

1. 开通阿里云 GPU 云服务器

1.1 搜索“GPU 云服务器”，点击“立即开通”



1.2 根据自身情况选择付费类型，推荐“按量付费”；

1.3 根据实际业务重心，选择地域；

1.4 选择网络（不是所有可用区都有所有机型的，建议结合机型的可用区进行选择）；

付费类型 ②

包年包月 先付费后使用，价格优惠

按量付费 先使用后付费，按需开通

抢占式实例 较按量付费最高可省90%

使用须知 按量付费实例不支持备案服务

地域 ②

如何选择地域

华北 2 (北京) 116 华东 1 (杭州) 56 华东 2 (上海) 43 请选择其他地域

使用须知 实例创建之后地域将无法更改，不同地域的实例之间内网互不相通；距离实例所在地域越近，对实例访问速度越快

网络及可用区 ②

如何选择可用区

专有网络 经典网络

华北 2 可用区 M | cn-shanghai-m / vsw-uf62l1ytgo65t1wscqmpp | 网段: 192.168.0.0...
创建专有网络 创建交换机

指定主网卡主私网IP地址

1.5 选择实例规格（选择 GPU 型号和数量）

The screenshot shows the 'Instance Specifications' section of the instance creation wizard. It includes a search bar for '模糊搜索规格名称' (Search for specification name) and filters for '筛选' (Filter), '选择 vCPU' (Select vCPU), '选择内存' (Select memory), and '架构' (Architecture). The 'GPU/FPGA/ASIC' tab is selected. Below the filters is a table listing two GPU instance types:

规格族	实例规格	vCPU	内存	GPU/FPGA	GPU显存	可售可用区	架构	参考价格
GPU 计算型 gn7i	ecs.gn7i-c8g1.2xlarge	8 vCPU	30 GiB	1 * NVIDIA A10	1 * 24 GB	23个可用区	GPU/F 加速	¥9.5326/时
GPU 计算型 gn7i	ecs.gn7i-c16g1.4xlarge	16 vCPU	60 GiB	1 * NVIDIA A10	1 * 24 GB	23个可用区	GPU/F 加速	¥10.0934/时

1.6 选择安装镜像

点击云市场镜像，搜索“deepgpu-llm-inference”，选择镜像进行安装；

The screenshot shows the 'Cloud Market Images' section. It includes tabs for '最近使用镜像' (Recently used images), '公共镜像' (Public images), '自定义镜像' (Custom images), '共享镜像' (Shared images), '云市场镜像' (Cloud Market Images), and '社区镜像' (Community images). A red box highlights the '从云市场获取更多选择 (含操作系统)' (Get more choices from the cloud market, including operating systems) button.

1.7 根据业务需求，配置云盘大小

The screenshot shows the 'Storage' configuration section. It includes a dropdown for '系统盘' (System disk) set to 'ESSD 云盘'. The table below shows the configuration details:

类型	容量	数量	IOPS	性能	操作
ESSD 云盘	500 GiB	1	7800	PLO (单盘IOPS性能上限1万)	<input checked="" type="checkbox"/> 随实例释放 <input type="checkbox"/> 加密

1.8 分配公网 IP，并配置带宽峰值（推荐按量付费，带宽拉满）

The screenshot shows the 'Public Network IP' and 'Bandwidth Configuration' sections. Under '公网 IP' (Public Network IP), there is a checkbox for '分配公网 IPv4 地址' (Assign public network IPv4 address) which is checked. Under '带宽计费模式' (Bandwidth Billing Mode), the '按使用流量' (Pay-as-you-go) option is selected. Under '带宽峰值' (Peak Bandwidth), a value of 100 Mbps is selected. A note at the bottom states: '阿里云免费提供最高 5Gbps 的恶意流量攻击防护。了解更多> | 提升防护能力>' (Alibaba Cloud provides free protection against up to 5Gbps of malicious traffic attacks. Learn more | Improve protection).

1.9 配置安全组

The screenshot shows the 'Security Groups' configuration page. At the top, there are tabs for 'Existing Security Groups' (selected) and 'Create New Security Group'. Below the tabs, there's a link 'How to Configure Security Groups' and a button 'Select Existing Security Group'. A note indicates that 1 instance + auxiliary network card is already assigned, with 1994 more available. A warning message states: 'Please ensure the selected security group opens port 22 (Linux) or 3389 (Windows), otherwise you cannot remotely log in to ECS. Go to settings'.

1.10 配置机器登录密码（推荐自定义密码）

The screenshot shows the 'Instance Configuration' page under 'Management Settings'. It includes sections for 'Login Certificate' (with options for 'Key Pair', 'Image Predefined Password', 'Custom Password' - selected, and 'Post-creation Settings'), 'Login Name' (set to 'root'), 'Login Password' (set to 'Temporary'), 'Confirm Password' (set to 'Temporary'), and 'Tags' (0/20). A note at the bottom says 'Labels are used to distinguish between uppercase and lowercase letters in key-value pairs. The labels you set will be applied to all instances and cloud disks in this order'.

1.11 勾选服务条款，并确认下单

The screenshot shows the final step of the order process. It displays the configuration fees: 'Configuration Fee: ¥ *** / hour' and 'Original Price: ¥ 10.62 / hour'. It also shows public network traffic fees: 'Public Network Traffic Fee: ¥ *** / GB' and 'Original Price: ¥ 0.800 / GB'. A checked checkbox indicates acceptance of the 'ECS Service Agreement'. A large orange button at the bottom right says 'Place Order'.

1.12 创建成功，点击“管理控制台”查看机器



1.13 查看机器状态并获取 IP 地址, SSH 远程登录

标签	监控	可用区 ▾	IP地址	状态 ▾	网络类型 ▾	配置	付费方式 ▾
u5klbhc test-	◆	⚠️	Ⓜ 上海 可用区 M	59.190.204.117 (公) 02.450.0.114 (私有)	运行中	专有网络	16 vCPU 60 GiB (I/O优化) GPU: NVIDIA A10 ecs.gn7i-c16g1.4xlarge 100Mbps (峰值) 按量 2023年10月18日 10:40 创建

2. 运行 LLM 模型推理

2.1 查看 DeepGPU-LLM 版本，确认是否需要升级

查看 DeepGPU-LLM 版本和安装路径

```
pip show -f deepgpu-llm
```

```
(base) root@iZuf6dxgo2te0ttu5klbhcZ:~# pip show -f deepgpu_llm
WARNING: Ignoring invalid distribution -ccelerate (/workspace/miniconda/lib/python3.10/site-packages)
Name: deepgpu_llm
Version: 0.9.7+pt2.0cu117
Summary: DeepGPU LLM inference package
Home-page:
Author:
Author-email:
License:
Location: /workspace/miniconda/lib/python3.10/site-packages
Requires: bfloat16, colorama, SentencePiece, transformers
Required-by:
Files:
    ..../..../bin/baichuan_cli
    ..../..../bin/baichuan_hf_cli
    ..../..../bin/chatglm_cli
    ..../..../bin/chatglm_hf_cli
    ..../..../bin/gpt_gemm
    ..../..../bin/huggingface_baichuan_convert
    ..../..../bin/huggingface_chatglm2_convert
    ..../..../bin/huggingface_glm_convert
    ..../..../bin/huggingface_llama_convert
    ..../..../bin/llama_cli
deepgpu_llm-0.9.7+pt2.0cu117.dist-info/INSTALLER
deepgpu_llm-0.9.7+pt2.0cu117.dist-info/METADATA
deepgpu_llm-0.9.7+pt2.0cu117.dist-info/RECORD
deepgpu_llm-0.9.7+pt2.0cu117.dist-info/REQUESTED
deepgpu_llm-0.9.7+pt2.0cu117.dist-info/WHEEL
deepgpu_llm-0.9.7+pt2.0cu117.dist-info/direct_url.json
deepgpu_llm-0.9.7+pt2.0cu117.dist-info/top_level.txt
deepgpu_llm/_init_.py
deepgpu_llm/_pycache_/_init_.cpython-310.pyc
deepgpu_llm/_pycache_/_baichuan_model.cpython-310.pyc
deepgpu_llm/_pycache_/_chatglm_model.cpython-310.pyc
deepgpu_llm/_pycache_/_deepgpu_utils.cpython-310.pyc
deepgpu_llm/_pycache_/_llama_model.cpython-310.pyc
deepgpu_llm/baichuan_model.py
deepgpu_llm/chatglm_model.py
deepgpu_llm/deepgpu_utils.py
deepgpu_llm/libdeepgpu_glm.so
deepgpu_llm/libdeepgpu_llama.so
deepgpu_llm/llama_model.py
```

查看最新版本: https://aiacc-inference-public-v2.oss-cn-hangzhou.aliyuncs.com/aiacc-inference-llm/deepgpu_llm.html

下载命令

```
wget https://aiacc-inference-public-v2.oss-cn-hangzhou.aliyuncs.com/aiacc-inference-llm/deepgpu_llm-0.9.7%2Bpt2.0cu117-py3-none-any.whl
```

安装命令

```
pip install xxx.whl
```

2.2 下载或上传 huggingface 标准的 LLM 模型

上传自己的模型，或者从 huggingface 下载开源模型。

下载命令

```
git-lfs clone https://huggingface.co/meta-llama/Llama-2-7b
git-lfs clone https://huggingface.co/THUDM/chatglm2-6b
git-lfs clone https://huggingface.co/baichuan-inc/Baichuan-13B-Chat
```

近期国内对 huggingface 基本全部墙了，有个迂回方案，不能保证所有模型适用

1. 从 huggingface 下载 LLM 模型相关的代码和配置文件（权重除外）

2. 从 modelscope 下载模型:

```
https://modelscope.cn/models/ZhipuAI/chatglm2-6b
https://modelscope.cn/models/ZhipuAI/ChatGLM-6B
https://modelscope.cn/models/baichuan-inc/Baichuan2-13B-Chat
https://modelscope.cn/models/baichuan-inc/Baichuan-13B-Chat
```

3. 将 huggingface 下载的代码和权重替换到 modelscope 下载的模型目录中

2.3 模型转换

转换命令

```
huggingface_baichuan_convert -in_file /root/deepGPU/models/Baichuan2-7B-Chat/ -saved_dir  
/root/deepGPU/models/deepgpu/baichuan2-7b-chat -infer_gpu_num 1 -weight_data_type  
fp16 -model_name baichuan2-7b-chat
```

其中:

- in_file 指明原始 huggingface 模型目录
- saved_dir 指明转换后的模型目录
- infer_gpu_num 指明转换后模型运行所需的 GPU 数量
- weight_data_type 指明转换后模型运行时的计算精度
- model_name 模型名称

模型转换脚本选择

转换脚本	模型
huggingface_baichuan_convert	baichuan 和 baichuan2 系列模型
huggingface_llama_convert	llama 和 llama2 系列模型
huggingface_glm_convert	chatglm 和 GLM-130b 模型
huggingface_chatglm2_convert	chatglm2 模型

2.4 运行模型推理

使用 DeepGPU-LLM 自带的运行脚本:

```
baichuan_cli --tokenizer_dir /root/deepGPU/models/Baichuan2-7B-Chat/ --model_dir  
/root/deepGPU/models/deepgpu/baichuan2-7b-chat/1-gpu/
```

可以复制该脚本进行代码修改，实现自己的模型加载和运行，增量开发其他功能。