Full CMOS-Memristor Implementation of a Dynamic Neuromorphic Architecture

Nathaniel Cady, Karsten Beckmann, Wilkie Olin-Ammentorp
Colleges of Nanoscale Science and Engineering
SUNY Polytechnic Institute
Albany, NY

Joseph Van Nostrand
Air Force Research Laboratory/RITB
Rome, NY

Gangotree Chakma, Sherif Amer, Ryan Weiss, Sagarvarma Sayyaparaju, Mussabir Adnan, John Murray, Mark Dean, James Plank, Garrett Rose
Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville
Knoxville, TN

***Abstract*—Neuromorphic computing systems seek to emulate biological neural functionality emulated in either software or electrical hardware. A key function for such systems is their ability to learn and adapt. In the human brain, such learning and adaptation is achieved via modulation of synaptic connections between different neurons. Memristors (implemented as resistive random access memory or ReRAM) have great potential to provide synaptic functionality for neuromorphic chip architectures. Under an AFRL-sponsored program, we have developed a unique memristor-CMOS hybrid system for implementing a dynamic adaptive neural network array, also known as mrDANNA. Most recently, our effort has moved from a software-based emulator, to FPGA implementation, and finally to the design, tapeout, and fabrication of this unique, adaptive approach to neuromorphic computing.**

Keywords—CMOS, neuromorphic computing, computer architecture

I. Project Description

Memristors, which can be implemented as resistive random access memory (RRAM) are a novel form of non-volatile memory expected to replace a variety of current memory technologies and enabling the design of new circuit architectures [1, 2]. Investigations of ReRAM as a storage technology have shown a combination of high storage density with fast access and write speeds. Recently, the endurance and reliability of ReRAM cells have reached the level at which they are competing with commercially available Flash memory and CMOS technologies, making ReRAM a viable candidate for data storage and novel logic and security architectures. To this end, we have demonstrated a vertically-integrated process flow for fabrication of hybrid CMOS logic and ReRAM [3, 4].

Our memristive neural network array (mrDANNA) is based on the Neuroscience-Inspired Dynamic Architecture (NIDA), developed by researchers at the University of Tennessee, Knoxville (UTK) as an approach to applying neuromorphic principles to a wide variety of applications. 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 including 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. The structure and simplicity of the NIDA architectural model has recently been leveraged in the development of a Dynamic Adaptive Neural Network Array (DANNA) [5], 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. The NIDA/DANNA/mrDANNA models also fit into a integrated hardware/software application development stack for demonstration, testing and benchmarking. Current benchmarks include static classification, time-series classification and real-time control applications [6].

Under support from the Air Force Research Laboratory, we have pursued the next generation of the NIDA/DANNA approach, implementing synaptic connections in array with hafnium oxide memristive devices (HfOx ReRAM) [7]. Recent work on this project has resulted in the generation of multiple version of mrDANNA "neurons", a fully digital DANNA circuit design, and implementation of these designs into a 65nm CMOS / ReRAM at the SUNY Polytechnic Institute's 300mm research foundry. This work builds on our existing efforts in developing a hybrid memristor-CMOS process flow (using a 300mm wafer platform) and developing multi-level resistive switching performance, which can be

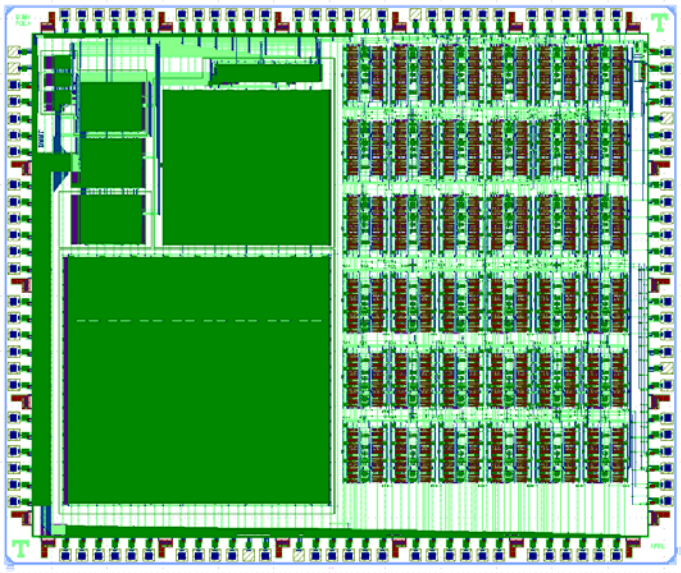

Fig. 1. Schematic of the memristive dynamic neural network array (mrDANNA), which includes memristor-CMOS hybrid neural network components, a fully-digital DANNA, a 512x512 addressable ReRAM (memristor) block, memristive reservoir computing circuits, and a wide array of individual ReRAM and transistor-ReRAM test circuits.

leveraged for setting multiple synaptic levels for our neural network arrays [3, 8]. We have fully taped out a reticle set for fabricating a full digital DANNA array, multiple mrDANNA test arrays, and individual neurons of differing types and configurations (digital, mixed analog/digital).

II. MrDanna Circuit Design

Each mrDANNA core consists of several memristive synapses and a single analog CMOS neuron. Inspired by biological synapses, we have modeled the synapses so that they are capable of representing multilevel synaptic weights. The memristors' non-volatility and capability for a high integration density within back-end-of-the-line metallization layers makes them an ideal storage element to be used in the synapse. For the mrDANNA design, a twin memristive synapse architecture is used to implement both positive and negative weights. Two memristive ReRAM devices are connected in the opposite polarity as shown in Figure 2 (left graphic) where the blue and red represent ReRAM devices with opposing polarity. This allows the use of ReRAM devices with an asymmetric switching behavior, such as HfO_2 based ReRAM devices developed at SUNY Polytechnic Institute. Each time there is a pre-synaptic fire, the neuron accumulates the weighted current from the synapses connected to it and compares it to a given threshold. As soon as the accumulated voltage crosses the set threshold voltage, the neuron generates a post-synaptic fire and resets to the initial condition for the next available pre-synaptic fires.

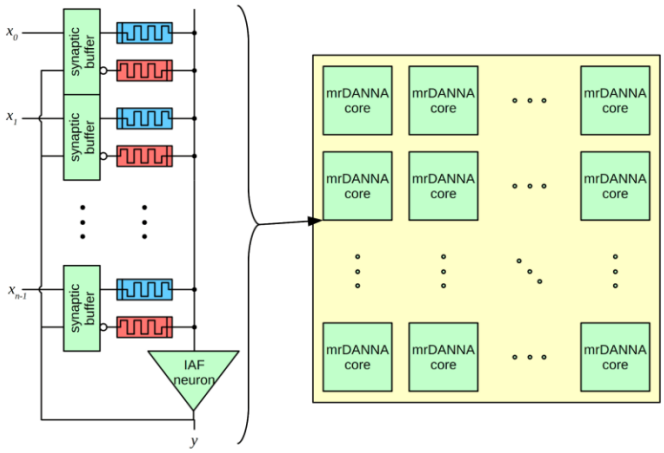

Fig. 2. Block diagram of the mrDANNA system, showing a single core on the left and the array network on the right.

III. Reservoir Computing Circuits

One of the alternate computing models being tested using this process is reservoir computing. In this mode of computing, inputs are applied to a large, dynamical system which can develop complex, time-dependent responses from simple inputs. The purpose of this 'reservoir' is to generate a unique internal state based on the inputs which have been applied to it. The state of the reservoir as a whole can then be examined by a simple output stage (such as a perceptron) to carry out tasks such as classification of the original inputs. Only this output stage must be trained to improve performance on a task, as long as the reservoir can act in a sufficiently complex manner.

We have implemented a mixed analog/digital circuit to implement reservoir computing, included with the other designs on the mask set. In this design, a one-dimensional cellular automata (CA) evolving in time is used to create a reservoir. CA are simple, cellular structures which change their behavior based on the past state of themselves and their neighbors. However, they develop very complex behaviors, and have even been shown as suitable substrates for universal computation [9]. The output layer examining the state of the reservoir is a memristive support-vector machine (SVM), linking the state of each cell to a memristor in a parallel read-out array. Each memristor can be individually addressed and programmed, allowing the weights of the SVM to be trained. This allows unique resistance states to be achieved for different reservoir states. The reservoir and SVM together create a full circuit for reservoir computing, the cellular memristive-output reservoir (CMOR). This design is a hardware implementation of a novel concept which has thus far only been demonstrated in software [10].

IV. Additional Circuits

In addition to the DANNA circuits, a number of other designs and features have been included in the taped-out mask set, as this integrated memristor/CMOS process provides a test-bed for a number of novel circuits and computing models. Large blocks of ReRAM memory (up to 512x512 cells) have been

included to demonstrate large-scale use of memristors as a storage element. A wide variety of test structures, for both CMOS and the integrated memristors, are also included (Table 1).

TABLE I. CIRCUIT ELEMENTS, MEMRISTIVE DEMONSTRATION CIRCUITS, AND FULL NEUROMORPHIC CIRCUITS IMPLEMENTED ON THE SUNY POLY / UT-KNOXVILLE MRDANNA DEVELOPMENT CHIP.

ID	Count/Die	Circuit
1	1	512x512 1T1R array (262,144x 100x100nm^2) w/ decoder (9 bit/ 9 word lines)
2	1	512x512 1T1R array (262,144x 100x100nm^2) w/ decoder (9 bit/ 9 word lines) and ESD contact pads
3	4	8x8 1T1R array (100x100nm^2) w/o decoder
4	16	12x12 1R array (100x100nm^2)
5	18	2x2 form and cut 1R structures
6	2	ReRAM time-delay PUF (http://ieeexplore.ieee.org/document/7484314/)
7	1	CMOR (Cellular Memristive Output Reservoir – reservoir computing circ.)
8	1	Full digital DANNA array (UT-Knoxville / Dean)
9	Multiple	mrDANNA "neurons" (various types/configurations)
10	100	1T1R (100x100nm^2) rVt NFET 2mA VCM w/ ESD contact pads
11	100	1T1R (100x100nm^2) rVt NFET 2mA VCM w/o ESD contact pads
12	100	1T1R (100x100nm^2) dgx NFET 2mA VCM w/o ESD contact pads
13	100	1T1R (100x100nm^2) dgx NFET 2mA VCM w/ ESD contact pads
14	100	1T1R (100x100nm^2) rVt NFET 2mA ECM w/ ESD contact pads
15	100	1T1R (100x100nm^2) rVt NFET 2mA ECM w/o ESD contact pads
16	100	1T1R (100x100nm^2) dgx NFET 2mA ECM w/o ESD contact pads
17	100	1T1R (100x100nm^2) dgx NFET 2mA ECM w/ ESD contact pads
18	48	RF 1T1R (100x100nm^2) dgx NFET 2mA VCM w/o ESD contact pads
19	16	1T1R (100x100nm^2) capacitive structures (10fF – 50pF)
20	80	1T1R on-chip pulse creation (1ns – 20ps)
21	24	Configurable XOR with pull down/up ReRAM (100x100nm^2)
22	8	1T dgxfet NFET 500uA test structure
23	8	1T dgxfet NFET 1mA test structure
24	8	1T dgxfet NFET 2mA test structure
25	8	1T dgxfet PFET 500uA test structure
26	8	1T dgxfet PFET 1mA test structure
27	8	1T dgxfet PFET 2mA test structure
28	8	1T rVt NFET 500uA test structure
29	8	1T rVt NFET 1mA test structure
30	8	1T rVt NFET 2mA test structure
31	8	1T rVt PFET 500uA test structure
32	8	1T rVt PFET 1mA test structure
33	8	1T rVt PFET 2mA test structure
34	1	Metal-Insulator-Metal capacitive test structures

V. CONCLUSIONS

Leveraging our past work in developing a combined CMOS/memristor process, we have designed a variety of approaches to neuromorphic computing and other novel circuits on a single mask-set. This demonstration chip will be used to demonstrate all-digital implementation of neuromorphic circuits that have emerged from original NIDA/DANNA efforts at UT-Knoxville, as well as novel hybrid memristor/CMOS neurons and small neuromorphic circuit blocks which are the result of collaborative efforts between UT-Knoxville and SUNY Polytechnic Institute. The resulting chips from this effort will be a valuable test vehicle for demonstrating low-power, highly dynamic neuromorphic circuits.

REFERENCES

[1] Chua, L., Memristor-The missing circuit element. IEEE Transactions on Circuit Theory, 1971. 18(5): p. 507-519.

[2] Strukov, D.B., et al., The missing memristor found. Nature, 2008. 453(7191): p. 80-83.

[3] Beckmann, K., et al., Nanoscale Hafnium Oxide RRAM Devices Exhibit Pulse Dependent Behavior and Multi-level Resistance Capability. MRS Advances, 2016. 1(49): p. 3355-3360.

[4] Beckmann, K., et al., Impact of Etch Process on Hafnium Dioxide Based Nanoscale RRAM Devices. ECS Transactions, 2016. 75(13): p. 93-99.

[5] Disney, A., et al., DANNA: A neuromorphic software ecosystem. Biologically Inspired Cognitive Architectures, 2016. 17: p. 49-56.

[6] Plank, J.S., et al. A Unified Hardware & Software Co-Design Framework for Neuromorphic Computing Devices and Applications. in 2017 IEEE International Conference on Rebooting Computing (ICRC). 2017.

[7] W. Olin-Ammentorp, K.B., J. E. Van Nostrand, G. S. Rose, M. E. Dean, J. S. Plank, G. Chakma, N. C. Cady, Applying Memristors Towards Low-Power, Dynamic Learning for Neuromorphic Applications. 42nd Annual GOMACTech Conference, Reno, NV, 2017.

[8] Alamgir, Z., et al., Pulse width and height modulation for multi-level resistance in bi-layer TaOx based RRAM. Applied Physics Letters, 2017. 111(6): p. 063111.

[9] Schiff, J.L., One-Dimensional Cellular Automata, in Cellular Automata 2007, John Wiley & Sons, Inc. p. 39-87.

[10] Stefano, N. and A. Molund, Deep Reservoir Computing Using Cellular Automata. CoRR, 2017. abs/1703.02806.