

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；

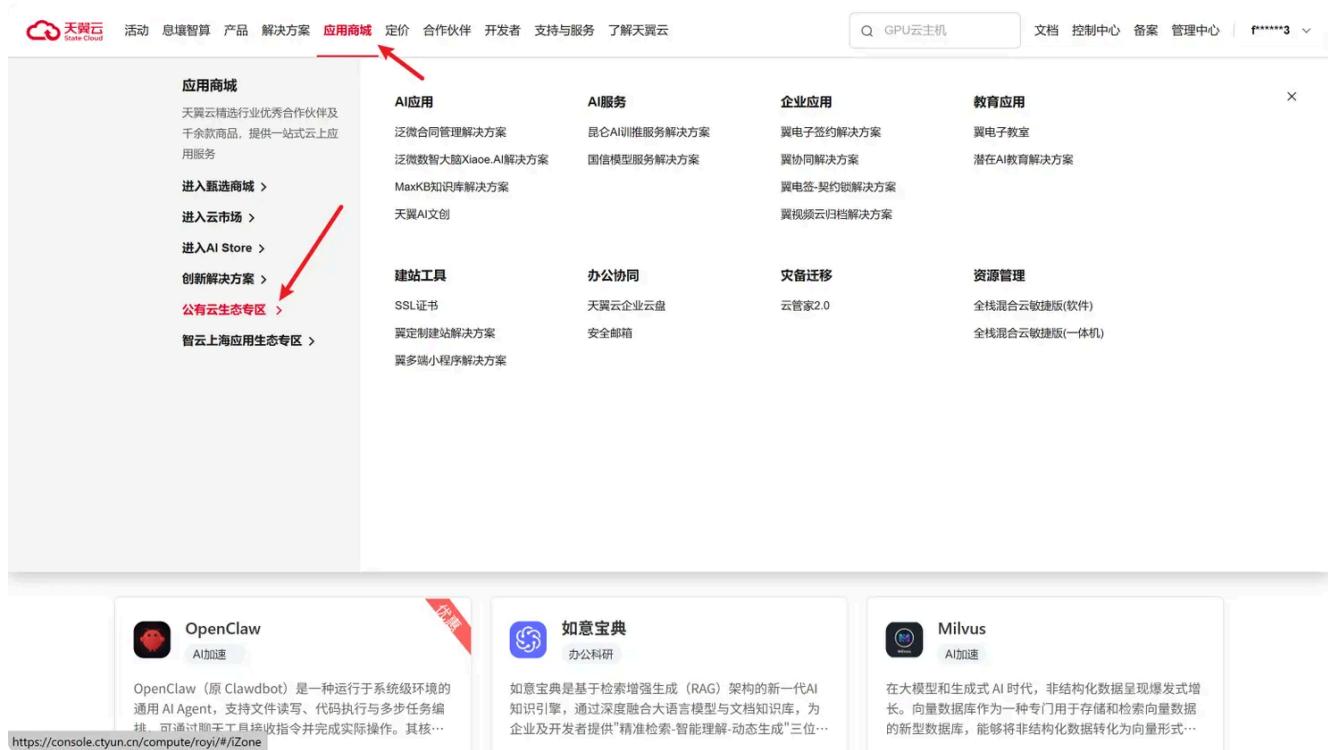
一键部署：支持从 Hugging Face、ModelScope、或本地直接部署，支持自动扩缩容、版本升级；

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。

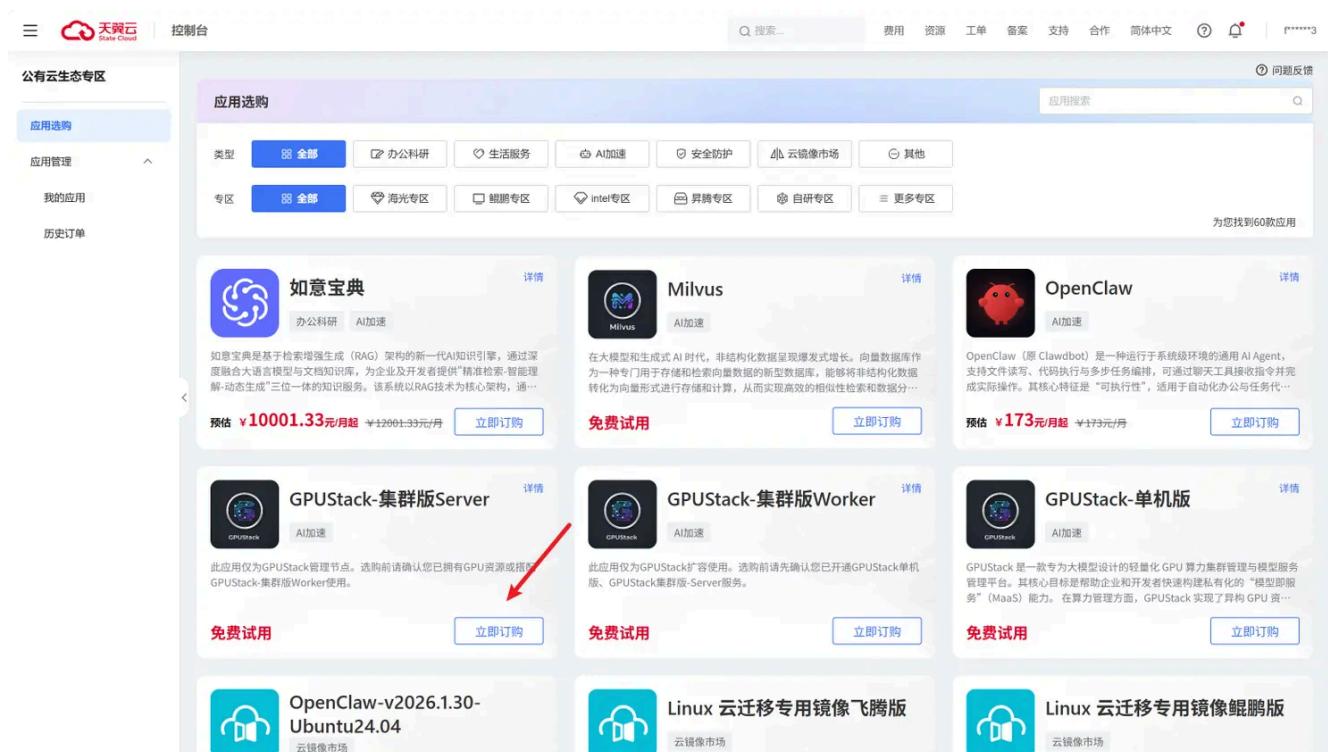
二、使用指南

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



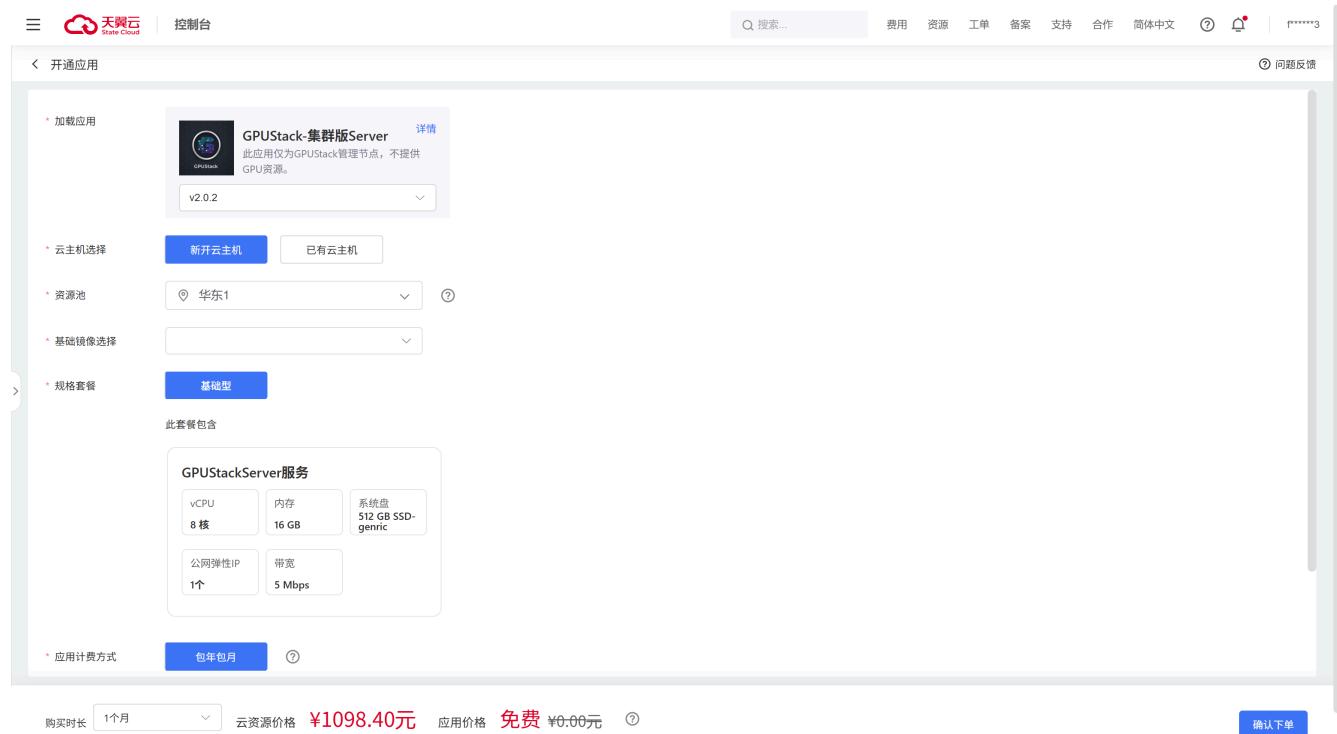
The screenshot shows the China Telecom Cloud Application Marketplace. At the top, there are several navigation tabs: 活动 (Activities), 息壤智算 (Xishang Intelligence Computing), 产品 (Products), 解决方案 (Solutions), 应用商城 (Application Marketplace), 定价 (Pricing), 合作伙伴 (Partners), 开发者 (Developers), 支持与服务 (Support and Services), and 了解天翼云 (Learn about China Telecom Cloud). The '应用商城' tab is highlighted with a red arrow. Below the tabs, there are several sections: '应用商城' (Application Marketplace), 'AI应用' (AI Applications), 'AI服务' (AI Services), '企业应用' (Enterprise Applications), '教育应用' (Education Applications), '建站工具' (Website Building Tools), '办公协同' (Office Collaboration), '灾备迁移' (Disaster Recovery Migration), and '资源管理' (Resource Management). A red arrow points to the '公有云生态专区' (Public Cloud Ecosystem Special Zone) link. At the bottom, there are three application cards: 'OpenClaw' (AI加速), '如意宝典' (Office Research), and 'Milvus' (AI加速). The 'GPUStack 集群版 Server' application is highlighted with a red arrow and a '立即选购' (Buy Now) button.

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



The screenshot shows the China Telecom Cloud Application Marketplace with the '应用选购' (Application Selection) tab selected. The 'GPUStack-集群版 Server' application is highlighted with a red arrow and a '立即订购' (Buy Now) button. The application details include a preview price of '预估 ¥10001.33元/月起' and a '立即订购' (Buy Now) button. Other applications listed include '如意宝典', 'Milvus', 'OpenClaw', 'GPUStack-集群版 Worker', 'GPUStack-单机版', 'OpenClaw-v2026.1.30-Ubuntu24.04', 'Linux 云迁移专用镜像飞腾版', and 'Linux 云迁移专用镜像鲲鹏版'. Each application has a '免费试用' (Free Trial) and '立即订购' (Buy Now) button.

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

The screenshot shows the Tianyun State Cloud application store interface. The top navigation bar includes the logo, a search bar, and links for Control Panel, Billing, Resources, Work Orders, Domains, Support, Cooperation, Simplified Chinese, and Help. A feedback link is also present. The main header is 'Public Cloud Application Zone' with a 'Problem Report' button. The left sidebar has sections for Application Selection, Application Management, My Applications, and History Orders. The main content area is titled 'Application Selection' with filters for Type (All, Office Research, Life Services, AI Acceleration, Security Protection, Cloud Image Market, Others) and Special Zone (All, Haigang Special Zone, Fengteng Special Zone, Intel Special Zone, Yiqing Special Zone, Self-developed Special Zone, More Special Zones). A search bar at the top right is labeled 'Application Search'. Below this, a message says 'We found 60 applications for you'. The page displays several service cards:

- 如意宝典** (如意宝典): AI Acceleration. Description:如意宝典是基于检索增强生成（RAG）架构的新一代AI知识引擎，通过深度融合大语言模型与文档知识库，为企业及开发者提供“精准检索-智能理解-动态生成”三位一体的知识服务。该系统以RAG技术为核心架构，通…
Prepaid: ¥10001.33/month. Instant Purchase button.
- GPUStack-集群版Worker** (GPUStack-Cluster Version Worker): AI Acceleration. Description:此应用仅为GPUStack扩容使用。选购前请先确认您已开通GPUStack单机版、GPUStack集群版-Server服务。
Free Trial button. Instant Purchase button.
- GPUStack-单机版** (GPUStack-Single Machine Version): AI Acceleration. Description:GPUStack 是一款专为大模型设计的轻量化 GPU 算力集群管理与模型服务管理平台。其核心目标是帮助企业和开发者快速构建私有化的“模型即服务”（MaaS）能力。在算力管理方面，GPUStack 实现了异构 GPU 资…
Free Trial button. Instant Purchase button.
- Milvus**: AI Acceleration. Description:在大模型和生成式 AI 时代，非结构化数据呈现爆发式增长。向量数据库作为一种专门用于存储和检索向量数据的新型数据库，能够将非结构化数据转化为向量形式进行存储和计算，从而实现高效的相似性检索和数据分…
Free Trial button. Instant Purchase button.
- OpenClaw**: AI Acceleration. Description:OpenClaw（原 Clawbot）是一种运行于系统级环境的通用 AI Agent，支持文件读写、代码执行与多步骤任务编排，可通过聊天工具接收指令并完成实际操作。其核心特征是“可执行性”，适用于自动化办公与任务代…
Prepaid: ¥173/month. Instant Purchase button.
- GPUStack-集群版Server** (GPUStack-Cluster Version Server): AI Acceleration. Description:此应用仅为GPUStack管理节点。选购前请确认您已拥有GPU资源或搭配GPUStack-集群版Worker使用。
Free Trial button. Instant Purchase button.
- OpenClaw-v2026.1.30-Ubuntu24.04**: Cloud Image Market. Description: Cloud Image Market.
Cloud Image Market button.
- Linux 云迁移专用镜像飞腾版**: Cloud Image Market. Description: Cloud Image Market.
Cloud Image Market button.
- Linux 云迁移专用镜像鲲鹏版**: Cloud Image Market. Description: Cloud Image Market.
Cloud Image Market button.

2.5 按提示订购“GPUStruct 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。

< 开通应用

* 加载应用

GPUStack-集群版Worker [详情](#)
此应用仅为GPUStack单机版、GPUStack集群版-Server扩容使用。

v2.0.2

* 云主机选择 [新开云主机](#) [已有云主机](#)

华东1

* 基础镜像选择 灵动-Skyreels大模型

* 规格套餐 基础型

此套餐包含

GPUStackWorker服务

vCPU 32 核	内存 256 GB	GPU型号 L20 * 2
显存 48 GB * 2	系统盘 512 GB SSD-generic	公网弹性IP 1个
带宽 5 Mbps		

* 应用计费方式 [包年包月](#) [?](#)

实例名称 App-GPUStack-Worker-m76ne 25 / 63 [?](#)

应用别名 GPUStack-集群版Worker-eslg 23 / 30 [?](#)

* 密码 [?](#)

购买时长 1个月 [?](#) 云资源价格 **¥15548.40元** 应用价格 **免费** **¥0.00元** [?](#)



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



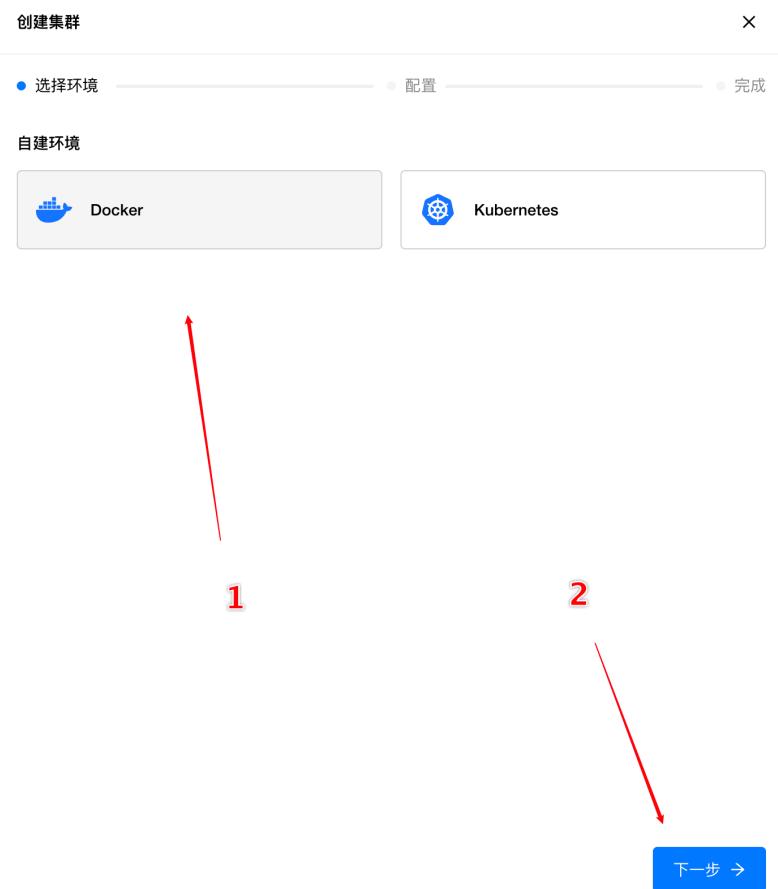
2.8 GPUstack 默认账号为 admin，密码为 gpustack。(可在应用中修改)

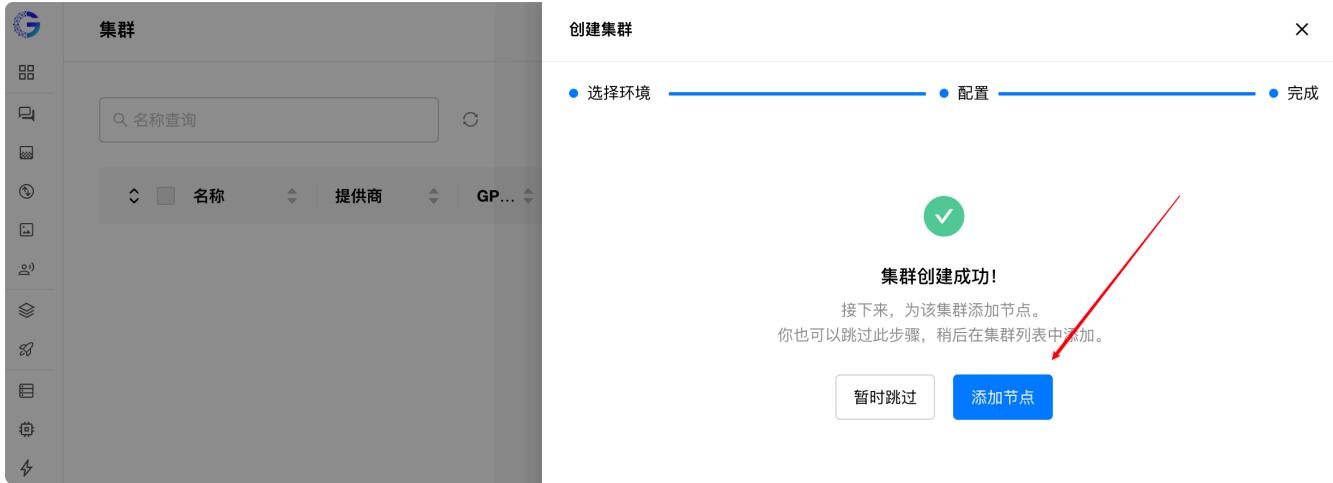


© 2026 软件名称 | 帮助 | v2.0.2

2.9 添加集群。参考以下操作，在 GPUStack 中创建新集群。

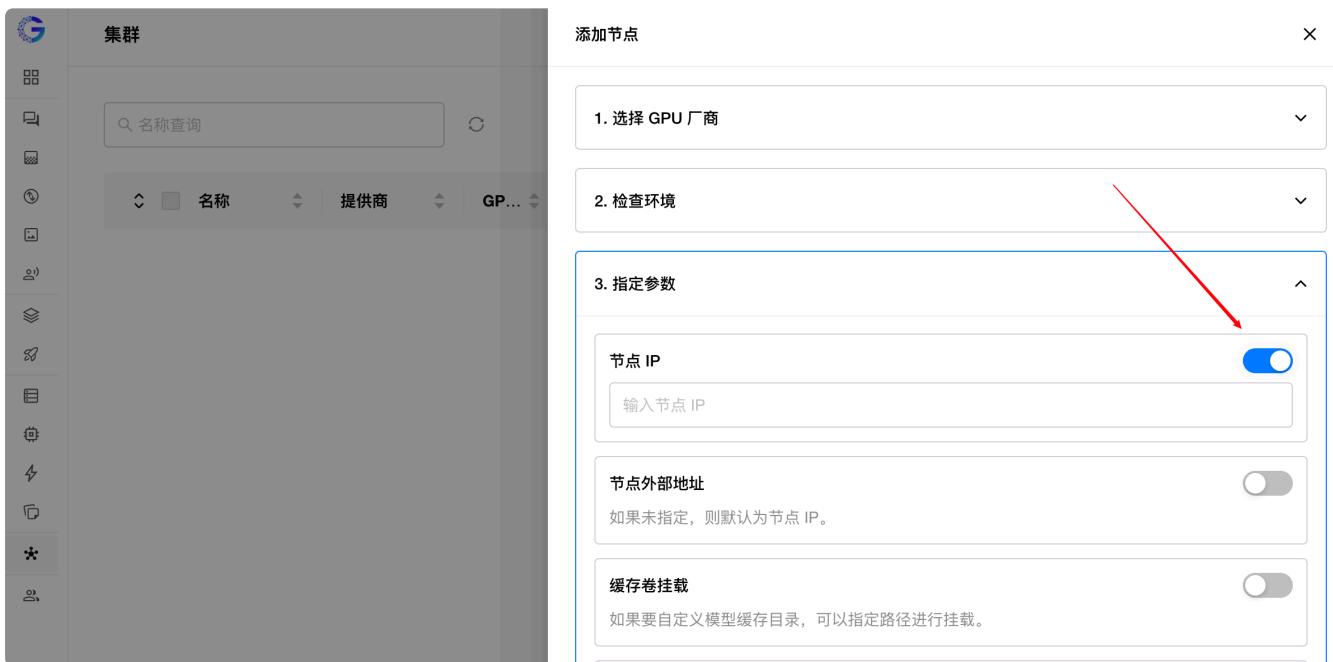
The image displays the GPUStack system load interface. At the top, there are five summary boxes: '集群' (Clusters) with 0, '节点' (Nodes) with 0, 'GPUs' with 0, '模型' (Models) with 0, and '副本数' (Replica Count) with 0. Below these is a '系统负载' (System Load) chart showing a single data series from 0 to 1. A modal window is overlaid on the chart. The modal has a light gray background and a white central area. It features a blue circular icon with a grid of four squares. The text '暂无集群' (No clusters available) is centered, followed by the sub-instruction '暂无可用集群, 请添加集群以开始使用。' (No available clusters, please add a cluster to start using it.). At the bottom of the modal are two buttons: a white button with black text '暂时跳过' (Skip for now) and a blue button with white text '创建您的第一个集群' (Create your first cluster). A red arrow points from the text '请添加集群以开始使用。' to the blue 'Create your first cluster' button.

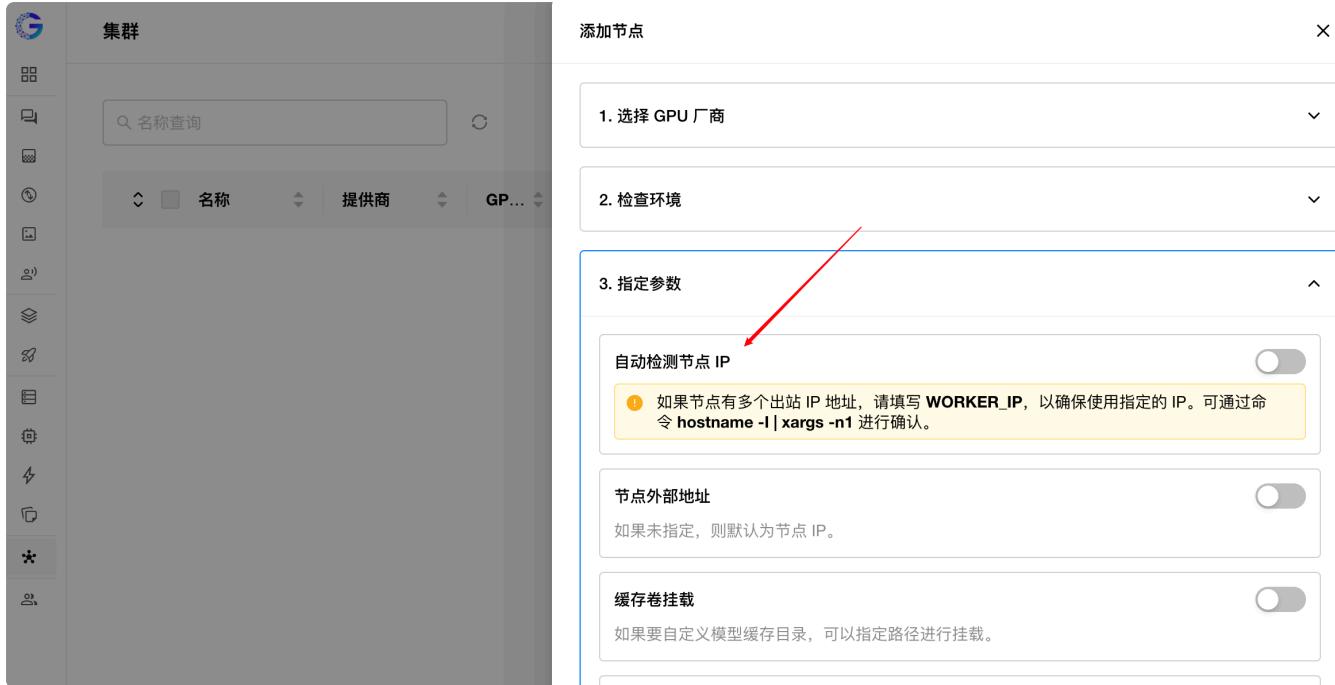




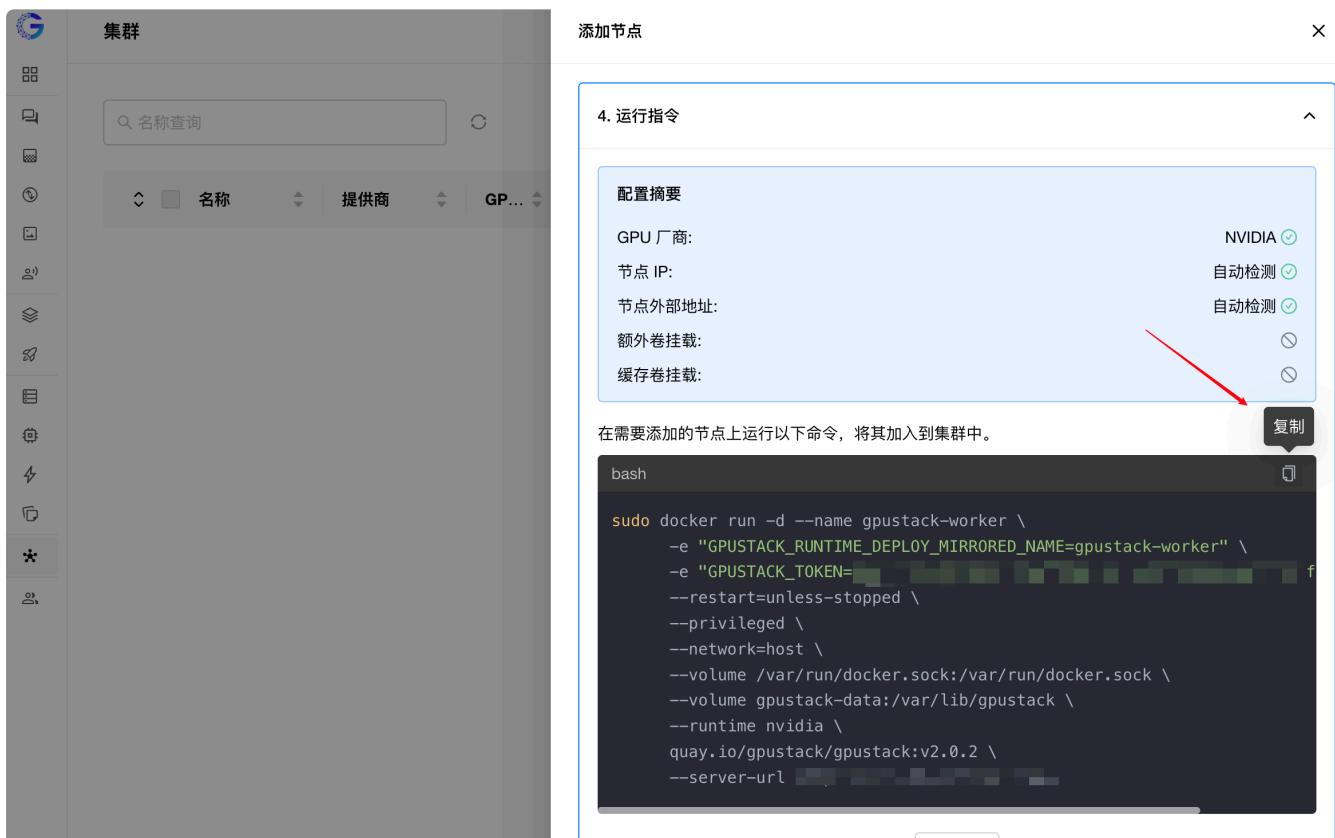
2.10 添加节点 (Worker)。集群创建完毕后，需添加 Worker 节点至集群，操作如下。

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

```
1 sudo docker run -d --name gpustack-worker \
2     -e "GPUSTACK_RUNTIME_DEPLOY_MIRRORED_NAME=gpustack-worker" \
3     -e "GPUSTACK_TOKEN=gpustack_XXXXXXXXXX" \
4     --restart=unless-stopped \
5     --privileged \
6     --network=host \
7     --volume /var/run/docker.sock:/var/run/docker.sock \
8     --volume gpustack-data:/var/lib/gpustack \
9     --runtime nvidia \
10    royi-hub-registry-huadong1.crs-internal.ctyun.cn/app-images/gpustack/gpustack:v2.0.2 \
11    --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。



The screenshot shows the Tianyi Cloud Control Panel interface. On the left, there is a sidebar with options like '计算控制台', '弹性云主机', '云主机启动模板', '物联网服务', '弹性伸缩', '镜像服务', and 'SSH密钥对'. The main area is titled '云主机列表' and shows a table of hosts. The table columns include '实例名称/ID', '镜像', '安全防护', '状态', '标签', '可用区', '企业项目', 'IPv4地址', 'IPv6地址', 'CPU架构', '核数', '付费方式/创建时间', '到期时间', '耗时时间', and '描述'. A red arrow points to the '操作' (Operation) column for the host 'App-GPustack'. The host details are: 实例名称/ID: App-GPustack, 镜像: CTyunOS 23.01..., 安全防护: 免费的IP, 状态: 运行中, 标签: 无, 可用区: default, 企业项目: 无, IPv4地址: 10.10.10.10, IPv6地址: 无, CPU架构: x86_64, 核数: 2, 付费方式/创建时间: 后付费/2024-01-15 10:00:00, 到期时间: 2024-01-15 10:00:00, 耗时时间: 00:00:00, 描述: 无。操作列显示了三个选项：远程登录、更多、和一个带有更多选项的按钮。

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。

若连接处于中断状态，则系统处于休眠状态，请刷新页面重新连接。 已连接 App-GPUStack-Worker-3t0r0 (992004bd-0a8b-179a-5336-167c7a24b05e) 帮助中心 粘贴输入 Send CtrlAltDel 发送远程命令 ▾

```

Authorized users only. All activities may be monitored and reported.
App-GPUStack-Worker-3t0r0 login: root
Password:
Last failed login: Thu Feb  5 11:26:58 CST 2026 on ttym
There were 4 failed login attempts since the last successful login.
Last login: Thu Feb  5 11:19:09 on

Authorized users only. All activities may be monitored and reported.

Welcome to 5.10.0-136.12.0.92.1.ctl3.x86_64
System information as of time: 2026-02-05T11:28:03 CST

System load: 0.00
Processes: 354
Memory used: .32
Swap used: 0.02
Usage on: 62
IP address: 192.168.48.5
IP address: 172.17.0.1
Users online: 1

[root@App-GPUStack-Worker-3t0r0 ~]# sudo docker run -d --name gpustack-worker \
> -e "GPUSTACK_RUNTIME_DEPLOY_MIRRORED_NAME=gpustack-worker" \
> -e "GPUSTACK_TOKEN=gpustack_3t0r0" \
> --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://:7000 --system-default-container-registry royi-hub-registry-huadong1.crs-internal.ctyun.cn/app-images

[root@App-GPUStack-Worker-3t0r0 ~]# _
```



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

集群

添加节点

配置摘要

GPU 厂商: NVIDIA (checked)

节点 IP: 未指定 (yellow triangle)

节点外部地址: 自动检测 (green circle)

额外卷挂载: (empty)

缓存卷挂载: (empty)

在需要添加的节点上运行以下命令, 将其加入到集群中。

```

bash

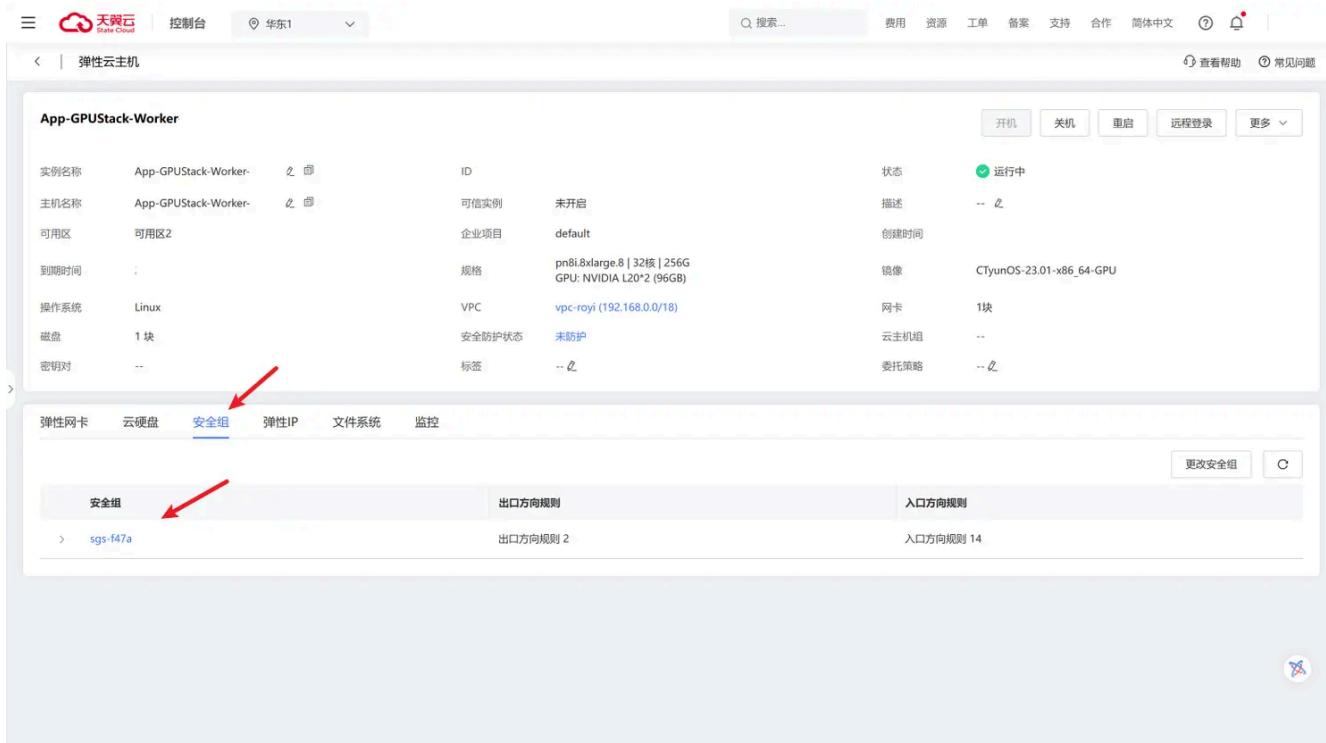
sudo docker run -d --name gpustack-worker \
-e "GPUSTACK_RUNTIME_DEPLOY_MIRRORED_NAME=gpustack-worker" \
-e "GPUSTACK_TOKEN=gpustack_3t0r0" \
--restart=unless-stopped \
--privileged \
--network=host \
--volume /var/run/docker.sock:/var/run/docker.sock \
--volume gpustack-data:/var/lib/gpustack \
--runtime nvidia \
quay.io/gpustack/gpustack:v2.0.2 \
--server-url http://:7000 --system-default-container-registry royi-hub-registry-huadong1.crs-internal.ctyun.cn/app-images

```

上一步

完成

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。



App-GPUStack-Worker

实例名称	App-GPUStack-Worker-1	ID	状态
主机名称	App-GPUStack-Worker-1	可信实例	未开启
可用区	可用区2	企业项目	default
到期时间	...	规格	pn8i.8xlarge.8 32核 256G GPU: NVIDIA L20*2 (96GB)
操作系统	Linux	VPC	vpc-royi (192.168.0.1/16)
磁盘	1 块	安全防护状态	未防护
密钥对	--	标签	--
			状态

弹性网卡 云硬盘 安全组 弹性IP 文件系统 监控

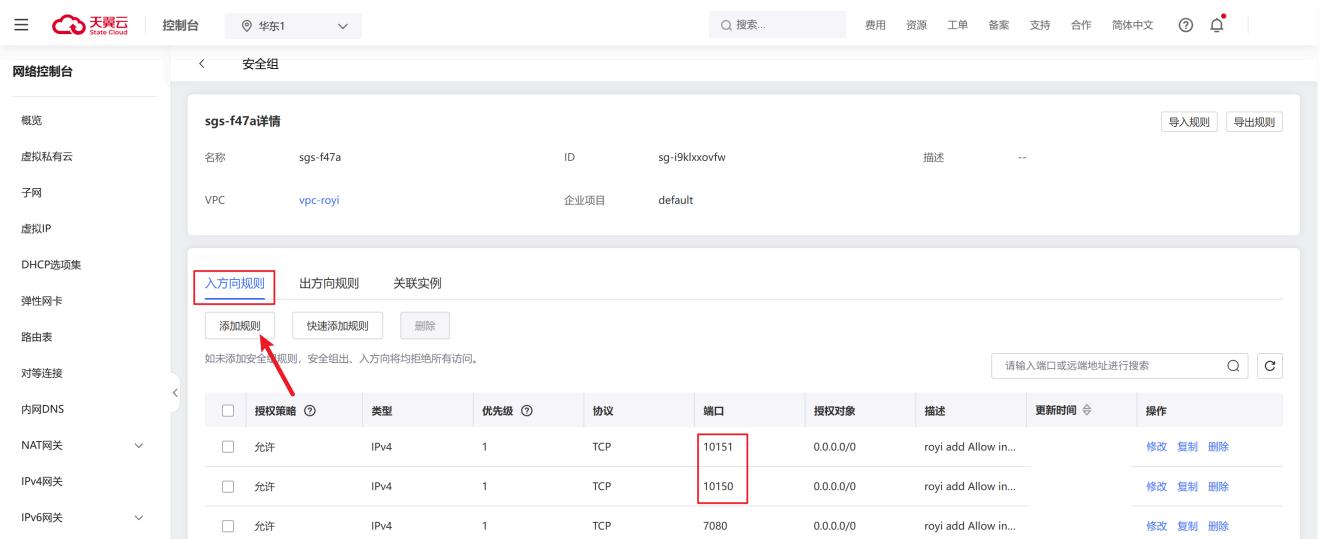
安全组

sgs-f47a

出口方向规则 入口方向规则

出口方向规则 2 入口方向规则 14

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。



sgs-f47a 详情

入方向规则

授权策略	类型	优先级	协议	端口	授权对象	描述	更新时间	操作
允许	IPv4	1	TCP	10151	0.0.0.0/0	royi add Allow in...	2023-07-10 10:00:00	修改 复制 删除
允许	IPv4	1	TCP	10150	0.0.0.0/0	royi add Allow in...	2023-07-10 10:00:00	修改 复制 删除
允许	IPv4	1	TCP	7080	0.0.0.0/0	royi add Allow in...	2023-07-10 10:00:00	修改 复制 删除

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。

GPUStack 集群

名称: test ★ 提供商: Docker GPU: 2 部署: 0 节点: 0 / 1 状态: Ready 操作

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。

GPUStack 概览

集群: 1 节点: 1 GPU: 2 模型: 1 副本数: 1

系统负载 (%)

平均 GPU 利用率: 0 % 平均 CPU 利用率: 0.3 % 平均 内存 利用率: 46.8 % 平均 显存 利用率: 3.5 %

使用量

201 Completion Tokens 38 Prompt Tokens 2 API Requests

用户排行

The screenshot shows the GPUStack interface with a sidebar on the left containing navigation links like '模型库', '部署', '资源', '节点', 'GPUS', '推理后端', '模型文件', '集群管理', '云凭证', '访问控制', and '用户'. The main area displays a grid of model cards. Each card includes a thumbnail, the model name, a brief description, and configuration details. A red arrow points to the 'Qwen3-0.6b' card. The right side of the interface features a detailed configuration panel for 'Qwen3-0.6b'. The '基本信息' tab is selected. The configuration fields include:

- 名称: Qwen3-0.6b
- 集群: test
- 模式: 标准
- 精度: BF16
- 后端: VLLM (highlighted with a red arrow)
- 后端版本: 0.11.0 (highlighted with a red arrow)
- 副本数: 2 (highlighted with a red arrow)
- 描述: (empty)
- 性能: (green box) 兼容性检查通过 (Compatibility check passed). The text states: 该模型大约需要消耗 40.49 GB 显存. (The model requires approximately 40.49 GB of GPU memory.)

2.13 等待模型部署成功后，便可进入试验场，进行验证。

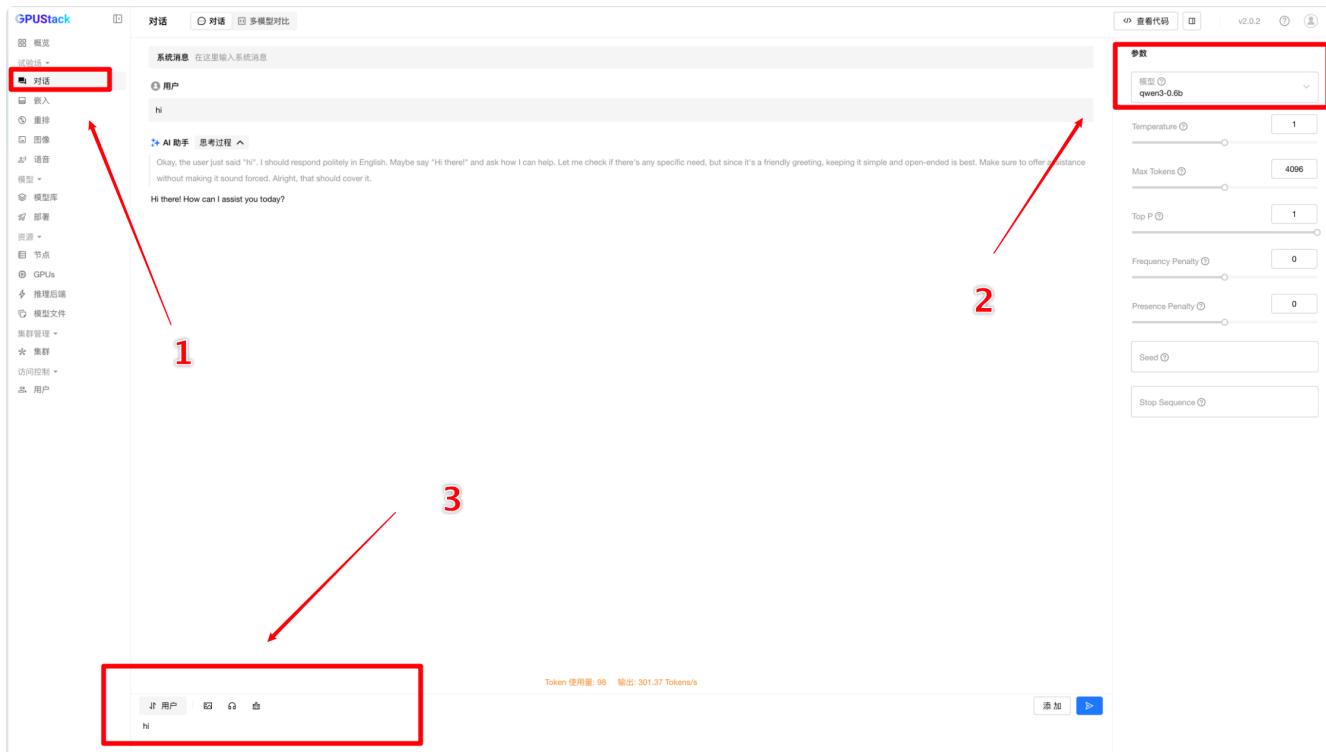
GPUStack

部署

v2.0.2

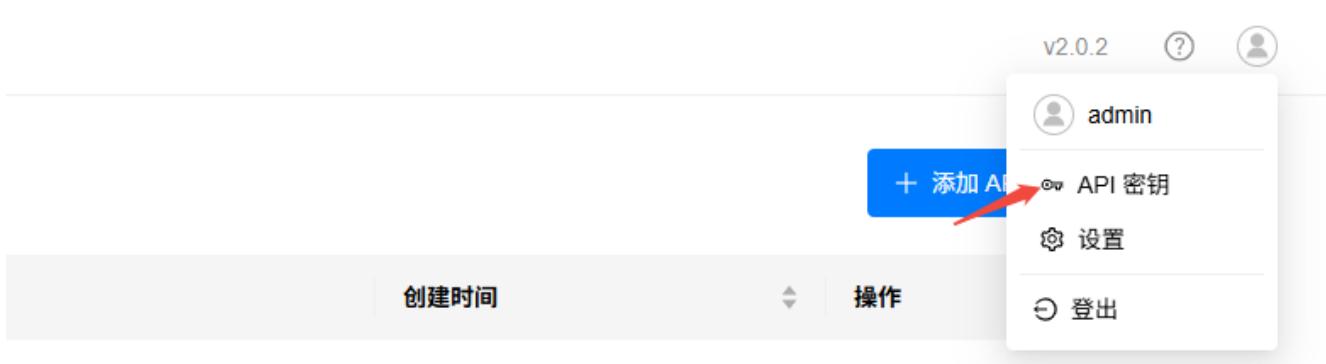
部署模型 启动

名称	集群	来源	副本数	创建时间	操作
qwen3-0.6b	test	ModelScope/Qwen/Qwen3-0.6B	1 / 1	2026-01-08 17:30:27	
qwen3-0.6b-llm			Running	2026-01-08 17:33:54	



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