EE 241B Spring 2020 Project

Compute-In-Memory SRAM Design for Accelerating Low-Precision Neural Network

Chuyao Cheng
Xuan Su

February 29, 2020

1 Abstract

With data volume growing exponentially and the popularities of today's edge devices, power and performance limits are being reached with frequent data movement in and out the memory system. There has been a lot of active research focusing on bringing computing as close as possible to the memory array such as, Compute-In-Memory(CIM) and Compute-Near-Memory Computing (CNM). CIM techniques focus on reducing data movement by integrating compute elements within or near the memory primitives in order to relax memory throughput, increase the performance of the applications and reduce energy.

Referencing state-of-the-art implementations of CIM memory, we aim to design a small sized CIM SRAM for faster MAC operation. As nowadays people are trying to develop smaller and more efficient networks to fit in edge devices such as phones, this technology could be very beneficial for fast and low-precision networks.

2 Reference

[1]. N. Jao, A. K. Ramanathan, S. Srinivasa, S. George, J. Sampson, and V. Narayanan, "Harnessing Emerging Technology for Compute-in-Memory Support," 2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2018. https://ieeexplore.ieee.org/abstract/document/8429408

[2]. A. Agrawal, A. Jaiswal, D. Roy, B. Han, G. Srinivasan, A. Ankit, and K. Roy, "Xcel-RAM: Accelerating Binary Neural Networks in High-Throughput SRAM Compute Arrays," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 66, no. 8, pp. 3064–3076, 2019. https://ieeexplore.ieee.org/abstract/document/8698312

[3]. R. Gauchi, M. Kooli, P. Vivet, J.-P. Noel, E. Beigne, S. Mitra, and H.-P. Charles, "Memory Sizing of a Scalable SRAM In-Memory Computing Tile Based Architecture," 2019 IFIP/IEEE 27th International Conference on Very Large Scale Integration (VLSI-SoC), 2019. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8920373

[4] Jia, H., Tang, Y., Valavi, H., Zhang, J., Verma, N. (2018). A microprocessor implemented in 65nm CMOS with configurable and bit-scalable accelerator for programmable in-memory computing. arXiv preprint arXiv:1811.04047. https://arxiv.org/abs/1811.04047

[5] Dong, q., Ersin Sinangil, M., Erbagci, B., Sun, D., Khwa, W., Liao, H., Wang, Y. and Chang, J. (2020). A 351 TOPS/W and 372.4 GOPS Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine Learning Applications. In: International Solid-State Circuits Conference. IEEE.