

团 体 标 准

T/SZ10T XXXX—2026

算力服务平台能力评估

Capability assessment of computing power service platforms

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX—XX—XX 发布

XXXX—XX—XX 实施

深圳市物联网产业协会 发 布

目 次

前 言 II

1 范围 3

2 规范性引用文件 3

3 术语和定义 3

4 缩略语 4

5 评估原则 4

 5.1 客观性原则 4

 5.2 系统性原则 4

 5.3 导向性原则 4

 5.4 可操作性原则 4

6 评估指标体系 4

 6.1 基础资源能力 5

 6.2 平台服务能力 5

 6.3 安全保障能力 6

 6.4 算电协同能力 6

7 评估方法 6

 7.1 基础资源能力评估方法 6

 7.2 平台服务能力评估方法 7

 7.3 安全保障能力评估方法 7

 7.4 算电协同能力评估方法 7

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。

本文件由深圳市物联网产业协会提出。

本文件由深圳市物联网产业协会归口。

本文件起草单位：深圳市标准技术研究院、中国质量认证中心有限公司、深圳市智慧城市大数据中心有限公司

本文件主要起草人：

算力服务平台能力评估

1 范围

本文件规定了智能算力服务平台能力评估的评估原则、评估指标体系、评估方法。

本文件适用于对提供智能算力资源服务化供给、调度与运营的平台（以下简称“平台”）进行能力评估，可用于平台的自我评估、第三方评估、行业对标及能力提升指导。

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
- GB/T 22239—2019 信息安全技术 网络安全等级保护基本要求
- GB/T 45958—2025 网络安全技术 人工智能计算平台安全框架
- GB 40879—2021 数据中心能效限定值及能效等级

3 术语和定义

下列术语和定义适用于本文件。

3.1

算力 `computing power`

计算系统的综合数据处理能力，从计算技术路线和计算体系结构可划分为通算算力、智算算力、超算算力等。通常用 FLOPS（Floating-point Operations Per Second，每秒浮点运算次数）作为计量单位。

3.2

通算算力 `general computing power`

基于通用处理器（如 CPU）构建的计算平台提供的计算能力，面向各种一般性计算任务，能够满足不同类型常规计算需求。

3.3

智算算力 `intelligent computing power`

基于人工智能计算体系结构（如 GPU、TPU、NPU、ASIC 等 AI 加速平台）提供、针对人工智能训练与推理等智能计算任务的计算能力。

3.4

超算算力 `high-performance computing (HPC) power`

由超级计算机或高性能计算系统提供的极大规模计算能力，用于执行复杂且计算密集型的科学计算

任务，是以高并行性、高性能为特征的算力类型。

3.5

算力服务平台 `computing power service platform`

对异构算力资源进行统一抽象、纳管、调度，并通过服务化接口向用户提供算力资源、算力任务或算力应用服务的软件平台。

4 缩略语

下列缩略语适用于本文件：

- AI：人工智能（Artificial Intelligence）
- GPU：图形处理器（Graphics Processing Unit）
- NPU：神经网络处理器（Neural Processing Unit）
- TPU：张量处理器（Tensor Processing Unit）
- ASIC：专用集成电路（Application Specific Integrated Circuit）
- FLOPS：每秒浮点运算次数（Floating-point Operations Per Second）
- SLA：服务等级协议（Service Level Agreement）
- MLPS：网络安全等级保护制度（Multi-Level Protection Scheme）
- CE：算效（Computational Efficiency）
- BOPS：每秒完成基本操作数（Basic Operations Per Second）
- PUE：电力使用效率（Power Usage Effectiveness）

5 评估原则

5.1 客观性原则

评估应基于可验证的数据和事实，评估方应确保评估过程的独立性与公正性，被评估方应提供真实、准确、完整的材料和数据。

5.2 系统性原则

评估应全面、系统，涵盖算力服务平台的资源供给、服务能力、安全保障、生态协同等多个维度，反映平台综合服务能力。

5.3 导向性原则

评估应引导算力服务平台向高效、绿色、智能、安全、可靠的方向发展，促进产业技术进步与服务模式创新。

5.4 可操作性原则

评估指标应定义清晰、可量化、可获取，评估方法应明确、具体，便于实施和复现。

6 评估指标体系

6.1 概述

算力服务平台能力评估指标体系包括4个一级指标：基础资源能力、平台服务能力、安全保障能力、算电协同能力。每个一级指标下设若干二级。具体评估指标体系框架见图1。

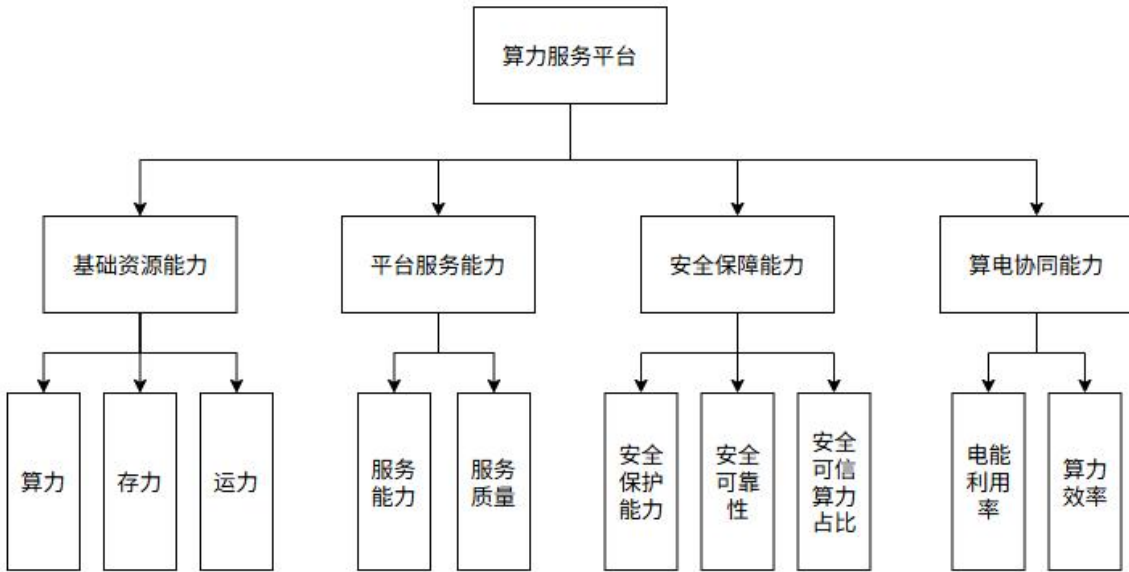


图1 算力服务平台评估指标体系

6.2 基础资源能力

6.2.1 算力供给能力

算力供给能力包含自建算力和调度算力两部分，为两部分算力的总和。

- a) 自建算力规模（FP16 PFLOPS）：平台可调度的自建智算算力的总和；
- b) 调度算力规模（FP16 PFLOPS）：平台基于协议或授权机制，对非自建算力资源进行接入并实现统一调度和使用的算力总和；

6.2.2 存力供给能力

反映算力中心对计算数据的承载规模和数据访问性能水平，是支撑高性能计算、人工智能训练与推理等算力应用的重要基础条件。

- a) 物理总容量（PB/EB）：存储系统所有硬件介质的标称容量总和（如硬盘、SSD、闪存的容量累加）；
- b) 可用容量（PB/EB）：物理总容量扣除 RAID 冗余、系统预留、坏块占用后的实际可分配容量；
- c) 容量利用率（%）：写入系统的数据量与可用容量的比值，一般控制在 70%–80%区间为理想状态；
- d) IOPS（ops/s）：数据存储设备的每秒最高可达的读写次数；
- e) 吞吐量（MB/s）：数据存力在具体的数据读写中的表现，衡量其对于业务产生各类数据应用需求满足的水平。

6.2.3 运力保障能力

运力指平台通过网络对数据与算力进行传输、调度和交付的能力，反映其支撑算力服务跨节点、跨区域运行的网络承载水平。

- a) 卡间带宽（GB/s）：卡间带宽是指同一计算节点内多块计算加速卡（GPU/TPU/NPU等）之间，单位时间内可双向传输的数据量，是衡量多卡并行计算场景下数据交互效率的核心指标，直接决定分布式AI训练任务的执行效率；
- b) 节点间带宽（GB/s）：平台服务器之间的互联速率。

6.3 平台服务能力

6.3.1 服务能力

服务能力主要考核弹性服务能力。弹性服务能力指平台根据用户业务负载的实时变化，自动、快速、按需调整算力、存力、运力等资源供给规模的能力。在扩缩容过程中，要求保持任务不间断。

- a) 扩容响应时间（s）：从触发扩容操作到新增资源完成配置、加入集群并可承载业务任务的总耗时；
- b) 缩容响应时间（s）：从触发缩容操作到闲置资源完成任务迁移、释放并退出集群的总耗时。

6.3.2 服务质量

- a) 可用性百分比（%）：服务可用性是指平台在规定时间内（如年度、月度），持续提供符合质量要求的算力服务的能力，可用性百分比是度量服务可用性的指标，计算公式如下公式（1）所示。

可用性百分比 = $\frac{\text{总时长}-\text{不可用时长}}{\text{总时长}} \times 100\%$ （1）

- b) 技术支持响应时间（h）：指从平台用户通过【工单系统/指定邮箱/专属对接人】提交服务请求，并提供完整故障信息的时刻起，至平台技术人员首次与用户取得联系、确认问题的时间间隔。表述可参考“2小时 5×8”。

6.4 安全保障能力

安全保障用于衡量平台的安全可信与自主可控水平。

6.4.1 安全保护能力

支持弱口令检测、漏洞检测、勒索病毒防护及20+类入侵检测等主机安全服务能力。日志记录采用GB/T 45958—2025第6.3条规定，要求平台需维护数据处理、模型训练和推理全流程的日志记录，并对日志访问进行控制。日志要求默认保留至少半年。

6.4.2 数据可靠性

- a) 持久性（%）：在给定时间周期内，数据保持完整且不丢失的概率，通常用多个“9”来表示，如99.99999999%（11个9）；
- b) 数据备份与恢复：平台需支持手动备份和恢复、定期自动数据备份、节点间数据同步备份、异常情况下数据恢复能力；

6.4.3 安全可信算力占比

采用通过安全可信认证的国产芯片提供的算力占总算力的比重（%）。

6.5 算电协同能力

算电协同能力用于衡量平台的绿色低碳与可持续发展能力。

- a) 电能利用效率，考核PUE，计算方式如公式（2）所示：
 $PUE = \text{数据中心总能耗} / \text{IT设备能耗}$ （2）
- b) 算力效率（算效比，单位：FLOPS/W）是衡量单位能耗算力产出的核心效率指标，如下公式（3）进行计算：
 $CE = \text{算力总规模（BOPS）} / \text{IT设备总功耗（W）}$ （3）

7 评估方法

7.1 基础资源能力评估方法

7.1.1 算力供给能力评估方法

- a) 自建算力规模评估 由被评估方提供加盖公章的算力部署采购服务资源到货清单复印件、采购凭证（包括但不限于发票、银行转账凭证）、加盖公章的算力清单汇总表等。
- b) 调度算力规模评估 由被评估方提供算力资源调度协议或授权书。

7.1.2 存力供给能力评估方法

被评估方提供系统相关指标的截图并加盖公章，并接受评估方的现场/远程视频抽查。

7.1.3 运力保障能力评估方法

被评估方可以在下述方法中选择一个进行：

- a) 提供硬件、软件说明材料校验卡间RDMA带宽、节点间带宽情况；
- b) 现场/远程视频演示。

7.2 平台服务能力评估方法

7.2.1 弹性服务能力评估方法

现场/远程视频演示扩缩容的操作动作，并计算节点级弹性扩缩容时间。

7.2.2 服务质量评估方法

- a) 可用性百分比
被评估方提供历史周期性服务系统监控数据证明，如：加盖公章的历史周期性服务系统监控数据（硬件层面、软件层面），计算整体服务在线时间占比。
- b) 技术支持响应时间
被评估方提供与客户签订的SLA协议中关于服务可用性的条款以及实际的达成情况说明。

7.3 安全保障能力评估方法

7.3.1 安全保护能力

被评估方提供资质认定证明，包括中国质量认证中心认定的CQC测评证明、网络安全等级保护证明（GB/T 22239—2019 MLPS 3级）、密码应用安全性评估证明等。被评估方提供至少留存半年的数据处理、模型训练和推理全流程的日志记录。

7.3.2 数据可靠性

- a) 数据持久性：
被评估方可以在下述方法中选择一个进行：
 - 1) 基于平台历史运行数据，对数据丢失事件发生情况进行统计分析；
 - 2) 基于故障场景分析与理论计算，被评估方提供持久性指标的数学推导过程。
- b) 数据备份与恢复：现场/远程视频演示各项备份恢复功能。

7.3.3 安全可信算力占比

根据被评估方提供的资料，计算7.1.1总算力中，国产芯片提供算力在总算力中的占比（%）。

7.4 算电协同能力评估方法

算电协同需要分别对自建算力和调度算力分别评估。

- a) PUE评估
依据GB 40879—2021，采用全年或分时标准工况测量，计算总能耗与IT能耗之比，并按国标能效等级进行判定。
- b) CE算效评估
将算力做FP16归一化处理，设备总功耗在负载率达到 30% - 80% 的典型运行状态下取均值。