Energy Efficient Neuromorphic Processing using Spintronic Memristive Device with Dedicated Synaptic and Neuron Terminology

Zoha Pajouhi

***Abstract*—Research towards brain-inspired computing based on beyond CMOS devices has gained momentum in recent years. The motivation beyond this vigorous research prevails in exploitation of the resemblance between the computing principles and the device characteristics. To this end, the devices are used to perform otherwise time-consuming and power hungry tasks required for brain-inspired computing. Due to their miniaturized dimensions, zero leakage and nonvolatility, spintronic devices are among the most promising class of beyond CMOS devices. In this paper, we propose a novel spintronic structure based on antiferrromagnetically coupled domain walls. The device structure enables dedicated terminology for synaptic and neuron connections. This characteristic enables more efficient design of neuromorphic systems by allowing larger design space for designers. Furthermore, thanks to the coupling between the domain walls, the device can potentially operate at higher speeds while maintaining the energy consumption of the device; this higher speed contributes to improved performance of the neuromorphic system. In order to evaluate our proposed device structure, we developed a cross-layer simulation framework. Our simulation framework analyzes the neuromorphic system at the device, circuit and algorithm levels. Our simulation results show an order of magnitude improvement in the energy consumption compared to CMOS and analog neurons and up to 2X performance improvement as well as 8% improvement in the energy over state-of-the-art neuromorphic platforms using spintronic devices.**

Keywords*— *Spintronics, memristor, domain wall, neuromorphic computing, static coupling, Spin-hall effect, Dyzaloshinshkii-Moryia, Perpendicular domain walls, neural network, pattern recognition, IoT.

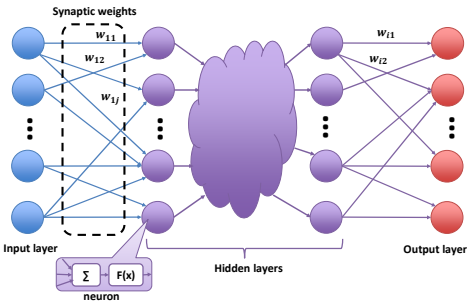


Fig. 1. Illustration of an Artificial Neural Network (ANN).

Zoha Pajouhi is with Intel Corporation, Email: zoha.pajouhi@intel.com. This article was prepared or accomplished by Zoha Pajouhi in her personal capacity. The opinions expressed in this article are the author's own and do not reflect the view of Intel Corporation.

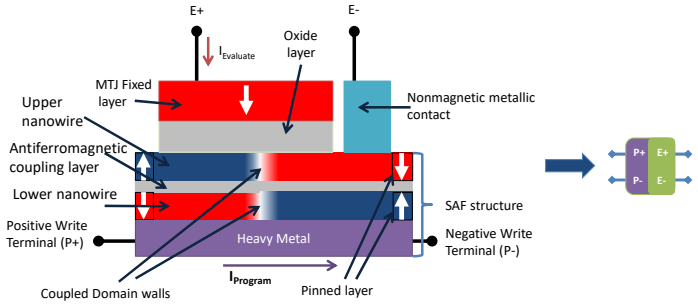
I. Introduction

The emergence of various computing devices and wearables connected to the internet (IoT) has motivated researchers to seek newer computing platforms that would extract meaningful information from the ever increasing data more efficiently and with reduced energy consumption. To this end, neuromorphic computing is among the most promising computing paradigms. The efficiency of neuromorphic computing platforms has motivated researchers towards the study of these platforms and investigation of implementing these types of systems in hardware. However, the implementation of neuromorphic computing in hardware on supercomputers based on typical Von-Neumann architectures has proven to be extremely inefficient [1-3]. Therefore, researchers have considered usage of non Von-Neumann architectures for neuromorphic computing. To this end, the advancements in the development of nanoscale devices with unique capabilities that resembles that of neural networks has broadened the horizon to increase the efficiency of such systems. Furthermore, various beyond CMOS nanoscale technologies [4-12] have been used to implement these architectures.

The vast majority of these devices are two terminal devices. They are typically operated in two different phases: a program/train phase where the conductance of the device is defined and a run/evaluate phase where the programmed conductance is exploited to evaluate the system. Therefore, there is a need to adjust the driving circuitry based on the requirements of the devices in each phase. This drawback potentially increases the design complexity and requires increased resources for operation.

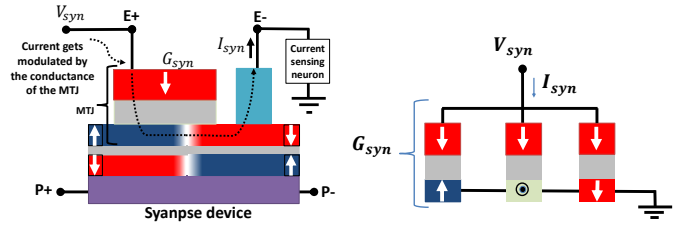
Furthermore, these devices have been utilized to perform either synapse or neuron operation and fail to integrate both functionality using the same device. This variegated usage of devices potentially increases the complexity of the fabrication of such systems. On the contrary, there are primitive proposals of devices that integrate the aforementioned functionalities [12]; however, the device does not provide dedicated terminology for each of the functions making the design of such systems challenging.

Despite the design challenges, spintronic neuromorphic systems have proven to be among the most efficient neuromorphic systems proposed so far [10-12]. Furthermore, spintronic memristors based on spin-hall effect have proved to be one of the most promising classes of such devices [13-15]. Recently, it has been shown that domain walls in synthetic antiferromagnetically coupled nanowires that are fabricated in



(a) The proposed device structure. (b) The device symbol.

Fig. 2. The proposed device structure.



(a) Usage of the device as a synapse. (b) Equivalent conductance of the evaluate path.

Fig. 3. Illustration of the device operation as a synapse.

the proximity of a heavy metal can be moved by passing a charge current through the nanowire(s) and the heavy metal [16].

In this paper, we propose a new four terminal device structure using spin-orbit and exchange coupled domain walls that is suitable for neuromorphic computing. Our device structure has dedicated and isolated terminals for different operations required for neuromorphic computing. This dedicated terminology expands the design space of these platforms and enables efficient operation of these platforms. Furthermore, the isolation between the programming and the evaluation ports provides significantly reduced complexity for the designers. Besides, we explain how the device can be used to realize both synaptic and neural functionalities and operate at a very low voltage. Furthermore, we developed a simulation framework to evaluate the effectiveness of our proposed device structure. Our simulation framework consists of a device simulation module, circuit analysis module and an algorithm interpretation module. Using our framework, we implemented a small-scale neural network for pattern recognition. Our results show up to 2X improvement in the performance as well as 8% reduction in the energy consumption compared to state of the art spintronic implementations.

The rest of the paper is organized as follows: in Section II, we propose the new device structure and our proposed cognitive computing system. In Section III we will explain our simulation framework and the methodology used to evaluate our proposed system. In Section IV, we will explain the simulation results and finally Section V concludes the paper.

II. Neural Network Implementation

There are different neural network structures and computing models. One of the most widely used type of neural networks are the Artificial Neural Networks (ANN). Thanks to their relative simplified computations, they are widely used in various applications. Therefore, we have focused on implementing this type of neural network. Fig. 1 illustrates a sample ANN. Without loss of generality, let us consider a feedforward ANN where the data flows from the inputs to outputs. To this end, the inputs are connected to neurons through synapses. Each synapse contains a weight; representing the importance of the input to a certain neuron. The sum of the weighted inputs are then summed up in every neuron and a transfer function translates the summed input to the output of the neuron. The output of the neuron in turn is the input to the neuron in the next stage through a synapse.

In order to use a feedforward ANN, there is a need to train the network; that is, to define the value of the weights associated with each synapse. There are various algorithms for this purpose that have their own pros and cons. Without loss of generality; and considering the training details, from structural point of view, neuromorphic platforms use beyond CMOS devices to implement synapses [17,18], neurons [9], or both functionalities [12]. The basics of such utilizations originates from the fact that the majority of these devices have an adjustable resistance. The resistance is adjusted during the training process and is fixed once the training is complete. Therefore, these devices are utilized in two different
modes: The first mode, where the training takes place, it is desired to change the resistance of the device. On the other hand, in the evaluation mode, it is generally desired to avoid any change in the value of the resistance. Researchers have made attempts to facilitate the design of neuromorphic systems by proposing a third terminal (see for example [12,19]); however, the program and evaluation phases still share a path to the ground. This shared path imposes restrictions on the driving circuits during the evaluation phase to avoid undesired modification in the resistance of the device. In the next subsection, we propose a four terminal device that has decoupled and isolated paths for programming and evaluation phases.

A. Proposed 4-terminal neuromorphic device

The four terminal device structure is based on spin-orbit and exchange coupled Domain Walls (DWs). Fig. 2 (a) shows the structure of this device and Fig. 2(b) shows the symbol we will be using throughout the rest of the paper to illustrate the device more easily. The device consists of a heavy metal layer adjacent to two magnetic layers: a lower magnetic layer (LM) and an upper magnetic layer (UM) separated by an antiferromagnetic insulating spacer. The LM and the UM have perpendicular magnetic anisotropy and are anti-ferromagnetically coupled to each other. Let us call the structure explained so far as the Synthetic Antiferromagnetic Structure (SAF). Furthermore, an oxide layer is fabricated adjacent to the UM and finally a pinned layer is positioned on top of the oxide layer.

The device has two programming pins (P+,P-) and two evaluate terminals (E+,E-). If a current is passed through the programming pins, (P+,P-), it passes through the HM and the LM, the reason being, its flow is blocked by the insulating layer from reaching the UM. On the other hand, thanks to the giant spin Hall effect, the current passing through the HM gets spin polarized and asserts a torque on the DWs contributing to their movement. The polarity of the current defines the direction of the movement; e.g. if $V_{P+}>V_{P-}$ ($V_{P-}>V_{P+}$) the DWs move to the left (right). Once the position of the DWs are defined through this method, the device maintains its state regardless of its connection to the power supply.

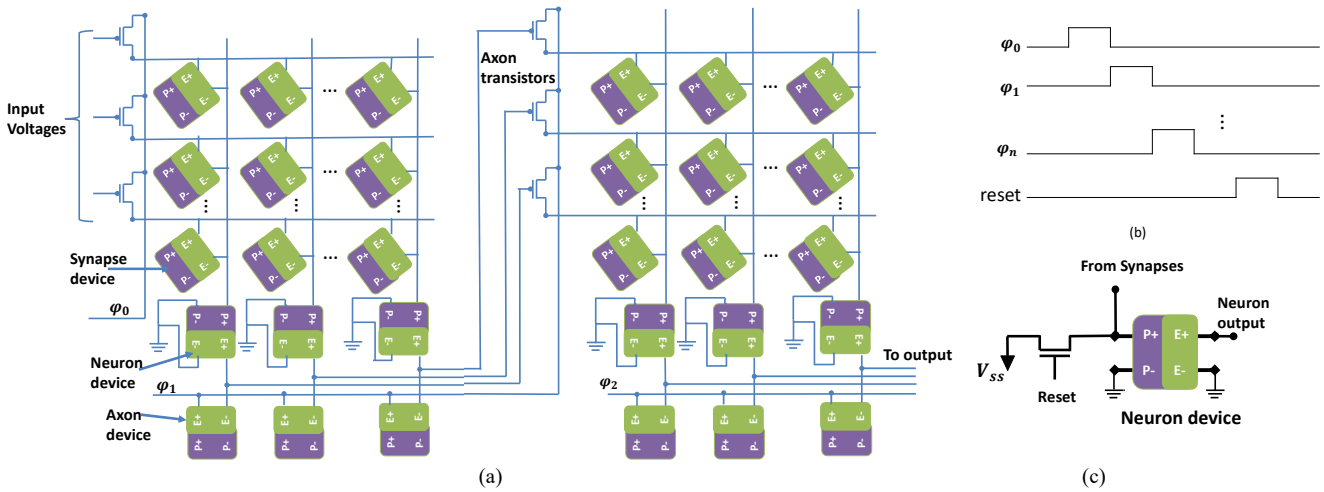


Fig. 5. Illustration of the connections in the Neural Network. (a) Illustration of the connections between the input layer and a hidden layer. (b) Illustration of the timing of signals required for correct operation. (c) The reset circuitry.

Moreover, observe in Fig. 2 that the fixed layer, the oxide and the UM form a Magnetic Tunnel Junction (MTJ) [20]. The conductance of the MTJ is defined by the position of the DWs. This conductance is sensed by passing a small current between the evaluate ports (E+,E-). The conductance between these two terminals spans between the minimum conductance; corresponding to the DWs located at the far right hand side (G_{ap}) and the maximum conductance corresponding to the DWs located at the far left hand side (G_p).

B. Synapse Implementation

Synapses connect each neuron to the following neuron through a weighting function. The weight of the neuron is defined through training of the neural network. Memristive devices have been used to implement the synaptic functionality, see for example [17,18]. To this end, the resistance of the memristive device is programmed based on the desired weight. Once the resistance of the memristive device is defined, it is used to evaluate the results of the neural network. It is noteworthy to mention that the vast majority of such memristive devices possess two terminals. Meaning that their programming and readout terminals are shared. To the best of our knowledge, this proposal is the first device that has fully decoupled programming and evaluation paths. Fig. 3 (a) illustrates the usage of the device to realize a synaptic function. For a fixed applied voltage between terminals E+ and E-, the current that passes through the device depends on the resistance of the device between these two nodes as illustrated in Fig. 3 (b). To this end, the device has three resistances in parallel, the first resistance, is the resistance that forms between the fixed layer and the portion of the nanowire on the left side of the DWs; the second resistance is the resistance formed between the fixed layer and finally the third resistance is the resistance formed between the fixed layer and the portion of the nanowire on the right side of the DWs. This resistance can be written as:

$$G_{syn} = R_{syn}^{-1},$$

$$G_{syn}(x) = (G_{p,mtj}\left(\frac{x}{L}\right)) + G_{ap,mtj}\left(1-\frac{x}{L}\right) + G_{DWs} \quad (1)$$

where $G_{p,mtj}$ is the conductance of the MTJ when the DWs are located at the far left side of the nanowire and $G_{ap,mtj}$ is the conductance of the MTJ when the DWs are located at the far right side of the nanowire. Also, L is the length of the oxide and x is the location of the DWs. Observe in Eq. 1 that the conductance of the structure is roughly linearly dependent on the position of the DWs. Therefore, for a fixed voltage of V_{syn}, the current I_{syn}=$G_{syn}V_{syn}$ is modulated by the conductance G_{syn} which is in turn dependent on the DWs position. On the other hand, the location of the DWs can be adjusted through terminals P+ and P- during the training.

C. Neuron functionality

Now let us explain the functionality of the device for neuron implementation. In an ANN, each neuron is connected to a set of synapses coming from its input neurons. The value of the input neurons are weighted through the synapses before reaching the neuron input. These weighted values are summed up and passed through a transfer function as illustrated in Fig.1. Our method of implementing a neuron as well as its connection to the axon is illustrated in Fig. 4. The voltages V_1 through V_n are the input voltages. These voltages are applied to the synaptic devices S_1 through S_n. Under such circumstances, currents I_1 through I_n are calculated by:

$$I_i = (V_i - V_{sense}) * G_i, I_{sense} = \sum_{i=1}^{n} I_i \quad (2)$$

where V_{sense} is the voltage of the sensing circuit and I_{sense} is the current passing through the sensing circuit. On the other hand, if we consider that the sensing device has a very small resistance compared to the synaptic devices; it would result in $V_{sense} << V_i$. Therefore, if we consider $V_{sense} \cong 0$, it can be concluded that:

$$I_i \cong V_i * G_i, I_{sense} \propto \sum_{i=1}^{n} V_i * G_i \quad (3)$$

Therefore, the I_{sense} is proportional to the sum of the weighted input voltages. In our implementation, we use the neuromorphic device as the sensing circuit. Specifically, the neurons are

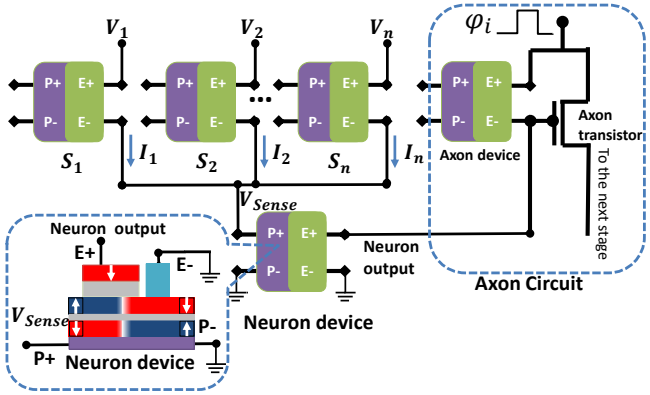


Fig. 4. Illustration of usage of the device as a neuron and its connection to the axon.

connected through their (P+,P-) terminals as illustrated in Fig. 4. To this end, once the current passes through these terminals, it gets spin polarized and inserts a torque on the DWs which causes their movement. Eventually, the position of the DWs is sensed through the axon circuitry. It is noteworthy to mention that the programming path is a fully metallic path. Therefore, the resistance of this path is relatively small. On the other hand, as we will explain in Section III, the resistance of the evaluate path depends on the dimensions of the MTJ. These dimensions will be adjusted to ensure that V_{sense} would be substantially smaller than the input voltages.

D. Axon circuitry

An axon in an ANN is used to connect every neuron to its outputs. This connection may implement a transfer function that relates the output value of the neuron to the weighted sum of its inputs. The axon circuitry consists of a PMOS transistor as well as another neuromorphic device. The position of the DWs in the axon device can be adjusted to tune V_{ax} which serves as the input to the PMOS transistor. Furthermore, the PMOS transistor is used to connect the output of the neuron to the next stage and acts as a buffer between the two stages. Specifically, as the position of the DWs in the neuron device changes, the gate voltage of the PMOS transistor changes. This change results in a change in the V_{ds} of the device. The drain of the PMOS is then connected to the input of the next stage. Therefore, the V_{ds} acts like the input voltage to the intermediate stage. It is noteworthy to mention that although the intermediate stage has loading effect on the PMOS. There are two main strategies that we have implemented to address such an effect. The first strategy, is to use an axon device to further calibrate the input to the axon transistor. The second strategy is to insert a dummy device is at the end of each row along with other synapses.

E. Evaluation procedure

So far, we have explained the neuromorphic implementation from the structural point of view. In this subsection, we will elaborate on the procedure at which the neuromorphic implementation should be operated to ensure correct evaluation. The first step for every neuromorphic implementation is the training of the neural network. There are various methods for implementing the training of neural network including online or offline. Although our implementation can be used for online training; for the sake of simplicity, we consider that the training has occurred offline and before going to the evaluation step. The evaluation scenario consists of 2 steps: the "execution", and "reset". Fig 5 illustrates the three different steps required for evaluation.

The first step is the execution step. For now, let us consider that the DWs are located at the far left side of the device. The execution step starts when input voltages are applied to pins V_1 through V_n, the voltage is modulated by the conductance of the branches G_1-G_n. All of these currents add up and pass through the *P*+,*P*- pins of the device. This current pass causes the movement of the DWs; displacing them from their original location. The execution step is concluded by disconnecting the inputs. This causes the DWs to stop and retain their position. Note that the execution step for the *ith* hidden layer is activated by pulling up the φ_i. Also, the input layer is activated by pulling up φ_0. We shall mention that the execution step for each layer acts as a read-out for the previous layer. Specifically, in the "read-out", the φ_{i+1} signal is activated and a current is passed across terminals (E+,E-) of the neuron of the *ith* layer to sense the resistance of the device and to connect it to the next stage through the axon circuitry.

Once all of the execution pulses are pulled up once, the neuron devices are reset by passing a negative current through the neuron device to restore the position of the DWs to the original position at the far left side of the nanowires. To this end, the reset transistor is activated by pulling up the *reset* signal as illustrated in Fig. 5 (c).

III. Simulation framework

In order to evaluate the proposed neuromorphic implementation, we developed a simulation framework. Fig. 6 illustrates our simulation framework. Our simulation framework consists of three main modules. The first module is the device analysis module. The device analysis module is responsible for simulation of the device behavior with respect to the current/voltage applied to it. The second module is the circuit analysis module. This module is responsible for analysis of the neuromorphic system at the circuit level. This include interfacing the neuromorphic device with the CMOS transistors and the power supply. Finally, the third module is the algorithmic and interpretation module. This module is responsible for interpreting the data embedded in the circuit and to embed it into the neural network algorithm for evaluation purposes. We will explain each module in detail in the following subsections.

A. Device analysis module

The device analysis module consists of two main submodules. The first submodule is responsible for analyzing the dynamics of the DWs. Specifically, this module consists of a magnetization dynamics solver based on Landau-Lifshitz-Gilbert (LLG) formalism. The second submodule consists of an electronic transport simulation analysis and implementation based on Nonequilibrium Green's Function (NEGF). This submodule is utilized to calculate the conductance of the evaluate path based on the position of the DWs.

1) The micromagnetic simulation submodule

In order to explain the micromagnetic simulation submodule, there is a need to understand the underlying physical mechanism that contributes to the movement of the DWs. Recently, it has been shown that DWs formed from two perpendicularly magnetized multilayers separated by an ultrathin antiferromagnetic layer can be moved by current very efficiently [16]. The movement is possible due to the interfacial chiral spin torques.

The DWs have a chiral Neel structure due to a Dzyaloshinskii-Moriya exchange interaction (DMI) [21-23] derived from strong spin-orbit coupling and the proximity of the heavy metal layer. When there is no current passing through the DWs, that is through terminals *P*+ and *P*- in Fig. 2, the magnetization rotates from up to down anticlockwise in a plane parallel to the length of the nanowire such that the magnetization in the middle of the lower DW (M_L) is aligned to the nanowire. If the exchange coupling between the two DWs is strong enough, the orientation of the magnetization at the middle of the upper layer (M_U) is opposite that of the lower layer.

However, when current passes in the direction along the nanowire (through terminals *P*+ and *P*-), if there is no exchange coupling between the two layers, the spin current generated by

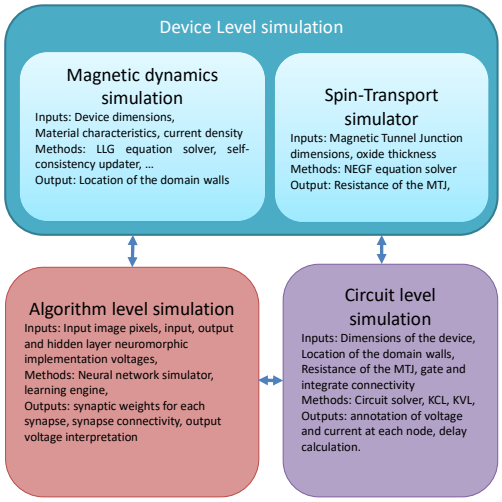


Fig. 6. Simulation framework used for neuromorphic system evaluation.

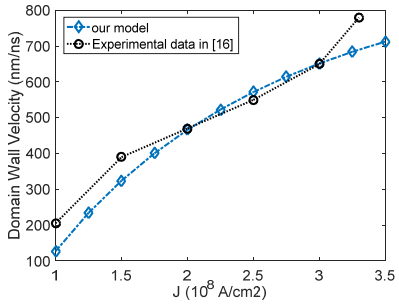


Fig. 7. Comparison of the results obtained by the simulation framework an the experimental data in [16].

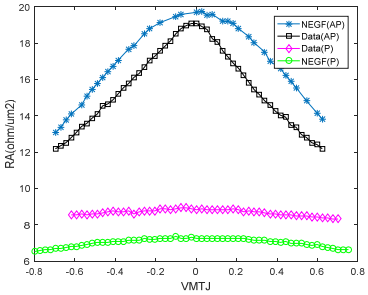


Fig. 8. NEGF calculation vs experimental conductance.

TABLE 1. DIMENSIONS OF NEUROMORPHIC DEVICE

PARAMETER	VALUE
Ferromagnet thickness	1nm
Heavy metal layer thickness	1.5nm
Domain wall width	4.3nm
Lower nanowire cross-section	16nm×1nm
Upper nanowire cross-section	16nm×1nm
Fixed layer cross-section	16nm×1nm
Nonmagnetic contact dimensions	16nm×16nm×1nm
Insulating coupling layer thickness	0.8nm
Oxide layer thickness	0.9 nm – 2 nm

the spin Hall effect, induces a torque on the DW in the lower layer that results in rotation of the magnetization of the DW towards the accumulated spin direction which is traverse to the length of the nanowire. This spin Hall torque together with the DMI field is the main contributor of the motion of the DW [16]. The magnitude of the spin Hall torque depends on the angle between M_L (M_U) and the spin accumulation direction. An increase in the current results in the rotation of M_L or M_U towards the spin accumulation direction resulting in a Bloch DW structure in which M_L (M_U) is oriented traverse to the nanowire and the spin Hall torque becomes zero. As a result, the DW velocity becomes saturated at a certain current.

On the other hand, if the coupling between the two layers is considered, the structure stabilizes the Neel DW structure. Therefore, the spin-orbit torque is much larger with the same current density resulting in a much higher DW velocities. In addition, this structure gives rise to a novel torque associated with the antiferromagnetic exchange-coupling field. This exchange coupling torque is roughly an order of magnitude larger than the DMI field and is the key contributor to the high velocity of the DW [16]. The enhanced velocity of the DWs causes larger displacement of the DWs at the same energy and performance of the DW without the antiferromagnetic layer causing more energy-efficient neuromorphic computing [16].

The dynamics of the spin-orbit and exchange coupled DWs can be expressed as three LLG equations that should be solved jointly as follows [16]:

$$\dot{q} = \frac{\alpha_L \alpha_U}{\alpha_U M_L(1+\alpha_L^2)+\alpha_L M_U(1+\alpha_U^2)}\xi \tag{4}$$

$$\dot{\psi}_L = \frac{1}{\Delta}\frac{\alpha_U}{\alpha_U M_L(1+\alpha_L^2)+\alpha_L M_U(1+\alpha_U^2)}\xi + \frac{\gamma}{\alpha_L}\{\frac{H_L^k}{2}\sin 2\psi_L - \frac{\pi}{2}H_L^{DM}\sin\psi_L - \frac{2J^{ex}}{M_L}\sin(\psi_L-\psi_U)\} + \frac{u_L}{\alpha_L\Delta} \tag{5}$$

$$\dot{\psi}_U = -\frac{1}{\Delta}\frac{\alpha_L}{\alpha_U M_L(1+\alpha_L^2)+\alpha_L M_U(1+\alpha_U^2)}\xi + \frac{\gamma}{\alpha_U}\{\frac{H_U^k}{2}\sin 2\psi_U - \frac{\pi}{2}H_U^{DM}\sin\psi_U - \frac{2J^{ex}}{M_U}\sin(\psi_U-\psi_L)\} + \frac{u_U}{\alpha_U\Delta} \tag{6}$$

$$\xi = [-M_L\left(\frac{1}{\alpha_L}+\beta_L\right)u_L - M_U\left(\frac{1}{\alpha_U}+\beta_U\right)u_U - \frac{\gamma\Delta M_L}{\alpha_L}\left\{\frac{H_L^k}{2}\sin 2\psi_L - \frac{\pi}{2}H_L^{DM}\sin\psi_L - \frac{2J^{ex}}{M_L}\sin(\psi_L-\psi_U) + \frac{\alpha_L\pi H_L^{SH}}{2}\cos\psi_L\right\} +$$

$$\frac{\gamma\Delta M_U}{\alpha_U}\{\frac{H_U^k}{2}\sin 2\psi_U - \frac{\pi}{2}H_U^{DM}\sin\psi_U - \frac{2J^{ex}}{M_U}\sin(\psi_U-\psi_L) + \frac{\alpha_U\pi H_U^{SH}}{2}\cos\psi_U\}] \tag{7}$$

where q is the position of the DW and $\dot{q}$ is its velocity. It is noteworthy to mention that due to the antiferromagnetic coupling between the DWs, they have the same position regardless of being in the UM or LM. Also, it is considered that the length of the DWs is equal to Δ. Besides, ψ_i (i=L: lower layer, U: upper layer) are the DW angles for the corresponding DWs. Furthermore, $\dot{\psi}_i$ is the derivative of the corresponding DW angle. Besides, γ is the gyromagnetic ratio and J^{ex} is the exchange coupling constant between the two DWs. Furthermore, α_i, β_i are the Gilbert damping constant and the non-adiabatic STT coefficients of the corresponding nanowires. Furthermore, H_i^k is the magnitude of in-plane anisotropy field ($H_i^k = 2K_i^u t_i/M_i$) derived from the shape anisotropy of the DW that favors a Block DW configuration ($\psi_i = \pm\pi/2$) over that of a Neel wall ($\psi_i = 0, \pi$). H_i^{DM} is the DMI exchange field. The volume spin transfer torque from the current within the i-th magnetic layer is derived from $u_i = \frac{\mu_B P_i J_i t_i}{e M_i}$, where μ_B is the Bohr magneton e is the electron charge and P_i is the spin polarization of the current and J_i is the current density of the i-th magnetic layer [16].

H_i^{SH} is the torque associated with the spin Hall. $H_L^{SH} = \frac{\hbar\theta_L^{SH} J_{HM}}{2eM_L}$ where J_{HM} is the current density in the heavy metal. Furthermore, the H_U^{SH} (for the upper layer) is smaller than H_L^{SH} since the spin Hall current is attenuated along the LM and the coupling spacer layer.

Our simulation framework was calibrated to the data represented in [16]. To this end, we used the same parameters explained in [16] to calibrate our model to the experimental results as shown in Fig. 7. Observe in Fig. 7 that our model matches the experimental data.

2) The Electron transport module and evaluation path calculation

In order to evaluate the position of the DWs, a small current is passed through ($E+$,$E-$) pins of the device. Note that the read mechanism is similar to that of a magnetic tunnel junction (MTJ) with the free layer replaced by the exchange-coupled DWs. Therefore, the conductance of the structure can be modeled as three different conductances in parallel. The first conductance is the conductance formed by the fixed layer and the nanowire at the left side of the DWs, the second conductance is the conductance between the fixed layer and the DWs; the third conductance is the conductance between the fixed layer and the nanowire on the right side of the DWs. This assumption is a viable assumption and is similar to the methods used in [24-25]. Our model is based on the NEGF formalism compact model [26,32,33]. The evaluate resistance of the neuromorphic device was considered to be [25]:

$$G \propto A(\exp(\alpha T_{ox} + \beta) + \exp(\gamma T_{ox} + \delta)) \quad (8)$$

where G is the conductance of the MTJ, A is the area of the fixed layer, T_{ox} is the oxide layer thickness and $\alpha,\beta,\gamma,\delta$ are fitting parameters. We have calibrated our model to the experimental results in [27]. Fig. 8 compares the NEGF simulation framework with the experimental results in [27].

It is noteworthy that, deterministic domain wall movement have been experimentally demonstrated in magnetic multilayer structures with a single nanowire [11,13,14,28].

Regarding the evaluate operation, one viable concern that can be raised is the disturbance of the DWs. To this end, an undesired scenario is that the spin polarized current passing through the upper layer could potentially change the position of the DWs. However, if the neuromorphic device structure is considered, the depinning current of the upper DW is substantially higher than that of the lower DW. The reason being, the lower nanowire is adjacent to the heavy metal. Due to the giant spin Hall effect in the heavy metal, the spin polarization is higher in the lower nanowire. On the other hand, the upper nanowire is deprived from such a proximity resulting in higher DW depinning.

TABLE 2. SIMULATION PARAMETERS FOR NEUROMORPHIC DEVICE

PARAMETER	VALUE
Saturation Magnetization lower nanowire (M_L)	7×10^{-5}emu/cm^2
Saturation Magnetization upper nanowire (M_U)	6.5×10^{-5}emu/cm^2
Spin-Hall angle	0.1
Perpendicular Magnetic Anisotropy	5×10^5 J/m^3
Effective DMI constant	1.2×10^{-3} J/m^2
Resistivity of Pt	200Ω.nm
Resistivity of the ferromagnet	150Ω.nm
Exchange coupling field J_{ex}	-0.5 erg/cm^2
Gilbert damping factor	0.1
Polarization	0.8
V_{dd}/V_{ss}	500mV/-200mV

B. Circuit simulation module

The circuit simulation module is responsible for simulation of CMOS transistors and its interface with the neuromorphic devices. To this end, the circuit simulation module is used to interface between both the synapse devices as well as the neurons. Furthermore, it interfaces with the device simulation module, namely, the micromagnetic simulation module and the electron transport module. Also, it is used for analyzing the required timing for each of the steps, namely, the execution, read-out and the reset steps. Furthermore, it is used to calculate the energy consumption of the circuit and to interface with the algorithm interpretation analysis module.

C. Algorithm interpretation module

The algorithm interpretation module is responsible for simulating the system at the algorithm level. To this end, the inputs and outputs of the neural network are passed to this module. This module performs simulation of the neural network as well as optimization of the network. Furthermore, this module is responsible for training the network and the trained weights are passed to the circuit simulation module. On the other hand, the outputs data from the circuit simulation module are transferred back to the algorithmic interpretation module to evaluate the correctness of the results.

IV. Results and Discussion

In order to evaluate our proposed neuromorphic implementation, we implemented a small scale ANN for character recognition. Specifically, we utilized a commonly used database for character recognition, namely MNIST data base [29]. To this end, the input images were considered to be 28x28 pixels. Therefore, there are a total of 784 inputs to the neural network. There is one hidden layer of neurons consisting of 15 neurons as well as 10 output neurons representing digits 0 to 9. The neuron transfer function was considered to be piecewise linear. At small inputs, the output was considered to be constant until the input reaches a certain value. If the input is above that value, the output changes linearly with the input. Finally, once a certain input voltage is reached, the output saturates.

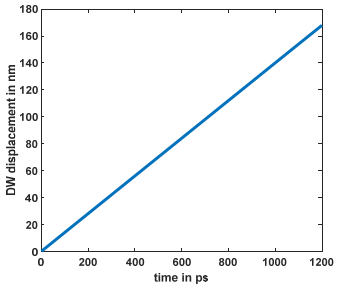


Fig. 9. Displacement of the domain wall for a current pass of 20 µA.

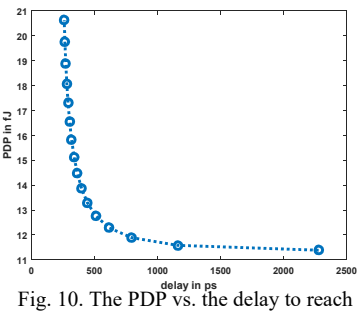


Fig. 10. The PDP vs. the delay to reach a displacement of 160 nm.

TABLE 3. COMPARISON BETWEEN NEURONS

NEURON DESCRIPTION	NEURON DELAY	POWER DELAY PRODUCT
Nanoscale silicon memristor [6]	500 ms	1.75 µJ
Lateral Spin Valve neuron [9]	500 ps	400 fJ
Spin Orbit neuron [10]	1 ns	15 fJ
Domain Wall neuron [12]	2 ns	6.1 fJ
Digital CMOS neuron [34]	10 ns	832 fJ
Analog CMOS neuron [8]	10 ns	700 fJ
This work	500 ps	5.73 fJ

Now let us explain the simulation results obtained from our framework. In order to ensure the manufacturability of our proposed neuromorphic system, we considered implementation in 16 nm Technology node. To this end, we used a Predictive Technology Model (PTM) of the CMOS devices in this node [30]. Furthermore, the transistors were upsized to reduce the drain source voltage drop (V_{ds}) of the transistors. The power supply for our implementation was considered to be *500 mV*. Eventually, the neuromorphic devices were designed to accommodate for the required precision. We considered that the

position of the DWs could be sensed with an accuracy of *5 nm*. Just about the size of the DWs. Our simulation results show that an accuracy of 5 bits is sufficient for our application; therefore, the length of the devices were considered to be *160 nm*. The dimensions of the neuromorphic devices used in our simulation is illustrated in Table 1. Note that the synaptic weights are tuned during training. Therefore, during normal operation, they are fixed to a certain value. Furthermore, we considered the critical current of the synapses as the required current to move the DWs for 5nm after *100 ps*. The critical current of the devices were measured to be *~43 μA*. It is noteworthy to mention that this current plays an important role in the performance of the system. Specifically, recall from Section III that each neuron is connected to several synaptic devices that are used to apply the correct weighting to the incoming currents. Therefore, the synaptic devices share a common path with the neuron devices. On the other hand, the position of the DWs should be constant during the execution for the synaptic devices as it is desired to change the position of the DWs in the neuron. This variegated requirement of the usage of the device is a limiting factor in common neuromorphic designs utilizing beyond-CMOS devices for both synaptic and neuron functionalities (see for example [10, 31]). A common practice is to limit the operating current to avoid disturbance of the synapses. This limitation crucially impacts the performance of the neuromorphic system. However, in our proposed implementation, due to the high critical current, the neuromorphic system can potentially operate at higher performance. As an example, Fig. 9 shows the displacement of the DWs for a programming current of *20 μA*.

The neural network can be optimized for performance or energy. Specifically, each neuron has an inherent trade-off between time and energy. The domain wall speed increases with an increase in the current density. This behavior is observable in Fig. 7 as well. This implies that if higher performance is desired (e.g. higher velocity), the power and energy consumption increases. To this end, let us consider the power delay product of a neuron. Specifically, we consider the entire power consumed by the neuron as well as the synapses attached to it together with the axon transistor. Furthermore, let us consider that a neuron should be programmed to a certain location; e.g. *160 nm* displacement after the first execution epoch. Fig. 10 plots the energy consumption of the neuron vs. time required to get there. Observe in Fig. 10 that the energy consumption decreases drastically with an increase in the delay. However, this decrease plateaus after *~500 ps*. Therefore, we considered *500 ps* as the required epoch for the execution time.

As explained earlier in Section III, our proposed implementation operates in 2 phases: the execution and the reset operations. In our implementation, we considered each execution and read-out epoch to be equal to *500 ps*. Since our network has two layers, the total execution time for the network is equal to *1 ns*. Specifically, in the first *500 ps*, the first layer is executed and in the second *500 ps*, the first layer is read-out by the second layers axon and the output layer is executed simultaneously. Finally, the output layer is read-out. The final read-out also takes *500 ps*. Once the outputs are read out, the neurons should be reset back to the original value. Thus, the entire cycle takes *2 ns* to complete.

Eventually, we would like to discuss the energy consumption of the network. The energy consumption has two main components: The execution component and the reset component. The execution component pertains to the time when the circuit is operating and the reset component pertains to the energy required to restore the DWs position for correct operation of the next evaluation cycle. We ran our simulations for pattern recognition and our results show an average neuron current of *10.3 μA* through the time epochs explained earlier. Furthermore, the energy consumption of each neuron can be calculated as *VIt* as explained earlier. Our results show an average energy consumption of *5.73 fJ*. On the other hand, the read-out of the final stage of the network is less than *0.001 fJ* making it negligible. Finally, the reset energy for our implementation is equal to *3 fJ*. The reason for this high energy consumption is that in the reset scenario, the reset transistor should be designed in such a way that the all of the DWs reach their original position regardless of their current location. Meaning that it should be high enough to move the DWs the entire *160 nm* resulting in increased energy consumption.

Finally, we would like to compare our proposed neuromorphic system with the implementations available in literature. Table 3 shows the energy consumption of different neurons using different neuromorphic devices. Specifically, we have used the same *VIt* formula as explained earlier for this comparison. Observe in Table 3 that our proposed method has an order of magnitude less energy consumption compared to an analog and a digital neuron.

One viable concern regarding the proposed system is the impact of process variations. Note that neural networks are inherently error-resilient. Therefore, the process variations can be addressed at the time of training the network. On the other hand, we have considered adjusting every neuron's strength by using the axon device. If process variations is not considered, one valid question for the essence of the axon device is that the neural network and the output can be adjusted by designing the neurons and the synapses of each layer through proper design. Therefore, the axon device could potentially be replaced by a simple MTJ or a fixed resistance. However, due to process variations and mismatch between the neurons, there is a need for a circuitry to adjust the output value with respect to the input. Therefore, the axon device enables tuning the output of the neuron for the next stage.

V. Conclusion

In this paper, we proposed a new device structure suitable for neuromorphic computing based on spin-orbit and exchange coupled nanowires. Furthermore, we explained how our 4-terminal device can be used to realize both synaptic and neuron functionality. Additionally, we developed a simulation framework based on the experimental results in literature and used it to evaluate a simple ANN for character recognition. Our results show that neuromorphic computing structures can be implemented very efficiently using our proposed device structure. Our simulation results show an order of magnitude improvement in the energy consumption compared to CMOS and analog neurons and up to 2X higher performance as well as 8% improvement in the energy consumption over state-of-the-art neuromorphic platforms using spintronic devices.

References

[1] IBM unveils a new brain simulator. IEEE Spectrum, (2009).

[2] E. Linn, R. Rosezin, S. Tappertzhofen, U. Böttger, and R. Waser. "Beyond von Neumann—logic operations in passive crossbar arrays alongside memory operations." *Nanotechnology* 23, no. 30 (2012): 305205.

[3] G. Indiveri, B. Linares-Barranco, R. Legenstein, G. Deligeorgis, and T. Prodromakis. "Integration of nanoscale memristor synapses in neuromorphic computing architectures." *Nanotechnology* 24, no. 38 (2013): 384010.

[4] B. L. Jackson, B. Rajendran, G. S. Corrado, M. Breitwisch, G. W. Burr, R. Cheek, K. Gopalakrishnan, S. Raoux, C. T. Rettner, A. Padilla et al., "Nanoscale electronic synapses using phase change devices," ACM Journal on Emerging Technologies in Computing Systems (JETC), vol. 9, no. 2, p. 12, 2013.

[5] D. Kuzum, R. G. Jeyasingh, B. Lee, and H.-S. P. Wong, "Nanoelectronic programmable synapses based on phase change materials for braininspired computing," Nano letters, vol. 12, no. 5, pp. 2179–2186, 2011.

[6] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," Nano letters, vol. 10, no. 4, pp. 1297–1301, 2010.

[7] T. Chang, Y. Y. Yang, and W. Lu. "Building neuromorphic circuits with memristive devices." *IEEE Circuits and Systems Magazine* 13, no. 2 (2013): 56-73.

[8] M. Sharad, D. Fan, and K. Roy, "Spin-neurons: A possible path to energy-efficient neuromorphic computers," Journal of Applied Physics, vol. 114, no. 23, p. 234906, 2013.

[9] M. Sharad, C. Augustine, G. Panagopoulos, and K. Roy, "Spin-based neuron model with domain-wall magnets as synapse," Nanotechnology, IEEE Transactions on, vol. 11, no. 4, pp. 843–853, 2012.

[10] A. Sengupta, S. H. Choday, Y. Kim, and K. Roy, "Spin orbit torque based electronic neuron," Applied Physics Letters, vol. 106, no. 14, p. 143701, 2015.

[11] S. Emori, U. Bauer, S.-M. Ahn, E. Martinez, and G. S. Beach, "Current driven dynamics of chiral ferromagnetic domain walls," Nature materials, vol. 12, no. 7, pp. 611–616, 2013.

[12] A. Sengupta, Y. Shim, and K. Roy. "Proposal for an All-Spin Artificial Neural Network: Emulating neural and synaptic functionalities through domain wall motion in ferromagnets." IEEE Trans. On Biomedical Cir. And Syst., early access, accessed 8/15/2016.

[13] E. Martinez, S. Emori, N. Perez, L. Torres, and G. S. Beach, "Current driven dynamics of Dzyaloshinskii domain walls in the presence of inplane fields: Full micromagnetic and one-dimensional analysis," Journal of Applied Physics, vol. 115, no. 21, p. 213909, 2014.

[14] S. Emori, E. Martinez, K.-J. Lee, H.-W. Lee, U. Bauer, S.-M. Ahn, P. Agrawal, D. C. Bono, and G. S. Beach, "Spin Hall torque magnetometry of Dzyaloshinskii domain walls," Physical Review B, vol. 90, no. 18, p. 184427, 2014.

[15] K.-S. Ryu, S.-H. Yang, L. Thomas, and S. S. Parkin, "Chiral spin torque arising from proximity-induced magnetization," Nature communications, vol. 5, 2014.

[16] S. H., Yang, See-Hun, K. S. Ryu, and S. Parkin. "Domain-wall velocities of up to 750 m s− 1 driven by exchange-coupling torque in synthetic antiferromagnets." *Nature nanotechnology* 10.3 (2015): 221-226.

[17] S. H. Jo, Sung Hyun, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu. "Nanoscale memristor device as synapse in neuromorphic systems." *Nano letters* 10, no. 4 (2010): 1297-1301.

[18] H. Kim, M. P. Sah, C. Yang, T. Roska, and L. O. Chua. "Neural synaptic weighting with a pulse-based memristor circuit." *IEEE Transactions on Circuits and Systems I: Regular Papers* 59, no. 1 (2012): 148-158.

[19] G. Srinivasan, A. Sengupta, and K. Roy. "Magnetic Tunnel Junction Based Long-Term Short-Term Stochastic Synapse for a Spiking Neural Network with On-Chip STDP Learning." *Scientific Reports* 6 (2016).

[20] Nishimura, Naoki, Tadahiko Hirai, Akio Koganei, Takashi Ikeda, Kazuhisa Okano, Yoshinobu Sekiguchi, and Yoshiyuki Osada. "Magnetic tunnel junction device with perpendicular magnetization films for high-density magnetic random access memory." *Journal of applied physics* 91, no. 8 (2002): 5246-5249.

[21] M. Heide, M., G. Bihlmayer, S. Blügel, "Dzyaloshinskii–Moriya interaction accounting for the orientation of magnetic domains in ultrathin films: Fe/W(110)," Phys. Rev. B 78, 140403 (2008).

[22] A. Thiaville, S. Rohart, E. Jue, V. Cros, A. Fert, "Dynamics of Dzyaloshinskii domain walls in ultrathin magnetic films," Europhys. Lett. 100, 57002 (2012).

[23] J. E. Hirsch, Spin Hall effect. Phys. Rev. Lett. 83, 1834–1837 (1999).

[24] D. Morris, D. Bromberg, J. J. Zhu, and L. Pileggi. "mLogic: Ultra-low voltage non-volatile logic circuits using STT-MTJ devices." In *Proceedings of the 49th Annual Design Automation Conference*, pp. 486-491. ACM, 2012.

[25] Fong, Xuanyao, and Kaushik Roy. "Robust low-power multi-terminal STT-MRAM." In *Non-Volatile Memory Technology Symposium (NVMTS), 2013 13th*, pp. 1-4. IEEE, 2013.

[26] Datta, Supriyo. "Nanoscale device modeling: the Green's function method." *Superlattices and microstructures* 28, no. 4 (2000): 253-278.

[27] Kishi, T., H. Yoda, T. Kai, T. Nagase, E. Kitagawa, M. Yoshikawa, K. Nishiyama et al. "Lower-current and fast switching of a perpendicular TMR for high speed and high density spin-transfer-torque MRAM." In *2008 IEEE International Electron Devices Meeting*, pp. 1-4. IEEE, 2008.

[28] K.-S. Ryu, L. Thomas, S.-H. Yang, and S. Parkin, "Chiral spin torque at magnetic domain walls," Nature nanotechnology, vol. 8, no. 7, pp. 527–533, 2013.

[29] LeCun, Yann, Corinna Cortes, and Christopher JC Burges. "The MNIST database of handwritten digits." (1998).

[30] Zhao, Wei, and Yu Cao. "New generation of predictive technology model for sub-45 nm early design exploration." *IEEE Transactions on Electron Devices* 53, no. 11 (2006): 2816-2823.

[31] Kim, Kuk-Hwan, Siddharth Gaba, Dana Wheeler, Jose M. Cruz-Albrecht, Tahir Hussain, Narayan Srinivasa, and Wei Lu. "A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications." *Nano letters* 12, no. 1 (2011): 389-395.

[32] S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki, and K. Ando, "Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions," Nature materials, vol. 3, no. 12, pp. 868–871, 2004.

[33] C. Lin, S. Kang, Y. Wang, K. Lee, X. Zhu, W. Chen, X. Li, W. Hsu, Y. Kao, M. Liu et al., "45nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell," in Electron Devices Meeting (IEDM), 2009 IEEE International. IEEE, 2009, pp. 1–4.

[34] A. Hirohata, H. Sukegawa, H. Yanagihara, I. Zutic, T. Seki, S. Mizukami, and R. Swaminathan, "Roadmap for emerging materials for spintronic device applications," Magnetics, IEEE Transactions on, vol. 51, no. 10,pp. 1–11, 2015.