



In-Memory Computing: The Emerging Computing Topic in the Post-von Neumann Era

Paolo Montuschi^{ID}, Polytechnic University of Turin

Yuan-Hao Chang^{ID}, Academia Sinica

Vincenzo Piuri^{ID}, University of Milan

In-memory computing is a paradigm to break the increasing gap between the processor and memory speeds by performing computation inside or near the memory. Here, we aim to stimulate the curiosity of readers toward this new, emerging area.

The von Neumann architecture has been the status quo for many decades. Computers built on the von Neumann architecture store programs and data in memory and load them to

the processor for computations. This architecture can efficiently process applications requiring complex computational operations with adequate data size. Nevertheless, the behaviors of some learning algorithms and data-intensive applications are entirely different from the past. They need to fetch numerous data and perform simple operations (for example, multiplication and addition). Therefore, the connection bus between the processor and memory of the von Neumann architecture becomes the bottleneck while processing the learning algorithms and data-intensive applications.


In recent years, the increasing gap between the processor and memory speeds and the skyrocketing amount of data in artificial intelligence/machine learning applications have caused the von Neumann architecture to have severe performance and power issues. To tackle these issues, in-memory computing, rising to the top of the material advancements and architecture innovations, has emerged to



break the von Neumann bottleneck by performing computations right inside or near the memory, thus opening up the post-von Neumann era.

In the post-von Neumann era, in-memory computing designs and technologies have been developed as accelerators to unite computation and data storage together, so as to resolve the severe performance and power issues in conventional von Neumann machines. Nonetheless, in-memory computing designs and architectures are facing a lot of new challenges in real-world applications ranging

for Real-World Applications,” edited by Yuan-Hao Chang and Vincenzo Piuri.¹ *TETC* accepted 11 articles, which appeared in the second issue of 2023 (<https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=10144811&punumber=6245516>). These articles cover four broad categories, including crossbar-based processing-in-memory designs, hardware solutions for memory-centric designs, architectural exploration for neural network acceleration with in-memory computing, and technologies considering memory-related issues.^{2,3,4,5,6,7,8,9,10,11,12}

and prospective authors to regularly check our calls for papers at <https://www.computer.org/csdl/journal/ec> and submit their work either to the ongoing special sections or to the journal’s technical tracks (<https://www.computer.org/digital-library/journals/ec/technical-tracks>). 

Systems and applications need to reconsider how to partition data and offload computation to the in-memory computing accelerators.

from Internet of Things to data center applications and at all levels of computer systems ranging from circuit levels to system levels. For example, the nonvolatile memory (NVM)-based crossbar design is ideal for computing vector-matrix multiplication and accelerating the neural network computation; nonetheless, due to the arbitrarily drifting of the current and resistance in memory circuits, the NVM materials [for example, resistive random access memory and 3D NAND flash] that conduct computations in analog reveal inherent imperfections, leading to inaccurate computations and serious scalability issues. Furthermore, systems and applications need to reconsider how to partition data and offload computation to the in-memory computing accelerators.

In response to the urgent need and research trend on tackling the issues of the emerging in-memory computing technology, in 2022, *IEEE Transactions on Emerging Topics in Computing (TETC)* launched an invited thematic section, “Memory-Centric Designs: Processing-in-Memory, In-Memory Computing, and Near-Memory Computing

The increasing interest in in-memory computing by *TETC*’s served communities has led the launch of another call for articles for an open special section, “Emerging In-Memory Computing Architectures and Applications,” edited by Alberto Bosio, Nima TaheriNejad, and Deliang Fan. This special section aims at promoting in-memory computing and its applications as a promising solution. It mainly focuses on the memory wall and power wall issues of emerging computer architectures and the reliability wall, leakage wall, and cost wall of the emerging device technologies. Several accepted papers are already available in the preprint section of *IEEE Xplore*.

TETC is a leading venue and reference point for emerging and impacting topics in computing, such as in-memory computing with its new opportunities and challenges, either through dedicated special sections or technical track submissions. As editors of *TETC*, we invite all researchers to visit *TETC*’s gateway to *IEEE Xplore* at <https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6245516>

REFERENCES

1. Y.-H. Chang and V. Piuri, “Guest editorial: IEEE transactions on emerging topics in computing thematic section on memory-centric designs: Processing-in-memory, in-memory computing, and near-memory computing for real-world applications,” *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 278–280, Apr./Jun. 2023, doi: 10.1109/TETC.2023.3267909.
2. A. Ding, Y. Qiao, and N. Bagherzadeh, “BNN an ideal architecture for acceleration with resistive in memory computation,” *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 281–291, Apr./Jun. 2023, doi: 10.1109/TETC.2023.3237778.
3. S. Ma, D. Brooks, and G.-Y. Wei, “A binary-activation, multi-level weight RNN and training algorithm for ADC-/DAC-free and noise-resilient processing-in-memory inference with eNVM,” *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 292–302, Apr./Jun. 2023, doi: 10.1109/TETC.2023.3241004.
4. B. K. Joardar, J. R. Doppa, H. Li, K. Chakrabarty, and P. P. Pande, “ReL-Prune: ReRAM crossbar-aware lottery ticket pruning for CNNs,” *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 303–317, Apr./Jun. 2023, doi: 10.1109/TETC.2022.3223630.
5. Y. Li et al., “A survey of MRAM-centric computing: From near memory to in memory,” *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 318–330, Apr./Jun. 2023, doi: 10.1109/TETC.2022.3214833.

6. B. Wu et al., "An energy-efficient computing-in-memory (CiM) scheme using field-free spin-orbit torque (SOT) magnetic RAMs," *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 331–342, Apr./Jun. 2023, doi: 10.1109/TETC.2023.3237541.

7. M. Hossain, A. Tatulian, S. Sheikhfaal, H. R. Thummala, and R. F. DeMara, "Scalable reasoning and sensing using processing-in-memory with hybrid spin/CMOS-based analog/digital blocks," *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 343–357, Apr./Jun. 2023, doi: 10.1109/TETC.2022.3212341.

8. M. Rios, F. Ponzina, A. Levisse, G. Ansaloni, and D. Atienza, "Bit-line computing for CNN accelerators co-design in edge AI inference," *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 358–372, Apr./Jun. 2023, doi: 10.1109/TETC.2023.3237914.

9. A. Balaji, P. K. Huynh, F. Catthoor, N. D. Dutt, J. L. Krichmar, and A. Das,

"NeuSB: A scalable interconnect architecture for spiking neuromorphic hardware," *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 373–387, Apr./Jun. 2023, doi: 10.1109/TETC.2023.3238708.

10. N. M. Ghiasi et al., "ALP: Alleviating CPU-memory data movement overheads in memory-centric systems," *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 388–403, Apr./Jun. 2023, doi: 10.1109/TETC.2022.3226132.

11. W. Qiao, L. Guo, Z. Fang, M.-C. F. Chang, and J. Cong, "TopSort: A high-performance two-phase sorting accelerator optimized on HBM-based FPGAs," *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 404–419, Apr./Jun. 2023, doi: 10.1109/TETC.2022.3228575.

12. G. Papadimitriou and D. Gizopoulos, "Anatomy of on-chip memory hardware fault effects across the

layers," *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 420–431, Apr./Jun. 2023, doi: 10.1109/TETC.2022.3205808.

PAOLO MONTUSCHI is a professor of computer engineering at the Polytechnic University of Turin, Turin, Italy. He is a Fellow of IEEE. Contact him at tetc-eic@computer.org.

YUAN-HAO CHANG is a research fellow at Academia Sinica, Taipei, Taiwan. He is a Fellow of IEEE. Contact him at johnson@iis.sinica.edu.tw.

VINCENZO PIURI is a professor of computer engineering at the University of Milan, Milan, Italy. He is a Fellow of IEEE. Contact him at vincenzo.piuri@unimi.it.

IEEE COMPUTER SOCIETY
Call for Papers

Write for the IEEE Computer Society's authoritative computing publications and conferences.

GET PUBLISHED
www.computer.org/cfp