综合信息
- 头条新闻
- 综合新闻
- 教育动态
- 合作交流
- 学术活动
- 图片新闻
- 业内信息
- 通知公告

当前位置 >> 首页 >> 综合信息 >> 综合新闻

综合新闻

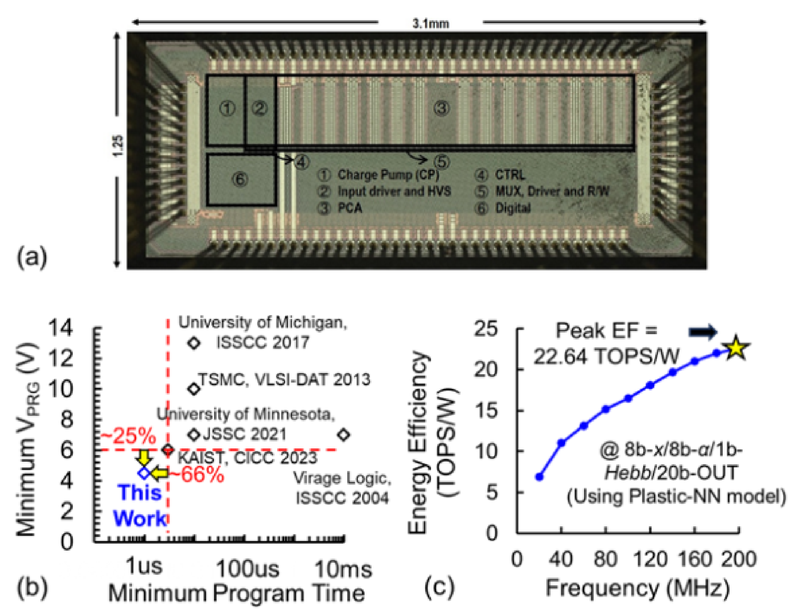
微电子所在片上学习存算一体芯片方面取得重要进展

稿件来源：重点实验室 王琳方、张康玮 发布时间：2024-02-26

当前，智能计算设备呈指数式增长，迫切需要低功耗与低延迟执行神经网络推理任务，以及不依赖云端的片上学习能力来动态适应边缘端复杂多变的应用场景。非易失存算一体技术可最大化减少数据搬运带来的功耗和延迟并消除静态功耗，为边缘智能计算设备提供了一种极具竞争力的方案。非易失存算一体芯片近年来持续快速发展，其在集成规模、能效、算力等方面均取得了长足的进步，实现了对片上推理任务的高效支持。但进行片上学习通常需要对模型参数进行海量次数的更新以及高精度的反向传播。受限于有限的擦写次数、较高的擦写功耗以及有限的计算精度，当前，非易失存算一体芯片仍然难以高效支持片上学习与训练。

针对以上问题，中国科学院微电子研究所刘明院士团队设计了基于非易失/易失存储融合型的片上学习存算一体宏芯片。团队在14nm FinFET工艺上验证了具有多值存储能力的5晶体管型逻辑闪存单元，编程电压（-25%）与编程时间（-66%）较同类型器件均获得有效降低；在此基础上，团队进一步提出了逻辑闪存单元与SRAM融合的新型阵列，不仅可以利用非易失与易失性存储单元的特点满足片上学习过程中长期与短期信息的存储，还能通过对矩阵-向量乘与矩阵元素乘的高效处理加速片上学习过程中所需的关键算子。团队还提出了一种与存储阵列深度融合的低硬件开销差分型模数转换电路，采用采样电容复用的方法节省面积，通过多元素稀疏感知的方案节省功耗。该芯片可以有效支持具有突触可塑性的神经网络，基于前馈过程动态更新短期信息，从而实现动态的片上学习。该存算一体宏芯片在14nm FinFET工艺下流片，可实现小样本学习等片上学习任务，8比特矩阵-矩阵-向量计算能效达到了22.64TOP/W。这一研究结果为基于存算一体架构的片上学习芯片提供了新思路。

近期，本工作以“A Flash-SRAM-ADC-Fused Plastic Computing-in-Memory Macro for Learning in Neural Networks in a Standard 14nm FinFET Process”为题发表在 ISSCC 2024国际会议上，微电子所博士生王琳方为第一作者、窦春萌研究员为通讯作者。参与本工作的主要研究人员还包括微电子所博士生李伟增以及硕士生周治道。该研究得到了科技部重点研发计划、国家自然科学基金、中国科学院战略先导专项等项目的支持。



14nm FinFET非易失-易失融合型片上学习存算一体芯片：（a）显微镜照片，（b）非易失逻辑闪存单元编程电压与时间对比，（c）工作频率与计算能效

中国科学院微电子研究所版权所有 邮编：100029
单位地址：北京市朝阳区北土城西路3号，电子邮件：webadmin@ime.ac.cn
京公网安备110402500036号，京ICP备05002781号-1