Guest Editors' Introduction: Near-Memory and In-Memory Processing

Hai "Helen" Li
Duke University, Durham, NC 27708 USA

Alaa R. Alameldeen
Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Onur Mutlu
ETH Zürich, 8092 Zürich, Switzerland

The rapid explosion in data-intensive applications is posing unprecedented demand for superior data-processing capabilities. With the end of Dennard scaling, the benefits of conventional approaches of moving stored data to processing cores and enhancing the computing capability of these cores through technology scaling have diminished. At the same time, advancements in new material and integration technologies have made the old concept of coupling compute units to memory more viable.

Near-memory processing (NMP) and in-memory processing (IMP), in general, cover a wide spectrum of computing capabilities embedded in close proximity to and/or inside the memory array. NMP/IMP approaches can reduce the latency and energy consumption associated with data movement throughout the cache and memory hierarchy. Moreover, NMP/IMP approaches can operate in parallel with CPU/GPU cores and other accelerators. Due to their potential to greatly reduce energy and improve execution time by reducing data movement, various NMP/IMP approaches have received substantial interest in both industry and academia.

Digital Object Identifier 10.1109/MDAT.2021.3124742
Date of current version: 28 February 2022.

In this context, this Special Issue on Near-Memory and In-Memory Processing introduces, explores, and investigates challenges and opportunities in developing innovative NMP/IMP computer architectures based on conventional and emerging technologies for a wide variety of modern applications. The aim of this special issue is to offer the readers a clear perspective of the rich landscape of both academic and industrial endeavors in architecting, designing, and testing NMP/IMP architectures and systems.

This special issue brings together six contributed articles covering hardware-, software- and algorithm-level techniques that enable and advance NMP and IMP systems. The first article, titled "Analog-to-Digital Converter Design Exploration for Compute-in-Memory Accelerators," by Jiang et al., investigates the analog shift-add analog-to-digital converter (ADC) design for compute-in-memory (CiM) array. The authors show that 6-bit ADC precision is sufficient for no accuracy degradation for a large array (512 × 512), meanwhile obtaining the best tradeoff between hardware performance and area overhead, compared to the prior state-of-the-art designs. The second article, titled "Computing-in-Memory Using Ferroelectrics: From Single- to Multi-Input Logic," by Huang et al., presents ferroelectric FETs (FeFETs)-based CiM designs that can

Copublished by the IEEE CEDA, IEEE CASS, IEEE SSCS, and TTTC

support various key operations—from bit-wise Boolean logic to multi-input majority vote (MAJ). These designs can be used for multiple learning tasks/paradigms, from classical machine learning to hyperdimensional computing, and achieve both compactness and efficiency.

The third article, titled "Executing Data Integration Effectively and Efficiently Near the Memory," by Zhao et al., explores the use of NMP to both accelerate the execution of data transformation workloads and reduce their energy needs. An NMP architecture was developed based on the observation that most data integration workloads have regular memory access patterns and varying computation intensity. The fourth article, titled "A 703.4 GOPs/W Binary SegNet Processor With Computing-Near-Memory Architecture for Road Detection," by Lyu et al., proposes an FPGA-based deep learning accelerator using the binary SegNet (BSegNet) with computing-near-memory (CNM) architecture for road detection at edges. The accelerator has optimized CNM architecture with massive bit-level parallel processing elements (PEs) and pipelines for low latency of the critical path.

The fifth article, titled "A Case for PIM Support in General-Purpose Compilers," by Sadeghi and Ejlali, makes a case for general support for PIM in compilers and put forth an approach to face it along with a simple model. Finally, "A Survey of Neuromorphic Computing-in-Memory: Architectures, Simulators, and Security," by Staudigl et al., provides a comprehensive survey of neuromorphic computing focusing on three essential aspects, including neuromorphic system architectures that evolved over time, simulation platforms concentrating on system, architecture, circuit, and device levels, and hardware security threats that are presented on already existing work in the CMOS domain to learn the lessons and identify the threats in neuromorphic platforms.

These six articles cover a diverse range of topics on NMP and IMP systems. We believe that the readers will find them interesting and gain new insights into this evolving area. We thank all those who submitted their research on this special issue. We also thank all the reviewers, the former EiC, Jörg Henkel, and Sara Dailey, without whose help this special issue would not have been possible. ■

Hai "Helen" Li is a Clare Boothe Luce Professor and the Associate Chair of the Electrical and Computer Engineering Department at Duke University, Durham, NC, USA. Her research interests include neuromorphic computing systems, machine learning and deep neural networks, memory design and architecture, and cross-layer optimization for low power and high performance. Li has a BS and an MS from Tsinghua University, Beijing, China, and a PhD from Purdue University, West Lafayette, IN, USA.

Alaa R. Alameldeen is an Associate Professor at the School of Computing Science, Simon Fraser University, Burnaby, BC, Canada. Alameldeen has a BSc and an MSc in computer science from Alexandria University, Alexandria, Egypt, and an MSc and a PhD in computer science from the University of Wisconsin, Madison, WI, USA.

Onur Mutlu is a Professor of computer science at ETH Zürich, Zürich, Switzerland. He is also a Faculty Member at Carnegie Mellon University, Pittsburgh, PA, USA, where he previously held the Strecker Early Career Professorship. His current broader research interests are in computer architecture, systems, hardware security, and bioinformatics. Mutlu has a BS in computer engineering and psychology from the University of Michigan, Ann Arbor, MI, USA, and an MS and a PhD in electrical and computer engineering from the University of Texas at Austin, Austin, TX, USA.

■ Direct questions and comments about this article to Hai "Helen" Li, Duke University, Durham, NC 27708 USA; hai.li@duke.edu.