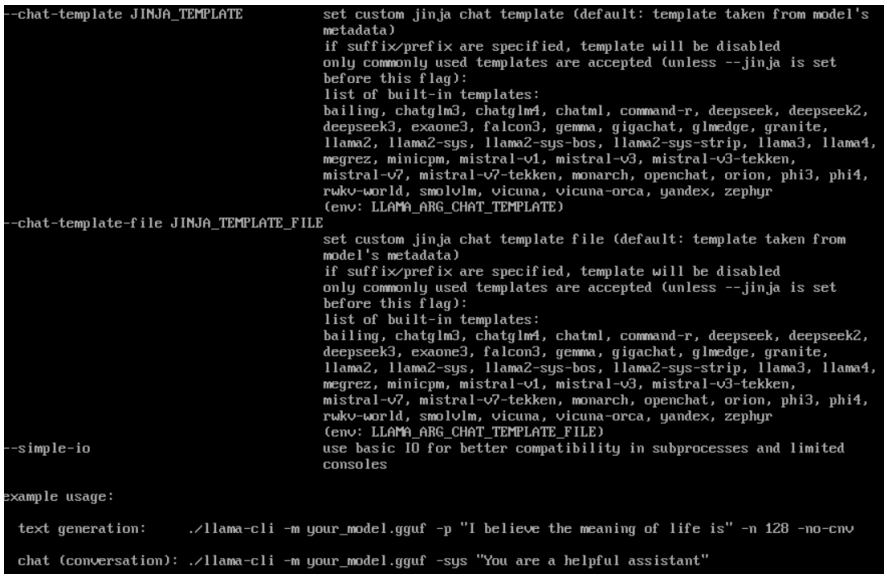
- VPC您可以自行创建
- 安全组选择3.1.1章节中配置的安全组
- 弹性公网IP选择现在购买，推荐选择“按流量计费”，带宽大小可设置为5Mbit/s
- 高级配置需要在高级选项支持注入自定义数据，所以登录凭证不能选择“密码”，选择创建后设置
- 其余默认或按规则填写即可。

4 商品使用

4.1 登录服务器查看 llama.cpp

通过命令 llama-cli -h / llama-server -h 查看服务状态。

下图为llama-cli -h的控制台输出。



4.2 体验模型

手工下载gguf格式的模型文件（例如Qwen3-0.6B-Q8_0.gguf 、bge-m3-q4_k_m.gguf、 bge-reranker-v2-m3-Q4_K_M.gguf） 然后根据路径修改下面的命令行。

聊天模型

服务端

llama-server -m Qwen3-0.6B-Q8_0.gguf --host 0.0.0.0

#客户端

curl http://localhost:8080/completion \

-H "Content-Type: application/json" \

-d '{"prompt": "hello", "n_predict":12}'

Embedding模型

服务端

llama-server -m bge-m3-q4_k_m.gguf --embedding --host 0.0.0.0

#客户端

curl http://localhost:8080/embedding \

```
-H "Content-Type: application/json" \
-d '{"content": "your text here"}'


# Rerank模型
#  服务端
llama-server -m bge-reranker-v2-m3-Q4_K_M.gguf --rerank --host 0.0.0.0
#客户端
curl http://localhost:8080/rerank \
-H "Content-Type: application/json" \
-d '{"query": "today is monday",
      "documents":[
            "Tomorrow is Tuesday",
            "I like English"
      ]}'
```

4.3 详细使用参考社区文档

llama-cli文档

llama-server文档