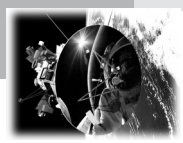

专栏

智算中心推理业务能力评估方法研究

武振宇[1]，赵占军[1]，卜忠贵[1]，刘鹏[1]，田雯[1]，王祎玮[1]，董江帆[1]，王猛[2]，张誌[3]
（1. 中国移动通信集团设计院有限公司，北京 100080；
2. 中国移动通信集团设计院有限公司安徽分公司，安徽 合肥 230031；
3. 中国移动通信集团设计院有限公司山东分公司，山东 济南 250101）

摘　要：人工智能推理中心的建设已成为当前智算中心建设的热点，按照智能算力的规模评估智算中心推理业务能力不再准确。通过建立时延不敏感业务模型、时延敏感业务模型、用户访问业务模型，提出对智算中心推理业务能力的量化评估方法，从而在建设环节做到建需匹配，提高投资效益。

关键词：智算中心；推理；业务能力评估
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025167

Research on evaluation methods for inference business capability of intelligent computing center

WU Zhenyu[1], ZHAO Zhanjun[1], BU Zhonggui[1], LIU Peng[1], TIAN Wen[1],
WANG Yiwei[1], DONG Jiangfan[1], WANG Meng[2], ZHANG Zhi[3]
1. China Mobile Group Design Institute Co., Ltd., Beijing 100080, China
2. Anhui Branch of China Mobile Group Design Institute Co., Ltd., Hefei 230031, China
3. Shandong Branch of China Mobile Group Design Institute Co., Ltd., Jinan 250101, China

Abstract: The construction of artificial intelligence inference centers has become a hotspot in the current development of intelligent computing centers. Evaluating the inference business capability of intelligent computing centers solely based on the scale of intelligent computing power is no longer accurate. A quantitative evaluation method for the inference business capability of intelligent computing centers was proposed by establishing three models: a delay-insensitive business model, a delay-sensitive business model, and a user access business model. This approach aims to achieve alignment between construction and requirements during the construction phase, thereby improving investment efficiency.

Key words: intelligent computing center, inference, business capability evaluation

收稿日期：2025−03−21；修回日期：2025−07−09
通信作者：武振宇，wuzhenyu@cmdi.chinamobile.com

0　引言

当前，新一轮科技革命和产业变革加速演进，数据成为新生产要素、算力成为新基础能源、人工智能（artificial intelligence，AI）成为新生产工具，三者共同构成新质生产力的重要驱动因素，并引发生产方式、生活方式和社会治理方式的深刻变革[1]。

以ChatGPT为代表的人工智能生成内容（artificial intelligence generated content，AIGC）技术的兴起，标志着AI的发展已迈入一个全新的、更具创新性的阶段。通用大模型在此阶段集中爆发，呈现跃迁式发展，推动了AI技术从“判断式”向“生成式”、从专用技术向通用性技术转变，这一转变已经成为推动全球科技创新的重要驱动力。AI的跃迁式发展将进一步带动大规模产业升级，成为各行各业提高效率、优化体验、降低成本的重要手段。

目前国产大模型已覆盖多个行业和领域，大模型应用场景不断拓展。国内AI技术加速突破应用，从赋能千行百业提质增效的辅助工具，跃升为使经济社会全面转型发展的基础设施和核心能力，推动生活、生产、治理等各领域发生深层次变革、系统性重塑，“AI+”时代正加速到来[2]。截至2024年12月31日，共302款生成式AI服务在中华人民共和国国家互联网信息办公室（简称“国家网信办”）完成备案[3]。国内主要的通用大模型，如深度求索（DeepSeek）、百度的文心一言、阿里的通义千问、字节跳动的豆包、华为的盘古、智谱AI的智普、科大讯飞的星火大模型、中国移动的九天等，以语言大模型为主，部分大模型还具备一定的多模态能力。模型的性能主要依赖于模型的规模，具体包括参数数量、数据集的大小以及计算量[4]。当前，大模型正从千亿参数向万亿、十万亿级别参数量发展，对算力的需求呈井喷式增长[5]。

随着GPT-o1及DeepSeek-R1的问世，借助于思维链等技术创新，大模型开始具备逻辑推理能力，输出准确率得到大幅提升[6]，正在重构人类社会的生产方式。DeepSeek-R1也标志着中国在大语言模型推理技术上取得重大突破，AI大模型领域迎来了突飞猛进的发展。未来，AI推理也将会在多模态基础上进一步向前发展，对智算资源的需求也将持续增长[7]。

1　智算中心的发展

中国信息通信研究院将算力划分为通用算力、智能算力、超算算力[8]。智能算力以智算中心为承载体，由智算中心内的大量AI加速卡提供。智算中心是基于先进的AI技术和计算架构，为AI应用提供所需的算力服务、数据服务和算法服务的公共算力新型基础设施。大模型时代，云计算与智算融合发展成为主流。面对大模型训练、推理和微调等多样化的应用需求，智算云技术体系继承云计算的三层体系架构[9]，应用层以生成式AI应用为核心，平台层以模型服务为核心，底层基础设施向大规模、多元异构、高效低时延转变。

作为AI发展的底座，智能算力已成为推动社会发展的核心动力之一，各行各业对智能算力的需求正以前所未有的速度增长，需求主要集中在模型训练、模型推理以及海量数据处理等方面。

近年来，国内外互联网巨头、头部云服务商、运营商等都在持续加大对智算资源的建设投入。2024年，美国政府设立智算中心基础设施工作组[8]，美国科技巨头则相继建成或规划了超大规模智算集群，如埃隆·马斯克旗下人工智能公司xAI的20万卡智算中心、Facebook母公司Meta的35万卡智算集群[10]、微软联合OpenAI提出的“星际之门”百万卡集群计划[11]。2025年1月，美国政府宣布投入5 000亿美元，将“星际之门”提升为美国AI基础设施建设的重要项目。我国则

从2021年开始出台政策推动建设进程[12]。政府、互联网公司、电信运营商纷纷开展智算中心建设，已经建立了多个万卡以上规模集群。2024年1月，工业和信息化部等七部门联合发文提出“建设超大规模智算中心，满足大模型迭代训练和应用推理需求”[13]。

智算中心的前期建设主要集中在训练资源的构建上，训练智算中心的建设规模是以计算能力来衡量的，如卡数、总体运算能力等。这些训练资源对于开发和优化AI模型至关重要，它们能够提供必要的数据和算力来训练复杂的深度学习网络。然而，随着AI大模型训练的逐步商用落地，尤其是DeepSeek-V3与DeepSeek-R1的横空出世，算法突破与架构创新之间形成了正向循环，推动了模型向“共享开放”的新形态迈进。一方面，这些模型展现的极具影响力的工程技术创新，大幅提升了训练效率，在一定程度上使得国内训练中心的建设节奏暂时放缓；另一方面，其开源特性以及跻身世界前列的出色性能，大大加速了AI应用落地的进程，各行各业对于快速、准确的推理服务的需求快速增长，使得推理应用即将迎来爆发期。这要求智算中心不仅要能够训练模型，还要能够快速地提供推理服务。因此，智算中心的建设重点正在逐渐从训练资源向推理资源建设转变。

2 智算中心推理业务能力评估

在智算中心的建设过程中，建设规模的确定，依据的是对智算中心业务能力的准确评估。因此，精准匹配集群规模与业务需求成为至关重要的考量点，一方面，要确保大模型开发应用在不同阶段的算力需求都能得到恰到好处的算力支持，另一方面，要能够让企业更合理地分配资源，包括人力、资金和技术等，以提高投资效益和运营效率。

业务能力评估结果的对比，可以帮助企业了解自建的智算中心在市场中的竞争水平，即与竞争对手相比的优势和劣势，进而制定有效的竞争策略，提升服务能力，更好地应对市场竞争。

当前对智算中心业务能力的评估仍然主要依据其计算能力，反映智能训练能力的指标较为单一，即算力规模。由于智能训练过程处于“关门”状态，不与外界客户直接发生交互，因此，通过算力规模即可评估智算中心的智能训练业务能力。

与智能训练不同的是，智能推理是面向实际业务场景的，智能推理的结果直接推送至终端客户，因此，智能推理中心的业务能力不能简单等同于计算能力。有别于AI训练中心，智能推理首先涉及能够支撑的用户访问量，还有一个重要的服务等级协议（service level agreement，SLA）指标是用户的业务响应时延[14]，如自动驾驶、工业控制、远程医疗、实时推荐、智能客服等应用场景均对业务响应时延有不同程度的要求，如果推理时延太高，可能会造成事故风险、用户流失等后果。因此，在智算中心推理业务逐步大规模开展之际，建立一套针对智算中心推理业务能力评估的方法至关重要。

3 推理业务能力评估模型

3.1 建立时延不敏感推理业务模型

在时延不敏感推理业务中，单张推理卡所能承载的推理会话数量与其可用内存大小密切相关。在智能推理的过程中，模型需要加载至AI加速卡的内存中，智能推理的中间结果也会在内存中缓存。

AI加速卡的总内存VR，减去操作系统占用部分OS，再减去模型本身占用部分LM（与模型参数量相关），剩下的内存空间用于在推理过程中运行，每个会话在推理过程中占用的内存大小为Cache（缓存推理过程的中间结果，与模型结构相关），那么单张推理卡可支撑的最大推理会

话数量$Session_0$则可以由式（1）计算得到。

$$Session_0 = \frac{VR - OS - LM}{Cache} \quad (1)$$

其中，$Session_0$表示单张AI加速卡支持的最大会话路数，单位为路；VR表示AI加速卡的内存大小，单位为GB；OS表示AI加速卡内存中系统的占用空间，单位为GB；LM表示模型的占用内存大小，与模型参数量相关，单位为GB；Cache表示每个会话在推理过程中占用的内存大小，与业务场景、所选模型密切相关。

3.2　建立时延敏感推理业务模型

在时延敏感推理业务中，单张推理卡所能承载的推理会话数量，还与内存带宽密切相关。在智能推理的过程中，图形处理器（graphics processing unit，GPU）会利用其显存不断读写推理过程的中间结果。内存读写的速度、计算单元的性能以及数据传输的效率等因素共同决定了智能推理的时延。

根据推理时延T的要求和访问内存的带宽B、带宽利用率r，可以计算出AI加速卡可以读取的输入数据量，该数据量减去模型本身的大小LM（与模型参数量相关），则为在时间T内可读取的推理过程中间结果的数据量，每个会话在推理过程中占用的内存大小为Cache，那么单张推理卡在在给定时延SLA要求内可支撑的最大推理会话数量$Session_1$则可以由式（2）计算得到。

$$Session_1 = \frac{T \times B \times r - LM}{Cache} \quad (2)$$

其中，$Session_1$表示在给定时延SLA要求下，单张AI加速卡可支持的会话路数，单位为路；T表示时延SLA要求，单位为s；B表示GPU卡访问内存带宽，单位为GB/s；r为推理卡产品特征值[15]，表示GPU卡访问内存带宽利用率。

与单张AI加速卡支持的最大会话路数$Session_0$比较，如果$Session_1<1$，表示该时延SLA要求下，该卡不适合该模型推理，需要升级AI加速卡；如果$Session_1<Session_0$，表示该时延SLA要求较为敏感，该AI加速卡支持的用户访问路数未达最大值，可支持有限的用户访问路数$Session_1$；如果$Session_1>Session_0$，表示该时延SLA要求不敏感，该AI加速卡适合该模型推理，但是限于显存大小，只能支持该卡的最大用户访问路数$Session_0$。

在具体时延SLA要求下，单张推理卡支持的用户访问路数取二者的最小值，即$\min(Session_0, Session_1)$。

3.3　建立用户访问模型，评估业务能力

一个AI推理集群由多张AI加速卡组成，承担着大量用户的访问请求，根据用户的在线时长、峰值集中、资源占用等特征的不同，推理集群可支撑的在线用户数也不同。用户访问模型的建模步骤如下。

步骤1　峰值小时内的在线用户数。

单日在线用户数UC乘以每用户平均每日在线时长LT，得出每日用户在线总时长，再除以系统服务的时间DT，则得出平均每小时的在线用户数，再乘以峰值集中系数PC[16]则可得到峰值小时内的在线用户数，表示为：

$$\frac{UC \times LT}{DT} \times PC \quad (3)$$

其中，UC表示单日在线用户数，单位为万；LT表示每用户平均每日在线时长，单位为h；DT表示智算中心日服务时长，单位为h；PC为峰值集中系数，表示在业务峰值的那个小时内，在线用户数是平均在线用户数的倍数。

步骤2　业务峰值小时内支撑在线用户数UC实际需要支撑的会话总数量。

资源占用率UR表示在业务峰值的那个小时内，用户虽然在线但并未发出推理请求，此时不占用推理资源（如用户在得到一个推理回复后并没有立刻继续发问，而在阅读思考）。因此，式（3）乘以资源占用率UR，可得到业务峰值小时内支

撑在线用户数UC实际需要支撑的会话总数量，表示为：

$$\frac{UC \times LT}{DT} \times PC \times UR \tag{4}$$

其中，UR表示处理用户请求的推理资源占用率，用户在线但不一定持续请求推理，而是可能在业务等待或者阅读思考。根据现网业务系统数据，从面向用户（toC）业务到面向企业（toB）业务，根据业务场景不同，资源占用率约0.3%～30%。

步骤3 业务峰值小时内支撑用户数UC需要的卡数。

式（4）计算得到的总用户数量需求（即用户会话总数），除以单张卡能够支撑的会话数Session，即可计算出业务峰值小时内支撑用户数UC需要的卡数，表示为：

$$AI = \frac{\frac{UC \times LT}{DT} \times PC \times UR}{Session} \tag{5}$$

其中，AI表示智算推理中心的资源规模，单位为卡。在时延不敏感的场景下，Session使用$Session_0$，在时延敏感的场景下，使用$Session_0$与$Session_1$的最小值min（$Session_0$,$Session_1$），通过式（5）进行反向推导，即可得到在已知卡数AI条件下，单日在线用户数UC的数值。

因此，在时延不敏感的业务场景下，一定规模的智算推理中心可支持的单日在线用户量为：

$$UC = \frac{\frac{AI \times Session_0}{PC \times UR}}{LT} \times DT \tag{6}$$

在时延敏感的业务场景下，一定规模的智算推理中心可支持的单日在线用户量为：

$$UC = \frac{\frac{AI \times \min(Session_0, Session_1)}{PC \times UR}}{LT} \times DT \tag{7}$$

本文通过以上3步建模，得到推理智算中心可支撑的单日在线用户量，实现对推理智算中心推理业务能力的量化评估。

4 某省运营商智能客服系统评估案例

以某省智能客服系统试点项目为例，该省内建有千卡推理资源池，需要评估资源池可支撑智能客服系统的用户数量规模。根据业务特征、行业经验、硬件参数等取相关值代入本模型进行计算。

（1）AI加速卡参数分析

现有资源池部署某国产AI加速卡1 024张，显存容量VR为32 GB，显存带宽B为800 GB/s，根据厂商数据，带宽利用率r约为90%。

（2）模型参数分析

本文拟部署模型大小为13 B，权重LM约为26 GB，经测试，其单路KV Cache（Key-Value Cache）约为1.63 GB/s。

（3）业务参数分析

本文对历史数据进行分析，传统客服系统平均在线时长为6～10 min，因智能客服系统效率更高，按照客户平均在线时长5 min进行考虑，智算中心日服务时长为24 h。

时延SLA要求为200 ms。根据省内IT支撑系统建设经验，峰值集中系数按3取值。

对于处理用户请求的推理资源占用率，不同业务场景数值不同。根据咨询分析机构Gartner发布的《2023年中国ICT技术成熟度曲线》，通用聊天大模型场景的资源占用率为1.7%～3.3%。根据2023年发布的《阿里云智能客服知识运营白皮书》，在电商平台的智能客服场景中，用户实际触发大模型推理的时间占比为1.7%～2.8%。综合考虑省内客户系统情况，推理资源占用率取3%。

某智能客服系统推理资源参数及业务参数见表1。

将各项取值代入模型进行计算，得出该推理

表 1　某智能客服系统推理资源参数及业务参数

参数	参数说明	值
VR	AI 加速卡的内存大小	32 GB
OS	AI 加速卡内存中，系统占用空间	0 GB
LM	模型占用内存大小，与模型参数量相关	26 GB
Cache	每个会话在推理过程中占用的内存大小	1.63 GB
$Session_0$	单张 AI 加速卡支持的最大会话路数	3.69 路
T	时延 SLA 要求	0.2 s
B	GPU 卡访问内存带宽	800 GB/s
r	GPU 卡访问内存带宽利用率，为推理卡产品特征值	90%
$Session_1$	在给定时延 SLA 要求下，单张 AI 加速卡可支持的会话路数	72.62 路
AI	智算推理中心的资源规模	1 024 卡
LT	用户每日平均在线时长	0.08 h
PC	峰值集中系数	3
DT	智算中心日服务时长	24 h
UR	处理用户请求的推理资源占用率	3.00%
UC	单日在线用户数	1 210 万

资源池可支撑超过 1 200 万用户使用，目前系统已上线。系统上线后，99.2% 的用户请求响应时延控制在 200 ms 以内，用户满意度调研显示，问题解决率从 65% 提升至 89%，用户评分从传统系统的 3.2 分（5 分制）提升至 4.1 分。现支撑用户数量已接近 1 100 万，系统运行平稳，未出现无法支撑的峰值流量，且多数服务器的 GPU 利用率处于合理水平，这表明本文提出的评估模型取得了较为良好的应用效果。

同理，如在某项目中，已知需要支撑的用户数，需要计算项目所需的推理资源规模，则使用中间式（5）即可得到。

5　结束语

推理能力评估模型通过推理业务建模、用户访问建模，评估智算中心可支撑的用户访问量，量化评估了智算中心的业务处理能力，相较单纯使用计算能力评估更加接近实际业务场景。模型基于智能推理时延 SLA 要求，针对时延敏感的推理请求进行业务建模，量化评估了智算中心在特定时延 SLA 下的业务能力，时延 SLA 不同，业务承载能力随之变化，使得业务能力评估更加具体化、实景化。

随着通用 AI 的发展，模型训练的落地应用，应用场景日渐成熟，AI 将深入赋能社会各个领域，加速全社会“AI+”转型进程，AI 将遍布各个业务场景，赋能千行百业。然而，不同的业务场景需要有不同的智算能力与之匹配，只有科学量化评估一个智算中心的业务承载能力，才能明确建设规模、推进业务发展、提高业务质量、优化客户体验。

本文的评估方法，可以科学量化一个推理智算中心的业务承载能力，使得智能算力与业务需求精准匹配，有的放矢，充分发挥智算资源的推理能力，提高投资效益，保障业务质量。

参考文献：

[1] 杨杰 . 西班牙巴塞罗那世界移动通信大会(MWC 2024)开幕式主旨演讲[EB]. 2024.
YANG J. Keynote speech at the opening ceremony of the Mobile World Congress (MWC 2024) in Barcelona, Spain[EB]. 2024.

[2] 人民网 . MWC25|中国移动总经理何飚: 拥抱“AI+”新时代 共谱数智新篇章[EB]. 2025.

People's Daily Online. MWC25|He Biao, general manager of China Mobile: embrace the "AI+" new era and co-create a new chapter of digital intelligence[EB]. 2025.

[3] 中华人民共和国国家互联网信息办公室. 国家互联网信息办公室关于发布2024年生成式人工智能服务已备案信息的公告[EB]. 2025.
Cyberspace Administration of Chian. Announcement of the Cyberspace Administration of Chian on the release of the recorded information of generative artificial intelligence services in 2024[EB]. 2025.

[4] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[J]. arXiv preprint, 2020: 2001.08361.

[5] 中国软件评测中心. 人工智能大语言模型技术发展研究报告(2024年)[R]. 2024
China Software Testing Center. Research report on the development of artificial intelligence large language model technology (2024)[R]. 2024.

[6] WEI J, WANG X, COHEN T, et al. Chain of thought prompting elicits reasoning in large language models[J]. arXiv preprint, 2022: 2201.11903.

[7] RADFORD A, METZ L, CHINTALA S. CLIP: contrastive language-image pre-training[J]. arXiv preprint, 2021: 2103.00020.

[8] 中国算力大会. 中国智算中心服务发展报告(2024年)[R]. 2024.
China Computational Power Conference. China artificial intelligence data center service development report (2024)[R]. 2024.

[9] NARAYANAN D, KUMAR A, LI Z. Cloud-native AI: principles and practices[M]. California: O'Reilly, 2022.

[10] META. Meta's AI infrastructure: scaling to 350,000 accelerator cards[EB]. 2024.

[11] MICROSOFT. Building and operating a million-scale AI cluster: lessons from Microsoft's project Stellar[R]. 2024.

[12] 中国电信. 智算产业发展研究报告(2024年)[R]. 2024.
China Telecom. Research report on the development of intelligent computing industry (2024)[R]. 2024.

[13] 工业和信息化部, 教育部, 科技部, 等. 工业和信息化部等七部门关于推动未来产业创新发展的实施意见[EB]. 2024.
Ministry of Industry and Information Technology, Ministry of Education, Ministry of Science and Technology, et al. Implementation opinions of seven departments including MIIT on promoting the innovation and development of future industries[EB]. 2024.

[14] JIANG Y, WANG Y, DUAN Y, et al. High-performance deep learning inference on GPU: a survey[J]. ACM Computing Surveys, 2021, 54(3): 1-35.

[15] JOHN S, MATHEW A, JOSE P. A comprehensive analysis of GPU performance for deep learning inference[J]. IEEE Micro, 2022, 42(3): 24-33.

[16] BERTSEKAS D P, GALLAGER R G. Queueing theory in computer networks: performance modeling and analysis[M]. New York: Prentice Hall, 1991.

[作者简介]

武振宇（1975- ），男，中国移动通信集团设计院有限公司高级工程师，主要研究方向为云计算、算力网络、智算中心等。

赵占军（1971- ），男，中国移动通信集团设计院有限公司高级工程师，主要研究方向为云计算、智算等。

卜忠贵（1976- ），男，中国移动通信集团设计院有限公司正高级工程师，主要研究方向为云计算、智算等。

刘鹏（1985- ），男，中国移动通信集团设计院有限公司高级工程师，主要研究方向为智算、支撑系统等。

田雯（1987- ），女，中国移动通信集团设计院有限公司高级工程师，主要研究方向为云计算、智算等。

王祎玮（1995- ），女，中国移动通信集团设计院有限公司助理工程师，主要研究方向为云计算、智算等。

董江帆（1987- ），男，中国移动通信集团设计院有限公司工程师，主要研究方向为业务网、智算等。

王猛（1981- ），男，中国移动通信集团设计院有限公司安徽分公司高级工程师，主要研究方向为云计算、智算等。

张誌（1978- ），男，中国移动通信集团设计院有限公司山东分公司高级工程师，主要研究方向为智算、支撑系统等。