
微电子所等在高能效浮点存内计算方面取得进展

作者：writer 来源：中国科学院

本文原地址：<https://www.iikx.com/news/progress/22259.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

微电子所等在高能效浮点存内计算方面取得进展

。近年来，存内计算架构成为人工智能加速芯片的重要研究方向。存内计算架构能够降低数据访问开销，显著提升神经网络等智能算法的能效。目前，针对存内计算芯片的研究集中在定点存内计算领域，现有浮点存内计算的实现方案采用近存电路、指数-底数分离、浮点-定点转换等方式，面临较低并行度或较多浮点运算周期的挑战，难以实现高能效和高性能。考虑到更大规模网络模型和复杂任务的需求以及浮点运算在神经网络训练过程中的必要性，浮点运算功能对于未来存内计算芯片非常重要。

针对高能效浮点存内计算的挑战，中国科学院院士、中科院微电子研究所研究员刘明团队与清华大学教授刘勇攀团队合作，提出了“稠密存内+稀疏数字”的新型混合存内计算架构。研究发现，神经网络各层的数据分布具有长尾特性，且其中大部分数据的指数位密集分布在一个较小的区间内。基于此，研究提出了将高能效存内计算核心用于执行密集指数位的数据运算，将高灵活度稀疏数字核心用于执行具有长尾特性的稀疏指数位的数据运算。高效的浮点-定点转换电路、灵活编码的稀疏数字核心以及节省高位计算的数字存内计算加法器树电路进一步提升了该设计方案的能效。该SRAM存内计算芯片在28nm工艺下流片，在4比特定点和16比特浮点情况下，存内计算核心的稠密网络峰值能效分别达到275TOPS/W和17.2TOPS/W，稀疏网络峰值能效分别达到1600TOPS/W和90TOPS/W。该成果有助于推动SRAM存内计算芯片在高准确率浮点神经网络及神经网络训练等方面的应用。

相关研究成果以 *A 28nm 16.9-300TOPS/W Computing-in-Memory Processor Supporting Floating-Point NN Inference/Training with Intensive-CIM Sparse-Digital Architecture* 为题，入选固态电路领域顶级会议ISSCC 2023。

图2.28nm高能效浮点存内计算芯片

研究团队单位：微电子研究所

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发