微电子所在28nm RRAM存内计算电路研究中获进展

2023-08-31 来源：微电子研究所

物联网与人工智能技术的发展对边缘节点计算平台的实时数据处理能力与能效提出了更高要求。基于新型存储器的非易失存内计算技术可实现数据的原位存储与计算，将数据搬运带来的功耗与延迟开销最小化，从而提升边缘设备的数据处理能力与效能比。然而，由于基础单元特性的非理想因素和阵列中的寄生效应以及模数转换电路的硬件开销，非易失存内计算面临着计算性能与能效方面的限制。中国科学院院士、微电子研究所研究员刘明团队采用跨层次协同设计的方法，提出了高并行与高效能比的新型RRAM存内计算结构。

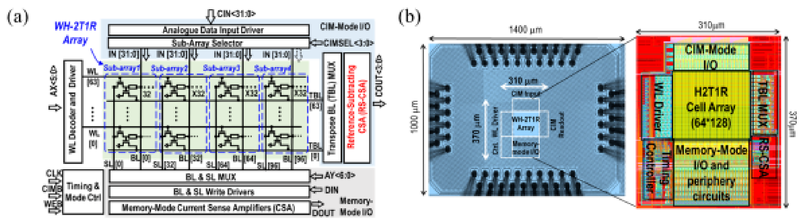
在器件层面，该研究提出了带权重二晶体管一忆阻器（WH-2T1R）的存算阵列结构。相较于1T1R结构，WH-2T1R结构使用core晶体管构成解耦的存算数据通路以减小寄生效应对计算电流的影响，只额外造成30.3%的面积开销。计算单元利用第二晶体管亚阈值区放大特性提高了计算13.5倍开关比的同时降低了88%的低阻态计算电流，从而实现了63.4%的乘加操作功耗降低。得益于计算开关比的提升，该RRAM存内计算结构可支持更高的输入并行度和多比特乘加操作。

在电路层面，该研究提出了参考电流减式电流型灵敏放大器的读出电路。由参考电流减支路根据上一次读出结果先对输入电流进行电流减再送到电流镜读出数据。参考电流减支路对半减小了电流镜输入电流范围，使RRAM存算结构支持的计算电流范围倍增，可实现更高输入并行和多比特乘加，并降低了79.5%的读出电路功耗。该研究通过进一步优化电流型灵敏放大器电流减配置，实现了积分非线性误差5倍提升以及微分非线性误差3.75倍提升。

在算法映射层面，该研究提出了高位数据冗余（MSB_RSM）的映射策略。RRAM存内计算结构配备不同的第二晶体管multiplier参数的多组阵列和额外的一组冗余阵列。不同的第二晶体管用于映射多比特权值的不同比特位。由于RRAM和晶体管非理想因素对计算电流的影响，冗余阵列用于额外映射权值对非理想因素补偿。科研人员经过对不同比特位补偿效果的分析发现，MSB-RSM对高位权值进行操作时可减小1σ误差40%。得益于更稳定的计算电流，在ResNet-18模型下的CIFAR-10和CIFAR-100任务的准确度提升了0.96%和2.83%。

上述方案在研究团队开发的嵌入式28nm工艺上得到验证。新型RRAM存内计算结构支持高并行的模拟域乘加操作，在1比特输入、3比特权值、4比特输出下ResNet-18任务中的平均能效达到了30.34TOPS/W，并可通过进一步优化将读出时序提升到154.04TOPS/W。该研究通过单元、电路以及系统面的系统设计，为高能效、高精度的模拟存内计算提供了新思路。

相关研究成果以*A 28nm RRAM Computing-in-Memory Macro Using Weighted Hybrid 2T1R Cell Array and Reference Subtracting Sense Amplifier for AI Edge Inference*为题，发表在《IEEE固态电路杂志》上。



（a）基于RRAM的存内计算芯片结构框图；（b）芯片晶圆照片及版图；（c）芯片测试数字波形；（d）芯片参数

责任编辑：侯茜