Session 15: Circuit and Device Interaction - Emerging Devices for Neural Network and IoT
Tuesday, December 4, 9:00 AM
Continental Ballroom 6
Co-Chairs: C-H Shen, National Nano Device Laboratories
J. Deng, Qualcomm

9:05 AM - 9:30 AM

For the first time, ultra-low power ferroelectric FinFET-based monolithic 3D+-IC technology was demonstrated for near memory computing (NMC) circuit. Key enablers are ICP-SiO2 interfacial layer, doped hafnia ferroelectric gate dielectric layer (HfZrO2), and far-infrared laser activation. The proposed stackable 3D NC-FinFETs thus fabricated exhibit record-low sub-threshold swing (NC-nFinFET: 45mV/dec and NC-pFinFET: 50mV/dec) and high Ion/Ioff (>10^6) that enable ultra-low power operation (Vdd=100mV) of CMOS inverter and SRAM. Moreover, above mentioned features of NC-FinFETs and the differential output of SRAM readout enable 50+% area reduction in the near-memory computing circuitry.

9:30 AM - 9:55 AM
15.2 First Demonstration of Ge Ferroelectric Nanowire FET as Synaptic Device for Online Learning in Neural Network with High Number of Conductance State and Gmax/Gmin, W. Chung, M. Si, and P. D. Ye, Purdue University

In this paper, optimum weight update scheme for improved linearity and asymmetry of channel conductance potentiation and depression in a Germanium ferroelectric (FE) nanowire FET (NWFET) was experimentally demonstrated and simulated for the first time. It was found that -5 V, 320 pulses and +5 V, 256 pulses both with 50 ns pulse width were the optimum pulsing conditions for potentiation and depression process, respectively. With the optimized scheme, non-linearity for potentiation and depression were extracted to be αp = 1.22 and αd = -1.75, respectively resulting in asymmetry (|αp - αd|) of 2.97 based on models embedded in MLP simulator and NeuroSim [1]. Gmax/Gmin ratio (few hundreds) and number of conductance states (> 256) are both very large. 9 alternating consecutive conductance updates (potentiation followed by depression) were executed to observe variability in conductance profiles. Multilayer perceptron neural network was simulated over 1 million MNIST images with extracted experimental parameters which yielded in online learning accuracy of ~ 88 %.

9:55 AM - 10:20 AM

The potential of embedded STT-MRAM technology for designing large-scale multiply-and-accumulation (MAC) array circuits are evaluated by comprehensive and holistic design-technology co-optimizations. After careful calibrations with experimental data, post-layout circuit simulations together with GPU-enabled massively parallel Monte Carlo evaluations are conducted to guarantee the designs at rare failure rates. With all critical device and design non-idealities included, architectural emulations are performed to examine the hardware neural network (HNN)’s accuracies and estimate system-level power, performance
and area specs. Results indicate the amount of process variation, parasites and error levels to control in order to achieve a feasible solution for STT-MRAM based HNNs.

10:20 AM Coffee Break

10:45 AM - 11:10 AM
15.4 A 68 Parallel Row Access Neuromorphic Core with 22K Multi-Level Synapses Based on Logic-Compatible Embedded Flash Memory Technology, M. Kim, J. Kim, G. Park, L. Everson, H. Kim, S. Song, **, S. Lee**, and C. H. Kim, University of Minnesota, **Anaflash Inc

A neuromorphic core utilizing logic-compatible embedded flash cells as non-volatile multi-level synapses is demonstrated in a 65nm CMOS process. A carefully-designed program-verify sequence and bitline voltage regulation scheme allow a record high number of rows to be activated in parallel without compromising the accuracy of the handwritten digital recognition application.

11:10 AM - 11:35 AM
15.5 Interchangeable Hebbian and Anti-Hebbian STDP Applied to Supervised Learning in Spiking Neural Network, C-C Chang, P-C Chen, B. Hudec, P-T Liu, and T-H Hou National Chiao Tung University

This work provides a complete framework, including device, architecture, and algorithm, for implementing bio-inspired supervised spiking neural networks (SNNs) on hardware. An analog synapse with atypical dual bipolar resistive-switching (D-BRS) modes demonstrates interchangeable Hebbian spiking-timing-dependent plasticity (STDP) and anti-Hebbian STDP, and it is capable of implementing supervised ReSuMe SNNs in crossbar arrays. By using an “exchange” update scheme, accurate supervised learning (~96% for MNIST) is achieved in a compact network.

11:35 AM - 12:00 PM
15.6 Stochastic Inference and Learning Enabled by Magnetic Tunnel Junctions (Invited), A. Sengupta, G. Srinivasan, D. Roy and K. Roy, Purdue University

Neuromorphic computational paradigms that exploit the stochastic switching behavior of devices in the presence of thermal noise is bringing about a wave of change in the way we perceive brain-inspired computing. In this article, we present proposals of spintronics enabled neuromorphic computing systems that perform probabilistic inference and online learning. Such stochastic neuromimetic hardware has the potential of enabling a new generation of state-compressed, low-power computing platforms, which can be significantly more efficient and scalable than their deterministic counterparts.