Approximate Computing: From Circuits to Applications

By WEIQIANG LIU, *Senior Member IEEE*
FABRIZIO LOMBARDI, *Life Fellow IEEE*
MICHAEL SCHULTE, *Fellow IEEE*

Computing systems at all scales (from mobile handheld devices to supercomputers, servers, and large cloud-based data-centers) have seen significant performance gains, mostly through the continuous shrinking of the complementary metal–oxide-semiconductor (CMOS) feature size that has doubled the number of transistors on a chip with every technology generation. However, power dissipation has become a fundamental barrier to scale computing performance across all platforms. As the classical Dennard scaling is coming to an end, reductions in on-chip power dissipation as well as throughput increase (as per Moore's Law) have encountered serious challenges. Computation at the nanoscales necessitates fundamentally different approaches; these approaches rely on different computational paradigms that exploit specific features in the targeted set of applications as well as interactions between hardware, software, and algorithms in a computing system.

This special issue explores the technological contributions and developments of approximate computing at disparate levels and provides insight into exciting directions for the future.

Approximate computing has been proposed as a novel paradigm for efficient and low power design at nanoscales. Efficiency is related to the computation of approximate results with at least comparable performance and lower power consumption compared to the fully accurate counterpart, i.e., approximate computing generates results that are good enough rather than always fully accurate. Although computational errors generally are not desirable, applications such as multimedia, signal processing, machine learning, pattern recognition, and data mining are tolerant to the occurrence of some errors; therefore, approximate computing is mostly driven by applications that are related to human perception/cognition and have inherent error resilience. Many of these applications are based on statistical or probabilistic computation, such as different approximations can be made to better suit the desired objectives. Therefore, it is possible to achieve not only energy efficiency but also simpler design and lower latency, while relaxing the strict accuracy requirement for these applications. The basic principles of approximate computing are found across the entire computing stack: devices exhibit approximate behavior at reduced nanoscale dimensions, circuits can be redesigned to better fit specific operational features, and modules and systems can be operated at a level to guarantee an acceptable accuracy and performance improvements.

Hence, approximate techniques have been studied at several levels, including nanoscale hardware (devices, circuits, and architectures), software/algorithms, programming languages, logic synthesis, automated design processes; various processing and memory architectures have been proposed for supporting approximate computing applications. At the hardware level, the design of approximate arithmetic units has received significant research interest. Design metrics and analytical approaches have been proposed for the evaluation of approximate circuits, such as adders and multipliers. Approximate algorithms and systems have been studied for emerging computing applications, such as deep neural networks (DNNs) and stochastic computation. Design

Digital Object Identifier 10.1109/JPROC.2020.3033361

automation, test, and security issues of approximate computing systems have also been investigated. It is widely accepted that the full potential of approximate computing cannot be fully exploited without considering the interactions between these different levels under often conflicting constraints, such as error analysis/control, power dissipation, and design complexity.

Approximate computing has received significant attention from both research and industrial communities in the past few years due to the challenge of designing power-efficient computing systems for emerging applications. Researchers from electronic engineering, computer science, and computer engineering, and many other related areas, have studied approximate techniques at different levels. It should also be noted that leading companies, such as IBM, Google, Intel, and ARM, are involved in experimental research and implementing commercial products and services based on approximate computing. For example, Google tensor processing units (TPUs) use approximate computing to reduce power consumption; IBM is applying approximate computing for on-chip artificial intelligence (AI) acceleration; and ARM is also considering design an approximate processor for low power. An increased number of funding agencies (such as NSF, DARPA, NSFC, and the EU Framework Horizon 2020) are sponsoring research in this field. Therefore, the proposed Special Issue covers an important and timely topic that will be welcomed by a wide readership.

Despite recent significant attention, approximate computing still requires additional and significant efforts to make it a mainstream computing paradigm for energy-efficient and high-performance systems. This Special Issue explores the technological contributions and developments at disparate levels of this new emerging paradigm, collecting state-of-the-art progress and pointing to exciting new directions in this important computing area.

I. OVERVIEW OF THE ISSUE

This special issue includes nine articles that cover most of the aspects of approximate computing circuits and systems. The first group of articles focuses on the approximate arithmetic and approximate memory in the circuit level. The second group of articles discusses specific domains and applicative areas, such as logic synthesis, testing, and security. The third group of articles presents the current endeavors in both digital and analog approximate computing into AI and, in particular, machine learning.

A. Approximate Arithmetic and Memory

Approximate Arithmetic Circuits: A Survey, Characterization, and Recent Applications
by H. Jiang, F. J. H. Santiago, H. Mo, L. Liu, and J. Han

Approximate arithmetic circuits have been widely due to their importance in computing systems. The authors have made great efforts to provide a comprehensive evaluation of recently proposed approximate arithmetic circuits, such as adders, multipliers, and dividers. These approximate circuits are compared under different design constraints and further applied in image processing and deep learning applications to examine their error characteristics. These results are applied to simple and more complex accumulative computations using approximate adders and multipliers.

Elementary Functions and Approximate Computing
by J.-M. Muller

Targeting error tolerant applications, this article presents the classical approaches of approximate elementary functions, such as sin, cos, tan, and so on. Mainstream techniques (such as shift-and-add algorithms, polynomial or rational approximations, table-based methods, and bit-manipulation) are discussed for improving performance versus a controlled accuracy loss. These techniques are further considered for software and hardware implementation in terms of speed and accuracy tradeoff.

Circuit-Level Techniques for Logic and Memory Blocks in Approximate Computing Systems
by S. Amanollahi, M. Kamal, A. Afzali-Kusha, and M. Pedram

The authors focus on circuit-level techniques for both approximate logic and memory blocks. Two main techniques, namely, voltage overscaling (VOS) and design complexity reduction (DCR), are comprehensively discussed for data-path circuit designs. Circuit-level techniques of bit truncation/trimming, read and write operations dropping, weak read and write operations, refresh rate reduction, and adjusting the restoration time for both volatile memory and nonvolatile memory (NVM) are also reviewed. Challenges and future directions are provided by considering the integration of approximate logic and memory using the same device technology under the required performance and output quality constraints.

B. Synthesis, Testing, and Security of Approximate Computing Circuits and Systems

A Survey of Testing Techniques for Approximate Integrated Circuits
by M. Traiola, A. Virazel, P. Girard, M. Barbareschi, and A. Bosio

Conventional testing methods are facing new challenges from emerging approximate designs. This article examines the test procedure of approximate integrated circuits in which initially catastrophic defects should be found and avoided. However, some defects can be tolerated in approximate computing circuits, thus creating new opportunities for testing. The authors identify three main approximation-aware testing phases, namely, fault classification, test pattern generation, and test set application. New metrics are proposed for an effective evaluation, and novel state-of-the-art test techniques are introduced. Moreover, it is estimated

that an approximate-aware test will also increase the yield compared with a conventional test flow.

Approximate Logic Synthesis: A Survey
by I. Scarabottolo, G. Ansaloni, G. A. Constantinides, L. Pozzi, and S. Reda

To achieve the best tradeoff between hardware complexity and accuracy for a specific approximate design, automatic tools based on approximate logic synthesis are needed. This article reviews the transformation methods for functional approximation that can achieve an approximate Boolean function from its exact designs. Three main categories have been identified and discussed in detail, namely, netlist transformation, Boolean rewriting, and approximate high-level synthesis. Open challenges in approximate logic synthesis, including scalability, runtime accuracy-configurable hardware, and cross-layer designs, are also presented.

Security in Approximate Computing and Approximate Computing for Security: Challenges and Opportunities
by W. Liu, C. Gu, M. O'Neill, G. Qu, P. Montuschi, and F. Lombardi

This article focuses on a new topical area of research that links security and approximate computing. There is strong and rather conclusive evidence that approximate computing circuits may introduce security vulnerabilities due to the uncertain and unpredictable nature of the intrinsic errors. These manifestations during an approximate execution may be indistinguishable from malicious modifications of the circuits or systems. On the other hand, approximate computing also presents new opportunities for hardware security and cryptography engineering to secure a system while accelerating computation. The authors have provided a classification of state-of-the-art works in this research field, including threat models and promising security approaches using approximate computing. A number of open questions and potential future research directions are also discussed.

C. Application of Approximate Computing for AI

Efficient AI System Design With Cross-Layer Approximate Computing
by S. Venkataramani, X. Sun, N. Wang, C.-Y. Chen, J. Choi, M. Kang, A. Agarwal, J. Oh, S. Jain, T. Babinsky, N. Cao, T. Fox, B. Fleischer, G. Gristede, M. Guillorn, H. Haynie, H. Inoue, K. Ishizaki, M. Klaiber, S.-H. Lo, G. Maier, S. Mueller, M. Scheuermann, E. Ogawa, M. Schaal, M. Serrano, J. Silberman, C. Vezyrtzis, W. Wang, F. Yee, J. Zhang, M. Ziegler, C. Zhou, M. Ohara, P.-F. Lu, B. Curran, S. Shukla, V. Srinivasan, L. Chang, and K. Gopalakrishnan

The application of approximate computing to DNNs is most appropriate as there is an abundance of redundancy in DNNs that can tolerate errors. This article presents IBM's RAPID, which is a multi-TOP DNN hardware accelerator core fabricated using 14-nm technology. This article starts from an introduction of AI workloads, their computational characteristics, and the summary of the different algorithmic approximation techniques for AI workloads. The RAPID architecture is then presented in detail with approximate computing techniques across different levels, including algorithms, architecture, programmability, and hardware. It shows that an algorithm-software-hardware codesign is critical for the design of efficient AI systems using approximate computing.

Deep In-Memory Architectures in SRAM: An Analog Approach to Approximate Computing
by M. Kang, S. K. Gonugondla, and N. R. Shanbhag

This article presents an analog form of approximate computing, namely, in-memory computing, for hardware acceleration of machine learning algorithms. The authors have provided an overview of recently proposed deep in-memory architectures (DIMAs) with several prototype chips using 65-nm technology; DIMA addresses data movement issues in von Neumann architectures. This article also focuses on the in-memory architectures realizing matrix computations on the bitlines of a memory bitcel array in a low compute-to-signal-noise ratio domain. A Shannon-inspired rationale for robustness to process, temperature and voltage variations, and design guidelines to manage analog nonidealities are provided. The authors also present a system-level framework for fully characterizing the fundamental energy-delay-accuracy tradeoff.

Resistive Crossbars as Approximate Hardware Building Blocks for Machine Learning: Opportunities and Challenges
by I. Chakraborty, M. Ali, A. Ankit, S. Jain, S. Roy, S. Sridharan, A. Agrawal, A. Raghunathan, and K. Roy

Resistive crossbars based on emerging resistive NVM devices, that are inherently approximate, have shown to be promising building blocks of in-memory computing systems for machine learning workloads. The authors present a comprehensive overview of the emerging paradigm of computing using NVM crossbars for accelerating machine learning workloads. This article starts from describing the design principles of resistive crossbars. Then, intrinsic approximations from device and circuit characteristics are discussed. An overview of spatial architectures that exploit the high storage density of NVM crossbars is presented. The authors demonstrate the frameworks that effectively capture device-circuit-architecture characteristics to evaluate large-scale DNNs using resistive crossbar and study the software- and hardware-based mitigation techniques for accurate recovery.

II. CLOSING REMARKS

As shown in the articles of this special issue, approximate computing has great potential for emerging applications. However, approximate

computing is not yet at a fully mature stage. We hope this special issue provides a comprehensive reference for current developments and promote future research on approximate computing.

This year has been challenging due to the global epidemic. We would like to thank all contributing authors for their articles and great efforts under these difficult circumstances. We are also sincerely thankful to all reviewers who provided valuable and constructive feedback on the manuscripts through a very selective and competitive review process. Finally, we want to thank Prof. Gianluca Setti (Editor-in-Chief), Jo Sun (Senior Publications Editor), and Vaishali Damle (Managing Editor) for their suggestions and supports during the entire process with the PROCEEDINGS OF THE IEEE. Their kind and valuable help is greatly appreciated. ■

ABOUT THE GUEST EDITORS

Weiqiang Liu (Senior Member, IEEE) received the B.Sc. degree in information engineering from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2006, and the Ph.D. degree in electronic engineering from Queen's University Belfast (QUB), Belfast, U.K., in 2012.

He is currently a Full Professor and the Vice Dean of the College of Electronic and Information Engineering, NUAA. He has published one research book by Artech House and over 100 leading journal and conference papers. One of his papers was selected as the Feature Paper of IEEE TRANSACTIONS ON COMPUTERS in the 2017 December issue. His research interests include approximate computing, emerging technologies in computing systems, computer arithmetic, and hardware security.

Dr. Liu has served as a Steering Committee Member of IEEE TRANSACTIONS ON MULTI-SCALE COMPUTING SYSTEMS. He has been a technical program committee member for several international conferences, including ARITH, DATE, ASAP, ISCAS, ASP-DAC, ICCD, ISVLSI, GLSVLSI, SiPS, NANOARCH, AICAS, NMDC, and ICONIP. He is a member of the Circuits and Systems for Communications (CASCOM) Technical Committee, the VLSI Systems and Applications (VSA) Technical Committee, and the IEEE Circuits and Systems Society. He received the prestigious Outstanding Young Scholar Award by the National Natural Science Foundation of China (NSFC) in 2020. He is the Program Committee Co-Chair of IEEE ARITH 2020. He has also served as an Associate Editor for IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART I: REGULAR PAPERS, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING, and IEEE OPEN JOURNAL OF COMPUTER SOCIETY; and the as a Guest Editor for PROCEEDINGS OF THE IEEE, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING, and *Microelectronics Journal* (Elsevier).

Fabrizio Lombardi (Life Fellow, IEEE) received the B.Sc. degree (Honors) in electronic engineering from the University of Essex, U.K., in 1977, the master's degree in microwaves and modern optics from the Microwave Research Unit, University College London, London, U.K., in 1978, and the Diploma degree in microwave engineering and the Ph.D. degree from the University of London, London, in 1978 and 1982, respectively.

He is currently the holder of the International Test Conference (ITC) Endowed Chair at Northeastern University, Boston, MA, USA. His research interests include emerging technologies (mostly nanoscale circuits and magnetic devices), memory systems, VLSI design, and fault/defect tolerance of digital systems. He has extensively published in these areas and coauthored/edited ten books.

Dr. Lombardi is a Fellow of the IEEE for his contributions to testing and fault tolerance of digital systems. He is a member of the Executive Board of the IEEE Nanotechnology Council (NTC) and the IEEE Computer Society. He is also a member of the IEEE Publication Services and Products Board (PSPB) for the duration of 2019–2023. He was a member of the Editorial Boards of *ACM Journal on Emerging Technologies in Computing Systems, IEEE Design and Test of Computers* magazine, and IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS. He was a recipient of the 1985/1986 Research Initiation Award from the IEEE/Engineering Foundation; the Silver Quill Award from Motorola, Austin, in 1996; the 2011 Meritorious Service Award and elevated to the Golden Core Membership by the IEEE Computer Society, in 2011; the 2019 NTC Distinguished Service Award; and the 2019 Spirit of the CS Award. He has received many professional awards, including the Visiting Fellowship at the British Columbia Advanced System Institute, University of Victoria, Victoria, BC, Canada, in 1988; twice the Texas Experimental Engineering Station Research Fellowship (1991–1992 and 1997–1998); the Halliburton Professorship in 1995; the Outstanding Engineering Research Award at Northeastern University in 2004; and the International Research Award from the Ministry of Science and Education of Japan (1993–1999). Together with his students, his manuscripts have been selected for the best paper awards at technical events/meeting, such as IEEE DFT and IEEE/ACM NANOARCH. He was the Chair of the 2016 and 2017 IEEE CS Fellow Evaluation Committee. He serves as the Chair for the Committee on Nanotechnology Devices and Systems of the Test Technology Technical Council of the IEEE. He has been involved in organizing many international symposia, conferences, and workshops sponsored by professional organizations as well as a Guest Editor of special issues in archival journals and magazines. He is the 2020 Vice-President for Publications (2007–2010). He has been appointed on the executive boards of many nonprofit organizations (such as Computing-in-the-Core, now code.org the nonpartisan advocacy coalition for K-12 computer science education) as well as the Computer Society (as an elected two-term member of its Board of Governors, from 2012 to 2017) and the IEEE (as an appointed member of the Future Directions Committee, from 2014 to 2017). He is also the 2021 President-Elect of IEEE NTC, and the 2022–2023 President; in 2021, he will be the second Vice President of IEEE Computer Society. He was a two-term Editor-in-Chief, an Associate Editor-in-Chief (2000–2006), an Associate Editor of the IEEE Transactions on Computers. He was the inaugural two-term Editor-in-Chief of IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING (2013–2017), and the Editor-in-Chief of IEEE TRANSACTIONS ON NANOTECHNOLOGY (2014–2019). He was twice a Distinguished Visitor of the IEEE Computer Society (in 1990–1993 and 2001–2004).

Michael Schulte (Fellow, IEEE) received the B.S. degree in electrical engineering from the University of Wisconsin–Madison, Madison, WI, USA, in 1991, and the M.S. and Ph.D. degrees in electrical engineering from The University of Texas at Austin, Austin, TX, USA, in 1993 and 1996, respectively.

He was a Faculty Member with the University of Wisconsin–Madison and Lehigh University, Bethlehem, PA, USA. He was a Principal Investigator of AMD's PathForward, FastForward-2 Node Architecture, and FastForward Extreme-Scale Computing projects. He is currently a Senior Fellow with AMD Research, Austin, TX, USA, where he leads research, advanced development, and technology transfer activities in high-performance computing, artificial intelligence, heterogeneous systems, and power-efficient processors. He is a Chief Engineer of AMD's Exascale Computing Technologies, including AMD technologies for the Frontier Exascale Supercomputer. He has published over 200 research articles. He is an inventor of over 20 patents and has over 20 patents pending. His research interests include computer arithmetic, computer architecture, power-efficient computing, and domain-specific systems.

Dr. Schulte is a Fellow of the IEEE for his contributions to computer architectures. He was a recipient of the NSF CAREER Award, the Alfred Noble Robinson Award, the AMD Way Award, the AMD Datacenter and Embedded Solutions Group Award of Excellence, and service awards from the National Society of Black Engineers and the IEEE. He has served as the Program Chair and the General Chair of the IEEE International Conference on Application-Specific Systems, Architectures and Processors (ASAP), the IEEE International Symposium on Computer Arithmetic (ARITH), and the Asilomar Conference on Signals, Systems, and Computers. He has served as an Associate Editor for IEEE TRANSACTIONS ON COMPUTERS, *Journal of Signal Processing Systems*, and *Journal of VLSI Signal Processing*. He has been a Guest Editor for special issues published in IEEE TRANSACTIONS ON COMPUTERS and *Journal of VLSI Signal Processing*.