Introduction to the Special Issue on Domain-Specific System-on-Chip Architectures and Run-Time Management Techniques

Domain-specific systems-on-chip (DSSoCs), a class of heterogeneous many-core systems, are recognized as a promising approach to narrowing down the performance and energy-efficiency gap between custom hardware accelerators and programmable processors. However, fulfilling this promise depends on successfully addressing a number of fundamental research questions. For instance, given a target domain, a designer must develop a suitable architecture and determine the set of appropriate hardware accelerators. While integrating too many accelerators would increase the design cost, missing critical accelerators can undermine the system performance and energy efficiency. Typically, a rich set of accelerators can dramatically lower the processing times. Hence, the rest of the system components, such as the on-chip communication, must also match the high performance requirements and enable nanosecond-level latencies between the IP blocks and accelerators. DSSoCs must also provide software tools, **application programming interfaces (APIs)**, and accelerator interfaces such that application developers can utilize them efficiently. Finally, a range of runtime management methodologies and algorithms are required to make the best use of the DSSoC resources and power budgets.

This Special Issue presents eleven research papers and a survey targeting these topics selected from over 40 submissions. It represents a remarkable collective effort involving both the academic and industrial research communities. The articles in this issue present novel and impactful solutions to important research problems, including novel device technologies, hardware accelerators, high-level synthesis techniques, design space exploration, scheduling, virtualization, compiler, and test techniques for domain-specific designs. The survey article titled *"Domain-Specific Architectures (DSAs): Research Problems and Promising Approaches"* provides a comprehensive overview of various research directions, outstanding challenges, and promising approaches in DSSoC system design.

Starting from the lowest level of abstraction, the article *"Experimental Demonstration of STT-MRAM-based Nonvolatile Instantly On/Off System: Case Studies"* presents a solution for IoT applications using nonvolatile STT-MRAMs. Experimental results show 15.1% lower power consumption with two orders of magnitude faster data restore time. At the hardware design level, the paper titled *"SHARP: An Adaptable, Energy-Efficient Accelerator for Recurrent Neural Network"* identifies adaptiveness as a key feature missing from existing RNN accelerators. It proposes an intelligent tiled-based dispatching mechanism to efficiently handle the data dependencies. The

ACM Reference format:
Umit Y. Ogras, Radu Marculescu, Trevor N. Mudge, and Michael Kishinevsky. 2023. Introduction to the Special Issue on Domain-Specific System-on-Chip Architectures and Run-Time Management Techniques. *ACM Trans. Embedd. Comput. Syst.* 22, 2, Article 27 (January 2023), 3 pages.
https://doi.org/10.1145/3567834

1539-9087/2023/01-ART27
https://doi.org/10.1145/3567834

proposed design achieves significant speedups compared to the state-of-the-art ASIC, FPGA, and GPU implementations.

This special issue also includes four novel frameworks that facilitate DSSoC design. The paper titled *"FARSI: Facebook AR System Investigator for Agile Domain-Specific System-on-Chip Exploration"* presents an early **design space exploration (DSE)** framework targeting the complex design space of DSSoCs. It also introduces FARSI, an agile system-level simulator with almost four orders of magnitude speedup and 98.5% accuracy compared to Synopsys Platform Architect. Similarly, the paper titled *"Early DSE and Automatic Generation of Coarse-Grained Merged Accelerators"* presents AccelMerger, the first automated methodology to create coarse-grained, control- and data-flow-rich merged accelerators. AccelMerger identifies functions to accelerate and merges accelerators under area and communication constraints, leading up to 16.7× speedup over software implementations and 1.91× speedup over current design space exploration tools. The paper titled *"QUIDAM: A Framework for Quantization-Aware DNN Accelerator and Model Co-Exploration"* presents a highly parameterized quantization-aware DNN accelerator and model co-exploration framework. QUIDAM identifies a wide range of design points where performance per area and energy varies more than 5× and 35×, respectively. It can speed up the design exploration process by 3–4 orders of magnitude while eliminating expensive synthesis and characterization requirements. Finally, the paper titled *"HLS-based High-Throughput and Work-Efficient Synthesizable Graph Processing Template Pipeline"* presents a high-level synthesis framework to generate FPGA-accelerated, high-throughput, synthesizable, and template-based graph processing pipelines. The proposed techniques achieve 50% higher throughput than an OpenCL baseline while improving execution time and power consumption by up to two orders of magnitude.

The next set of papers presents novel software tools and test techniques for domain-specific architectures. The article titled *"AHA: An Agile Approach to the Design of Hardware Accelerators and Compilers"* observes that domain specialization typically entails significant modifications to the software stack to leverage the updated hardware. To address this need, it presents a new approach to enable flexible and evolvable domain-specific hardware specialization based on **coarse-grained reconfigurable arrays (CGRAs)**. The proposed methodology can significantly improve the productivity of hardware-software engineering teams and allow faster deployment of complex accelerator-rich computing systems. Similarly, the paper titled *"CEDR - a Compiler-integrated Extensible DSSoC Runtime"* addresses the architecture, system software, and application development challenges of domain-specific architectures in a unified compile- and runtime workflow. The capabilities of CEDR are demonstrated using real-life signal processing applications on Xilinx Zynq MPSoC-ZCU102 FPGA, x86 systems, Odroid-XU3 hardware, and Nvidia Jetson Xavier platforms. Finally, the article titled *"AdaTest: Reinforcement Learning and Adaptive Sampling for On-chip Hardware Trojan Detection"* presents an adaptive test pattern generation framework for efficient and reliable **Hardware Trojan (HT)** detection using a hardware-software codesign principle. AdaTest achieves up to two orders of test generation speedup and test set size reduction over prior works while achieving the same or higher Trojan detection rate.

The last set of papers addresses virtualization and application design in the context of DSSoCs. The article titled *"Virtualizing a Post-Moore's Law Analog Mesh Processor: The Case of a Photonic PDE Accelerator"* illustrates virtualization using an innovative **Reconfigurable Optical Computer (ROC)**, designed to solve partial differential equations in one shot instead of existing iterative computations. Virtualization of ROCs is required to combat prohibitively large photonic arrays and underutilization. The authors present a lightweight ROC virtualization architecture and methodology that achieves approximately 2× speedup with minimal virtualization overhead. Finally, the paper titled *"A Predictable QoS-aware Memory Request Scheduler for Soft Real-Time Systems"* targets QoS-aware memory controller design in domain-specific soft real-time systems. The

proposed memory controller generates an urgency-based schedule for the contending memory requests based on the allowable response time latencies associated with each request. Extensive experiments using real memory traces demonstrate the practical efficacy of the proposed memory controller design.

We hope the state-of-the-art domain-specific system design solutions compiled in this issue will catalyze further research in this promising area. We thank the authors for their contributions and the anonymous reviewers for their valuable comments and timely work.

Umit Y. Ogras
University of Wisconsin

Radu Marculescu
University of Texas

Trevor N. Mudge
University of Michigan

Michael Kishinevsky
Intel Corporation

Guest Editors