中国科学: 信息科学 2024 年 第 54 卷 第 12 期: 2905–2906
SCIENTIA SINICA Informationis

刊讯

大模型芯片与系统专题简介

尹首一[1*], 唐溪[2], 涂锋斌[3]

1. 清华大学, 北京 100084
2. 华东理工大学, 上海 200237
3. 香港科技大学, 香港 999077
* 通信作者. E-mail: yinsy@tsinghua.edu.cn

大模型给人工智能发展带来了历史性变革, 已经在机器翻译、人机交互、医学诊断、自动驾驶等智能任务中取得重大突破. 大模型巨大的计算量和参数量, 对芯片与系统的算力需求也急剧增加. 为了应对这一挑战, 学术界和工业界纷纷将目光投向大模型芯片与系统的研究, 以期突破传统计算架构的瓶颈, 实现更高性能解决方案以及更高效的设计方法. 大模型芯片与系统是推动人工智能技术进一步发展的关键技术之一, 被多个国家和地区列为未来科技发展的重要方向. *SCIENCE CHINA Information Sciences*在 2024 年 67 卷第 10 期组织出版了 “大模型芯片与系统专题” (Special Topic: AI Chips and Systems for Large Language Models), 共收录了 1 篇综述、1 篇观点和 4 篇研究论文, 涵盖了多项关键设计方向. 综述文章从高算力大模型芯片与系统的先进集成出发, 对基于芯粒的系统架构设计问题的国内外研究现状进行归纳与展望. 观点文章讨论了大模型技术对电子设计自动化（Electronic Design Automation, EDA）领域的影响, 提出大电路模型以应对高算力芯片在设计流程上的技术挑战. 研究论文以自顶而下的视角, 分别介绍了大模型芯片与系统在算法、编译、软硬件协同设计, 芯片架构设计等方面的最新成果. 这些研究描绘了大模型芯片与系统设计的主要挑战和解决方案.

基于芯粒的设计方法将系统芯片分解成多个较小的芯粒, 并通过先进封装重新组装成一个新的系统芯片, 是实现高算力大模型芯片与系统的重要技术. 这种方法在后摩尔定律时代备受关注, 其在成本、性能和敏捷设计方面具有明显的优势. 尽管芯粒设计作为一种新兴技术受到了广泛关注, 但仍然面临诸多挑战. 清华大学尹首一和李翔宇团队的综述 “Review of chiplet-based design: system architecture and interconnection” 从芯片设计者的角度全面综述了现有的芯粒设计, 对基于芯粒的系统架构设计问题的国际研究现状和最新进展归纳, 并给出关于发展趋势的分析. 该论文系统性地总结了芯粒设计的系统架构、互连拓扑和路由方案, 并对芯粒的发展趋势进行了展望, 如异构 MPSoC 设计将受益于芯粒设计方法、芯粒设计的平台化设计趋势, 以及封装 – 架构 – 互连的协同优化趋势等. 与以往专注于底层技术或某一单一领域的分析不同, 本综述采用了一种更全面、系统性的策略来研究这些方法, 旨在为设计人员提供系统性的、纵向比较的观点.

集成电路的 EDA 技术是全球范围内极具专业化和技术密集度的行业. 高性能芯片短生命周期和严格的性能、功耗、面积要求给 EDA 设计带来了挑战. 大模型技术的发展为解决这些挑战带来了新的机遇. 大模型能够分析大量设计案例, 在广阔的设计空间中进行预测和分析, 这有望对 EDA 工具产生革命性的影响. 尽管目前 “AI for

引用格式: 尹首一, 唐溪, 涂锋斌. 大模型芯片与系统专题简介. 中国科学: 信息科学, 2024, 54: 2905–2906, doi: 10.1360/SSI-2024-0322
Yin S Y, Tang Y, Tu F B. Special Topic: AI Chips and Systems for Large Language Models (in Chinese). Sci Sin Inform, 2024, 54: 2905–2906, doi: 10.1360/SSI-2024-0322

EDA" 已经成为 EDA 领域的研究热点之一, 但这些方法大多专注于单一任务的优化, 未能充分考虑电路设计的整体连贯性和多阶段复杂性, 限制了 "AI for EDA" 对设计流程的整体提升. 香港中文大学、北京大学、东南大学与华为等 11 家单位合作的观点文章 "Large circuit models: opportunities and challenges" 提出构建一个专为电路设计与优化的大模型: 大电路模型 (Large Circuit Model, LCM). 本文将 EDA 设计流程视为一个多模态转换过程: 从自然语言规格说明到模块化架构设计, 再到硬件描述语言编写的 RTL 代码, 直至最终的物理设计布局. 大电路模型将专注于解决 EDA 流程中最为重要的优化与验证问题, 推动 EDA 工具的能力向前迈进一大步.

为更好地从海量数据中学习特征分布, 神经网络的的计算开始呈现动态性, 即模型执行过程随输入数据变化而变化. 这种动态性为神经网络编译器的编译优化带来巨大挑战. 复旦大学尚笠和上海交通大学张宸团队的研究论文 "TSCompiler: efficient compilation framework for dynamic-shape models" 提出了面向动态神经网络的编译框架 TSCompiler, 采用基于参数的张量形状表示, 并将算子拆解、规约为基本算子, 基于基本算子语义构建数据流分析框架进行参数形式张量形状传播. 同时, TSCompiler 提出基于硬件架构约束构建的调度空间, 并采用基于信赖域的贝叶斯优化算法快速搜索高性能的调度组合生成张量程序. 相较于现有系统, TSCompiler 在编译生成张量程序的执行效率上有显著提升, 同时在端到端时延方面也表现出较大优势.

在边缘计算的快速发展中存内计算 (Computing-In-Memory, CIM) 技术以其卓越的能效比和计算并行性, 成为推动智能设备发展的关键力量. 然而, 面对边缘场景对运算精度、模型密度和能耗效率的不同需求, 如何实现在不同应用场景的神经网络架构定制化, 仍然是当前亟待解决的问题. 与此同时, 混合专家 (Mixture of Experts, MoE) 模型作为大模型算法的新兴研究方向, 其在边缘端部署仍待优化. 南方科技大学王中锐和中科院微电子所尚大山团队的研究论文 "CMN: a co-designed neural architecture search for efficient computing-in-memory-based mixture-of-experts" 提出了一种软硬件协同优化的神经网络架构搜索框架 CMN. 该框架以实现高效计算存储为目标, 通过软件层面的 MoE 模型设计与硬件层面的 CIM 系统配置的紧密协同, 在软件层面针对 MoE 结构的专家位置、数量和维度进行优化搜索, 在硬件层面针对存算阵列大小、外围电路设计等进行优化搜索, 两者构成嵌套搜索以适应各边缘场景下的不同需求.

目前已经有大量关于神经网络中非线性模块硬件架构设计的研究, 但大多局限于卷积神经网络和小规模的 Transformer, 对于当前热门的大模型中最常见的 softmax 和层归一化模块仍缺乏充分的研究. 特别是关于层归一化的硬件架构设计, 因其同时需要计算平方根与除法, 且近似层归一化对模型性能的影响大于 softmax, 导致其高效通用硬件架构设计成为难题. 上海交通大学贺光辉和徐宁仪团队的研究论文 "Hardware-oriented algorithms for softmax and layer normalization of large language models" 提出了面向硬件的近似 softmax 和层归一化算法, 然后设计出相应的高效硬件架构. 相较于已有的同领域最优设计, 本文提出的架构在面积和功耗方面具有显著节省. 这些设计无需微调, 几乎不影响模型性能, 能够作为大模型芯片的即插即用非线性单元, 有助于加速整体硬件设计与部署.

随着大规模预训练 Transformer 模型的迅速发展, 神经网络在自然语言处理领域取得了显著的进展. 然而, 这些模型的成功往往依赖巨大的计算资源和能源消耗, 因此, 开发低能耗、高效率的计算模型成为了研究者们关注的重点. 浙江大学唐华锦团队的研究论文 "SpikingMiniLM: energy-efficient spiking transformer for natural language understanding" 提出了一种适用于自然语言理解的脉冲 Transformer 模型 – SpikingMiniLM. 该模型通过一系列创新性的方法, 包括多步脉冲编码、改进的注意力机制和残差连接、稳定脉冲发放速率的参数初始化, 以及人工神经网络到脉冲神经网络的知识蒸馏技术, 克服了训练脉冲神经网络语言模型的挑战, 展示了脉冲神经网络在自然语言理解任务上的可行性. SpikingMiniLM 有效降低了自然语言理解任务所需的整体能源开销, 特别是在计算功耗上具有显著的优势.

以上是本期专题收录的全部文章, 衷心感谢所有为本专题撰稿的作者, 并诚挚感谢所有匿名审稿人对稿件及时且细致的评审工作.