True In-memory Computing with the CRAM: From Technology to Applications

Masoud Zabihi[1], Zhengyang Zhao[1], Zamshed I. Chowdhury[1], Salonik Resch[1], Mahendra DC[2], Thomas Peterson[2], Ulya R. Karpuzcu[1], Jian-Ping Wang[1], Sachin S. Sapatnekar[1]
[1] Department of Electrical and Computer Engineering [2] School of Physics and Astronomy
University of Minnesota, Minneapolis, MN (USA)

ABSTRACT

Big data applications are memory-intensive, and the cost of bringing data from the memory to the processor involves large overheads in energy and processing time. This has driven the push towards specialized accelerator units that can perform computations close to where the data is stored. Two approaches have been proposed:

- *near-memory computing* places computational units at the periphery of memory for fast data access.
- *true in-memory computing* uses the memory array to perform computations through simple reconfigurations.

Although there has been a great deal of recent interest in the area of in-memory computing, most solutions that are purported to fall into this class are really near-memory processors that perform computation near the edge of memory arrays/subarrays rather than inside it. We discuss a true in-memory computation platform in this presentation, the Computational Random Access Memory (CRAM). The CRAM enables this capability by making a small modification to a standard spintronics-based memory array. The CRAM-based approach is digital, unlike prior analog-like in-memory/near-memory solutions, which provides more robustness to process variations, particularly in immature technologies than analog schemes.

Our solution is based on spintronics technology, which is attractive because of its robustness, high endurance, and its trajectory towards fast improvement [2, 4]. The outline of the CRAM approach was first proposed in [3], operating primarily at the technology level with some expositions at the circuit level. The work was developed further to show system-level applications and performance estimations in [1] based on a spin-transfer-torque (STT) magnetic tunnel junction (MTJ). Next, in [5], a bridge was built between the two to provide an explicit link between CRAM technology, circuit implementations, and operation scheduling. Most recently, in [6], a redesigned CRAM was designed around a new MTJ based on the spin-Hall effect (SHE), providing greatly improved energy efficiency.

This talk provides an overview of several years of effort in developing the CRAM concept and surveys all of these efforts. The presentation covers alternatives at the technology level, followed by a description of how the in-memory computing array is designed, using the basic MTJ unit and some switches, to function both as a memory and a computational unit. This array is then used to build gates and arithmetic units by appropriately interconnecting memory cells, allowing high degrees of parallelism. Next, we show how complex arithmetic operations can be performed through appropriate scheduling (for adders, multipliers, dot products) and data placement of the operands. Finally, we demonstrate how this approach can be used to implement sample applications, such as neuromorphic inference engine and a 2D convolution, presenting results that benchmark the performance of these CRAMs against near-memory computation platforms. The performance gains can be atrributed to (a) highly efficient local processing within the memory, and (b) high levels of parallelism in rows of the memory.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
GLSVLSI '19, May 9–11, 2019, Tysons Corner, VA, USA
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6252-8/19/05…$15.00
https://doi.org/10.1145/3299874.3319451

CCS CONCEPTS

• **Computer systems organization → Special purpose systems**; • **Hardware → Non-volatile memory**; **Arithmetic and datapath circuits**; **Spintronics and magnetic technologies**.

KEYWORDS

Spintronics, In-memory computing, CRAM, Memory bottleneck, STT-MRAM, SHE-MRAM, Spin-Hall effect, Neuromorphic computing, Adders, Multipliers, Nonvolatile memory, Convolution

ACM Reference Format:
Masoud Zabihi[1], Zhengyang Zhao[1], Zamshed I. Chowdhury[1], Salonik Resch[1], Mahendra DC[2], Thomas Peterson[2], Ulya R. Karpuzcu[1], Jian-Ping Wang[1], Sachin S. Sapatnekar[1]. 2019. True In-memory Computing with the CRAM: From Technology to Applications. In *Great Lakes Symposium on VLSI 2019 (GLSVLSI '19), May 9–11, 2019, Tysons Corner, VA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3299874.3319451

ACKNOWLEDGMENTS

This work was supported in part by the DARPA Non-Volatile Logic program, NSF SPX Award CCF-1725420, and by C-SPIN, one of the six SRC STARnet Centers, sponsored by MARCO and DARPA.

REFERENCES

[1] Z. Chowdhury, S. K. Khatamifard, M. Zabihi, J. D. Harms, Y. Lv, A. P. Lyle, J. P. Wang, S. Sapatnekar, and U. Karpuzcu. 2017. Efficient in-memory processing using spintronics. *IEEE Computer Architecture Letters* 17, 1 (Jan–Jun 2017), 42–46.
[2] A. Hirohata, H. Sukegawa, H. Yanagihara, I. Zutic, T. Seki, S. Mizukami, and R. Swaminathan. 2015. Roadmap for emerging materials for spintronic device applications. *IEEE Transactions on Magnetics* 51, 10 (Oct. 2015), 1–11.
[3] J. P. Wang and J. D. Harms. 2015. General structure for computational random access memory (CRAM). https://www.google.com/patents/US9224447 US Patent 9,224,447 B2.
[4] J. P. Wang, S. S. Sapatnekar, C. H. Kim, P. Crowell, S. Koester, S. Datta, K. Roy, A. Raghunathan, X. S. Hu, M. Niemier, A. Naeemi, C.-L. Chien, C. Ross, and R. Kawakami. 2017. A pathway to enable exponential scaling for the beyond-CMOS era. In *Proceedings of the ACM/ESDA/IEEE Design Automation Conference.* ACM, New York, NY.
[5] M. Zabihi, Z. Chowdhury, Z. Zhao, U. R. Karpuzu, J.-P. Wang, and S. S. Sapatnekar. 2018. In-memory processing on the spintronic CRAM: From hardware design to application mapping. *IEEE Transactions on Computers* (2018). https://doi.org/10.1109/TC.2018.2858251 (currently in IEEEXplore Early Access).
[6] M. Zabihi, Z. Zhao, Z. Chowdhury, M. Resch, T. Peterson, Mahendra DC, J.-P. Wang, U. Karpuzcu, and S. S. Sapatnekar. 2019. Using spin-Hall MTJs to build an energy-efficient in-memory computation platform. In *Proceedings of the IEEE International Symposium on Quality Electronic Design.* IEEE, Piscataway, NJ.