Guest Editorial Cross-Layer Designs, Methodologies, and Systems to Enable Micro AI for On-Device Intelligence

THIS Special Issue of the IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS (JETCAS) aims to investigate the latest research in the domain of cross-layer design approaches including algorithms, architectures, hardware, and system integration for micro-intelligent systems processing. In order to avoid compute-intensive algorithms running on the cloud and allow low latency and savings in communication link for bandwidth, on-device sensor analytics and micro-AI deployment of various applications with extremely low power is gaining more traction every day.

This Special Issue covers four comprehensive topics about cross-layer designs, methodologies, and systems to enable micro-AI for on-device intelligence. The areas of interest are as follows:

- AI model design,
- hardware design methodology,
- memory hierarchy and data movement, and
- flexibility and reliability.

We begin with [A1], Mazumder *et al.* where the Guest Editors of this issue provide a brief tutorial of on-device AI for efficient inference in resource-constrained micro-AI devices. The tutorial goes into detail on the optimization strategies for both network selection and consequent replication of the networks onto micro-AI platforms. The study of this tutorial will give the reader an exhaustive idea regarding neural network exploration, help understand quantization and sparsification methods, and allow the ability to configure different networks to be deployed onto micro-AI platforms.

In consequence, this Special Issue selects articles that address the issue of efficient AI model design. Furthermore, this Special Issue also seeks to bridge the gap between the generation of AI networks and their hardware representations with relevant research articles from the domain of hardware design methodology. In addition to the hardware design methods, this Special Issue further extends its reach to incorporate articles that look into exploitation strategies on data movement and memory hierarchy to allow efficient and low power deployments. Finally, the Special Issue aims to collect the research on novel methods that lead to enhanced reliability and flexibility with AI computing. A brief summary of the works according to their broader niche is elaborated in the next few sections.

I. AI MODEL DESIGN

AI model design conforms to the techniques and approaches that lead to efficient processing of AI networks on resource-constrained devices with a negligible compromise for accuracy. The relevant techniques to address this problem can include strategies for computation reduction, neural architecture search for network selection, and alteration in training to reduce timing and power overhead. In this Special Issue, several articles are introduced that aim to reduce network parameters for generating lightweight models which are friendly for micro-AI deployment.

In [A2], Niculescu *et al.* address the issue of automated deployment of vision-based CNN navigation for nano drones. Nano-UAVs (unmanned aerial vehicles) have limited memory capacity and require significant processing of network parameters to make vision applications deployable on to them. To this extent, this article focuses on the deployment of PULP-Dronet on a Crazyflie 2.1 nano-UAV. The proposed approach improves the behavior of the nano-UAV in relation to obstacle avoidance, free flight, and lane following without compromising prediction accuracy.

In [A3], Safayenikoo *et al.* delve deeper into optimizing training time with skip connections in weight. This article follows the principle that there are temporal variations in accuracy improvement during training and when these variations are insignificant, one can selectively skip updating the weights and update the bias only to allow the model to train and avoid overfitting. The corresponding upshot with this results in accuracy improvement with considerably less computation and training time.

Additional articles on optimization of AI models are as follows:

- "SWANN: Small-World Architecture for Fast Convergence of Neural Networks" by Javaheripi *et al.* [A4]
- "Advanced Design Methods From Materials and Devices to Circuits for Brain-Inspired Oscillatory Neural Networks for Edge Computing" by Carapezzi *et al.* [Invited submission from the 3rd IBM IEEE CAS/EDS - AI Compute Symposium (AICS 2020)] [A5]
- "QS-NAS: Optimally Quantized Scaled Architecture Search to Enable Efficient On-Device Micro-AI" by Hosseini *et al.* [A6]
- "AutoRank: Automated Rank Selection for Effective Neural Network Customization" by Javaheripi *et al.* [A7]
- "TempDiff: Feature Map-Level CNN Sparsity Enhancement at Near-Zero Memory Overhead via Temporal Difference" by De Alwis *et al.* [A8]

Digital Object Identifier 10.1109/JETCAS.2021.3130743

2156-3357 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

- “An Overview of Sparsity Exploitation in CNNs for On-Device Intelligence With Software-Hardware Cross-Layer Optimizations” by Kang *et al.* [A9]

II. Hardware Design Methodology

This topic covers novel approaches of task scheduling, tiling scheme, and data movement with various hardware design objectives such as energy, power, and timing. Implementation styles such as in-memory computing, near-memory processing, and systolic array architectures, as well as techniques considering tradeoffs among computation and communication, arrangement of processing engines, selection of fixed/floating-point/flexible hardware datawidth, and their impact on timing, power, performance, and energy efficiency are worthy of consideration in this topic.

In [A10], Kang *et al.* present a genetic algorithm based energy-aware convolutional neural network (CNN) quantization framework (EGQ) for processing-in-memory (PIM) architectures. EGQ predicts layer-wise dynamic energy consumption based on the number of ADC access. Also, EGQ automatically optimizes layerwise weight/activation bitwidth that can reduce the total dynamic energy with negligible accuracy loss.

In [A11], Pinkham *et al.* explore optimal mapping of DNN models on an AR/VR compute platform that consists of on-sensor and edge processors to minimize energy and latency.

In [A12], Shiau *et al.* propose a low-cost and learning-based interpolation method to reconstruct high-resolution images. The proposed method generates reconstructed pixels by processing reference pixels with optimal weights, which are pre-trained by solving the minimum mean square error problem for real images.

In [A13], Shi *et al.* outline and analyze the possible methods for handling residual connections of residual neural networks, in combination with line buffer depth-first (LBDF) processing, which is a recent method to reduce memory usage and off-chip memory accesses in high-resolution CNN processing.

Additional articles included in this Special Issue which address the hardware design challenges are as follows:

- “PIM-DRAM: Accelerating Machine Learning Workloads Using Processing in Commodity DRAM” by Roy *et al.* [A14]
- “A Lego-based Neural Network Design Methodology With Flexible NoC” by Chen *et al.* [A15]
- “An Energy-efficient Deep Belief Network Processor Based on Heterogeneous Multi-core Architecture With Transposable Memory and On-chip Learning” by Wu *et al.* [A16]
- “A Multiplier-less Convolutional Neural Network Inference Accelerator for Intelligent Edge Devices” by Hsieh *et al.* [A17]
- “A 16nJ/Classification FPGA-Based Wired-Logic DNN Accelerator Using Fixed-Weight Non-Linear Neural Net” by Kosuge *et al.* [A18]

III. Memory Hierarchy and Data Movement

The memory hierarchy of AI models usually dominates the silicon area in taped-out designs. Furthermore, the cost for memory communication dwarfs that of computation for AI model implementation. Hence, compression techniques, parallel operation, and efficient data movement in relation to different memory levels are of considerable research interest. This Special Issue takes a deep dive into these topics with several articles.

In [A19], Hossain *et al.* address the issue of storage and movement of bulky data through resource-constrained edge devices. A heterogeneous DNN accelerator is proposed that can process multiple workloads using different power-performance operating points. With near-memory computing and leakage reuse, the monolithic architecture experiences an increase to 3.26 TOPS/W energy efficiency from 0.048 TOPS/W energy efficiency for conventional monolithic architectures.

In [A20], Lu *et al.* provide a modification to the implementation of hybrid deep neural networks (DNNs). Hybrid DNNs occupy fewer resources compared to traditional DNNs. However, these hybrid networks are difficult to be replicated on edge devices. This article provides a solution to this problem with a flexible interconnect architecture, 3-D cross-ring, and an efficient dataflow. The proposed modifications allow 90% higher PE utilization and reduce DRAM accesses by 6× when compared to state-of-the-art accelerators.

Additional articles focused on memory-aware optimization are as follows:

- “A TinyML Platform for On-Device Continual Learning With Quantized Latent Replays” by Ravaglia *et al.* [A21]

IV. Flexibility and Reliability

On-device versatile machine learning is highly desired in applications such as difficult-to-reach sensors powered from energy harvesting, smart battery-powered always-on applications, wearables, implantables, and a broad range of ultra-low-power tinyML devices. To enable a power-efficient and secure execution of advanced ML architectures on the same flexible hardware, novel power-aware circuit architectures, and design methodologies are required. This topic calls for novel methods considering flexible and reliable hardware design for versatile modern AI applications.

For example, in [A22], Wang *et al.* propose an accurate and cost-efficient micro AI-enabled countermeasure for securing modern edge devices against emerging cyber-attacks (malware and side-channels attacks) at the hardware level by monitoring applications’ hardware performance counter (HPC) features.

In [A23], Kenarangi *et al.* developed a practical modeling and training flow, and a single-MOSFET high-resolution analog multiplier for ML inference. A multi-bit multiplication is facilitated within a *single* transistor by feeding the features and feature weights into, respectively, the body and gate inputs. The utility and versatility of this fundamental single-transistor block are demonstrated based on the design process of a binary classifier, deep neural network (DNN), and convolutional neural network (CNN). Furthermore, the authors present a novel linearization approach and training flow guided, in a closed loop, by SPICE simulated data. This article provides a vital insight into low-power on-chip ML compute using flexible and robust ML circuits and training flow.

In addition, this Special Issue comprises the following article that explores flexible and reliable hardware for AI applications:

- "Towards Real-Time, At-Home Patient Health Monitoring Using Reservoir Computing CMOS IC" by Chandrasekaran *et al.* [Invited submission from the 3rd IBM IEEE CAS/EDS—AI Compute Symposium (AICS 2020)] [A24]

Acknowledgment

The Guest Editors appreciate the contributions of the authors and would like to thank them for their efforts in contributing to this Special Issue. They would also like to thank all the reviewers for their volunteer and generous service in conforming to the high standards of JETCAS. The Guest Editors would like to mention the constant guidance and support received from the Editor-in-Chief, Prof. An-Yeu (Andy) Wu, Associate-Editor-in-Chief, Prof. Herbert Iu, and the Senior Editorial Board as a whole. Finally, the Guest Editors would like to recognize the great efforts extended by the IEEE Publishing Operations personnel for the successful completion of this Special Issue.

Tinoosh Mohsenin, *Corresponding Guest Editor*
Department of Computer Science and Electrical Engineering
University of Maryland
Baltimore, MD 21250 USA

Inna Partin-Vaisband, *Guest Editor*
Department of Electrical and Computer Engineering
University of Illinois at Chicago
Chicago, IL 60607 USA

Houman Homayoun, *Guest Editor*
Department of Electrical and Computer Engineering
University of California at Davis
Davis, CA 95616 USA

Jae-Sun Seo, *Guest Editor*
School of Electrical, Computer and Energy Engineering
Arizona State University
Tempe, AZ 85287 USA

Xin Zhang, *Guest Editor*
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598 USA

Appendix: Related Articles

[A1] A. N. Mazumder *et al.*, "A survey on the optimization of neural network accelerators for micro-AI on-device inference," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3129415.

[A2] V. Niculescu, L. Lamberti, F. Conti, L. Benini, and D. Palossi, "Improving autonomous nano-drones performance via automated end-to-end optimization and deployment of DNNs," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3126259.

[A3] P. Safayenikoo and I. Akturk, "Weight update skipping: Reducing training time for artificial neural networks," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3127907.

[A4] M. Javaheripi, B. D. Rouhani, and F. Koushanfar, "SWANN: Small-world architecture for fast convergence of neural networks," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3125309.

[A5] S. Carapezzi *et al.*, "Advanced design methods from materials and devices to circuits for brain-inspired oscillatory neural networks for edge computing," *J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3128756.

[A6] M. Hosseini and T. Mohsenin, "QS-NAS: Optimally quantized scaled architecture search to enable efficient on-device micro-AI," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3127932.

[A7] M. Javaheripi, M. Samragh, and F. Koushanfar, "AutoRank: Automated rank selection for effective neural network customization," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3127433.

[A8] U. De Alwis and M. Alioto, "TempDiff: Feature map-level CNN sparsity enhancement at near-zero memory overhead via temporal difference," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3128326.

[A9] S. Kang, G. Park, S. Kim, S. Kim, D. Han, and H.-J. Yoo, "An overview of sparsity exploitation in CNNs for on-device intelligence with software-hardware cross-layer optimizations," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3120417.

[A10] B. Kang, A. Lu, Y. Long, D. Kim, S. Yu, and S. Mukhopadhyay, "Genetic algorithm based energy-aware CNN quantization for processing-in-memory architecture," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3127129.

[A11] R. Pinkham, A. Berkovich, and Z. Zhang, "Near-sensor distributed DNN processing for augmented and virtual reality," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3121259.

[A12] Y.-H. Shiau, K.-Y. Huang, P.-Y. Chen, and C.-Y. Kuo, "A low-cost hardware design of learning-based one-dimensional interpolation for real-time video applications at the edge," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3121070.

[A13] M. Shi, P. Houshmand, L. Mei, and M. Verhelst, "Hardware-efficient residual neural network execution in line-buffer depth-first processing," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3120103.

[A14] S. Roy, M. Ali, and A. Raghunathan, "PIM-DRAM: Accelerating machine learning workloads using processing in commodity DRAM," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3127517.

[A15] K.-C. Chen, C.-K. Tsai, Y.-S. Liao, H.-B. Xu, and M. Ebrahimi, "A Lego-based neural network design methodology with flexible NoC," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3125399.

[A16] J. Wu *et al.*, "An energy-efficient deep belief network processor based on heterogeneous multi-core architecture with transposable memory and on-chip learning," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3114396.

[A17] M.-H. Hsieh, Y.-T. Liu, and T.-D. Chiueh, "A multiplier-less convolutional neural network inference accelerator for intelligent edge devices," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3116044.

[A18] A. Kosuge, M. Hamada, and T. Kuroda, "A 16 nJ/classification FPGA-based wired-logic DNN accelerator using fixed-weight non-linear neural net," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3114179.

[A19] M. S. Hossain and I. Savidis, "Leakage reuse for energy efficient near-memory computing of heterogeneous DNN accelerators," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3121687.

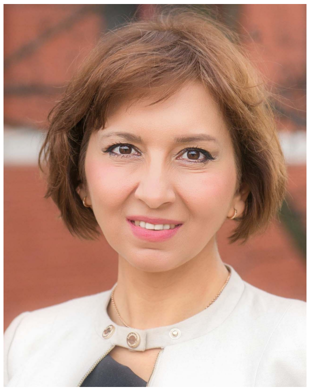
[A20] W. Lu, P.-T. Huang, H.-M. Chen, and W. Hwang, "An energy-efficient 3D cross-ring accelerator with 3DSRAM cubes for hybrid deep neural networks," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3116611.

[A21] L. Ravaglia *et al.*, “A TinyML platform for on-device continual learning with quantized latent replays,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3121554.

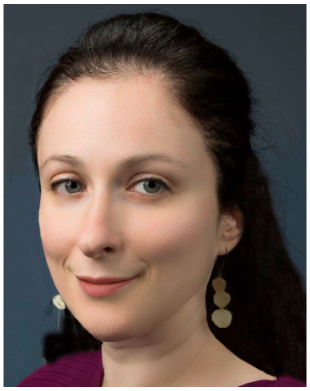
[A22] H. Wang, H. Sayadi, S. M. P. Dinakarrao, A. Sasan, S. Rafatirad, and H. Homayoun, “Enabling micro AI for securing edge devices at hardware level,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3126816.

[A23] F. Kenarangi and I. Partin-Vaisband, “A single-MOSFET analog high resolution-targeted (SMART) multiplier for machine learning classification,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3124940.

[A24] S. T. Chandrasekaran, S. P. Bhanushali, I. Banerjee, and A. Sanyal, “Towards real-time, at-home patient health monitoring using reservoir computing CMOS IC,” *J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, Dec. 2021, doi: 10.1109/JETCAS.2021.3128587.



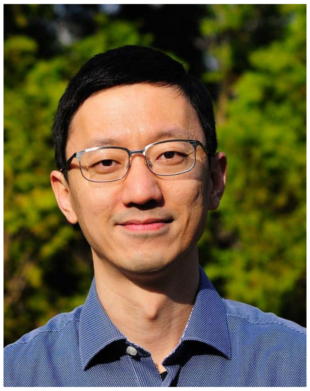
Tinoosh Mohsenin received the M.Sc. degree from Rice University in 2004 and the Ph.D. degree from the University of California at Davis in 2010, both in electrical and computer engineering. She is currently an Associate Professor with the Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, where she is also the Director of the Energy Efficient High Performance Computing Laboratory. She has over 150 peer-reviewed journals and conference publications. Her research focus is on designing energy-efficient embedded processors for machine learning and knowledge extraction computing techniques used in autonomous systems, wearable smart health monitoring, and the Internet of Things. She was a recipient of the NSF CAREER Award in 2017, the Best Paper Award from the ACM Great Lakes VLSI Conference in 2016, and the Best Paper Honorable Award from the IEEE Circuits and Systems Symposium in 2017 for developing processors in biomedical and deep learning. She was a recipient of the ISSCC 2020 Evening Session Award for co-organizing an evening session titled “The Smartest Designer in the Universe.” She was an invited Keynote Speaker of the IEEE AI Circuits and Systems Conference (AICAS), the 14th IEEE Dallas Circuits and Systems Conference (DCAS), and the 27th IEEE International Conference on Electronics Circuits and Systems (ICECS) in 2020. She received ACM Service Award for her contributions as the Program Chair and the General Chair of the 29th and 30th ACM Great Lake VLSI Symposium in 2019 and 2020, respectively. She is currently the Corresponding Guest Editor IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS. She has previously served as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS and IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS.



Inna Partin-Vaisband received the B.Sc. degree in computer engineering and the M.Sc. degree in electrical engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 2006 and 2009, respectively, and the Ph.D. degree in electrical engineering from the University of Rochester, Rochester, NY, USA, in 2015. From 2003 to 2009, she held a variety of software and hardware research and development positions at Tower Semiconductor Ltd., G-Connect Ltd., and IBM Ltd., all in Israel. She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Illinois at Chicago. Her research is currently focused on innovation in the areas of artificially intelligent hardware, hardware security, electronic design automation, and integrated power delivery and management. She is an Associate Editor of the *Microelectronics Journal* and has served for the technical program and organization committees of various conferences.

Houman Homayoun is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of California at Davis. He is also the Director of the National Science Foundation Center for Hardware and Embedded Systems Security and Trust (CHEST). He conducts research in hardware security and trust, applied machine learning and AI, data-intensive computing, and heterogeneous computing, where he has published more than 200 technical papers in prestigious conferences and journals on the subject and directed over U.S. $8 million in research funding from NSF, DARPA, AFRL, NIST, U.S. Congress, and various industrial sponsors. He served as a member for the Advisory Committee, Cybersecurity Research and Technology Commercialization working group in the Commonwealth of Virginia. He is also serving as a Core Group Member for the Hardware Security Body of Knowledge Development Team supported by the Department of Defense. His work received several best paper awards and nominations in various conferences, including ACM GLSVLSI 2016, IEEE ICDM and ICCAD 2019, ISVLSI 2020, and IEEE DCAS 2021. His CHEST Center received congressional support for research in HW security which was included in the 2021 National Defense Authorization Act. He was a recipient of the 2010 National Science Foundation Computing Innovation Fellow Award by CCC/CRA. He chaired and co-chaired major conferences in ACM, including Great Lake Symposium on VLSI. Since 2017, he has been serving as an Associate Editor for IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS.

Jae-Sun Seo received the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, in 2010. From 2010 to 2013, he was with the IBM Thomas J. Watson Research Center, where he worked on cognitive computing chips under the DARPA SyNAPSE Project and energy-efficient integrated circuits for high-performance processors. In 2014, he joined the School of Electrical, Computer and Energy Engineering, Arizona State University, where he is currently an Associate Professor. In 2015, he was a Visiting Faculty at the Intel Circuits Research Laboratory. His current research interests include efficient hardware design of machine learning and neuromorphic algorithms and integrated power management. He was a recipient of the Samsung Scholarship from 2004 to 2009, the IBM Outstanding Technical Achievement Award in 2012, the NSF CAREER Award in 2017, and the Intel Outstanding Researcher Award in 2021.



Xin Zhang (Senior Member, IEEE) received the B.S. degree in electronics engineering from Xi’an Jiaotong University, Xi’an, China, in 2003, and the Ph.D. degree in microelectronics from Peking University, Beijing, China, in 2008. In 2008, he joined the Institute of Industrial Science, The University of Tokyo, Tokyo, Japan, as a Project Researcher. In 2012, he was a Visiting Scholar at the University of California at Berkeley, and then a Project Research Associate at the Institute of Industrial Science, The University of Tokyo. In 2013, he was with the Institute of Microelectronics (IME), Agency for Science, Technology and Research (A*STAR), Singapore, as a Scientist. Since 2014, he has been a Research Staff Member at the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. Since 2021, he has been an Adjunct Professor with the Department of Electrical Engineering, Columbia University, New York, NY, USA. He has authored or coauthored over 50 technical papers and has over 30 filed or issued patents. His research interests include analog circuits, power management circuits, DC-DC converters, AC-DC converters, power devices, magnetics, machine learning hardware/accelerators, computer system architecture, and server system power delivery/packaging/cooling. He is currently serving as a Technical Program Committee Member for the Applied Power Electronics Conference (APEC), the IEEE VLSI Symposium on Technology and Circuits, the IEEE Custom Integrated Circuits Conference (CICC), and the IEEE International Solid-State Circuits Conference (ISSCC). He is an Organizing Committee Member for the IBM IEEE CAS/EDS—AI Compute Symposium. He is also serving as a Technical Advisory Board Member for the Analog-Mixed Signal Circuits, Systems, and Devices (AMS-CSD), and Semiconductor Research Corporation (SRC). He has served as a Guest Editor for IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS and IEEE SOLID-STATE CIRCUITS LETTERS.