Digital Object Identifier 10.1109/JXCDC.2021.3111236

Special Topic on Monolithic 3-D Integration for Energy-Efficient Computing

AS THE traditional 2-D scaling is approaching its physical limit, there is a great motivation to explore the third dimension for future integrated circuit design. The memory industry has already adopted monolithic 3-D integration (e.g., in 176-layer 3-D NAND Flash), while the 3-D vertical integration structure of logic transistors (e.g., 3-D stacked nanosheets, NMOS on top of PMOS) is emerging for sub-3-nm logic nodes. The other trend is to stack the embedded nonvolatile memories [e.g., resistive random access memory (RRAM), phase change memory (PCM), magnetic random access memory (MRAM), and ferroelectric field-effect-transistor (FeFET)] on top of CMOS using the back-end-of-line (BEOL) processing. Taking one step further, the integration of multiple tiers of active transistors with embedded memories is expected to offer significant improvements in the throughput and energy efficiency thanks to the massive connectivity between logic and memories. Besides the technological breakthroughs, circuit design automation methodologies have become key enablers to optimize the tier partitioning in monolithic 3-D architectures. In addition, heat dissipation should be taken care of by accurate thermal modeling in these monolithic 3-D architectures. New heat spreading materials and advanced embedded cooling techniques are also important.

To constrain the scope and make a differentiation with the concurrent research attempts on advanced packaging with a smaller pitch, here the monolithic 3-D integration is defined as multitier integration of logic transistors and/or memory devices using fine-pitch interconnect vias (<500 nm), as distinguished from the conventional μm-scale through the silicon via (TSV) approach. The monolithic 3-D integration could embrace sequential processing as well as the layer transfer and die stacking as long as the inter-via density is high ($>10^6/\text{mm}^2$).

This Special Topic of the IEEE JOURNAL ON EXPLORATORY SOLID-STATE COMPUTATIONAL DEVICES AND CIRCUITS (JXCDC) called for the recent research advances in the area of the monolithic 3-D integration spanning from materials/devices toward circuits/architectures for energy-efficient computing. The topics of interest of this Special Topic included but were not limited to the following.

1) monolithic 3-D transistors and their circuit and system implications (e.g., complementary FET, NMOS on top of PMOS);
2) BEOL compatible transistors (e.g., based on semiconducting oxides or 2-D materials);
3) BEOL compatible memories (e.g., RRAM, PCM, MRAM, and FeFET);
4) computing-in-3-D NAND or 3-D NOR Flash;
5) silicon recrystallization methods for top tiers;
6) monolithic 3-D integration fabrication methods (e.g., layer transfer, sequential processing, and nano-TSV);
7) prototypes of monolithic 3-D circuit primitives (e.g., 3-D SRAM);
8) design automation of monolithic 3-D architectures (e.g., electronic design automation (EDA) flow for 3-D physical layout);
9) thermal modeling and simulations of monolithic 3-D architectures;
10) heat spreading materials and embedded cooling for monolithic 3-D architectures;
11) system-level design and benchmarking for monolithic 3-D architectures;
12) system-circuit-device co-design for energy-efficient monolithic 3-D architectures.

After the open submission and a rigorous peer-reviewed process, seven articles were selected for this Special Topic. The topics of these seven articles range from EDA flow for 3-D tier partition, 3-D standard cell synthesis, 3-D NAND-based compute-in-memory (CIM), 3-D SRAM-based compare-in-memory, and 3-D systolic machine learning accelerator.

The content of the Special Topic articles is discussed as follows.

A. 3-D EDA DESIGN FLOW

1) Kim *et al.* from Georgia Institute of Technology contributed [A1], where the authors presented a design methodology named pin-in-the-area that assigns block pins at any position inside the boundary of a block using commercial 2-D placement-and-route tools and enables an efficient block implementation and integration for a block-level monolithic 3-D integrated circuits, with examples of implementation at 28 nm.
2) Pentapati *et al.* from Georgia Institute of Technology contributed [A2], where the authors proposed a machine-learning-based prediction algorithm to decrease the discrepancy between the pre- and post-partitioned 3-D design using regression models, achieving a significant reduction in the total negative slack of the test design using the machine-learning model integrated pseudo-3-D flow.
3) Cheng *et al.* from the University of California at San Diego contributed [A3] where the authors proposed a standard cell synthesis framework for multi-row complementary field-effect-transistor (CFET), which

stacks P-FET on N-FET or vice versa and simultaneously solves place-and-route to minimize the cell area by considering single-row and multi-row placement together.

4) Lee *et al.* from the University of California at San Diego contributed [A4] where the authors proposed a standard cell synthesis framework for multi-tier vertical gate-all-around nanowire FET, which aims to obtain the maximum-achievable power, performance, area, and cost (PPAC) benefits.

B. 3-D HARDWARE ACCELERATORS

1) Shim *et al.* from Georgia Institute of Technology contributed [A5], where the authors presented a design of a 3-D NAND-CIM accelerator based on the macro parameters from an industry-grade prototype chip. The deep neural network (DNN) inference performance is evaluated using the DNN+NeuroSim framework.
2) Mathur *et al.* from the University of Texas at Austin contributed [A6], where the authors presented a systematic framework for performing system-level design space exploration of 3-D systolic accelerators with the SRAM for machine-learning workloads. A thermal-aware design is performed to constrain the temperature rise.
3) Ramanathan *et al.* from Pennsylvania State University contributed [A7], where the authors presented SRAM-based 3-D-content addressable memory (CAM) circuit designs for realizing beyond-Boolean in-memory compare operation without any area overheads. Silicon macro has been demonstrated.

APPENDIX: RELATED ARTICLES

[A1] J. Kim, B. W. Ku, J. Yoon, and S. K. Lim, "An effective block pin assignment approach for block-level monolithic 3-D ICs," *IEEE J. Explor. Solid-State Comput.*, vol. 7, no. 1, pp. 26–34, Jun. 2021.

[A2] S. Pentapati, B. W. Ku, and S. Lim, "Machine learning integrated pseudo-3-D flow for monolithic 3-D ICs," *IEEE J. Explor. Solid-State Comput.*, vol. 7, no. 1, pp. 35–42, Jun. 2021.

[A3] C.-K. Cheng, C.-T. Ho, D. Lee, and B. Lin, "Multirow complementary-FET (CFET) standard cell synthesis framework using satisfiability modulo theories (SMT)," *IEEE J. Explor. Solid-State Comput.*, vol. 7, no. 1, pp. 43–51, Jun. 2021.

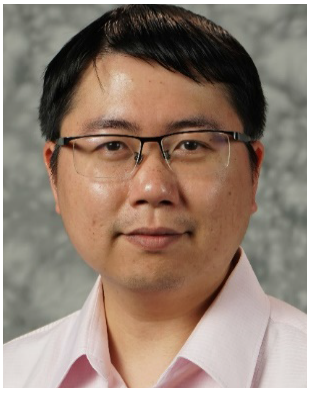
[A4] D. Lee, C.-T. Ho, I. Kang, S. Gao, B. Lin, and C.-K. Cheng, "Many-tier vertical gate-all-around nanowire FET standard cell synthesis for advanced technology nodes," *IEEE J. Explor. Solid-State Comput.*, vol. 7, no. 1, pp. 52–60, Jun. 2021.

[A5] W. Shim and S. Yu, "System-technology codesign of 3-D NAND flash-based compute-in-memory inference engine," *IEEE J. Explor. Solid-State Comput.*, vol. 7, no. 1, pp. 61–69, Jun. 2021.

[A6] R. Mathur, A. K. A. Kumar, L. John, and J. P. Kulkarni, "Thermal-aware design space exploration of 3-D systolic ML accelerators," *IEEE J. Explor. Solid-State Comput.*, vol. 7, no. 1, pp. 70–78, Jun. 2021.

[A7] A. K. Ramanathan *et al.*, "CiM3D: Comparator-in-memory designs using monolithic 3-D technology for accelerating data-intensive applications," *IEEE J. Explor. Solid-State Comput.*, vol. 7, no. 1, pp. 79–87, Jun. 2021.

SHIMENG YU
Georgia Institute of Technology
Atlanta, GA 30332 USA
e-mail: shimeng.yu@ece.gatech.edu



SHIMENG YU (Senior Member, IEEE) received the B.S. degree in microelectronics from Peking University, Beijing, China, in 2009, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 2011 and 2013, respectively.

From 2013 to 2018, he was an Assistant Professor with Arizona State University, Tempe, AZ, USA. He is currently an Associate Professor of electrical and computer engineering with Georgia Institute of Technology, Atlanta, GA, USA. His research interests are nanoelectronic devices and circuits for energy-efficient computing systems. His expertise is on the emerging nonvolatile memories (e.g., resistive random access memory (RRAM) and ferroelectrics) for different applications such as deep learning accelerator, neuromorphic computing, monolithic 3-D integration, and hardware security.

Prof. Yu was a recipient of the National Science Foundation (NSF) Faculty Early CAREER Award in 2016, the IEEE Electron Devices Society (EDS) Early Career Award in 2017, the ACM Special Interests Group on Design Automation (SIGDA) Outstanding New Faculty Award in 2018, and the Semiconductor Research Corporation (SRC) Young Faculty Award in 2019.