Software Engineering and Applications 软件工程与应用, 2026, 15(2), 137-142
Published Online April 2026 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2026.152014

Hans汉斯

智算算力交易分层定价模型研究

梁秉豪，张传刚*，袁明明

浪潮通信信息系统有限公司，山东 济南

收稿日期：2026年1月7日；录用日期：2026年4月7日；发布日期：2026年4月16日

摘　要

随着生成式人工智能、大模型训练、智能视频分析等应用的快速发展，全球智能算力需求呈现指数级增长。传统的算力定价模式主要基于硬件配置或使用时长，难以适应多样化、高并发的智算任务需求，存在任务适配性差与服务等级协议覆盖不足等问题。本文提出一种面向智算算力交易的分层定价模型，将定价模型划分为基础服务与增值服务两部分。基础层基于任务类型(如图像检测、自然语言处理、语音识别等)的核心先验参数构建定价模型；增值层则根据用户对时效、并发、多节点协同等服务等级需求进行浮动。结合真实智算任务日志开展实证研究，通过对比传统计费与分层计费的成本、任务完成率等指标，验证了模型的优越性。该模型在提升资源利用率、满足差异化服务需求方面具有显著优势，为算力资源的市场化配置提供了可行的技术路径。

关键词

算力交易，定价模型，基础服务，增值服务，分层定价

Research on the Tiered Pricing Model for Intelligent Computing Power Trading

Binghao Liang, Chuangang Zhang*, Mingming Yuan

Inspur Communication Information System Co., Ltd., Jinan Shandong

Received: January 7, 2026; accepted: April 7, 2026; published: April 16, 2026

Abstract

With the rapid development of applications such as generative AI, large model training, and intelligent video analytics, the global demand for intelligent computing power is growing exponentially. Traditional computing power pricing models, primarily based on hardware configuration or usage duration, struggle to accommodate the requirements of diverse, highly concurrent intelligent computing tasks.

*通讯作者。

文章引用: 梁秉豪, 张传刚, 袁明明. 智算算力交易分层定价模型研究[J]. 软件工程与应用, 2026, 15(2): 137-142.
DOI: 10.12677/sea.2026.152014

These models face challenges including poor task adaptability and insufficient Service Level Agreement (SLA) coverage. This paper proposes a tiered pricing model for intelligent computing power transactions, dividing the pricing structure into basic services and value-added services. The basic service tier establishes pricing models based on core prior parameters of task types (such as image detection, natural language processing, and speech recognition). The value-added tier incorporates dynamic adjustments according to users' SLA requirements including timeliness, concurrency, and multi-node collaboration. Empirical research is conducted using real intelligent computing task logs, and the superiority of the model is verified by comparing costs, task completion rates, and other indicators between traditional billing and layered billing. This model demonstrates significant advantages in enhancing resource utilization and meeting differentiated service demands, providing a viable technical pathway for the market-based allocation of computing power resources.

Keywords

Computing Power Trading, Pricing Model, Basic Services, Value-Added Services, Tiered Pricing

1. 引言

随着生成式人工智能等技术在近年来取得了突破性的进展，全球数智化迎来了前所未有的变革。从人机对话、图像生成到代码编写，在颠覆传统生产模式的同时对千亿甚至万亿参数模型有了更多的需求，智能算力需求也因此呈现指数级增长。根据中国信通院等权威机构研究数据显示，我国智算算力规模预计在 2026 年可以突破 ZFLOPS 量级，远超通算算力增长速度。为应对算力需求的发展趋势，构建高效、灵活的算力供给与交易市场，以实现算力资源的市场化最优配置，已成为缓解供需矛盾、促进产业健康发展的关键路径。目前，智算算力交易主要采用基于硬件配置和使用时长为核心的传统定价模式[1]，在高度复杂和多样化的智算任务面前，日益暴露出其固有的局限性，主要体现在两大核心问题上：

首先，任务适配性差。不同类型的智算任务对资源的消耗存在数量级差异。传统基于“卡时”或“机时”的定价模式无法精准量化这种因任务类型、参数规模、数据量等内在属性造成的成本差异，导致需求方难以评估真实使用成本，供给方也无法实现按量计费的精细化管理，造成资源错配与效率损失。

其次，SLA 覆盖不足。智算任务对服务质量，尤其是时效性、并发能力和可靠性等方面的要求差异显著。一个离线模型训练任务可能对完成时间不敏感，而一个在线实时推理服务则要求毫秒级响应和高并发支持。传统定价模式通常仅提供单一、标准化的服务等级协议，缺乏针对不同 SLA 需求的弹性定价机制，难以满足市场对高性能、高保障服务的差异化需求。

为解决上述痛点，推动智算交易市场向更高效、更公平、更精细化的方向发展，本文提出一种面向智算算力交易的分层定价模型。该模型旨在将定价结构从“一刀切”的粗放模式，演进为需求驱动的精细模式。通过引入真实智算任务日志开展实证分析，完善定价公式的理论支撑与参数设计，补充交易流程与差异处理机制，并结合主流云厂商定价策略明确差异化优势，提升模型的科学性与可落地性。

2. 相关工作

2.1. 定价模型

随着云计算的普及，算力定价模型经历了多个阶段的发展，传统云计算市场按资源规格和按使用时

长的定价模型简单透明，但未能充分考虑用户使用习惯的差异。随着模型即服务(MaaS)等新兴产品形态的不断成熟，业界开始探索更加精细化的定价模型。

第一代模型主要基于服务器或加速卡等硬件实体，其定价直接与 GPU 型号、内存等硬件配置挂钩[2]。例如 AWS 早期的 EC2 实例定价，仅根据 CPU 核心数、内存容量等硬件参数划分档位。尽管直观，但此模型无法反映同一硬件上不同 AI 任务的资源消耗差异，尤其在多租户混部场景下会导致成本分摊不公。

第二代模型转向基于计算量的定价，其中以 FLOPS 作为计价单位最为典型。Google Cloud 的 TPU v4 定价采用每 TFLOPS/秒计费模式，但计算量难以准确计量端到端任务的实际执行成本，且未考虑数据传输、存储开销等隐性成本。

近年来，基于任务特征的定价模型成为研究热点。针对计算机视觉任务，已有研究通过分析图像复杂度与不确定性方差，建立了分辨率与任务性能的关联模型，为基于分辨率的定价提供了核心依据——简单图像任务适配低分辨率，复杂任务则需要高分辨率支持，这种适配关系直接影响计算资源消耗量级[3]。而 Serverless AI Inference 定价作为近年重点研究方向，其核心优势在于通过自动扩缩容实现资源按需分配，平衡灵活性与成本可控性。Amazon SageMaker 的无服务器推理方案可根据流量自动扩缩资源，无请求时缩减至零资源占用，极大降低闲置成本，其定价同时兼顾计算时长、数据量与内存配置，为多场景适配提供了实践参考[4]。总体上看，定价模型的精细化演进与底层算力调度技术的成熟度高度相关。

2.2. 定价策略

智能算力定价是一个复杂的动态过程，简单的固定价格模型难以涵盖实际需求。运营者为了在满足用户算力需求的同时，实现收益最大化与资源利用率最优化，广泛采用了博弈论、拍卖算法和机器学习等技术对定价策略进行优化[5]。

基于博弈论的方法[6]，将算力提供商与用户视为理性且智能的决策参与者，他们之间的策略互动构成了一个非合作博弈。定价过程即是为这个博弈寻找一个稳定的均衡点，在此点上，任何一方都无法通过单方面改变自身策略而获得更多利益。Li 等人[7]构建了云边协同算力交易的 Stackelberg 博弈模型，通过逆向归纳法求解均衡价格，实现了提供商收益与用户效用的帕累托最优。

基于拍卖算法的方法[8]，将算力资源作为拍品，用户通过出价来竞争资源的使用权。价格由用户的出价和平台的拍卖规则共同决定，能够直接反映资源的实时稀缺程度和市场供需关系。

基于机器学习的方法[9]，利用历史交易数据、资源使用情况、用户行为和市场供需等信息，通过机器学习模型来预测最优价格或动态调整定价策略，是目前使用最为广泛的方法。

3. 分层定价模型

3.1. 模型整体框架

本文提出的分层定价模型主要包括两个层次：基础服务定价层：主要基于智算任务的技术参数(如图像分辨率、Token 数量、音频时长等)构建定价模型，体现面向用户通用需求的定价策略。增值服务定价层：基于用户对服务等级协议的需求，提供可选的增值服务，体现面向用户个性化需求的定价策略。该框架通过将任务资源消耗与服务质量要求分离定价，既保证了基础价格模型的可解释性，又为个性化需求提供了灵活的服务选择。

3.2. 基础服务定价模型

基础服务部分主要聚焦智算任务的固有参数，屏蔽底层技术差异，反映不同类型智算任务在理想条

件下的基本资源消耗，针对训练与推理两大类场景，分别构建了参数化的定价模型。

3.2.1. 训练场景定价模型

训练任务通常具有使用周期长、资源消耗大和业务数据复杂等特点。本文针对图像、文本和语音三类典型的智算训练任务进行了建模，如表 1 所示。目标检测任务定价主要取决于数据规模、复杂度和模型大小。核心参数包括：反应数据规模的图片数量和分辨率。分辨率越高，单张图片处理所需的计算量和内存占用显著增加。目标框数量反映了任务的复杂度和标注密度。目标框越多，模型在回归和分类子任务上的计算负担越重。模型参数量是决定计算和内存需求的关键因素。参数越多的模型，单次前向/反向传播的计算量越大，对显存的要求也越高。算力资源消耗随参数量增长通常呈现亚线性趋势，因此可以采用对数函数进行拟合。自然语言处理任务定价模型主要考虑序列建模复杂度。Token 总量代表了训练数据的总体规模，而序列长度对计算复杂度影响较大。音频时长和采样率共同定义了原始音频数据的体量和信息密度。采样率越高，音频信号的时间分辨率越高，处理所需的计算资源越多。

Table 1. Pricing model for model training scenario
表 1. 训练场景定价模型

任务类型	核心参数	计量公式
目标检测	图片数量(N/张)、分辨率(M/像素)、目标框数量(K/个)、训练轮数(E/轮)、模型参数量(P/亿)	$C_{base}=(\alpha\times N\times M+\beta\times K\times E)\times\log(P)$
自然语言处理	Token 总量(T/万个)、序列长度(L/个)、训练轮数(E/轮)、模型参数量(P/亿)	$C_{base}=(\gamma\times T\times L\times E)\times\log(P)$
语音识别	音频时长(H/小时)、采样率(S/kHz)、训练轮数(E/轮)、模型参数量(P/亿)	$C_{base}=(\delta\times H\times S\times E)\times\log(P)$

其中 $\alpha,\beta,\gamma,\delta$ 为单位资源的价格系数，可以通过历史任务数据进行回归，或者通过市场竞价机制进行确定，此外 $\log(P)$ 反映不同模型在相同任务下的资源消耗量差异。其中，模型参数量 P 与计算资源消耗的关系，主要基于对大规模深度学习模型训练经验性规律得到，为定价模型提供一个平滑、可解释且能反映边际成本递减效应的函数形式。从理论计算复杂度看，深度学习模型的训练与推理主要取决于浮点运算次数。对于卷积神经网络，其核心操作的浮点运算次数与参数量 P 近似呈线性关系。然而，在实际硬件执行与系统调度层面，最终的资源消耗与理论浮点运算次数增长并非完全一致。主要受限于内存带宽、设备间通信开销、并行化效率以及任务调度策略等因素，导致资源消耗随参数量增长呈现亚线性趋势[10]。这种亚线性关系在业界对大规模模型训练的观察中得到验证。研究表明，在给定计算预算下，模型性能的提升与参数量、数据量和计算量之间存在幂律关系[11]。同时，为达到特定性能，所需的计算量增长远快于参数量增长[12]。单纯增加参数量带来的边际性能收益(或边际资源需求)是递减的。采用对数函数可以有效在定价模型中刻画这种边际递减效应，使得定价在面对参数量指数级增长时保持相对稳定与合理，避免超大模型的定价过高，从而鼓励更高效的模型使用与资源配置。

3.2.2. 推理场景基础定价

推理场景分为实时推理与批量推理，其定价模型如表 2 所示。实时推理核心在于请求频次和单次请求的数据量。请求次数对应服务的调用量，单请求数据量对于图像分类可以是图片的大小；对于文本生成可以是输入 Token 的数量。数据量越大，单次请求消耗的计算资源越多。批量推理适用于对时效性要求不高的离线处理任务，任务数据量和单任务数据量共同定义了批量任务的总规模。

Table 2. Pricing model for model ınference scenario
表 2. 推理场景定价模型

任务类型	核心参数	计量公式
实时推理	Q 为请求次数，D 为单请求数据量	$C_{base}=(\varphi\times Q\times D)\times\log(P)$
批量推理	N 为任务数据量，D 为单任务数据量	$C_{base}=(\omega\times N\times D)\times\log(P)$

其中 φ,ω 为单位资源价格系数，同样可以通过历史数据进行回归或市场竞价机制确定；

3.3. 增值服务定价模型

基础服务定价模型定义了标准服务质量下的定价方法。然而，在实际业务中，用户对算力服务的需求是多样化的。许多应用场景对任务的完成时效、系统的并发处理能力或计算的协同规模有着超出基础规格的特殊要求。增值服务定价模型正是为了满足这类差异化需求而设计，根据用户所选择的更高等级的服务水平协议进行动态调整，具体模型如表 3 所示。

Table 3. Pricing model for value-added services
表 3. 增值服务定价模型

服务类型	触发条件	计量公式
时效约束	用户要求完成时间 T_{req} < 标准预估完成时间 T_{est}	$C_{total}=C_{base}\times\left(\frac{T_{req}}{T_{est}}\right)^{\mu}$
并发量	用户要求 QPS_{req} > 标准服务 QPS_{std}	$C_{total}=C_{base}\times\left(\frac{QPS_{req}}{QPS_{std}}\right)^{\sigma}$
多节点协同	计算节点数量 M	$C_{total}=C_{base}\times\log(M)$

其中时效约束代表用户要求的任务完成时间短于系统基于当前资源状况预估的标准完成时间。为了满足时效需求需要为其分配更多计算资源、启用更高优先级的调度策略或预留专用资源通道，显著提升运营成本。高并发要求意味着系统需要在单位时间内处理更多的请求，对单节点处理能力、负载均衡以及内存带宽和网络 I/O 等提出更高要求。通常需要通过水平扩展或分配更高性能的硬件来满足。当任务无法在单台服务器上完成时，可以采用分布式计算，这样引入了额外的成本，包括节点间的通信开销、同步等待时间、更复杂的任务调度与管理成本。增值服务定价模型将用户个性化需求转化为量化模型，实现了服务差异化的精准建模。

4. 总结与展望

本文针对当前智算算力交易定价模型存在的问题，提出了一种面向智算任务的分层定价模型。该模型将定价结构划分为基础服务与增值服务两个层次：基础服务层依据任务类型的核心参数构建定价模型，覆盖图像、文本、语音等多种典型智算任务；增值服务层则根据用户对时效、并发、多节点协同等服务等级需求进行动态价格调整。该模型不仅提升了算力资源的适配性与利用率，也为用户提供了更加灵活、透明的成本控制方式。后续需要进一步探索模型参数的自适应优化机制，结合实时市场数据与用户反馈，实现定价策略的动态校准。

基金项目

泰山产业领军人才项目(tscx202312006)；山东省博士后创新项目(SDCX-ZG-202400307)。

参考文献

[1] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., *et al*. (2010) A View of Cloud Computing. *Communications of the ACM*, **53**, 50-58. https://doi.org/10.1145/1721654.1721672

[2] Li, Z., OBrien, L., Cai, R. and Zhang, H. (2012) Towards a Taxonomy of Performance Evaluation of Commercial Cloud Services. 2012 *IEEE Fifth International Conference on Cloud Computing*, Honolulu, 24-29 June 2012, 344-351. https://doi.org/10.1109/cloud.2012.74

[3] Luo, W.Q., Tian, Z., Li, Y.F., *et al*. (2025) Task-Aware Resolution Optimization for Visual Large Language Models.

[4] Amazon Web Services (2025) Using Amazon SageMaker Serverless Inference. https://docs.aws.amazon.com/zh_cn/sagemaker/latest/dg/serverless-endpoints.html

[5] 黄潇洁. 面向算力网络的算网资源定价方法研究[D]: [硕士学位论文]. 北京: 北京邮电大学, 2024.

[6] Hu, J., Li, K., Liu, C. and Li, K. (2021) A Game-Based Price Bidding Algorithm for Multi-Attribute Cloud Resource Provision. *IEEE Transactions on Services Computing*, **14**, 1111-1122. https://doi.org/10.1109/tsc.2018.2860022

[7] Chen, Y., Li, Z., Yang, B., Nai, K. and Li, K. (2020) A Stackelberg Game Approach to Multiple Resources Allocation and Pricing in Mobile Edge Computing. *Future Generation Computer Systems*, **108**, 273-287. https://doi.org/10.1016/j.future.2020.02.045

[8] Lloret-Batlle, R. and Jayakrishnan, R. (2017) Envy-Free Pricing for Collaborative Consumption of Supply in Transportation Systems. *Transportation Research Procedia*, **23**, 772-789. https://doi.org/10.1016/j.trpro.2017.05.043

[9] Lu, R., Hong, S.H. and Zhang, X. (2018) A Dynamic Pricing Demand Response Algorithm for Smart Grid: Reinforcement Learning Approach. *Applied Energy*, **220**, 220-230. https://doi.org/10.1016/j.apenergy.2018.03.072

[10] Narayanan, D., Shoeybi, M., Casper, J., *et al*. (2021) Efficient Large-Scale Language Model Training on GPU Clusters. https://doi.org/10.1145/3458817.3476209

[11] Kaplan, J., Mccandlish, S., Henighan, T., *et al*. (2020) Scaling Laws for Neural Language Models.

[12] Hoffmann, J., Borgeaud, S., Mensch, A., *et al*. (2022) Training Compute-Optimal Large Language Models.