2023 IEEE 66th International Midwest Symposium on Circuits and Systems (MWSCAS)
Phoenix, Arizona, USA, August 6-9, 2023

Neuromorphic Dendritic Synapse Integrating Gated-RRAM

1st Siddharth Barve
Dept. of Electrical and Computer Engineering
University of Cincinnati
Cincinnati, USA
barvesh@mail.uc.edu

2nd Rashmi Jha
Dept. of Electrical and Computer Engineering
University of Cincinnati
Cincinnati, USA
jhari@ucmail.uc.edu

2023 IEEE 66th International Midwest Symposium on Circuits and Systems (MWSCAS) | 979-8-3503-0210-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/MWSCAS57524.2023.10405939

***Abstract*—Artificial neural networks (ANNs) are becoming increasingly widespread yet the current hardware implementations exceed the power, area, and time budget of many applications. Additionally, the current ANN neurons are a highly simplified model of their biological counterparts. In this work, we propose a neuromorphic dendritic synapse that improves the computational complexity of individual neurons as a foundation for more efficient implementation of complex ANNs.**

***Index Terms*—Neuromorphic, Dendrite, Synapse, Resistive Random-Access Memory**

I. Introduction

Artificial neural networks (ANNs) are becoming increasingly widespread from use in computer vision, speech recognition, medical diagnosis, natural language processing, etc. [1], [2]. However, many of these applications may require low power and real-time computation in a small footprint. Unfortunately, current computer architectures are unable to provide the high memory bandwidth for real-time compute while additionally costing high power and area [1], [3], [8]. To combat these issues, higher memory bandwidth architectures such as field-programmable gate arrays (FPGAs) and application-specific integrated circuits (ASICs) have been employed [1], [8]. These architectures rely on parallelism and near- or in-memory to drastically improve the memory bandwidth while also reducing the power consumption and area.

Biological neural networks have been the inspiration for many ASICs developed for accelerating neural networks. However, while biological neural neurons utilize their dendrites, synapses, and somas for computation, ANNs primarily focus on the latter two and still at their most basic states [2]–[4]. Many current ANNs utilize artificial neurons or perceptrons which consist of three parts: synaptic weights, accumulation, and activation. The synaptic weights combine all the attenuation of the amplitude of the signal in a neuron into a single net synaptic weight [2], [3], [8]. This synaptic weight is then used for a weighted sum of inputs which is accumulated as a total stimulus similar to the functionality of the soma [2], [3]. This stimulus then produces an output signal via an activation function acting like the action potential threshold in biological neurons [8]. Hardware accelerators have been able to take inspiration from these biological neurons to produce the multiply-accumulate (MAC) crossbar seen in Fig. 1. The conductive elements are able to act as the net synaptic weights via device physics and governing circuit laws allow for accumulation of weighted stimulus. However, there is more to the computational capabilities of neurons than the functionality described by the net synaptic weight and accumulation.

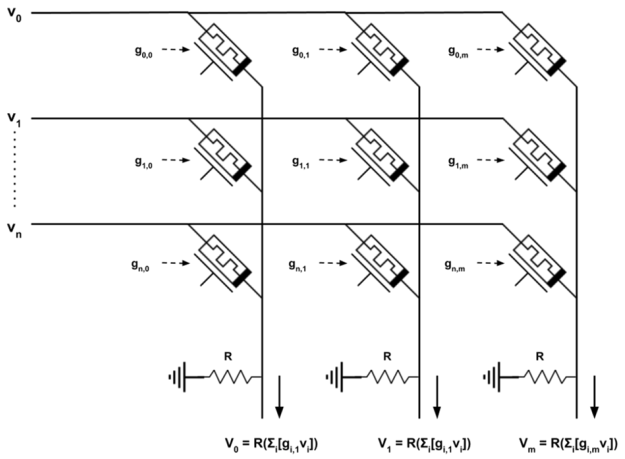


Fig. 1. Gated-memristive multiply-accumulate crossbar. The current produced at each gated-synaptic device is a product of the input voltage and conductance of the synaptic device. The total current for each neuron is the sum of products of the inputs and corresponding weights.

The dendrites in neurons have many functions that allow biological neurons to have drastically more computational power than many artificial neurons [2]–[6]. There exist both active and passive dendrites in biological neurons both of which have their own characteristics and functionality. Active dendrites are able to provide amplification of signals, feedback loops, as well as regulating signals from distal neurons among other functions [2]–[6]. Recent work in developing and integrating artificial dendrites, at the time of this writing, has been focused primarily on these active dendrite functions [2]–[6]. Many of these works primarily focus on a single function of active dendrites and produce hardware specialized for that specific function. Additionally, many of these implementations lack reconfigurability of these dendritic functions which is important moving from different applications and datasets [2]–

Funded by National Science Foundation and University of Cincinnati.

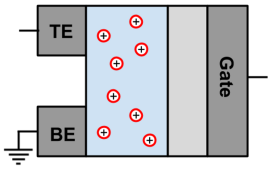


Fig. 2. llustration of gated-RRAM. Oxygen vacancies in the oxide channel can drift towards or away from the top electrode (TE) and bottom electrode (BE) interface of the channel depending on the bias on the gate.

[6]. The active aspect of these implementations also requires more power and signal complexity.

However, there also exist passive dendrites with many crucial functions. Passive dendrites are responsible for encoding delay into inputs as a function of their distance [6]. This allows for the neuron to label the input location via its encoded delay. Additionally, these delays aid in detecting sequences of inputs. By delaying the most distal input, the distal stimulus is still retained while a more proximal stimulus occurs resulting in higher total stimulus [6]. However, reversing the order from proximal to distal, the proximal stimulus diminishes by the time the distal stimulus occurs [6]. These delays also aid in filtering signals by their frequency [6]. In addition to encoding delay, passive dendrites induce sublinear summation of stimulus to ensure that a single stimulus or set of stimuli do not overpower the neuron [6]. In this paper, we propose a new biologically inspired reconfigurable dendrite-synaptic element that is able to harness the properties of gated-resistive random-access memory (gated-RRAM) to implement the functionalities of passive dendrites and synapses in-memory for improve bio-inspired hardware acceleration. The remainder of this paper is organized as follows: section II describes the functionality of the gated-RRAM device, section III details the dendrite-synaptic element and its functionality, and section IV illustrates and describes the results.

II. Gated-RRAM Device

Gated-RRAM is a type of gated-memristive device that has been recently reported [1], [8], [9]. As shown in Fig. 2, gated-RRAM devices consist of a gate, a top electrode, a bottom electrode, a channel oxide containing oxygen vacancies (V_o^{2+}) through which current can flow between top and bottom electrodes. When a positive bias is applied on the gate with respect to the top and bottom electrodes, the V_o^{2+} in the drift towards the top and bottom electrode interface increasing the conductance of the device, to a maximum conductance at it low resistance state (LRS), measured between top and bottom electrodes. Conversely, when a negative bias is applied on the gate with respect to the top and bottom electrodes, the V_o^{2+} in the drift away from the top and bottom electrode interface decreasing the conductance of the device, to a minimum conductance at it high resistance state (HRS), measured between top and bottom electrodes. The increase of conductance of the device is called potentiation while the decrease of conductance of the device is called depression.

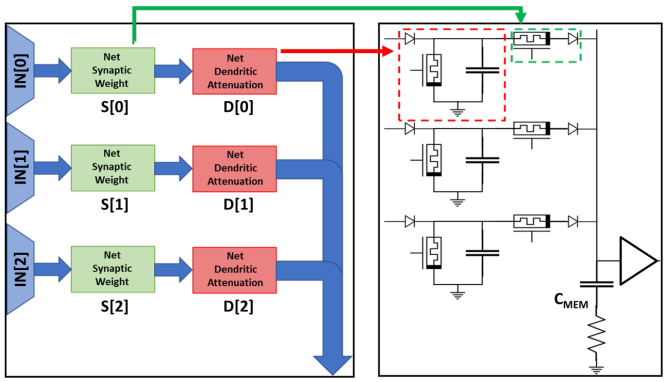


Fig. 3. llustration of dendritic synapse. Each input (IN[0]-IN[2]) has its own net dendritic attenuation element (D[0]-D[2]) and net synaptic weight element (S[0]-S[2]). The remaining crossbar structure of the neuron is similar to that of the neurons in Fig. 1 in which the total stimulus is accumulated in C_{MEM} as membrane potential.

A. Gated-RRAM Crossbar

Gated-RRAM have been investigated as a synaptic device for in-memory computing of the MAC operation shown in Fig. 1. Additionally, integrating multi-state gated-RRAM allows for a more dense memory crossbar since each gated-RRAM device can store multiple bits. The reconfigurability of gated-RRAM allows for synaptic weights to be programmed into the neuromorphic crossbar. In contrast to other two-terminal memristive synaptic devices, the additional terminal provided by the gate allows the synaptic device to be simultaneously programmed and read for in-memory computation [1], [8]–[11]. This ability to program via a third terminal can additionally reduce the programming circuitry required by the architecture.

III. Dendritic Synapse

Current ANN implementations couple the attenuation of the amplitude of the signal due to neuronal elements into the aforementioned net synaptic weight. This results in many modern neuromorphic architectures and synapses to use a single synaptic device with programmable conduction [8], [11], [12]. This drastically simplifies the computation of the signal attenuation in comparison to complexity in biological neurons. We decided to enable similar computation simplicity while retaining the computational capabilities of passive dendrites. We developed a dendritic synapse containing two sub-components shown in Fig. 3: a net synaptic weight and a net dendritic attenuation.

A. Net Synaptic Weight

The net synaptic weight is handled by the gated-RRAM device as a conventional memristive synaptic weight, similar to those seen in Fig. 1. The net synaptic weight is used to attenuate the amplitude of the input signal. However, it is important to note that we placed the net synaptic weight after the net dendritic attenuation element. This is to reduce the coupling effects of the synapse on the dendritic attenuation. We wanted the synaptic and dendritic elements to be able to

be tuned independently to simplify the programmability of the dendritic synapse.

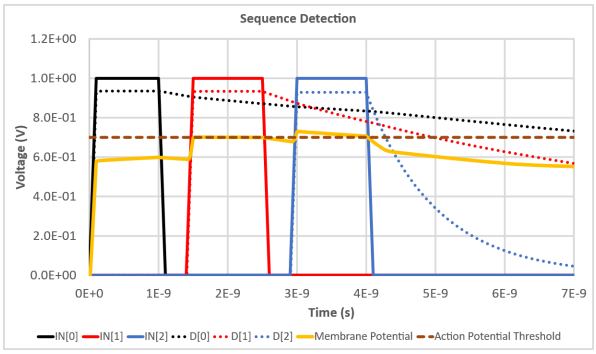
B. Net Dendritic Weight

The net dendritic attenuation component consists of a gated-RRAM device in parallel to a capacitive element. This is similar to the leaky integrate-and-fire RC component used in neuromorphic spiking neural networks (SNNs). However, by implementing a gated-RRAM device we allow for the rate of leakage to be tuned by programming the conductance of the gated-RRAM device. Potentiating the gated-RRAM device allows for reduction in the signal delay while depressing the gated-RRAM device increases the signal delay. The net dendritic attenuation component is used to attenuate the frequency of the input pulses via delay. This is similar to the behavior observed in lossy transmission lines. Therefore, the programmability of the gated-RRRAM results in programmability of signal delay. This allows for tuning the effective distance of the stimulus from the soma of the neuron. Combining both the dendritic and synaptic elements allow for complete control over the delay and amplitude of the signals observed by the soma or accumulation. We demonstrate that with these two programmable parameters, we can observe the various functions of passive dendrites in biological neurons.

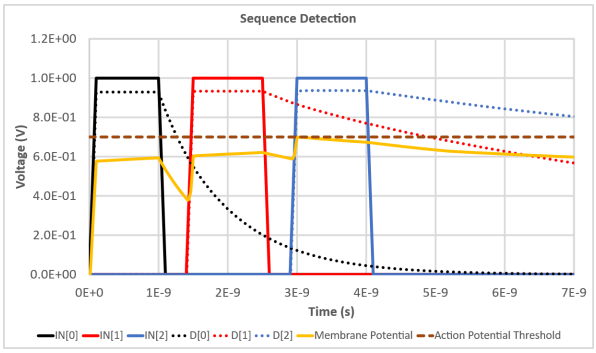
IV. Results and Discussion

We simulated neurons integrating the dendritic synapses using SPICE. We modeled the gated-RRAM device using the gated-synaptic device model from [9] that was fitted to fabricated gated-RRAM devices.

We first demonstrated the ability of dendritic elements to be used for detecting sequences of stimuli using only a single neuron as shown in Fig. 4. We first ensured that the synaptic weights were kept constant for all the input stimuli. Then the net dendritic element for each input was programmed such that the first input (IN[0]) would have the largest delay while the last input (IN[2]) would have the smallest delay as shown in Fig. 4a. This resulted in the neuron becoming sensitive to the sequence of IN[0], IN[1], IN[2]. We see that by applying the correct sequence of inputs that membrane threshold was able to exceed the action potential threshold thus correctly detecting the sequence. This is because due to the dendritic delay D[0] and D[1] are still retained by the time IN[2] arrives allowing for summation of the three stimuli. However, in Fig. 4b, we reversed the order of the delays due to the dendritic element. Therefore, the neuron would now be sensitive to the sequence of IN[2], IN[1], IN[0]. We see this change by applying the same sequence of inputs as before, IN[0], IN[1], IN[2], the membrane potential no longer exceeds the action potential threshold. This is because the dendritic delay of D[0] is not too small for it to be properly retained by the time IN[2] arrives. By tuning these delays we can have the neuron be sensitive to different sequences of stimuli. The synaptic weight can be used in conjunction to allow for more or less inputs to be required in the sequence by varying the attenuation of each stimulus. Conventionally, large networks of multiple neurons



(a)

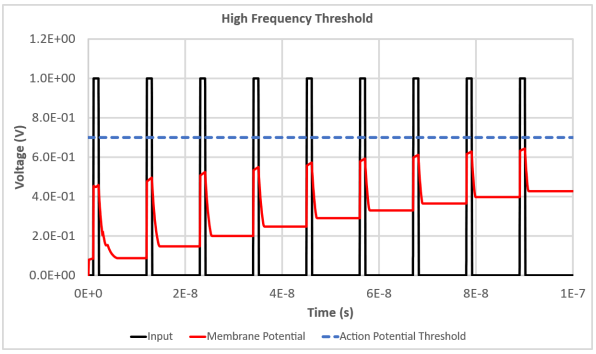


(b)

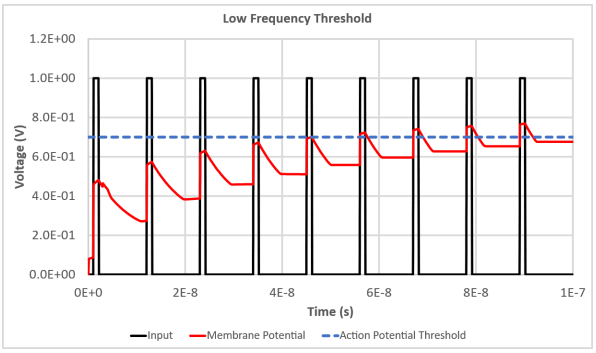
Fig. 4. SPICE simulation of sequence of three inputs IN[0]-IN[2]. Each input has the same synaptic weight but different dendritic attenuations D[0]-D[2]. The total stimulus of the neuron is accumulated as the membrane potential. a) The signal retention or delay is decreasing from D[0]-D[2]. b) The signal retention or delay is increasing from D[0]-D[2].

like LSTMs or other recurrent networks have been used for detection of temporal sequences [7]. However, we demonstrate that the integration of these programmable dendrites allows for detection of sequences in a single neuron. Additionally, the tuning of these delays can be used for coincidence detection of stimuli. The delay and retention of these signals can be tuned to increase or decrease the window of coincidence. This is especially useful with the increasing interest in SNNs which utilize spike-timing dependent plasticity (STDP) [8].

We also demonstrated the ability of passive dendrites to allow for filtering signals via their frequency as shown in Fig. 5. We demonstrate in Fig. 5a that by lowering the delay of the stimulus the membrane potential rises much slower and is unable to exceed the action potential. This is because the stimulus has decayed much faster and is not sufficiently retained by the time the next pulse arrives. This in conjunction with the leakage of the accumulation resulting in a lower total membrane potential. However, by increasing the delay of the stimulus in Fig. 5b the membrane potential is able to exceed the action potential threshold with the same frequency of pulses. This is because each pulse is retained more by the time the next pulse arrives allowing for better accumulation of the pulses. Therefore, we are able to tune the frequency threshold for a spiking stimulus by programming the delay of the passive dendritic element. Additionally, we can observe

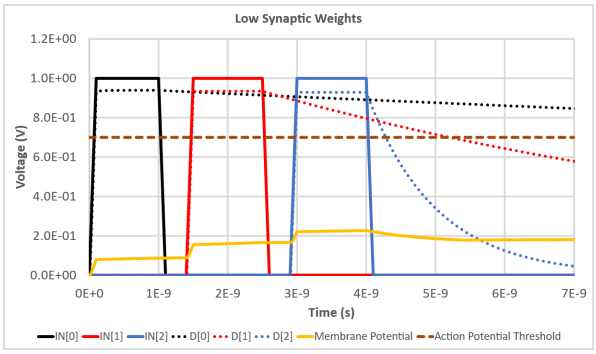


(a)

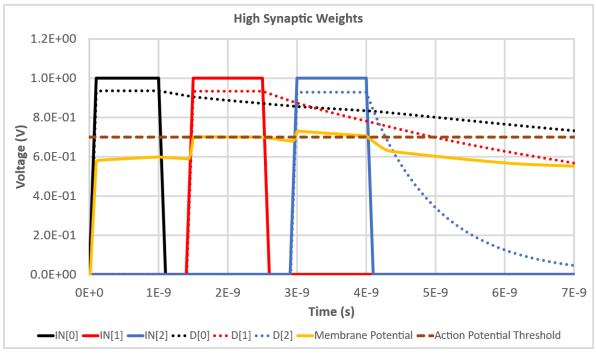


(b)

Fig. 5. SPICE simulation of sequence of three inputs IN[0]-IN[2]. Each input has the same synaptic weight but different dendritic attenuations D[0]-D[2]. The total stimulus of the neuron is accumulated as the membrane potential. a) The signal retention or delay is decreasing from D[0]-D[2]. b) The signal retention or delay is increasing from D[0]-D[2].



(a)



(b)

Fig. 6. SPICE simulation of sequence of three inputs IN[0]-IN[2]. Each input has the same dendritic attenuations D[0]-D[2] as Fig. 4a. a) The inputs have lower synaptic weights (HRS). b) The inputs have higher synaptic weights (LRS)..

that the accumulation of these decaying stimuli is sublinear as observed in biological passive dendrites as well.

To ensure that the functionality of our passive dendritic elements was independent of the functionality of the passive synaptic elements. In Fig. 6, we demonstrate that by simulating the same neuron as in Fig. 4 with varying synaptic weights. We see in Fig. 6a that by lowering the synaptic weight to the HRS of the gated-RRAM, even though the sequence of the inputs is correct the high attenuation of those stimuli prevents the membrane potential from exceeding the threshold. In Fig. 6b, we see that by potentiating the synaptic weight back to LRS, the correct sequence is once again able to be detected by the neuron. Additionally, we can observe in Fig. 6 that moving from HRS to LRS in the synaptic weight had little effect on the delay and amplitude of the signal coming from the passive dendritic elements. Therefore, the synaptic weight and dendritic attenuation parameters can be independently tuned for their corresponding functionality with significant confidence.

V. Conclusion

In conclusion, we propose a reprogrammable dendritic-synapse architecture for neuromorphic neurons that integrated gated-RRAM to combine the functionality of synapses and passive dendrites in biological neurons. These dendritic-synapses can be used to drastically increase the computational capabilities of individual neurons as demonstrated via passive in-memory computation. The temporal aspects these dendrites introduce may allow for reduction of large recurrent neural networks into few power efficient neurons as evident by the single neuron sequence detection. Furthermore, the reprogrammability of gated-RRAM devices can allow for easier implementation of these dendritic-synapses for acceleration of inferencing and even development of algorithms for on-chip learning that needs to be investigated. We demonstrate by implementing more bio-inspired features into hardware neurons we are able to improve computational power of individual neurons. However, further work is required to capitalize on this additional functionality. Training algorithms must be co-designed with peripheral architectures to dynamically tune these available parameters. These algorithms and their corresponding peripheri will need to take into account the dataset, timing, power budget, and scalability of the target application. However, our goal in this work is to provide the foundational structure of these new algorithms by integrating the additional functions and tuning parameters we have proposed.

Acknowledgment

This project is funded by National Science Foundation under award number 1926465 and Rindsberg Fellowship provided by the University of Cincinnati.

REFERENCES

[1] S. Barve et al., "NeuroSOFM: A Neuromorphic Self-Organizing Feature Map Heterogeneously Integrating RRAM and FeFET," in IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, doi: 10.1109/JXCDC.2021.3119489.

[2] Li, X., Tang, J., Zhang, Q. et al. Power-efficient neural network with artificial dendrites. Nat. Nanotechnol. 15, 776–782 (2020). https://doi.org/10.1038/s41565-020-0722-5.

[3] J. Schemmel, L. Kriener, P. Müller and K. Meier, "An accelerated analog neuromorphic hardware system emulating NMDA- and calcium-based non-linear dendrites," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017, pp. 2217-2226, doi: 10.1109/IJCNN.2017.7966124.

[4] Spratling MW. Cortical region interactions and the functional role of apical dendrites. Behav Cogn Neurosci Rev. 2002 Sep;1(3):219-28. doi: 10.1177/1534582302001003003. PMID: 17715594.

[5] Bull L. Are Artificial Dendrites Useful in Neuro-Evolution? Artif Life. 2021 Nov 2;27(2):75-79. doi: 10.1162/artl_a_00338. PMID: 34727155.

[6] London M, Häusser M. Dendritic computation. Annu Rev Neurosci. 2005;28:503-32. doi: 10.1146/annurev.neuro.28.061604.135703. PMID: 16033324.

[7] H. Wan, X. Tian, J. Liang and X. Shen, "Sequence-Feature Detection of Small Targets in Sea Clutter Based on Bi-LSTM," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-11, 2022, Art no. 4208811, doi: 10.1109/TGRS.2022.3198124.

[8] T. J. Bailey, A. J. Ford, S. Barve, J. Wells and R. Jha, "Development of a Short-Term to Long-Term Supervised Spiking Neural Network Processor," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, doi: 10.1109/TVLSI.2020.3013810.

[9] A. Jones and R. Jha, "A Compact Gated-Synapse Model for Neuromorphic Circuits," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, doi: 10.1109/TCAD.2020.3028534.

[10] E. Herrmann, A. Rush, T. Bailey and R. Jha, "Gate Controlled Three-Terminal Metal Oxide Memristor," in IEEE Electron Device Letters, vol. 39, no. 4, pp. 500-503, April 2018, doi: 10.1109/LED.2018.2806188.

[11] Pedró, Marta et al. "Self-Organizing Neural Networks Based on OxRAM Devices under a Fully Unsupervised Training Scheme." Materials (Basel, Switzerland) vol. 12,21 3482. 24 Oct. 2019, doi:10.3390/ma12213482

[12] M. Jerry et al., "Ferroelectric FET analog synapse for acceleration of deep neural network training," 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2017, pp. 6.2.1-6.2.4, doi: 10.1109/IEDM.2017.8268338.