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

微电子所在数模混合存算一体芯片方面取得重要进展

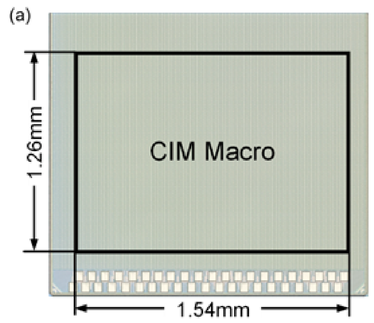
稿件来源：重点实验室 袁易扬、张康玮 发布时间：2024-02-26

当前，基于边缘智能计算设备运行的人工智能应用日趋复杂及高精度，为降低边缘设备运行的延迟和功耗，存算一体技术被应用在边缘设备端，通过减小数据搬运的开销最大化减少边缘设备上的延迟与功耗。但传统的存算一体宏仅支持使用整数型数据计算，难以支持日趋高精度、高复杂度以及片上训练的边缘端智能计算任务。且仅使用单一模拟或数字方案的存算一体宏，在能量效率、面积效率和精度上难以取得最优化。如何有效结合模拟存算与数字存算模式优势，在总体上取得更高的能量效率和面积效率，同时尽可能保证高精度，以及如何探索数模混合方案的设计空间，仍然是存算一体宏领域继续解决的问题。

针对以上问题，中国科学院微电子研究所刘明院士团队研发出基于外积运算的数模混合存算一体宏芯片，设计了一种数模混合浮点 SRAM 存内计算方案，提出了模拟与数字存算宏的混合方法，结合了使用模拟存算方案进行高效阵列内位乘法和使用数字存算方案进行高效阵列外多位移位累加的优点，达到整体上高能量效率与面积效率。通过残差式数模转换器架构，使数模转换器所需分辨率仅为输入位精度的对数，实现了高吞吐率和低开销。通过基于矩阵外积计算数学原理的浮点/定点存算块架构，矩阵-矩阵-向量计算可通过累加器元件完成。同之前的数字存算方案使用矩阵内积原理的大扇入、多级加法器树相比，吞吐率更高。该架构还支持细粒度的非结构激活稀疏性以进一步提升总体能效。该存算一体宏芯片在28nm CMOS工艺下流片，可支持BF16浮点精度运算以及INT8定点精度运算，BF16浮点矩阵-矩阵-向量计算峰值能效达到了72.12TFLOP/W，INT8定点矩阵-矩阵-向量计算峰值能效达到了111.17TFLOP/W。这一研究结果为采用数模混合方案的存算一体架构芯片提供了新思路。

近期，本工作以“A 28nm 72.12TFLOPS/W Hybrid-Domain Outer-Product Based Floating-Point SRAM Computing-in-Memory Macro with Logarithm Bit-Width Residual ADC”为题发表在 ISSCC 2024国际会议上，微电子所博士生袁易扬为第一作者，张锋研究员与北京理工大学王兴华教授为通讯作者。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。

(a)



* Performance in 90% input sparsity (ResNet-18)

** Using ResNet-18 model, the software baseline was 75.31%

(b)

Chip Summary		
Technology(nm)	28	
Macro Area(mm^2)	1.94	
SRAM Capacity(Kb)	192Kb	
Supply Voltage(V)	0.7-0.95	
Number of Input Channels	256	
Input Precision	BF16	INT8
Weight Precision	BF16	INT8
Throughput	1.98-4.28 TFLOPS	2.89-5.31 TOPS
Energy Efficiency (Average performance)	16.55-32.78 TFLOPS/W	22.78-50.53 TOPS/W
Energy Efficiency* (Peak performance)	36.41-72.12 TFLOPS/W	50.12-111.17 TOPS/W
Accuracy Loss** (@Cifar100)	-0.05%	-1.57%

28nm 基于外积的数模混合浮点存算一体宏芯片：（a）芯片显微镜照片，（b）芯片特性总结表

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