GPUStack- 集群版 用户手册

一、产品概述

1.1 产品介绍

GPUStack 云服务是基于开源 GPUStack 构建的托管式 AI 模型部署平台，让您无需管理基础设施，即可在高性能 GPU 集群上轻松部署和运行各类 AI 模型。

1.2 产品核心能力

资源管理：提供自动化 GPU 资源调度与集群管理，支持异构 GPU 设备统一纳管，实现资源利用率最大化与成本最优化；

模型部署：支持主流开源大模型一键部署，兼容 Hugging Face、ModelScope 等模型源，集成 vLLM、SGLang 和 TensorRT-LLM 等高性能推理引擎，满足不同场景性能需求；

智能运维：内置自动扩缩容、故障转移与负载均衡机制，提供实时性能监控与告警，确保服务高可用性与稳定性；

安全管控：提供完善的认证授权体系与网络隔离策略，支持私有化部署与数据安全保障，满足企业级安全合规要求。

1.3 产品优势

自动处理底层 GPU 资源调度、模型优化和扩展，让您专注于应用开发而非运维。

零运维负担：无需管理 GPU 驱动、CUDA 版本或集群配置；

开箱即用： 集成 vLLM、SGLang 和 TensorRT-LLM 等高性能推理引擎，支持自定义推理框架；

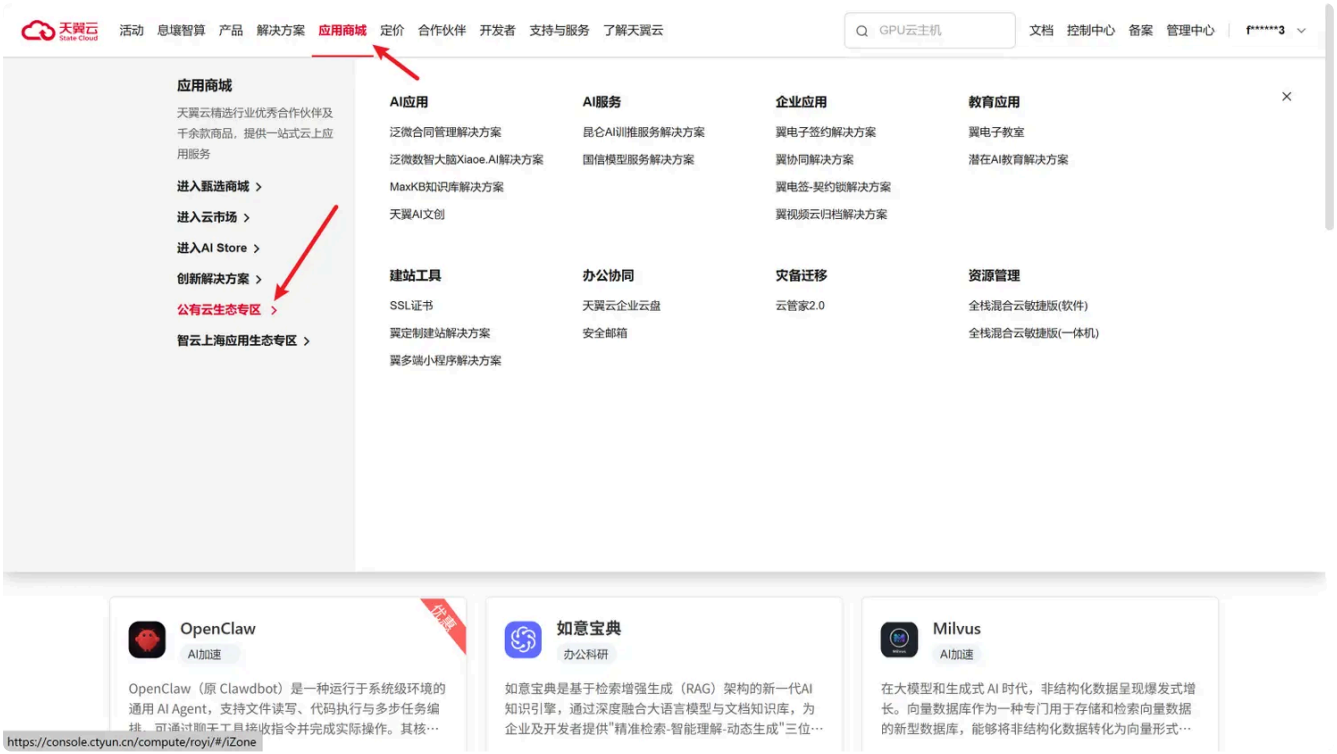
一键部署：支持从 Hugging Face、ModelScope、或本地直接部署，支持自动扩缩容、版本升降级；

性能优化配置：提供预调优模式，用于低延迟或高吞吐量；

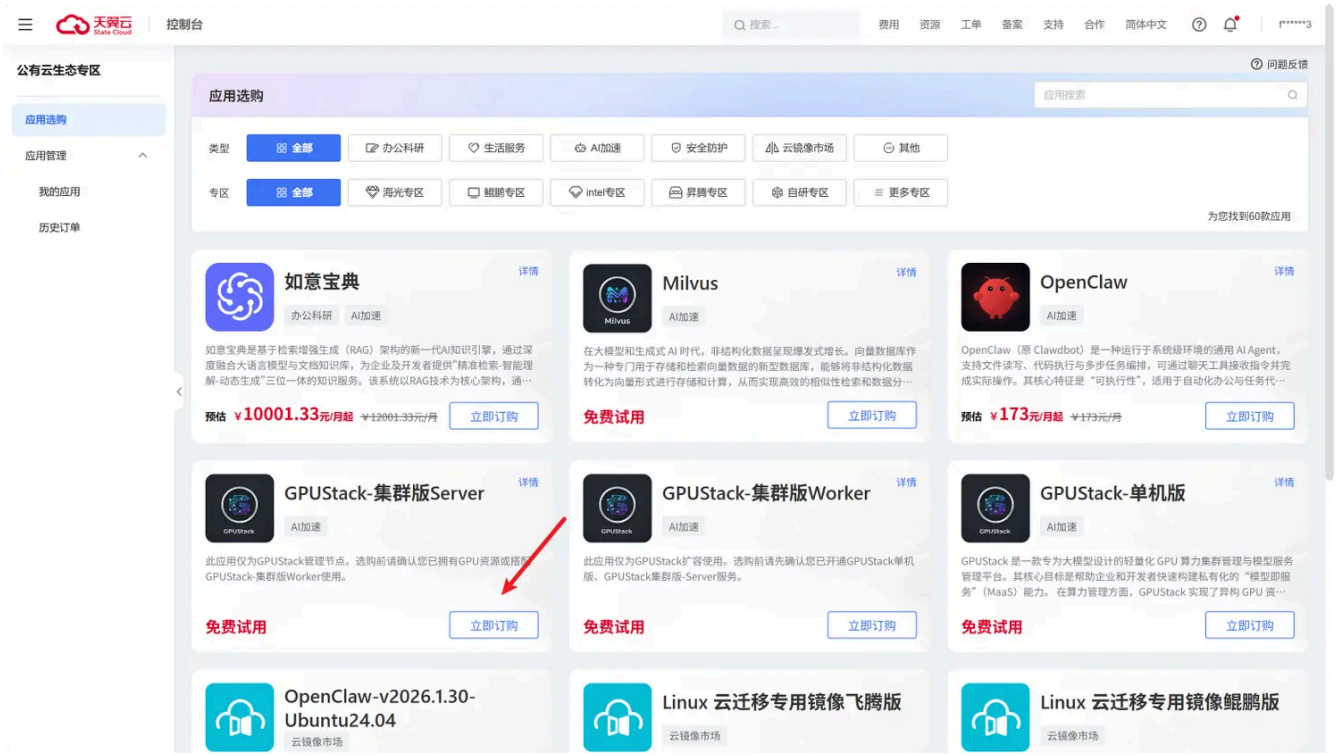
运维能力： 支持自动故障恢复、负载均衡、监控、认证和访问控制。

二、使用指南

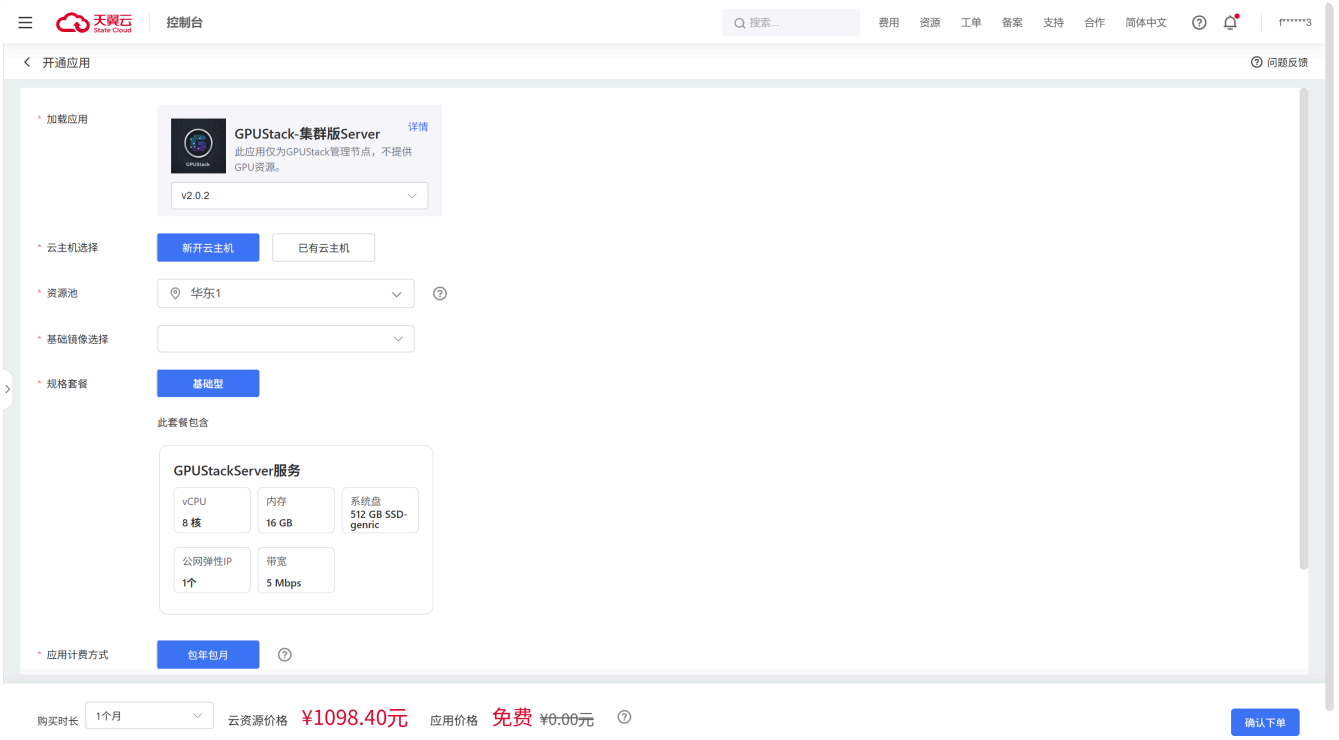
2.1 登录天翼云官网，选择【应用商城】-【公有云生态专区】，点击立即选购，进入应用专区页



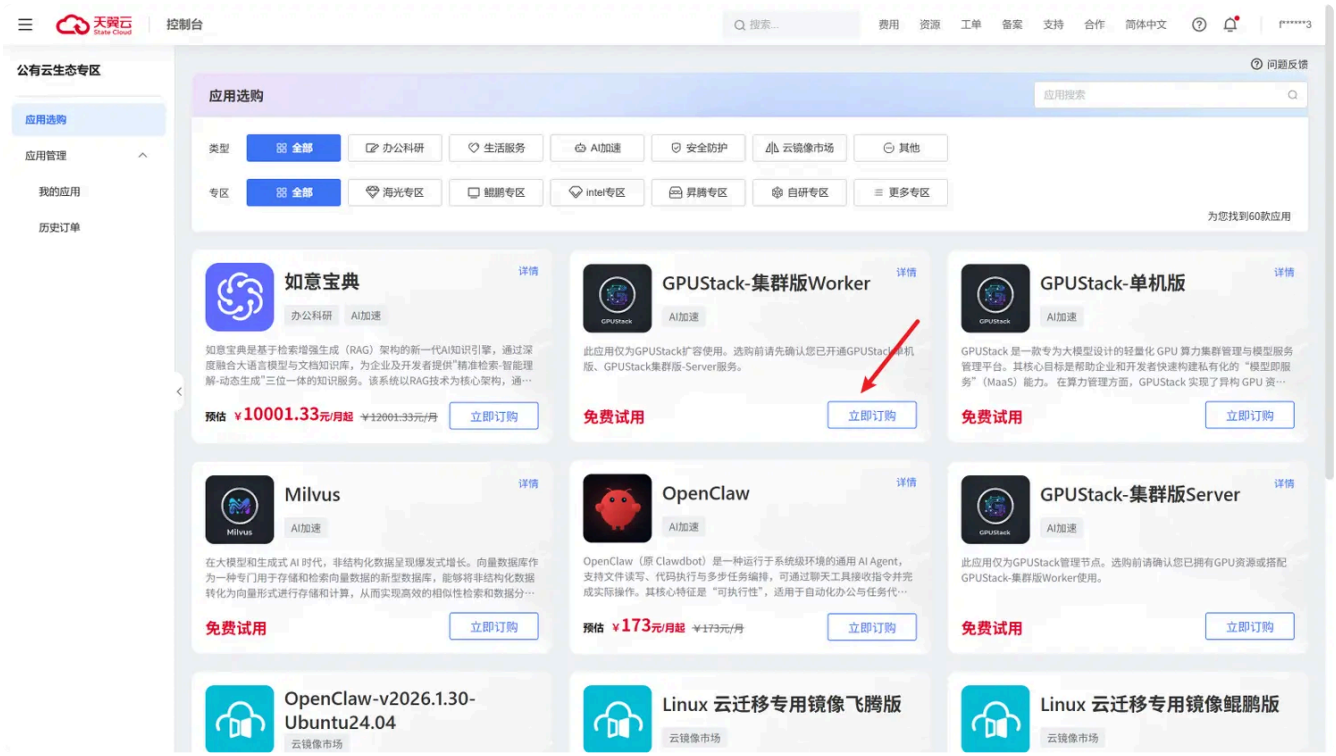
2.2 订购 Server 节点。在应用专区中选择“GPUStack 集群版 Server”，点击立即订购



2.3 按提示订购“GPUStack 集群版 Server”，点击确认下单即可成功购买服务。

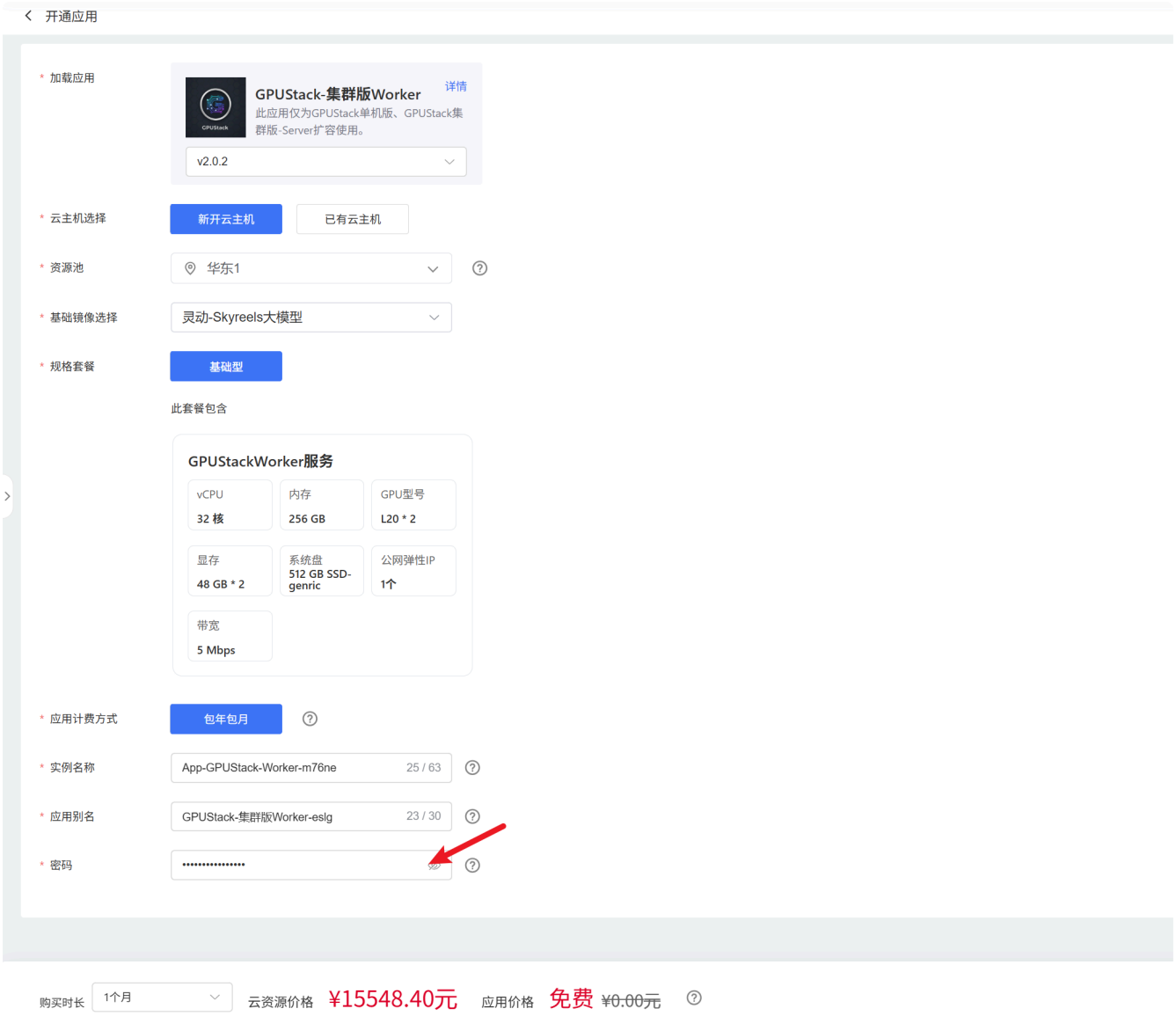


2.4 订购 Worker 节点。在应用专区中选择“GPUStack 集群版 Worker”，点击立即订购



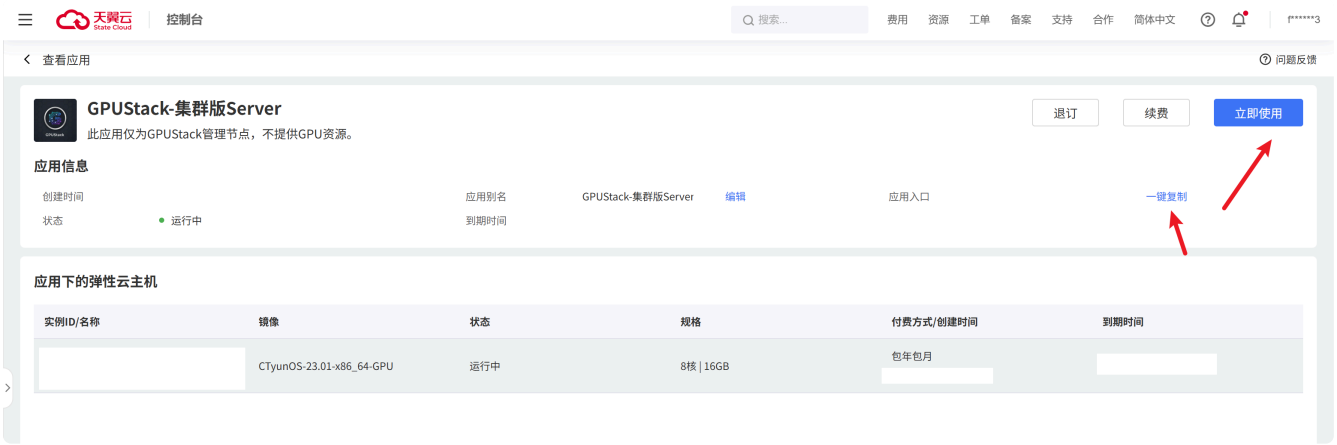
2.5 按提示订购“GPUStack 集群版 Worker”，点击确认下单即可成功购买服务。

说明：请重新设置或妥善保存云主机密码，在后续配置中使用。



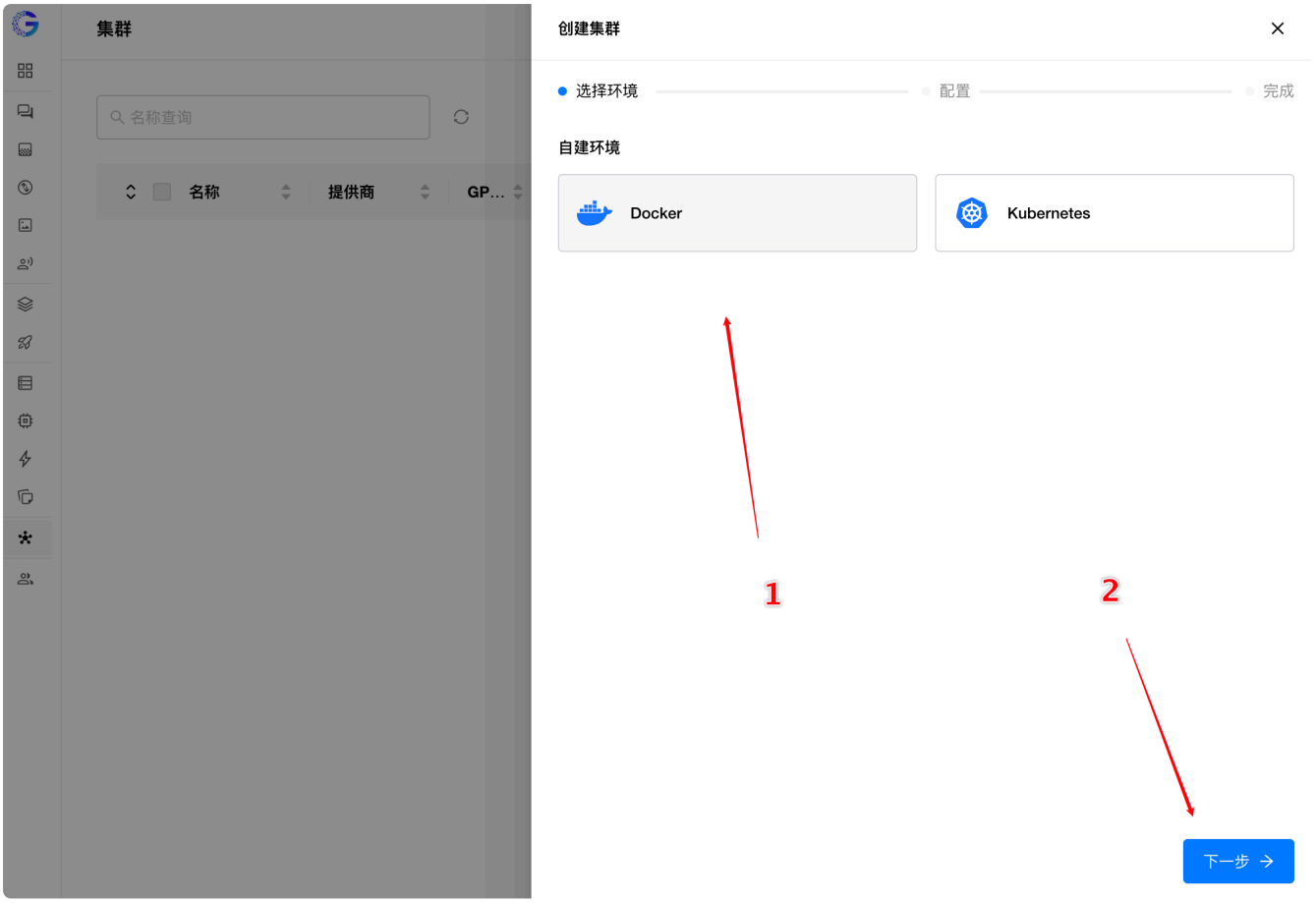
2.6 在【我的应用】页，查看应用状态，当 Server 应用状态为运行中时，代表 Server 服务部署完成。

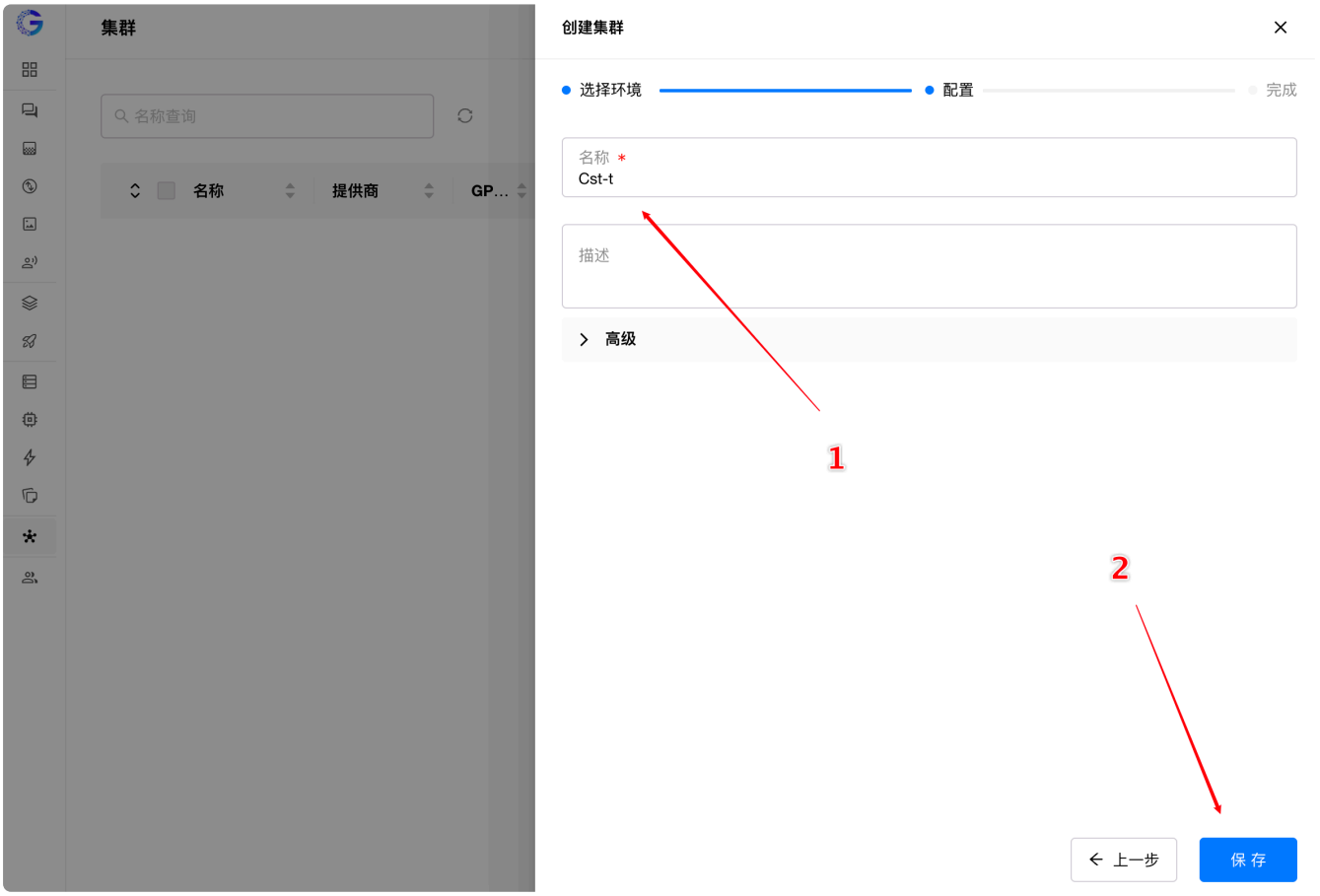
2.7 点击查看应用按钮，在浏览器中输入复制的应用入口，即可访问 GPUstack



2.8 GPUstack 默认账号为 admin，密码为 gpustack。（可在应用中修改）

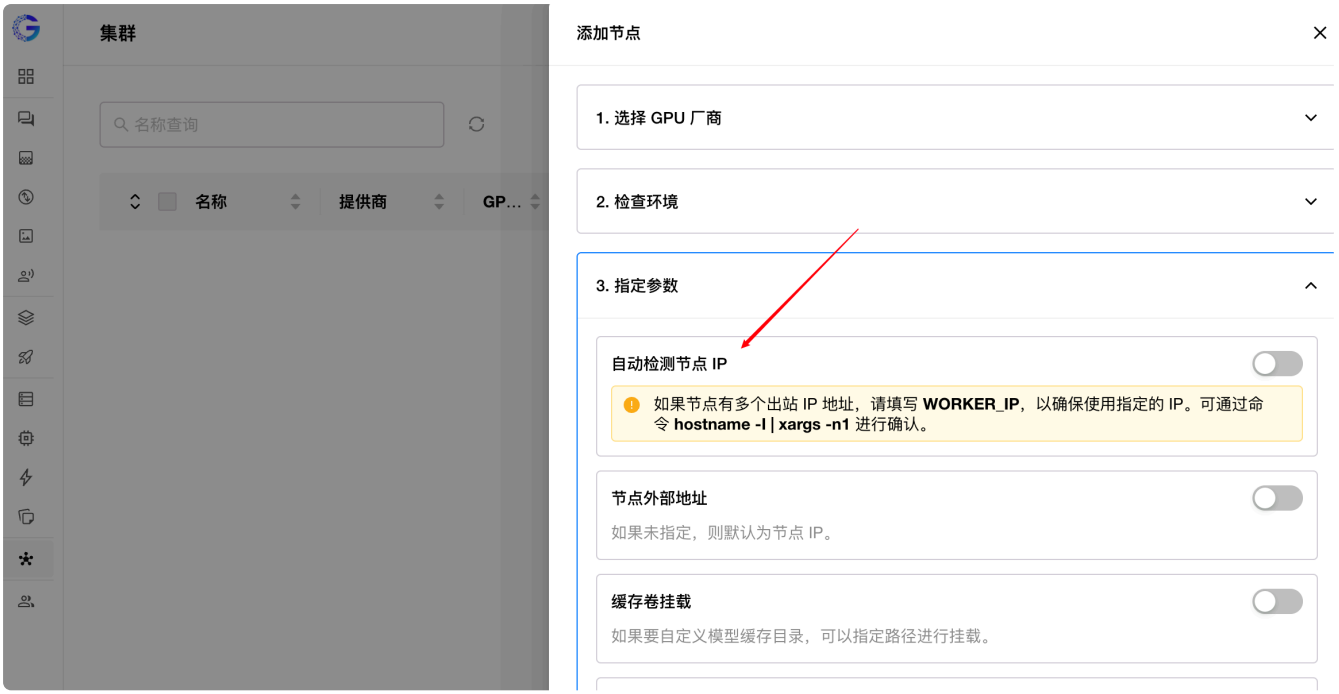
2.9 添加集群。参考以下操作，在 GPUStack 中创建新集群。


集群
名称查询
名称
提供商
GP...
创建集群
选择环境
配置
完成
自建环境
Docker
Kubernetes
1
2
下一步

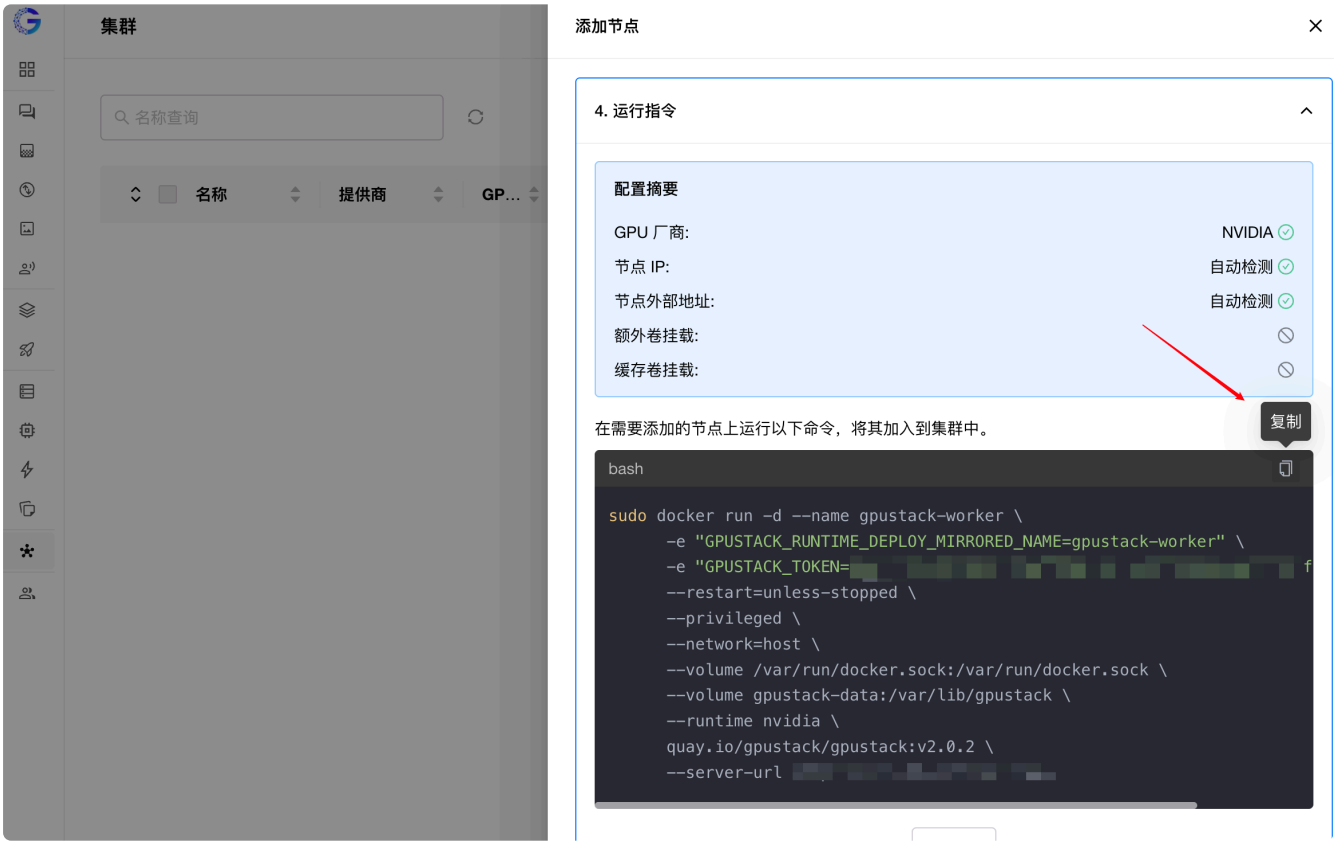

集群
名称查询
名称
提供商
GP...
创建集群
选择环境
配置
完成
名称 *
Cst-t
描述
高级
1
2
上一步
保 存

2.10 添加节点（Worker）。集群创建完毕后，需添加 Worker 节点至集群，操作如下。

2.10.1 取消手动配置节点 IP，选择自动检测节点 IP



2.10.2 复制服务 Worker 节点启动命令



若您使用公有云生态专区购买的 Worker 节点，需修改命令：

- 替换命令“quay.io/gpustack/gpustack:v2.0.2 \”为“royi-hub-registry-huadong1.crs-internal.ctyun.cn/app-images/gpustack/gpustack:v2.0.2 \”
- 在命令末尾直接增加如下配置：--system-default-container-registry royi-hub-registry-huadong1.crs-internal.ctyun.cn/app-images

示例如下：

Bash

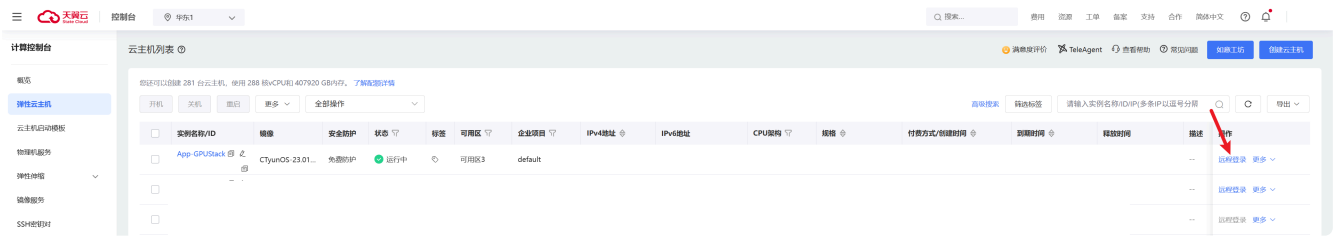
```
sudo docker run -d --name gpustack-worker \
      -e "GPUSTACK_RUNTIME_DEPLOY_MIRRORED_NAME=gpustack-worker" \
      -e "GPUSTACK_TOKEN=gpustack_XXXXXXXXX" \
      --restart=unless-stopped \
      --privileged \
      --network=host \
      --volume /var/run/docker.sock:/var/run/docker.sock \
      --volume gpustack-data:/var/lib/gpustack \
      --runtime nvidia \
      royi-hub-registry-huadong1.crs-internal.ctyun.cn/app-images/gpustack/gpustack:v2.0.2 \
      --server-url http://XXX.XXX.XXX.XXX:7080 --system-default-container-registry royi-hub-registry-huadong1.crs-internal.ctyun.cn/app-images
```

说明：

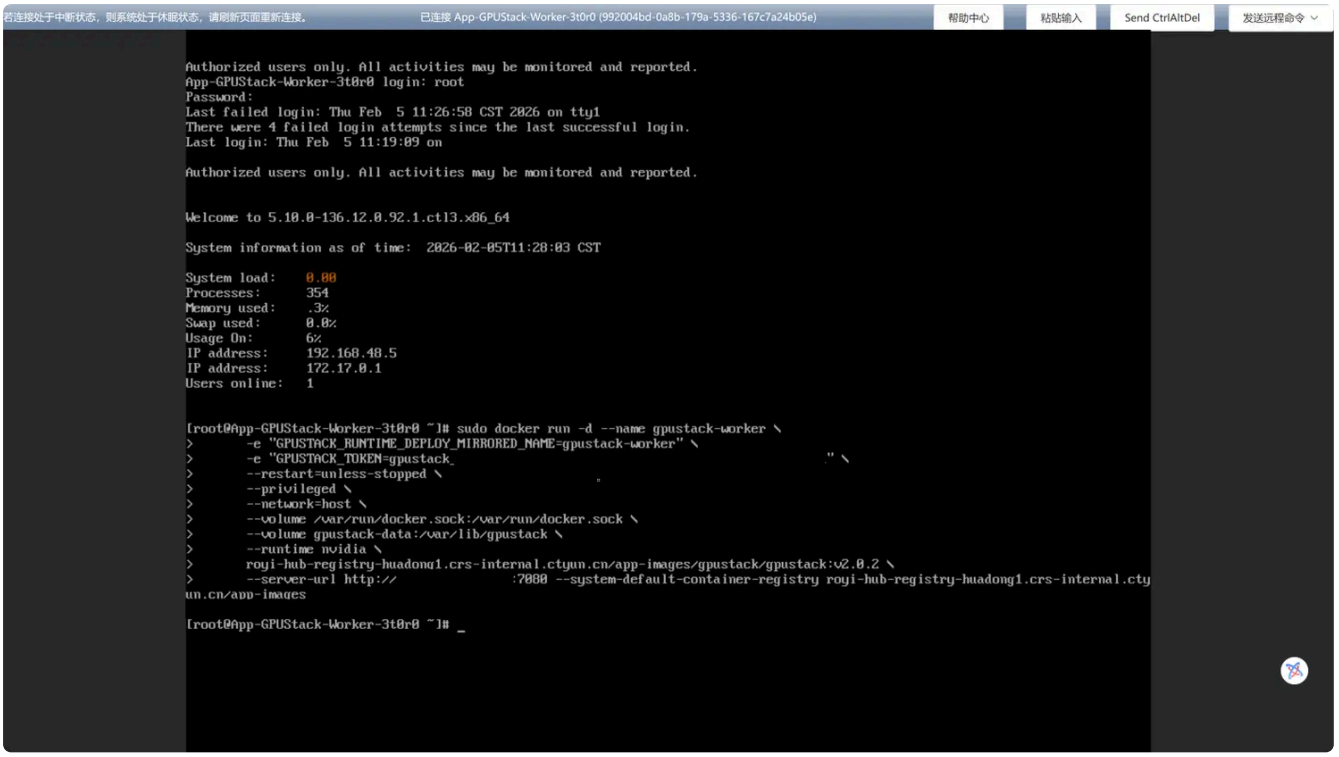
- 第 3 行中 TOKEN 信息由 GPUstack 生成，可复制 GPUstack 生成的命令中的 TOKEN 信息
- 11 行中，server-url http://XXX.XXX.XXX.XXX:7080 需替换为 Server 地址，即应用详情页中的应用入口地址

2.11 启动 Worker 服务。

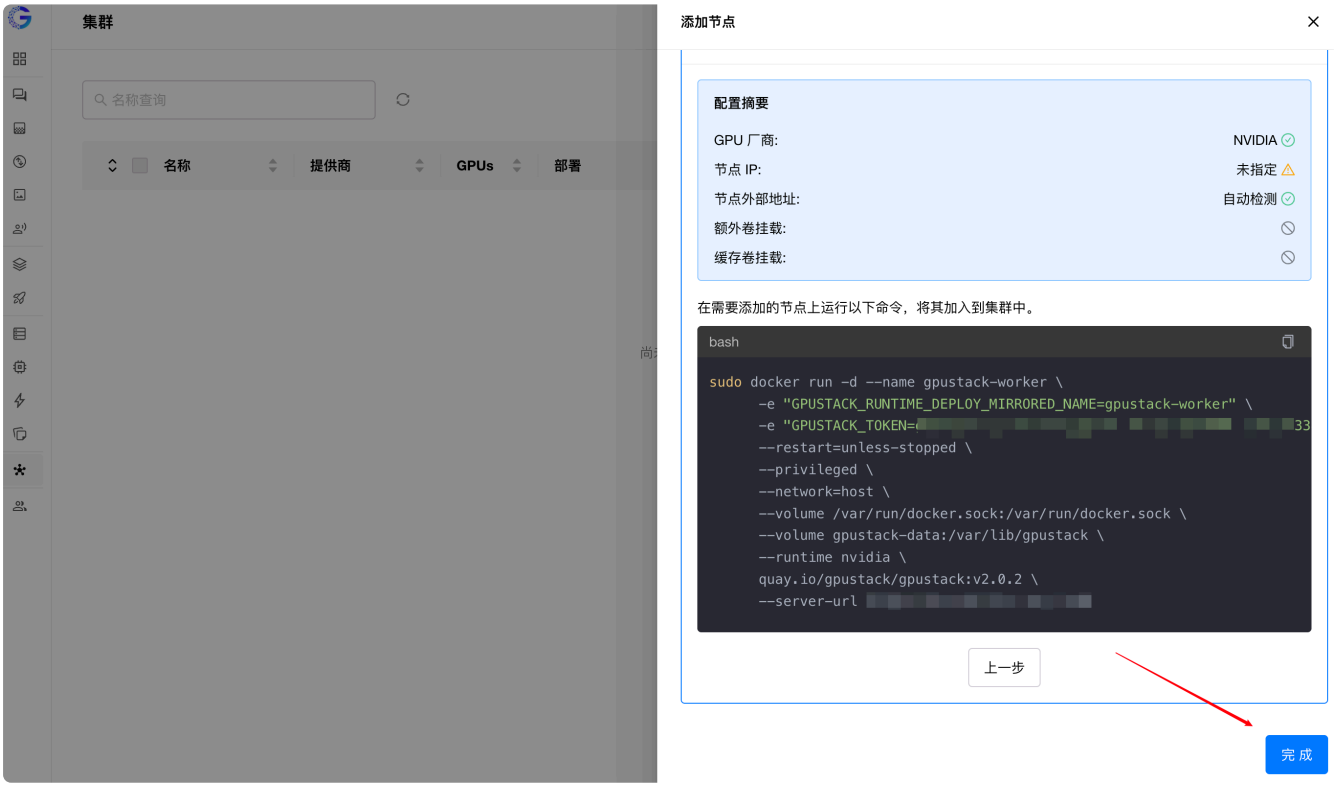
2.11.1 通过天翼云云主机控制台，使用“步骤 2.5”中设置的密码，登录 Worker 节点云主机。



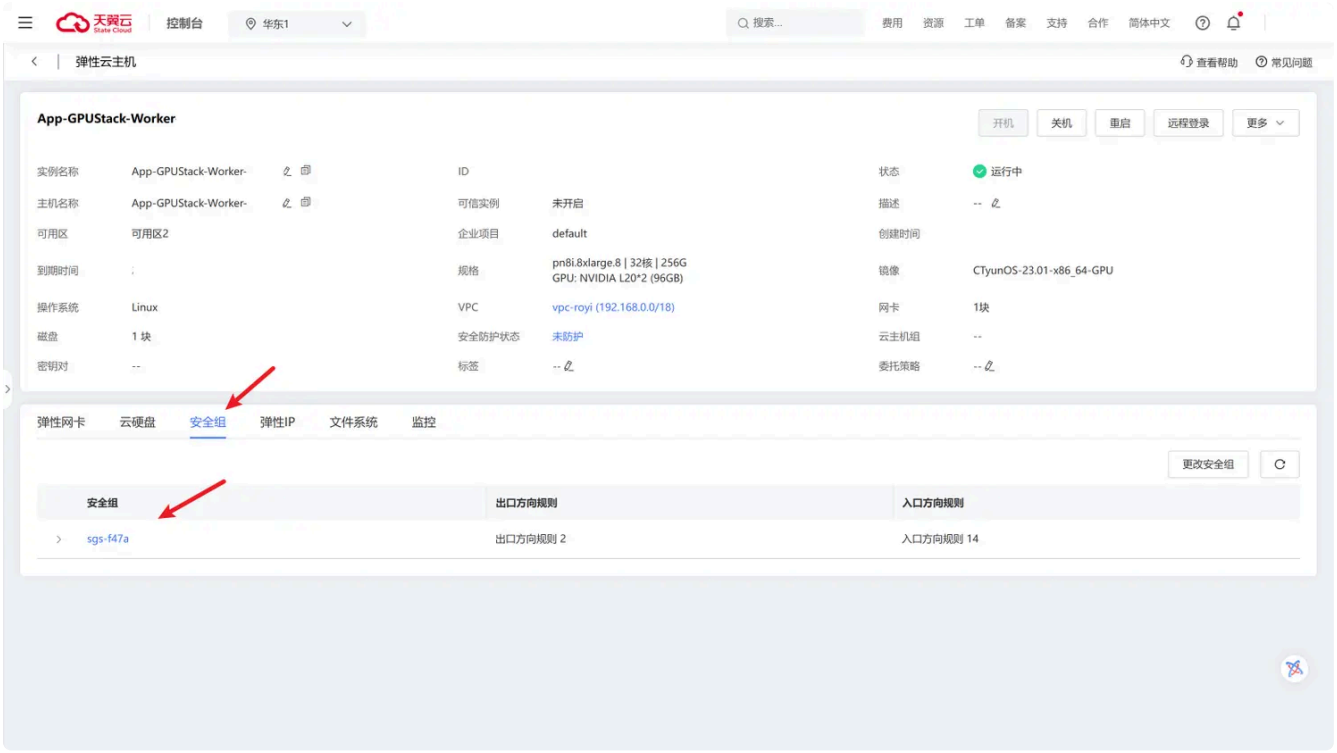
2.11.2 在命令行中执行步骤 2.10.2 复制的 Worker 节点启动命令。



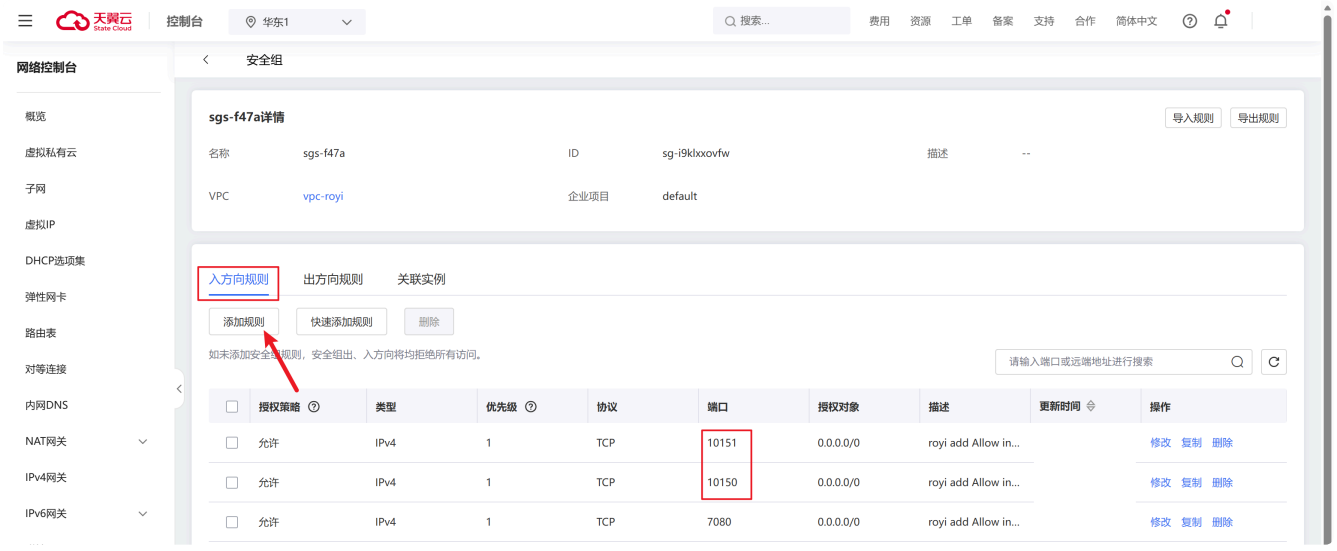
2.11.3 执行完成后，返回 GPUStack 应用 Web 页面点击完成，节点成功添加



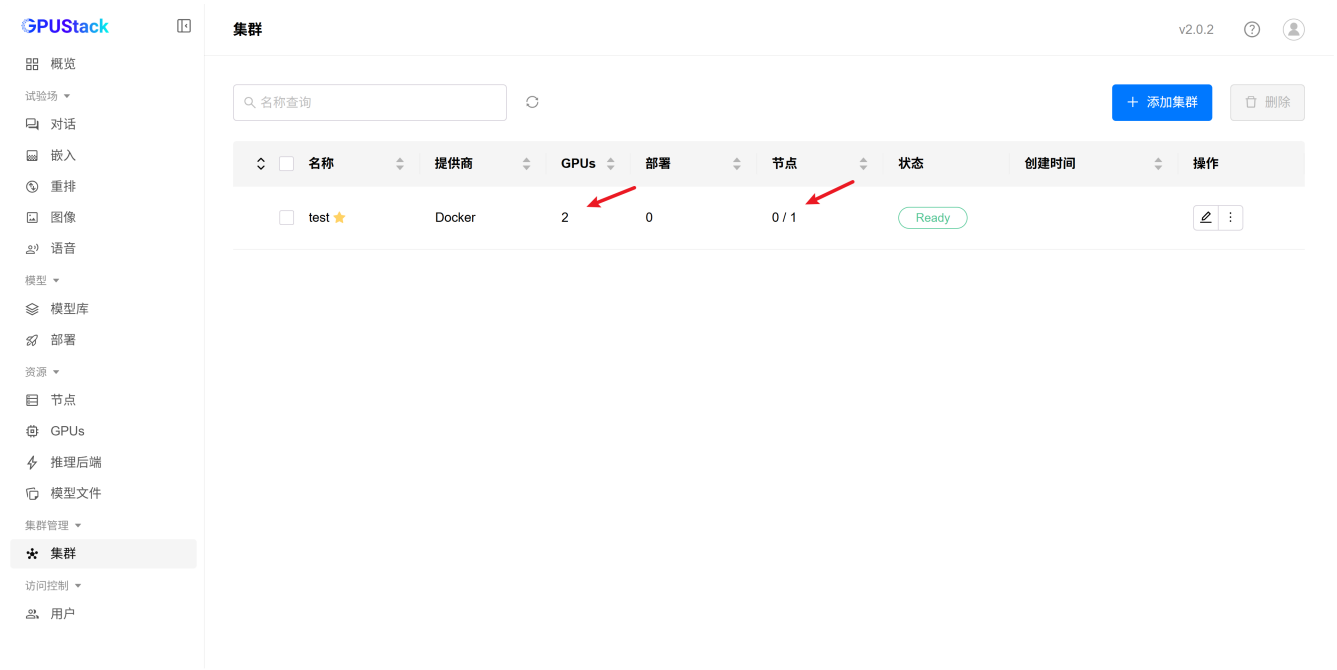
2.11.4 修改 Worker 节点安全组规则。通过云主机列表页，进入 Worker 节点云主机详情页，选中“安全组” Tab，点击安全组名称超链接。



2.11.5 开放 10050、10051 端口。在入方向规则 Tab 中，点击添加规则按钮，新增端口 10050 和 10051 允许策略。

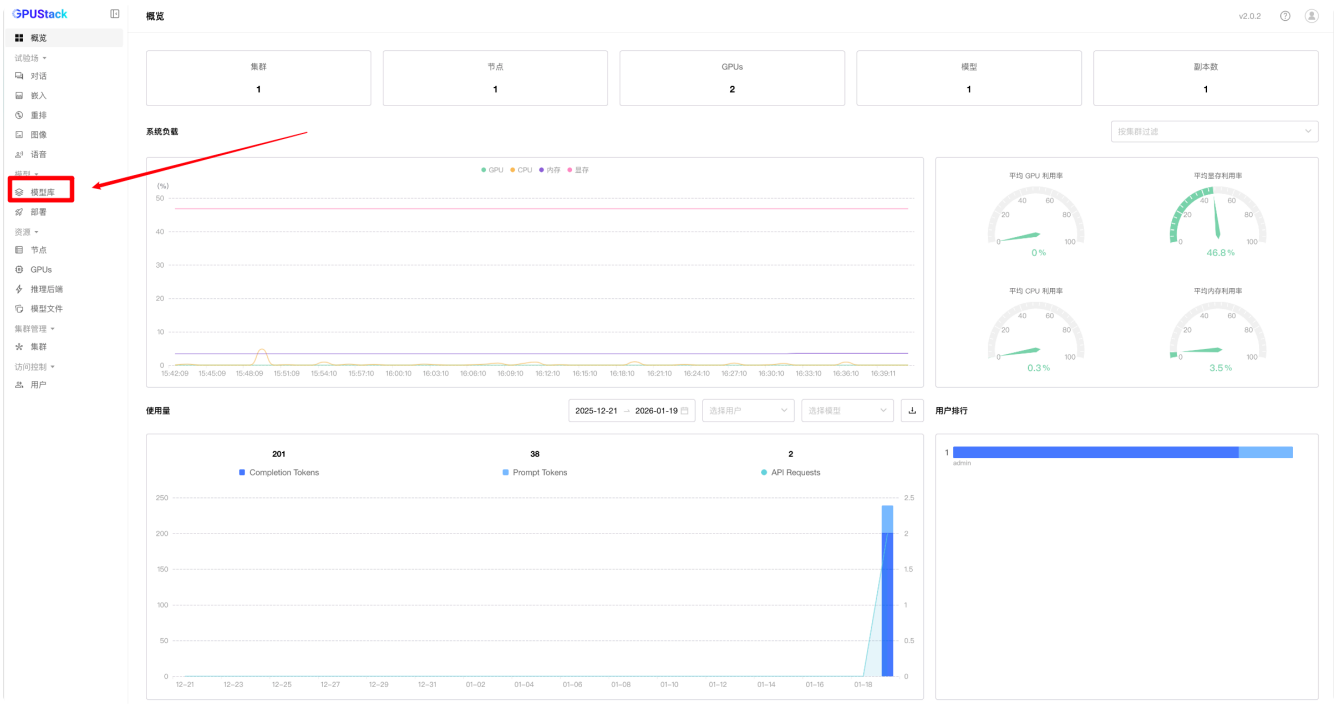


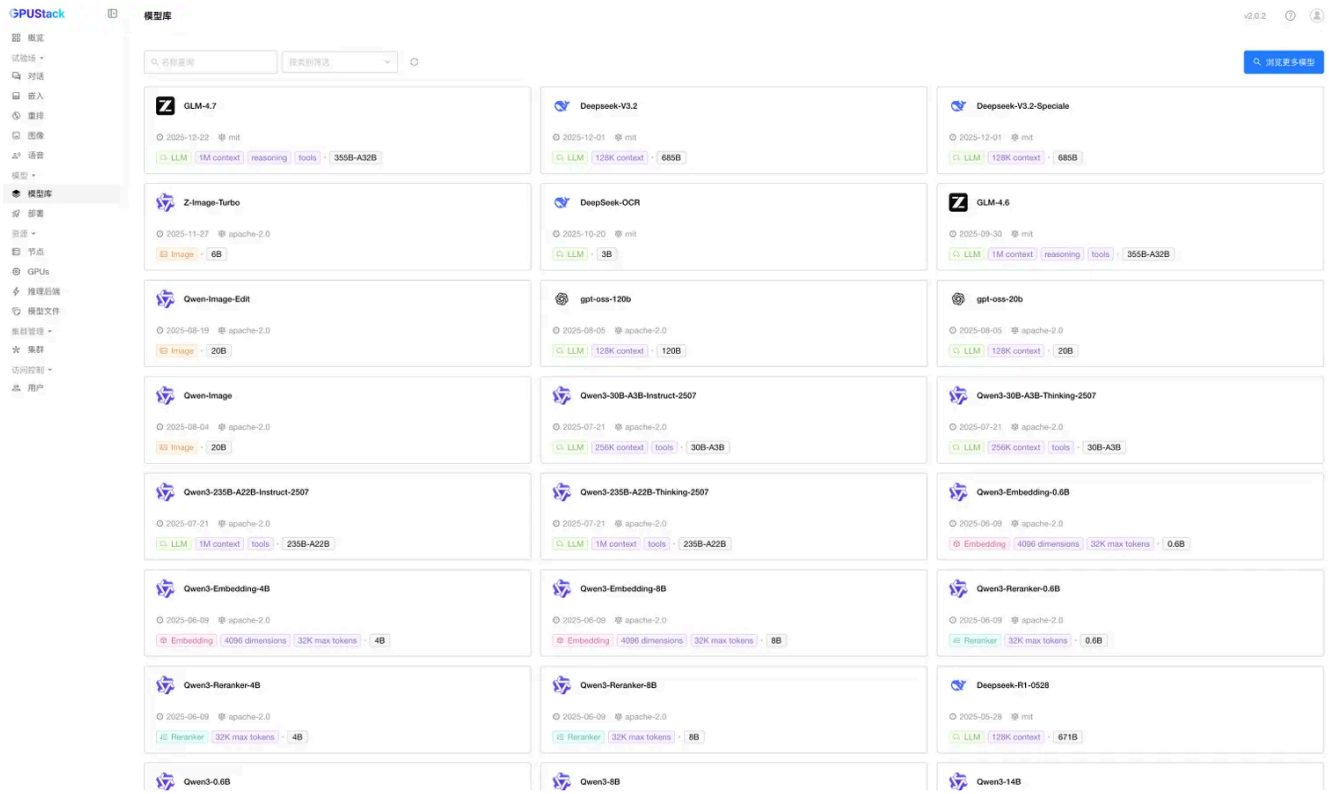
2.11.16 添加成功后，可在集群列表查看节点添加结果，节点列表中可查看 Worker 节点状态。

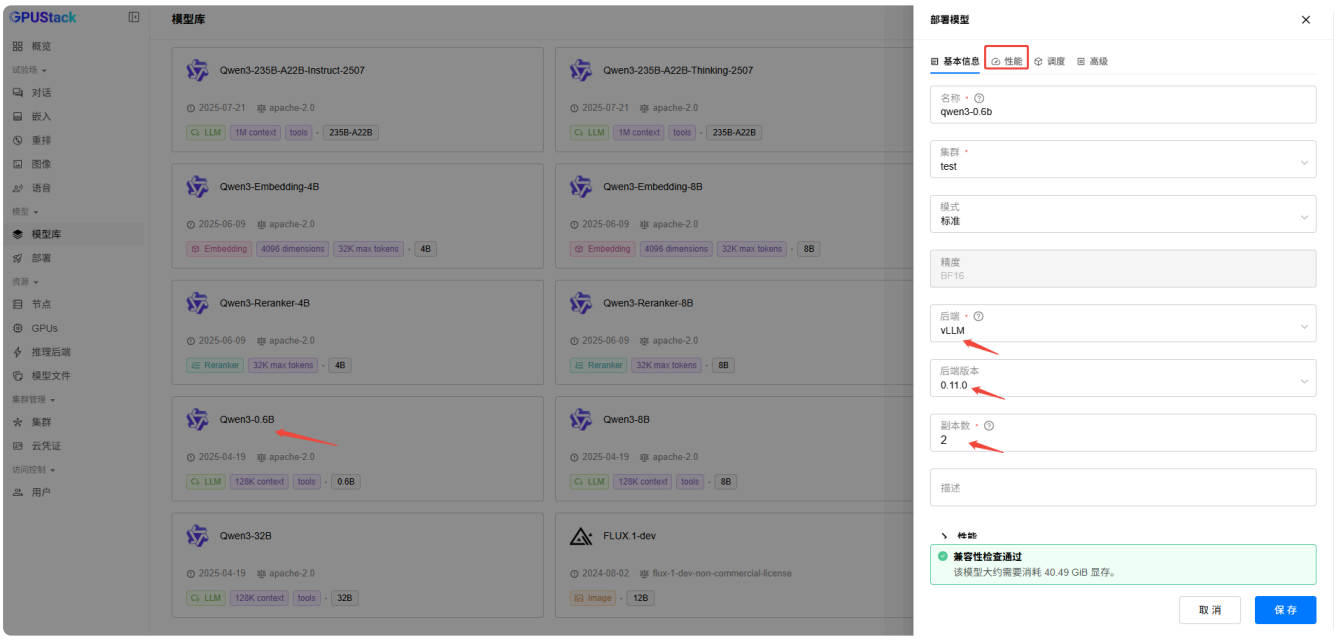


说明：如需添加多个 Worker 节点，在订购后，重复执行步骤 2.10、2.11 即可。

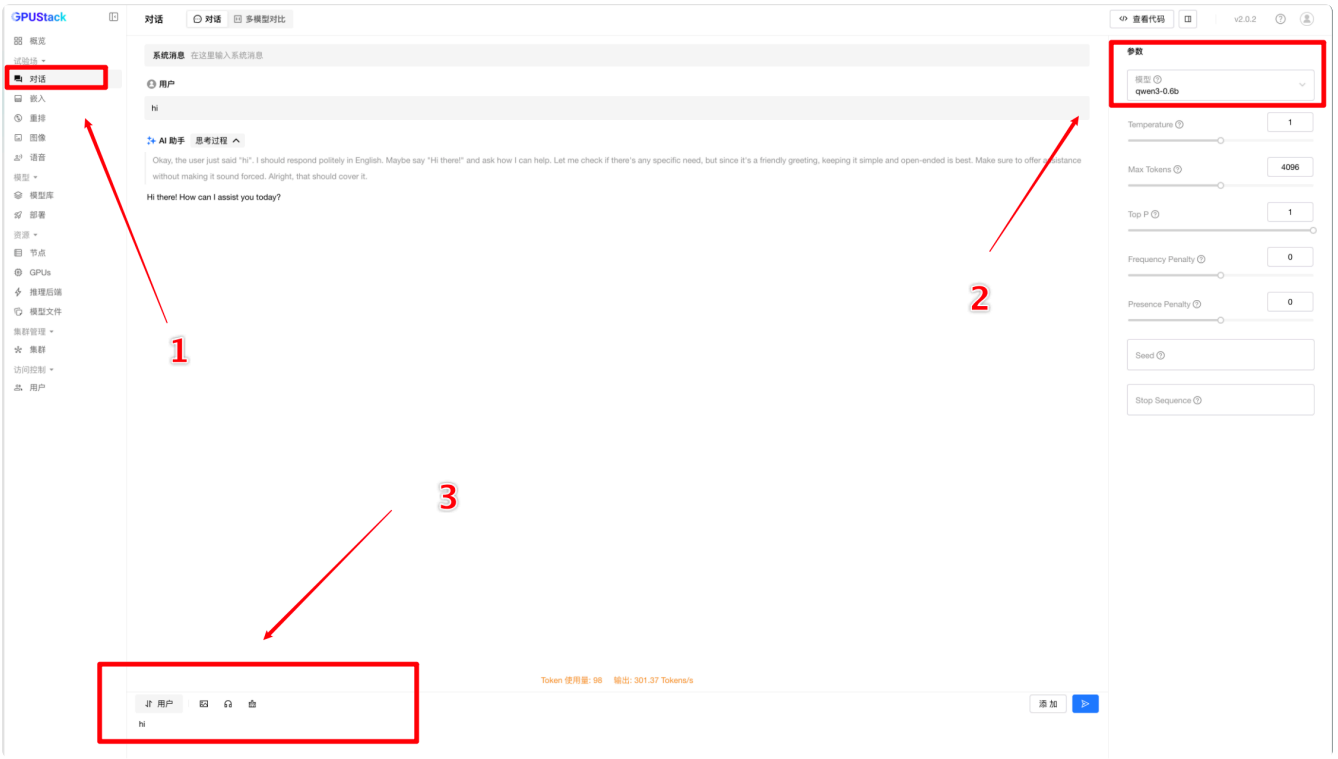
2.12 部署模型。进入模型库菜单，选择合适的模型开始部署。





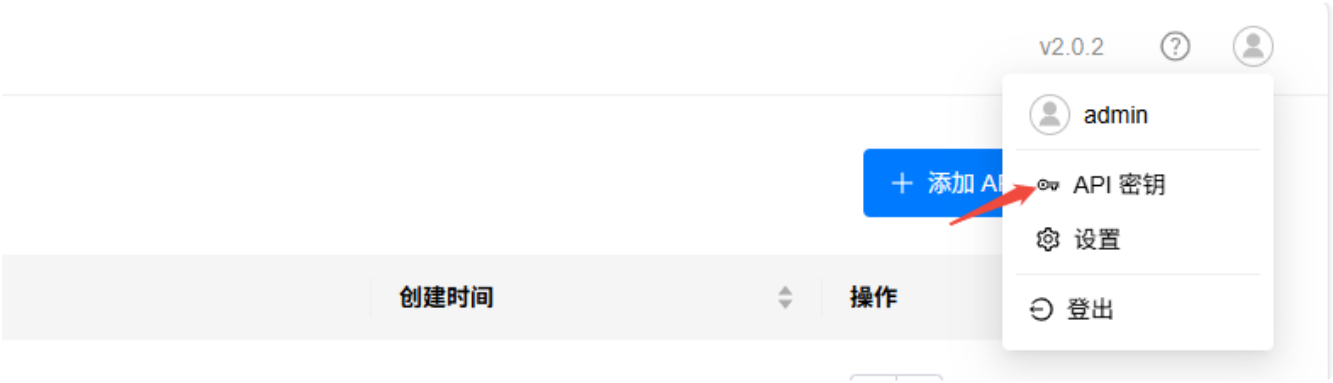


2.13 等待模型部署成功后，便可进入试验场，进行验证。



2.14 API 密钥管理

1. 将鼠标悬停在用户头像上，选择"`API 密钥`"
2. 点击"`创建密钥`"生成新密钥
3. **安全提示**：该密钥仅在创建时可见一次，请妥善保存



三、常见问题

Q1: 如何升级或降级已部署的模型版本?

A: GPUStack 支持无缝模型版本切换

1. 在模型详情页点击"编辑"按钮
2. 在"模型版本"下拉菜单中选择目标版本

3. 手动删除原来的示例，即可自动创建新的副本
4. 支持新旧副本共存，可以在用户无感知情况下切换版本

Q2: 本地模型文件如何导入 GPUStack?

A：挂载共享存储，需要在集群创建之前完成挂载并存储，否则 worker 节点无法访问

1. 根据页面提示在创建集群时，进行缓存卷挂载，会自动生成对应的 worker 指令
2. 此时在部署模型时，可以选择本地模型进行部署