ICS 35.020
CCS L70

团 体 标 准

T/GBA XXX-2023

大湾区算力网络总体技术要求

General requirements for computing network of the Greater Bay Area

（征求意见稿）

XXXX-XX-XX 发布 XXXX-XX-XX 实施

粤港澳大湾区标准创新联盟 发布

目　　次

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》的规定起草。

本文件由鹏城实验室提出。

本文件由粤港澳大湾区标准创新联盟归口。

本文件授权粤港澳大湾区标准创新联盟组织伙伴和所有成员单位使用，联盟组织伙伴需等同采用转化为自身团体标准，并在全国团体标准信息平台上公开标准基本信息。

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。

本文件起草单位：鹏城实验室、中国联合网络通信有限公司广东省分公司、中国联合网络通信有限公司研究院。

本文件主要起草人：

本文件为首次发布。

大湾区算力网络总体技术要求

1 范围

本文件规定了大湾区算力网络的术语、系统架构、功能要求、各功能模块间的接口要求。

本文件适用于大湾区算力网络的总体设计和建设。

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

YD/T 4255-2023 算力网络 总体技术要求

3 术语和定义

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

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 network

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。

[参考：ITU-T Y.2501,有修改]

3.5

算力感知 computing awareness

算力感知是网络对算力资源和算力服务的部署位置、实时状态、负载信息、业务需求等多维度感知。
[来源：YD/T 4255-2023]

3.6

算网编排管理 computing & network orchestration and management

算网编排管理是对算力资源和网络资源进行统一管理和编排，包括注册、OAM等。
[来源：YD/T 4255-2023]

4 缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）
CPU：中央处理单元（Central Processing Unit）
DetNet：确定性网络（Deterministic Networking）
DNS RR：域名系统资源记录（Domain Name System Resource Record）
FCFS：先来先服务（First Come First Service）
IP：网际互连协议（Internet Protocol）
I/O：输入/输出（Input/Output）
OTDR：光时域反射仪（Optical Time Domain Reflectometer）
OTN：光传输网络（Optical Transport Network）
OXC：光交叉连接（Optical Cross Connect）
OMSP：光复用段保护（Optical Multiplex Section Protect）
OSU：光服务单元（Optical Service Unit）
QKD：量子密钥分发（Quantum Key Distribution）
QoE：体验质量（Quality of Experience）
QoS：服务质量（Quality of Service）
QUIC：快速 UDP 互联网连接（Quick UDP Internet Connection）
RDMA：远端内存直接访问(Remote Direct Memory Access)
ROADM：可重构光分插复用器（Reconfigurable Optical Add-Drop Multiplexer）
SDWAN：软件定义广域网（Software Defined WideArea Network）
SRV6：基于 IPv6 的段路由（Segment Routing IPv6）
TCP：传输控制协议（Transmission Control Protocol）
UDP：用户数据报协议（User Datagram Protocol）
WDM：波分复用(Wavelength Division Multiplexing)
WSON：波分交换光网络（Wavelength Switched Optical Network）

5 总体架构

5.1 总体架构

大湾区算力网络连接分散在大湾区域内的智算中心、超算中心以及通用云计算中心，汇聚和共享算力、数据、模型和应用等资源。各类计算中心通过加入大湾区算力网络实现资源共享，并通过算力网络统一调度，提高全网资源利用率，满足算力需求。

大湾区算力网络的业务分层和系统设计见图1。

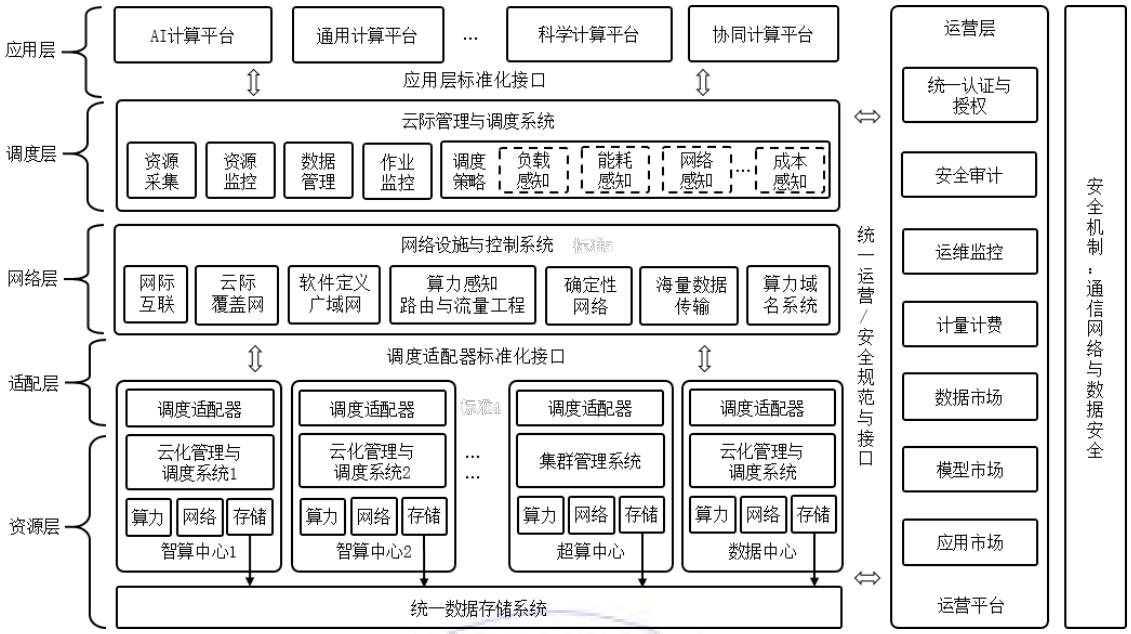

图1 大湾区算力网络总体架构

5.2 部署架构

大湾区算力网络中各计算中心的互联参考部署架构见图2，其中：

a) 计算中心之间可通过以下两种方式进行互联：

1） IP 专网：主要用于路由控制面信令，计算中心间低通量数据交互且对安全防护、网络质量有一定要求的传输场景；

2） OTN 高速网络：主要用于计算中心间高通量数据交互且对安全防护、网络质量有一定要求的传输场景；

3） 互联网：主要用于计算中心间低通量数据交互且对安全防护、网络质量无要求的传输场景。

b) 枢纽节点为某区域内的大型/重要智算、超算或通用云计算中心，枢纽节点之间的互联方式优先级由高到低为：OTN 高速网络，IP 专网，互联网；

c) 其它计算中心之间的互联方式优先级由高到低为：IP 专网，OTN 专线高速网络，互联网。

d) 大湾区算力网络平台包括云际管理和调度平台、运营平台，可部署在某个枢纽节点计算中心内，或部署于独立的服务集群。

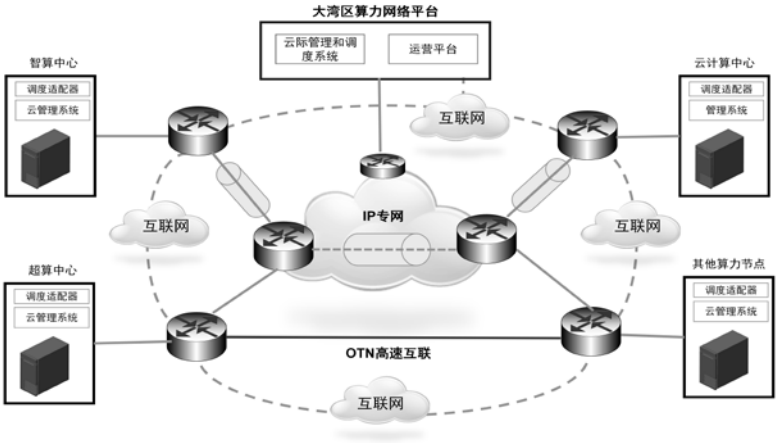

图 2 大湾区算力网络部署架构

6 功能要求

6.1 概述

大湾区算力网络总体架构可分为资源层、适配层、网络层、调度层、应用层、运营层和安全机制，各业务层的具体功能要求如下。

6.2 资源层

大湾区算力网络资源层包含各计算中心的算力、存储、网络、数据等资源，应具备以下功能：

a) 算力资源提供：应提供计算、存储等资源，并支持按需扩缩容；
b) 计算中心互联与开放：应实现各计算中心算力和数据资源的互联互通，支持各计算中心算力和数据资源的对外开放；
c) 统一数据存储：应在现有计算中心的存储资源上构建统一的支持块/文件/对象的融合存储服务，支持跨计算中心的数据管理和迁移；
d) 用户认证：应提供用户创建、删除、修改、鉴定等功能，保障算力网络中授权的用户登录计算中心进行资源使用；
e) 作业管理：提供对作业的提交、取消、查询、变更等功能；
f) 数据中心内资源管理：对计算中心资源（如CPU、GPU、内存等）进行实时监控，为作业分配资源，并实现资源隔离，以确保敏感数据的安全和不同作业之间的隔离；
g) 账号管理：提供计算中心云化管理调度系统账号创建、删除、修改等功能；
h) 数据中心内调度策略：计算中心云化管理调度系统应支持多种调度策略，以面对算力网络复杂需求场景。可根据需要选择不同调度策略，也可以自定义调度策略；
i) 稳定性：用户认证系统、云化管理调度系统等均应保障持续稳定运行，从而为算力网络提供稳定可靠的算力资源。

6.3 适配层

大湾区算力网络适配层通过调度适配器，实现调度层与资源层之间的数据交互。调度适配器部署在各计算中心内，对接各计算中心内异构的云化管理与调度系统，负责异构集群资源采集和上报，实现作业转发和管理。调度适配器应支持从各计算中心采集集群资源、负载、能耗、费率等信息并上报给调度层，由调度层根据这些信息进行作业调度。适配层应包含以下功能：

a) 应支持智算中心、超算中心、数据中心三类大型计算中心对外接口的统一适配，屏蔽异构集群技术栈差异；
b) 应支持接口扩展，满足其它类型的计算中心接入大湾区算力网络的要求；
c) 作业代理：应支持将调度层下发的作业发送到计算中心本地调度系统，周期性采集作业状态的资源使用信息，并上报到调度层；
d) 算力资源代理：应支持周期性采集所属计算集群的资源信息并上报给调度层；
e) 数据管理与缓存：应支持适配层和调度层之间的数据管理，包括数据上传、下载、断点续传等。支持作业调度时的数据下发，和作业执行结束后的数据取回。应支持缓存作业数据，避免使用重复数据时数据的再次传输，提高数据利用率；
f) 账号映射：应支持算力网络统一用户账号到各个计算中心云化管理调度系统的用户账号之间的映射。

6.4 网络层

6.4.1 概述

网络层主要负责多个异构计算中心的网络接入和互联，实现算力数据的路由控制和高速转发。网络层主要包含光层和IP层。

6.4.2 光层

大湾区算力网络应支持以全光网络为底座，为重要枢纽节点间的数据交互提供低时延、高通量的传输通道。光层具体包含以下能力：

a) 应支持100G/400G/800G等WDM/OTN长距离大容量传输能力，实现算力资源互联和海量数据高速传输；
b) 应支持ROADM/OXC等全光交换和调度能力，例如9维，20维和32维等；
c) 应支持OTN/OSU/WDM等多种方式的全光算力接入能力；
d) 应支持基于WSON/OMSP等多种方式的保护和协同机制，提升网络层生存性和可靠性；
e) 宜支持光层和IP层跨层协同，包括业务、路径、运维、保护等多方面的协同。

6.4.3 IP层

大湾区算力网络需要构建一张IP专网，支持全面覆盖大湾区的各大重要算力枢纽节点，支持口字型、交叉型等多种组网方式，支持IPv6+、软硬切片、SDWAN等技术。IP层具体包含以下能力：

a) 用户接入：应支持个人用户和政企用户通过互联网、VPN专线等方式访问算力网络的资源；
b) 算力感知：应支持通过监测网络中的计算资源（如 CPU、内存、存储等）的使用情况，并根据这些信息进行路由计算来选择最佳的路径；应支持对当前计算资源和网络资源的感知和理解，能够通过实时监测和分析计算资源的性能指标（如 CPU 利用率、内存占用率、磁盘 I/O 等）和网络资源的性能指标（如带宽、时延、抖动等），来识别系统的瓶颈和热点，并根据用户的需求和优先级来动态调整资源分配，以最大程度地提高算力资源的性能和效率；
c) 应用感知：宜支持对网络流量进行深度分析和识别，实现对网络应用的感知和优化，针对性地进行网络优化和调整，包括流量路径的调整、优先级的调整等；宜支持利用网络协议报文（例如SRv6报文）的可编程空间，将应用信息（标识和/或网络性能需求）携带进入网络，使能网络感知应用及其需求，为其提供精细的网络服务和精准的网络运维；支持为算力提供差异化的服务能力，将流量引导进入相应的SRv6 Policy隧道、网络切片、DetNet（确定性网络）路径、服务功能链路径等，实现应用分流和灵活选路；
d) 服务感知：宜支持通过用服务标识的方式来对具体的服务进行表示，并可用于应用的任务调用、端到端连接、网络服务路由、算力的寻址和调度；宜支持通过对服务标识的感知，实现服务级颗粒度的SLA需求感知；支持通过服务标识索引关联对应的转发表项，从而完成对服务的精细化感知和流量转发；
e) 软件定义广域网：应支持广域网的自动化管理和控制，实现网络可编程；应支持灵活地部署和管理多个分支机构之间的网络连接；应支持业务快速部署能力；应支持通过动态路径选择和故障转移等技术，提高网络的可靠性；
f) 确定性网络：应支持以算力资源的需求为导向，通过软/硬切片能力将不同业务进行切片承载，实现时延、带宽的区隔；应支持基于SRV6技术实现业务的端到端快速开通；应支持各类数据传输协议，包括UDP、TCP、QUIC等，以保证数据的传输质量和可靠性；
g) 海量数据传输：宜支持数据压缩和分片技术，减小数据的大小，降低网络带宽的占用；宜支持数据快递服务，充分利用现有网络的冗余带宽，通过控制器，实现高效按需的弹性数据上云服务。

6.5 调度层

大湾区算力网络调度层主要实现跨计算中心之间的资源管理和全局作业调度，应包含以下功能：

a) 跨数据中心算力资源管理与监控：应支持各计算中心算力、存储、网络资源的实时信息采集与监控，支持各类算力资源的统一度量和管理；
b) 算网编排：应支持按需实现算力、网络等资源的编排；
c) 数据资源管理：应支持算力网络上数据资源的监控、管理，支持数据的上传、下载、断点续传和缓存等；
d) 作业管理：应支持提交作业，查看作业列表，以及对作业的管理操作；

e) 计算资源管理：应支持全网算力、网络、存储等资源的管理，包括调度适配器上报的资源和负载信息；
f) 跨数据中心调度策略：应支持根据算力网络的算力资源、数据资源情况以及业务需求，选择合适的计算中心运行计算作业，应至少支持以下调度策略中的一种：
   1) 手动调度策略：人工指定计算中心运行作业；
   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，选择最合适的计算中心运行作业并通过调度将数据搬移到该计算中心；对于不适合数据迁移的大数据或者隐私数据场景，感知数据所在计算中心并将作业调度到该计算中心运行；
   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；
g) 计算作业调度：应支持根据调度策略，将计算作业分发到相应计算中心；
h) 算力域名系统：宜支持基于域名解析机制的算力网络资源发布与订阅，基于DNS RR表达算网资源信息，实现算网资源的发布与订阅。

6.6 应用层

应用层负责大湾区算力网络与用户应用平台的信息交互。用户应用平台包括AI计算平台、通用计算平台、科学计算平台等。大湾区算力网络应用层应包含以下功能：

a) 应支持用户通过查询，获取大湾区算力网络可用的算力节点、算力资源、网络资源、公开数据集等；
b) 应支持用户向大湾区算力网络发起资源申请请求和业务接入请求；
c) 应支持用户计算作业创建、查询、终止、删除、获取计算结果等功能。

6.7 运营层

运营层实现多个计算中心（含超算、智算和通用云计算中心）的算力、网络、数据等资源的统一运营，应包含以下功能：

a) 统一认证与授权：应支持账户注册、认证和授权，确保账户全局统一，可在多个计算中心使用；
b) 统一计量计费：应支持对算力资源和服务的使用情况统计，并生成订单，用于费用结算；
c) 资源运营可视化：应支持对算力资源、网络资源、基础设施可视化管理；
d) 资源监控和告警：应支持对计算中心的网络、算力等资源进行实时的信息采集和故障检测，并可以生成告警日志；
e) 统一算力控制台：应支持各大计算中心的算力统一注册和接入，支持计算中心、网络资源的配置、加入、退出算力网络；
f) 算力交易中心：应支持全网算力资源供给情况的搜索和查看，支持用户采用包年包月、计量等方式订购算力服务；
g) 产品管理：应支持算网产品，解决方案的发布、上下架、打标签等能力；
h) 数据交易中心：应支持数据提供商发布数据，支持用户订阅交易和使用数据服务；
i) 算网度量：应支持针对计算中心不同规格型号和品牌的基础算力设备、存储设备和网络条件进行算力、存力和运力进行评估以及等级定义，并面向用户提供统一的抽象描述，实现对外服务的算力、存力和运力逻辑单元产品的标准化。

6.8 安全机制

大湾区算力网络安全机制需要从数据安全、网络安全、传输安全、安全评估、安全管理等多个方面为算力网络提供安全保护，应包含以下功能：

a) 数据安全：数据安全能力可为算力网络业务提供数据隐私保护、关键数据存证、计算结果验证等数据安全服务，具体应支持以下能力：
   1） 应支持重要数据备份和恢复，避免数据丢失或损坏，确保系统数据的完整性和可靠性。
   2） 应支持数据存储加密，防止重要数据被篡改；
   3） 应支持敏感数据隔离存储，并采用访问控制技术，仅允许授权用户或节点访问，避免非法访问和数据泄露；
   4） 应支持数据追溯功能，可以追溯数据的来源、流转路径和处理过程，提高数据的可信度；
   5） 应支持引入机密计算的算力，构建高可信计算环境，为高安全需求业务服务；
   6） 宜支持引入安全多方计算可保证数据在计算节点的机密性；
   7） 宜支持区块链，通过区块链对账和存证可保证账单数据可信、不可篡改，保障算力网络业务各参与方对业务交易的可信；
b) 网络安全：网络安全能力支持网络攻击发现、收到攻击及时相应处置等功能，以保障算力网络正常运转，具体应支持以下能力：
   1） 应支持算力网络边界访问控制，并对算力网络的业务流量检查，防止未授权的恶意访问；
   2） 应支持网络边界入侵检测和防御，并对计算节点和业务进行攻击监控；
   3） 应支持态势感知，实时展示算力网络中存在安全风险状况和发展趋势，并能够对安全风险进行处置响应；
c) 传输安全：传输安全能力提供各大计算中心之间的数据传输加密和认证，可以支持多种加密传输方式，具体应支持以下能力：
   1） 应支持通过混沌算法、噪声扰动等方式实现物理层信息安全传输；
   2） 应支持基于OTNsec方式的加密传输；
   3） 宜支持基于QKD的量子加密传输；
   4） 宜支持通过OTDR等方式实现对光网络物理层的状态感知，防泄漏和窃听等；
d) 跨境安全:跨境安全能力提供省内和香港、澳门两地数据交互传输的数据治理和安全防护能力，保障跨境数据流通符合国家相关法律法规要求；
e) 安全评估：安全评估能力提供对算力节点的安全测评、基线核查、漏洞扫描等能力，保障计算节点的安全性，具体应支持以下能力：
   1） 应支持安全测评，检测评估计算节点的主机安全、虚拟化安全、网络安全等安全防护机制和安全配置情况，以及计算节点是否有机密计算、隐私计算、或其它安全服务能力；
   2） 应支持基线核查，核查计算节点的安全防护机制和安全配置是否满足安全要求；
   3） 应支持漏洞扫描，对计算节点进行漏洞扫描，及时检测和发现计算节点的脆弱性；
f) 安全管理：安全管理能力应支持能力管理、安全业务管理视图、操作审计等功能，并能呈现安全态势等供安全管理员分析与决策。具体应支持以下能力：
   1） 应支持安全能力的注册、能力的管理和配置；
   2） 应支持算力网络和业务调用安全能力层资源进行管理和协调；
   3） 应支持操作审计，通过对算力网络产生的日志进行分析，对正常流程、异常状态和安全事件等进行记录和监控，形成操作审计信息并及时进行安全分析和评估；
   4） 宜支持对安全需求分析和安全能力感知功能进行管理、配置和呈现。

7 接口要求

7.1 概述

大湾区算力网络各业务层之间通过不同的接口进行相互访问，完整的算力网络系统应包含以下接口。

7.2 资源层与适配层间接口

资源层与适配层间通过命令行/RESTful/Socket等接口方式实现智算中心资源和负载信息的获取，并实现作业的提交和管理。

7.3 适配层与调度层间接口

调度层通过适配层提交、查询和管理作业。适配层向调度层上报智算中心资源和负载信息。

7.4 运营层与调度层间接口

运营层从调度层获取大湾区各计算中心的资源和负载信息，并提交、查询和管理作业。

7.5 应用层与运营层间接口

应用层通过调用运营层接口进行统一认证和授权。
运营层通过调用应用层接口获取模型和数据信息。

7.6 应用层与调度层间接口

应用层可以直接连接调度层并通过调度层接口提交和管理作业，通过调度层的调度策略，将应用层的计算需求准确和快速地路由到对应的计算节点。

附 录 A

（资料性）

无人机监管城市应用案例

在该应用场景下，无人机按预先规划好的路线对城市特定区域进行巡视航拍后，基于大湾区算力网络，将航拍的图片和视频数据快速回传给部署在云计算中心的后台。后台基于算力网络调度层选择合适的智算中心推理节点，把相关的图片和视频数据快速传送给智算中心进行推理分析，智算中心再把分析结果反馈给后台。后台将分析结果同步给监管部门相关平台，若分析结果存在异常情况（如商店起火、杂物堵路等），则及时向监管人告警。

由于无人机采集的图片和视频数据量比较大，为了保障时效性，需要为图片和视频数据回传后台，以及跨计算中心数据交互等场景，提供高速和低时延的网络环境。同时，需要支持各类调度策略，实现各个 AI 推理分析节点的算力灵活调度。

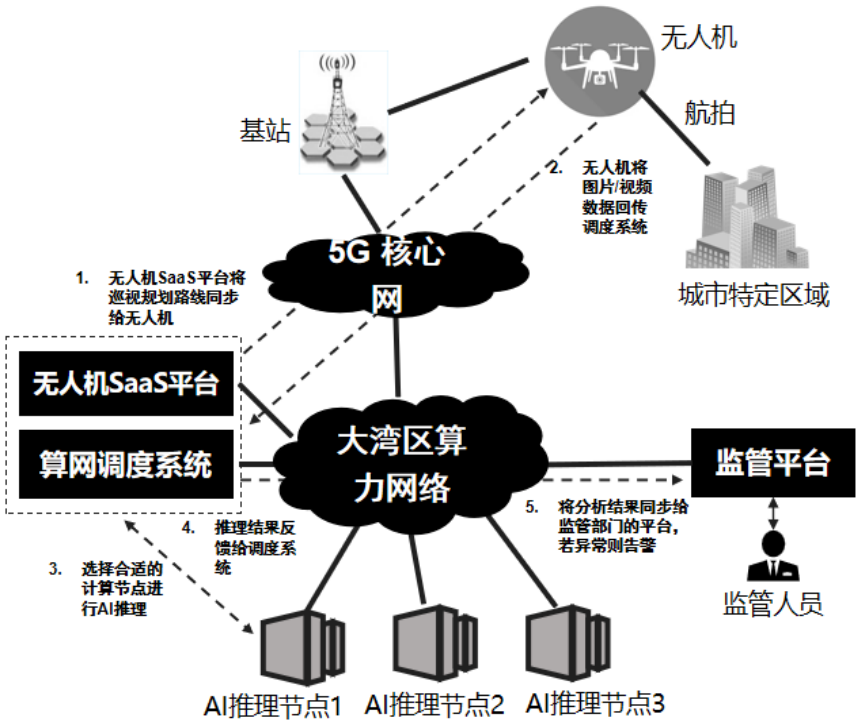

图 A.1 无人机监管城市示意图

附 录 B

（资料性）

智能安防应用案例

在该应用场景下，用户通过 R1 节点接入网络访问 MEC 资源，正常情况下，视频业务流接入到最近的 MEC-1 节点。但是，假设在该时间点 MEC-1 的资源利用率接近 90%，处于高负载状态，视频会频繁出现卡顿现象。通过大湾区算力网络的按需调度能力，自动将视频业务引流至最优计算节点 MEC-2 处理，实现视频业务流不中断，视频更加流畅，从而大幅提升用户体验。

由于各 MEC 节点存在算力资源分配差异、算力负载不均等问题，可能导致本地 MEC 无法完全满足业务实时性和确定性的需求。通过算力感知协议，将所需要的算力资源要求，嵌入网络的控制层信息中，通过网络控制平面分发网络中的算力资源信息，并根据算力资源信息将计算任务调度到满足用户需求的节点，实现对算力资源感知和优化调度。

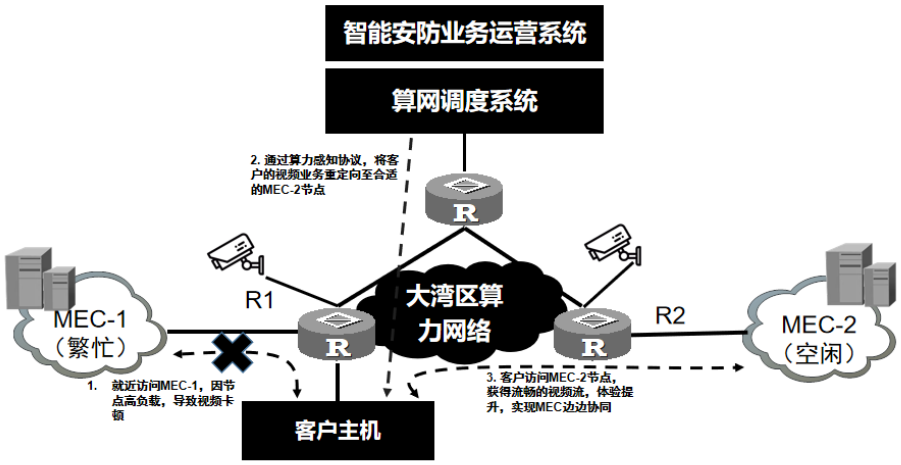

图 B.1 智能安防应用示意图

附 录 C

（资料性）

3D 实时渲染应用案例

在该应用场景下，用户登录实时云渲染门户，选择算力规格，提交其设计的 3D 模型，算网调度系统根据用户的算力需求以及用户接入位置，就近选择渲染边缘节点完成云端渲染。渲染完成后，门户会返回一个发布 url 给用户，用户访问该 url 可以获取 3D 实时渲染效果，并可以实时对模型进行互动操作（如放大、缩小、拉近、拉远等）。

由于用户需要对模型进行实时互动操作，要求网络的传输时延达到毫秒级，用户才能获得较好的体验。因此，需要基于算力网络的感知能力，为 3D 实时渲染应用选择具备相应算力资源且靠近用户的渲染边缘节点提供服务，保障用户对模型实时互动操作过程中不会出现掉帧、卡顿等问题。

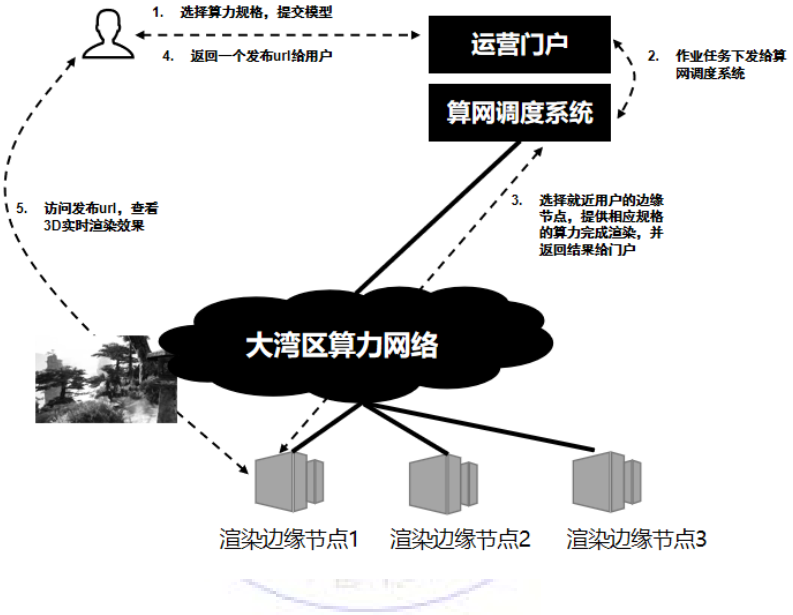

图 C.1 3D 实时渲染应用示意图

附 录 D

（资料性）

AI 算力产品标准化应用案例

在传统方式下，用户在选择购买 AI 算力产品时，需要根据 GPU 厂商提供的浮点算力值选择物理 GPU 的形式，对用户的专业技术要求较高。AI 算力产品标准化应用方案是自定义一个 AI 算力的基准，然后隐藏真实的 GPU 设备型号，只是基于参考的算力基准提供虚拟 GPU 售卖。用户可以直观地基于标准的 AI 算力模型应用的参考性能值去判断选择，令用户更容易判断。这样做的好处就是云厂商可以利用 GPU 虚拟化软件自定义 GPU 设备，可以合理的利用新旧不同 GPU 设备池化，将整个 GPU 资源集群的利用率大大提高若干倍，从而令单位算力成本最低。

举个例子，基于典型的普适性的 AI 模型 Resnet50 为基准，对资源池里的 GPU 设备进行测试，并基于测试的数据将不同规格的 GPU 设备划分不同算力等级，例如分为基础版 B（Basic）、标准版 S（Standard）、高级版 P（Premium）三大等级，而且支持在该 B/S/P 三大等级里面具体根据运营需求和用户需求细分了更小颗粒度的算力等级实例，最终售卖的 AI 算力产品规格的样例示意如下（以基础版为例）：

表 D.1 AI 算力产品规格的样例示意表

第一级分类	第二级分类	第三级分类	实例规格	算力规格
基础版 - B (Basic)	B1	small	# GPU 卡数：1 # GPU 每卡显存：6 GB # CPU 核数：4 # 内存：12 GB	B1.small ： 吞吐量约为 145 images/sec # ResNet50 模型，训练精度 fp16 # ImageNet2012 数据集 # BatchSize = 112
		medium	# GPU 卡数：1 # GPU 每卡显存：12 GB # CPU 核数：4 # 内存：12 GB	B1.medium: 吞吐量约为 390 images/sec # ResNet50 模型，训练精度 fp16 # ImageNet2012 数据集 # BatchSize = 224
		large	# GPU 卡数：1 # GPU 每卡显存：24 GB # CPU 核数：8 # 内存：24 GB	B1.large： 吞吐量约为 801 images/sec # ResNet50 模型，训练精度 fp16 # ImageNet2012 数据集 # BatchSize = 448

附 录 E

（资料性）

算力网原生的大模型应用案例

在该应用场景下，大湾区算力网络在区域内大型算力中心（即枢纽节点）部署基座大模型，通过调用多地算力及数据资源，支撑面向各行业应用的大模型微调、蒸馏、推理部署等全流程开发工作。用户登陆模型开发平台，选择基座模型、下游任务数据以及算力规格，提交模型微调请求，大湾区算力网络调度系统根据算力需求以及数据位置，选择合适的节点完成模型微调。微调完成后，用户可下载微调模型进行使用，也可通过模型开发平台提交模型部署请求，算力网络调度系统根据模型规模及用户需求选择合适的一个或多个计算中心，完成模型的推理服务部署，并发布模型调用 API。模型使用者通过 API 提交模型推理请求，算力网根据模型部署位置及各计算中心资源状态进行任务编排与服务调度，选择合适的部署节点完成模型推理服务。同时，在获取用户许可的前提下，模型开发平台通过 API 收集用户反馈信息，将相关数据反馈至模型提供者及基座模型，对大模型进行持续更新。

由于用户的应用需求，大模型的持续学习、微调、推理可能在不同计算中心完成，因此需要算力网络能够有效屏蔽各计算中心从底层芯片、存储、网络到上层系统接口的异构性，使得同一模型可以快速迁移到不同的计算中心完成相应的训练、推理操作，支撑大模型应用生态建设。

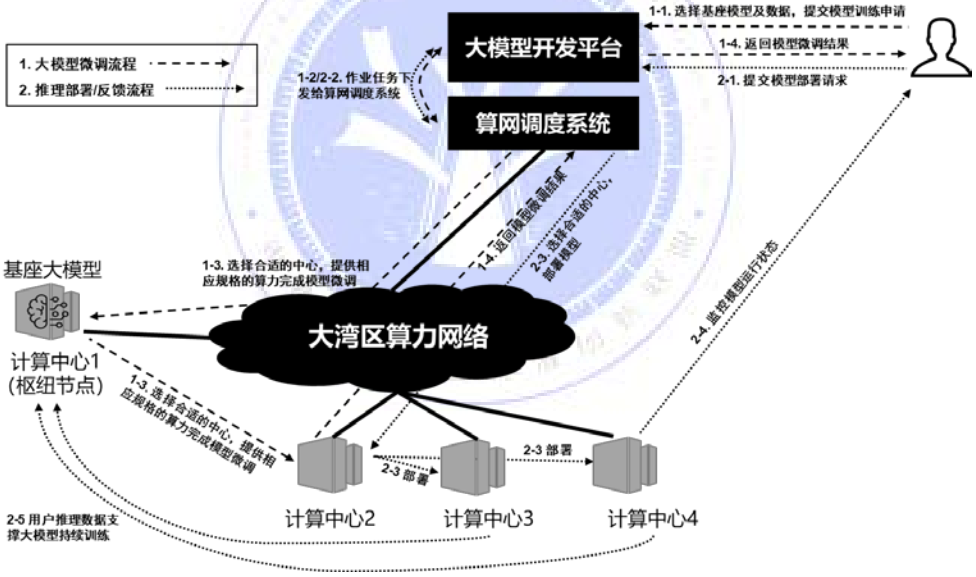

图 E.1 算力网原生的大模型应用示意图

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

[4] YD/T 4255-2023 算力网络 总体技术要求