



人工智能大模型发展带来的智能算力需求测算研究

李双杰, 杜春香, 肖子玉

(中国移动通信集团设计院有限公司, 北京 100080)

摘要: 以 ChatGPT 为代表的人工智能大模型的爆发带动智能算力需求快速增长, 电信运营商、云服务商等纷纷加大智算中心布局, 大模型的发展对智能算力有多大规模需求、对智算中心的布局与建设有着十分重要的影响。从大模型对智能算力的消耗特点出发, 对影响大模型算力消耗的关键影响因素进行深入分析, 并基于不同参数量级不同行业类型构建了分层次分阶段的大模型智能算力需求测算模型, 对大模型对智能算力的需求进行系统分析, 为智算中心的建设提供明确的业务输入与支撑。

关键词: 大模型; 智能算力; 需求预测

中图分类号: TP301

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2024215

Research on intelligent computing power prediction based on AI large models

LI Shuangjie, DU Chunxiang, XIAO Ziyu

China Mobile Group Design Institute Co., Ltd., Beijing 100080, China

Abstract: AI large models, represented by ChatGPT, have driven a rapid growth in intelligent computing power demand. Telecom operators and cloud service providers are increasingly expanding their intelligent computing center layouts. The future trends of large models and the scale of their computing power demand are of great importance for the layout and construction of intelligent computing centers. Starting from the consumption characteristics of computing power by AI large models, the key influencing factors of computing power consumption were analyzed. Based on multi-level models with different parameter scales and industry types, the intelligent computing power demand of AI large models was systematically analyzed, providing clear business inputs and support for the construction of intelligent computing centers.

Key words: AI large model, intelligent computing power, demand forecasting

0 引言

随着数字经济的快速发展, 以算力为基础的数字信息基础设施逐渐成为经济社会数智化转型

发展的底座, 不仅改变了人类的生产方式、生活模式和科研范式, 也代表人类智慧新的发展水平^[1]。算力逐渐成为数字经济时代的核心生产力, 国际数据公司 (International Data Corpora-

tion, IDC) 2023 年发布的研究报告指出, 算力指数平均每提高 1 点, 国家的数字经济和 GDP 将分别增长 3.6‰ 和 1.7‰^[2]。

按照中国信息通信研究院 (以下简称“信通院”) 2021 年发布的《中国算力发展指数白皮书》的定义, 算力是数据中心的服务器通过对数据进行处理后实现结果输出的一种能力^[3]。按照类型划分, 算力可以分为通用算力、智能算力和超级算力, 具体定义与适用场景如下。

(1) 通用算力: 基于中央处理器 (central processing unit, CPU) 芯片的服务器所提供的算力, 主要用于云计算、边缘计算等基础通用计算。

(2) 智能算力: 由专为人工智能 (artificial intelligence, AI) 设计的芯片, 如图形处理单元 (GPU)、现场可编程门阵列 (FPGA) 和专用集成电路 (ASIC), 所构建的加速计算系统, 主要被用于支持人工智能大模型的训练过程和执行推理任务。

(3) 超算算力: 由高性能计算集群 (如超级计算机等) 提供的计算能力, 主要服务于科学研究前沿领域的复杂计算任务, 包括药物分子结构设计、行星模拟、基因序列分析等。

1 大模型发展情况

人工智能大模型 (以下简称大模型) 是指在大规模数据集上训练, 具备十亿 (或百亿、千亿或更多) 级别参数的深度神经网络模型, 能够捕捉和学习非常复杂的数据模式, 能够同时学习多种不同任务, 这也有助于模型学习更广泛和泛化的语言理解能力^[4]。2023 年 ChatGPT 的火热出圈激发全球各大科技企业积极拥抱生成式人工智能 (artificial intelligence generated content, AIGC), 直接引发了国内大模型企业的“百模大战”。截至 2024 年 8 月, 国内完成备案并上线、能为公众提供服务的生成式人工智能服务大模型已达 190 多个, 注册用户超过 6 亿。按照 Gartner 新兴技术演进的扩散曲线, 预计大模型发展将经历以下 3 个阶段。

(1) 期望膨胀期 (2023—2024 年): 主要特征是大模型数量快速增加, 涌现出各类大模型相关企业, 包括通用大模型、行业专用大模型、应用开发等, 大模型发展处于以算法和数据训练为核心的技术驱动阶段, 开源可商用的模型数量少, 模型能力处于初步可用的水平, 尚未形成成熟可持续的商业模式。在这个阶段, 各类大模型数量快速增长、技术持续迭代是智能算力需求增长的主要驱动力, 智能算力需求将呈现短期内激增态势。

(2) 生态培育期 (2025—2028 年): 主要特征是大模型数量和企业呈现下降趋势, 大模型行业进入稳步发展和整合阶段。具有充裕资金、较强技术实力和独特商业模式的大模型企业将能够生存。在这个阶段, 大模型技术演进不断进步, 头部大模型参数和数据量持续上升, 智能算力训练需求将仍保持持续增长; 商业模式趋向成熟, 面向个人的 AIGC 用户规模和交互量持续上升, 将促使智算推理资源需求快速增长; B/G 端需要针对行业领域和业务场景进行大量工程工作, 尤其是传统行业的知识获取和积累需要较长时间, 行业大模型稳步发展, 不断适配行业发展需求。智能算力需求将呈现较快增长趋势。

(3) 商业落地期 (2028 年后): 主要特征是基于现有算法理论的基础大模型技术成熟, 商业模式逐渐清晰, 具备规模落地条件。在该阶段, 行业专用大模型规模发展, 业务场景和商业模式更为清晰, 企业将把精力更多从训练转移到推理工作负载上, 智能算力需求结构将发生变化, 推理资源需求与占比将快速增长, 推动整体智能算力保持稳步增长。

2 智算需求研究概述

大模型从建立到正式推出需要经过训练和推理两个过程, 对智能算力有着巨大需求。信通院发布的《中国综合算力指数 (2023 年)》显示,



截至2023年6月月底,国内算力规模中,通用算力规模占比达74%;智能算力规模占比达25%,同比增长60%,预计未来智能算力规模与占比将随着大模型的深入应用持续提升^[5]。大模型的爆发掀起新一轮人工智能应用热潮,进一步加剧了智能算力的供需不平衡,成为了推动智能算力需求增长的关键因素。为满足大模型快速发展带来的巨大智能算力需求,各地政府积极出台相关规划政策,促进智算中心的建设。当前,中国算网基础设施正在经历以下3个转变。

一是资源分布转变:从分散的资源集聚向形成集群转变,构建热点集约、跨区辐射、边缘覆盖的基础设施体系,以满足不同地区和不同行业场景的差异化需求。

二是功能性能转变:基础设施的功能性能从以通用计算为主转向集成通用计算、智能计算、超级计算、量子计算等多元计算能力,以支持多样化的计算需求。

三供给模式转变:从计算和网络分离的模式向计算与网络共生的模式转变,推动基础设施的综合集成和提供一体化服务,以实现算网资源的高效便捷利用。

在这种情况下,电信运营商、云服务商也积极响应政府号召,纷纷加大智算中心的布局。中国移动规划“N+X”智算中心布局体系,实现“东数西算”枢纽节点基本覆盖,其在内蒙古构建的亚洲最大单体智算中心拥有万片级AI加速芯片,智能算力规模高达6.7 EFLOPS。阿里云的张北智算中心集群总建设规模达到12 EFLOPS,超过谷歌的9 EFLOPS和特斯拉的1.9 EFLOPS,成为全球最大智算中心。

虽然目标与方向比较清晰,但从资源建设角度看,大模型带来的智能算力需求未来能够达到何种规模、在不同行业的分布、推理与训练对算力需求结构等均会对智算中心的建设带来至关重要的影响。因此,有必要从大模型出发构建一套

量化的测算模型对未来智能算力的发展进行系统分析,为智算中心的建设提供输入与指引。

3 智能算力测算模型

3.1 研究概述

针对大模型与智能算力之间的关系,有两个著名的定律。一个是由OpenAI团队研究发现的“规模定律(scaling law)”特征,即模型性能与模型规模、数据集大小和训练用的计算量之间存在幂律关系,性能会随着这3个因素的指数增加而线性提高,因此,大模型消耗的智能算力等于模型规模、数据集大小和单位消耗的乘积^[6]。另一个是由美国谷歌联合斯坦福大学提出的“涌现能力(emergent abilities)”,即如果一种能力不存在于较小模型中,而存在于较大模型中,那么这种能力就是涌现出来的,通俗地讲就是当模型规模或复杂性达到一定程度时,新的行为或能力突然出现^[7]。李国杰院士在研究中也指出大模型核心特征是“涌现能力”^[8]。大模型参数规模远大于传统深度学习模型,传统模型参数量通常在数万至数亿之间,大模型参数规模通常达到数亿,并且已经扩展到超过万亿。如OpenAI的GPT-1到GPT-3,参数量从1.1亿大幅提升至1 746亿,GPT-4则包含1.8万亿参数,实现性能的突破。

目前国内相关研究主要是在这两个定律基础上针对人工智能大模型发展趋势及对智能算力需求方面开展:

(1) 文献[9]提出大模型是通过深度学习算法和人工神经网络训练出的具有庞大规模参数的人工智能模型,是建立在网络能力、数据能力和计算能力三者之上的创新应用;

(2) 文献[10]认为实现高效训练已成为影响大模型应用普及的关键要素之一,大模型训练需要大量数据、资源和时间,涉及服务器中所有资源的综合利用;

(3) 文献[11]提出参数优化与数据集构建是

大模型产品普及与迭代的核心问题，并对类 Chat GPT 大模型在教育、医疗、金融等不同领域的应用进行了分析：

(4) 文献[12]提出到 2025 年我国算力规模将超过 300 EFLOPS，智能算力与总算力的比例将达到 35%；

(5) 文献[13]认为未来我国智算中心的建设进程将会进一步加速，以满足更高性能的技术需求；

(6) 文献[14]对当前各地区出台的鼓励智算中心建设的政策进行了梳理，对智算中心从数据上传、数据计算、模型推理等方面如何支撑大模型的过程进行了分析；

(7) 文献[15]提出算力时代的三定律，其中算力第二定律（增长定律）指出，算力每 12 个月增长一倍。

3.2 测算思路

根据“规模定律”，影响大模型智能算力消耗的因素主要是大模型的参数量、数据集大小、每参数每 token 需要的算力，大模型智能算力需求的计算式为：

大模型智能算力消耗=大模型参数×

数据集大小×每参数每 token 需要的算力 (1)

虽然每个大模型可以按照式 (1) 测算其智算消耗需求，但对某个区域来说，大模型千差万别，若按照统一的某种类型进行测算，结果难免误差较大。为了使测算更加科学精准，可以综合考虑参数量级、基础与行业分类等因素，分类开展精细化测算，汇总后得到未来总的智能算力需求。具体分类如下。

(1) 分参数量级：将大模型分为百亿级、千亿级和万亿级 3 类。每个参数量级的大模型选取一个典型模型作为该类的样本，以样本大模型的参数作为该类大模型的参数基准，然后测算该类大模型的总智能算力消耗。

(2) 分模型类型：将大模型分为基础大模型与行业大模型两类。在参数量级相同的情况下基

础大模型和行业大模型的差别重点在训练阶段所需要的数据集不同，基础大模型训练阶段需要的数据集要远高于行业大模型。

因此，按照参数量级和模型类型组合可以得到 6 种类型的大模型，大模型 6 种类型情况如图 1 所示。

行业大模型	类型 II (百亿, 行业)	类型 IV (千亿, 行业)	类型 VI (万亿, 行业)
基础大模型	类型 I (百亿, 基础)	类型 III (千亿, 基础)	类型 V (万亿, 基础)
	百亿大模型	千亿大模型	万亿大模型

图 1 大模型 6 种类型情况

针对每类大模型，分别测算其训练和推理两个阶段所消耗的智能算力，然后汇总得到该类大模型的智能算力需求。智能算力需求规模测算思路如图 2 所示。

训练 (training) 阶段：通过大量数据和算法，大模型学会识别和生成规律。模型参数在此过程中不断调整，以最小化预测与实际值之间的误差，从而使其具备适应各种任务的学习能力，涵盖图像识别到自然语言处理等多个领域。训练阶段需要反向传播算法，推理阶段需要正向传播算法，反向传播需要的运算次数是正向传播 2 倍，故训练时每个词的运算量是推理情况的 3 倍^[7]。该阶段智能算力消耗计算式为：

训练阶段大模型算力=大模型参数×

训练数据集×3×每参数每 token 所需算力 (2)

推理 (inference) 阶段：建立在训练完成的基础上，将训练好的模型应用于新的、未见过的数据。模型利用先前学到的规律进行预测、分类或生成新内容，使得 AI 在实际应用中能够做出有意义的决策。推理阶段数据集大小与大模型访问人次和单次访问所需 token 量有关。该阶段智能算力的消耗计算式为：

推理阶段大模型算力=模型参数数量×

单日访问人次×单次 token 量×

每参数每 token 所需算力 (3)



因此,大模型带来的智能算力需求等于训练与推理之和,即:

$$\text{大模型智能算力消耗} = \text{训练阶段算力消耗} + \text{推理阶段算力消耗} \quad (4)$$

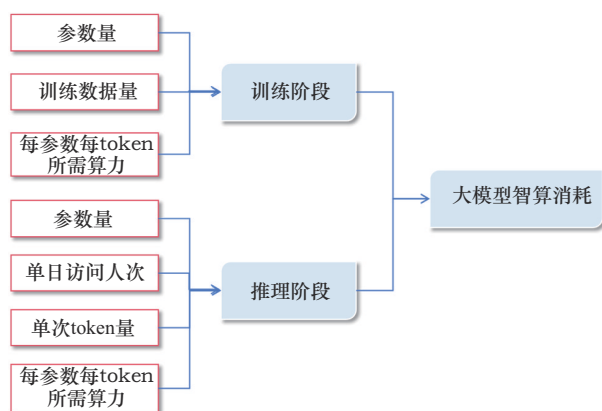


图2 智能算力需求规模测算思路

AI大模型的爆发式增长带动了智能算力需求的快速增长,预计最近5年内智能算力的需求主要来自与大模型相关的研究和应用,因此,可以从大模型智能算力需求占比来测算未来智能算力的需求:

$$\text{智能算力需求} = \frac{\text{大模型智能算力需求}}{\text{大模型智能算力占比}} \quad (5)$$

3.3 结果测算

本文基于某市智算中心规划项目,在模型搭建过程中,针对阿里云、华为云、智谱清言等行业大模型重点企业进行了走访调研,当前国内大

模型发展呈现以下特征。

一是大模型参数规模加速增长。图2所示为当前大模型参数发展变化情况,模型参数规模每年按照十倍左右速度高速增长,从2020年百亿级别逐渐发展到2023年万亿级别。目前,具有千亿参数的大模型已经成为了行业主流。按照参数量级可以将目前市面主流大模型分为3类:百亿级、千亿级、万亿级。

二是从类别看,目前大模型可分为基础大模型与行业大模型。行业大模型在通用大模型的基础上,通过结合特定行业的数据进行进一步训练而形成的。这类模型主要服务于特定的行业应用,其在训练过程中所需的计算资源消耗相对较低。随着大模型技术的成熟,自然而然出现了从通用模型向行业定制模型的过渡。只有少数具备强大技术和资金实力的公司会去开发基础大模型,而更多企业可以基于这些基础模型,针对具体应用场景进行调整和优化,从而开发出多样化的行业专用模型和针对细分场景的模型,进而推动人工智能技术在各个行业的应用和创新。

三是大模型将分批在重点行业场景推广应用。人工智能技术正处于从“理论可用”向“实际应用”的过渡时期,其在商业领域的应用仍处于早期发展阶段。目前,该领域面临的挑战包括应用场景需求的碎片化、高昂的人力开发和计算

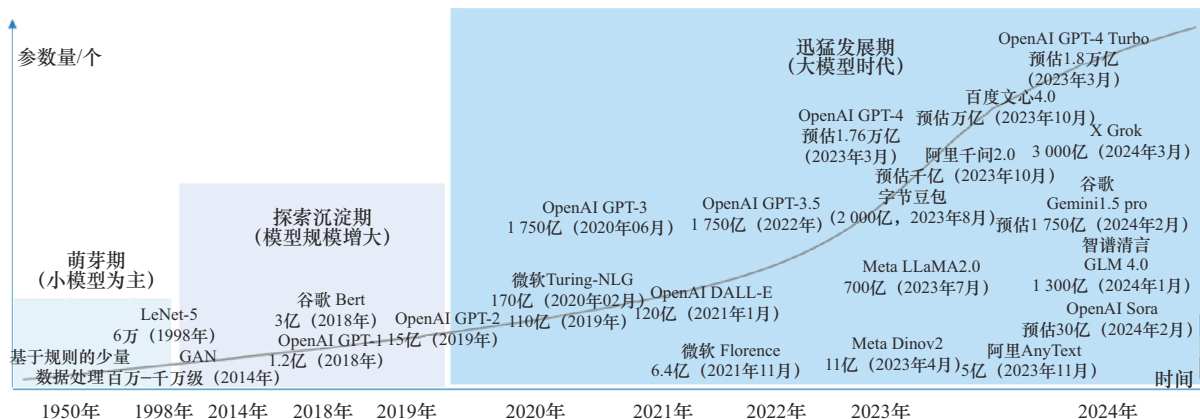


图3 大模型发展参数变化情况

成本，以及长尾场景数据稀缺导致模型训练精度不足，还有从实验室环境到实际应用场景中模型算法的适应性差异。预计在数据资源、资金实力和数字化基础方面具备优势的行业，如政府服务、金融、医疗、通信和互联网等，将率先实现大规模模型的应用。随着技术门槛降低和成本减少，商业、教育等其他领域也将加速采用大模型技术。同时，不同地区会根据其特色产业和优势，制定相应政策来促进大模型在本地产业中的应用，以此来加强和提升这些产业的竞争力。

在参数方面，参考OpenAI在2020年发表的针对GPT-3的论文，GPT-3的模型参数为1 750亿，训练数据量为45 TB，折算3 000亿token，消耗算力为 3.15×10^8 PFLOPS^[7]。而从GPT-4公开的信息来看，GPT-4的模型参数为1.8万亿左右，训练数据量为13万亿，消耗算力约为 2.15×10^{10} PFLOPS，在大约25 000个A100上训练了90到100天。

综合考虑国内外大模型发展情况，大模型智能算力模型关键参数取值见表1。

按照上述模型和参数取定对某地区智能算力需求进行了测算，2023—2026年智能算力预测结果见表2，结果显示，未来3年智能算力需求保持高速增长趋势，具体呈现以下特征。

(1) 智能算力需求总量保持高速增长：预计

表1 大模型智能算力模型关键参数取值				
指标名称	单位	百亿	千亿	万亿
单模型参数量	亿个	100	1 000	10 000
基础大模型单日使用人数	万	30	50	100
行业大模型单日使用人数	万	1	4	5
大模型单人平均使用次数	次	15	10	10
大模型峰值总计访问量	亿人次	0.045	0.05	0.1
每次访问对话次数	次	12	12	12
每次对话token量	个	150	260	300
单token训练所需算力	FLOPS-day	5.64×10^{-5}	2.31×10^{-5}	3.94×10^{-6}

2023年为8.9 EFLOPS，2025年为25.3 EFLOPS，2026年为37.4 EFLOPS，2024—2026年每年平均复合增速为61%。

(2) 推理阶段消耗算力占比持续提升：预计2023年训练与推理算力占比分别为64%与36%，2025年变化为32%与68%，2026年变化为27%与73%。

(3) 行业大模型消耗算力占比稳步增长：2023年基础大模型与行业大模型算力消耗占比分别为84%与16%，到2025年分别变化为54%与46%，2026年变化为45%与55%。

(4) 万亿大模型算力消耗占比不断增加：2023年万亿级、千亿级、百亿级大模型算力消耗占比分别为55%、40%、5%，2026年分别变化为61%、37%、3%。

表2 2023—2026年智能算力预测结果

指标名称		2023年	2024年	2025年	2026年
智能算力总需求/EFLOPS		8.9	15.1	25.3	37.4
增速		—	69%	67%	48%
大模型智能算力总需求/EFLOPS		54	106	179	273
大模型智能算力需求占比		60%	70%	77%	82%
算力结构	基础大模型算力占比	84%	65%	54%	45%
	行业大模型算力占比	16%	35%	46%	55%
	训练算力占比	64%	43%	32%	27%
	推理算力占比	36%	57%	68%	73%
	万亿模型算力占比	55.2%	56.9%	58.2%	60.6%
	千亿模型算力占比	40.0%	38.8%	38.2%	36.6%
	百亿模型算力占比	4.8%	4.2%	3.5%	2.8%



4 结束语

本文研究成果已经在某市智能算力规划项目中进行了应用,基于大模型的智能算力需求也成为该市智算中心规划建设工作的重要依据和参考。对后续相关企业智算中心的布局、政府相关产业规划等均有一定的借鉴意义。下一步,针对大模型带来的智算需求测算研究有如下建议。

一是结合智算中心建设进一步丰富需求测算结果的输出。尤其是面向国家“东数西算”工程的布局,要重点考虑大模型发展对智能算力位置需求,在本模型基础之上,综合考虑成本、数据安全、业务时延等因素,围绕不同行业不同场景的特点进一步进行梳理,研究大模型智能算力不同区域分布需求及变化。

二是根据大模型发展的新情况对本文提出的参数进一步优化。当前大模型技术发展日新月异,本文是在之前的研究基础上结合大模型当下发展提出的一种测算思路,未来随着大模型规模越来越大,本文中提到的一些参数也需要进一步更新调整。

参考文献:

- [1] 李正茂,雷波,孙震强,等.云网融合:算力时代的数字信息基础设施[M].北京:中信出版社,2022.
LI Z M, LEI B, SUN Z Q, et al. Cloud network convergence: computing era's digital information infrastructure[M]. Beijing: CITIC Press Group, 2022.
- [2] IDC. 2022-2023全球算力发展指数评估报告[R]. 2023.
IDC. 2022-2023 global computing index[R]. 2023.
- [3] 中国信息通信研究院.中国算力发展指数白皮书[R]. 2021.
CAICT. White paper on China's computing index[R]. 2021.
- [4] 孙长秋,杜长斌,李菲,等.人工智能及大模型技术研究[J].通信管理与技术,2024(3): 43-47.
SUN C Q, DU C B, LI F, et al. Research on artificial intelligence and large model technology[J]. Communications Management and Technology, 2024(3): 43-47.
- [5] 中国信息通信研究院.中国算力发展指数白皮书(2023年)[R]. 2023.
CAICT. White paper on China's computing index(2023)[R]. 2023.
- [6] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[J]. 2020, arXiv:2001.08361.
- [7] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[J]. 2022, arXiv: 2206.07682.
- [8] 李国杰.大数据与计算模型[J].大数据,2024,10(1): 9-16.
LI G J. Big data and computing models[J]. Big Data Research, 2024, 10(1): 9-16.
- [9] 孙柏林.大模型评述[J].计算机仿真,2024,41(1): 1-7,24.
SUN B L. Review of large models[J]. Computer Simulation, 2024, 41(1): 1-7,24.
- [10] 田海东,张明政,常锐,等.大模型训练技术综述[J].中兴通讯技术,2024(8): 21-28.
TIAN H D, ZHANG M Z, CHANG R, et al. A survey on large model training technologies[J]. ZTE Technology Journal, 2024(8): 21-28.
- [11] 严昊,刘禹良,金连文,等.类ChatGPT大模型发展、应用与前景[J].中国图象图形学报,2023,28(9): 2749-2762.
YAN H, LIU Y L, JIN L W, et al. The development, application, and future of LLM similar to ChatGPT[J]. Journal of Image and Graphics, 2023, 28(9): 2749-2762.
- [12] 温小振,常金凤,吴美希.综合算力发展现状与趋势分析[J].信息通信技术与政策,2024,50(2): 7-11.
WEN X Z, CHANG J F, WU M X. Analysis of current situation and trend of comprehensive computing power development[J]. Information and Communications Technology and Policy, 2024, 50(2): 7-11.
- [13] 王月,盛凯,常金凤,等.数据中心基础设施关键技术应用及发展[J].信息通信技术与政策,2021,47(4): 27-31.
WANG Y, SHENG K, CHANG J F, et al. Application and development of key technologies in data center infrastructure[J]. Information and Communications Technology and Policy, 2021, 47(4): 27-31.
- [14] 殷凯凯.基于电信运营商视角的智算中心规划建设关键点分析及建议[J].现代信息科技,2024,8(1): 22-27.
YIN K K. Analysis and suggestions of key points for planning and construction of artificial intelligence computing center based on telecom operators' perspectives[J]. Modern Information Technology, 2024, 8(1): 22-27.

- [15] 李正茂, 王桂荣. 论算力时代的三定律[J]. 电信科学, 2022, 38 (6): 13-17.

LI Z M, WANG G R. Laws of computing power in the computing power era[J]. Telecommunications Science, 2022, 38 (6) : 13-17.

[作者简介]



李双杰 (1987-), 男, 中国移动通信集团设计院有限公司高级工程师, 主要研究方向为算力经济、产业发展洞察。



杜春香 (1974-), 女, 中国移动通信集团设计院有限公司高级工程师, 主要研究方向为算力经济、投资管理。



肖子玉 (1969-), 女, 中国移动通信集团设计院有限公司正高级工程师, 主要研究方向为算力网络、5G/6G 网络。