1 INTRODUCTION

In recent years, the semiconductor industry has begun to experience a significant slowdown in the performance improvements gained from technology scaling. While this is due in part to the impending end of Moore’s Law scaling, power consumption and architectural limitations have also become critical limiting factors for the level of performance achievable. Neuromorphic or neuro-inspired computer architectures are particularly worthwhile given the increasing number of big data problems requiring techniques and systems that can capture knowledge from an abundance of data. Here, we present a memristor-based Dynamic Adaptive Neural Network array (mrDANNA) with the potential to address contemporary application challenges while also enabling continued performance scaling.

The work leverages a hybrid CMOS-Memristor process developed and tested at the SUNY Polytechnic Institute, College of Nanoscale Science and Engineering (CNSE) for its design and eventual hardware realization. The process integrates metal-oxide memristors in the metal layers of the IBM 65nm 10LPe CMOS process, leading to a seamless CMOS/Memristor integration process.

The specific neuromorphic architecture on which the mrDANNA is based is the Neuroscience-Inspired Dynamic Architecture (NIDA) [5], [6], [8], as an approach to applying neuromorphic principles to a wide variety of applications. The structure and simplicity of the NIDA architectural model has been leveraged in the development of a Dynamic Adaptive Neural Network Array (DANNA) [1], an efficient digital system constructed from a basic element that can be configured to represent either a neuron or a synapse. Unique characteristics of the NIDA/DANNA approach over other neuromorphic or neuroscience-inspired systems include: a simplified neuron model, a higher functionality synapse model, real-time dynamic adaptability, configurability for the overall neuromorphic structure (e.g. number of neurons, number of synapses and connections), and scalability for element performance and system capacity.

2 MEMRISTIVE DEVICES

The memristive devices considered for this work were designed and fabricated in-house at the SUNY Polytechnic Institute’s Center for Semiconductor Research (CSR). Devices were manufactured on a 300mm wafer platform and integrated with the IBM 65nm 10LPe process technology. It should be noted that facilities in-house allow for an area-efficient and seamless flow of front-end CMOS and back-end memristive and metallization processes. The seamless integration of CMOS with memristive technology is a unique feature as compared to related efforts where memristive devices are integrated post-fabrication on an existing CMOS chip [4]. A custom, cost-effective build embeds a HfO$_2$ memristor device between metal 1 (M1) and metal 2 (M2) metallization layers. Fig. 1 (right) illustrates the device cross-section, with Fig. 1 (left) showing the TEM cross-section of the fabricated device.

The I-V characteristics of a fabricated HfO$_2$ memristor device are shown in Fig. 2. An average low resistance state (LRS) and high resistance state (HRS) of 10K and 150K, respectively, were observed during pulsing measurements. The on/off ratio and LRS are likely to increase, which is critical for low power operation, by manipulating the thicknesses and stoichiometries in the memristor film stack as well as moving the transistor in the 1T1M configuration from a discrete to an on-chip component effectively reducing parasitic capacitance. The devices show an excellent readout stress insensitivity, with the resistive states being insensitive to trillions of nanosecond pulse readouts. In addition, a low positive temperature dependence (5.9e-4 1/C) results in little change to the circuit performance over a large range of temperatures. Recent results also indicate controllable analog/multi-level switching in the memristive devices, which is key for their application as synaptic weights in the mrDANNA system. The model used for the simulation...
work in this paper is based on a model first developed and presented by McDonald et al. in [2], [3]. While the original model could be used for unipolar, nonpolar and bipolar behavior, this work is restricted to bipolar behavior following the observed characteristics of the device.

3 MEMRISTIVE NEURAL CIRCUITS

Synapses in the mrDANNA circuit must represent either a positive or negative weight and include delay distance as prescribed by the NIDA/DANNA model [1], [7]. Unlike the digital DANNA, in mrDANNA the synaptic weights are represented using memristors where the memristance is proportional to the desired weight. For the neuron, we implement an integrate-and-fire circuit similar to that described by Wu [9]. Here the design allows the neurons to operate in two different phases, integration and firing. When the neuron operates in its integration phase, the op amp acts as an integrator such that charge accumulates resulting in the membrane potential. A comparator circuit compares the membrane potential with the threshold voltage and generates a driving voltage. A resulting pulse on driving voltage then drives a “firing flop” which generates the corresponding output spike of the neuron.

The synaptic buffer drives voltage inputs across the two memristors of a synapse, one buffered and one complemented to allow for negative weights. It also controls the Long Term Potentiation (LTP) and Long Term Depression (LTD) mechanisms used for online learning and adaptation. The feedback signal from the IAF neuron connected to the synaptic buffers regulate the LTP/LTD event. Based on a firing event from the post-synaptic IAF neuron, as sensed by the synaptic buffer, the weight will be updated (increase/decrease) accordingly. A high level view of mrDANNA is provided in Fig. 3.

4 SHAPE RECOGNITION EXAMPLE

To showcase the usefulness of this type of network with synapses and integrate and fire neurons, a circuit for recognizing four basic shapes, triangle, square, diamond and plus, has been constructed. Besides recognizing noiseless images, some imperfect noisy images of triangles, squares, diamonds and also plus signs have been considered to determine the accuracy level of recognition of this network. Zero, one, two and three noise bits were considered for simulation and results for percentage of accuracy are shown in Fig. 4.

The results show that the network recognizes most of the cases with noisy bits up to 3 bits among the 25 bits of the image. The network recognizes images with one noisy bit with a hundred percent accuracy for all noise bits, the accuracy level slightly goes down but the percentage of accuracy is higher than 80 percent at 3 noise bits, which makes the circuit worthy enough in recognizing a particular shape. Average power calculated for this application is approximately 1.6mW.

5 CONCLUSION

This work demonstrates the efficiency of recognizing different patterns and also presents the power consumption for different LRS and HRS levels. In future work, more complex networks for spatiotemporal data applications and large pattern recognition will be solved with this efficient computing architecture.
REFERENCES


