云骁智算公有云

用户手册

天翼云科技有限公司

2024-10-30

版权声明

请在阅读或使用本文档之前仔细阅读并理解本声明中的内容。您的阅读或使用行为将被视为对本声明的认可。

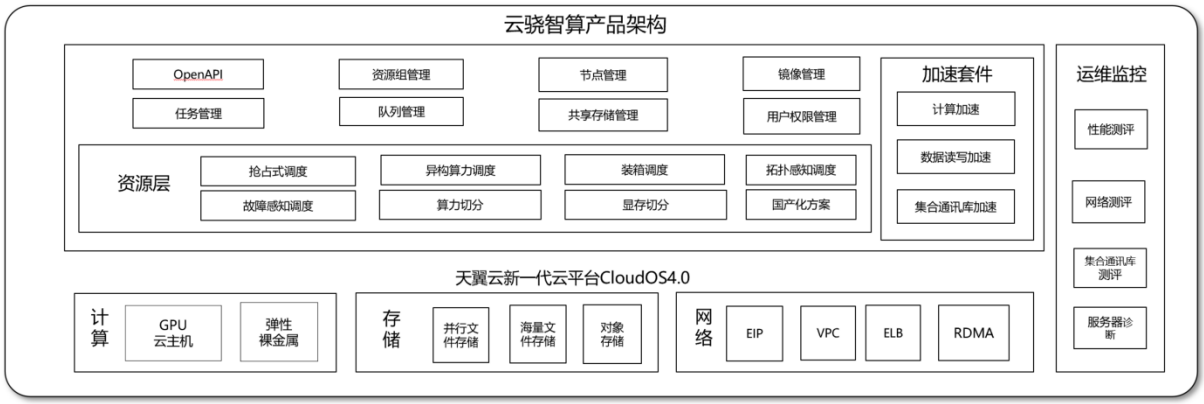
- 您应当通过天翼云网站或天翼云提供的其他授权通道获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为天翼云的保密信息，您应当严格遵守保密义务；未经天翼云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。
-
- 由于产品版本升级、调整或其他原因，本文档内容有可能变更。天翼云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在天翼云授权通道中不定期发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过天翼云授权渠道获取该文档的最新版本。
- 您购买的产品、服务或特性等应受天翼云商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，天翼云对本文档内容不做任何明示或暗示的声明或保证。
- 本文档仅作为用户使用天翼云产品及服务的参考性指引，天翼云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。
- 如若发现本文档存在任何错误，请与天翼云取得直接联系。

目录

1.产品介绍

1.1 产品定义

云骁智算是提供高性能计算、存储、网络服务的智能计算加速平台，可提供异构算力的管理与调度，计算与存储间的高效互联，跨域监控和故障感知，一键自助诊断及智能加速套件等能力，通过云骁智算平台可大幅提升数据加载、训练和推理效率。



云骁智算平台底层主要由高性能计算、存储和网络组成：

- 计算侧支持多种规格的高性能裸金属，实现灵活、稳定、易用的高性能计算。
- 存储侧支持高性能并行文件存储搭配 RDMA 无损网络，存储用户读写数据时延低至亚毫秒。
- 网络侧支持 TCP/IP 和 RDMA 等多种通信协议，支持单服务器上连多个 leaf 交

换机的组网方式，出现连接故障可自动切换。单机最大带宽可达 3.2T，实现超大规模、高效并行通信。

云骁智算平台包括资源管理、系统运维监控和加速套件等多个部分。

- 资源管理部分，云骁标准资源组提供基于 GPU 物理机和 GPU 云主机的集群化开通与管理，云骁扩展资源组在标准资源组基础上提供全托管和高可用控制面板的标准 Kubernetes 集群服务，支持以云骁计算节点作为 Kubernetes 集群的工作节点。支持一键提交训练任务、日志查看、支持主流训练框架（如：PyTorch 等）。
- 系统运维监控，提供从服务器检测、RDMA 性能检测到集合通讯库性能检测的全方位一键式环境健康检测，以及多维度资源使用情况的实时监控。
- 加速套件，支持数据及通信层面的加速能力。例如，支持高性能 Checkpoint 框架 CTFlashCkpt，将训练阻塞时间降低到最小；支持高性能通讯库 CTCCL，基于天翼云网络进行深度的定制优化。

1.2 产品优势

1.安全稳定的算力底座：

a. 高性价比国产化算力支持。

b. 大规模分布式训练支持。

c. 万卡规模集群管理与实践。

2.多用户场景支持：

a. 使用标准资源组，用户可直接登录节点操作。

b. 扩展资源组，为用户预装 k8s 集群及相应控制器，用户直接进行自定义任务创建与管理。

3.全流程监控与故障感知：

a. 训前环境健康一键检测。

b. 训中多维度指标实时监控。

c. 多场景故障感知与断点续训。

4.智算加速套件：

a. 高性能集合通信库提高拥塞条件下的通信性能与故障感知。

b. 高性能 Checkpoin 框架，实现接近于 0 的模型状态保存时间开销。

1.3 功能特性

1. 计算资源管理： 支持创建标准资源组、扩展资源组等多种算力集群模式；支持英伟达、昇腾等多种智能芯片；支持 GPU 云主机、GPU 裸金属等灵活算力形态；支持包年包月、按量计费等多种灵活计费形式。
2. 配套高性能网络：支持高性能 RDMA 网络，可提供最高单机 3.2T 带宽，实现超大规模、高效并行通信。
3. 高性能存储便捷接入：支持自动连接并便捷使用包括对象存储、高性能并行文件存储在内的多种存储类型，提供百万级 IOPS、亚毫秒级时延；支持丰富的大容量非结构化数据保存和分析场景。
4. 自定义任务管理: 支持用户通过队列对算力额度进行细粒度划分；支持一键配置和执行自定义任务，并查看任务运行记录。
5. 高效调度：支持节点创建时根据底层网络拓扑，进行网络拓扑亲和性开通；支持万卡规模的异构算力调度能力；支持 binpack、gang 调度等多种调度策略。
6. AI 监控与运维：支持智算场景下的多维度监控指标展示；支持一键诊断功能对节点软硬件配置、多节点一致性配置、RDMA 网络性能、集合通讯库性能等进行自助诊断。

1.4 应用场景

1. 大模型：支持万张 GPU 规模的资源弹性，支持 3D 并行分布式训练、数据加速等算力调度赋能层能力，大大提升 AI 任务效率降低成本。

2. 政务场景：通过增量预训练和模型微调训练政务行业大模型，实现政策咨询、公文助手、智能导办、坐席辅助等功能，缩短群众办事时长。
3. 科研教育：人工智能驱动的科学研究（AI for Science, AI4S）融合科学原理和大数据，打造新一代科学技术服务平台，实现数据与算力、算法融合应用。将基于人工智能技术算法、大数据对科学计算与工业范式进行创新。

1.5 术语解释

1. 地域（Region）：是指物理数据中心所在的不同地理地域，不同地域之间内网完全隔离，保证不同地域间最大程度的稳定性和容错性。
2. 可用区 AZ（Availability Zone）：是指在同一地域（Region）内，电力和网络互相独立的物理区域。用户提交的训练任务、在线服务、计算节点以及存放数据的云盘和对象存储均在该可用区中。控制台 Header 上显示的是可用区。
3. 资源组：一个资源组是一组不同计算节点的集合，资源组内可以有不同规格的节点。用户可以根据自己的需求对资源组进行扩容、缩容。
4. 队列：队列是一批用于特定计算任务的固定配额的资源，用户使用队列中的资源处理特定工作负载。一个队列中的节点规格是一致的。
5. 节点：节点是集群的组成单元，每个节点对应一台物理机，按包年包月售卖。

1.6 规格

分类	典型配置	显卡型号
物理机	physical.h7ns.4xlarge3	8*A800
物理机	physical.h8ns.6xlarge8	8*H800
物理机	physical.h8ne.5xlarge9	8*H800
物理机	physical.h7es.4xlarge3	8*A100
物理机	physical.h6ns.2xlarge11	8*L40S
物理机	physical.acas910b.2xlarge1	8*910B
物理机	physical.lcas910b.2xlarge1	8*910B
云主机	pi7.4xlarge.4	1×A10
云主机	pi7.8xlarge.4	2×A10
云主机	pi7.16xlarge.4	4×A10
云主机	p8a.6xlarge.4	1×A100
云主机	p8a.12xlarge.4	2×A100
云主机	p8a.24xlarge.4	4×A100

1.7 使用限制

1. 不支持跨 AZ 创建资源组。

2. IB 网络只支持租户级隔离，不支持子账号级隔离。

3. 请勿通过其他产品控制台删除云骁智算平台为用户创建的资源，如资源组管理节点云主机，elb 和 vpce。

1.8 与其他服务的关系

物理机

云骁在资源组开通时，可以选择 xPU 物理机作为资源组的机器类型。

云主机

云骁在资源组开通时，可以选择 xPU 云主机作为资源组的机器类型。

VPC

云骁创建资源组时可以选择已有虚拟私有云，也可以去虚拟私有云订购页创建虚拟私有云。

ZOS

云骁可以使用对象存储服务存储用户的数据和模型。

HPFS

云骁可以使用 HPFS 并行文件服务存储用户的数据和模型。

2. 计费说明

2.1.计费项

云骁智算平台使用方式包括标准资源组和扩展资源组两种，计费项包括资源组费用和

计算节点费用，资源组计费项具体如下表 1、2，节点见表 3。

资源组包括标准资源组和扩展资源组两种，具体如下：

表 1 标准资源组

计费项	计费项说明	计费模式	计费公式
资源组管理费	当前免费	-	-
卡管理费	当前免费	-	-

表 2 扩展资源组

计费项	计费项说明	计费模式	计费公式
资源组管理费	当前免费	-	-
卡管理费	当前免费	-	-
资源组管控面资源费用	使用管控面部署 K8S 集群对应资源的用量	包年包月	规格单价*管控节点数量*购买时长
		按量计费	规格单价*管控节点数量*使用时长

表 3 计算节点费用

计费项	计费项说明	计费模式	计费公式
计算节点	使用计算资源的用量	包年包月	规格单价*管控节点数量*购买时长

		按量计费	规格单价*管控节点数量*使用时长

2.2.计费模式

云骁智算计费模式概述

云骁智算节点提供包年包月和按量计费两种计费模式，可满足不同场景的业务需求，在选择计费模式时，应结合业务需求和实际情况来做出合适的选择。

- 包年包月：一种预付费模式，先付费再使用，适用于多种场景，尤其是需要稳定资源并长期使用的情况，购买周期越长，享受的优惠折扣越大。
- 按量付费： 一种后付费模式，先使用再付费，根据使用时长计费，按小时出账，计费颗粒度可精确到秒级。适用于需要灵活调整资源、业务不稳定或资金有限的场景。

表 1 列出了两种计费模式的区别

计费模式	包周期	按量付费
付费方式	预付费 按照购买周期结算	后付费 按照实际使用时长计费
计费周期	按照购买周期计费	秒级计费，按小时出账
适用计费项	扩展资源组、节点资源	扩展资源组、节点资源
适用资源池	一类节点	一类节点

适用功能模块	资源组管理、节点管理	资源组管理、节点管理
变更规格	暂不支持	暂不支持
适用场景	可预估资源量，并长期稳定使用的场景，价格比按需计费更优惠，对于长期使用者，推进该方式	资源使用量存在波动，短期使用，且需要随时开通，随时删除的场景

2.3.续费

包年/包月的扩展资源组或节点到期后会影响云骁智算业务正常使用。如果用户想继续使用，需要在指定的时间内为相关实例进行续费，否则实例资源会自动释放，数据丢失且不可恢复。

续费操作仅针对包年包月的实例资源，按量计费模式的资源不需要计费，只需要保证账户余额充足。

包年包月的实例资源续费相关功能如表 1 所示：

表 1 续费相关的功能

功能	说明
手动续费	订购扩展资源组或节点等实例资源时选择手动续费，从购买到被自动删除之前，用户可以随时在云骁智算控制台为实例续费，延长实例到期时间，确保业务正常运行。
自动续费	订购扩展资源组或节点等实例资源时选择自动续费，实例会在每次到期前自动续费，避免因忘记续费而导致资源被自动删除。

限制

- 未完成订单中的资源不允许续订，如开通中的资源、规格变更中的资源、退订中的资源。
- 已退订或释放的资源不可续费。
- 若资源到期后续费，续费周期自资源续订解冻开始，计算新的服务有效期，按照新的服务有效期计算费用。例如，客户资源 2020 年 9 月 30 号到期，10 月 11 号续订 1 个月，那么资源新的服务开始时间为 10 月 11 号，到期时间为 11 月 10 号。相关费用自 10 月 11 号开始计算。

2.4.退订

客户（天翼云用户）可根据需要，在符合天翼云退订规则的前提下，灵活退订资源。

3. 快速入门

此章节主要目标为帮助用户了解云骁智算平台的基本使用流程，帮助用户快速使用云骁智算服务。本文主要基于管理控制台从 2 种类型账号 （主账号、子账号）分别介绍端到端的使用方式。

3.1.准备工作

在使用云骁智算平台之前需要完成天翼云账号的注册、实名认证以及开通相关服务。云骁智算平台使用时分为主账号和子账号。

主账号

- 主账号可创建管理用户组和用户将创建的子用户划分到用户组。
- 主账号可以创建和管理 IaaS 资源，包括创建管理 GPU 弹性裸金属节点，创建管理 VPC 和 vpce，创建管理 HPFS 文件系统，对象管理存储 Bucket。
- 主账号可创建和管理资源组，对资源组进行扩容和缩容（增加计算节点和解绑计算节点）

子账号

- 子账号可创建和管理训练镜像，创建和管理任务，查看训练任务监控指标。

3.1.1 注册账号

在创建和使用天翼云骁智算之前，用户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，如果用户拥有天翼云的账号，可登录后直接使用天翼云骁智算。

1. 打开天翼云门户网站，点击【免费注册】；

2. 在注册页面，请填写“邮箱地址”、“登录密码”、“手机号码”，

并点击“同意协议并提交” 按钮，如 1 分钟内手机未收到验证码，请再次点击“免费获取短信验证码”按钮；

3. 注册成功后，可到邮箱激活用户的账号，即可体验天翼云。

4. 如需实名认证，请参考会员服务-实名认证：

   a. 个人实名认证

   b. 企业实名认证

   c. 个人实名认证变更

   d. 企业实名认证变更

3.1.2 创建子账号

1. 登录云骁智算控制台，点击左侧【用户管理】菜单栏，进入【组织管理】下的子账号管理，点击”创建用户子账号"。

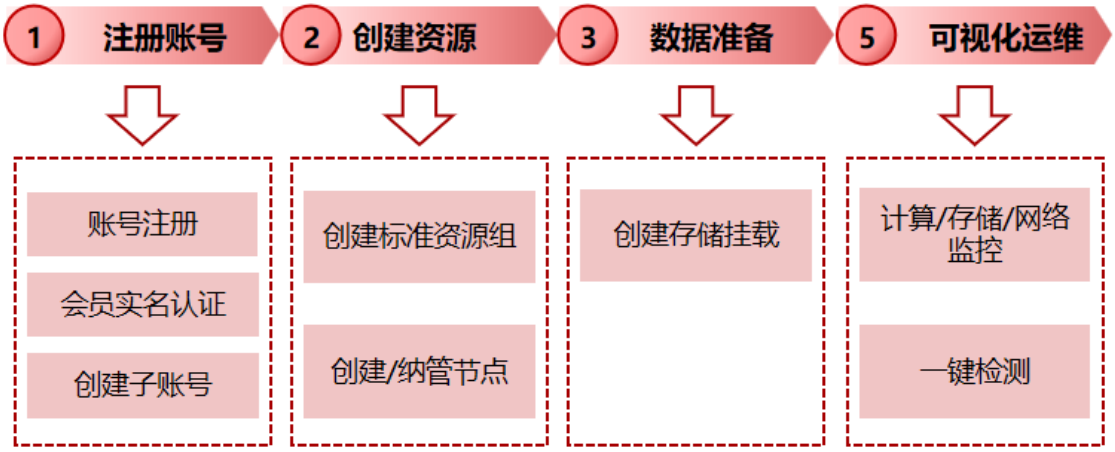
2. 输入子账号信息：子账号名，员工名称、密码、邮箱、手机号等信息，输入后，点击“确定”完成创建。

3.2.主账号使用流程

主账号使用主要聚焦运维管理和资源全流程创建，根据使用资源不同，分为两类使用

方式：标准资源组、扩展资源组，资源组差异详情请参考：资源组

标准资源组使用流程



扩展资源组使用流程

使用流程说明

流程	子任务	说明	详细指导
注册账号	账号注册	首次登录云骁智算平台需要先完成主账号注册	注册账号
	会员实名认证（可选）	账号注册完成后，可选择进行个人/企业实名认证	会员服务-实名认证
	创建子账号	子账号无法进行运维管	创建子账号

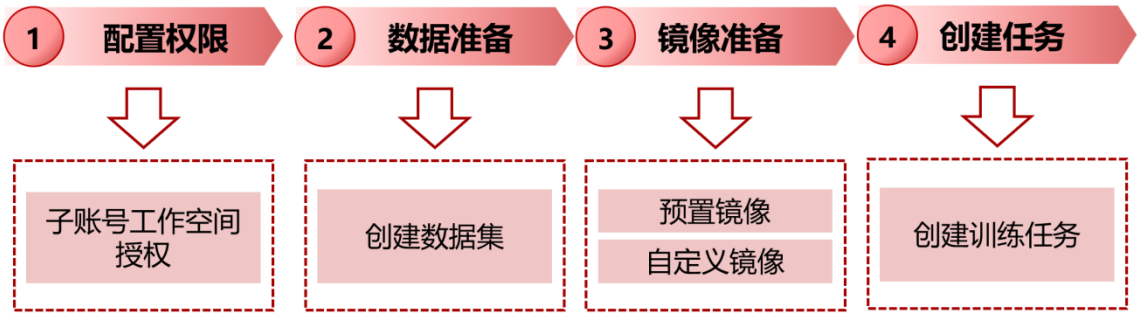
		理，仅可创建和管理训练镜像，创建和管理任务，查看训练任务监控指标。	
创建资源组	创建标准资源组	标准资源组提供基于 GPU 物理机和 GPU 云主机的集群化开通与管理	创建标准资源组
	创建扩展资源组	标准资源组提供全托管和高可用控制面板的标准 Kubernetes 集群服务	创建扩展资源组
	创建/纳管节点	资源组创建完成后需要创建/纳管节点用于承载任务所需算力运行。	创建/纳管节点
	创建队列	队列是资源配额、以及任务运行的隔离单元，在运行训练或推理任务时，通过将任务绑定到队列进行资源的排队和使用申请	创建队列
	创建工作空间	工作空间可对等于项目，不同项目可进行资源隔离。在创建工作空	创建工作空间

		间时需要添加工作空间成员和关联队列资源。 同项目成员（开发人员）可以分享 AI 资产（数据集、镜像、训练任务），进行协作	
数据准备	创建存储挂载	通过存储挂载，可支持用户将 ZOS 或 HPFS 实例批量挂载到相应的节点上，并且管理挂载目录。	创建存储挂载
	创建 K8S 共享存储	可对训练中用到海量数据的进行准备与管理，用户实现动态弹性调度，支持多种数据来源，支持开启数据加速访问。	创建 K8S 共享存储
创建任务	准备数据集	数据集支持多种数据来源，目前支持 ZOS 对象存储和 HPFS 并行文件系统，支持开启数据加速访问。 前置条件：已创建工作空间并进入工作空间内。	准备数据集

	准备模型镜像	镜像仓库中提供预置镜像和自定义镜像两种能力： ● 预置镜像即平台预先设置的完整镜像 ● 自定义镜像即可上传本地自有镜像 前置条件：已创建工作空间并进入工作空间内。	准备模型镜像
	创建训练任务	支持创建自定义创建训练任务。 前置条件：已创建工作空间并进入工作空间内。	创建训练任务
可视化运维	计算/存储、网络监控	云骁智算平台为用户提供资源监控（资源组监控、节点监控）、HPFS 监控、RoCE 监控、任务监控，多种维度查看监控指标的变化情况。	计算、存储、网络监控
	一键检测	一键诊断功能能够帮助云骁资源组管理的节点、网络等主要资源进	一键检测

		行有效的检测和运维。	

3.3.子账号使用流程



使用流程说明

流程	子任务	说明	详细指导
配置权限	子账号工作空间授权	子账号需要完成空间授权才能进入工作空间进行操作。	子账号工作空间授权
数据准备	创建数据集	数据集支持多种数据来源，目前支持 ZOS 对象存储和 HPFS 并行文件系统，支持开启数据加速访问。 前置条件：子账号已被授予工作空间权限并进入工作空间内。	创建数据集
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。 前置条件：子账号已被授予	预置镜像

		工作空间权限并进入工作空间内。	
	自定义镜像	自定义镜像即可上传本地自有镜像，上传完成后可在创建自定义训练任务时选择并使用。 前置条件：子账号已被授予工作空间权限并进入工作空间内。	自定义镜像
创建任务	创建训练任务	支持创建自定义创建训练任务。 前置条件：子账号已被授予工作空间权限并进入工作空间内。	创建训练任务

4. 用户指南

4.1.权限管理

云骁智算平台和天翼云云管系统采用统一的单点登录，并通过统一的 AK/SK 秘钥鉴权访问 OpenAPI。云骁智算平台的权限管理主要是通过天翼云统一身份认证实现，以不同用户身份登录云骁，具有不同的菜单权限。

用户在天翼云门户注册的账号即是主账号，一个主账号代表一个租户，也称为租户管理员。云平台支持多租户访问，租户可以创建子用户。云骁平台结合智算业务使用场景及安全考虑，在系统侧对主账号及子账号的权限做了默认定义。云骁公有云版本

中，主账号可以创建和管理算力、网络、存储资源，可以创建工作空间，并将子账号加为工作空间的成员。子账号没有直接创建和访问资源的能力，子账号可以在工作空间中使用分配的资源，创建数据集、创建任务、管理任务相关的容器镜像等。

系统默认主账号的权限如下：

- 可以创建管理用户组和用户；
- 可以创建和管理工作空间，包括将队列添加到工作空间下作为工作空间的可用计算资源，将其他子用户加入到工作空间中作为工作空间的成员等；
- 可以创建和管理计算资源，包括创建和管理资源组，对资源组进行扩容和缩容（增加计算节点和释放计算节点）；创建和管理 GPU 计算节点，创建和管理 VPC 及 VPCE；创建和管理队列；创建和管理自定义脚本；
- 可以创建存储挂载及 k8s 共享存储，包括创建和挂载 HPFS 并行文件系统、ZOS 对象存储等数据源；
- 可以创建和管理一键检测任务；
- 可以查看资源的监控指标等。

系统默认子账号的权限如下：

子账号被添加为工作空间成员后，

- 可以在工作空间下创建和管理数据集；
- 可以在工作空间下创建和管理容器镜像；
- 可以在工作空间下创建和管理训练任务、查看任务监控指标等。

4.2.资源

智算业务中训练或推理的高效运行以高性能的计算、存储、网络资源的有效组织为基础。云骁智算平台资源管理与调度功能面向多租户，提供资源的集群化开通与管理。关键能力包括：

- 算力节点间 RDMA 无损高速网络连接
- 高性能自研集合通信库 CTCCL
- 高性能自研 Checkpoint 框架 CTFlashCkpt

4.2.1 资源组

资源组是指运行所需要的资源组合。云骁智算提供两种资源组类型：云骁标准资源组和云骁扩展资源组。云骁标准资源组提供基于 GPU 物理机和 GPU 云主机（部分资源池支持）的集群化开通与管理，云骁扩展资源组在标准资源组基础上提供全托管和高可用控制面板的标准 Kubernetes 集群服务，支持以云骁计算节点作为 Kubernetes 集群的工作节点。用户可在云骁智算产品控制台便捷地完成购买、使用的全流程，如果用户已购买单独的裸金属资源，也可在资源组创建时选择已有节点将裸金属添加至云骁智算资源组，用于后续的训练任务使用。

4.2.1.1 新建资源组

使用前提

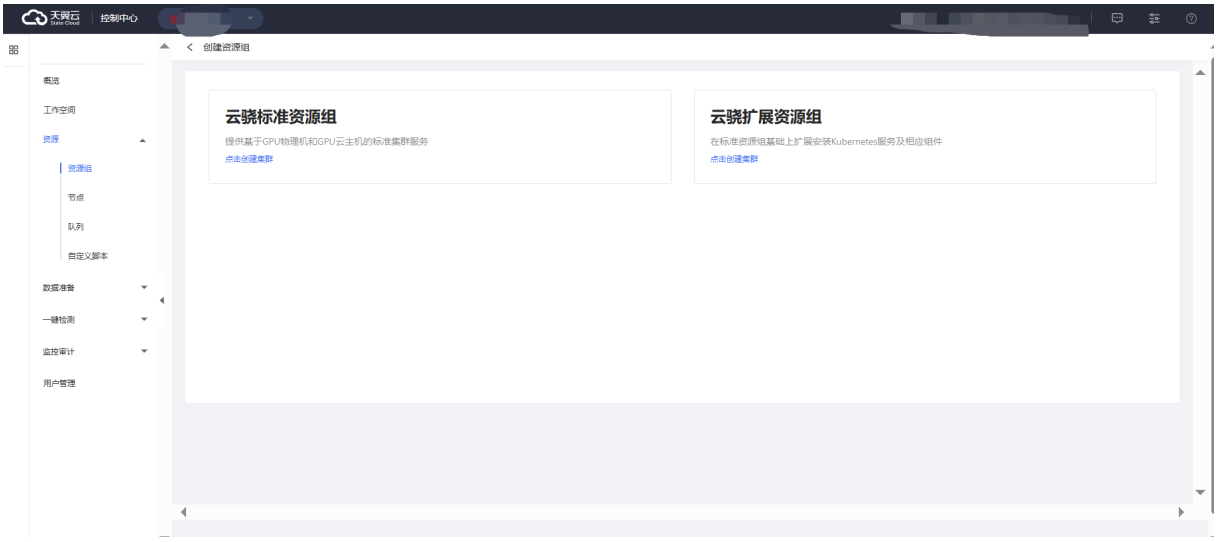
当前用户是主账号。

操作步骤

1. 登录云骁智算，单击左侧导航栏中的“资源组“，进入资源组列表页。

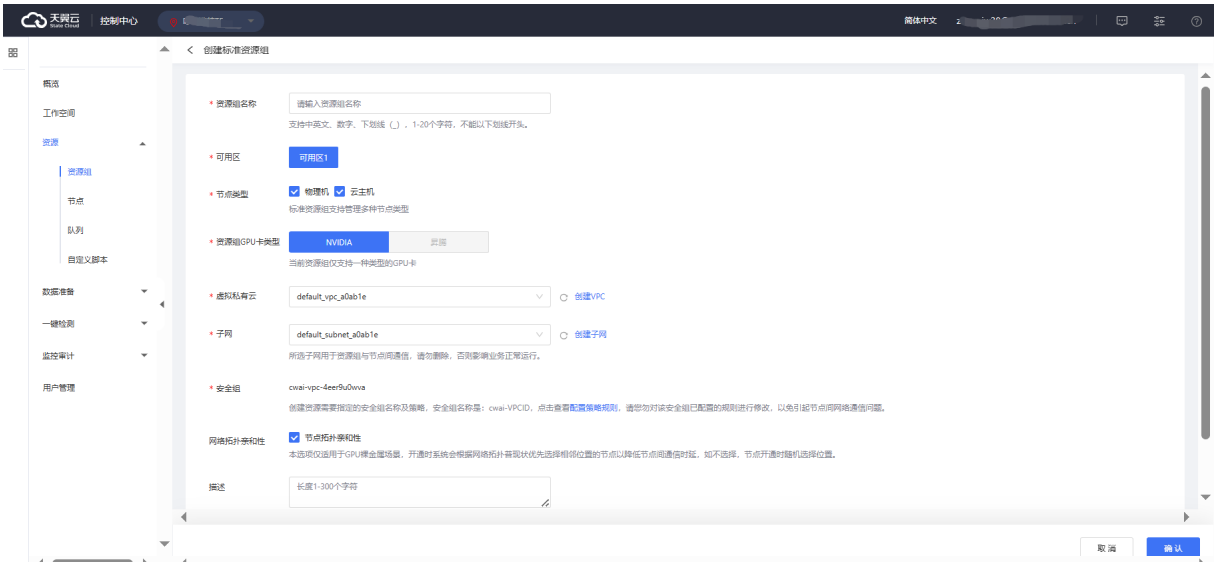
2. 单击列表页左上方的上方的“创建资源组”，进入创建页面。

3. 在创建页面进行新建资源组流程，根据自己的需求选择创建”云骁标准资源组“或者”云骁扩展资源组“。



4. 云骁标准资源组创建需要输入资源组信息：输入资源组名称、可用区、网络等基本信息。

5. 标准资源组配额默认 10 个，可通过工单升级，单资源池最多升级为 50 个。



字段说明：

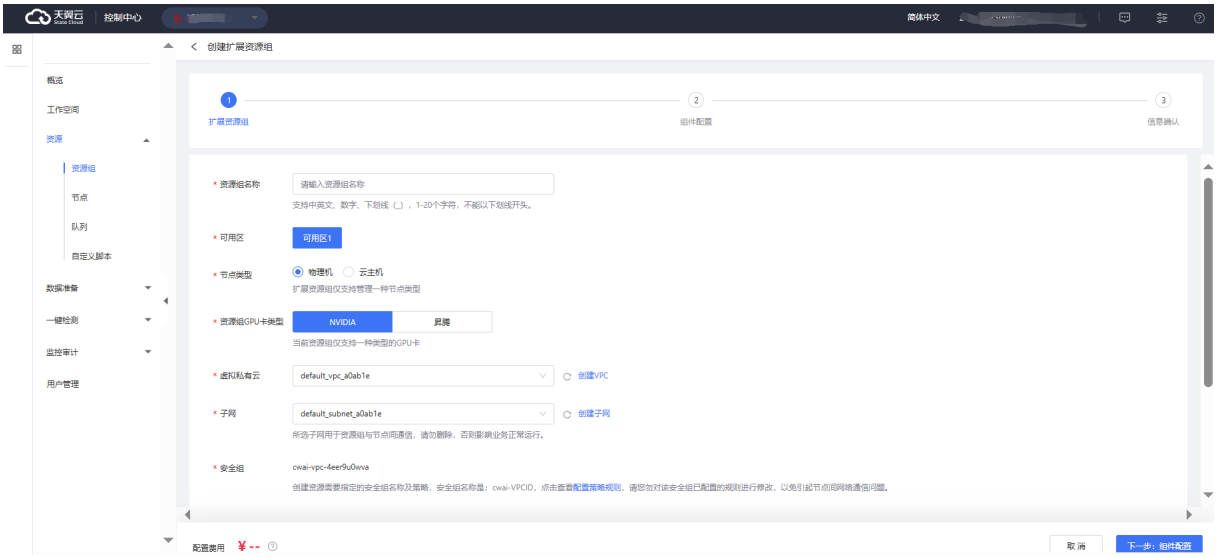
资源组信息

字段名称	类型	是否必填	长度	说明
资源组名称	输入框	是	20 字符	支持中英文、数字、下划线（_），1-20 个字符，不能以下划线开头。资源组名称不能重复
可用区	单选	是		默认第一个可用区，根据各资源池可用区情况显示
节点类型	多选	是		包括物理机和云主机两类，默认全部选中，至少选择一种
资源组 GPU 卡类型	单选	是		包括 NVIDIA 和昇腾两个类型，默认选中 NVIDIA
虚拟私有云	下拉单选	是		点击可刷新 VPC 列表，点击“创建 VPC”新打开页面跳转至 VPC 创建页面
子网	下拉单选	是		筛选子网下的 普通子网类型，点击选择 VPC 子网，点击“创建子网”新打开页面跳转至 VPC 创建页面
安全组	显示	是		默认查询是否有对应的安全组，如有则展示，如无则需要点击自动创建按钮进行创建，管理节点的安全组名称是：cwai-<VPCID>。可点击自动创建按钮
网络拓扑亲和性	勾选项	否		该功能仅适用于 GPU 裸金属场景，开通时系统会根据网络拓扑现状优先选择同一交换机下的节点以降低节点间通信时延。 该功能目前处于试用阶段。
描述	输入框	否	1-300	
协议	链接	是		点解链接至协议页面，勾选之后可点击确认按钮

6. 云骁扩展资源组创建的具体步骤如下：

扩展资源组配额默认 10 个，可通过工单升级，单资源池最多升级为 50 个。

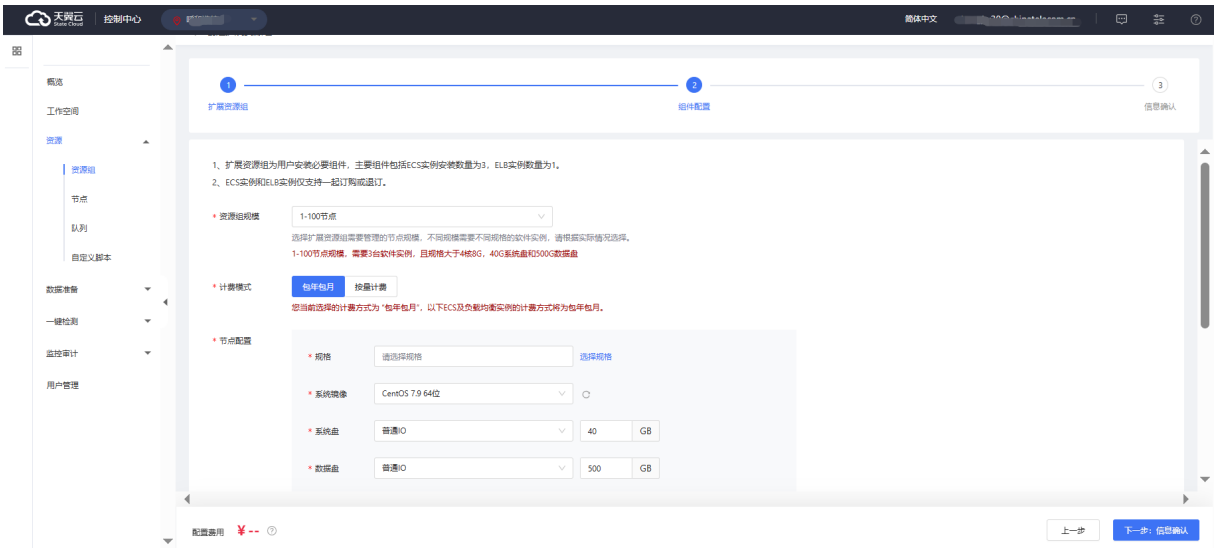
1）输入资源组名称、可用区、网络等基本信息。



字段名称	类型	是否必填	长度	说明
资源组名称	输入框	是	20 字符	支持中英文、数字、下划线（_），1-20 个字符，不能以下划线开头。资源组名称不能重复
可用区	单选	是		默认第一个可用区，根据各资源池的可用区显示
节点类型	单选	是		包括物理机和云主机两个选项，只能选择一个，默认选中物理机
资源组 GPU 卡类型	单选	是		包括 NVIDIA 和昇腾两个类型，默认选中 NVIDIA
虚拟私有云	下拉单选	是		点击可刷新 VPC 列表，点击“创建 VPC”新打开页面跳转至 VPC 创建页面
子网	下拉单选	是		显示普通类型的子网信息，点击创建子网，可跳转至创建子网页面
安全组	显示			默认查询是否有对应的安全组，如有则展示，如无则需要点击自动创建按钮进行创建，管理节点的安全组名称是：

				cwai-<VPCID>。可点击自动创建按钮
网络拓扑亲和性	勾选项	否		该功能目前处于试用阶段，仅适用于 GPU 裸金属场景，开通时系统会根据网络拓扑现状优先选择相邻位置的节点以降低节点间通信时延
调度策略	多选项	否		支持 DRF，Binpack ,Gang 三种调度策略，可以多选
描述	输入框	否	0-300	

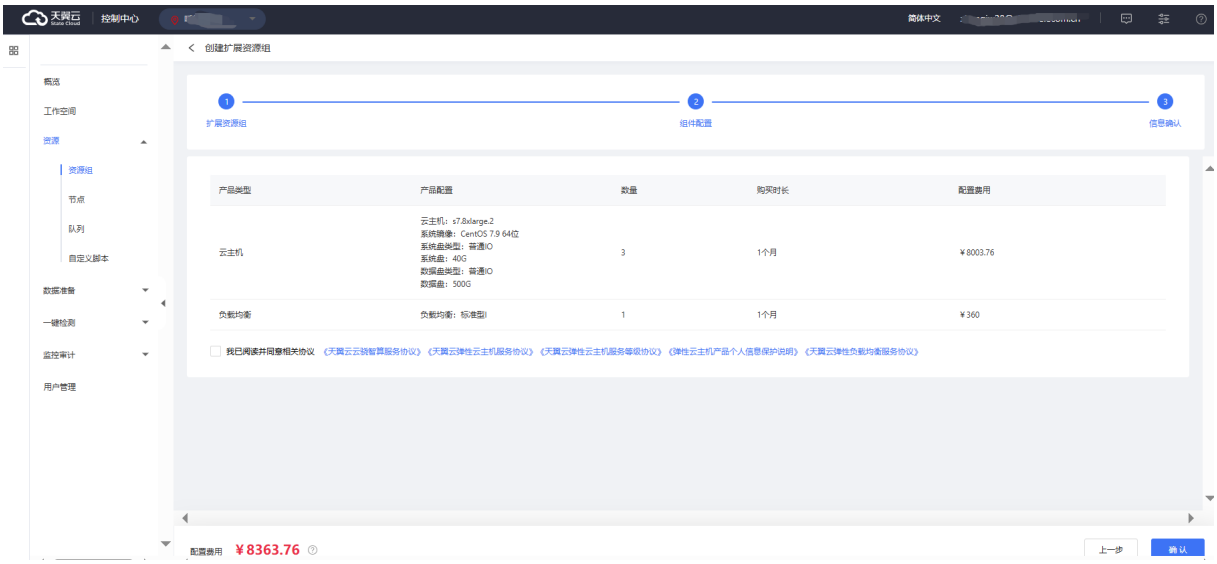
2）输入组件配置信息：



字段名称	类型	是否必填	长度	说明
资源组规模	单选	是		包括 4 种，1-100 节点，101-300 节点，301-500 节点，500 节点以上，默认选择 1-00 节点
计费模式	单选	是		支持包周期或按需两种模式，默认包周期
规格	单选	是		选择资源池可选的规格
操作系统	单选	是		选择管理节点的操作系统
系统盘	单选	是		仅支持超高 IO 类型，最小 40G，系统盘规格范围 40-2048
数据盘	单选	是		选择一块数据盘，仅支超高 IO 类型，最小

				500G，数据盘规格范围 500-32768，
数量	输入框	是		默认为 3
API Server	选择	是		标准 I 型，增强 I 型，高阶 I 型，默认标准 I 型
使用 EIP 暴露 API server	选择	否		默认未勾选
EIP	选择	否		当使用 EIP 暴露 API server 为选中状态时，需要选择已有的 EIP
时长	选择	是		包周期时显示，支持按年、按月，按月支持 1-11 月，按钮支持 1-3 年
续订方式	选项	是		包括自动续订，手动续订

3）开通信息确认：勾选协议，点击确认按钮完成扩展资源组创建。完成支付后即完成资源组的创建，后续资源组管理员便可在资源组列表/详情页中对资源组进行管理。



4.2.1.2 资源组列表

使用前提

当前用户是主账号。

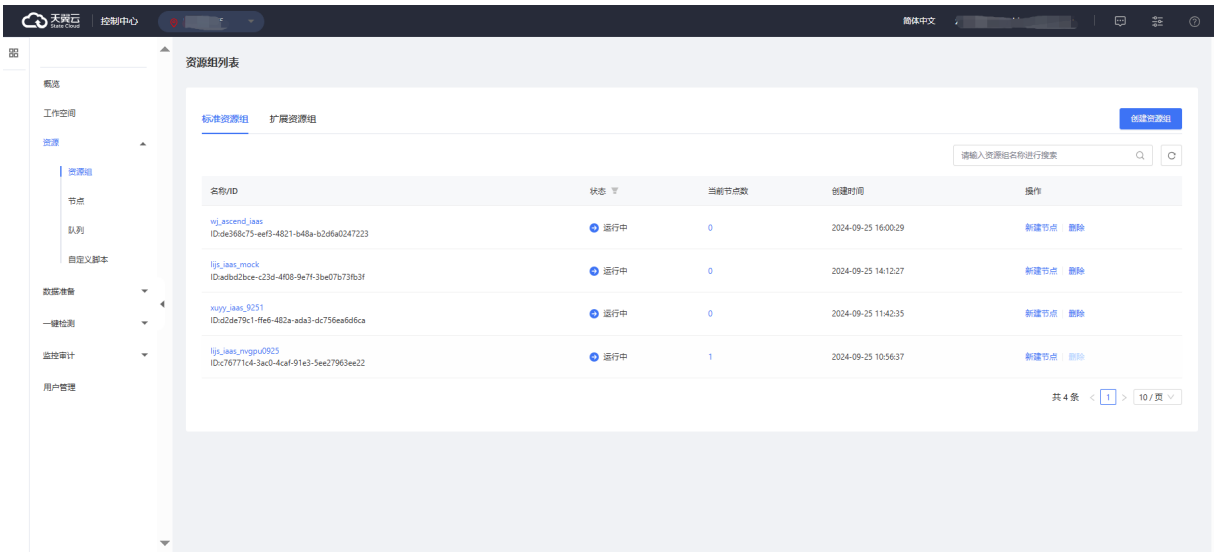
操作步骤

1. 登录云骁智算，单击左侧导航栏中的资源组，进入资源组列表页。
2. 资源组列表页可以查看资源组的状态、当前节点数、创建时间、描述和操作。

3. 可以通过输入资源组的名称对资源组进行过滤。

4. 点击“新建节点”可以对资源组扩容。

5. 点击“删除”，可以删除当前资源组。



6. 资源组“重试“，可以对配置失败的资源组重新进行配置。

7. 资源组状态

控制台状态	状态属性	状态说明	可进行操作
创建中	中间状态	用户可见，指资源组创建中的中间状态	
创建失败	稳定状态	用户可见，指资源组控制面云主机、elb 下单失败	
配置中	中间状态	用户可见，指资源组组件创建过程中的状态，包括创建 VPCE，配置相关组件	
配置失败	稳定状态	用户可见，创建失败的状态，可以重试	重试
运行中	稳定状态	用户可见，资源组组件启动成功，各组件正常运行，资源组处于运行中的状态	
删除中	中间状态	用户可见，删除过程中	

		的状态	

4.2.1.3 资源组详情

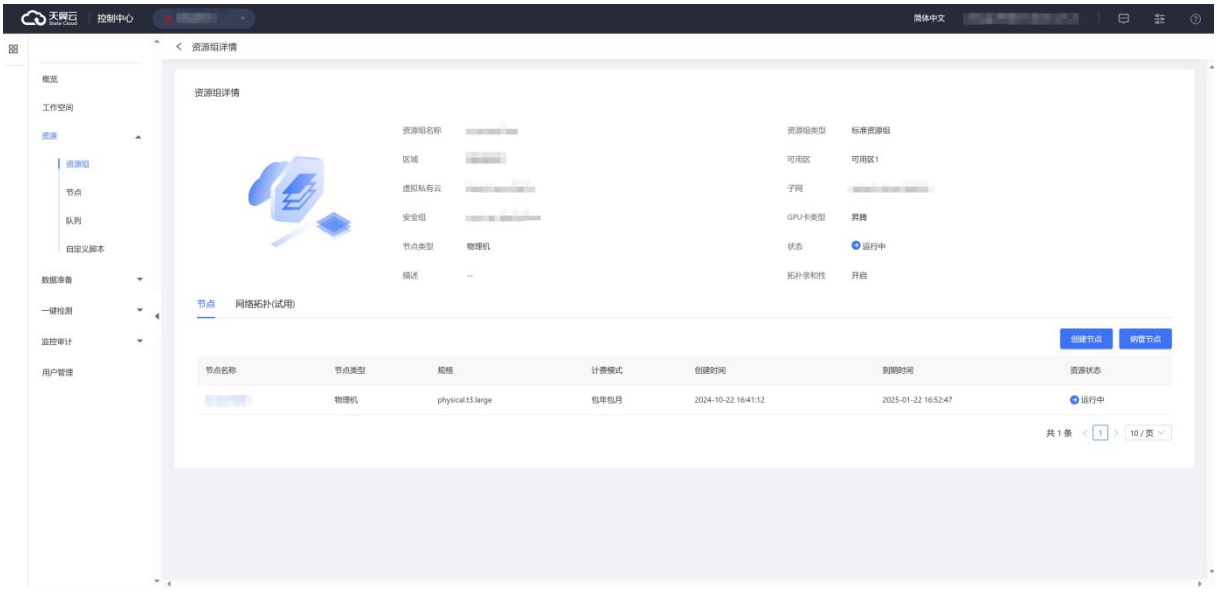
使用前提

当前用户是主账号。

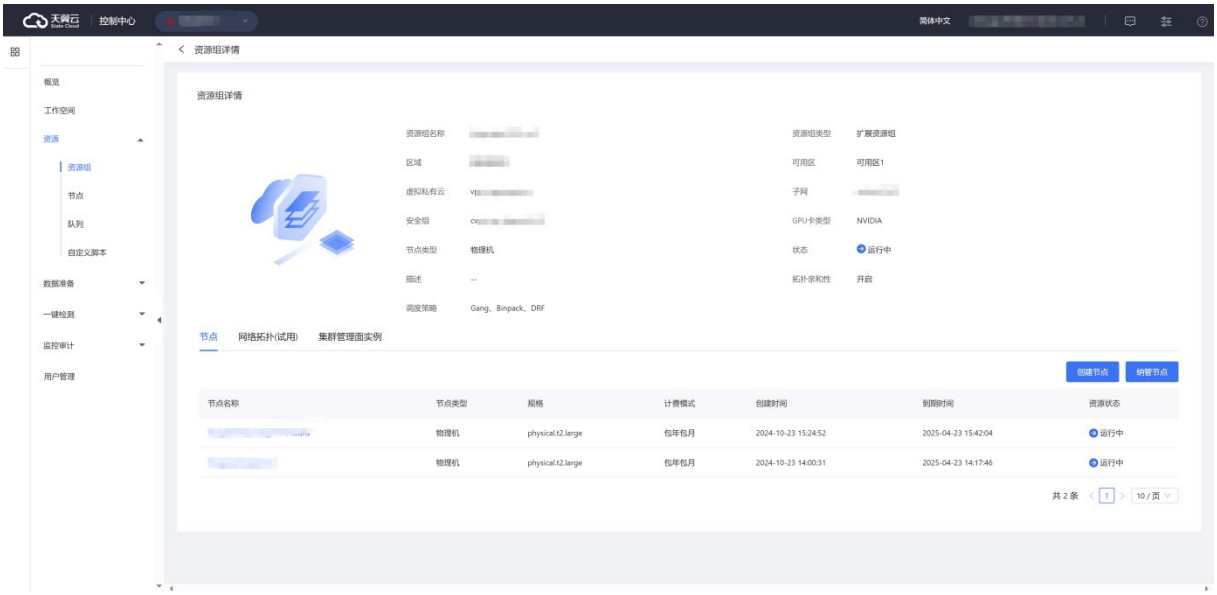
操作步骤

1. 登录云骁智算，单击左侧导航栏中的资源组，进入资源组列表页。
2. 单击列表页中资源组名称，进入资源组详情页面。
3. 标准资源组可以查看资源组的基本信息、节点列表和网络拓扑结构（试用）。

节点信息仅展示该资源组下状态为已绑定的的节点。节点列表可点击创建节点、纳管节点操作，分别跳转到对应创建节点和纳管节点页面，并默认显示对应资源组信息。



4. 扩展资源组可以查看资源组的基本信息、节点列表、网络拓扑和集群管理面的实例信息。



4.2.1.4 删除资源组

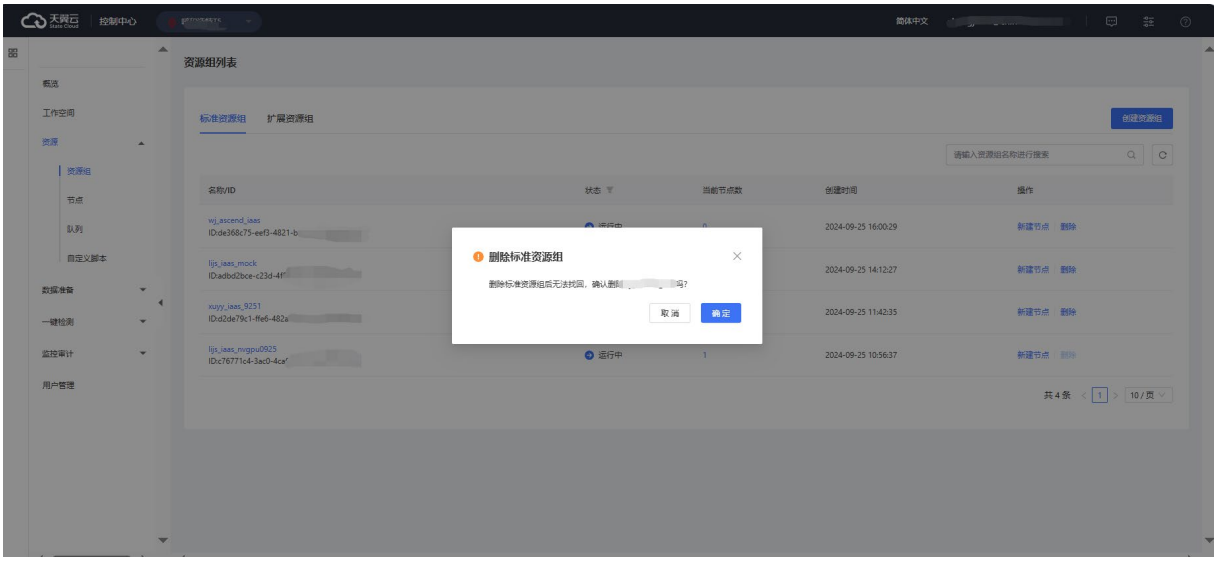
使用前提

当前用户是主账号。

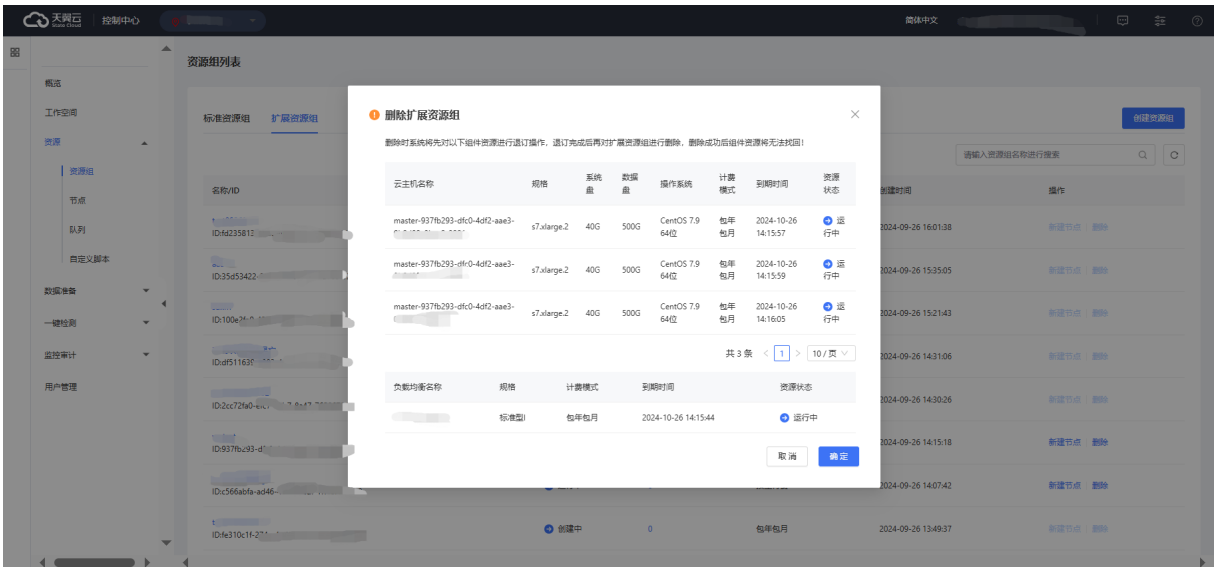
操作步骤

1. 登录云骁智算，单击左侧导航栏中的资源组，进入资源组列表页。

2. 标准资源组：删除资源组之前，需要将资源组下的全部节点解绑。当资源组状态为运行中且节点数为零时，该资源组可以进行删除操作，删除成功后资源组在控制台不可见，无法找回。点击标准资源组列表页的“删除”按钮，弹出提示框。点击“确认”删除资源组，点击“取消”保持在资源组列表页。



3. 扩展资源组：当资源组状态为运行中且组件状态为运行中，资源组下节点为零时，点击扩展资源组列表页的“删除”按钮，弹出提示框展示当前资源组管理面的资源，包含三台云主机和一个负载均衡实例，提示“删除时系统将先对以下组件资源进行退订操作，退订完成后再对扩展资源组进行删除，删除成功后组件资源将无法找回。”当资源组下的组件信息已退订或已销毁时，不需要再次对组件进行退订，可直接完成扩展资源组的删除。



4.2.1.5 续订资源组

使用前提

当前用户是主账号。

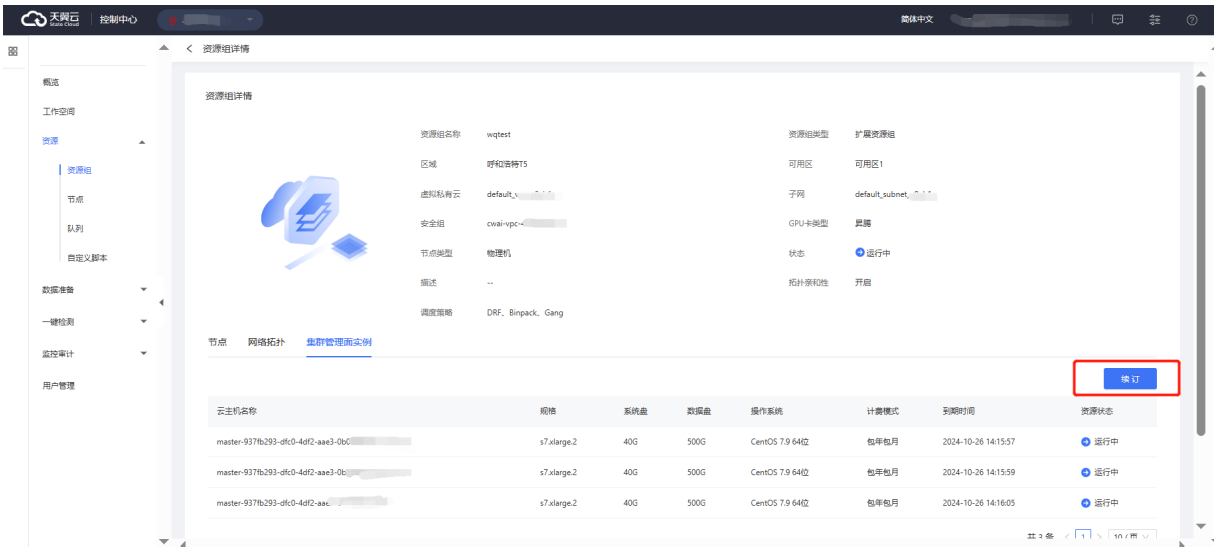
当前资源组是扩展资源组，仅管理面资源为运行中或已到期时支持续订。

操作步骤

1. 登录云骁智算，单击左侧导航栏中的资源组，进入资源组列表页。

2. 只有扩展资源组可以续订，续订指扩展资源组的管理面资源进行续订。单击扩展资源组列表页中资源组名称，进入资源组详情页面，续订资源组的管理面资源。

3. 点击集群管理面实例，查看当前集群管理面的云主机和负载均衡的信息。点击“续订”按钮进入订单续订管理页面，根据资源 ID 进行资源续订。续订时请务必保证云主机和负载均衡实例一同续订，以防资源组不能正常使用。



4.2.2 节点

节点即计算节点，目前云骁智算支持 GPU 云主机（部分资源池支持）和物理机作为资源组的计算节点。通过云骁平台可以对节点进行生命周期管理和批量操作。

云骁智算的计算节点支持 2 种来源：

- 新建节点：新订购节点加入资源组
- 纳管节点：将已开通的存量节点加入资源组

不同来源节点的操作不同，详情参见 FAQ-退订节点和移除节点有什么区别?

云骁支持 2 种节点计费模式：

- 包年包月：云主机和物理机都支持。
- 按量计费：云主机支持，物理机在部分资源池可支持。

4.2.2.1 新增节点

使用前提

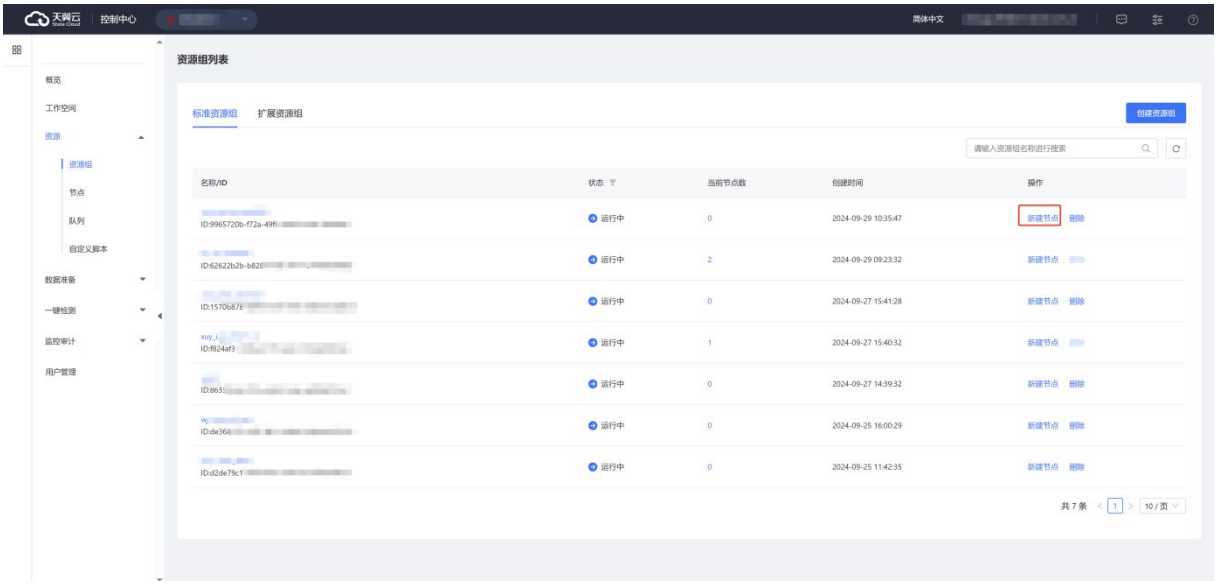
已创建标准资源组或扩展资源组

操作步骤

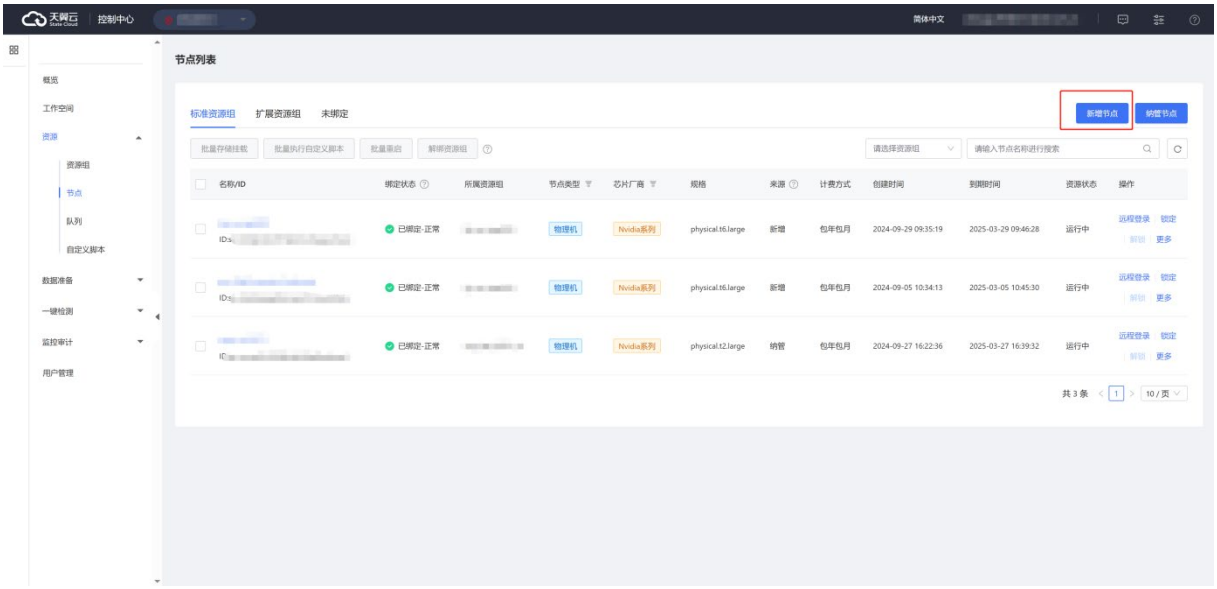
1. 进入控制台，选择云骁智算

2. 可以创建新节点加入到已有的资源组中，新建节点有三处入口：

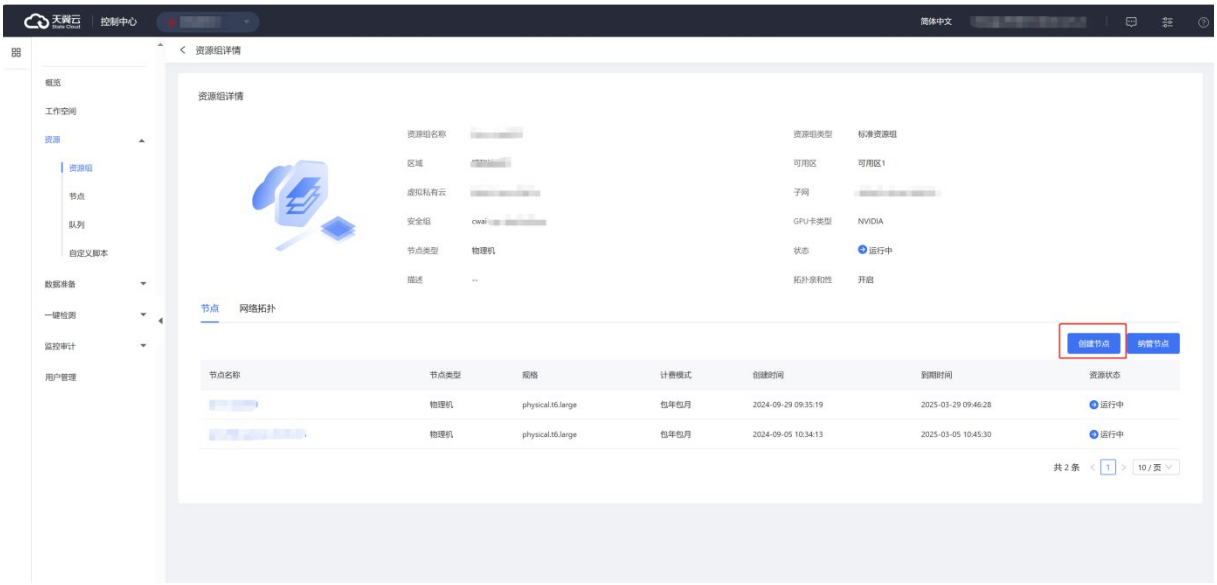
1）在资源组列表点击新建节点按钮



2）在节点列表点击新增节点按钮

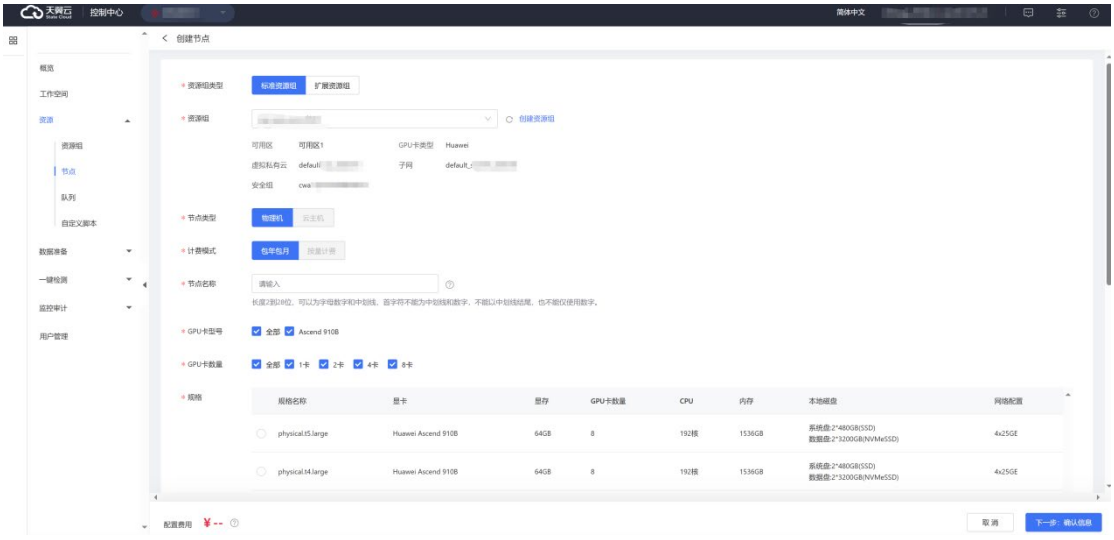


3）在资源组详情页点击新增节点按钮

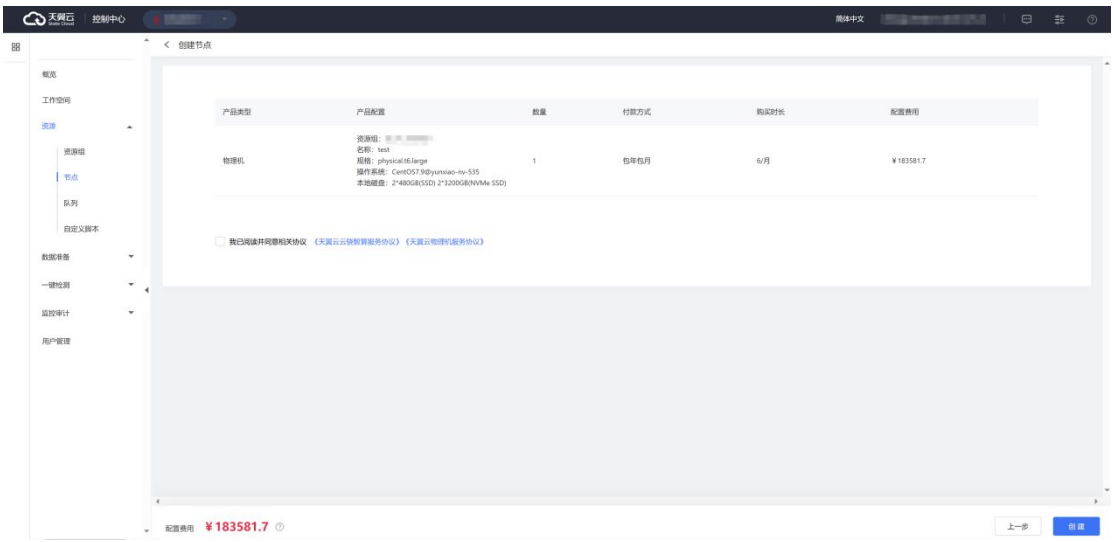


3. 在创建节点页面填写开通参数信息

1）选择节点类型、计费模式、节点名称、节点配置，完成节点的信息填写。云主机批量创建的上限为 50 台，物理机批量创建的上限为 50 台。



2）在信息确认页完成信息确认后，点击“创建”即可完成节点创建。



4.2.2.2 纳管节点

使用前提

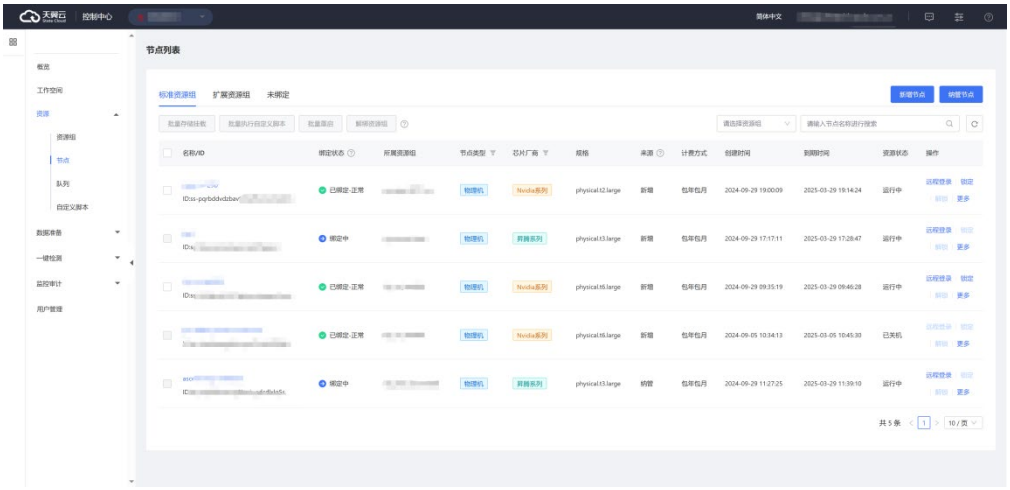
已创建标准资源组或扩展资源组，且资源组状态为运行中。

操作步骤

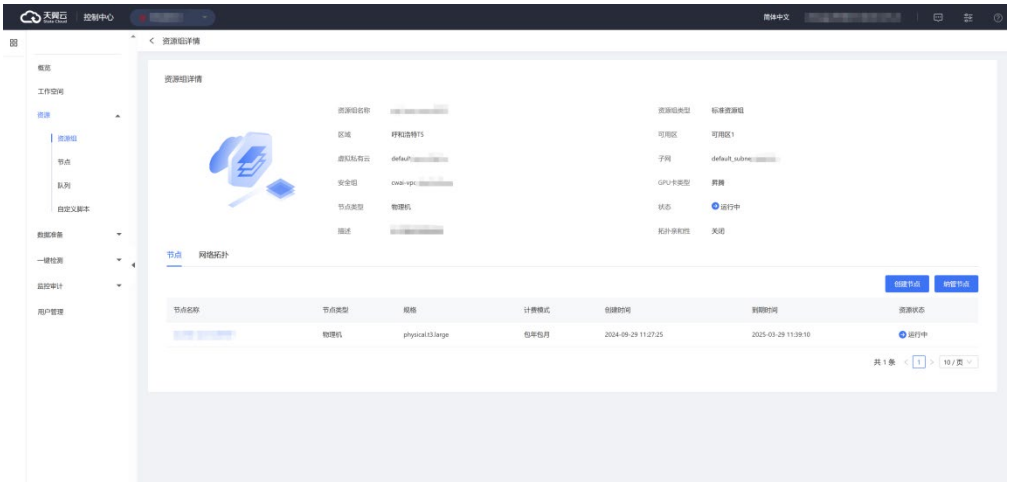
1. 进入控制台，选择云骁智算。资源组纳管已有节点，包括的从云骁解绑的节点以及云管未纳管的节点。

2. 可以纳管已经创建的节点到已有资源组，纳管节点有三处入口：

1）在节点列表点击纳管节点按钮



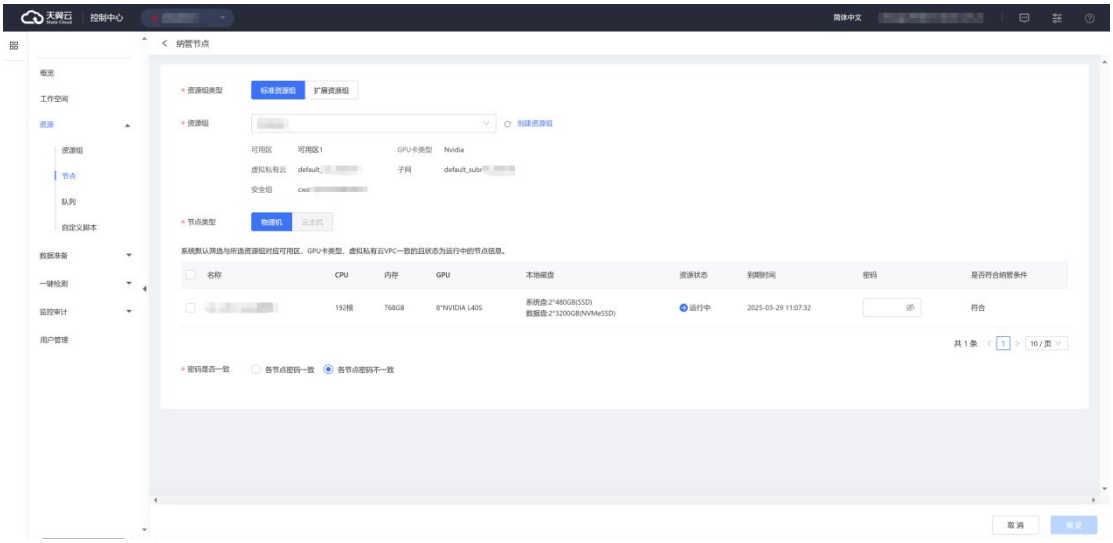
2）在资源组详情页点击纳管节点按钮



3. 选择资源组、规格，系统判断是否符合纳管条件，点击确定按钮，完成纳管，节点列表中显示所选资源组所在的可用区、VPC、GPU 卡类型一致的节点信息，包括未绑定状态和未纳管到云骁的节点。判断节点是否符合纳管条件的原则有：

（1）标准资源组：判断镜像、安全组是否符合纳管条件

（2）扩展资源组：判断镜像、安全组是否符合纳管条件。



4.2.2.3 节点列表

使用前提

节点数量>=1。

操作步骤

节点列表展示资源组的计算节点，云骁智算支持的云主机（部分资源池支持）和物理机作为资源组的计算节点。

1. 点击进入节点列表页。

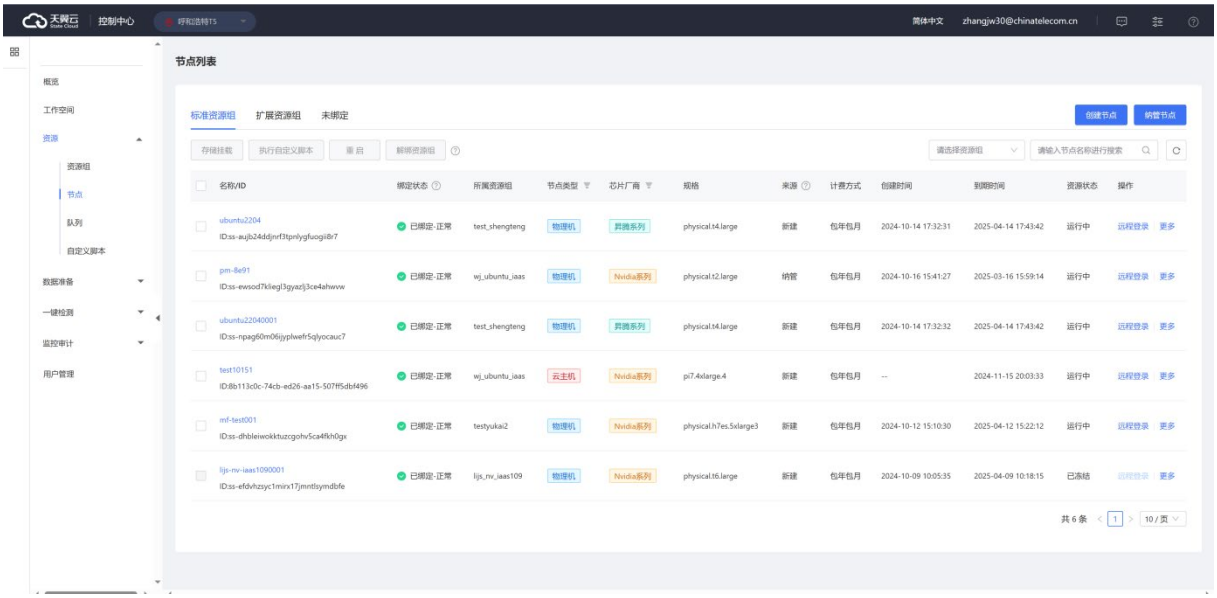
2. 节点列表可以查看标准资源组的节点、扩展资源组的节点和未绑定的节点，展示节点的基本信息。

3. 未绑定节点根据来源分为新建节点和纳管节点，都可以再次与已有资源组进行绑定。

4. 纳管节点：其中纳管节点可以进行移除，将非云骁平台开通的节点资源与资源组解绑并移除出节点列表。

5. 新建节点：新建节点则不支持移除操作，提示“移除仅支持来源为纳管的节点”。新建节点可以点击”退订“完成资源退订。

6. 根据节点所属的资源组和节点名称，对节点列表进行过滤。

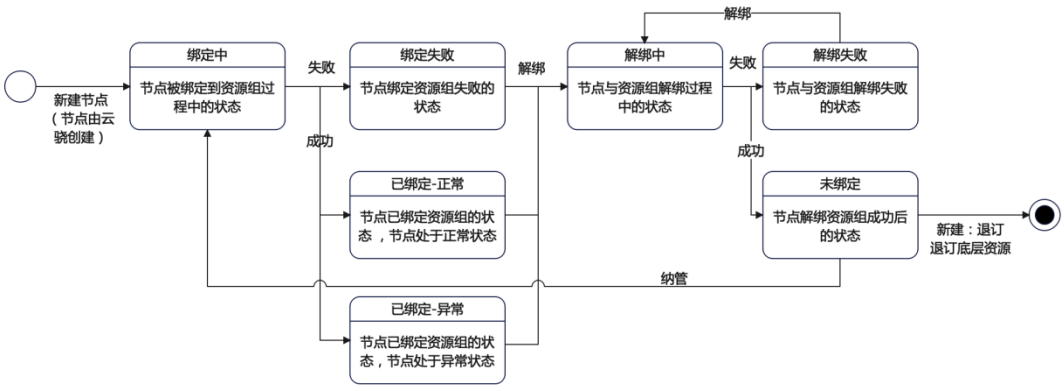


7. 节点绑定状态说明

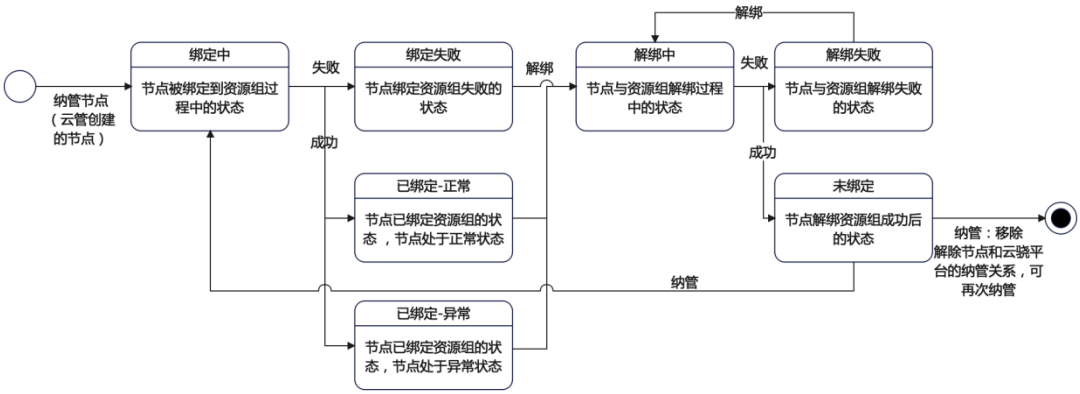
控制台状态	状态属性	状态说明	可进行操作
绑定中	中间状态	用户可见，指节点被绑定到资源组过程中的状态	
绑定失败	稳定状态	用户可见，指节点绑定资源组失败的状态	解绑资源组
已绑定-正常	稳定状态	用户可见，节点已绑定资源组的状态 ，节点处于正常状态	解绑资源组
已绑定-异常	稳定状态	用户可见，节点已绑定资源组的状态，节点处于异常状态	解绑资源组
解绑中	中间状态	节点与资源组解绑过程中的状态	
解绑失败	稳定状态	节点与资源组解绑失败的状态	解绑资源组
未绑定	稳定状态	节点解绑资源组成功后的状态	

节点状态及操作：

- 新建节点：



- 纳管节点：



4.2.2.4 节点详情

使用前提

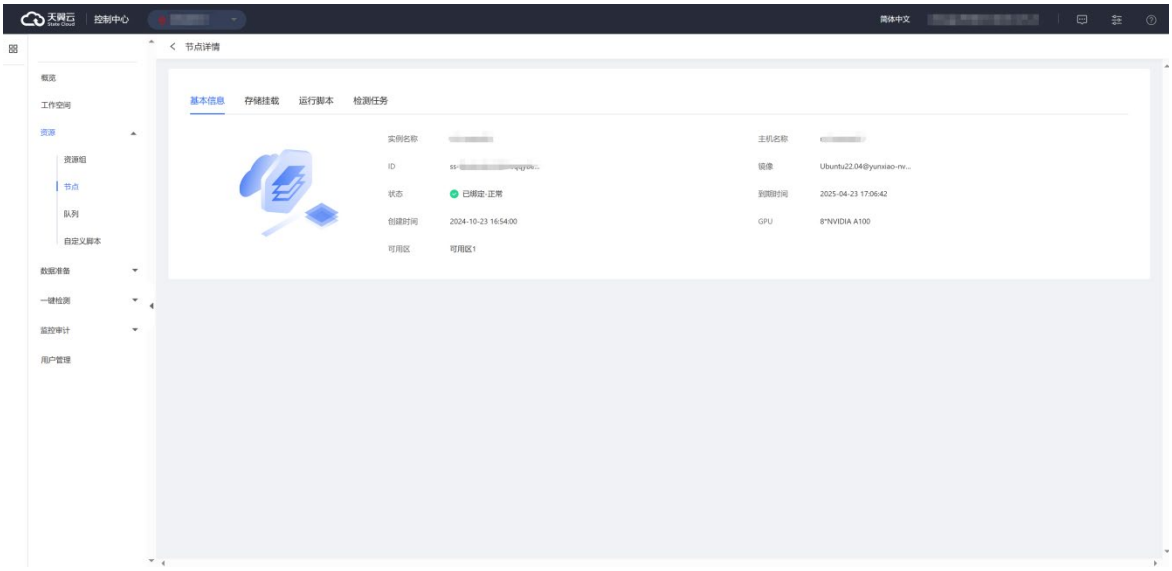
节点数量>=1。

操作步骤

节点详情页展示节点的基本信息以及与节点关联的存储、脚本、检测任务信息。

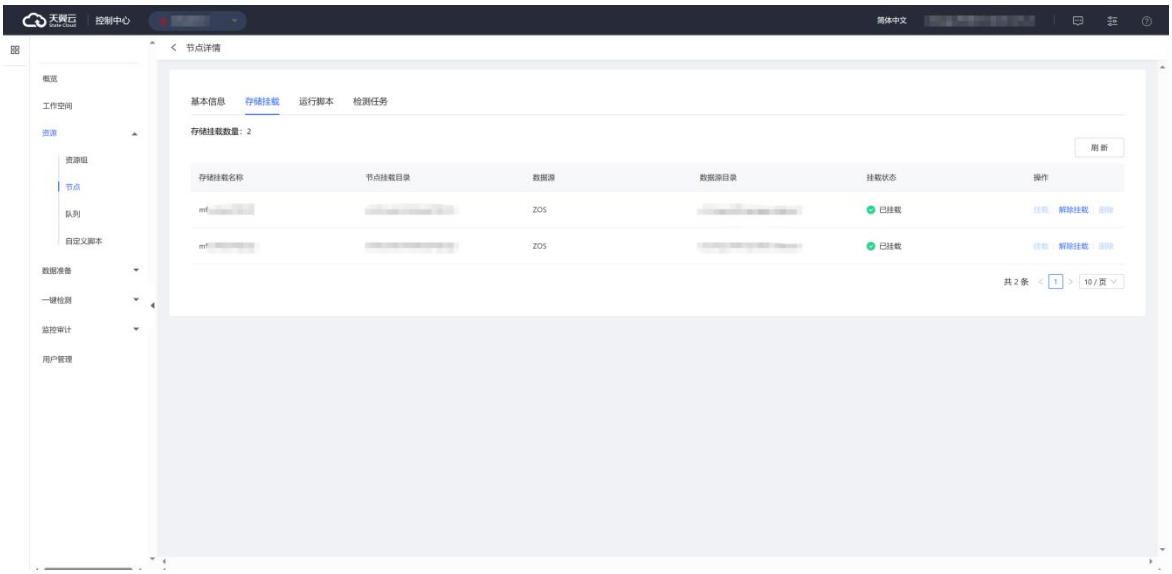
1、 点击节点名称进入节点详情页。

2、 基本信息菜单下展示当前节点的名称、状态和到期时间等基本信息。

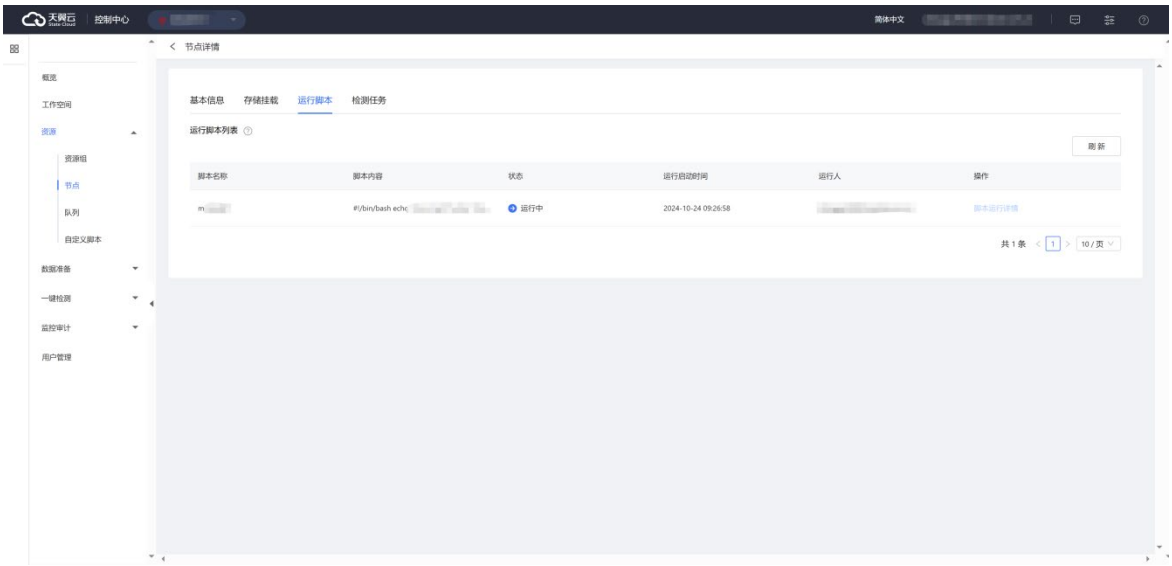


3、 存储挂载菜单下展示当前节点上存储挂载的名称、目录和数据源信息。可以通

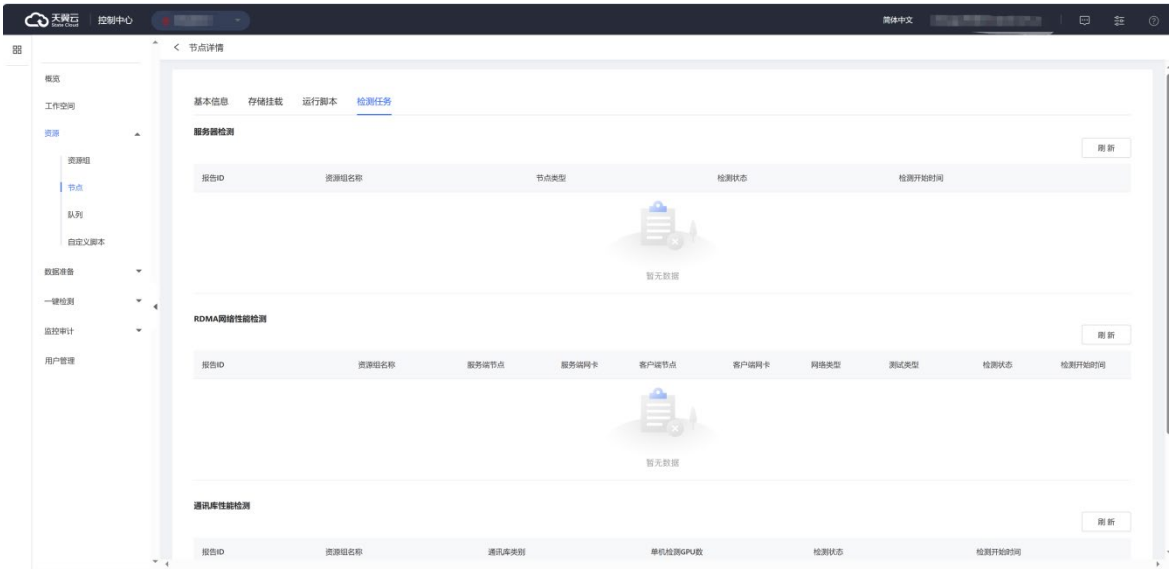
过按钮对存储挂载进行挂载、解除挂载和删除操作。



4、 运行脚本菜单下可以查看当前节点上运行的脚本。脚本运行完成后可以查看脚本运行详情。



5、 检测任务菜单下可以查看当前节点上运行的检测任务详情。



4.2.2.5 节点解绑

使用前提

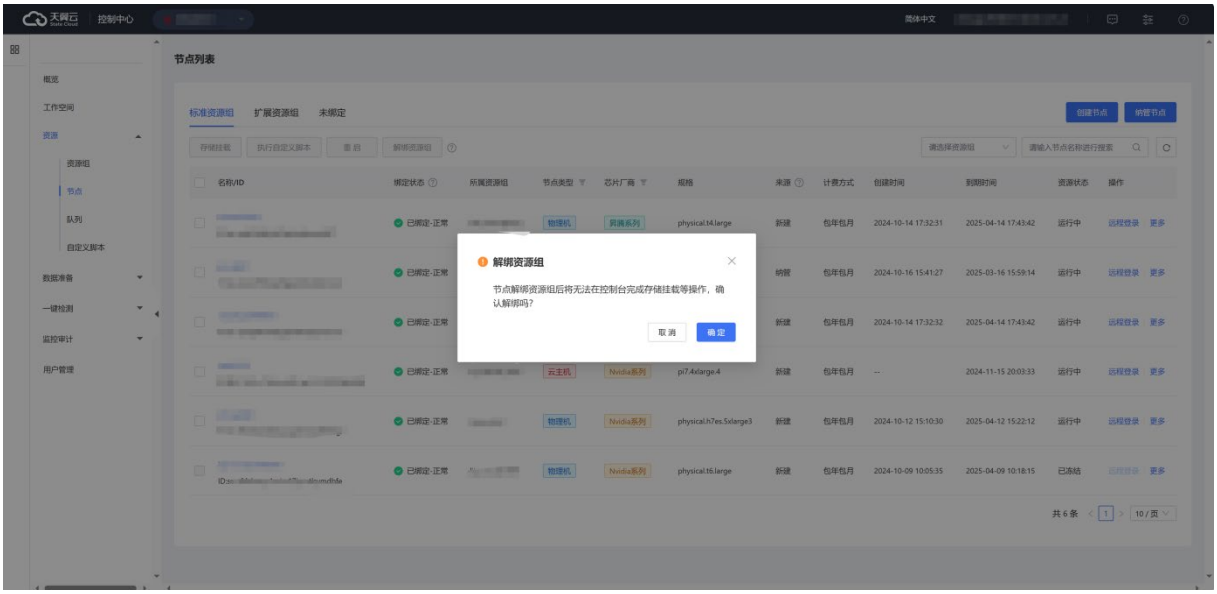
1. 当前资源组状态为运行中
2. 节点的绑定状态已绑定-正常，绑定失败，解绑失败

操作步骤

节点列表展示资源组的计算节点，可以解绑不再使用的计算节点。

1、点击进入节点列表页。

2、当不再使用该节点时，可以点击“解绑资源组”，将该节点与资源组解绑，节点进入到未绑定界菜单。解绑操作只是解除该节点和资源组的绑定关系，不包含节点退订动作。如需退订请根据节点类型（云主机、物理机）去相应的产品页面进行操作。



4.2.2.6 节点移除

使用前提

1. 节点处于未绑定状态
2. 节点来源为纳管节点

操作步骤

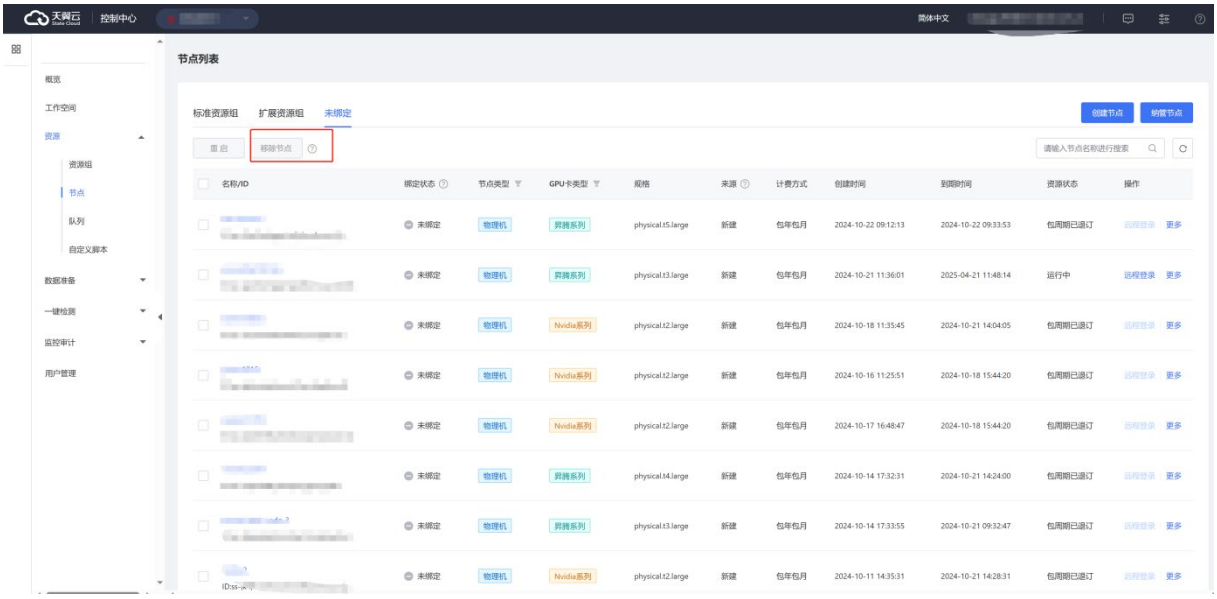
将非云骁平台开通的资源与资源组解绑并移除出节点列表。

1、进入节点列表页。

2、 点击左侧菜单资源-节点，选择未绑定页签。

3、 选择节点，点击”退订“按钮进行退订。

4、 选择节点，点击“移除“按钮，将纳管节点移除出云骁智算平台。



4.2.2.7 节点远程登陆

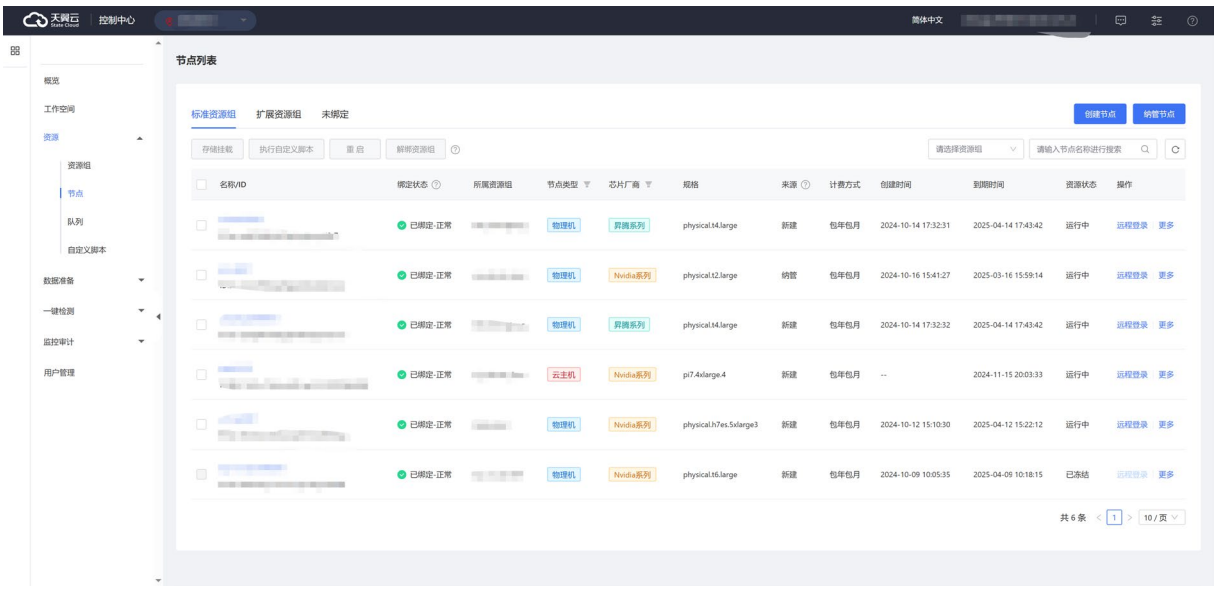
使用前提

节点的资源状态为运行中

操作步骤

节点列表展示资源组的计算节点，可以远程登陆到计算节点。

1. 点击进入节点列表页。
2. 节点资源状态为“运行中”时，平台支持节点的“远程登陆”，在节点列表页的节点操作栏，点击“远程登陆”，进入到节点的远程登陆界面。



4.2.2.8 节点批量管理

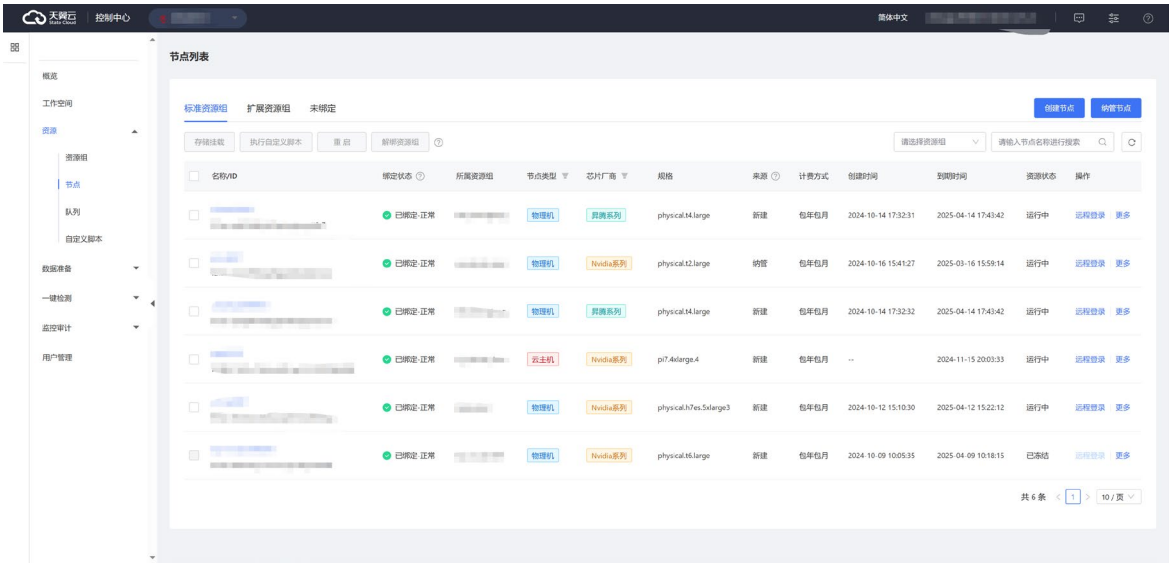
使用前提

节点的绑定状态已绑定-正常

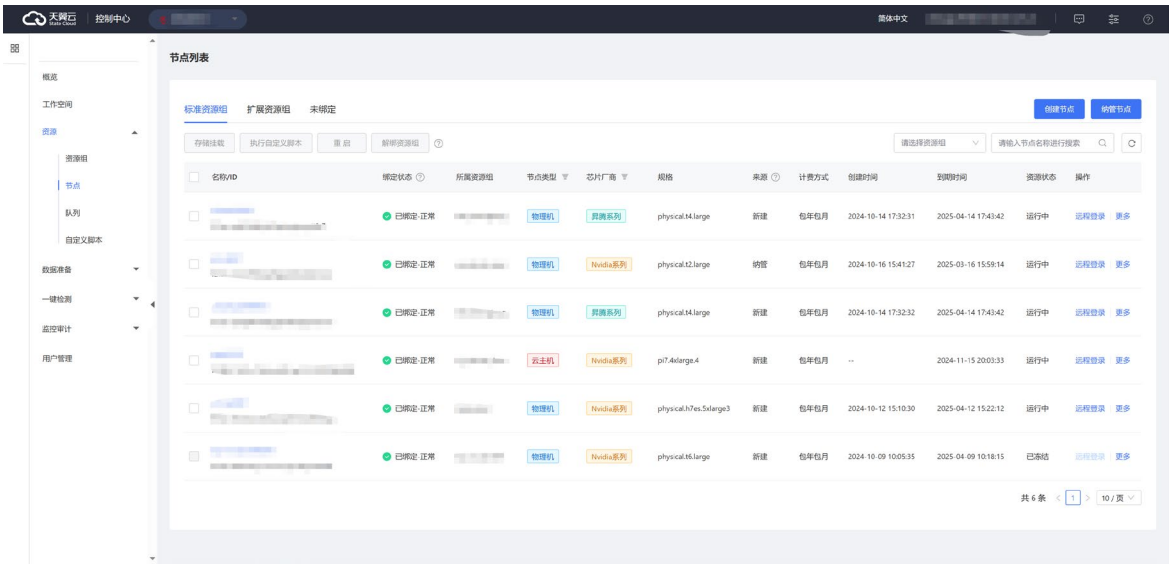
操作步骤

节点列表展示资源组的计算节点，可以通过批量操作选项对计算节点进行批量操作。

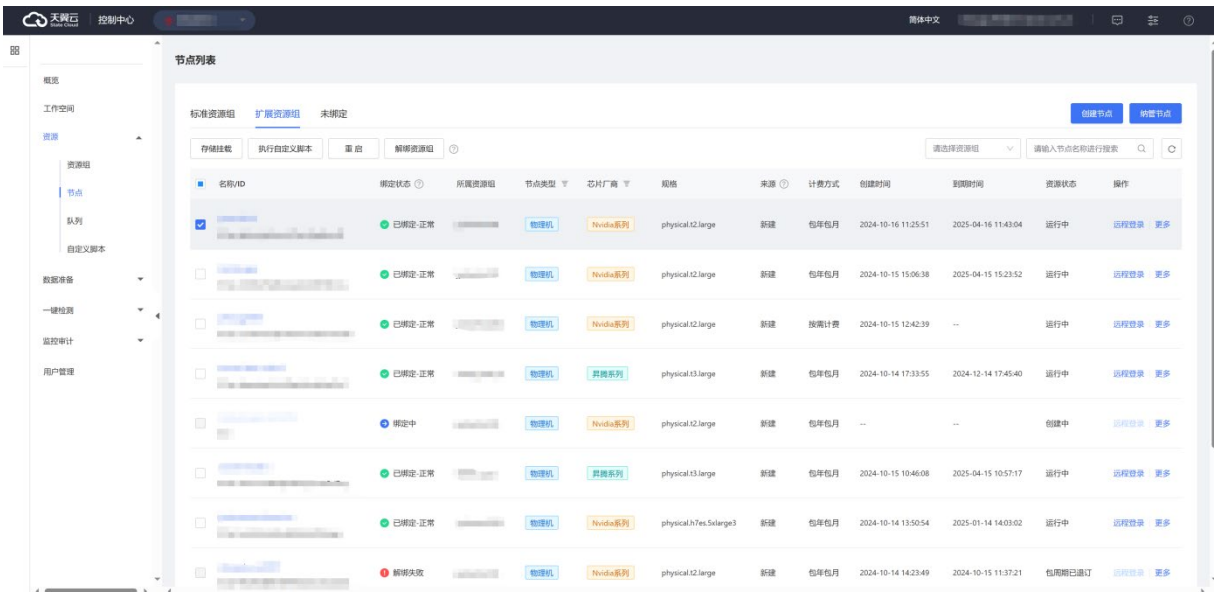
1. 点击进入节点列表页。
2. 根据当前节点所属的资源组类型，选择“标准资源组”或者“扩展资源组”筛选出目标节点。



3. 标准资源组：支持“批量存储挂载”、“批量执行自定义脚本”，可对选中的目标节点一键进行批量存储挂载、批量执行脚本运行，在弹出框中选择目标存储挂载或者目标脚本，点击“确认”开启节点的批量操作。可以对选中的节点“批量重启”，或者进行“解绑资源组”的操作。



4. 扩展资源组：支持“批量存储挂载”、“批量执行自定义脚本”、“批量重启”，可对选中的目标节点一键进行批量存储挂载、批量执行脚本运行和批量重启。



4.2.3 队列

队列是资源组内部分资源配额的集合，一个资源组可创建多个队列。在运行训练或推理任务时，通过将任务绑定到队列进行资源的排队和使用申请。只有云骁扩展资源组可以用来创建队列。

使用前提

- 当前用户是主账号。
- 资源组列表中存在>=1 个云骁扩展资源组。

操作步骤

1. 登录云骁智算，进入队列列表页。

2. 单击左上方“新建”，进入队列创建页面。

3. 在创建页面填写相关参数，具体参数如下：

    a. 资源组：必填，选择队列所属的扩展资源组

b. 队列名称：必填

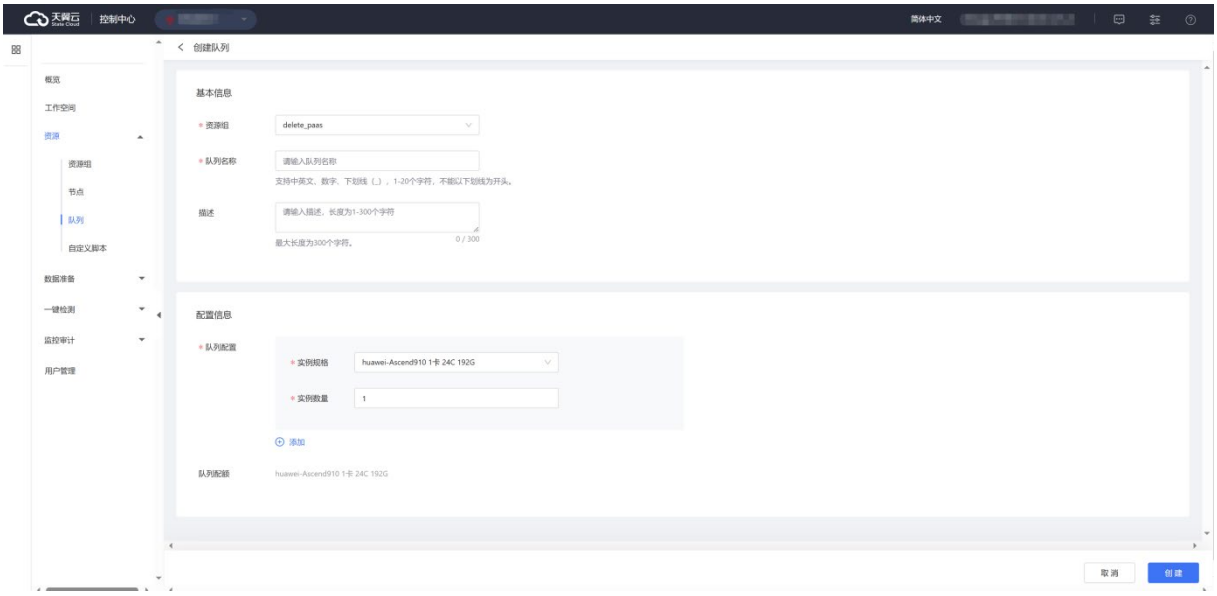
c. 描述：选填

d. 实例规格：必填

e. 实例数量：必填

4. 完成上述配置后，单击右下角“创建”即完成扩展资源组中队列的创建。

5. 队列删除：当前队列不再需要使用时可以删除当前队列。



4.2.4 自定义脚本

自定义脚本是方便用户对节点的配置进行自动化统一管理，例如对系统参数进行修改，对驱动版本进行升级等。自定义脚本结合节点的批量运行功能，可以支持用户对较大规模的节点进行自动化统一配置。

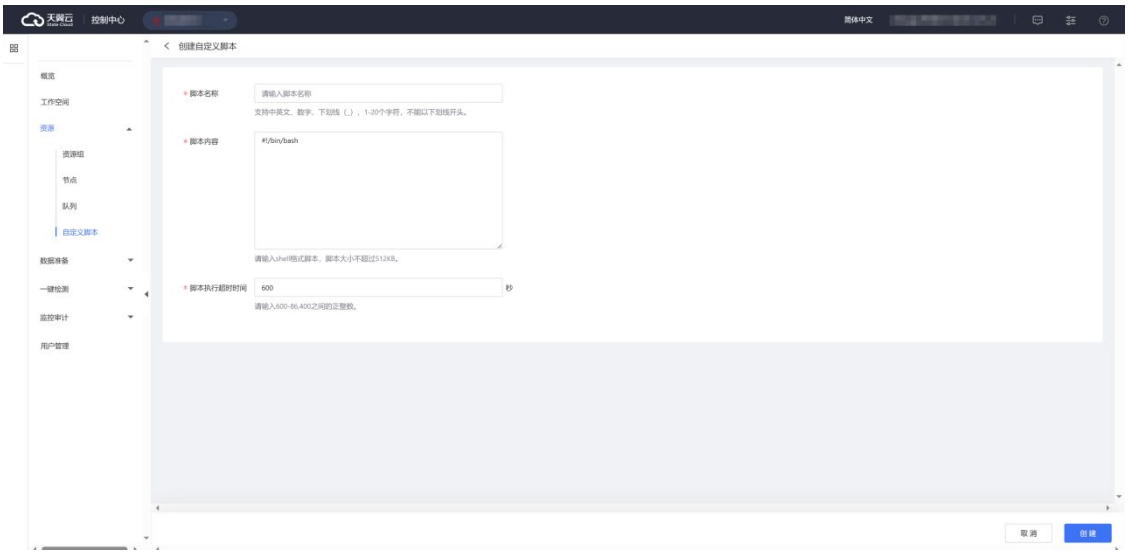
4.2.4.1 新建脚本

使用前提

当前用户是主账号。

操作步骤

1. 登录云骁智算控制台，单击左侧菜单栏的菜单项”资源” - ” 自定义脚本 “，点击页面“创建自定义脚本”按钮。



2. 输入脚本名称、shell 脚本内容、脚本执行超时时间。默认脚本超时时间为 600 秒。用户可以修改，但不能小于 600 秒。

3. 点击“创建”，完成脚本创建。

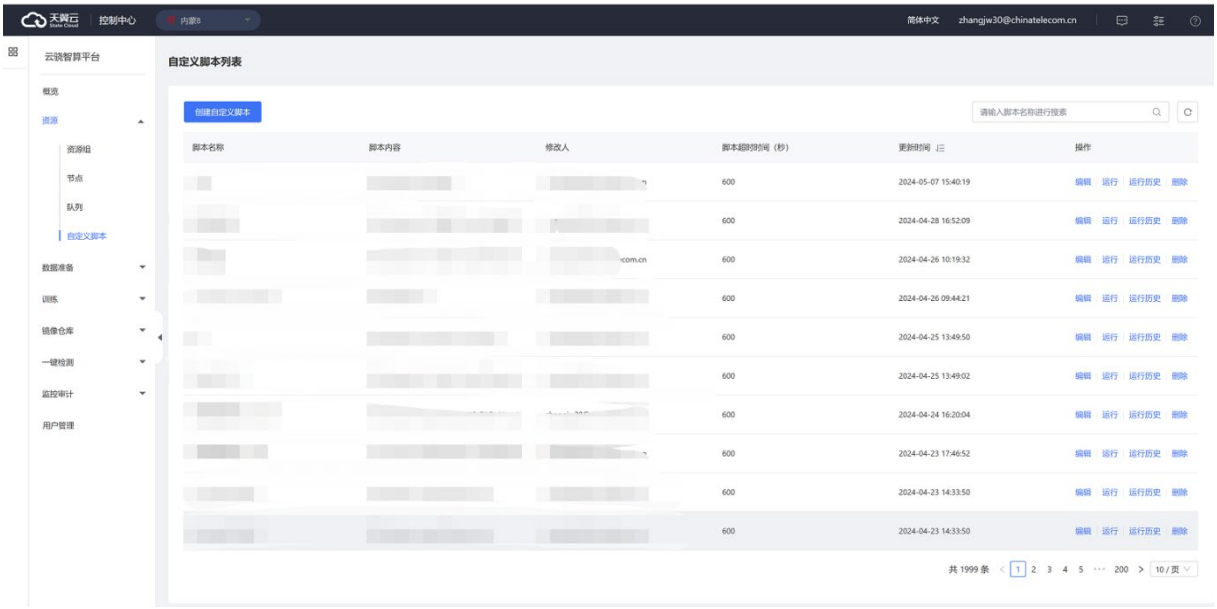
4.2.4.2 脚本列表

使用前提

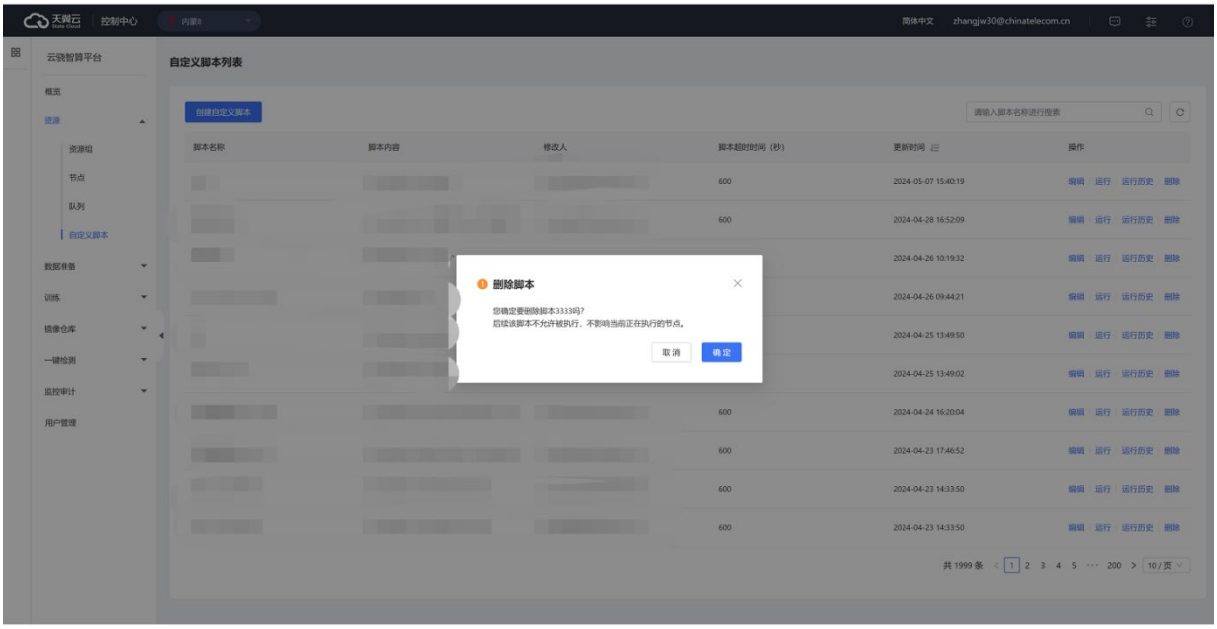
当前用户是主账号。

操作步骤

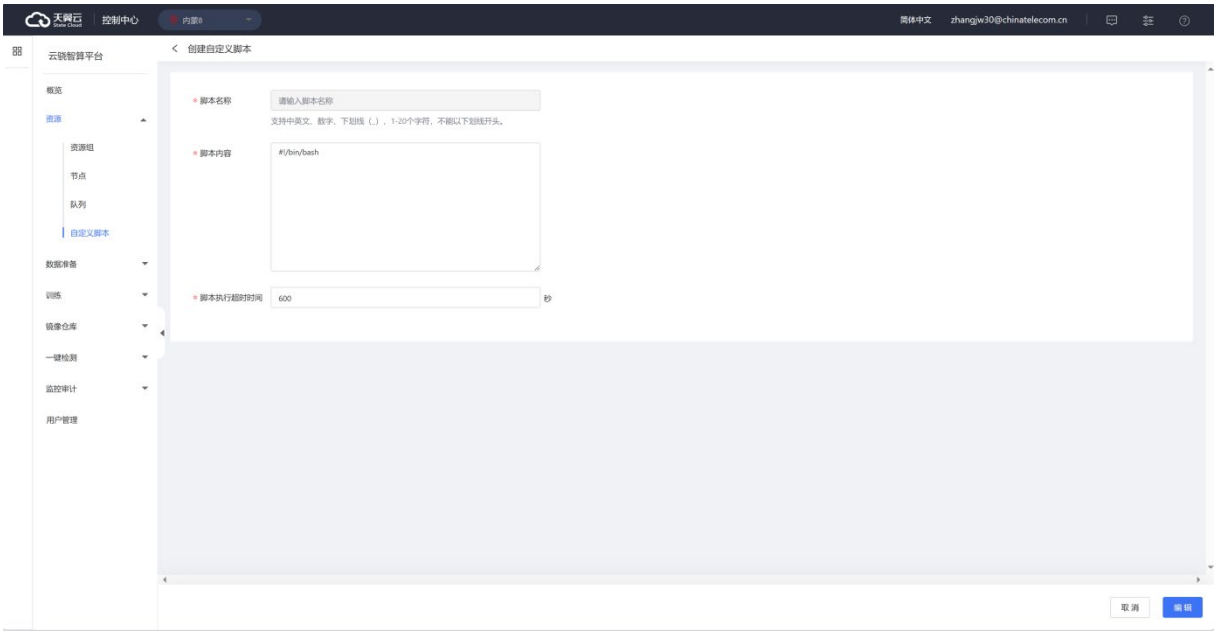
1. 登录云骁智算控制台，单击左侧菜单栏的菜单项【资源】- 【自定义脚本】，查看当前已经创建的自定义脚本列表。



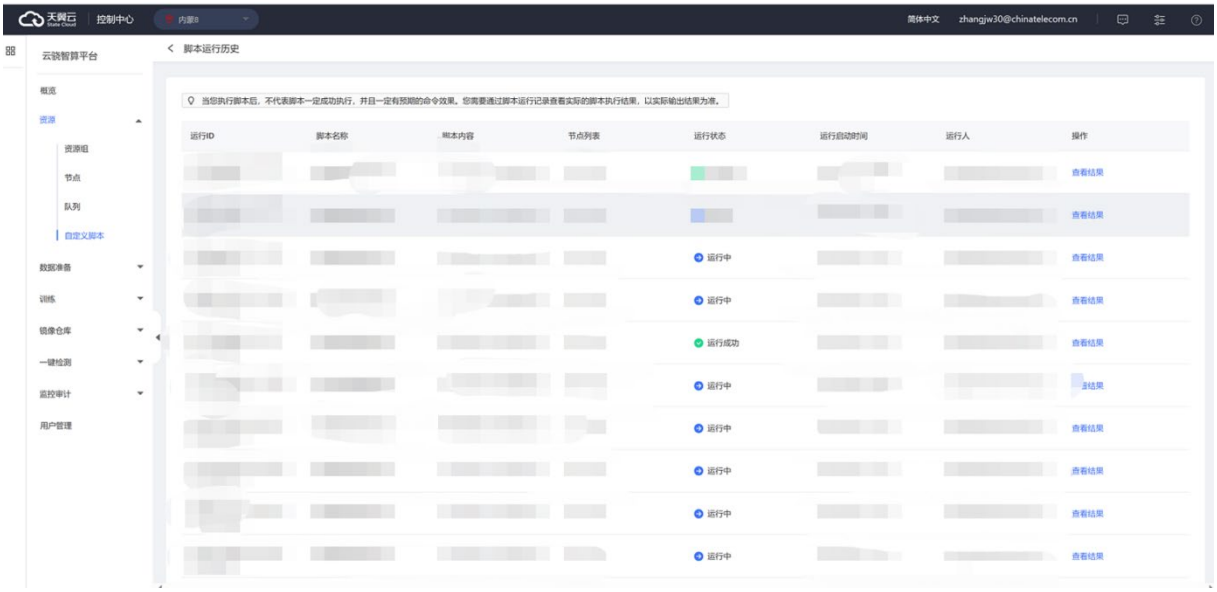
2. 对于脚本列表中不再使用的脚本，可以点击“删除”。后续该脚本不允许被执行，删除该脚本不影响当前正在执行的节点。



3. 点击“编辑”，可以修改脚本的内容和脚本超时时间。



4. 点击“运行历史”可以查看脚本运行的历史记录。



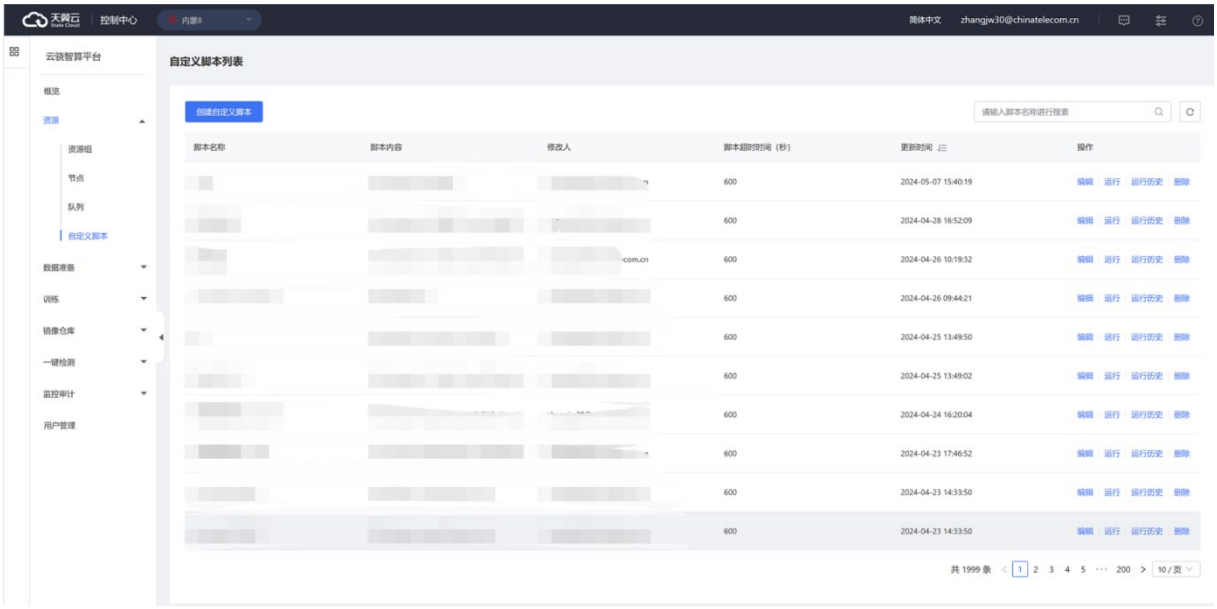
4.2.4.3 运行脚本

使用前提

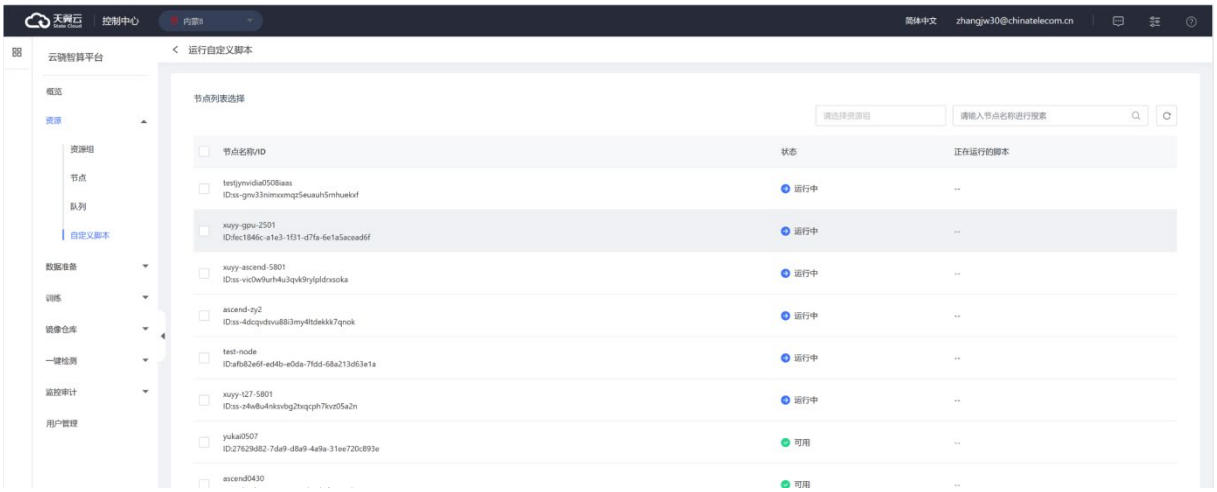
当前用户是主账号。

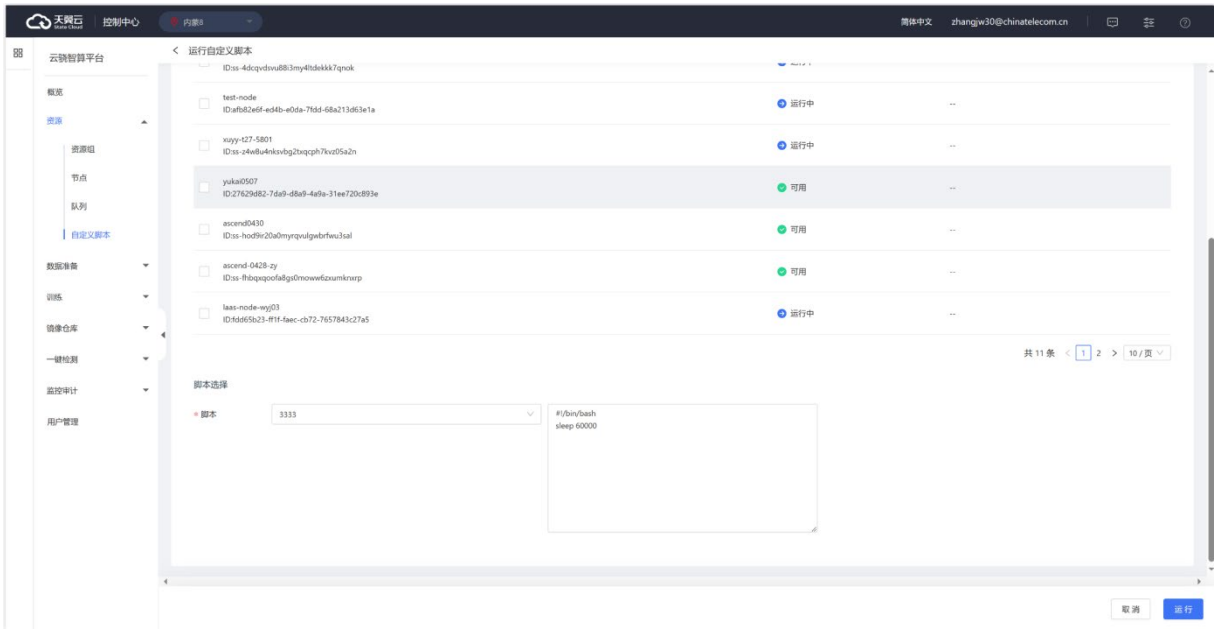
操作步骤

1. 登录云骁智算控制台，单击左侧菜单栏的菜单项”资源“- “自定义脚本 “，查看当前已经创建的自定义脚本列表，点击“运行”创建一个脚本运行。

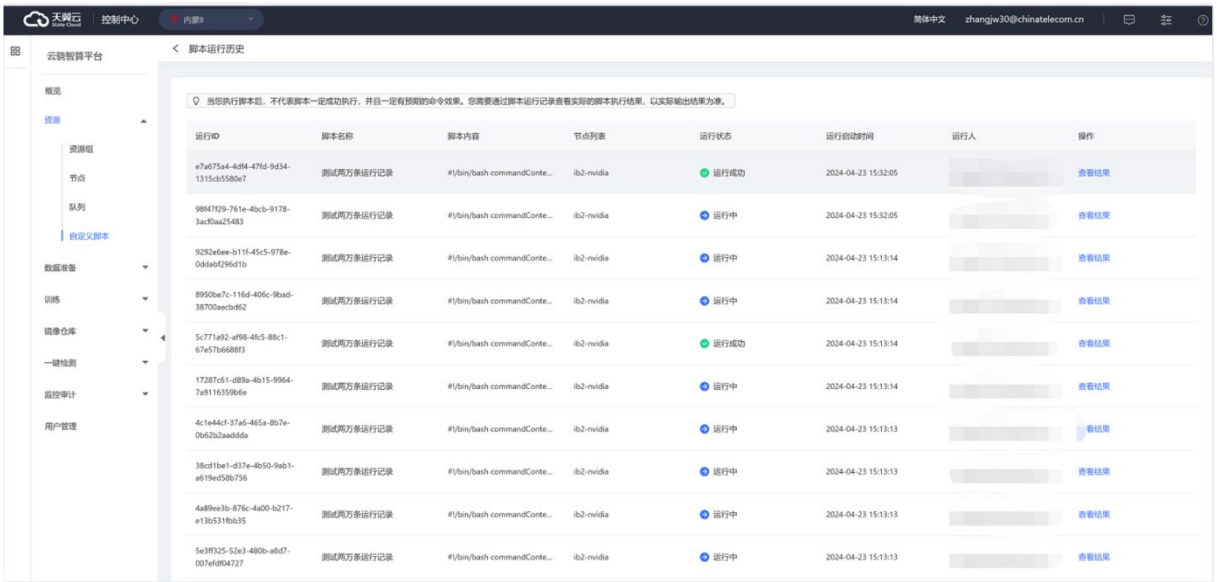


2. 选择运行自定义脚本的节点，查看当前的脚本名称和脚本内容。

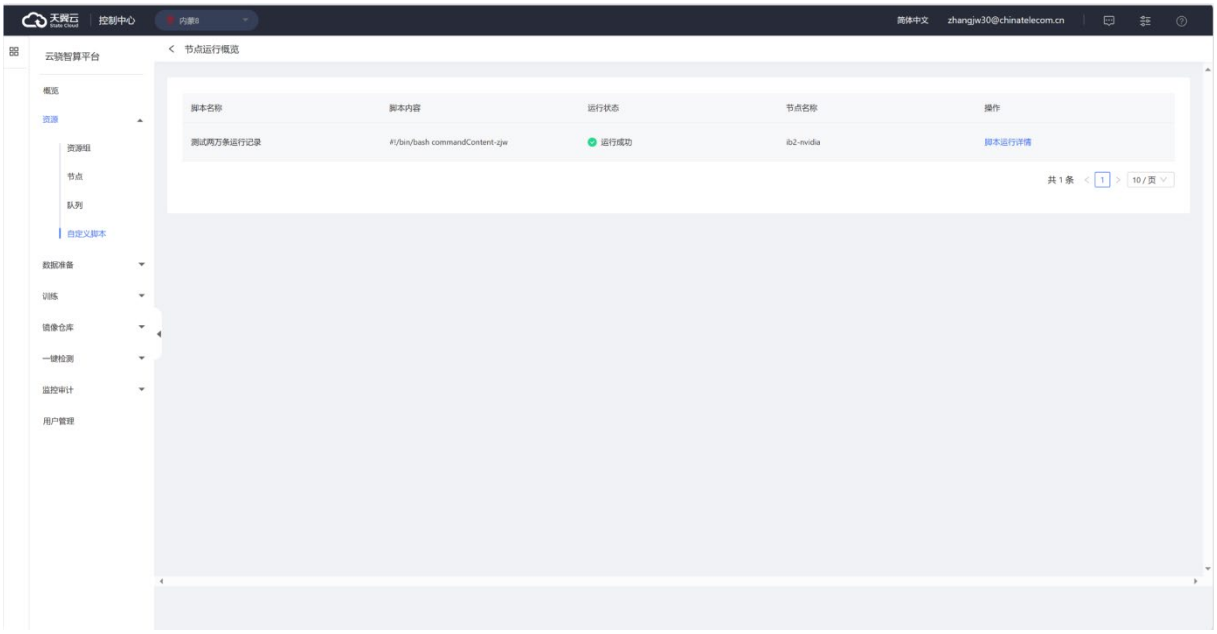




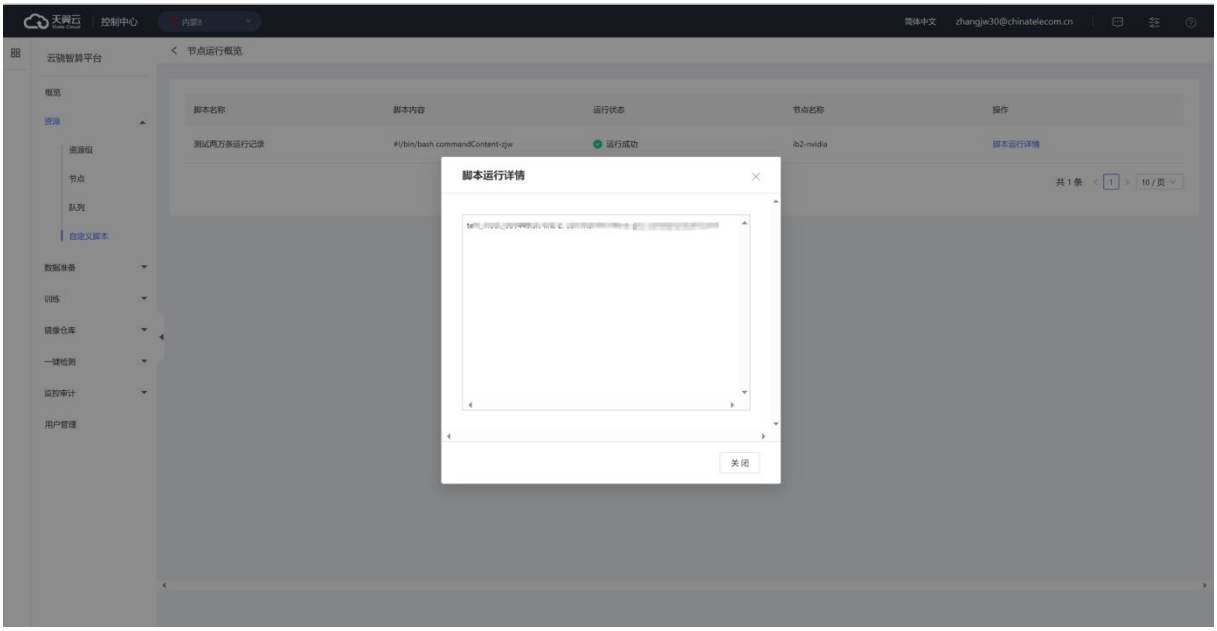
3. 点击“运行”，完成脚本运行的创建。可以在运行历史中查看详情。



4. 点击“查看结果”进入到节点运行概览页面。



5. 点击“脚本运行详情”查看脚本运行详细结果。



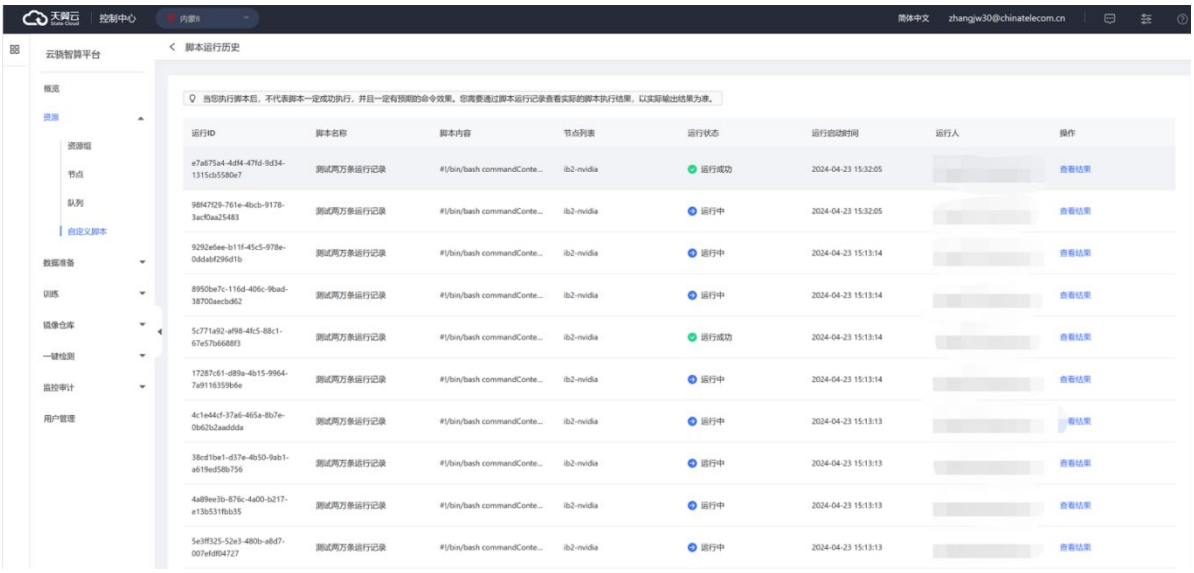
4.2.4.4 脚本执行记录

使用前提

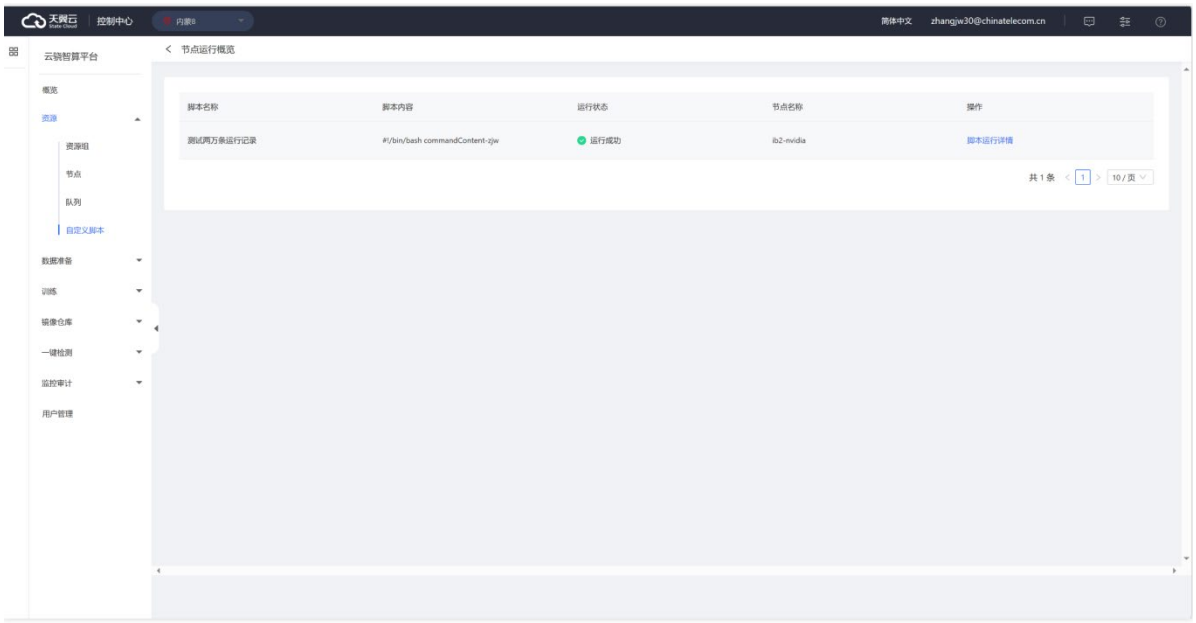
当前用户是主账号。

操作步骤

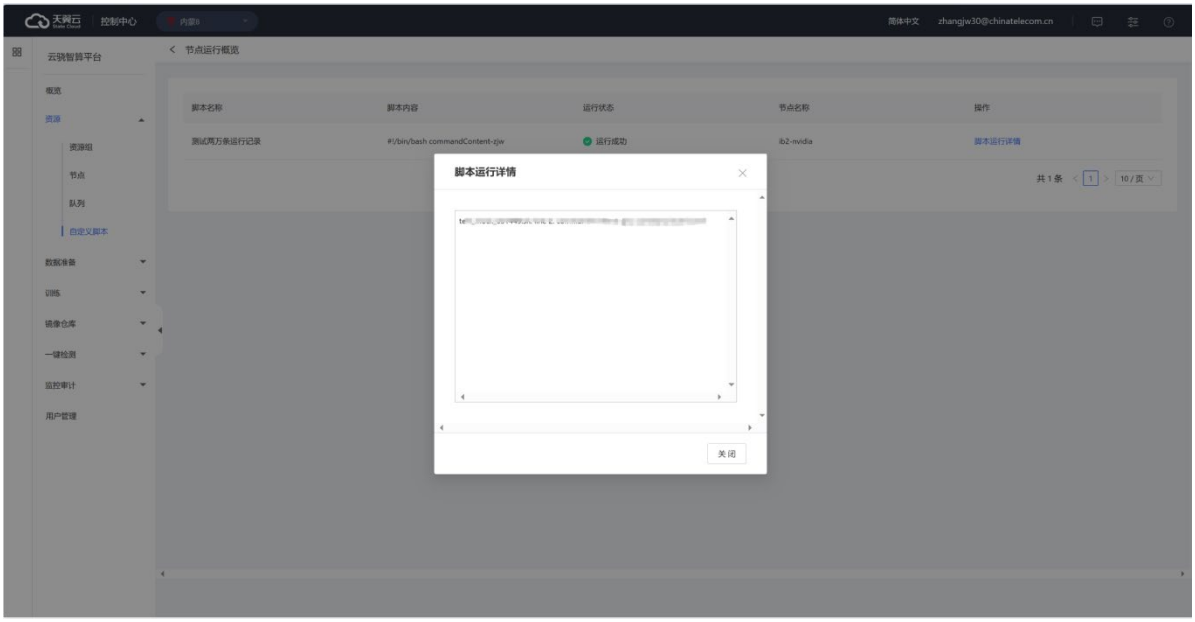
1. 点击脚本运行历史，可以在运行历史中查看脚本运行详情。



2. 点击“查看结果”进入到节点运行概览页面。



3. 点击“脚本运行详情”查看脚本运行详细结果。



4.3.数据准备

4.3.1.创建存储挂载

通过存储挂载，可支持用户将 ZOS 或 HPFS 实例批量挂载到相应的节点上，并且管理挂载目录。

使用前提

当前用户是主账号。

操作步骤

1. 登录云骁智算控制台，单击左侧菜单栏的菜单项“数据准备”>“存储挂载”，点击页面“创建存储挂载”按钮。

2. 进入“创建存储挂载”页面，根据要求配置相关信息。

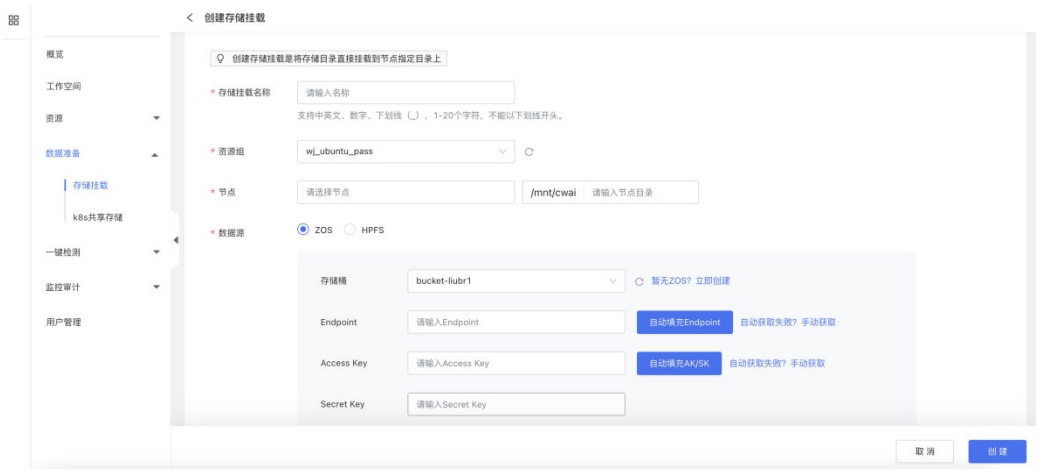
   a. 存储挂载名称：填写将要创建的存储挂载的名称，支持中英文、数字、下划线（_），1-20 个字符，且不能以下划线开头。

   b. 资源组：下拉选择需要挂载存储的资源组，支持搜索。

   c. 节点：选择资源组后，在节点处可选择该资源组下需挂载存储的节点。支持同时挂载多个节点。选择完节点后，添加需挂载的节点目录，目录须以/开头，且不能为空。

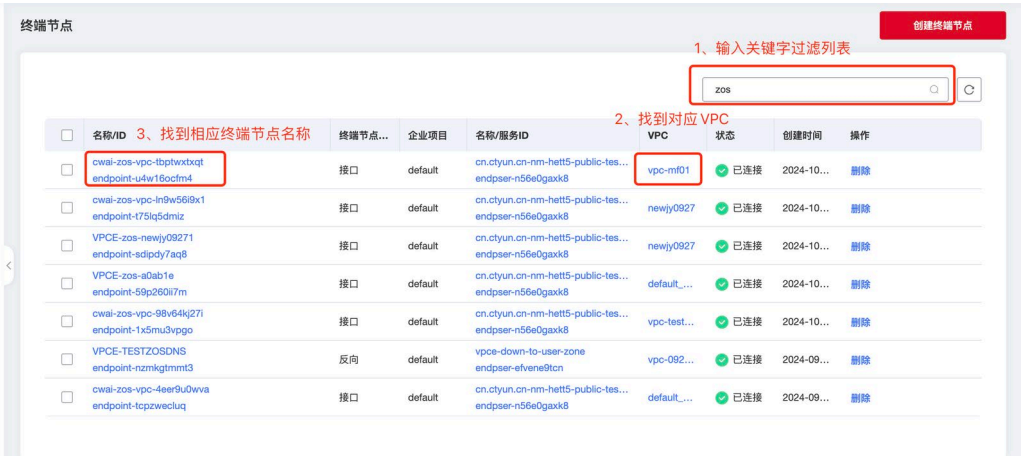
   d. 数据源：选择数据来源，支持 ZOS 与 HPFS 两种数据来源

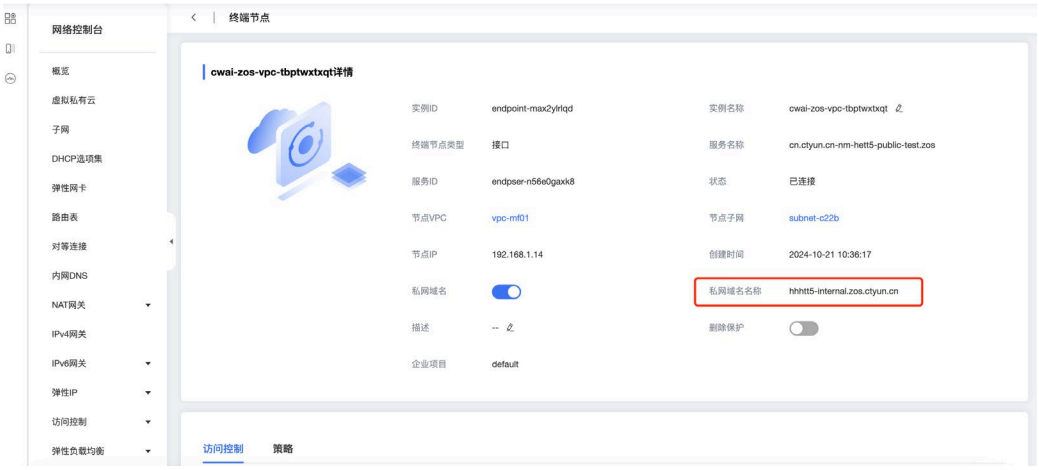
      i. 选择数据源为 ZOS：



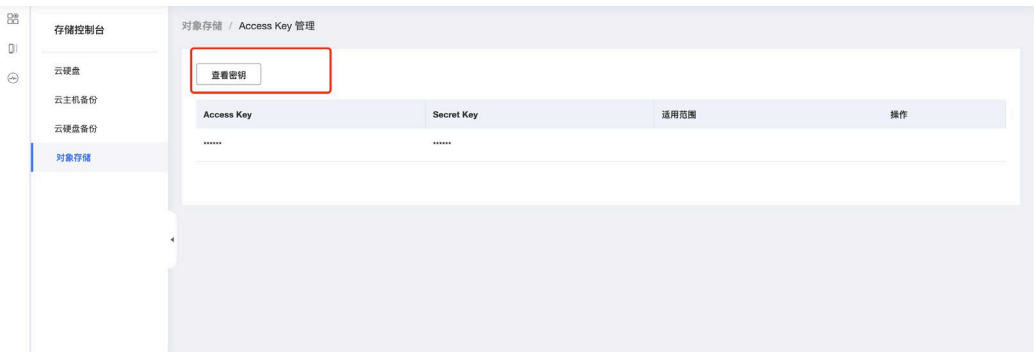
1. 存储桶：在下拉框中选择可用的 ZOS 名称。若当前无存储对象，点击“暂无 ZOS？立即创建”可跳转至存储控制台创建对象存储 创建桶。创建完成后，返回此页面，点击 ZOS 选项旁的“刷新”按钮，下拉列表出现已创建的对象存储桶，选择需要挂载的 ZOS。

2. Endpoint：选择“自动填充 Endpoint”，可进行自动填充；若填充失败，点击蓝字“自动获取失败？手动获取”跳转至网络控制台终端节点页面，可以在搜索框输入关键字过滤列表，找到该资源组对应的 VPC（若不清楚资源组 VPC 名称，可进入“资源”>“资源组”页面，找到对应资源组名称，点击名称进入详情页，虚拟私有云后即是 VPC 名称），从而找到相应的终端节点名称，点击名称进入详情页，复制"hppt://私网域名名称"到云骁相应页面。

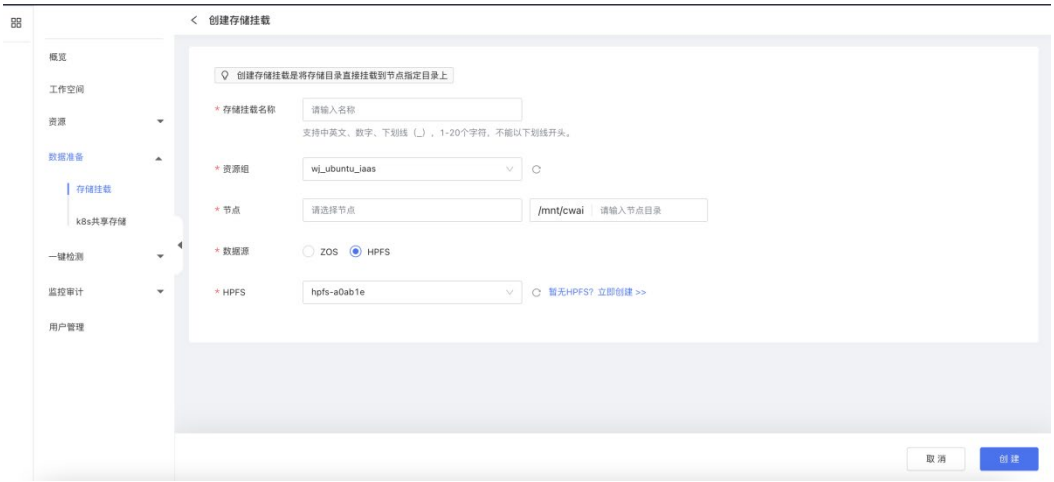




3. Access Key、Secret Key：选择“自动填充 AK/SK”，可进行自动填充；若填充失败，点击蓝字“自动获取失败？手动获取”跳转至存储控制台 Access Key 管理页面，点击“查看密钥”，复制 Access Key、Secret Key 到云骁相应页面。



ii. 选择数据源为 HPFS：下拉选择可用的 HPFS 名称。若当前无可用 HPFS 文件，点击“暂无 HPFS？立即创建”可跳转至 HPFS 控制台，在 HPFS 控制台先完成新建并行文件系统 创建文件系统。创建完成后，返回此页面，点击 HPFS 选项旁的“刷新”按钮，下拉列表出现已创建的文件，选择需要挂载的 HPFS 对象。



3. 页面右下角点击“创建”完成存储挂载的创建流程。

常见问题及说明

- 若没有开通存储相关服务，跳转后可能需要付费开通。
- 创建时提示“用户选择的节点所处在子网未全部配置天翼云内网 DNS 服务器……”/修改标准裸金属子网 DNS 配置后导致的 ZOS 存储挂载异常问题 解决方式

4.3.2. 管理存储挂载

创建存储挂载以后，用户可对存储挂载进行管理，包括挂载、解除挂载、删除等操作。

使用前提

当前用户是主账号。

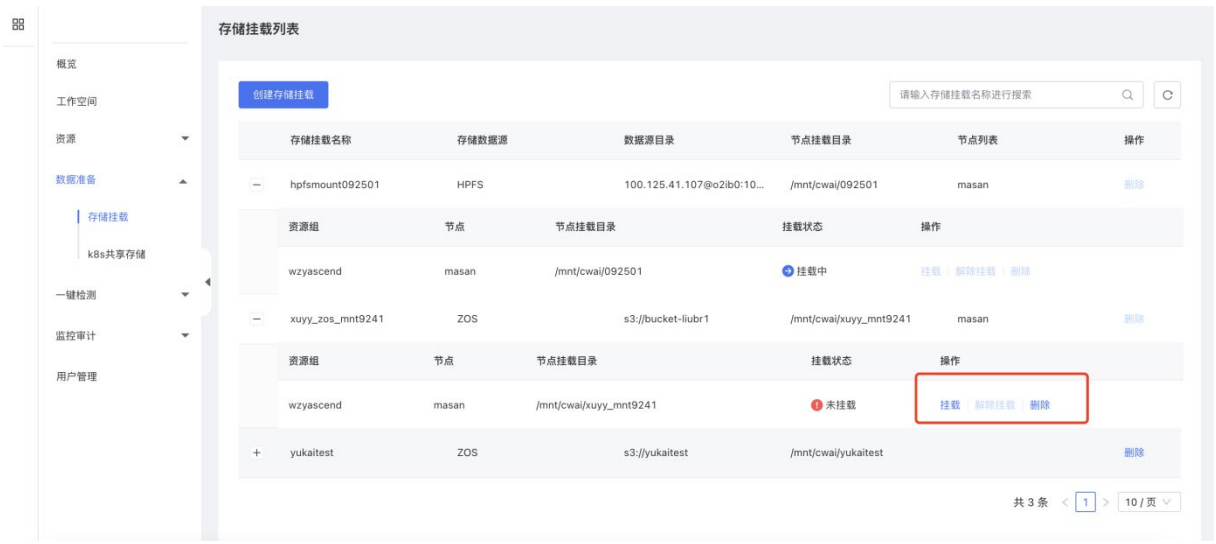
操作说明

查看数据集列表

登录云骁智算控制台，单击左侧菜单栏的菜单项“数据准备”>“存储挂载”，查看存储挂载列表。

查看挂载信息

存储挂载列表可以查看每个存储挂载的详细信息，点击+号展开节点列表，可以查看每个节点下的挂载情况。



对存储挂载进行操作

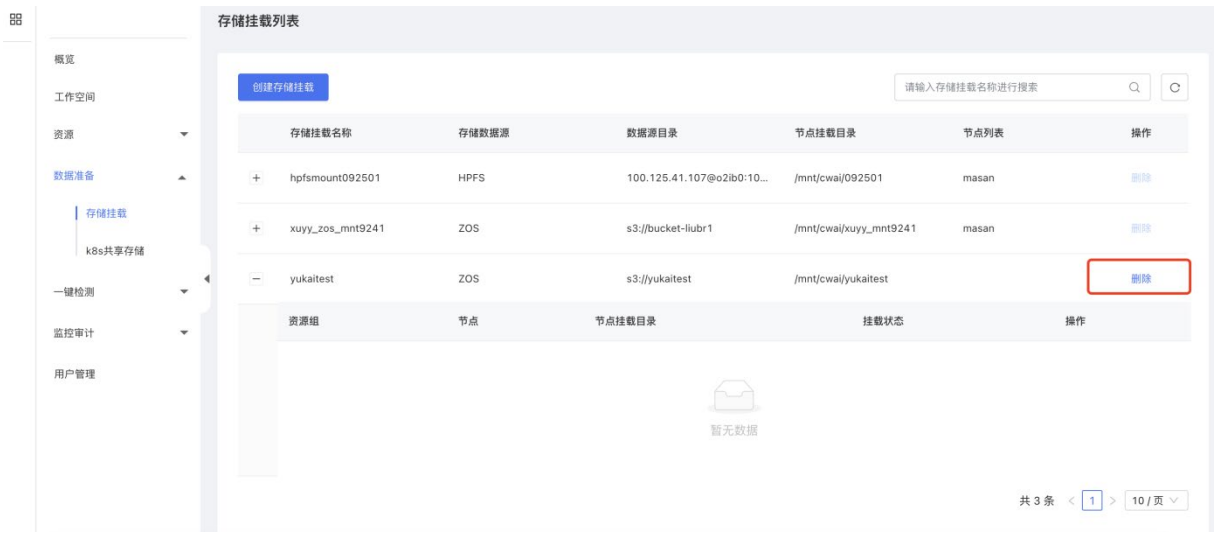
1. 挂载：对于状态为未挂载的节点，可以点击“挂载”进行挂载。

2. 解除挂载：对于状态为已挂载的节点，可以点击“解除挂载”，取消单个节点的挂载；

3. 删除：对于状态为未挂载的节点，可以点击“删除”，删除本次挂载；对于状态为已挂载的节点，需先解除挂载，后可点击“删除”进行删除；

删除存储挂载

解除所有节点挂载之后，可以点击“删除”删除本条存储挂载。



4.3.3.创建 k8s 共享存储

k8s 共享存储仅针对云骁扩展资源组使用。通过 k8s 共享存储可以将 ZOS 或 HPFS 实例挂载到内 k8s 集群内部，后续在工作空间内可以进一步将 k8s 共享存储挂载到任务中进行数据读写操作。

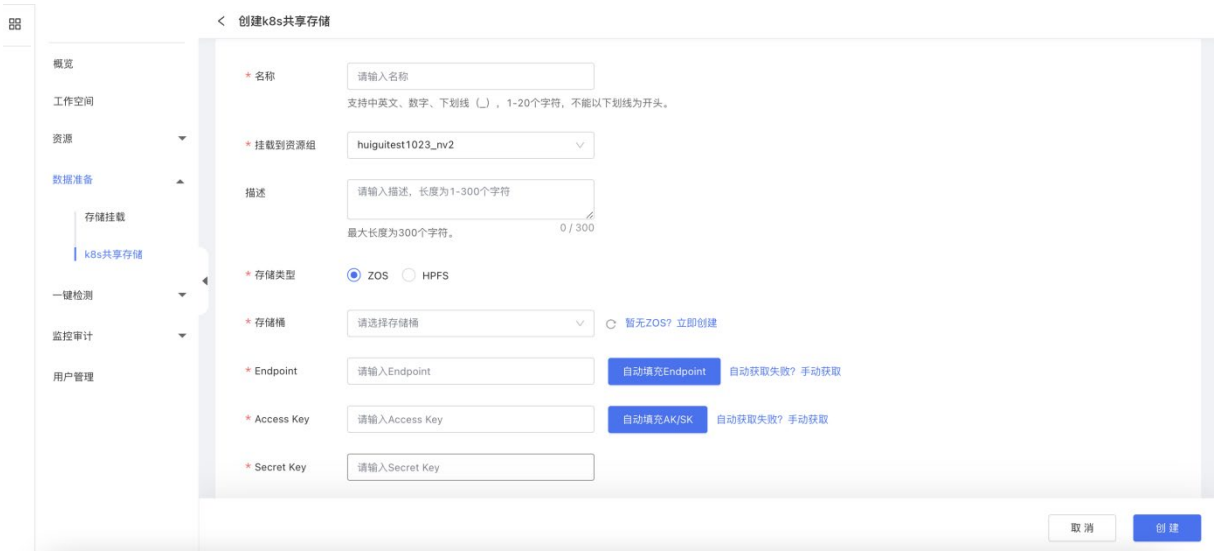
使用前提

当前用户是主账号。

操作步骤

1. 登录云骁智算控制台，单击左侧菜单栏的菜单项“数据准备” > “k8s 共享存储”，点击页面“创建 k8s 共享存储”按钮。

2. 进入“创建 k8s 共享存储”页面，根据要求配置相关信息。



a. 名称：填写将要创建的存储挂载的名称，支持中英文、数字、下划线（_），1-20 个字符，且不能以下划线开头。

b. 挂载到资源组：下拉选择将要挂载的资源组，这里的资源组只能选择扩展资源组。

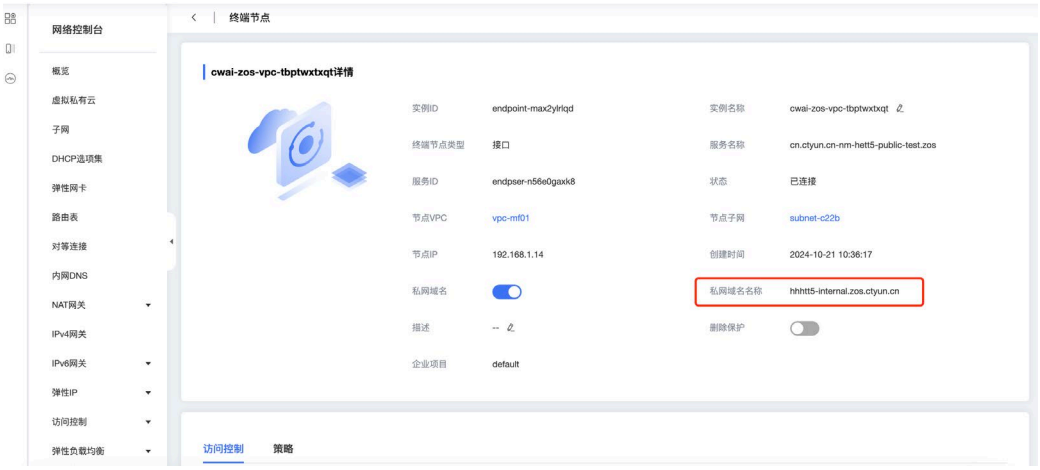
c. 描述：输入该存储挂载描述，最大长度为 300 个字符。

d. 存储类型：选择存储类型，支持 ZOS 与 HPFS 两种数据类型

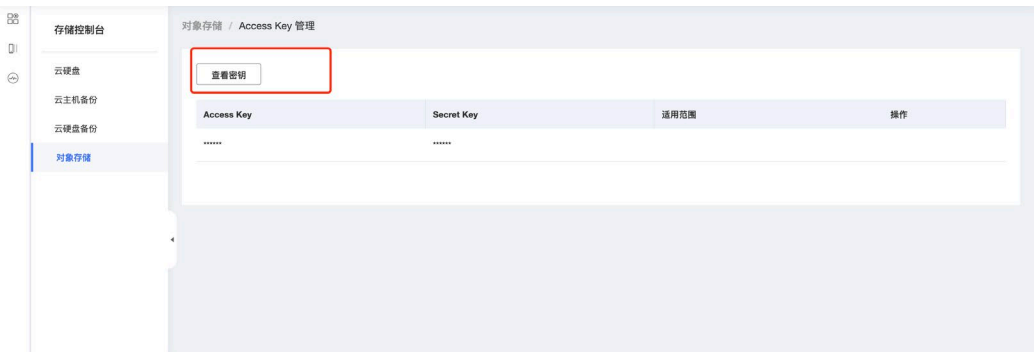
i. 选择存储类型为 ZOS：

1. 存储桶：在下拉框中选择可用的 ZOS 名称。若当前无存储对象，点击“暂无 ZOS？立即创建”可跳转至存储控制台创建对象和桶 创建桶。创建完成后，返回此页面，点击 ZOS 选项旁的“刷新”按钮，下拉列表出现已创建的对象存储桶，选择需要挂载的 ZOS。

2. Endpoint：选择“自动填充 Endpoint”，可进行自动填充；若填充失败，点击蓝字“自动获取失败？手动获取”跳转至网络控制台终端节点页面，可以在搜索框输入关键字过滤列表，找到该资源组对应的 VPC（若不清楚资源组 VPC 名称，可进入“资源”>“资源组”页面，找到对应资源组名称，点击名称进入详情页，虚拟私有云后即是 VPC 名称），从而找到相应的终端节点名称，点击名称进入详情页，复制"hppt://私网域名名称"到云骁相应页面。



3. Access Key、Secret Key：选择“自动填充 AK/SK”，可进行自动填充；若填充失败，点击蓝字“自动获取失败？手动获取”跳转至存储控制台 Access Key 管理页面，点击“查看密钥”，复制 Access Key、Secret Key 到云骁相应页面。



ii. 选择存储类型为 HPFS：

1. 在“存储实例“处，下拉选择可用的 HPFS 名称，若当前无可用 HPFS 文件，点击“暂无 HPFS？立即创建”可跳转至 HPFS 控制台，在 HPFS 控制台先完成新建并行文件系统 创建文件系统。创建完成后，返回此页面，点击 HPFS 选项旁的“刷新”按钮，下拉列表出现已创建的文件，选择需要挂载的 HPFS 对象。

3. 点击“创建”，完成共享存储创建。

4.3.4. 管理 k8s 共享存储

创建 k8s 共享存储以后，用户可对 k8s 共享存储进行管理，包括查看、HPFS 子目录管理、编辑、删除等操作。

使用前提

当前用户是主账号。

操作说明

查看 k8s 共享存储列表

登录云骁智算控制台，单击左侧菜单栏的菜单项“数据准备”>“k8s 共享存储”，查看已创建的 k8s 共享存储列表。

查看 k8s 共享存储详情

在 k8s 共享存储列表页，点击想要查看的 k8s 共享存储名称，可以看到 k8s 共享存储的详细信息。

HPFS 子目录管理

在 HPFS 类型的存储详情页面可进行共享存储的子目录管理。

1. 添加子目录：点击“添加”按钮，填写子目录名称、授权用户，只有被授权的用户才能在工作空间看见此子目录。

2. 删除子目录：对于已创建的子目录，可在操作列点击“删除”操作，删除子目录。

说明：

1. 该页面无法自动获取 HPFS 实例上已创建的目录列表，需要用户通过手动创建的方式手动同步 HPFS 目录，以便在云骁平台上分配和使用。
2. 在该页面创建 HPFS 子目录并不意味着在 HPFS 实例上同步创建，创建的目录仅是在云骁平台内部进行子目录的可视化管理，直到该目录被挂载到任务实例中，并实际执行读写操作的时候，才会对不存在目录进行创建。

编辑 k8s 共享存储

在 k8s 共享存储列表页，找到想要编辑的 k8s 共享存储名称，在操作列点击“编辑”进入编辑页面，支持更改名称与描述。

删除 k8s 共享存储

如果不需要某个 k8s 共享存储，可在操作栏单击“删除”按钮删除。

4.4.监控

云骁智算平台为用户提供资源监控（资源组监控、节点监控）、HPFS 监控、RoCE 监控、任务监控，多种维度查看监控指标的变化情况。

4.4.1.资源监控

资源监控提供了资源组和节点级别的监控能力，支持查看 CPU 和内存、网络、GPU/NPU、磁盘等资源利用情况。

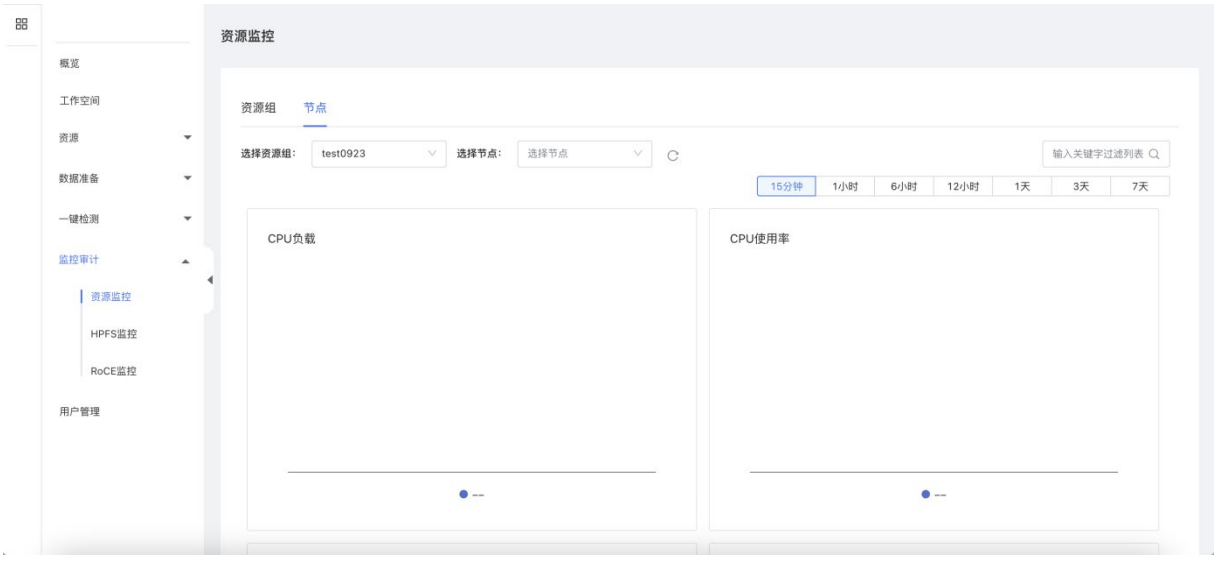
使用前提

当前用户是主账号。

操作说明

点击“监控审计”>“资源监控”菜单，进入资源监控页面，支持查看资源组和节点两种维度的资源监控指标，支持模糊搜索。

在选择资源组与节点时，仅支持查询状态为“运行中”的资源组与“已绑定-正常”的节点。对于资源组维度，图像展示所有卡的聚合值；对于节点维度，图像展示每张卡的指标值。



指标说明

表 1: 资源组级别的指标

类别	指标	单位	说明
CPU 与内存	CPU 负载	数值	资源组所有节点的 CPU 负载的平均值
	CPU 使用率	%	资源组所有节点的 CPU 使用率的平均值
	内存使用率	%	资源组所有节点的内存使用率的平均值
网络	网络吞吐	Kbps	资源组所有节点的网络吞吐的平均值
GPU/NPU	GPU/NPU 使用率	%	资源组所有卡的使用率的平均值
	GPU/NPU 显存使用率	%	资源组所有卡的显存使用率的平均值
	GPU/NPU 显存使用量	GB	资源组所有卡的显存使用量的总量
	GPU/NPU 最高温度	°C	资源组所有卡的温度的最大值
	GPU/NPU 最大功耗	W	资源组所有卡的功耗的最大值

表 2: 节点级别的指标

类别	指标	单位	说明
CPU 与内存	CPU 负载	数值	节点的 CPU 负载（1 分钟）
	CPU 使用率	%	节点的 CPU 使用率

	内存使用率	%	节点的内存使用率
网络	网络吞吐	Kbps	节点的网络吞吐，包括网络的接收速率和发送速率
GPU/NPU	GPU/NPU 使用率	%	节点每张卡的使用率
	GPU/NPU 显存使用率	%	节点每张卡的显存使用率
	GPU/NPU 显存使用量	GB	节点每张卡的显存使用量
	GPU/NPU 温度	°C	节点每张卡的温度
	GPU/NPU 功耗	W	节点每张卡的功耗
	NPU 芯片健康状态	数值	节点每张卡的 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。
磁盘	本地磁盘使用率	%	节点的本地磁盘使用率
	本地磁盘读速率	KB/s	节点的本地磁盘读速率
	本地磁盘写速率	KB/s	节点的本地磁盘写速率

4.4.2. HPFS 监控

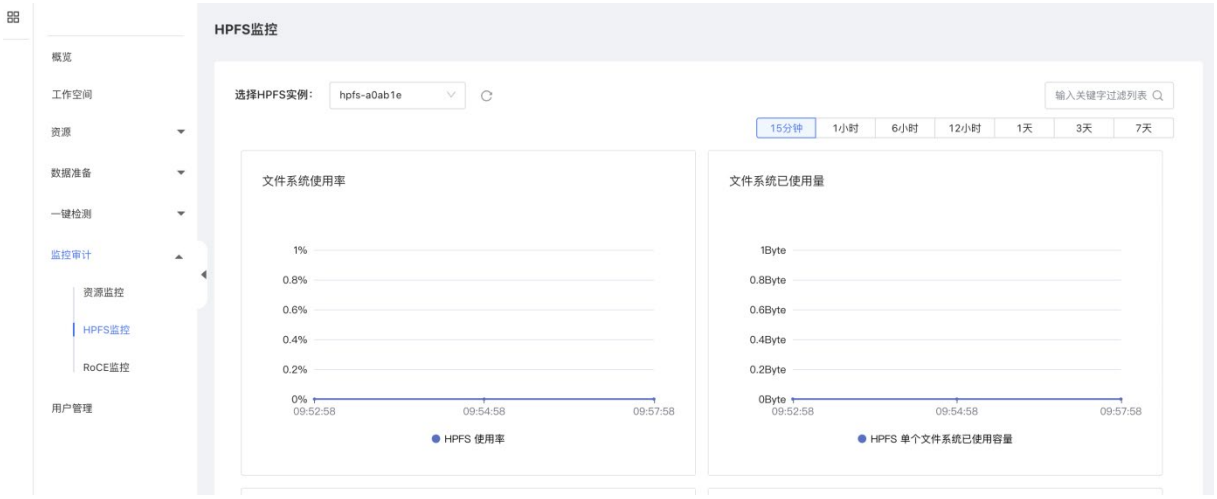
HPFS 监控提供了 HPFS 实例级别的监控能力，支持查看文件系统的资源利用情况。

使用前提

当前用户是主账号。

操作说明

点击“监控审计”>“HPFS 监控”菜单，进入 HPFS 监控页面，支持查看 HPFS 实例级别的文件系统监控指标。



指标说明

表 1: HPFS 相关指标

指标	单位	说明
文件系统使用率	%	HPFS 容量使用率
文件系统已使用量	MB	HPFS 容量使用量
文件系统写 IOPS	次	HPFS 单个文件系统写 IOPS
文件系统读 IOPS	次	HPFS 单个文件系统读 IOPS
文件系统写带宽	bps	HPFS 单个文件系统写带宽
文件系统读带宽	bps	HPFS 单个文件系统读带宽

4.4.3. RoCE 监控

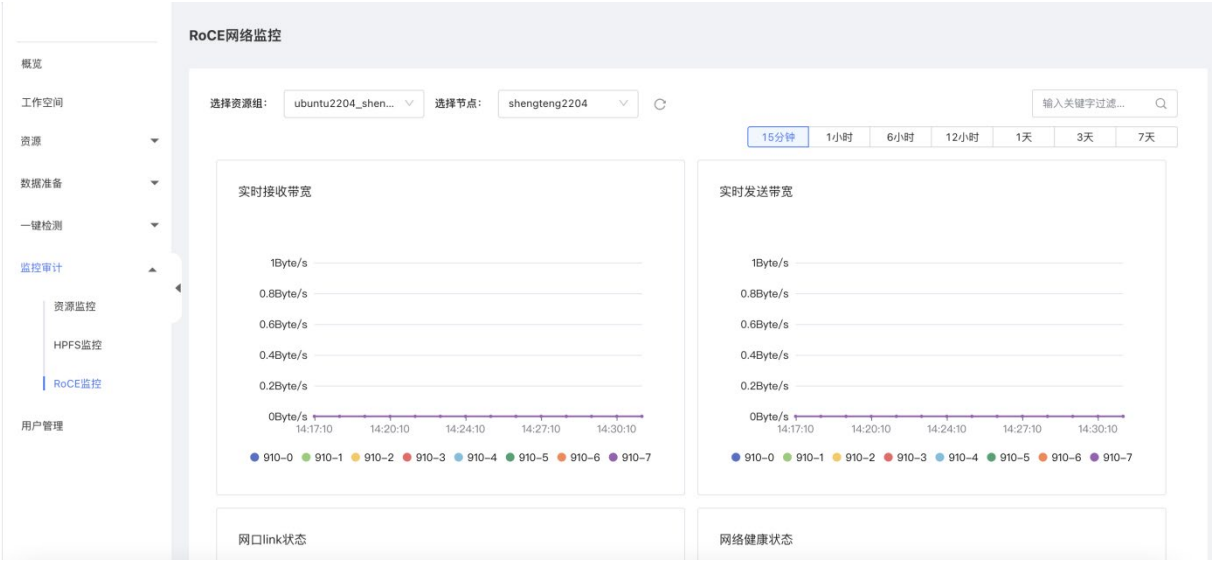
RoCE 监控提供了节点级别的监控能力，支持查看每张 RoCE 网卡相关的资源利用情况。

使用前提

当前用户是主账号。

操作说明

点击“监控审计”>“RoCE 监控”菜单，进入 RoCE 监控页面，支持查看节点级的每张 RoCE 网卡的监控指标。



指标说明

表 1: RoCE 相关指标

指标	单位	说明
实时接收带宽	MB/S	节点每张 RoCE 网卡的实时接收带宽

实时发送带宽	MB/S	节点每张 RoCE 网卡的实时发送带宽
网口 link 状态	数值	节点每张 RoCE 网卡的网口 link 状态 取值范围：{0，1} 1：表示在过去一段时间间隔内处于 link up 状态。 0：表示在过去一段时间间隔内网口出现了 link down 状态。
网络健康状态	数值	节点每张 RoCE 网卡的网络健康状态 取值范围：{0，1} 1：表示在过去一段时间间隔内网络处于健康（可以联通）状态。 0：表示在过去一段时间间隔内网络出现了不健康（无法联通）状态。
RoCE 接收总报文数	个	节点每张 RoCE 网卡的 RoCE 接收总报文数
RoCE 发送总报文数	个	节点每张 RoCE 网卡的 RoCE 发送总报文数
RoCE 接收坏包报文数	个	节点每张 RoCE 网卡的 RoCE 接收坏包报文数
RoCE 发送坏包报文数	个	节点每张 RoCE 网卡的 RoCE 发送坏包报文数
RoCE 接收 CNP 报文数	个	节点每张 RoCE 网卡的 RoCE 接收 CNP 报文数
RoCE 发送 CNP 报文数	个	节点每张 RoCE 网卡的 RoCE 发送 CNP 报文数

RoCE 重试报文数	个	节点每张 RoCE 网卡的 RoCE 重试报文次数
调度队列接收的 PFC 帧报文数	个	节点每张 RoCE 网卡的调度队列接收的 PFC 帧报文数
调度队列发送的 PFC 帧报文数	个	节点每张 RoCE 网卡的调度队列发送的 PFC 帧报文数

4.5.一键检测

一键检测提供云骁算力环境的健康检测能力，从服务器健康、RDMA 网络性能、集合通讯库性能等多方面进行检测。用户可以在运行训练任务前先对环境健康进行全面测评，为训练的长稳运行提供保障。在训练任务出现故障或性能问题是也可以结合一键检测工具进行问题排查。

一键检测的主要功能包括：创建检测任务、查看检测历史及检测详情。

4.5.1.服务器检测

服务器检测提供针对资源组的 GPU、NPU 节点的检测能力，主要检测节点的关键软硬件是否安装，关键配置是否开启，参与训练的多节点配置是否一致，配置是否符合用户设定等方面。

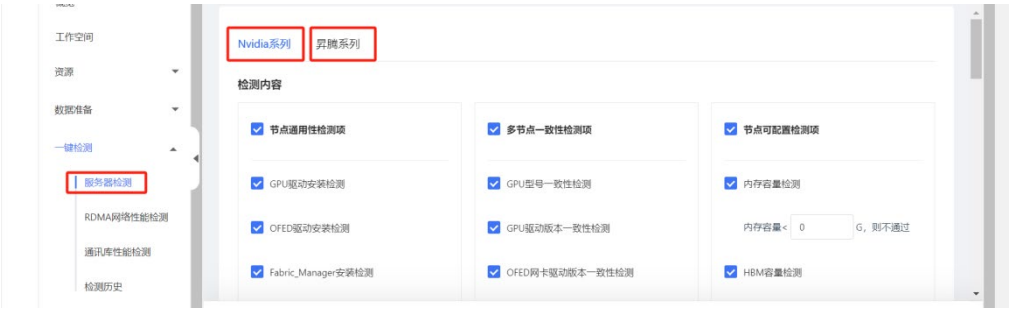
使用前提：

当前用户是主账号。

操作步骤：

1. 选择检测类型：

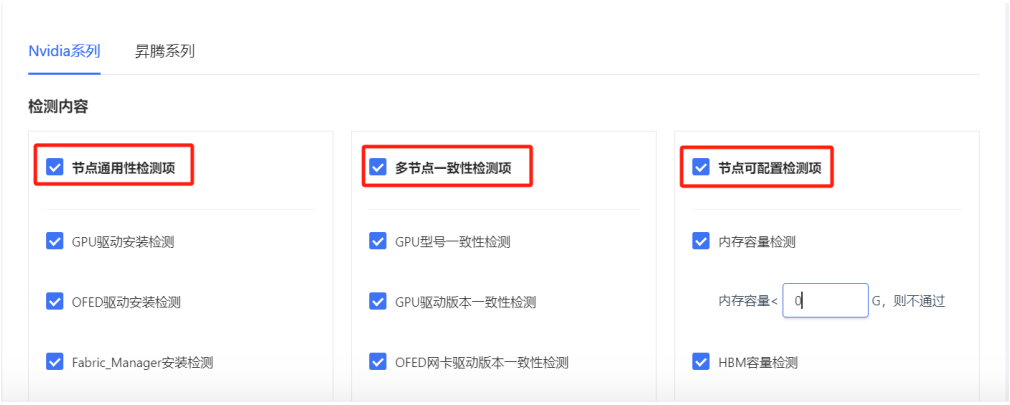
   - 点击【服务器检测】菜单，进入服务器检测任务新建页。
   - 选择“Nvidia 系列”或者"昇腾系列"。



2. 选择检测内容：

   - 节点通用检测项：该项为系统内置检测项，用来判断单个节点的关键软件和配置是否符合预期，用户可以根据自己的业务需求进行检测项的选择，检测结果为“通过”或“不通过”。
   - 多节点一致性检测项：系统内置检测项，用来判断参与训练的多节点关键配置是否一致。主要分为两种场景：
     - 第一种场景：用户选择其中一个节点的配置作为基线，其他节点均和基线节点进行对比，如果结果一致，则检测结果为“通过”，不一致，则检测结果为“不通过”。
     - 第二种场景：用户没有设置基线节点，则将对所有节点的安装配置结果进行统计，将每项检测的所有检测结果详细列出，结果“不涉及”是否通过。

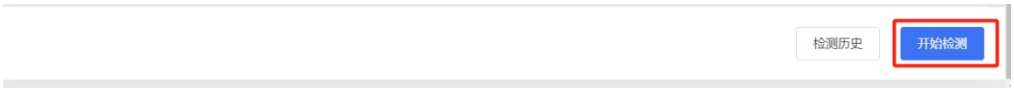
- 节点可配置检测项：系统内置检测项和用户自定义检测参数。针对具体检测项，用户可自行定义检测标准，检测结果为“通过”或“不通过”。



3. 选择检测目标：
   - 资源组：根据选择的系列（Nvidia 或昇腾）列出相关资源组供用户选择（单选）。
   - 节点：左侧选择资源组内单个或多个节点作为目标，将其移动到右侧成为已选节点。
   - 开启基线节点设置：选择一个节点作为多节点一致性检测的基线节点，此项为非必选。如果选择“开启节点基线设置”，则需要在右侧已选节点列表中选择一个节点作为基线节点。
   - 输入节点密码：输入创建该节点时设置的密码。注意：选择多个节点需要保证所有节点的密码一致，节点密码只有一个输入框，如不一致会检测失败。

4. 开始检测：

   ○ 点击“开始测试”，启动检测。也可以点击检测历史查看节点的历史检测报告。



4.5.2. RDMA 网络性能检测

RDMA 网络性能检测提供 IB 或 RoCE 网络中网卡到网卡的带宽和延时检测。

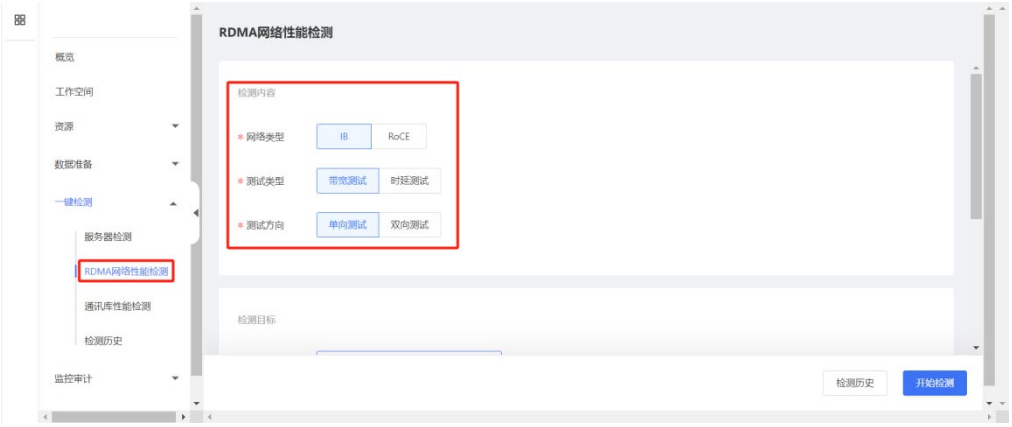
使用前提：

当前用户是主账号。

操作步骤：

1. 选择检测内容：

- 点击【RDMA 网络性能检测】菜单，进入 RDMA 网络性能检测新建页，选择检测内容。
- 网络类型：IB 和 RoCE。
- 测试类型：带宽测试、时延测试。
- 测试方向：单向测试、双向测试。



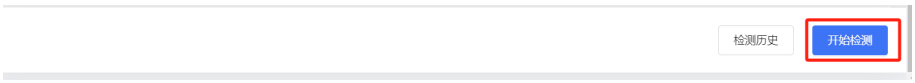
2. 选择检测目标：

- 资源组名称：根据 IB（英伟达）和 RoCE（昇腾）类型筛选出资源组供供用户选择（单选）。
- 服务端网卡：选择一个服务端网卡。
- 客户端网卡：选择一个客服端网卡。
- 节点密码：输入资源组下节点的密码。注意：资源组下各节点密码需要保持一致，该输入框只能输入一个节点密码，不一致会检测失败。

- 节点端口：指定端口供测试命令使用。

3. 开始检测：

- 点击“开始测试”，启动检测，也可以点击检测历史查看节点的历史检测报告。
- 启动之后，进行检测确认。确认之后，跳入检测历史页面进行检测结果查看。



4.5.3 通讯库性能检测

通讯库性能检测对两种典型的集合通讯库，即英伟达系列的 NCCL 和昇腾系列的 HCCL 进行多种通信模型的性能检测，可输出算法带宽，辅助用户判断环境健康。

使用前提：

当前用户是主账号。

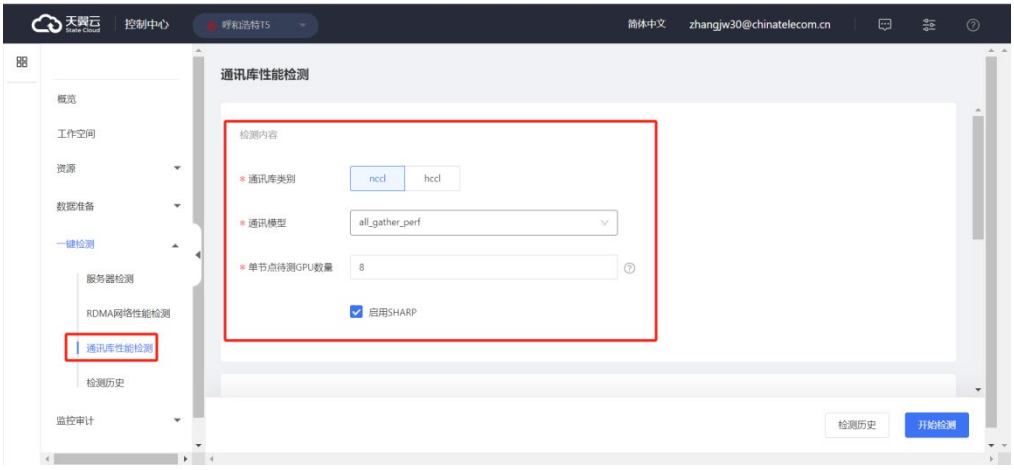
操作步骤：

1. 选择检测内容：

- 点击【通讯库性能检测】菜单，进入通讯库性能检测新建页，选择检测内容。
- 通讯库类别：nccl（英伟达）、hccl（昇腾）。
- 通讯模型：选择相关通讯模型（单选）。
- 单节点待测 GPU 数量：1-8，默认为 8。

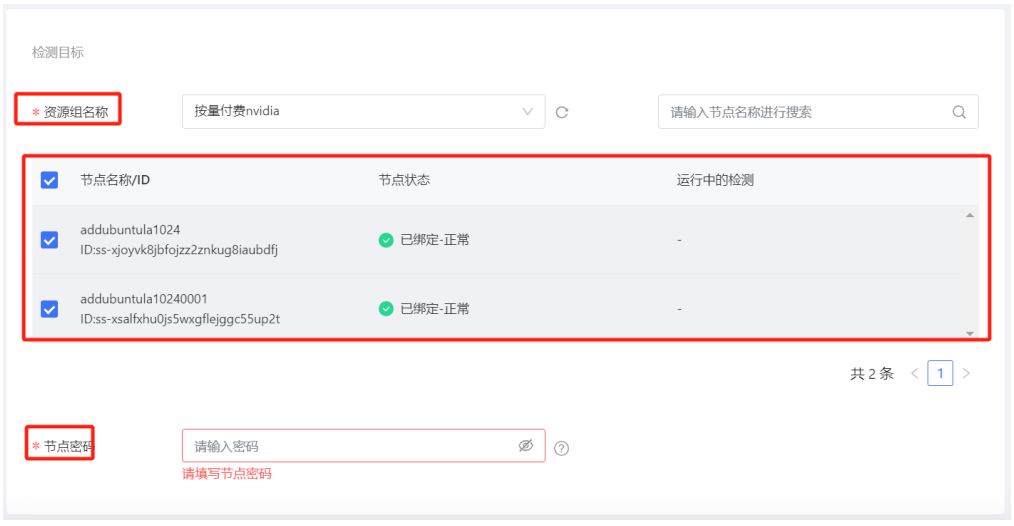
  目前云骁支持的节点规格单节点不会超过 8 卡。
- 启用 SHARP：选择 nccl，默认勾选，选择 hccl，无此选项。

  备注：SHARP 是随 IB 网络一起推出的，可将集合通信运算（如 all-reduce、reduce 和 broadcast）从服务器的计算引擎卸载到网络交换机的插件。通过直接在网络结构中执行归约（求和、平均等），勾选 SHARP 在配套软硬件支持的基础上可改进这些运算和整体应用程序性能。



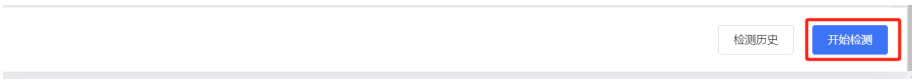
2. 选择检测目标：

- 资源组名称：根据选择的通讯库（nccl 或 hccl）列出相关资源组供用户选择。如选择 nccl，则列出英伟达资源组；选择了 hccl，列出昇腾资源组。
- 选择节点：根据选择的资源组列出资源组下方的节点，对节点进行勾选，可多选。
- 节点密码：输入资源组下节点的密码。*注意：资源组下各节点密码需要保持一致，该输入框只能输入一个节点密码，不一致会检测失败。



3. 开始检测：

- 点击“开始测试”，启动检测，也可以点击检测历史查看节点的历史检测报告。
- 启动之后，进行检测确认。确认之后，跳入检测历史页面进行检测结果查看。



4.5.4 检测历史

使用前提：

当前用户是主账号。

操作步骤：

1. 查看检测历史

- 点击【检测历史】菜单，进入检测历史查看页面。
- 检测类型切换：选择上方导航栏，可以对检测类型经常切换。（检测类类型为：服务器检测、RDMA 网络性能检测、通讯库性能检测）

- 查看报告：点击列表检测项中查看报告按钮，可对该检测项目详情进行查看。
- 重新运行：点击列表检测项中重新运行按钮，会跳入检测项目发起页面，可对检查项进行，发起重新检测。
- 更多：可以对该检测项进行导出报告、终止检测（运作中才可以终止，检测完成之后该按钮无法点击）、删除记录操作。

2. 各类型检测报告详情

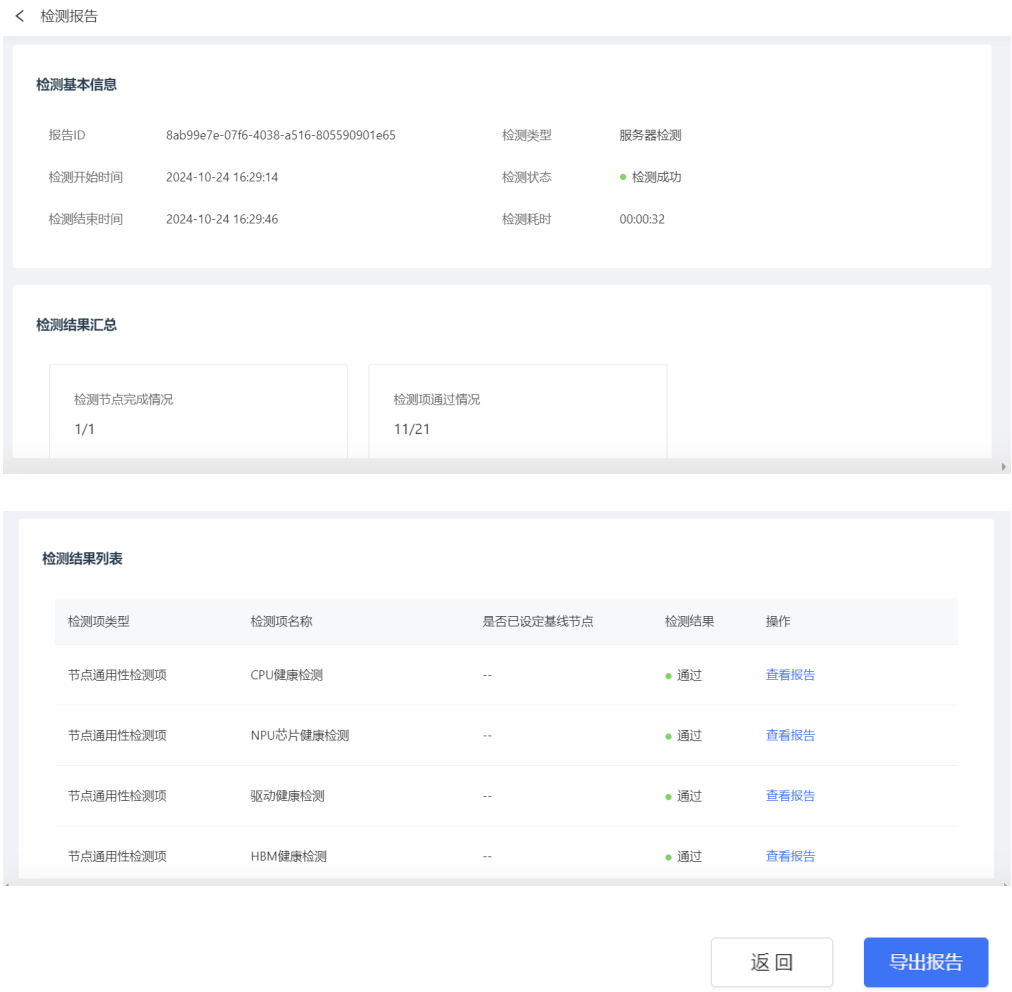
- 服务器检测报告详情

点击服务器检测类型下的项目查看报告，进入服务器检测报告详情，详情主要包括：

检测基本信息：报告 ID、检测类型、检测开始时间、检测结束时间、检测状态、检测耗时。

检测结果汇总：检测结果完成情况、检测项通过情况。

检测结果列表：列出各项检测结果。



点击"查看报告"， 可查看检测详情，包括： 检测项名称、检测方法、判断标准、处理建议、详细结果等信息。

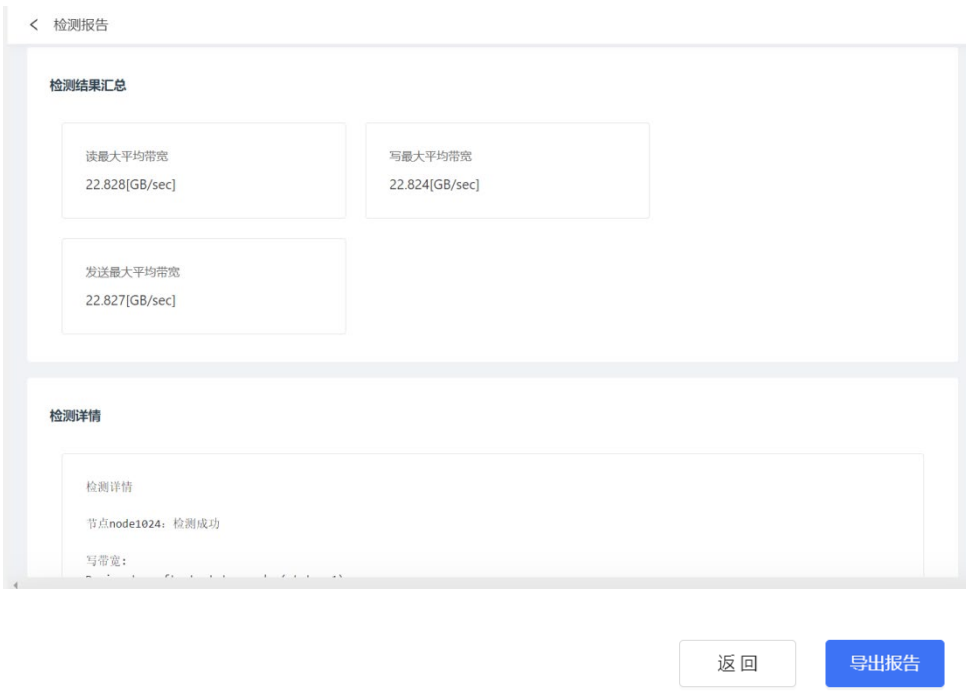
○ RDMA 网络性能检测报告详情

点击 RDMA 网络性能检测类型下的项目查看报告，进入 RDMA

网络性能检测报告详情，详情主要包括：

检测基本信息	报告 ID、检测类型、检测开始时间、检测结束时间、检测状态、检测耗时。
网络性能检测信息	网络类型、客户端网卡名称、服务端网卡名称、客户端网卡节点、服务端网卡节点、测试方向、测试类型。
检测结果汇总	读最大平均带宽、写最大平均带宽、发送最大平均带宽。

检测详情：查看检测报告详情。



○通讯库性能检测报告详情

点击通讯库性能检测类型下的项目查看报告，进入通讯库性能检测报告详情，详情主要包括：

检测基本信息	报告 ID、检测类型、检测开始时间、检测结束时间、检测状态、检测耗时。
通讯库检测信息	通讯库类型、通讯模型、单节点测试 GPU 数量、是否启动 SHARP。
检测结果汇总	最大算法带宽。

检测详情：查看检测报告详情。

4.6.工作空间

工作空间是由主账号（管理员）创建的项目空间，使得工作空间成员（开发人员）可以分享 AI 资产（数据集、镜像、训练任务等），进行协作。

工作空间创建完成后，工作空间管理员可以对其进行管理，包括编辑工作空间描述，给工作空间添加同一个资源组中的队列，给工作空间添加成员（子账号），以及给成员授权某个队列的使用权限。

4.6.1.工作空间管理

- 工作空间的创建

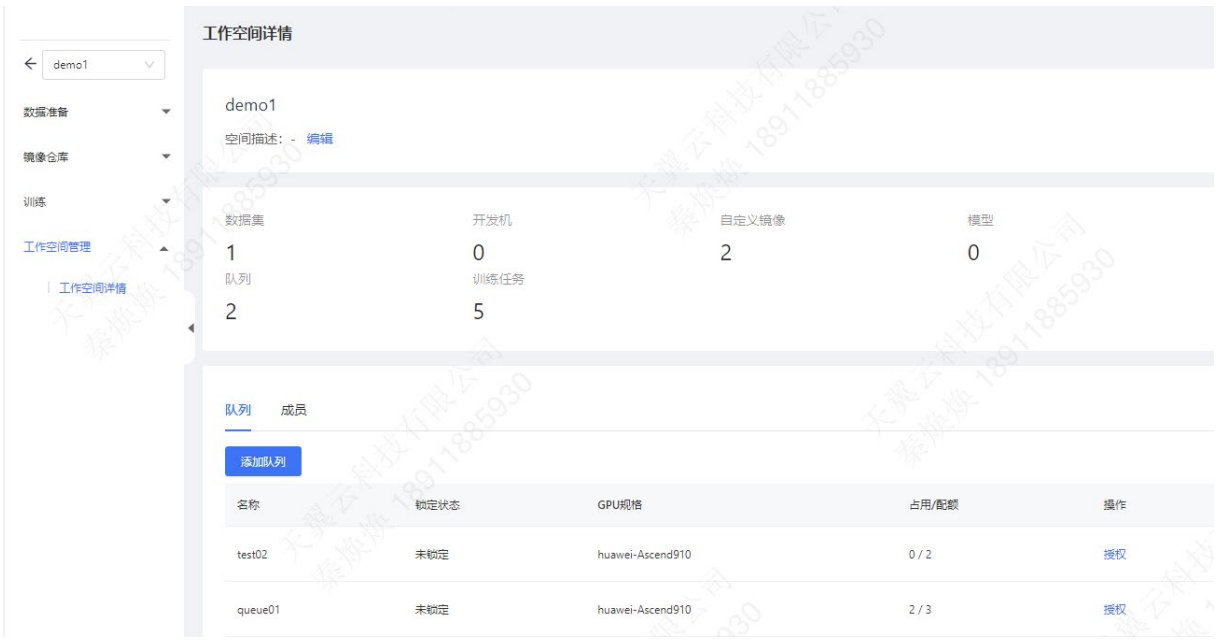
1、 当前租户已创建至少一个队列

2、 进入工作空间创建页，填写工作空间名称和描述，添加工作空间成员

3、 关联资源：在关联队列列表中选择某个之前已经创建的队列，点击【创建】完成资源关联并创建工作空间。

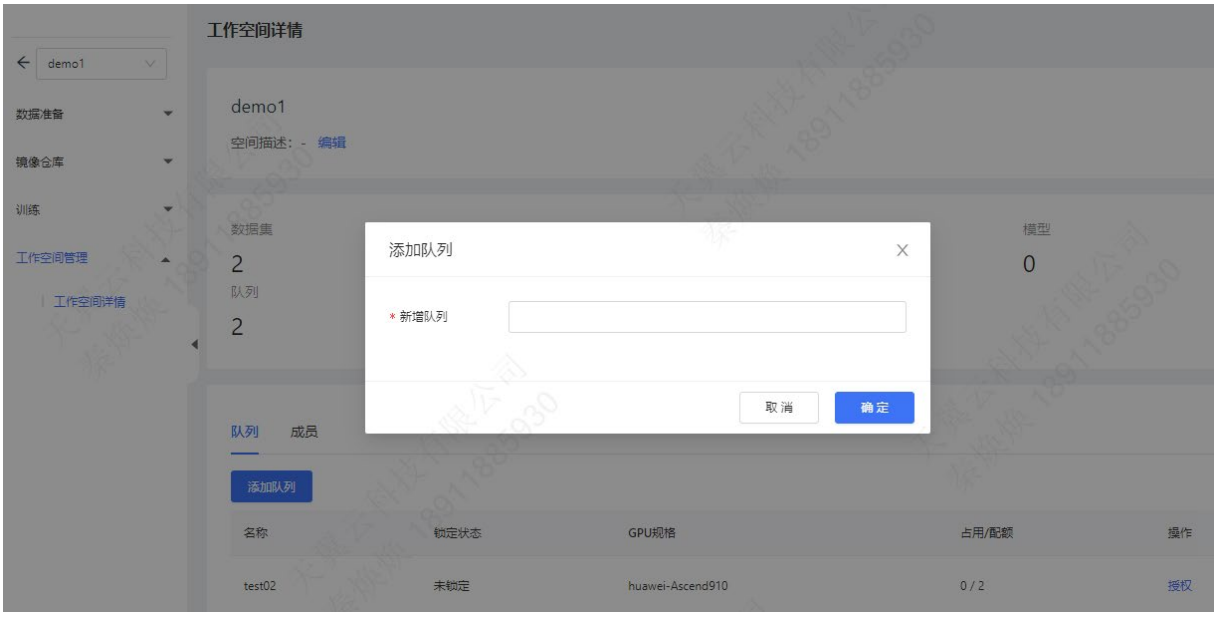
- 工作空间的管理

1、进入工作空间：点击工作空间列表页中某个工作空间的名称，可以进入该工作空间



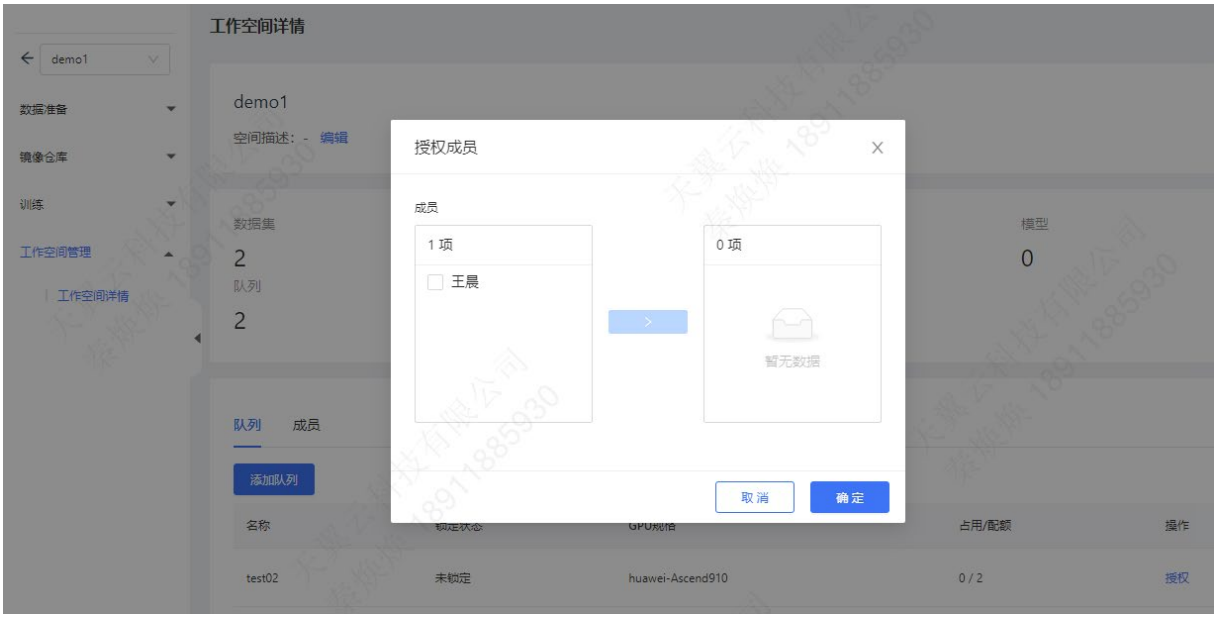
2、修改工作空间描述：点击工作空间管理-工作空间详情页的【编辑】按钮，修改工作空间描述。

3、添加队列：点击【添加队列】按钮，可以添加队列以便在该工作空间中可以使用新增的队列，只能添加同一个资源组的队列



4、添加成员：点击【成员】tab 页，点击【添加成员】，可以添加工作空间成员

5、授权工作空间成员使用特定队列：点击【队列】tab 页队列列表中某个队列的【授权】操作，弹出授权成员对话框，可以选择某些工作空间成员，点击【确定】，将当前队列授权给选择的工作空间成员使用



4.6.2.数据集

对于主账号在“数据准备”>“k8s 共享存储中”已创建成功的 k8s 共享存储，数据集功能可以对其进行进一步的划分。通过数据集，用户可在 k8s 共享存储的基础上，为特定的数据、模型、算法代码指定特定的目录并为之命名，起到标签管理的作用。同时对于 ZOS 数据集，支持在英伟达系列集群开启数据加速功能。

使用前提

在数据准备模块已成功创建 k8s 共享存储。

操作说明

1. 创建数据集

1. 进入数据集列表，点击“创建数据集”。

2. 进入“创建数据集”页面，按要求配置相关信息。

a. 名称：填写将要创建的数据集名称，支持中英文、数字、下划线（_），1-20个字符，且不能以下划线开头。

b. 描述：填写数据集对描述。

c. 可见范围：选择可见范围，

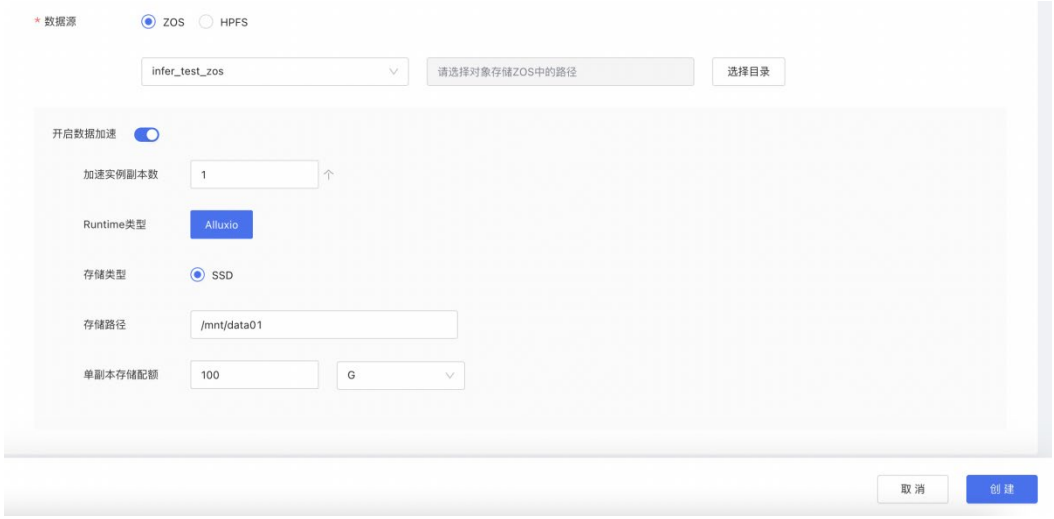
   i. “仅自己可见”：对主账号与本账号可见；

   ii. “工作空间公开可见”：工作空间内对成员都可见。

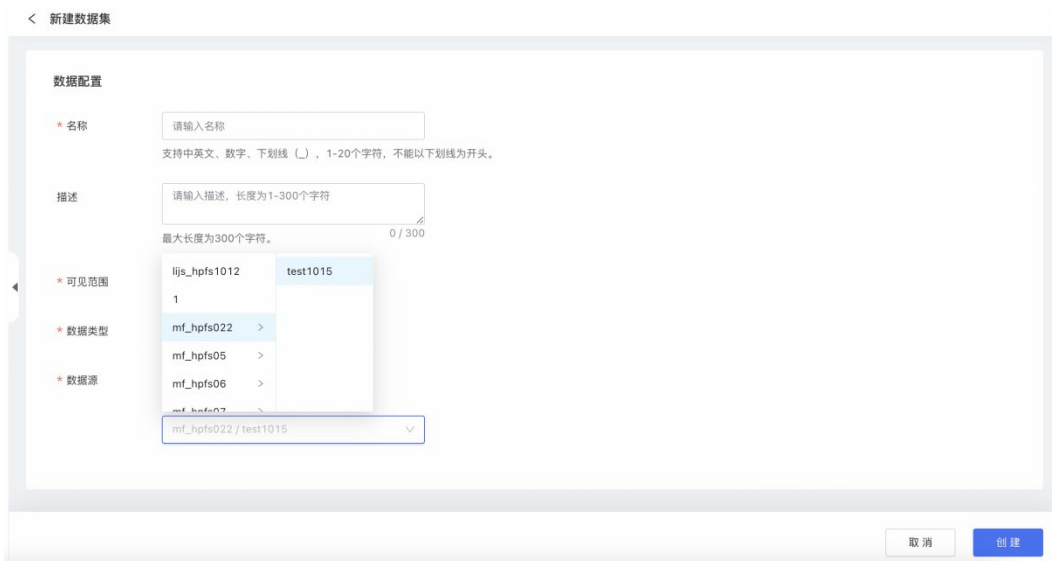
d. 数据类型：选择数据类型，支持文本与图像；

e. 数据源：选择数据来源，支持 ZOS 与 HPFS 两种数据来源，

   i. 选择 ZOS：选择已挂载的 ZOS 存储桶，点击“选择目录”，可选择桶下的子目录；对于 ZOS 存储，英伟达资源组支持开启数据加速。



ii. 选择 HPFS：选择已挂载的 HPFS 共享存储实例，支持选择整个 HPFS 实例，以及 HPFS 实例下已创建的子目录。



2. 管理数据集

- 查看：创建完成后，返回数据集列表，可以查看已创建的数据集。
- 编辑：点击“编辑”，可以对数据集对名称、描述、可见范围进行编辑，点击更新保存更改。
- 删除：点击“删除”，可删除对应的数据集。

4.6.3.镜像仓库

云骁智算平台中镜像用于提供开发所需的环境。镜像仓库中提供预置镜像和自定义镜像两种能力。预置镜像即平台预先设置的完整镜像，可以在创建自定义训练时直接使用。自定义镜像即可上传本地自有镜像，上传完成后可在创建自定义训练任务时选择并使用。

4.6.3.1 预置镜像

预置镜像即平台预先设置的完整镜像，可以在创建自定义训练时直接使用。

操作流程

1. 登录云骁智算控制台。
2. 进入对应工作空间。
3. 在左侧导航栏中，选择“镜像仓库>镜像列表”进入“预置镜像”页面。
4. 在“预置镜像”页面中可查看预置镜像。

预置镜像列表

镜像名称	机器类型	操作系统	框架	模型	训练任务类型

chatglm2-6b:DeepSpeed0.10.2-Pytorch2.0.1-cuda11.6-Ubuntu18.04-x64	英伟达	Ubuntu18.04	Pytorch2.0.1	ChatGLM2-6B	ptuning v2 微调
chatglm2-6b:DeepSpeed0.10.2-Pytorch2.1.0-cuda12.2-Ubuntu18.04-x64	英伟达	Ubuntu18.04	Pytorch2.1.0	ChatGLM2-6B	ptuning v2 微调
chatglm2-6b:v2-DeepSpeed0.9.2-Pytorch1.11.0-cann7.0-Ubuntu22.04-arm64	昇腾	Ubuntu22.04	Pytorch1.11.0	ChatGLM2-6B	ptuning v2 微调

查看镜像详情

在预置镜像的列表页单击对应镜像行的“+”可查看预置镜像的操作系统、镜像类型、创建时间、镜像大小、描述、构建历史信息。

4.6.3.2 自定义镜像

自定义镜像即可上传本地自有镜像，上传完成后可在创建自定义训练任务时选择并使用。上传的本地自有镜像仅该工作空间内可见和使用。

前提条件

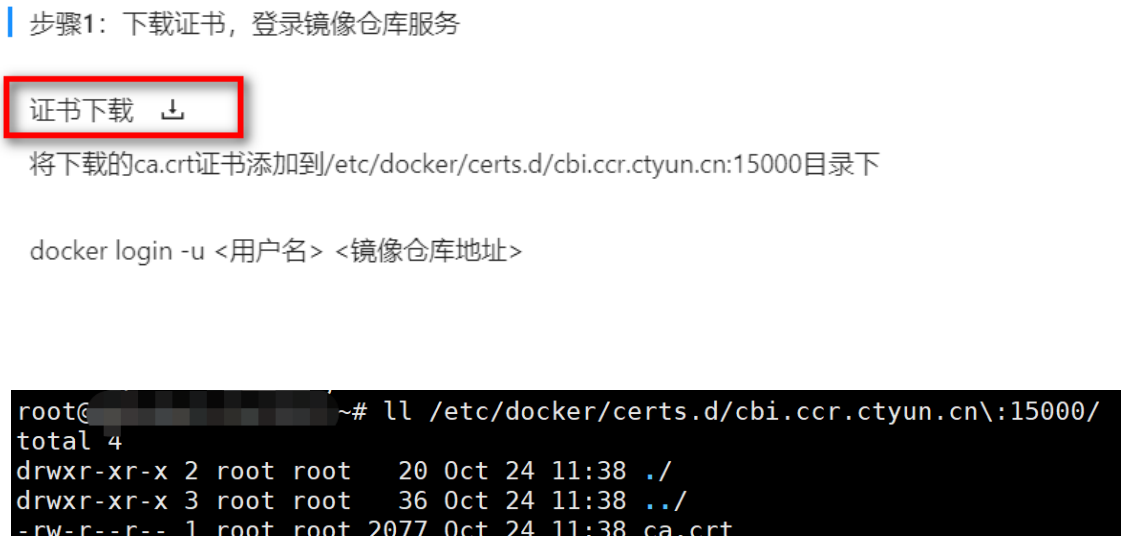
自定义镜像的仓库地址为内网地址，无法直接从公网访问，需先设置镜像仓库的访问网络，支持使用云主机、裸金属或云专线的方式进行网络配置，可结合实际情况进行选择。配置方式详见最佳实践-如何上传镜像到云骁智算的私有镜像仓库。

操作流程

1. 登录云骁智算控制台。
2. 进入对应工作空间。
3. 在左侧导航栏中，选择“镜像仓库>镜像列表”进入“自定义镜像”页面。
4. 在“自定义镜像”页面中点击“上传镜像”按钮进入上传镜像页面。

上传镜像步骤

- 步骤 1：下载证书，并将下载的 ca.crt 证书添加对应目录下。



- 步骤 2：登录 harbor，获取“上传镜像”页面的用户名和密码进行登录。

```
root@[redacted] ~# docker login -u user-testubuntu1024 cbi.ccr.ctyun.cn:15000
Password:
WARNING! Your password will be stored unencrypted in /root/.docker/config.json.
Configure a credential helper to remove this warning. See
https://docs.docker.com/engine/reference/commandline/login/#credentials-store

Login Succeeded
```

- 步骤 3：重新命名本地镜像

  docker tag <本地镜像 ID> <镜像仓库地址>/<镜像仓库名称>/<镜像名称>:<镜像版本号>

  示例：

```
docker tag testimage:latest cbi.ccr.ctyun.cn:15000/project-testubuntu1024/imagename:latest
```

- 步骤 4：推送至镜像仓库

  docker push <镜像仓库地址>/<镜像仓库名称>/<镜像名称>:<镜像版本号>

  示例：

```
docker push cbi.ccr.ctyun.cn:15000/project-testubuntu1024/imagename:latest
```

- 步骤 5：在“镜像仓库列表>自定义镜像”查看上传的自定义镜像。

删除镜像

单击镜像列表右侧操作栏“删除”操作可删除上传的自有镜像。

4.6.4.训练

4.6.4.1.创建自定义训练任务

模型训练过程需要不断迭代和优化参数设置，寻找最优的模型结构和权重。云骁智算

训练模块支持创建自定义训练、管理自定义训练和查看训练详情功能，以更方便地寻找到最优的结果。

前提条件

- 训练任务运行需要消耗资源，请确保账户内资源未被冻结（未欠费）。
- 创建自定训练任务前，请确定该工作空间关联的队列可用。
- 用于训练的数据、模型等已全部上传至存储。具体上传方法请参见上传数据到 ZOS 存储和上传数据到 HPFS 存储。
- 如需保存训练输出数据需建立单独的文件夹用于训练日志保存。

操作步骤

1. 进入新建训练任务页面。
2. 设置自定义训练任务参数：填写训练任务的基本信息、资源信息和环境信息等。
3. 保存并运行自定义训练任务。

进入新建训练任务页面

1. 登录云骁智算控制台。
2. 进入对应工作空间。
3. 在左侧导航栏中，选择“训练>自定义训练”进入训练任务列表。
4. 单击“新建训练任务”，进入“新建训练任务”页面。

设置自定义训练任务参数

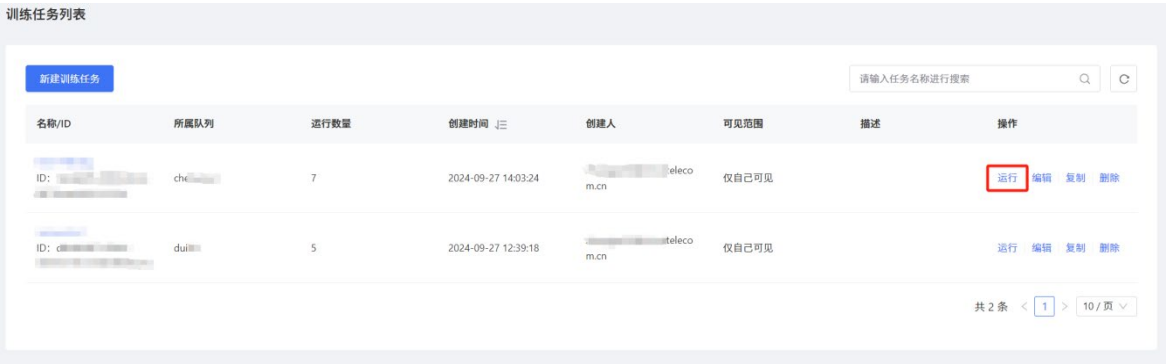
训练过程中需要从存储中获取输入数据和算法进行模型训练，训练结果也支持存储至存储桶中。新建训练任务时需设置的基本信息、资源信息和环境信息参数如下：

参数名称	参数说明
任务名称	必填，训练任务的名称。 支持 1-20 个字符，可以包含中英文、数字、下划线（_），不能以下划线为开头。
所属队列	必选，选择运行训练任务的队列。创建和管理队列参见队列。
镜像来源	必选，选择预置镜像或已上传自定义镜像中的镜像名称。详见镜像仓库。
训练框架	必选，选择预置框架，目前支持 PyTorch 和 TensorFlow 训练框架。 ● 支持 PyTorch 的 DDP 模式和 DDP_deepspeed 模式。 ● 支持 TensorFlow 的 PS 模式。
存储	添加存储路径，训练任务启动时，系统将自动获取路径中的数据和算法到训练运行容器中。 最多添加 10 个存储挂载路径。详见数据准备和数据集。
可见范围	选择哪些账号可见该训练任务。 ● 仅自己可见：仅任务创建人有权查看该任务。 ● 工作空间内公开可见：该工作空间内所有账号均可以查看该任务。
启动命令	必填，指定代码的执行命令。 训练命令必须有程序启动指令，例如：/bin/bash -c；python -e。 支持一次输入多条命令，多条命令需以换行符分隔。

环境变量	将被注入到训练容器中的环境变量。 可配置多个。 说明：为保证数据安全，请勿输入敏感信息，例如明文密码。
资源配置	必填，配置训练任务可用的资源。 ● 应用 PyTorch 框架时需要配置 Worker 节点资源。 ● 应用 TensorFlow 框架时需要配置 Worker 节点、PS 节点、Chief 和 Evaluator 的资源。 说明：资源配置内节点规格内容包含 GPU 卡类型、显存、GPU 卡数、CPU 核心数、内存信息。例如，“Ascend910B-Full-Mesh-64G 2 卡 48C 384G”中，Ascend910B-Full-Mesh 表示 GPU 卡型号、64G 表示单张 GPU 卡的显存为 64G、2 卡表示 GPU 卡数、48C 表示 CPU 核心数、384G 表示内存大小。
Tensorboard 日志路径	选择是否采集 Tensorboard 日志。开启后需要指定日志读取路径。
任务描述	可填，训练任务的简介，便于在训练任务列表快速了解训练任务信息。 支持 1~300 字符。

保存并运行自定义训练任务

- 完成参数设置后，单击“保存任务”。保存成功后跳转回训练任务列表页面，但训练任务保存后不会自动执行训练。
- 在训练任务列表操作栏单击“运行”，训练任务在成功调度所需资源后开始执行。



- 每单击 1 次“运行”即启动 1 次训练任务执行，支持多次运行。运行成功后状态变为“运行中”，当队列内资源不足时训练任务状态为“排队中”。训练任务状态详见训练任务生命周期。

4.6.4.2 管理自定义训练任务

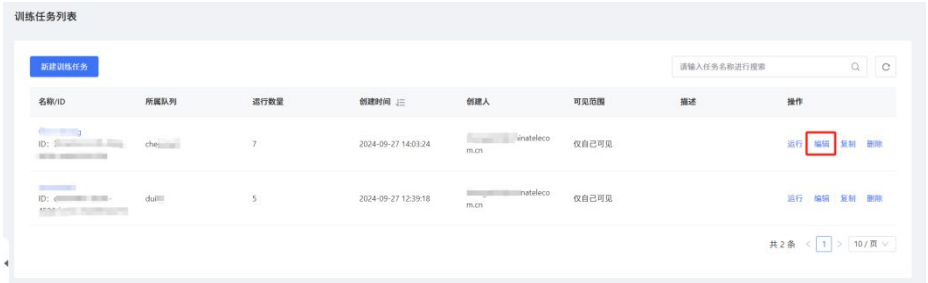
云骁智算平台提供对训练任务进行复制、编辑、删除的能力。

编辑训练任务

训练任务创建完成后，用户可通过编辑训练任务的功能修改已保存训练任务的参数，再次保存后覆盖原训练任务。

操作步骤

1. 在“训练任务列表”页面的操作栏中单击“编辑”进入编辑训练任务页面。



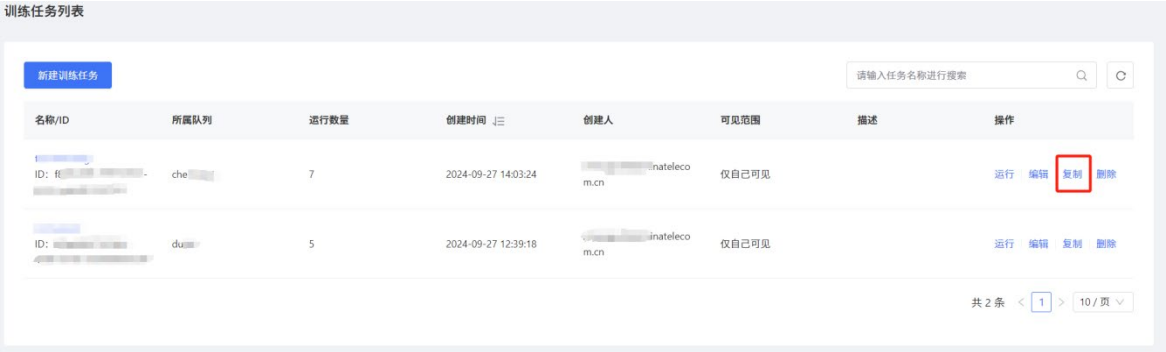
2. 除任务名称和训练框架外均可修改。
3. 更改相关参数并确认后，点击“保存任务”，即可完成编辑。

复制训练任务

在模型训练中，需要调整参数以获得最佳训练结果。云骁智算提供训练任务复制功能，可一键复制当前训练任务参数，并在当前任务基础上修改相关参数再次保存，提高模型训练效率。

操作步骤

1. 在“训练任务列表”页面的操作栏中单击“复制”进入复制训练任务页面。



2. 修改任务名称，不可与原任务同名。
3. 修改相关参数并确认后，点击“保存任务”，即可完成复制。

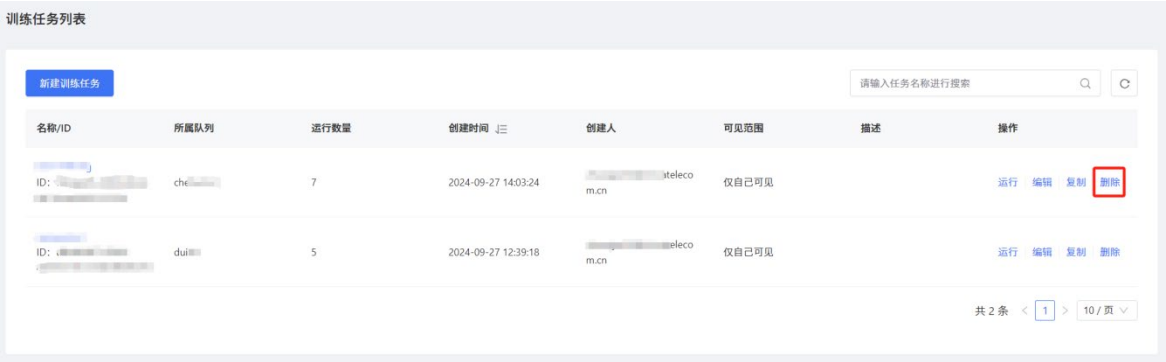
删除训练任务

如果不再需要使用此训练任务，建议清除相关资源，避免产生不必要的费用。

> 注意：训练任务删除后无法恢复，请谨慎操作。

操作步骤

1. 在训练任务列表页面的操作栏中单击“删除”。



2. 在删除确认弹窗中点击“确认”，即完成删除。

4.6.4.3 查看自定义训练任务详情

云骁智算平台可查看训练任务及其运行实例的基础信息和运行信息等，有助于更全面地了解训练任务信息。

操作步骤

1. 登录云骁智算控制台。
2. 进入对应工作空间。
3. 在左侧导航栏中，选择“训练>自定义训练”进入训练任务列表。
4. 在训练任务列表中，单击训练任务名称，进入自定义训练任务详情页。
5. 在任务详情页面可查看基本信息和运行情况。

基本信息

可查看任务的资源、环境配置的信息等，具体见下表：

参数	说明
任务名称	训练任务名称。
ID	训练任务唯一标识。

任务描述	训练任务的描述。
所属队列	训练任务占用此队列内的资源。
镜像名称	该训练任务使用镜像的名称。
存储	该训练任务的数据、算法的存储路径。
训练框架	该训练任务使用的训练框架名称。
训练模型	该训练任务使用的训练框架的具体模式。
启动命令	训练任务代码的执行命令。
环境变量	被注入到训练容器中的环境变量。
资源配置	训练任务占用的资源规格*数量。
Tensorboard	Tensorboard 日志是否采集，若是展示存储路径。
创建人	记录训练任务创建人账户名。
可见范围	显示训练任务的可查看权限范围。

运行情况

- 查看每一次训练任务运行记录的运行 ID、创建来源、所属队列、状态、运行时长、训练框架、开始时间信息。训练任务状态详见训练任务生命周期。
- 查看运行记录的详情，包括运行记录的基本信息、实例、日志、TimeLine 和监控。

运行记录详情	说明
基本信息	查看该运行记录的任务名称、ID、任务描述、所属队列、镜像名称、可见范围、存储、训练框架、训练模型、启动命令、环境变量、资源配置、Tensorboard、创建人信息。

实例	查看该运行记录下占用的实例信息，包括角色、实例名称、状态、重启次数、实例 IP、Host IP、运行时长、开始时间、查看日志等。
日志	选择实例，查看对应实例的标准输出日志。 注意：运行记录停止后日志消失，请在停止前保存所需日志数据。 如果需要长期查看日志数据，需在存储内单独建立目录保存日志数据。
TimeLine	查看该运行记录的时间线。
监控	选择实例，查看对应实例的监控信息。

- 对运行记录进行停止、删除操作。停止和删除操作后训练任务释放使用的资源，但停止操作保存运行记录。

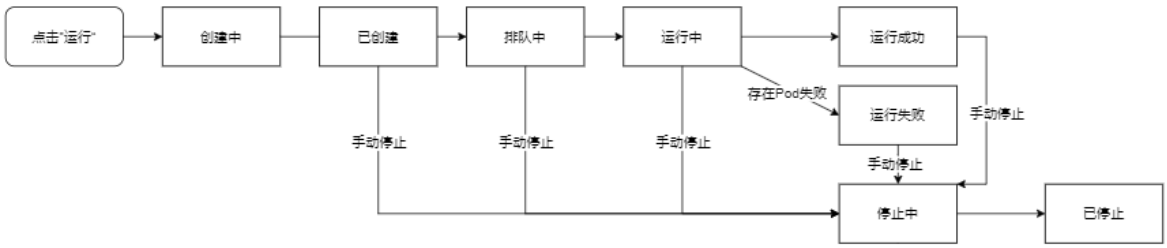
4.6.4.4 训练任务生命周期

训练任务状态说明

任务状态	说明
创建中	完成训练任务定义后，点击“运行”，创建运行记录的过程。
已创建	完成运行记录创建，已生成可运行任务。
排队中	任务正在等待资源分配和节点调度。
运行中	执行训练任务代码过程中。
运行成功	训练任务执行完成。
运行失败	≥1 个运行实例（Pod）失败。

停止中	用户触发“停止”操作，正在停止训练任务。
已停止	训练任务已停止运行。
异常	实例在排队、部署及停止的过程出现系统错误。

训练任务状态流转图



4.7.AI 加速

4.7.1.CTCCL 优化套件

CTCCL（CTyun Collective Communication Library）是天翼云自研的集合通信库。CTCCL 针对天翼云自身特点持续优化，提升性能并提供额外的可靠性保障。

4.7.1.1.CTCCL 介绍

CTCCL 是基于 NCCL 并结合天翼云云骁平台架构开发的集合通信库。和 NCCL 相比，CTCCL 提升了通信效率并提供额外的可靠性保障，为用户节约时间和经济成本。

CTCCL 新增关键特性：

- 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。

- 故障定位，第一时间识别故障点并上报，结合调度器修复或剔除故障节点，快速恢复硬件故障。

4.7.1.2.安装 CTCCL 库

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。本文档以 CTyunOS2.0+CUDA12.2 为示例，安装使用 ctccl-cuda12.2-0.2.0-1.x86_64.rpm ，其他环境类似。

1. 安装基础环境：

安装网卡驱动、openmpi、Nvidia-driver、CUDA12.2 等基础环境。如果要在容器内运行，则需要额外安装 docker、nvidia-container-toolkit 等。

2. 下载 rpm 包，并安装。如果在容器内运行则将 rpm 包复制到容器内安装。

```
wget https://jiangsu-10.zos.ctyun.cn/ctccl/nvidia/ctyunos2.0/ctccl-cuda12.2-0.2.0-1.x86_64.rpm
rpm -ivh ctccl-cuda12.2-0.2.0-1.x86_64.rpm
```

```
[root@host-p3bdep home]# rpm -ivh ctccl-cuda12.2-0.2.0-1.x86_64.rpm
Verifying...                          ################################# [100%]
Preparing...                          ################################# [100%]
Updating / installing...
   1:ctccl-0.2.0-1+cuda12.2           ################################# [100%]
[root@host-p3bdep home]#
```

3. 默认安装目录在/usr/lib64 下。

```
[root@host-2ucjk lib64]# ls | grep "libnccl*"
libnccl.so.2
libnccl.so.2.19.4
[root@host-2ucjk lib64]#
```

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

以下命令查找并根据需要替换 libnccl.so 文件

```
find / -name "libnccl*"     #获得 NCCL_PATH
cd /usr/lib64
cp libnccl.so.2.19.4 $NCCL_PATH
```

```
[root@host-p3bdep lib64]# ls | grep libnccl*
Binary file libnccl.so.2.19.4 matches
```

4. 根据需要设置 CTCCL 的环境变量，其他 NCCL 的环境变量也均有效。

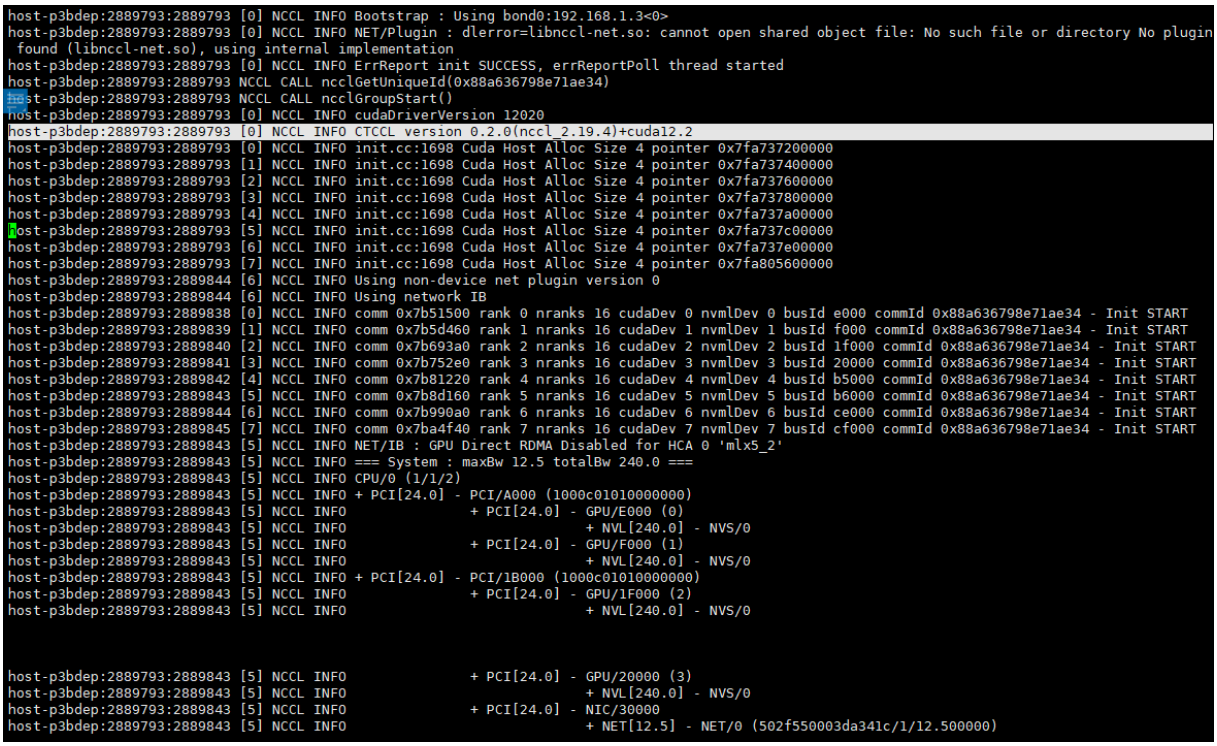
方法一：在训练脚本中配置环境变量

```
export NCCL_IB_QPS_PER_CONNECTION=8    #使用 8QP 并行传输
export NCCL_DEBUG="WARN"               #设置日志级别为 WARN
```

方法二：在节点上或容器内配置/etc/nccl.conf 文件

```
NCCL_IB_QPS_PER_CONNECTION=8
NCCL_DEBUG=WARN
```

5. 其他使用方式和 NCCL 完全适配，运行时可以看到对应 CTCCL 的版本信息。



4.7.1.3.启动 CTCCL 容器

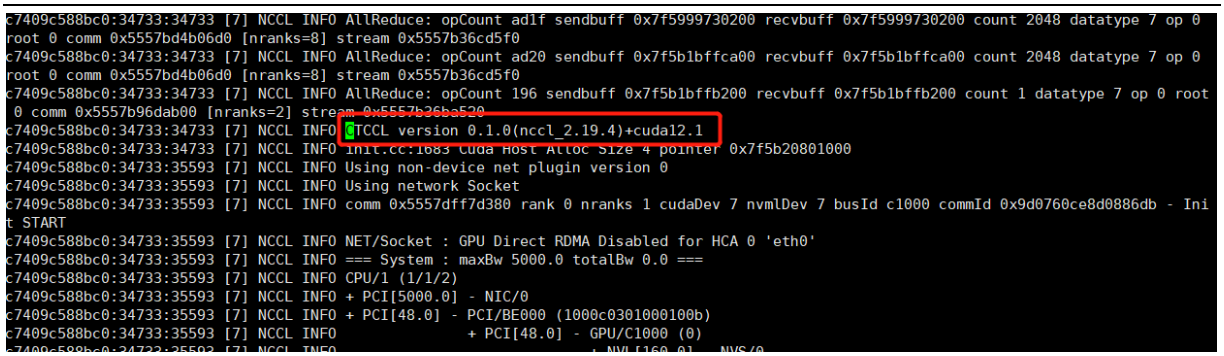
CTCCL 完全兼容 NCCL，用户使用只需要保证版本兼容即可。CTCCL 和 NCCL 版本对应关系：

通信库	CTCCL	NCCL
版本对应	0.1.0	2.19.4

启动 CTCCL 容器

可以通过包含 CTCCL（替换 NCCL）的云骁预置容器镜像来使用 CTCCL 功能， 镜像名为：ctccl0.1.0-pytorch2.2.0-cuda12.0-ubuntu20.04-x64:v1。

可通过 Docker 启动容器，开启日志，打印版本信息如下时，则确认使用 CTCCL：



4.7.1.4.使用 CTCCL 容器

在 NVIDIA 平台上，可以使用 NCCL-Test 工具测试 CTCCL 的性能。

1. 测试代码下载路径：

https://github.com/NVIDIA/nccl-tests.git

2. 编译：

make MPI=1 MPI_HOME={{MPI 路径}} CUDA_HOME={{CUDA 路径}} NCCL_HOME={{NCCL 路径}} -j 40

3. 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

4.7.1.5.附：CTCCL 环境变量

CTCCL 兼容 NCCL 环境变量，NCCL 环境变量参考：Environment Variables — NCCL 2.20.3 documentation (nvidia.com)。

CTCCL 容器已修改环境变量如下：

环境变量	说明	取值
NCCL_IB_QPS_PER_CONNECTION	单连接使用的 QP 数量	配置范围 1-128，默认值 8
NCCL_DEBUG	打印日志级别	VERSIONWARN(默认值)INFOTRACE

开启 CTCCL 自研特性，需要添加如下变量：

环境变量	说明	取值
CTCCL_IB_RETRY_DISABLE	开启故障重传机制	默认开启

4.7.2.CTFlashCkpt 加速包

CTFlashCkpt 是由云骁一体化智算加速平台提供的针对大模型训练场景提供的高性能 checkpoint 框架，实现接近于 0 的模型状态保存时间开销 ，将训练阻塞时间降低到最小。

4.7.2.1.CTFlashCkpt 介绍

CTFlashCkpt 是由云骁一体化智算加速平台提供的针对大模型训练场景提供的高性能 checkpoint 框架，实现接近于 0 的模型状态保存时间开销 ，将训练阻塞时间降低到最小。目前 CTFlashCkpt 支持原生 pytorch 训练、英伟达训练框架 Megatron-LM 和华为昇腾 ModelLink 训练框架，本文为您介绍 CTFlashCkpt 相关技术原理和接入操作。

背景信息

在大规模分布式训练中，由于软硬件故障的影响，任务可能会遭遇中断或需要重启。

为了应对这种情况，通常会采用定期保存 Checkpoint 的方法来记录和恢复训练进度。由于 Checkpoint 本身的耗时与模型的大小成正比，随着大模型参数量和训练数据量的增长，训练的时间开销也在不断增长。例如，对于百亿、千亿参数的大模型，单次 Checkpoint 的保存时间开销通常在几分钟到十几分钟之间。并且使用英伟达发布的 Megatrong-LM 或者原生的 Pytorch 训练模型的时候，需要中断训练进程，造成算力资源的浪费。因此，在训练过程中需要以一种可靠的方式来减少时间消耗和算力浪费。

CTFlashCkpt 采用异步存储机制加快训练速度，减少训练中断带来的影响，提升 GPU 的有效使用率。

4.7.2.2.安装 CTFlashCkpt

源码安装 CTFlashCkpt:

1. python 包下载地址：https://huabei-2.zos.ctyun.cn/huabei2-cwai-images/ctflashckpt.tar
2. 下载后进入到工程根目录
3. `sh scripts/build_wheel.sh build`
4. `pip install dist/eagle-0.1.0-py3-none-any.whl`

至此，CTFlashCkpt 的软件包安装完成，使用原生 pytorch 的话，已经可以使用了（具体使用方法见“使用 CTFlashCkpt”章节）。

CTFlashCkpt 扩展使用:

如果需要为英伟达 Megatron-LM 或华为 ModelLink-megatrton 的存储加速的话，

还分别需要下面的步骤：

英伟达 Megatron-LM

1. 下载 Megatron-LM 代码，可以参考 https://github.com/NVIDIA/Megatron-LM

2. 将 CTFlashCkpt 工程内 scripts 文件夹内的 replace_megatron_checkpointing_methods.sh 拷贝到 nvida Megatron-LM 的根目录,假设是/app/Megatron

3. `cd /app/Megatron-LM && git checkout core_r0.5.0 && sh replace_megatron_checkpointing_methods.sh && pip3 install -e .`

华为 ModelLink

1. 下载 ModelLink 代码，可以参考 https://gitee.com/ascend/ModelLink/blob/master/examples/README.md

2. 将 CTFlashCkpt 工程内 scripts 文件夹内的 replace_megatron_checkpointing_methods.sh 拷贝到昇腾 ModelLink 的根目录,假设是/app/ModelLink

`cd /app/ModelLink && sh replace_megatron_checkpointing_methods.sh && pip3 install -e .`

4.7.2.3.使用 CTFlashCkpt

原生 pytorch checkpoint

在训练代码内需要保存 checkpoint 的 py 文件内，进行加载和 checkpointer 初始化：

```
from eagle.framework_train.pytorch.async_checkpoint.ddp_checkpointer import DdpCheckpointer
checkpointer = DdpCheckpointer("/tmp/flash_checkpoint_example_ckpt")
```

将 pytorch 代码里用来存储/读取 checkpoint 的代码，例如 torch.save、torch.load，修改为

```
checkpointer.save_checkpoint(step, state_dict)
checkpointer.load_checkpoint()
```

可以参考工程内的示例：examples/ddp_example.py

CTFlashCkpt 扩展使用:

英伟达 Megatron-LM

按照上述安装方法安装 CTFlashCkpt 后，修改的英伟达 Megatron-LM 的默认存储已经被 CTFlashCkpt 替换，只要按照标准 Megatron 的 pretrain_gpt.py 训练，得到的 checkpoint 即是异步存储的。

华为 ModelLink

按照上述安装方法安装 CTFlashCkpt 后，修改的 ModelLink 的默认存储已经被 CTFlashCkpt 替换，只要按照标准 ModelLink 的 pretrain_gpt.py 训练，得到的 checkpoint 即是异步存储的。

镜像使用

平台的公共镜像中我们已经分别为安装了英伟达 Megatron-LM 和昇腾 ModelLink 的镜像集成好了 CTFlashCkpt 的功能，可以直接运行 megatrong/modellink：

- 英伟达镜像：ctflashtckpt:v0.1.0-pytorch2.0.0-cuda12.2-ubuntu22.04-x64-llama
- 昇腾镜像：ctflashtckpt:v0.1.0-pytorch2.1.0-cann7.0.0-ubuntu22.04-aarch64-llama

我们提供了运行训练 llama 的示例，可以自行阅读并修改相关配置进行训练：

- 可参考 examples/llama 文件夹的 run_llama.sh 文件，这是训练启动程序；
- var_example.sh，这是启动 run_llama.sh 的参数示例（不要运行此文件）。

使用效果观察

昇腾 ModelLink 的 CTFlashCkpt 写加速

数据集大小	ckpt 文件大小	训练环境配置	原生 ckpt 单次保存到存储 训练中断时间	优化后 flash ckpt 单次保存耗时 训练中断时间
7B	76G	机器配置：昇腾 910B 显存 512G，1 node，8 卡 4TP 2PP 1DP	10.45min	3.458s
13B	146G	机器配置：昇腾 910B 显存 512G，1 node，8 卡 4TP 4PP 1DP	20.43min	3.33s

英伟达 megatron 的 CTFlashCkpt 写加速

数据集大小	ckpt 文件大小	训练环境配置	原生 ckpt 单次保存到存储 训练中断时间	优化后 flash ckpt 单次保存耗时 训练中断时间
7B	76G	机器配置：Nvidia L40s 显存 384G，1 node，8 卡	3.9min	3.971s

		8TP 1PP 1DP		
13B	146G	机器配置：Nvidia L40s 显存 384G，2node，16 卡 8TP 2PP 1DP	4.71min	3.5s

注：昇腾及英伟达的测试数据来自于各自独立的测试环境，两组数据相互之间不具备可比性。

5. 最佳实践

5.1.如何上传数据到 ZOS 存储

最佳实践-ZOS 存储

5.2.如何上传数据到 HPFS 存储并使用

最佳实践-HPFS 存储

5.3.如何上传镜像到云骁智算的私有镜像仓库

登录

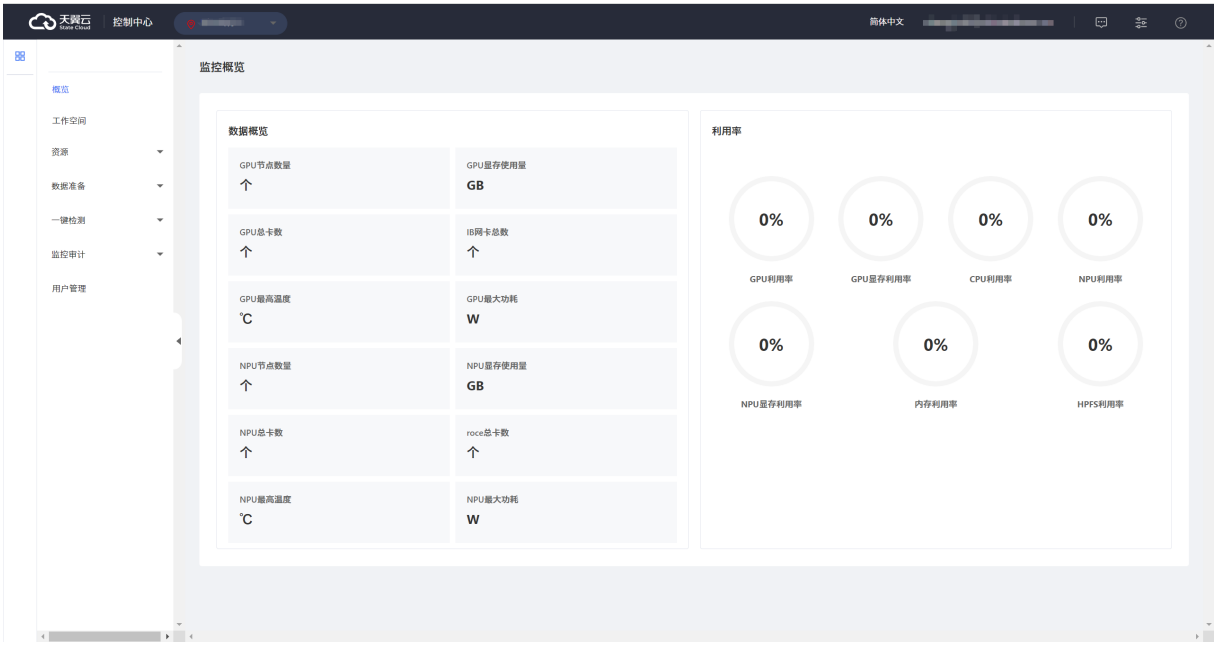
创建和使用天翼云骁智算之前，需要先注册天翼云门户的账号。如果已拥有天翼云的账号，可登录后直接使用天翼云骁智算。

1. 进入天翼云官网，选择【产品-计算-高性能计算-云骁智算】，点击【管理控制

台】。

2. 输入用户名密码，登录云骁智算平台。

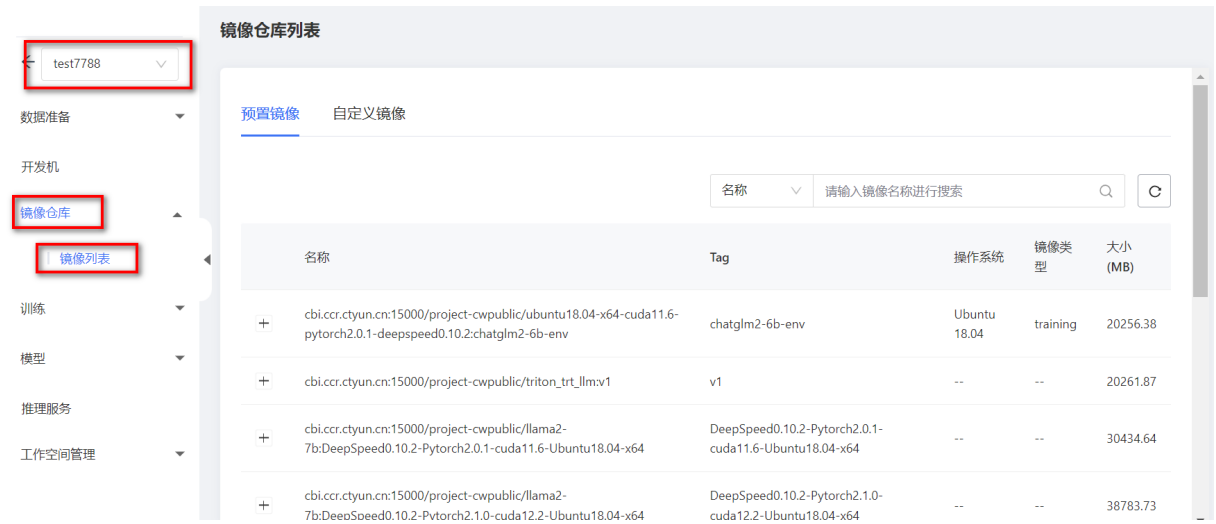
3. 登陆成功后，跳转至云骁智算产品首页。



镜像仓库

1. 进入镜像仓库需要先进入【工作空间】。在左侧导航栏选择【工作空间】，再选择需要进入的工作空间。

2. 进入工作空间后，在左侧导航栏选择【镜像仓库-镜像列表】，即可查看系统预置镜像和自定义镜像



镜像上传

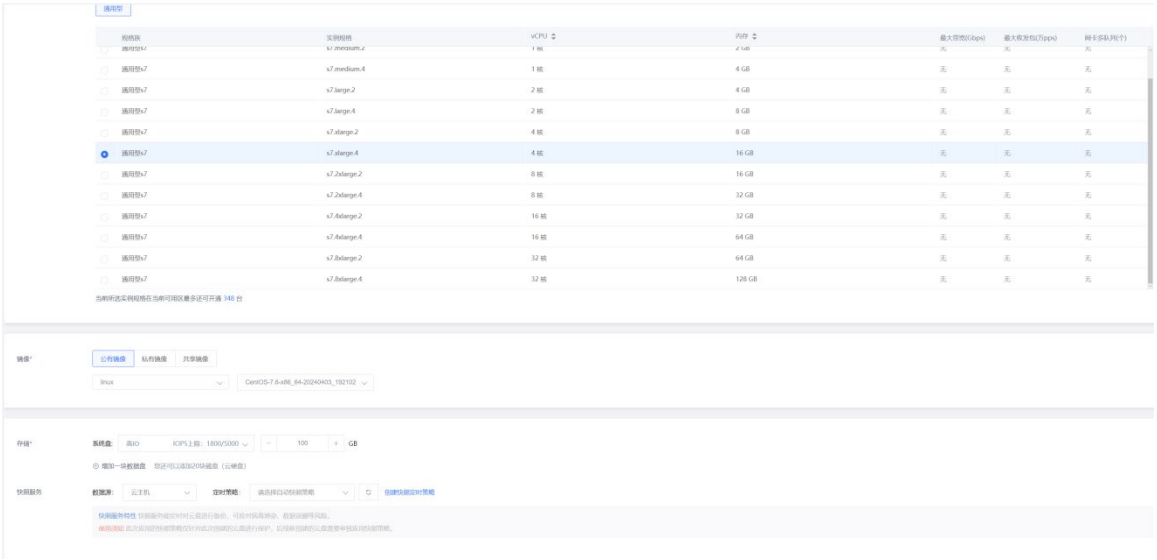
1. 在“自定义镜像”中，允许用户上传自己制作的镜像：选择“自定义镜像”标签，点击“上传镜像”按钮。

2. 需要提前说明的是，自定义镜像的仓库地址为内网地址，无法直接从公网访问，所以需要先设置镜像仓库的访问网络，总共有三种方式可以进行网络配置，请用户结合自己实际情况进行选择，具体如下：

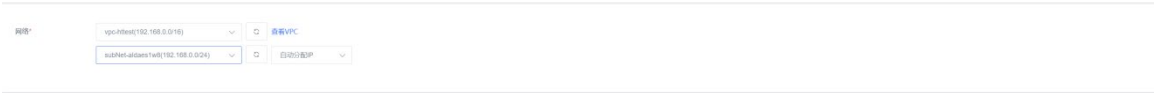
方法一：使用云主机配置

第一步：创建云主机

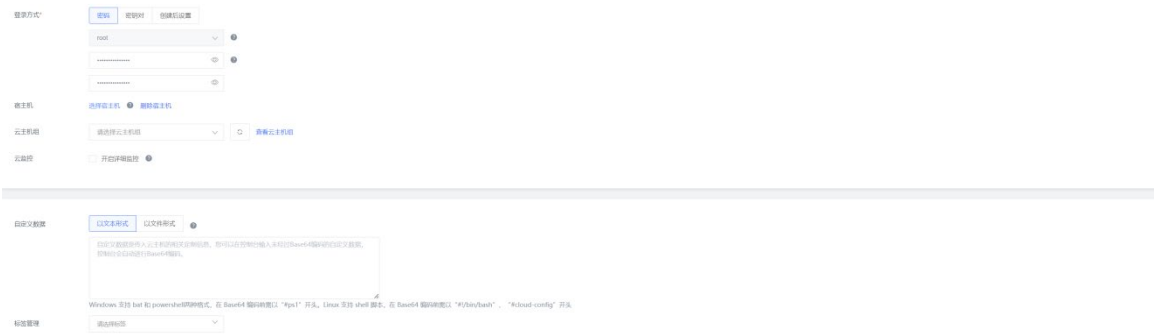
1) 在已经创建了资源组的账号下创建一台云主机，注意，必须是已经有资源组创建的账号下，这样资源组对应的 vpc 下的云主机才能与 harbor 仓库网络连通。请根据需要选取云主机规格以及镜像，系统盘大小请根据需要上传的镜像大小选取



2) 网络选取云骁资源组对应的 VPC，安全组默认即可



3) 设置密码后创建



第二步：配置环境

1) 登录云主机安装 docker，可选择外网 yum 安装或者下载安装包本地安装

2) 登录云骁控制台查看私有镜像仓库地址

3) 查看 harbor 地址对应的 vpce 地址

用户创建的云主机无法直接连通 harbor 仓库地址，需要通过云骁创建资源组时创建的终端节点 vpce 进行访问，所以需要先找到 vpce 地址

4) 找到 vpce 后，将该 ip 和对应域名配置在所在节点的/etc/host 中

```
[root@master-c7b1a80f-a06c-4c5f-91df-628841185ff5-0002 ~]# cat /etc/hosts
127.0.0.1 localhost localhost.localdomain localhost4 localhost4.localdomain4
::1 localhost6 localhost6.localdomain6 localhost6.localdomain
10.0.0.221 cbi.ccr.ctyun.cn
```

方法二：使用裸金属配置

云骁纳管的裸金属上自带 docker 和 harbor 相关的证书及网络配置，可将镜像文件传到裸金属上，然后直接上传

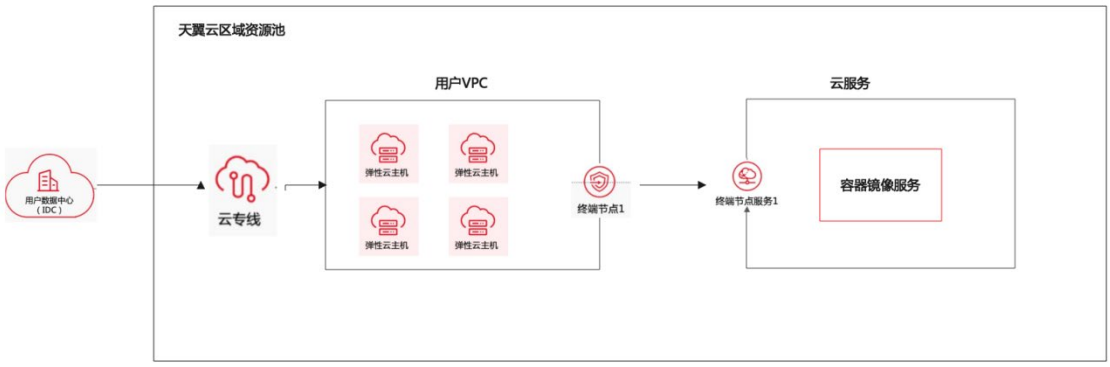
方法三：通过云专线配置

第一步：操作前提

1) 已经通过云专线打通用户网络与天翼云 VPC 网络

2) 用户在天翼云 VPC 网络内已经创建好镜像服务对应的 VPC 终端节点 IP（通过云骁创建标准资源组或扩展资源组，其所在 VPC 内已创建云骁镜像服务对应的 VPC 终端节点，并将改地址与镜像服务域名的映射配置在资源组节点中）

3) 用户网络内已存在机器，它可访问天翼云 VPC 内的终端节点 IP



第二步：配置环境

1) 登录云主机安装 docker，可选择外网 yum 安装或者下载安装包本地安装

2) 登录云骁控制台查看私有仓库地址

3) 查看 harbor 地址对应的 vpce 地址

用户创建的云主机无法直接连通 harbor 仓库地址，需要通过云骁创建资源组时创

建的终端节点 vpce 进行访问，所以需要先找到 vpce 地址

4) 找到 vpce 后，将该 ip 和对应域名配置在所在节点的/etc/host 中

```
[root@master-c7b1a80f-a06c-4c5f-91df-628841185ff5-0002 ~]# cat /etc/hosts
127.0.0.1 localhost localhost.localdomain localhost4 localhost4.localdomain4
::1 localhost6 localhost6.localdomain6 localhost6.localdomain
10.0.0.221 cbi.ccr.ctyun.cn
```

3. 通过以上配置，已经将访问镜像仓库地址的访问网络已配置好，接下来可以按照上传镜像弹窗所示内容上传镜像了。

1）下载证书到本地，并拷贝到对应目录下


```
#拷贝到对应目录下
mkdir -p /etc/docker/certs.d/cbi.ccr.ctyun.cn:15000
cp ca.crt /etc/docker/certs.d/cbi.ccr.ctyun.cn:15000
```

```
root@                    ~# ll /etc/docker/certs.d/cbi.ccr.ctyun.cn\:15000/
total 4
drwxr-xr-x 2 root root   20 Oct 24 11:38 ./
drwxr-xr-x 3 root root   36 Oct 24 11:38 ../
-rw-r--r-- 1 root root 2077 Oct 24 11:38 ca.crt
```

2） 登录 harbor，拷贝红框对应的登录命令和密码进行登录


```
root@[redacted]~# docker login -u user-testubuntu1024 cbi.ccr.ctyun.cn:15000
Password:
WARNING! Your password will be stored unencrypted in /root/.docker/config.json.
Configure a credential helper to remove this warning. See
https://docs.docker.com/engine/reference/commandline/login/#credentials-store

Login Succeeded
```

3）重新命名本地镜像

```
[root@[redacted]# docker tag testimage:latest cbi.ccr.ctyun.cn:15000/project-testubuntu1024/imagename:latest
```

```
docker tag testimage:latest cbi.ccr.ctyun.cn:15000/project-testubuntu1024/imagename:latest
```

4）推送镜像

```
docker push cbi.ccr.ctyun.cn:15000/project-testubuntu1024/imagename:latest
```

5）在自定义镜像中查看新上传的镜像信息

5.4.训练最佳实践- 昇腾+Pytorch+ChatGLM-6B

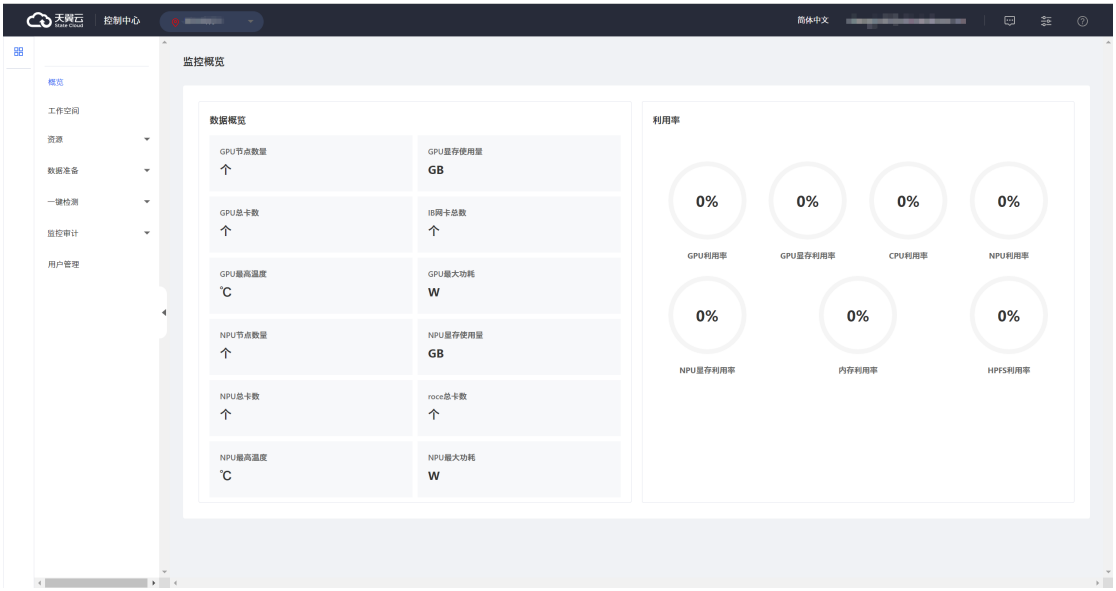
创建和使用天翼云骁智算之前，需要先注册天翼云门户的账号。如果已拥有天翼云的

账号，可登录后直接使用天翼云骁智算。

1. 进入天翼云官网，选择“产品>计算>高性能计算>云骁智算”，点击“管理控制台”。

2. 输入用户名密码，登录云骁智算平台。

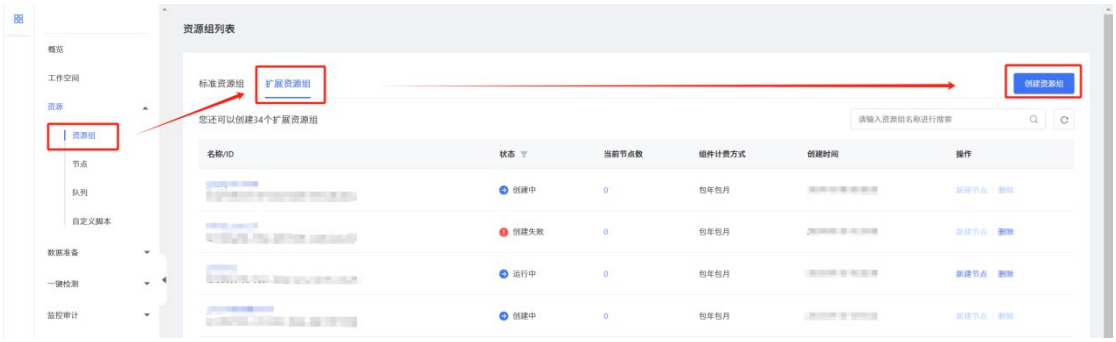
3. 登录成功后，跳转至云骁智算产品概览页。



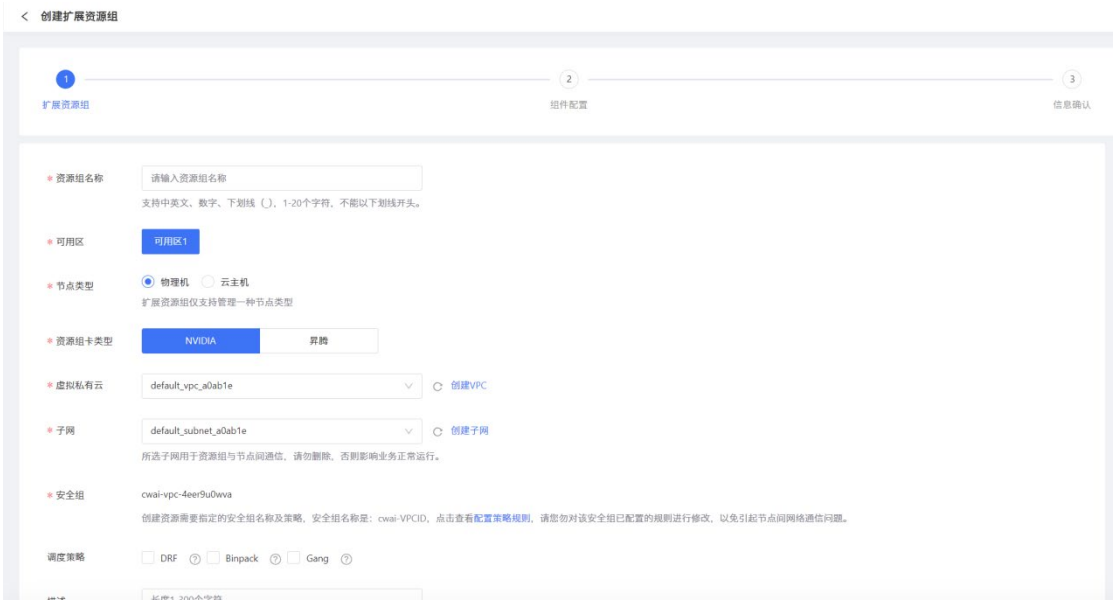
资源组

资源组是运行所需要的资源组合，资源组内可以有不同规格的节点。

1. 在左侧导航栏选择“资源>资源组”，进入“扩展资源组”页面，点击“创建资源组”按钮，选择“云骁扩展资源组”。



2. 在“创建扩展资源组”页面，依次完成“扩展资源组>组件配置>信息确认”操作，点击“确认”按钮，创建扩展资源组。

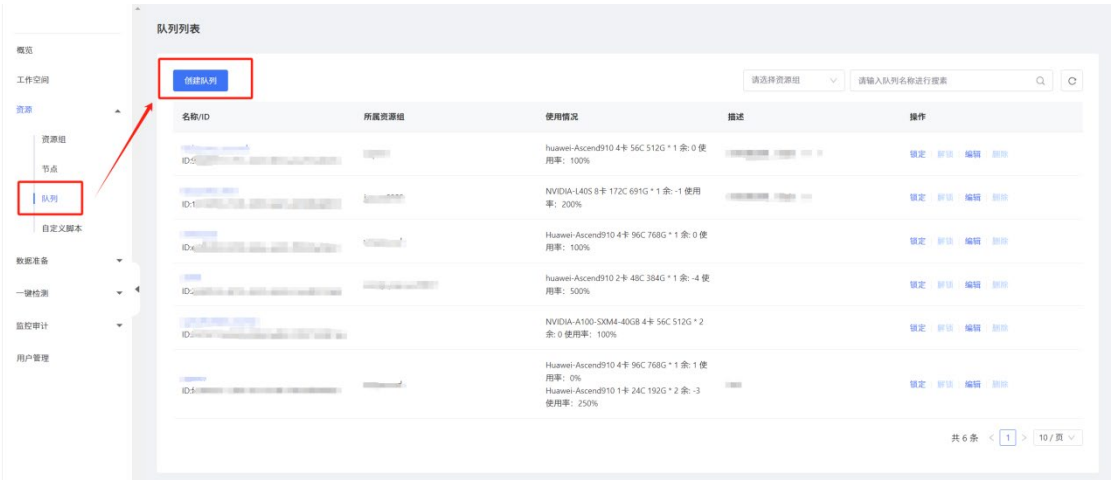


3. 在资源组列表中查看新创建的资源组。

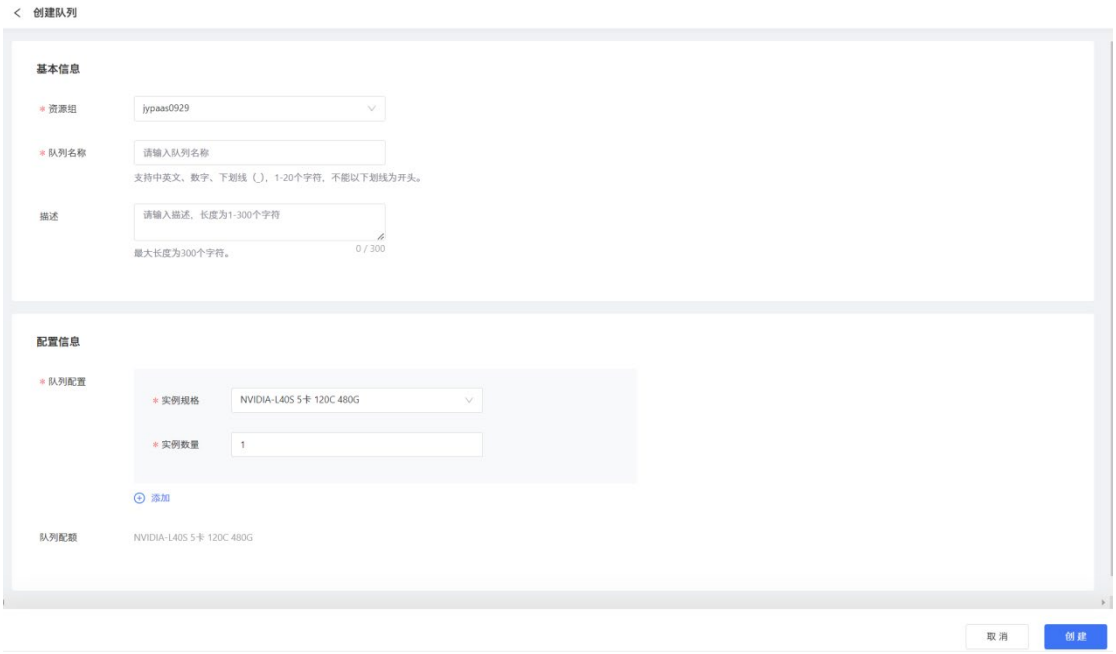
队列

队列是对资源配额的细粒度管理，用户通过队列划分配额的粒度和数量，队列是资源组内部分资源配额的集合，一个资源组可创建多个队列。在运行训练任务时，通过将任务绑定到队列进行资源的排队和使用申请。只有云骁扩展资源组可以用来创建队列。

1. 点击左侧导航栏选择“资源>队列”，点击“创建队列”按钮，创建训练使用的队列。



2. 选择已创建的资源组，输入队列名称，选择训练所需要的实例规格和实例数量，点击“创建”，完成队列创建。



3. 在队列列表中查看已创建的队列。

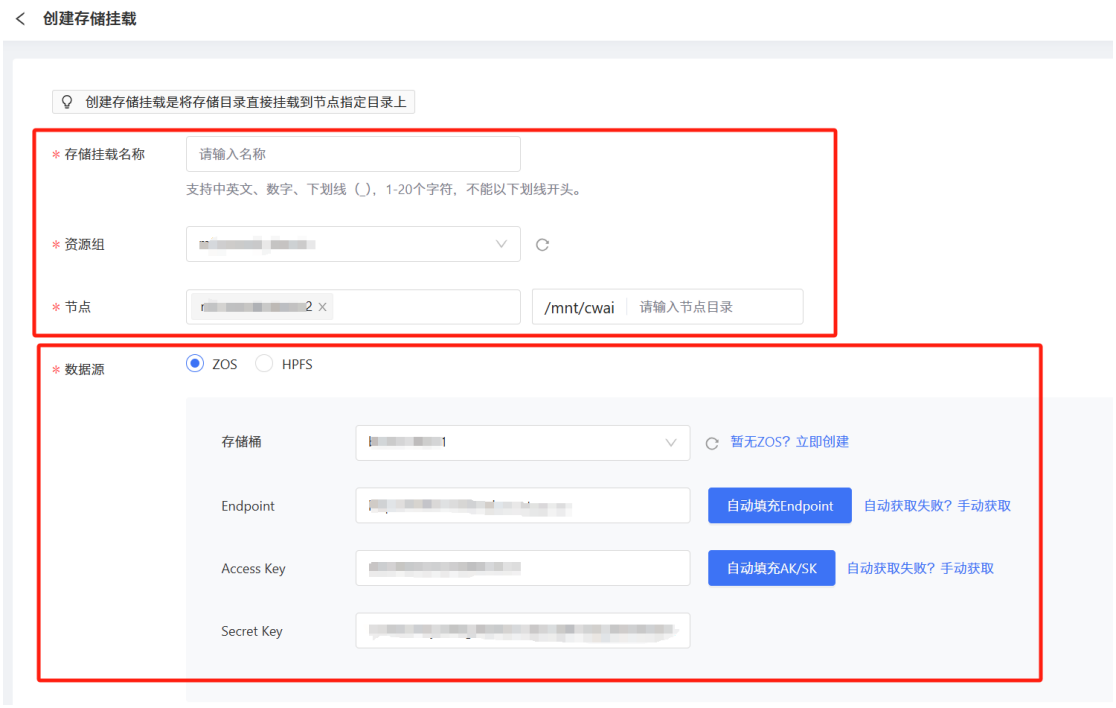
数据准备

通过存储挂载，将 ZOS 实例批量挂载到相应的节点上，并且管理挂载目录。

创建存储挂载

1. 在左侧导航栏选择“数据准备>存储挂载”进入存储挂载列表页面，点击“创建存储挂载”按钮。

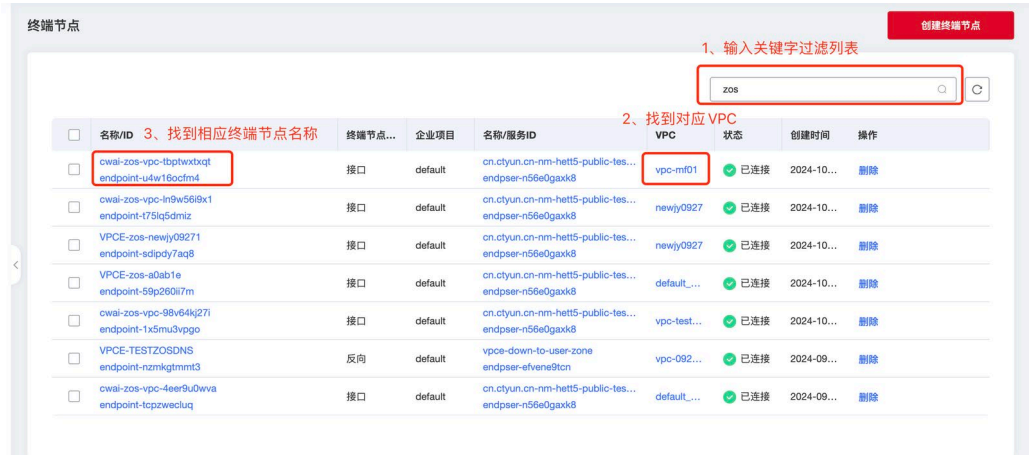
2. 配置存储挂载所需的名称、资源组、节点和数据源等相关参数。后续登录机器上传数据到 ZOS 时，请与此处所选节点设置的目录保持一致（/mnt/cwai/xxx）。存储挂载数据源选择“ZOS”。选择存储桶和填写 ZOS 的 Endpoint 和 Access Key/Secret Key。

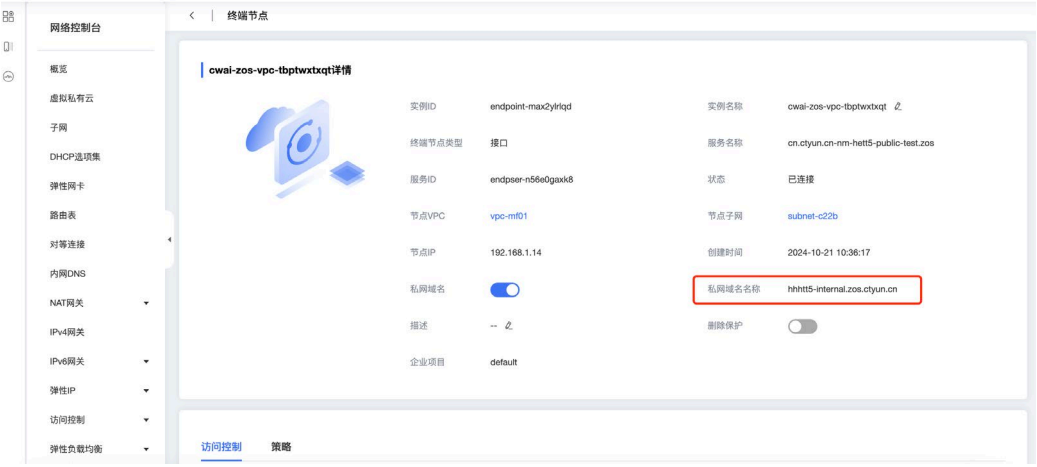


a. 存储桶：在下拉框中选择可用的 ZOS 名称。若当前无存储对象件，点击“暂无 ZOS？立即创建”可跳转至存储控制台创建对象存储 创建桶。创建完成后，返回此页面，点击 ZOS 选项旁的“刷新”按钮，下拉列表出现已创建的对象存储桶，选择需要挂载的 ZOS。

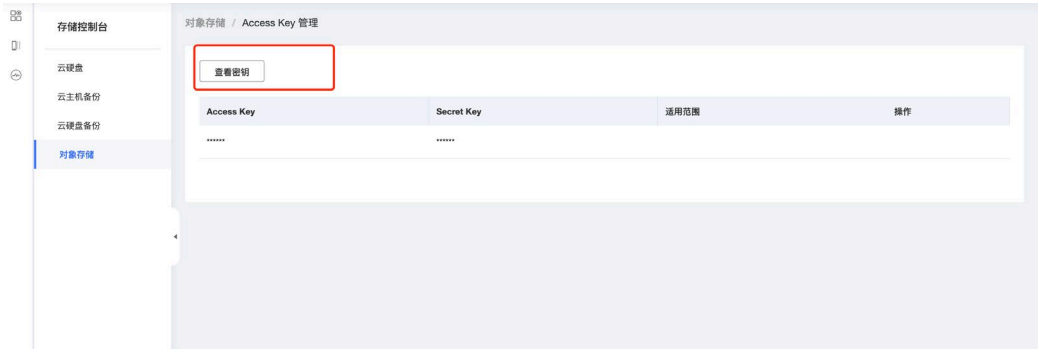
b. Endpoint：选择“自动填充 Endpoint”，可进行自动填充；若填充失败，点击蓝字“自动获取失败？手动获取”跳转至网络控制台终端节点页面，可以在搜索框输入关键字过滤列表，找到该资源组对应的 VPC（若不清楚资源组 VPC 名称，可进入“资源>资源组”页面，找到对应资源组名称，点击名称进入详情页，虚拟私有云后即是 VPC 名称），从而找到相应的终端节点名称，点击名

称进入详情页，复制"hppt://私网域名名称"到云骁相应页面。





c. Access Key/Secret Key：选择“自动填充 AK/SK”，可进行自动填充；若填充失败，点击蓝字“自动获取失败？手动获取”跳转至存储控制台 Access Key 管理页面，点击“查看密钥”，复制 Access Key、Secret Key 到云骁相应页面。



3. 点击“保存”按钮，成功创建存储挂载。

上传数据到 ZOS

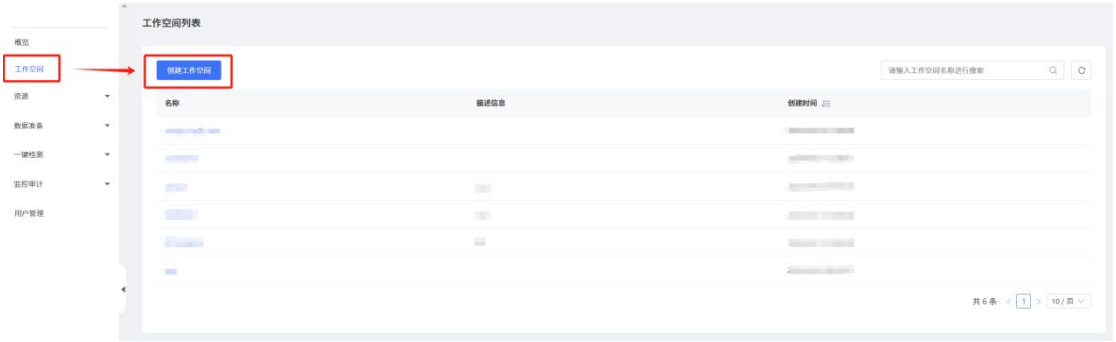
下载微调 ChatGLM2-6B 模型所需文件，并将 wget 的文件解压后传到物理机 ZOS 挂载目录（/mnt/cwai/xxxx）。

```
wget https://huabei-2.zos.ctyun.cn/huabei2-cwai-dataset/chatglm2-6b-cwai-ascend.tar
tar -xvf chatglm2-6b-cwai-ascend.tar
```

工作空间

工作空间可实现项目资源隔离、多项目分开工作。工作空间的创建需要在队列创建操作之后。

1. 在左侧导航栏选择“工作空间”进入工作空间列表页面，点击“创建工作空间”按钮。



2. 完成“基础信息”和“关联资源”配置，点击“创建”即可完成工作空间。其中关联资源列表仅显示未关联其他工作空间的队列，若无队列资源可选择，可创建新队列或将目标队列与其他工作空间解绑后即可选择。

3. 在工作空间列表可查看已创建的工作空间。

数据准备

创建训练任务所需要的数据集，通过云骁智算数据集模块实现训练中用到海量数据的准备与管理。

创建数据集

1. 进入数据集列表，点击“创建数据集”。

2. 进入“创建数据集”页面，按要求配置相关信息。

数据源：选择 ZOS。选择已挂载的 ZOS 存储桶，点击“选择目录”，可选择桶下的子目录。

训练

创建自定义训练任务实现一键运行任务，且支持同一任务多次运行。

1. 进入工作空间后在左侧导航栏选择“训练>自定义训练”进入训练任务列表页面，点击“新建训练任务”按钮。

2. 在“新建训练任务”页面，配置训练任务相关参数。

a. “镜像来源”选择预置镜像。具体预置镜像通过下拉菜单选择：公共镜像名称为 chatglm2-6b，镜像版本为 DeepSpeed0.9.2-Pytorch1.11.0-cann7.0-Ubuntu22.04-arm64

b. “存储”选择“+数据集”，挂载目录选择上述步骤创建的数据集，成功后会将目录挂载到容器内目录（默认为/data）。

c. 在启动命令栏中填入启动训练任务的执行命令：

```
/bin/bash
-c
cd /opt/ml/input/data/data &&chmod +x ./run.sh
&&BS_PER_DEVICE=1GRAD_ACC=8LR=2e-2
MAX_STEPS=1000SAVE_STEPS=200MODEL_PATH=/root/chatglm2-6b-files/weights/
OUTPUT_DIR=/opt/ml/output/test-adgen-chatglm2-6b-pt
LOG_DIR=/opt/ml/log/tb-test-adgen-chatglm2-6b-pt ./run.sh
```

ChatGLM2-6B 模型微调启动命令说明：

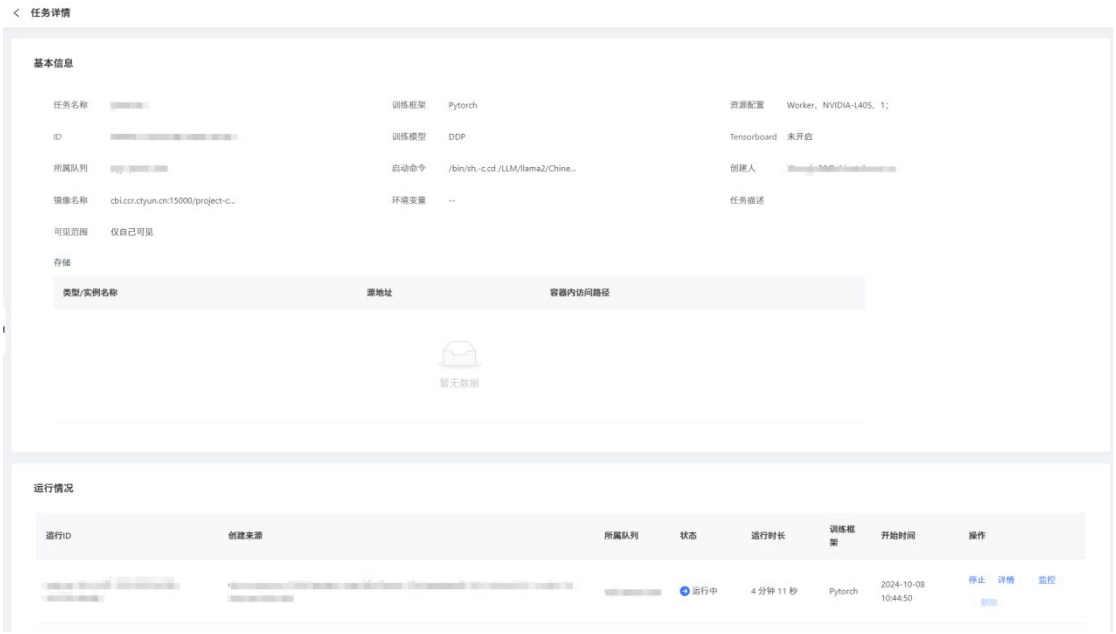
```
BS_PER_DEVICE #每张卡的 batchsize
GRAD_ACC #梯度累积步数
LR #学习率
MAX_STEPS #总训练步数
SAVE_STEPS #checkpoint 保存间隔
MODEL_PATH #预训练权重路径
```

```
OUTPUT_DIR #checkpoint 保存路径
LOG_DIR #tensorboard 日志保存路径
```

1. 确认所配置参数，点击“保存任务”。

2. 保存创建的训练任务后，需在训练任务列表页面单击“运行”操作，训练任务才会开始调度资源开始运行。支持多次运行。

3. 运行后在训练任务列表页面点击任务名称，进入训练任务详情页查看运行的详细信息。



4. 单击运行记录的“详情”操作可查看本次运行的实例、TimeLine、监控、日志等信息。

5.5.断点续训练

断点续训练指因硬件故障、系统问题、连接错误以及其他未知问题导致训练任务中断后，下一次训练可以在上一次的训练基础上继续执行。断点续训练可以减少需要长时间训练大模型的时间成本。

断点续训练是通过 checkpoint 机制实现。Checkpoint 机制是在模型训练的过程中，不断地保存训练结果（包括但不限于 EPOCH、模型权重、优化器状态、调度器状态）。即便模型训练中断，也可以基于 Checkpoint 接续训练。当训练任务发生中断后需要接续训练，只需要加载 Checkpoint，并使用断点前最近一次 Checkpoint 存储的信息初始化训练状态即可。

说明：目前仅支持使用昇腾芯片训练的任务断点续训，其他芯片正在开发中，敬请期待。

前提条件

- 资源组内存在空闲节点，即未被其他训练任务占用的机器。
- 训练任务所属队列内有多余配额可用且足够时，训练任务可自动触发断点续训；训练任务所属队列内无多余配额或多余配额不足时，需扩容该队列，成功后，训练任务可触发断点续训。

训练过程

云骁智算平台会识别中断任务并使用断点前最近一次 Checkpoint 存储的信息将训练任务重新调度并拉齐训练任务。您只需在训练开始前设置分布式存储/读取 Checkpoint，设置成功后将

模型代码上传至存储或自定义镜像，并在”创建训练任务”页面选择对应存储或自定义镜像即可。

Checkpoint 有以下两种工具：

工具一：使用原生 Pytorch Checkpoint。

需在 pytorch 中设置分布式存储/读取 Checkpoint，例如设置 torch.save、torch.load 等参数的值。

工具二：在原生 Pytorch Checkpoint 基础上使用云骁智算平台提供自研 CTFlashCkpt 加速包，其采用异步存储机制加快训练速度，详见 CTFlashCkpt 介绍。

CTFlashCkpt 加速包的安装和使用方法参见安装 CTFlashCkpt 和使用 CTFlashCkpt。

说明：如果 latest_checkpointed_iteration.txt 内上一次训练最后保存点和下一次训练日志中开始点不一致，可能是写入存储硬盘速度较慢导致，以日志中开始点为准。

6. 常见问题

6.1.资源类

Q：为什么新建节点时选择不到预期的物理机规格？

GPU 物理机规格默认不可见，需要联系客户经理进行加白名单流程。

Q：标准资源组和扩展资源组的区别是什么？

标准资源组提供基于 GPU 物理机和 GPU 云主机的标准集群服务；
扩展资源组是在标准资源组的基础上安装 Kubernetes 服务及相应组件。

Q：退订节点和移除节点有什么区别？

- 退订节点：只有新建节点支持在云骁平台进行退订操作，退订操作会导致资源回收和清理，节点上的数据将无法恢复。
- 移除节点：只有纳管节点支持在云骁平台进行移除操作，将非云骁平台开通的

节点与资源组解绑并移除出节点列表，不涉及底层资源的退订。

Q：队列“锁定”操作会影响队列内正在运行的训练任务吗?

“锁定”队列后，只是禁止后续训练任务在该队列上调度，不会影响目前队列内已创建完成的任务。

Q：云骁智算都有哪些资源池上线售卖?

目前云骁智算已上线 13 个资源池，其中 7 个为全网可见资源池，其余资源池需开通白名单查看：

序号	资源池	属性
1	华东 1	全网可见
2	华南 2	全网可见
3	上海 15	
4	芜湖 4	
5	西南 2-贵州	全网可见
6	杭州 7	全网可见
7	华北 2	全网可见
8	北京 9	
9	武汉 41	全网可见
10	长沙 42	全网可见
11	成都 11	
12	北京 11	
13	湖北武汉一城一池 1	

后期将增加更多资源池供用户选择，敬请期待。

6.2.数据类

Q：创建时提示“用户选择的节点所处在子网未全部配置天翼云内网 DNS 服务器……”/修改标准裸金属子网 DNS 配置后导致的 ZOS 存储挂载异

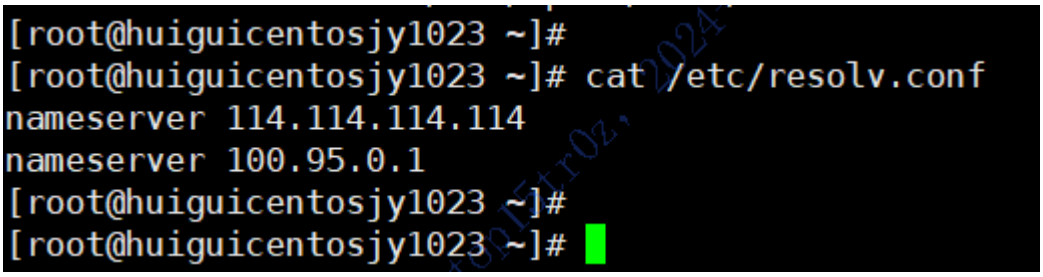
常问题

说明：在 ZOS 存储挂载及使用时，GPU 节点与 ZOS 实例之间通过 TCP 网络通信，因 GPU 节点与 ZOS 实例分别处在不同的 VPC 内，需要通过 VPC 创建端点服务来建立连接。云骁平台会自动为用户处理相关操作，但需要用户所选节点所在的子网已配置天翼云内网 DNS 服务器地址作为子网的 DNS 服务器地址，以便节点能正常使用天翼云内网 DNS 服务并解析 ZOS 端点服务的私网域名。

若用户在创建 ZOS 存储挂载时，出现“用户选择的节点所处在子网未全部配置天翼云内网 DNS 服务器……”弹窗，可能是由于用户修改了节点所在的子网 DNS 配置或所选节点默认 DNS 配置不满足此处要求。

用户可以通过以下方式进行修改：

- 云主机、弹性裸金属节点：前往子网列表对相关节点的子网进行 DNS 服务器配置，修改子网 DNS 服务配置后，重启节点即可生效，通过登录节点查看 /etc/reslov.conf 文件可查看 DNS 配置详情。
- 标准裸金属节点：修改子网的 DNS 服务配置，只会对后续新建节点生效，对于修改前已创建节点，可以通过以下方式修改：
  1. 重装此台裸金属，或者销毁此台裸金属重开；
  2. 修改此裸金属的/etc/resolv.conf 文件，增加 `nameserver 100.95.0.1`



Q：上传数据到 ZOS 存储时对数据大小有何限制?

参见对象存储数据上限。

6.3.训练类

Q：训练任务处于已完成状态时占用资源吗?

训练完成的任务会继续占用资源，如需释放资源请备份相关训练结果数据点击停止任务资源即可被释放。

Q：训练任务运行时报错“创建资源任务失败：队列设备获取报错:8bd12065-f643-4dbe-8685-3bf82dc5b521”？

训练任务所属队列中被删除，需进入编辑训练任务页面重新选择队列。

6.4.计费类

Q：云骁智算平台有哪些计费项?

云骁智算服务公测期为免费提供，但其涉及使用的产品如物理机、弹性负载均衡、弹性云主机、并行文件服务 HPFS、对象存储等，按照对应产品使用的收费标准另行收费。具体收费情况以下单页面费用为准。

计费说明参见:

- 物理机计费说明
- 弹性负载均衡计费说明
- 弹性云主机计费说明
- 并行文件服务 HPFS 计费说明
- 对象存储计费说明

Q：操作系统是否需要收费?

当前云骁智算平台默认提供的镜像基于 CentOS，该镜像无需支付额外费用。

Q：物理机到期冻结后，如何解冻?

参见物理机到期处理。

6.5.其他

Q：云骁智算的使用流程有哪些关键步骤?

- 创建资源组：在资源组内可扩缩容节点。
- 创建队列：资源组与队列的关系为一对多，即一个资源组内可有≥1 个队列。
- 检测环境与性能：检测资源组环境和硬件性能情况，为训练任务的顺利运行做准备。
- 创建自定义训练任务：关联队列后设置相关信息即可运行训练任务。
- 查看监控告警信息：支持以不同时间间隔维度检测资源和任务的运行情况，并根据系统告警信息及时调整。

  详情请参见用户指南。

Q：云骁智算使用时有什么限制?

- 建议不要自行升级节点的内核版本和操作系统版本。
- 禁止修改网络相关的配置，否则可能导致无法连接节点。
- 不支持跨 AZ 创建资源组。
- IB 网络只支持租户级隔离，不支持子账号级隔离。
- 请勿通过其他产品控制台删除云骁智算平台为用户创建的资源，如资源组管理节点云主机 ELB 和 VPCE。