Guest Editorial: Communication-Aware Designs and Methodologies for Reliable and Adaptable On-Chip AI SubSystems and Accelerators

THIS Special Issue of the IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS (JETCAS) is dedicated to investigate the latest research about the topic of communication-aware AI subsystems and accelerators. Because of the complex communication, extensive computations, and massive storage requirements, the demand of communication-aware AI designs has been increased in recent years.

This Special Issue covers three comprehensive topics about future AI designs in circuits and systems, such as follows:

1) efficient data movement in contemporary AI subsystems;
2) new design methodology for considering trade-offs among computing engine and data movement/storage unit from energy/power/timing point of view;
3) flexible and reliable communication-aware AI subsystems for future on-chip adaptive learning application.

Starting with the first topic, this Special Issue selects four papers to cover some recent in/near-memory computing techniques to accommodate the computing/processing-in-memory issue. Furthermore, this Special Issue aims to collect the research about efficient task scheduling and data movement methodology from an energy/power/timing optimization point of view. Finally, this Special Issue will collect the research about the novel computing flow, together with flexible data movements and communications, to address the problem of AI computing with flexibility and reliability.

Nine papers are selected to widely cover the aspects of communication-aware AI subsystems. We briefly introduce them according to the three different topics in the next sections.

I. Efficient Data Movement in AI Systems

Memory access latency and overhead of neural network connections have already become the performance bottleneck, and the memory area dominates the total area cost of the memory access dominates power consumption in accelerators of deep neural networks. In this Special Issue, two papers are selected to introduce in-memory computing techniques. In addition, two papers are included to present new techniques to reduce the data movement in AI computing systems.

The first article by Zhu *et al.*, "A communication-aware DNN accelerator on ImageNet using in-memory entry-counting based algorithm-circuit-architecture co-design in 65nm CMOS," presents a communication-aware processing-in-memory deep neural network accelerator, which implements an in-memory entry-counting scheme for low bit-width quantized multiplication-and-accumulations (MACs). To maintain good accuracy on ImageNet, the proposed design adopts a full-stack co-design methodology, from algorithms, circuits, to architectures.

The second article by Agrawal *et al.*, "CASH-RAM: Enabling in-memory computations for edge inference using charge accumulation and sharing in standard 8T-SRAM arrays," introduce an in-memory computing primitive for accelerating dot-products within standard 8T-SRAM caches, using charge-sharing. The inherent parasitic capacitance of the bitlines and sourcelines is used for accumulating analog voltages, which can be sensed for an approximate dot product. The charge sharing approach involves a self-compensation technique which reduces the effects of nonidealities, thereby reducing the errors.

The third article by Samal *et al.*, "Attention-based activation pruning to reduce data movement in real-time AI: A case-study on local motion planning in autonomous vehicles," presents an attention-based feedback for controlling input data, referred to as the activation pruning, that reduces activation maps in early layers of a DNN network which are critical for reducing data movement in real-time AI processing. The proposed approach is demonstrated for coupling RGB and Lidar images to perform real-time perception and local motion planning in autonomous systems.

The fourth article by Lee *et al.*, "SRNPU: An energy-efficient CNN-based super-resolution processor with tile-based selective super-resolution in mobile devices," introduces an energy-efficient convolutional neural network (CNN) based super-resolution (SR) processor, super-resolution neural processing unit (SRNPU), for mobile applications. The SRNPU can support the tile-based selective super-resolution (TSSR) which dynamically selects the proper sized CNN in a tile-by-tile manner and reduce the computational workload of CNN-based SR.

Digital Object Identifier 10.1109/JETCAS.2020.3023568

II. Design Methodology to Optimize Performance and Power Consumption

In the conventional AI systems, the efficiency of the task scheduling depends on the data movement between the memory and the AI computing engine. To investigate the research about efficient task scheduling and find a new design methodology for timing optimization, three papers are selected in this Special Issue.

The first article by Dinelli *et al.*, "MEM-OPT: A scheduling and data re-use system to optimize on-chip memory usage for CNNs on-board FPGAs," introduces MEM-OPT, a scheduling algorithm and data reuse system that aims to optimize on-chip memory usage on-board FPGAs for what concerns input feature maps storage and processing elements multiply and accumulation process. The work presents MEM-OPT implementations results on a Xilinx XC7Z020, including hardware resources, maximum clock frequency, and power consumption.

The second article by Carreras *et al.*, "Optimizing temporal convolutional network inference on FPGA-based accelerators," presents an accelerator architecture, NEURAghe, supporting Temporal Convolutional Networks (TCN), implemented on FPGA-based SoCs. The TCN can freely selectable support kernel sizes and dilated convolutions, with freely selectable dilation rates and stride values. In addition, an optimal execution/scheduling of data-transfers exploiting the specific sequence-based structure of data in TCNs is proposed to optimize the performance.

The third article by Manda *et al.*, "A Latency-optimized reconfigurable NoC for in-memory acceleration of DNNs," points out a problem that crossbar-based in-memory computing may significantly increase the volume of on-chip communication since the weights and activations are on-chip. Hence, the authors propose a methodology to generate an NoC architecture along with a scheduling technique customized for different DNNs. The authors prove mathematically that the generated NoC architecture and corresponding schedules achieve the minimum possible communication latency for a given DNN.

III. Flexible Communication-Aware AI Systems

Due to the fixed data path for the dedicated applications (or neural network models), the computing flows of the contemporary AI systems are usually designed based on a certain interconnection or design parameters, which leads to lower computing flexibility. In this Special Issue, two papers are collected to introduce the novel reconfigurable design for the flexible AI systems.

The first article by Hsiao *et al.*, "Design of a sparsity-aware reconfigurable deep learning accelerator supporting various types of operations," exploits the sparsity in current DNN models and design a sparsity-aware deep learning hardware accelerators that can support efficient computation of various DNN operations, including convolution, de-convolution, depthwise convolution, pointwise convolution, fully connected layer operation, and long-short-term memory (LSTM). Through reconfiguring dataflow and parallelizing different operations, the proposed designs not only improve system performance but also increase hardware utilization with a significant reduction of power consumption in memory accesses and arithmetic computations.

The second article by Guo *et al.*, "Hybrid fixed point-binary deep neural network design methodology for low power object detection," proposes a hybrid fixed point/binary deep neural network design methodology for object detection to achieve low-power consumption by taking advantage of both the fixed-point and binary deep neural networks, which allocates enough bit-width to design the hardware datapath in different layers of deep neural network. The proposed method combines dynamic fixed-point quantization and binarization techniques together to extremely compress the object detection model to result in a compact hybrid fixed-point/binary detection neural network, which achieves lower bandwidth and lower computational complexity.

Acknowledgment

The Guest Editors would like to thank the authors for preparing and contributing their excellent techniques to this Special Issue. They also appreciate all the volunteer reviewers for their generous service to professionally comply with the high-quality standards of JETCAS. We are truly grateful for the guidance and support provided by JETCAS Editors-in-Chief (EiCs), Prof. An-Yeu (Andy) Wu, Associate EiC, Prof. Ho Ching (Herbert) Iu, and Associate EiC for Digital Communications, Prof. Wen-Hsiao Peng, as well as the professional comments made by the Senior Editorial Board. The Guest Editors also thank the IEEE Publishing Operations personnel for their great efforts and patience in finalizing this Special Issue.

Kun-Chih Chen, *Corresponding Guest Editor*
Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung 80421, Taiwan

Masoumeh Ebrahimi, *Guest Editor*
School of Electrical Engineering and
Computer Science
KTH Royal Institute of Technology
16440 Stockholm, Sweden

Maurizio Palesi, *Guest Editor*
Department of Electrical, Electronics and
Computer Engineering
University of Catania
95124 Catania, Italy

Tim Kogel, *Guest Editor*
Virtual Prototyping Verification Group
Synopsys Inc.
52072 Aachen, Germany

Kun-Chih (Jimmy) Chen (Member, IEEE) received the Ph.D. degree from the Graduate Institute of Electronics Engineering, Nation Taiwan University, Taiwan, in 2013. From October 2014 to January 2015, he served as a Post-Doctoral Fellow for the Intel-NTU Connected Context Computing Center working on the development of Green Sensing Platform for Internet of Things (IoT), Reliable Thermoelectric Converter, and Power-Aware Software Defined Network (SDN). From February 2015 to July 2016, he joined the faculty of Electronic Engineering Department, Feng Chia University. He is currently an Assistant Professor with the Computer Science and Engineering Department, National Sun Yat-Sen University. His research interests include multiprocessor SoC (MPSoC) design, neural network learning algorithm design, reliable system design, and VLSI/CAD design. He is a member of ACM. He received the Best Paper Award of International Symposium on VLSI Design, Automation, and Test (VLSI-DAT 2014), the Best Paper Award of International Joint Conference on Convergence (IJCC 2016), and the Best Ph.D. Dissertation Award of the IEEE Taipei Section in 2014. He also received the TCUS Young Scholar Innovation Award, the NSYSU New Faculty Award (three times), the NSYSU Excellent Tutor Award, and the NSYSU Excellent Teaching Award. He served as the Technical Program Committee (TPC) Chair for the International Workshop on Network on Chip Architectures (NoCArc 2018), the General Chair for the International Workshop on Network on Chip Architectures (NoCArc 2019), and a Guest Editor for the *Journal of Systems Architecture* (JSA) and *Nano Communication Network* (NanoComNet). Besides, he also served as the TPC for some major IEEE international conferences, such as ISCAS and SOCC.

Masoumeh (Azin) Ebrahimi (Senior Member, IEEE) received the Ph.D. degree (Hons.) from the University of Turku, Finland, in 2013, and the joint M.B.A. degree from the University of Turku and the EIT-ICT School in 2015. She has led several national and international projects, such as EU-MarieCurie-Vinnova, Academy of Finland, and Vetenskapsrådet (VR). She is currently a Senior Researcher (docent) with the KTH Royal Institute of Technology, Sweden, and an Adjunct Professor with the University of Turku, Finland. Her scientific work contains more than 100 publications, including journal articles, conference papers, book chapters, edited proceedings, and edited Special Issue of journals. Her main areas of interests include interconnection networks and neural network accelerators. She is a member of HiPEAC. She actively acts as a Guest Editor, an Organizer, and the Program Chair in different venues and conferences.

Maurizio Palesi (Senior Member, IEEE) is currently an Associate Professor in computer engineering with the Department of Electrical, Electronics, and Computer Engineering, Università degli Studi di Catania, Catania, Italy. His research activity is focused in the area of embedded systems with particular emphasis on single-chip implementations based on the network-on-chip design paradigm. He is a coauthor of 50 articles in international journals, six book chapters, 70 papers in international conferences/symposium/workshops, and a coauthor of a book. He is a member of the European Network on High Performance and Embedded Architecture and Compilation (HiPEAC). He was a recipient of the Best Paper Award at the Design Automation and Test in Europe (DATE 2011) and the HiPEAC Paper Award 2014. He has served as the General Chair and the TPC Co-Chair for several international conferences and workshops. He served as a Guest Editor for 17 Special Issues in top-level journals, including, *IET Computers and Digital Techniques*, the *ACM Transactions on Embedded Computing Systems*, and the *International Journal of High Performance Systems Architecture*. He serves as an Associate Editor for 12 international journals.

Tim Kogel received the diploma and Ph.D. degrees (Hons.) in electrical engineering from the Aachen University of Technology (RWTH), Aachen, Germany, in 1999 and 2005, respectively. He is currently a Principal Engineer of virtual prototyping with the Synopsys Verification Group. He has authored a book and numerous technical and scientific publications on system-level modeling of SoC platforms. He is also leading a team of applications engineering specialists, responsible for the definition, realization, and deployment of Synopsys' Virtual Prototyping solutions.