# 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

THE von Neumann architecture has been the status quo since the dawn of modern computing. Computers built on the von Neumann architecture are composed of an intelligent master processor (e.g., CPU) and dumb memory/storage devices incapable of computation (e.g., memory and disk). However, the skyrocketing data volume in modern computing is calling such status quo into question. The excessive amounts of data movement between processor and memory/storage in more and more real-world applications (e.g., machine learning and AI applications) have made the processor-centric design a severe power and performance bottleneck. The diminishing Moore's Law also raises the need for a memory-centric design, which is rising on top of the recent material advancement and manufacturing innovation to open a paradigm shift. By doing computation right inside or near the memory, the memory-centric design promises massive throughput and energy savings.

Although many memory-centric designs and technologies have been proposed to resolve severe power and performance bottleneck in traditional processor-centric designs, there are still a lot of new challenges for adopting memory-centric designs in real-world applications. These applications are ranged from IoT to data-center applications, and applications in different domains usually have diversified requirements and constraints. In addition to the challenges brought by the diversified real-world applications, the new memory-centric designs also create challenges to the designs at multiple levels of computer systems, ranging from circuit/device levels to architecture and system levels. Thus, there is an urgent need for technology, innovation, modeling, analysis, design, and application, ranging from circuit/device level to architecture/system level and application level.

This Thematic Section accepted 11 articles and involved numerous reviewers who were selected for their expertise on the precise topics of each manuscript. Thus, it represents a collective effort from the research community on an international scale. The manuscripts appearing in this Thematic Section tackle some of the most recent and impactful memory-centric design issues spanning levels of design abstraction, ranging from circuit/device level to architecture/system level and application level. The works address different issues related to memory-centric designs and can be grouped in four broad categories:

1) crossbar-based processing-in-memory designs,
2) hardware solutions for memory-centric designs,
3) architectural exploration for neural network acceleration,
4) technologies considering memory-related issues.

Considering *crossbar-based processing-in-memory (PIM) designs*, the article "BNN an Ideal Architecture for Acceleration with Resistive in Memory Computation" by Andrew Ding, Ye Qiao, and Nader Bagherzadeh provides a case study to deploy trained binary neural network (BNN) on crossbar-based processing-in-memory architectures. This work proposes a BNN with binarized weights, which are ideally mapped to fewer memristive devices, so as to achieve higher tolerance to the computational noise of crossbar arrays. Meanwhile, the article "A Binary-Activation, Multi-Level Weight RNN and Training Algorithm for ADC-/DAC-Free and Noise-Resilient Processing-in-Memory Inference with eNVM" by Siming Ma, David Brooks, and Gu-Yeon Wei further proposes a training algorithm, i.e., a new noisy neuron annealing (NNA) algorithm, to enable an ADC-/DAC-free PIM design with binary activation and multi-level weight to achieve high inference accuracy. Finally, the article "ReaLPrune: ReRAM Crossbar-aware Lottery Ticket Pruning for CNNs" by Biresh Kumar Joardar, Janardhan Rao Doppa, Hai (Helen) Li, Krishnendu Chakrabarty, and Partha Pratim Pande developes a crossbar-aware pruning strategy, called RealPrune, to reduce the crossbar demand without accuracy drop by taking the crossbar structure into considerations.

Considering *hardware solutions for memory-centric designs*, the article "A Survey of MRAM-Centric Computing: From Near Memory to In Memory" by Yueting Li, Tianshuo Bai, Xinyi Xu, Yundong Zhang, Bi Wu, Hao Cai, Biao Pan, and Weisheng Zhao provides a comprehensive survey regarding MRAM-based circuit designs to support memory-centric computing. It

Digital Object Identifier 10.1109/TETC.2023.3267909

outlines the background, trends, and challenges involved in the development of magnetic random-access memory-centric computing and highlights its recent prototypes and advances in applications. The article "An Energy-Efficient Computing-in-Memory (CiM) Scheme using Field-Free Spin Orbit Torque (SOT) Magnetic RAMs" by Bi Wu, Haonan Zhu, Zhaohao Wang, Ying Wang, Ke Chen, Weiqiang Liu, Xiaobo Sharon Hu, and Fabrizio Lombardi proposes an FF-SOT MRAM-based computing-in-memory scheme, which supports efficient XNOR/XOR logic operations and cascading adder with extensive simulations. The results show that the proposed FF-SOT-CiM achieves up to 3.1x (2.6x) latency (energy) reduction compared to SRAM-based CiM, with negligible hardware overhead when performing in-memory XOR. The article "Scalable Reasoning and Sensing Using Processing-in-Memory with Hybrid Spin/CMOS-Based Analog/Digital Blocks" by Mousam Hossain, Adrian Tatulian, Shadi Sheikhfaal, Harshavardhan R. Thummala, and Ronald F. Demara implementes a 2D based-array to perform processing in memory. This work proposes a novel design with the combination of non-volatile memory and CMOS logic to enable more functions in memory, targeting efficient analog activation. The proposed design, called SCAPE, embeds analog arithmetic capabilities providing a selectable thresholding functionality to realize generalized neuron activation functions.

Extended to *architectural exploration for neural network acceleration*, the article "Bit-Line Computing for CNN Accelerators Co-Design in Edge AI Inference" by Marco Rios, Flavio Ponzina, Alexandre Levisse, Giovanni Ansaloni, and David Atienza developes a new architectural design with bit-line computing circuits for neural-network convolutions with its associated hardware/software co-design framework to optimize and deploy convolution. The proposed design results in a 91% energy saving (for a 1% accuracy degradation constraint) regarding the state-of-the-art bit-line computing approaches. In addition, the article "NeuSB: A Scalable Interconnect Architecture for Spiking Neuromorphic Hardware" by Adarsha Balaji, Phu Khanh Huynh, Francky Catthoor, Nikil D. Dutt, Jeffrey L. Krichmar, and Anup Das proposes a new interconnect architecture, called NeuSB, for Spiking Neural Networks (SNN)s. NeuSB partitions a bus lane into several slices and merges them into segments to reduce energy and transmission latency because partitioned buses have a smaller inter-spike interval, need a smaller buffer for data packages, and require less energy and chip size.

Regarding *technologies considering memory-related issues*, the article "ALP: Alleviating CPU-Memory Data Movement Overheads in Memory-Centric Systems" by Nika Mansouri Ghiasi, Nandita Vijaykumar, Geraldo F. Oliveira, Lois Orosa, Ivan Fernandez, Mohammad Sadrosadati, Konstantinos Kanellopoulos, Nastaran Hajinazar, Juan Gómez Luna, and Onur Mutlu proposes a programmer-transparent hardware-software cooperative mechanism, called ALP. ALP is a compiler-based technique that alleviates the data-movement in the memories between the host and the near-data processing (NDP) units by proactively and accurately transferring data between segments. Furthermore, the article "TopSort: A High-Performance Two-Phase Sorting Accelerator Optimized on HBM-based FPGAs" by Weikang Qiao, Licheng Guo, Zhenman Fang, Mau-Chung Frank Chang, and Jason Cong proposes a two-phase sorting (TopSort) accelerator to improve the sorting performance by fully utilizing the capability of high-bandwidth memory (HBM) on FPGA. The evaluation shows that TopSort is 6.7x and 2.2x faster than the state-of-the-art CPU and FPGA sorters. Finally, the article "Anatomy of On-Chip Memory Hardware Fault Effects Across the Layers" by George Papadimitriou and Dimitris Gizopoulos" extends to study the memory fault issue. This work points out that the evaluation should take into consideration all faults that arrive from the underlying on-chip memory hardware as well as their distribution. By having a unique full-system setup, this work explores how the particular system attributes can affect the estimation of Program Vulnerability Factor (PVF) and proves that the PVF estimation alone can deliver contradicting results against the full-stack Architectural Vulnerability Factor (AVF).

The Guest Editors thank the reviewers for their valued time, expertise, and constructive feedback in their reviews. We also thank all of the authors for their submissions and their accommodation of the publication deadlines and constraints. Finally, we are indebted to the Editor-in-Chief of *IEEE Transactions on Emerging Topics in Computing*, Professor Paolo Montuschi, who has collectively made this Thematic Section possible.

Sincerely,

YUAN-HAO CHANG, *Guest Editor*
Institute of Information Science
Academia Sinica
Taipei, 115, Taiwan
johnson@iis.sinica.edu.tw

VINCENZO PIURI, *Guest Editor*
Dipartimento di Informatica
Università degli Studi di Milano
20133, Milano, MI, Italy
vincenzo.piuri@unimi.it

**Yuan-Hao Chang** (Fellow, IEEE) received the PhD in computer science from the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. He is currently a deputy director and research fellow (professor) of Institute of Information Science (IIS), Academia Sinica, Taipei, Taiwan. His research interests include memory/storage systems, operating systems, embedded systems, and real-time systems. He has published more than 60 research papers in ACM/IEEE Transactions (e.g., *IEEE Transactions on Emerging Topics in Computing*, *IEEE Transactions on Computers*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *IEEE Transactions on Very Large Scale Integration*, *ACM Transactions on Embedded Computing Systems*, *ACM Transactions on Design Automation of Electronic Systems*, and *ACM Transactions on Storage*). His work received best paper awards from premier conferences ACM/IEEE CODES+ISSS 2022, ACM/IEEE ISLPED 2020, and ACM/IEEE CODES+ISSS 2019. He is an associate editor of *IEEE Transactions on Emerging Topics in Computing (TETC)*, *ACM Transactions on Storage (TOS)*, and *ACM Transactions on Cyber-Physical Systems (TCPS)*.

**Vincenzo Piuri** (Fellow, IEEE) received the PhD degree in computer engineering from Politecnico di Milano, Italy, in 1989. He has been a full professor of computer engineering with the Università degli Studi di Milano, Italy, since 2000. He has been an associate professor with the Politecnico di Milano; a visiting professor with the University of Texas, Austin, USA; and a visiting researcher with George Mason University, USA. His research interests include digital processing architectures, arithmetic architectures, fault tolerance, dependability, and cloud computing infrastructures, artificial intelligence, machine learning, pattern analysis and recognition, signal and image processing, He has published more than 400 articles in international journals, proceedings of international conferences, books, and book chapters. He has been IEEE Vice President for Technical Activities, IEEE Director and the President of the IEEE Systems Council and the IEEE Computational Intelligence Society. He has been the editor-in-chief of the IEEE Systems Journal and an associate editor of *IEEE Transactions on Computers*. He is also a Distinguished Scientist of ACM.