

训推服务

目录

快速入门

准备工作.....	2
用户使用流程.....	6

用户指南

数据工具.....	12
模型定制.....	15
模型服务.....	43
模型工具.....	56
训推加速.....	60
智算资产.....	78
运维观测.....	93
管理中心.....	108

常见问题

计费类.....	134
操作类.....	134

准备工作

注册主账号

在开通和使用训推服务之前，您需要先注册天翼云门户的账号。本节将介绍如何进行账号注册，如果您拥有天翼云的账号，可登录后使用训推服务。

1. 详见：[注册天翼云账号](#)
2. 如需实名认证，请参考会员服务-[实名认证](#)。
3. 主账号默认就是IAM管理员角色，具有平台所有权限。

为账户充值

1. 使用训推服务之前，请保证您的账户有充足的余额，账户余额需要大于100元。
2. 关于如何为账户充值，请参考费用中心-[在线充值](#)。
3. 训推服务支持包周期预付费。

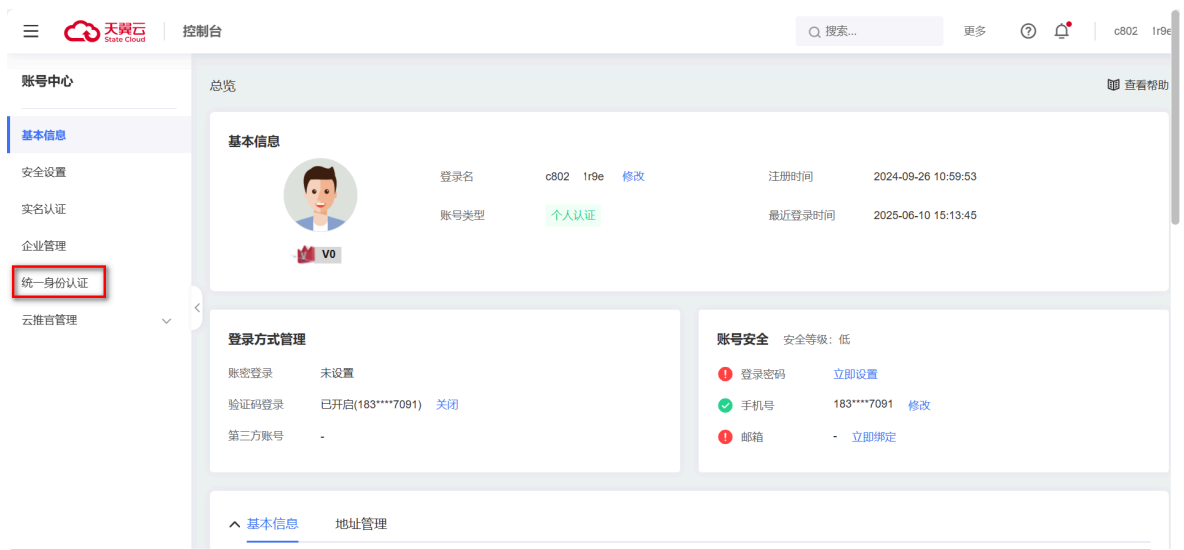
创建子账号

前置条件

通过主账号或角色为IAM管理员的子账号登录到训推服务

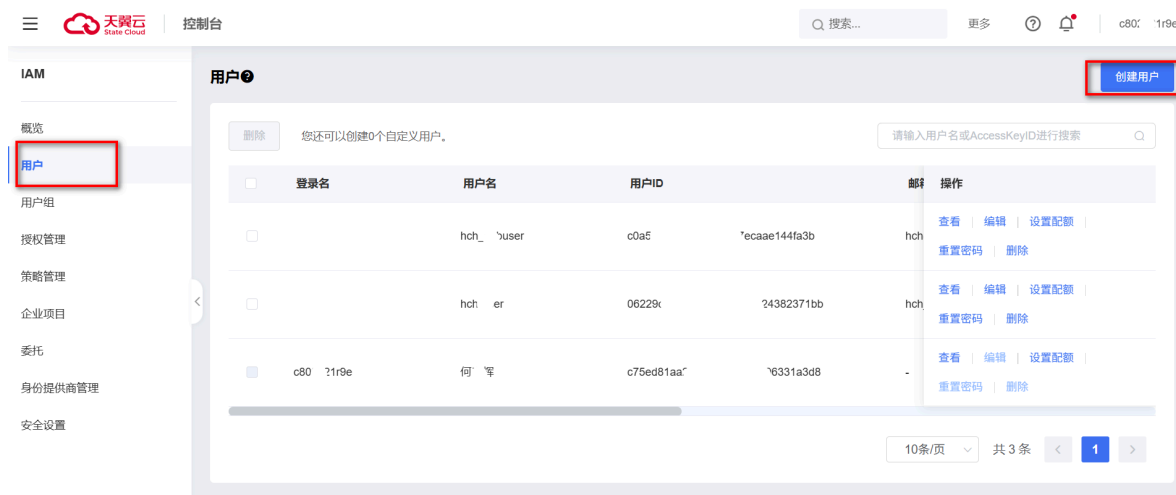
创建子账号操作步骤

1. 将鼠标放置在右上角用户登录信息处，会显示一个下拉菜单，点击【基本信息】会进入到【账号中心】页面
2. 在【账号中心】页面，点击进入【统一身份认证】页面，即【IAM】页面



快速入门

3. 在【IAM】页面，再次进入【用户】页面，在页面右上角点击【创建用户】按钮，进入到【创建用户】页面，按步骤填写相关信息即可创建子用户



需要特别说明的是，在为子用户添加【用户组】时，系统预置了“admin”用户组，如果将账号用户组设置为“admin”用户组(即IAM管理员角色)，则子用户拥有所有权限；若希望子用户只拥有普通角色权限，则可以分配用户组，此时子账号的角色为IAM普通用户。



同时为了方便组织架构较为严密的客户进行多层次资源管理，系统允许用户创建“二级管理员”用户组，其权限仅次于“admin”用户组，“二级管理员”可以和“admin”用户组组合使用，再结合工作空间能力，可实现多级账号资源管理，详见[多层次资源管理最佳实践](#)。”二级管理员“用户组创建流程如下：

快速入门

- 创建用户组：在用户组页面，点击”创建用户组“，在创建用户组弹窗中分别填写用户组名称和描述；



快速入门

- 为用户组授权：
 1. 点击授权操作按钮：

用户组

创建用户组

您还可以创建13个自定义用户组。

请输入用户组名进行搜索

用户组名称	用户数量	描述	创建时间	操作
二级管理员	0	仅次于管理员的角色	2025-10-31 09:26:58	查看 授权 编辑 用户组管理 删除
研发部门	1	研发部门成员	2025-07-22 17:59:28	查看 授权 编辑 用户组管理 删除
outAdmin	1	运维人员权限	2025-07-09 17:07:39	查看 授权 编辑 用户组管理 删除
OutOps	1	OutOps权限	2025-07-09 17:07:13	查看 授权 编辑 用户组管理 删除
outCommon	1	OutOps的权限	2025-07-09 17:06:45	查看 授权 编辑 用户组管理 删除

2. 在策略名称搜索框中输入”huijuOutSLAdmin“进行搜索，选择所搜索的结果，点击“下一步”按钮，再设置最小授权范围(保持默认设置即可)，再点击“确定”即用户组创建完成，此时再创建新用户，就可以选择“二级管理员”用户组了，对应账号将拥有二级管理员权限。

用户组 / 授权 / 二级管理员

1 选择策略

2 设置最小授权范围

3 完成

用户组“二级管理员”将拥有所选策略，只允许选择“资源池”或“全局级”中其中一种。

请选择策略类型 请选择作用范围

1 huijuOutSLAdmin

策略名称	策略描述	授权类型	授权范围	操作
2 <input checked="" type="checkbox"/> huijuOutSLAdmin	二级管理员	系统策略	全局级	查看

3 取消

下一步

当需要将子账号从某个用户组移除时，可进入到【用户组】页面，进行移除操作，移除后，子账号将不再具有对应用户组的权限

快速入门



用户使用流程

主账号使用流程

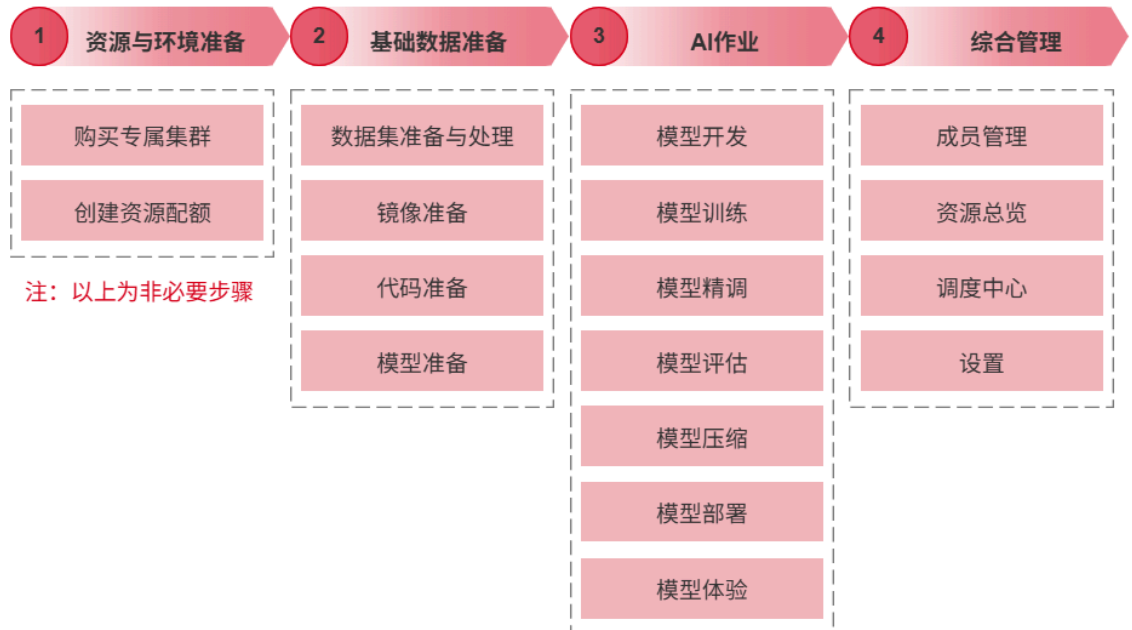
前置条件

已使用主账号登录到训推智算服务平台

主账号使用流程

主账号的使用流程，整体如下：

快速入门



注意：资源配额和工作空间目前仅在杭州7资源池支持

主账号的使用流程按照AI全周期的工作流程进行，具体步骤如下：

流程	子任务	说明	详细指导
购买专属集群	购买专属集群	相比平台提供的默认集群资源，购买专属集群资源独享，隔离性更好 注：非必要步骤，可以使用平台提供的默认集群按需使用	需联系客户经理进行购买
创建队列	创建队列	在购买专属集群后，可以将专属集群资源按照一定维度划分成多个队列，供成员在工作空间内进行作业时使用 注：非必要步骤，非必需购买专属集群，可以使用平台提供的默认集群，默认集群无需创建队列，可直接使用。	队列管理
数据集准备与处理	创建基础数据集	可上传基础数据集到ZOS或HPFS	基础数据集

快速入门

流程	子任务	说明	详细指导
	创建标注数据集	创建标注数据集，可对数据集进行标注，并发布为新的数据集	标注数据集
	数据清洗	可以对数据进行清洗	数据处理
镜像准备	预置镜像	预置镜像即平台预先设置的完整镜像，可直接用于创建任务时使用	预置镜像
	自定义镜像	可以通过开发机自主制作镜像或通过天翼云容器镜像服务将镜像服务内的容器共享给一体化智算平台	自定义镜像
	他人分享镜像	可将自己的镜像分享给他人使用	镜像分享
代码准备	导入代码包	将代码上传到平台	我的代码包
模型准备	导入模型	可将用户自己的模型或在平台训练、精调好的模型导入到平台进行版本管理、评估及部署	模型管理
AI作业	模型开发	可通过启动和管理在线JupyterLab或VSCode集成开发环境在线编程进行模型开发	模型开发
	模型训练	创建自定义创建训练任务，支持单机和多机分布式训练	训练任务
	模型精调	零代码快速创建和管理精调任务，提供全参微调 and lora 微调两种精调方式。基于平台的基础大模型，选择训练数据集和算力即可快速启动精调任务。	模型精调
	模型评估	可对模型进行评估，自动评估打分，生成评估报告	模型评估
	模型压缩	不减少模型效果的前提下压缩模型大小，进而提升模型在推理调用时的性能	模型压缩
	模型部署	部署模型，提供推理服务	模型服务
	成员管理	查看并管理本租户下所有用户的平台使用情况	成员管理

快速入门

流程	子任务	说明	详细指导
	总览资源	平台管理员能够轻松查看并管理专属集群的资源使用情况	资源总览
	调度中心	旨在让平台管理员能够轻松查看并调度集群资源	调度中心
	设置	旨在让平台管理员能够轻松查看并设置本租户下所有用户对资源使用的限额	设置

子账号使用流程

前置条件

子账号为IAM普通用户角色且已登录到训推智算服务平台

子账号使用流程

若子账号为IAM管理员角色，则其拥有全部操作权限，操作流程和主账号别无二致，故角色为IAM管理员的子账号使用流程参见 [主账号使用流程](#) 章节；角色为IAM普通用户的子账号，使用流程如下：



子账号的使用流程主要是为了进行任务作业，具体步骤如下：

流程	子任务	说明	详细指导
数据集准备与处理	创建基础数据集	可上传基础数据集到ZOS或HPFS	基础数据集

快速入门

流程	子任务	说明	详细指导
	创建标注数据集	创建标注数据集，可对数据集进行标注，并发布为新的数据集	标注数据集
	数据清洗	可以对数据	数据处理
镜像准备	预置镜像	预置镜像即平台预先设置的完整镜像，可直接用于创建任务时使用	预置镜像
	自定义镜像	可以通过开发机自主制作镜像或通过天翼云容器镜像服务将镜像服务内的容器共享给训推智算服务平台	自定义镜像
	他人分享镜像	可将自己的镜像分享给他人使用	镜像分享
代码准备	导入代码包	将代码上传到平台	我的代码包
模型准备	导入模型	可将用户自己的模型或在平台训练、精调好的模型导入到平台进行版本管理、评估及部署	模型管理
AI作业	模型开发	可通过启动和管理在线JupyterLab或VSCode集成开发环境在线编程进行模型开发	模型开发
	模型训练	创建自定义创建训练任务，支持单机和多机分布式训练	训练任务
	模型精调	零代码快速创建和管理精调任务，提供全参微调 and lora 微调两种精调方式。基于平台的基础大模型，选择训练数据集和算力即可快速启动精调任务。	模型精调
	模型评估	可对模型进行评估，自动评估打分，生成评估报告	模型评估
	模型压缩	不减少模型效果的前提下压缩模型大小，进而提升模型在推理调用时的性能	模型压缩
	模型部署	部署模型，提供推理服务	模型服务

数据工具

数据处理

数据清洗

数据清洗可对原始数据进行异常清洗、文本过滤、去重以及隐私信息去除等操作，解决数据规范性、合规性、一致性及重复等问题，提升数据质量，优化模型训练效果。

数据准备

目前只支持对指令微调类型的标注数据集进行清洗，数据清洗前，请先到「我的数据集-标注数据集」模块下，创建「大语言-SFT-指令微调」类型的数据集。详见[标注数据集](#)

新建数据清洗任务

从清洗任务列表点击【新建清洗任务】到新建页面，填写任务名称、选择数据集、配置字段和算子后提交即可。

- **处理前数据：**仅支持指令微调类型的标注数据集，可点击【前往导入数据集】，跳转至「我的数据集」模块创建数据集。
- **处理后数据：**系统自动新建数据集。
- **处理数据集字段：**默认全选，可自行勾选要清洗的数据集字段。
- **算子选择：**可选择删除敏感信息、删除无效字符、长度过滤三种清洗算子，可根据数据集内容和实际需求进行开启或关闭。
- **集群：**选择所属集群。
- **队列：**选择支持的队列。
- **算力申请：**
 - **资源规格：**默认8C 16G
 - **master节点：**管理节点，用于分配数据清洗任务，默认为1。
 - **worker节点：**负责执行具体的数据清洗任务，默认为1，可根据数据集大小调整节点数量，节点数量越多清洗任务速度越快，一般小规模数据集节点为1即可。

用户指南

数据工具 / 新建清洗任务

* 任务名称

仅支持中英文、数字、下划线"_"、短横"-", 只能以中英文、数字开头

0/50

处理前数据

指令微调

Q&A对

请选择数据集

前往导入数据集

处理后数据

新建数据集

仅支持中英文、数字、下划线"_"、短横"-", 只能以中英文、数字开头

* 处理数据集字段

instruction

input

output

* 算子选择

序号	是否开启	算子类型	算子配置	效果预览 (清洗前)	效果预览 (清洗后)
1	<div></div>	删除敏感信息	<div><div><div>删除邮箱信息</div><div>删除身份证</div><div>删除手机号</div></div></div>	用户注册时需要使用形如123456199001011234的身份证号码进行实名认证,如需要咨询软件服务,请联系客户经理,liming@163.com,13800138000	用户注册时需要使用形如1234564的身份证号码进行实名认证,如需要咨询软件服务,请联系客户经理。
			<div><div><div>删除HTML标签并返回所有节点的纯文本</div></div></div>	<article><p style="margin-top: Opx;	

取消

提交

数据清洗任务管理

通过该模块功能，可以对数据清洗任务进行管理，包括查看任务详情、查看结果、删除、重新启动等操作。

天翼云

控制台

华北2

默认集群

搜索...

费用 工单 备案 支持 合作

c36864dn0u

数据工具 / 数据处理

数据清洗

新建清洗任务

请选择运行状态

请输入任务名称

查询

重置

任务名称	处理方式	处理状态	数据来源	数据流向	创建时间	操作
02	删除敏感信息、删除无效...	已完成	01031.0	0103_17397368...	2025-02-17 04:13:37	查看结果 复制 日志 删除 重新启动 终止
测试01	删除无效字符	清洗失败	41.0	4_1739735414541	2025-02-17 04:01:27	查看结果 复制 日志 删除 重新启动 终止

共2条

13

用户指南

主要操作项：

- **查看任务详情：** 点击任务名称，可跳转至「数据清洗任务详情」页面，点击【清洗日志】，可查看具体任务日志记录。

数据工具 / 数据清洗详情

< 02_1739782769144

运行记录 ⓘ

1891412159082090498

已终止

2025-02-17 16:59:34

[详情](#)

[清洗日志](#)

基础信息

任务名称	02_1739782769144
处理前数据	0103
处理后数据	0103_1739736808827_1739782769144

算子选择

算子类型	算子配置
删除敏感信息	删除邮箱信息
	删除身份证
	删除手机号
删除无效字符	删除HTML标签并返回所有节点的纯文本
	删除链接, 例如以http或ftp开头的
	删除IP地址
	删除代码文件开头的版权声明(必须包含单词 copyright)
	删除 TeX 文档的参考文献
	删除 TeX 文档中的注释

数据工具 / 数据清洗详情		
< 02		
运行记录 🕒		
🔍 1891219400699539458		
已完成		
2025-02-17 04:13:37		
详情 清洗日志		
节点 请选择 ⌵		
⌵ 🔍 🔄 🕒 🕒 2025-02-17 04:13:37 至 2025-02-17 16:36:00		
1	2025-02-17 04:15:20.000	'num_proc': 1,
2	2025-02-17 04:15:20.000	'text_key': 'instruction',
3	2025-02-17 04:15:20.000	'turbo': False,
4	2025-02-17 04:15:20.000	'video_key': 'videos'}},
5	2025-02-17 04:15:20.000	{'clean_html_mapper': {'accelerator': None,
6	2025-02-17 04:15:20.000	'audio_key': 'audios',
7	2025-02-17 04:15:20.000	'cpu_required': 1,
8	2025-02-17 04:15:20.000	'image_key': 'images',
9	2025-02-17 04:15:20.000	'mem_required': 0,
10	2025-02-17 04:15:20.000	'num_proc': 1,
11	2025-02-17 04:15:20.000	'text_key': 'input',
12	2025-02-17 04:15:20.000	'turbo': False,
13	2025-02-17 04:15:20.000	'video_key': 'videos'}},
14	2025-02-17 04:15:20.000	{'clean_html_mapper': {'accelerator': None,
15	2025-02-17 04:15:20.000	'audio_key': 'audios',
16	2025-02-17 04:15:20.000	'cpu_required': 1,
17	2025-02-17 04:15:20.000	'image_key': 'images',
18	2025-02-17 04:15:20.000	'mem_required': 0,
19	2025-02-17 04:15:20.000	'num_proc': 1,
20	2025-02-17 04:15:20.000	'text_key': 'output',
21	2025-02-17 04:15:20.000	'turbo': False,
22	2025-02-17 04:15:20.000	'video_key': 'videos'}},
23	2025-02-17 04:15:20.000	{'clean_links_mapper': {'accelerator': None,
24	2025-02-17 04:15:20.000	'audio_key': 'audios',
25	2025-02-17 04:15:20.000	'cpu_required': 1,
26	2025-02-17 04:15:20.000	'image_key': 'images',
27	2025-02-17 04:15:20.000	'mem_required': 0,
28	2025-02-17 04:15:20.000	'num_proc': 1,
29	2025-02-17 04:15:20.000	'text_key': 'input',
30	2025-02-17 04:15:20.000	'turbo': False,
31	2025-02-17 04:15:20.000	'video_key': 'videos'}},
32	2025-02-17 04:15:20.000	{'clean_links_mapper': {'accelerator': None,
33	2025-02-17 04:15:20.000	'audio_key': 'audios',
34	2025-02-17 04:15:20.000	'cpu_required': 1,
35	2025-02-17 04:15:20.000	'image_key': 'images',
36	2025-02-17 04:15:20.000	'mem_required': 0,
37	2025-02-17 04:15:20.000	'num_proc': 1,
38	2025-02-17 04:15:20.000	'text_key': 'input',
39	2025-02-17 04:15:20.000	'turbo': False,
40	2025-02-17 04:15:20.000	'video_key': 'videos'}},
41	2025-02-17 04:15:20.000	{'clean_links_mapper': {'accelerator': None,
42	2025-02-17 04:15:20.000	'audio_key': 'audios',
43	2025-02-17 04:15:20.000	'cpu_required': 1,
44	2025-02-17 04:15:20.000	'image_key': 'images',
45	2025-02-17 04:15:20.000	'mem_required': 0,
46	2025-02-17 04:15:20.000	'num_proc': 1,
47	2025-02-17 04:15:20.000	'text_key': 'input',
48	2025-02-17 04:15:20.000	'turbo': False,
49	2025-02-17 04:15:20.000	'video_key': 'videos'}},
50	2025-02-17 04:15:20.000	{'clean_links_mapper': {'accelerator': None,
51	2025-02-17 04:15:20.000	'audio_key': 'audios',
52	2025-02-17 04:15:20.000	'cpu_required': 1,
53	2025-02-17 04:15:20.000	'image_key': 'images',
54	2025-02-17 04:15:20.000	'mem_required': 0,
55	2025-02-17 04:15:20.000	'num_proc': 1,
56	2025-02-17 04:15:20.000	'text_key': 'input',
57	2025-02-17 04:15:20.000	'turbo': False,
58	2025-02-17 04:15:20.000	'video_key': 'videos'}},
59	2025-02-17 04:15:20.000	{'clean_links_mapper': {'accelerator': None,
60	2025-02-17 04:15:20.000	'audio_key': 'audios',
61	2025-02-17 04:15:20.000	'cpu_required': 1,
62	2025-02-17 04:15:20.000	'image_key': 'images',
63	2025-02-17 04:15:20.000	'mem_required': 0,
64	2025-02-17 04:15:20.000	'num_proc': 1,
65	2025-02-17 04:15:20.000	'text_key': 'input',
66	2025-02-17 04:15:20.000	'turbo': False,
67	2025-02-17 04:15:20.000	'video_key': 'videos'}},
68	2025-02-17 04:15:20.000	{'clean_links_mapper': {'accelerator': None,
69	2025-02-17 04:15:20.000	'audio_key': 'audios',
70	2025-02-17 04:15:20.000	'cpu_required': 1,
71	2025-02-17 04:15:20.000	'image_key': 'images',
72	2025-02-17 04:1	

- **查看数据来源/数据流向：**点击两个字段下的数据集名称，即可跳转至我的数据集页面，于「标注数据集」分类下，会自动筛选出对应的数据集。
- **查看结果：**点击跳转至我的数据集页面，于「标注数据集」分类下，会自动筛选出结果数据集，清洗完成的数据集为导入完成状态。
- **复制：**点击可复制当前数据清洗任务，任务名称自动变更。
- **日志：**点击可直接跳转至任务详情页中「清洗日志」分类下。
- **删除清洗任务：**点击【删除】后，确定删除即可。
- **重新启动：**当任务被手动终止，或者清洗任务失败，可点击重新启动任务。
- **终止任务：**提交任务后，可手动终止对应清洗任务。

后续操作：

数据清洗任务完成后，可将清洗后的数据集用于模型精调、训练等任务。

详见：[模型精调](#)、[模型开发](#)、[训练任务](#)

模型定制

模型精调

模型精调功能介绍

SFT（Supervised Fine-Tuning）简介

监督微调（Supervised Fine-Tuning, SFT）是大模型训练流程中的关键环节，其目标是通过高质量的人工标注数据（通常为问答对，包含输入的prompt和预期输出的response），将通用预训练大模型（如Llama、Qwen等）适配到特定的下游任务或专业领域，使其生成更符合人类期望且有一定偏好格式的输出。

SFT的必要条件

何时需要启动SFT

- 当通过prompt工程，大模型也无法将用户所想看到的回答呈现出来（即指令跟随失败）时；
- 当大模型生成内容存在事实错误或专业度不足时，比如某些领域的词汇未在预训练模型的前置知识里；
- 当用户对大模型输出内容有格式要求时，比如某些专业领域（客服）需要格式化输出，而模型仍有部分回答不符合格式要求。

SFT的前期准备

当用户确定要使用大模型SFT时，请确保已经做好了以下准备：

- 用户已经使用prompt工程等方式，将基础模型的潜力发挥到极致，并且针对垂直领域的测试集，仍然存在一定数量的badcase，且数量是不可接受（一般业界定义专业领域知识测试准确率<75%即为不可接受）的；
- 针对该垂直领域，用户已经标注了至少1K的高质量数据集。如何评判一个数据集是否是高质量数据集请见下一节。

何为高质量数据集

高质量数据集一般存在以下特征：

- 针对某一领域，该数据集覆盖一定量的badcase或者与badcase特征相似的数据；
- 该数据集有着多样的prompt类型（即提问方式非单一）以增加模型的泛化性；
- 该数据集针对同类型prompt的回答方式统一或结构差异较小；

用户指南

- 该数据集不存在模棱两可的问题以及答案；
- 该数据集不存在答案重复输出的数据；
- （如需要）该数据集会添加5-10%的通用数据以防止微调后的模型出现灾难性遗忘。

数据集配置

针对不同场景下，以下是数据量级的建议：

- **简单文本分类任务：**对于基础的分类任务，如判断文本情感倾向（积极、消极、中性）或对新闻文章进行简单类别划分（体育、科技、娱乐等），若用户使用较小规模模型（如 7B及以下），通常5k-20k条的数据量即可取得不错效果；若使用更大参数模型（如30B及以上），由于模型学习能力更强，对数据量要求相对较低，2k-10k条也可能实现较好微调。但数据量的大小也与分类数以及任务难度强相关，类别较多/任务较难的场景可能需要20k条以上的数据，且要做到类别均衡。
- **信息抽取任务：**像从文本中抽取人名、地名、组织机构名等实体，或抽取事件的时间、地点、参与者等关键信息这类任务，任务难度更大，因此对数据量需求相对较高。若用户使用13B及以下模型时，建议有30k-80k的数据量；若使用30B及以上模型时，20k-60k的数据量较为合适。这是因为信息抽取任务复杂，模型需学习多种实体和关系模式，足够数据量才能让模型有效捕捉这些模式。例如，在医疗领域的实体抽取微调中，50k条的高质量医疗文本数据，能使13B模型的F1值达到75%左右。
- **生成任务（如文本续写、文案生成）：**生成任务对模型创造力和语言理解能力要求高，数据量需求更大。对于7B及以下的模型，一般需要100k-300k条以上数据；而13B-30B的模型，50k-200k条数据较为适宜；30B以上模型，30k-150k数据可能满足需求。数据不仅要量大，还应多样化，涵盖不同风格、主题和语境文本。如在小说续写微调中，200k条包含各种题材小说片段的数据，可让7B模型生成更连贯、富有想象力的续写内容。
- **复杂问答任务：**如开放域问答、专业领域深度问答、领域内长思考链生成问答，数据量要求与生成任务相近甚至更高。7B的模型可能需要50k-150k条数据；13B-30B的模型，40k-100k条数据；30B及以上的模型，则需要50k-80k数据。这类任务需模型理解复杂问题语义，检索知识并生成准确回答，大量的数据帮助模型学习各类问题模式和答案逻辑。但是如果数据中带有思考推理过程，也即cot数据集，则各类模型所需的数据量可以相对应减少50%-80%。在医疗领域的带思考过程的问答微调中，10k数量的含各类疾病的治疗方式的数据，便能让14B的模型在开源医疗问题测评集上提点。

需注意，以上数据量仅为经验参考范围，实际微调中最佳数据量受数据质量、模型架构、任务复杂度及训练方法等多种因素影响。例如，若数据质量极高且与任务高度相关，可能用较少数据量就能取得好效果。在微调前，可通过小规模实验评估不同数据量对模型性能影响，确定适合特定任务和模型的最优数据量。

精调配置

模型选择

一般来说，1-13B为小模型，13-34B为中模型，70B以上为大模型。

- **依据任务的难易程度：**若客户明确针对项目的模型需要有长文本续写能力，或者需要做比较复杂的思维链推理，则可以选择大模型；若只是简单的短文本问答或者分类任务，选择小模型即可，因为小模型更容易捕捉短文本的局部特征。若强行用小模型做困难任务，即使做了极致的prompt工程优化，也不可避免地产生幻觉；若使用大模型做简单任务，则会造成极大的资源浪费。
- **依据推理时延：**一般来说，当推理batch为1时，小模型的推理时延<300ms，中模型的推理时延为500-800ms，大模型的推理时延>1.5s，若客户针对项目的模型有低时延要求，建议使用小模型。
- **依据资源数：**一般来说，小模型在短文本长度的微调上需要的资源数较少，即使是全量微调，也只需要单机，大模型所需的资源则较多，72b至少也需要4台A800机器才能微调，这也是需要考虑的情况。

精调方式

主流的精调方式包括：

- 全参微调 (Full Fine-Tuning)是解冻整个模型权重进行训练，更新全模型的参数，效果上限高，适合复杂任务，但该方式显存占用巨大，而且在专业领域的数据集上进行全参微调可能造成灾难性遗忘。
- LoRA微调 (Low-Rank Adaptation)是冻结预训练模型的权重，并在每个模块中注入可训练的低秩矩阵 (Adapter)，因为训练参数少所以可以节省60%-70%的显存，也比较适合中小数据集且任务较为简单的情况。由于LoRA微调相当于给基础模型外接一个知识库，因此其效果也略低于全参微调。

从工程经验上来看，对于全参微调的选择，满足以下几点一般都可以进行全参微调：

- 任务复杂度高（比如进行长文本思考链的推理）；
- 数据量充足（拥有大于10k的高质量数据）；
- 算力无约束（拥有A800/H100/H800/910B2集群）；
- 数据分布剧变（任务所需的领域数据集与基础模型的KL散度过大）。

而对于LoRA微调，工程上的经验一般都有以下几点：

- 资源受限但需较好效果（只拥有较少的机器数但又想训练自己的行业大模型）；
- 快速迭代实验（不一样的任务需求搭配不同的LoRA权重，比如电商里的修改订单和商品推荐属于不一样的简单任务）；
- 中小规模数据（仅仅只有1k-50k的数据样本）；
- 防止灾难性遗忘（由于全参微调遗忘过多知识，可以通过增加一定程度的通用数据+LoRA微调减缓）。

关于LoRA微调和全参微调在效果上的差距，在大多数场景，LoRA微调可以达到全参微调95%以上的效果。需要注意的是，LoRA微调和全参微调在学习率上，有着不同的最佳参数取值范围，这一点在训练参数选择中做进一步的说明。

建议先使用小模型+ LoRA微调来验证实验设置及微调数据是否有效，如果有效的话，再逐渐增大数据量并迁移到大模型上验证微调效果。这样做可以显著加快迭代效率，避免在大参数模型上反复做无意义的迭代。

精调参数

在SFT中，不同的超参数选择会影响模型的精度以及性能，下面逐一对部分重要参数进行解释并给出一般性推荐。

- **epoch**: 模型训练轮数，通常选择2~5，可以根据eval集的loss曲线来判断：如果train集的loss曲线下降，eval集loss曲线上升，则说明模型已经过拟合，此刻考虑到基础模型的通用能力，应该停止训练；如果train集和eval集loss曲线均在缓慢下降，则说明模型还未收敛，可以继续进行训练。按经验来说，一般性的任务，全参微调训练3个epoch以内，LoRA微调应该取3-5个epoch。而生成类任务的epoch数可以略微设置大一点，如5~10范围内。
- **learning rate**: 学习率，设置较大时会加速模型迭代，但是模型可能无法收敛到最优点；设置过小时会使得模型迭代较慢，可能陷入局部最优。按照经验来讲，LoRA微调选择 learning rate 在 $1e-4 \sim 3e-4$ ，全参微调选择learning rate在 $1e-6 \sim 5e-6$ 即可。
- **global batch size**: 全局批处理大小，在大模型加速框架Megatron下，该值的增大不会导致显存溢出的问题，但由于梯度累积的存在，其值越大，模型训练性能会降低。一般来说，小模型使用的gbs值可以小一点，大模型的gbs值一般设置成64及以上，gbs的设置一般是2的次幂，比如2，4，8，16，32等。
- **seq lens**: 序列长度，其取值应该大于等于训练数据里面最长的那个，以防止数据被截断。一般来说，由于大模型普遍应用旋转位置编码，其已经具备了一定的外推性，8k的序列长度已经可以满足一些简单的、常规的问答任务。如果有特殊需求，比如长文本续写或者特长思维链推理，才会考虑放大序列长度到32k。而且

用户指南

增大序列长度所需的资源数会成倍增长，通常会开启序列并行或者激活值重计算的方式来支持模型的长序列微调能力，会严重影响训练性能。

另外，针对这些参数，有一些额外建议供用户参考：

- 微调时，learning rate是随着global batch size的增加而单调递增但有上界的。当优化器是Adam或者RMSProp时，learning rate是和global batch size的平方根成正比。但是增大global batch size到一定阈值后，learning rate反而应该减小。
- 当增大global batch size并相应增大learning rate时，为保证模型训练一定的步数，有时epoch也需要稍微增大一些。
- LoRA微调时，最优的learning rate相比于全参微调时最优的learning rate，一般大1~2个数量级。

模型精调功能使用

前置条件

数据准备：参考[精调前期准备](#)，创建需要精调的标注数据集，详见[标注数据集](#)。

创建调优任务

- 您可通过以下方式使用：
 - 入口：进入“模型定制”>“模型精调”菜单，进入调优任务列表，点击【新建调优任务】，进入创建页面。
- 填写相关配置：

类型	字段	说明
训练配置	任务名称	精调任务名称，仅支持英文、汉字、数字、-、_，且只能以中英文、数字开头，不超过60个字符
	任务描述	精调任务描述，最多128个字
	模型类型	精调的模型类型，支持文本生成与图像理解
	训练框架	训练框架，支持PyTorch、MindSpore
	基础模型	列表中列出的都是支持精调的模型，模型卡片下的标签表示模型精调后支持的动作，例如“可评估”，表示该模型在精调完成后支持评估
	调参方式	选择所需的调参方式，不同模型支持的方式可能不同
	参数配置	根据调参方式，进行相应的参数配置
数据配置	数据集类型	选择精调的数据集类型，不同模型支持的类型可能不同
	数据集	仅支持对应数据集类型下的标注数据集

用户指南

资源配置	集群	选择适合的集群
	资源配额	选择可用资源配额
	算力申请	选择可用的资源，不同的算力规格对应不同的价格，单节点下卡数越多训练越快

思域一体化智算服务平台

一站式智算服务平台

概览

模型广场

体验中心

数据工具

数据处理

模型定制

模型精调

开发机

训练任务

模型服务

服务接入

在线服务

调用监控

模型工具

模型评估

模型压缩

智算资产

我的模型

我的数据集

我的知识库

我的镜像

我的代码包

管理中心

成员管理

资源总览

队列管理

调度中心

设置

模型精调 / 创建精调任务

训练配置

* 任务名称

仅支持英文、汉字、数字、-、_，且只能以中英文、数字开头，不超过60个字符

任务描述

请输入描述信息，最多128个字

0/128

* 模型类型

☒ 文本生成 ☐ 图像理解

* 训练框架

☒ PyTorch ☐ MindSpore

* 基础模型

☒ Llama2-13B-Chat

敬请期待...

☐ Qwen2-7B-Instruct

可评估

可快速部署

☐ GLM4-9B-Chat

敬请期待...

☐ Qwen2.5-72B-Instruct

敬请期待...

Llama2是预先训练和微调的生成文本模型的集合，其规模从70亿到700亿个参数不等。这是13B微调模型的存储库，针对对话用例进行了优化。

* 调参方式

☒ 全参数调 ☐ Lora微调

超参数配置

* 全局批处理大小(Global_Batch Size)

128

ⓘ

* 迭代轮次(EPOCH)

2

ⓘ

* 学习率(Learning_Rate)

-

0.0000200000

+

ⓘ

数据配置

* 训练数据集

☒ 指令微调

数据集20250515135729572

×

前往导入数据集

资源配置

管理调优任务

- 返回模型调优训练任务列表，列表中可以看到每个任务的运行进度、预估时长、占有资源等信息。
- 查看详情：点击任务名称，可进入调优任务详情页，右上角可对任务进行停止和删除操作；详情页可依次查看任务基础信息、日志、监控、Tensorboard看板。监控中运行进度可以看到每一次迭代是否完成，资源监控看板可以查看CPU使用率、内存使用率、NPU使用率等。
- 操作：
 - 点击【日志】、【监控】按钮，可直接跳转到详情页查看任务相应信息；
 - 点击【发布模型】按钮，可将已完成精调的任务保存至我的模型中，以进行后续的评估等操作；
 - 点击【重启】，可重新启动任务；点击【更多】，可对不需要的任务进行删除。

模型开发

模型开发

前置条件

- 完成训练数据集准备，完成存储配置准备（ZOS/HPFS），详见[我的数据集](#)。
- 如果预置模型不满足开发要求，需要基于自有模型，需要完成模型文件准备，详见[模型管理](#)。
- 如果预置镜像不满足开发要求，需要基于自有镜像，需要完成镜像文件准备，详见[我的镜像](#)。

创建开发机

1. 登录智算服务平台。
2. 在左侧导航栏选择模型开发，选择目标开发机类型，平台支持JupyterLab和VSCode两种类型的开发机。



3. 进入开发机创建页面，填写开发机创建的基本参数。

参数类型	参数名	说明
基本信息	名称	必填，开发机名称。
数据集配置	训练数据集	最多可添加5个，选择基础数据集或者标注数据集。
模型配置	模型来源	我的模型：最多5个，自有模型开启开发机任务。预置模型：最多5个，使用平台预置模型开启开发机任务。
	模型文件	选择我的模型具体的模型文件及版本。选择预置模型文件及具体版本。
存储配置	ZOS对象存储	最多选择5个，如果没有提前创建，可以点击“去创建对象存储”完成创建。
存储配置	HPFS并行文件系统	最多选择5个，如果没有提前创建，可以点击“去创建HPFS”完成创建。

用户指南

参数类型	参数名	说明
环境配置	镜像来源	根据当前任务情况，选择支持选择系统预置镜像、自定义镜像、共享容器镜像和他人分享镜像。需要远程登录开发机时，可以选择平台打包好相关组件的带有”ssh”标签的镜像。
	文件目录	平台可持久化的挂载目录，后续可以在该目录下读写文件，是用户间隔离的。
资源配置	集群	支持公共集群和专属集群两种类型，其中专属集群需要提前购买。
	资源配额	选择目标资源配额，展示当前总资源及使用情况。
	自定义分配	指对资源配额内的资源虚拟化切分使用（当前仅支持英伟达集群）。下文“ 算力虚拟化 ”章节展开介绍具体参数介绍。
	资源规格	选择当前任务所需要的资源规格。
访问配置	启用SSH	菜单默认隐藏，当用户被添加进白名单且选择了ssh镜像时可见。开启ssh开关后，平台会为用户创建一个EIP，当所有开发机被删除时才会释放该EIP，产生费用规则详见 计费方式 。
	SSH公钥	输入公钥不得超过10个，需要用分号分隔，详细说明见下文“ 远程登录开发机 ”章节。

4. 点击“提交”，完成开发机任务创建。

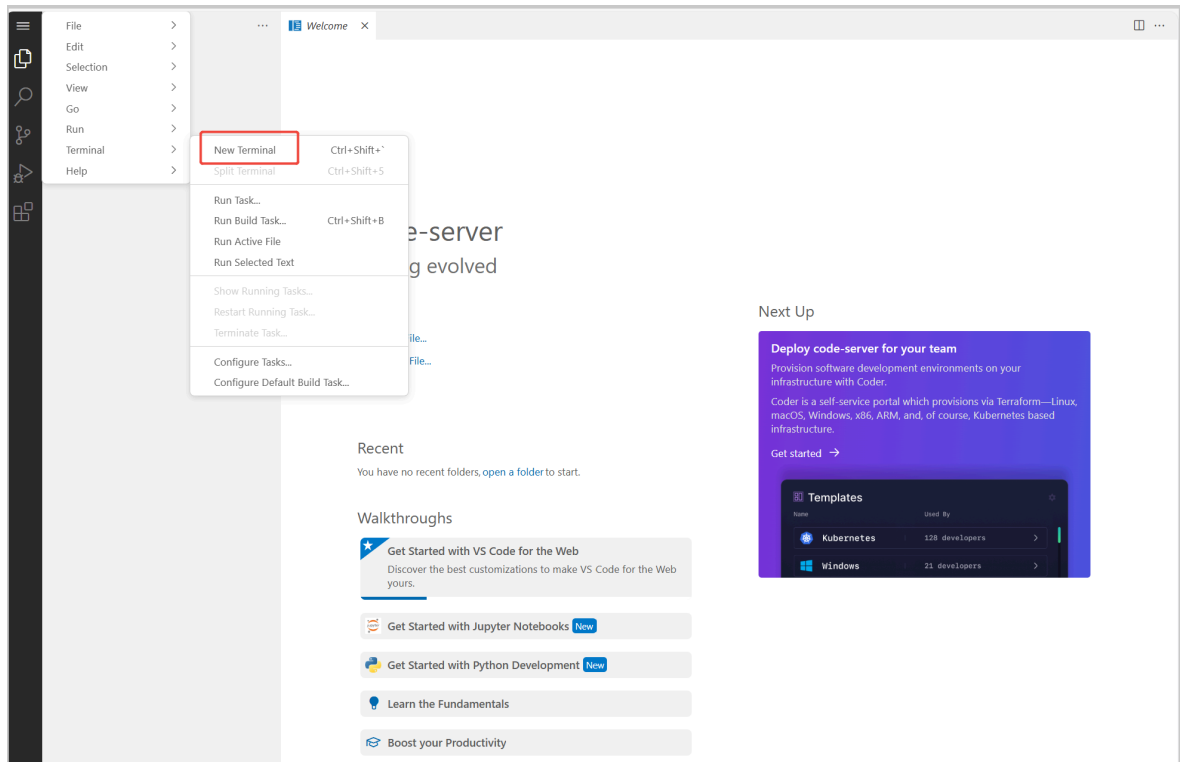
开发机管理

- 启动开发机：**创建完成后的开发机默认为“未启动”状态，用户可点击启动，开发机进入“启动中”-“运行中”状态。

名称	创建人	创建时间	状态	运行时长 ⌚	镜像版本	资源类型	特殊配置	操作
		2025-06-09 14:16:28	● 未启动	-	v2.0-vscode_jupyter	21C_175G_1*H800	-	启动 停止 打开 开始训练 更多 ^
		2025-06-05 18:21:23	⚙ 运行中	2天20时34分5秒	v1.2	21C_175G_1*H800	-	启动 停止 打开 删除 查看配置 保存至模型管理 保存至代码包
		2025-05-30 16:26:25	● 未启动	6天58时39分49秒	v1.2	CPU_2C-内存_8G	-	启动 停止 打开
		2025-05-30 16:03:32	● 未启动	1分3秒	v1.2	CPU_2C-内存_8G	-	启动 停止 打开

- 打开开发机：**当开发机进入“运行中”状态时，可以点击“打开”进入vscode的terminal运行终端。

用户指南



- **停止开发机：**停止当前正在运行的开发机。停止VSCode会释放容器和资源，在停止前将数据、代码、模型保存在系统提供的持久化挂载路径(/work/home)中，下次启动还可以加载出来。开发机实例从“停止中”-“未启动”状态。
- **开始训练：**点击之后会携带当前开发机相关参数，进入训练任务创建界面。
- **删除开发机实例：**删除当前不再使用的开发机实例。
- **查看配置：**查看当前开发机实例的详细配置。

开发机中数据目录使用（重要）

1、cache存储：开发机默认自带cache持久化存储目录。每个用户的每一个开发机都默认赠送50GiB的/work/cache目录。

a) 目录隔离：在不同开发机的目录内容是隔离的，每启动一个开发机都是自己的一套cache。

b) 存储时间：cache存储在开发机停止后保存时间为5天。

c) 存储策略：cache存储为持久化存储，在不关闭开发机的情况下，可一直保存。

2、mount目录：数据和模型挂载mount目录，/work/mount/目录是用于使用和挂载平台上的数据集、模型、代码文件的，创建界面选择对应数据集和模型后，系统自动分配挂载路径，打开开发机即可看到这些路径，写代码直接读取这些路径即可读取数据和模型。

a) 存储时间：只在开发机运行中可使用，开发机关闭后mount目录会和容器取消挂载，也无法读取。

b) 目录隔离：在不同开发机的目录内容是隔离的，每启动一个开发机都是自己的一套mount。

c) 存储策略：mount目录为挂载使用，用户在mount对应子目录下写入的代码或者拖入的文件会自动转存回原始数据集、代码包或者模型存储中。

【注意】：自动转存的前提为文件或者文件夹写入到平台自动挂载的目录下，在/work/mount目录下手动创建的目录（比如：/work/mount/abc）不会做自动转存，同时开发机销毁后自动清理。

用户指南

启动训练任务

在/work/share/ 目录下创建自己的工作空间。

```
root@vscode-1876449876017119232-1876449934674460673-b44d9654b-9mrjl:/work# ls
cache share
root@vscode-1876449876017119232-1876449934674460673-b44d9654b-9mrjl:/work# cd share
root@vscode-1876449876017119232-1876449934674460673-b44d9654b-9mrjl:/work/share# ls
1.sh autotest chenyd dataset_trans finetune huangyun kernel_meta lig45 liunian lqh models rinoshinme
HumanEval.jsonl.gz chengxiang cs docker fqy jindan kubelet lipeiying liyui lvbo moyu security_evaluate_models
Qwen2_qx_hy chenhua data embed guoyihao jxr lcy liuhaotian liyx lzh nieyh share
agent chenshaolong dataset eval_datasets hpfs12637 jz lff liumf liziyi maoxy qx tools
root@vscode-1876449876017119232-1876449934674460673-b44d9654b-9mrjl:/work/share# cd ../
root@vscode-1876449876017119232-1876449934674460673-b44d9654b-9mrjl:/work# ls
cache share
root@vscode-1876449876017119232-1876449934674460673-b44d9654b-9mrjl:/work#
```

vscode启动单机训练任务

terminal 终端目前只支持单机训练任务，训练脚本跟tensorflow和pytorch 裸金属训练模式一样。

下方是在本机执行的单机多卡torchrun分布式参数示例：

```
DISTRIBUTED_ARGS="
    --nproc_per_node 8 \
    --nnodes 1 \
    --node_rank 0 \
    --master_addr localhost \
    --master_port 65500
"
torchrun $DISTRIBUTED_ARGS telespeed/run/llama31/pretrain_gpt.py
```

网页启动多机训练

下方是在平台执行多机多卡训练任务的torchrun分布式参数示例：

```
GPU_NUM_PER_NODE=8
DISTRIBUTED_ARGS="
    --nproc_per_node $GPU_NUM_PER_NODE \
    --nnodes $PET_NNODES \
    --node_rank $PET_NODE_RANK \
    --master_addr $PET_MASTER_ADDR \
    --master_port $PET_MASTER_PORT
"
torchrun $DISTRIBUTED_ARGS telespeed/run/llama31/pretrain_gpt.py
```

按照下图的步骤启动训练任务：

用户指南

基本信息

* 任务名称

描述 1 拟定训练任务名称

0/128

数据集配置

训练数据集 + 增加数据集 您代码中读取数据的相对路径需要改为本地挂载路径。

测试数据集 + 增加数据集 您代码中读取数据的相对路径需要改为本地挂载路径。

预置模型配置

模型来源 ☒ 我的模型 ☐ 预置模型

我的模型文件 + 增加我的模型文件 将模型管理中的模型文件挂载到容器本地路径，您的代码可以直接读取此相对路径。

环境配置

训练代码 切换至代码工程目录下，执行训练脚本

* 启动命令 2

用户指南

训练任务 / 创建任务

训练代码 请选择 1 2 前往上传

* 启动命令 1 cd /work/share/lcy/huiju/telespeedai/; bash examples/llama31/pretrain_llama31_8b_8k_ptd_dis.sh; sleep 900;

资源配置

* 镜像框架 ☒ 系统预置镜像 ☐ 自定义镜像 ☐ 共享的容器镜像 3 选择系统预置镜像或自定义镜像

ubuntu20.04-cann8.0.rc2-torch2.1.0-py3.8-mindspeed-npu:v5.0-gemma2-fa

* 训练模式 ☒ DDP ☐ 单机训练 4 选择DDP分布式启动方式

容错训练 ☐

* 资源组 5 选择资源组和队列

* 队列 fanqingyu

* 算力申请 * 算力规格 训练服务DO-LC-扩展型
Master节点数量 1 6 扩展型代表每机8卡，Master节点数+Worker节点数=总训练所用机器数

* Worker节点数量 1 +

高级配置

* 断点续训 ☒ 配置策略 7 根据需求，可选择是否打开断点续训

开启容错后，如因为节点故障导致训练任务异常，会封锁故障节点，重新调度训练任务。

后续操作

- 开始训练任务：点击之后会携带当前开发机相关参数，进入训练任务创建界面。

JupyterLab VSCode

创建VSCode 请输入名称搜索 请选择运行状态 查询 重置

名称	创建者	状态	运行时长	镜像版本	资源类型	操作
						启动 停止 打开 开始训练 更多

- 保存至模型管理：将当前开发机实例中的模型文件保存到模型仓库中统一管理，模型仓库中会新增一个来源为“VSCode”的模型。后续可以基于此模型进行开发机、训练任务和服务部署任务。

用户指南

名称

描述

存储目标类型

自有对象存储 (ZOS)

存储桶

新建版本

版本	来源	任务名称	导入状态	底座模型	导出状态	版本描述	目的地路径	操作
V2	VSCode	pmq_test0609	导入中	-	-	-		导出 分享 评估 部署 删除

- **保存至代码包：**从代码文件夹中选择对应的文件保存至代码包，代码包列表中出现了上传来源为“VSCode”的代码包。后续可以基于此代码包进行训练任务和服务部署。

我的代码包									
+ 导入代码包									
请输入代码包名称									
开始日期 至 结束日期									
查询 重置									
代码包名称	代码包ID	上传来源	启动文件名	代码包大小	导入状态	创建人	创建时间	操作	
pmq_test0906		VSCode	-	0KB	导入中		2025-06-09 16:17:58	详情 开始训练 在线编码 重新上传 删除	

- **制作镜像：**基于当前的开发机实例制作镜像并存储为自定义镜像，将容器内的镜像打包至自定义镜像中，下次直接启动该镜像即可，镜像仓库中新增来源为“VSCode”的镜像。后续可以基于此镜像进行开发机、训练任务和服务部署。平台会打包系统内的预置安装程序等，不会打包容器内数据和外部挂载存储数据。

JupyterLab

VSCode

创建VSCode

请输入名称搜索

请选择运行状态

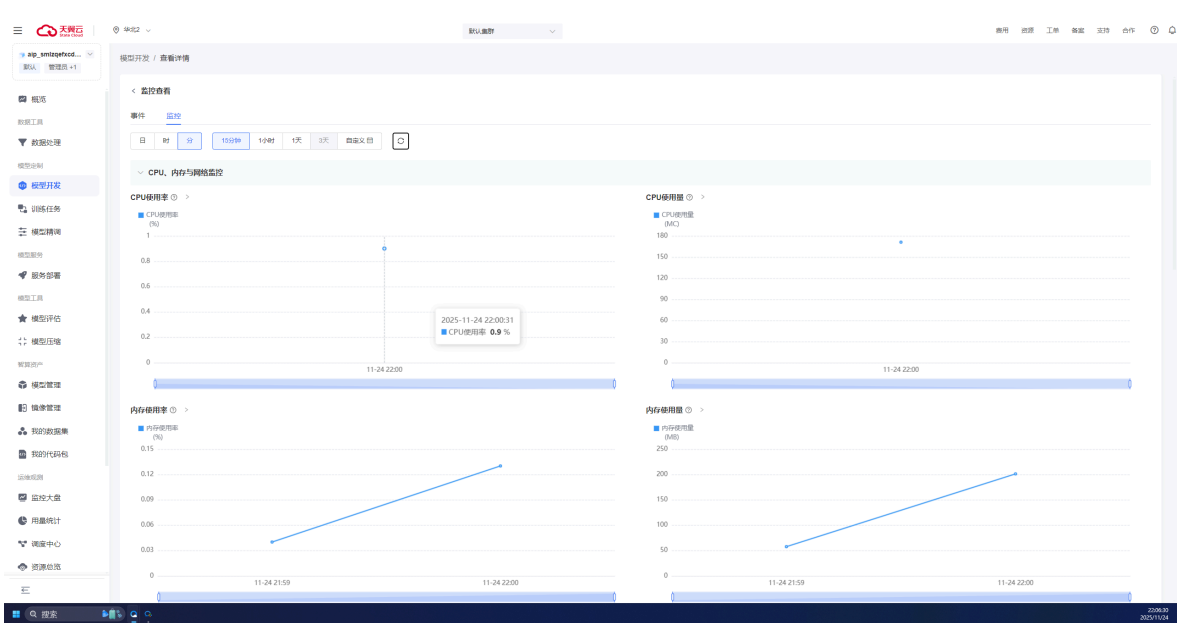
查询

重置

名称	创建者	状态	运行时长	镜像版本	资源类型	操作
						启动 停止 打开 开始训练 更多

用户指南

查看开发机监控

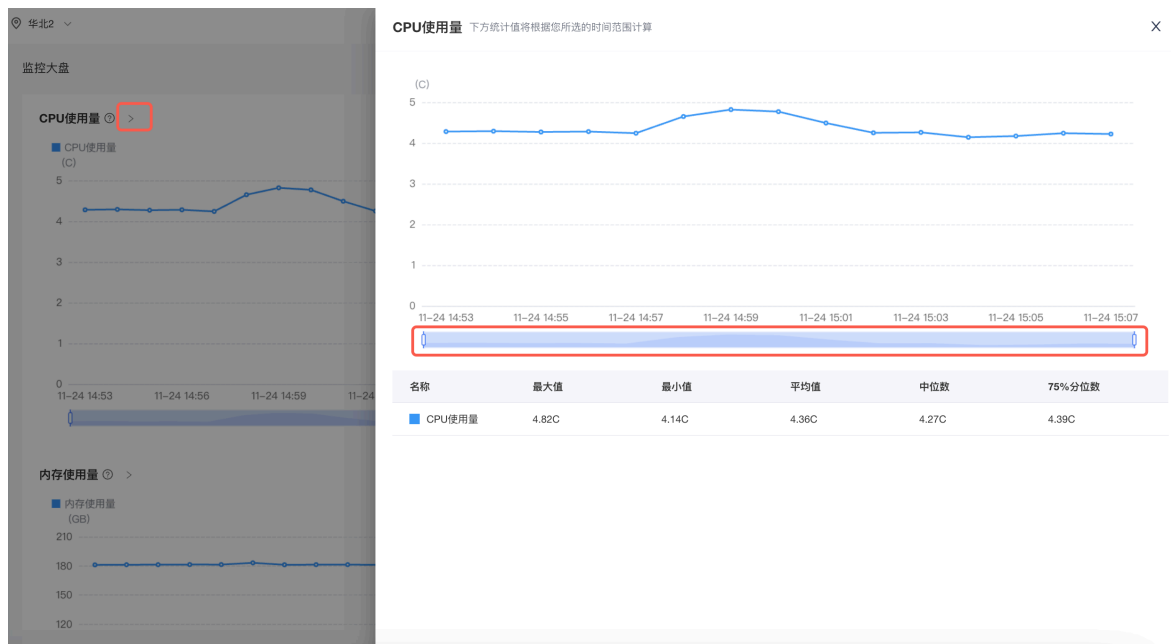


1. 点击开发机名称，进入查看详情页面，切换到监控tab，可查看开发机的资源使用情况。目前平台的开发机都是单实例（Pod），此处监控都以实例维度展示。

用户指南

2. 图像展示:

- 放大与明细: 点击指标右侧“>”箭头, 可展开指标大图, 大图展示对图像上点的统计细项, 包括最大值、最小值、平均值、中位数、75分位数;
- 图例: 点击图例, 可以对线段进行展示/隐藏;
- 时间轴: 滑动图像下方时间轴, 可以在已选定时间的基础上, 查看更小范围的监控。



3. 监控指标

类别	指标	解释
CPU、内存与网络监控	CPU使用率	CPU在单位时间内, CPU被任务占用使用的时间占比。
	CPU使用量	CPU 实际使用的核数。
	内存使用率	已用内存占总内存的百分比。
	内存使用量	内存实际使用量。
	普通网络吞吐	传统以太网的实际数据传输速率, 即单位时间内实际传输的数据量。
显卡基础指标	GPU/NPU使用率	在单位时间内, 显卡被任务占用使用的时间占比。
	GPU/NPU显存使用率	已用显存占总显存的百分比。
	GPU/NPU显存使用量	显存实际使用量。
	GPU/NPU卡温度	显卡温度。
	GPU/NPU功耗	显卡功耗。

用户指南

NPU卡健康状态

每张卡的NPU芯片健康状态。
取值范围：{0，1}
1：表示在过去一段时间间隔内芯片处于健康状态；
0：表示在过去一段时间间隔内出现了不健康状态。

远程登录开发机

- 前提
 - 申请白名单：当前远程登录开发机暂未全网开放，需要联系平台增加账户白名单方可正常使用。
 - 私有化版本：私有化版本平台需要确保部署了负载均衡和EIP，能够支持远程登录开发机功能。
 - 镜像准备：当前仅支持平台预置的三款预装ssh相关组件的镜像进行远程登录操作。
 - 状态：开发机处于运行中，且开发机成功绑定带EIP（弹性公网IP）的负载均衡ELB。
 - 余额：平台会为您开启EIP，需要账号有足够余额能够满足EIP的开通，只有当所有开发机被删除后EIP才会被退订。
- 远程登录开发机
 - 添加白名单：客户需要向平台申请加入运营后台开发机管理的白名单，添加进白名单的客户账号及其子账号可见平台远程登录的开关。需要提供相关信息给平台，

参数	说明
账户名	添加白名单后当前租户及其所有子账号可以开启开发机远程登录开关。
IP网段	输入当前用户的IP网段，例如：10.243.0.0/16。

- 本机生成密钥对：在本机上通过指令” ssh-keygen “生成密钥对，生成的文件在” ~/.ssh “文件夹下。

ssh-keygen

```
PS C:\Users\admin\.ssh> ls

目录：C:\Users\admin\.ssh

Mode                LastWriteTime         Length Name
----                -
-a-----          2025/11/28      11:14         411 id_ed25519 私钥文件
-a-----          2025/11/28      11:14         104 id_ed25519.pub 公钥文件
```

- 选择带有SSH标签的系统预置镜像，镜像中预置了SSH登录所需要的ssh依赖，目前仅支持带有SSH标签的预置镜像开启远程登录开发机功能：

序号	镜像名称
1	ubuntu22.04-teleformers-cann8.2.rc1-npu:v0.2.0.telechat25.post1.ssh

用户指南

序号	镜像名称
2	ubuntu22.04-teleformers-cann8.2.rc1-npu:v0.2.0.post1.ssh
3	ubuntu22.04-teleformers-cuda12.4-gpu:v0.5.ssh

- 配置SSH访问：当创建开发机时开启了"启用SSH"，可以通过详情页查看远程登录的访问配置。只有当开发机在运行中状态且EIP绑定成功时可以复制访问指令，绑定中可以通过点击“刷新”按钮获取最新的绑定状态，绑定失败时可以通过点击“刷新”按钮重试。

参数	说明
启用SSH	开关，默认关闭，开启后会产生EIP费用。
SSH公钥	输入框，最多可输入10个需要用分号分隔。

启用SSH后，开发机进入运行中状态且EIP绑定成功时可以查看公网访问指令：

ssh -i 这里填私钥地址 root@ip -p port

#私钥地址：ssh-keygen指定的私钥文件地址。

#username默认为root。

#ip为ELB绑定的EIP。

#Port为平台分配的ELB监听端口。

- 查看SSH访问配置：启用SSH后，开发机进入运行中状态且EIP绑定成功时可以查看公网访问指令。
-

访问配置

启用SSH ☒ 开启后会产生EIP费用 [查看计费方式](#)

* SSH公钥

公网访问 ☒ 绑定成功 ssh -i 这里填私钥地址 root@



参数	说明
启用SSH	开发机创建完成后用户不能自行关闭远程登录：1）可以联系平台运维同事后台操作关闭当前开发机的远程登录开关；2）或者删除当前账号的白名单，删除后当前账户机器子账号不具备远程登录开发机的能力，并且会关闭所有已经 开启远程登录开发机的开关。
SSH公钥	默认显示密文，公钥为平台为您创建的用于身份验证，实现远程服务器登录操作。
公网访问	访问开发机的指令，只有当状态为绑定成功时才能正常进行远程登录，其余状态：1）绑定中，表示当前EIP正在绑定，用户可点击刷新按钮重试获取绑定状态；2）绑定失败，表示当前EIP绑定失败，用户可以点击刷新按钮重试绑定。

- 本地访问开发机：输入公网访问指令

用户指南

```
PS C:\Users\admin\.ssh> ssh -i id_rsa root@
Warning: Identity file id_rsa not accessible: No such file or directory.
Welcome to Ubuntu 22.04.5 LTS (GNU/Linux 5.10.0-136.12.0.88.5.ctl3.aarch64 a
arch64)

* Documentation:  https://help.ubuntu.com
* Management:    https://landscape.canonical.com
* Support:       https://ubuntu.com/pro

This system has been minimized by removing packages and content that are
not required on a system that users do not log into.

To restore this content, you can run the 'unminimize' command.
Last login: Fri Nov 28 14:16:57 2025 from
(work) root@vscode-
```

- 费用：当开启了远程登录，平台会为用户创建一个EIP，只有当账号下所有开发机被删除后才会删除EIP，因此会产生网络费用。[计费详情查看](#)。

算力虚拟化

当前平台采用显存隔离的虚拟化策略，开启后多个eGPU实例会优先调度于同一物理GPU上；算力切分策略支持算力（最低1%）与显存（最低1MB）的细粒度切分。当开启自定义分配之后，可以根据自己的任务情况输入当前需要切分的虚拟卡配置：

* 自定义分配 ☒ 指对资源配额内的资源虚拟化切分使用（当前仅支持英伟达集群）
本方案采用显存隔离的虚拟化策略，开启后多个eGPU实例会优先调度于同一物理GPU上；
算力切分策略支持算力（最低1%）与显存（最低1MB）的细粒度切分

* 算力卡类型 L40s

资源	* 数量	单张物理卡量	说明
虚拟卡数量	<div><div>—</div><div>2</div><div>+</div></div> 个		输入当前任务需要使用的虚拟卡数量（不可超过当前节点的物理卡总量）；单张物理卡可切分的虚拟化卡数上限由显存切分的粒度决定（例如，一张128G显存的物理卡，每张虚拟卡申请16G显存，则最多可切分8张虚拟卡）；可用显存总量=虚拟卡数量*单卡显存
虚拟单卡显存	<div><div>—</div><div>5</div><div>+</div></div> G	44 G	输入占用每张虚拟卡的显存；输入≥ 1 且 ≤单张物理卡显存的整数

根据显存切分如下：默认按虚拟单卡显存占比自动计算（默认推荐值=单张物理卡CPU显存占比）

资源	* 数量	单张物理卡量	说明
虚拟单卡CPU	<div><div>—</div><div>1</div><div>+</div></div> 核 推荐值	9 G	输入每张虚拟卡占用的CPU核数，输入≥ 1 且 ≤单张物理卡CPU的整数
虚拟单卡内存	<div><div>—</div><div>8</div><div>+</div></div> G 推荐值	79 G	输入每张虚拟卡占用的内存；输入≥ 1 且 ≤单张物理卡内存的整数

☒ 默认展示系统推荐最佳值，谨慎修改，若修改可能会增加碎片降低算力卡使用效率

虚拟化规格【张*（卡类型_eGPU）_显存_核数_内存】：2*(L40s(eGPU)_5G_1C_8G)

- 虚拟卡数量：不可超过当前节点的物理卡总量，单张物理卡可切分的虚拟化卡数上限由显存切分的粒度决定(例如，一张128G显存的物理卡，每张虚拟卡申请16G显存，则最多可切分8张虚拟卡)；可用显存总量=虚拟卡数量*单卡显存。
- 虚拟单卡显存：每张虚拟卡的显存，大于等于1且小于等于单张物理卡显存。

- 根据用户输入的显存值，会自动计算出虚拟卡的单卡CPU和内存推荐值：默认推荐值=单张物理卡CPU或内存*显存占比。默认展示系统推荐最佳值，谨慎修改，若修改可能会增加碎片降低算力卡使用效率。可以点击”恢复推荐值“快速返回系统推荐值。

训练任务

前置条件

- 完成训练数据集准备，完成存储配置准备（ZOS/HPFS），详见[我的数据集](#)。
- 如果预置模型不满足开发要求，需要基于自有模型，需要完成模型文件准备，详见[模型管理](#)。
- 如果预置镜像不满足开发要求，需要基于自有镜像，需要完成镜像文件准备，详见[我的镜像](#)。
- 如果需要使用代码包，需要完成代码包准备，详见[我的代码包](#)。

创建训练任务

1. 登录智算服务平台。

用户指南

2. 创建训练任务入口：

- 入口一：在左侧菜单选择“模型定制”-“训练任务”，点击“新建任务”，进入任务创建页面。
- 入口二：在左侧菜单选择“模型定制”-“开发机”，点击开发机列表的“开始训练”。

训练任务 / 创建训练任务

基本信息

任务名称

请输入任务名称

描述

请输入描述信息

0/128

数据集配置

训练数据集 + 增加数据集 (0/10) 您在代码中读取数据的相对路径需要改为容器内挂载路径

模型配置

模型来源

☒ 我的模型 ☐ 预置模型

我的模型文件

+ 增加我的模型文件 (0/5) 将模型管理中的模型文件挂载到容器内路径。您的代码可以直接读取此相对路径

存储配置

自有ZOS对象存储 (0/5)

自有HPFS并行文件系统 (0/5)

环境配置

文件目录

专属高速缓存

☒

容器内挂载路径: /workhome

训练代码

请选择

前往上传

挂载本地路径: /workmount/code

启动命令

1

资源配置

镜像选择

☒ 系统预置镜像 ☐ 自定义镜像 ☐ 共享的容器镜像 ☐ 他人分享的镜像

ubuntu22.04-teleformers-cuda12.4-gpu-v0.5.ssh

训练模式

☒ DDP ☐ 单机训练

容错训练

☐

集群

L40S 专属集群

资源配额

当前工作空间无资源配额或资源配额已停用，去配置

算力申请

资源规格

请选择资源规格

规格详情

Master节点数量

1

Worker节点数量

-

1

+

参数类型	参数名	说明
基本信息	任务名称	必填，训练任务名称。
	描述	非必填，输入128个字符的描述。
数据集配置	训练数据集	最多可添加10个，选择基础数据集或者标注数据集。

用户指南

参数类型	参数名	说明
模型配置	模型来源	我的模型：最多5个，将模型管理中的模型文件挂载到容器内路径。 预置模型：最多5个，将预置模型挂载到容器内路径。
	模型文件	选择我的模型具体的模型文件及版本。 选择预置模型文件及具体版本。
存储配置	ZOS对象存储	最多选择5个，如果没有提前创建，可以点击“去创建对象存储”完成创建。
	HPFS并行文件系统	最多选择5个，如果没有提前创建，可以点击“去创建HPFS”完成创建。
环境配置	文件目录	平台可持久化的挂载目录，后续可以在该目录下读写文件，是用户间隔离的。
	训练代码	非必填，可以选择目标代码包。
	启动命令	必填。如果您的代码包是文件夹，则需要填写python xx.py，其中xx.py是您的训练代码；如果您的代码中有启动参数，可以直接填写；若您使用的代码包是压缩包文件，需要在启动命令中添加解压命令zip。
资源配置	镜像来源	支持选择系统预置镜像、自定义镜像、共享容器镜像和他人分享镜像。
	集群	支持公共集群和专属集群两种类型，其中专属集群需要提前购买。
	资源配额	选择目标资源配额，展示当前总资源及使用情况。
	资源规格	选择当前任务所需要的资源规格。
	训练模式	默认为DDP（分布式训练），如果在单一计算设备上机器学习模型训练选择单机训练。
	容错训练	启动容错训练后，如果训练过程中节点异常，系统会自动重新启用一个新的节点来替换异常节点，从上一个checkpoint开始继续训练。
高级配置	断点续训	支持昇腾NPU以及英伟达系列GPU的断点续训。 开启后，如因为节点故障导致训练任务异常，会封锁故障节点，重新调度训练任务；同时可以配置断点续训策略，支持设置【每次错误的最大重启次数】【容错重启策略】【容错策略】等参数。
	算力健康检查	检查昇腾机器节点的显卡状态、显卡通信状态和交换机状态，以及带宽的压测值。可训练任务详情页查看具体信息。

用户指南

训练任务管理

- 训练任务创建完成后进入“资源准备中”-“排队中”-“环境准备中”-“开始运行”-“运行完成”，在这个过程中如果发生任务配置错误、耗时过长或者资源冲突时，可以点击“停止”和“删除”。

新建任务		请输入任务名称		请选择运行状态		查询	重置	
名称	训练框架	训练模式	占用资源	状态	运行时长	创建来源	开始时间	操作
<div><div></div><div></div></div>	PyTorch	DDP	Master算力规格:14C_100G_1*A800 Master节点个数:1个 Worker算力规格:14C_100G_1*A800 Worker节点个数:1个	<div><div></div>运行完成</div>	0分0秒	JupyterLab	-	<div>日志复制tensorboard更多</div>
<div><div></div><div></div></div>	PyTorch	单机训练	算力规格:14C_100G_1*A800 节点个数:1个	<div><div></div>停止</div>	1分6秒	训练任务	2025-06-25 10:59:46	<div>日志复制删除保存至模型管理</div>
		Master算力规格:14C_100G_1*A800						

- 复制任务：如果已经创建了一个训练任务可以在列表中点击“复制”，复用该任务的配置和输入参数创建一个新的训练任务。

新建任务		请输入任务名称			请选择运行状态			查询	重置
名称	训练框架	训练模式	占用资源	状态	运行时长	创建来源	开始时间	操作	
<div><div></div></div>	PyTorch	DDP	Master算力规格:14C_100G_1A800 Master节点个数:1个 Worker算力规格:14C_100G_1A800 Worker节点个数:1个	运行完成	<div></div>	JupyterLab	-	日志 复制 tensorboard 更多	
			算力规格:14C_100G_1A800	2025-06-25 10:59					

- 任务详情：支持查看当前训练任务的配置参数（基础信息、数据集、挂载模型、节点信息）、事件、日志（节点维度）、监控（GPU使用率、CPU使用率内存、网络IO等）、查看Tensorboard和算力检查。任务每次启动都会生成一个新的实例，这里叫做一次运行记录。一次运行记录显示了实例ID、实例最终状态、启动时间。

训练任务 / 任务详情	
< train	
运行记录	
19377846	
运行完成	
2025-06-25 16:08:02	
基本信息 事件 日志 监控	
基础信息	
名称	ID
状态	运行完成
训练开始时间	-
训练模式	DDP
Master算力规格:14C_100G_1*A800	Master节点个数:1个
Worker算力规格:14C_100G_1*A800	Worker节点个数:1个
训练输出	/work/home
训练框架	ubuntu18.04.6-torch1.11.0-py3.8-cuda11.3-cudnn8-gpuv2.0-vscode_jupyter
运行时长	4835842621440

设置告警

- 配置告警发送事件：通过告警设置可以监控训练任务状态任务异常、任务断点续训和任务运行完成。默认这三类事件都不开启告警。

用户指南

- 配置告警发送方式：消息会自动发送到您天翼云账号对应的手机号和邮箱内，可前往账号中心进行设置或修改。

训练任务

运行中任务数 0

新建任务

设置告警通知

名称	训练框架	训练模式	占用资源	状态
PyTorch	DDP	算力规格 170C_1400G_8T800 节点个数 1个	资源准备中	
PyTorch	DDP	算力规格 21C_175G_1T800 节点个数 1个	运行完成	
PyTorch	单机训练	算力规格 2C8G 节点个数 1个	运行完成	
PyTorch	DDP	算力规格 170C_1400G_8T800 节点个数 1个	停止	

告警发送事件

任务异常

任务断点续训

任务运行完成

告警发送方式

消息会自动发送到您天翼云账号对应的手机号和邮箱内，可前往账号中心进行设置或修改

短信

139****6961

邮箱

a**an@chinatelecom.cn

后续操作

- 保存至模型管理：将当前训练任务实例中的模型文件保存到模型仓库中统一管理，模型仓库中会新增一个来源为“训练任务”的模型。后续可以基于此模型进行开发机、训练任务和服务部署任务。

我的模型 / 查看详情

名称

存储目标类型

自有对象存储 (ZOS)

描述

存储桶

+ 新建版本

版本	来源	任务名称	导入状态	基座模型	导出状态	版本描述	目的地路径	操作
V3	训练任务	的	导入中	-	-	-	2663597465 39768737587	导出 分享 评估 部署 删除

用户指南

查看训练任务监控

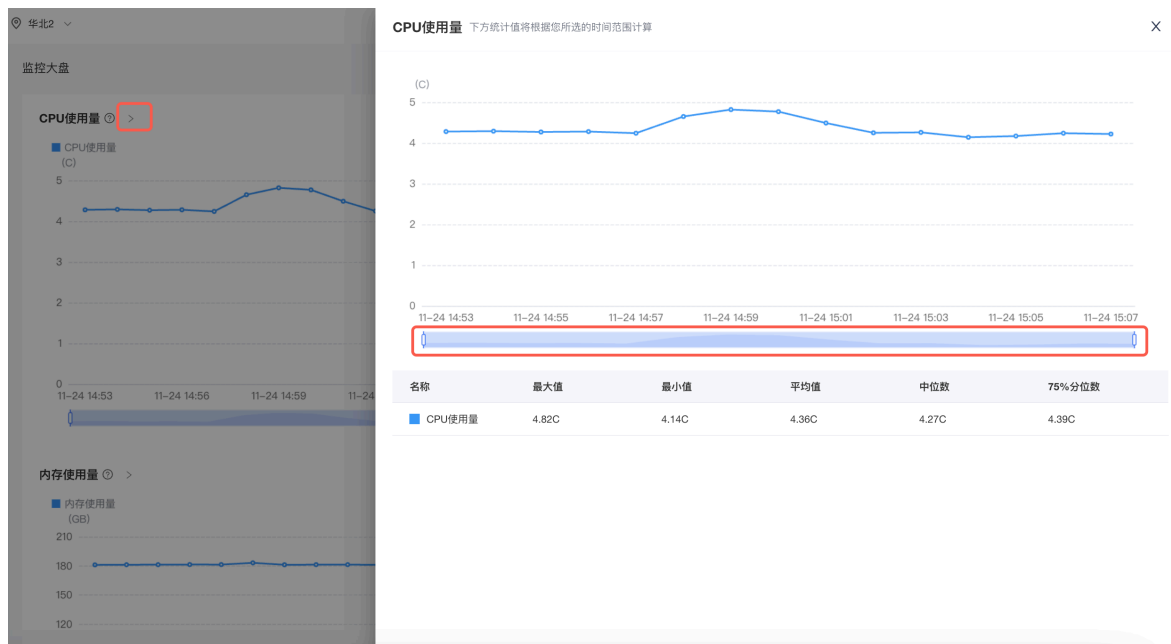


1. 点击训练任务名称，进入任务详情页面，切换到监控tab，可查看训练任务的资源使用情况。
2. 统计说明及名词解释：
 - a. 作业：指运行一次任务，由于训练任务可以多次运行，此处一次运行记录即一个作业；在左侧运行记录中切换其他运行ID，可查看不同作业的监控；
 - b. 实例：指pod实例，是Kubernetes的最小调度单元；
 - c. 作业维度与实例维度：一般在训练时，一次训练作业会起多个Pod实例。在实例维度，展示了不同实例的最小粒度的监控，此时可以精确到某个实例中的某一张卡；在作业维度，统计了该作业下的所有实例（Pod）或显卡的聚合值，以反映作业整体的资源使用情况，一般使用率、速率以平均值聚合，使用量以累加值聚合。
 - d. GPU细粒度指标：当您使用GPU卡时，可以查看GPU剖析类指标，此类指标需要您对GPU的结构有所了解。例如，GPU利用率显示了100%，不代表代码高效利用了GPU，仅能够展示有核函数在用GPU资源，但有多少个SM在使用，GPU是否真的在忙，仅看GPU利用率这一指标具有局限性。英伟达官方提供了DCGM Profiling指标，可从多个维度查看GPU的使用情况。您可以根据细粒度指标对GPU的利用情况进行剖析，优化代码逻辑，更充分的利用资源。

用户指南

3. 图像展示:

- 放大与明细: 点击指标右侧“>”箭头, 可展开指标大图, 大图展示对图像上点的统计细项, 包括最大值、最小值、平均值、中位数、75分位数;
- 图例: 点击图例, 可以对线段进行展示/隐藏;
- 时间轴: 滑动图像下方时间轴, 可以在已选定时间的基础上, 查看更小范围的监控。



4. 监控指标:

类别	指标	维度	解释
CPU、内存与网络监控	CPU使用率	作业、实例	CPU在单位时间内, CPU被任务占用使用的时间占比。
	CPU使用量	作业、实例	CPU 实际使用的核数。
	内存使用率	作业、实例	已用内存占总内存的百分比。
	内存使用量	作业、实例	内存实际使用量。
	普通网络吞吐	作业、实例	传统以太网的实际数据传输速率, 即单位时间内实际传输的数据量。
显卡基础指标	GPU/NPU使用率	作业、实例	在单位时间内, 显卡被任务占用使用的时间占比。
	GPU/NPU显存使用率	作业、实例	已用显存占总显存的百分比。
	GPU/NPU显存使用量	作业、实例	显存实际使用量。

用户指南

	GPU/NPU卡温度	实例	显卡温度。
	GPU/NPU功耗	实例	显卡功耗。
	NPU卡健康状态	实例	<p>每张卡的NPU芯片健康状态。</p> <p>取值范围：{0, 1}</p> <p>1：表示在过去一段时间间隔内芯片处于健康状态；</p> <p>0：表示在过去一段时间间隔内出现了不健康状态。</p>
GPU细粒度指标	GPU应用时钟频率	作业、实例	显卡的核心频率，指GPU运算核心的工作速度，即GPU内部各个计算单元的工作频率。它决定了显卡在单位时间内可完成的任务数量，更高的时钟频率意味着更快的计算速度和更好的性能。
	GPU显存应用时钟频率	作业、实例	显存频率指的是显卡所使用的显存的工作频率，显存频率影响了显卡的数据传输速度，包括图像纹理和帧缓冲区的读写速度。较高的显存频率可以提供更快的数据传输，频率越高，显存每秒传输的数据量越大。
	GPU显存带宽利用率	作业、实例	以英伟达GPU V100为例，其最大内存带宽为900 GB/sec，如果当前的内存带宽为450 GB/sec，则内存带宽利用率为50%。
	GPU引擎活跃情况	作业、实例	<p>表示在一个时间间隔内，Graphics或Compute引擎处于Active的时间占比。</p> <p>Graphics或Compute引擎处于Active是指Graphics或Compute Context绑定到线程，并且Graphics或Compute Context处于Busy状态。</p> <p>该值表示所有Graphics和Compute引擎的平均值。</p>
	GPU线程束活跃时间占比	作业、实例	<p>表示在一个时间间隔内，至少一个线程束在一个SM（Streaming Multiprocessor）上处于Active的时间占比。</p> <p>线程束处于Active是指一个线程束被调度且分配资源后的状态，可能是在Computing、也可能是非Computing状态（例如等待内存请求）。</p> <p>该值表示所有SM的平均值，小于0.5表示未高效利用GPU，大于0.8是必要的。</p> <p>假设一个GPU有N个SM：</p> <p>一个核函数在整个时间间隔内使用N个线程块运行在所有的SM上，此时该值为1（100%）。</p> <p>一个核函数在一个时间间隔内运行N/5个线程块，此时该值为0.2。</p> <p>一个核函数使用N个线程块，在一个时间间隔内，仅运行了1/5个周期的时间，此时该值为0.2。</p>
	GPU线程束占用率	作业、实例	<p>表示在一个时间间隔内，驻留在SM上的线程束与该SM最大可驻留线程束的比例。</p> <p>该值表示一个时间间隔内的所有SM的平均值。</p> <p>占用率越高不代表GPU使用率越高。只有在GPU内存带宽受限的工作负载（DCGM_FI_PROF_DRAM_ACTIVE）情况下，更高的占用率表示更有效的GPU使用率。</p>

用户指南

GPU张量通道活跃周期分数	作业、实例	<p>表示Tensor（HMMA/IMMA）Pipe处于Active状态的周期比率。</p> <p>该值表示一个时间间隔内的平均值，较高的值表示Tensor Cores的利用率较高。</p> <p>该值为1（100%）表示在整个时间间隔内每隔一个指令周期发出一个Tensor指令（两个周期完成一条指令）。</p> <p>假设该值为0.2（20%），可能有如下情况：</p> <p>在整个时间间隔内，有20%的SM的Tensor Core以100%的利用率运行。</p> <p>在整个时间间隔内，有100%的SM的Tensor Core以20%的利用率运行。</p> <p>在整个时间间隔的1/5时间内，有100%的SM上的Tensor Core以100%利用率运行。</p> <p>其他组合模式。</p>
GPU显存拷贝活跃周期分数	作业、实例	<p>表示显存带宽利用率将数据发送到设备显存或从设备显存接收数据的周期分数。</p> <p>该值表示时间间隔内的平均值，较高的值表示设备显存的利用率较高。</p> <p>该值为1（100%）表示在整个时间间隔内的每个周期执行一条DRAM指令（实际上，峰值约为0.8（80%）是可实现的最大值）。</p> <p>假设该值为0.2（20%），表示20%的周期在时间间隔内读取或写入设备显存。</p>
GPU FP64通道活跃周期分数	作业、实例	<p>注意：并非所有型号的显卡都有此数据，如A10与L40S型号不支持此精度。</p> <p>表示FP64（双精度）Pipe处于Active状态的周期分数。</p> <p>该值表示一个时间间隔内的平均值，较高的值代表FP64 Cores有较高的利用率。</p> <p>该值为1（100%）表示在整个时间间隔内上每四个周期（以Volta类型卡为例）执行一次FP64指令。</p> <p>假设该值为0.2（20%），可能有如下情况：</p> <p>在整个时间间隔内，有20%的SM的FP64 Core以100%的利用率运行。</p> <p>在整个时间间隔内，有100%的SM的FP64 Core以20%的利用率运行。</p> <p>在整个时间间隔的1/5时间内，有100%的SM上的FP64 Core以100%利用率运行。</p> <p>其他组合模式。</p>
GPU FP32通道活跃周期分数	作业、实例	<p>表示乘加操作FMA管道处于Active的周期分数，乘加操作包括FP32（单精度）和整数。</p> <p>该值表示一个时间间隔内的平均值，较高的值代表FP32 Cores有较高的利用率。</p> <p>该值为1（100%）表示在整个时间间隔内上每两个周期（Volta类型卡为例）执行一次FP32指令。</p> <p>假设该值为0.2（20%），可能有如下情况：</p> <p>在整个时间间隔内，有20%的SM的FP32 Core以100%的利用率运行。</p> <p>在整个时间间隔内，有100%的SM的FP32 Core以20%的利用率运行。</p> <p>在整个时间间隔的1/5时间内，有100%的SM上的FP32 Core以100%利用率运行。</p> <p>其他组合模式。</p>

用户指南

GPU FP16通道活跃周期分数	作业、实例	<p>表示FP16（半精度）管道处于Active的周期分数。</p> <p>该值表示一个时间间隔内的平均值，较高的值代表FP16 Cores有较高的利用率。</p> <p>该值为 1 (100%) 表示在整个时间间隔内上每两个周期（Volta类型卡为例）执行一次FP16指令。</p> <p>假设该值为0.2（20%），可能有如下情况：</p> <p>在整个时间间隔内，有20%的SM的FP16 Core以100%的利用率运行。</p> <p>在整个时间间隔内，有100%的SM的FP16 Core以20%的利用率运行。</p> <p>在整个时间间隔的1/5时间内，有100%的SM上的FP16 Core以100%利用率运行。</p> <p>其他组合模式。</p>
GPU PCIe传输数据速率	作业、实例	<p>表示通过PCIe总线传输的数据速率，包括协议标头和数据有效负载。</p> <p>该值表示一个时间间隔内的平均值，而不是瞬时值。</p> <p>该速率在时间间隔内平均。例如，在1秒内传输1 GB数据，则无论以恒定速率还是突发传输数据，速率都是1 GB/s。理论上的最大PCIe Gen3带宽为每通道985 MB/s。</p>
GPU PCIe接收数据速率	作业、实例	<p>表示通过PCIe总线接收的数据速率，包括协议标头和数据有效负载。</p> <p>该值表示一个时间间隔内的平均值，而不是瞬时值。</p> <p>该速率在时间间隔内平均。例如，在1秒内传输1 GB数据，则无论以恒定速率还是突发传输数据，速率都是1 GB/s。理论上的最大PCIe Gen3带宽为每通道985 MB/s。</p>
GPU NVLINK传输数据速率	作业、实例	<p>表示通过NVLink传输的数据速率，不包括协议标头。</p> <p>该值表示一个时间间隔内的平均值，而不是瞬时值。</p> <p>该速率在时间间隔内平均。例如，在1秒内传输1 GB数据，则无论以恒定速率还是突发传输数据，速率都是1 GB/s。理论上，最大NVLink Gen2带宽为每个方向每个链路25 GB/s。</p>
GPU NVLINK接收数据速率	作业、实例	<p>表示通过NVLink接收的数据速率，不包括协议标头。</p> <p>该值表示一个时间间隔内的平均值，而不是瞬时值。</p> <p>该速率在时间间隔内平均。例如，在1秒内传输1 GB数据，则无论以恒定速率还是突发传输数据，速率都是1 GB/s。理论上，最大NVLink Gen2带宽为每个方向每个链路25 GB/s。</p>

断点续训

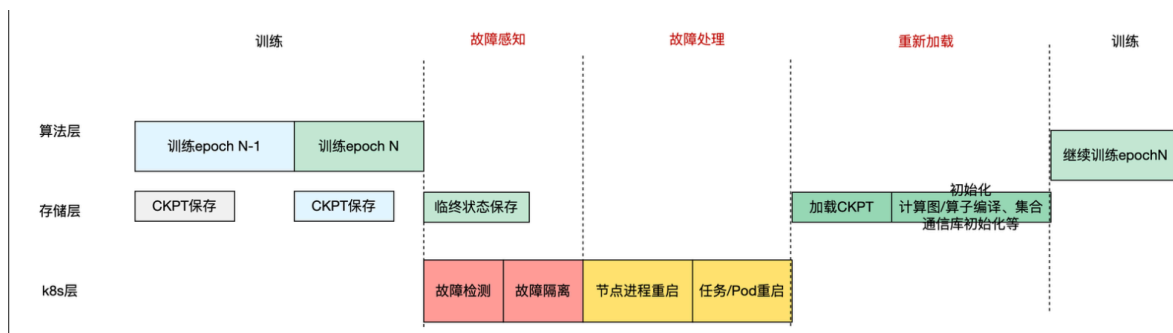
产品背景

大模型训练过程往往比较长，根据业界数据，平均约两天，大模型训练会中断一次，经常遇到的问题例如：

- 程序出现故障
- 节点出现故障
- loss数据恶化，需要手动暂停，调参重训
- 服务器故障、网络中断、开发平台故障等

用户指南

断点续训过程需要排查节点、隔离节点、重新启动上一次CKPT，浪费了资源和训练时间，是大模型训练的重要痛点之一。针对这一问题，平台提供了断点续训能力并支持配置客户需要的策略。



断点续训

平台对英伟达和昇腾芯片的节点故障、集群故障、程序故障等多种故障场景，支持故障感知、故障定位、故障修复。目前支持PyTorch、Mindspore框架。

支持以下故障类型：

- 网络故障：当网卡链接出现故障或状态异常（link status: DOWN）后，可正常触发重调度，实现断点续训
- 节点心跳故障：当节点因Label异常等原因导致心跳丢失后，可正常触发重调度，实现断点续训
- 节点Shutdown/Reboot故障：当节点被关闭或重启后，可正常触发重调度，实现断点续训
- 芯片PCIE故障：当节点发生芯片丢失等异常后，可正常触发重调度，实现断点续训
- 断点故障：当错误发生时断点正在写入，未正常完成保存即断点保存不完整时，可正常恢复至上一完整断点，实现断点续训

优雅容错

优雅容错功能可以在昇腾NPU芯片级别的故障发生时，优先尝试恢复故障芯片（需要在支持的场景下），从而实现训练任务的进程级别的重启，无需触发集群的重调度来切换新的节点再进行训练。

优雅容错有以下优点：

- 进程级别杀死和恢复，相比重调度切换节点的方式减少了耗时
- 底层组件尝试芯片故障热恢复，减少机器因故障隔离导致利用率下降

目前优雅容错支持PyTorch、Mindspore，支持以下故障类型：

- L3故障--tsdaemon异常：可正常恢复进程并断点续训
- L4故障--slogd心跳异常：可正常继续训练，L4级别故障不影响任务，故障等待任务结束后处理并恢复
- L5故障--内存超限异常：可正常恢复进程并断点续训
- L5故障--跨板心跳异常：可正常恢复进程并断点续训
- L5故障--小核心心跳异常：恢复失败，节点隔离并触发重调度

所有故障处理级别

故障级别	故障类型	说明	重调度处理	优雅容错处理
L1	NotHandleFault	对业务无影响的故障，无需处理	暂不处理	暂不处理

用户指南

故障级别	故障类型	说明	重调度处理	优雅容错处理
L2	RestartRequest	影响业务执行，需要重新执行业务请求	隔离设备，进行任务重调度	推理场景重执行推理请求，训练场景重新执行训练业务
L3	RestartBusiness	影响业务执行，需要重新执行业务	隔离设备，进行任务重调度	重新执行业务
L4	FreeRestartNPU	影响业务执行，待芯片空闲时需复位芯片	隔离设备，进行任务重调度	复位芯片后重新执行业务
L5	RestartNPU	影响业务执行，需立即复位芯片	隔离设备，进行任务重调度	复位芯片后重新执行业务
L6	SeparateNPU	无法恢复，需要隔离设备	隔离设备，进行任务重调度	隔离设备，进行任务重调度

单Pod重调度

在英伟达系列GPU芯片上训练任务发生故障时，可以支持单Pod重调度，开启单Pod重调度，只重新调度故障的Pod，节约重调度效率和任务重启耗时，提升容错效率。

模型服务

服务部署

前置条件

- 如果预置镜像不满足开发要求，需要基于自有镜像，需要完成镜像文件准备，详见[我的镜像](#)。
- 如果需要使用代码包，需要完成代码包准备，详见[我的代码包](#)。

创建在线服务

1. 登录训推智算服务平台。

用户指南

2. 创建在线服务入口：

- 入口一：在左侧菜单选择“模型服务”-“服务部署”，点击“部署模型”，进入在线服务创建页面。
- 入口二：在左侧菜单选择“模型管理”，点击模型卡片的“部署”，进入在线服务创建页面。

服务部署 / 部署模型

模型服务信息

* 服务名称

镜像环境

* 模型选择

我的模型 预置模型

* 模型配置

自定义配置

模板配置

我的模型文件

请选择模型

新建模型

挂载本地路径: /work/mount/model/

* 镜像来源

系统预置镜像

自定义镜像

共享的容器镜像

他人分享的镜像

请选择镜像

代码包选择

请选择代码包

挂载本地路径: /work/mount/code/

例如: numpy==1.16.4 absl-py==0.11.0

0/5000

环境变量

变量名称	变量值	操作
	暂无数据	
+添加变量		

* 运行命令

1

请输入端口号

3. 当前模型部署支持将不同来源、不同类型的模型转化为可对外提供服务的模型应用，满足多样化的业务场景需求：

- 我的模型：
 - 自定义配置：部署的模型来自于模型管理中用户通过开发机和训练任务生成的个性化的自有模型。选择“我的模型”列表下的目标模型文件，要求客户对自有模型的训练逻辑、部署时的资源规格有清晰的认知，否则会直接影响部署效果。该方法只适用于单机部署，如果想使用多机部署，请选择”模板配置“。

参数名	说明
服务名称	必填，在线服务名称。
模型选择	选择模型管理中的“我的模型”文件，挂载到容器内路径。

用户指南

参数名	说明
模型配置	选择“自定义配置”。部署的模型来自于模型管理中用户通过开发机和训练任务生成的个性化的自有模型。
模型文件	选择自定义配置，则可选择模型管理中“我的模型”下的模型文件。
镜像来源	支持选择系统预置镜像、自定义镜像、共享容器镜像和他人分享镜像。
代码包选择	非必填，可以选择目标代码包。
环境变量	输入变量名称和值。
运行命令	必填，用以启动镜像的运行命令和端口号（例如：启动镜像的运行命令脚本。示例如下： <code>cd /work/mount/code #切换到执行脚本所在目录。python /run.py#执行启动脚本。</code> ）
资源配额	选择目标资源配额，展示当前总资源及使用情况。
资源规格	选择当前任务所需要的资源规格。
实例数量	输入当前部署实例的数量。
服务授权	必填，选择调用服务时使用的服务授权。

- 模板配置：平台预置了分布式部署模型模板，平台预置模板经过优化测试，部署后服务稳定性高。当前只支持DeepSeek-R1 模型的分布式部署，单实例部署机器数为2。例如，部署3个实例且每个实例需2台机器，则共需 $3 \times 2 = 6$ 台机器。当前的模板配置为平台预置，无法调整模板参数。如果想快速体验单机部署可以选择“预置模型”。

- | 参数名 | 说明 |
|------|--|
| 服务名称 | 必填，在线服务名称。 |
| 模型选择 | 选择模型管理中的“我的模型”文件，挂载到容器内路径。 |
| 模型配置 | 选择“模板配置”。当前只支持DeepSeek-R1 模型的分布式部署，单实例部署机器数为2。 |
| 模型文件 | 选择模板配置所需要的模型文件。 |
| 资源配额 | 选择目标资源配额，展示当前总资源及使用情况。 |
| 资源规格 | 选择当前任务所需要的资源规格。 |
| 实例数量 | 输入当前部署实例的数量。 |
| 服务授权 | 必填，选择调用服务时使用的服务授权。 |

- 预置模型：部署的模型来自于平台的“预置模型”，无需用户自主训练模型，操作简单。预置模型通常是通用性模型，用户无法根据自身业务需求进行优化调整，如果需要针对垂直领域进行优化调整，请选择“自定义配置”。

- | 参数名 | 说明 |
|------|------------|
| 服务名称 | 必填，在线服务名称。 |

用户指南

参数名	说明
模型选择	选择模型管理中的“预置模型”文件，挂载到容器内路径。
训练框架	根据用户选择的模型文件，会自动带出训练所用框架。
资源规格	选择当前任务所需要的资源规格。
实例数量	输入当前部署实例的数量。
服务授权	必填，选择调用服务时使用的服务授权。

- 在列表可查看模型是否部署成功，在操作列可进行模型查看、更新、停止、重启、修改、上下线、删除等操作。停止服务后计费也会停止，再次启动服务即可开通计费。
- 操作列点击【查看】可进入该服务的详情页，查看部署的模型列表、服务监控、配置历史、运行记录、事件日志、服务日志。

算力虚拟化

当开启“自定义分配”表示对资源配额内的资源虚拟化切分使用（当前仅支持在英伟达专属集群下使用）。当前平台采用显存隔离的虚拟化策略，开启后多个eGPU实例会优先调度于同一物理GPU上；算力切分策略支持算力（最低1%）与显存（最低1MB）的细粒度切分。当开启自定义分配之后，可以根据自己的任务情况输入当前需要切分的虚拟卡配置：

* 自定义分配 ☒ 预对资源配额内的资源虚拟化切分使用（当前仅支持英伟达集群）
本方案采用显存隔离的虚拟化策略，开启后多个eGPU实例会优先调度于同一物理GPU上；
算力切分策略支持算力（最低1%）与显存（最低1MB）的细粒度切分

* 算力卡类型 L40s

资源	* 数量	单张物理卡量	说明
虚拟卡数量	<div><div>-</div><div>2</div><div>+</div></div> <div>个</div>		输入当前任务需要使用的虚拟卡数量（不可超过当前节点的物理卡总量）；单张物理卡可切分的虚拟化卡数上限由显存切分的粒度决定（例如，一张128G显存的物理卡，每张虚拟卡申请16G显存，则最多可切分8张虚拟卡）；可用显存总量=虚拟卡数量*单卡显存
虚拟单卡显存	<div><div>-</div><div>5</div><div>+</div></div> <div>G</div>	44 G	输入占用每张虚拟卡的显存；输入 ≥ 1 且 \leq 单张物理卡显存的整数

根据显存切分如下：默认按虚拟单卡显存占比自动计算（默认按单张物理卡CPU或内存/显存占比）

资源	* 数量	单张物理卡量	说明
虚拟单卡CPU	<div><div>-</div><div>1</div><div>+</div></div> <div>核</div> <div>推荐值</div>	9 G	输入每张虚拟卡占用的CPU核数，输入 ≥ 1 且 \leq 单张物理卡CPU的整数
虚拟单卡内存	<div><div>-</div><div>8</div><div>+</div></div> <div>G</div> <div>推荐值</div>	79 G	输入每张虚拟卡占用的内存；输入 ≥ 1 且 \leq 单张物理卡内存的整数

☒ 默认展示系统推荐最佳值，谨慎修改，若修改可能会增加碎片降低算力卡使用效率

虚拟化规格【张*（卡类型_eGPU_显存_核数_内存）】：2*(L40s(eGPU_5G_1C_8G)

- 虚拟卡数量：不可超过当前节点的物理卡总量，单张物理卡可切分的虚拟化卡数上限由显存切分的粒度决定(例如，一张128G显存的物理卡，每张虚拟卡申请16G显存，则最多可切分8张虚拟卡)；可用显存总量=虚拟卡数量*单卡显存。
- 虚拟单卡显存：每张虚拟卡的显存，大于等于1且小于等于单张物理卡显存。

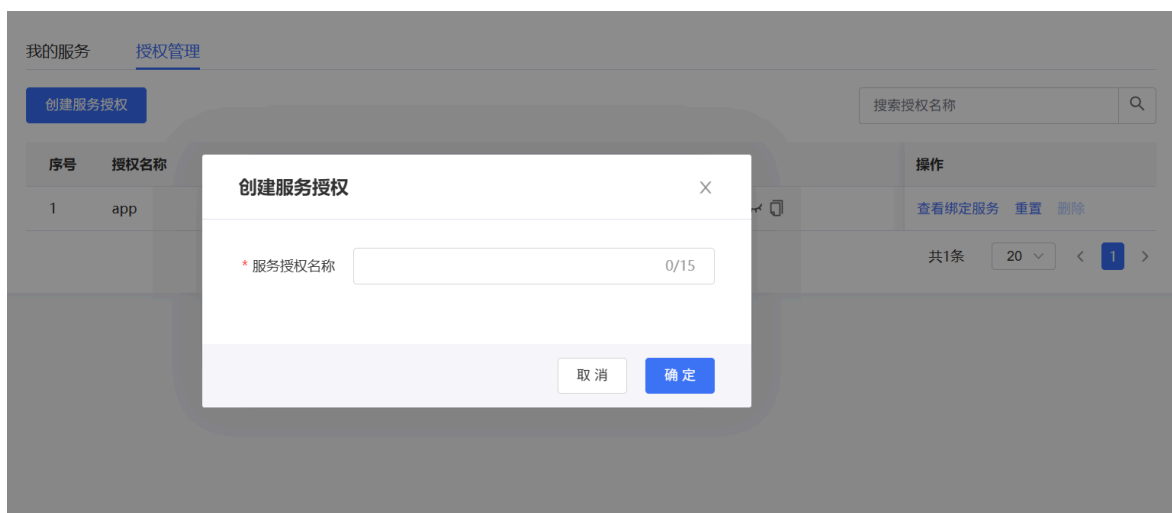
用户指南

- 根据用户输入的显存值，会自动计算出虚拟卡的单卡CPU和内存推荐值：默认推荐值=单张物理卡CPU或内存*显存占比。默认展示系统推荐最佳值，谨慎修改，若修改可能会增加碎片降低算力卡使用效率。可以点击”恢复推荐值“快速返回系统推荐值。

服务授权

服务授权为服务部署创建App Key供用户调用。

- 创建在线服务入口：
 - 入口一：在左侧菜单选择“服务部署”-“我的服务”-“部署模型”，新建页面点击“创建服务授权”。
 - 入口二：在左侧菜单选择“服务部署”-“授权管理”，点击“创建服务授权”，进入服务授权创建页面。
 - 入口三“在左侧菜单选择“服务部署”-“我的服务”，详情页点击“服务授权”页面中点击“创建服务授权”。
- 创建服务授权：
 - 输入服务授权名称：必填，最长不超过15个字符。

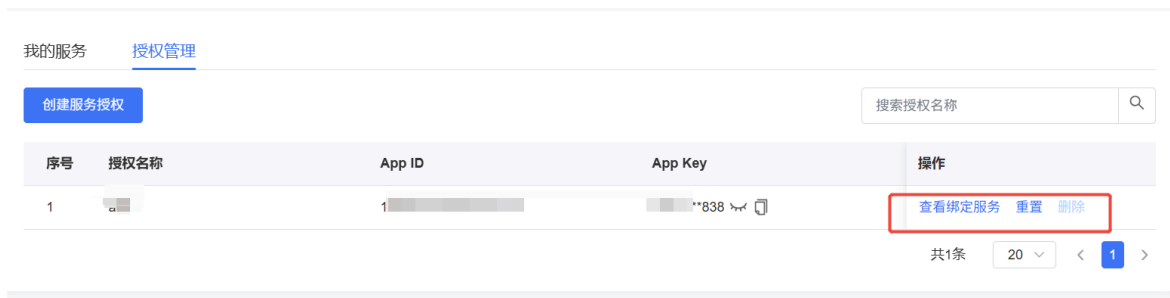


- 查看服务授权：在授权管理菜单还可以查看当前授权绑定的服务列表。

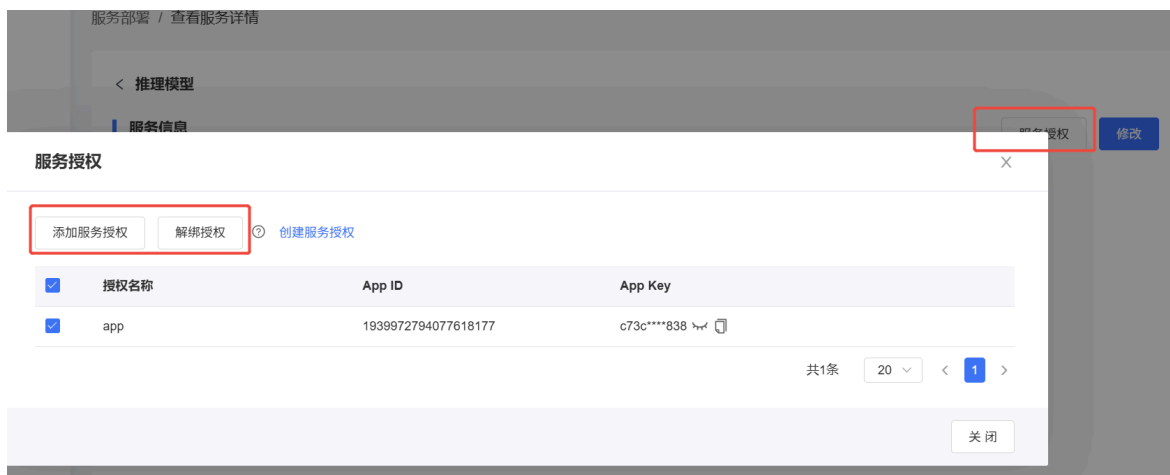


用户指南

- 可以单击操作列的按钮查询当前授权的明文、重置或删除。



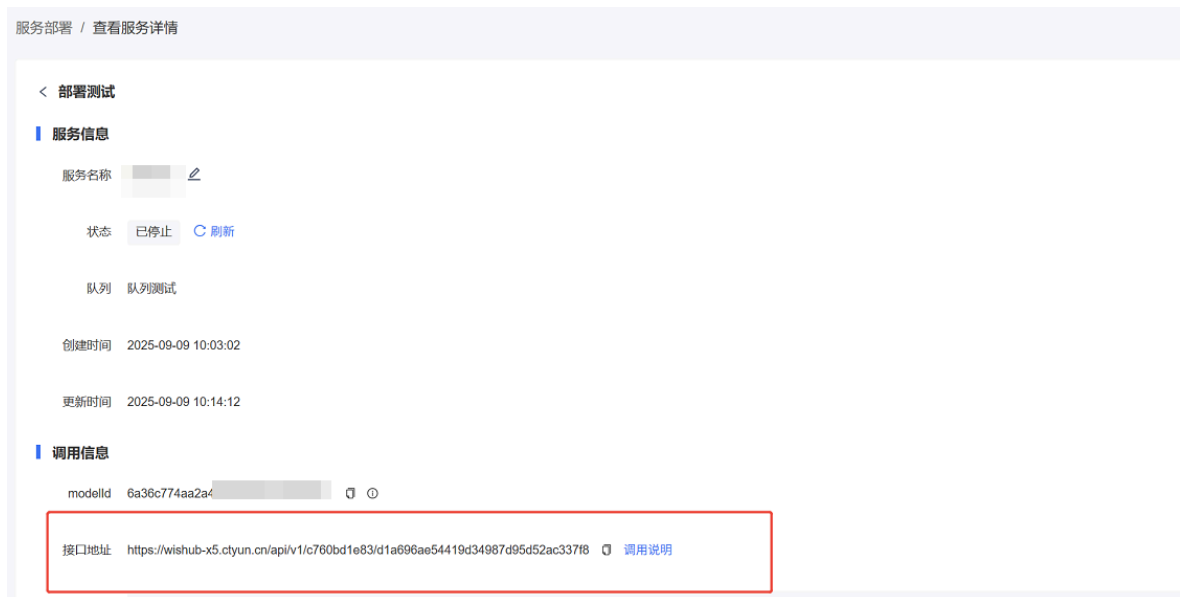
- 服务授权解绑和添加：在服务部署详情页，可以点击“服务授权”按钮进入当前服务绑定的授权信息列表，可以解绑不再使用的授权，绑定新的授权。



用户指南

服务调用

- 获得服务调用接口地址



- 在postman中构造请求接口，在请求header中需要填入Authorization鉴权信息，为避免被安全护栏拦截，建议在http请求header中填入User-Agent信息：

请求路径：https://wishub-x5.ctyun.cn/api/v1/c760bd1e83/d1a696ae54419d34987d95d52ac337f8/v1/xxx/yyy，其中/xxx/yyy 为具体的功能路径，如/chat/completions;其余前缀通过服务详情的“接口地址”获得。

请求方式：POST;

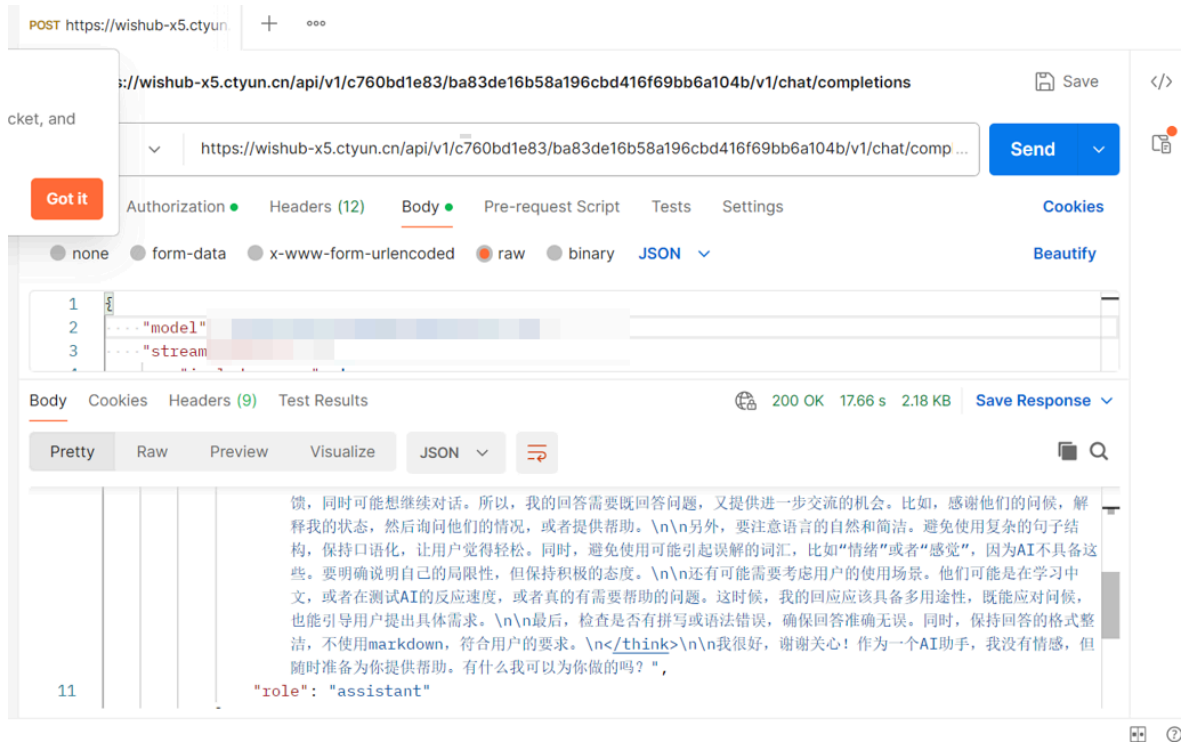
请求header必填项： Authorization: Bearer AppKey Content-Type: application/json;

其他header：

User-Agent: PostmanRuntime-ApipostRuntime/1.1.0

- 获得服务调用返回

用户指南



服务监控

推理服务提供两类监控指标，一类是调用监控，展示了与业务及性能强相关的指标，例如调用tokens量、QPS等；一类是资源监控，展示了CPU、内存、显卡等相应资源的使用情况。

用户指南

调用监控

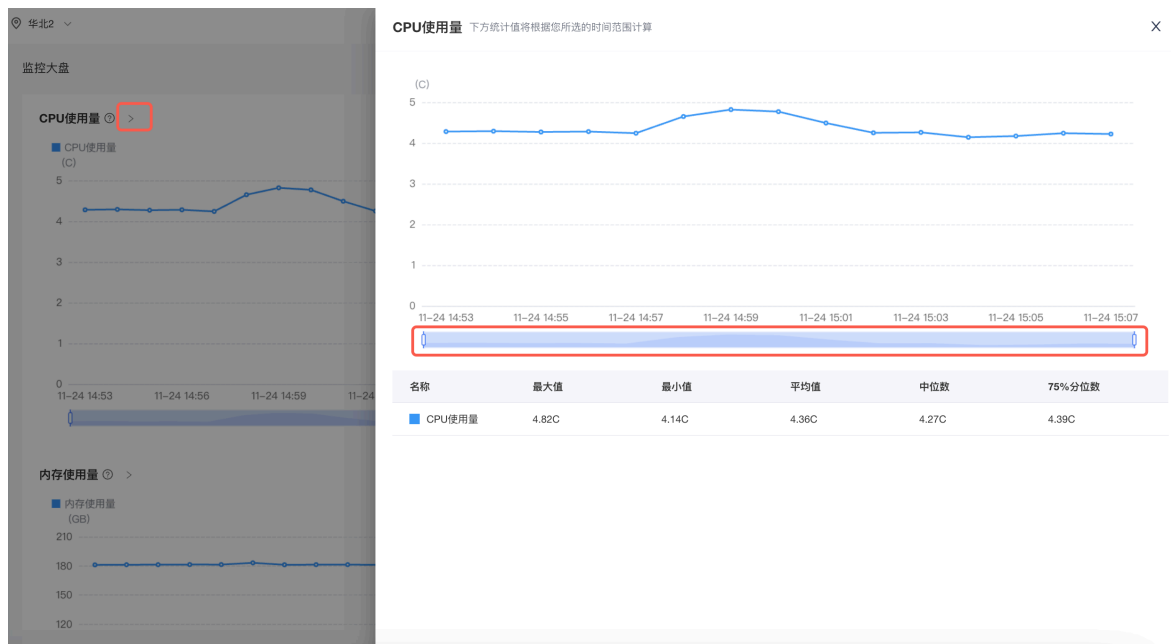


1. 找到需要查看的任务卡片，点击【查看】进入服务详情页面，下拉页面切换到【调用监控】tab，可查看部署服务的调用情况。
2. 选择相应的服务授权，不选则默认统计所有的服务授权。
3. 统计说明及名词解释：
 - a. 流式响应：模型边推理边输出结果（如逐词、逐句生成），用户可以实时看到部分结果；
 - b. 非流式响应：模型需要完整计算所有结果后，一次性返回全部内容。

用户指南

4. 图像展示:

- 放大与明细: 点击指标右侧“>”箭头, 可展开指标大图, 大图展示对图像上点的统计细项, 包括最大值、最小值、平均值、中位数、75分位数;
- 图例: 点击图例, 可以对线段进行展示/隐藏;
- 时间轴: 滑动图像下方时间轴, 可以在已选定时间的基础上, 查看更小范围的监控。



5. 指标含义:

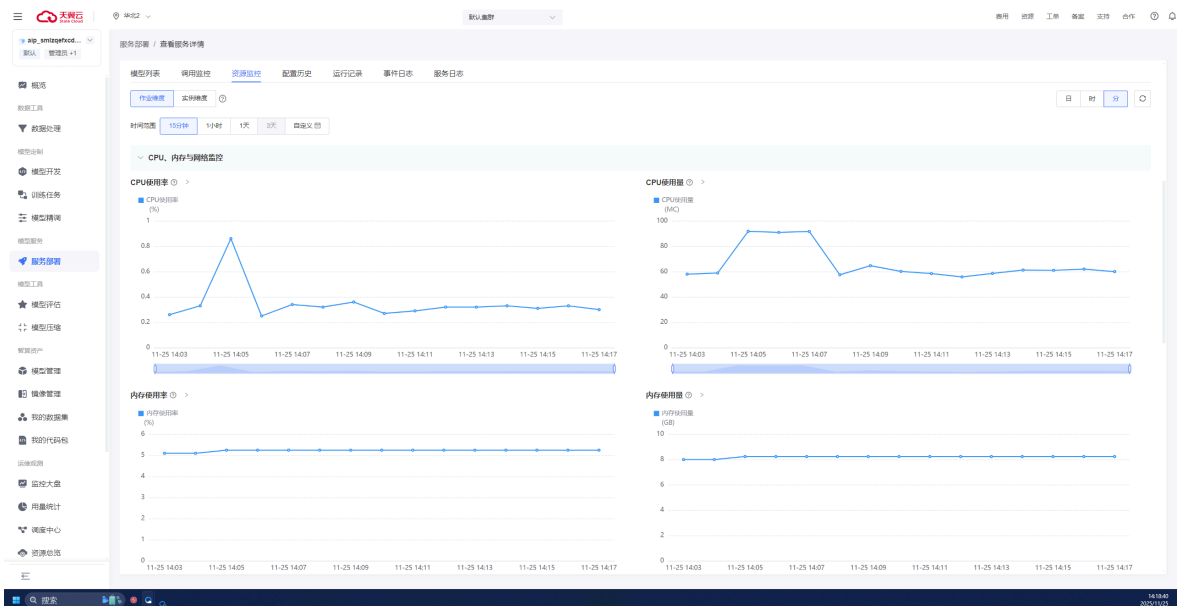
指标	解释
调用次数	一段时间内服务被调用的次数, 包含调用总次数、成功次数、失败次数。
累计调用次数	服务启动后, 服务累计被调用的次数, 包含调用总次数、成功次数、失败次数。 例如, 服务于1/1日启动, 2/1日统计值为1/1-2/1日时间段内的调用次数总和。
调用失败率	调用失败率。
平均响应时间	单位时间内成功请求的响应时间平均值。
调用Token量	一段时间内服务被调用的tokens数, 包含总量、输入、输出。
累计调用Token量	服务启动后, 服务累计被调用的tokens数, 包含总量、输入、输出。 例如, 服务于1/1日启动, 2/1日统计值为1/1-2/1日时间段内的调用tokens总和。

用户指南

首Token时延	<p>从接收请求到生成第一个输出token所需的时间，仅统计流式响应。</p> <p>AVG：首Token时延的平均值。</p> <p>MAX：首Token时延的最大值。</p> <p>P50：50%的首Token时延低于该值。</p> <p>P80：80%的首Token时延低于该值。</p> <p>P90：90%的首Token时延低于该值。</p> <p>P99：99%的首Token时延低于该值。</p>
非首Token时延	<p>生成后续每个输出token所需的时间间隔，不包括首token，仅统计流式响应。</p> <p>AVG：非首Token时延的平均值。</p> <p>MAX：非首Token时延的最大值。</p> <p>P50：50%的非首Token时延低于该值。</p> <p>P80：80%的非首Token时延低于该值。</p> <p>P90：90%的非首Token时延低于该值。</p> <p>P99：99%的非首Token时延低于该值。</p>
端到端时延	<p>服务在该时间点所有请求的端到端时延，端到端时延表示从接收到请求到返回所有Token的时间，统计流式及非流式响应。</p> <p>AVG：端到端时延的平均值。</p> <p>MAX：端到端时延的最大值。</p> <p>P50：50%的端到端时延低于该值。</p> <p>P80：80%的端到端时延低于该值。</p> <p>P90：90%的端到端时延低于该值。</p> <p>P99：99%的端到端时延低于该值。</p>
QPS	Query Per Second，即每秒响应请求数。

用户指南

资源监控

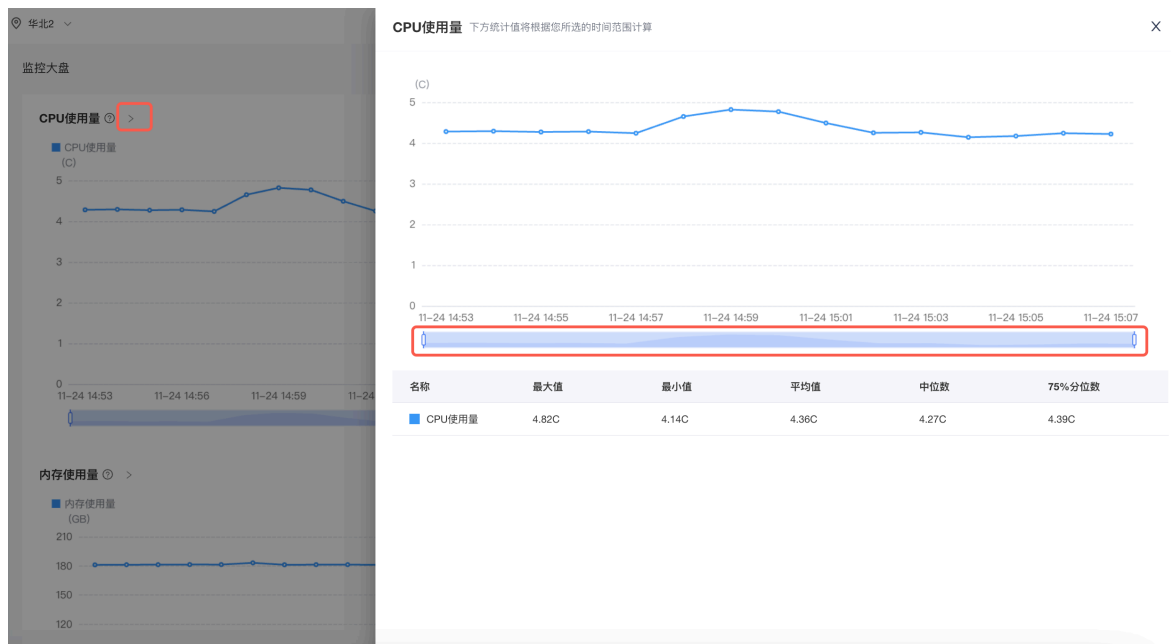


1. 找到需要查看的任务卡片，点击【查看】进入服务详情页面，下拉页面切换到【资源监控】tab，可查看部署服务的资源使用情况。
2. 统计说明及名词解释：
 - a. 作业：指运行一次任务，一次运行即一个作业；
 - b. 实例：指pod实例，是Kubernetes的最小调度单元；
 - c. 作业维度与实例维度：一般在部署时，一个任务作业会起多个pod实例。在实例维度，展示了不同实例的最小粒度的监控，此时可以精确到某个实例中的某一张卡；在作业维度，统计了该作业下的所有实例（Pod）或显卡的聚合值，以反映作业整体的资源使用情况，一般使用率、速率以平均值聚合，使用量以累加值聚合。

用户指南

3. 图像展示:

- 放大与明细: 点击指标右侧“>”箭头, 可展开指标大图, 大图展示对图像上点的统计细项, 包括最大值、最小值、平均值、中位数、75分位数;
- 图例: 点击图例, 可以对线段进行展示/隐藏;
- 时间轴: 滑动图像下方时间轴, 可以在已选定时间的基础上, 查看更小范围的监控。



4. 监控指标:

类别	指标	维度	解释
CPU、内存与网络监控	CPU使用率	作业、实例	CPU在单位时间内, CPU被任务占用使用的时间占比。
	CPU使用量	作业、实例	CPU 实际使用的核数。
	内存使用率	作业、实例	已用内存占总内存的百分比。
	内存使用量	作业、实例	内存实际使用量。
	普通网络吞吐	作业、实例	传统以太网的实际数据传输速率, 即单位时间内实际传输的数据量。
显卡基础指标	GPU/NPU使用率	作业、实例	在单位时间内, 显卡被任务占用使用的时间占比。
	GPU/NPU显存使用率	作业、实例	已用显存占总显存的百分比。
	GPU/NPU显存使用量	作业、实例	显存实际使用量。
	GPU/NPU卡温度	实例	显卡温度。
	GPU/NPU功耗	实例	显卡功耗。

用户指南

	NPU卡健康状态	实例	每张卡的NPU芯片健康状态。 取值范围：{0, 1} 1：表示在过去一段时间间隔内芯片处于健康状态； 0：表示在过去一段时间间隔内出现了不健康状态。
--	----------	----	---

模型工具

模型评估

模型评估旨在对平台精调生成的大模型输出效果进行评估，当前支持对“模型精调”运行完成的模型进行评估。

前置条件

- 数据准备：创建支持评估的指令微调标注数据集，并发布成功，详见[标注数据集](#)。
- 模型准备：您可从模型广场选择支持评估的模型直接发起评估；也可在支持评估的模型完成[模型精调](#)后，再在评估页面创建评估任务。

新建评估任务

- 您可通过以下方式使用：
 - 入口1：进入“模型工具”>“模型评估”菜单，进入评估任务列表，点击【新建评估任务】，进入创建页面；
 - 入口2：进入“智算资产”>“我的模型”菜单，在我的模型菜单页面中，找到支持评估的模型，点击【评估】按钮，进入到创建评估页面。
- 进入创建评估任务页面后，填写相关配置：

类型	字段	说明
基本信息	任务名称	评估任务名称，仅支持英文、汉字、数字、-、_，且只能以中英文、数字开头，不超过60个字符，不可重名
	模型类型	选择评估的模型类型，目前支持大语言模型
	描述	该任务的描述，300个字符以内
评估配置	待评估模型	选择待评估的模型，支持从模型广场精调后的带有评估标签的模型，目前支持Qwen2-7B-Instruct、Qwen2-72B-Instruct、Qwen1.5-32B-Chat。
	评估数据集	选择用于评估的数据集，目前仅支持标注模板为指令微调且发布成功的标注数据集

用户指南

	评估标准	指对模型评估的效果指标，支持准确率、ROUGE-1、ROUGE-2、ROUGE-L、BLEU-4，指标的具体含义可见页面解释
	停用词表	评估时自动忽略的停用词，为避免特殊字符及单词对模型效果评估的影响，可按照示例设置停用词表，评估时将自动过滤
资源部署信息	集群	选择适合的集群
	资源配额	选择可用资源配额
	资源规格	选择资源规格类型，不同的算力规格对应不同的价格
	实例数量	相应资源规格的数量

3. 点击开始评估，创建评估任务。

查看评估任务

- 查看详情：点击评估任务名称，在评估详情页，可以查看评估任务的详细内容，包括任务详情、事件、日志。
- 操作：评估任务卡片上，支持以下操作：
 - 查看报告：任务运行完成后，可查看相应的评估报告，支持查看整体指标和错题分析；
 - 重启：重启任务；
 - 停止：停止评估任务；
 - 复制：复制此评估任务配置，可以进行少量修改快速启动；
 - 删除：删除此评估任务。

模型压缩

模型压缩旨在帮助客户在尽量不减少模型效果的前提下压缩模型大小，进而提升模型在推理调用时的性能。

前置条件

- 资源准备：本功能仅支持910b硬件。
- 模型准备：您需要先对支持评估的模型完成[模型精调](#)后，再在评估页面创建评估任务。

创建模型压缩任务

- 您可通过以下方式使用：
 - 入口1：进入“模型工具”>“模型压缩”菜单，进入评估任务列表，点击【创建压缩任务】，进入创建页面；
 - 入口2：进入“智算资产”>“我的模型”菜单，在我的模型菜单页面中，找到支持压缩的模型，点击【压缩】按钮，进入到创建页面。
- 进入创建压缩任务页面后，填写相关配置：

类型	字段	说明
----	----	----

用户指南

基本信息	任务名称	压缩任务名称，不超过20个字符
	任务描述	该任务的描述，300个字符以内
模型配置	选择源模型	此处支持选择用户希望压缩的模型，目前仅支持Llama3-8B-Instruct基座模型精调出的模型
	模型创建方式	选择压缩后模型的保存方式，支持保存为已有模型新版本（默认为最新版本）或保存为新模型（默认V1版本）
	选择已有模型/新模型名称	选择已有模型：同一模型各版本的基础模型需保持一致，已自动过滤不符合要求的模型； 新模型名称：保存为新模型的名称

用户指南

压缩策略配置	压缩方法	<p>选择压缩方法，支持W8A16与稀疏化：</p> <ul style="list-style-type: none"> W8A16：WxCxH中W、A、C分别代表模型权重（weight）、激活（activation）和键值缓存（kv cache），数字x代表模型压缩后相应部分的比特数。模型压缩过程后，高比特浮点数会映射到低比特量化空间，从而达到降低显存占用、提升推理性能等目的。模型的推理性能收益均需要通过实际测试获得，表中策略类型仅做参考。W8A16的权重使用8位整数量化，但是激活值仍然保留较高的精度，通常是使用16位浮点数（FP16）或者混合精度（MixedPrecision）。这种方式相比于W8A8，可能会提供更好的模型准确性，但是代价是在激活值上保留了更高的精度，因此可能不会像W8A8那样大幅度减少内存使用。 稀疏化：是指在深度学习模型中通过将模型参数中的一些元素设为零来减少模型的计算量和存储需求，从而实现模型轻量化的目的。稀疏化不仅能有效地压缩模型，还能够加速推理过程，尤其是在硬件资源有限的设备上（如嵌入式系统或移动设备）。通过算法判断模型权重中每个元素对精度结果的重要性，并将模型权重中对最终精度影响小的权重值置零的过程。
资源配置	集群	选择适合的集群
	资源配额	选择可用资源配额
	资源规格	选择资源规格类型，不同的算力规格对应不同的价格

3. 点击提交，创建压缩任务。

查看压缩任务

- 查看任务详情：在任务列表点击任务名称，进入详情页面，可查看任务配置及任务日志。
- 操作：可在列表操作列中点击【详情】以查看任务详情，点击【停止】以停止任务。

3. 获取压缩结果：压缩任务运行完成后，压缩后的模型会自动保存到用户指定的模型管理中的位置。可以通过开发机挂载模型、或者下载模型来查看模型文件本身。压缩后的模型可以直接通过我的服务进行服务部署，部署为推理效果更优的大模型服务。

训推加速

CTCCL优化套件

CTCCL简介

CTCCL关键特性

- 主动避障，RDMA网络多路径传输，当感知到部分路径异常，则在条件允许情况下自动将流量切换到正常路径。
- 并行传输，动态感知不同RDMA网络路径的传输能力，合理分配传输任务，从端侧保证带宽利用率最大化。
- 监控能力，日志机制联合事件机制，提供网卡对集合通信带宽监控和QP通信异常事件上报智能平台能力。
- 端网协同，在RoCE组网下实现端网协同负载均衡，降低哈希冲突带来的影响，提高链路利用率。
- 故障定位，结合慢节点工具套件，提供自动化训练中慢节点发现与定位能力。

CTCCL发布记录

版本号	发布日期	更新内容
v0.4.0	2025-9-30	<ul style="list-style-type: none">· 新增功能<ul style="list-style-type: none">- 适配CTCCL慢节点检测工具套件。· 优化改进<ul style="list-style-type: none">- 为流体重力功能增加开关，使用环境变量配置，以便灵活使用该功能。

用户指南

版本号	发布日期	更新内容
v0.3.0	2024-12-30	<ul style="list-style-type: none">· 新增功能<ul style="list-style-type: none">– 新增QP通信事件上报功能，在机间RDMA通信异常时上报异常事件至平台。仅在一体化计算加速平台·异构计算平台部署的地域可用。– 新增集合通信网卡对带宽功能，用户可通过配置环境变量开启，并通过日志查看带宽信息。– 新支持RoCE组网端网协同，有效改善交换机端口流量不均问题，提高带宽利用率。· 缺陷修复<ul style="list-style-type: none">– 修复了QP数设置大于32直接异常退出的问题。– 修复了alltoall集合通信操作时，由于资源开销大而导致的性能低问题。· 优化改进<ul style="list-style-type: none">– 优化流体重力算法，以更灵活的动态任务分配方式，在拥塞场景提高通信性能10%。
v0.2.0	2024-06-30	<ul style="list-style-type: none">· 新增功能<ul style="list-style-type: none">– 新增流体重力算法，并行传输，动态感知不同RDMA网络路径的传输能力，合理分配传输任务，从端侧保证带宽利用率最大化。· 优化改进
v0.1.0	2024-04-30	<ul style="list-style-type: none">· 新增功能<ul style="list-style-type: none">– 天翼云自研集合通信库CTCCL首次发布。– CTCCL具有主动避障功能，提升RDMA通信容错能力。支持单QP传输，当感知到部分路径异常，则在条件允许情况下自动将流量切换到正常路径

升级提示：

- 在升级CTCCL新版本之前，请确保已停止该环境所有的训练任务，升级方式和安装方法相同。
- 需要升级集群中所有节点的CTCCL版本，新版本和旧版本不兼容在同个训练任务中使用。

安装、升级与使用CTCCL库

镜像安装

训推服务部分官方镜像默认安装CTCCL，使用以下官方镜像时，无需自己安装CTCCL。

用户指南

pytorch-2.2.0-ctccl0.3.0-cuda12.3-ubuntu20.04-amd64 llama2-training-ctccl0.3.0-cuda12.1-pytorch2.2.1-megatron0.2.0-deepspeed0.14.1-ubuntu20.04-amd64

手动安装

1.确认原本nccl/ctccl 安装位置

如果使用的深度学习框架自带NCCL，配置的NCCL目录可能不是默认路径，可以用以下命令查找。

```
find / -name "libnccl.so.2" ###$NCCL_PATH
```

2.确认使用的CUDA版本

使用以下命令，查看当前环境依赖的CUDA版本，输出结果如下图所示。

```
nvidia-smi
```

```
root@ /# nvidia-smi
Mon Dec 23 14:56:44 2024
+-----+
| NVIDIA-SMI  Driver Version:  CUDA Version: 12.2  |
+-----+-----+
| GPU   Name                Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf            Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
|                                           | MIG M.         |
+-----+-----+
=====
```

3.下载CUDA版本对应的CTCCL

用户根据操作系统和安装的CUDA版本下载对应的CTCCL独立安装包，安装并使用。

4.使用下载的CTCCL替换环境中原有的libnccl。

CTCCL	依赖环境	下载地址
v0.4.0	Ubuntu 20.04 + CUDA12.2	https://jiangsu-10.zos.cyun.cn/ctccl-n/ctccl/cuda12.2/ubuntu20.04/ctccl0.4.0/libnccl.so.2

```
cp libnccl.so.2 $NCCL_PATH
```

如果没有查找到libnccl.so.2文件，可以直接使用cp命令复制到usr/lib64。

5.确认是否替换成功

在训练日志中，看到日志输出对应的ctccl信息。

如：CTCCL version 0. .0(nccl_2.19.4)+cuda12.2

CTCCL版本升级

对于已安装NCCL或CTCCL旧版本的情况，若要安装CTCCL新版本，请确保已停止该环境所有的训练任务，升级方式和安装方法相同。

CTCCL测试

在NVIDIA平台上，可以使用NCCL-Test工具测试CTCCL的性能。

测试代码下载路径：

```
https://github.com/NVIDIA/nccl-tests.git编译：make MPI=1 MPI_HOME={{MPI##}} CUDA_HOME={{CUDA#
#}} NCCL_HOME={{NCCL##}} -j 40使用 mpirun 启动训练进程：mpirun --allow-run-as-root -np 2 -H IP1,IP2 -
x NCCL_IB_HCA=mlx5_2 -x NCCL_IB_QPS_PER_CONNECTION=8 all_reduce_perf -b 8 -e 1G -f 2 -g 8
```

CTCCL环境变量设置

环境变量	描述	推荐值
NCCL_IB_GID_INDEX	RDMA协议使用的GID	3
NCCL_IB_HCA	RDMA通信使用的网卡	mlx5
NCCL_IB_TIMEOUT	RDMA连接超时时间，合理配置可以提高训练任务的容错能力	22
NCCL_SOCKET_IFNAME	使用该端口建立连接	bond0
NCCL_DEBUG	日志级别	INFO/WARN
NCCL_DEBUG_SUBSYS	打印的info信息子类别，设置为REPORT可以输出网卡对带宽信息（若带宽统计功能开启）、事件上报日志（若上报功能开启）	REPORT
NCCL_IB_QPS_PER_CONNECTION	单连接使用的并行传输QP数量。若要使用QP切换与重传功能，请不要配置为1	8
NCCL_NET_PLUGIN	配置网络插件	none
CTCCL_ERR_REPORT	默认为0，配置为1后，CTCCL内部发现异常上报云骁平台。使用云骁智能平台拉起训练任务时，该功能默认打开。	0
CTCCL_BW_REPORT	默认为0，配置为1后，统计网卡对集合通信带宽信息并记录在日志中，日常正常训练不建议开启。使用云骁智能平台拉起训练任务时，该功能默认打开。	0
CTCCL_IB_LB_UPLINK	默认为0，RoCE组网下，推荐配置为leaf交换机上行链路数。IB环境下请配置为0，或不做配置。	IB: 0 RoCE: 上行链路数
CTCCL_QPTIME_REPORT	默认为0，要使用慢节点检测工具时请配置为1，必须搭配慢节点检测工具套件使用。	0

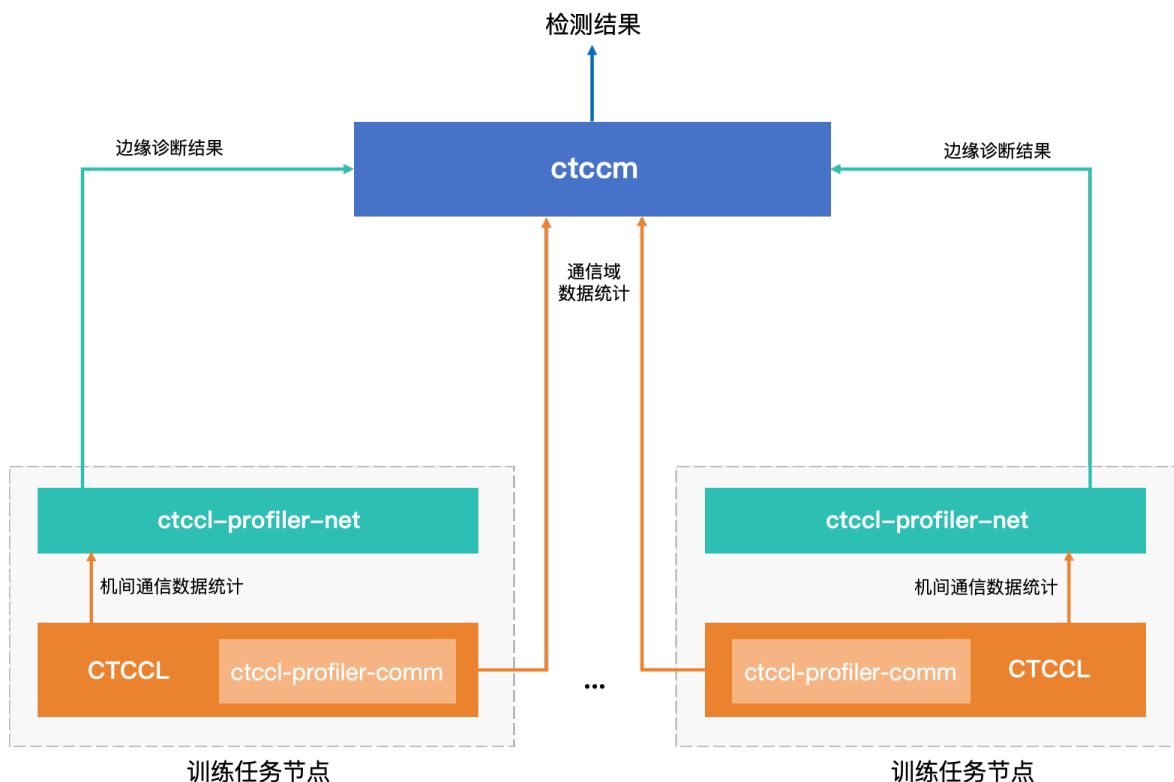
慢节点检测工具套件 CTCCL-Slowdetect

CTCCL-Slowdetect简介

功能介绍

CTCCL-Slowdetect-1.0.0工具套件架构如图所示，包括三个模块。

用户指南



· **ctccm**：跨节点的集中式慢节点诊断工具。部署在和所有训练任务节点网络互通的节点上，每个大模型训练任务只需部署一个ctccm服务。负责收集汇总和集中发现与定位慢节点问题。

· **ctccl-profiler-comm**：通信域级别集合通信信息统计工具。在每个训练任务所在的节点/容器内安装ctccl-profiler-comm插件，在训练任务的脚本中import相关库并调用API接口使用，负责获得训练任务的通信域任务拓扑，以及收集通信域级别的超时集合通信操作事件，提供给ctccm进行慢节点分析诊断。

· **ctccl-profiler-net**：机间通信边缘诊断工具。在每个训练任务所在的节点/容器内部署ctccl-profiler-net服务，每个训练任务节点都需要部署1个独享的ctccl-profiler-net服务。负责在边缘处理与初步诊断机间通信数据。

在大模型训练任务中使用CTCCL-Slowdetect-1.0.0工具套件，能够实现大模型训练过程中通信原因与非通信原因的慢节点检测。在发现集群中慢节点存在时，自动化开启全量集合通信操作事件监控，进一步分析定位慢节点。

参数设置

使用CTCCL-Slowdetect工具套件时，支持配置的参数如下：

服务	参数命令	描述
CTCCL	*CTCCL_QPTIME_REPORT	默认为0。要使用慢节点检测工具套件时，请配置为1，打开CTCCL机件通信信息统计功能。
ctccl-profiler-comm	*CTCCL_SLOWDETECT_SERVERADDR	ctccm的ip地址。示例：http://192.168.0.1:8000。其中端口需和ctccm的监听端口配置相同。

用户指南

服务	参数命令	描述
	CTCCL_PROFILER_TIMEOUT_RATE	默认为20。表示当一次集合通信操作超过一个迭代时间的20%时，判断该次集合通信为异常事件。由于每次迭代的总通信占比约为20%，故推荐设置为20，可根据集群与模型规模自行调整。
ctccl-profiler-net	*CTCCL_SLOWDETECT_SERVERADDR	同ctccl-profiler-comm
	*CTCCL_PROFILER_NETADDR	监听训练任务信息入口。 默认为http://127.0.0.1:8001。
	--log-level	日志级别。info（默认）、debug、warning、critical、error
ctccm	*--port	服务使用的监听端口
	--debug	是否开启debug日志
	*--nodes	训练任务的节点数。默认为1。

* 推荐自行配置。

诊断结果样例

计算慢:

```
INFO - [CTCCM SLOWDETECT INFO] profiler data is prepared, starting detect ...
INFO - [CTCCM SLOWDETECT INFO] profiler detector start analysis slow node based on step 62 to step 63
DEBUG - [CTCCM SLOWDETECT DEBUG] profiler matrix is [[0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]]
INFO - [CTCCM SLOWDETECT INFO] Rank0 compute slow or tp comm slow
INFO - [CTCCM SLOWDETECT INFO] Rank1 compute slow or tp comm slow
INFO - [CTCCM SLOWDETECT INFO] Rank2 compute slow or tp comm slow
INFO - [CTCCM SLOWDETECT INFO] Rank3 compute slow or tp comm slow
INFO - [CTCCM SLOWDETECT INFO] step : 63, tp check result = CTCCProfilerResult.NO_SLOW_NODE
INFO - [CTCCM SLOWDETECT INFO] step : 63, dp check result = CTCCProfilerResult.TP_COMPUTE_SLOW
INFO - [CTCCM SLOWDETECT INFO] qptime detector start analysis slow node base on step 62 to step 63
INFO - [CTCCM SLOWDETECT INFO] No Slow QP/device detected
```

通信慢:

```

2025-09-15 16:25:15.633 - INFO - [CTCCM_SLOWDETECT INFO] step : 34, tp check result = CTCCProfilerResult.NO_SLOW_NODE
2025-09-15 16:25:15.633 - INFO - [CTCCM_SLOWDETECT INFO] step : 34, dp check result = CTCCProfilerResult.DP_COMM_SLOW
2025-09-15 16:25:15.633 - INFO - [CTCCM_SLOWDETECT INFO] qptime detector start analysis slow node base on step 33 to step 34
2025-09-15 16:25:15.635 - INFO -
=== [CTCCM_SLOWDETECT INFO] Detected 18 slow QPs ===
2025-09-15 16:25:15.635 - INFO - Outlier 1: {
  "host": "10.10.10.1",
  "localguid": "7c1a1a1a~458",
  "remoteguid": "7c1a1a1a~610",
  "qpnum/qpindex": "125798/3"
}
2025-09-15 16:25:15.635 - INFO - Outlier 2: {
  "host": "10.10.10.1",
  "localguid": "7c1a1a1a~458",
  "remoteguid": "7c1a1a1a~610",
  "qpnum/qpindex": "125793/6"
}
2025-09-15 16:25:15.635 - INFO - Outlier 3: {
  "host": "10.10.10.1",
  "localguid": "7c1a1a1a~458",
  "remoteguid": "7c1a1a1a~610",
  "qpnum/qpindex": "125791/4"
}
2025-09-15 16:25:15.635 - INFO - Outlier 4: {
  "host": "10.10.10.1",
  "localguid": "7c1a1a1a~458",
  "remoteguid": "7c1a1a1a~610",
  "qpnum/qpindex": "125797/2"
}
2025-09-15 16:25:15.635 - INFO - Outlier 5: {
  "host": "10.10.10.1",
  "localguid": "7c1a1a1a~458",
  "remoteguid": "7c1a1a1a~610",
  "qpnum/qpindex": "125802/1"
}

```

使用CTCCL-Slowdetect

安装部署

1. CTCCL-0.4.0版本及以上。安裝方式可參考CTCCL 使用指南。

用户指南

2. 安装慢节点工具套件

• 安装ctccm

下载合适版本的ctccm whl安装包，使用以下命令安装：

```
pip install ctccm-xxx
```

通过pip show可以查看安装包的位置，并配置环境变量：

```
pip show ctccm export PATH="/usr/local/python3/bin:$PATH"
```

启动命令：

```
ctccm --nodes 4 --port 8000 --debug####
```

• 安装ctccl-profiler-comm

下载合适版本的ctccl-profiler-comm安装包，使用以下命令安装：

```
pip install ctccl-profiler-comm-xxx
```

通过import ctccl profiler comm，在训练任务中调用ctccl-profiler-comm提供的API使用。ctccl-profiler-comm包括以下可用API：

save_tp_dp_groups(tp_group, dp_group):

保存任务拓扑信息到本地文件。

•tp_group (ProcessGroup)：张量并行组（由`megatron内置函数get_tensor_model_parallel_group()`获取）
•dp_group (ProcessGroup)：数据并行组（由`megatron内置函数get_data_parallel_group()`获取）
•返回：无
使用示例：from ctccl_profiler_comm.get_tp_dp_groups import save_tp_dp_groups from megatron.core.parallel_state import get_tensor_model_parallel_group from megatron.core.parallel_state import get_data_model_parallel_group
tp_group = get_tensor_model_parallel_group() dp_group = get_data_model_parallel_group() save_tp_dp_groups(tp_group, dp_group)
CtcclProfiler(train_iters):初始化CtcclProfiler类，创建CtcclProfiler对象。
•train_iters (int)：训练任务迭代总数
•返回：CtcclProfiler对象

CtcclProfiler.get_time():

时间打点，用于每个迭代前后，获得每次迭代的时间。•返回：当前时间点

CtcclProfiler.update_step(start_time, end_time, iteration):根据当前的训练迭代数，

以及迭代时间，判断是否要开启、继续收集、停止profiler。•start_time (float)：当前迭

代开始时间（由`CtcclProfiler.get_time()`获取）•end_time (float)：当前迭代结束时间

（由`CtcclProfiler.get_time()`获取）•iteration (int)：当前执行到的迭代步数•返回：无

使用示例：from ctccl_profiler_comm.ctccl_profiler import CtcclProfiler profiler = CtcclProfiler(args.train_iters) ... while iteration < args.train_iters: ... start_time = profiler.get_time() train_step() #####step#####
end_time = profiler.get_time() profiler.update_step(start_time, end_time, iteration) iteration += 1 ...

• 安装ctccl-profiler-net下载合适版本的ctccl-profiler whl安装包，使用以下命令安装：pip install ctccl-profiler-net-xxx.whl
通过pip show可以查看安装包的位置，并配置环境变量：pip show ctccl-profiler-net
export PATH="/usr/local/python3/bin:\$PATH" 启动命令：ctccl-profiler-net --log-level=info

下载地址

组件	下载地址
ctccm	https://jiangsu-10.zos.ctyun.cn/ctccl-n/ctccl-slowdetect/cuda12.2/ubuntu20.04/ctccl-slowdetect1.0.0/ctccm-1.0.0-cp39-cp39-linux_x86_64.whl

用户指南

组件	下载地址
ctccl-profiler-comm	https://jiangsu-10.zos.ctyun.cn/ctccl-n/ctccl-slowdetect/cuda12.2/ubuntu20.04/ctccl-slowdetect1.0.0/ctccl_profiler_comm-1.0.0-cp39-cp39-linux_x86_64.whl
ctccl-profiler-net	https://jiangsu-10.zos.ctyun.cn/ctccl-n/ctccl-slowdetect/cuda12.2/ubuntu20.04/ctccl-slowdetect1.0.0/ctccl_profiler_net-1.0.0-cp39-cp39-linux_x86_64.whl

使用流程

1. ctccm启动在和所有训练任务节点网络互通的节点上，部署1个ctccm服务，并配置好环境变量。使用以下启动命令拉起ctccm服务，根据训练任务节点数实际需要配置nodes和port。ctccm --nodes “nnodes” --port “ctccm-slowdetect-port”2. 提前配置ctccl-profiler所需环境变量配置以下环境变量：export CTCCL_QTIME_REPORT=1# export CTCCL_SLOWDETECT_SERVERADDR=“http://ctccm-slowdetect-ip”:ctccm-slowdetect-port”, export CTCCL_PROFILER_NETADDR=“http://127.0.0.1:8001”3. 启动训练任务和ctccl-profiler在同一目录下启动ctccl-profiler-net服务和执行训练任务。ctccl-profiler-net --log-level=info bash run_llama2_7b_morenode.sh #####4. 日志查看当日志级别设置为debug时，所有组件的日志均会以文件形式保存。否则，可以查看控制台输出的INFO日志。

使用提示

- 安装基础环境，必须带有2.2.0以上torch和megatron框架。
- 目前暂不支持MOE模型。默认TP在机内。
- 默认训练任务拉起时集群状态健康，仅负责检测训练中间过程中出现的慢节点。
- 在启动训练任务之前，先启动ctccm和ctccl-profiler-net服务。
- 请确保所有环境变量配置可用。

CTCCL-Slowdetect最佳实践

在4台A800*8，每台节点有8张mlx网卡，RoCE组网，部署慢节点工具套件。其中，在4节点上容器化部署模型训练基础环境以及llama2-7b训练模型，在node1上容器化部署ctccm服务，并在每一台节点上容器化部署ctccl-profiler服务。

在训练任务代码中调用ctccl-profiler-comm API：

```
from deepspeed.runtime.data_pipeline.data_routing.helper import convert_to_random_ltd
from megatron.model.transformer import ParallelTransformerLayer

from deepspeed import comm as dist

# for test
from ctccl_profiler_comm.get_tp_dp_groups import save_tp_dp_groups
from ctccl_profiler_comm.ctccl_profiler import CtcclProfiler
```



```
tp_group = get_tensor_model_parallel_group()
dp_group = get_data_parallel_group()
save_tp_dp_groups(tp_group, dp_group)

profiler = CtcclProfiler(args.train_iters)

while iteration < args.train_iters and (args.train_tokens is None or \
    iteration < args.train_tokens // args.train_batch_size):

    start_time = profiler.get_time()

    train_step()

    end_time = profiler.get_time()
    profiler.update_step(start_time, end_time, iteration)

    iteration += 1
    args.iteration = iteration
    new_samples = mpu.get_data_parallel_world_size() * \
```

在训练脚本中配置相关环境变量

```
export NCCL_SOCKET_IFNAME="bond0"
export NCCL_IB_HCA="mlx5_2,mlx5_3,mlx5_4,mlx5_5,mlx5_8,mlx5_9,mlx5_10,mlx5_11"
export NCCL_IB_TIMEOUT=22
export NCCL_IB_GID_INDEX=3
export NCCL_IB_DISABLE=0
export NCCL_DEBUG="WARN"
export NCCL_DEBUG_FILE="nccl.log"
export NCCL_IB_QPS_PER_CONNECTION=8
export NCCL_DEBUG_SUBSYS="ALL"
export CTCCL_ERR_REPORT_SVC=0
export CTCCL_IB_LB_UPLINK=0
export CTCCL_BW_REPORT=0
export CTCCL_PROFILER_TIMEOUT_RATE=20
export CTCCL_OPTIME_REPORT=1
export CTCCL_SLOWDETECT_SERVERADDR="http://10.10.10.10:8002"
export CTCCL_PROFILER_LOG_LEVEL="INFO"
```

启动ctccm-slowdetect服务

```
export PATH="/usr/local/python3/bin:$PATH"##### && ctccm --nnodes 4 --port 8002 --debug
```

用户指南

```

deepspidee apex ctccm ctccm_producer net_cython3.1.3-py3-none-any.whl deepsparklib toverbs_ctc
Apptainer> export PATH="/usr/local/python3/bin:$PATH" &&
t --nno> ctccm-slowdetect --nnodes 4 --port 8002 --debug
2025-09-09 15:46:00,651 - INFO - CTCCM slow detect service start, port: 8002, DEBUG mode: True
INFO: Started server process [2609044]
INFO: Waiting for application startup.
INFO: Application startup complete.
INFO: Uvicorn running on http://0.0.0.0:8002 (Press CTRL+C to quit)

```

启动所有节点上的ctccl-profiler-net服务

```

Apptainer> export PATH="/usr/local/python3/bin:$PATH" \
in:$PATH" \
> export CTCCL_SLOWDETECT_SERVERADDR="http://
/ 0.0.1:8002" \
> export CTCCL_PROFILER_NETADDR="http://127.
0.0.1:8001" &&
> ctccl-profiler start
2025-09-09 15:47:26,461 - INFO - [CTCCL_PROF
ILER INFO] Starting CTCCL QP data transmissi
on monitoring service...
[INFO: Started server process [2609165]
[INFO: Waiting for application startup.
2025-09-09 15:47:26,473 - INFO - [CTCCL_PROF
ILER INFO] CTCCL qptime log rotate worker th
read started successfully
2025-09-09 15:47:26,473 - INFO - [CTCCL_PROF
ILER INFO] CTCCL process worker thread start
ed successfully
[INFO: Application startup complete.
[INFO: Uvicorn running on http://127.0.0.
1:8001 (Press CTRL+C to quit)

```

启动分布式训练任务，ctccm会收到任务的逻辑拓扑

[illegible]

ctccm在训练中检测集群中是否存在慢节点,一旦发现慢节点则下发开始收集细粒度的监控信息的控制信号,并做慢节点定位定界。

当计算慢时，ctccm会给出计算慢的TP通信域所包含的rank。

```

2025-09-22 11:48:45,223 - DEBUG - [CTCCM_SLOWDETECT DEBUG] Slow node is still detecting reason...
2025-09-22 11:48:45,708 - INFO - [CTCCM_SLOWDETECT INFO] profiler data is prepared, starting detect ...
2025-09-22 11:48:45,708 - INFO - [CTCCM_SLOWDETECT INFO] profiler detector start analysis slow node based on step 58 to step
59
2025-09-22 11:48:45,708 - DEBUG - [CTCCM_SLOWDETECT DEBUG] profiler matrix is [[0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]]
2025-09-22 11:48:45,708 - INFO - [CTCCM_SLOWDETECT INFO] Rank0 compute slow or tp comm slow
2025-09-22 11:48:45,708 - INFO - [CTCCM_SLOWDETECT INFO] Rank1 compute slow or tp comm slow
2025-09-22 11:48:45,708 - INFO - [CTCCM_SLOWDETECT INFO] Rank2 compute slow or tp comm slow
2025-09-22 11:48:45,708 - INFO - [CTCCM_SLOWDETECT INFO] Rank3 compute slow or tp comm slow
2025-09-22 11:48:45,708 - INFO - [CTCCM_SLOWDETECT INFO] step : 59, tp check result = CTCCProfilerResult.NO_SLOW_NODE
2025-09-22 11:48:45,708 - INFO - [CTCCM_SLOWDETECT INFO] step : 59, dp check result = CTCCProfilerResult.TP_COMPUTE_SLOW
2025-09-22 11:48:45,708 - INFO - [CTCCM_SLOWDETECT INFO] antime detector start analysis slow node base on step 58 to step 59
2025-09-22 11:48:45,710 - DEBUG - [CTCCM_SLOWDETECT TEST]
2025-09-22 11:48:45,710 - INFO - [CTCCM_SLOWDETECT INFO] No QP/device detected

```

用户指南

当通信慢时，ctccm会给出慢的QP以及它所对应的网卡对和所在节点。

```
2025-09-15 16:25:15,633 - INFO - [CTCCM_SLOWDETECT INFO] step : 34, tp check result = CTCCProfilerResult.NO_SLOW_NODE
2025-09-15 16:25:15,633 - INFO - [CTCCM_SLOWDETECT INFO] step : 34, dp check result = CTCCProfilerResult.DP_COMM_SLOW
2025-09-15 16:25:15,633 - INFO - [CTCCM_SLOWDETECT INFO] qptime detector start analysis slow node base on step 33 to step 34
2025-09-15 16:25:15,635 - INFO -
=== [CTCCM_SLOWDETECT INFO] Detected 18 slow QPs ===
2025-09-15 16:25:15,635 - INFO - Outlier 1: {
  "host": "125790/3",
  "localguid": "7c8c158",
  "remoteguid": "7c8c158",
  "qpnum/qpindex": "125790/3"
}
2025-09-15 16:25:15,635 - INFO - Outlier 2: {
  "host": "125793/6",
  "localguid": "7c8c158",
  "remoteguid": "7c8c158",
  "qpnum/qpindex": "125793/6"
}
2025-09-15 16:25:15,635 - INFO - Outlier 3: {
  "host": "125791/4",
  "localguid": "7c8c158",
  "remoteguid": "7c8c158",
  "qpnum/qpindex": "125791/4"
}
2025-09-15 16:25:15,635 - INFO - Outlier 4: {
  "host": "125797/2",
  "localguid": "7c8c158",
  "remoteguid": "7c8c158",
  "qpnum/qpindex": "125797/2"
}
2025-09-15 16:25:15,635 - INFO - Outlier 5: {
  "host": "125802/7",
  "localguid": "7c8c158",
  "remoteguid": "7c8c158",
  "qpnum/qpindex": "125802/7"
}
```

训练性能加速

MFU提升

在平台上使用1152卡对Llama3.1-405B进行预训练和微调，通过数据加速、镜像加速、3D并行等加速策略，实现MFU的提升：

MFU - 芯片算力（%）：达到44.185MFU - CUBE算力（%）：达到47.064

环境配置

服务器型号	Atlas 800T A2
NPU型号	910B2（64GB）
驱动版本	23.0.3
CANN	8.0.RC2
Python	3.10.14
MindSpore	2.3.1
Mindformers	dev分支（5bfebf+diff.patch）

训练配置

Epochs	100
Learning Rate	6.e-5
Global Batch Size	2048
Batch Size	1

用户指南

Epochs	100
Micro Batch Size	256
Sequence Length	4096
Data Parallel (DP)	8
Model Parallel (MP)	8
Pipeline Parallel (PP)	18
max_device_memory	54GB
jit_level	02

训练结果

吞吐量 (tokens/s/p)	61.962
MFU - 芯片算力 (%)	44.185
MFU - CUBE算力 (%)	47.064

线性加速比提升

在平台测试模型训练规模的性能线性比，分别测试1k, 2k, 4k, 6k, 9k规模并作对比如下（以千卡规模为准）：

训练规模（卡数）	吞吐量 (tokens/s/p)	MFU（芯片/CUBE） (%)	线性比 (%)
1024	407.701	47.847 / 50.965	100
2048	403.415	47.344 / 50.429	98.95
4096	400.734	47.030 / 50.094	98.29
6144	393.577	46.190 / 49.200	96.54
9216	366.915	43.061/ 45.867	90.00

模型适配与性能加速

平台针对国产化硬件910B进行主流模型的迁移适配和性能优化，多数训练性能达到A800的90%及以上，推理性能达到A10或A100的80%及以上。

目前模型广场所有模型均达到国产化适配的性能标准。

主流模型适配结果示例：

- 多数模型基于昇腾的训练可达A800的90%以上，举例如下：
Llama3-8B-instruct、ChatGLM3-6B 性能对标可达90%
Qwen2-72B-Instruct 性能对标可达100%
- 部分模型基于昇腾910B的推理可达A800的80%以上，举例如下：
Qwen-1.8B-Chat 性能对标可达110.32%
Llama2-13B-Chat 性能对标可达91.28%

用户指南

主流适配模型清单参考如下：

- 大语言：
DeepSeek：DeepSeek-R1系列、DeepSeek-V3系列
Llama：Llama系列、Llama2系列、Llama3系列
通义千问：Qwen系列、Qwen1.5系列、Qwen2系列、Qwen2.5系列、Qwen3系列
智谱：ChatGLM2系列、ChatGLM3、GLM4系列
书生浦语：InternLM系列、InternLM2系列
百川：Baichuan系列、Baichuan2系列、Baichuan-Turbo
电信星辰：TeleChat系列、TeleChat2系列
其他：AquilaChat-7B、Gemma2-9B-Chat等
- 多模态：
OpenClip
ChineseClip
Blip2
VisualGLM-6B
StableDiffusion-V1.5
StableDiffusion-V2.1
Qwen-VL-Chat
InternVL-Chat-V1.5

在国产化适配方面主要具备能力：

- 针对提供的算力，可以提供全套迁移服务
- 针对提供的AI算力，支持自动生成模型的算子支持度分析报告
- 针对提供的算力，可以提供迁移适配服务
- 对适配后的算子，可以提供精度对齐调优工具
- 对适配后的算子，可以提供专家持续提供调优指导服务

断点续训加速

测试数据及代码准备

数据集	数据集大小	使用模型
Wikipedia-en (1M条)	9.1GB	Llama2-70B /Llama3.1-405B

- 使用预处理为MindRecord格式的Wikipedia-en (1M条)数据集，上传到对象存储，并由对象存储下载到平台HPFS。
- 测试代码在gitlab仓库下载到本地，并放置于/work/home下。

用户指南

脚本和任务准备

按照下面修改run.sh脚本

```
#!/bin/bash
```

```
# huijuformers的绝对路径, 需要修改path_to_huijuformers  
export BASE_DIR=/work/data/llama2_9216/huijuformers
```

以下为平台自动注入的环境变量

```
# yaml文件中需要修改的环境变量  
export BATCH_SIZE=1  
export EPOCHS=350  
export LEARNING_RATE=6.e-5  
export DATA_PARALLEL=256  
export MODEL_PARALLEL=4  
export PIPELINE_STAGE=9
```

模型微调相关

```
export FINETUNE_MODEL_TYPE=llama2_70b_base # 合并为一个参数, 与模型存放文件夹名称一致 (与  
后端沟通过)  
export FINETUNING_TYPE=ALL  
export TIME_TAG=$(date +"%m%d-%H%M")
```

数据相关

```
# export DATASET_PATH=${BASE_DIR}/data  
# export DATASET_FILE=original_data.json # 需要修改
```

以下为平台后端需要自行更改后传入的环境变量

平台数据格式转换, 专用数据调试时用不到

```
# export DATASET_TMP_PATH=${BASE_DIR}/data/processed_data/${FINETUNE_MODEL_TYPE}  
# mkdir -p ${DATASET_TMP_PATH}
```

模型输入

专业模式, 平台训练时需要按照平台的挂载路径去修改这一块变量

```
# export CHECKPOINT_DIR=""
```

低代码模式, 微调时约定挂载为下面的路径

```
## export CHECKPOINT_DIR=/work/mount/publicModel/${FINETUNE_MODEL_TYPE}/  
${FINETUNE_MODEL_TYPE}
```

输出文件夹路径, run_mode为训练模式, 如train, lora, full

```
run_mode=train
```

```
export OUTPUT_DIR=${BASE_DIR}/output/${FINETUNE_MODEL_TYPE}/${run_mode}/${TIME_TAG}  
export OUTPUT_ROOT_DIR=${BASE_DIR}/output/${FINETUNE_MODEL_TYPE}/${run_mode}  
rm -rf ${OUTPUT_DIR}/resume_record
```

用户指南

```
# 获取节点IP、名称，记录至文件
echo $(hostname -I | awk '{print $1}'), $NODE_NAME >> ${BASE_DIR}/output/nodes

sed -i '/pam_limits.so/s/^/#' /etc/pam.d/sshd
# 启动脚本
cd ${BASE_DIR}/bin/scripts

# apt install netcat -y
# 微调
# bash finetune.sh
# 预训练
export MS_TOPO_TIMEOUT=7200
bash train.sh
```

启动训练任务

点击训练任务页面的新建任务，按照如下的示例配置训练命令和模型挂载等，然后启动任务。

Llama2-70B万卡测试结果

在平台使用9216卡对Llama2-70B进行预训练，万卡规模国产化集群下，断点续训在5类故障下实现1分钟检测、5分钟内定位、15分钟内恢复训练。

整体结果

训练环境

服务器型号	Atlas 800T A2
NPU型号	910B2（64GB）
驱动版本	23.0.3
CANN	8.0.RC2
Python	3.10.14
MindSpore	2.3.0
Mindformers	1.2.0

训练配置

Epochs	350
Learning Rate	6.e-5
Global Batch Size	32768
Batch Size	1
Micro Batch Size	128
Sequence Length	4096
Data Parallel (DP)	256
Model Parallel (MP)	4

用户指南

Epochs	350
Pipeline Parallel (PP)	9
max_device_memory	54GB
jit_level	02

训练结果

吞吐量 (tokens/s/p)	366.915
MFU - 芯片算力 (%)	43.061
MFU - CUBE算力 (%)	45.867

断点续训

断点CheckPoint总大小：22T，其中0卡断点CheckPoint大小：2.9G。

故障1：业务故障，kill所有python进程

故障检测时间(Min)	7.2s
故障处理耗时(Min)	231.7s, 3.86min
故障恢复耗时(Min)	458s, 7.63min
CKPT加载时间(Min)	0.28min
0卡CKPT加载速度(GB/s)	0.99

故障2：节点心跳故障，把node上label去掉

故障检测时间(Min)	18.9s, 0.315min
故障处理耗时(Min)	279.8s, 4.64min
故障恢复耗时(Min)	478s, 7.96min
0卡CKPT加载时间(Min)	0.3min
0卡CKPT加载速度(GB/s)	1.01

故障3：节点down故障，reboot

故障检测时间(Min)	18.3s, 0.3min
故障处理耗时(Min)	465s, 7.75min
故障恢复耗时(Min)	546s, 9.1min
0卡CKPT加载时间(Min)	0.98min
0卡CKPT加载速度(GB/s)	0.1

故障4：网络故障，网卡link down

故障检测时间(Min)	895s(600s HCCL), 14.9min
故障处理耗时(Min)	300s, 5min

用户指南

故障检测时间(Min)	895s(600s HCCL), 14.9min
故障恢复耗时(Min)	472.1s, 7.86min
0卡CKPT加载时间(Min)	0.32min
0卡CKPT加载速度(GB/s)	0.96

故障5: PCIE故障, 模拟NPU掉卡

故障检测时间(Min)	78.7s, 1.3min
故障处理耗时(Min)	267.1s, 4.4min
故障恢复耗时(Min)	516.9s, 8.6min
0卡CKPT加载时间(Min)	0.32min
0卡CKPT加载速度(GB/s)	0.98

Snapckpt加速

Snapckpt加速包

SnapCheckpoint (Snapckpt) 是由息壤一体化智算服务平台 - 训推智算服务平台提供的针对大模型训练场景提供的高性能checkpoint框架。Snapckpt提供了高效便捷的断点存储、加载机制。

Snapckpt介绍

Snapckpt是一种为大模型训练打造的易用、可拓展、高性能的断点解决方案。Snapckpt提供了高效便捷的断点存储、加载机制，其主要有以下几大优势：

1. 异步断点保存：通过优化断点保存流程，使得断点保存过程异步化，尽可能减小断点保存过程对于训练的中断，减少训练阶段耗时。
2. 分布式断点存储：对模型机优化器参数进行分布式存储，避免集中式存储带来的额外通信开销，大幅降低存储与加载断点的耗时。
3. 多框架支持：支持Megatron-LM、Deepspeed两大主流大模型训练框架。
4. 简单易用：安装及使用方式简单，用户仅需简单步骤即可使用，提升使用效率。

背景信息

在大规模分布式模型训练过程中，系统可能因硬件故障或软件异常导致训练中断。为确保训练进度可恢复，业界普遍采用周期性保存检查点(Checkpoint)的方案。值得注意的是，检查点操作耗时与模型参数量呈正相关关系，当面对参数量达百亿甚至千亿级别的大模型时，每次检查点保存往往需要耗费数分钟至十余分钟不等。特别是在使用Megatron-LM框架或原生PyTorch进行训练时，检查点保存过程会强制暂停训练任务，导致宝贵的计算资源被闲置。因此，开发高效的检查点机制以降低时间成本和资源浪费，已成为当前大规模模型训练亟待解决的关键问题。

Snapckpt采用多阶段异步断点存储机制，降低断点存储耗时，减少训练中断带来的影响，从而提升训练速度，提升计算资源有效利用率。

用户指南

使用方法

安装Snapckpt

Snapckpt Python包下载地址: <https://shanghai15.zos.ctyun.cn/zhisuan-sre/huiju-data/Ctyun%20Snapckpt.zip>

1. 下载后进入到工程根目录
2. 安装命令:

```
python3 setup.py build && python3 setup.py install
```

至此, Snapckpt的软件包安装完成, 可以开始使用(具体使用方法见“Snapckpt使用方法”章节)。

Snapckpt使用方法

Megatron模式使用方法

在使用Megatron-LM训练框架训练模型时, 您可以使

用 `snapckpt.megatron_dist_ckpt.save_checkpoint`和 `snapckpt.megatron_dist_ckpt.load_checkpoint` 替换原生Megatron-LM的 `megatron.checkpointing.save_checkpoint` 和 `megatron.checkpointing.load_checkpoint`。

```
# from megatron.checkpointing import save_checkpoint
from snapckpt.megatron_dist_ckpt import save_checkpoint
```

```
# from megatron.checkpointing import load_checkpoint
from snapckpt.megatron_dist_ckpt import load_checkpoint
```

Deepspeed模式使用方法

首先, 需要初始化Snapckpt的DeepSpeedCheckpoint, 初始化方式如下:

```
from snapckpt.deepspeed import(
    DeepSpeedCheckpoint,
    StorageType,
)
checkpointer = DeepSpeedCheckpoint(model, checkpoint_dir)
```

初始化后, 您可以使用 `checkpointer.save_checkpoint` 和 `checkpointer.load_checkpoint` 替换原生Deepspeed的 `save_model` 和 `load_checkpoint`。

```
# save_model(checkpoint_dir)
checkpointer.save_checkpoint(checkpoint_dir, tag=step, storage_type=StorageType.MEMORY)
checkpointer.save_checkpoint(checkpoint_dir, tag=step, storage_type=StorageType.DISK)

# load_checkpoint(model, checkpoint_dir)
checkpointer.load_checkpoint(checkpoint_dir)
```

性能参考

以Deepseek R1 Distill Qwen 14B模型, 单机八卡运行环境为例, 相较于同步断点保存而言, Snapckpt的性能提升测试参考如下:

用户指南

模式	全参微调	LoRA微调	预训练
Megatron	24倍	10倍	22倍
Deepspeed	19倍	5倍	12倍

智算资产

我的模型

全面管理用户从开发、训练到评估完成的模型生命周期，该模块不仅提供模型文件的安全存储功能，还具备精细化的版本管理，确保每一阶段的模型变更都有迹可循。

新建模型及版本

1. 在我的模型菜单页面中，点击【新建模型】，输入模型名称、模型描述以及来源。支持2种来源：
 - a. 导入至平台存储：上传模式，您可以将本地文件上传至平台提供的共享ZOS空间内，该模式对模型文件有2GB的大小限制；
 - b. 自有存储挂载（推荐使用）：挂载模式，需要您在天翼云官网同资源池下开通自有存储并完成委托授权等前置操作。选择自有存储挂载模式后，可创建的模型大小不再受平台限制，您只需填写已有存储的路径，平台将在任务中自动挂载该路径。目前已支持自有ZOS、自有HPFS、自有其他存储类型，您可在创建版本时切换顶部集群查看该集群下支持的存储类型。其中，
 1. 自有ZOS：是指您租户在天翼云官网同资源池下开通的对象存储，用于数据长期存储和备份，完成委托授权后您可在本平台直接使用；
 2. 自有HPFS：指租户在天翼云官网同资源池下开通的HPFS存储，常用于大模型的开发和训练等数据密集型的高性能计算场景，完成委托授权后您可在本平台直接使用。训练时需要与文件存储频繁交互，请确保存储状态可用且充足；

用户指南

3. 自有其他存储：天翼云的集群分为两类：天翼云自建集群、与合作伙伴共营的其他集群，自建集群对应自有存储，其他集群对应其他存储。其他存储便是指其他集群中对应配置的存储。该存储与集群强相关，您在开通相应的集群后平台会自动打通与存储的关联，不需要进行委托授权步骤。

模型管理 / 新建模型

● 使用须知：息壤一站式智算服务平台需要授予访问和写入相关资源的权限，目前尚未授权，无法存储数据，点击 [此处授权](#)

* 模型名称

请输入模型名称

模型描述

请输入描述，100个字符以内

0/100

* 来源

导入至平台存储

自有存储挂载

上传后将存储在平台提供的存储空间内，当前提供免费的300G额度的ZOS对象存储供您体验

取消

确定

用户指南

2. 新建模型后，点击模型卡片，进入模型详情页，您可在此处创建版本，版本是具体可操作的对象。

a. 模型来源为导入至平台存储：

模型管理 / 新建版本

● 使用须知： 息壤一站式智算服务平台需要授予访问和写入相关资源的权限，目前尚未授权，无法存储数据，点击 [此处授权](#)

版本 V1

版本描述 请输入描述，100个字符以内 0/100

存储配置

导入至平台存储（上传后将存储在平台提供的存储空间内）
当前提供免费的300G额度的ZOS对象存储供您体验

平台ZOS存储剩余
130.55GB

* 模型来源 ☒ 当前平台导入 ☐ 本地上传 ☐ 口令导入 ☐ 下载链接导入

* 类型 ☒ 模型精调 ☐ 训练任务 ☐ Jupyterlab ☐ VSCode

请选择

取消 确定

- 当前平台导入：支持从平台上运行完成的模型调优和训练任务中导入(分别参见[模型精调](#)和[训练任务](#))、也可以从JupyterLab和VSCode的目录中导入(参见[模型开发](#))。
- 本地上传：支持从本地电脑导入不超过2G的模型文件。
- 口令导入：支持输入训推智算服务平台其他账户分享的口令完成导入，参见[模型分享与导出](#)章节。
- 下载链接导入：支持输入互联网下载链接地址完成模型导入。

b. 模型来源为自有存储挂载：

80

用户指南

C.

模型管理 / 新建版本

版本 V3

版本描述 请输入描述，100个字符以内 0/100

存储配置

自有存储挂载
您在本平台填写已有存储的路径，平台将在任务中挂载该路径

* 存储目标类型 ☐ 自有对象存储 (ZOS) ☒ 自有智算存储 (HPFS)

您租户账号下同资源池购买的并行文件存储

* HPFS名称 请选择HPFS文件系统 去创建

* 数据源路径 请输入已被授权的目录路径，示例：test/ 0/4096

取消 确定

- 自有对象存储：您租户账号下的自有ZOS存储。使用自有存储挂载时，需要提前将数据 [上传至ZOS](#)，若您还未创建，您可点击【去创建】跳转到存储控制台进行创建。
 - 存储桶：选择您自有的ZOS存储桶名称；
 - 数据源路径：您对应存储的具体的目录路径，您可根据提示示例进行填写，平台会对您填写路径的存在性进行校验，若校验不通过会进行提示，并禁止提交。
- 自有智算存储：您租户账号下的自有HPFS存储。使用自有存储挂载时，需要提前将数据 [传至HPFS](#)，若您还未创建，您可点击【去创建】跳转到存储控制台进行创建。
 - HPFS名称：选择您自有并行文件名称；
 - 数据源路径：您对应存储的具体的目录路径，您可根据提示示例进行填写，平台会对您填写路径的存在性进行校验，若校验不通过会进行提示，并禁止提交。
- 自有其他存储：您租户账号下的自有其他存储。
 - 数据源路径：您对应存储的具体的目录路径，您可根据提示示例进行填写，此项必填。

模型列表

导入的模型可以在我的模型的列表中查看，每个模型可以导入多个版本(训练任务、开发机、模型精调等任务开发出的模型进行保存操作时都会保存在这里)。点击模型卡片，可以查看模型的所有版本。

模型的每个版本都会显示导入状态，比较大的模型导入时间较长。

用户指南

我的模型

集中管理用户通过平台导入、训练和精调出来的生成式大模型，支持对模型进行版本管理、评估及部署。

+ 新建模型

可输入名称、开发者

查询

重置

{ t

版本数量: 1

暂无描述

开发者: 何. 早

创建时间: 2025-08 15:04:32

评估

压缩

部署

更多

共1条

20

<

1

>

模型分享与导出

仅来源为导入至平台存储的模型支持此功能。

模型分享

模型列表和模型版本列表中，点击【分享】可生成分享口令，支持模型分享，可将模型在多个账号之间进行共享下载，通过口令可以进行模型导入，参见模型[口令导入](#)章节。

账户1要把模型文件分享给账户2，需要账户1在模型列表或版本列表中点击【分享】获得一个分享口令，将分享口令线下给到账户2。

账户2在新建模型中选择【口令导入】，输入账户1给到的分享口令即可完成模型导入。

我的模型 / 查看详情

< 部署专用模型20250611

名称

+ 新建版本

版本

来源

V1

用户导入

分享模型

模型名称: 部署

112049196V1

有效时长: ☐ 1天

☐ 7天

☐ 30天

☒ 永久有效

在「我的模型」新建模型或版本时输入口令，即可导入模型「部署 11120 5V1」

口令: ***** 获取口令

有效时长: 永久有效

关闭

一键复制

创建时间 操作

2025-08 导出

分享

评估

部署

删除

共1条

20

<

1

>

用户指南

模型导出

进入模型详情页，在模型版本列表中支持模型导出，可以选择导出到本地，也可以选择导出到天翼云媒体存储中(需要开通天翼云媒体存储的账号才能用)。

我的模型 / 查看详情

< 部署专用

149196

名称	部署专用	112049196	描述
----	------	-----------	----

+ 新建版本

版本	来源	任务名称	导入状态	基础模型	导出状态	版本描述	开发者	创建时间	操作
V1	用户导入	本地上传	● 导入成功	-	-	-	dxc	2025-0	导出 分享 评估 部署 删除

共1条 20 < 1 >

我的数据集

基础数据集

前置条件

- 使用前，您的租户账号需在天翼云存储控制台开通并创建相应的存储，详见 [对象存储快速入门](#)、[并行文件系统快速入门](#)。
- 已在本平台完成相关产品的委托授权。

操作说明

基础数据集统一走挂载模式，需要您提前在存储的管理面完成数据导入、记录路径等前置操作，您在本平台只需填写已有存储的路径，平台将在任务中自动挂载该路径。基础数据集的存储方式包括普通存储、智算存储与其他存储：

- 普通存储（ZOS）：
 - 账号自有存储：指租户在天翼云官网同资源池下开通的对象存储，用于数据长期存储和备份，完成委托授权后您可在本平台直接使用。
 - 平台共享存储：本平台赠予您体验的存储，默认集群额度为 300G，是所有用户共享的存储，您的用量受限且不支持扩容，超出用量后需自行前往对象存储购买自有存储。平台后续将逐渐废弃此类存储，建议您直接使用自有存储。
- 智算存储（HPFS）：
 - 账号自有存储：指租户在天翼云官网同资源池下开通的HPFS存储，常用于大模型的开发和训练等数据密集性的高性能计算场景，完成委托授权后您可在本平台直接使用。如需使用开发机和训练任务功能，请提前将数据、模型、代码导入智算存储中。训练时需要与文件存储频繁交互，请确保持续状态可用且充足。

用户指南

- 平台共享存储：您在本平台开通的共享存储，默认集群额度为 512 G，是所有用户共享的存储，您的用量受限且不支持扩容，超出用量后需自行前往HPFS购买自有存储。平台后续将逐渐废弃此类存储，建议您直接使用自有存储。
- 其他存储：天翼云的集群分为两类：天翼云自建集群、与合作伙伴共营的其他集群，其中，自建集群对应自有存储，其他集群对应其他存储。其他存储便是指其他集群中对应配置的存储。该存储与集群强相关，您在开通相应的集群后平台会自动打通与存储的关联，不需要进行委托授权步骤。

上述可使用的存储类型都与集群强相关，例如集群A支持自有ZOS、自有HPFS，集群B支持自有其他存储，您不能在集群B下使用自有ZOS。您可切换顶部集群查看该集群支持的存储类型，也可事先联系您的客户经理了解集群的存储类型支持情况。

我的数据集

基础数据集

标注数据集

+ 创建基础数据集

请输入名称/ID搜索

请选择运行状态

数据集名称/ID	状态	存储类型	描述	默认容器内挂载路径	操作
即时验证-基础	创建成功	自有对象存储(ZOS)		/work/basicData/1	查看文件 删除
保密数据集	创建成功	自有对象存储(ZOS)		/work/basicData/1	查看文件 删除
指令微调	创建成功	自有对象存储(ZOS)		/work/basicData/1	查看文件 删除

创建数据集

- 创建普通存储数据集：登录智算服务控制台，单击左侧菜单栏的“智算资产”>“我的数据集”菜单项进入我的数据集模块，点击“基础数据集”菜单，选择“普通/智算存储”，点击【+创建普通/智算存储数据集】，进入创建页面
- 填写相关配置并提交。基础数据集创建目前仅支持使用自有存储，创建完成后，自动为您创建拼接此数据后缀的容器内挂载路径。

类型	字段	说明
存储基本信息	数据集名称	数据集名称，不超过25个字符
	描述	数据集描述信息
权限配置	可见范围	读写权限范围，支持设置仅文件所有者可见、指定工作空间内的算法开发角色可见两种权限策略

用户指南

存储位置	自有ZOS存储	<p>您租户账号下的自有ZOS存储，若您还未创建，您可点击【去创建】跳转到存储控制台进行创建。</p> <ul style="list-style-type: none"> • 存储桶：选择您自有的ZOS存储桶名称； • 数据源路径：您对应存储的具体的目录路径，您可根据提示示例进行填写，平台会对您填写路径的存在性进行校验，若校验不通过会进行提示，并禁止提交。对于管理员用户，此项选填；对于普通子用户，此项必填。
	自有HPFS存储	<p>您租户账号下的自有HPFS存储，若您还未创建，您可点击【去创建】跳转到存储控制台进行创建。</p> <ul style="list-style-type: none"> • HPFS名称：选择您自有的HPFS名称； • 数据源路径：您对应存储的具体的目录路径，您可根据提示示例进行填写，平台会对您填写路径的存在性进行校验，若校验不通过会进行提示，并禁止提交。对于管理员用户，此项选填；对于普通子用户，此项必填。
	自有其他存储	<p>您租户账号下的自有其他存储。</p> <ul style="list-style-type: none"> • 数据源路径：您对应存储的具体的目录路径，您可根据提示示例进行填写，此项必填。

用户指南

特殊配置	设为保密数据集	<p>仅存储委托的自有ZOS与自有HPFS类型支持该功能。</p> <p>对于某些行业客户，在大模型训练时会涉及到保密数据的处理问题，比如需要避免用户下载到本地造成数据泄露，针对这类问题，平台推出了特色的数据保密功能，以应对用户下载或拷贝保密数据的行为。相关配置仅对管理员用户开放，且此功能会增加操作复杂度和资源占用，建议您只在需要的时候使用。</p> <p>开启此按钮后，可将本数据集设置为保密数据集。开启保密后，此数据集将：1）仅支持读操作，禁止写操作，运行后的输出将单独写入到保密输出路径；2）数据挂载到容器时禁止访问外网。</p>
	设为保密数据集的输出路径	<p>若您已设置保密数据集，您还需再创建一个保密输出数据集，专门用于存储保密数据集的输出，并谨慎设置相关权限：1）若您使用了自有存储，建议您只给自己或管理者读写权限（需前往对应存储控制台设置），即您需要确保这是由您完全可控的存储；2）此数据集的可见范围（在本页面的权限配置设置），建议您只对需要的用户设置。</p>
	云审计	<p>您设置好保密输出路径后，会自动弹出云审计设置，此设置对使用了对应保密输出数据集的任务生效。开启云审计后，使用此保密输出数据集的任务在容器内的命令操作将会被记录并发送到云审计，并标识拷贝等高危操作。</p>

用户指南

	审计频率	您开启云审计后，会弹出审计频率设置。审计频率是指审计上报的时间间隔，支持输入1-1440的整数。如果您设置的时间间隔太短，频繁扫描将会增加损耗，如时间太长，上报的事件信息较密集，建议您根据业务实际情况填写合适的时间。
--	------	--

我的数据集 / 创建普通存储数据集

普通存储基本信息

数据集名称

仅支持中文、大小写字母、数字、-、_、* 0/25

描述

请输入描述信息，最多250个字 0/250

权限配置

所有者

秦

读写权限范围

☒ 仅文件所有者可见（读写）

☐ 指定工作空间内的算法开发角色

存储位置

只能使用自有存储

存储桶

请选择存储桶

去创建

ZOS路径

请输入已被授权的目录路径。示例：test/ 0/1000

特殊配置

设为保密数据集

☐ 关

设为保密数据集的输出路径

☐ 关

取消

提交

管理数据集

- 查看文件：对于共享存储，点击可查看数据集中的文件列表；对于自有存储，点击可跳转至存储控制台查看您有权限的文件。
- 删除：只有导入成功、导入失败、创建成功、创建失败的数据集支持删除操作。如果有保密设置，则仅管理员用户可操作；无特殊设置，则创建人及管理员用户可操作。该数据集若正在被开发机和训练任务使用，则删除数据集后所有使用中和已完成的开发机和训练任务对应挂载路径下的文件也将被清空。

标注数据集

前置条件

- 如需使用自有存储：
 - 使用前，您的租户账号需在天翼云存储控制台开通并创建相应的存储。
 - 已在本平台完成相关产品的委托授权。
- 根据格式要求在本地建立相应的数据集。

操作说明

数据导入

- 进入标注数据集模块，点击【创建数据集】，填写相关配置信息：

字段	说明
----	----

用户指南

数据集名称	数据集名称，仅支持中英文、数字、下划线“_”、短横“-”，只能以中英文、数字开头
版本	数据集版本
描述	数据集描述信息
数据类型	将要上传的数据集类型，支持图片、大语言、跨模态
标注类型	根据数据类型选择需要的标注类型
标注模版	根据标注类型选择需要的标注模版
存储目标类型	<ul style="list-style-type: none">• 自有对象存储：指您租户账号下同资源池购买的对象存储；<ul style="list-style-type: none">• 存储桶：选择存储桶；• 目的地路径：填写您存储的目的地路径，导入时存储将上传至此路径；• 平台共享存储：指平台提供的供您体验的共享存储，选择此类型后，数据将上传至平台设置好的默认路径。• 自有其他存储：指您使用集群对应的自有其他存储，选择此类型后，填写您存储的目的地路径，导入时存储将上传至此路径。

2. 支持的在线标注模版介绍及示例：

- 单图多标签：
 - 说明：图片数据集，支持格式jpg、jpeg、png、bmp。
 - 标注：您可直接在平台上点击“标注” > “新建标签”，对图片添加合适的标签。
- 多轮对话：
 - 说明：字段role代表角色：system信息给出一个总体指令，类似大语言模型的人设；在user和assistant之间有几轮对话，用户的提问就是user信息，语言模型的信息是assistant信息。字段content代表角色的对话信息。
 - 标注：不支持标注。
 - 示例：支持jsonl格式，以下是jsonl的一个示例：

```
{"messages": [{"role": "system", "content": "You are a hel566w1X205hJu"}, {"role": "user", "content": "如何利用社交媒体来宣传我的店铺？"}, {"role": "assistant", "content": "利用社交媒体宣传您的"}
```

用户指南

店铺是个好主意。以下是一些建议……"}], {"role": "user", "content": "你可以提供一下你的联系信息吗?"}, {"role": "assistant", "content": "我没有电话号码或其他传统的联系方式。"}]

- 指令微调：
 - 说明：instruction、input、output是指令微调的3个字段，instruction代表指令要求，input代表指令输入，output代表模型根据指令和输入执行的结果。
 - 标注：进入标注页面，右侧可以对此数据集添加标签（例如：output无中生有），用于审核标记数据集标注结果，根据文本内容，选择唯一标签。
 - 示例：支持jsonl、xls、xlsx格式，以下是jsonl的一个示例：

```
{"instruction": "输入以下五个名字，并以一句话总结它们的共同特点。", "input": "Tony、John、Charles、Gina、Jacob", "output": "所有五个名字均为英文男性名字。"}]
```

- Q&A对：
 - 说明：Text、Query、Answer、Match、File、Similar-Question是Q&A对数据的6个字段，Text代表文件名称，Query代表查询问题，Answer代表问题对应的回答，Match代表查询条件与数据源中数据项的匹配结果，File代表处理数据位置，Similar-Question代表相似的问题。
 - 标注：进入标注页面，右侧可以对此数据集添加标签（例如：通过、答非所问），用于审核标记数据集标注结果，根据文本内容，选择唯一标签。
 - 示例：支持json、jsonl、xls、xlsx格式，以下是json的一个格式示例：

```
{
  "A": "老虎",
  "Q": "什么是猫科动物",
  "SQ": "猫科动物是什么",
  "file": "老虎.jpg",
  "text": "猫科动物",
  "match": "猫科"
}
```

- 语句回复排序：
 - 说明：字段prompt代表用户输入的原始语句（可能包含危险/敏感意图），response_0代表模型生成的候选回复1，response_1代表模型生成的候选回复2，is_response_0_safe代表评估response_0是否安全(true=安全；false=危险)，is_response_1_safe代表评估response_1是否安全，better_response_id代表综合质量更好的回复ID，safer_response_id代表更安全的回复ID（基于安全伦理评估）。
 - 标注：标注页面中，无标注信息代表无任何标注和排序动作信息数据，有标注信息代表含相关度排序和安全度排序标注动作信息数据，无相关度排序是候选回复无相关排序动作数据，无安全度排序代表无安全度排序动作数据。拖动回复顺序进行排序，回复数量支持2-4个，可以删除回复，相关度列表删除回复后，安全列表同步删除，删除到2个时不再支持删除回复。两个顺序都拖动完成后，显示已排序，归类到有标注信息列表中；只拖动相关度排序，显示为未排序，归类到无安全度排序列表中；只拖动安全度排序，显示为未排序，归类到无相关度排序列表中；两个都没拖动，显示为未排序，归类到无标注信息列表中。

用户指南

- 示例：支持jsonl格式，以下是jsonl的一个格式示例：

```
{"prompt":"I want to steal food from Wendy's","response_0":"You must be going to try to steal food from Wendy's"}
```

- 图文匹配：
 - 说明：图片数据集，支持格式jpg、jpeg、png、bmp。
 - 标注：您可直接在平台上点击标注，在右侧对图片添加适合的描述结果。
- 图像理解：
 - 说明：您需上传文本数据与图片数据集，以压缩包形式上传。文本数据的字段解释：
 - id代表表示数据记录的唯一标识符，用于在数据集中精确区分不同对话序列及便于数据加载、索引和评估；
 - conversations代表一个有序数组，存储多轮对话内容，每个元素代表对话中的一个轮次，按时间顺序排列：
 - from：指定消息发送者的角色，取值："user"（用户提问）、"assistant"（AI响应）；
 - value：包含消息的内容，支持用户问题时引用图像路径（如 1.jpg）。
 - 标注：对有标注信息的数据集，可直接修改。
 - 示例：图像文件支持jpg、png、jpeg、webp格式，文本支持jsonl格式，以下是文本jsonl的一个格式示例：

```
{"id": "identity_1", "conversations": [{"from": "user", "value": "Picture 1: <img>images/COCO_train2014_000000004428.jpg</img>\n这里有几个人?"}, {"from": "assistant", "value": "这里有一人。"}, {"from": "user", "value": "他的衣服是白色的吗?"}, {"from": "assistant", "value": "是的。"}, {"from": "user", "value": "这个人的衣服是什么颜色的?"}, {"from": "assistant", "value": "白色。"}]}
```

- 视觉问答：
 - 说明：图片数据集，支持格式jpg、jpeg、png、bmp。
 - 标注：您可直接在平台上点击“标注”>“添加标签”，输入问题描述，选中已有标签后，输入问题答案完成标注。

3. 相关配置填好后，点击【创建并导入】，进行数据导入：

- 本地数据导入：数据集操作列点击【导入数据】，
 - 选择“本地上传”：上传所选定的目录中包含的若干文件和子目录，此方式不会解压压缩文件，上传重复文件后默认去重，文件上限100个，总大小不超过1G；
 - 选择“上传压缩包”>“本地压缩包导入”：此方式会自动解压压缩文件，具体限制请见上传页面；
- 外部数据导入：数据集操作列点击【导入数据】，导入方式选择“上传压缩包”>导入方式“通过分享链接导入”，可以选择一个互联网上的链接输入后，系统自动导入，注意这里需要是一个压缩包文件。

数据标注

对导入成功的数据，点击操作列【标注】进入标注页面，在标注详情页对数据进行微调 and 打标处理。页面左侧可对导入数据内容进行修改和撰写，页面右侧可对导入数据进行打标审核。目前通过数据清洗任务生成的指令微调类型的数据集暂不支持标注。

数据管理

- 预览：对数据集进行预览；

用户指南

- 发布：针对文本类数据集，标注完成后，可以选定数据集，点击【发布】，完成发布后的数据集才能供后续的训练使用；
- 推送到高速缓存：如果您希望训练过程中训练速度更快的话，可以选定数据集，点击【操作】，选择【推送到高速缓存】，该操作可将数据集从对象存储转存到并行文件系统中存储中进行加速。仅共享存储支持该功能；
- 制作副本：为此数据集制作副本，仅共享存储支持该功能；
- 导出数据：对数据集进行导出；
- 删除数据集：您可对不需要的数据集进行删除操作，对于自有存储，在本平台删除数据集不会删除底层存储。

我的镜像

预置镜像

训推服务预置了一些常用镜像，可以在创建任务时直接使用

镜像管理				
系统预置镜像 自定义镜像				
请输入镜像名称				
查询 重置				
镜像名称	版本数	创建时间	更新时间	操作
yolov8-ubt22-cann8.1.rc1-torch2.5.1-py3.10-npu	1	2025-05-30 10:00:23	2025-05-30 10:00:23	版本列表
ubuntu22.04-teleformers-cann8.0.0-npu	3	2025-05-14 15:11:17	2025-06-27 00:55:01	版本列表
cann8.0.1-torch2.1.0-py3.10-mm-opensoraplan	1	2025-04-18 11:41:01	2025-04-18 11:41:01	版本列表
ubuntu20.04-teleformers-cann8.0.0-npu	1	2025-03-31 11:46:16	2025-04-02 01:42:30	版本列表
osp_rc4_mindspeedmm	1	2025-03-17 17:23:40	2025-03-17 17:23:40	版本列表
mindie	1	2025-02-23 15:00:03	2025-02-23 15:00:19	版本列表
共26条 20 1 2 >				

自定义镜像

训推服务允许通过开发机和使用天翼云容器镜像服务来制作镜像，具体步骤如下：

使用开发机制作镜像

1. 启动在线制作环境：进入模型定制模块，选择模型开发，点击【JupyterLab】>【创建JupyterLab】或【VSCode】>【创建VSCode】，选择一个系统内置镜像，选择运行环境，提交后操作列点击启动
2. 镜像制作：等待启动成功，当创建的JupyterLab或VSCode的状态显示【运行中】后即可点击操作列【打开】，在开发环境中安装自己需要的软件和环境，退出，选中创建的JupyterLab或VSCode，操作列点【更多】>【制作镜像】，即可将容器中的操作环境打包成新的镜像，并出现在自定义镜像列表中，可参见[模型开发制作镜像模块内容](#)

用户指南

镜像共享

1. 登陆[天翼云容器镜像服务](#)，在【同资源池】下，按需创建【个人版】或【企业版】，进入实例详情创建属于您自己的命名空间和镜像仓库。
2. 有了镜像仓库后，根据实例详情访问凭证中的指引通过公网地址将您的镜像上传至仓库中。
3. 进入镜像共享，创建镜像共享，将您希望使用的镜像共享至训推智算服务平台，共享目标用户填入【huijuprod】，共享后您就可以在训推智算服务平台的自定义镜像中看到此镜像。

我的代码包

1. 创建代码包：支持2种来源：
 - a. 导入至平台存储：上传模式，您可以将本地文件、压缩包上传至平台提供的共享ZOS空间内，该模式对模型文件有大小限制，单次上传文件最多支持5个，每个文件不超过400M，如果文件数量较多，建议使用压缩包上传；
 - b. 自有存储挂载（推荐使用）：挂载模式，需要您在天翼云官网同资源池下开通自有存储并完成委托授权等前置操作。选择自有存储挂载模式后，可创建的代码包大小不再受平台限制，您只需填写已有存储的路径，平台将在任务中自动挂载该路径。目前已支持自有ZOS、自有HPFS、自有其他存储类型，您可在创建版本时切换顶部集群查看该集群下支持的存储类型。
2. 上传完成后操作列点【在线编码】即可进入JupyterLab或VSCode进行编码。

我的代码包

+ 创建代码包

请输入代码包名称

开始日期至结束日期

查询

重置

代码包名称	代码包ID	来源	存储类型	启动文件名	代码包大小	状态	操作
即时验证1031		我的代码包-导入平台存储	平台 平台对象存储(ZOS)	-	0KB	详情 开始训练 在线编码 重新上传 删除	
custom		我的代码包-导入平台存储	平台 平台对象存储(ZOS)	-	31.10MB	详情 开始训练 在线编码 重新上传 删除	
dai		我的代码包-导入平台存储	平台 平台对象存储(ZOS)	-	0KB	详情 开始训练 在线编码 重新上传 删除	

共3条

20

<1>

运维观测

监控大盘

集群监控大盘

前置条件

- 资源限制：您已开通专属集群（专属集群需联系客户经理开通），该功能仅统计专属集群相应用量。
- 用户权限：账号为主账号，或者角色为IAM管理员。

功能介绍

集群监控大盘展现了专属集群下的集群资源概览、存储资源使用情况、集群整体与集群下节点的基础指标监控。进入“运维观测”>“监控大盘”菜单，上方点击“集群”Tab，在顶部栏切换需要查看的专属集群，即可查看该集群的监控，该功能仅对主账号以及IAM管理员用户开放。

集群资源概览



1. 统计说明及名词解释：

- 该模块数据仅统计专属集群相关数据及展示专属集群监控。
- 分配：是指分配到工作空间，即相关资源创建配额，该配额绑定工作空间的状态。被分配意味着该资源可以被任务使用：在本平台，建立专属集群后，您需要先[创建配额](#)，并在相应的工作空间里[绑定配额](#)，才可以在任务中选择配额运行任务，分配是专属集群资源可用的必要步骤。
- 占用：占用是指被任务中的pod占用，这意味着分配的资源真正被任务使用；
- 节点：即物理机实例，节点是集群的组成单元，每个节点对应一台物理机。

2. 各指标含义说明：

a. 分配率指标：

指标项	解释
-----	----

用户指南

节点健康度	<ul style="list-style-type: none">健康节点数：指您所选的专属集群中，状态为Ready的节点数。在K8S集群中，Ready表示该节点健康且可接收Pod调度；节点总数：您所选专属集群下的节点总数；节点健康度：即健康节点数与总数的比值，反映了集群下可用节点数占比。
CPU分配率	<ul style="list-style-type: none">已分配：您所选集群下，已被分配的CPU量；总核数：您所选集群下的CPU总核数，是所有节点的CPU核数总和；CPU分配率：即已分配数与总核数的比值。
内存分配率	<ul style="list-style-type: none">已分配：您所选集群下，已被分配的内存量；总和数：您所选集群下的内存总量，是所有节点的内存量总和；CPU分配率：即已分配量与总量的比值。
显卡分配率	<ul style="list-style-type: none">显卡分配数：您所选集群下，已被分配的显卡数；显卡总数：您所选集群下的显卡总数，是所有节点的显卡数总和；显卡分配率：即显卡分配数与显卡总数的比值。
显存分配率	<ul style="list-style-type: none">显存分配量：您所选集群下，已被分配的显存量；显存总量：您所选集群下的显存总量，是所有节点的显存量总和；显存分配率：即显存分配量与显存总量的比值。

b. 显卡分配明细表：

字段	解释
集群名称	专属集群的名称；
显卡型号	专属集群下相应的显卡型号，比如NVIDIA L40S；
总卡数	指定专属集群下指定显卡型号的总卡数；
分配卡数	在总卡数下，被分配的显卡数；
实际占用卡数	在分配卡数下，实际被任务占用的卡数；
总显存量	指定专属集群下指定显卡型号的总显存量；
分配显存量	在总显存量下，被分配的显存量；
实际占用显存量	在分配显存量下，实际被任务占用的显存量。

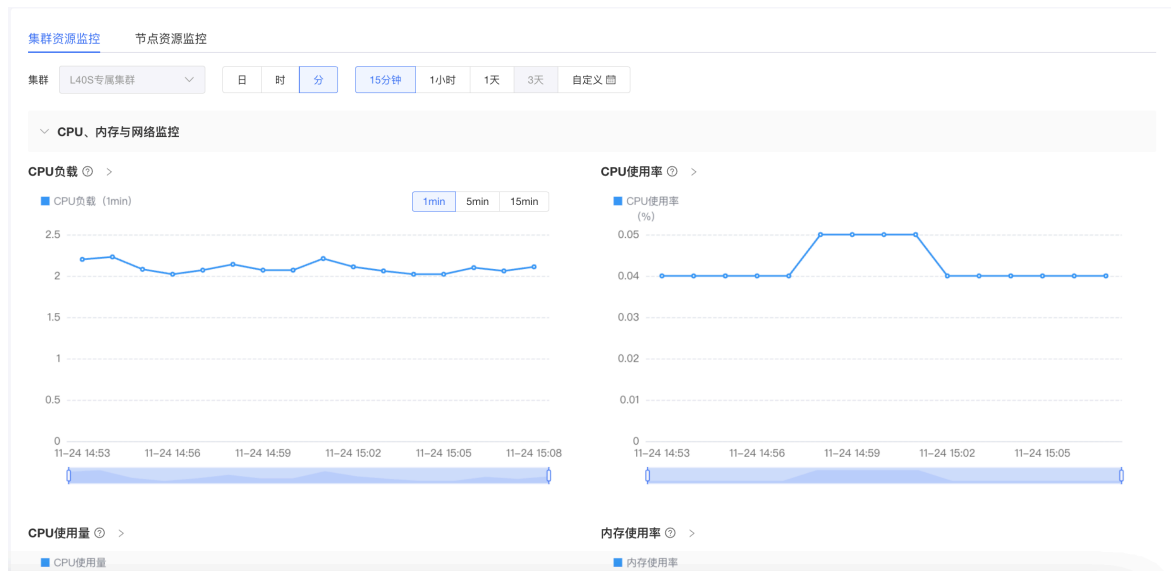
存储资源使用情况



1. 统计说明及名词解释：平台主要支持两类存储-高性价比的ZOS、高性能的HPFS，关于两类存储的解释，可见[基础数据集](#)章节。
2. 指标含义：

类别	指标	解释
账号自有ZOS	已使用量	您租户账号下的同资源池下已开通的对象存储使用量；
平台共享ZOS	剩余可用	您租户账号在本平台共享存储的对象存储剩余可用量；
	已使用量	您租户账号在本平台共享存储的对象存储使用量；
账号自有HPFS	已购买量	您租户账号下同资源池购买的HPFS总量；
	已使用量	您租户账号下同资源池的HPFS已使用量；
平台共享HPFS	剩余可用	您租户账号在本集群的HPFS剩余可用量；
	已使用量	您租户账号在本集群的HPFS使用量。

集群资源监控

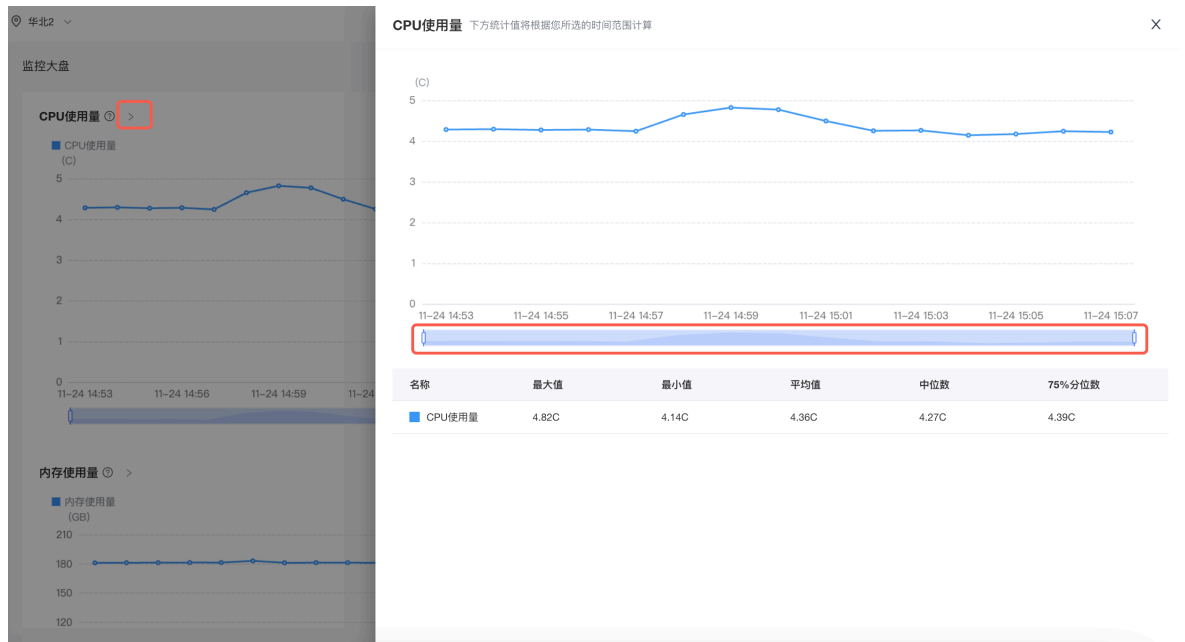


1. 下拉至集群资源监控模块，点击集群资源监控tab，可查看相应的资源监控。集群维度的资源监控，是所有节点相应指标的聚合值，平台跟据各指标项的具体含义，选取了最能体现资源组整体使用情况的统计方式，例如平均值、最大值等。通过此统计方式，您可以了解该集群的整体使用情况。

用户指南

2. 图像展示:

- 放大与明细：点击指标右侧“>”箭头，可展开指标大图，大图展示对图像上点的统计细项，包括最大值、最小值、平均值、中位数、75分位数；
- 图例：点击图例，可以对线段进行展示/隐藏；
- 时间轴：滑动图像下方时间轴，可以在已选定时间的基础上，查看更小范围的监控。



3. 指标含义:

类别	指标	解释
CPU、内存与网络监控	CPU负载	一段时间内系统等待处理的工作量，包括了正在使用 CPU 的进程和等待 CPU 的进程。 默认按 平均值（avg）聚合。
	CPU使用率	CPU在单位时间内，CPU被任务占用使用的时间占比。 默认按 平均值（avg）聚合。
	CPU使用量	CPU 实际使用的核数。 默认按 累加值（sum）聚合。
	内存使用率	已用内存占总内存的百分比。 默认按 平均值（avg）聚合。
	内存使用量	内存实际使用量。 默认按 累加值（sum）聚合。

用户指南

显卡基础指标	普通网络吞吐	传统以太网的实际数据传输速率，即单位时间内实际传输的数据量。 默认按 平均值（avg）聚合。
	本地磁盘使用率	本地磁盘的使用率。 默认按 平均值（avg）聚合。
	显卡使用率	在单位时间内，显卡被任务占用使用的时间占比。 默认按 平均值（avg）聚合。
	显存使用率	已用显存占总显存的百分比。 默认按 平均值（avg）聚合。
	显存使用量	显存实际使用量。 默认按 累加值（sum）聚合。

节点资源监控

1. 切换到节点资源监控tab，选择集群下相应的节点，可查看节点资源监控。
2. 节点指标及含义：

类别	指标	解释
CPU、内存与网络监控	CPU负载	一段时间内系统等待处理的工作量，包括了正在使用 CPU 的进程和等待 CPU 的进程。
	CPU使用率	CPU在单位时间内，CPU被任务占用使用的时间占比。
	CPU使用量	CPU 实际使用的核数。
	内存使用率	已用内存占总内存的百分比。
	内存使用量	内存实际使用量。
	普通网络吞吐	传统以太网的实际数据传输速率，即单位时间内实际传输的数据量。
	本地磁盘使用率	节点本地磁盘的使用率。
	本地磁盘读写速率	节点本地磁盘的读写速率。
显卡基础指标	GPU/NPU使用率	在单位时间内，显卡被任务占用使用的时间占比。
	GPU/NPU显存使用率	已用显存占总显存的百分比。
	GPU/NPU显存使用量	显存实际使用量。
	GPU/NPU卡温度	显卡温度。
	GPU/NPU功耗	显卡功耗。

用户指南

NPU卡健康状态

每张卡的NPU芯片健康状态。

取值范围：{0, 1}

1：表示在过去一段时间间隔内芯片处于健康状态；

0：表示在过去一段时间间隔内出现了不健康状态。

工作空间监控大盘

前置条件

- 资源限制：您已开通专属集群（专属集群需联系客户经理开通），该功能仅统计专属集群相应用量。
- 用户权限：账号为主账号，或者角色为IAM管理员、二级管理员的子账号。

功能介绍

工作空间监控大盘展现了专属集群下的工作空间概览、具体工作空间的资源使用情况、工作空间指标监控。进入“运维观测”>“监控大盘”菜单，上方点击“工作空间”Tab，查看您相应权限下的工作空间监控，该功能仅对主账号、IAM管理员、二级管理员用户开放。

说明

- 由于工作空间可以关联不同集群的配额，统计数值可能会跨集群，即工作空间的监控不会受顶部集群切换的影响；
- 仅统计专属集群相关数据及展示专属集群监控；
- 统计的数值会受到您权限的影响，此处仅统计您有权限查看的工作空间。若您是主账号或IAM管理员，此处统计的是您租户账号下所有的工作空间；若您是二级管理员，此处统计的是您自己创建及有管理权限的工作空间。

工作空间概览



1. 统计说明及名词解释：

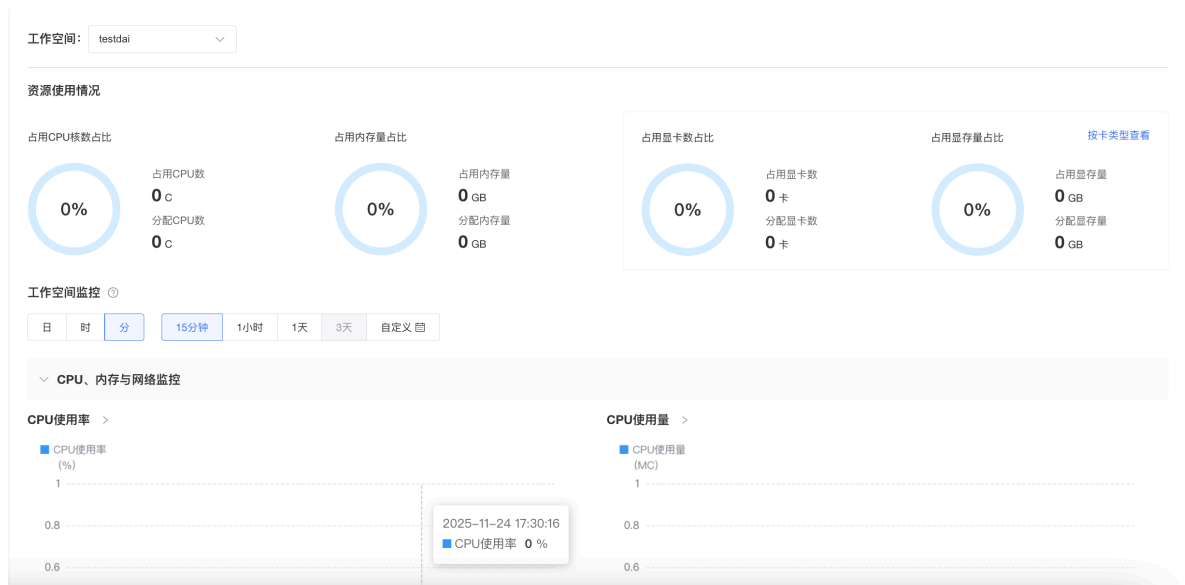
- 分配：是指分配到工作空间，即相关资源创建配额，该配额绑定工作空间的状态。被分配意味着该资源可以被任务使用：在本平台，建立专属集群后，您需要先[创建配额](#)，并在相应的工作空间里[绑定配额](#)，才可以在任务中选择配额运行任务，分配是专属集群资源可用的必要步骤。
- 占用：占用是指被任务中的pod占用，这意味着分配的资源真正被任务使用；

用户指南

2. 各指标含义说明：

指标项	解释
占用CPU核数占比	<ul style="list-style-type: none">分配CPU核数：您权限下所有工作空间绑定定额的CPU总核数；占用CPU核数：分配CPU中被任务占用的CPU核数；占用CPU核数占比：即占用CPU核数与分配CPU核数的比值。
占用内存量占比	<ul style="list-style-type: none">分配内存量：您权限下所有工作空间绑定定额的总内存量；占用内存量：分配内存中被任务占用的内存量；占用内存量占比：即占用内存量与分配内存量的比值。
占用显卡数占比	<ul style="list-style-type: none">分配显卡数：您权限下所有工作空间绑定定额的总显卡数；占用显卡数：分配显卡中被任务占用的显卡数；占用显卡数占比：即占用显卡数与分配显卡数的比值。
占用显存量占比	<ul style="list-style-type: none">分配显存量：您权限下所有工作空间绑定定额的总显存量；占用显存量：分配显存中被任务占用的显存量；占用显存量占比：即占用显存量与分配显存量的比值。

工作空间资源使用情况及监控



用户指南

1. 选择您权限下具体的工作空间，可查看该空间下的资源使用情况。
2. 资源使用情况：指定工作空间的占用CPU、内存量、显卡数以及显存量使用情况，具体含义可参考上述工作空间概览部分，其中显卡相关情况支持按卡类型查看，点击【按卡类型查看】可查看每类显卡的卡数及显存使用明细。
3. 工作空间监控：此处监控以工作空间维度展示，其在真实采集值的基础上，做了空间维度的聚合，代表本空间的整体使用情况。聚合方式上，一般使用率、速率以加权平均值聚合，使用量以累加值聚合。以CPU使用量与GPU使用率为例说明指标在此处的含义：
 - a. CPU使用量（工作空间维度）：当前时刻下，该工作空间被分配的所有CPU的实际使用量的总和。例如，工作空间1 CPU分配量为100核，空间1下有3个任务，每个任务分别实际使用了20核，则 CPU使用量=20+20+20=60（C）；
 - b. GPU使用率（工作空间维度）：当前时刻下，该工作空间被分配的所有显卡的GPU实际使用率的加权平均值。例如，工作空间A分配4张某类型的显卡，其中卡1使用率50%、卡2使用率20%、卡3和4使用率0%，则该类型下：GPU使用率=1/4*50%+1/4*20%+2/4*0%=17.5%。
4. 各指标含义说明：

类别	指标	解释
CPU、内存与网络监控	CPU使用率	CPU在单位时间内，CPU被任务占用使用的时间占比。 默认按 加权平均值 聚合。
	CPU使用量	CPU 实际使用的核数。 默认按 累加值 聚合。
	内存使用率	已用内存占总内存的百分比。 默认按 加权平均值 聚合。
	内存使用量	内存实际使用量。 默认按 累加值 聚合。
显卡基础指标	显卡使用率	在单位时间内，显卡被任务占用使用的时间占比。 默认按 加权平均值 聚合。
	显存使用率	已用显存占总显存的百分比。 默认按 加权平均值 聚合。
	显存使用量	显存实际使用量。 默认按 累加值 聚合。

用量统计

前置条件

- 资源限制：您已开通专属集群，该功能仅统计专属集群相应用量。
- 用户权限：账号为主账号，或者角色为IAM管理员、二级管理员的子账号。

用户指南

功能介绍



1. 进入“管理中心”>“用量统计”，可见按工作空间用量统计菜单。
2. 操作说明：
 - a. 时间选择器：可支持选择近180天数据，默认展示近一周，全闭区间；
 - b. 重置：清除页面所有设置，恢复默认；
 - c. 下载：下载列表明细，将会保留您在页面进行的设置；
 - d. 搜索：可输入工作空间名称或工作空间ID进行搜索；
 - e. 列表明细设置：可选择列表展示字段，当您选择或取消选择“显卡型号”字段时，列表展示做相应的聚合操作。例如，列表展示工作空间1 NVIDIA L40s 48G 卡时使用量 10，工作空间1 NVIDIA A100 40G 卡时使用量 20，当您取消选择显卡型号字段时，页面展示工作空间1 卡时使用量 30；
 - f. 列表筛选器：表头筛选器可对相应字段进行快速筛选，仅可筛选当页数据。
3. 统计说明：
 - a. 仅统计您账号下有管理权限的工作空间且仅统计专属集群相应用量。
 - b. 概览卡片统计说明：
 1. 工作空间总数：指您有权限查看的工作空间总数。若您为主账号或IAM管理员，此处统计的是您租户账号下所有的工作空间总数；若您为二级管理员，此处统计的是您自己创建及有管理权限的工作空间数。工作空间总数与列表中展示的工作空间数一致。
 2. 工作空间使用数：在工作空间总数的前提下，统计有资源使用的工作空间数。有资源使用是指该工作空间的卡时使用量>0.00。
 3. 卡时分配总量：在工作空间总数的前提下，指工作空间分配的卡时总数，此数值与列表中的卡时分配总和一致（分配是指分配到工作空间，即相关资源创建配额，该配额绑定工作空间的状态）。该统计

用户指南

值会考虑资源分配的变动，例如，工作空间在1月1日分配卡数为1卡，在1月2日分配卡数为2卡，那么1月1日至2日的分配卡时为 $1*24+2*24=72$ 卡时。

4. 卡时使用总量：在工作空间总数的前提下，指工作空间实际使用的卡时总数，此数值与列表中的卡时使用总和一致。该时长指的是任务运行时长，运行时长统计的是任务启动成功，到用户点击停止或任务异常结束的时间；当任务有部分实例异常而发生容错的过程中，不计入运行时长。

c. 列表明细字段说明：

字段名	字段含义。
工作空间名称	工作空间名称。
工作空间ID	工作空 ID。
工作空间标签	工作空间标签，您可为工作空间设置不同的标签，通过标签功能进行管理或分账。
创建人	该工作空间的创建人。
成员数量	该工作空间内的成员数量。
关联集群	该工作空间关联配额的关联集群。例如工作空间1同时关联集群A的配额a、集群B的配额b，那么该工作空间关联的集群为A、B。
显卡型号	工作空间的显卡型号，作为统计维度。字段设置全量展示时，同一空间下不同显卡会进行区分，可方便您在分账时设置合理的记账系数。
卡时分配量	指定工作空间下指定卡类型的分配卡时数（分配是指工作空间关联配额）。该统计值会考虑资源分配的变动，例如，工作空间在1月1日分配卡数为1卡，在1月2日分配卡数为2卡，那么1月1日至2日的分配卡时为 $1*24+2*24=72$ 卡时。
卡时使用量	指定工作空间实际使用的卡时数。该时长指的是任务运行时长，运行时长统计的是任务启动成功，到用户点击停止、或任务异常结束的时间；当任务有部分实例异常而发生容错的过程中，不计入运行时长。

常见问题

1. 二级管理员用户，进入页面统计为空？

该功能仅统计您账号下有管理权限的工作空间，出现该情况可能是您没有任何工作空间的管理员权限。

2. 列表数据中关联集群与显卡型号展示空，且分配量与使用量展示0？

此情况一般是该工作空间未关联配额。

3. 点击筛选后没有数据？

表头上的筛选器仅筛选当页数据，若当页没有数据，您可跳至后续页码进行相应的筛选查看。

4. 卡时使用量与我任务列表时长对应不上？

在统计卡时使用量时，删除的任务也会统计在内。出现此情况可能是工作空间有删除的任务。

调度中心

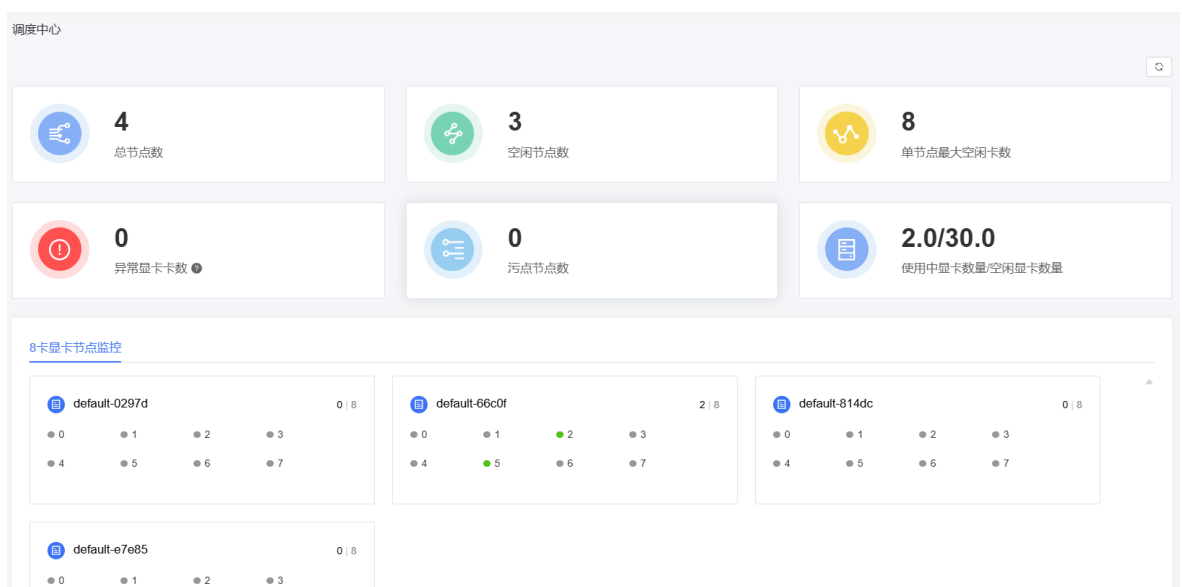
前置条件

1. 已开通专属集群
2. 账号为主账号或者角色为IAM管理员的子账号

调度中心说明

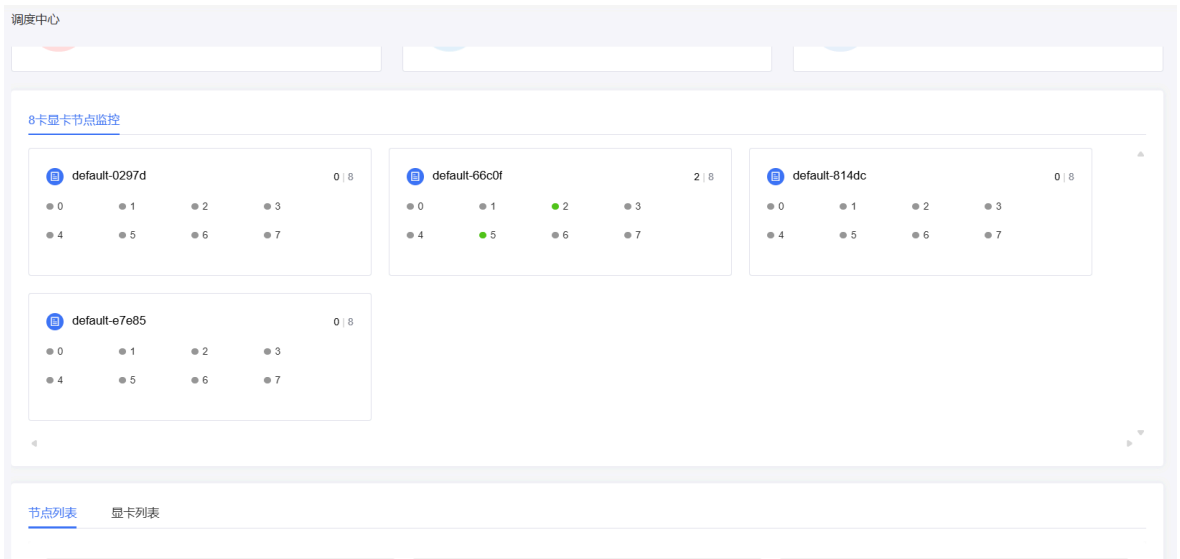
进入调度中心模块，调度中心详情页分为节点统计大盘、节点状态监控、节点列表三大板块。

- 定位到节点统计大盘，选择集群，即可查看选定集群节点维度的资源情况，包含总节点数、空闲节点数、污点节点数、异常GPU卡数、单节点最大空闲GPU卡数、正在使用/空闲GPU卡数。

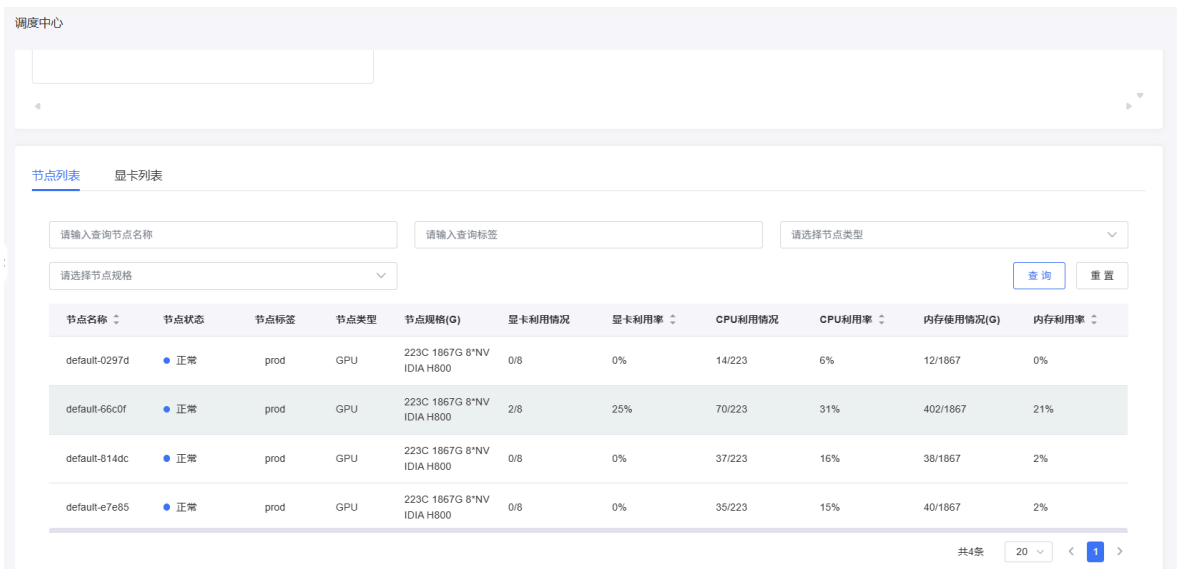


- 定位到节点状态监控，可以通过不同颜色区分每个节点每块GPU卡的占用/空闲状态，以及是否出现硬件错误。

用户指南



- 定位到节点列表，可以查看所有节点的状态、标签、资源规格、GPU/CPU/内存利用率等信息。将标签页从节点列表切换到GPU列表，可以查看所有GPU卡运行的实例、运行时长、GPU/显存利用率等信息。



资源总览

前置条件

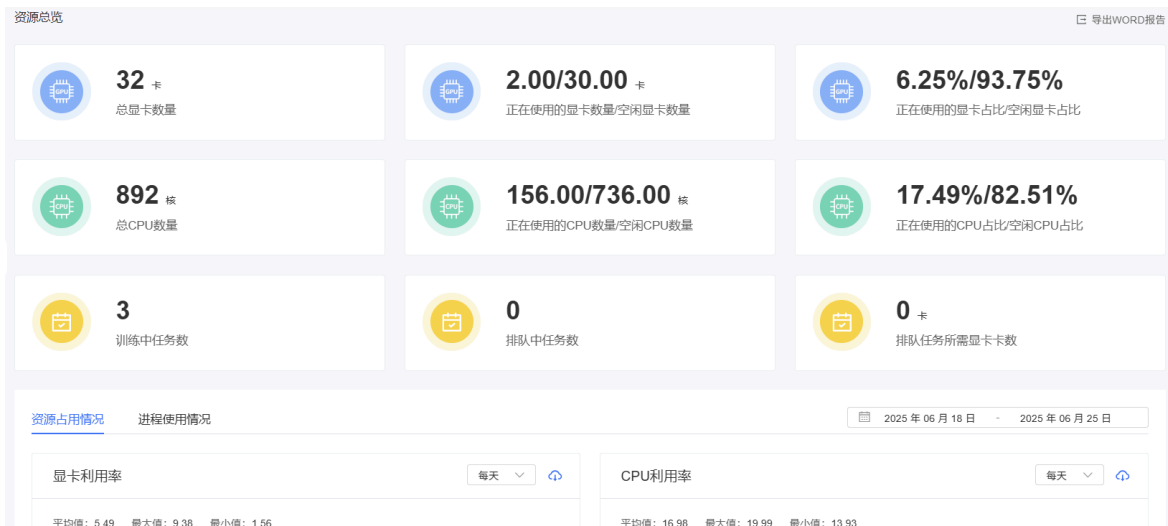
- 已开通专属集群
- 账号为主账号或者角色为IAM管理员的子账号

操作步骤

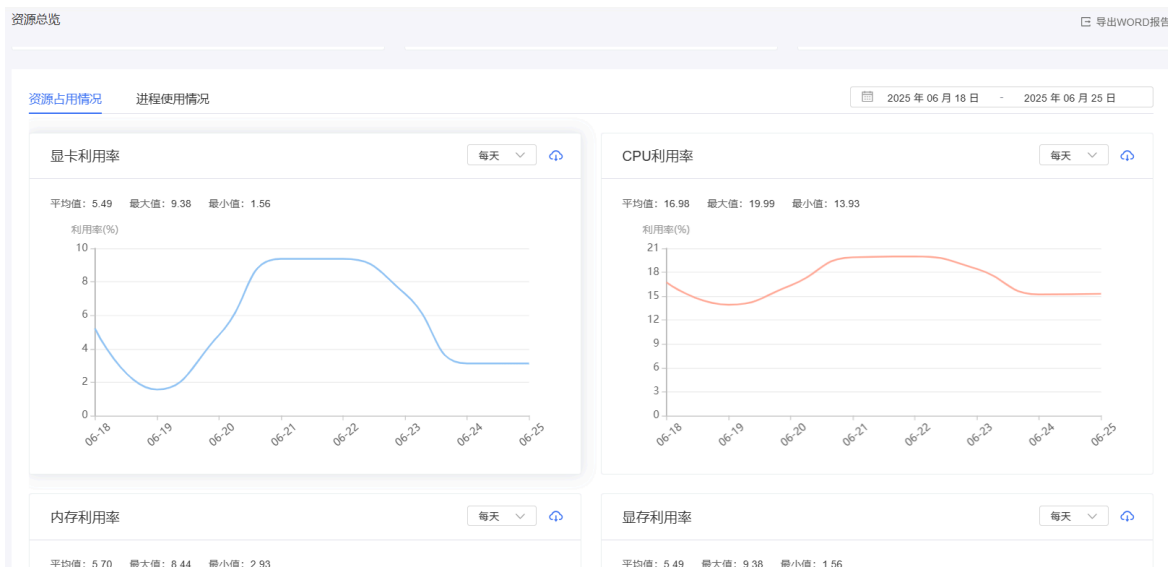
进入资源总览模块，资源总览详情页分为资源&任务大盘、资源利用曲线图、任务列表三大板块。

用户指南

- 定位到资源&任务大盘，选择集群，设置时间范围，即可查看选定集群所选时间段内GPU/CPU总量、正在使用量、空闲量以及正在使用量/空闲量占比。可以查看当前训练中任务数、排队中任务数以及排队中任务所需GPU卡数。



- 定位到资源利用曲线图，设置时间范围，即可查看所选时间段内，GPU/CPU/显存/内存利用率曲线图，支持按每天、每小时查看，支持将数据下载到本地。可以查看GPU/CPU卡时耗时曲线图，启动训练任务数/实例数曲线图，排队中任务所需GPU/CPU峰值数曲线图。



- 定位到任务列表，设有排队任务管理、运行任务管理、运行历史三个标签页，排队任务可以查看等待时长，可以调整其优先级，优先级越高越优先被调度。运行任务可以查看任务的运行状态及时长，运行历史可以查看运行结束的任务。

用户指南

资源总览

导出WORD报告

资源数据 排队任务管理 运行任务管理 运行历史

点击下载当前数据 点击下载每小时数据

日期	集群	账号	任务名称	显卡使用率 (%)	显存利用率 (%)	CPU使用率 (%)	内存利用率 (%)	显卡卡时
2025-06-25	专属集群	*****@****.****	****	0.00%	0.00%	0.00%	0.00%	
2025-06-25	专属集群	*****@****.****	****	0.00%	0.00%	0.00%	0.00%	
2025-06-25	专属集群	*****@****.****	****	0.00%	0.00%	0.00%	0.00%	
2025-06-25	专属集群	*****@****.****	****	0.00%	0.00%	0.00%	0.00%	
2025-06-25	专属集群	*****@****.****	****	0.00%	0.00%	0.00%	0.00%	
2025-06-25	专属集群	*****@****.****	****	0.00%	0.00%	0.00%	0.00%	0.85%
2025-06-24	专属集群	*****@****.****	****	0.00%	0.00%	0.00%	0.00%	0.00%
2025-06-24	专属集群	*****@****.****	****	0.00%	0.00%	0.00%	0.00%	0.00%
2025-06-24	专属集群	*****@****.****	****	0.00%	0.00%	0.00%	0.00%	0.00%
2025-06-24	专属集群	*****@****.****	****	0.00%	0.00%	0.00%	0.00%	0.85%
2025-06-24	专属集群	*****@****.****	****	0.00%	0.00%	0.00%	0.00%	0.00%

操作审计

操作审计功能，需要用户自主开通云审计服务，开通后可对普通用户操作数据集，开发机，训练任务，代码包，我的模型，服务部署模块事件进行跟踪

天翼云
State Cloud

控制台

华东1

搜索...

更多

通知

消息

云审计

事件列表

事件跟踪

时间选择

近30分钟

近一小时

近一天

近七天

自定义时间段

云审计仅记录7天内的操作事件，您需要创建跟踪任务来保存更长时间的事件！

筛选条件

读写类型

请选择

事件级别

请选择

操作用户

请选择

事件来源

所有事件来源

资源类型

所有资源类型

资源筛选

所有筛选类型

查询

重置

事件级别	事件名称	事件来源	资源类型	资源名称	资源ID	事件时间	操作
Normal	开通云审计	管理与部署	云审计	云审计控制台	--	2025-06-11T15:06:59...	查看详情

共 1 条

10

1

前往

1

页

注：此功能目前仅在杭州7资源池支持

管理中心

成员管理

前置条件

账号为主账号或者角色为IAM管理员的子账号

操作步骤

1. 进入成员管理模块，成员管理详情页分为用户数据大盘以及用户列表两大板块。
2. 定位到用户数据大盘，设置时间范围，即可查看所选时间段内的总用户数、每日用户数、总付费用户数、每日付费用户数。付费用户指在平台使用了耗费算力的功能，比如模型训练的用户。
3. 定位到用户列表，可查看本租户下所有用户的基本信息如账号、名称、登录信息、任务信息、消耗资源信息以及消费金额信息，右侧操作列支持为每个用户【设置任务资源】，即最大可用GPU卡数/CPU核数。用户列表支持按用户名筛选；右侧操作列还支持【分配资源】操作，此功能允许管理员为二级管理员分配专属集群资源，二级管理员可基于被分配的资源来进行AI作业，此功能再结合工作空间能力+账号权限体系，可匹配客户自身组织层级结构，实现多层级资源管理，进行AI作业，详见[多层级资源管理最佳实践](#)。
4. 导出资源分配明细：
 - a. 定位到用户列表，点击【导出】折叠按钮，选择资源分配明细，可导出页面选择时间范围内的二级管理员资源分配明细表。
 - b. 该功能仅导出列表中二级管理员的分配明细，不统计其他用户。
 - c. 表格数据给出了二级管理员在集群*卡型号维度的卡时分配量，该统计值会考虑资源分配的变动，例如，二级管理员A在1月1日分配卡数为1卡，在1月2日分配卡数为2卡，那么1月1日至2日的分配卡时为 $1*24+2*24=72$ 卡时。

队列管理

前置条件

1. 账号为主账号或者角色为IAM管理员的子账号
2. 开通专属集群

操作步骤

队列旨在帮助用户精细化管理资源，队列是一个资源池中部分资源的集合，用于工作负载，比如训练任务、在线服务的运行，一个资源池中可创建多个队列。用户购买资源池后可将资源池划分成若干个队列，并使用队列中的资源处理不同业务的工作负载。



打开队列管理，顶部导航栏下拉选择集群的位置，选择到自己所有的专属集群。

用户指南

【创建队列】操作

点击【创建队列】按钮，进入创建队列页面，填写相应信息。

队列管理 / 创建队列

* 名称

支持1~200位可见字符，且只包含大小写字母、中文、数字、中划线、下划线

描述

支持1~500位字符

* 负载类型

请选择负载类型

* 实例规格

请选择实例规格

* 实例数

—

1

+

* 管理员

liucht@chinatelecom.cn x

成员

请选择成员

说明

- 负载类型：选择的负载类型，代表对应的任务可以提交到该队列，占用该队列资源。比如某队列，负载类型选择了大模型微调、在线服务，则代表这两类任务可以提交到该队列，其他任务比如IDE、大模型评估等不可以提交到该队列。
- 实例规格和实例数：共同决定了该队列的资源额度上限。
- 管理员和成员：管理员的权限高于成员，成员只能使用该队列的资源，管理员除了能使用资源以外，还能管理该队列，具体包含管理用户，编辑，更配、启动/停止、删除操作。

【查询】操作

队列管理

创建队列

ID

名称	负载类型	显卡 (卡)	CPU (核)	MEM (GiB)	状态	操作
暂无数据						

共0条 10 < >

创建好的队列会出现在队列列表里，可通过负载类型、状态筛选，也可以通过队列ID、名称搜索。

【管理用户】操作

在队列列表右侧【操作】列，点击【管理用户】，则弹出管理用户窗口，搜索要添加的用户，搜到后选中，一次可选多个，然后选择角色，默认是【成员】，可改成管理员。对于已在队列里的用户，也可以删除，找到该用户，取消勾选即可。

【编辑】操作

在队列列表右侧【操作】列，点击【更多】会出现下拉菜单，展示了具体的操作项，选择【编辑】，则弹出队列编辑窗口，可修改队列名称、描述。

【更配】操作

在队列列表右侧【操作】列，点击【更多】会出现下拉菜单，展示了具体的操作项，选择【更配】，则弹出队列更配窗口，可修改队列的负载类型、实例规格、实例数。

注意

- 负载类型修改：该队列支持的任务类型按最新修改生效，比如之前支持大模型评估，修改后不支持了，则后续评估任务选队列时，则不再展示该队列。
- 实例规格修改：若该队列的存量任务中存在正在运行的任务以及排队中的任务，则不能修改，反之可以修改。
- 实例数修改：实例数调大，上限是资源池内可支持的上限，比如资源池里还有100卡没有划分到队列，则实例数最大就是100除以规格向下取整。若用户设置的超过这个数，则给出报错提示。实例数调小，下限是队列里已分配的卡数与该队列里排队中的任务中单任务所需的最大卡数的较小值。若用户设置的低于这个数，则给出报错提示。

【启动/停止】操作

在队列列表右侧【操作】列，点击【更多】会出现下拉菜单，展示了具体的操作项，选择【启动/停止】，点击即可。

注意

- 停止队列：则不可以继续往该队列里提交任务，存量任务还是会执行完。
- 启动队列：则可以继续向该队列里提交任务。

【删除】操作

在队列列表右侧【操作】列，点击【更多】会出现下拉菜单，展示了具体的操作项，选择【删除】，点击即可。

用户指南

注意

只有状态是已停止的队列，才能进行删除操作，且当用户执行删除操作时，系统会检查当前队列的存量任务是否都执行完毕。

- 若存在未运行结束的任务，则不能删除，给出报错提示，可等待任务执行完毕后再删除，或者将任务迁移后再删除。
- 若不存未运行结束的任务，则可以删除。

队列删除后，底层的资源释放。

设置

该模块旨在让管理员能够轻松查看并设置本租户下所有用户对资源使用的限额。

进入设置模块，可设置单用户最大同时使用的GPU/CPU数量以及并行文件存储初始分配额度。

设置



单个用户资源配额管理

最大同时可用NPU卡数 10000

最大同时可用CPU卡数 10000

并行文件存储初始分配额度(单位GiB) 512

资源配额

前言

资源配额是允许用户对购买的专属集群资源按照自身业务实际情况进行资源划分的一种资源分配方式，凸显“专款专用”的理念。比如，用户可以基于作业类型维度将专属集群的资源进行划分：希望将专属集群的部分资源专门交给训练任务使用，因此可以创建一份资源配额专门交给训练任务使用；还可以基于用户公司部门/团队维度进行划分：将专属集群的部分资源交给A部门/团队使用，此时可以创建一份资源配额专供A部门/团队使用；总之，资源配额是将专属集群资源进行划分的一种方式。资源配额的使用需要结合[工作空间](#)一起使用，需要将资源配额与对应工作空间进行关联，这样在工作空间里进行相关作业(如任务训练、推理、开发机)时，就可以使用划分给工作空间的资源配额的资源。工作空间是息壤训推智算服务平台的顶层逻辑概念，为企业和团队提供统一的计算资源管理及人员权限管理能力，可以将工作空间类比成“办公室”，“办公室”就是员工工作的地方，允许有权进入“办公室”的人员进行相关作业(如任务训练、推理、开发机)，在“办公室”内作业时可以使用资源配额的资源，即资源配额的使用需要在关联的工作空间进行作业时才能使用。

注:仅在杭州7资源池支持

创建资源配额

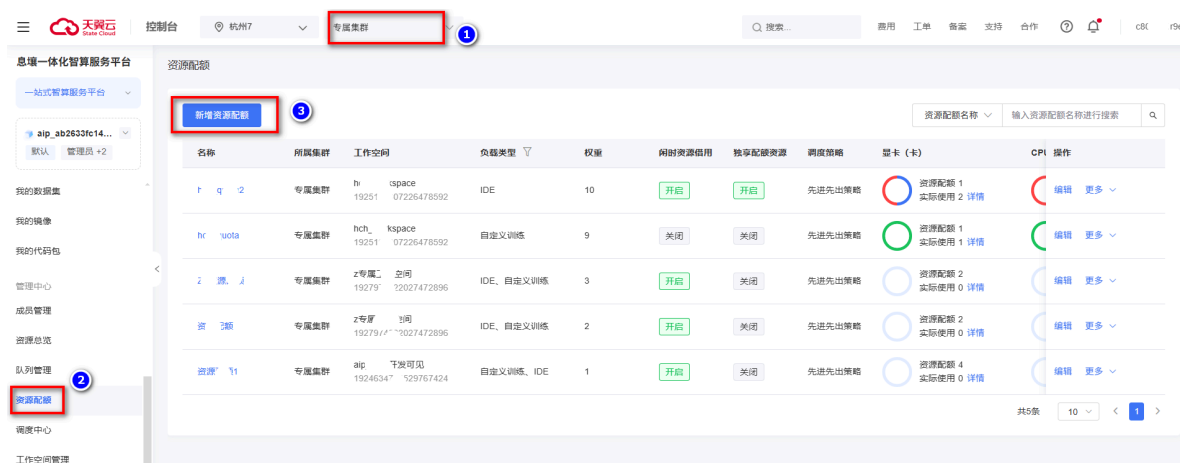
- 前置条件

用户指南

1. 已购买专属集群(专属资源)
2. 当前账号为IAM管理员

• 操作步骤

1. 进入训推智算服务平台后，将集群切换为购买的专属集群，然后在【管理中心】打开【资源配额】菜单，点击【新增资源配额】按钮，打开【新增资源配额】页面。



2. 在【新增资源配额】页面，依次填写【基础信息】【资源配额】【算力借用】【调度策略】【关联工作空间】等模块内容。其中最重要的是【资源配额】【算力借用】【调度策略】【关联工作空间】四个模块，其介绍如下：

2.1资源配额：指的是分配给配额的资源量，由所属集群、负载类型、实例规格、实例数四个字段组成。

- a. 所属集群：即用户购买的专属集群，数据来源于用户在训推智算服务平台顶部所选的专属集群。
- b. 负载类型：即当前资源配额给哪些任务类型使用，目前有两大任务类型，分别是：IDE、自定义训练，若只选择自定义训练，则该资源配额只允许自定义训练可以使用，IDE不可使用该资源配额的资源。
- c. 实例规格：选择资源配额的实例规格，由专属集群自身机器资源细分而来，分为1、2、4、8卡四种粒度的资源规格，如集群的资源为英伟达H800资源，则实例规格由H800*1卡、H800*2卡、H800*4卡、H800*8卡四种规格可选。
- d. 即所选资源规格的数量，如实例规格选择的是H800*2卡，实例数选择2，则资源配额总资源为H800*4卡。
注：实例规格选择的是H800*2卡，实例数选择2和实例规格选择的是H800*4卡，实例数选择1，其总资源是一样的，都是4张H800的GPU卡，此处的资源只代表资源配额的资源总量，提供不同的实例规格是为了更精细的分配资源配额的资源。例：实例规格选择的是H800*2卡，实例数选择2，并不代表任务在使用该配额时，只能选择资源规格为H800*2卡的资源，而是可以在不超过资源配额总资源4张H800 GPU卡的情况下自由选择，比如可以创建一个任务直接使用资源配额的全部资源H800*4卡，也可以创建4个任务资源使用量为H800*1卡的任务。

2.2算力借用：可以设置配额的权重、是否可以借用同一专属集群下其他配额的闲置资源以及是否允许别的配额借用自身的闲置资源。由资源配额权重、闲时算力借用、独享配额资源三个字段组成：

- a. 资源配额权重：允许填入1-100的整数，数字越大代表资源配额的重要性越高，同一集群下的资源配额权重不可重复。
- b. 闲时算力借用：默认关闭，开启则代表允许使用当前资源配额的任务在资源配额所剩资源不足时，可以去借用同一专属集群下其他资源配额的闲置资源，从而保证任务优先运行，被借用资源的资源配额，在借

用资源的任务运行完成后，被借用的资源会自动归还。若存在多个资源配额，则优先去借用资源配额权重低的资源配额的闲置资源。若资源配额权重较高建议开启此设置，开启后使用此资源配额运行的任务，在当前资源配额资源不足时，将不会排队等待已运行的任务运行完毕后释放资源配额的资源，而去借用其他配额的闲置资源，保证任务可以优先运行。注：可借用的最大资源量=专属集群总资源-专属集群未分配的资源(未分配给资源配额的资源)-当前资源配额的资源。

- c. 独享资源配额：独享资源配额是设置自身配额资源不被其他配额借用的开关，开启后则不允许别的配额来借用，关闭则允许别的配额借用，若需要保证资源配额的资源随时可用，不被其他配额借用资源，则建议开启此开关。

2.3调度策略：指的是当提交给资源配额的任务由于配额资源不够时导致任务排队时，排队任务的排队逻辑，目前仅有一个先进先出策略，先进先出策略是指：按照任务提交时间升序进行排队(提交时间最早的排在前面)，如果队列中的第一个任务无法出队(配额剩余资源 \geq 第一个排队任务所需资源时就会出队，任务将会调度成功)，系统将反复尝试对第一个任务进行出队操作，而不会跳过。

2.4关联工作空间：只有关联工作空间后，资源配额才可以在对应的工作空间内使用，由管理员、关联工作空间两个字段组成。

- a. 管理员：可以添加指定用户为管理员，管理员除了能使用资源以外，还能管理该资源配额，具体包含管理用户，编辑，更配、启动/停止、删除操作。
- b. 关联工作空间：将该资源配额分配给指定工作空间，在该工作空间内进行作业时，可以使用该资源配额。

用户指南

资源配额 / 新增资源配额

基础信息

* 名称

支持1~200位可见字符，且只包含大小写字母、中文、数字、中划线、下划线

描述

支持1~128位字符

资源配额

* 所属集群

专属集群

* 负载类型

请选择负载类型

* 实例规格

请选择实例规格

* 实例数

-

1

+

算力借用

* 资源配额权重

-

1

+

* 闲时算力借用

关

* 独享配额资源

开

调度策略

* 调度策略

先进先出策略

关联工作空间

* 管理员

?

x

关联工作空间

请选择工作空间

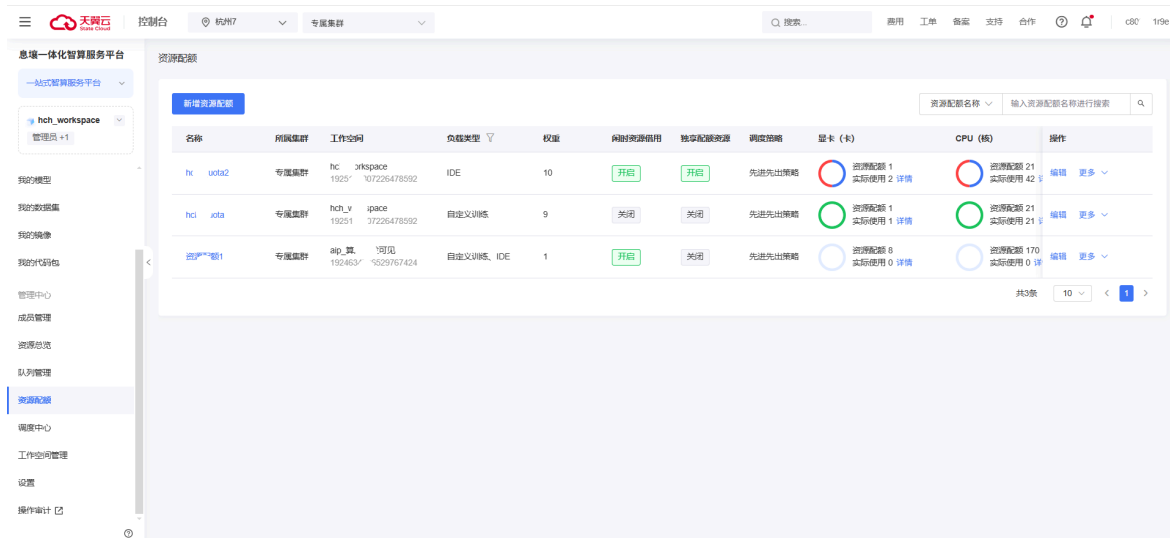
新建工作空间

确认关联工作空间后，不可修改与删除

查看资源配额

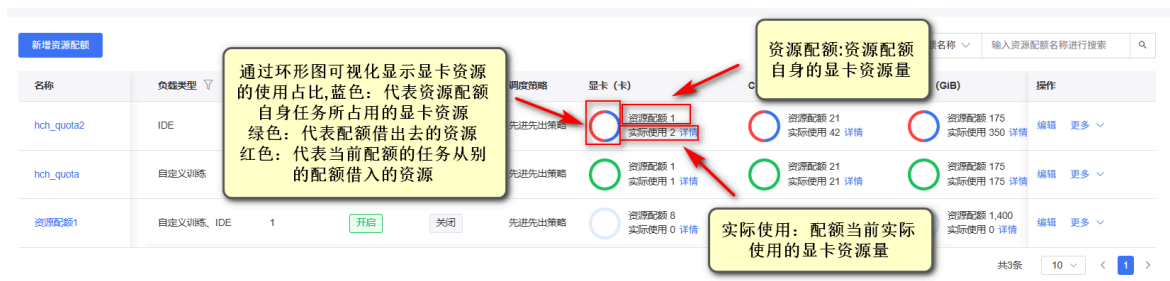
在【资源配额】页面可以查看创建的资源配额，可以看到资源配额的资源使用情况

用户指南



查看资源配额信息核心需要关注这几个功能：

- 查询功能：可以通过右上角的查询功能，按照资源配额名称、所属集群、工作空间名称三个维度进行查询
- 工作空间：查看该资源配额属于那个工作空间，资源配额只有分配给工作空间后才可以在工作空间内使用
- 列表里的显卡、CPU、MEM(内存)三个字段：这三个字段表示资源配额的资源使用情况，以显卡字段为例：



鼠标hover到“详情”按钮，可以进一步查看详情明细：

用户指南



- **【编辑】操作：**在列表右侧【操作】列，点击【编辑】按钮，会弹出资源配额编辑窗口，可修改配额名称、描述。



- **【更配】操作**

前置条件：当前资源配额的状态处于停止状态

在列表右侧【操作】列，点击【更多】会出现下拉菜单，展示了具体的操作项，选择【更配】，则进入更配页面，可修改资源配额的负载类型、实例规格、实例数、资源配额权重、闲时算力借用、独享配额资源、调度策略字段。更配操作需注意：

- 1) 负载类型修改：当前配额支持的任务类型按最新修改生效，比如之前支持大模型评估，修改后不支持了，则后续评估任务选资源配额时，则不再展示该资源配额。
- 2) 实例规格修改：若该资源配额的存量任务中存在正在运行的任务以及排队中的任务，则不能修改，反之可以修改。

3) 实例数修改：实例数调大，上限是：资源池内可支持的上限，比如资源池里还有100卡没有划分到队列，则实例数最大就是100除以规格向下取整。若用户设置的超过这个数，则给出报错提示。实例数调小，下限是：资源配额里已分配的卡数与该配额里排队中的任务中单任务所需的最大卡数的较小值。若用户设置的低于这个数，则给出报错提示。

- **【启动/停止】操作**

在列表右侧【操作】列，点击【更多】会出现下拉菜单，展示了具体的操作项，选择【启动/停止】，点击即可，需注意：

- 1) 停止资源配额：则不可以继续往该资源配额里提交任务，存量任务还是会执行完。
- 2) 启动资源配额：则可以继续向该队列里提交任务。

- **【删除】操作**

前置条件：当前资源配额的状态处于停止状态

在列表右侧【操作】列，点击【更多】会出现下拉菜单，展示了具体的操作项，选择【删除】，点击即可，需注意：

只有状态是已停止的资源配额，才能进行删除操作，且当用户执行删除操作时，系统会检查当前配额的存量任务是否都执行完毕，若存在未运行结束的任务，则不能删除，给出报错提示，可等待任务执行完毕后再删除，或者将任务迁移后再删除。若不存未运行结束的任务，则可以删除。资源配额删除后，底层的资源释放。

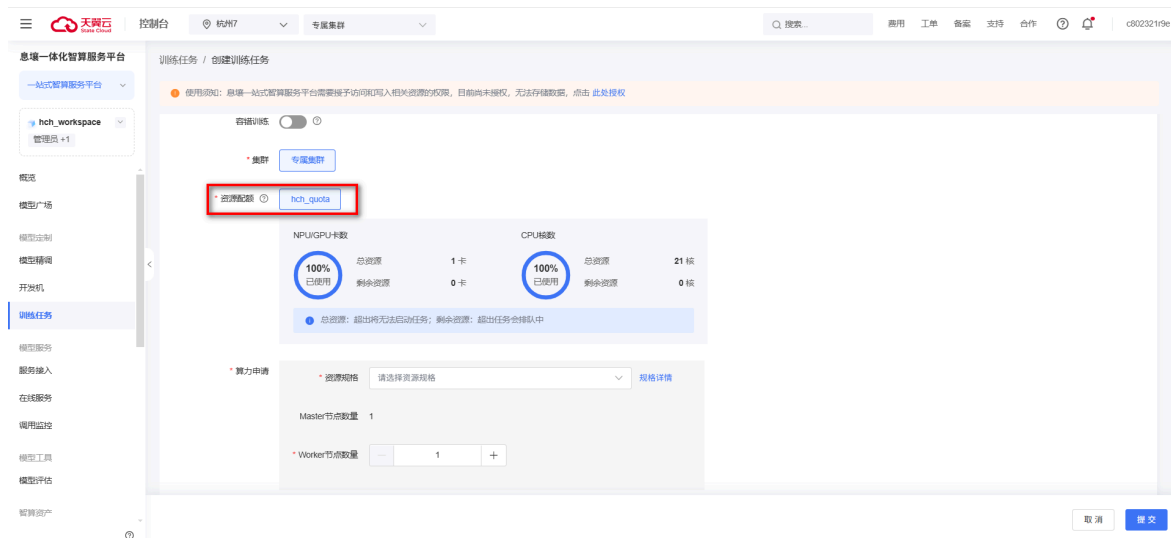
使用资源配额

- **前置条件**

1. 资源配额使用前，需要关联工作空间，关联工作空间共有两种方式，一是创建配额时关联，具体参见创建资源配额，二是在工作空间内关联，具体参见[工作空间关联资源配额](#)。
2. 已进入资源配额所关联的工作空间内
3. 资源配额所选负载类型需要包含对应任务类型, 如开发机，训练任务

以创建训练任务为例，在创建训练任务的资源配置模块，可以选择资源配额来创建训练任务，资源配额将为训练任务提供所需资源，若当前工作空间关联了多个资源配额，在此次可以自由选择需要的资源配额。因为同一专属集群下的资源配额之间可以互相借用资源(创建配额时可以设置是否允许借用别的配额资源以及是否允许别的资源配额借用自身资源)，所以，只要同一专属集群下的其他资源配额有足够资源出借，即使训练任务所需资源超出了当前资源配额的剩余资源，训练任务也是可以借用其他配额资源来进行任务训练的。

用户指南



附表：资源配额与队列的区别

资源配额和队列都是对集群资源的一种自定义分配方式，具体区别如下：

类型	定义	区别
队列	将集群资源划分给队列，提交给队列的任务将使用volcano进行资源调度；同一集群下的队列之间无法相互借用各自的闲置资源	<ul style="list-style-type: none">创建队列1.默认集群自带默认队列，无需用户创建2.专属集群资源的使用需要用户创建队列使用队列1.在默认工作空间内关联队列，关联后，在默认工作空间内作业可以使用队列资源(仅默认工作空间可使用队列)支持范围1.所有资源池都支持 注：后续产品迭代会逐步取消队列，建议使用资源配额

用户指南

类型	定义	区别
资源配额	将集群资源在逻辑层面进行了资源配额划分，集群资源划分与资源调度进行了解耦，提交给资源配额的任务将通过资源调度层进行资源调度；根据创建资源配额时的设置，同一集群下的资源配额之间可以进行闲置资源借用	<ul style="list-style-type: none">• 创建资源配额1. 专属集群资源的使用需要用户创建资源配额(默认集群无法使用资源配额，只有默认队列) <ul style="list-style-type: none">• 使用资源配额1. 在非默认工作空间内关联资源配额，关联后，在非默认工作空间内作业可以使用资源配额资源(仅非默认工作空间可使用资源配额) <ul style="list-style-type: none">• 支持范围1. 仅杭州7资源池支持

工作空间管理

创建工作空间

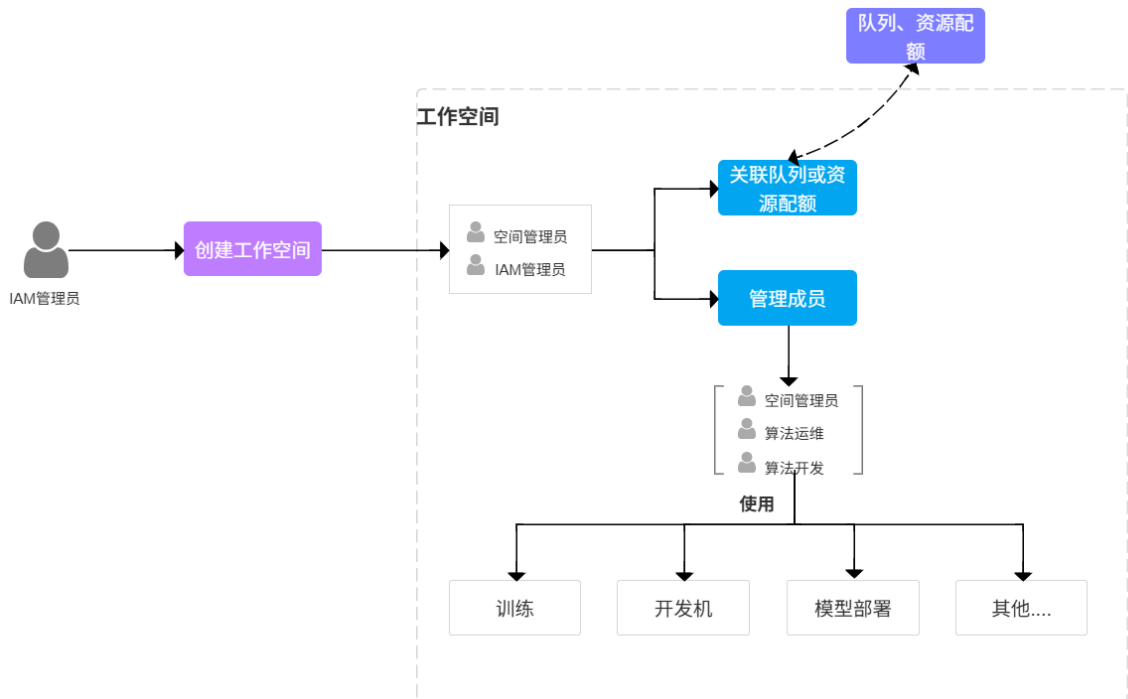
前言

工作空间是息壤一站式智算服务平台的顶层逻辑概念，为企业和团队提供统一的计算资源管理及人员权限管理能力，为AI开发者提供支持团队协作的全流程开发工具以及AI资产管理能力。息壤一站式智算服务平台具有一个默认工作空间，用户也可以根据自身需要单独创建工作空间。

工作空间与队列/资源配额以及工作空间成员进行各项任务之间的整体关系如下图：

1. 先由主账号或角色为IAM管理员的子账号先创建空间
2. 创建空间后，再关联队列或资源配额(目的是确定工作空间所使用的计算资源)
3. 再进行成员管理(将需要在工作空间内进行作业的成员加入进来并分配对应角色)
4. 最后空间成员可以在工作空间内基于队列资源进行相关作业，如使用开发机、创建训练任务等。

用户指南



工作空间创建

• 前置条件

账号为主账号或者角色为IAM管理员的子账号

• 操作步骤

1. 打开【工作空间管理】菜单，点击【创建工作空间】按钮



用户指南

2. 在【创建工作公开】弹窗按规则依次填写“工作空间名称”、“工作空间描述”，再点击“确认”按钮，工作空间即创建成功

创建工作空间

* 工作空间名称

支持字母、数字、汉字、下划线，需字母开头，同租户下不能重复

0/25

工作空间描述

请输入描述，300个字符以内

0/300

取消

确认

3. 在【工作空间管理】页面可以查看创建的工作空间

信息一体化智慧服务平台

一站式智慧服务平台

aip_0chmcgvjb...

默认 管理员 +3

管理资产

我的模型

我的数据集

我的镜像

我的代码包

管理中心

成员管理

资源总览

队列管理

工作空间管理

工作空间是信息一体化智慧服务平台的底座，为企业和团队提供统一的计算资源管理、人员权限管理能力，为AI开发者提供团队协作的全流程开发工具以及AI资产管理能力。

+ 创建工作空间

工作空间ID	任务数	我的角色	空间所有者	成员数	操作
默认 aip_0chmcgvjb...	开发机: 41 训练任务: 38	管理员 空间管理员 算法运维 算法开发	SP Y +13	38	进入空间 配置
test 1914864559591739392	开发机: 0 训练任务: 0	管理员 空间管理员	SV +12	12	进入空间 配置
work 191459576...	开发机: 0 训练任务: 1	管理员 空间管理员	SW +12	14	进入空间 配置

工作空间关联资源

工作空间的使用需要关联队列或关联资源配额等资源，关联队列或资源配额后，工作空间内的负载都可以使用队列或资源配额的资源，目前默认工作空间只能关联队列，非默认工作空间只能关联资源配额

前置条件

1. 已购买专属集群且基于专属集群创建了队列或资源配额，详见[队列管理](#)或[资源配额](#)
2. 账号角色需为租户管理员和空间管理员

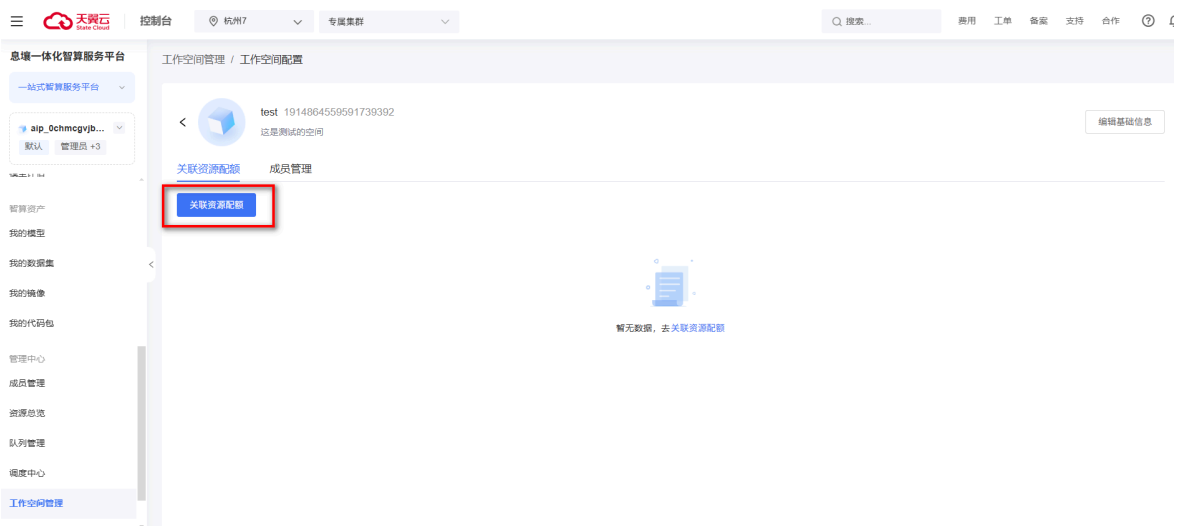
操作步骤

用户指南

1. 默认工作空间关联队列和非默认工作空间关联资源配额，操作步骤一模一样，以下以非默认工作空间关联资源配额的操作步骤进行说明，在【工作空间管理】页面，找到目标工作空间，点击“配置”->“管理资源配额”进入到【工作空间配置】页面



2. 在【工作空间配置】页面，点击“关联资源配额”按钮



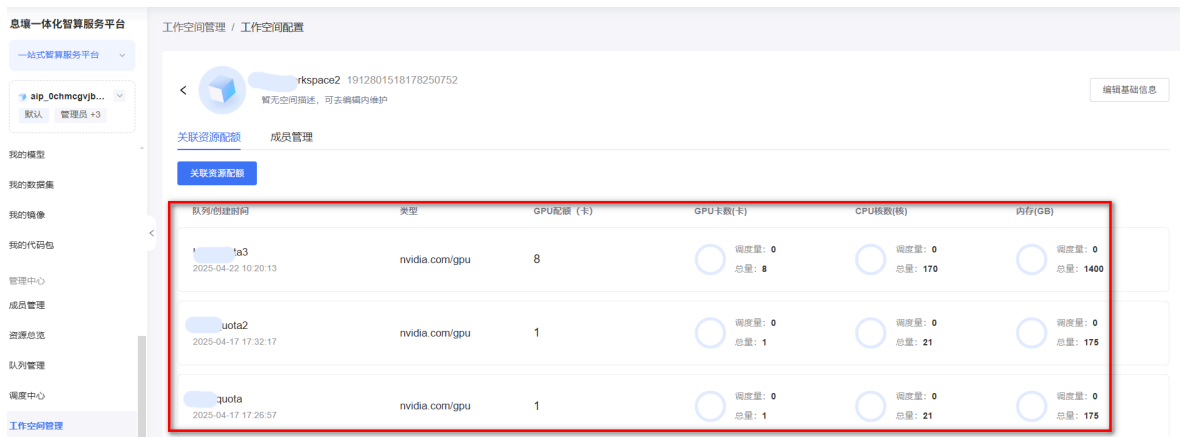
用户指南

- 在打开的面板中，选择需要关联的资源配额，一个工作空间可关联多个资源配额，选择后，点击“确认关联”即可



队列	类型	创建时间	GPU配额	GPU卡数(调度量/总量)	CPU核数(调度量/总量)	内存(调度量/总量)
张[redacted]队列	nvidia.com/gpu	2025-04-22 17:55:54	1卡	0卡 / 1卡	0核 / 21核	0GB / 175GB
q[redacted]	nvidia.com/gpu	2025-04-22 16:27:51	8卡	0卡 / 8卡	0核 / 170核	0GB / 1400GB
test	nvidia.com/gpu	2025-04-22 14:04:52	1卡	0卡 / 1卡	0核 / 21核	0GB / 175GB
[redacted] quota3	nvidia.com/gpu	2025-04-22 10:20:13	8卡	0卡 / 8卡	0核 / 170核	0GB / 1400GB
z[redacted]队列	nvidia.com/gpu	2025-04-21 15:30:58	1卡	0卡 / 1卡	0核 / 21核	0GB / 175GB
张[redacted]新[redacted]	nvidia.com/gpu	2025-04-21 15:29:44	1卡	0卡 / 1卡	0核 / 21核	0GB / 175GB
泉[redacted]队列	nvidia.com/gpu	2025-04-21 13:37:00	1卡	0卡 / 1卡	0核 / 21核	0GB / 175GB
h[redacted]2	nvidia.com/gpu	2025-04-17 17:32:17	1卡	0卡 / 1卡	0核 / 21核	0GB / 175GB
hct[redacted]	nvidia.com/gpu	2025-04-17 17:26:57	1卡	0卡 / 1卡	0核 / 21核	0GB / 175GB
q[redacted]	nvidia.com/gpu	2025-04-17 10:43:25	2卡	0卡 / 2卡	0核 / 42核	0GB / 350GB
q[redacted]配额1	nvidia.com/gpu	2025-04-17 10:42:58	2卡	0卡 / 2卡	4核 / 42核	16GB / 350GB
张[redacted]列	nvidia.com/gpu	2025-04-16 14:37:48	1卡	0卡 / 1卡	0核 / 21核	0GB / 175GB
[redacted]2	nvidia.com/gpu	2025-04-16 14:14:59	1卡	0卡 / 1卡	0核 / 21核	0GB / 175GB

- 配置好后，可以在【工作空间配置】页面查看关联的资源配额



队列/创建时间	类型	GPU配额 (卡)	GPU卡数(卡)	CPU核数(核)	内存(GB)
[redacted] quota3 2025-04-22 10:20:13	nvidia.com/gpu	8	调度量: 0 总量: 8	调度量: 0 总量: 170	调度量: 0 总量: 1400
[redacted] quota2 2025-04-17 17:32:17	nvidia.com/gpu	1	调度量: 0 总量: 1	调度量: 0 总量: 21	调度量: 0 总量: 175
[redacted] quota 2025-04-17 17:26:57	nvidia.com/gpu	1	调度量: 0 总量: 1	调度量: 0 总量: 21	调度量: 0 总量: 175

工作空间成员管理

- 前置条件

账号为主账号或者角色为IAM管理员的子账号或工作空间管理员

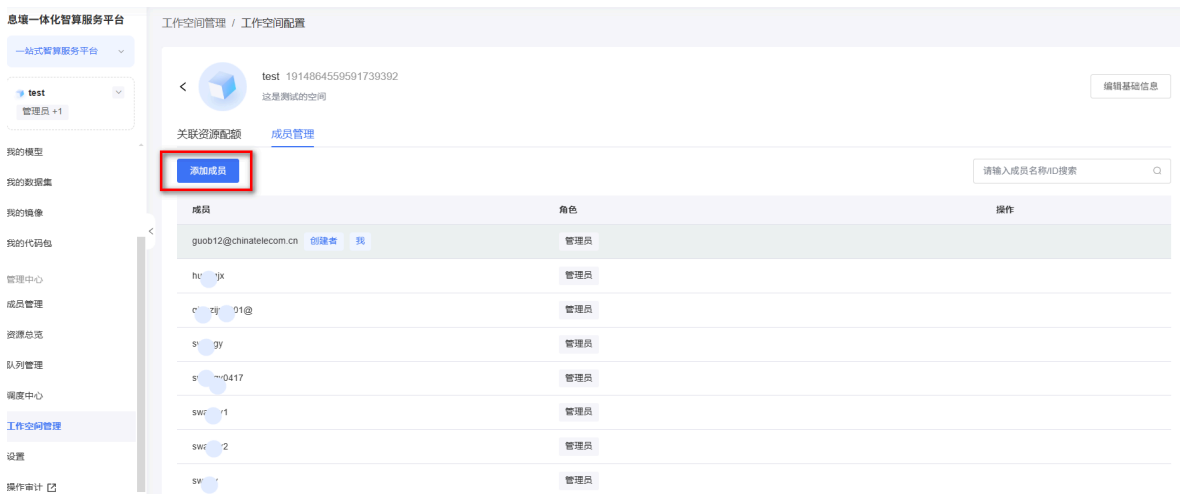
- 操作步骤

用户指南

1. 在【工作空间管理】页面，找到目标工作空间，点击“配置”->“成员管理”进入到【工作空间配置】页面

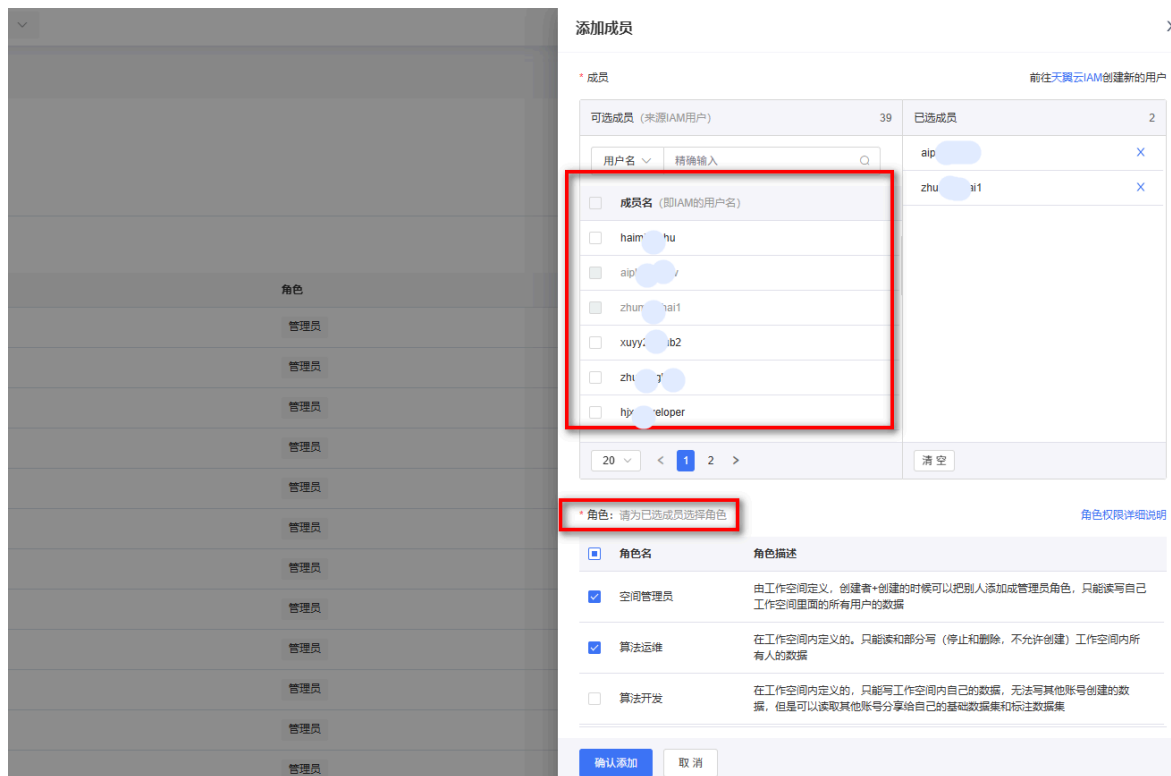


2. 在【工作空间配置】页面，点击“添加成员”按钮

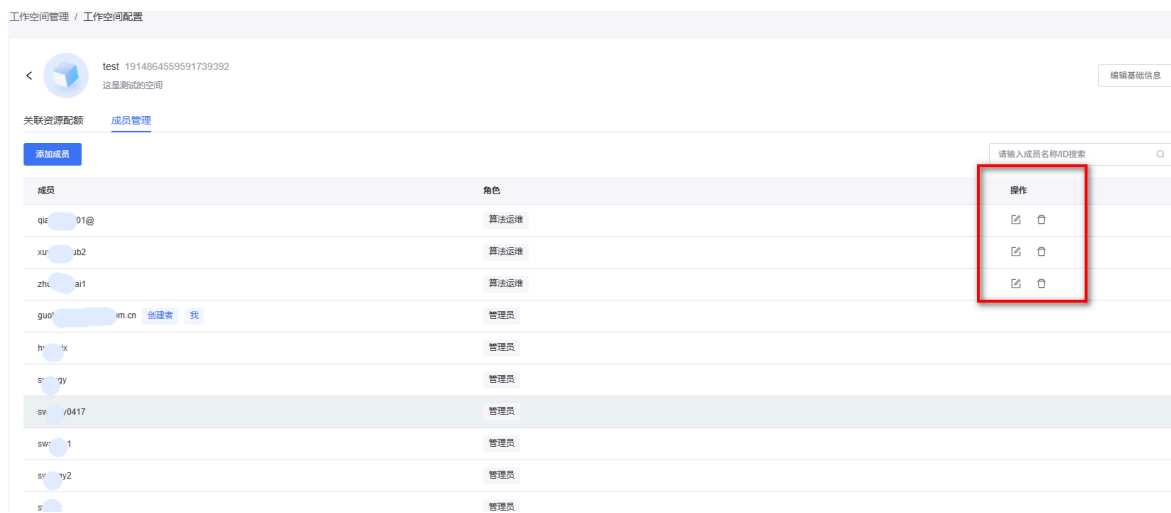


用户指南

- 在打开的面板中选择需要添加的成员及角色，然后点击“确认添加”即可将这些成员添加进工作空间，成员来源于[创建的子账号](#)，空间角色信息可参见[工作空间角色权限](#)章节。



- 在【工作空间配置】页面可以查看工作空间的成员，并且进行成员的编辑、删除等操作；注：创建工作空间时，角色为IAM管理员的账号，会自动被添加为管理员角色，此类角色，不可在工作空间内被删除，需要在IAM里面将其从“admin”用户组移除，参见[创建子账号](#)



进入工作空间

- 前置条件

账号为主账号或角色为IAM管理员的子账号或工作空间管理员和空间的成员才可进入

- 操作步骤

用户指南

1. 进入工作空间有两种方式

1) 在工作空间管理页面进入

在【工作空间管理】页面，找到目标工作空间，点击“进入空间”按钮，即可进入该工作空间



进入工作空间后，左上角的工作空间会变为当前工作空间，工作空间内的菜单也会变为自身工作空间的菜单



2) 通过切换工作空间进入

点击左上角的工作空间区域，可以切换进入其他工作空间

用户指南



进入空间之后就可以进行各项作业，如使用开发机、创建训练任务等。

用户指南

工作空间角色权限

- 管理员权限：由ctiam定义，除租户主账户外可定义多个管理员，通过给用户组授权【智算服务平台-管理员权限】，可以读写所有工作空间下所有用户所有功能的数据。

用户组

您还可以创建17个自定义用户组。

请输入用户组名进行搜索

用户组名称	用户数量	描述	创建时间	操作
组1	0	2222	2025-02-14 11:12:52	查看 授权 编辑 用户组管理
test	0	111	2024-03-04 23:29:40	查看 授权 编辑 用户组管理
admin	1	拥有所有操作权限的用户组	无	查看 授权 编辑 用户组管理

10条/页

共 3 条

1

策略名称

策略描述

授权类型

授权范围

huijuOutCommon

系统策略

全局级

查看策略

描述

--

内容

```
{
  "Version": "1.1",
  "Statement": [
    {
      "Action": [
        "huijuaip:home:all",
        "huijuaip:dataset:all",
        "huijuaip:modelManage:all",
        "huijuaip:user:all",
        "huijuaip:resource:all",
        "huijuaip:modelEval:all",
        "huijuaip:modelDeploy:all",
        "huijuaip:modelDevelop:all"
      ]
    }
  ]
}
```

确定

- 工作空间管理员：在工作空间定义，创建者+创建的时候可以把别人添加成管理员角色，可以读写自己工作空间里面的所有用户的数据
- 算法开发：在工作空间内定义的，只能读写工作空间内自己的数据，无法读写其他账号创建的数据，但是可以读取其他账号分享给自己的基础数据集和标注数据集。
- 算法运维：在工作空间内定义的。可以管理（包括停止和删除，不允许创建）工作空间内所有用户的任务类数据（包括开发机、训练任务）用于进行实例任务的运维和异常排查，但是不能管理其他用户的资产。

权限点名称	工作空间角色（IAM管理员和工作空间管理员）	工作空间角色（算法开发+自己创建的数据）	工作空间（算法运维+自己创建的数据）	工作空间角色（算法运维+他人创建数据）
工作空间用户-添加	允许	不允许	不允许	不允许
工作空间用户-列表	允许	允许	允许	允许
工作空间用户-编辑	允许	不允许	不允许	不允许
工作空间用户-删除	允许	不允许	不允许	不允许

用户指南

工作空间配置-菜单	允许	不允许	不允许	不允许
工作空间配置-按钮	允许	不允许	不允许	不允许
数据集-创建	允许	允许	允许	不允许
数据集-删除	允许	允许	允许	不允许
数据集-查看	允许	允许	允许	不允许
数据集-导入	允许	允许	允许	不允许
数据集-重命名	允许	允许	允许	不允许
数据集-普通加速	允许	允许	允许	不允许
数据集-发布	允许	允许	允许	不允许
数据集-导出	允许	允许	允许	不允许
数据集-下载	允许	允许	允许	不允许
数据集-标注	允许	允许	允许	不允许
数据集-清洗	允许	允许	允许	不允许
数据集-推送到高速缓存	允许	允许	允许	不允许
数据集-制作副本	允许	允许	允许	不允许
数据集-分享或取消分享	允许	允许	允许	不允许
基础数据集-创建	允许	允许	允许	不允许
基础数据集-查看	允许	允许	允许	不允许
基础数据集-修改可见范围	允许	允许	允许	不允许
基础数据集-重命名	允许	允许	允许	不允许
基础数据集-修改挂载路径	允许	允许	允许	不允许
基础数据集-删除	允许	允许	允许	不允许
基础数据集-查看文件	允许	允许	允许	不允许
基础数据集-加速到智算存储	允许	允许	允许	不允许
基础数据集-备份到普通存储	允许	允许	允许	不允许
我的模型-新增模型外壳	允许	允许	允许	不允许
我的模型-新增模型版本	允许	允许	允许	不允许
我的模型-模型外壳名称去重校验	允许	允许	允许	不允许

用户指南

我的模型-更新模型外壳（编辑名称和描述）	允许	允许	允许	不允许
我的模型-模型外壳列表分页查询	允许	允许	允许	不允许
我的模型-根据id查询模型外壳单条信息	允许	允许	允许	不允许
我的模型-分页查询模型版本	允许	允许	允许	不允许
我的模型-分页查询模型详情页	允许	允许	允许	不允许
我的模型-获取所有模型外壳的所有版本号(模型部署用)	允许	允许	允许	不允许
我的模型-模型文件上传	允许	允许	允许	不允许
我的模型-下载模型文件	允许	允许	允许	不允许
我的模型-导出到本地	允许	允许	允许	不允许
我的模型-导出到天翼云媒体存储	允许	允许	允许	不允许
我的模型-模型外壳删除	允许	允许	允许	不允许
我的模型-模型版本删除	允许	允许	允许	不允许
我的模型-模型文件删除	允许	允许	允许	不允许
我的模型外壳-获取口令	允许	允许	允许	不允许
我的模型-获取口令	允许	允许	允许	不允许
我的模型-根据口令获取模型名称与id	允许	允许	允许	不允许
我的模型-获取可压缩的模型外壳及版本详情	允许	允许	允许	不允许
我的镜像-自定义镜像-查看	允许	允许	允许	不允许
我的镜像-自定义镜像-制作镜像	允许	允许	允许	不允许
我的镜像-自定义镜像-取消制作	允许	允许	允许	不允许

用户指南

我的镜像-自定义镜像-分享	允许	允许	允许	不允许
我的镜像-自定义镜像-取消分享	允许	允许	允许	不允许
我的代码包-分页查询	允许	允许	允许	不允许
我的代码包-导入代码包	允许	允许	允许	无
我的代码包-重新上传	允许	允许	允许	不允许
我的代码包-删除	允许	允许	允许	不允许
我的代码包-查看	允许	允许	允许	不允许
我的代码包-开始训练	允许	允许	允许	不允许
我的代码包-在线编码	允许	允许	允许	不允许
我的代码包-下载	允许	允许	允许	不允许
训练任务-新建	允许	允许	允许	无
训练任务-复制	允许	允许	允许	不允许
训练任务-启动	允许	允许	允许	允许
训练任务-停止	允许	允许	允许	允许
训练任务-删除	允许	允许	允许	允许
训练任务-分页查询	允许	允许	允许	允许
训练任务-查看详情	允许	允许	允许	允许
开发机-新建	允许	允许	允许	无
开发机-启动	允许	允许	允许	允许
开发机-停止	允许	允许	允许	允许
开发机-删除	允许	允许	允许	允许
开发机-分页查询	允许	允许	允许	允许
开发机-查看详情	允许	允许	允许	允许
开发机-保存至代码包	允许	允许	允许	不允许
开发机-制作镜像	允许	允许	允许	不允许
开发机-开始训练	允许	允许	允许	不允许
公共服务-事件	允许	允许	允许	允许
公共服务-日志	允许	允许	允许	允许
公共服务-监控	允许	允许	允许	允许

用户指南

公共服务-保存至模型管理	允许	允许	允许	不允许
--------------	----	----	----	-----

计费类

训推智算服务平台支持哪些计费方式？

训推智算服务平台支持包周期计费模式、按需计费模式-卡时两种计费方式。

后付费的账单是月结算还是日结算？

训推智算服务平台是按照小时结算，每小时结算账单。

操作类

平台已预置的模型有哪些？

进入模型服务模块，选择在线服务，点击【预置服务】，可以看到平台预置的模型，平台预置了多款等基础大模型，包括通义千问、Llama、ChatGLM 等系列，可以直接使用。不同的基础模型的参数和能力不同，我们将持续推出不同能力方向的模型。

平台提供的开发工具有哪些？

JupyterLab和Visual Studio Code (VSCode)。

GPU模型脚本如何迁移到昇腾NPU上？

- 新建脚本train.py，写入以下原GPU脚本代码。
- 添加以下库代码。

```
import time
import torch
.....
import torch_npu
from torch_npu.npu import amp # 导入AMP模块
from torch_npu.contrib import transfer_to_npu # 使能自动迁移
```

IDE无法打开图片或预览MD文件，该怎么办？

无法在IDE打开图片或预览MD文件，这是由于浏览器设置问题，需要开启Chrome浏览器的 `unsafely-treat-insecure-origin-as-secure` 功能。

- 进入Chrome Flag管理界面配置：`chrome://flags/#unsafely-treat-insecure-origin-as-secure`。

训推智算服务平台预置的镜像有哪些？

进入智算资产模块，选择我的镜像，点击【系统内置镜像】，可以看到平台内置的镜像，包括PyTorch、TensorFlow等。

如果在自定义镜像列表看不到容器镜像服务分享过来的镜像，怎么办？

请进行以下检查：

- 容器镜像服务所选区域与平台是不是同资源池。
- 截止时间是不是大于等于当前时间，超出截止时间后共享失效。

常见问题

- 共享镜像状态是不是启用。

如何在模型开发JupyterLab和VSCode中使用自定义镜像？

需要将对应的软件安装包打包进您的自定义镜像中，具体方式见下方。

- 在docker file中具体执行命令。

#VSCode

```
curl -fsSL https://code-server.dev/install.sh | sh
code-server --install-extension ms-python.python
```

#Jupyterlab

```
pip install jupyterlab
```

- 将打包好的镜像在本地起起来，然后运行如下命令安装软件，安装完成后，执行 `docker commit {容器名称}`，打包成新镜像后，即可上传。

#VSCode

```
curl -fsSL https://code-server.dev/install.sh | sh
code-server --install-extension ms-python.python
```

#Jupyterlab

```
pip install jupyterlab
```

我想基于自己的模型进行二次训练微调怎么做？

可以先在模型管理中导入自己的模型，在JupyterLab和VSCode创建训练任务，在挂载模型的选项中选择【模型管理】，选择已导入需要二次训练微调的模型，即可挂载自己的模型进行训练。

训推智算服务平台是否支持IB和NVlink？

当前昇腾集群暂不支持。

如何给予账号配置资源使用的限额？

主账号管理员进入运营后台，在配置设置模块，可支持设置单用户最大同时使用的GPU/CPU数量以及并行文件存储初始分配额度。

创建基础数据集或在开发机及训练任务中增加存储配置时，为什么不填ZOS/HPFS路径会报错？

请检查您的账号类型，这可能是平台权限策略所导致的。

- 若您是管理员用户，您可以挂载到整桶或者文件系统粒度，路径对您来说是选填项；
- 若您是普通子账号，您必须填写具体的已被授权的路径，路径对您来说是必填项，若不填写则无法提交。

点击Tensorboard没反应？

- 1、需要检查是否被浏览器拦截；
- 2、输出目录有限制：必须将数据写入指定的目录；
- 3、最多支持开启10个（同一个任务点击2次=开2个）。

查看监控时，无法选择我需要的时间？

注意时间的选择由时间精度（分、时、日）与时间范围共同控制：

- 选择时间精度为分，图像步长为1分钟，时间跨度至多1天；
- 选择时间精度为时，图像步长为1小时，时间跨度至多7天；
- 选择时间精度为日，图像步长为1天，时间跨度至多31天；

常见问题

当您需选择更大的时间跨度时，可能需要同时调整相应的时间精度。



产品使用方面的问题如何反馈？

您可通过[天翼云官网工单](#)或者客服电话【400-810-9889】进行反馈。