

T/ISC

团 体 标 准

T/ISC XXXX—XXXX

面向大模型训练与推理的算力费用核算模型

Comprehensive pricing and billing model for computing power trading
in large model training

（征求意见稿）

XXXX – XX – XX 发布

XXXX – XX – XX 实施

中国互联网协会 发 布

目 次

前 言 II

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 总体架构 2

5 核算资源要素、核算单元与数据计量 4

6 核算数据采集与处理 7

7 算力费用核算模型 9

8 训练算力费用核算模型 10

9 推理算力费用核算模型 11

10 核算规则 13

11 核算结果调整及综合定价应用 14

12 计费方式与核算结果应用 15

13 服务等级协议与费用核算关联 16

14 结算与支付 18

15 试验方法 19

附 录 A （资料性） 主流 GPU/NPU 性能系数参考表 22

附 录 B （资料性） 大模型训练计算精度类型折算系数表 23

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归口。

本文件起草单位：

本文件主要起草人：

面向大模型训练与推理的算力费用核算模型

1 范围

本文件规定了大模型训练与推理场景下算力费用核算模型的总体架构、核算资源要素、核算单元与数据计量、核算数据采集与处理、算力费用核算模型、训练算力费用核算模型、推理算力费用核算模型、核算规则、核算结果调整及综合定价应用、计费方式与核算结果应用、服务等级协议与费用核算关联、结算与支付以及试验方法等内容。

本文件适用于智算中心、云计算平台、算力调度平台等为大规模深度学习模型训练与推理提供的GPU、NPU、TPU等异构算力资源（含裸金属、容器、虚拟化实例）的交易定价与计费，也可作为算力交易双方签订合同的参考。

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 35274 信息安全技术 大数据服务安全能力要求
GB/T 36074 信息技术服务 服务管理
GB/T 36326 信息技术 云计算 云服务运营通用要求
GB/T 37739 信息技术 云计算 平台即服务部署要求
YD/T 3213 内容分发网络服务信息安全管理系统及接口测试方法

3 术语和定义

下列术语和定义适用于本文件。

3.1

大模型训练 large model training

基于深度学习框架，对包含海量参数（通常指参数规模超过十亿量级）的人工神经网络模型，利用大规模并行异构计算资源进行数据学习、权重优化和收敛的过程。

3.2

大模型推理 large model inference

基于已训练模型，对输入数据进行计算并生成预测结果、分类结果、向量表示、Token序列或其他模型输出的过程。

3.3

算力费用 compute expense

大模型训练或推理过程中因使用算力资源及与其直接或间接相关的存储、网络、软件、平台、数据处理、运维和服务等资源形成的货币化核算金额。

注：算力费用可按照成本核算、内部结算或交易结算等不同口径形成。涉及交易价格时，其价格确定方式应与费用核算结果相区分。

3.4

算力费用核算模型 compute expense accounting model

针对规定的大模型训练或推理场景，以明确的核算对象和核算边界为基础，对算力及相关资源消耗进行计量，对直接费用进行归集、对间接费用进行合理分摊，并依据核算单元、费用参数、核算规则和调整机制计算形成算力费用核算结果的方法体系。

3.5

算力计费 compute billing

基于费用核算结果、资源价格、交易规则和合同约定，计算算力服务使用方应支付费用的过程。

3.6

算力价格 **compute price**

单位算力资源、核算单元或者计费单元在交易中的货币价格。

3.7

算力交易 **compute power trading**

算力资源提供方与使用方通过交易平台或者其他方式，对算力资源进行配置、购买、调度、使用、计量、核算、计费、结算和服务交付的活动。

3.8

服务等级协议 (SLA) **Service Level Agreement**

算力提供方与使用方之间就提供的算力服务的质量、可用性、性能、服务响应、故障回复等指标所达成的正式承诺协议，以及当服务未达到承诺标准时的补偿或赔偿机制。

3.9

训练任务 **training job**

为实现规定模型训练目标而提交并执行的工作负载。

训练任务可包括数据准备、预训练、继续预训练、监督微调、强化学习、模型评估、Checkpoint保存和恢复等相关阶段。

3.10

有效算力 **effective compute**

在规定任务和软件环境下，能够实际产生模型训练或推理工作成果的计算能力。

3.11

算力性能系数 **performance coefficient**

表示特定算力资源相对于规定基准算力资源在特定训练或推理工作负载下实际处理能力差异的无量纲系数。

4 总体架构

4.1 总体要求

4.1.1 大模型训练与推理算力费用核算应建立以资源消耗为基础、以核算对象为载体、以可计量数据为依据、以费用归集和分摊为核心的费用核算模型。

4.1.2 费用核算宜按照下列逻辑实施：

- a) 应用场景识别；
- b) 核算对象和核算边界确定；
- c) 算力及相关资源识别；
- d) 资源使用量计量；
- e) 直接费用和间接费用识别；
- f) 核算单元确定；
- g) 费用归集和分摊；
- h) 训练或推理费用计算；
- i) 异常事项及 SLA 调整；
- j) 形成最终费用核算结果；
- k) 用于定价、计费、合同结算及管理分析。

4.1.3 费用核算模型应能够支持 GPU、NPU、TPU 等异构算力资源，并应支持裸金属、虚拟机、容器等不同资源交付形态。

4.2 核算基本原则

算力费用核算应符合以下原则：

- a) 真实性：核算结果应基于实际资源配置、实际资源使用记录或者经验证的工作量数据；
- b) 一致性：同类核算对象在同一核算周期内应采用一致的核算口径和规则；
- c) 可追溯性：资源使用量、费用参数、分摊依据和调整记录应能够追溯；

- d) 场景适配性：训练和推理应采用与其资源消耗特点相适配的核算单元和核算方法；
- e) 避免重复核算：同一资源费用不应在多个费用项目中重复归集；
- f) 可扩展性：核算模型应能够适应异构硬件、新型模型架构及新的算力服务模式；
- g) 可验证性：核心计量数据和计算过程应能够通过试验、日志或第三方数据进行验证；
- h) 核算与定价适度分离：基础费用核算结果宜与市场折扣、营销优惠及交易价格调整分别记录。

4.3 核算对象

4.3.1 训练场景的核算对象可包括：

- a) 单次训练任务；
- b) 训练任务的一个阶段；
- c) 训练作业或工作负载；
- d) 项目；
- e) 租户；
- f) 训练集群或资源池。

4.3.2 推理场景的核算对象可包括：

- a) 推理请求；
- b) 批量推理任务；
- c) 模型实例；
- d) 模型服务；
- e) 模型；
- f) 项目；
- g) 租户；
- h) 推理资源池。

4.3.3 对于共享算力资源，应建立从资源池到任务、模型服务、模型、项目或租户的费用分摊关系。

4.4 核算边界

4.4.1 训练任务核算边界宜根据实际场景包括：

- a) 任务资源分配；
- b) 数据预处理；
- c) 训练环境初始化；
- d) 模型加载；
- e) 正式训练；
- f) 模型评估；
- g) Checkpoint 保存；
- h) Checkpoint 恢复；
- i) 任务清理及资源释放。

4.4.2 推理服务核算边界宜包括：

- a) 模型装载；
- b) 模型实例运行；
- c) 推理请求处理；
- d) KV Cache 等推理缓存资源；
- e) 动态批处理；
- f) 服务扩缩容；
- g) 请求排队；
- h) 输出生成；
- i) 服务空闲但因容量保障而持续占用的资源；
- j) 资源释放。

4.4.3 服务初始化、模型下载、冷启动等过程是否纳入具体核算对象，应根据资源责任归属和核算目的明确。

4.5 核算维度

4.5.1 费用核算模型应支持按一个或多个维度形成核算结果，包括但不限于：

- a) 时间维度；
- b) 设备维度；
- c) 算力类型维度；
- d) 资源规格维度；
- e) 任务维度；
- f) 模型维度；
- g) 模型版本维度；
- h) 项目维度；
- i) 租户维度；
- j) 地域维度；
- k) 资源池维度；
- l) 训练阶段维度；
- m) 推理请求维度；
- n) 模型服务维度；
- o) 服务等级维度。

4.5.2 不同核算维度的数据应具有可关联的唯一标识。

5 核算资源要素、核算单元与数据计量

5.1 核算资源要素

算力费用核算涉及的资源要素包括计算资源、存储资源、网络资源、软件 and 平台资源、数据处理资源、运维服务资源以及其他必要资源。

核算时不要求所有资源均单独形成交易计费项目，但应明确其是否纳入费用核算，以及采用直接归集、间接分摊或者包含于其他费用项目中的处理方式。

5.2 计量资源

5.2.1 计算资源包括但不限于：

- a) GPU；
- b) NPU；
- c) TPU；
- d) CPU；
- e) 其他专用 AI 加速器。

5.2.2 对于计算资源，应至少记录：

- a) 资源类型；
- b) 设备型号或规格；
- c) 设备数量；
- d) 资源实例数量；
- e) 资源分配时间；
- f) 资源释放时间；
- g) 实际运行时间；
- h) 资源使用率；
- i) 显存或内存配置及使用情况。

5.3 显存和内存资源

显存、主机内存等资源可根据业务形态独立核算，也可纳入加速卡或实例综合资源费用。

推理场景下，当多个模型实例共享同一加速设备时，显存占用及显存占用时长宜作为共享资源费用分摊的重要依据。

对于长上下文推理、持续会话以及大量KV Cache占用场景，宜记录因上下文长度、缓存时间等引起的显存和内存使用变化。

5.4 存储资源

5.4.1 存储资源包括：

- a) 本地存储；
- b) 共享文件存储；
- c) 对象存储；
- d) 模型仓库；
- e) 镜像仓库；
- f) Checkpoint 存储；
- g) 推理缓存和日志存储。

5.4.2 存储资源可采用 GB·h、TB·h、实际存储容量周期或者其他约定单位核算。

5.4.3 读写请求量、IO 吞吐或高性能存储服务等级对费用有显著影响时，可单独纳入核算。

5.5 网络资源

5.5.1 网络资源包括：

- a) 节点内和节点间高速互联；
- b) InfiniBand 网络；
- c) RoCE 网络；
- d) 以太网；
- e) 跨资源池网络；
- f) 公网流量；
- g) 数据传输服务。

5.5.2 网络费用可根据网络资源预留、使用时间、带宽、数据传输量或者服务等级进行核算。

5.5.3 训练场景宜重点考虑分布式训练产生的节点间通信资源。

5.5.4 推理场景宜重点考虑高并发请求、跨节点推理、模型服务入口流量和输出数据流量。

5.6 软件 and 平台资源

5.6.1 软件和平台资源包括但不限于：

- a) 操作系统；
- b) AI 计算框架；
- c) 容器和虚拟化平台；
- d) 集群调度平台；
- e) 模型服务平台；
- f) 推理引擎；
- g) 监控、日志和审计服务；
- h) 商业软件许可；
- i) 模型管理及 MLOps 服务。

5.6.2 能够直接归属于单一任务或服务的，应作为直接费用归集；无法直接归属的，可作为间接费用进行分摊。

5.7 数据处理资源

训练和推理过程中因数据清洗、预处理、数据转换、Tokenization、Embedding生成、数据加载、数据缓存及其他数据处理形成的资源使用量，可根据核算边界纳入综合费用核算

5.8 运维和服务资源

5.8.1 运维及服务费用可包括：

- a) 集群运行维护；
- b) 故障处置；

- c) 技术支持;
- d) 资源调度服务;
- e) 专属服务;
- f) 安全保障服务;
- g) 其他合同约定的运营服务。

5.8.2 运维服务无法直接归属单一核算对象时，宜按照合理、稳定、可验证的分摊依据进行分摊。

5.9 核算单元选择原则

5.9.1 核算单元应根据应用场景、资源使用方式、数据可获得性和核算目的确定。

5.9.2 核算单元选择应满足：

- a) 能够合理反映核算对象的资源消耗;
- b) 能够通过可靠数据进行计量;
- c) 不同核算周期具有一致性;
- d) 能够支持费用追溯和验证;
- e) 不应因核算单元选择造成明显的重复核算;
- f) 业务工作量指标与资源消耗之间应具有可解释的关联关系。

5.9.3 不得将某一种核算单元作为所有训练和推理场景下的唯一核算单位。

5.10 训练场景核算单元

5.10.1 训练场景可采用下列一个或多个核算单元：

- a) GPU · h、NPU · h、TPU · h;
- b) 加速卡实例 · h;
- c) SCU · h;
- d) CPU 核 · h;
- e) 内存 GB · h;
- f) 存储 GB · h 或 TB · h;
- g) 网络流量或带宽时间;
- h) 训练任务 · h;
- i) 训练 Token 数量;
- j) 训练 step 或迭代次数;
- k) 其他能够反映训练资源消耗的指标。

5.10.2 训练 Token 数量、训练 step、Batch Size 等指标宜作为任务工作量、性能分析或费用分摊依据，不宜在缺少资源消耗映射关系时单独作为综合算力费用核算的唯一依据。

5.11 推理场景核算单元

5.11.1 推理场景可采用：

- a) GPU · h、NPU · h、TPU · h;
- b) 实例 · h;
- c) 模型服务运行时间;
- d) 请求数量;
- e) 输入 Token 数量;
- f) 输出 Token 数量;
- g) 总 Token 数量;
- h) 有效推理 Token 数量;
- i) 推理运行时间;
- j) 加速设备活动时间;
- k) 显存占用量及占用时间;
- l) 并发请求量;
- m) 吞吐量;
- n) 其他能够合理表征推理资源消耗的指标。

- 5.11.2 在线推理、实时推理等场景宜综合考虑持续资源预留和实际业务工作量。
- 5.11.3 离线推理和批量推理宜优先依据任务实际占用资源及任务工作量进行核算。

5.12 标准算力单位

- 5.12.1 使用 SCU 进行异构算力折算时，应明确：
 - a) 基准设备；
 - b) 基准模型；
 - c) 任务类型；
 - d) 模型规模；
 - e) 计算精度；
 - f) 软件框架及版本；
 - g) 并行策略；
 - h) 测试数据；
 - i) 节点及网络配置；
 - j) 评价指标；
 - k) 折算方法。
- 5.12.2 训练 SCU 与推理 SCU 可采用不同性能标定方法。
- 5.12.3 未经等价测试验证，不宜直接将训练场景性能系数用于推理场景。

6 核算数据采集与处理

6.1 一般要求

费用核算所使用的数据应能够完整支持资源计量、费用归集、费用分摊、异常处理和结果复核。数据应具有统一的时间基准、核算对象标识和资源标识。

6.2 数据类型

应根据场景采集下列数据中的适用项：

- a) 计算资源配置数据；
- b) 算力资源分配和释放数据；
- c) GPU、NPU、TPU 及 CPU 实际使用数据；
- d) 显存和内存使用数据；
- e) 资源利用率数据；
- f) 存储使用量和 IO 数据；
- g) 网络资源及数据传输数据；
- h) 任务开始、暂停、恢复和结束时间；
- i) 模型及模型版本信息；
- j) Token 数据；
- k) 推理请求数据；
- l) 任务运行状态；
- m) 并发量和 Batch 信息；
- n) SLA 监测数据；
- o) 资源价格或单位费用参数；
- p) 平台服务费用参数；
- q) 其他核算所需数据。

6.3 训练数据采集

训练任务宜采集：

- a) 模型参数规模；
- b) 训练数据规模；

- c) 训练 Token 数量;
- d) Batch Size;
- e) 序列长度;
- f) 训练 step;
- g) 训练时间;
- h) 各类加速设备数量及运行时间;
- i) GPU、NPU 或 TPU 利用率;
- j) 显存占用;
- k) 训练吞吐量;
- l) 通信效率;
- m) 数据并行、模型并行、流水线并行及其他并行方式;
- n) Checkpoint 生成、保存及恢复信息;
- o) 模型评估时间;
- p) 任务暂停、中断、失败和重试记录。

6.4 推理数据采集

推理服务宜根据业务模式采集:

- a) 输入 Token 数量;
- b) 输出 Token 数量;
- c) 请求数量;
- d) 请求开始和结束时间;
- e) 请求处理时长;
- f) TTFT;
- g) Token 生成速率;
- h) 并发请求数;
- i) Batch Size;
- j) 上下文长度;
- k) 模型实例数量;
- l) 实例运行时间;
- m) 加速设备实际活动时间;
- n) GPU、NPU 或 TPU 利用率;
- o) 显存占用;
- p) KV Cache 等缓存资源占用;
- q) 动态批处理信息;
- r) 服务扩缩容信息;
- s) 峰值负载和平均负载;
- t) 请求成功、失败及重试状态;
- u) SLA 等级及相关性能数据。

6.5 数据采集频率和精度

6.5.1 采集频率应与最小核算粒度相适应,并能够识别资源短时使用、任务启动和终止、服务扩缩容及异常中断等情况。

6.5.2 对于设备利用率、显存使用量、服务时延等变化较快的数据,宜采用秒级或能够满足核算精度要求的更细粒度采样方式。

6.5.3 对通过事件日志能够准确确定的任务开始时间、资源分配时间、请求数量、Token 数量等数据,宜优先采用事件记录。

6.6 时间同步

6.6.1 参与资源计量和核算的系统应使用统一时间源。

6.6.2 关键资源事件、请求事件和费用事件的时间戳应能够建立对应关系。

6.7 数据清洗和校验

核算数据应进行完整性、逻辑一致性和异常值校验。至少应校验：

- a) 任务时间与资源分配时间关系；
- b) 资源配置与实际实例关系；
- c) Token 数量和请求数量完整性；
- d) 资源使用量是否存在重复统计；
- e) 跨系统时间戳是否一致；
- f) 异常设备数据是否影响费用计算；
- g) 价格或费用参数版本是否与核算周期匹配。

6.8 数据聚合与归集

核算数据应进行完整性、逻辑一致性和异常值校验。至少应校验：

- a) 任务时间与资源分配时间关系；
- b) 资源配置与实际实例关系；
- c) Token 数量和请求数量完整性；
- d) 资源使用量是否存在重复统计；
- e) 跨系统时间戳是否一致；
- f) 异常设备数据是否影响费用计算；
- g) 价格或费用参数版本是否与核算周期匹配。

6.9 异常数据处理

核算数据应进行完整性、逻辑一致性和异常值校验。至少应校验：

- a) 任务时间与资源分配时间关系；
- b) 资源配置与实际实例关系；
- c) Token 数量和请求数量完整性；
- d) 资源使用量是否存在重复统计；
- e) 跨系统时间戳是否一致；
- f) 异常设备数据是否影响费用计算；
- g) 价格或费用参数版本是否与核算周期匹配。

6.10 数据可信性

6.10.1 核算数据应具备防止未经授权修改的技术措施。

6.10.2 关键核算数据宜保留原始记录、聚合结果和计算结果之间的对应关系。

6.10.3 数据保存期限应覆盖合同结算、对账、争议处理及法律法规要求的期间。

7 算力费用核算模型

7.1 一般要求

费用核算模型应根据核算对象分别识别直接费用和间接费用，并根据资源计量结果形成基础费用核算结果。综合算力费用按照公式（1）计算：

$$C = C_1 + C_2 + C_3 + C_4 \quad (1)$$

式中：

C ——最终算力费用核算结果；

C_1 ——直接费用；

C_2 ——间接费用分摊金额；

C_3 ——核算边界内其他费用；

C_4 ——异常事项、SLA等形成的费用调整金额，可为正值、负值或零。

7.2 直接费用核算

直接费用宜按照资源实际使用量与相应单位费用计算，资源类别可包括计算、显存或内存、存储、网络、软件、平台、数据处理和其他能够直接归属的资源。

当单位费用已经包含某项配套资源时，该项资源不应重复计算。

7.3 间接费用分摊

7.3.1 公共平台、公共网络、共享存储、调度平台、资源池管理和运维服务等不能直接归属于单一核算对象的费用，可进行间接费用分摊。费用分摊权重可采用：

- a) 计算资源占用时间；
- b) 标准算力使用量；
- c) 显存或内存容量时；
- d) 存储容量时；
- e) 数据传输量；
- f) 请求量；
- g) Token 工作量；
- h) 项目资源配额；
- i) 多个指标形成的组合权重。

7.3.2 分摊规则应具有稳定性和可解释性，不宜频繁变化。

8 训练算力费用核算模型

8.1 一般要求和构成

训练任务综合费用需要指训练任务最终费用核算结果。包括：

- a) GPU、NPU、TPU 等加速算力资源费用；
- b) CPU、内存、存储、网络等配套资源费用；
- c) 数据处理相关费用；
- d) 训练平台、软件和相关服务费用；
- e) 间接费用分摊；
- f) 失败重试、异常中断、SLA 等训练相关调整金额。

8.2 按资源实际占用量核算

8.2.1 采用资源实际占用方式时，第 h 类加速资源使用量按照公式（2）计算：

$$Q(h) = N(h) \times T(h) \quad (2)$$

式中：

$Q(h)$ ——第 h 类加速资源使用量，单位可为卡时；

$N(h)$ ——实际分配的加速卡数量；

$T(h)$ ——资源实际占用时间。

8.2.2 实际加速资源费用即把第 h 类加速资源单位费用与对应使用量相乘，再求和。

8.3 按有效算力核算

8.3.1 需反映不同异构资源的实际训练贡献时，可采用有效算力进行核算。

8.3.2 采用有效算力核算时，应保证基准性能、性能系数及单位费用属于同一核算体系。

8.3.3 不应同时对同一性能差异在资源使用量和资源单价中重复调整。

8.4 模型规模和任务工作量

8.4.1 模型参数规模、训练 Token 数量、数据集规模、Batch Size、序列长度和训练 step 可作为训练任务工作量分析及资源费用归集的辅助依据。

8.4.2 上述指标不宜脱离实际资源使用数据直接推导最终费用。

8.4.3 在资源遥测数据不可获得且需要按任务工作量进行核算时，应通过基准测试建立任务工作量与资源使用量之间的映射关系。

8.5 并行训练

8.5.1 采用数据并行、模型并行、流水线并行、专家并行或组合并行时，应计量全部参与训练的加速设备、CPU、内存、存储和网络资源。

8.5.2 由并行策略导致的额外通信资源应纳入训练任务核算。

8.5.3 通信效率可用于有效算力评价，但与已经单独核算的网络费用不应产生重复费用调整。

8.6 Checkpoint 和模型评估

8.6.1 Checkpoint 生成、保存和恢复产生的计算、存储和网络资源可计入训练任务费用。

8.6.2 模型评估与训练任务属于同一项目且使用相同核算边界时，可归入训练任务；需独立评价费用时，可建立单独核算对象。

8.7 资源利用率

8.7.1 训练过程中应记录 GPU、NPU、TPU 及配套资源利用率。资源利用率可用于：

- a) 判断资源使用效率；
- b) 计算有效算力；
- c) 识别异常资源空转；
- d) 确定共享资源分摊权重；
- e) 分析费用合理性。

8.7.2 采用资源预留方式核算时，低利用率不必然导致费用自动降低。

8.7.3 因算力服务提供方资源调度、硬件故障或平台异常造成的明显资源低效，应按照责任归属及 SLA 调整规则处理。

8.8 失败和重试

8.8.1 训练任务失败或重试时，应记录：

- a) 失败时间；
- b) 失败原因；
- c) 已产生的资源使用量；
- d) Checkpoint 状态；
- e) 重试所产生的资源使用量；
- f) 责任归属。

8.8.2 由使用方模型、代码、数据或配置原因导致的失败和重试，其实际资源消耗可根据约定纳入核算。

8.8.3 由服务提供方硬件、平台、网络或其他服务故障导致的失败和重复计算，不宜直接由使用方承担全部重复资源费用，应依据 SLA 和合同约定进行调整。

8.9 断点续训

采用Checkpoint恢复后继续训练的，应避免对已完成并经核算的训练阶段重复计量。恢复Checkpoint所产生的资源使用可按照实际使用量核算。

9 推理算力费用核算模型

9.1 一般要求

9.1.1 推理费用核算不应仅依据“设备数量×运行时间”形成单一模型。

9.1.2 推理核算应建立资源消耗与请求、Token、并发、实例运行时间及服务等级等业务工作量之间的关联，并根据推理服务形态选择资源占用核算、业务工作量核算或混合核算。

9.2 推理费用构成

9.2.1 推理最终费用核算结果：

- a) GPU、NPU、TPU 及 CPU 等计算费用；

- b) 显存和内存费用，未包含在计算实例费用时适用；
- c) 模型文件、缓存及其他存储费用；
- d) 网络及数据传输费用；
- e) 模型服务平台、推理引擎及其他平台服务费用；
- f) 间接费用分摊；
- g) 异常事项及 SLA 调整金额。

9.3 基于资源占用的推理核算

9.3.1 专属模型实例、离线推理、批量推理或资源能够直接归属于核算对象时，可采用资源实际占用方式。

9.3.2 模型服务持续运行并为满足 SLA 要求预留实例时，实例空闲时间属于提供该模型服务必要的容量成本，可纳入模型服务核算。

9.4 基于请求和 Token 工作量的核算

9.4.1 多个请求共享同一模型实例或资源池，且难以将加速设备使用量直接归属到单一请求时，可采用请求及 Token 工作量作为费用归集或分摊依据。

9.4.2 Token 核算宜至少区分输入 Token 和输出 Token。输入和输出 Token 对算力、显存访问和生成时延的影响存在差异时，不宜默认其单位资源消耗完全相同。

9.4.3 采用 Token 核算时，可包含输入 Token 数量和输出 Token 数量。经资源消耗标定形成的单位输入 Token 费用，经资源消耗标定形成的单位输出 Token 费用。

9.4.4 系数宜通过规定模型、硬件、软件环境和服务配置下的实际资源消耗进行标定，不宜仅依据市场销售价格反推资源费用。

9.5 基于请求数量的核算

9.5.1 请求处理资源消耗相对稳定，且请求间工作量差异较小时，可采用请求数量作为核算单元。

9.5.2 请求间输入长度、输出长度、上下文长度或者模型复杂度差异较大时，不宜单独采用请求数量进行费用归集。

9.6 混合核算

9.6.1 在线模型服务宜采用“资源容量费用+业务工作量费用”组合核算。

9.6.2 采用该方式时，应防止同一计算资源同时在容量费用和工作量费用中重复核算。

9.7 共享推理资源费用分摊

多个模型、租户或请求共享同一加速资源时，宜优先使用可以直接测量的资源指标进行费用分摊。可采用：

- a) 加速设备活动时间；
- b) 显存容量时；
- c) 计算内核活动量；
- d) 请求执行时长；
- e) 输入和输出 Token；
- f) 请求量；
- g) 多个指标构成的组合权重。

9.8 长上下文推理

长上下文推理应考虑输入Token增加导致的计算量、显存或内存使用、KV Cache规模和单请求持续时间变化。仅按请求次数核算无法合理反映资源消耗时，应增加Token、显存容量时或请求处理时间等核算指标。

9.9 多轮交互和流式推理

9.9.1 多轮交互场景应能够识别每轮输入、历史上下文及输出 Token。

9.9.2 流式推理场景宜记录：

- a) TTFT;
- b) 输出 Token 数量;
- c) Token 生成速率;
- d) 连接持续时间;
- e) 计算资源持续占用情况。

9.9.3 不得因输出采用流式返回形式而忽略实际推理资源消耗。

9.10 高并发和动态批处理

9.10.1 高并发推理应记录并发请求数、排队时间、Batch Size 和资源利用率。

9.10.2 采用动态批处理时，多个请求共享一次计算批次的资源费用宜按照能够反映各请求资源贡献的规则进行归集或分摊。

9.11 多模型和多实例推理

9.11.1 同一计算设备部署多个模型时，应记录各模型实例的显存占用、服务运行时间和实际计算使用情况。

9.11.2 同一模型部署多个实例时，应将扩缩容形成的资源变化纳入模型服务核算。

9.12 空闲资源

9.12.1 为满足在线推理低时延、高可用或峰值负载要求而长期保留的资源，其空闲期间资源费用可计入模型服务综合费用。

9.12.2 如资源因服务提供方调度效率不足产生非必要空闲，不宜全部作为使用方业务工作量费用进行分摊。

9.12.3 规定账单周期（小时、天、月）、账单内容（资源明细、调整系数、SLA 赔偿、费用汇总）的规范格式。结算系统应依据统一计量结果生成账单，并保证账单内容完整、准确、可追溯。账单生成、支付、退款、补偿等全过程应保留完整记录。

10 核算规则

10.1 按实际资源使用量核算

资源能够准确计量且直接归属于核算对象时，应优先采用实际资源使用量核算。

10.2 按资源预留量核算

使用方要求独占或预留算力资源时，可按预留规格和预留时间核算。资源预留量、实际使用量及两者差异应分别记录。

10.3 按实际运行时长核算

采用资源实例方式提供服务时，可按实例实际运行时间核算。运行时间起止条件应明确，不应仅依赖订单创建时间。

10.4 按任务工作量核算

训练Token、训练step、批处理数据量等任务工作量可在建立资源消耗映射关系后用于费用核算或分摊。

10.5 按 Token 核算

Token核算主要适用于生成式大模型推理。输入Token、输出Token和其他需要独立识别的Token类型应分别计量。Token统计应采用确定的Tokenizer及模型版本。

10.6 按请求量核算

请求资源消耗较稳定时，可按成功处理的请求数量核算。失败请求是否计入应结合失败原因确定。

10.7 按实例核算

固定规格模型服务、专属推理实例或训练实例可按实例数量和实例运行时间核算。

10.8 按综合资源使用量核算

当CPU、加速设备、存储和网络等多个资源均显著影响费用时，宜采用综合资源核算模型。

10.9 混合核算

混合核算可组合预留资源、实际资源使用量、Token、请求、实例和其他核算单元。组合核算应明确各组成部分的费用边界，防止重复核算。

10.10 最低资源使用量和保底资源

为满足资源启动、最小实例规格、模型显存装载或者SLA容量保障而存在最低资源需求时，可设置最低核算资源量。最低资源量及其适用条件应事先明确。

10.11 超额资源使用

实际资源使用量超过约定配额时，超额部分应单独记录。超额资源的交易价格可以不同，但基础费用核算应保持可追溯。

10.12 资源利用率调整

使用有效算力核算或共享资源费用分摊时，可采用资源利用率进行调整。如已按照资源实际占用量完整核算基础成本，不宜仅因利用率偏低自动扣减资源费用，应结合资源预留方式及责任归属处理。

10.13 资源共享分摊

共享资源费用应采用能够反映资源消耗或资源受益程度的指标进行分摊。分摊规则发生变化时，应保留规则版本和生效时间。

10.14 任务失败和重试

失败和重试费用应根据故障原因、责任归属及实际资源消耗确定。服务提供方原因造成的重复资源消耗，应按照SLA或合同规则进行费用调整。

10.15 异常中断

因停电、硬件故障、网络故障、平台故障或其他异常导致任务中断时，应记录异常期间实际资源使用量及任务有效工作量。异常期间资源费用应依据责任和SLA规则确定。

10.16 SLA 未达标

SLA未达到约定要求时，应在基础费用核算完成后实施核算结果调整。试验应验证综合定价模型的准确性、公平性、一致性及可重复性。试验环境应保持一致。

11 核算结果调整及综合定价应用

11.1 基础费用核算结果

基础费用核算结果应基于实际资源消耗、直接费用和间接费用分摊形成。市场供需、营销优惠、长期合同折扣等交易性因素宜在基础费用核算结果形成后另行处理。

11.2 算力性能调整

异构硬件采用标准算力或者有效算力核算时，可使用性能系数。性能系数应根据实际训练或推理性能标定，不宜仅依据理论峰值算力确定。

11.3 时间因素

峰谷时段、资源紧张程度等可用于交易价格调整。时间因素属于定价或计费应用时，不宜反向改变已发生的实际物理资源使用量。

11.4 服务质量因素

服务质量可通过SLA调整金额或者服务等级对应价格实现。对同一服务质量因素，不应同时在基础单价和SLA补偿中重复调整。

11.5 合约周期和购买规模

长期合约、资源预留、大规模采购等因素可用于交易价格调整。相关调整应与实际算力费用核算记录分离保存。

11.6 安全及合规服务

专属资源、安全隔离、可信计算、专项审计等额外服务形成的资源或服务成本，可纳入费用核算。仅属于商业定价策略而无对应资源或服务消耗的，应作为价格调整项目处理。

11.7 综合交易价格

11.7.1 用于交易时，可在费用核算结果基础上形成综合交易价格。

- a) 综合交易单价包括：
- b) 综合交易单价；
- c) 基准价格；
- d) 性能调整系数；
- e) 时间调整系数；
- f) 服务等级调整系数；
- g) 合约调整系数；
- h) 市场因素调整系数。

11.7.2 上述系数为交易定价参数，并非所有项目均应设置。

11.7.3 不能合理量化的因素不宜强行设置调整系数。

11.7.4 算力交易平台应建立符合本标准要求的综合定价管理体系。

11.7.5 应保证数据真实性、一致性和完整性。

12 计费方式与核算结果应用

12.1 一般要求

计费应基于可验证的费用核算结果、交易价格及合同约定实施。计费方式可以改变费用支付和交易表达形式，但不应造成核算数据不可追溯。

12.2 按量计费

可按照实际计算资源使用量、实例使用量、Token数量、请求数量或其他约定计费单元进行计费。

12.3 包周期计费

可采用包月、包年等周期方式。包周期模式下应明确：

- a) 预留资源规格；
- b) 资源保障等级；
- c) 合同期限；
- d) 超额资源使用规则；
- e) 提前释放或变更规则。

12.4 任务计费

可按照一次完整训练任务或批量推理任务形成任务级结算费用。任务费用应能够追溯至其基础资源使用量和相关费用构成。

12.5 Token 计费

模型推理服务可采用输入Token、输出Token或其他经约定的Token计费方式。Token计费价格可在Token资源费用核算结果基础上结合交易策略确定。

12.6 请求计费

适用于请求间资源消耗差异较小的推理服务。请求资源消耗差异明显时，宜结合Token、模型类型或实例资源使用情况实施分级或混合计费。

12.7 有效算力计费

训练或者适合进行有效算力标定的推理场景，可按有效算力使用量进行计费。

12.8 抢占式或竞价实例

抢占式或者竞价资源可采用动态交易价格。应明确：

- a) 资源回收条件；
- b) 任务中断通知方式；
- c) 已消耗资源的核算；
- d) Checkpoint 处理；
- e) 中断后的恢复和费用处理规则。

12.9 混合计费

12.9.1 混合计费可采用：

- a) 保底资源加弹性资源；
- b) 预留实例加按量实例；
- c) 实例运行费用加Token费用；
- d) 基础容量费用加请求费用；
- e) 其他组合方式。

12.9.2 混合计费应与第 10.9 规定保持一致。

12.10 辅助计费项目

存储、网络、数据处理、软件许可、技术支持、安全服务等资源可单独计费，也可包含在综合服务价格中。采用综合价格时，应在费用核算系统中明确其费用归属，防止重复计费。

13 服务等级协议与费用核算关联

13.1 一般要求

13.1.1 算力服务提供方应建立与算力费用核算和计费模式相匹配的SLA。

13.1.2 SLA评价宜作为基础费用核算结果形成后的调整机制。

13.1.3 总体逻辑宜为：基础费用核算结果→SLA评价→核算调整→最终费用核算结果。

13.2 服务等级分类

训练和推理服务可根据资源保障、性能保障、可用性和响应要求划分不同服务等级。不同等级对应的指标和阈值应在服务协议中明确。

13.3 训练SLA指标

训练场景SLA指标可包括：

- a) 算力资源可用率；
- b) 训练任务成功完成率；
- c) GPU、NPU或TPU可用率；
- d) 网络通信可用率；

- e) 分布式训练同步成功率；
- f) 存储访问成功率；
- g) Checkpoint 保存成功率；
- h) Checkpoint 恢复成功率；
- i) 训练吞吐性能；
- j) 任务完成时间；
- k) 平均故障恢复时间；
- l) 资源交付及时性。

13.4 推理 SLA 指标

推理场景SLA指标可包括：

- a) 模型服务可用率；
- b) 请求成功率；
- c) TTFT；
- d) 单请求响应时间；
- e) Token 生成速率；
- f) 单位时间吞吐量；
- g) 并发请求处理能力；
- h) P95、P99 等约定分位时延；
- i) 模型实例启动或扩容时间；
- j) 资源性能达标率；
- k) 故障恢复时间。

13.5 SLA 指标测量

SLA指标应具有确定的统计周期、数据来源和计算方法。对于TTFT、Token生成速率等推理指标，应明确测量范围、模型版本、请求特征和统计方法。

13.6 SLA 核算调整

服务补偿时通常为负值。SLA调整规则可根据以下内容确定：

- a) 不可用时间；
- b) 性能偏差程度；
- c) 受影响资源数量；
- d) 受影响请求或任务数量；
- e) 任务重算资源消耗；
- f) 合同约定的补偿规则。

13.7 训练任务失败

因服务提供方原因造成训练任务失败、有效训练成果丢失或Checkpoint不可恢复的，应根据实际影响对费用核算结果进行调整。调整范围宜与故障实际造成的无效资源消耗相对应。

13.8 推理服务性能不达标

推理服务未达到约定可用性、TTFT、响应时间、Token生成速率或者吞吐能力时，可根据受影响请求数量、时间范围或服务费用实施核算调整。

13.9 补偿方式

13.9.1 SLA 补偿可包括：

- a) 费用减免；
- b) 费用返还；
- c) 服务时长补偿；
- d) 等额资源补偿；

- e) 算力额度补偿;
- f) 合同约定的其他方式。

13.9.2 非货币形式补偿的，应在核算记录中明确其金额或权益对应关系。

13.10 数据安全和专属服务

13.10.1 高安全等级环境可包括：

- a) 专属资源环境;
- b) 安全隔离环境;
- c) 可信计算环境;
- d) 满足特定安全合规要求的环境。

13.10.2 相应安全费用应依据实际采用的安全资源、软件、运营或服务进行核算。

13.10.3 不得在缺乏相应服务内容的情况下仅以安全等级名称重复增加费用。

13.11 服务质量系数

依据网络带宽、RDMA支持、节点间通信拓扑、故障恢复时间等确定SLA等级系数，低SLA折扣、高SLA溢价。可参考公式（3）计算：

$$K_s = \frac{A+N+R}{3} \tag{3}$$

式中：
A——资源可用率；
N——网络性能；
R——恢复能力。

13.12 合约周期与购买量折扣系数

长期合约（≥1年）、大额度预付或批量购买可享受的折扣率设定原则。

13.13 多维度综合定价公式

给出最终结算单价的计算公式：
综合单价=基准单价×（性能系数）×（时间系数）×（服务系数）×（合约系数）×…
最终费用=Σ（综合单价×计量数量）+辅助计费项费用。

14 结算与支付

14.1 账单生成

结算系统应基于最终费用核算结果及适用交易价格生成账单。账单应保证完整、准确、可追溯。

14.2 账单内容

账单宜至少包括表1所列信息。

表1 账单信息

类别	内容
基本信息	用户或租户、项目 ID、任务 ID 或模型服务 ID、订单编号、核算周期
场景信息	训练或推理、任务类型、服务模式
资源信息	GPU/NPU/TPU 型号和数量、CPU、内存、显存、存储、网络、实例规格
训练信息	模型、模型规模、训练 Token、训练 step、训练阶段、训练时长、Checkpoint 等
推理信息	模型、模型版本、请求量、输入 Token、输出 Token、实例运行时间、并发、服务等级等

类别	内容
计量信息	卡时、实例时、SCU 时、资源使用量及其他适用核算单元
费用信息	直接费用、间接费用、辅助资源费用
参数信息	单位费用、适用费用参数及版本
SLA 信息	SLA 等级、实际达成情况、调整金额
定价信息	基准价格、交易调整因素、优惠或折扣
结果信息	基础核算费用、调整金额、最终核算费用、优惠金额、应付金额

14.3 结算周期

结算周期可采用小时、日、月、任务周期或双方约定的其他周期。推理服务采用月度等长周期结算时，应能够向下追溯至请求、模型实例或资源使用记录。

14.4 支付方式

支付方式可包括：

- a) 预付费；
- b) 后付费；
- c) 包周期；
- d) 信用支付；
- e) 算力额度或算力券；
- f) 双方约定的其他方式。

14.5 欠费处理

14.5.1 欠费处理可包括：

- a) 停止新增任务；
- b) 限制资源扩容；
- c) 暂停模型服务；
- d) 冻结资源；
- e) 释放计算资源。

14.5.2 涉及用户数据、模型文件和 Checkpoint 的，不得仅因计算资源欠费立即删除，其保存和处置应按照合同及相关数据管理要求执行。

14.6 争议处理

14.6.1 计量或核算出现争议时，应能够提供：

- a) 原始计量记录；
- b) 任务或请求日志；
- c) 核算规则版本；
- d) 价格和费用参数版本；
- e) 费用计算过程；
- f) SLA 记录；
- g) 异常调整记录。

14.6.2 经复核发现核算错误的，应按照确认后的数据实施退补。

15 试验方法

15.1 总体要求

试验应验证费用核算模型的准确性、一致性、可重复性和可追溯性。试验应分别覆盖训练和推理典型场景。

15.2 计量准确性试验

应构建已知资源使用量的测试任务，对系统采集结果与基准记录进行比较。可采用相对误差，允许误差阈值应由平台技术能力、核算粒度或合同要求确定。

15.3 核算结果验证

应基于预设资源使用量、单位费用和分摊参数独立计算理论费用，并与系统核算结果进行比较。验证内容至少包括：

- a) 直接费用；
- b) 间接费用分摊；
- c) 总费用；
- d) SLA 调整；
- e) 异常费用处理。

15.4 训练核算试验

15.4.1 训练场景测试宜覆盖：

- a) 单机单卡训练；
- b) 单机多卡训练；
- c) 多节点分布式训练；
- d) 不同并行方式；
- e) Checkpoint 保存和恢复；
- f) 任务失败及重试；
- g) 异构算力；
- h) 资源低利用率；
- i) 不同训练精度。

15.4.2 试验应验证卡时、有效算力、Token 吞吐、通信效率等指标与费用核算结果之间的一致性。

15.5 推理核算试验

15.5.1 推理测试宜覆盖：

- a) 在线实时推理；
- b) 离线批量推理；
- c) 流式推理；
- d) 多轮交互；
- e) 高并发；
- f) 长上下文；
- g) 动态批处理；
- h) 多模型共享；
- i) 多实例部署；
- j) 自动扩缩容。

15.5.2 应验证请求数量、输入 Token、输出 Token、实例运行时间、设备使用量、显存占用以及费用核算结果之间的一致性。

15.6 异构算力折算试验

异构算力性能系数应通过统一软件环境、统一模型和统一测试方法进行验证。训练和推理应分别建立基准工作负载。

15.7 SLA 试验

15.7.1 训练场景宜验证：

- a) 资源中断;
- b) 网络异常;
- c) 性能下降;
- d) Checkpoint 恢复;
- e) 任务恢复。

15.7.2 推理场景宜验证:

- a) 服务中断;
- b) TTFT 超限;
- c) Token 生成速率下降;
- d) 请求失败;
- e) 高并发性能下降;
- f) 实例扩容不及时。

15.7.3 应验证 SLA 监测结果和费用调整结果的一致性。

15.8 数据异常试验

应测试数据缺失、重复采样、时间漂移、计量服务中断等异常情况。系统不应因异常数据自动形成明显失真的费用核算结果。

附录 A
(资料性)
主流 GPU/NPU 性能系数参考表

A.1 总则

性能系数主要用于反映不同硬件平台在典型大模型训练任务中的综合训练能力，不作为硬件理论峰值算力评价指标。性能系数宜综合考虑以下因素：

- a) 理论计算能力（FP16、BF16、FP8 等）；
- b) 显存容量及显存带宽；
- c) 高速互联带宽（NVLink、IB、HCCS 等）；
- d) 典型训练模型吞吐率；
- e) GPU/NPU 利用率；
- f) 软件生态成熟度；
- g) 训练稳定性。

A.2 标准算力性能系数计算原则

性能系数宜采用标准算力单位（SCU）进行统一折算。推荐以某一典型训练硬件作为基准设备，其他设备按照标准化测试结果折算得到性能系数。

性能系数可按公式（4）计算：

$$K_p = \frac{P_i}{P_b} \tag{4}$$

式中：

P_i ——被测设备典型训练性能；

P_b ——基准设备典型训练性能。

典型训练性能宜采用单位时间内有效训练吞吐率（如Token/s、Samples/s或Effective FLOPS）进行评价。

A.3 推荐测试模型

建议采用以下典型模型进行性能标定：

- a) Transformer：Llama、Qwen、GPT；
- b) MoE：DeepSeek、Qwen3.5；
- c) 视觉模型：ResNet、ViT 等；
- d) 多模态模型：Qwen2.5-VL、InternVL。

最终性能系数宜采用多个典型模型测试结果加权平均获得。

A.4 主流GPU/NPU性能系数参考表

表A.1 主流GPU/NPU性能系数参考表

硬件类型	训练精度	参考性能系数
英伟达A100 80GB	BF16	1.00
英伟达H100 SXM	BF16	1.80~2.20
英伟达H200	BF16	2.20~2.60
英伟达A6000	BF16	0.70~0.80
华为910B	BF16	0.75~0.85
沐曦C500	BF16	0.65~0.75

附录 B
(资料性)
大模型训练计算精度类型折算系数表

- B.1 总则**
不同计算精度在训练速度、显存占用、计算效率及训练收敛能力方面存在差异。综合定价时，可根据计算精度采用相应折算系数，对基础价格进行修正。
- B.2 折算原则**
计算精度折算系数宜综合考虑：
a) 理论计算性能；
b) 实际训练吞吐率；
c) 显存占用；
d) 能耗；
e) 训练收敛质量；
f) 软件支持程度。
对于相同训练任务，应优先采用实际训练性能作为折算依据。
- B.3 精度分类**
训练精度包括但不限于：FP64、FP32、FP16、BF16、FP8、INT8（训练辅助）、INT4（训练辅助）。
- B.4 折算系数参考**

表B.1 折算系数参考表

计算精度	适用场景	参考折算系数
FP64	科学计算	0.40-0.60
FP32	传统深度学习	0.70-0.90
FP16	混合精度训练	1.00
BF16	大模型训练基准	1.00
INT8、INT4	量化训练	依据模型确定