计算机大模型的算力支撑技术研究

王冀远
杭州蚂蚁酷爱科技有限公司　浙江　杭州　310012

摘　要：计算机大模型算力需求呈现规模化、高强度持续性、不均衡性等特征，对高效运算、海量数据处理等有核心诉求，且算力支撑技术需满足多方面适配性要求。本文详细阐述了计算机大模型算力支撑核心技术体系，剖析了硬件、软件协同、大规模调度管理等方面的关键瓶颈，并针对性地提出了硬件算力优化、软件-硬件协同优化、大规模分布式算力支撑创新、能效与算力协同提升等优化路径与创新方向。

关键词：计算机大模型；算力支撑技术；硬件瓶颈；软件协同；能效优化

引言：随着计算机大模型在各领域的广泛应用，其规模不断扩大，参数数量呈指数级增长。大模型训练与推理过程对算力的需求愈发严苛，不仅需要强大的运算能力处理海量数据，还对算力的稳定性、可扩展性等提出高要求。在此背景下，深入研究计算机大模型的算力支撑技术至关重要，它是保障大模型高效运行、推动相关领域发展的关键所在，对提升整体技术水平和应用效果意义重大。

1 计算机大模型的算力需求特征与核心诉求

1.1 计算机大模型的算力消耗核心特征

计算机大模型的算力消耗呈现规模化增长态势，随模型参数规模、训练数据量的提升呈现指数级攀升趋势[1]。算力消耗具有高强度持续性，模型训练过程需长时间维持满负荷运算状态，运算间隙的算力闲置率极低。算力消耗存在明显的不均衡性，训练过程中不同运算环节对算力的需求差异较大，矩阵运算、特征提取等核心环节占用绝大部分算力资源，数据预处理、结果校验等环节算力消耗相对较低。算力消耗对稳定性要求极高，短暂的算力波动或中断会影响模型训练的连续性，导致训练周期延长、模型精度受损。算力消耗还具有高并行性需求，单一运算单元难以支撑大规模模型的高效运算，需依托多运算单元协同工作，实现算力的并行释放与高效利用。

1.2 大模型训练与推理的算力核心诉求

大模型训练的算力诉求集中在高效运算能力与海量数据处理能力两方面。训练过程需对海量标注数据进行反复迭代运算，需算力支撑快速完成数据读取、运算、存储的闭环流程，提升训练迭代速度。训练阶段对算力的精度要求较高，需精准支撑复杂的神经网络运算，减少运算误差对模型收敛效果的影响。大模型推理的算力诉求侧重低延迟与高吞吐量，推理过程需快速响应输入请求，输出精准结果，满足实际应用场景中的实时性需求。推理阶段需在保证运算精度的基础上，实现算力资源的高效调配，避免算力浪费，提升单位算力的推理效率。无论是训练还是推理，算力都需具备良好的可扩展性，能够根据模型规模、数据量的变化灵活调整，适配不同阶段的算力需求。

1.3 算力支撑技术与大模型的适配性核心要求

算力支撑技术与大模型的适配性需满足运算效率适配要求，算力支撑技术需匹配大模型的运算逻辑，优化运算架构，减少运算过程中的算力损耗，实现算力与模型运算的高效协同。适配性需满足规模弹性适配要求，算力支撑技术需具备灵活扩容、缩容能力，能够根据大模型参数规模、训练任务量的变化，动态调整算力资源供给，保障模型训练与推理的顺畅推进。适配性需满足精度适配要求，算力支撑技术需精准匹配大模型运算过程中的精度需求，在保障运算精度的前提下，实现算力资源的高效利用，平衡算力消耗与运算效果。适配性还需满足稳定性适配要求，算力支撑技术需具备高可靠性，能够长时间稳定运行，避免因技术故障导致算力中断，保障大模型训练与推理任务的顺利完成。

2 计算机大模型算力支撑核心技术体系

2.1 芯片算力支撑技术

通用计算芯片算力优化技术聚焦芯片运算架构升级，通过优化指令集设计，提升芯片对大模型运算任务的适配能力，降低运算过程中的冗余消耗，强化通用芯片在多场景下的算力输出效率[2]。专用加速芯片设计与架构优化围绕大模型核心运算需求展开，简化非必要运算模块，强化矩阵运算、特征处理等关键环节的算力支撑，打造贴合大模型运算逻辑的专用架构，提升算力输出密度。芯片间协同算力调度技术通过优化芯片间数据传输路径，协调多芯片运算节奏，实现多芯片算力的高效整合，避

免单芯片负载过高、多芯片算力脱节等问题，充分释放集群芯片的整体算力潜力。

2.2 算力存储支撑技术

高带宽存储架构设计技术优化存储层级布局，缩短存储与运算单元的距离，提升存储带宽，满足大模型运算过程中海量数据的高速传输需求，缓解存储瓶颈对算力释放的限制。存储与算力的协同适配技术实现存储速度与算力输出的精准匹配，根据算力运算节奏动态调整存储响应效率，避免存储速度滞后于算力运算或存储资源闲置等情况。海量数据高速读写支撑技术通过优化数据编码方式与读写算法，提升存储系统的数据读写速率，保障大模型训练与推理过程中数据读取的连续性，减少因数据读写延迟导致的算力浪费。

2.3 算力网络支撑技术

高带宽低延迟网络架构搭建优化网络节点布局，采用高速传输协议，降低网络传输延迟，提升网络带宽承载力，保障多运算节点、存储节点间的数据传输高效顺畅，为分布式算力集群提供稳定支撑。分布式算力网络协同技术协调不同节点的算力资源，实现分布式节点间的算力联动，打破地域与节点限制，整合分散算力形成规模化算力集群，支撑大规模大模型运算任务。网络传输与算力调度的协同优化结合算力负载变化，动态调整网络传输优先级，优先保障高负载运算节点的数据传输需求，实现网络资源与算力资源的协同高效利用，提升集群运算效率。

2.4 算力调度与管理技术

动态算力资源分配技术根据大模型训练与推理的实时需求，灵活分配算力资源，针对不同运算环节的算力需求差异，精准调配资源，提升算力资源利用率，避免资源浪费。算力负载均衡优化技术实时监测各运算节点的负载情况，将过高负载合理分流至空闲节点，避免部分节点过载、部分节点闲置，维持整体算力集群的负载均衡，保障算力稳定输出。算力资源虚拟化管理技术将物理算力资源虚拟化处理，整合分散的物理算力，形成可灵活调配的虚拟算力资源池，简化算力管理流程，提升算力资源的灵活适配能力，适配不同规模的运算任务。

2.5 能效优化支撑技术

算力节点能效调控技术实时监测算力节点的能耗状态，根据运算负载动态调整节点运行参数，在保障算力输出的前提下，降低节点能耗，减少运行成本。闲置算力回收与再利用技术识别算力集群中的闲置算力，通过合理调度将闲置算力分配至有需求的运算任务，减少算力浪费，提升整体算力资源的利用效率，实现资源高效配置[3]。高算力场景下能效平衡技术兼顾算力输出与能耗控制，优化高算力场景下的运算流程与节点运行模式，在保障大模型高效运算的基础上，实现算力输出与能效消耗的平衡，推动算力支撑技术绿色可持续发展。

3 计算机大模型算力支撑技术的关键瓶颈

3.1 硬件层面算力支撑瓶颈

硬件层面的算力支撑瓶颈主要体现在核心芯片的性能提升受限，芯片运算架构的升级速度难以跟上大模型参数规模的增长速度，算力输出效率的提升出现放缓趋势。芯片制造工艺的突破面临技术壁垒，高端芯片的产能不足，难以满足大规模大模型集群化运算的硬件需求。存储硬件与运算硬件的性能不匹配，存储带宽的提升速度滞后于算力提升速度，导致海量数据传输过程中出现卡顿，限制算力充分释放。硬件设备的能耗控制难度较大，高算力输出场景下硬件能耗大幅增加，既提升运行成本，也对硬件设备的稳定运行带来挑战，硬件集群的扩展成本较高，难以实现低成本、高效益的算力扩容。

3.2 软件与硬件协同适配瓶颈

软件与硬件协同适配不足形成明显瓶颈，软件层面的运算算法、调度程序多基于通用硬件设计，未能充分贴合专用加速硬件的架构特点，导致硬件算力无法充分发挥。软件更新迭代速度与硬件升级节奏不匹配，硬件性能提升后，对应的软件优化未能及时跟进，出现硬件资源闲置的情况。不同厂商的硬件设备与软件系统存在兼容性问题，缺乏统一的适配标准，导致多品牌硬件集群协同运行时，软件调度难度增加，影响整体算力输出效率。软件层面的优化聚焦单一环节，未实现与硬件运算、存储、网络等环节的深度协同，难以形成算力支撑的合力，进一步加剧适配瓶颈。

3.3 大规模算力调度与管理瓶颈

大规模算力调度与管理的瓶颈集中在调度算法的优化不足，面对分布式集群中的海量算力资源，调度算法难以实现资源的精准分配，无法根据不同运算任务的需求快速调整算力分配方案。算力负载的实时监测精度不足，难以精准捕捉各节点的负载波动，导致负载分流不及时，部分节点过载运行、部分节点闲置浪费。算力资源的虚拟化管理存在漏洞，虚拟算力资源与物理算力资源的匹配度不高，调度过程中出现延迟，影响算力响应速度。大规模算力集群的管理缺乏高效的统一管控机制，不同节点的管理标准不统一，运维难度较大，一旦出现故障，难以快速排查处理，影响算力支撑的稳定性，无法满足大模型长时间连续运算的需求。

4 算力支撑技术的优化路径与创新方向

4.1 硬件算力支撑技术优化方向

硬件算力支撑技术优化需聚焦核心芯片性能突破，升级芯片运算架构，优化指令集与运算单元设计，提升芯片算力输出效率，适配大模型参数规模增长需求[4]。突破芯片制造工艺技术壁垒，提升高端芯片产能，降低芯片生产成本，满足大规模集群化运算的硬件供给。优化存储硬件性能，提升存储带宽与读写速度，实现存储硬件与运算硬件的性能匹配，缓解数据传输卡顿问题。优化硬件设备能耗设计，研发低功耗硬件架构，降低高算力场景下的能耗损耗，平衡算力输出与能耗控制，降低硬件集群扩展成本，实现算力扩容的高效益与低成本兼顾。

4.2 软件-硬件协同算力优化路径

软件-硬件协同算力优化需立足专用加速硬件架构特点，优化运算算法与调度程序设计，充分发挥硬件算力潜力，避免硬件资源闲置。同步推进软件更新与硬件升级，根据硬件性能提升及时优化软件适配能力，形成软件与硬件协同迭代的良性循环。建立统一的软件-硬件适配标准，解决不同厂商设备与系统的兼容性问题，降低多品牌硬件集群的调度难度，提升整体算力输出效率。推动软件与硬件运算、存储、网络等环节的深度协同，打破单一环节优化的局限，形成算力支撑合力，缓解协同适配瓶颈。

4.3 大规模分布式算力支撑创新方向

大规模分布式算力支撑创新需优化调度算法设计，提升算法对海量算力资源的精准分配能力，根据运算任务需求快速调整算力分配方案，提升调度灵活性与高效性。提升算力负载实时监测精度，优化监测技术与手段，精准捕捉各节点负载波动，及时完成负载分流，避免节点过载与闲置浪费。完善算力资源虚拟化管理体系，优化虚拟算力与物理算力的匹配机制，降低调度延迟，提升算力响应速度。建立高效统一的算力集群管控机制，统一各节点管理标准，简化运维流程，提升故障排查与处理效率，保障算力支撑稳定性，满足大模型长时间连续运算需求。

4.4 能效与算力协同提升创新路径

能效与算力协同提升创新需研发智能能效调控技术，根据算力负载动态调整硬件运行参数，在保障算力输出的基础上最大限度降低能耗。创新闲置算力利用模式，优化闲置算力识别与调度机制，充分挖掘闲置算力潜力，提升算力资源整体利用效率[5]。优化高算力场景运算流程，创新节点运行模式，平衡算力输出与能耗消耗，实现能效与算力的同步提升。推动低功耗技术与算力支撑技术的深度融合，研发高效低耗的算力支撑方案，降低算力运行成本，实现算力支撑的可持续发展，兼顾高效算力输出与绿色低碳需求。

结束语

计算机大模型算力支撑技术研究是推动大模型发展的关键。面对硬件、软件协同及大规模算力调度管理等方面的瓶颈，需从硬件优化、软件-硬件协同、分布式算力创新及能效与算力协同提升等多方面发力。通过不断探索创新，突破现有技术限制，提升算力支撑能力，为计算机大模型在各领域的深入应用提供坚实保障，促进人工智能技术的持续进步与发展。

参考文献

[1]史天运,李国华,代明睿,等.铁路计算机视觉大模型研究[J].铁路计算机应用,2024,33(11):8-16.

[2]吴田军,骆剑承,李子琪,等.面向土地空间参数大规模计算的遥感大模型研究[J].遥感学报,2025,29(7):2305-2327.

[3]马春燕,陈晶,姚鼎,等.嵌入式智能计算机计算能力评测方法[J].计算机学报,2023,46(11):2279-2301.

[4]付慧敏,郑刚.基于计算机视觉和本体模型构建知识图谱的不安全作业识别[J].沈阳工业大学学报,2025,47(4):501-508.

[5]邢卫.计算机视觉驱动的多维数据智能分析模型研究[J].黑龙江科学,2025,16(12):130-133.