AI 大模型快速发展推动算力供给结构改革

路宇浩 *，王童，吴琦婕，姚雪冰，徐影
中移数智科技有限公司，北京 100080
DOI: 10.61369/TACS.2025010040

摘　　要： 当前，人工智能发展已进入以大模型为代表智能时代，由于其参数规模大、数据量大等特点，导致传统规模效应下的算力服务已难以满足需求。本文基于人工智能技术发展趋势与大模型的特点，分析了大模型对存储、计算和网络的需求，剖析了当前算力供给所存在的异构算力纳管与适配程度低、算力集约程度和利用率低、算力资源供给缺乏弹性、算力资源配置不合理、算力使用不够便捷等痛点问题。最后提出了大模型背景下的算力供给体系框架建议。

关 键 词： 算力；算力供给；大模型

The Rapid Development of LLMs Drives Reform of Computing Power Structure

Lu Yuhao*, Wang Tong, Wu Qijie, Yao Xuebing, Xu Ying
Chinamobile Digitization Technology, Beijing 100080

Abstract: Currently, the development of artificial intelligence has entered the era of intelligence represented by large models. The traditional arithmetic service has been difficult to meet the demand. Based on the development trend of AI technology and the characteristics of large models, this paper analyzes the demand of large models for storage, computation and network. This paper summarizes five points of problems and puts forward a proposal for the framework of the arithmetic supply system in the context of big models.

Keywords: computing power; computing power service; large language model

前言

大模型是具有大规模参数的复杂机器学习模型，其展现出的涌现能力[1]使得大模型具有更强的感知、认知与生成能力。由于需要依赖大规模的训练数据和参数，大模型对于算力的需求呈爆发式增长。

我国十分重视 AI 产业发展与算力基础设施建设。近年来我国算力规模不断扩大，截止到2024年底我国智能算力规模达到90 EFLOPS，预计到2026年将进入每秒十万亿次浮点运算时代[2]。

然而我国算力供给能力与大模型需求之间存在差距。建设侧，异构算力设施由多家企业提供，独立建设与运营，资源协调配合难；供需侧，全国数据中心闲置率较高，算力资源供不应求；场景侧，缺乏针对不同场景的协调机制，算力普惠应用受限。上述问题严重制约了算力供给效能。

本文重点围绕人工智能发展历程与算力需求变化关系，从大模型的特点入手分析大模型对于算力的需求以及面临的问题和挑战，并提出大模型背景下的算力供给体系框架建议。大模型对算力服务的新需求。

AI 技术发展可归纳为四个阶段。第一阶段是基于规则的，依靠专家知识和预定义规则，通常部署在单体计算机中。第二阶段是基于统计的，应用统计学方法构建模型，数据集小、迭代次数少，单体计算机即可完成计算任务。第三阶段是基于深度学习的，通过神经元连接提取特征实现高维抽象表达，数据量和计算复杂度显著提升，需多个 CPU、GPU 参与，通常使用服务器集群。第四阶段是基于大模型的，可跨任务、跨模态建模，其性能取决于模型大小、数据集大小和计算量，对算力供给模式提出新挑战。

通讯作者简介：路宇浩，中移数智科技有限公司（中国移动设计院有限公司）数智化咨询总监、高级工程师，研究方向为云原生体系下算力网络、云计算、智算技术架构设计与场景化实践。

一、大模型对算力服务的新需求

大模型对于存储、计算与网络等算力服务新需求如图1所示。

（一）存储需求

大模型对于存储的需求主要为可扩展的存储容量、多源异构数据存储以及高性能的跨设备读写。

大模型的数据量和参数量与日俱增，需要可扩展的存储容量。以GPT-3为例，在预训练阶段，GPT-3使用的语料库体量约为45TB[4]。单个节点的检查点往往也达上百GB，随着模型规模的扩大还会持续上涨。

大模型的训练数据呈现出多模态的趋势，对多元异构数据的存取具有需求。多模态大模型对文本、语音、图像、视频等多种模态数据进行协同处理[5]，具有更强的知识获取能力和推理性能，能够加速大模型能力涌现，是未来的一大发展趋势。

大模型的数据来源广且读写频繁，对于高性能的跨设备读写具有较强需求。一方面需要保证跨设备、跨数据中心的读写性能。另一方面存储节点和计算节点间存在着频繁的数据交互，需实现数据高吞吐和高并发以缩短数据准备的等待时间。

（二）计算需求

大模型对于计算的需求主要为灵活的分布式并行计算、加速芯片间的异构计算以及自动化的算力分配。

大模型对芯片的显存容量和计算能力提出挑战，分布式并行计算可有效解决。AI训练过程中使用的计算量每3.4个月翻一倍[6]，远超摩尔定律。灵活的分布式并行计算可自适应生成训推策略，拆分模型和数据，提升整体性能。

为提升大模型计算速度，需要多种加速芯片间的超异构算力协同。当前常用的异构计算方案多为两层异构架构，未来可能形成“CPU+GPU+FPGA+ 其他”的多层超异构协同计算，充分发挥各类高性能芯片的优势，以获得更优的计算速率和计算精度。

大模型训练和推理阶段的计算量不同，需要计算资源能够随需求的变化而灵活调整。研究表明大模型训练阶段与推理阶段所需的算力相差约三倍[2]。算力资源应动态调整，在训练时调度较大算力，训练结束后将空闲资源再分配给其他计算任务，防止海量算力资源闲置，提高资源利用率。

（三）网络需求

大模型对于网络的需求主要为超高速率、超大规模、超常稳定、超低时延、负载均衡和自动化网络部署。

训练阶段，算力底座需满足超大规模、超高速率和超长稳定要求。网络规模随算力集群的扩大而增加。数据并行、张量并行和流水线并行的通信量与模型参数量、数据批量大小及层间连接数相关，提升速率可降低传输时间。由于算力集群规模大，小故障发生率也会成为大概率事件，因此大模型训练对网络稳定性要求极高。

推理阶段，大模型对负载均衡和超低时延具有较强需求。为避免过长的用户等待时间，需要通过合理的调度算法，使网络具备处理多事务高并发请求的能力。在自动驾驶、金融交易等实时业务场景中，往往要求在几毫秒内接收反馈，面向大模型的网络传输需要具备超低时延的特性。

不同场景对算力服务的需求各异，网络应实现自动化配置。大模型业务在规模、任务类型、业务场景、安全属性等方面要求各异，导致对网络资源的需求差异性很大，网络应具备自动化配置能力，随任务需求弹性伸缩，自动感知业务类型和负荷，自适应网络参数，做到多台并行的灵活部署，防止资源浪费或过载。

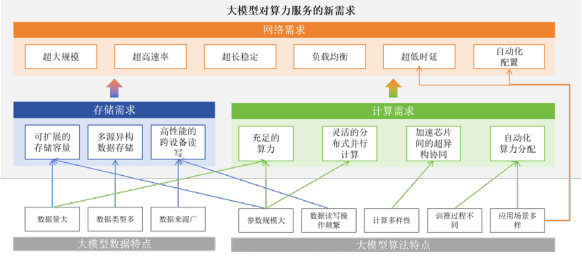

图1 大模型特点以及对算力服务的新需求

二、当前算力供给存在的问题与挑战

大模型出现后算法对于硬件设施和计算架构的依赖性变强，算力体系对大模型应用和发展起着决定性作用。

第一，异构算力纳管与适配问题突显。不同类型的异构芯片具有不同的体系结构和编程模型，不同厂家的AI芯片在通信协议、计算精度等方面也具有差异，未来更多异构芯片的发明以及更多异构资源组合的出现，将使得不同框架、不同硬件、不同厂商间的统一化接入更加困难，异构算力的纳管与适配问题日益突显。应当构建统一的异构算力资源池，保障异构芯片计算的高效协同。

第二，算力集约程度和利用率较低。我国算力节点呈高度分散状态，但相互之间尚未建立有效连接。另外，尽管我国算力规模持续增长，但存在大量闲散算力。据统计，我国数据中心的平均利用率只有55%，许多数据中心甚至不到20%[7]。因此，应当建立算力间的泛在连接，形成一体化算力体系，调动全部算力的使用，提高算力的集约程度和整体利用率。

第三，算力资源供给缺乏弹性。当前大多数的算力交付主要以提供指定规格的算力资源为主[8]，这需要使用方对设备型号、算法适配、模型部署等方面进行一定的研究。并且算力资源一旦锁定，即使在用户未使用时也无法将闲置资源重新分配给其他用户，使用率低。算力服务应当从资源式的刚性供给向任务式演进[9]，根据作业需求对算力进行实时的扩张或收缩，完成算力的动态调度和分发。

第四，算力资源配置不合理。存储、算力和网络带宽三者相辅相成，任何一项都会影响算力集群的整体性能。例如，只单独提升单卡计算能力和设备总数，而不提升通信性能，会由于通信瓶颈的存在而使设备间加速比急速下降，导致集群性能提升出现折损。同样，存储空间不够也会导致大模型部署时出现存储溢出的问题，导致训推失败。应当寻找一个科学的资源配置方法，合理配置算力资源，充分释放资源潜力。

第五，算力使用不够便捷。自建算力集群需要考虑资源类

型、配置、适配及兼容等问题，建设和运营成本巨大。而租用算力服务商资源，需要使用者估算所需的资源类型，并考虑更换服务商的代码迁移问题，为算力的使用带来了极大的不变。算力无法普惠大众是制约大模型研究发展的主要问题。应当从便捷算力的交易和运营出发，建立算力共享机制，实现算力的高效利用和长效运营。

三、结论与建议

大模型时代对算力服务提出了新要求，亟需优化算力供给架构来支撑大模型产业发展。算力供给体系的建设是一个系统性工程，可以从算力体系框架和技术研发两个方面推进。算力体系框架方面，以芯片层异构算力统一纳管为基础，整合数据中心级与区域算力基础设施，构建“超级计算机”，建立国家、区域、数据中心分级算力调度机制与能力，真正形成算力供给一盘棋。同时在运营管理层推动分层分级算力资源配置策略建立，集约化高效投入，明确算力并网、交易制度流程，据此构建算力服务的长效运营机制，保证算力供给体系正向循环。算力关键技术研发方面，推动异构资源纳管与调度、算－存－网协同优化、大规模分布式训练、零信任安全等技术落地，与算力供给体系架构相辅相成，共同赋能 AI 产业发展。

参考文献

[1]Kaplan J, Mccandlish S, Henighan T, et al. Scaling Laws for Neural Language Models[J]. 2020.

[2]2024年政府工作报告.[EB/OL].https://www.gov.cn/gongbao/2024/issue_11246/202403/content_6941846.html.

[3] 陈晓红，曹廖滢，陈姣龙，等. 我国算力发展的需求、电力能耗及绿色低碳转型对策. 中国科学院院刊，2024，39(3)：528. 539.

[4] 中国通信工业协会数据中心委员会 . 中国算力产业高质量发展白皮书（2023年）. [EB/OL].[2023-12-12].https://mp.weixin.qq.com/s/6i3eDS_OG7OAkOyn-q3qew.

[5]Brown T B, Mann B, Ryder N, et al.Language Models are Few-Shot Learners[J]. 2020.DOI:10.48550/arXiv.2005.14165.

[6] 刘建伟，丁熙浩，罗雄麟 . 多模态深度学习综述 [J]. 计算机应用研究 ,2020,37(6):1601-1614.

[7]AI and compute[EB/OL]. https://openai.com/research/ai-and-compute.

[8] 郭凤仙，孙耀华，彭木根 . 6G 算力网络：体系架构与关键技术 [J]. 无线电通信技术 ,2023,49(1):21-30.

[9] 陈晓红，许冠英，徐雪松，等 . 我国算力服务体系构建及路径研究 [J]. 中国工程科学 ,2023,25(6):49-60.

[10] 中国信息通信研究院云计算与大数据研究所 . 中国算力服务研究报告（2023年）[R].2023.

