Demonstration of model-driven approach to science to systems research for Emerging Logic and Memory Technologies

NEEDS 2012-2017 (needs.nanohub.org)

PI: H.S.P. Wong, Stanford University
May 2017

Background and Goals:
While device technologies are the building blocks for future high performance/energy efficient systems, it is crucial to understand system applications and how device technology advances will impact system-level performance, such that device and system can be co-designed for the optimal performance. The goal of our work in the NEEDS program, was to demonstrate a model-driven approach to science to systems research for emerging logic and memory technologies by using (i) the Stanford NEEDS RRAM model for hyper-dimensional (HD) computing system optimization, and (ii) the MIT NEEDS VS model and the Stanford NEEDS CNFET model to evaluate a 32-bit ARM microprocessor performance at a projected 5-nm technology node.

What was accomplished?
We use a model-driven approach for next-generation system explorations. We have been using the developed RRAM V2.0 compact model to facilitate device-architecture co-design for new neural-inspired computation models. In particular, we have performed extensive system analysis of using 3D RRAM for HD computing, which represents and processes information in high dimensionality. Instead of computing with numbers, HD computing encodes data with high-dimensional (e.g., kilo-bit length) vectors, inspired by the remarkable correspondence of mathematical properties of high-dimensional space to human's perception, memory, and cognition.

The RRAM's unique properties enable efficient HD computing system designs that require inherent randomness in vectors and memory-intensive vector operations. Hence, valuable insights can be further obtained from HD system analysis with RRAM compact model employed. This allows us to dig deep into the interaction between intrinsic device properties (e.g., variability, endurance) and the new computing architecture. Fig. 1(a) shows the modeling of stochastic SET of RRAMs, which is used to produce random vectors in an HD system. By simulating a 36-layer 3D RRAM array, the impact of sparsity on energy efficiency and system recognition accuracy (for language recognition tasks) is explored (Fig. 1(b)). Furthermore, system-level reliability analysis shows that RRAM-centric HD systems are resilient to hard memory errors induced by endurance failures (Fig. 1(c)).

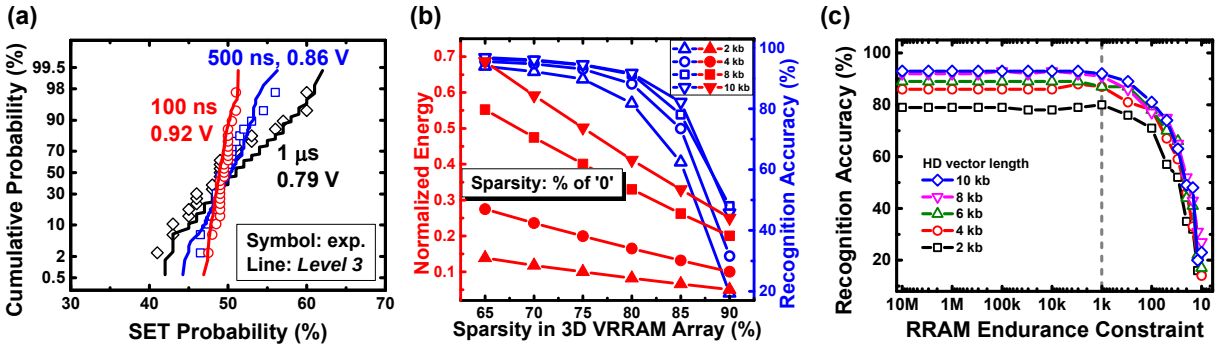

Fig. 1. (a) Measured and modeled RRAM SET probability (P_{SET}) distributions obtained from multiple RRAM's andvarious programming conditions. (b) Simulated energy and accuracy vs. sparsity (% of '0's) in 3D VRRAM array initialized by tuning RRAM SET probabilities, which is supported by the RRAM model. (c) Simulated recognition accuracy of the HD language recognition system as a function of RRAM endurance and HD vector length.

In a practical VLSI system, transistors are connected through the interconnects, so properly taking into account the wire parasitic resistance and capacitance is important. In collaboration with ARM, we developed a script-based tool to implement VLSI circuit modules (here we used an ARM microprocessor as an example) from underlying transistor and interconnect technologies through layers of abstraction (e.g. layout-dependent parasitic extraction and chip-level signal routing optimization) to obtain the system performance (Fig. 2a). With such an automatic design tool, device-system co-optimization is enabled to harvest the most benefit out of a technology. Fig. 2b shows the optimization of the transistor extension length for the minimal core-level energy-delay product with a fixed gate pitch of 36 nm for Si FinFET and MoS_2 FET based on NEGF simulations, using the MVS model as the core of the transistor compact model. Fig. 2c shows the optimal core-level energy-frequency tradeoff for the projected Si FinFET and theoretical MoS_2-FET with different specific contact resistivities, indicating the importance of improving the metal-to-MoS_2 contacts.

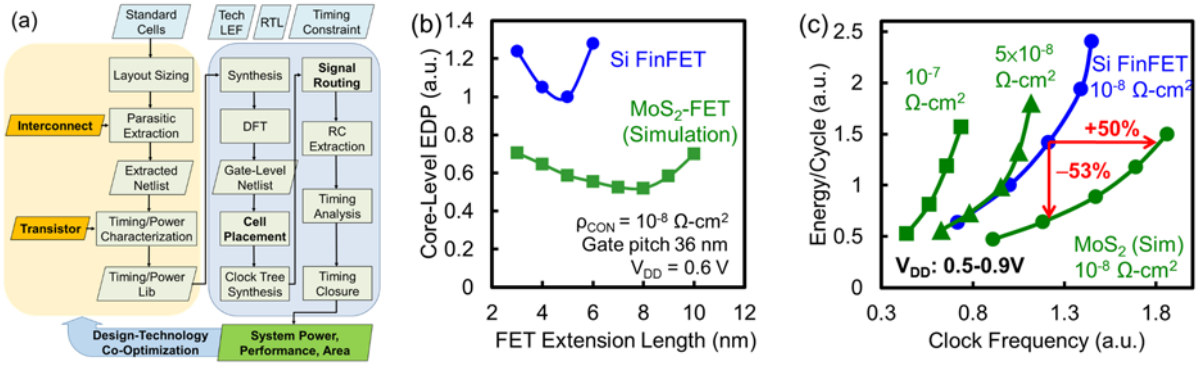

Fig. 2. (a) Device-to-system implementation and optimization flow. (b) Optimization of transistor extension and contact lengths with fixed gate pitch of 36 nm for Si FinFET and MoS_2 planar FET for minimal core-level energy-delay product. (c) Core-level energy per cycle vs. clock frequency for Si FinFET and MoS_2-FET with different specific contact resistivities labeled besides the curves (unit: Ω-cm^2).

Why was it important?
Through these demonstrations, we showed the importance of device-system co-optimization, which requires an understanding of both systems and device properties through compact modeling. Without the consideration of systems and applications, the impact of device performance and variability cannot be well understood; whereas insightful system-level analysis would have been impossible without an accurate physical models.

Three students (Chi-Shuen Lee, Haitong Li, Zizhen Jiang) have been directly involved in this work and several other students from our collaborators have been crucial to our success (Gage Hills and Tony Wu from Prof. Subhasish Mitra's group, and Abbas Rahimi from Prof. Jan Rabaey's group). We have strong international collaborations with the National Nanodevice Laboratory (Taiwan) and with National Chiao Tung University (Taiwan) and Peking University (China). Our work was presented and discussed at the Computing Community Consortium (CCC) 2016 Nanotechnology-Inspired Information Processing Systems of the Future (NPS) Workshop in August 2016 (report available at http://cra.org/ccc/wp-content/uploads/sites/2/2016/04/15591-CRA-Nanotech-workshop-report-v4.pdf). Our results have been incorporated in the research plans and directions of ARM and IMEC.

References (* denotes work supported or partially supported by NEEDS)

[1] H. Li, T. F. Wu, A. Rahimi, K.-S. Li, M. Rusch, C.-H. Lin, J.-L. Hsu, M. M. Sabry, S. B. Eryilmaz, J. Sohn, W.-C. Chiu, M.-C. Chen, T.-T. Wu, J.-M. Shieh, W.-K. Yeh, J. M. Rabaey, S. Mitra, and H.-S. P. Wong, "Hyperdimensional computing with 3D VRRAM in-memory kernels: device-architecture co-design for energy-efficient, error-resilient language recognition" IEEE International Electron Devices Meeting (IEDM), pp. 16.1, 2016. (*)

[2] H. Li, K.-S. Li, C.-H. Lin, J.-L. Hsu, W.-C. Chiu, M.-C. Chen, T.-T. Wu, J. Sohn, S. B. Eryilmaz, J.-M. Shieh, W.-K. Yeh, and H.-S. P. Wong, "Four-layer 3D vertical RRAM integrated with FinFET as a versatile computing unit for brain-inspired cognitive information processing" Symposium on VLSI Technology (VLSI), pp. 1-2, 2016.(*)

[3] C.-S. Lee, B. Cline, S. Sinha, G. Yeric, H.-S. P. Wong, "32-bit Processor Core at 5-nm Technology: Analysis of Transistor and Interconnect Impact on VLSI System Performance" pp. 28.3, *IEEE International Electron Devices Meeting* (*IEDM*), 2016.(*)

[4] C.-S. Lee, E. Pop, A. Franklin, W. Haensch, and H.-S. P. Wong, "A Compact Virtual-Source Model for Carbon Nanotube Field-Effect Transistors in the Sub-10-nm Regime—Part I: Intrinsic Elements" *IEEE Trans. Electron Devices*, vol. 62, no. 9, pp. 3061-3069, Sep. 2015. (*)

[5] C.-S. Lee, E. Pop, A. Franklin, W. Haensch, and H.-S. P. Wong, "A Compact Virtual-Source Model for Carbon Nanotube Field-Effect Transistors in the Sub-10-nm Regime—Part II: Extrinsic Elements and Performance Assessment" *IEEE Trans. Electron Devices*, vol. 62, no. 9, pp. 3070-3078, Sep. 2015.(*)

[6] Z. Jiang, Z. Wang, X. Zheng, S. Fong, S. Qin, H. -Y. Chen, C. Ahn, J. Cao, Y. Nishi, and H. -S. Philip Wong, “Microsecond Transient Thermal Behavior of HfOx-based Resistive Random Access Memory Using a Micro Thermal Stage (MTS)" 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 21.3.1-21.3.4.(*)

[7] Z. Jiang, Y. Wu, S. Yu, L. Yang, K. Song, Z. Karim, and H.-S. P. Wong, "A Compact Model for Metal–Oxide Resistive Random Access Memory With Experiment Verification" *IEEE Trans. Electron Devices*, vol. 63, no. 5, pp. 1884-1892, May 2016.

[8] Z. Jiang, S. Yu, Y. Wu, J. H. Engel, X. Guan and H. S. P. Wong, "Verilog-A compact model for oxide-based resistive random access memory (RRAM)" 2014 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), Yokohama, 2014, pp. 41-44. (*)