Stochastic Spiking Neural Networks Enabled by Magnetic Tunnel Junctions: From Nontelegraphic to Telegraphic Switching Regimes

Chamika M. Liyanagedera, Abhronil Sengupta, Akhilesh Jaiswal, and Kaushik Roy

Phys. Rev. Applied **8**, 064017 — Published 15 December 2017

DOI: 10.1103/PhysRevApplied.8.064017

Magnetic Tunnel Junction Enabled Stochastic Spiking Neural Networks: From Non-Telegraphic to Telegraphic Switching Regime

Chamika M. Liyanagedera,* Abhronil Sengupta, Akhilesh Jaiswal, and Kaushik Roy
Purdue University, West Lafayette, IN 47906
(Dated: November 27, 2017)

Stochastic Spiking Neural Networks based on nanoelectronic spin devices can be a possible pathway at achieving "brain-like" compact and energy-efficient cognitive intelligence. The computational model attempt to exploit the intrinsic device stochasticity of nanoelectronic synaptic or neural components to perform learning or inference. However, there has been limited analysis on the scaling effect of stochastic spin devices and its impact on the operation of such stochastic networks at the system level. This work attempts to explore the design space and analyze the performance of nanomagnet based stochastic neuromorphic computing architectures for magnets with different barrier heights. We illustrate how the underlying network architecture must be modified to account for the random telegraphic switching behavior displayed by magnets with low barrier heights as they are scaled into the superparamagnetic regime. We perform a device to system level analysis on a deep neural network architecture for a digit recognition problem on the MNIST dataset.

I. INTRODUCTION

Emulating the computational primitives of neural network based machine learning approaches by the inherent device physics of nanoelectronic components have proven to be useful in reducing the area and energy requirements of the underlying hardware fabrics. To that effect, several post-CMOS technologies like phase change memories [1], Ag-Si devices [2], spintronic devices [3] among others have shown to exhibit neural and synaptic functionalities at the intrinsic device level. In this work, we focus on spintronic technologies, in particular, due to the low current and energy requirements of such devices in comparison to traditional memristive technologies.

While traditional neuromorphic computing models have been based on deterministic neural and synaptic primitives, recent efforts have been directed towards adapting such computing schemes to stochastic models. This has been driven primarily by two factors: (1) Deterministic neural or synaptic models are characterized by multi-bit resolution. However, as device dimensions of nanoelectronic neurons or synapses are scaled down, they might lose the multi-bit resolution capacity. In conjunction, such devices are expected to exhibit increased stochasticity during the switching process. For instance, spintronic devices exhibit stochasticity due to thermal noise at non-zero temperatures. Consequently, computational models that leverage the underlying device stochasticity are being recently explored. Information encoding over time due to probabilistic synaptic or neural updates also enables state compression of neural and synaptic units, thereby allowing them to be implemented by single-bit technologies. (2) The human brain, the main inspiration behind such neuromorphic computing models, is characterized by stochastic neural and synaptic units. As a matter of fact, neuroscience studies have indicated cortical neurons generate spikes probabilistically over time [4]. Consequently, stochastic neural computing models can potentially enable "brain-like" cognitive computing. In this work, we will focus on stochastic neural inference in deep neural networks for typical pattern recognition tasks [5]. However, the analysis can be easily extended to stochastic synaptic units [6], or even other unconventional computing platforms that require stochastic switching elements like Ising computing [7, 8], Bayesian inference, among others.

Spintronic devices have recently found wide application in large scale neuro computing hardware owing to their scalability and low power requirements. Spin-torque memristors with magnetic domain walls have been shown to be a suitable candidate for implementing multi-level neuro-synapses [9] and integrate and fire spiking neurons [3].Another study demonstrated that the inherent magnetic dynamics of a MTJs can be used to emulate the functionality of biologically inspired leaky integrate and fire spiking neurons [10]. In [11], spin-transfer torque Magnetic Tunnel Junctions (MTJs) are used as stochastic binary synapses, where the stochastic effects of the devices are used to performed unsupervised learning. It was also demonstrated that MTJs can be used as binary elements to implement long-term short-term stochastic synapses to improve the learning efficiency of a neural network [6]. A review on bio-inspired neuromorphic computing platforms based on spintronic devices can be found in [12].

As mentioned previously, spintronic devices display a stochastic switching nature due to thermal noise. Given a particular duration of write current flowing through the device, a magnet exhibits a particular probability of switching during that corresponding write cycle. Consecutive write and read cycles can be used to generate an output pulse stream whose average value depends on the magnitude of the input stimulus. While stochastic neural networks based on spintronic devices have been explored previously [5, 13], there has been limited analysis on the scaling effects of these devices. It is generally

* cliyanag@purdue.edu

expected that as the magnet dimensions scale down, the device would exhibit increased stochasticity. Further, the operating current or voltage ranges required for operating such devices in the probabilistic regime would reduce. However, as the scaling tends to the superparamagnetic regime the magnets undergo random telegraphic switching with low data retention time, making the device practically volatile in nature. Utilizing such a device as a biased random generator require re-thinking of the peripherals and the underlying network architecture, since parallel read and write operations of the nano-magnets are now required. However, adaptation of such low energy superparamagnets as neural components come at the expense of reduced error resiliency. This is mainly because the gradient or the rate of change of switching characteristics of such magnets in response to input current magnitude is extremely high. This article attempts to address the different schemes of operation of stochastic Spiking Neural Networks (SNNs) for magnets in non-telegraphic to telegraphic regime and analyze its associated energy-accuracy tradeoffs at the system level.

II. MAGNETIC TUNNEL JUNCTION AS A STOCHASTIC SWITCHING ELEMENT

An MTJ is a magneto-resistive device that consists of a tunneling oxide sandwiched between two magnetic contacts. One of the contacts is magnetically hardened and is called the *pinned* layer, while the direction of magnetization of the other contact, called the *free* layer, can be switched. In a spin-Hall effect based MTJ (SHE-MTJ), the direction of the free layer is switched by passing a charge current through an underlying heavy metal (HM), as shown in Fig. 1. The passage of the charge current (I_{charge}) through the HM layer induces a resulting spin current (I_{spin}) flowing perpendicular to the planes of the magnetic layers of the MTJ. This spin current can switch the direction of magnetization of the free layer, making it parallel (P) or anti-parallel (AP) to that of the pinned layer, through the well known spin-orbit torque mechanism [14, 15]. Due to the magneto-resistance effect, the SHE-MTJ exhibits a lower resistance (R_P), when in the P state and a higher resistance (R_{AP}), when in the AP state. Thus, the SHE-MTJ shown in Fig. 2, exhibits decoupled read and write current paths. Write operation can be achieved by a charge current flowing through the HM layer, while the read operation can be accomplished by sensing the resistance of the MTJ in a direction transverse to the plane of the magnetic layers.

It is to be noted that the switching process of the nanoscale free layer is influenced by thermal noise at non-zero temperatures. Thermal noise results in a stochastic switching behavior, wherein, for a given current flowing through the HM layer, the MTJ switches with a certain probability. Moreover, the probability of switching can be controlled by the magnitude of the current flowing through the HM. The dynamics of the magnetization

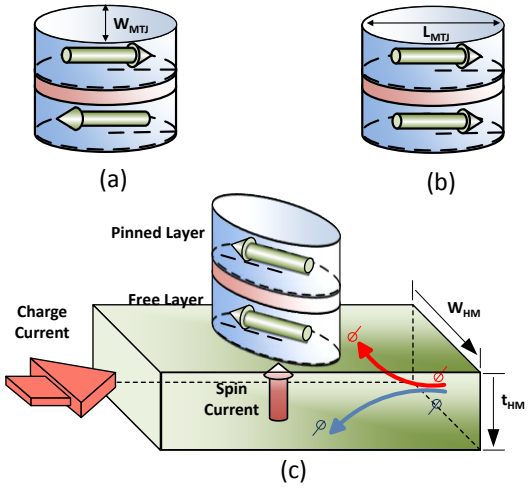

FIG. 1. (a) High resistive anti-parallel state of an MTJ, (b) Low resistive parallel state of an MTJ, and (c) A SHE-MTJ device structure where the MTJ is switched by passing charge current through the underlying heavy metal. The charge current flowing through the heavy metal leads to spin splitting, thereby creating a perpendicular spin current, switching the magnetization direction of the free layer.

vector in presence of the HM layer current is given by the stochastic Landau-Lifshitz-Gilbert-Slonczewski (LLGS) equation and can be written as [16],

$$\frac{\partial \widehat{m}}{\partial \tau} = -\widehat{m}\times\vec{H}_{EFF} - \alpha\widehat{m}\times\widehat{m}\times\vec{H}_{EFF} + \frac{1}{|\gamma|}(\alpha\widehat{m}\times S\vec{T}T + S\vec{T}T) \tag{1}$$

where τ is $\frac{|\gamma|}{1+\alpha^2}t$.

Here, α is the Gilbert's damping constant, γ is the gyromagnetic ratio, $\widehat{m}$ is the unit vector in the direction of the magnetization, t is the simulation time and H_{EFF} is the effective magnetic field including the demagnetization field and the interface anisotropy field. A detailed description of the various fields included in H_{EFF} can be found in [16]. $S\vec{T}T$ in equation (1) is the term representing the torque due to the SHE effect (modeled as a spin-transfer torque term) and can be written as follows [17],

$$S\vec{T}T = |\gamma|\beta(\widehat{m}\times(\epsilon_{she}\widehat{m}\times\widehat{mp})),\ \beta = \frac{\hbar J_q}{2e\mu_o M_S t_{FL}} \tag{2}$$

where, $\widehat{mp}$ is the magnetization of the pinned layer, e is charge of an electron, μ_o is the permeability of vacuum, $\hbar$ is modified Planck's constant, t_{FL} is the thickness of the free layer and M_S is saturation magnetization. J_q is the charge current density flowing through the heavy metal. ϵ_{she} is the spin polarization efficiency (defined as the ratio of the spin current generated due to the charge current flowing through the HM layer) and can be written as [18],

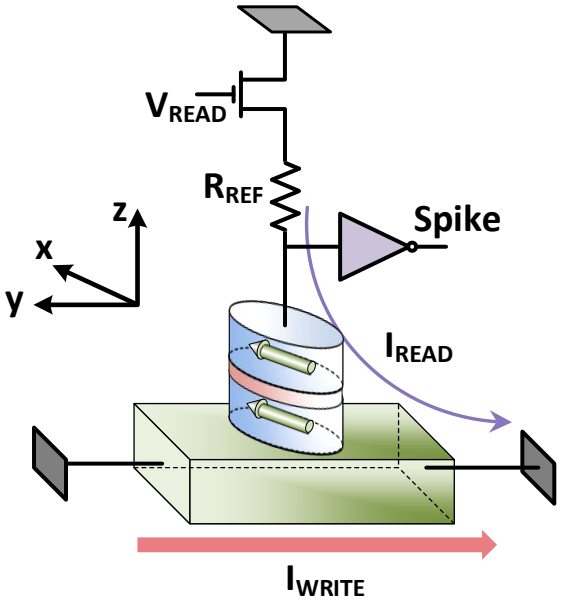

FIG. 2. Decoupled read and write current paths of the MTJ with HM. Output of the inverter will be high if the MTJ is in the P state, and low if the MTJ is in the AP state.

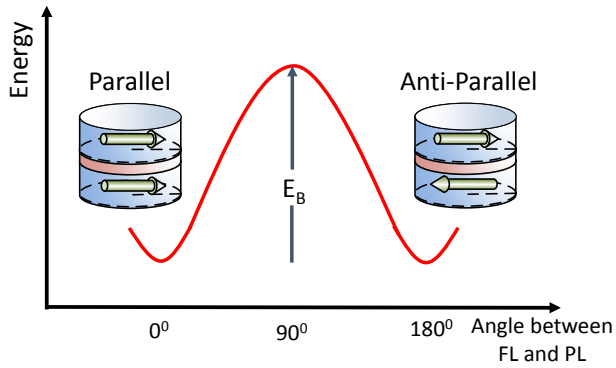

FIG. 3. The two operating states of an MTJ. The two states are thermally stable if the barrier height of the magnet, E_B, is large enough.

$$\epsilon_{she} = \frac{I_{spin}}{I_{charge}} = \frac{\pi w}{4t}\theta_{she}\left(1 - sech\left(\frac{t}{\lambda_{sf}}\right)\right) \tag{3}$$

where, w is width of free layer, t is thickness of heavy metal, θ_{she} is spin hall angle, λ_{sf} is spin flip length.

The random switching process due to the effect of the thermal noise can be included in the LLGS equation through a stochastic field $\vec{H}_{thermal}$ in $\vec{H}_{EFF}$ [19],

$$\vec{H}_{thermal} = \vec{\zeta}\sqrt{\frac{2\alpha k_B T}{|\gamma| dt M_S Vol}} \tag{4}$$

where, k_B is the Boltzmann constant, T is the temperature, Vol is volume of the free layer magnet and dt is the simulation time step. The $\vec{\zeta}$ term in eqn. 4 is a Gaussian random variable with zero mean and a standard deviation equal to 1. The inclusion of thermal noise turns the LLG equations into a stochastic differential equation (SDE). We used the Heun's method to integrate the stochastic LLG equation. The details of applying Heun's method to stochastic LLG equation can be found in [19], [20]. The total Field acting on the nano-magnet H_{EFF} is given by,

$$\vec{H}_{EFF} = \vec{H}_{thermal} + \vec{H}_{aniso} + \vec{H}_{external} \tag{5}$$

where $\vec{H}_{aniso}$ is the anisotropy field and in in-plane magnets its is dominated by the demagnetization field arising due to the shape of the magnet and is given by

$$\vec{H}_{demag} = -M_S[N_{xx}m_x\hat{x}, N_{yy}m_y\hat{y}, N_{zz}m_z\hat{z}] \tag{6}$$

where N_{xx}, N_{yy}, N_{zz} are the demagnetization factors that were calculated based on the analytical equations presented in [21], and m_x, m_y, m_z are the magnetization components of the nano-magnet in the $\hat{x}, \hat{y}$ and $\hat{z}$ directions. The presence of any external field can be included through the term $\vec{H}_{external}$.

A. Stochasticity in Non-Telegraphic Regime

The parallel and anti-parallel states of the MTJ is stabilized by an energy barrier, E_B, that is defined as the product of the magnetic anisotropy and volume (Fig. 3). The retention time for the magnetic state of a nano-magnet is given by [22],

$$T_{RETENTION} = \tau_0 exp(\frac{E_B}{k_B T}) \tag{7}$$

where, τ_0 is a characteristic time constant in the range $1ps - 100ps$ [22]. The retention time or the lifetime of the magnet varies exponentially with the barrier height. The non-volatility of the magnet enables such devices to be used in synchronous clocked systems where the device is operated in successive write and read phases. During the write cycle, a current pulse of fixed duration is passed through the HM layer, that can switch the MTJ from one state over the barrier to the other stable state. The switching probability of the magnet varies with the magnitude of the current pulse flowing through the underlying HM layer. During the read phase, a small current is passed through the MTJ-R_{ref} (can be implemented by another MTJ whose state is not disturbed by the small read current) voltage divider circuit (refer Fig. 2) and the MTJ state is read at the output of the inverter. The read current should be sufficiently small such that it does not disturb the state of the MTJ during the read phase.

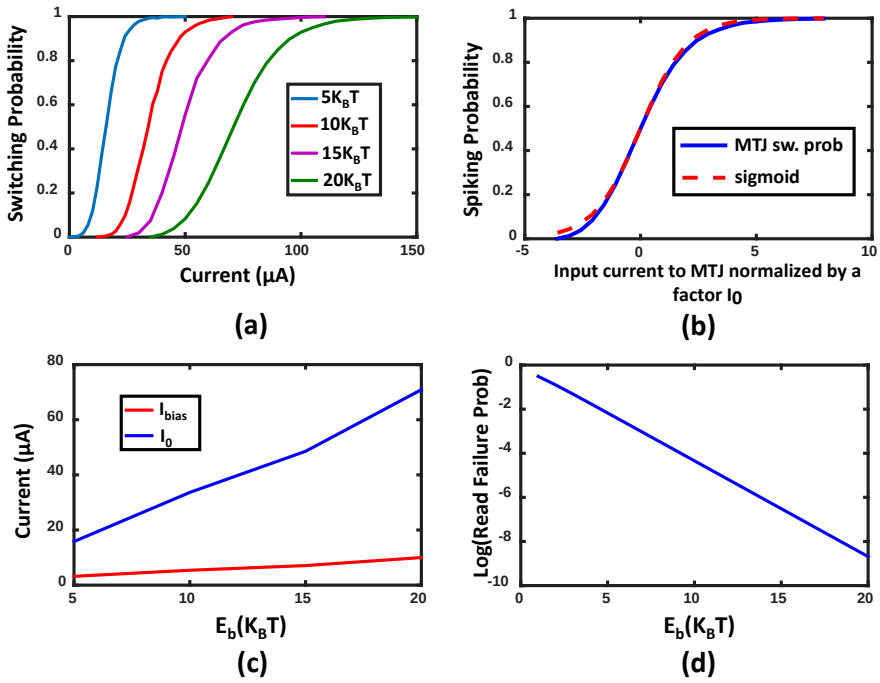

FIG. 4. (a) Switching characteristics of an MTJ with varying E_B at $T = 300K$ for a write cycle duration of $0.5ns$, (b) MTJ switching probability characteristics as a function of $I - I_{bias}$, normalized by a factor I_o. The data closely resembles the sigmoid function, (c) Variation of the bias current, I_{bias}, and the normalizing factor, I_o, with varying E_B. Both I_{bias} and I_o decrease with decreasing E_B, (d) Failure probability during a read cycle of $1ns$ (in logarithm scale) with varying E_B.

TABLE I. Device Parameters

Parameter	Values			
	$1K_BT$	$2K_BT$	$10K_BT$	$20K_BT$
Free Layer Width, W_{MTJ}	$10nm$	$17nm$	$30nm$	$40nm$
Free Layer Length, L_{MTJ}	$25nm$	$42.5nm$	$75nm$	$100nm$
Free Layer thickness	0.8 nm		1.2 nm	
Saturation magnetization, M_s	750 KA/m		1000 KA/m	
Heavy metal thickness	$2nm$			
Spin-Hall Angle, θ_{she}	0.3 [15]			
Gilbert's damping factor, α	0.0122 [15]			
Temperature, T	$300K$			

Since the voltage difference at the voltage divider output for the parallel and anti-parallel states is generally small, multiple stages of inverters are required to obtain a full swing at the output.

Fig. 4(a) illustrates the variation of the MTJ switching probability with the amplitude of the current pulse being passed through the HM layer for different E_B. The device parameters used for simulations are enlisted in Table I. Note that the barrier height of the magnet was varied by scaling the area of the magnets appropriately. It can be shown that the probabilistic switching characteristics of the MTJ holds a sigmoidal relationship to the write current by describing the SHE layer current I, with two different parameters, namely I_{bias} and I_o. I_{bias} is the dc current required to bias the switching probability of the MTJ to 0.5, and I_o is the scaling factor used to map the swing of the switching probability around the bias current to the sigmoid curve. Fig. 4(b) depicts the variation of the switching probability of the MTJ with $I - I_{bias}$, normalized by a factor I_o. I_o can be found by fitting the switching probability characteristics ($Psw()$) to the sigmoid function such that (refer Fig.4 b),

$$sigmoid(\frac{I - I_{bias}}{I_o}) \approx Psw(I) \tag{8}$$

As shown in Fig. 4(a), when E_B and hence, the device dimensions are scaled down, the current range required for stochastic switching decreases, thereby reducing the write current requirements of the device. Fig. 4(c) depicts that both the components, I_{bias} and I_o, reduce with reduction in the barrier height. Reduction in I_o implies that the current range that can be utilized for stochastic MTJ switching reduces, thereby increasing the rate of change of switching probability with varying input current. Consequently the computing system becomes more prone to variations in the MTJ input current and exhibits less error resiliency with the reduction of I_o. These considerations will be highlighted in the next section.

Note that, if E_B is not sufficiently large, the state of the magnet can switch during the read operation due to very small $T_{RETENTION}$. The retention failure probability $P_{F,RETENTION}$, of an MTJ within a given read access time is given by,

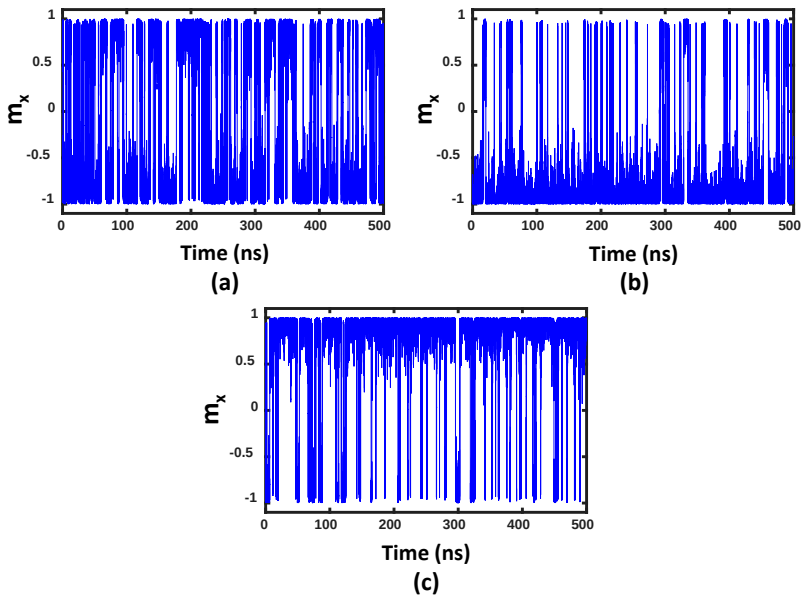

FIG. 5. Switching characteristics of an MTJ with $1k_BT$ barrier height: (a) When the current flowing through the HM is zero, the MTJ is equally likely to be in the parallel or anti-parallel state, (b) When $-1.5\mu A$ is flowing through the HM layer, the MTJ is more likely to be in the anti-parallel state, (c) When $1.5\mu A$ is flowing through the HM layer, the MTJ is more likely to be in the parallel state.

$$P_{F,RETENTION} = 1 - exp(-t_{read}/exp(\Delta)) \qquad (9)$$

where, $P_{F,RETENTION}$ is the retention failure probability of the MTJ during a read time of t_{read} in nanoseconds, and Δ is the E_B of the MTJ in k_BT. In order to find the necessary t_{read} for correct read operation, SPICE simulations (with a Verilog A model for the MTJ [23]) were performed in IBM $45nm$ technology node. Simulation results show that the required read time is around $0.2ns$ for the nominal corner and $1ns$ for the worst case corner (with 2σ variations in the threshold voltage of the CMOS transistors). Hence, for retention failure probability calculations the required read time is taken to be $1ns$ to ensure that a correct read can be achieved even at the worst corner. As illustrated in Fig. 4(d), retention failure probability increases exponentially as the MTJ is scaled down. In order to keep the retention failure probability smaller than 1%, the E_B of the magnet should be kept greater than $4.6k_BT$. When the MTJs are scaled further they enter the superparamagnetic regime where the magnets are no longer thermally stable during the read cycle. Hence, parallel read-write operations are required for magnets in the superparamagnetic regime ($E_B < 5k_BT$) to realize stochastic switching elements.

B. Stochasticity in the Telegraphic Regime

For low barrier height nano-magnets ($E_B \sim 1k_BT$), even with zero charge current flowing through the HM layer, the MTJ will exhibit random telegraphic switching between the two equilibrium states (Fig. 5(a)) due to thermal noise. The random switching characteristics of such scaled devices in the superparamagnetic regime can be still manipulated by passing a charge current through the HM layer. For instance, Fig. 5(a)-(c) represents the in-plane magnetization of the MTJ in presence of $0, 1.5, -1.5\mu A$ write current flowing through the HM layer of a $1k_BT$ magnet. The dwell time of the MTJ in either of the two stable states can be modulated by the magnitude and direction of the input write current.

The volatility of these devices entails a rethinking of the manner in which such nano-magnets can be operated with peripherals to realize a stochastic computing element. Due to device volatility and low retention time, such devices cannot be operated with separate write and read phases. Consequently, the write and read terminals of the MTJ are activated simultaneously and the device state is read while an input bias current flows through the underlying HM layer of the MTJ. For high energy-barrier MTJs the effect of read current on the switching characteristics is not a design issue since read and write cycles are de-coupled in time. However, for MTJs in the telegraphic switching regime, the read current can bias the switching characteristics since the read and write operations occur in parallel. Further, since the devices are highly scaled, the write (for stochastic switching) and read currents fall in the same order of magnitude (unlike high barrier height magnets where the write current for stochastic switching is higher). Hence the resistive divider of the read circuit (Fig. 2) needs to be highly op-

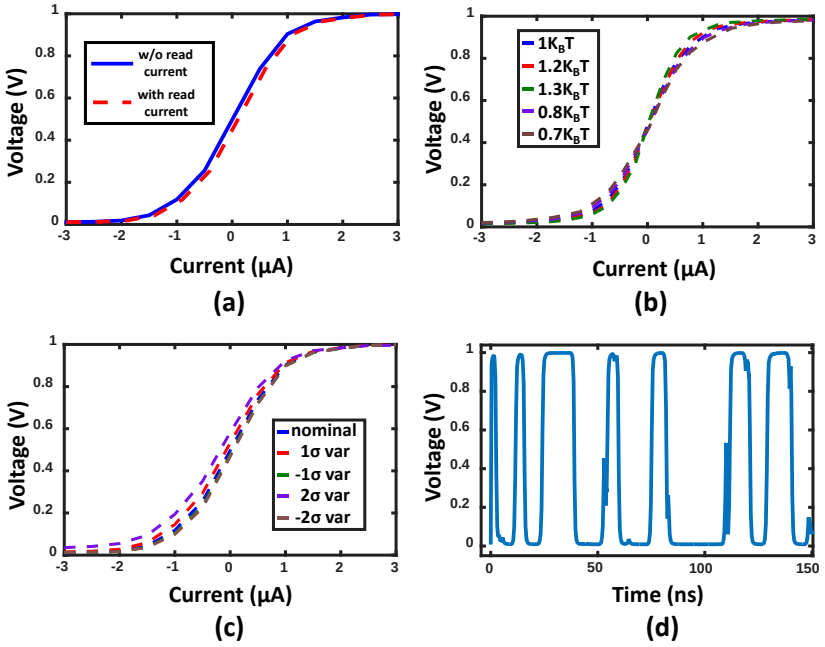

FIG. 6. (a) Average inverter output over a duration of $2\mu s$ with and without the impact of the read current, (b) Variation of the inverter average output over a duration of $2\mu s$ with magnitude of the write current for different E_B values, (c) Inverter average output over a duration of $2\mu s$ for nominal corner and for the worst case conditions of $\pm 1\sigma$ and $\pm\ 2\sigma$ variations in the threshold voltages of the transistors, (d) A typical plot of the output voltage of the inverter stage of the read circuit as a function of time under zero external input current.

timized such that the read current is maintained at the minimal value. SPICE simulations reveal that the read current can be minimized to $100nA$ while having minimal effect on the MTJ switching characteristics. Fig. 6(a) depicts the average output of the inverter stage over a duration of $2\mu s$ with and without the read current. The case "with read current" is simulated by considering the additional spin-orbit torque induced by the $100nA$ read current flowing through the HM layer while the case "without read current" ignores the effect of the additional read current. As can be observed from Fig. 6, the read current has minimal impact on the MTJ switching probability. Further, effect of device dimension variations (or equivalently E_B variations) and read circuit variations ($\pm 1\sigma$ and $\pm\ 2\sigma$ variations in the threshold voltages of the CMOS transistors) was shown to have minimal effect on the stochastic switching behavior of the nano-magnets (Figs. 6(b)-(c)). Fig. 6(d) represents a typical plot of the voltage output of the inverter stage as a function of time with no input current flowing through the underlying HM of the MTJ.

Note that the switching characteristics of superparamagnetic MTJs are highly sensitive to any change in the magnitude of the write current. As depicted in Fig. 6(a), the switching probability of the MTJ shifts from 0.5 to 0.85 for a $1\mu A$ change in the write current. Hence, the impact of variations in the input current provided to a network of such scaled MTJs can be significant, and will be analyzed in more details in the next section. We

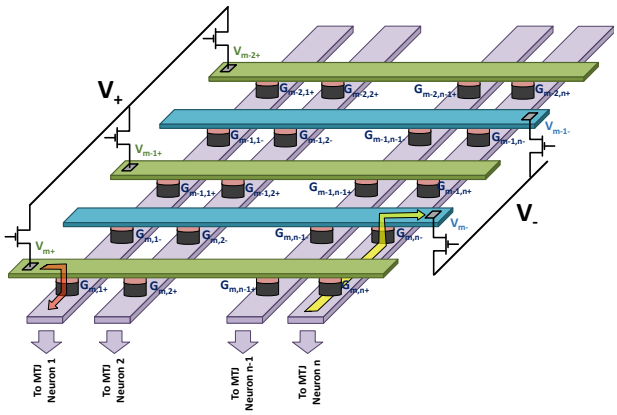

FIG. 7. Crossbar architecture connecting the inputs of one layer to the neurons of the corresponding layer. Horizontal bars provide the input voltage for the synapses. The summation of weighted synaptic currents along the columns of the crossbar array are then provided as inputs to the MTJ neurons.

would like to conclude this section by mentioning that parallel read-write operation is not suited for magnetization switching in the non-telegraphic regime ($10-20k_BT$ barrier height magnets) since the telegraphic switching would occur in timescales of $\sim \mu s - ms$, thereby, resulting in enhanced delay for the computing process.

C. Stochastic Neuromorphic Computing

A neural network is essentially a collection of layers of neurons interfaced through a network of weighted synapses. A particular input to a neuron is first scaled by the corresponding synaptic weight of the synapse before they are accumulated and processed by the neuron. Neurons with sigmoid like transfer functions have been shown to be appealing for implementing deep spiking neural networks [5], making SHE-MTJ structures ideal for realizing energy efficient neuromorphic hardware. In the stochastic neural network being considered in this work, the MTJ neurons generates an output spike probabilistically depending on the instantaneous magnitude of the resultant weighted synaptic input [5]. This computing framework can be directly translated to the resistive crossbar architecture illustrated in Fig. 7, where the synaptic weights are mapped into the resistive elements between the horizontal and vertical metal lines. Note that resistive crossbar arrays based on memristive devices like phase change materials [1], Ag-Si devices [2] and spintronic devices [24] have been proposed and experimentally demonstrated [25]. Two horizontal lines are used for each input connected to the crossbar array to implement the functionality of positive and negative weights. An input spike provided to the network will activate the corresponding access transistors supplying a voltage to the horizontal lines V_+ (positive voltage) and V_- (negative voltage), which is translated to a current through the vertical columns (weighted by the conductances of the resistive elements). The current accumulated in the vertical columns are then supplied as the write currents to the stochastic neurons of the corresponding layer. If the weight connecting an input m to a neuron n is negative, then the corresponding resistive element connecting the positive horizontal line and the vertical column ($G_{m,n+}$) is programmed to a high resistive 'off' state and the weight connecting the vertical column and the negative horizontal line is programmed to a conductance given by $G_{m,n-} = w_{m,n}G_o$ and vice versa. Here, $w_{m,n}$ is the synaptic weight between the input m and neuron n and G_o is the mapped conductance for unity weight. The conductances of the resistive elements are selected by scaling the synaptic weights by a factor G_o given by, $\frac{I_o}{\delta V}$, where δV is the magnitude of the supply voltage driving the rows of the crossbar array and I_o is the current scaling factor of the stochastic MTJ mentioned previously. Assuming that the magneto-metallic spin devices have low input resistance in comparison to the cross-point resistances of the crossbar array, the neurons will receive a weighted summation of spike inputs in a particular layer and produce output spikes probabilistically over time that will drive the fan-out neurons of the next layer. For magnetic neurons operating in the non-telegraphic regime, the read circuit can be interfaced with a latch that stores the inverter output during the read cycle, which will drive the next stage of neurons during the following write cycle (hence synchronous operation). For magnetic neurons operating in superparamagnetic regime, the inverter output can directly drive the neurons in the next stage (hence asynchronous operation). Note that the high barrier-height magnets are also driven by a current source to bias it at a switching probability of 0.5 unlike MTJs in the superparamagnetic regime. Due to the small input current and the zero bias current of magnetic neurons operating in the superparamagnetic regime, asynchronous architectures will grant significant power savings in the neurons and the resistive crossbar array. However, as shown later, asynchronous implementation will incur significant power loss at the read circuit, owing to the continuous switching activity of the inverters.

III. DESIGN CONSIDERATIONS: SYNCHRONOUS AND ASYNCHRONOUS NEUROMORPHIC SYSTEMS

A. Device to System Simulation Framework

In order to analyze the design considerations for synchronous and asynchronous stochastic SNNs, a hybrid device-circuit-system co-simulation framework was developed for this work. Stochastic LLGS simulation for MTJs with different barrier heights was used to evaluate the probabilistic switching behavior of magnets operating in non-telegraphic to telegraphic regime. In this work, we use magnets of barrier height $10k_BT$ and $20k_BT$ for non-telegraphic regime and magnets of barrier height $1k_BT$ and $2k_BT$ for telegraphic regime. The device parameters used for simulations are summarized in table I. SPICE level simulations based on a Verilog-A model of the MTJ was used to evaluate the performance of the stochastic MTJ along with associated peripherals.

In order to perform a system-level analysis, the performance of the network was assessed for a large scale deep learning network architecture (28x28-6c5-2s-12c5-2s-10o) on a standard digit recognition problem based on the MNIST dataset [26]. The network consists of alternate layers of convolutional and subsampling operations. The dimensions of the input MNIST images are 28x28, which were applied as input to the convolutional layer consisting of 6 convolutional kernels of size 5x5. The subsampling kernel was of size 2x2, and was followed by another convolutional layer comprising of 12 output maps, which in turn, was followed by another subsampling layer. The final layer consisted of 10 neurons, each of which represented one of the ten digit classes. Once the training was accomplished, the learnt weights are mapped to the synaptic conductances using a value of $G_o = 5\mu S$ which is a typical resistance range of memristive synaptic devices. The same resistive crossbar array was used for all the different barrier height neuronal devices. The supply voltage δV was adjusted in each case to satisfy the relationship, $\delta V = \frac{I_o}{G_o}$, as ex-

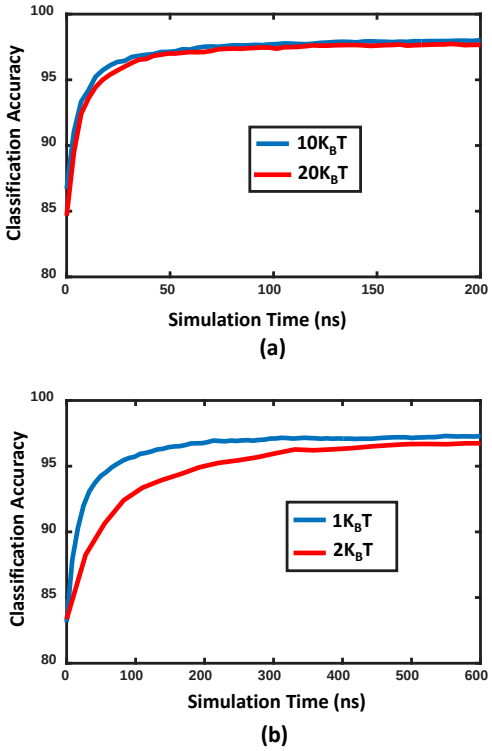

FIG. 8. Variation of classification accuracy of the proposed network with time for (a) Synchronous, and (b) Asynchronous implementations.

plained previously. The supply voltages δV, was calculated to be $0.1V$,$0.11V$,$1.05V$ and $2V$ for nano-magnets of barrier height $1K_BT$,$2K_BT$,$10K_BT$ and $20K_BT$ respectively. The sigmoid characteristic curves for the magnets operating in the telegraphic regime was obtained by averaging the output voltage of the read inverter circuit over a period of $2\mu s$ (for $1k_BT$) and $5\mu s$ (for $2k_BT$). More information about the structure of the simulated network , the training methodology [27], and a brief introduction to neural networks [28][29][30] can be found in the supplementary section provided with this article [31].

B. Performance and Energy Estimation

Fig. 8 depicts the temporal evolution of the classification accuracy of the stochastic SNN for the synchronous and asynchronous designs. For the $10K_BT$ and the $20K_BT$ synchronous designs the classification accuracy reaches 98.1% and 97.6% respectively, while it saturates at 97.5% and 97.2% for the $1K_BT$ and $2K_BT$ asynchronous designs. Both synchronous networks surpass an accuracy of 95% just under 20ns, whereas the two asynchronous networks require 80ns (for $1K_BT$) and 250ns (for $2K_BT$) to reach the same accuracy. In the asynchronous implementation, the high frequency telegraphic switching of the nano-magnets are translated into voltage spikes at a lower frequency due to gate capacitance charge delays of the CMOS devices, which explains the slower response of the asynchronous networks compared to the synchronous designs. Also as the E_B of the nano-magnets are increased (for the superparamagnetic regime), the retention time of the nano-magnets increase, decreasing the spiking frequency at the output of the inverters. Hence as the results show, for asynchronous designs, the time required for a network to reach a target accuracy increases with the E_B of the nano-magnets used in the design. For the synchronous networks the duration of one time-step was selected to be $4ns$, which includes a write time of $0.5ns$, a rest period of $2ns$, a read time of $1ns$ followed by a reset period of $0.5ns$. The duration of the time-step for the asynchronous networks were determined by measuring the average duration of a voltage pulse at the output of the inverter read circuit at zero write current, and was calculated to be $8.2ns$ and $27.5ns$ for the $1K_BT$ and $2K_BT$ networks, respectively.

Fig. 9 summarizes the energy consumption observed for different components of the network (both synchronous and asynchronous) corresponding to a target classification accuracy of 96%. Neuron energy (Fig. 9(a)) refers to the energy dissipated in the MTJ neuron due to the write/reset currents flowing through the HM layer. The neuron energy consumption is lowest for the $1K_BT$ asynchronous design with an energy consumption of $1.15pJ$ per image classification, and increases with the size of the magnets, up to $37.8pJ$ per image classification for the $20K_BT$ synchronous design. This trend can be explained by the increasing write current requirements of the nano-magnets as their sizes are increased. Since the current flowing through the HM layer are first routed through the resistive cross-bar network (synapses), the energy consumption in the synapses (Fig. 9(c)) show a similar trend, increasing with the size of the magnets. Also the bias current required in the synchronous designs to bias the switching probability of the MTJs to 0.5, adds to the power dissipation in the HM layer and the synapses. The energy consumption in the synapses per image classification are $0.27nJ$ and $0.74nJ$ for the $1K_BT$ and $2K_BT$ asynchronous designs and, $1.3\ nJ$ and $6.5nJ$ for the $10K_BT$ and $20K_BT$ synchronous designs. The read energy consumption, illustrated in Fig. 9(b), is the summation of the power dissipated in the MTJ due to the read current passing through and the power dissipated in the CMOS interface circuitry. As the results indicate, the read energy consumption per image classification are larger for the asynchronous implementations ($3.3\ nJ$ for the $1K_BT$ and $8.95nJ$ for the $2K_BT$) compared to the synchronous implementations ($2.1nJ$ for the $10K_BT$ and $2.75nJ$ for the $20K_BT$). The majority of the read power dissipation in asynchronous networks occur at the CMOS inverters, which are required to operate continuously due to the parallel read/write nature of the neurons. In synchronous networks, however, the CMOS inverters are only required to operate during

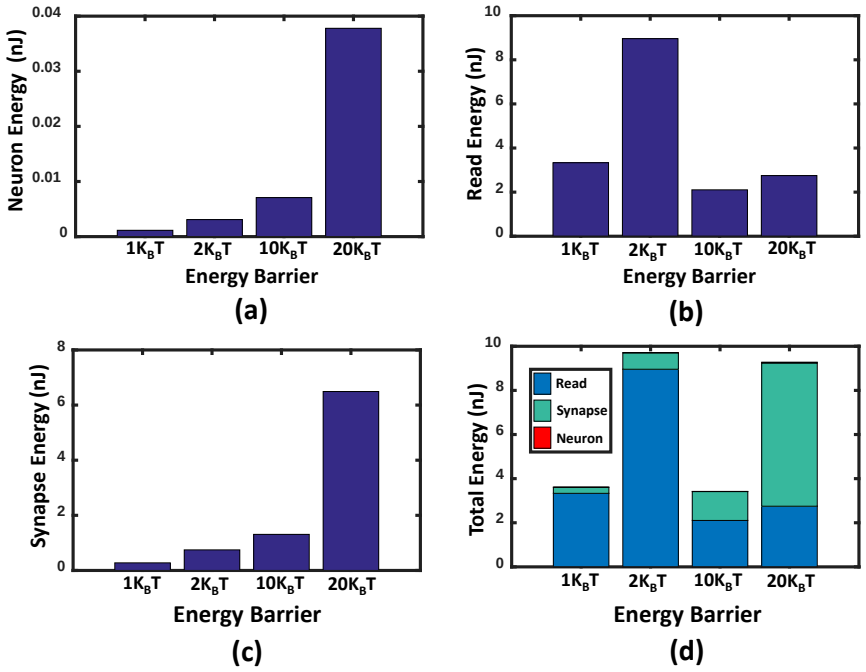

FIG. 9. (a) Energy consumption of the MTJ neuron, (b) Energy consumption of the read circuit, (c) Energy consumption of the synapses, (d) Total energy consumption per image classification (for an accuracy of 96%) for the asynchronous ($1k_BT$ & $2k_BT$) and synchronous ($10k_BT$ & $20k_BT$) networks.

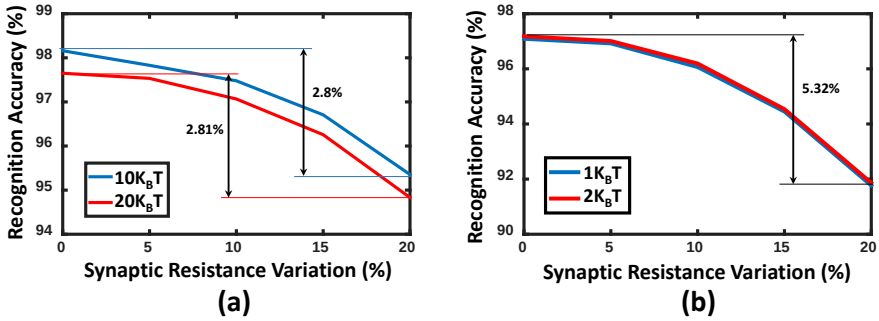

FIG. 10. Average classification accuracy (measured over 50 independent Monte Carlo simulations) with variations in the resistive synapses (%σ variations) for the (a) synchronous design, (b) asynchronous designs.

the read cycle, and can be deactivated at other times using access transistors to save power. For both designs the power dissipated in the neurons are an order of magnitude smaller compared to the power dissipated in the synapses and the read circuit, owing to the low resistance of the HM layer. As depicted by Fig. (9(d)), the $10K_BT$ synchronous network shows the minimum energy requirement per image classification ($3.4nJ$), closely followed by the $1K_BT$ asynchronous network ($3.6nJ$). The $2K_BT$ asynchronous network exhibit an energy consumption of $9.7nJ$ per image classification followed by the $20K_BT$ synchronous network with an energy consumption of $9.28nJ$. For the synchronous networks, the energy consumption associated with the clocking circuitry is negligible, especially since a classification accuracy of 96% can be achieved under 10 clock cycles, and hence is not considered in this analysis.

C. Effect of Variations

Most of the computations of the proposed network occurs in the resistive cross bar array. Hence, any variations in the resistive elements of the crossbar array can result in a significant degradation of the classification accuracy. To measure the effect of such variations, separate experiments were performed allowing variations with a

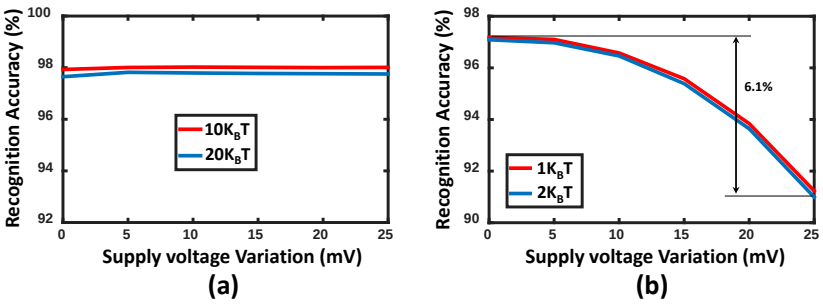

FIG. 11. Average classification accuracy (measured over 50 independent Monte Carlo simulations) with variations in the supply voltage (up to $25mV$ variations) for the (a) synchronous, and (b) asynchronous designs.

standard deviation up to 20% in the resistive elements. According to the results (refer Fig. 10), for variations in the synapses with a standard deviation of 20%, the accuracy loss is only 2.8% for the synchronous designs and 5.32% for the asynchronous designs. The slightly higher accuracy degradation observed in the asynchronous designs in comparison to the synchronous designs can be explained by the increased sensitivity of the MTJ switching probability in response to the write current at the superparamagnetic regime.

Due to the low operating currents of the nano-magnets used in the asynchronous design, the operating voltage of the crossbar architecture given by $\delta V = \frac{I_o}{G_o}$ can be very small for low K_BT magnets. Hence any variation in the supply voltage can potentially result in a large deviation in the write current magnitude, influencing the classification accuracy of the network. Fig. 11 depicts the behavior of the classification accuracy of the two designs in the presence of supply voltage variation. As shown by Fig. 11(a), due to the larger supply voltages used in the synchronous designs, $10K_BT$ and $20K_BT$ synchronous implementations are resilient to supply voltage variations up to $25mV$. The asynchronous implementations, on the order hand, exhibit an accuracy degradation of 6.1% under $25mV$ variation in the supply voltage.

As explained in section II, the CMOS inverter read circuit for the asynchronous implementation must be designed carefully so that the average magnetization of the nano-magnet is properly reflected on the average output of the inverter. Any variation in the CMOS circuitry can offset the average output of the inverters, adversely affecting the classification accuracy of the network. As depicted by Fig. 12, the classification accuracy of the $1K_BT$ asynchronous network decrease by 3% and the accuracy of the $2K_BT$ asynchronous network decrease by 0.7% at the worst case corner with 2σ variations in the CMOS read circuit. The synchronous networks are resilient towards such CMOS variations since the read time is selected to be adequate for a correct read even at the worst cell corner.

D. Effect of Temperature

In this work, the switching characteristics of the MTJs were varied between the telegraphic and non-telegraphic regime by adjusting the width of the FL appropriately. However, the switching characteristics of the MTJs can significantly deviate from design values as the operating temperature changes. Fig. 13) depicts how the classification accuracy of the two designs vary as the operating temperatures are changed from $200K$ to $400K$. As observed by the simulation results, the two synchronous networks are resilient to variations in temperature and shows an error degradation less than 0.4% at 400K. The two asynchronous networks on the other hand are not as resilient to variations in temperature. The $1K_BT$ network display an accuracy degradation of 0.71% at 400K and 0.6% at 200K, while the $2K_BT$ network display an accuracy degradation of 2.8% at 400K and 3.2% at 200K. The higher temperature dependency of the $2K_BT$ network can be explained by the change in the switching characteristics of the MTJs at different temperatures. As illustrated by Fig. 14) the average inverter output of the $2K_BT$ magnet displays a larger shift compared to the $1K_BT$ magnet with temperature, resulting in a higher accuracy degradation.

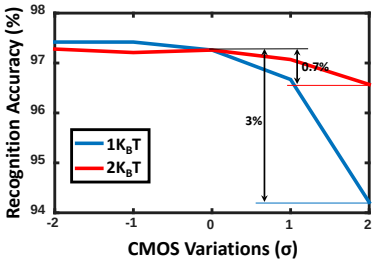

FIG. 12. Average classification accuracy for the worst case corner, with variations in the CMOS read circuit (upto $\pm 2\sigma$ variation) for the asynchronous design.

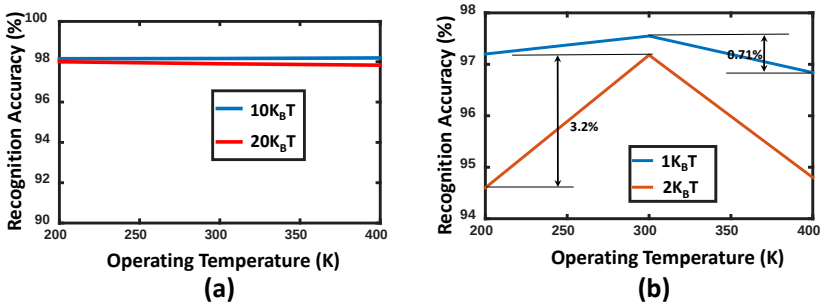

FIG. 13. Classification accuracy with varying operating temperature for the (a) synchronous, and (b) asynchronous designs.

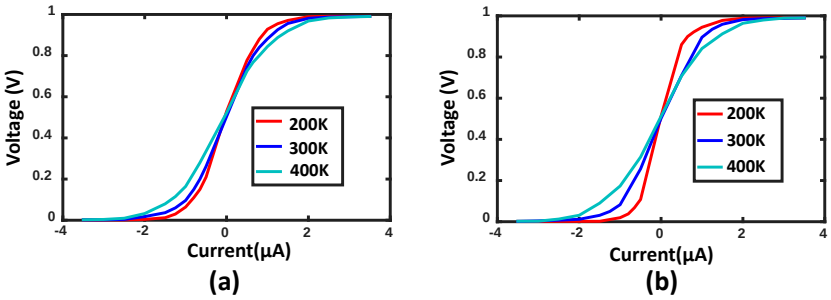

FIG. 14. Average inverter output under different temperatures for (a) $1K_BT$ magnet (b) $2K_BT$ magnet .

IV. SUMMARY

In conclusion, we outline the design considerations for MTJ based stochastic SNNs with varying barrier heights. We showed that the reduced energy consumption of low barrier height magnets is achieved at the expense of reduced error and variation tolerance and constrained design space of CMOS peripherals. We further showed that, in contrast to the popular belief that superparamagnetic MTJs would be more energy-efficient in comparison to high barrier-height magnets, parallel and always ON "read" and "write" operations in superparamagnets causes the peripheral "read" circuit energy consumption to dominate the network energy consumption profile. While scaling in the peripheral CMOS technology will reduce the peripheral energy consumption, reduced error tolerance might still be a concern for spin-based neuromorphic hardware design. The analysis performed in this work can be easily extended to other applications that require probabilistic inference, for example Bayesian networks and Ising computing.

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 (SRC), the National Science Foundation (NSF), Intel Corporation and DoD Vannevar Bush Fellowship.

[1] Duygu Kuzum, Rakesh GD Jeyasingh, Byoungil Lee, and H-S Philip Wong, "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing," Nano letters **12**, 2179–2186 (2011).

[2] Sung Hyun Jo, Ting Chang, Idongesit Ebong, Bhavitavya B Bhadviya, Pinaki Mazumder, and Wei Lu, "Nanoscale memristor device as synapse in neuromorphic systems," Nano letters **10**, 1297–1301 (2010).

[3] Abhronil Sengupta and Kaushik Roy, "A vision for all-spin neural networks: A device to system perspective," IEEE Transactions on Circuits and Systems I: Regular Papers **63**, 2267–2277 (2016).

[4] Rubén Moreno-Bote, "Poisson-like spiking in circuits with probabilistic synapses," PLoS Comput Biol **10**, e1003522 (2014).

[5] Abhronil Sengupta, Maryam Parsa, Bing Han, and Kaushik Roy, "Probabilistic deep spiking neural systems

enabled by magnetic tunnel junction," IEEE Transactions on Electron Devices **63**, 2963–2970 (2016).
[6] Gopalakrishnan Srinivasan, Abhronil Sengupta, and Kaushik Roy, "Magnetic tunnel junction based long-term short-term stochastic synapse for a spiking neural network with on-chip stdp learning," Scientific Reports **6**, 29545 (2016).
[7] Yong Shim, Akhilesh Jaiswal, and Kaushik Roy, "Ising spin model using spin-hall effect (she) induced magnetization reversal in magnetic-tunnel-junction," Journal of Applied Physics **121** (2017).
[8] Brian Sutton, Kerem Yunus Camsari, Behtash Behin-Aein, and Supriyo Datta, "Intrinsic optimization using stochastic nanomagnets," Scientific Reports **7** (2017).
[9] Steven Lequeux, Joao Sampaio, Vincent Cros, Kay Yakushiji, Akio Fukushima, Rie Matsumoto, Hitoshi Kubota, Shinji Yuasa, and Julie Grollier, "A magnetic synapse: multilevel spin-torque memristor with perpendicular anisotropy," Scientific reports **6**, 31510 (2016).
[10] Abhronil Sengupta, Priyadarshini Panda, Parami Wijesinghe, Yusung Kim, and Kaushik Roy, "Magnetic tunnel junction mimics stochastic cortical spiking neurons," Scientific Reports **6**, 30039 (2016).
[11] Adrien F Vincent, Jérôme Larroque, Nicolas Locatelli, Nesrine Ben Romdhane, Olivier Bichler, Christian Gamrat, Wei Sheng Zhao, Jacques-Olivier Klein, Sylvie Galdin-Retailleau, and Damien Querlioz, "Spin-transfer torque magnetic memory as a stochastic memristive synapse for neuromorphic systems," IEEE transactions on biomedical circuits and systems **9**, 166–174 (2015).
[12] Julie Grollier, Damien Querlioz, and Mark D Stiles, "Spintronic nanodevices for bioinspired computing," Proceedings of the IEEE **104**, 2024–2039 (2016).
[13] Gopalakrishnan Srinivasan, Abhronil Sengupta, and Kaushik Roy, "Magnetic tunnel junction enabled all-spin stochastic spiking neural network," in *2017 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (IEEE, 2017) pp. 530–535.
[14] Luqiao Liu, Chi-Feng Pai, Y Li, HW Tseng, DC Ralph, and RA Buhrman, "Spin-torque switching with the giant spin hall effect of tantalum," Science **336**, 555–558 (2012).
[15] Chi-Feng Pai, Luqiao Liu, Y Li, HW Tseng, DC Ralph, and RA Buhrman, "Spin transfer torque devices utilizing the giant spin hall effect of tungsten," Applied Physics Letters **101**, 122404 (2012).
[16] Akhilesh Jaiswal, Xuanyao Fong, and Kaushik Roy, "Comprehensive scaling analysis of current induced switching in magnetic memories based on in-plane and perpendicular anisotropies," IEEE Journal on Emerging and Selected Topics in Circuits and Systems **6**, 120–133 (2016).
[17] Jiang Xiao, Andrew Zangwill, and Mark D Stiles, "Boltzmann test of slonczewskis theory of spin-transfer torque," Physical Review B **70**, 172405 (2004).
[18] Sasikanth Manipatruni, Dmitri E Nikonov, and Ian A Young, "Energy-delay performance of giant spin hall effect switching for dense magnetic memory," Applied Physics Express **7**, 103001 (2014).
[19] William Fuller Brown Jr, "Thermal fluctuations of a single-domain particle," Physical Review **130**, 1677 (1963).
[20] Werner Scholz, Thomas Schrefl, and Josef Fidler, "Micromagnetic simulation of thermally activated switching in fine particles," Journal of Magnetism and Magnetic Materials **233**, 296–304 (2001).
[21] Amikam Aharoni, "Demagnetizing factors for rectangular ferromagnetic prisms," Journal of applied physics **83**, 3432–3434 (1998).
[22] L Lopez-Diaz, L Torres, and E Moro, "Transition from ferromagnetism to superparamagnetism on the nanosecond time scale," Physical Review B **65**, 224406 (2002).
[23] Xuanyao Fong, Sumeet K Gupta, Niladri N Mojumder, Sri Harsha Choday, Charles Augustine, and Kaushik 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.
[24] Abhronil Sengupta, Aparajita Banerjee, and Kaushik Roy, "Hybrid spintronic-cmos spiking neural network with on-chip learning: Devices, circuits, and systems," Physical Review Applied **6**, 064003 (2016).
[25] Mirko Prezioso, Farnood Merrikh-Bayat, BD Hoskins, GC Adam, Konstantin K Likharev, and Dmitri B Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," Nature **521**, 61–64 (2015).
[26] Rasmus Berg Palm, "Prediction as a candidate for learning deep hierarchical models of data," Technical University of Denmark **5** (2012).
[27] Michael A Nielsen, "Neural networks and deep learning," (2015).
[28] Anil K Jain, Jianchang Mao, and K Moidin Mohiuddin, "Artificial neural networks: A tutorial," Computer **29**, 31–44 (1996).
[29] Hamid Soleimani, Arash Ahmadi, and Mohammad Bavandpour, "Biologically inspired spiking neurons: Piecewise linear models and digital implementation," IEEE Transactions on Circuits and Systems I: Regular Papers **59**, 2991–3004 (2012).
[30] Paul A Merolla, John V Arthur, Rodrigo Alvarez-Icaza, Andrew S Cassidy, Jun Sawada, Filipp Akopyan, Bryan L Jackson, Nabil Imam, Chen Guo, Yutaka Nakamura, *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," Science **345**, 668–673 (2014).
[31] See Supplemental Material at [URL] for more information about the structure of the simulated network , the training methodology, and a brief introduction to neural networks.