

GPUStack- 单机版 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's a navigation bar with links like '活动', '息壤智算', '产品', '解决方案', '应用商城' (highlighted with a red arrow), '定价', '合作伙伴', '开发者', '支持与服务', and '了解天翼云'. Below the navigation is a grid of application categories: AI应用, AI服务, 企业应用, 教育应用; 建站工具, 办公协同, 灾备迁移, 资源管理. On the left sidebar, there are links: '进入甄选商城 >', '进入云市场 >', '进入AI Store >', '创新解决方案 >', '公有云生态专区 >' (highlighted with a red arrow), and '智云上海应用生态专区 >'. In the center, there are three application cards: 'OpenClaw' (AI加速), '如意宝典' (办公科研), and 'Milvus' (AI加速). Each card has a '立即订购' button.

2.2 选择“GPUMStack 单机版”，点击立即订购

This screenshot shows the 'GPUMStack-单机版' application card within the '应用选购' section of the marketplace. The card includes a preview image, the name '如意宝典', a brief description, a price of '预估 ¥10001.33元/月起', and two buttons: '立即订购' and '免费试用'. Below this card, there are other application cards for 'GPUStack-集群版Worker', 'GPUStack-单机版' (with a red arrow pointing to its '立即订购' button), 'Milvus', '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'. The page header also features a '立即订购' button.

2.3 按提示订购“GPUMstack 单机版”，点击确认下单即可成功购买服务。

The screenshot shows the Tianyi Cloud Control Panel interface for purchasing a GPUStack service. The top navigation bar includes links for '搜索...', '费用', '资源', '工单', '备案', '支持', '合作', '简体中文', and user account information. The main form is titled '开通应用' (Launch Application) and contains fields for '资源池' (Region: 华东1), '基础镜像选择' (Image Selection: 灵动-Skyreels大模型), and '规格套餐' (Specification Package: 基础型). Below these are sections for '此套餐包含' (What's included in this package) and 'GPUStack服务' (GPUStack Service), which details configurations like 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), and 带宽 (5 Mbps). Further down are fields for '应用计费方式' (Billing Method: 包年包月), '实例名称' (Instance Name: App-GPUMstack-Standalone-nlv91), '应用别名' (Alias: GPUStack-单机版-requ), and '密码' (Password: masked). At the bottom, it shows '购实时长' (Purchase Duration: 1个月), '云资源价格' (Cloud Resource Price: ￥15548.40元), '应用价格' (Application Price: 免费 ￥0.00元), and a '确认下单' (Confirm Purchase) button.

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

The screenshot shows the 'My Applications' page in the Tianyi Cloud Control Panel. The left sidebar includes '公有云生态专区' (Public Cloud Ecosystem Special Zone), '应用选购' (Application Purchase), '应用管理' (Application Management), '我的应用' (My Applications, highlighted in blue), and '历史订单' (Order History). The main area is titled '我的应用' (My Applications) and features filters for '类型' (Type: 全部, 办公科研, 生活服务, AI加速, 安全防护, 其他) and '专区' (Special Zone: 全部, 海光专区, 鲲鹏专区, intel专区, 昇腾专区, 自研专区, 更多专区). Two application cards are displayed: 'OpenClaw' (status: 运行中, 'Running') and 'GPUStack-单机版' (status: 运行中, 'Running'). A red arrow points to the 'GPUStack-单机版' card. Both cards include a '查看应用' (View Application) button and a '续费' (Renew) link.

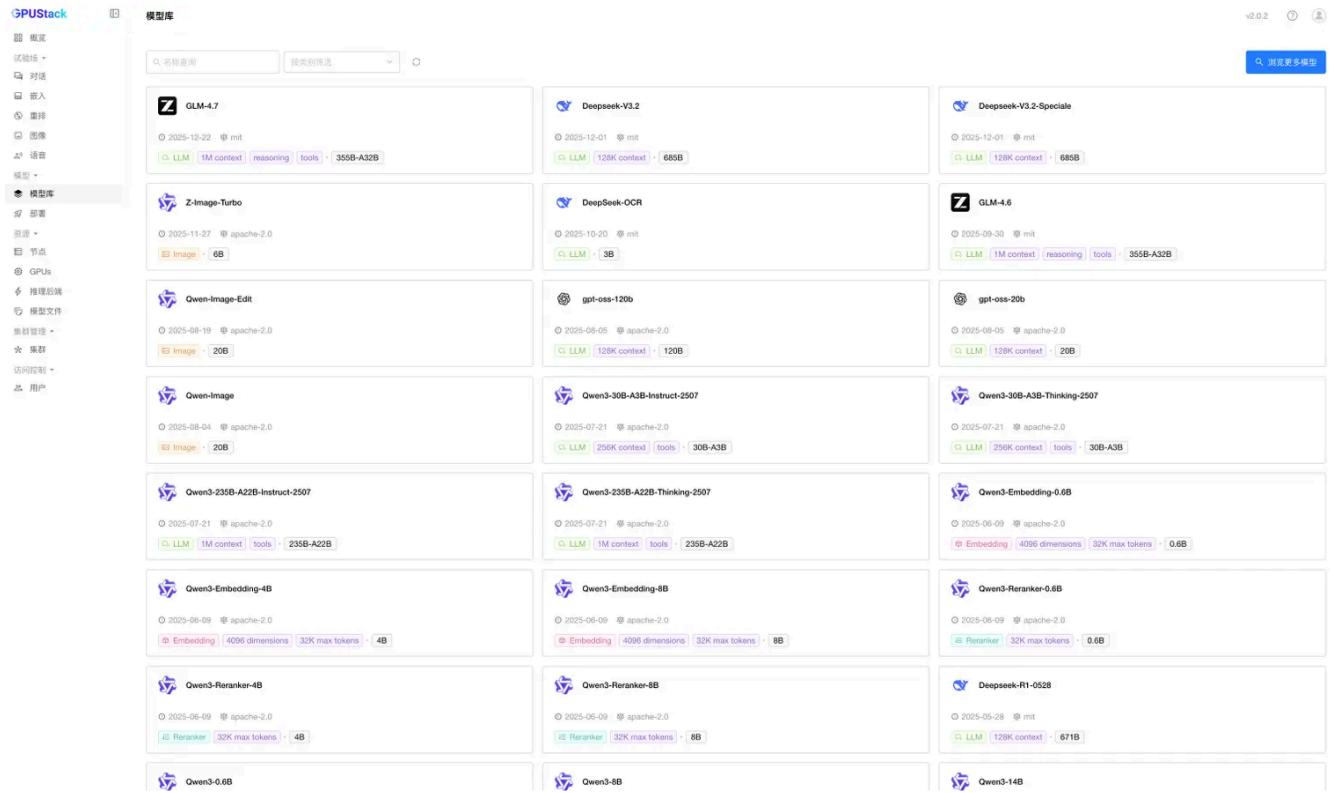
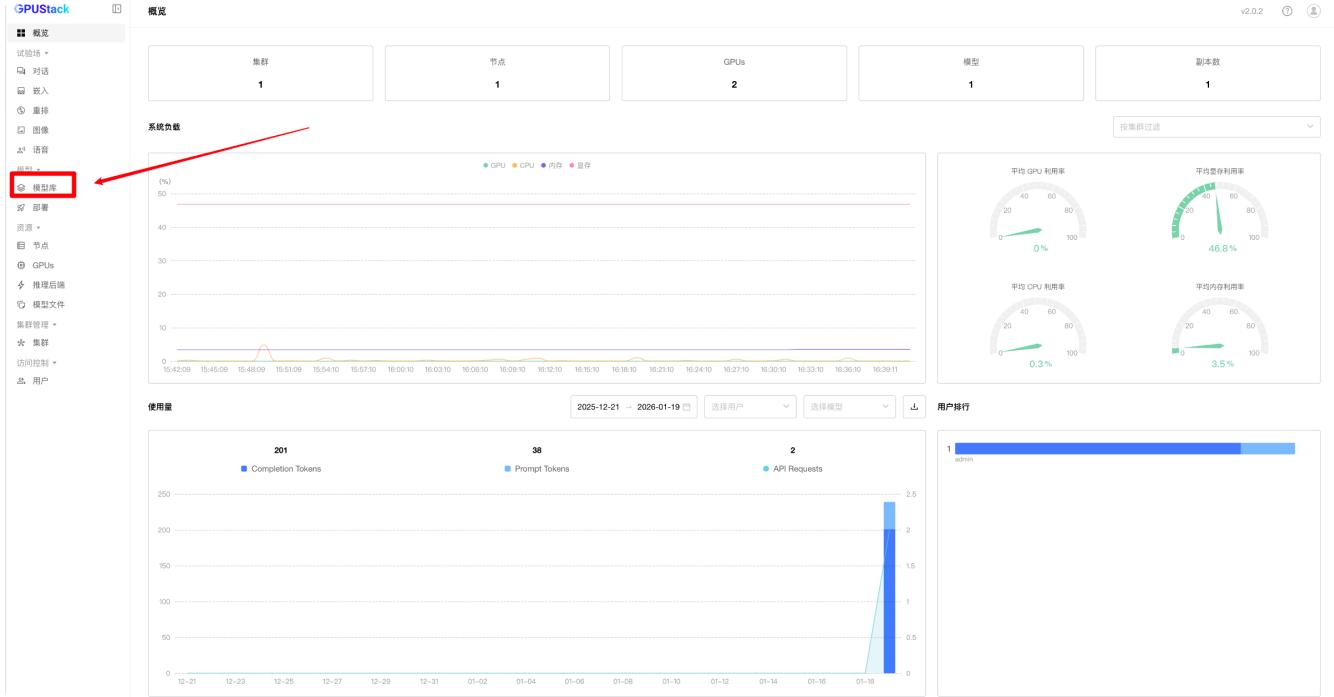
2.5 点击查看应用按钮，点击【立即使用】或在浏览器中输入复制的应用入口，即可访问 GPUstack

The screenshot shows the Tianyun Cloud Control Panel interface. In the center, there's a card for the 'GPUStack-单机版' application. At the top right of this card, there are three buttons: '退订' (Cancel), '续费' (Renew), and a prominent blue button labeled '立即使用' (Launch Now). Below the application name, there's a brief description: 'GPUstack是一款轻量化算力集群管理与模型服务管理平台，可快速构建私有MaaS，实现异构GPU资源的统一管理' (GPUstack is a light-weighted算力集群管理与模型服务管理平台, which can quickly build a private MaaS,实现异构GPU resources unified management). Underneath the description, there's a section titled '应用信息' (Application Information) with fields for '创建时间' (Creation Time), '状态' (Status - currently '运行中' (Running)), '应用别名' (Alias - 'GPUStack-单机版'), '到期时间' (Expiration Time - '2026-03-03 10:33:26'), and '编辑' (Edit). To the right of these fields is a blue button labeled '一键复制' (One-click Copy). Below this information is a table titled '应用下的弹性云主机' (Elastic Cloud Hosts under Application) with one row of data. The row includes columns for '实例ID/名称' (Instance ID/Name - 'CTyunOS-23.01-x86_64-GPU'), '镜像' (Image - '运行中' (Running)), '状态' (Status), '规格' (Specification - '32核 | 256GB | L20 96GB * 2'), '付费方式/创建时间' (Payment Method/Created Time - '包年包月 | 2026-02-03 10:33:04'), and '到期时间' (Expiration Time - '2026-03-03 10:33:26').

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



2.7 部署模型。进入模型库菜单，选择合适的模型开始部署；



The screenshot shows the GPUStack interface. On the left, the sidebar has '模型库' (Model Library) selected. The main area displays a list of models:

- Qwen3-235B-A22B-Instruct-2507
- Qwen3-235B-A22B-Thinking-2507
- Qwen3-Embedding-4B
- Qwen3-Embedding-8B
- Qwen3-Ranker-4B
- Qwen3-Ranker-8B
- Qwen3-0.6B** (highlighted with a red arrow)
- Qwen3-8B
- Qwen3-32B
- FLUX 1-dev

To the right, a detailed deployment configuration window is open for 'Qwen3-0.6B'. It includes fields for '基本信息' (Basic Information), '性能' (Performance), '调度' (Scheduling), and '高级' (Advanced). The '性能' tab is selected. The configuration includes:

- 名称: qwen3-0.6b
- 集群: test
- 模式: 标准 (Standard)
- 精度: BF16
- 后端: vLLM
- 后端版本: 0.11.0
- 副本数: 2
- 描述: 兼容性检查通过 (Compatibility check passed)

Buttons at the bottom right are '取消' (Cancel) and '保存' (Save).

2.8 等待模型部署成功后，便可进入试验场，进行验证；

The screenshot shows the GPUStack interface. The sidebar has '部署' (Deployment) selected. The main area shows a deployment list table:

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	Running
qwen3-0.6b-llsp0				2026-01-08 17:33:54	Running

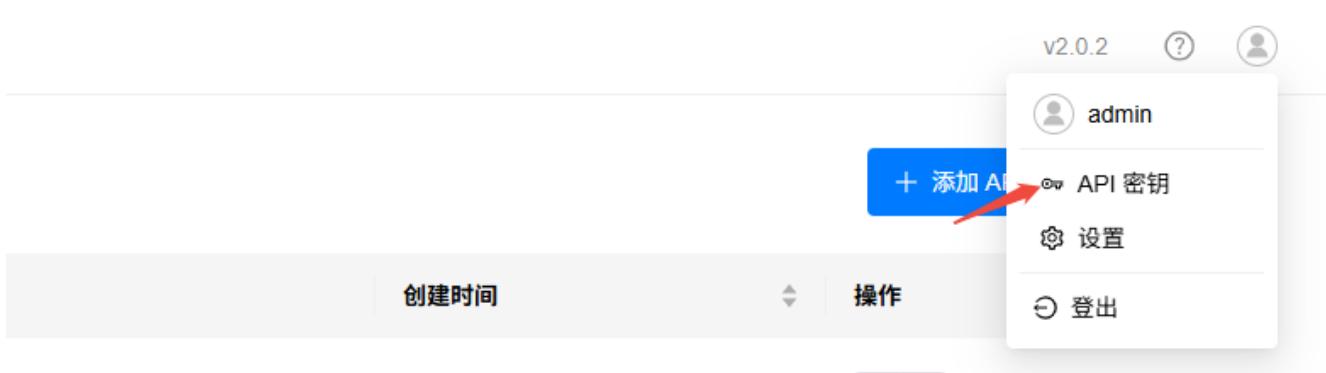
The screenshot shows the GPUStack interface. The sidebar has '对话' (Conversation) selected. The main area shows a conversation window:

- A red arrow points from the sidebar '对话' button to the conversation input field.
- A red arrow points from the '参数' (Parameters) section to the '模型' (Model) dropdown, which is set to 'qwen3-0.6b'.
- A red arrow points from the conversation input field to the conversation output area, which displays the message 'hi'.

At the bottom, a status bar shows 'Token 使用量: 98' and '输出: 301.37 Tokens/s'.

2.9 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

版权与声明

© [2025] [天翼云科技有限公司]。保留所有权利。

本文档内容仅供参考。