Session 28: Circuit Device Interaction – Emerging Memory System and Advanced Heterogeneous Integration
Wednesday, December 17, 9:00 a.m.
Continental Ballroom 5
Co-Chairs: Yves LaPlanche, ARM
Robert Wu, Broadcom

9:05 a.m.
28.1 Low Power and High density STT-MRAM for Embedded Cache Memory using Advanced Perpendicular MTJ Integrations and Asymmetric Compensation Techniques, K. Ikegami, H. Noguchi, C. Kamata, M. Amano, K. Abe, K. Kushida, E. Kitagawa, T. Ochiai, N. Shimomura, S. Itai, D. Saida, C. Tanaka, A. Kawasumi, H. Hara, J. Ito and S. Fujita, Toshiba Corporation

We investigate mobile CPU power reduction by perpendicular spin transfer torque magnetic random access memory (STT-MRAM) with regard to write power redution and integration process. Damping constant of MTJ is known to determine write current. We develop MTJ with low damping constant and showed that it can reduce area and power of cache memory of processor compared SRAM or eDRAM. In terms of integration process, MTJ is known to lost switching characteristic above sintering temperature. To integrate high-performance CMOS and low power STT-MRAM, we propose MTJ-Last process and showed that area and delay overhead of fabricating MTJ on upper metal layer is negligible for cache application. To further reduce power with circuit technique, we investigate asymmetric compensation of reference layer (RL) magnetic field. By utilizing low power MTJ and MTJ-Last process, we can reduce cache power by 60% with only 7 % performance degradation compared to SRAM cache.

9:30 a.m.
28.2 Challenge of MOS/MTJ-Hybrid Nonvolatile Logic-in-Memory Architecture in Dark-Silicon Era, T. Hanyu, D. Suzuki, A. Mochizuki, M. Natsui, N. Onizawa, T. Sugibayashi, S. Ikeda, T. Endoh and H. Ohno, Tohoku University, NEC Corporation

In this presentation, I introduce "nonvolatile logic-in-memory architecture" to solving performance-wall and power-wall problems in the present CMOS-only-based logic-LSI processors. The use of magnetic tunnel junction devices combined with a CMOS-gate style makes it possible to achieve a high-performance and ultra-low-power logic LSI. Some concrete examples such as nonvolatile ternary content-addressable memory (TCAM) and nonvolatile field-programmable gate array (FPGA) using the proposed method allow you to achieve the desired performance improvement compared to a corresponding CMOS-only-based realization.

9:55 a.m.
28.3 Technology and Circuit Optimization of Resistive RAM Memory Arrays for Low-Power, Reproducible Operation, D. Sekar, B. Bateman, U. Raghuram, S. Bowyer, Y. Bai, M. Calarrudo, P. Swab, J. Wu, S. Nguyen, N. Mishra, R. Meyer, M. Kellam, B. Haukness, C. Chevallier, H. Wu, H. Qian, F. Kreupl and G. Bronner, Rambus, *Tsinghua University, **Technical University of Munich

For Resistive RAM (RRAM), reproducibility in large arrays requires control of capacitive surge currents during programming. In this work, we present results from a 256kb RRAM chip which demonstrate how device optimization in conjunction with innovative circuits can control surge currents due to inherent cell and array parasitics. We propose a fab-friendly TiN/conductive TaOx/ HfO_2/TiN bi-layer RRAM that gives 2x lower power and improves variability and switching yield compared to conventional HfO_2 RRAMs. Our studies reveal new insight that current crowding in RRAMs with conductive metal oxide electrodes improves thermal efficiency and damps surge currents, leading to the improved characteristics. A novel circuit to control surge current is proposed and demonstrated that improves write current by 40% and endurance by 63%. Switching, endurance and data retention results for the 256kb chip are presented.

10:20 a.m.
28.4 Variability-tolerant Convolutional Neural Network for Pattern Recognition Applications Based on OxRAM Synapses, D. Garbin, O. Bichler*, E. Vianello, Q. Rafhay**, C. Gamrat*, L. Perniola, G. Ghibaudo** and B. DeSalvo, CEA, LETI, Minatec, *CEA, LIST, **IMEP-LAHC

Software implementations of artificial Convolutional Neural Networks (CNNs), taking inspiration from biology, are at the state-of-the-art for Pattern Recognition (PR) applications and they are successfully used in commercial products. However, they require power-hungry CPU/GPU to perform convolution operations based on computationally expensive

sums of multiplications. This hinders their integration in portable devices. Some full CMOS based hardware implementations of CNN have been suggested, but they still require the computation of multiplications. In this work, we present for the first time to our knowledge a spike-based hardware implementation of CNN using HfO_2 based OxRAM devices as binary synapses. OxRAM devices are chosen for their low switching energy and promising endurance performance. We perform an experimental and theoretical study of the impact of programming conditions at both device and system levels. A complex visual pattern recognition application is demonstrated with a spike-based hierarchical CNN, inspired from the mammalian visual cortex organization. A high accuracy (pattern recognition rate >94%) is obtained for all the tested programming conditions, even if the variability associated to weaker programming conditions is larger.

10:45 a.m.
28.5 3D Synaptic Architecture with Ultralow sub-10 fJ Energy per Spike for Neuromorphic Computation, I-T. Wang, Y.-C. Lin, Y.-F. Wang, C.-W. Hsu and T.-H. Hou, National Chiao Tung University

A high-density 3D synaptic architecture based on self-rectifying $Ta/TaO_x/TiO_2/Ti$ RRAM is proposed as an energy- and cost-efficient neuromorphic computation hardware. The device shows excellent analog synaptic features that can be accurately described by the physical and compact models. Ultra-low energy consumption comparable to that of a biological synapse (<10 fJ/spike) has been demonstrated for the first time.

11:10 a.m.
28.6 High Dependable 3-D Stacked Multicore Processor System Module Fabricated using Reconfigured Multichip-on-wafer 3-D Integration Technology, K.-W. Lee, H. Hashimoto, M. Onishi, Y. Sato, C. Nagai, M. Murugesan, J.-C. Bea, T. Fukushima, T. Tanaka and M. Koyanagi, Tohoku University

A high dependable 3-D stacked multicore processor module composed of the 4-layer 3-D stacked multicore processor chip and the 2-layer 3-D stacked cache memory chip is implemented using novel reconfigured multichip-on-wafer 3-D integration and reliable backside TSV technologies for the first time. Tier boundary scan, self-repair circuits, and memory BIST circuits in the 4-layer 3-D stacked multicore processor chip and the basic read/write functions of memory circuits in the 2-layer 3-D stacked cache memory chip are successfully evaluated. X-ray computed tomography (CT) scanning technology is proposed as a non-destructive failure analysis method to characterize high-density TSVs integration and bump joining qualities in the 3-D stacked chip.

11:35 a.m.
28.7 Pairwise Coupled Hybrid Vanadium Dioxide-MOSFET (HVFET) Oscillators for Non-Boolean Associative Computing, N. Shukla, A. Parihar*, M. Cotter, M. Barth, X. Li, N. Chandramoorthy, D.G. Schlom**, V. Narayanan, A. Raychowdhury* and S. Datta, The Pennsylvania State University, *Georgia Institute of Technology, **Cornell University

In this work, we provide (i) the first experimental demonstration of coupled Hybrid VO_2 FET oscillators with input programmable synchronization; (ii) hardware platform capable of efficiently computing a fractional distance norm for associative data computing in high dimensional space (iii) projection of a ~20X reduction in power consumption over CMOS.

12:00 p.m.
28.8 Hybrid CMOS/BEOL-NEMS Technology for Ultra-Low-Power IC Applications, N. Xu, J. Sun, I-R. Chen, L. Hutin, Y. Chen, J. Fujiki, C. Qian and T.-J. King Liu, University of California, Berkeley

3-D NEM relays are proposed in order to leverage an advanced CMOS BEOL technology to reduce the die area and power consumption of digital logic and memory circuits.