ICS 35.020
CCS L 70

团　　体　　标　　准

T/AI 123.6—2024

人工智能算力网络
第 6 部分：云际管理与调度

Artificial intelligence computing net –
Part 6: JointCloud management and scheduling

T/AI 123.6-2024

2024－11－29 发布　　2024－11－29 实施

中关村视听产业技术创新联盟　发布

T/AI 123.6-2024

T/AI 123.6-2024

版权保护文件

版权所有归属于该标准的发布机构，除非有其他规定，否则未经许可，此发行物及其章节不得以其他形式或任何手段进行复制、再版或使用，包括电子版，影印件，或发布在互联网及内部网络等。使用许可可于发布机构获取。

T/AI 123.6-2024

目　　次

前言 II
引言 III
1 范围 1
2 规范性引用文件 1
3 术语和定义 1
4 缩略语 3
5 软件参考架构 3
6 功能要求 4
6.1 算力资源管理 4
6.2 镜像管理 7
6.3 数据管理 7
6.4 作业管理 9
6.5 作业调度 11
6.6 调度策略 12
6.7 调度性能评估 14
7 接口要求 14
7.1 作业管理接口 14
7.2 调度策略接口 14
7.3 调度评估模型接口 15
附录 A（资料性）接口说明 16
参考文献 22

T/AI 123.6-2024

前　言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规则起草。

本文件是T/AI 123《人工智能算力网络》的第6部分，T/AI 123已经发布了以下部分：

——第1部分：总体要求；

——第6部分：云际管理与调度。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由新一代人工智能产业技术创新战略联盟AI标准工作组提出。

本文件由中关村视听产业技术创新联盟归口。

本文件起草单位：鹏城实验室、北京大学、国防科技大学、华为技术有限公司、中科寒武纪科技股份有限公司、广东博华超高清创新中心有限公司。

本文件主要起草人：余跃、徐春香、杨晓珍、林冯军、王晖、张叶红、刘敏、曾炜、樊春、聂永丰、张鹏、朱宗卫、赵轩、张伟民、赵海英、崔晓冉。

T/AI 123.6-2024

引　言

T/AI 123为人工智能算力网络规定了技术要求，接口要求，调度适配协同要求，网络设施要求、运营服务要求等。

T/AI 123《人工智能算力网络》拟由以下部分组成：

——第1部分：总体要求。目的在于确立智算网络的系统参考架构、总体功能要求及接口要求。

——第2部分：异构资源统一度量和标识。目的在于确立数据、算力、模型、存储、网络等异构资源统一抽象度量和标识的方法与转换要求。

——第3部分：云际数据统一存储应用接口。目的在于确立全局共享数据存储的统一技术架构及其功能、接口、性能、可靠性要求。

——第4部分：调度适配器南向接口规范。目的在于确立调度适配器与各智算中心云化管理调度系统之间的接口，实现智算中心资源获取和作业下发。

——第5部分：资源感知路由协议接口。目的在于确立通过智算网络资源感知路由协议获取资源动态信息的接口。

——第6部分：云际管理与调度。目的在于确立异构云互联的云际资源管理与调度系统功能及智能优化调度技术与策略。

——第7部分：网络设施与互联。目的在于确立智算中心网络互联需要的网络种类、网络规格、网络功能等。

——第8部分：应用层接口。目的在于确立运营层、调度层为应用层提供的服务功能及接口。

——第9部分：多中心协同计算。目的在于确立多中心协同计算的应用场景，抽象建模协同计算作业，作业流程、作业评估及协同计算参考架构。

——第10部分：基础运营服务。目的在于确立运营层基础功能，包括用户/智算中心认证与授权、行为审计、运维监控等。

——第11部分：算力运营。目的在于确立算力资源的计量和费用结算方法，算力运营分析等。

——第12部分：数据资源流通交易统一规范与接口。目的在于确立运营层数据市场与模型市场资源流通交易过程中的功能接口，包括数据资源发布、确权、计量计费、运营分析等。

——第13部分：通信网络安全。目的在于确立智算网络中通信网络认证、传输等过程中的安全技术要求。

——第14部分：数据与模型安全。目的在于确立智算网络运营及应用中数据与模型隐私安全技术要求。

T/AI 123.6-2024

T/AI 123.6-2024

人工智能算力网络 第6部分：云际管理与调度

1 范围

本文件确立了人工智能算力网络（简称“智算网络”）的云际管理与调度的软件参考架构，规定了云际资源管理和作业调度的技术要求和调度策略要求。

本文件适用于人工智能算力网络中云际管理与调度软件系统的设计和开发。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41867-2022 信息技术 人工智能 术语

T/AI 123.1-2023 人工智能算力网络 第1部分：总体要求

3 术语和定义

GB/T 41867-2022界定的以及下列术语和定义适用于本文件。

3.1

数据中心 data center

一种能够提供容纳、互联和操作的结构，或结构组。它使用信息技术、电信网络设备提供的数据存储、处理、迁移服务及其它所有功能，并集成能量供应、环境控制和为保证服务可用性而制定的必要的韧性、安全性级别定义。

注1：数据中心结构一般包含数个楼宇或空间，用以支撑数据中心主要功能。

注2：包含数据中心中信息及通信技术设备及支撑环境控制设备边界或空间，定义于更大的结构或楼宇中。

[来源：ISO/IEC 22237-1:2021，3.1.8]

3.2

计算中心 computing center

为多用户提供计算服务的设施。用户的操作通过对计算设备及辅助硬件的操作及中心人员的服务实现。

[来源：ISO/IEC/IEEE 24765:2017，3.741]

3.3

人工智能计算中心 artificial intelligence computing center

智算中心

T/AI 123.6-2024

一种能够为多用户提供人工智能计算服务、数据容纳的结构或结构组。使用信息技术、电信网络设备提供的数据存储、处理、迁移，人工智能计算加速等功能，并集成能量供应、环境控制和为服务可用性而制定的必要的可靠性组件。

注1：人工智能计算中心一般包含数据中心可能涉及的楼宇或空间，用以支撑人工智能计算中心主要功能。

注2：人工智能计算中心中的服务器，一般包含人工智能服务器和通用服务器等，服务器称为“节点”。

[来源：ISO/IEC 22237-1:2021，3.1.8和ISO/IEC/IEEE 24765:2017，3.741，有修改]

3.4

算力网络 computing net

一种为用户提供计算资源的设施。通过网络技术将各地的计算中心连接起来，进而统筹分配和调度计算任务的网络。算力网络基于统一的算网资源视图，为业务和应用提供算力和网络融合保障和调度。

[来源：ITU-T Y.2501，6，有修改]

3.5

人工智能算力网络 artificial intelligence computing net

智算网络

一种为多用户提供人工智能计算资源及服务的平台。通过新型网络技术将各地分布的人工智能计算中心连接起来，构成多个中心间感知算力、数据、算法资源，进而统筹分配和调度人工智能计算任务的网络。

3.6

云际 jointcloud

通过网络将不同云实体进行整合、供应和管理。

[来源：GB/T 40690-2021，3.1 有修改]

3.7

资源 resource

系统具有的有限可用性的组件，常见的资源类型包括：CPU、GPU、内存、显存、磁盘容量、磁盘带宽、网络带宽、网络端口、I/O操作等。

3.8

作业 job

人工智能训练或推理任务的逻辑组合。

注1：一个作业包括一个或多个任务。

[来源：ISO/IEC 2382:2015，1.2122879 有修改]

3.9

任务 task

被调度的训练/推理对象。

T/AI 123.6-2024

注1：任务用于完成一个相对独立的业务功能。一个任务属于且仅属于一个作业。

[来源：ISO/IEC 2382:2015，1.2122482 有修改]

3.10

子任务 sub-task

根据任务内在逻辑关系拆分形成的较小任务单元。

3.11

作业负载 workload

智算中心正在运行的作业数量。

4 缩略语

下列缩略语适用于本文件。

CPU 中央处理单元（Central Processing Unit）
GPU 图形处理单元（Graphics Processing Unit）
MLU 机器学习处理器（Machine Learning Processing Unit）
NPU 神经网络处理单元（Neural-Network Processing Unit）
QoE 体验质量（Quality of Experience）
QoS 服务质量（Quality of Service）

5 软件参考架构

智算网络连接分散在各地的智算中心节点，汇聚和共享算力、数据、模型和应用等资源。智算中心通过加入算力网络实现资源共享，并通过智算网络统一调度，提高全网资源利用率，满足业务和应用的算力和网络需求。T/AI 123.1—2023 规定了智算网络的总体架构，见图 1。智算网络调度层的云际管理与调度系统（简称“云际系统”）主要实现跨智算中心之间的资源管理和全局作业调度。云际系统将存在平台异构性和分布泛在性的智算中心的算力资源和数据资源映射到统一的量纲维度，形成业务层可理解、可阅读的统一算力资源池，并综合考量 AI 训练作业和资源的特性而形成合适的调度策略，将作业调度至最优的智算中心。云际系统软件参考架构见图 2。

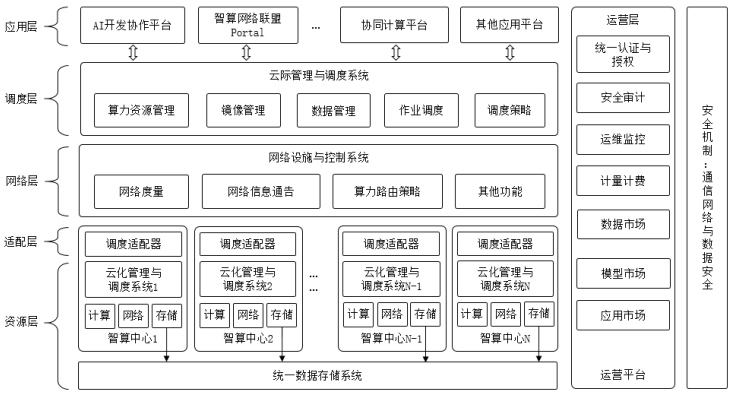

图 1 智算网络总体架构

T/AI 123.6-2024

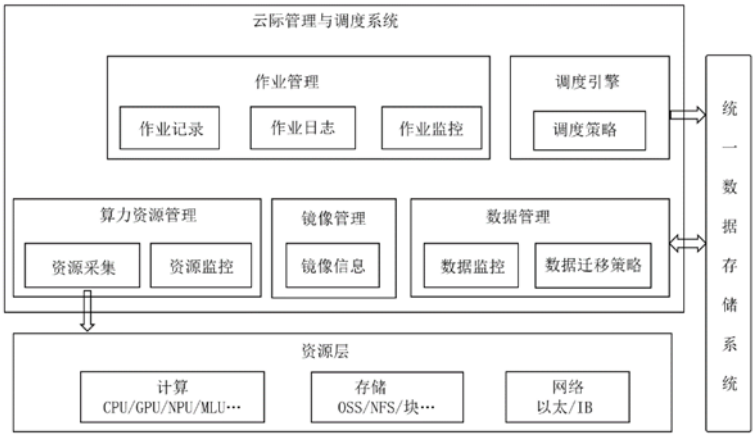

图 2 云际系统软件参考架构

云际系统应包含以下功能：

a) 算力资源管理：支持各智算中心计算、存储、网络资源的信息采集与监控，周期性地调用各智算中心的接口获取相应信息，支持各类算力资源的统一管理；

b) 镜像管理：支持对各智算中心可用的镜像信息进行管理；

c) 数据管理：支持智算网络上数据资源的监控，根据数据迁移策略选择合适的数据迁移源列表，提供给统一数据存储系统进行数据迁移；

d) 计算作业管理：算力使用者所提交作业记录的增删改查、日志管理、运行负载监控等基础功能；

e) 计算作业调度：根据调度策略，将计算作业分发到相应智算中心；

f) 调度策略：根据智算网络的算力资源、数据资源情况，选择合适的智算中心运行计算作业，至少支持以下调度策略中的一种：

1) 手动调度策略：人工指定智算中心运行作业；

2) FCFS 调度策略：先来先服务的调度策略；

3) 负载感知调度策略：根据各集群负载状况，优先选择负载低和有资源的集群调度作业；

4) 能耗感知调度策略：根据各集群总体能耗水平调度作业，优先选择能耗低的集群调度作业；

5) 价格感知调度策略：根据各集群资源费率调度作业，优先选择费率较低的集群调度作业；

6) 网络感知调度策略：根据各集群网络带宽、时延、抖动、丢包等性能参数调度作业，优先选择网络性能较优的集群调度作业；

7) 数据感知调度：对于数据量小的场景，选择最合适的智算中心运行作业并通过调度将数据迁移到该智算中心；对于不适合数据迁移的大数据或者隐私数据场景，感知数据所在智算中心并将作业调度到该智算中心运行；

8) 算力业务 QoS/QoE 调度策略：以作业的 QoS/QoE 为最高优先级决策依据，均衡优选网络路径和集群调度作业。

6 功能要求

6.1 算力资源管理

T/AI 123.6-2024

6.1.1 算力资源采集

云际系统对接入智算网络的所有智算中心的基础信息进行管理，包括智算中心的ID、名称、所属区域、创建时间、可用状态和位置信息等，如表1所示。

表1 智算中心信息表

名称	标识符	类型	长度	备注
编号	ai_center_id	string	小于 64 字节	智算中心编号
名称	name	string	小于 64 字节	智算中心名称
所属区域	region	string	小于 64 字节	智算中心所属区域
状态	status	int		智算中心状态，0：不可用；1：可用；2：未知
位置信息	geometry	float，float		智算中心地理位置坐标[经度，维度]
创建时间	created_at	timestamp		创建时间
更新时间	updated_at	timestamp		更新时间

云际系统通过接口对各个智算中心可被调度的算力资源信息进行信息采集。算力资源包括CPU、GPU、NPU、MLU、内存、存储和自定义资源，如表2所示。

表2 算力资源字段表

名称	标识符	类型	长度	备注
CPU	cpu	int		CPU 数量
GPU	gpu	int		GPU 数量
NPU	npu	int		NPU 数量
MLU	mlu	int		MLU 数量
内存	memory	int		单位 MB
存储	storage	int		单位 GB
自定义资源	user_define	int		自定义

6.1.2 算力资源整合

云际系统通过将各类资源统一为资源规格的形式进行资源整合，统一量纲，并创建一张资源规格与各智算中心资源情况的映射表。资源规格的种类和数量由智算网络系统管理员根

T/AI 123.6-2024

据各个智算中心的算力资源情况进行配置。当用户使用智算网络提交计算作业时，通过指定资源规格的形式来使用算力资源。资源规格的定义如表3所示。

表3 资源规格字段表

名称	标识符	类型	长度	备注
规格标识	resource_spec_id	string	小于64字节	资源规格在智算网络调度系统内的唯一标识
规格名称	resource_spec_name	string	小于64字节	规格名称
资源数量	resource_quantity	json		各类资源的数量，样例：{“cpu”:“1”, “memory”:“2Gi”, “gpu”:“1”}
可支持作业类型	service_type	string	小于64字节	该资源能够支持的作业类型：交互式作业、批处理作业、流式作业等
规格价格	price	int		规格包价格
智算中心列表	computing_center_list	string		资源规格对应的智算中心列表，用“，”号分隔，样例：center1，center2
创建时间	created_at	datestamp		创建时间
更新时间	updated_at	datestamp		更新时间

6.1.3 算力资源监控

云际系统定时通过接口获取各智算中心的各种计算资源的使用情况，包括各中心作业负载数量、CPU、GPU、NPU、MLU 和内存等资源的使用率和空闲数量，对智算网络算力资源进行总体监控。监控的资源信息如表 4 所示。

表4 资源监控字段表

名称	标识符	类型	长度	备注
智算中心名称	center_name	string	小于 64 字节	
运行任务数	running_task_num	int		某个智算中心正在运行的任务数
等待任务数	waiting_task_num	int		某个智算中心等待的任务数
CPU 使用量	cpu_used	int		某个智算中心已经使用的 CPU 核数
CPU 空闲量	cpu_free	int		某个智算中心空闲的 CPU 核数
NPU 使用量	npu_used	int		某个智算中心已经使用的 NPU 个数
NPU 空闲量	npu_free	int		某个智算中心空闲的 NPU 个数
GPU 使用量	gpu_used	int		某个智算中心已经使用的 GPU 个数
GPU 空闲量	gpu_free	int		某个智算中心空闲的 GPU 个数
MLU 使用量	mlu_used	int		某个智算中心已经使用的 MLU 个数
MLU 空闲量	mlu_free	int		某个智算中心空闲的 MLU 个数

T/AI 123.6-2024

表 4 资源监控字段表（续）

名称	标识符	类型	长度	备注
内存使用量	memory_used	int		某个智算中心已经使用的内存数
内存空闲量	memory_free	int		某个智算中心空闲的内存数
存储使用量	storage_used	int		某个智算中心已经使用的存储大小
存储空闲量	storage_free	int		某个智算中心空闲的存储大小
自定义资源使用量	defineres_used	int		某个智算中心已经使用的自定义资源大小
自定义资源空限量	defineres_free	int		某个智算中心空闲的自定义资源大小
安全等级	security_grade	int		某个智算中心的安全等级

6.2 镜像管理

云际系统对接入智算网络的各个智算中心可用的镜像信息进行管理，包括镜像ID、名称、所属智算中心、操作系统、驱动版本、可支持开发框架、已安装软件包、更新时间等信息，如表5所示。

表5 镜像信息字段表

名称	标识符	类型	长度	备注
编号	image_id	string	小于 64 字节	镜像唯一标识
名称	name	string	小于 64 字节	镜像名称
智算中心	ai_center_id	string	小于 64 字节	所属智算中心
基础环境	base_os	json		基础环境信息，样例：{“base_os”:“”,“cuda_version”:“”}
支持框架	framework	string		支持的算法框架，样例：TensorFlow，PyTorch
匹配资源	resource_spec	string		该镜像可适配的计算资源规格列表，样例：resource_spec_id1，resource_spec_id2
创建时间	created_at	timestamp		创建时间
更新时间	updated_at	timestamp		更新时间

6.3 数据管理

T/AI 123.6-2024

6.3.1 数据监控

云际系统通过查询接口向统一数据存储系统获取数据的标识、名称、版本、存储分布和访问权限等相关信息，进行数据的监控，并将该信息作为作业调度和数据迁移源选取的依据之一。数据资源信息如表6所示。

表6 数据资源字段表

名称	标识符	类型	长度	备注
数据名称	data_name	string	小于64字节	数据名称
数据版本	data_version	string	小于64字节	数据版本
上传时间	created_at	datestamp		上传时间
更新时间	updated_at	datestamp		更新时间
上传人	created_user	string	小于64字节	上传的用户

存储位置	location	json		可能在多个智算中心存储了该数据，在各智算中心的存储情况,样例如下：
				[
				{
				“dataset_url”: “”, //存储路径
				“access”: “r/rw”, //读写权限
				“ai_center_id”:“”, //智算中心
				“download_state”: “”, //之前的下载状态
				“download_rate_limit”: “”, //下载速率的限制
				“concurrency_limit”: “”, // 支持的最大下载并发
				“current_downloads”: “”, // 当前下载数
				“network_topology”: “”, // 智算中心网络拓扑
				“download_cost”, “” //之前的下载平均耗时
				},
				{
				……
				}
				]

6.3.2 数据迁移

云际系统在完成作业调度后，通过监控的数据情况（字段定义见表 6）检查被调度的智算中心是否存储作业运行所需的数据，如果没有，需要从其他的智算中心进行迁移。这时，需要云际系统根据监控的数据情况和数据迁移策略选定迁移的数据源（存储数据的智算中

T/AI 123.6-2024

心）列表，提供给统一存储系统完成数据迁移，见图3。

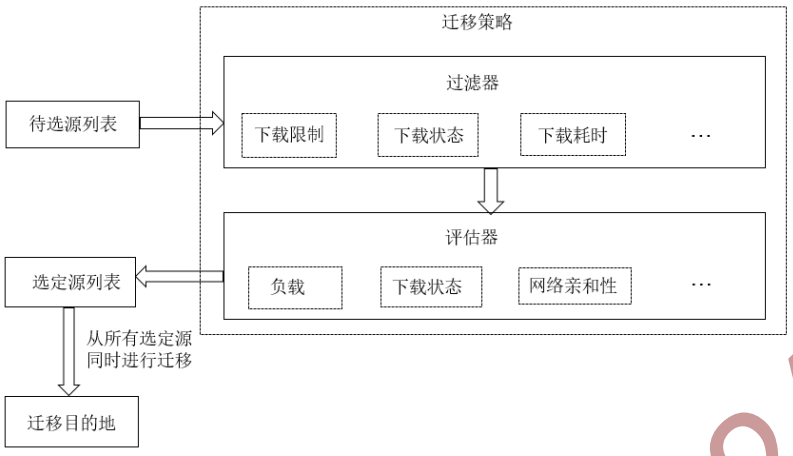

图3 数据迁移过程

6.3.3 数据源特性

数据迁移从一个或者几个数据源向目标数据中心迁移，该过程汇总数据源的选择需要考虑不同数据源的特性。以下列出一部分数据源特性以供参考，系统开发者可选择其中一部分，也可自定义其它特性：

a) 读取权限：其它智算中心是否对本数据有读取权限；
b) 下载状态：本智算中心的本数据的历史下载情况，下载成功、中断和失败的概率；
c) 下载速率限制：本智算中心限制的最高下载速率；
d) 下载并发数：本智算中心限制的最大下载并发数；
e) 当前下载数：本智算中心的当前下载数；
f) 网络拓扑：本智算中心在智算网络中的位置分布信息；
g) 平均下载耗时：本智算中心的本数据的历史下载平均耗时。

6.3.4 数据迁移策略

数据迁移策略将若干数据源和数据源的特性作为输入，经过过滤处理和评估器的评估决策，最终选出最优的数据源列表。云际系统可选择的数据迁移评估策略根据实际的场景需求和实际情况确定。以下列出一部分评估策略以供参考，系统开发者可选择其中一部分，也可自定义其它评估策略：

a) 负载最小优先：根据当前下载数和最大下载并发数的比值，比值小的智算中心评估分数高；
b) 耗时最少优先：历史下载平时耗时短的智算中心评估分数高；
c) 网络拓扑位置亲和性：逻辑位置相近的智算中心评估分数高；
d) 下载状态最优优先：历史下载中，成功率高的智算中心评估分数高；
e) 下载速度限制最大优先：下载速率限制最大的智算中心评估分数高。

6.4 作业管理

6.4.1 统一作业类型

T/AI 123.6-2024

云际系统对各个智算中心的作业进行管理，通过定义统一的作业类型来表示各个智算中心的作业，并在云际系统中保存统一类型的作业对象。云际系统能够对各智算中心发起创建作业、停止作业、获取作业信息、获取作业运行日志操作，并实时监听各个作业的运行状态更新事件。

由于各智算中心运行的原生作业类型不一致，云际系统为每一个智算中心创建一个适配器，见图4。通过调用适配器的统一接口，云际系统间接地与智算中心交互。每个智算中心的适配器通过调用智算中心的原生接口来实现云际系统的操作，并为云际系统屏蔽了各智算中心的接口差异。

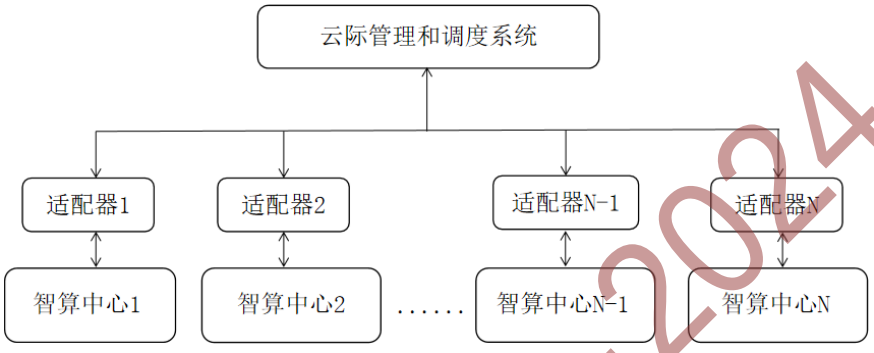

图 4 智算中心适配器

云际系统的作业信息如表7。

表7 云际系统作业定义字段表

名称	标识符	类型	长度	备注
作业id	id	string	小于64字节	作业唯一标识
作业名称	name	string	小于128字节	作业名称
用户id	user_name	string	小于64字节	创建作业的用户id
作业描述	desc	string	小于128字节	作业描述
上传人	created_user	string	小于64字节	上传的用户
子任务	tasks	json		作业包括的子任务信息

云际系统的作业中包括的子任务信息如下表8：

表8 云际系统作业子任务定义字段表

名称	标识符	类型	长度	备注
任务名称	name	string	小于64字节	任务名称
启动命令	command	string		启动命令
镜像标识	image_id	string	小于64字节	镜像标识
资源规格标识	resourcespec	string	小于64字节	资源规格标识

6.4.2 作业生存周期管理

T/AI 123.6-2024

云际系统对作业进行生存周期管理，根据监听或轮询获得智算中心的作业运行状态信息，对云际系统的作业状态信息进行更新。根据实际应用场景，作业管理模块需要输入作业状态策略、重试策略和作业所包括的各个子任务在智算网络中的实时状态信息，见图5。

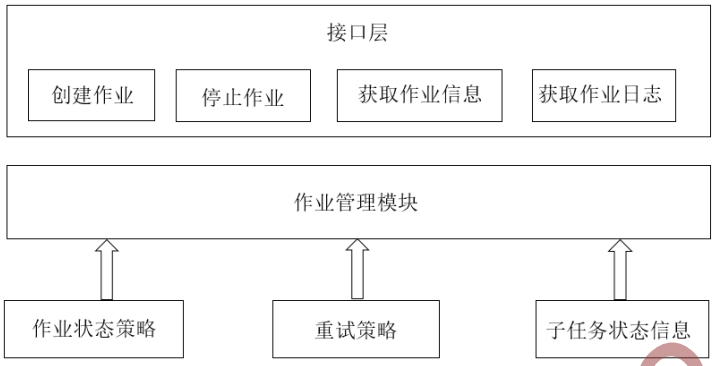

图 5 作业管理

6.4.3 作业状态策略

云际系统可选择的作业状态策略根据实际的场景需求和实际情况确定。以下列出一部分作业状态策略以供参考，系统开发者可选择其中一些，也可自定义其它作业状态策略：

a) 所有子任务运行成功，则作业状态为成功；
b) 任意一个子任务运行成功，则作业状态为成功；
c) 所有子任务运行失败，则作业状态为失败；
d) 任意一个子任务运行失败，则作业状态为失败。

6.4.4 作业重试策略

云际系统可选择的作业重试策略根据实际的场景需求和实际情况确定。以下列出一部分作业重试策略以供参考，系统开发者可选择其中一些，也可自定义其它作业重试策略：

a) 作业状态为成功，不重试；
b) 作业状态为失败，最多重试 N 次，N 为大于 1 的正整数。

6.5 作业调度

云际系统负责对用户提交的作业进行调度，根据各智算中心的资源负载、数据位置、计费因素、通信效率、作业资源需求等因素（简称“调度因子”）将作业所包括的各个子任务调度到合适的智算中心运行。云际系统对智算网络各智算中心具有弹性感知功能，能及时感知各个智算中心的加入与退出、各个智算中心的资源及各种调度参考信息的实时更新。

云际系统下发到调度的智算中心的每个子任务的格式如下表9所示：

表9 子任务字段定义

名称	标识符	类型	长度	备注
任务名称	name	string	小于64字节	任务名称
启动命令	command	string		启动命令
镜像id	image_id	string	小于64字节	镜像id
资源规格id	resourcespec_id	string	小于64字节	资源规格id

T/AI 123.6-2024

云际系统通过流水线的方式调度作业，作业在用户提交之后进入作业队列中进行排队，在每个作业调度周期中，系统从队列中取出作业后使用调度策略对作业和候选智算中心进行处理，见图6。调度策略的实现逻辑由系统开发者定义，调度策略所需输入数据通过智算中心的适配器获取。系统开发者需要创建一个调度结果评估模型，模型对各个调度策略的输出结果进行综合决策，选出最优的智算中心作为作业的调度结果。

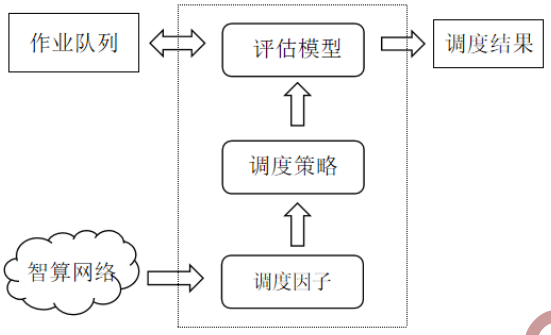

图 6 作业调度过程

6.6 调度策略

6.6.1 调度因子

云际系统可选择的调度因子根据实际的场景需求和实际情况确定。以下列出一部分调度因子以供参考，系统开发者可选择其中一部分，也可自定义其他调度因子：

a) 作业要求：作业对资源的明确要求，如需要保证作业所有的子任务请求的资源都得到满足时才能将各个子任务绑定到相应的智算中心、已明确指定待调度的智算中心等；

b) 资源规格：资源规格字段如表3定义；

c) 数据拓扑：数据资源字段如表5定义。云际系统考虑数据在各个智算中心之间的分布情况，分布式作业子任务在智算网络中的位置分布对子任务间的通信效率和作业的整体运行效率有较大影响；

d) 计费信息：资源规格字段如表3定义。云际系统考虑各个智算中心的资源规格价格因素，各智算中心对作业所使用的计算资源收费不同；

e) 集群负载：资源监控字段如表4定义。当智算中心的集群负载较高时，集群的性能可能出现下降，并且加大了智算中心的运维压力；

f) 能源损耗：各个智算中心由于使用了异构的集群硬件和软件，对能源的损耗并不相同。能源损耗包括：被调度的智算中心在数据传输期间的设备静态能耗和任务执行时的设备计算能耗，见式1。

$$E = \int_{t0}^{ts} P_{static}(t)\,dt + \int_{ts}^{te} P_{exe}(t)\,dt \quad \text{.....(1)}$$

式中：

t_s --为任务开始执行的时间；

t_0 --为任务开始数据拷贝的时间；

$P_{static}(t)$ --设备的静态功率；

t_e --为任务结束执行的时间；

$P_{exe}(t)$ --为任务执行时设备的功率。

T/AI 123.6-2024

g) 算力性能：各个智算中心由于使用了异构的集群硬件和软件，以及不同的管理策略。在使用相同的资源规格的情况下依旧会有不同的算力差异。智算中心性能相关指标可参考智算中心建设相关通用标准；
h) 安全等级：智算中心的安全等级见表4资源监控字段定义表中的安全等级字段。不同的用户和作业对作业的运行环境的安全要求不同，部分保密级别较高的数据和算法代码需要安全等级较高的运行环境。

6.6.2 调度策略

调度策略将若干调度因子作为输入，使用一定的处理逻辑对候选的智算中心进行运算处理，输出每个智算中心对应该调度策略的规范化评分。各个调度策略的输出将作为评估模型的输入，进行进一步的综合决策。云际系统可选择的调度策略根据实际的场景需求和实际情况确定。以下列出一部分调度策略以供参考，系统开发者可选择其中一部分，也可自定义其他调度策略：

a) 集群负载最小优先：集群作业负载数量小的智算中心评分高，负载数量大的智算中心评分低；
b) 资源规格空闲优先：根据作业申请的资源规格，该资源规格空闲的智算中心评分高，繁忙的智算中心评分低；
c) 数据拓扑位置亲和性：对于分布式作业，在满足作业资源规格的前提下，逻辑位置相近的智算中心评分高，位置较远的评分低；
d) 计费价格最低优先：根据作业申请的资源规格，计费价格低的智算中心评分高，价格高的智算中心评分低；
e) 能源损耗最低优先：根据能源损耗公式计算，能源损耗低的智算中心评分高，损耗高的智算中心评分低；
f) 算力性能最高优先：算力性能高的智算中心评分高，性能低的智算中心评分低；
g) 安全等级最高优先：安全等级高的智算中心评分高，等级低的智算中心评分低；
h) 集群负载公平性：优先考虑将作业调度到使得所有智算中心作业负载数量的方差最小的节点上。使得集群负载数量方差越小的节点评分越高。

6.6.3 调度评估模型

多个调度策略的输出将作为评估模型的输入，经过评估模型的计算得出最优的智算中心，获得最终的调度结果。评估模型由系统开发者根据场景需求确定和实现。默认可以采用以下的评估模型：

对于一个等待调度的作业，分别为调度策略$A_1, A_2, \dots, A_n$赋予权重$W_1, W_2, \dots, W_n$，对于智算中心C_k，假设在调度策略$A_1, A_2, \dots, A_n$下为其计算得到的规范化评分分别为$G_{k1}, G_{k2}, \dots, G_{kn}$，则智算中心$C_k$的最终评估得分为本智算中心在各调度策略下的规范化评分与权重值的加权和，见式2。

$$G_k = \sum_{i=1}^{n} G_{ki} \times W_i \qquad (2)$$

式中：

n --正整数，调度策略的个数；
i --正整数，调度策略和权重的编号；
k --正整数，智算中心的编号；
G_{ki} --在调度策略A_i下，智算中心C_k计算得到的规范化评分；
W_i --为调度策略A_i赋予的评估权重。

T/AI 123.6-2024

在所有的智算中心$C_1, C_2, \ldots, C_m$中选出最终评估得分最高的智算中心做为最终的调度结果。

6.7 调度性能评估

6.7.1 调度用时

调度用时T是从云际系统收到作业请求的时点t_r到云际系统下发所有子任务到智算中心完成的时点t_s的时间间隔，见公式3。

$$T = t_s - t_r \quad \ldots\ldots(3)$$

式中：

t_s --云际系统下发所有子任务到智算中心完成的时点；

t_r --云际系统收到作业请求的时点。

6.7.2 调度吞吐率

调度吞吐率是云际系统在单位时间内，针对特定的作业，完成调度到智算中心的作业数量，等于云际系统完成调度的作业数量N与完成作业调度的总用时T_{total}的比值，见公式4。

$$Th = \frac{N}{T_{total}} \quad \ldots\ldots(4)$$

式中：

N --云际系统完成调度的作业数量；

T_{total} --云际系统完成调度的总用时。

7 接口要求

7.1 作业管理接口

7.1.1 创建作业

向云际系统提交创建作业请求，需包括作业名称和各子任务的镜像标识、资源规格、启动命令等参数。

接口要求见附录A.1。

7.1.2 停止作业

向云际系统请求停止某个已经提交的运行作业。

接口要求见附录A.2。

7.1.3 获取作业信息

向云际系统请求获取某个已经运行的作业的详细信息，包括作业运行的状态、所在的智算中心名称、作业标识和资源规格等。

接口要求见附录A.3。

7.1.4 获取作业日志

向云际系统请求获取某个已经运行的作业的日志信息。

接口要求见附录A.4。

7.2 调度策略接口

T/AI 123.6-2024

7.2.1 调度策略评分

调度策略评分接口为每一个智算中心评分，输出为一个整数型的分数。对于每个智算中心，所有的调度策略接口的评分结果将被输入调度评估模型，以此进一步地评估决策。

接口要求见附录 A.5。

7.3 调度评估模型接口

7.3.1 评估模型评分

对于每一个作业，将所有的智算中心在所有调度策略的评分结果作为评估模型的输入。经过评估模型的决策输出最优的智算中心作为作业的调度结果。

接口要求见附录 A.6。

T/AI 123.6-2024

附 录 A
（资料性）
接口说明

A.1 创建作业接口

创建作业接口基本信息见表 A.1

表A.1 创建作业接口基本信息

URI	/openapi/v1/trainjob
HTTP方法	POST
功能	向云际系统提交作业
输入参数	见表A.2、A.3
返回信息	见表A.4、A.5

创建作业接口参数列表见表 A.2

表A.2 创建作业接口参数列表

名称	标识符	类型	长度	是否必选	备注
作业名称	name	string	32	是	
子任务信息列表	tasks	json		是	
子任务信息1	+task1	object			详细参数见表A.3
子任务信息2	+task2	object			
…	…	…			
子任务信息 N	+taskN	object			

task 参数列表见表 A.3

表A.3 task参数列表

名称	标识符	类型	长度	是否必选	备注
子任务名称	name	string	32	是	
启动命令	command	string		是	作业启动命令
镜像ID	imageId	string	32	是	
资源规格ID	resourceSpecId	string	32	是	
副本数量	replicaNumber	int		是	

创建作业返回信息见表 A.4

表A.4 创建作业返回信息

名称	标识符	类型	长度	是否必选	备注
状态码	status	int		是	
错误码	code	string	256	是	
错误信息	errorMsg	string	128	是	

T/AI 123.6-2024

表A.4 创建作业返回信息（续）

返回数据	data	json		是	
作业ID	+jobId	string		是	
子任务运行信息	+taskInfos	json		是	
子任务运行信息1	++taskInfo1	object			详细参数见表A.5
子任务运行信息2	++taskInfo2	object			
…	…	…			
子任务运行信息N	++taskInfoN	object			

taskInfo 参数列表见表 A.5

表A.5 taskInfo参数列表

名称	标识符	类型	长度	是否必选	备注
子任务名称	name	string	32	是	
智算中心ID列表	centerIDs	array		是	副本运行所在的各个智算中心的 ID 列表

A.2 停止作业接口

停止作业接口基本信息见表 A.6

表A.6 停止作业接口基本信息

URI	/openapi/v1/trainjob/{jobId}/stop
HTTP方法	POST
功能	向云际系统停止作业
输入参数	见表A.7
返回信息	见表A.8

停止作业接口参数列表见表 A.7

表A.7 停止作业接口参数列表

名称	标识符	类型	长度	是否必选	备注
作业ID	jobId	string		是	停止的作业ID

停止作业接口返回信息见表 A.8

表A.8 停止作业接口返回信息

名称	标识符	类型	长度	是否必选	备注
状态码	status	int		是	
错误码	code	string	256	是	
错误信息	errorMsg	string	128	是	

T/AI 123.6-2024

A.3 获取作业信息接口

获取作业信息接口基本信息见表 A.9

表A.9 获取作业信息接口基本信息

URI	/openapi/v1/trainjob/{id}
HTTP方法	GET
功能	向云际系统获取作业信息
输入参数	见表A.10
返回信息	见表A.11、A.12

获取作业信息接口参数列表见表 A.10

表A.10 获取作业信息接口参数列表

名称	标识符	类型	长度	是否必选	备注
作业ID	jobId	string		是	获取信息的作业ID

获取作业信息接口返回信息见表 A.11

表A.11 获取作业信息接口返回信息

名称	标识符	类型	长度	是否必选	备注
状态码	status	int		是	
错误码	code	string	256	是	
错误信息	errorMsg	string	128	是	
返回数据	data	json		是	
作业ID	+jobId	string		是	
作业名称	+name	string		是	
作业状态	+status	string		是	pending#等待执行 running#正在执行 completed#执行完成 failed#执行失败
子任务信息	+taskInfos	json		是	
子任务信息1	++taskInfo1	object			详细参数见表A.12
子任务信息2	++taskInfo2	object			
…	…	…			
子任务信息N	++taskInfoN	object			

taskInfo 参数列表见表 A.12

表A.12 taskInfo参数列表

名称	标识符	类型	长度	是否必选	备注
子任务名称	name	string	32	是	
启动命令	command	string		是	作业启动命令

T/AI 123.6-2024

表A.12 taskInfo参数列表（续）

镜像ID	imageId	string	32	是	
资源规格ID	resourceSpecId	string	32	是	
副本数量	replicaNumber	int		是	
智算中心ID列表	centerIDs	array		是	副本运行所在的各个智算中心的ID列表
副本运行状态列表	replicaStatus	array		是	pending#等待执行 running#正在执行 completed#执行完成 failed#执行失败

A.4 获取作业日志接口

获取作业日志接口基本信息见表 A.13

表A.13 获取作业日志接口基本信息

URI	/openapi/v1/trainjob/{id}/task/{taskIndex}/replica/{replicaIndex}/log
HTTP方法	GET
功能	向云际系统获取作业日志
输入参数	见表A.14
返回信息	见表A.15

获取作业日志接口参数列表见表 A.14

表A.14 获取作业日志接口参数列表

名称	标识符	类型	长度	是否必选	备注
作业ID	jobId	string		是	获取日志的作业ID
子任务ID	taskId	string		是	获取日志的子任务ID
副本ID	replicaId	string		是	获取日志的副本ID

获取作业日志接口返回信息见表 A.15

表A.15 获取作业日志接口返回信息

名称	标识符	类型	长度	是否必选	备注
状态码	status	int		是	
错误码	code	string	256	是	
错误信息	errorMsg	string	128	是	
返回数据	data	json		是	
日志文件URL	+logUrl	string		是	作业日志文件的下载URL

A.5 调度策略评分接口

T/AI 123.6-2024

调度策略评分接口基本信息见表 A.16

表A.16 调度策略评分接口基本信息

URI	/scheduler/algorithm/{algorithmId}/score/{centerId}
HTTP方法	GET
功能	根据输入的策略ID和智算中心ID，获取该智算中心评分
输入参数	见表A.17
返回信息	见表A.18

调度策略评分接口参数列表见表 A.17

表A.17 调度策略评分接口参数列表

名称	标识符	类型	长度	是否必选	备注
策略ID	algorithmId	string	32	是	
智算中心ID	centerId	string	32	是	

调度策略评分接口返回信息见表 A.18

表A.18 调度策略评分接口返回信息

名称	标识符	类型	长度	是否必选	备注
状态码	status	int		是	
错误码	code	string	256	是	
错误信息	errorMsg	string	128	是	
返回数据	data	json		是	
分数	+score	int		是	

A.6 评估模型评分接口

评估模型评分接口基本信息见表 A.19

表A.19 评估模型评分接口基本信息

URI	/scheduler/evaluate/score/{centerId}
HTTP方法	GET
功能	根据输入的调度策略ID列表、调度策略评分列表和智算中心Id，获取该智算中心的最终评分
输入参数	见表A.20
返回信息	见表A.21

评估模型评分接口参数列表见表 A.20

T/AI 123.6-2024

表A.20 评估模型评分接口参数列表

名称	标识符	类型	长度	是否必选	备注
调度策略ID列表	algorithmIds	array	32	是	
调度策略评分列表	algorithmScores	array	32	是	
智算中心ID	centerId	string	32	是	

评估模型评分接口返回信息见表 A.21

表A.21 评估模型评分接口返回信息

名称	标识符	类型	长度	是否必选	备注
状态码	status	int		是	
错误码	code	string	256	是	
错误信息	errorMsg	string	128	是	
返回数据	data	json		是	
分数	+score	int		是	

T/AI 123.6-2024

参 考 文 献

[1] ISO/IEC 22237-1:2021 Information technology — Data centre facilities and infrastructures — Part 1: General concepts

[2] ISO/IEC/IEEE 24765:2017 Systems and software engineering — Vocabulary

[3] ITU-T Y.2501 Computing power network — Framework and architecture

T/AI 123.6-2024