Hybrid Spintronic-CMOS Spiking Neural Network with On-Chip Learning: Devices, Circuits, and Systems

Abhronil Sengupta, Aparajita Banerjee, and Kaushik Roy

Phys. Rev. Applied **6**, 064003 — Published 8 December 2016

DOI: 10.1103/PhysRevApplied.6.064003

Hybrid Spintronic-CMOS Spiking Neural Network With On-Chip Learning: Devices, Circuits and Systems

Abhronil Sengupta,* Aparajita Banerjee, and Kaushik Roy
School of Electrical & Computer Engineering, Purdue University, West Lafayette, IN 47907, USA

Over the past decade Spiking Neural Networks (SNN) have emerged as one of the popular architectures to emulate the brain. In SNN, information is temporally encoded and communication between neurons is accomplished by means of spikes. In such networks, spike-timing dependent plasticity mechanisms require the online programming of synapses based on the temporal information of spikes transmitted by spiking neurons. In this work, we propose a spintronic synapse with decoupled spike transmission and programming current paths. The spintronic synapse consists of a ferromagnet-heavy metal heterostructure where programming current through the heavy metal generates spin-orbit torque to modulate the device conductance. Low programming energy and fast programming times demonstrate the efficacy of the proposed device as a nanoelectronic synapse. We perform a simulation study based on an experimentally benchmarked device-simulation framework to demonstrate the interfacing of such spintronic synapses with CMOS neurons and learning circuits operating in transistor sub-threshold region to form a network of spiking neurons that can be utilized for pattern recognition problems.

I. INTRODUCTION

Brain-inspired computing models have emerged as one of the most powerful tools for pattern recognition and classification problems over the past few decades [1]. Such schemes attempt to develop abstract models of the communication and functionalities involved in the neurons and synapses in the human brain in order to construct computing tools efficient at recognition and cognitive tasks. However, implementation of such non-von Neumann computing schemes on general-purpose supercomputers have not been able to harness the energy efficiency of the human brain. The sequential fetch, decode and execute cycles involved in traditional von-Neumann computing are in complete contrast to the parallel, event driven processing involved in the mammalian cortex. For instance, the IBM *Blue Brain* project [2] utilized the Blue Gene supercomputer to simulate brain activity in animals and consumed orders of magnitude more energy than the brain, even at neuron firing rates much slower than the biological time scale.

Custom CMOS analog and digital VLSI neurocomputing platforms have been also utilized to implement neuron and synapse functionalities. The *BrainScaleS* [3], *SpiNNaker* [4] and the IBM *TrueNorth* [5] are instances of such neurocomputers based on conventional CMOS technology. However, the significant mismatch between the neuroscience mechanisms involved in the brain and the CMOS transistors have limited the capability of such computing technologies to achieve the area or power efficiency of the brain. For example, four 8-T SRAM cells (32 CMOS transistors) are required to implement the functionality of a single 4-bit synapse in a digital CMOS implementation [6].

Recently neurocomputing architectures based on emerging post-CMOS technologies have gained popularity as they offer a direct mapping to many of the neuroscience mechanisms involved in biological synapses [7–11] and neurons [12–14]. In order to achieve an integration density similar to the brain, neuromorphic computing architectures aim to achieve a fan-out of 10,000 for each neuron, thereby requiring orders of magnitude more synapses than neurons. Additionally, unsupervised learning using Spike-Timing Dependent Plasticity (STDP), or other Hebbian learning rules, require online programming of synapses during spike transmission. Hence, a nanoelectronic device emulating synaptic functionalities is an essential component of spiking neuromorphic architectures.

In this work, we propose a ferromagnet (FM)-heavy metal (HM) multilayer structure where spin-orbit torque induced by the programming current flowing through the HM is the main underlying physical mechanism for gen-

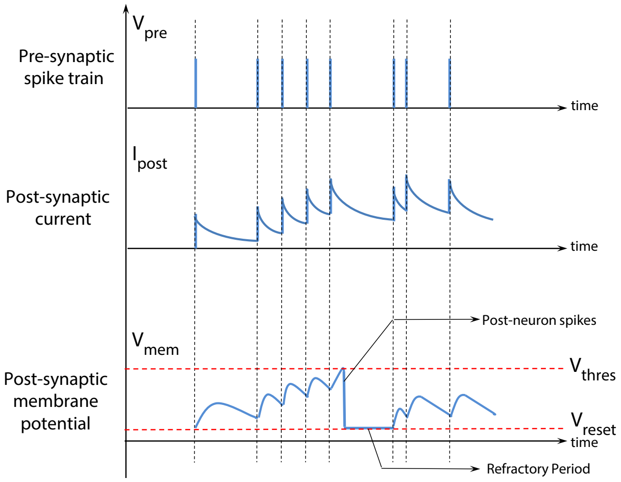

FIG. 1. Neuron and synapse dynamics in response to a spike train.

[*] asengup@purdue.edu

erating synaptic plasticity. The ferromagnet is part of a Magnetic Tunneling Junction (MTJ) structure where spike voltage transmitted through the MTJ gets modulated by the MTJ conductance. The proposed three-terminal device structure offers the advantage of decoupled spike transmission and programming current paths thereby leading to efficient implementation of on-chip learning. Further, the proposed synapse can be programmed at low current magnitudes and small programming time durations and thereby consumes orders of magnitude lower programming energy in comparison to other state-of-the-art emerging synaptic devices. We discuss a comprehensive framework for simulating such spintronic synapse based spiking neural systems from the device (including calibration to experimental results) to the system level for performing recognition tasks.

II. SPIKING NEURAL NETWORKS: PRELIMINARIES

II.1. Neuron and Synapse dynamics in Spiking Neural Networks

A synapse is a junction connecting two neurons. The transmitting neuron is termed as the pre-neuron while the receiving neuron is termed as the post-neuron. The pre-neuron transmits a train of voltage spikes which may be represented by a set of Dirac-delta functions at time instants t_f,

$$V_{pre} = \sum_f \delta(t - t_f) \tag{1}$$

The synapse response to such a spike train is modelled by,

$$\tau_{post}\frac{dI_{post}}{dt} = -I_{post} + w\sum_f \delta(t - t_f) \tag{2}$$

where, I_{post} is the post-synaptic current produced by the synapse characterized by weight w and τ_{post} is the time-constant of the post-synaptic current. Hence, the post-synaptic current increases by an amount modulated by the synapse conductance (weight) at each spike instant and then starts decaying exponentially. The temporal dynamics of the leaky-integrate-fire neuron in response to such a post-synaptic current is given by,

$$\tau\frac{dV_{mem}}{dt} = -V_{mem} + R_{mem}\sum_i I_{post,i} \tag{3}$$

where, V_{mem} is the membrane potential, R_{mem} is the membrane resistance, $I_{post,i}$ is the post-synaptic current input from the i-th neuron, and τ is the membrane time-constant. Fig. 1 shows the temporal characteristics of the neuron and synapse in response to a series of voltage spikes transmitted from the pre-neuron. When the neuron's membrane potential V_{mem} crosses the threshold V_{thres}, the membrane potential gets reset to V_{reset} and does not vary for a time duration termed as the refractory period.

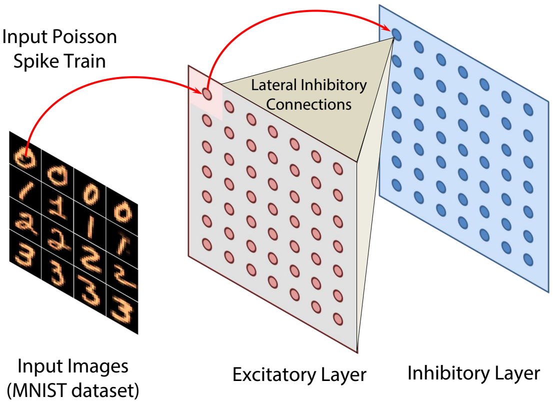

FIG. 2. Network connectivity utilized for pattern recognition. Neurons with lateral inhibitory connections receive input Poisson spike trains with average rate proportional to pixel intensity.

II.2. Learning: STDP

According to the theory of Hebbian Learning [15], synaptic weight or conductance is modulated depending on the spiking patterns of the pre-neuron and post-neuron. STDP, a form of Hebbian learning, states that the weight of the synapse increases (decreases) if the pre-neuron spikes before (after) the post-neuron. Intuitively, this signifies that the synapse strength should increase if the pre-neuron spikes before the post-neuron as the pre-neuron and post-neuron appear to be temporally correlated. The relative change in synaptic strength decreases exponentially with the timing difference between the pre-neuron and post-neuron spikes. The STDP characteristics have been formulated in a mathematical framework based on measurements for rat hippocampal glutamatergic synapses [16],

$$\begin{aligned}\Delta w &= A_+ \exp\left(\frac{-\Delta t}{\tau_+}\right), \Delta t > 0\\ &= -A_- \exp\left(\frac{\Delta t}{\tau_-}\right), \Delta t < 0\end{aligned} \tag{4}$$

Here, A_+, A_-, τ_+ and τ_- are constants and $\Delta t = t_{post} - t_{pre}$, where t_{pre} and t_{post} are the time-instants of pre- and post-synaptic firings respectively. We will refer to the case of $\Delta t > 0$ ($\Delta t < 0$) as the positive (negative) time window for learning.

II.3. Spike Frequency Adaptation

In order to model spike frequency adaptation mechanisms observed in biological neurons, an additional slowly varying adaptation parameter a is introduced in the temporal dynamics of the neuron as,

$$\tau \frac{V_{mem}}{dt} = -V_{mem}(1+a) + R_{mem} \sum_i I_{post,i} \qquad (5)$$

The adaptation parameter a increases every time the neuron spikes, otherwise it decays exponentially. This implies that in case a neuron starts spiking at a high frequency, the leak parameter starts to increase to reduce its spike frequency.

II.4. Network Connectivity

Fig. 2 shows the network connectivity of spiking neurons utilized for pattern recognition problems. Such a network topology has been shown to be efficient in several pattern recognition problems like digit recognition [17] and sparse encoding [18]. Input image pixels are encoded as Poisson spike trains with average rate directly proportional to the pixel intensity. These input spike trains are received by all neurons in an excitatory layer through synapses whose weights are learnt using STDP. Each neuron in the excitatory layer is connected to a corresponding neuron in an inhibitory layer such that a spike in the excitatory neuron triggers a spike in the corresponding neuron in the inhibitory layer. Each neuron in the inhibitory layer is connected to all neurons in the excitatory layer except the neuron from which it received the input. This connectivity helps to implement lateral inhibitory connections in the excitatory layer such that when one neuron starts to spike in response to some input pattern, it prohibits the other neurons from spiking. However, in order to prevent a particular neuron from dominating the spiking pattern due to lateral inhibitory connections, spike frequency adaptation mechanism is also implemented in each neuron. The neurons in the excitatory layer are assigned classes based on their highest response (spike frequency) to input training patterns.

III. SPINTRONIC SYNAPSE

III.1. Spin-orbit torque driven motion of Dzyaloshinskii domain walls

In this section we provide a brief discussion on the underlying physical phenomena involved in current induced domain wall motion in heavy metal (HM) - ferromagnet (FM) - insulator (I) multilayer structures.

Recent experiments on magnetic nanostrips of Pt/CoFe/MgO and Ta/CoFe/MgO have revealed high domain wall velocities due to charge current densities that are two orders of magnitude lower than that achievable by conventional spin-transfer torque (STT) [19]. Additionally, domain wall motion was observed to be against the direction of electron flow (i.e. in the direction of current flow) in multilayer structures with Pt as the underlayer, thereby suggesting that current induced spin-orbit torque is the main mechanism of domain wall motion in such multilayer structures (with negligible contribution from conventional STT) [19]. In such magnetic heterostructures with high perpendicular magnetocrsytalline anisotropy (PMA), spin orbit coupling and broken inversion symmetry leads to the stabilization of homochiral domain walls through the Dzyaloshinskii-Moriya exchange interaction (DMI) [20]. We restrict our analysis for Pt/CoFe/MgO multilayer structures for this text due to the possibilities of achieving high domain wall velocities ($\sim 400m/s$) [21–23]. However, the analysis can be easily extended to other magnetic heterostructures with different underlayers.

Such interfacial DMI at the FM-HM interface leads to the formation of a Néel domain wall with left-handed chirality for Pt/CoFe/MgO multilayer structures [19, 21–23]. The DMI strength in such structures with HM underlayers has been observed to be sufficiently strong to impose a Néel wall configuration in FMs where conventional magnetostatics would have yielded a Bloch configuration [19]. When an in-plane charge current is injected through the HM, a transverse spin-current is generated due to deflection of opposite spin-polarizations on the top and bottom surfaces of the HM. This phenomena is termed as spin-Hall effect [24] and arises as a consequence of spin-orbit torque. The accumulated spins at the FM-HM interface leads to DMI stabilized Néel domain wall motion. The direction of domain wall motion is in the direction of charge current flow and the final magnetization of the ferromagnet is given by the cross-product of the direction of injected spins at the FM-HM interface and the magnetization direction of the FM at the domain wall location.

III.2. Device proposal for spintronic synapse

Such spin-orbit torque driven domain wall motion in FMs due to charge current flow through a HM underlayer leads to the possibility of a device structure that can manifest decoupled spike transmission (read) and programming (write) current paths. We propose a three-terminal device structure consisting of a FM lying on top of a HM (Fig. 3). The FM is part of an MTJ structure where the FM is separated from a Pinned layer (magnetic region whose magnetization is fixed) by a Tunneling Oxide barrier (MgO). The FM has two additional Pinned layers on either side to ensure that the domain wall stabilizes at the extreme locations of the FM for sufficiently large values of the programming current. While the spike current flows through the MTJ structure between terminals T1 and T3, the programming current flows through the HM layer between terminals T2 and T3. Note that a preliminary synaptic device proposal based on Bloch domain wall motion due to spin-orbit torque was explored

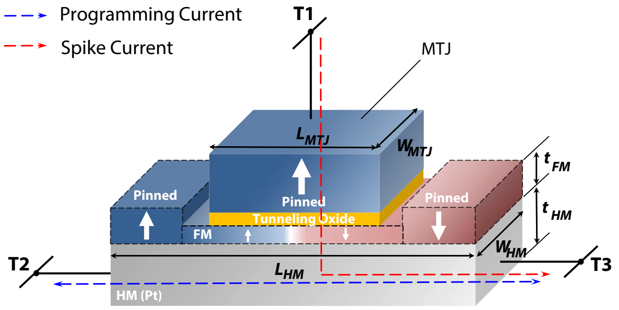

FIG. 3. Device structure for a spintronic synapse with decoupled spike transmission and programming current paths. Spike current flows through the MTJ structure between terminals T1 and T3. Programming current flows through the HM between terminals T2 and T3.

previously in Ref. [25]. However, an external magnetic field was required to modulate the device conductance during learning. Further the magnet width was not scalable beyond 100nm to ensure Bloch wall orientation. The current device proposal based on Néel wall motion is not only more energy efficient, but also requires no external magnetic field for domain wall motion due to the inherent interfacial DMI. Further this work provides a synergistic device-circuit-system perspective for the implementation of STDP in SNNs utilizing the proposed spintronic device as the core building block.

The location of the domain wall in the FM encodes the resistance of the device lying in the path of the spike current between terminals T1 and T3 and thereby implements the synaptic functionality. On the other hand, the programming current path is completely decoupled (between terminals T2 and T3) and the resistance in the path of the programming current is mainly determined by the HM resistance. It is worth noting here that although some amount of spike current will flow through the HM, the magnitude of this current can be maintained to sufficiently low values below the domain wall depinning current since the synapses are required to drive CMOS neurons operating in the subthreshold regime.

III.3. Synaptic plasticity mechanism

Programming current flowing from terminal T2 to terminal T3 results in domain wall motion in the same direction so that the +z domain in the FM starts to expand and vice versa. For a given duration of the programming current pulse, the domain wall displacement is directly proportional to the magnitude of the programming current.

On the other hand, the device conductance between terminals T1 and T3 varies linearly with the domain wall position. Let us denote the conductance of the device when the entire FM magnetization is parallel (antiparallel) to the Pinned layer as $G_P(G_{AP})$, i.e. the domain wall is at the extreme right (left) of the FM. Thus, for an intermediate position of the domain wall at a position x from the left-edge of the MTJ, the device conductance between terminals T1 and T3 is given by,

$$G_{eq} = G_P.\frac{x}{L} + G_{AP}.\left(1 - \frac{x}{L}\right) + G_{DW} \qquad (6)$$

where, L denotes the length of the MTJ excluding the domain wall width and G_{DW} represents the conductance of the wall region. It is worth noting here, that L, G_{DW}, G_P and G_{AP} are all constants (for constant voltage drop across the MTJ). Due to such a linear relationship between domain wall position and device conductance, the programming current is directly proportional to the change in device conductance (which encodes the synaptic weight) for a fixed duration of the programming signal.

III.4. Spiking neuromorphic architecture based on spintronic synapse

Fig. 4 represents possible arrangement of a spintronic synapse with access transistors $M_{A1} - M_{A4}$ to decouple the programming and spike current paths. The access transistors act as switches to select the appropriate terminals of operation for the device. The operating mode of the synapse, i.e. the spike transmission mode or programming mode is accomplished by the control signal POST. The POST signal is activated during the programming mode of operation of the synapse.

The PRE line is used to pass the necessary amount of programming current required for the corresponding weight change involved due to the delay between the pre-neuron and post-neuron spikes. A negative (positive) current should flow through the HM for the negative (positive) time window duration. Since the programming current amplitude is directly proportional to the amount of weight change, the current signal flowing through the HM should vary in a similar fashion as the STDP learning curve (exponentially) with the time delay between the pre-neuron and post-neuron spikes.

For simplicity, let us discuss the case for the positive time window. The exponential variation of current

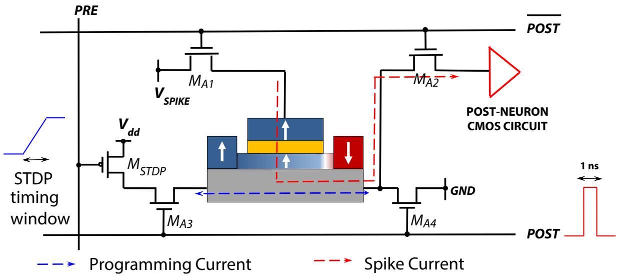

FIG. 4. Spintronic synapse with access transistors to decouple the programming and spike current paths.

through the HM can be obtained by a transistor operating in sub-threshold regime since the current flowing through the transistor will vary exponentially with the gate to source voltage. Thus for a linear increase of voltage of the PRE line with time, the transistor M_{STDP} will be driven from cut-off to saturation regime when the POST signal is activated and an appropriate programming current should flow through the HM. It is worth noting here that the HM resistance $\sim$ a few hundred ohms and the maximum programming current required is $\sim$ a few tens of μA, thereby leading to a very small voltage drop across the device when the POST signal is activated. Fig. 4 shows the interface circuits involved in the synapse programming for the positive time window. A similar approach can be adopted to program the synapses for the negative time window (by utilizing an NMOS operating in sub-threshold saturation driven by a linearly increasing gate voltage to pass programming current from terminals T3 to T2) and the two learning circuits for the negative and positive timing windows have to be activated sequentially everytime the pre-neuron spikes. Since the time duration involved in programming is $\sim$ a few ns in comparison to learning time constants used in this work $\sim$ μs, the POST signal essentially samples the necessary amount of programming current from the PRE line (programming current magnitude determined by M_{STDP} transistor).

In our proposed programming scheme, we program the synapses only when the post-neuron spikes. Hence, in order to account for the negative and positive time windows involved in STDP learning, the POST signal should be activated with a delay corresponding to the time duration of the negative timing window in order to sample the programming current contributions from the learning circuits for both the timing windows.

Arrangement of synapses in an array fashion as shown in Fig. 5, interfaced with CMOS neurons can lead to dense spiking neuromorphic architectures. Please note that the access transistors M_{A2} and M_{A4} for terminal T3 of the device (Fig. 4) can be shared across the row such that the corresponding horizontal line connecting terminals T3 for the devices in a particular row are driven to GND (POST signal is HIGH) or the post-neuron circuit (POST signal is LOW). Details of the CMOS circuits involved in the programming scheme and neuron implementation will be discussed in the next section.

IV. CMOS LEARNING AND NEURON CIRCUITS

IV.1. Sub-threshold circuit for STDP learning

The circuit involved in generating the PRE signal is discussed in this section. Fig. 6 shows the sub-threshold CMOS circuit used to generate the PRE signal for pre-neuron A connecting to post-neurons C and D. We discuss the mechanism for generating the signal for the positive time window. A similar design can be used to generate the programming current for the negative time window. The circuit was originally proposed in [26] as a reset and discharge synapse. However it failed to emulate the post-synaptic dynamics of biological synapses as the circuit response depends only on the previous input spike [27]. In this work, we employ this circuit to implement STDP learning in our proposed device.

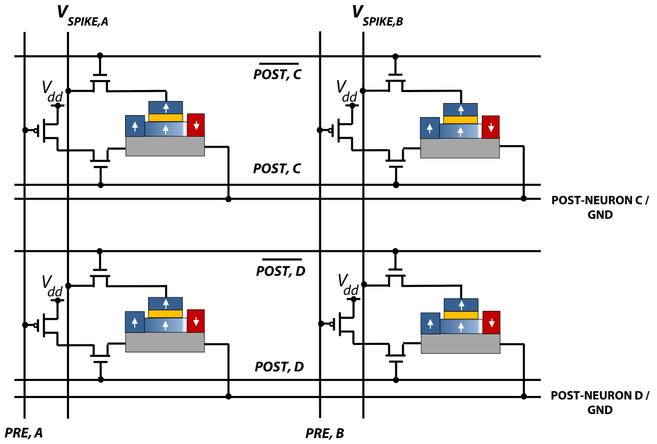

FIG. 5. Possible arrangement of synapses in an array interfaced with CMOS neurons and programming circuits. The figure shows synapses connecting pre-neurons A and B to post-neurons C and D.

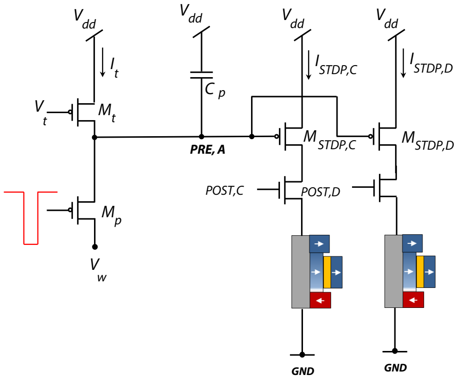

FIG. 6. Sub-threshold CMOS circuit utilized for generating the programming current involved in STDP learning (circuit for positive time window shown) for pre-neuron A connecting to post-neurons C and D.

The transistor M_p acts as a switch. When the positive time window starts, the transistor M_p receives a low-active pulse and gets turned ON. As a result, the node PRE, A is set to the bias voltage V_w. After the transistor M_p is switched OFF, the transistor M_t, operating in sub-threshold saturation regime, provides a constant current to linearly charge the capacitor C_p at a rate $\frac{I_t}{C_p}$. Hence, if the transistor M_{STDP} is operated in sub-threshold saturation, exponential dynamics will be observed in the output current I_{STDP}. The current flow-

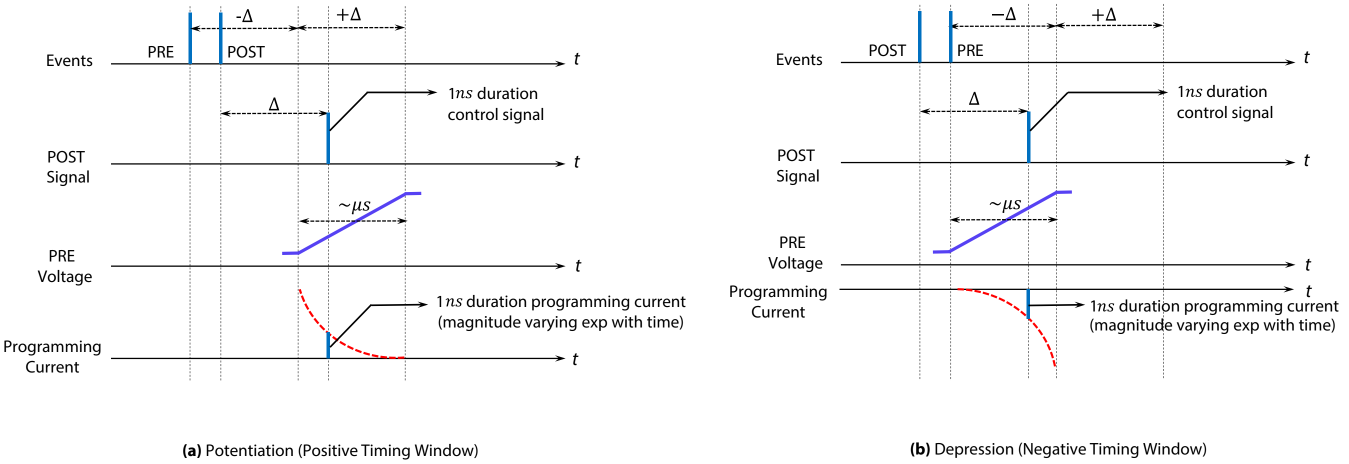
-Δ
+Δ
Events
PRE
POST
t
Δ
1ns duration control signal
POST Signal
~μs
PRE Voltage
1ns duration programming current (magnitude varying exp with time)
Programming Current
(a) Potentiation (Positive Timing Window)
−Δ
+Δ
Events
POST
PRE
Δ
1ns duration control signal
POST Signal
~μs
PRE Voltage
Programming Current
1ns duration programming current (magnitude varying exp with time)
(b) Depression (Negative Timing Window)

FIG. 7. Detailed timing diagrams demonstrating the implementation of (a) potentiation (positive timing window) and (b) depression (negative timing window) in the spintronic synapse. POST is the control signal that is activated during programming while PRE is the gate voltage of the M_{STDP} transistor that implements synaptic plasticity. Duration of the programming current is determined by the duration of the POST signal while the magnitude is determined by the value of the PRE signal when the POST signal is high.

ing through transistor M_{STDP} for an input pulse at time $t = t_n$ is given by,

$$I_{STDP} = I_0 e^{\frac{-U_T C_p (t-t_n)}{kI_t}} \qquad (7)$$

where, k is the sub-threshold slope factor and U_T is the thermal voltage. Hence, whenever the pre-neuron spikes, the circuits for generating the STDP characteristics for the negative and positive time windows are activated sequentially. When learning starts for the positive timing window, a short pulse is applied to the gate of the transistor M_p so that the circuit is reset and the node PRE, A is charged to V_w. When the post-neuron does not spike, the transistor M_{STDP} is in cut-off since the POST signal is deactivated and the access transistors for programming are turned OFF. Once the post-neuron spikes, the programming current path gets activated and the transistor M_{STDP} switches to the sub-threshold saturation regime and transmits the necessary amount of programming current through the device. Note that apart from the transistor M_{STDP} (one transistor for each of the positive and negative timing windows), the entire learning circuitry can be shared across the column of the crossbar array.

The operation is discussed in details in Fig. 7. Let us first describe the case for the positive timing window, i.e. post-neuron spiking after the pre-neuron (Fig. 7(a)). $(-\Delta)/(+\Delta)$ represents the duration during which the learning circuits for the negative/positive timing windows are activated sequentially for the corresponding preneuronal firing event. The control signal POST is activated after a duration (Δ) the post-neuron spikes. As described in the figure, magnitude of the programming pulse is determined by the current being passed by the programming transistor M_{STDP} (value of the PRE voltage when the POST signal is active) and the duration is determined by the duration of the POST signal. Since the PRE signal varies in $\sim \mu s$ time scale and does not almost change during the programming time duration ($\sim ns$ time scale), it ensures that the programming current magnitude is almost constant and is equal to the sampled value from the exponential STDP dynamics corresponding to the appropriate spike timing difference. As mentioned previously, since the programming current magnitude is directly proportional to the amount of change in the MTJ conductance, exponential STDP characteristics is implemented in the spintronic device. Similar discussions are valid for the negative timing window (Fig. 7(b)) where the post-neuron spikes before the preneuron. In this case, the POST signal is activated during the negative window $(-\Delta)$ and the NMOS transistor passes an appropriate amount of programming current in the opposite direction through the device. Circuit-level simulations confirming the proposal have been demonstrated in Fig. 11(b).

IV.2. Differential Pair Integrator circuit for post-synaptic current generation

The Differential Pair Integrator (DPI) circuit has been a popular mechanism for generating synaptic dynamics [28] and integration of such DPI circuits with memristor synapses has been recently proposed [29]. Fig. 8(a) shows how such DPI circuits can be integrated with our proposed spintronic synapses to generate exponential post-synaptic currents in response to input spikes. Assuming all transistors are in sub-threshold saturation and using the translinear principle [28, 29] it can be shown that the output current I_{syn} exhibits temporal dynamics

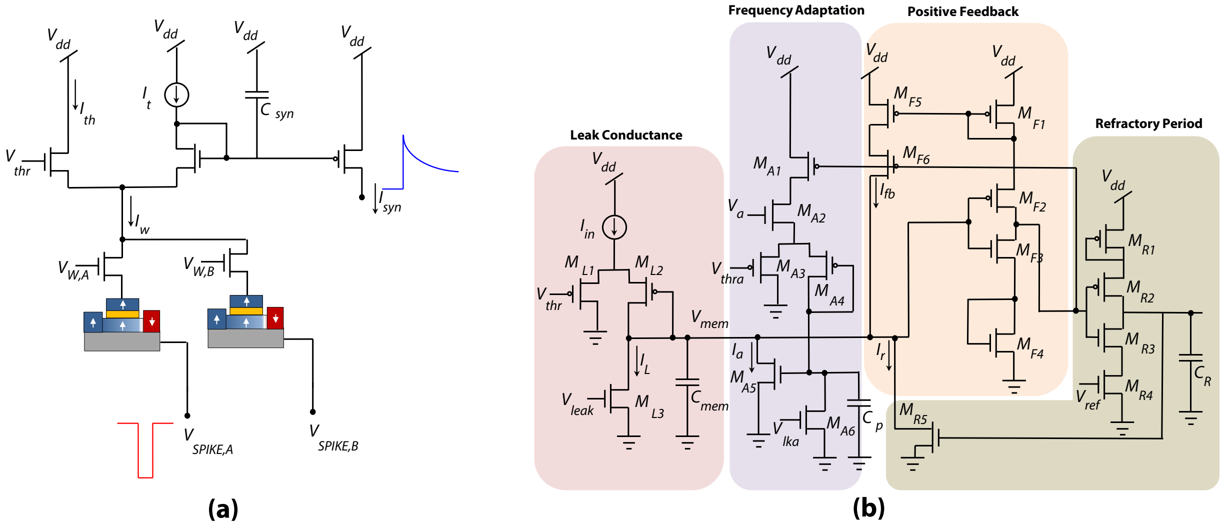

FIG. 8. (a) DPI circuit interfaced with spintronic synapses to emulate synaptic dynamics. (b) Sub-threshold CMOS neuron with leak conductance, spike frequency adaptation, positive feedback and refractory period implementation blocks [28].

of the form,

$$\tau_{syn}\frac{dI_{syn}}{dt}+I_{syn}=\frac{I_w I_{th}}{I_t} \tag{8}$$

where, $\tau=\frac{CU_T}{kI_t}$. The above relationship is valid if the circuit is operated in the linear region ($I_t \ll I_w$). The bias voltage V_w acts as a scaling gain factor for the post-synaptic current. On the arrival of an input spike, the current I_w gets modulated by the MTJ conductance and thereby causes I_{syn} to increase by an amount governed by the synaptic weight. When there is no spike transmission, I_{syn} decreases exponentially thereby emulating the synaptic dynamics discussed earlier. The access transistors driven by $\overline{\text{POST}}$ signal have not been shown in Fig. 8 but are present in the design to ensure that the programming current path is deactivated when spike transmission path is enabled.

IV.3. Sub-threshold CMOS neuron

CMOS circuits operating in sub-threshold (Fig. 8(b)) have been shown to replicate a wide range of temporal dynamics observed in biological neurons like spike frequency adaptation and refractory period generation [28, 30, 31]. When operated in the sub-threshold regime, the main mechanism of carrier transport in CMOS transistors is diffusion, thereby emulating the mechanism of ion flow in biological neuron channels [28].

I_{in} represents the input current provided to the neuron. Using the translinear principle and assuming all transistors in sub-threshold saturation, it can be shown that the temporal dynamics of I_{mem} is given by [28],

$$\tau_{mem}\frac{dI_{mem}}{dt}+I_{mem}\left(1+\frac{I_a}{I_t}\right)=\frac{I_{in}I_{th}}{I_t} \tag{9}$$

where, $\tau=\frac{C_{mem}U_T}{kI_t}$. The above relation is again valid when the DPI circuit operates in the linear region (i.e. $I_t \ll I_{in}$).

We would like to conclude this section by relating the computing models discussed in Section II to circuit implementations discussed in Section IV. Postsynaptic and neuron dynamics (referred in Eqs. 2 and 5) can be directly mapped to the DPI circuit and subthreshold CMOS neuron circuit (referred in Eqs. 8 and 9) respectively. Readers are referred to Ref. [28] for details on neuromorphic chips utilizing such analog CMOS neurons and interfacing such circuits with post-CMOS synaptic crossbar arrays. Our proposal in this work includes the implementation of plasticity mechanism (referred in Eq. 4) in the spintronic device structure utilizing the device concepts (presented in Section III) and learning circuit primitives (presented in Section IV.1).

V. SIMULATION RESULTS

V.1. Simulation Framework

In order to simulate the SNN implementation based on the proposed spintronic synapse, a hierarchical simulation framework was utilized. Device-level simulations of the spin-orbit torque induced domain wall motion was performed in MuMax [32], a GPU accelerated micromagnetic simulation tool. A behavioral model of the device was developed for subsequent simulation of such synapses interfaced with CMOS neurons and learning circuits. The circuit level simulations were performed in HSPICE using a standard cell library in commercial 45nm CMOS technology. The device and circuit simulations were utilized to generate models of the plastic synapses and spiking neurons to perform system level simulations of a network of spiking neurons using Brian simulator [33].

V.2. Device Level Simulations

The magnetization dynamics of the ferromagnet can be described by solving Landau-Lifshitz-Gilbert equation with additional term to account for the spin-orbit torque generated by spin-Hall effect at the FM-HM interface [21, 34],

$$\frac{d\widehat{\mathbf{m}}}{dt} = -\gamma(\widehat{\mathbf{m}} \times \mathbf{H}_{eff}) + \alpha(\widehat{\mathbf{m}} \times \frac{d\widehat{\mathbf{m}}}{dt}) + \beta(\widehat{\mathbf{m}} \times \widehat{\mathbf{m}}_P \times \widehat{\mathbf{m}}) \tag{10}$$

where, $\widehat{\mathbf{m}}$ is the unit vector of FM magnetization at each grid point, $\gamma = \frac{2\mu_B\mu_0}{\hbar}$ is the gyromagnetic ratio for electron, α is Gilbert's damping ratio, $\mathbf{H}_{eff}$ is the effective magnetic field, $\beta = \frac{\hbar\theta J}{2\mu_0 e t M_s}$ ($\hbar$ is Plancks constant, J is input charge current density, θ is spin-Hall angle [21], μ_0 is permeability of vacuum, e is electronic charge, t is FL thickness and M_s is saturation magnetization) and $\widehat{\mathbf{m}}_P$ is direction of input spin current. The effective field $\mathbf{H}_{eff}$ also includes the field due to DMI and is given by,

$$\mathbf{H}_{DMI} = -\frac{2D}{\mu_0 M_s}\left[\frac{\partial m_z}{\partial x}\widehat{x} + \frac{\partial m_z}{\partial y}\widehat{y} - \left(\frac{\partial m_x}{\partial x} + \frac{\partial m_y}{\partial y}\right)\widehat{z}\right] \tag{11}$$

Here, D represents the effective DMI constant and determines the strength of DMI field in such multilayer structures. A positive sign of D implies right-handed chirality and vice versa. In the presence of DMI, the boundary conditions at the edges of the sample is given by,

$$\frac{\partial\widehat{\mathbf{m}}}{\partial n} = \frac{D}{2A}\widehat{\mathbf{m}} \times (\widehat{\mathbf{n}} \times \widehat{z}) \tag{12}$$

where, A is the exchange correlation constant and $\widehat{\mathbf{n}}$ represents the unit vector normal to the surface of the FM. The simulation parameters are given in Table I and was used for the rest of this work, unless otherwise stated. The parameters were obtained experimentally from magnetometric measurements of Ta(3nm)/Pt(3nm)/CoFe(0.6nm)/MgO(1.8nm)/Ta(2nm) nanostrips [22]. Current density was estimated by assuming that the current flow is mainly through the FM-HM layers in the stack structure [22].

Fig. 9(a) shows the domain wall displacement in a CoFe sample with cross-section of $160nm \times 0.6nm$ for a charge current density of $J = 0.1 \times 10^{12} A/m^2$. The grid size was taken to be $4 \times 4 \times 0.6nm^3$. Fig. 9(b) depicts the variation of the domain wall velocity with input charge current density. The velocity increases linearly with the current density and ultimately reaches a saturation velocity. The graphs are in good agreement with results illustrated in [21] for the same multilayer structure described in this section. Fig. 9(c) illustrates the fact that the domain wall displacement is directly proportional to the magnitude of the programming current (for domain wall velocities below the saturation regime). For a duration of $1ns$, a maximum current of $\sim 80\mu A$ is required to displace the domain wall from one edge of the FM to the other edge.

TABLE I. Device Simulation Parameters

Parameters	**Value**
Ferromagnet Dimensions	$320 \times 20 \times 0.6nm^3$
Grid Size	$4 \times 1 \times 0.6nm^3$
Heavy Metal Thickness	$3nm$
Domain Wall Width	$7.6nm$
Saturation Magnetization, M_s	$700\ KA/m$
Spin-Hall Angle, θ	0.07
Gilbert Damping Factor, α	0.3
Exchange Correlation Constant, A	$1 \times 10^{-11} J/m$
Perpendicular Magnetic Anisotropy	$4.8 \times 10^5 J/m^3$
Effective DMI constant, D	$-1.2 \times 10^{-3} J/m^2$

Non-Equilibrium Green's Function (NEGF) based transport simulation framework [37] was used to model the variation of the MTJ resistance with oxide thickness (Fig. 10(a)) and applied voltage (Fig. 10(b)) respectively. In order to determine the MTJ resistance for a FM with a domain wall separating two oppositely polarized magnetized domains, the NEGF based simulator [37] was modified by considering the parallel connection of three MTJs. The magnetization direction of the FL of the three MTJs were considered parallel, anti-parallel and perpendicular (domain wall) to the pinned layer magnetization. The length of the first two MTJs was varied according to the position of the domain wall while the width of the third MTJ was taken to be equal to the domain wall width. Fig. 11(a) depicts the variation of the device conductance with domain wall position (origin at the middle of the FM). In order to ensure proper synaptic functionality, it is also essential that the device resistance (for a particular position of the domain wall) does not vary with the voltage drop across the device. This is ensured by appropriately interfacing the device with the DPI circuit discussed earlier to generate the synaptic dynamics. The range of synapse resistances are in the $M\Omega$ range while the current flowing through the MTJ is in the range of a few nAs. Hence the voltage drop across the MTJ should be $\sim$ a few mV ($< 100mV$). It is apparent from Fig. 10(b) that the operating range of V_{MTJ} is low enough to ensure negligible variation of the device conductance with device voltage drop for a particular domain wall position. As explained in the earlier section, such a linear variation of the device conductance with domain wall position results in the programming current being directly proportional to the relative conductance (weight) change involved. Hence the temporal profile of the necessary programming current also follows the STDP characteristics.

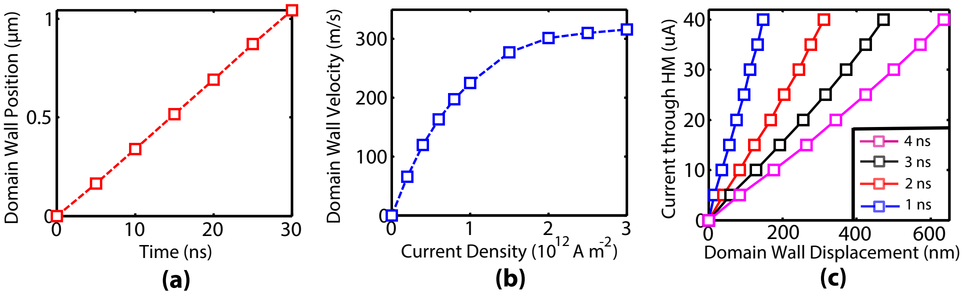

FIG. 9. (a) Domain wall displacement as a function of time for a CoFe strip of cross-section $160nm \times 0.6nm$ due to the application of a charge current density, $J = 0.1 \times 10^{12} Am^{-2}$. (b) Domain wall velocity as a function of current density. The results are in good agreement with [21]. (c) Domain wall displacement is directly proportional to the programming current for a fixed duration of the programming pulse.

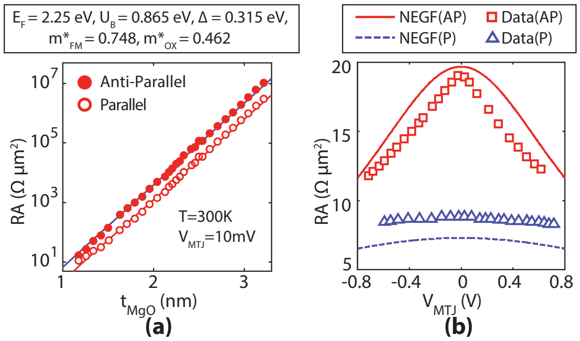

FIG. 10. The NEGF based transport simulation framework was calibrated to experimental results illustrated in [35, 36]. MTJ resistance varies with (a) oxide thickness and (b) applied voltage.

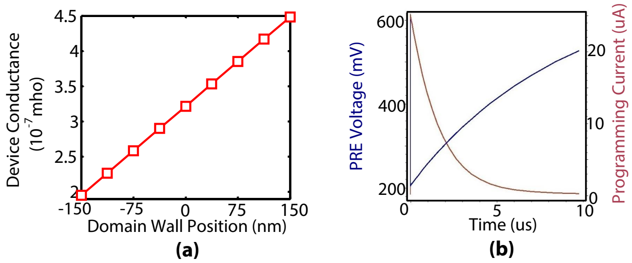

FIG. 11. (a) Linear variation of device conductance with domain wall position. (b) Programming circuit simulation to generate the STDP characteristics in the proposed spintronic synapse.

V.3. Circuit Level Simulations

The programming and neuron circuits were simulated using a standard cell library in 45nm commercial CMOS technology. Although biological time scales are in the range of ∼ ms, it is not essential to limit the processing speed of the circuit to such slow time constants for implementing pattern recognition systems [6]. The circuits were designed to operate at time constants in the range of $\sim \mu s$.

Fig. 11 (b) shows the response of the programming circuit for the case when the programming current path is active throughout the simulation time. The gate voltage of the transistor M_{STDP} increases linearly and is reset at each input pulse leading to exponential sub-threshold current dynamics. The average power consumption of the circuit is $0.46\mu W$ for the entire positive time window. The duration of the time window can be varied by changing the capacitance value. Further, this programming circuit can be shared by synapses in a particular column. It is worth noting here, that this power consumption does not include the power consumed in the M_{STDP} transistor as current will flow through it only when the programming current path is activated for $1ns$. The supply voltage for M_{STDP} transistor was maintained at $600mV$ and hence the maximum amount of energy consumption involved in synapse programming is $\sim 48fJ(600mV \times 80\mu A \times 1ns)$ per synaptic event.

Fig. 12 depicts the response of the CMOS neuron to a constant input current. As explained earlier, spike frequency adaptation scheme reduces the spike frequency to a steady state value. For a membrane capacitance of $50fF$, the average power consumption of the circuit was $\sim 5.7pJ/$ spike.

V.4. System Level Simulations

The device and circuit behavioral models were used to simulate an SNN for digit recognition problems. The input images (28×28 pixels) used for training was taken from the MNIST dataset [38]. The images were rate encoded and an array of 100 excitatory neurons was used to simulate the self-learning functionality of synapses in SNNs. Fig. 13 (a) demonstrates the SNN topology used for the recognition problem arranged in a crossbar array fashion. Synapses present at the crosspoints joining the inputs to the excitatory neurons can be programmed depending on the temporal spiking patterns of the pre- and post-neuron. Note that a synapse is absent at the crosspoint joining the excitatory to the inhibitory neu-

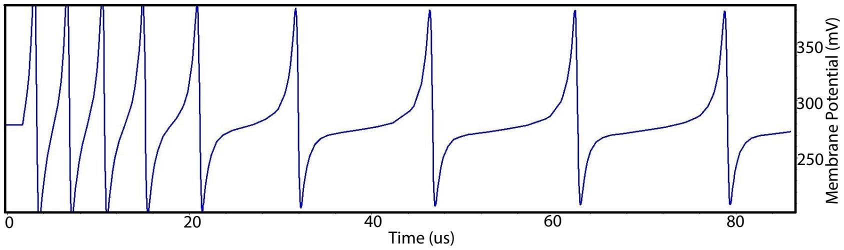

FIG. 12. CMOS neuron response to a constant input current with positive feedback, spike frequency adaptation and refractory period implementations.

ron. Inhibitory neurons are exactly similar to the excitatory neurons except that the output voltage spikes are negative.

Fig. 13 (b)-(c) depicts synapse weights plotted in 28×28 array (same as input images) for each of the 100 neurons used for the recognition purpose. Initially all the weights are random. However, as learning progresses the synapses of each neuron start learning generic representations of the various digits. Thus a particular neuron becomes more sensitive to the digit whose generic representation is being stored in its synapse weights since it will fire more if input spike trains are received at the pixel locations corresponding to high synaptic weights. The various system level simulation parameters have been outlined in Table II. The parameters were tuned to achieve learning ability in the synapses. The units of the time constants are with respect to the duration of each timestep in the simulation. For this work, the circuits were designed to operate in $\sim \mu s$ time scale as mentioned before. It is worth noting here that the manner in which the time constants and other parameters can be tuned in the circuit level simulations have been discussed in the previous section. The numbers in braces represent the value corresponding to the inhibitory neuron.

TABLE II. System Simulation Parameters

Parameters	**Value**
No. of excitatory/inhibitory neurons	100
Probability of input spike per timestep	$0 - 0.06375$
Number of timesteps per image	350
STDP time constants	100(1)
Neuron time constants	10(10)
Post-synaptic current time constants	1 (2)

Additionally, we would like to mention here, that such neuromorphic systems are significantly robust to imprecision due to device mismatch, variability and noise effects due to the adaptive nature of such computations involving plasticity, homeostasis and feedback mechanisms [28]. Further, authors in Ref. [39] demonstrate the immunity of such single layer SNNs based on crossbar arrays of resistive synapses with lateral inhibition and homeostasis effects to variations and non-idealities in typical resistive synaptic devices and CMOS neuron circuits. In particular, we performed an analysis of the impact of variations in the oxide thickness/MTJ synaptic conductances on the classification accuracy of the system. Almost no degradation in classification accuracy was observed for the 100-neuron network even with 25% variation in the resistances of the spintronic synapses.

VI. CONCLUSIONS

While prior proposals have investigated mono-domain spintronic devices for implementing spiking neurons [41] and short-term plasticity effects [42], to the best of our knowledge this is the first work to propose a hybrid spintronic-CMOS SNN design with self-learning (from the device to the system level) based on a three-terminal multi-domain spintronic synapse device structure consisting of decoupled spike transmission and programming current paths. This is advantageous for implementation of neuromorphic systems capable of on-chip learning since the programming current path is independent of the read current path. Interface CMOS circuit design for self-learning is highly simplified since the resistance in the programming current path is constant and determined mainly by the HM resistance and independent of the synapse conductance.

Table III provides a comparative analysis of our spintronic synapse (calibrated to experiments performed in Ref. [22]) with other proposed synaptic devices. Synaptic device structures based on emerging post-CMOS technologies [7, 8, 11] are usually two-terminal devices and do not offer de-coupled programming and read current paths. Additionally they are usually characterized by relatively high programming energies. In contrast, our proposed synapse offers low programming energy and requires very small programming time. A maximum programming energy of $\sim 48fJ$ is consumed per synaptic event due to the highly energy-efficient spin-orbit torque induced synaptic plasticity. Three terminal synaptic devices based on FeFET [10] and floating gate transistors [9] have been also proposed. However, the programming in such devices is usually accomplished through the gate terminal and a high gate voltage is usually applied across a very thin oxide [9, 10] leading to reliability issues, in

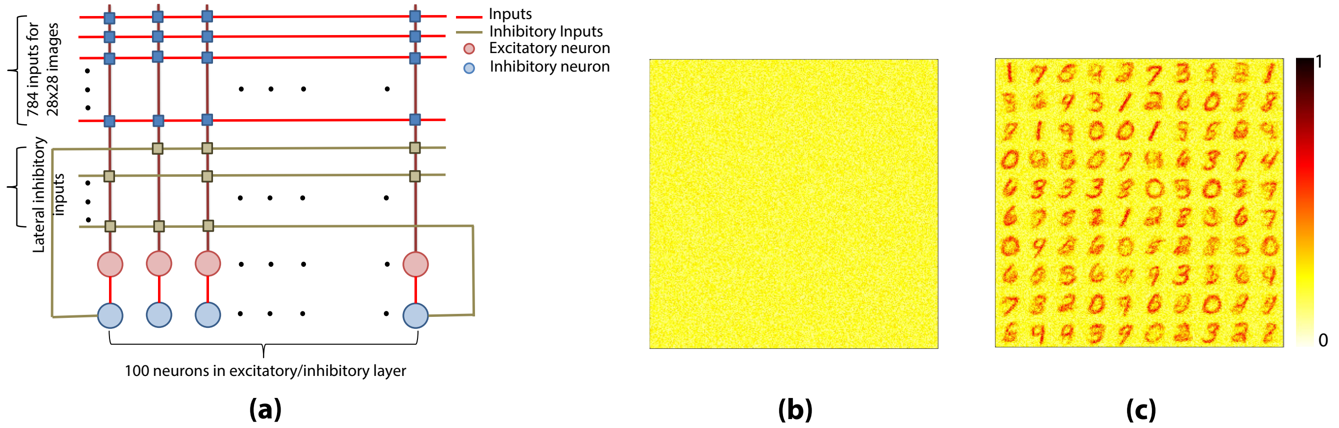
784 inputs for 28x28 images
Lateral inhibitory inputs
Inputs
Inhibitory Inputs
Excitatory neuron
Inhibitory neuron
100 neurons in excitatory/inhibitory layer
1
0
(a)
(b)
(c)

FIG. 13. (a) SNN topology used for digit recognition arranged in a crossbar array fashion. (b) Initial random synapse weights plotted in a 28×28 array for 100 neurons in the excitatory layer. (c) Representative digit patterns start getting stored in the synapse weights for each neuron after 1000 learning epochs.

TABLE III. Comparison with other proposed synapses

Device	**Dimensions**	**Programming Energy/ Operating Voltage**	**Programming Time**	**Terminals**	**Programming Mechanism**
GeSbTe memristor [7]	$40nm$ mushroom and $10nm$ pore	Average $2.74pJ$/ event	$60ns$	2	Programmed by Joule heating (Phase change)
GeSbTe memristor [11]	$75nm$ electrode diameter	$50pJ$ (reset) & $0.675pJ$ (set)	$10ns$	2	Programmed by Joule heating (Phase change)
Ag/AgInSbTe/Ag chalcogenide memristor [40]	$100\mu m$ x $100\mu m$	Threshold voltage - $0.3V$	$5\mu s$	2	Programmed by Joule heating (Phase change)
Ag-Si memristor [8]	$100nm$ x $100nm$	Threshold voltage - $2.2V$	$300\mu s$	2	Movement of Ag ions
FeFET [10]	Channel length - $3\mu m$	Maximum gate voltage - $4V$	$10\mu s$	3	Gate voltage modulation of ferroelectric polarization
Floating gate transistor [9]	$1.8\mu m/0.6\mu m(0.35\mu m$ CMOS technology)	$V_{dd} - 4.2V$ & Tunneling Voltage -$15V$	$100\mu s$ (injection) & $2ms$ (tunneling)	3	Injection and tunneling currents
SRAM synapse [6]	$0.3\mu m^2$ ($10nm$ CMOS technology)	Average $328fJ$ (4-bit synapse)	-	-	Digital counter based circuits
Spintronic synapse	Ferromagnet dimensions - $320nm$ x $20nm$	Maximum $48fJ$/ event	$1ns$	3	Spin-orbit torque

addition to associated high power consumption. Programming is also relatively slow in such three terminal synaptic devices [9, 10]. It is worth noting here, that the current flowing through the oxide in the MTJ structure for our proposed synapse is the read current which is $\sim nA$ and drives sub-threshold CMOS circuits. SRAM based synapses have been also proposed for digital CMOS based SNN design [6]. However, for implementing 1 bit of the synapse, an 8-T SRAM cell has to be used, thereby leading to significant area overhead for implementation of a single synapse [6]. In addition, learning circuits will involve multiple digital counters and will be more area/power consuming than our proposed design.

Interested readers are referred to Ref. [43] for a discussion on the practical implementation of arrays of such spintronic devices interfaced with CMOS transistors. The size limitation of crossbar arrays of such spintronic devices is determined by the driving capabilities of rows of the array by input voltages in the presence of parasitics. In addition, sneak paths also become a potential issue for large crossbar arrays in order to implement on-chip learning. These are concerns that are equally valid for spin-devices and other memristive technologies, in general. However, it is worth noting here that computation occurring in a large crossbar can be distributed easily among smaller crossbar arrays by simply replacing the large unit by an equivalent number of smaller crossbar units using peripheral control circuitry.

In conclusion, we formulated a device, circuit and algorithm co-simulation framework calibrated to experimental results to validate the functionalities and performance of the proposed hybrid spintronic-CMOS based SNN design with on-chip learning. We proposed circuit primitives for generating STDP in the proposed synapse and demonstrated how such synaptic devices could be arranged in a crossbar fashion leading to an area and power efficient SNN implementation that is capable of recognizing patterns in input data. Simulation studies indicate the efficiency of the proposed hybrid spintronic-CMOS based SNN design as an ultra-low power neuromorphic computing platform capable of online learning.

ACKNOWLEDGMENT

The work was supported in part by, Center for Spintronic Materials, Interfaces, and Novel Architectures (C-SPIN), a MARCO and DARPA sponsored StarNet center, by the Semiconductor Research Corporation, the National Science Foundation, Intel Corporation and by the National Security Science and Engineering Faculty Fellowship.

[1] S. Ghosh-Dastidar and H. Adeli, Spiking neural networks, *International journal of neural systems* 19, 295 (2009).

[2] H. Markram, The blue brain project, *Nature Reviews Neuroscience* 7, 153 (2006).

[3] J. Schemmel, J. Fieres, and K. Meier, Wafer-scale integration of analog neural networks, in *Neural Networks (IJCNN), 2008 IEEE International Joint Conference on* (IEEE, 2008), pp. 431–438.

[4] X. Jin, M. Lujan, L. A. Plana, S. Davies, S. Temple, and S. Furber, Modeling spiking neural networks on SpiNNaker, *Computing in Science & Engineering* 12, 91 (2010).

[5] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura *et al.*, A million spiking-neuron integrated circuit with a scalable communication network and interface, *Science* 345, 668–673 (2014).

[6] B. Rajendran, Y. Liu, J.-s. Seo, K. Gopalakrishnan, L. Chang, D. J. Friedman, and M. B. Ritter, Specifications of nanoscale devices and circuits for neuromorphic computational systems, *Electron Devices, IEEE Transactions on* 60, 246 (2013).

[7] 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)* 9, 12 (2013).

[8] 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* 10, 1297 (2010).

[9] S. Ramakrishnan, P. E. Hasler, and C. Gordon, Floating gate synapses with spike-time-dependent plasticity, *Biomedical Circuits and Systems, IEEE Transactions on* 5, 244 (2011).

[10] Y. Nishitani, Y. Kaneko, M. Ueda, E. Fujii, and A. Tsujimura, Dynamic observation of brain-like learning in a ferroelectric synapse device, *Japanese Journal of Applied Physics* 52, 04CE06 (2013).

[11] D. Kuzum, R. G. Jeyasingh, B. Lee, and H.-S. P. Wong, Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing, *Nano letters* 12, 2179 (2011).

[12] M. Sharad, C. Augustine, G. Panagopoulos, and K. Roy, Spin-based neuron model with domain-wall magnets as synapse," *Nanotechnology, IEEE Transactions on* 11, 843 (2012).

[13] S. G. Ramasubramanian, R. Venkatesan, M. Sharad, K. Roy, and A. Raghunathan, SPINDLE: SPINtronic deep learning engine for large-scale neuromorphic computing, in *Proceedings of the 2014 international symposium on Low power electronics and design* (ACM, 2014), pp. 15–20.

[14] A. Sengupta, S. H. Choday, Y. Kim, and K. Roy, Spin orbit torque based electronic neuron, *Applied Physics Letters* 106, 143701 (2015).

[15] R. Morris, The organization of behavior, Wiley: New york; 1949, *Brain research bulletin* 50, 437 (1999).

[16] G.-q. Bi and M.-m. Poo, Synaptic modification by correlated activity: Hebb's postulate revisited, *Annual review of neuroscience*, 24, 139 (2001).

[17] P. U. Diehl and M. Cook, Unsupervised learning of digit recognition using spike-timing-dependent plasticity, *Frontiers in Computational Neuroscience* 9, 99 (2015).

[18] P. Knag, J. K. Kim, T. Chen, and Z. Zhang, A sparse coding neural network asic with on-chip learning for feature extraction and encoding, *Solid-State Circuits, IEEE Journal of* 50, 1070 (2015).

[19] S. Emori, U. Bauer, S.-M. Ahn, E. Martinez, and G. S. Beach, Current-driven dynamics of chiral ferromagnetic domain walls, *Nature materials*, 12, 611 (2013).

[20] G. Chen, J. Zhu, A. Quesada, J. Li, A. NDiaye, Y. Huo, T. Ma, Y. Chen, H. Kwon, C. Won *et al.*, Novel chiral magnetic domain wall structure in Fe/Ni/Cu (001) films, *Physical review letters*, 110, 177204 (2013).

[21] E. Martinez, S. Emori, N. Perez, L. Torres, and G. S. Beach, Current-driven dynamics of Dzyaloshinskii domain walls in the presence of in-plane fields: Full micromagnetic and one-dimensional analysis, *Journal of Applied Physics* 115, 213909 (2014).

[22] 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* 90, 184427 (2014).

[23] N. Perez, L. Torres, and E. Martinez-Vecino, Micromagnetic Modeling of Dzyaloshinskii–Moriya Interaction in Spin Hall Effect Switching, *Magnetics, IEEE Transactions on* 50, 1 (2014).

[24] J. Hirsch, Spin hall effect, *Physical Review Letters* 83, 1834, (1999).

[25] A. Sengupta, Z. Al Azim, X. Fong, and K. Roy, Spin-orbit torque induced spike-timing dependent plasticity, *Applied Physics Letters* 106, 093704 (2015).

[26] J. Lazzaro and J. Wawrzynek, *Low-power silicon neurons, axons and synapses.* Springer, 1994.

[27] C. Bartolozzi and G. Indiveri, Synaptic dynamics in analog VLSI, *Neural computation* 19, 2581 (2007).

[28] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri, Neuromorphic electronic circuits for building autonomous cognitive systems, *Proceedings of the IEEE* 102, 1367 (2014).

[29] G. Indiveri, R. Legenstein, G. Deligeorgis, T. Prodromakis *et al.*, Integration of nanoscale memristor synapses in neuromorphic computing architectures, *Nanotechnology* 24, 384010 (2013).

[30] G. Indiveri, A low-power adaptive integrate-and-fire neuron circuit, in *Circuits and Systems (ISCAS), 2003 IEEE International Symposium on* (IEEE, 2003), pp. 820–823.

[31] P. Livi and G. Indiveri, A current-mode conductance-based silicon neuron for address-event neuromorphic systems, in *Circuits and Systems (ISCAS), 2009 IEEE International Symposium on* (IEEE, 2009), pp. 2898–2901.

[32] A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. Van Waeyenberge, The design and verification of mumax3, *AIP Advances* 4, 107133 (2014).

[33] D. F. Goodman and R. Brette, The brian simulator, *Frontiers in neuroscience* 3, 192 (2009).

[34] J. C. Slonczewski, Conductance and exchange coupling of two ferromagnets separated by a tunneling barrier, *Physical Review B* 39, 6995 (1989).

[35] 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* 3, 868 (2004).

[36] 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.

[37] X. Fong, S. K. Gupta, N. N. Mojumder, S. H. Choday, C. Augustine, and K. Roy, KNACK: A hybrid spin-charge mixed-mode simulator for evaluating different genres of spin-transfer torque MRAM bit-cells, in *Simulation of Semiconductor Processes and Devices (SISPAD), 2011 International Conference on* (IEEE, 2011), pp. 51–54.

[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE* 86, 2278 (1998).

[39] D. Querlioz, O. Bichler, P. Dollfus, and C. Gamrat, Immunity to device variations in a spiking neural network with memristive nanodevices, *Nanotechnology, IEEE Transactions on* 12, 288 (2013).

[40] Y. Li, Y. Zhong, J. Zhang, L. Xu, Q. Wang, H. Sun, H. Tong, X. Cheng, and X. Miao, Activity-dependent synaptic plasticity of a chalcogenide electronic synapse for neuromorphic systems, *Scientific reports* 4, 4906 (2014).

[41] A. Sengupta, P. Panda, P. Wijesinghe, Y. Kim, and K. Roy, Magnetic tunnel junction mimics stochastic cortical spiking neurons, *Scientific reports* 6, 30039 (2016).

[42] A. Sengupta and K. Roy, Short-term plasticity and long-term potentiation in magnetic tunnel junctions: Towards volatile synapses, *Physical Review Applied* 5, 024012 (2016).

[43] H. Noguchi, K. Ikegami, K. Kushida, K. Abe, S. Itai, S. Takaya, N. Shimomura, J. Ito, A. Kawasumi, H. Hara *et al.*, A 3.3 ns-access-time 71.2μw/mhz 1mb embedded stt-mram using physically eliminated read-disturb scheme and normally-off memory architecture, in *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers* (IEEE, 2015), pp. 1–3.