

大模型推理中的存算技术 专题导读



专题策划人



郑纬民



陆游游

随着大模型的落地应用，大模型推理服务的部署规模持续扩大，对人工智能基础设施提出了更高的要求。智能体的多轮迭代、长程任务执行等特点，推动上下文长度进一步增加，这对系统的计算与存储能力提出了新的挑战。本期专题以“大模型推理中的存算技术”为主题，聚焦大模型推理的存算技术，邀请该领域的专家学者撰写了7篇文章。这些文章从键值（KV）Cache的复用与存储管理、近存计算架构与存算协同、大规模向量管理以及大模型推理在端侧具身场景下的应用等角度，对相关技术进行了深入讨论。本专题旨在梳理大模型推理场景中的存算技术研究进展，促进中国大模型推理系统方向的学术交流和思想碰撞，助力相关技术的创新与工程落地。

《BidKV：显存受限推理的KV缓存回收机制》提出了一种代价感知的显存KV回收机制BidKV，通过对可释放容量与抢占扰动进行效用信号编码，选择回收单位代价收益高的对象。该机制以轻量适配器接入了虚拟大型语言模型（vLLM）和SGLang主流框架，并在英伟达和昇腾等加速卡上验证了效果。

《存储层级感知的大模型前缀键值缓存系统》提出了一种面向显存—内存—固态硬盘（SSD）多层存储架构的大模型前缀键值存储系统PCache，介绍了访存代价感知的重要KV选择、页面级连续KV加载、动态KV重排和变序计算等技术，并通过实验验证了其能够在保证推理精度稳定的同时降低首Token时延。

《面向智能体时代大语言模型推理的新型存算解聚合架构》提出了一种基于光互联的以内存为中心的新型存算解聚合架构，以实现计算和存储资源的独立伸缩、池化共享和按需协同。该架构通过光互联技术连接计算侧和存储侧资源，利用了光互联低延迟、高带宽和可扩展能力，介绍了存储侧多层级管理和近存计算的技术，并通过仿真验证了架构的有效性。

《新架构大模型推理的存算特征演变与近存计算适配技术研究》分析了大模型架构中多头潜在注意力（MLA）和混合专家（MoE）等机制对硬件存算的影响。文章借助Roofline模型分别对MLA和MoE机制在算术与访存强度上进行了分析，讨论了已有近存计算的不足，并展望了MLA与MoE组合架构下的分层近存计算框架。

《面向大模型推理的存算协同理论综述》对大模型推理中的计算与存储之间的协同关系进行了综述，从以算换存、

以存换算和KV存储路径这3个角度展开，具体包括输入/输出（I/O）重组、压缩、裁剪与重算等以算换存技术，KV布局重用和推测解码等以存换算技术，以及多路径选择等KV存储路径技术，并展望了未来的存算协同技术。

《大规模向量存储系统研究综述》对大规模向量存储技术进行了综述，概述了内存图索引与聚类索引机制，分析了大规模向量场景下向量量化压缩、固态硬盘上索引、异构存算硬件加速等技术，讨论了向量检索与大模型检索增强的协同技术，总结并展望了大规模向量存储技术的未来发展。

《端侧具身智能系统：挑战与展望》对端侧具身智能的算法、系统和芯片优化进行了综述，介绍了端侧智能在步骤压缩、权重与激活精度量化、扩散特征缓存、分块异步等通用算法方面的优化，以及视觉—语言—动作模型（VLA）和世界—动作模型（WAM）两种模型的特定算法优化，对比了当前端侧具身推理系统的优劣，并展望了端侧智能系统的未来发展。

本期作者来自知名高校和企业，这些作者或参与国家科技重大专项、国家重点研发计划，或参与国家实验室、国家人工智能学院的智算集群相关工作，或负责企业前沿技术与系统的研发工作，均具有一线研发经验，从大模型系统的存

储管理、存算系统、向量存储和端侧具身智能等方面介绍了最新的研究成果。希望本期专题能为读者提供技术参考，也感谢所有作者和审稿专家对本期专题的支持！

策 划 人 简 介

郑伟民，清华大学计算机系教授、博士生导师，中国工程院院士，曾任中国计算机学会第十届理事长；长期从事并行/分布处理、大规模数据存储系统领域的科研与教学工作；曾获国家科技进步奖一等奖1次、国家科技进步奖二等奖2次、国家发明奖二等奖1次，2016年获何梁何利基金科学与技术进步奖，2016年获ACM戈登·贝尔奖，长期讲授计算机系统结构课程，2008年该课程被评为国家级精品课程；发表论文530余篇，拥有著作10部。

陆游游，清华大学计算机系长聘副教授、博士生导师，清华大学（计算机系）—软通计算机人工智能系统联合研究中心主任，青年北京学者，FAST、SOSP、OSDI、USENIX ATC、EuroSys等国际会议组织委员会或程序委员会委员；主要研究方向为存储与人工智能系统；曾获国家优秀青年科学基金项目资助，作为项目负责人主持了新一代人工智能国家科技重大专项项目、国家重点研发计划项目、国家自然科学基金重点项目等；6次获得最佳国际论文奖项，参与研制的高性能文件系统SuperFS部署于鹏城云脑II，2023年—2025年该系统多次蝉联世界超算存储IO500榜单第一名；发表论文80余篇。