ICS 35.240.01
CCS L 70

团　　体　　标　　准

T/CESA 1463—2025
T/CIIA 063—2025

信息技术 智算服务 异构算力虚拟化及池化系统要求

Information technology — Intelligent computing services —Technical requirements for heterogeneous computing power virtualization and pooling

2025－12－29 发布　　2025－12－29 实施

中国电子工业标准化技术协会
中国信息协会　　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。

本文件由中国电子技术标准化研究院提出。

本文件由中国电子工业标准化技术协会和中国信息协会共同归口。

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：中石油（北京）数智研究院有限公司、国家电网有限公司信息通信中心、北京睿思智联科技有限公司、中信建投证券股份有限公司、中国农业发展银行、中国铁塔股份有限公司、华夏银行股份有限公司、新华人寿保险股份有限公司、中国电子技术标准化研究院、山东省计算中心（国家超级计算济南中心）、中信证券股份有限公司、广发证券股份有限公司、中国光大银行股份有限公司、中国人寿财产保险股份有限公司、中国大地财产保险股份有限公司、太保科技有限公司、泰康科技有限公司、国联民生证券股份有限公司、龙盈智达（北京）科技有限公司、中原证券股份有限公司、山西证券股份有限公司、中煤（深圳）研究院有限责任公司、中智管理咨询有限公司、中国盐业集团有限公司、山东睿创未来技术有限公司、第四范式（北京）技术有限公司、云尖信息技术股份有限公司、北京百度网讯科技有限公司。

本文件主要起草人：李禹宏、李昕哲、白俊杰、刘军、闫龙川、牛佳宁、金萌、杨世琪、傅帅、李剑戈、周立斌、刘达、马骁、曲嘉彬、徐常智、李惠民、龚伟华、王彦博、关杏元、刘硕、陆鸣、谢志豪、陈志峰、秦龙、王继彬、王春晓、徐崚峰、邓华良、苏兆聪、解培、孔宇飞、路则明、涂炯、王辉、段红帅、刘雪华、吴凌智、金建新、张月、杨璇、杨林杰、喻荟铭、徐得景、许慧青、周晶、庞丽敏、问梁军、张秋萍、贾士朋、卢冕、柏青、张俭锋、谢伟光、高宏玲、艾文思、翟腾、刘牧鑫。

信息技术 智算服务 异构算力虚拟化及池化系统要求

1 范围

本文件确立了异构算力虚拟化及池化系统架构，规定了异构算力虚拟化、异构算力池化、异构算力调度等方面的通用要求。

本文件适用于为智算中心和模型研发机构等在异构算力的统筹、调度和管理方面提供技术指导，并为上述组织选型算力服务产品提供参考。算力服务商也可参照本文件进行产品规划与设计。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

异构算力 heterogeneous computing power

由不同类型的计算硬件（如 CPU、GPU、FPGA、NPU、ASIC 等）组成的计算能力。

3.2

异构算力虚拟化 heterogeneous computing power virtualization

通过虚拟化技术，将不同类型的物理计算硬件抽象为统一的逻辑资源，向上层应用屏蔽底层硬件的差异性，实现资源的灵活分配和共享。

3.3

异构算力池化 heterogeneous computing power pooling

将经过虚拟化的异构算力资源整合到一个统一的资源池中进行集中管理，根据任务需求动态分配资源。

3.4

跨平台模型部署 cross-platform model deployment

将已在某个硬件平台训练好的模型部署到其他不同的硬件平台上运行，以实现模型在不同环境下的应用。

3.5

资源调度 resource scheduling

根据任务的优先级、资源需求和硬件资源状态，将计算任务分配到合适的计算资源上执行的过程。

3.6

资源利用率 resource utilization

实际使用的计算资源量与可用计算资源量的比率。

3.7

负载均衡 load balancing

将计算任务均匀分配到多个计算资源上，以避免部分资源过载而其他资源闲置。

3.8

任务 task

在云计算环境中由用户提交的工作单元，包含特定的计算需求、输入数据和处理目标。

注：任务可以是计算任务、数据处理任务、机器学习训练任务等。

3.9

多租户 multi-tenancy

在同一套硬件和软件基础设施上，为多个不同的用户或组织提供服务。

注：每个用户或组织被称为一个租户。通过资源隔离和访问控制等机制，保障各租户之间的数据和资源安全。

3.10

云原生 cloud native

基于云计算架构设计和构建应用程序的技术集合和方法

注：利用云原生构建的应用具备弹性、敏捷、松耦合、易交付、易观测等特征。

[来源：GB/T 44158-2024,术语和定义 3.2]

4 缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

API：应用程序编程接口(Application Programming Interface)

ASIC：专用集成电路 (Application-Specific Integrated Circuit)

CPU：中央处理器(Central Processing Unit)

CUDA：统一计算设备架构(Compute Unified Device Architecture)

FPGA：现场可编程门阵列(Field-Programmable Gate Array)

GPU：图形处理单元(Graphics Processing Unit)

K8S：一种开源的容器编排引擎（Kubernetes）

MiB：兆字节(Mebibyte)

ML：机器学习(Machine Learning)

NPU：神经网络处理单元(Neural Processing Unit)

ONNX：开放神经网络交换(Open Neural Network Exchange)

RBAC：基于角色的访问控制(Role-Based Access Control)
SoC：片上系统(System on Chip)

5 异构算力虚拟化及池化系统架构

异构算力虚拟化及池化系统指通过接入组件整合硬件算力、云算力、边缘算力等异构算力资源，经池化技术汇聚形成统一资源池，利用虚拟化技术分割为灵活分配的虚拟算力单元，结合异构算力调度机制与标准化接口、运维监控体系，为上层模型层和应用层提供高效算力支持，并通过算力运营，实现对异构算力资源的全流程管理与优化，提升整体算力资源利用率。

异构算力虚拟化及池化系统架构见图1。包括异构算力接入、异构算力虚拟化、异构算力池化、异构算力调度、异构算力接口、运维运营等核心技术能力，为GB/T 45288.1-2025中规定的资源池建设、工具链整合、数据资源管理及行业应用落地提供底层技术支撑。

本文件不涉及图中虚线框中模型层和应用层的内容。

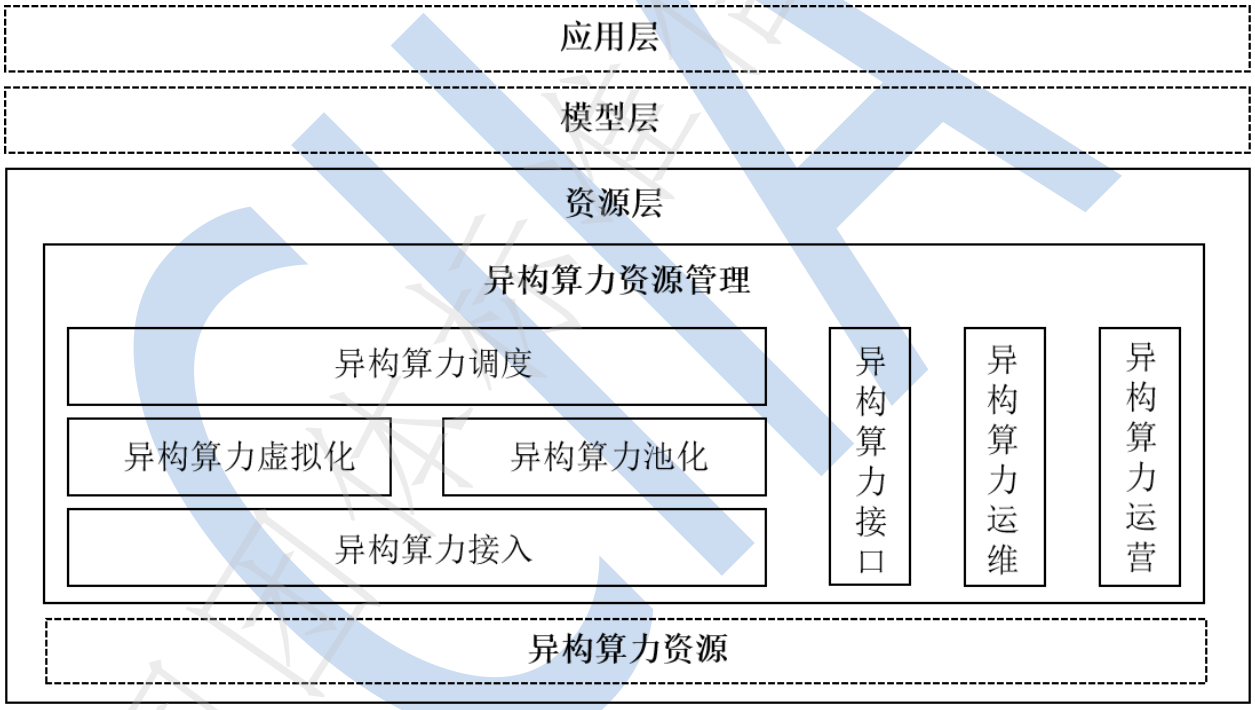

图 1 异构算力虚拟化及池化系统架构

6 异构算力接入要求

异构算力接入要求包括：

a) 应支持算力接入组件管理能力；
b) 应支持多架构硬件算力与云化算力接入，具备广泛的处理器架构兼容性，包括 x86 架构、ARM 架构、RISC-V 架构等，同时应支持对主流云服务商算力资源的统一接入，构建混合异构算力池；
c) 应支持多类型处理器接入，包括 CPU、GPU、NPU、FPGA、ASIC 等多种类型处理器；
d) 应支持多形态算力接入，包括硬件算力、云算力、边缘算力等多形态接入方式；

e） 宜支持多样化通信协议接入，如 PCIe（用于服务器内部高速设备互联）、InfiniBand（适用于数据中心内高性能计算集群互联，提供低延迟、高带宽连接）、Ethernet（以太网，常见通用网络连接协议）等，以满足不同算力设备的通信需求；
f） 宜具备算力设备自动发现与识别能力，可快速准确识别新接入的算力设备，自动获取设备基础信息，如型号、规格、性能参数等，并在系统中完成注册与初始化配置；
g） 宜支持算力设备的热插拔接入与移除，在设备接入或移除过程中，系统可自动进行资源状态更新与配置调整，不影响其他已接入算力设备的正常运行；
h） 宜具备对算力接入链路的监测与故障诊断能力，实时监控链路带宽、延迟、丢包率等指标，当链路出现异常时，能快速定位故障点并提供故障预警与修复建议；
i） 宜支持对多架构、多类型、多形态的异构资源进行标准化识别，建立统一资源标签体系，如算力类型、算力架构、性能指标等，形成全局资源视图。

7 异构算力池化要求

异构算力池化要求包括：
a） 应支持多架构算力池化，包括 x86 架构、ARM 架构、RISC-V 架构等的算力池化；
b） 应支持多类型处理器算力池化，包括 CPU、GPU、FPGA、ASIC 等多种类型处理器的算力池化；
c） 应支持多形态算力池化，包括硬件算力、云算力、边缘算力等多形态的池化；
d） 应支持算力池化资源分配能力，动态且精准地将池化后的算力分配至不同应用或用户，保障各类任务的高效执行；
e） 应具备算力池化资源隔离与安全保障机制，确保不同用户在使用池化算力时，资源相互隔离，数据安全不被泄露，防止恶意抢占或干扰行为；
f） 宜支持算力资源的快速检索与匹配，当有新的算力需求时，能在短时间内从算力池中筛选出符合要求的资源；
g） 宜增加算力池数据交换性能要求，如单节点：≥400Gbps 吞吐，端到端延迟：≤1.5us；
h） 可增加多种算力精度（FP32/FP16/BF16/INT8）的动态匹配兼容性要求。

8 异构算力虚拟化要求

异构算力虚拟化要求包括：
a） 应支持多架构算力池化，包括 x86 架构、ARM 架构、RISC-V 架构等多种类型处理器的算力虚拟化；
b） 应支持多类型处理器算力池化，包括 CPU、GPU、FPGA、ASIC 等多种类型处理器的算力虚拟化；
c） 应支持对国内和国外主流显卡进行细粒度切分；
d） 应支持多形态算力池化，包括硬件算力、云算力、边缘算力等多形态的算力虚拟化；
e） 应具备高效的资源抽象与封装能力，将不同架构、类型和形态的底层算力资源抽象为统一的虚拟资源对象，屏蔽硬件差异，为上层应用提供简洁、标准的资源访问接口；
f） 应支持细粒度的资源分配，如显存、算力、vGPU 数等，能够根据应用的具体需求，精确地划分和分配虚拟算力资源，提高资源利用率，避免资源浪费；
g） 应具备算力资源动态调整机制，可根据应用负载的实时变化，灵活调整虚拟算力资源的配置，如增加或减少虚拟核心数量、调整内存分配等，确保应用始终获得最优的算力支持；
h） 应提供完善的资源监控与管理功能，实时监测虚拟算力资源的使用情况，包括利用率、性能指标、运行状态等，并提供可视化的管理界面，方便管理员进行资源管理和故障排查；

i） 应具备资源隔离与安全保障能力，通过硬件虚拟化技术或软件隔离机制，确保不同虚拟算力资源之间的独立性和安全性，防止资源相互干扰和数据泄露；
j） 宜支持虚拟算力资源的迁移，在不中断应用运行的情况下，实现虚拟算力资源在不同物理节点之间的迁移，保证应用的连续性；
k） 宜增加迁移过程中的数据一致性保障机制要求，确保资源迁移时应用状态无损，避免中断或数据丢失。

9 异构算力调度要求

9.1 基础调度能力要求

基础调度能力要求包含：
a） 应支持优先级驱动调度策略，提供自定义任务优先级规则（如业务等级、 deadlines 约束），确保高优先级任务优先获取算力资源，低优先级任务错峰执行；
b） 应支持多租户公平分配机制，基于 RBAC 的多用户资源配额管理，按用户设置资源使用上限，如 GPU 显存、CPU 核心数配额等，保障资源分配公平性，避免单租户抢占导致系统过载；
c） 宜提供可视化调度策略配置界面，支持自定义调度规则，如“优先使用边缘算力处理低时延任务”“优先使用低成本云算力处理非紧急任务”；
d） 宜支持按时间片为轻量级任务动态分配资源，提升算力共享资源的时间利用率；
e） 宜增加算力资源故障隔离机制，避免为任务分配故障节点。

9.2 动态调度能力要求

动态调度能力要求包括；
a） 应支持网络拓扑感知调度策略，通过统一的网络拓扑 API 和智能调度策略，解决大规模数据中心 AI 训练任务的网络通信性能问题；
b） 应支持基于实时负载数据，动态调整调度策略，提供任务跨节点、跨域迁移；
c） 应支持实时感知任务运行时资源需求变化，提供资源动态扩缩容，如 GPU 显存分配量实时调整；
d） 应支持定义紧急任务触发规则，提供中断低优先级任务并优先调度紧急任务；
e） 宜支持跨域协同调度策略，进行云边端协同调度，根据任务类型、网络时延动态等选择执行节点；
f） 可支持基于算法的调度策略自优化，通过历史调度数据，如任务完成率、资源利用率等，持续迭代算法参数；
g） 可支持分布式环境协同调度，万级节点规模的分布式调度架构，通过分层调度实现跨数据中心、跨云厂商的资源协同。

10 异构算力接口要求

10.1 南向接口要求

用于与底层硬件资源的交互，实现对硬件设备的管理和控制，接口具备良好的兼容性和扩展性，具体要求应包括：
a） 支持硬件设备接入接口；
b） 支持资源状态接入接口；

c） 支持算力资源分配接口；
d） 支持边缘算力接入接口；
e） 支持云算力协同接口；
f） 支持超分资源分配接口；
g） 支持超分算法调用接口；
h） 支持资源回收接口；
i） 支持数据预处理与后处理接口；
j） 支持监控数据上报接口；
k） 支持设备控制接口；
l） 支持硬件健康状态监测接口。

10.2 北向接口要求

用于与上层应用和管理工具的交互，实现对算力资源的管理和控制，接口具备良好的兼容性和扩展性，具体要求应包括：

a） 支持跨池协同调度接口；
b） 支持推理服务调用接口；
c） 支持资源查询接口；
d） 支持弹性伸缩控制接口；
e） 支持运维管理接口；
f） 支持国产化适配接口；
g） 支持跨平台通信接口；
h） 支持服务发现与注册接口；
i） 支持版本管理与兼容性接口；
j） 支持任务调度与管理接口；
k） 支持数据交互接口；
l） 支持模型交互接口；
m） 支持用户与权限管理接口；
n） 支持算力资源计费接口。

11 异构算力运维要求

11.1 监控要求

异构算力监控要求包括：

a） 应支持对异构算力资源的算力数据的采集，如处理器算力、资源负载等；
b） 应支持异构算力的实时监控能力，如关键指标、数据采集频率，确保及时反映资源状态变化；
c） 应支持异构算力指标的统一标准化处理，通过适配器将不同异构资源的私有指标转换为统一格式，实现跨架构数据兼容；
d） 应提供历史数据存储与查询能力，如系统运行日志、硬件系统日志、任务运行日志等；
e） 应支持自定义告警规则，紧急告警多通道通知；
f） 宜支持可视化展示界面，展示如资源热力图、任务看板、趋势分析等；
g） 宜支持算力任务执行过程的全面监控；
h） 宜支持算力调度的监控，如调度策略的执行情况、任务队列长度、等待时间等数据；

i） 可支持异构算力资源的多维度性能指标监控，覆盖硬件层、虚拟层及链路层；
j） 可支持对人工智能应用的监控；
k） 可支持算力网络链路监控，展示网络状态、网络类型等信息。

11.2 安全合规要求

安全合规要求包括：

a） 应支持在异构算力虚拟化和池化过程中，用户数据的安全策略，包括数据的加密存储、传输安全、访问控制等；
b） 应提供全链路安全审计日志，记录用户对算力资源的访问行为、数据流向等；
c） 可支持数据分类分级保护，根据用户数据敏感等级（如公开数据、商业机密和个人隐私等），自动匹配加密强度与访问控制策略；
d） 应支持 GPU 资源按多租户模式进行隔离，支持独占模式与共享模式，通过硬件级隔离机制防止侧信道攻击；
e） 可支持合规性审计。定期扫描算力环境的安全配置，生成合规性报告并对接监管平台。

11.3 故障处理与容灾

故障处理与容灾要求包括：

a） 应具备故障检测能力，如硬件、软件、网络等；
b） 应具备数据的定期备份和快速恢复策略，确保在发生数据丢失或损坏时能够迅速恢复；
c） 应具备容错机制，当出现硬件故障或系统异常时，能够自动进行资源恢复和重新分配；
d） 应具备高可用性架构，确保关键组件在发生故障时能够快速切换，避免单点故障；
e） 可支持跨域容灾切换，如主数据中心与边缘节点间的任务无缝迁移。

11.4 拓展性

拓展性要求应包括：

a） 具备良好的扩展性，能够灵活应对异构计算设备的增加、业务规模的扩大以及技术的更新迭代；
b） 支持快速接入新型异构算力设备及算力资源，只需在算力资源中部署相应的接入组件，即实现资源的自动识别与资源整合；
c） 采用插件化架构，遵循统一的接口规范，允许通过添加或更新插件来扩展功能；
d） 支持接入自动化运维工具，实现计算资源的自动监控、扩展和管理；
e） 支持自定义调度算法的扩展，允许用户根据业务需求开发并集成新的调度策略；
f） 预留未来技术适配接口，支持新型算力形态的接入框架，通过抽象层设计屏蔽底层技术差异，最小化技术迭代对上层应用的影响。

11.5 隔离性

隔离性要求应包括：

a） 提供严格的资源隔离机制，确保不同用户之间的计算资源、内存资源、网络资源等不会相互干扰；
b） 支持性能隔离，确保不同用户或者任务在高负载下仍能保持稳定的性能，不受其他用户或者任务的影响；
c） 支持安全隔离，通过访问控制和数据保护机制，确保不同任务或用户之间的数据安全。

12 异构算力运营要求

异构算力运营要求包括：

a) 应支持异构算力资源目录管理能力，明确各类算力的技术参数、算力形态及服务属性，形成标准化资源描述体系；

b) 应支持多租户资源配额管理，基于 RBAC（基于角色的访问控制）为不同使用者分配独立的算力使用限额；

c) 宜支持多维度计量计费标准，包括按资源使用量、时长、性能等级及任务类型等进行计量计费；

d) 宜支持清晰的计费账单与费用明细查询能力，包括按租户、时间周期、资源类型等生成账单报告；

e) 宜支持对接第三方支付平台，实现算力服务费用的自动结算与扣款；

f) 宜支持阶梯式折扣、预付费优惠等商业化运营策略；

g) 可支持用户分级管理，根据用户规模、付费能力及业务需求划分普通用户、企业用户、战略合作伙伴等层级，提供差异化服务与技术支持；

h) 可支持运营数据统计分析，采集资源利用率、任务调度成功率、用户满意度等关键指标，通过数据驱动持续优化资源分配策略、调度算法及服务流程；

i) 可支持引入 AI 运营辅助工具，利用机器学习算法对历史运营数据进行分析，预测资源使用峰值、优化调度策略。

参 考 文 献

[1] GB/T 35293-2017 信息技术 云计算 虚拟机管理通用要求
[2] GB/T 42018-2022 信息技术 人工智能 平台计算资源规范
[3] GB/T 43782-2024 信息技术 人工智能 机器学习系统技术要求
[4] GB/T 44158-2024 信息技术 云计算 面向云原生的应用支撑平台功能要求
[5] GB/T 45288.1-2025 人工智能 大模型 第1部分：通用要求
[6] GB/T 45401.1-2025 人工智能 计算设备调度与协同 第1部分：虚拟化与调度