Session 3: Circuit and Device Interaction - Device and Algorithm Co-design for Neuromorphic and In-memory Computing
Monday, December 3, 1:30 PM
Grand Ballroom B
Co- Chairs: R. Tetzlaff, TU Dresden
T-H Hou, National Chiao Tung University

1:35 PM - 2:00 PM
3.1 Exploiting Hybrid Precision for Training and Inference: A 2T-1FeFET Based Analog Synaptic Weight Cell, *X. Sun, P. Wang, K. Ni*, S. Datta*, and S. Yu**, Arizona State University, *University of Notre Dame, **Georgia Institute of Technology*

In-memory computing with analog non-volatile memories (NVMs) can accelerate both the in-situ training and inference of deep neural networks (DNNs) by parallelizing multiply-accumulate (MAC) operations in the analog domain. However, the in-situ training accuracy suffers from unacceptable degradation due to undesired weight-update asymmetry/nonlinearity and limited bit precision. In this work, we overcome this challenge by introducing a compact Ferroelectric FET (FeFET) based synaptic cell that exploits hybrid precision for in-situ training and inference. We propose a novel hybrid approach where we use modulated "volatile" gate voltage of FeFET to represent the least significant bits (LSBs) for symmetric/linear update during training only, and use "non-volatile" polarization states of FeFET to hold the information of most significant bits (MSBs) for inference. This design is demonstrated by the experimentally validated FeFET SPICE model and co-simulation with the TensorFlow framework. The results show that with the proposed 6-bit and 7-bit synapse design, the in-situ training accuracy can achieve ~97.3% on MNIST dataset and ~87% on CIFAR-10 dataset, respectively, approaching the ideal software based training.

2:00 PM - 2:25 PM
3.2 Analog Computing for Deep Learning: Algorithms, Materials & Architectures (Invited), *W. Haensch, IBM Research*

Analog, or neuromorphic, computing for Deep Learning (DL) utilizes the fact that matrix manipulations that are inherent in the back-propagation algorithm, can be performed at constant time, in parallel, on arrays with nonvolatile memory (NVM) elements in which the weights are encoded. We discuss the NVM material requirements that need to be met to achieve a classification accuracy on par with the conventional digital approaches, discuss advantages and drawbacks, and highlight opportunities that can take advantage using analog arrays.

2:25 PM - 2:50 PM
3.3 Hardware Acceleration of Simulated Annealing of Spin Glass by RRAM Crossbar Array, *J. H. Shin, Y. Jeong, M. A. Zidan, Q. Wang, and W. D. Lu, University of Michigan*

Simulated annealing was successfully accelerated by in-memory computing hardware/software package using RRAM crossbar arrays to solve spin glass problems. Ta2O5-based RRAM array and stochastic Cu-based CBRAMs were utilized for calculation of the Hamiltonian and decision of spin-flip events, respectively. A parallel spin-flip strategy was demonstrated to further accelerate simulated annealing.

2:50 PM Coffee Break

3:15 PM - 3:40 PM
3.4 Demontration of Generative Adversarial Network by Intrinsic Random Noises of Analog RRAM Devices, *Y. Lin, H. Wu, B. Gao, P. Yao, W. Wu, Q. Zhang, X. Zhang, X. Li,*

F. Li, J. Lu*, G. Li**, S. Yu***, and H. Qian, Tsinghua University, *Hunan University, **Huawei Technologies Co., LTD., ***Georgia Institute of Technology*

For the first time, Generative Adversarial Network (GAN) is experimentally demonstrated on 1kb analog RRAM array. After online training, the network can generate different patterns of digital numbers. The intrinsic random noises of analog RRAM device are utilized as the input of the neural network to improve the diversity of the generated numbers. The impacts of read and write noises on the performance of GAN are analyzed. Optimized methodology is developed to mitigate the excessive noise effect on RRAM based GAN. This work proves that RRAM is suitable for the application of GAN. It also paves a new way to take advantage of the non-ideal effects of RRAM devices.

3:40 PM - 4:05 PM

3.5 Error-Resilient Analog Image Storage and Compression with Analog-Valued RRAM Arrays: An Adaptive Joint Source-Channel Coding Approach, *X. Zheng, R. Zarcone*, D. Paiton*, J. Sohn, W. Wan, B. Olshausen* and H. -S. Philip Wong, Stanford University, *University of California, Berkeley*

We demonstrate by experiment an image storage and compression task by directly storing analog image data onto an analog-valued RRAM array. A joint source-channel coding algorithm is developed with a neural network to encode and retrieve natural images. The encoder and decoder adapt jointly to the statistics of the images and the statistics of the RRAM array in order to minimize distortion. This adaptive joint source-channel coding method is resilient to RRAM array non-idealities such as cycle-to-cycle and device-to-device variations, time-dependent variability, and non-functional storage cells, while achieving a reasonable reconstruction performance of ~ 20 dB using only 0.1 devices/pixel for the analog image.