Manufacturing Science and Technology Group
Room 202B - Session MS+MI+RM-TuM

IoT Session: Challenges of Neuromorphic Computing and Memristor Manufacturing (8:00-10:00 am)/Federal Funding Opportunities (11:40 am-12:20 pm)

Moderators: Christopher L. Hinkle, University of Texas at Dallas, Sean Jones, National Science Foundation, Alain C. Diebold, SUNY College of Nanoscale Science and Engineering

8:00am **MS+MI+RM-TuM-1 ReRAM – Fabrication, Characterization, and Radiation Effects***, **David Hughart**, R Jacobs-Gedrim, K Knisely, N Martinez, C James, B Draper, E Bielejec, G Vizkelethy, S Agarwal,* Sandia National Laboratories*; H Barnaby,* Arizona State University*; M Marinella,* Sandia National Laboratories **INVITED**

Resistive switching properties in transition metal oxides and other thin films have been an active area of research for their use in nonvolatile memory systems as Resistive Random Access Memory (ReRAM). ReRAM is a candidate for storage class memory technologies, and studies have also revealed a high degree of intrinsic radiation hardness making digital ReRAM a candidate for radiation-hardened memory applications. Analog ReRAM has also generated interest from the neuromorphic computing community for use as a weight in neural network hardware accelerators.

One of the manufacturing challenges for the valence change memory (VCM) type of ReRAM has been the development of substoichiometric switching layer films. Physical vapor deposited (PVD) substoichiometric TaO_x films are an attractive option for a VCM switching layer because they are complementary-metal-oxide-silicon (CMOS) compatible and are deposited at low temperatures. However, control of the oxygen partial pressure to produce substoichiometric TaO_x films cannot be directly achieved through flow control because the oxygen consumption by the Ta target and chamber surfaces is nonlinear as the chamber transitions from metal to insulator conditions. The oxygen partial pressure can be controlled using a feedback system, though feedback-assisted deposition techniques are difficult to regulate, making them ill-suited to production. One alternative to a feedback system is to deposit a higher stoichiometry TaO_x film, deposited in a more stable flow-partial pressure chamber regime, and use annealing to drive Ta into the film to achieve the desired stoichiometry. Here, we compare switching layers fabricated using both techniques, and discuss the relative merits of each technique. The devices are manufactured in crossbar arrays to be testable by automatic probers, enabling the collection of large scale yield and performance data sets across process splits.

Manufacturing improvements enabled fabrication of analog ReRAM with characteristics suitable for neuromorphic computing applications. The performance of a TaO_x ReRAM based hardware accelerator at image classification accuracy after training was evaluated. The classification accuracy showed little degradation in initial radiation tests, suggesting analog ReRAM may be suitable for neuromorphic computing applications in radiation environments as well.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

8:40am **MS+MI+RM-TuM-3 Memristive Synapses – Tuning Memristors for Performance and CMOS Integration***, **Nathaniel Cady**,* SUNY Polytechnic Institute **INVITED**

Neuromorphic computing systems can achieve learning and adaptation in both software and hardware. The human brain achieves these functions via modulation of synaptic connections between neurons. Memristors, which can be implemented as Resistive Random Access Memory (ReRAM), are a novel form of non-volatile memory expected to replace a variety of current memory technologies and enable the design of new circuit architectures. Memristors are a prime candidate for so-called "synaptic devices" to be used in neuromorphic hardware implementations. A variety of challenges persist, however, for integrating memristors with CMOS, as well as for tuning device electrical performance. My research group has developed a fully CMOS-compatible integration strategy for ReRAM-based memristors on a 300 mm wafer platform, which can be implemented in both front-end-of-line (FEOL) and back-end-of-line (BEOL) configurations. With regard to memristor performance, we are focusing on strategies to reduce stochastic behavior during both binary and analog device switching. This is a key metric for neuromorphic applications, as variability in device conductance state directly influences the ultimate number of levels (weights) that can be implemented per synapse. Using a two pronged approach, we have developed device operational parameters to maximize analog performance, while also tuning the ReRAM materials stack and processing conditions to reduce stochasticity and optimize switching parameters (forming, set, and reset).

9:20am **MS+MI+RM-TuM-5 Analog In-Memory Computing for Deep Neural Network Acceleration***, **Hsinyu Tsai**, S Ambrogio, P Narayanan, R Shelby, G Burr,* IBM Almaden Research Center **INVITED**

Neuromorphic computing represents a wide range of brain-inspired algorithms that can achieve various artificial intelligence (AI) tasks, such as classification and language translation. By taking design cues from the human brain, such hardware systems could potentially offer an intriguing Non-Von Neumann (Non-VN) computing paradigm supporting fault-tolerant, massively parallel, and energy-efficient computation.

In this presentation, we will focus on hardware acceleration of large Fully Connected (FC) DNNs in phase change memory (PCM) devices [1]. PCM device conductance can be modulated between the fully crystalline, low conductance, state and the fully amorphous state by applying voltage pulses to gradually increase the crystalline volume. This characteristic is crucial for memory-based AI hardware acceleration because synaptic weights can then be encoded in an analog fashion and be updated gradually during training [2,3]. Vector matrix multiplication can then be done by applying voltage pulses at one end of a memory crossbar array and accumulating charge at the other end. By designing the analog memory unit cell with a pair of PCM devices as the more significant weights and another pair of memory devices as the less significant weights, we achieved classification accuracies equivalent to a full software implementation for the MNIST handwritten digit recognition dataset [4]. The improved accuracy is a result of larger dynamic range, more accurate closed loop tuning of the more significant weights, better linearity and variation mitigation of the less significant weight update. We will discuss what this new design means for analog memory device requirements and how this generalizes to other deep learning problems.

1. G. W. Burr et al., "Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element," IEDM Tech. Digest, 29.5 (2014).

2. S. Sidler et al., "Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: impact of conductance response," ESSDERC Proc., 440 (2016).

3. T. Gokmen et al., "Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations," Frontiers in Neuroscience, 10 (2016).

4. S. Ambrogio et al., "Equivalent-Accuracy Accelerated Neural Network Training using Analog Memory," Nature, to appear (2018).

11:00am **MS+MI+RM-TuM-10 Computation Immersed in Memory: Integrating 3D vertical RRAM in the N3XT Architecture***, **Weier Wan**, W Hwang, H Li, T Wu, Y Malviya,* Stanford University*; M Aly,* Nanyang Technological University, Singapore*; S Mitra, H Wong,* Stanford University **INVITED**

The rise of data-abundant computing, where massive amount of data is processed in applications such as machine learning, computer vision and natural language processing, demands highly energy-effcient computing systems. However, the limited connectivity between separated logic and memory chips in conventional 2D system results in majority of program execution time and energy spent at memory access.The Nano-Engineered Computing Systems Technology (N3XT) [1] approach overcomes these memory bottlenecks by monolithically integrating interleaving layers of memory and logic on the same chip, and leveraging nano-scale interlayer vias (ILVs) to provide ultra-dense connectivity between logic and memory.

The metal oxide resistive switching memory (RRAM) [2] offers non-volatility, good scalability, and monolithic 3D integration, making it a good candidate as on-chip high-capacity main memory and storage in the N3XT system. Our experimentally calibrated studies show that a N3XT system with RRAM as digital storage and CNFET as logic devices could achieve 2-3 orders of magnitude improvement in energy efficiency (product of execution time and energy) in a wide range of applications (e.g. PageRank, deep neural network inference) compared to a conventional 2D system. Such 3D nano-system has also been experimentally demonstrated with

RRAM, CNFET and CMOS monolithically integrated to perform in-situ ambient gas classification [3] and hyper-dimensional computing [4].

Besides offering substantial benefits for conventional digital systems, the monolithic integration of RRAM and logic devices also enables "in-memory computing", where computation is performed in the memory itself without explicitly moving data between memory and logic. Various types of in-memory computing operations could be performed using RRAM arrays, including analog multiply-accumulate and bit-wise logical operations. We perform system modeling that models program scheduling, communication and routing, and memory array and its peripheral circuits design on various operations to study their benefits and bottlenecks from application level. In particular we analyze the in-memory vector-matrix multiplication for deep neural network inference and bit-wise operations in 3D vertical-RRAM for hyper-dimensional computing. We show that with algorithm-architecture co-design, RRAM-based in-memory computing could further improve energy and area efficiency compared to digital implementation in a 3D monolithically integrated system.

[1] M.M.S. Aly et al., IEEE Computer, 2015. [2] H.-S P. Wong et al., Proc. IEEE, 2012. [3] M.M. Shulaker et al., Nature, 2017. [4] T. Wu et al., ISSCC, 2018.

11:40am **MS+MI+RM-TuM-12 Materials for the Second Quantum Revolution***, **Tomasz Durakiewicz,*** Los Alamos National Laboratory

Onset of the second quantum revolution is marked by proliferation of quantum technologies. Still mostly in the laboratory R&D phase, but likely to emerge soon as a growing sector of general consumer technology, quantum devices require constant supply of novel functional quantum materials. The current paradigm of meticulous long-term studies to understand fundamental properties in detail and be able to model them ab initio is unlikely to disappear; however, the rapid growth of technology may require modification of classical approach by accelerated discovery process aided by machine learning, data mining, and ability to model, synthesize and test novel materials quickly. In this presentation we will discuss opportunities and current developments in select classes of quantum materials, like low-dimensional materials, strongly correlated systems and topological insulators, and the role NSF plays in this rapidly growing area.

12:00pm **MS+MI+RM-TuM-13 SynBio(medicine): The Intersection Biomaterials and Living Systems***, **David Rampulla,*** National Institute of Health

The National Institute for Biomedical Imaging and Bioengineering (NIBIB) has long supported the development of biomaterials as platform technologies with broad biomedical application and has recently started a program in Synthetic Biology. This presentation will discuss the biomaterials portfolio at NIBIB with a specific focus on the use of synthetic biology approaches to engineer next generation materials for biomedicine. The talk will also highlight specific funding opportunities of interest and discuss some strategies for navigating the NIH application process.

Author Index

Bold page numbers indicate presenter