ICS 35. xxx
CCS Lxx

团　　体　　标　　准

T/ISC XXX—XXXX

智能算力服务等级协议

Intelligent Computility Service Level Agreement

在提交反馈意见时，请将您知道的相关专利与支持性文件一并附上。

（征求意见稿）

2025-12-19

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中 国 互 联 网 协 会　　发布

目　　次

前　　言 III
引　　言 V
智能算力服务等级协议 1
1　范围 1
2　规范性引用文件 1
3　术语和定义 1
3.1　算力　computility 1
3.2　算力服务　computility service 1
3.3　算力资源服务　computility resource service 1
3.4　算力资源调度　computility resource scheduling 2
3.5　服务等级指标　service level metric 2
4　符号和缩略语 2
5　智能算力服务等级协议概述 2
5.1　CSLA 的内容和使用原则 2
5.2　CSLA 管理 2
5.3　CSLA 编制指引 2
6　智能算力服务等级协议要素 2
6.1　必备要素 2
6.2　可选要素 5
7　智能算力服务等级描述 6
8　智能算力服务等级协议管理流程 7
8.1　流程概述 7
8.2　CSLA 的设计 7
8.3　CSLA 的签署 8
8.4　CSLA 的执行 8
8.5　CSLA 的变更 8
附　录　A　（资料性）　算力服务等级 9
A.1 服务等级指标 9
A.2 服务等级指标项权重 14
附　录　B　（资料性）　监测指标 16
B.1 动态监测指标 16
B.2 静态统计指标 17

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。

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。

本文件由中国互联网协会提出并归口。

本文件起草单位：

本文件主要起草人：

本文件及其所代替文件的历次版本发布情况为：

——

引　言

随着数字经济的深度发展，算力已成为支撑人工智能、大数据、云计算等新一代信息技术创新应用的核心基础设施，是推动产业数字化转型、保障关键领域运行的战略资源。然而，当前智能算力服务市场存在服务标准不统一、质量评估维度缺失、供需双方权责界定模糊等问题，导致算力资源配置效率偏低、服务纠纷频发，难以满足大模型研发、科研创新、企业数字化升级等场景对高质量算力服务的需求。

本标准立足智能算力产业发展特点，以“提升服务质量、优化资源配置、保障双方权益”为目标，系统界定了算力、算力服务、算力资源调度等核心术语，明确了算力服务等级协议的构成要素（含必备要素与可选要素）及全生命周期管理流程，并通过附录形式提供了可操作的服务等级指标、权重划分、等级描述及监测指标，为算力服务供需双方签订协议、评估服务质量提供统一参考。

智能算力服务等级协议

1 范围

本文件规定了智能算力服务等级协议的构成要素，包括算力服务等级协议的概述、要素和管理流程。

适用于：

a）为算力服务提供商和用户之间制定服务等级协议提供参考依据；

b）为客户对算力服务提供商服务质量进行考评提供参考依据；

c）为算力算法服务平台对算力服务提供商等级协议监测提供参考依据。

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 37092 信息安全技术密码模块安全要求

GB/T 36325 信息技术云计算云服务等级协议基本要求

YDB 144 云计算服务协议参考框架

3 术语和定义

3.1

算力 computility

计算机设备或计算/数据中心处理信息的能力，是计算机硬件和软件配合共同执行某种计算需求的能力。

3.2

算力服务 computility service

计算机设备或计算/数据中心处理信息的能力，是计算机硬件和软件配合共同执行某种计算需求的能力。以多样性算力为基础，以算力互联网为连接，以供给有效算力为目标的算力产业新领域，通过全新技术实现异构算力统一输出，并与云、大数据、人工智能等技术交叉融合，最终将算力、存储、网络等资源统一封装，以服务形式完成交付。

3.3

算力资源服务 computility resource service

包括通用计算能力、超级计算能力和智能计算能力服务，以及数据存储能力。

3.4

算力资源调度 computility resource scheduling

主要实现算力供给方和需求方间算力供需匹配，包括算力度量、算力编排、算力标识转换和算力资源选择等算力服务，但不包括弹性网络调度。

3.5

服务等级指标 service level metric

用于评估和度量算力服务提供商服务能力的指标。

4 符号和缩略语

下列符号和缩略语适用于本文件。

CSLA 算力服务等级协议（computility service level agreement）

5 智能算力服务等级协议概述

5.1 CSLA 的内容和使用原则

CSLA明确了算力服务的范围、内容、服务等级等内容，是双方（是指算力服务提供商和用户）对服务相关约定的一致理解和认可，也是进行服务考核、改善服务质量的有效依据。

明确定义的服务内容和服务等级指标是CSLA的重要组成部分。

CSLA应具备的服务等级指标可参考附录A；监测指标可参考附录B。

当在使用CSLA时，遵循以下原则：

a) 双方应就服务等级协议达成共识；

b) 算力服务等级指标应是可获取、可理解。

5.2 CSLA 管理

双方应以一种受控的方式来管理与CSLA有关的活动，包括CSLA的设计、CSLA的签署、CSLA的执行和CSLA的变更。

5.3 CSLA 编制指引

拟定服务等级协议时，双方应按第6章内容进行选择、裁剪和调整，形成双方共同认可的服务等级协议。

6 智能算力服务等级协议要素

6.1 必备要素

6.1.1 算力服务客户

算力服务客户是在服务等级协议中，有算力服务需求的组织机构或个人。内容满足以下要求：

a) 机构客户的主要内容宜包括组织机构名称、相关部门、联络人等，并明确接口人的岗位、职责、姓名、工作地点、联系方式等；

b) 个人客户的主要内容宜包括姓名、联系方式、通讯地址、有效身份证明等。

6.1.2 算力服务提供商

算力服务提供商是在服务等级协议中，提供算力服务的机构。提供商的主要内容宜包括组织机构名称、相关部门、接口人等，并明确参与服务人员的岗位、职责、姓名和联系方式等。

6.1.3 算力算法服务平台

算力算法服务平台兼具监督管理能力与产业服务能力。面向监管端实现算力标识检索、算力运行监测等核心服务能力，面向市场端实现算网参数调度、身份认证管理、算力资源服务和算力交易结算等基础服务能力，统筹汇聚通算、智算、超算、云计算等公共算力资源，以市场化手段形成优质高效、国内领先的智能算力运行服务和资源配置机制，为大模型企业、科研机构提供普惠智算资源。

6.1.4 服务内容

服务内容是在服务等级协议中，双方达成共识的服务内容。针对达成共识的算力服务类别，宜说明具体服务的服务模式、服务功能和服务方式，以及算力服务提供商在服务部署、运行和退出阶段所提供的其他服务，如培训、业务数据的迁入迁出等。

6.1.5 服务期限

服务期限是在服务等级协议中，双方达成共识的服务起止时间。服务期限应包括服务等级协议生效的起始时间点和服务等级协议终止的截止时间点。

6.1.6 服务时间

服务时间是算力服务提供商为客户提供算力服务的时间窗口，宜依据业务需求确定，如可从5×8h、7×12h、7×24h等不同选项中做出选择。

6.1.7 服务计量标准度

服务计量标准度是算力服务提供商应承诺用户将按用户实际的购买量或者使用量计费，并基于实际运行性能和使用情况为客户说明计费规则。

算力服务提供商应根据用户的特定需求和不同的使用场景，如训练场景或推理场景等，设定明确的计费规则。计费方式宜按照核时或卡时进行；对于需要独占队列的服务，宜考虑采取独占节点的计费模式，比如按节点设定包月或包年的费用；对于存储资源产品，宜实行按需计费模式，如元/TB*月或年；网络资源服务产品宜采用按需计费模式，如包月或包年；如果业务倾向智算/大模型的话，增加MaaS服务产品，宜考虑按需计费模式，如按次、按功能模块、按核时等方式。

算力算法服务平台基于以下规则形成算力资费标准指导结算价：平台基于同类型销售产品报价计算平均价(平均价：平台采取抽样或询价等方式获取一定数量的算力服务商报价计算平均值），低于均价的50%或高于均价的150%的价格需和平台协商。

6.1.8 服务等级指标

在服务等级协议中，双方商定的衡量算力服务提供商服务能力的参数，用于衡量算力服务提供商服务质量。

服务等级指标宜包含指标名称、指标描述等内容，用于评估算力服务提供商在可用性、可靠性、性能、绿色低碳等方面的服务质量，适用于各类型算力服务。算力服务等级指标见附录A。

6.1.9 服务监测指标

在服务等级协议中，算力算法服务平台宜对算力服务提供商服务质量进行监测。

服务监测指标宜包含指标名称、指标描述等内容，可用于监测算力服务提供商性能。适用算力算法服务平台监测算力服务提供商指标见附录B。

6.1.10 软件支持的丰富度

为算力服务客户提供硬件设备加速设计的软件支持，包括多样的开发框架、高效的加速库、实用的开发工具及丰富的预训练模型、推理服务等。

a) 开发框架支持的多样化：平台应支持主流AI开发框架，并对框架进行优化，支持灵活的模型设计和自动化计算图优化。支持简化开发者的工作流程，提高模型开发效率和性能；
b) 加速库集成的高效性：平台应集成优化加速库，显著提升深度学习推理和训练的效率。加速库为多种AI应用（包括计算机视觉、自然语言处理、语音识别等），使AI模型能够在更短的时间内实现高性能推理；
c) 开发工具支持的全面性：平台应包含开发工具，提供全面的调试和优化功能，帮助开发者高效开发和部署AI应用。工具的易用性和功能全面性确保开发团队能够随时对模型进行深入调试和优化；
d) 预训练模型库丰富性：平台宜融合预训练模型库，涵盖从图像识别到自然语言处理的各种AI应用。预训练模型可以帮助企业快速启动项目，缩短开发周期和产品上市时间，减少从零开始训练模型的复杂性和成本；
e) 自动化部署的高效性：平台宜提供自动化和智能化管理AI推理工作负载的能力。快速将AI模型部署到生产环境中，在运行过程中自动监控和优化模型性能。能够减少人工干预，通过智能调度和资源分配确保AI应用以最佳状态运行。简化部署流程、实时监控系统状态、自动调整计算资源，自动化部署让企业能够专注于业务目标，而无需担心技术细节，显著提升了AI应用的整体效率和可靠性。

6.1.11 服务交付物

服务交付物是在交付算力服务的过程中，双方约定提供的有形或无形的成果。

应明确服务过程中需提交的各类交付成果，如：

a) 软件、平台、硬件、数据等有形成果；
b) 状态恢复、性能提升、业务优化、知识资产等无形成果；
c) 由第三方出具的适配验证报告等测试报告类成果。

6.1.12 责任和义务

双方应履行的责任与义务，应满足以下要求：

a) 算力服务提供商的主要责任和义务包括：按照双方约定的交付内容，符合国家法律法规、相关标准等方面的要求；
b) 算力服务客户的主要责任和义务包括：应遵守相关法律法规、规章制度，在双方协议约定条件下使用相应服务并及时支付相关费用。

6.1.13 违约责任

双方发生违约行为时应承担的责任，以及免除责任的相关情况说明。免责条款应包括计划停机、自然灾害和算力服务提供商不可抵抗的其他因素等情形。

6.1.14 保密要求

双方对于算力服务过程中所产生数据 / 信息、个人隐私及商业秘密的保密要求。保密要求应包含保密信息的范围、双方的保密责任、保密期限等。

6.1.15 投诉渠道

算力服务提供商提供的有效投诉渠道。投诉渠道应包括电话、来访、信件及其他。

6.1.16 补偿

双方共识的服务内容未达到时，算力服务提供商向客户提供的补偿。双方宜根据责任协商补偿的具体内容，如现金、免费延长服务时间等形式。

6.2 可选要素

6.2.1 第三方

与服务相关的第三方组织机构或个人，宜包括联系人名称或姓名，联系方式,与算力服务提供商和客户的关系、责任、权利与义务等。

6.2.2 服务优先级

双方达成共识的服务优先级定义，服务优先级应由服务影响度及紧急度来决定：

a) 影响度：根据故障、问题、变更等发生后的影响范围进行确定；

b) 紧急度：根据服务是否涉及关键业务应用来进行确定。

6.2.3 变更管理流程

当服务发生变化或因某种原因导致服务等级协议需要发起变更时，双方应遵循的变更管理流程，包括变更类型的划分、变更的申请、变更的响应、变更的影响评估、变更的批准、变更的执行、变更的确认和变更的总结等。

6.2.4 服务交付相关流程

支撑服务交付的管理流程，宜包括服务交付发起流程、故障管理流程、服务请求流程和服务报告流程等。

6.2.5 资源条件

保证服务交付所需的支持性资源，宜包括场地条件、通信条件、工具、硬件设施和软件平台等。在CSLA中宜明确资源条件的提供商。

6.2.6 争议解决机制

当服务等级协议执行发生争议时，双方商定的争议解决机制，宜包括协商、仲裁、起诉等。

6.2.7 服务考核要求

双方商定的服务考核要求，宜包括考核内容、考核评价标准、考核方式、考核周期、考核报告、考核结果等。

6.2.8 服务费用和支付

服务费用和支付宜包括服务计费方式、计费标准和支付方式等。

6.2.9 知识产权

在服务过程中，宜针对双方所提供的任何信息、数据和文件做出知识产权说明。

6.2.10 通知和送达

在服务过程中，宜由一方发出书面形式的通知，以及通知送达时间和方式。

7 智能算力服务等级描述

智能算力的服务等级被分为一到五级。

7.1.1 五级智能算力服务描述

五级智能算力服务应满足以下指标：

a) 支持3000亿规模以下和3000亿以上的参数模型微调，在调试阶段，可用性和可靠性远超行业平均水平；
b) 具备行业领先的软硬件协同加速效果，任务运行过程中硬件资源利用充分、专用加速指令调用高效，连续长时间运行无故障，且故障恢复速度极快；
c) 符合国家最严格的安全性和合规性要求，持续的漏洞管理和定制的安全解决方案；
d) 绿色低碳指标远超国内平均水平；
e) 内置大模型从技术分类层面至少覆盖视觉、语音、自然语言、时间序列、多模态、大模型、文生图和遥感等8大类；线上开发环境至少提供3种模式；提供算力资源、模型算法以及平台全面多维度监测和分析能力。
f) 具有行业领先的性能指标，并配备完善的软件服务支持。

7.1.2 四级智能算力服务描述

四级智能算力服务应满足以下指标：

a) 支持1000亿规模以下和1000~3000亿规模的参数模型微调，在调试阶段，可用性和可靠性超过行业平均水平；
b) 软硬件协同加速效果显著优于行业平均，硬件核心利用率较高，专用加速指令调用占比突出，连续长时间运行故障极少，故障恢复速度快；
c) 采用最先进的安全技术，确保数据的完整性和隐私；
d) 绿色绿色低碳指标显著优于国内平均水平；
e) 内置大模型从技术分类层面至少覆盖视觉、语音、自然语言、时间序列、多模态、大模型等6大类；线上开发环境至少提供2种模式；提供算力资源、模型算法以及平台多指标综合监测和分析能力。
f) 具有显著优于行业的性能指标，并配备一定的软件服务支持。

7.1.3 三级智能算力服务描述

三级智能算力服务应满足以下指标：

a) 支持500亿规模以下和500~1000亿的参数模型微调，在调试阶段，可用性和可靠性达到行业平均水平；

b) 软硬件协同加速效果达到行业平均水平，硬件资源利用合理，专用加速指令可有效调用，连续长时间运行故障较少，故障恢复及时；
c) 有一定的高级的安全防护，确保数据的完整性和隐私；
d) 绿色低碳指标优于国内平均水平；
e) 内置大模型从技术分类层面至少覆盖视觉、语音、时间序列、多模态等4大类；线上开发环境至少提供1种模式；提供算力资源、模型算法以及平台指标基础监测能力。
f) 优于行业平均的性能指标，配备一些软件服务支持。

7.1.4 二级智能算力服务描述

二级智能算力服务应满足以下指标：
a) 支持100亿参数以下和100~500亿的参数模型微调，在调试阶段，可保障服务可用性和可靠性，故障较少产生；
b) 具备基础的软硬件协同加速效果，硬件资源能满足基本使用需求，专用加速指令可部分调用，连续长时间运行偶有故障，故障可在合理时间内恢复；
c) 具备增强的安全措施，确保数据的完整性和隐私性；
d) 绿色低碳指标达到国内平均水平；
e) 达到行业平均的性能指标，配备至少1项软件服务支持。

7.1.5 一级智能算力服务描述

一级智能算力服务应满足以下指标：
a) 支持一些模型的微调，可提供基本的稳定运行时间，可能偶尔出现停机和小幅度的性能波动；
b) 软硬件协同加速效果处于基础水平，硬件资源利用效率一般，专用加速指令调用有限，连续长时间运行可能出现故障，故障恢复时间较长；
c) 包括基础的数据安全措施，如定期备份和基本的用户认证；
d) 绿色低碳指数低于国内平均水平；
e) 低于业平均的性能指标，没有软件服务支持。

8 智能算力服务等级协议管理流程

8.1 流程概述

双方应遵循下述流程对CSLA 进行管理，以确保按一种受控的方式管理与 CSLA 有关的活动。

8.2 CSLA 的设计

CSLA 的设计是指算力服务提供商根据客户的要求，结合智能算力服务等级协议要求，设计出CSLA文本。在设计 CSLA 的过程中，算力服务提供商应：
a) 识别客户对算力服务的需求；
b) 设定合理的服务级别指标，服务级别指标应反映客户对服务内容和服务级别的要求，并得到服务提供商组织内部的一致理解；
c) 参考本标准第 6 章中的CSLA要素，编制出具体的CSLA文本。

8.3 CSLA 的签署

CSLA的签署是指算力服务提供商与客户签署约定服务级别的协议文档，应满足以下要求：

a) 算力服务提供商与客户针对已有CSLA进行协商或确认；

b) 算力服务提供商与客户签署双方确认后CSLA文档，CSLA 的协议文档以纸质或电子形式提供。

8.4 CSLA 的执行

CSLA的执行是指算力服务提供商按CSLA中的要求向客户交付算力服务，应满足以下要求：

a) 算力服务提供商做好相应的人员和资源的准备工作；

b) 算力服务提供商按照CSLA中规定的服务内容和服务级别为客户提供算力服务；

c) 必要时算力算法服务平台对算力服务提供商的算力运行质量及安全进行监测，以供双方参考。

8.5 CSLA 的变更

CSLA的变更是指双方对CSLA的变更进行管理和控制，应满足以下要求：

a) 双方协商CSLA变更流程，并建立变更管理机制；

b) 算力服务提供商根据CSLA变更方案进行实施，并对变更结果进行跟踪；

c) 在完成CSLA变更后，算力服务提供商及时通知客户，并由客户对变更结果进行确认

附 录 A
（资料性）
算力服务等级

在智能算力服务等级协议中，双方可参考表A.1服务等级指标进行商定的算力服务提供商服务能力的参数，并参考表A.2服务等级指标项权重对算力提供商服务能力，分成五个等级，等级描述参考表A.3。

A.1 服务等级指标

服务等级指标如表A.1所示：

表 A.1　服务等级指标

指标的名称	描述
服务可用度	1）服务可用度指的是算力服务客户发起服务请求后服务可访问的时间占总服务时间的比例。
计算节点的可用性	1）计算节点在线运行的时间占总时间比例。在线时间率越高，表示计算节点的可用性越好。 2）计算节点发生故障后恢复正常运行所需的时间范围。较短的故障恢复时间能够提高计算节点的可用性。
存储服务的可用性	1）存储服务的读写速度。高速的读写速度有助于提高存储服务的可用性； 2）存储服务是否提供数据冗余备份功能，以防止数据丢失或损坏。冗余备份能够提高数据的可用性，即使某个存储节点发生故障，数据仍然可恢复； 3）明确存储服务在发生故障后恢复正常运行所需的时间范围。
网络服务可用性	1）网络服务发生故障的概率。较低的网络故障率意味着网络服务更加稳定可靠； 2）网络发生故障后恢复正常运行所需的时间。较短的故障恢复时间能够提高网络服务的可用性。
服务的可靠性	1）计算节点在一定时间内发生故障的次数。 2）当服务出现中断时，修复故障并恢复正常服务所需的平均时间长度； 3）算力服务连续出现故障之间的平均时间间隔。
数据存储的可靠性	1）数据的一致性：在多个用户或系统同时访问同一份数据时，确保数据在逻辑上保持正确和可靠的状态； 2）数据持久性：数据在存储后能够在一定条件下保持不丢失，即使在系统故障或灾难情况下也能够恢复； 3）数据可用性：数据能够在需要时被正常访问和使用，系统能够执行其功能并及时响应请求； 4）数据完整性：采取措施确保数据在存储、传输和处理过程中不受损坏或篡改，保持其原始状态； 5）错误检测与纠正：系统能够检测和纠正在存储和传输过程中可能发生的数据错误，保证数据的准确无误； 6）备份与恢复：定期对数据进行备份，并确保在数据丢失或损坏时能够快速恢复，减少业务中断的时间； 7）数据加密：对敏感数据进行加密存储，防止未经授权的访问和泄露，确保数据的机密性； 8）系统冗余：通过建立冗余系统，如使用多个数据中心或服务器，来提高数据的可靠性和可用性； 9）灾难恢复计划：制定和实施灾难恢复计划，以应对自然灾害、人为错误或其他突发事件，确保数据的持续可靠存储； 10）保证存储服务可有效地保存和保护数据，降低数据丢失或损坏的风险。
数据持久性	数据不丢失的概率。
数据私密性	不同用户的数据是否互不可见。

指标的名称	描述
容量与弹性	1）描述提供最大并发任务数； 2）描述最大支持用户数； 3）描述资源弹性扩展时间； 4）描述资源收缩时间。
浮点计算能力	浮点计算能力是指计算机在单位时间内能够执行的浮点运算次数，通常用每秒浮点运算次数（FLOPS）来衡量。对于CPU和GPU来说，其浮点计算能力主要由以下几个因素决定： 核心数量：CPU或GPU的核心数量越多，可以同时进行更多的浮点运算任务，从而提高整体的计算能力。 主频：核心的频率越高，单个核心每秒能执行的浮点运算次数就越多。 单周期浮点计算值：即每个时钟周期内一个核心能完成的浮点运算次数。 精度：每秒浮点运算次数（FLOPS）应标明计算精度，如FP64双精度浮点数，即 64 位浮点数；FP32单精度浮点数，即 32 位浮点数；FP16半精度浮点数，即 16 位浮点数；FP8八位浮点数。不同精度等级的浮点数在计算机系统中的选择取决于具体的应用需求。
网络运力	该指标定义是以数据通信网网络基础设施为基础，通过自动化、智能化网络技术和运营管理平台为支撑，实现数据在不同用户、算力设施间以及算力设施内高效流动的网络运载力。它是构建综合性算力服务的重要一环，要求如下： 1）描述网络支持的最大数据传输速率。较大的网络带宽可以支持更多的数据传输，提高网络运力； 2）确定吞吐量指网络在单位时间内能够传输的数据量。高吞吐量意味着网络能够处理更多的数据传输请求； 3）描述网络延迟的时间范围； 4）描述网络拥塞率指网络中出现拥塞的概率范围。较低的网络拥塞率有助于提高网络的运力； 5）描述网络是否能够有效地分担网络流量； 6）网络带宽可以保证数据传输的速度范围和稳定性情况； 7）算力服务资源的提供方的时延、抖动、丢包情况。

指标的名称	描述
服务响应性能	算力服务提供商应根据用户的业务需求明确告知性能响应性能情况，包括资源池利用率、响应时间等。 1）资源池利用率：描述不同时间和地点，能够被有效利用的算力资源池情况； 2）响应时间：算力服务对用户请求的响应时间。较短的响应时间可以提高用户体验，加快数据处理和计算任务的执行速度； 3）接入速度：分配用户算力资源到用户可正常适用算力资源的速度。快速、稳定的接入速度可以提高用户的工作效率和体验； 4）并发处理能力：算力服务的并发处理能力指其同时处理多个用户请求的能力。较强的并发处理能力可以保证多用户同时接入时系统仍能保持高效稳定； 5）网络通信稳定性：稳定的网络通信可以保证数据传输的可靠性和速度； 6）数据传输速率：数据传输速率指用户在接入算力服务时数据传输的速度。高速的数据传输速率有助于提高接入性能和工作效率。
服务资源弹性	该指标定义是根据需求动态分配和管理计算、存储和网络资源的能力，通过资源动态扩缩能力，应对中小企业等特定行业的服务需求，服务资源调配能力要求如下： 1）能够根据实际需求动态调整计算和存储资源，实现弹性伸缩； 2）通过负载均衡技术，将请求均匀地分配到不同的计算节点或存储节点上，避免某些节点负载过重而导致性能下降或服务中断； 3）利用容器化技术，将应用程序及其依赖项打包成独立的容器，实现快速部署和资源隔离，提高资源利用率和服务部署的灵活性； 4）通过自动化运维工具和平台，实现服务资源的自动化监控、调配和管理，减少人工干预，提高资源利用效率和服务稳定性。
服务安全	1）算力服务提供商应确保服务安全，定期检查所使用的各项软件及时安全更新等； 2） 访问控制、攻击防范、网络审计、安全检测等应符合GB/T 22239中第三级的要求； 3）提供网络访问控制使算力服务客户实现网络分段、网络隔离和网络过滤功能，服务客户不应访问未授权的VPC地址； 4）用户客户端到算力服务平台之间的远程数据传输应采取保护和隔离措施； 5）除了部署传统的基于特征库的防御手段外，算力服务平台环境应具备针对APT、零日漏洞利用、定向攻击等高级威胁检测能力，识别在环境中的渗透、扩散、数据窃取等行为； 6）算力服务平台具备基于大数据技术的威胁检测、判断和关联分析能力，能够从全攻击链整体对安全威胁的发生和发展进行识别、分析和评估； 7）应具备防火墙与安全监控系统联动及防火墙策略迁移的能力，便于算力服务客户及时阻断违规 行为，消除安全威胁； 8)应具备安全分析和可视化能力，应提供安全事件展示、攻击路径展示及多维分析展示（例如： IP 地址、邮件、文件、域名、受威胁资产等多个维度）。
电能利用效率	电能利用效率（PUE）：PUE计算公式为：PUE=Pt/PIT，其中Pt为数据中心全年总耗电量， 单位是KWh;PIT为数据中心的IT设备全年耗电量， 单位也是KWh。

指标的名称	描述
可再生能源利用率	可再生能源利用率(R) 可参照下列公式计算，相关能耗计算可参照《综合能耗计算通则》(GB/T2589)执行（电力折标系数按当量值计算，下同）：R=ER/[E总能耗-(E 总电耗－E 自发自用- E 绿电交易)×a] ER：计算年可再生能源利用量（不含电网既有可再生能源 量），单位：标准煤； E总能耗：数据中心项目年能源消耗总量，单位：标准煤； E总电耗：数据中心项目年用电量，单位：标准煤； E自发自用：数据中心项目自建分布式可再生能源设施年利用 量（仅限于自发自用），单位：标准煤； E绿电交易：数据中心建设运营单位通过参与绿色电力交易用 于该数据中心项目年使用量，单位：标准煤； a可再生能源占比：数据中心项目被评价年度接入电网既有可再生能源占比，单位：百分比（%）。
内置算法模型数量	内置算法模型数量是衡量智能算力服务平台在算法模型资源储备维度的核心量化指标，指算力服务提供商在其智能算力服务平台中，预先集成并可供用户直接调用、二次开发或适配训练的算法模型总个数
内置算法模型类型覆盖度	内置算法模型类型覆盖度是评估智能算力服务平台算法模型资源多样性与场景适配能力的核心指标，指平台内预先集成的算法模型在技术领域、行业场景两大维度的类型覆盖范围占比，反映平台对不同技术需求（如视觉、语音处理）与行业应用（如农业、安防）的支撑能力
算法模型服务监测能力	算法模型服务监测能力是评估智能算力服务平台对内置算法模型全生命周期运行状态与服务质量的动态管控能力的核心指标，指平台通过技术手段实时采集、分析、预警模型服务关键数据（含调用频次、响应时延、错误率、资源占用、Token 消耗等），并生成可视化监测报告、支持异常追溯与性能优化的能力，反映平台对模型服务稳定性、可靠性及资源利用效率的管控水平
线上开发环境多样性	线上开发环境多样性是评估智能算力服务平台对不同技术栈、开发习惯及任务场景适配能力的核心指标，指平台为用户提供的线上开发环境在部署模式（如容器化、云原生）、工具链支持（如 IDE 类型、调试工具）、框架兼容性（如主流 AI 开发框架适配）及协作模式（单人开发、团队协同）等维度的差异化覆盖能力，反映平台满足用户多样化开发需求、降低开发门槛的水平
模型全生命周期开发配套能力	模型全生命周期开发配套能力是评估智能算力服务平台支撑算法模型从设计、训练、调试、部署到运维全流程开发效率与稳定性的核心指标，指平台为用户提供的覆盖模型开发各阶段的工具链、资源支持及流程管理能力，涵盖数据预处理工具、训练参数调优组件、模型版本管理系统、自动化部署模块及运行监控面板等配套服务，反映平台降低模型开发门槛、保障开发全流程顺畅性的水平

指标的名称	描述
软硬件协同加速能力	协同加速比：衡量硬件+定制化软件组合相对通用软硬件方案的性能提升效果，计算公式为：协同加速比=（通用方案完成大模型训练 / 推理时长）÷（协同方案完成同一任务时长），不同服务等级对应明确加速比要求。
硬件资源适配效率	评估软件对硬件核心特性的利用程度，通过硬件核心利用率和专用指令调用占比量化，核心利用率指AI芯片计算核心在任务运行中的实际占用比例，专用指令调用占比指任务执行过程中调用硬件专用加速指令的次数占总指令次数的比例。
大模型任务适配兼容性	验证协同方案对不同规模大模型的适配能力，覆盖100亿/500亿/1000亿/3000亿参数级模型，通过无适配报错率量化，计算方式为适配过程中出现兼容性错误的次数占总适配次数的比例。
国产化软硬件协同稳定性	连续72小时大模型训练/推理任务运行中，协同方案的故障中断次数和故障恢复时间。

A.2 服务等级指标项权重

服务等级指标项权重如表A.2所示：

表 A.2　服务等级指标项权重

一级指标	权重/%	二级指标	权重/%
可用性	20	服务可用度	40
		计算节点的可用性	20
		存储服务的可用性	10
		网络服务可用性	30
可靠性	15	数据存储的可靠性	25
		数据持久性	25
		数据私密性	25
		服务的可靠性	25
服务安全	5	服务安全	100
绿色低碳	5	电能利用效率	80
		可再生能源利用率	20
内容服务能力	25	内置算法模型数量	20
		内置算法模型类型覆盖度	20
		算法模型服务监测能力	20
		线上开发环境多样性	20
		模型全生命周期开发配套能力	20
软硬件协同加速能力	10	协同加速比	25
		硬件资源适配效率	25
		大模型任务适配兼容性	25
		国产化软硬件协同稳定性	25
其他	20	容量与弹性	20
		浮点计算能力	20
		网络运力	20
		服务响应性能	20
		服务资源弹性	20

附 录 B
（资料性）
监测指标

在智能算力服务等级协议中，如需对部分指标进行监测可参考表B.1和表B.2。

B.1 动态监测指标

动态监测指标如表B.1所示：

表 B.1 动态监测指标

主要类型	监测指标的名称	描述
CPU计算性能	CPU利用率	CPU的使用程度，高利用率意味着系统正在处理大量任务或存在资源瓶颈。
GPU计算性能	GPU利用率	反映了GPU的使用情况。
存储访问速度	延迟时间	指硬件或软件请求从发出到响应的时间间隔。
网络吞吐量	数据传输速率	衡量网络在单位时间内传输数据的能力。
	网络带宽利用率	展示了当前网络带宽的使用比例。
CPU资源使用情况	CPU利用率	显示总体、租户、项目维度，CPU整体资源的使用情况，值越大，代表对CPU的利用率越高。
GPU资源使用情况	GPU利用率	显示总体、租户、项目维度，GPU整体资源的使用情况，值越大，代表对GPU的利用率越高。
内存资源使用情况	内存利用率	显示内存整体资源的使用情况，值越大，代表对内存的利用率越高。
存储资源使用情况	存储利用率	显示存储整体资源的使用情况，值越大，代表对存储的利用率越高。
算法模型使用分析	调用次数总数	按照时间区间显示每一个模型服务的调用次数统计
	调用次数最大值	统计时间区间内每一个模型服务的调用次数的最大次数
	调用次数最小值	统计时间区间内每一个模型服务的调用次数的最小次数
	调用次数平均值	统计时间区间内每一个模型服务的调用次数的平均次数
	请求时长	统计每一个模型每次请求的请求时长
	Token消耗量	统计每一个模型每次请求的Token消耗量
算法服务平台状态监测	算法服务平台状态	显示实时算法服务平台的运行状态，是否可用

B.2 静态统计指标

静态统计指标如表B.2所示：

表 B.2 静态统计指标

主要类型	指标名称	描述
CPU计算性能	每秒操作数	衡量CPU在一秒钟内可以执行的操作数量，这直接反映了其处理能力。高性能的CPU通常具有较高的每秒操作数，这对于需要大量计算的任务尤其重要。
GPU计算性能	每秒浮点运算次数	衡量GPU在一秒钟内可以执行的浮点运算次数，这是评估其性能的重要指标。
存储访问速度	存储I/O性能	包括读写速度，反映了存储设备的数据传输能力。
CPU资源使用情况	CPU总量	显示总体、租户、项目维度，CPU总体资源的大小。
GPU资源使用情况	GPU总量	显示总体、租户、项目维度，GPU总体资源的大小。
内存资源使用情况	内存总量	显示内存总体资源的大小。
存储资源使用情况	存储总量	显示显示总体、租户、项目维度，存储总体资源的大小。