

“面向高效人工智能的新型架构与电子系统”专栏

编者语

以深度学习和大模型为代表的人工智能技术正在从单一的算法创新迈向模型、数据与硬件系统协同演进的新阶段。随着模型规模持续增长、任务链条不断延长,以及人工智能由云端走向边缘端和物理世界,芯片算力、存储带宽、能效与实时性正成为制约智能技术进一步发展的关键因素。具身智能、低空无人机、可穿戴医疗、智能射频和科学计算等新兴应用,不仅要求电子计算机提供高吞吐量、高精度的智能计算能力,还对低功耗、低延迟、在线学习和环境适应能力提出了更加严苛的要求。

传统冯·诺依曼计算架构面临日益突出的“存储墙”和“功耗墙”问题,仅依赖工艺微缩和计算单元堆叠已难以持续满足高效人工智能的发展需求。近年来,面向高效人工智能的研究正由单点技术创新走向器件、电路、架构与系统的跨层协同:一方面,通过新型器件与新型计算范式提升计算密度和能效;另一方面,通过专用加速、异构系统集成及软硬件协同设计优化数据流和整体系统效率,为突破传统计算架构瓶颈提供了多层次技术路径。与此同时,人工智能本身也在深刻改变电子系统的设计方式:深度学习可用于复杂电磁结构优化,大语言模型开始参与硬件描述语言生成、电路优化、设计验证和电子设计自动化流程。

本专栏围绕“面向高效人工智能的新型架构与电子系统”这一主题,聚焦人工智能与电子设计、信息器件、计算架构及系统平台的交叉融合,收录7篇论文,内容涵盖人工智能赋能电子设计、存内计算与新型器件计算、高能效人工智能加速器以及数据驱动的异构电子系统等研究方向,展现了从算法、器件和电路到架构与系统的多层次创新。

在人工智能赋能电子设计方面,本专辑收录了大语言模型辅助电子设计自动化和深度学习驱动阵列优化两项研究工作。《大语言模型在电子设计自动化中的应用综述》系统梳理了大语言模型在寄存器传输级代码生成与优化、高层次综合与逻辑优化、后端实现与验证、模拟电路设计以及EDA流程自动化等方向的研究进展,并分析了领域数据不足、多模态信息处理、工具链集成、知识产权和数据安全等关键挑战,为构建面向集成电路设计的专业大模型和智能化EDA平台提供了系统参考。《基于MAAM-LKN模型的稀疏阵列双子阵RCS缩减方法》针对实际系统中关闭阵元雷达散射截面难以降为零的问题,引入关闭阵元RCS因子,构建工作子阵与关闭子阵联合优化模型,并结合多模态特征和增强注意力机制,实现稀疏阵列布局的高效预测与优化,体现了人工智能方法在复杂电磁系统设计中的应用潜力。

在存内计算与新型器件计算方面,专辑收录了面向组合优化和微分方程求解的两项研究工作。《基于多值阻变存储器与子句折叠策略的并行梯度布尔可满足性求解器》面向传统存内可满足性求解器阵列利用率低、存储空间开销大的问题,提出基于多值阻变存储器的

Fold-SAT 架构,通过子句折叠策略将多个子句压缩存储于同一阵列列中,并将该策略应用于并行梯度算法的前向与反向传播阵列,在保持并行求解能力的同时显著降低阵列规模,为复杂组合优化问题的高效硬件求解提供了新的实现路径。《面向高效微分求解电解质晶体管基储备池计算系统》利用铟镓锌氧化物电解质晶体管的非线性映射、短时记忆和低功耗特性,构建物理储备池计算系统,实现混沌常微分方程和偏微分方程的高效求解,并通过器件动力学建模、轻量化算法和 FPGA 系统验证,展示了新型器件计算在端侧科学计算中的应用前景。

在高效人工智能加速器方面,专辑收录了面向卷积神经网络推理和脉冲神经网络在线学习的两项研究工作。《基于层间特征图压缩的卷积神经网络加速器》针对中间特征数据传输带来的片上存储和片外带宽压力,提出统计感知的混合压缩方法,通过“离线分析—在线压缩”的软硬件协同机制,分别利用增强型零游程编码和动态基准—增量编码处理零值与非零值数据,在提高压缩率的同时降低硬件开销,为资源受限边缘设备上的高效神经网络推理提供了新的解决方案。《FF-STDP:一种基于前向-前向与 STDP 融合的低功耗在线学习 SNN 加速器》将前向-前向学习的局部监督机制与脉冲时序依赖可塑性相结合,构建无需误差反向传播的半监督在线学习方法,并设计事件驱动的 SNN 加速器,实现推理与学习的高效协同,为边缘人工智能系统获得低功耗、持续学习和环境适应能力提供了硬件基础。

在数据驱动的异构电子系统方面,《基于 DPU 的机载射频综合处理系统研究与设计》针对传统 FPGA 与 DSP 定制架构中计算资源与任务紧耦合、异构算力难以聚合以及系统扩展能力有限等问题,提出以数据处理单元为核心的机载射频综合处理架构。该系统通过高精度时钟同步、远程直接内存访问、数据直通和异构算力互联等机制,将射频接入、智能计算、并行计算、通用计算、可重构计算和分布式存储纳入统一算力网络,为高吞吐量、低时延和可扩展的智能射频系统提供了系统级实现路径。

本专栏从不同层次展示了面向高效人工智能的新型电子技术:既包括大语言模型和深度学习赋能电子设计的新方法,也涵盖多值阻变存储器、电解质晶体管和脉冲神经网络等新型计算范式;既关注数据压缩、在线学习和存内计算等核心架构,也延伸至异构算力互联和复杂电子系统集成。相关研究共同体现了从“器件—电路—架构—系统—设计方法”开展跨层协同创新的发展趋势。期望本专栏能够为相关领域的研究人员提供有益的学术参考,促进人工智能与电子技术的深度交叉融合,并推动我国人工智能硬件器件、芯片与系统的自主创新和协同发展。

客座编委: 杨玉超 北京大学
单伟伟 东南大学
涂锋斌 香港科技大学
王中锐 南方科技大学