Guest Editorial: IEEE Transactions on Computers Special Section on Emerging Non-Volatile Memory Technologies: From Devices to Architectures and Systems

Yuan-Hao Chang, *Senior Member, IEEE*, Jingtong Hu, *Member, IEEE*,
Mehdi B. Tahoori, *Senior Member, IEEE*, and Ronald F. DeMara, *Senior Member, IEEE*

EMERGING non-volatile memory (NVM) technologies (e.g., 3D NAND flash, STT-MRAM, ReRAM, PCM, and FeRAM) have attracted significant interest in recent years because of the fast-growing performance and capacity demands on memory and storage in the big data era. Well-known examples include the 3D XPoint memory and various NVDIMM hybrid memory technologies. They have shown potential towards larger memory and storage capacities with nearly zero leakage power, while extending memory/system architecture design approaches. The unique characteristics of NVM technologies not only introduce new opportunities, but simultaneously create challenges to the designs at multiple levels of abstraction in computer systems, including those of device management, CPU cache management, memory/storage architecture, and system design. Furthermore, emerging NVM technologies also drive the development of techniques which perform computing operations in memory, i.e., processing-in-memory (PIM), by taking advantage of crossbar-based accelerators using NVMs. Thus, for the emerging NVM technologies, there is an urgent need for technology innovation, modeling, analysis, design, and application, ranging from the device-level to the system-level. The manuscripts appearing in this Special Section focus on one or more of those innovations.

As enabled through advances in manufacturing technologies, various types of high-density and low-leakage-power NVMs are developed as the media of CPU register/cache, main memory, secondary storage, and even computational logic. For example, STT-MRAM (resp. PCM) can be used as part of the CPU cache (resp. main memory) to provide a large memory capacity at low cost and low leakage power. However, while placing NVMs into the memory hierarchy, the special characteristics of NVMs, such as non-volatility, read/write asymmetry, limited write endurance, and diversified control features, introduce new design challenges, including the device control/management, cache/memory/storage architecture, and system software designs, all of which are worthy of extended treatment in this timely Special Section.

This Special Section received nearly 60 submissions 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 and industry participants on an international scale. From the many excellent submissions received, twelve manuscripts were accepted. Within the space available, six manuscripts appear in this Special Section with the remainder to appear in subsequent issues of transactions. The manuscripts appearing in this Special Section tackle some of the most recent and impactful design issues of NVMs spanning levels of design abstraction, ranging from device to architecture and system levels, including three categories of works herein: (1) memory architectural approaches, (2) NVM-enhanced storage schemes, and (3) novel PIM approaches using NVM.

Considering aspects of memory architecture, the article "An Analytical Model for Performance and Lifetime Estimation of Hybrid DRAM-NVM Main Memories" by Reza Salkhordeh, Onur Mutlu, and Hossein Asadi considers DRAM-NVM hybrid main memories and proposes an efficient analytical model to accurately analyze the performance and lifetime of hybrid memory. This model can be further applied to aid designers on tuning hybrid memory configurations. Meanwhile, the article "HyVE: Hybrid Vertex-Edge Memory Hierarchy for Energy-Efficient Graph Processing" by Guohao Dai, Tianhao Huang, Yu Wang, Huazhong Yang, and John Wawrzynek exploits ReRAM to replace DRAM on graph processing, and presents a hybrid memory hierarchy to improve the energy efficiency on graph processing by avoiding random access and data

- *Y. Chang is with the Institute of Information Science at Academia Sinica, Taipei, Taiwan. E-mail: johnson@iis.sinica.edu.tw.*
- *J. Hu is with the Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15260. E-mail: jthu@pitt.edu.*
- *M. Tahoori is with the Department of Computer Science, Karlsruhe Institute of Technology, Karlsruhe 76131, Germany. E-mail: mehdi.tahoori@kit.edu.*
- *R. DeMara is with the Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 4000. E-mail: ronald.demara@ucf.edu.*

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TC.2019.2923033

writes to ReRAM. These works concentrate on the significant issues of advancing NVM architectures by balancing performance, write energy, and resilience from lifetime perspectives.

Extended to consider storage systems, NVM can be adopted to improve the energy profile and update performance of storage systems. The article "Efficient and Consistent NVMM Cache for SSD-based File System" by Youmin Chen, Youyou Lu, Jiwu Shu, and Pei Chen adopts NVM as the buffer cache of file systems to boost the storage/file system performance by resolving the frequent synchronization operation issue. The authors design a fined-grained cache management to absorb the synchronization operations in the NVM cache adaptively, so that the write traffic to the storage device can be significantly reduced while the file system consistency is still preserved.

This Special Section also examines trends toward executing various computational operations within memory, i.e., processing-in-memory (PIM), with a special focus on PIM with NVMs. The article "In-Memory Processing on the Spintronic CRAM: From Hardware Design to Application Mapping" by Masoud Zabihi, Zamshed Chowdhury, Zhengyang Zhao, Ulya R. Karpuzu, Jian-Ping Wang, and Sachin S. Sapatnekar presents the computational RAM (CRAM) platform, where each CRAM is similar to one STT-MRAM cell with one additional transistor. Based on the CRAM platform, the authors develop a PIM design with computing logic functions, including addition and multiplication, in CRAM. This PIM design with CRAM can map the problems of the 2D convolution on multibit numbers and the inference engine for binary neuromorphic digit recognition, and thus can significantly improve the performance and energy efficiency of 2D convolution tasks and neural network applications. The article "Optimal Application Mapping and Scheduling for Network-on-Chips with Computation in STT-RAM based Router" by Lei Yang, Weichen Liu, Nan Guan, and Nikil Dutt further considers the problem on how to map task graph and schedule tasks of applications, e.g., neural networks, on Network-on-Chips (NoCs) with PIM-based routers. This design utilizes PIM techniques to improve the application performance by offloading the execution from processors to routers that support STT-RAM-enabled PIM. Finally, the article "SPARE: Spiking Neural Network Acceleration Using ROM-Embedded RAMs as In-Memory-Computation Primitives" by Amogh Agrawal, Aayush Ankit, and Kaushik Roy extends PIM methods to accelerate the next-generation neural networks such as spiking neural networks (SNNs) using CMOS-based ROM-embedded SRAMs and STT-MRAM-based ROM-embedded MRAMs as the computation primitives. Based on the ROM-embedded RAM technologies, the authors propose an in-memory, distributed processing architecture that uses the ROM-embedded RAMs as the storage of the lookup-table for neuro-synaptic models and leverages the input data sparsity in SNNs with event-driven processing in each processing unit. Thus, the proposed architecture can accelerate the SNN without additional area overhead. It is worth noting that the above papers focus on developing PIM techniques using NVMs to accelerate neural networks, and how these techniques can form cornerstones to enable broader PIM techniques with NVMs in other application domains as well.

Forthcoming manuscripts span topics including: (1) NVM lifetime and performance enhancement (2) access support enabling NVM as a primary memory technology, (3) data security of NVM-enhanced storage systems, and (4) neural network acceleration with NVM-based PIM approaches.

Considering the NVM lifetime, the article "Improving the Lifetime of Non-Volatile Cache by Write Restriction" by Sukarn Agarwa and Hemangee K. Kapoor strives to improve the lifetime of NVM when NVM is used as the CPU cache. The authors propose to partition the cache into equal-sized windows, restrict the usage of windows, and distribute writes uniformly across cache sets, so that the write variation to NVM cells can be minimized. While considering NVM write performance, the paper "Quick-and-Dirty: An Architecture for High-Performance Temporary Short Writes in MLC PCM" by Mingzhe Zhang, Lunkai Zhang, Lei Jiang, Frederic T. Chong, and Zhiyong Liu focuses on improving the write performance of systems. This paper adopts short writes to reduce the write latency, but short writes have remarkably brief retentions and require frequent refresh operations. To reduce refresh overheads, the authors present a lightweight scheme that performs short writes for the data with frequent updates, and uses idle-memory intervals to refresh the data of short writes with nominal writes.

With NVM as main memory, the main-memory non-volatility enables new system applications while introducing new system design challenges. The article "NICO: Reducing Software-transparent Crash Consistency Cost for Persistent Memory" by Xueliang Wei, Dan Feng, Wei Tong, Jingning Liu, and Liuqing Ye looks at the data integrity/consistency issue on persistent memory, i.e., NVM main memory. Persistent memory enables a system to support resuming the execution from the point that the system crashed, but a risk is the loss of data consistency if the system crash or power outage occurs after some partial datasets were updated. In their paper, the authors design a lightweight checkpointing scheme to create a consistent snapshot of persistent memory data with a persistent buffer. This scheme uses backend operations to achieve the software-transparent crash consistency with only a small amount of performance overhead. On the other hand, the persistent memory provides a lightweight solution to enable complex applications on transiently-powered systems, which lose power frequently and need to retain information between power losses. The article "Sytare: a Lightweight Kernel for NVRAM-Based Transiently-Powered Systems" by Gautier Berthou, Tristan Delizy, Kevin Marquet, Tanguy Risset, and Guillaume Salagnac discusses the transiently-powered systems with peripheral devices from the viewpoint of the operating system kernel. The authors investigate the peripheral state volatility, peripheral access atomicity, and interrupt handling issues, and propose a kernel-oriented solution to resolve these issues on transiently-powered systems with minimal impact on the programming model and the performance of applications. These techniques resolving the information retention and crash recovery issues are also the key techniques to advance the development of IoT systems and energy-harvesting systems.

Furthermore, NVM is adopted to accelerate the performance of cryptographic file systems in the paper "NV-

eCryptfs: Accelerating Enterprise-level Cryptographic File System with Non-Volatile Memory" by Chunhua Xiao, Lei Zhang, Weichen Liu, Linfeng Cheng, Pengda Li, Yanyue Pan, and Neil Bergmann. In this paper, an asynchronous software stack is presented to utilize NVM as a fast storage tier for parallelizing encryption and data I/O. The ciphertext is buffered in the NVM cache and flushed back to the storage device with background threads. Although the above techniques focus on improving the performance of storage/file systems with NVM, they envision the potentials of NVM on improving a wide variety of capabilities of storage systems while supplementing existing storage devices to enable new application scenarios.

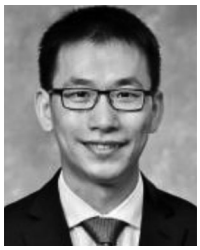
Regarding the neural network acceleration with NVM, the manuscript "NNPIM: A Processing In-Memory Architecture for Neural Network Acceleration" by Saransh Gupta, Mohsen Imani, Harveen Kaur, and Tajana Rosing proposes a PIM architecture, call NNPIM, to accelerate a neural network's inferencing phase inside the memory. The authors design a crossbar memory architecture to support fast addition, multiplication, and search operations. Meanwhile, this architecture maps various neural network functionalities using parallel in-memory components with optimization techniques to achieve high computation speedup and energy efficiency by significantly reducing the amount of data movement between the memory element and the processing cores.

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 two consecutive Editors-in-Chief of *IEEE Transactions on Computers*, Drs. Paolo Montuschi and Ahmed Louri, who have collectively made this Special Section possible.

Sincerely,

Yuan-Hao Chang
Jingtong Hu
Mehdi B. Tahoori
Ronald F. DeMara
Guest Editors and Topical Editor

Yuan-Hao Chang (S'07-M'09-SM'14) received the PhD degree in computer science from the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, in 2009. He is currently a research fellow (equal to professor) with tenure at Institute of Information Science, Academia Sinica, Taipei, Taiwan. His research interests include memory/storage systems, operating systems, embedded systems, and real-time systems. In these fields, he has published more than 120 articles on archival journals and peer-reviewed conferences. He serves as a member of many conference program committees and as a reviewer for various IEEE/ACM transactions and highly cited conferences. He is an associate editor of *ACM Transactions on Cyber-Physical Systems (ACM TCPS)*, and was the program co-chair and general co-chair of IEEE Non-Volatile Memory Systems and Applications Symposium (NVMSA) 2017 and 2018 respectively. He is a senior member of the IEEE.



Jingtong Hu (S'07-M'13) received the PhD degree in computer science from the University of Texas, Dallas, in 2013. He is currently an assistant professor with the Department of Electrical and Computer Engineering, University of Pittsburgh. His research interests include emerging non-volatile memory, embedded systems, and FPGA. His works received Best Paper Award Nomination from DAC 2017, 2019 and ISQED 2018. He served as program committee members for many conferences such as DAC, DATE, ASP-DAC, RTSS, CASES. He is also a reviewer for various IEEE/ACM Transactions. He is a member of the IEEE.

Mehdi B. Tahoori (S'02-M'04-SM'08) is a full professor and chair of Dependable Nano-Computing (CDNC) at the Institute of Computer Science & Engineering (ITEC), Department of Computer Science, Karlsruhe Institute of Technology (KIT), Germany. He holds several pending and granted US and international patents. He has authored more than 250 publications in major journals and conference proceedings on a wide range of topics, from dependable computing and emerging nanotechnologies to system biology. His current research interests include nanocomputing, reliable computing, VLSI testing, reconfigurable computing, emerging nanotechnologies, and systems biology. He is a recipient of the National Science Foundation Early Faculty Development (CAREER) Award. He has been a program committee member, organizing committee member, track and topic chair, as well as workshop, panel, and special session organizer of various conferences and symposia in the areas of VLSI design automation, testing, reliability, and emerging nanotechnologies, such as ITC, VTS, DAC, ICCAD, DATE, ETS, ICCD, ASP-DAC, GLSVLSI, and VLSI Design. He is a senior member of the IEEE.

Ronald F. DeMara (S'87-M'93-SM'04) received the PhD degree in computer engineering from the University of Southern California, in 1992. Since 1993, he has been a full-time faculty member with the University of Central Florida where he is a professor of Electrical and Computer Engineering, joint faculty of Computer Science, and Digital Learning Faculty fellow. His research interests are in adaptive and resilient computer architectures with emphasis on reconfigurable logic devices, evolvable hardware, and post-CMOS devices, on which he has published more than 275 articles and holds one patent. He has served on the Editorial Boards of *IEEE Transactions on Emerging Topics in Computing, IEEE Transactions on Computers, IEEE Transactions on VLSI Systems*, *Journal of Circuits, Systems, and Computers*, the journal *Microprocessors and Microsystems*, and as Associate Guest Editor of *ACM Transactions on Embedded Computing Systems*, as well as a Keynote Speaker of the *International Conference on Reconfigurable Computing and FPGAs (ReConFig)* and the *IEEE Reconfigurable Architectures Workshop (RAW)*. He received *IEEE's Joseph M. Biedenbach Outstanding Engineering Educator Award* in 2008. He is a senior member of the IEEE.