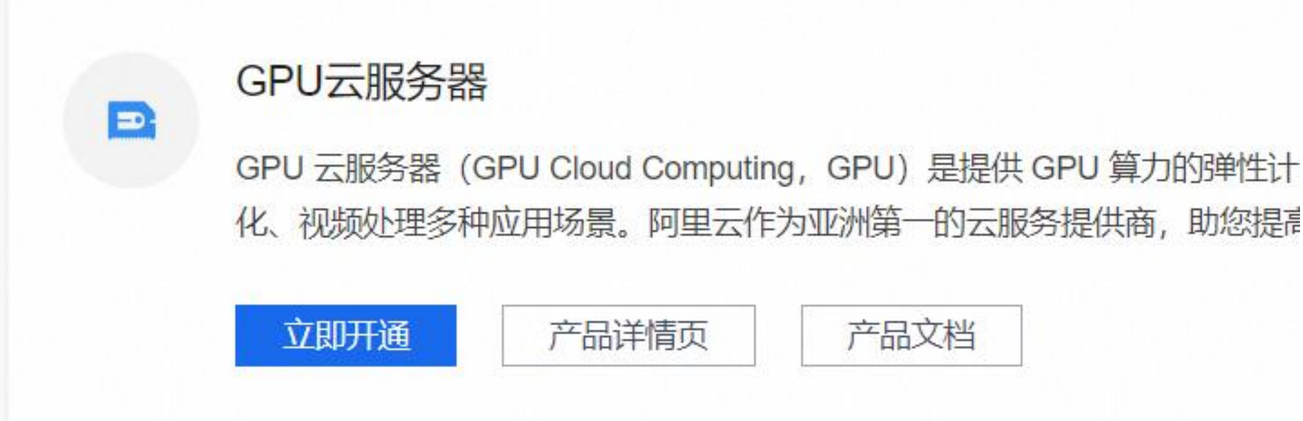
1. 开通阿里云 GPU 云服务器

1.1 搜索“GPU 云服务器”，点击“立即开通”



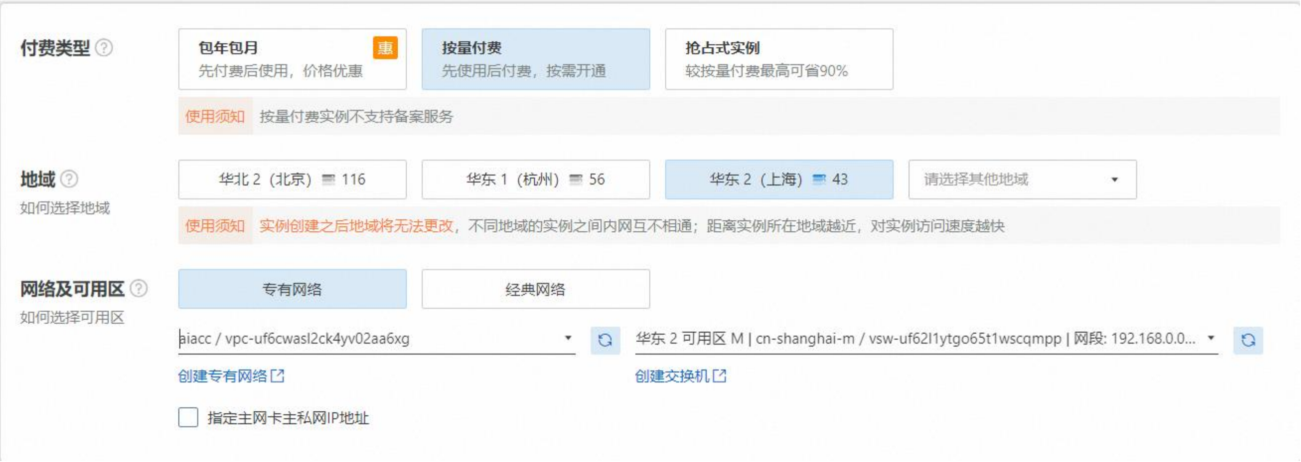
GPU云服务器

GPU 云服务器（GPU Cloud Computing，GPU）是提供 GPU 算力的弹性计
化、视频处理多种应用场景。阿里云作为亚洲第一的云服务提供商，助您提高

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：包年包月（先付费后使用，价格优惠）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（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

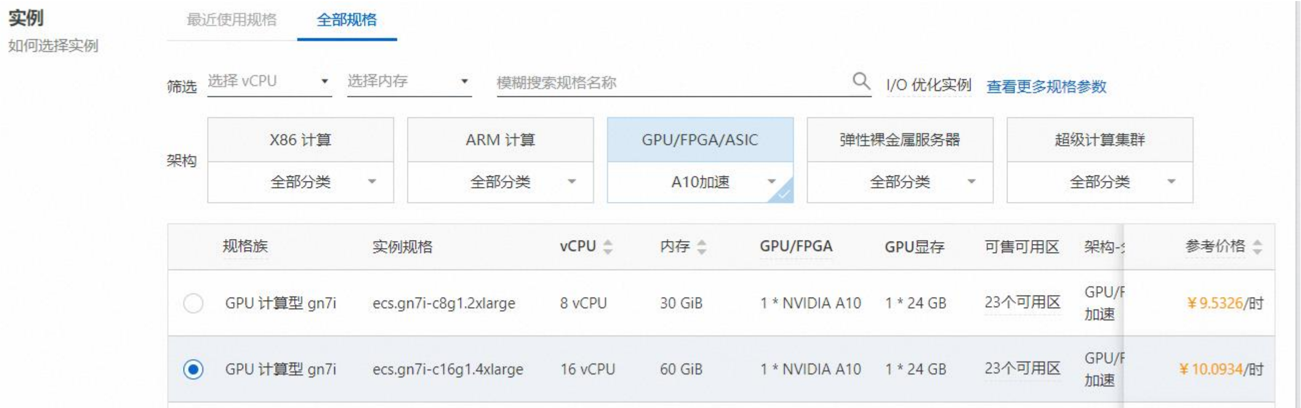
网络及可用区（如何选择可用区）：[✓ 专有网络] | 经典网络

aiacc / vpc-uf6cwasl2ck4yv02aa6xg

华东 2 可用区 M | cn-shanghai-m / vsw-uf62l1ytgo65t1wscqmpp | 网段: 192.168.0.0...

□ 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）



实例

如何选择实例

最近使用规格 | 全部规格

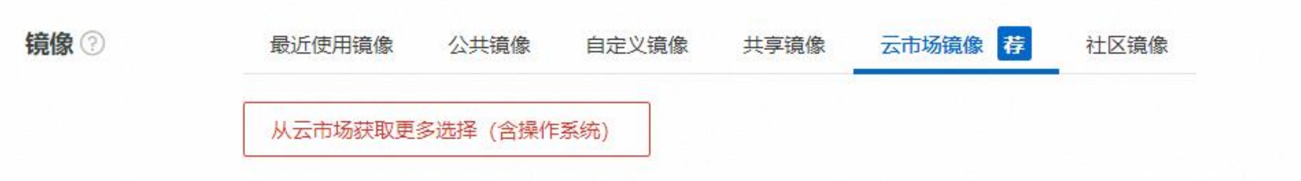
筛选 选择 vCPU 选择内存 模糊搜索规格名称 I/O 优化实例 查看更多规格参数

架构 X86 计算 全部分类 | ARM 计算 全部分类 | GPU/FPGA/ASIC A10加速 | 弹性裸金属服务器 全部分类 | 超级计算集群 全部分类

	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；

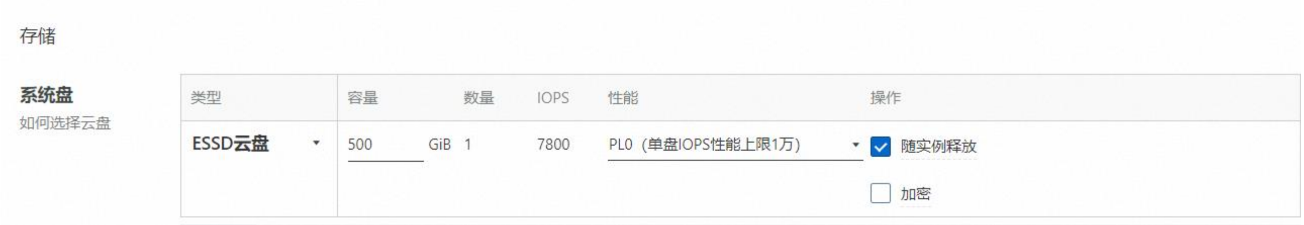


镜像

最近使用镜像 | 公共镜像 | 自定义镜像 | 共享镜像 | 云市场镜像 荐 | 社区镜像

从云市场获取更多选择（含操作系统）

1.7 根据业务需求，配置云盘大小



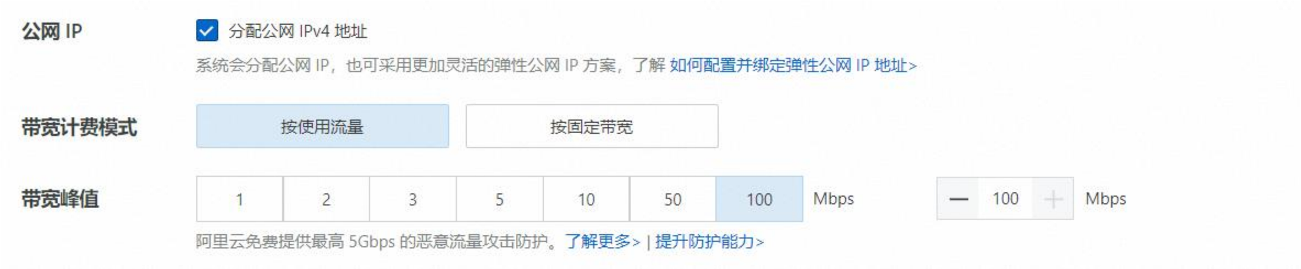
存储

系统盘

如何选择云盘

类型	容量	数量	IOPS	性能	操作
ESSD云盘	500 GiB	1	7800	PL0（单盘IOPS性能上限1万）	☑ 随实例释放 □ 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）



公网 IP ☑ 分配公网 IPv4 地址

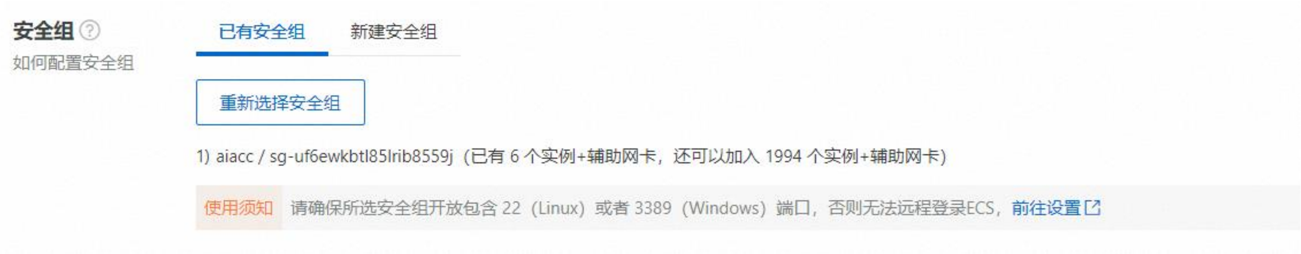
系统会分配公网 IP，也可采用更加灵活的弹性公网 IP 方案，了解 如何配置并绑定弹性公网 IP 地址>

带宽计费模式 按使用流量 | 按固定带宽

带宽峰值 1 | 2 | 3 | 5 | 10 | 50 | 100 Mbps — 100 + Mbps

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>

1.9 配置安全组

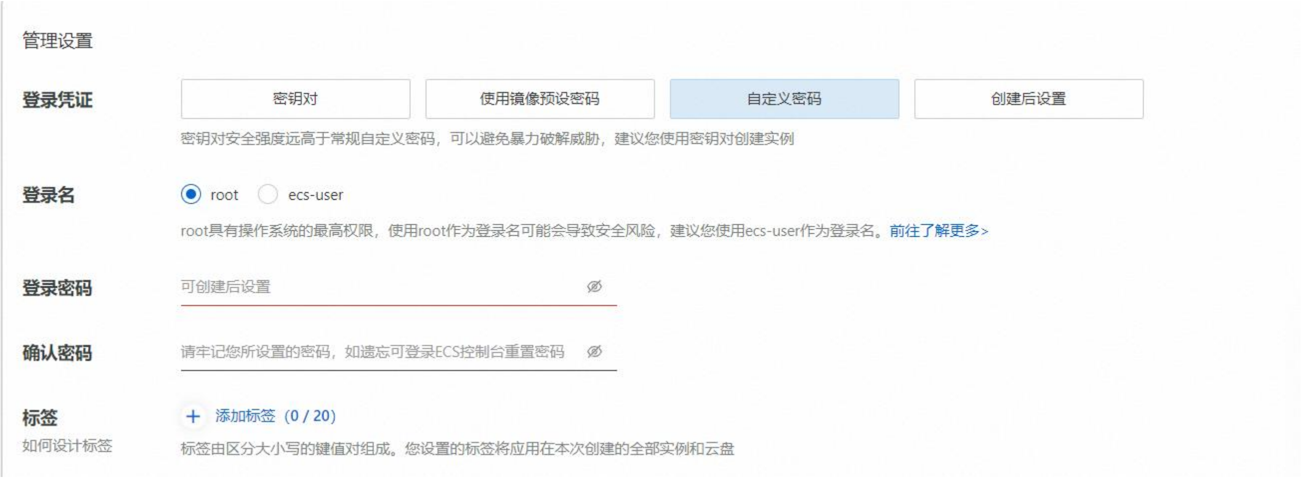


安全组 ⓘ
如何配置安全组

1) aiacc / sg-uf6ewkbtl85lrib8559j（已有 6 个实例+辅助网卡，还可以加入 1994 个实例+辅助网卡）

使用须知　请确保所选安全组开放包含 22（Linux）或者 3389（Windows）端口，否则无法远程登录ECS，前往设置

1.10 配置机器登录密码（推荐自定义密码）



管理设置

登录凭证　密钥对　使用镜像预设密码　[✓ 自定义密码]　创建后设置

密钥对安全强度远高于常规自定义密码，可以避免暴力破解威胁，建议您使用密钥对创建实例

登录名　● root　○ ecs-user

root具有操作系统的最高权限，使用root作为登录名可能会导致安全风险，建议您使用ecs-user作为登录名。前往了解更多>

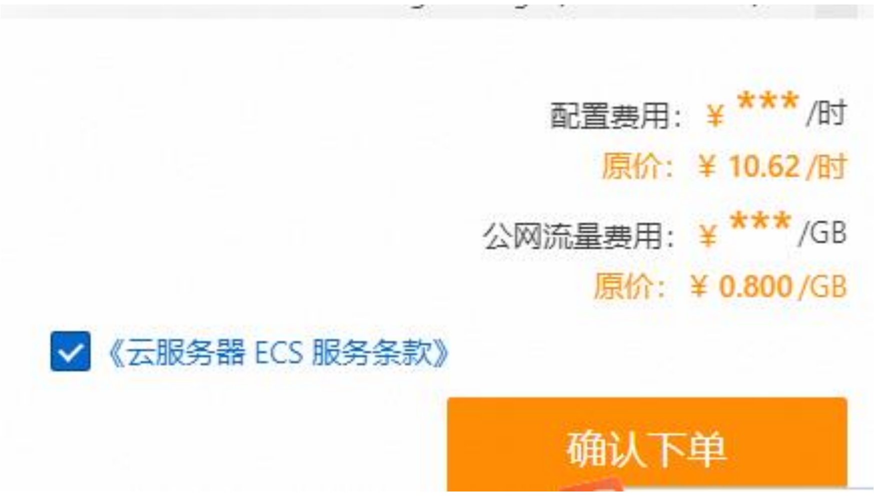
登录密码　可创建后设置

确认密码　请牢记您所设置的密码，如遗忘可登录ECS控制台重置密码

标签　＋ 添加标签（0 / 20）
如何设计标签

标签由区分大小写的键值对组成。您设置的标签将应用在本次创建的全部实例和云盘

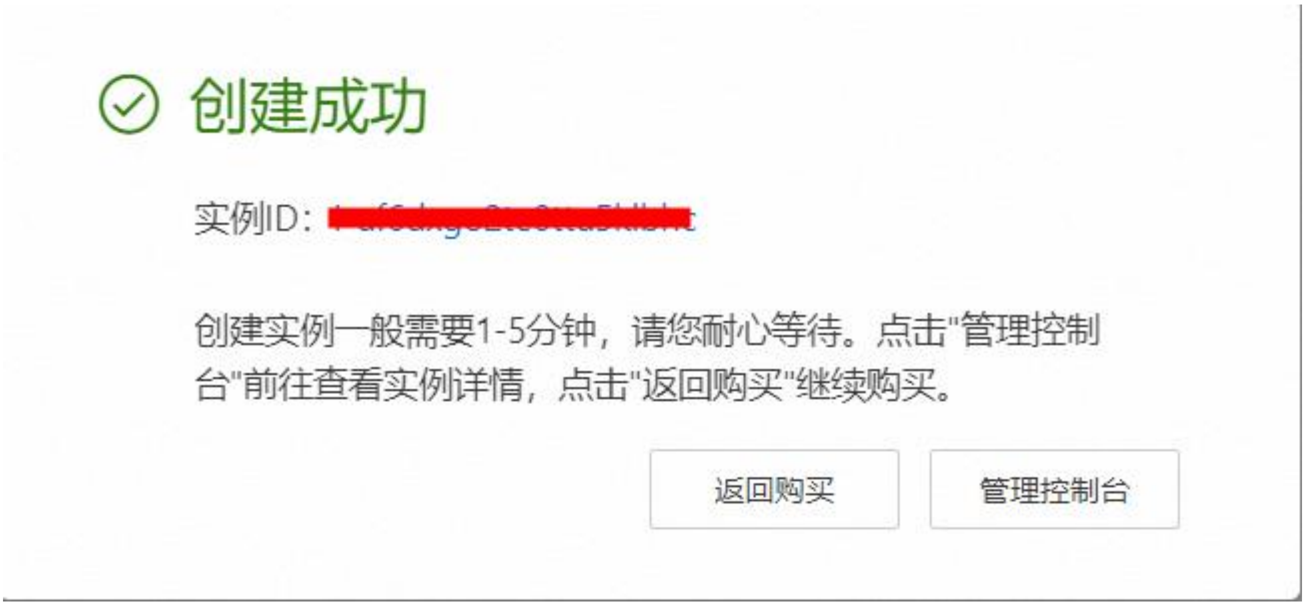
1.11 勾选服务条款，并确认下单



配置费用：￥ *** /时
原价：￥ 10.62 /时
公网流量费用：￥ *** /GB
原价：￥ 0.800 /GB

☑ 《云服务器 ECS 服务条款》

1.12 创建成功，点击“管理控制台”查看机器

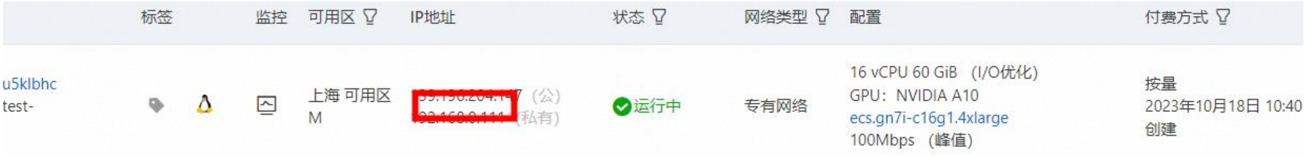


创建成功

实例ID: [redacted]

创建实例一般需要1-5分钟，请您耐心等待。点击"管理控制台"前往查看实例详情，点击"返回购买"继续购买。

1.13 查看机器状态并获取 IP 地址，SSH 远程登录



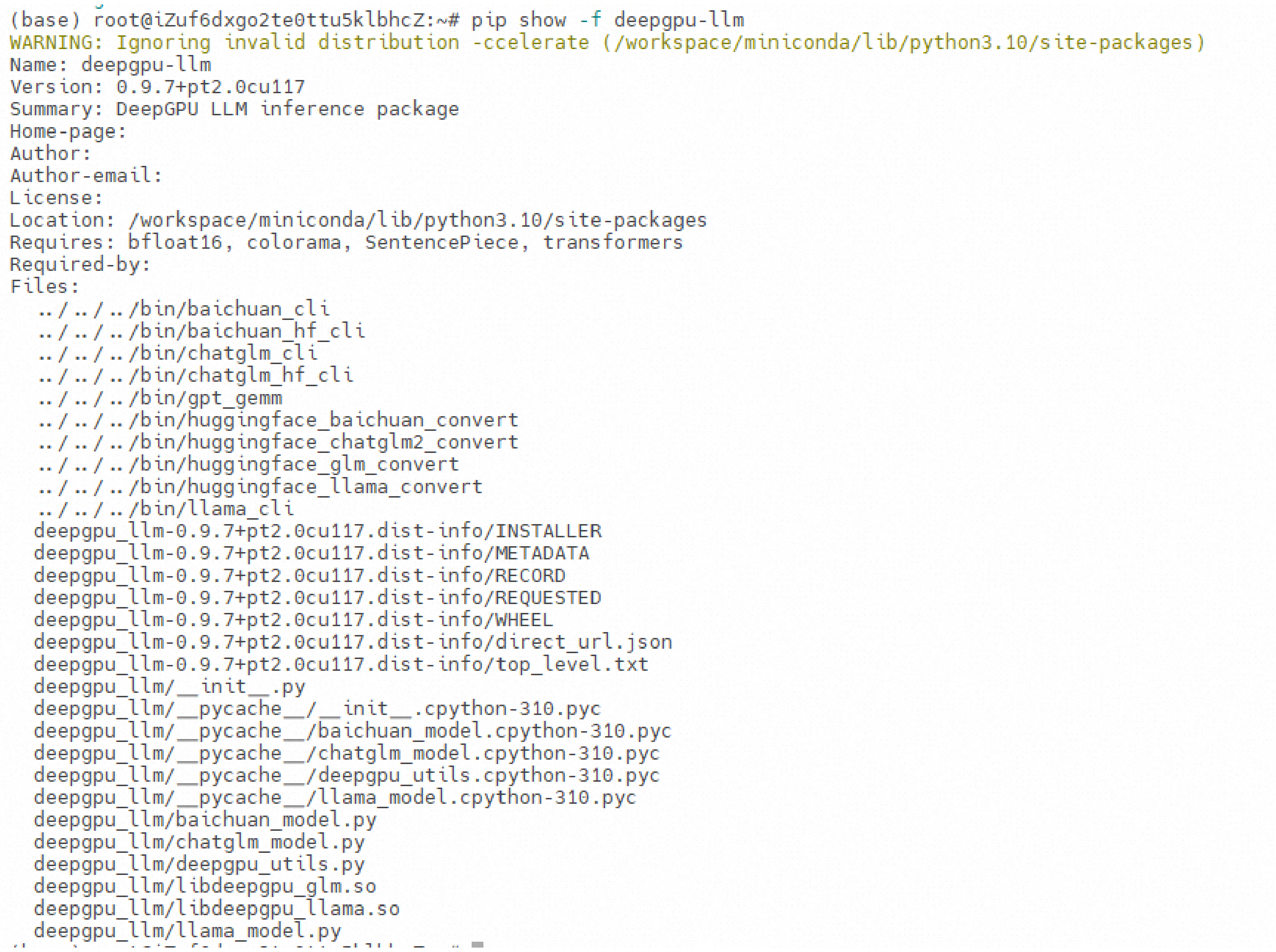
	标签		监控	可用区	IP地址	状态	网络类型	配置	付费方式
u5klbhc test-				上海 可用区 M	[redacted] (公) (私有)	运行中	专有网络	16 vCPU 60 GiB (I/O优化) GPU：NVIDIA A10 ecs.gn7i-c16g1.4xlarge 100Mbps (峰值)	按量 2023年10月18日 10:40 创建

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。