Applying Memristors Towards Low-Power, Dynamic Learning for Neuromorphic Applications


March, 2017

42nd Annual GOMACTech Conference

https://www.gomactech.net/2017/index.html

You may get this paper from http://neuromorphic.eecs.utk.edu/
Applying Memristors Towards Low-Power, Dynamic Learning for Neuromorphic Applications

Wilkie Olin-Ammentorp¹, Karsten Beckmann¹, Joseph E. Van Nostrand², Garrett S. Rose³, Mark E. Dean³, James S. Plank³, Gangotree Chakma³, Nathaniel C. Cady¹

1. SUNY Polytechnic Institute, Colleges of Nanoscale Science and Engineering, Albany, NY, United States
2. Air Force Research Laboratory/RIITB, Rome, NY, United States
3. University of Tennessee, Knoxville, Department of Electrical Engineering and Computer Science, Knoxville, TN, United States

Abstract:
While neuromorphic computing offers methods to solve complex problems, current software-based networks offer limited flexibility and potential for low-power implementations. The memristive dynamic adaptive neural network array (mrDANNA) is a flexible hardware-based system, with applications including, but not limited to real-time speech recognition and spatio-temporal navigation. We present simulations of the mrDANNA system using physically integrated memristors (aka ReRAM) to encode synaptic weights, based on an empirical model characterizing our memristors.

Section I - Introduction
The semiconductor industry is experiencing a significant slowdown in scaling and performance improvements, with predictions that the traditional lateral scaling of transistors will end in 2021 [1]. While economic factors and physical boundaries have contributed to the decline of scaling, improvements in power consumption have become a critical factor in the design of circuits. The research described herein aims to enable future generations of computing systems to solve new and complex problems, by using a low-power device to implement an alternative computational architecture. The architecture in question is neuromorphic computation, which has gained widespread acceptance for its success in areas such as speech recognition, artificial intelligence, and processing big data.

While neuromorphic computing has demonstrated utility, it is still limited in its applications by the hardware and software required to carry out computations. Neural networks are often implemented in graphical processing units (GPUs) and other dedicated hardware, which are often non-portable, power hungry, and have performance gains limited by traditional scaling [2]. By creating dedicated hardware which leverages alternative circuit components and novel approaches, neuromorphic computing can be realized as a low-cost, low-power component of computing systems.

This paper explores the development and operation of a hardware-based neuromorphic system, the memristive Dynamic Adaptive Neural Network Array (mrDANNA), developed collaboratively by researchers at the University of Tennessee, Knoxville (UTK) and SUNY Polytechnic Institute. mrDANNA is based on the Neuroscience-Inspired Dynamic Architecture (NIDA) [3], which applies neuromorphic computing to a wide variety of problems, including spatio-temporal navigation and control problems. mrDANNA focuses on implementing the strong learning capabilities of NIDA into a mixed analog/digital system, incorporating memristors as the synaptic element. This replacement of a synaptic CMOS block improves performance, while reducing power consumption and increasing device density. The proposed system utilizes a custom, nanoscale hybrid CMOS/ReRAM process developed and tested at the SUNY Polytechnic Institute's Colleges of Nanoscale Science and Engineering (CNSE).

Section II – Theory and Design of mrDANNA and Memristors
The Dynamic Adaptive Neural Network Array (DANNA) architecture is a biologically-inspired learning system, designed to be a flexible, low-cost platform with powerful learning capabilities. The fundamental structure of the DANNA architecture is a two-dimensional grid of elements, each of which can process inputs and provide outputs. Grid points can behave as neurons, synapses, or a pass-through elements.

Synapses are simple elements, which have a single input and weighted output, and can connect to either neurons, or other synapses. Neuron elements have an internal charge, which is modified by the weight of any connected, firing synapse. Once a neuron's charge exceeds its threshold, it fires, sending a signal through synapses which receive its output. In this manner, information travels through, and is processed by, the neural network.

Synapses are simple elements, which have a single input and weighted output, and can connect to either neurons, or other synapses. Neuron elements have an internal charge, which is modified by the weight of any connected, firing synapse. Once a neuron's charge exceeds its threshold, it fires, sending a signal through synapses which receive its output. In this manner, information travels through, and is processed by, the neural network.

The flexible nature of this underlying architecture allows DANNA to carry out a wide variety of calculations.
However, the large number of configurations available for the computing network means that finding the optimal arrangement of neurons and synapses within the network is not a trivial issue. As is the case for neuromorphic computing in general, finding an optimal network configuration is often a large part of solving the problem. For DANNA, network configuration optimization is performed by an evolutionary algorithm (EO). In contrast to other types of neuromorphic computing, DANNA does not require hand-picked networks. Instead, DANNA utilizes EO to automatically search for and select networks which provide outputs that are useful to solve a problem. Using EO, networks can be modified (mutated), or combined (crossover) to create a new generation of networks. Each of these network’s usefulness in solving the desired problem (fitness) is evaluated, and the top-performing networks continue to be optimized and modified in the next generation. This continues until a network with satisfactory performance is found. Evolutionary optimization has been used to find DANNA networks which solve an inverted pole-balancing problem, and a 2D navigation problem. An example network is shown in Figure 1b. Future applications which are being investigated include three-dimensional spatio-temporal navigation and speech recognition. DANNA has been implemented in hardware using field-programmable gate arrays (FPGAs), but further hardware implementations are being considered. mrDANNA is a modification of the DANNA system, which uses memristors to implement the network's synapses. This can reduce the number of digital components required to construct the system, and allows for simultaneous analog processing of inputs. Memristors are well-suited to act as synapses, as they are two-terminal devices which can act as a resistor with a programmable resistance. The resistance acts as the weight of a synapse, implemented in a single device with a low footprint and power consumption. Initial estimates show that with an optimized memristor, mrDANNA can use less than half the power of a CMOS implementation [5].

Section III – Fabrication and Operation of Memristors
Memristors are two-terminal devices whose resistance depends on the accumulative current passed through the device. This general definition encompasses a broad variety of emerging memories, which use different mechanisms and materials. However, the majority of memristors are based on a metal-insulator-metal (MIM) structure, where the insulator is a transition-metal oxide (TMO). While many MIM-TMO devices have demonstrated memristive behavior, most do not meet the requirements for usage in neuromorphic circuits, lacking properties such as manufacturability, endurance, data retention, and multi-level programmability. With these goals in mind, we have developed CMOS-integrated ReRAM that can function as memristors using a customized variant of the IBM 65nm 10LPe process technology at SUNY Polytechnic’s CNSE fabrication facilities. In this process, ReRAM’s are embedded between the interconnects of Metal 1 (M1) and Metal 2 (M2). A film stack of HfO$_2$, Ti, and TiN is deposited over the tungsten Via 1 (W-V1) electrode. This film stack is then etched to create individual memristors and capped with SiO$_2$. The normal process flow continues with Via 2 (V2). This creates ReRAM in the 1T1R configuration, where each ReRAM is connected in series to a transistor. This transistor can serve as a current limiting device, as well as an addressing element.
Figure 3: a) The current flowing through a memristor after consecutive reset pulses. b) Average absolute resistance change from the application of consecutive pulses at differing voltages. [7]

Figure 4: Potentiating a memristive neuron in a DANNA network

These ReRAM have displayed good switching characteristics, high endurance, and multi-level operation and fulfill the required memristor characteristics. By cycling the ReRAM with +2 V set pulses and -1.2 V reset pulses, their high resistance states (HRS) and low resistance states (LRS) were measured and found to be 70 kΩ, and 4 kΩ, respectively. The 'set' operation, changing the ReRAM from high to low resistance state, is very fast. However, the 'reset' operation, moving from low to high resistance, is not as fast, due to the kinetics of the resistance switch. This means that using pulsed biasing, the resistance of the device may be incrementally increased, allowing for the multi-level operation necessary for neuromorphic computation. Figure 3 shows the multi-level operation of a ReRAM device, paving the way towards using ReRAM as synaptic elements [7].

Section IV – Simulation of Learning and Operating with mrDANNA
Using electrical characteristics measured from the CMOS-integrated RRAM, we create virtual model for use in simulating DANNA networks which employ RRAM elements to encode synaptic weights. Using the multi-level capability of RRAM, the synaptic weight of a synapse in the DANNA network can be incrementally adjusted, following the rules of Hebbian learning which are used in the system.

There is a strong correlation between the pre-synaptic and post-synaptic neurons firing; therefore the weight of the synapse is increased as the simulation continues. Physically, this takes the form of gradually lowering the resistance of the RRAM element, which decreases the attenuation of the voltage signal traveling through the synapse.

Section V – Conclusions
As scaling traditional computing methods becomes more difficult and less rewarding, alternative methods of computation must be found. Neuromorphic computing is currently one area which is being investigated for expansion into low-power systems with advanced capabilities. The DANNA system has shown its applicability to many learning tasks, and the memristive adaptation of DANNA (mrDANNA) using CMOS-integrated RRAM reduces its power usage and hardware complexity. Simulations of mrDANNA using empirical
RRAM models have been shown, and further work will focus on improving RRAM to reduce power usage and increase reliability before integrating it into learning systems as a synaptic element.

References:
3. C. A. Schuman, J. D. Birdwell, and M. Dean, “Neuroscience-inspired dynamic architectures,” in Biomedical Science and Engineering Center Conference (BSEC), 2014 Annual Oak Ridge National Laboratory, May 2014