DOI: 10.7544/issn1000-1239.qy20260801 CSTR: 32373.14.issn1000-1239.qy20260801

前　言

面向AI的算力网作为新一代智能信息基础设施，通过高速网络将分布各地的智能算力资源整合为可统一调度的算力资源池，为大规模智能应用提供高效、灵活、可靠的弹性算力服务体系。从大规模算力集群到分散式边缘算力节点，再到广域自组织的联盟算力网，算力网正从“连得通”向“算得优”“用得智”快速演进。算力网的内涵正在从连接与调度扩展为融合AI原生能力的智能体网络与协作基础设施。为此，《计算机研究与发展》联合第三届CCF分布式计算大会暨中国算力网大会（CCF Computility 2026）组委会，组织专题“AI算力网：系统与关键应用”，收录的6篇文章从多个维度切入算力网的关键技术问题，内容涵盖联盟算力网服务质量控制、算力网多智能体安全、边缘算力网络模型协同推理、学习驱动模型量化、面向多Chiplet加速器的多DNN负载在线调度等方向。

联盟算力网协调跨域异构计算资源，同时保留各独立提供商的管理自治权。这种自治性限制了联盟层的全局可观测性，使各域网关在进行路由决策时无法获得实时域内状态，只能依赖粗粒度且存在滞后的性能摘要。在这种弱可观测和状态滞后的条件下，时延敏感型任务的超时风险容易在少数域和少数窗口内发生结构性聚集。针对这一问题，周钰雯等人的“广域联盟算力网的服务质量反馈控制机制”提出了一种域级SLO反馈机制PDSF，其利用窗口级统计摘要动态更新时延敏感型任务的路由代价，从而对基于状态快照的自路由过程施加偏置调节。理论分析表明，在路由响应满足弱单调性的条件下，PDSF具有有界性、遗忘性以及延迟反馈下的局部稳定性。与现有可部署基线方法相比，PDSF能够显著缓解长尾过载，降低最差域暴露风险，并在保护时延敏感型任务的同时避免尽力而为型任务性能出现明显退化。

多智能体辩论作为一种分布式多专家协同范式，依托算力网络连接多个大语言模型智能体，以差异化角色进行多轮迭代讨论，可有效提升协同推理质量。这种基于算力网的分布式多专家协同范式在面对越狱攻击等安全威胁时，其内生安全性仍缺乏深入探索。为此，邹逸飞等人的“基于结构化提示改写的多智能体辩论越狱攻击”在半黑盒威胁模型下提出了一种针对多智能体辩论的结构化提示改写方法：通过叙事封装规避安全检测；角色驱动升级制造对抗压力；迭代精化逐步诱导有害细节；修辞混淆掩盖恶意意图；协同利用辩论交互机制生成有害输出。在4种主流多智能体辩论框架上使用GPT-4o、GPT-4、GPT-3.5-turbo及DeepSeek开展实验，结果表明该方法使生成内容的平均有害性提升28.14%~80.34%，攻击成功率在特定场景下高达80%。在此基础上，本文进一步围绕输入、生成和传播3个环节开展了初步防御实验，初步验证了多环节轻量级防护思路在多智能体辩论系统中的可行性。

在算力网中部署大语言模型推动了边缘推理需求的快速增长，但模型参数规模的持续膨胀给边缘侧的存储、计算和能耗带来了巨大压力。模型量化作为模型压缩的关键技术，能够有效降低模型的资源开销。然而，在极低比特（≤4 bit）条件下，基于固定规则或经验启发的

传统量化方法在精度保持、训练稳定性以及跨架构与应用场景适配性方面仍存在明显局限。针对这一挑战，黄馨怡等人的“学习驱动的模型量化：方法、挑战与展望”系统梳理了学习驱动量化的核心方法与关键机制，以后训练量化（PTQ）与量化感知训练（QAT）两类典型范式为主线，总结了可学习参数建模、重构与近似优化、辅助信号监督以及数据和硬件约束下的学习驱动量化性能增强方法。从基于误差解析的量化参数与学习信号优化、学习驱动量化中精度与资源的权衡，以及跨模型架构的学习驱动量化统一框架等方面，对学习驱动量化的未来发展趋势进行了展望。

在网络边缘有大量闲置的算力、内存容量与I/O带宽资源，这为汇聚分布式算力资源形成边缘算力网，进而突破单设备性能瓶颈提供了重要机遇。基于混合专家（MoE）架构的大语言模型参数量巨大，在单一边缘设备上部署面临严峻挑战，通常采用专家交换策略将非活跃专家卸载至二级存储并按需加载。随着上下文窗口不断扩展，预填充阶段面临关键瓶颈：长输入序列几乎激活所有专家，全量专家交换产生大量I/O流量，导致首词时延（TTFT）显著延长。为此，叶盛源等人的“面向边缘算力网络的混合专家模型协同推理系统”提出了一种面向边缘算力网络的协作MoE推理系统Sirius。该系统通过跨节点的协同调度汇聚聚合资源，加速预填充阶段并降低TTFT。在真实边缘算力网络平台上，针对Mixtral-8x7B和Qwen3-30B-A3B的实验表明，Sirius相比现有基线方法实现了1.2~4.0倍的推理吞吐量提升。

基于芯粒（Chiplet）架构的加速器成为支撑大规模深度神经网络（DNN）推理服务的关键设施。然而，并行多DNN负载的实时性与动态性特征给调度算法带来了严峻挑战，现有算法在决策开销、服务水平协议（SLA）保障及动态适应性方面存在局限。为此，王晨赳等人的“基于多智能体强化学习的多Chiplet加速器多DNN负载的在线调度”提出了一种面向多Chiplet加速器的多智能体强化学习多DNN负载在线调度算法（MAOSA），以最大化整体SLA满足率为目标。该算法将DNN任务划分为DNN块并建模为智能体，利用注意力机制并行生成包含执行顺序与PE级资源分配的调度策略，显著降低了决策开销；设计了考虑时间裕量的协作奖励函数以优化SLA，并引入周期性统一抢占机制来适应负载动态变化。实验结果表明，与当前最先进算法相比，MAOSA将SLA满足率提升了18%，系统吞吐量提升了31.9%，尾时延降低了28.7%，调度执行时间减少了约11倍。

在移动边缘计算环境中部署大语言模型面临端侧算力与存储受限、无线带宽紧张以及云端模型持续演进等挑战。现有基于投机解码的边云协同推理虽可借助边缘草稿模型与云端目标模型的并行协作降低生成时延，但通常要求二者紧密匹配。当云端模型频繁更新时，往往需要反复同步边缘模型，导致额外通信开销、时延增加，并削弱系统可扩展性。为此，李豫晨等人的“面向演进式大语言模型的带宽高效边云协同推理方法”提出了ECSpec通信高效边云协同推理框架。其核心在于共享骨干设计，使静态边缘草稿模型可适配一族不断演进的云端目标模型，从而避免端侧重复训练与权重反复下载，显著降低通信与维护成本。进一步结合时变信道与设备异构特性，设计了信道感知自适应投机机制，依据实时信道状态和能耗预算动态调整草稿长度，在效率与能耗之间实现平衡。实验结果表明，ECSpec在通信受限的边缘环境中可实现更稳定、高效的大语言模型协同推理。

面向AI的算力网是个极具理论价值和应用前景的研究方向，正吸引学术界与工业界日益广泛的关注。本专题所收录的工作从资源调度、推理加速、模型量化、协同推理到安全治理，

多角度地展现了算力网领域的最新研究成果。受篇幅所限，本专题难以全面呈现这一方向的全部最新进展，但我们希望所收录的工作能够为相关研究和实践提供参考与启发，进一步促进学术交流与创新发展。

在此，我们衷心感谢《计算机研究与发展》为本专题提供宝贵平台，感谢作者、审稿专家和编辑部老师为本专题顺利出版所付出的辛勤努力。谨向所有关心和支持本专题工作的专家学者致以诚挚谢意！

郭得科（中山大学）
顾荣（南京大学）
仇超（天津大学）
2026 年 7 月