中国科学: 信息科学 2021 年 第 51 卷 第 6 期: 1041–1042
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS

刊讯

近存/存内计算专题简介

黄如[1*], Yiran CHEN[2*], 蔡一茂[1*]

1. 北京大学, 北京 100871, 中国
2. Duke University, Durham NC 27708, USA
* 通信作者. E-mail: ruhuang@pku.edu.cn, yiran.chen@duke.edu, caiyimao@pku.edu.cn

高性能计算和数据处理芯片是软件算法功能实现的硬件载体, 软件算法和高性能计算芯片有着相互依存、相互促进的关系. 近年来, 集成电路蓬勃发展带来的算力提升对基于深度神经网络的智能应用取得巨大成功有着不可磨灭的贡献. 同时, 大数据和人工智能时代的来临也给传统的计算/存储分离的硬件架构带来了的存储墙及功耗墙挑战. 为了降低计算和存储单元间数据频繁交换带来的延迟和功耗, 以存储为中心的计算架构逐渐兴起. 近存计算和存内计算技术可以有效压缩计算单元和存储单元的时间与空间距离, 提高面向智能应用芯片的能效和性能, 受到了学术界和工业界广泛关注. 国际和国内学者报道了众多近存/存内计算技术领域的重要进展, 包括基于传统动态随机存储器 (DRAM) 等的近存计算和存内计算芯片, 以及基于阻变存储器 (ReRAM)、相变存储器 (PCM)、磁存储器 (MRAM) 等新型存储器件的近存/存内计算范式. 该领域已经逐渐发展到从单纯的器件研究向电路、架构、算法综合研究的新阶段. 研究人员开始更加关注器件、电路及系统稳定性与可靠性、新网络和新算法等对芯片的影响及对应的优化方法, 协同设计开始得到关注. 为了更好地将近存/存内计算技术的最新研究成果介绍给读者, *SCIENCE CHINA Information Sciences*在 2021 年第 6 期组织出版了 (Special Focus on Near-memory and In-memory Computing), 该专题共收录了 9 篇文章, 包括 2 篇综述、1 篇进展和 6 篇研究论文.

基于新存储技术的图计算和机器学习的加速器体系结构可以有效地减少加速器体系结构的数据流动来获得高性能和降低功耗. 南加州大学 Xuehai Qian 教授在综述文章 “Graph processing and machine learning architectures with emerging memory technologies: a survey” 中对基于新存储技术的图计算和机器学习的加速器体系结构的最新研究成果进行了全面的总结: 对于 3D 内存技术, 主要问题是如何将计算映射到分布式的内存模块和相应的计算单元, 从软硬件协同设计的角度减少通信; 对于可变电阻式存储器, 主要问题是如何将基本的计算映射到可以进行矩阵向量乘的交叉式内存控制器结构, 并对当前和未来的热点研究问题进行了展望.

存内计算范式可减少数据转移并提升计算效率. 基于自旋磁存储器 (MRAM) 的存内计算设计具有非易失、高能效、易于集成等特点. 东南大学 Jun Yang 教授团队与合作者在综述 “A survey of in-spin transfer torque MRAM computing” 一文中从存储单元、电路设计和系统协同优化 3 个方向综述了基于 MRAM 的存内计算技术, 解析了当前的设计挑战及潜在的解决方案.

人工智能处理器存在计算与存储单元之间的大规模数据移动, 给传统的计算架构带来了严重的存储墙及功耗墙挑战. 通过在计算芯片内部署大量存储单元, 让存储单元具有计算能力, 可以大幅减少

引用格式: 黄如, Yiran CHEN, 蔡一茂. 近存/存内计算专题简介. 中国科学: 信息科学, 2021, 51: 1041–1042, doi: 10.1360/SSI-2021-0176

或避免数据移动, 从而解决以上问题. 电子科技大学 Jun Zhou 教授团队在综述文章 “Energy-efficient computing-in-memory architecture for AI processor: device, circuit, architecture perspective” 中分析了现有人工智能处理器对数据移动和低功耗的需求, 介绍了存内计算的背景和实现方法, 分析了现有的基于数模混合的存内计算设计方案和其他设计方案, 总结了近几年存内计算芯片的研究情况, 并对存内计算未来的发展进行了讨论.

存内计算技术被认为是打破 “冯 · 诺依曼瓶颈” 的有效方法. 中科院计算所的 Yinhe Han 教授团队与合作者在研究进展 “Breaking the von Neumann bottleneck: architecture-level processing-in-memory technology” 中从体系结构的角度出发, 对基于 DRAM 的存内计算 (PIM) 技术所面临的主要挑战进行了总结和介绍. 分析了其在数据一致性、PIM 兼容性和 PIM 透明性等方面所面临的问题, 并对应介绍了当前的研究进展和解决这些问题的方法. 同时, 讨论了现有 PIM 模拟器的相对局限性, 以及 4 种传统的 PIM 仿真器. 最后, 展望了体系结构级 DRAM PIM 技术的前景和发展方向.

发现神经网络内部的连接方式对了解脑的工作机理及类脑计算的研究有重要的意义. 密歇根大学 Wei D. Lu 教授团队在论文 “Neural connectivity inference with spike-timing dependent plasticity network” 中提出一个基于二阶忆阻器的人工神经网络, 可以实时、高效地处理生物神经网络产生的数据并映射出生物神经网络内部的连接方式.

相比于传统计算芯片, 基于忆阻器的存算一体芯片在神经网络计算方面具有更高的能效, 然而存算一体芯片的计算精度往往受到忆阻器的非理想特性影响. 清华大学 Bin Gao 教授和 Huaqiang Wu 教授团队在论文 “Array-level boosting method with spatial extended allocation to improve the accuracy of memristor based computing-in-memory chips” 中提出一种阵列空间复制的识别率提升方法, 并针对神经网络不同层的识别率重要性不同的特点, 提出了贪婪空间分配算法. 实验结果表明, 该方法可以有效抑制忆阻器非理想特性的影响, 为提升忆阻器存算一体芯片的计算精度提供了新思路和方法.

基于 RRAM 的存算一体神经网络加速器有高并发、低功耗等优点, 但自身限制对所处理的神经网络提出一定要求. 一方面, 神经网络需要考虑模拟计算带来的噪声问题和数模转换等问题; 另一方面, 神经网络的权重数量受 RRAM 单元数量的限制, 权重和中间激活值需要进行量化以部署到硬件. 因此, 通常为 RRAM 存算一体加速器设计和部署的神经网络开销教大, 且模型性能较差. 北京大学 Guangyu Sun 教授团队在论文 “NAS4RRAM: neural network architecture search for inference on RRAM-based accelerators” 中提出针对基于 RRAM 加速器的神经网络搜索框架 —NAS4RRAM, 在设计约束下, 自动化地探索设计空间, 寻找符合约束的高性能神经网络, 从而显著提升设计效率.

西南大学 Shukai Duan 教授团队在论文 “Bayesian neural network enhancing reliability against conductance drift for memristor neural networks” 中针对忆阻器神经网络硬件实现中的电导漂移等导致网络精确度产生严重退化的问题, 提出了一种基于贝叶斯神经网络的权值优化方法, 将权值分布与电导偏差相关联, 使网络在权值发生较大改变时仍然能够维持高精度, 显著提高了网络的鲁棒性. 该方案可以集成到神经形态计算的硬件化中, 为其大规模应用提供了保障.

存内计算技术可以有效地降低内存访问延迟, 加速传统神经网络的推理运算. 然而新型的图卷积神经网络 (GCN) 混合了计算和内存访问特性, 限制了存内计算架构的访存优势. 深圳大学 Yi Wang 教授团队在论文 “Towards efficient allocation of graph convolutional networks on hybrid computation-in-memory architecture” 中面向异构存内计算架构, 提出了一种 GCN(graph convolutional networks) 的任务调度方法. 该调度方法能够有效地分配 GCN 中具有不同计算和存储资源需求的任务, 显著提高异构架构的利用率并降低 GCN 的推理延迟. 该工作为进一步加速图神经网络的处理效率提供了新的思路和技术途径.