Towards Memristive Dynamic Adaptive Neural Network Arrays

Nathaniel C. Cady¹, Karsten Beckmann¹, Harika Manem¹, Mark E. Dean², Garrett S. Rose², Joseph E. Van Nostrand³

Colleges of Nanoscale Science and Engineering¹
SUNY Polytechnic Institute
Albany, NY 12203

Department of Electrical Engineering and Computer Science²
University of Tennessee, Knoxville
Knoxville, TN 37996

Air Force Research Laboratory
RITB³
Rome, NY 13441

Abstract: We present the design and underlying device technology for a mixed-mode (analog and digital circuits) neuromorphic computing system built for rapid configuration, dynamic adaptation, low-power operation, that is well suited for processing spatio-temporal data. 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. Thus, the memristive Dynamic Adaptive Neural Network Array (mrDANNA) presented here addresses contemporary application challenges while also enabling continued performance scaling.

Keywords: Memristor; ReRAM; Neuromorphic Computing; Neural Networks

I. 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 has also become a critical limiting factor for the level of performance achievable. The research proposed here aims to enable future generations of computing systems by (1) leveraging an emerging, power-efficient device technology (i.e. the memristor) and (2) considering an alternative architectural model (i.e. neuromorphic) that promises to overcome many of the performance limitations of conventional von Neumann systems. It is worth noting that 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. Thus, the proposed memristor-based dynamic adaptive neural network array (mrDANNA) addresses contemporary application challenges while also enabling continued performance scaling.

The specific neuromorphic architecture on which the proposed mrDANNA is based is the Neuroscience-Inspired Dynamic Architecture (NIDA) [1], developed by researchers at the University of Tennessee, Knoxville (UTK) as an approach to applying neuromorphic principles to a wide variety of applications. In this work we leverage the features of the NIDA architecture in the construction of a mixed-mode, analog/digital memristor-based DANNA (mrDANNA) system. The proposed mrDANNA will maintain several important features of the digital representation of DANNA, including real-time dynamic adaptability and synaptic functionality that includes both weight and delay distance information [2]. The proposed mrDANNA system will improve performance and power consumption by leveraging memristor technology for analog operation and increased device density.

The proposed system will utilize a custom cost-effective hybrid CMOS/Memristor process developed and tested at the SUNY Polytechnic Institute’s College of Nanoscale Science and Engineering (CNSE) in the production of a test chip and first generation prototype of the mrDANNA fabric. The process integrates metal-oxide memristors in the metal layers of the CMOS process, leading to a seamless CMOS/Memristor integration process. The final system will utilize the area and power-efficient memristive devices for the implementation of synapses in the mrDANNA.

The remainder of the paper is organized as follows. Section II will provide background information on the memristors and NIDA and DANNA systems that enabled the design of the proposed system. Section III describes the device fabrication methodology employed at CNSE for the manufacture of the memristors. The proposed mrDANNA system will be explained in section IV with concluding remarks and future prospects provided in section V.

II. Background

A. Memristor

Resistive Random Access Memory (ReRAM), aka – memristor, is a novel form of non-volatile memory expected to replace a variety of current memory technologies and enabling the design of new circuit architectures. Investigations of memristor as a storage technology have shown a combination of high storage density with fast access and write speeds [3,4]. In addition,
the implementation of new circuit architectures, including neuromorphic systems, has attracted much attention. Recently, the endurance and reliability of memristor cells have reached a level at which they are competing with commercially available flash and CMOS memory technologies, making memristors a viable candidate for data storage and novel memory and logic architectures [3,4].

Memristive synapses have been interfaced with CMOS neurons in [5] and a cellular non-linear network based on memristors was proposed in [6]. Memristors were used to build non-volatile two-level and multi-level memories in [7] and [8], respectively. Memristors can also be used in digital logic as programmable switches in switching blocks [9] and to build block memories. To implement the mrDANNA system, we are using a hybrid CMOS/Memristor process that we have developed at CNSE/SUNY Polytechnic Institute. The process integrates metal-oxide memristors in the metal layers of the 65nm 10LPe CMOS process from IBM, leading to a seamless CMOS/Memristor integration process. 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 [10].

B. NIDA and DANNA

Our memristive Dynamic Adaptive Neural Network Array (mrDANNA) is based on the Neuroscience-Inspired Dynamic Architecture (NIDA) [1], a novel approach to applying neuromorphic principles to a wide variety of applications. The primary component of the NIDA architecture and corresponding DANNA structure is a single programmable element that can be configured as a neuron, synapse or pass-thru function, and be replicated across a grid with nearest neighbor or “small world” connectivity [1,2]. To maximize performance and minimize size and power consumption very simple state-machines and logic circuits are used, avoiding the use of digital signal processors, floating-point units, arithmetic-logic units, memory arrays and other common microprocessor units.

In a neural network, the neuron’s function is to accumulate “charge” by adding the weighted inputs to an existing charge level until that level reaches a programmable threshold. Upon reaching this threshold the neuron fires its outputs and resets its charge to a predetermined bias level. A weighted-sum threshold activation function is the primary candidate for the neuron design due to its simplicity and functionality. However, other activation functions (such as linear, sigmoid or Gaussian) might also be implemented for more realistic approximations of biological neural networks.

The synapse function captures selected features of both axons and synapses found in biological neural networks. Specifically, a NIDA/DANNA synapse contains both the associative delay distance between two neurons and the weight (or strength) of the synaptic connection. Two unique characteristics of the synapse model are: 1) the weight value held by the synapse can automatically potentiate (long-term potentiation, LTP) or depress (long-term depression, LTD) depending on the firing condition of its output neuron and 2) a string of firing events can be stored in a “distance FIFO” to simulate a synapse transmitting a set of firing events down its length (thus constituting temporal storage of events) [1,2].

Key features of the NIDA architecture include: 1) a spiky representation of data, 2) the ability for the system to adapt during run-time, and 3) a synaptic representation that includes delay distance as well as weight information. The inclusion of delay distance (i.e. a programmable delay between pre- and post-synaptic neurons) is expected to be of particular benefit in the processing of spatio-temporal data.

An evolutionary optimization (EO) environment has been developed at UTK to configure the neural networks in a DANNA [2]. The EO trains over parameters of the network (weights and delay distances on synapses and thresholds on neurons) as well as the structure (the number and placement of neurons and synapses) and the dynamics of the network. The dynamics of the network are directly embedded in the structure itself (delays in the synapse and charges in the neuron). Most other artificial neural network implementations have a predefined, fixed structure rather than one determined by a dynamic optimization method as in our approach. The EO requires the user to specify the inputs and outputs to the network, and a fitness function that rates how well a network's outputs fare with the given inputs. The EO then generates an initial population of random networks, and gradually evolves the population using mutation and crossover operations, until it generates a network that exceeds a predefined threshold for fitness. These operations occur on the direct representation of the network.

III. Device Fabrication

The memristive devices considered for this work were designed and fabricated in-house at the SUNY Polytechnic Institute's Center for Semiconductor Research (CSR) [11]. 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. A custom, cost-effective build embeds the memristor devices between metal 1 (M1) and metal 2 (M2) metallization layers. This is achieved by implementing an extra tungsten via (W-V1) below the memristor device. The composition of the M1 and V1 layers was altered from copper to tungsten to allow for the use of front-end-of-the-line (FEOL) tools to deposit the HfO2 layer, which serves as the...
active layer (metal-oxide layer) for the memristor device. Fig. 1 (right) illustrates the device cross-section, with Fig. 1 (left) showing the TEM cross-section of the fabricated device.

The current state of device development shows highly reliable devices performing with an endurance of over 450k cycles in a two resistive state switching mode, as evidenced from Fig. 2. An average low resistance state (LRS) and high resistance state (HRS) of 7kΩ and 40kΩ, 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 device film stack. The devices show an excellent readout stress insensitivity, with the resistive states being unsusceptible 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. Fig. 3 illustrates the IV measurements for the memristive device for over >450k cycles.

Recent results also indicate capability for controllable analog/multi-level switching in our memristive devices, which is key for their application as synapses 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 [12]. While the original model could be used for unipolar, nonpolar and bipolar behavior, this work is restricted to bipolar behavior following the characteristics of the device.

**IV. Proposed System**

The mrDANNA array will be composed of multiple neural network elements that can be trained for a variety of potential end applications (image/signal processing, anomaly detection, etc.). This architecture will reflect the tiled nature of the original DANNA system [1,2], but the various elements will be constructed using memristor-based circuits designed for improved performance, lower power consumption and improved density.
Conclusions and Future Work

This work provides a motivation towards memristive dynamic adaptive neural network arrays. The proposed neuromorphic system is an effective alternative to traditional Von Neumann architectures, and will improve performance and power consumption by leveraging memristor technology for analog operation and increased device density. Initial circuit-level simulations have provided proof of concept of the mrDANNA design elements, with future work being invested into developing the full-scale system that can enable evaluation of larger scale spatio-temporal data such as: video and audio classification, autonomous control of robotic systems, and real-time anomaly detection in network traffic.

Acknowledgements

The authors acknowledge the Center for Semiconductor Research at CNSE for wafer development and the AFRL for the funding received through the grand award number FA87501110008.

References


