Quantitative Evaluation of Hardware Binary Stochastic Neurons

Orchi Hassan,[1,*] Supriyo Datta,[2] and Kerem Y. Camsari[3]

[1] *Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh*

[2] *School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana 47906, USA*

[3] *Department of Electrical and Computer Engineering, University of California, Santa Barbara, Santa Barbara, CA 93106, USA*

(Received 15 January 2021; revised 12 April 2021; accepted 24 May 2021; published 17 June 2021)

Recently, there has been increasing activity to build dedicated Ising machines to accelerate the solution of combinatorial optimization problems by expressing these problems as a ground-state search of the Ising model. A common theme of such Ising machines is to tailor the physics of the underlying hardware to the mathematics of the Ising model to improve some aspect of performance that is measured in speed to solution, energy consumption per solution, or area footprint of the adopted hardware. One such approach to build an Ising spin, or a binary stochastic neuron (BSN), is a compact mixed-signal unit based on a low-barrier nanomagnet-based design that uses a single magnetic tunnel junction (MTJ) and three transistors (3T-1MTJ design) where the MTJ functions as a stochastic resistor (1SR). Such a compact unit can drastically reduce the area footprint of BSNs while promising massive scalability by leveraging the existing magnetic RAM technology that has integrated 1T-1MTJ cells in approximate Gbit densities. The 3T-1SR design however can be realized using different materials or devices that provide naturally fluctuating resistances. Extending previous work, we evaluate hardware BSNs from this general perspective by classifying necessary and sufficient conditions to design a fast and energy-efficient BSN that can be used in scaled Ising machine implementations. We connect our device analysis to systems-level metrics by emphasizing hardware-independent figures of merit such as *flips per second* and dissipated *energy per random bit* that can be used to classify any Ising machine.

DOI: 10.1103/PhysRevApplied.15.064046

I. INTRODUCTION

In the era of the internet of things, combinatorial optimization problems are ubiquitous [1]. In fact, most of the real problems that quantum computers are aiming to solve can be formulated as combinatorial optimization problems. From directing traffic flow [2] to routing interconnections in integrated circuit design [3,4], to making financial decisions [5], drug discoveries [6], etc.—all involve solving a form of combinatorial optimization problems. The demand for solving these problems faster and more efficiently is ever increasing. But such problems typically fall into the NP-hard or NP-complete class in complexity theory [7], with no known polynomial time solution, making them notoriously difficult to solve in digital computers using traditional computing methods. This has given rise to an alternative paradigm in computing, namely Ising computing. Ising computing maps combinatorial optimization problems to an Ising model, and solves it by searching for the ground state of the system described by [8,9]

$$E = -I_0\left(\frac{1}{2}\sum_{i,j=1}^{N} J_{ij}\, m_i m_j + \sum_{i=1}^{N} h_i m_i\right), \tag{1}$$

where m denotes the Ising spin, J is the coupling coefficient, h is the external bias, and I_0 is the annealing parameter that is proportional to the inverse of the temperature. In the machine learning field, the same underlying principle is used for Boltzmann machines with the annealing parameter being 1. The binary stochastic neurons (BSNs) [10] of stochastic neural networks are well suited to function as a "spin" is such systems, described mathematically by

$$m_i = \mathrm{sgn}[\tanh(I_i) - r_i], \tag{2}$$

where r_i is a random number between $+1$ and -1, and $I_i = -\partial E/\partial m_i$ is the input to the neuron.

Given the importance of optimization problems, a lot of research has gone into developing algorithms and identifying appropriate hardware for Ising computing. Various approaches including quantum computers based on quantum annealing or adiabatic quantum optimization implemented with superconducting circuits [11],

[*] orchi@eee.buet.ac.bd

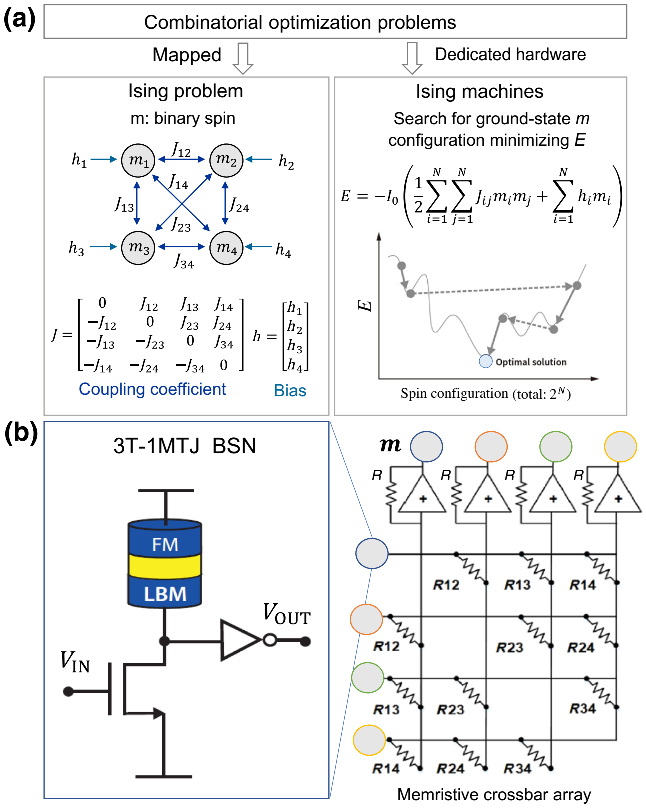

(a)
Combinatorial optimization problems
Mapped
Dedicated hardware
Ising problem
m: binary spin
h_1 h_2 h_3 h_4
m_1 m_2 m_3 m_4
J_{12} J_{14} J_{13} J_{24} J_{23} J_{34}
$J = \begin{bmatrix} 0 & J_{12} & J_{13} & J_{14} \\ -J_{12} & 0 & J_{23} & J_{24} \\ -J_{13} & -J_{23} & 0 & J_{34} \\ -J_{14} & -J_{24} & -J_{34} & 0 \end{bmatrix}$ $h = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix}$
Coupling coefficient
Bias
Ising machines
Search for ground-state m configuration minimizing E
$E = -I_0\left(\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij}m_im_j + \sum_{i=1}^{N} h_im_i\right)$
E
Optimal solution
Spin configuration (total: 2^N)
(b)
3T-1MTJ BSN
FM
LBM
$V_{\rm OUT}$
$V_{\rm IN}$
$\boldsymbol{m}$
R
R12 R13 R14 R23 R24 R34
Memristive crossbar array

FIG. 1. The 3T-1MTJ compact BSN hardware that utilizes the natural physics of low-barrier nanomagnets holds the promise to accelerate the simulated annealing processors. (a) The underlying working principle of Ising machines. (b) An implementation scheme utilizing MTJ and memristive crossbar arrays, where the BSN is the Ising spin m_i, memristors (R_{ij}) implement the weight and bias coefficients, and the feedback resistor R can control the annealing temperature electrically.

coherent Ising machines implemented with laser pulses [12], phase-change oscillators [13], or CMOS oscillators [14–17] and digital annealers based on simulated annealing [18] implemented with digital circuits [1,19–24] are being explored.

In this paper we comprehensively evaluate and characterize a stochastic magnetic tunnel junction (SMTJ) based realization of the Ising spin [Eq. (2)] where random numbers are generated using the natural physics of low-barrier nanomagnets [25] in a compact design. A network of these BSN units can be coupled with a memristive crossbar array [26–28] to perform the synaptic operation shown in Fig. 1, drastically improving the area requirements and accelerating the computation speed of Ising machines. We evaluate the performance of the BSN device in terms of its energy and delay metrics and connect these to the problem and substrate-independent metric of *flips per second* that the probabilistic system makes [29].

Our evaluation of the 3T-1MTJ BSN design considers different types of low-barrier nanomagnet realizations of MTJs. As the MTJ essentially functions as a two-terminal stochastic resistor (SR), we take a general 3T-1SR design approach, classifying necessary and sufficient conditions for achieving the BSN operation for different types of SR in Sec. II. We relate these conditions to the different SMTJ realizations in Sec. III. We report the timescale of operation, power, and energy for each case based on benchmarked SPICE simulations of the BSN hardware consisting of spintronic elements from a modular circuit framework [30] coupled to 14 nm FinFET predictive technology models [31], and provide analytical results for relevant quantities in Sec. IV. Lastly, we use these device performance metrics to project onto hardware performance figures of merit such as flips per second that a probabilistic sampler makes. Our projections indicate orders of magnitude improvement potential over current digital implementations.

II. GENERAL APPROACH TO THE DESIGN OF A BSN

BSNs are well suited to function as a “spin” in Ising machines for solving combinatorial optimization problems [10,32]. A compact and efficient hardware realization of the BSN leveraging the natural physics of stochastic nanomagnets can be made by using unstable MTJs [33–37], as shown in Fig. 1.

The compact design of a BSN based on low-barrier magnet (LBM) SMTJs was proposed in 2017 [25]. Using magnet and circuit physics to analyze the performance, it was reported that using a LBM in a circular disk geometry with energy barriers below k_BT as the free layer of a MTJ results in subnanosecond response times requiring only a few femtojoules of energy per random bit [32]. The proposed design and the performance analysis considers a very specific type of SMTJ that had circular in-plane magnetic anisotropy (IMA) whose fluctuations are undisturbed by the current in the circuit for typical current drive conditions. However, in 2019, a version of the BSN design that was implemented in hardware to solve an 8-bit factorization problem [38] consisted of a SMTJ with perpendicular anisotropy (PMA) and a barrier of a few k_BT as its free layer. Unlike the circular in-plane design, the PMA design relied on its resistance being tunable by the spin-transfer-torque effect in order to achieve the BSN operation. This has called for an extension of our initial analysis presented in Ref. [32] that we systematically perform in this paper.

As the MTJs in the BSN circuit effectively act as a fluctuating resistor R [39] and the design principle is independent of this realization, for establishing the fundamental design rules, we approach it from a general perspective and we hope that these design rules stimulate discussion in the realization of different stochastic resistors that use different mechanisms [40–45].

A. Types of fluctuating resistance

We categorize the fluctuating R into four types. Based on the fluctuating nature, it can be continuous or bipolar (telegraphic). Second, it can be tunable or nontunable depending on whether it is affected by the current that is flowing through it.

A continuous resistor can have any resistance between $[R_P \rightarrow R_{AP}]$, while a bipolar resistor only assumes the two values R_P and R_{AP}, as shown in Fig. 2(a). The distribution of continuous resistances can be of different types as well. It can be uniform or follow the slightly bimodal distribution in the case of a MTJ, as shown in the figure. Different distributions typically result in different average R values, slightly bimodal or uniform distributions are better suited for BSN realizations than Gaussian distributions.

The current I flowing in the circuit can tune the probability distribution of the resistance fluctuations, and we call such resistors tunable resistors. When designing a BSN with current tunable R, we need to know the current where fluctuations are equal between the two extreme states (I_{50}) [39] and the current required to pin the resistance to one of those states. An important parameter in this case is the bias current I_0, which is the slope of the R versus I curve at the 50:50 point. Typically, about $3I_0$–$5I_0$ current is required to pin the fluctuating resistance to one of its states. In Sec. III 2 (see Fig. 10) we provide analytical expressions for I_0 for four cases of resistors that can be obtained by various MTJs.

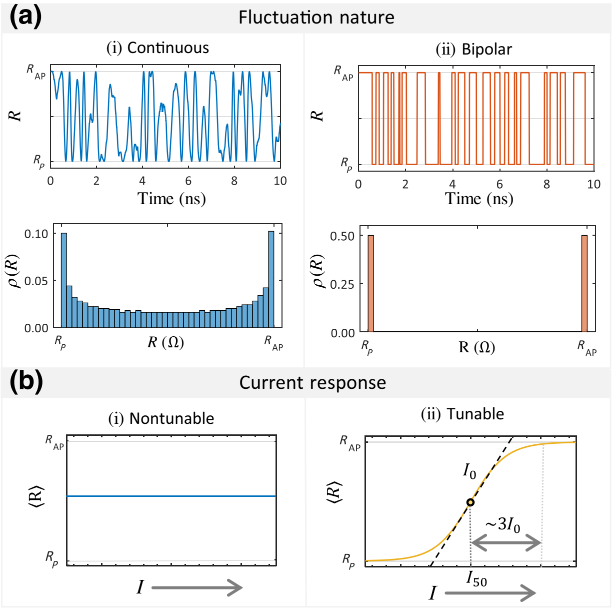


FIG. 2. Categorizing resistances: (a) Fluctuating nature: they can be continuous or bipolar. The time dynamics and distribution are shown for each category. (b) Current tunability: the fluctuations could be unaffected by I or they could be a function of I as indicated by their transfer characteristics. Here I_{50} is the current at the 50:50 point where the resistance spends equal time in the R_P and R_{AP} states, and I_0 is the biasing current defined as the slope of the (R versus I) curve at the 50:50 point. The pinning current is typically about $3I_0$–$5I_0$.

Based on this analysis, we categorize the fluctuating resistance into four types: nontunable continuous (NTC), nontunable bipolar (NTB), tunable continuous (TC), and tunable bipolar (TB).

B. Performing the BSN function

We take a look at the transfer characteristics of the device to see whether the four types of resistance can faithfully mimic the BSN operation described by Eq. (2). The fluctuating R is a physical realization of the random variable r_i, the NMOS acts as a constant current source that provides tunability, and the inverter performs the sgn operation in Eq. (2).

Figure 3 shows that, while all other resistance types are able to reproduce the desired sigmoidal average curve $\langle m_i \rangle = \tanh(I_i)$, the nontunable bipolar resistor gives a staircaselike function instead. This is because of the fixed delta functionlike resistance distribution at the two extreme states [see Fig. 2(a)(ii)]. As there is no continuity in the resistance distribution and no additional means of tuning the delta distribution itself has been introduced to the structure, the BSN output fluctuations are equal until either of the threshold points are crossed, resulting in the staircaselike function.

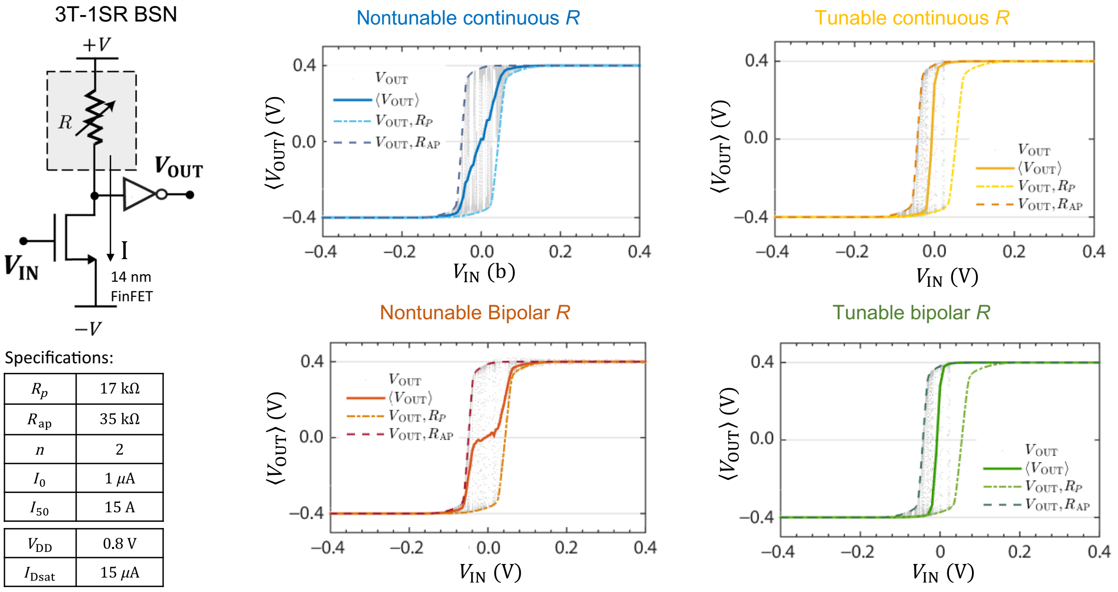
Mathematically, when the resistance is bipolar, then r_i is ± 1. Thus, for any input I_i where $|\tanh(I_i)| < 1$, the output $\langle m \rangle$ is equal to zero. In Fig. 4(b), if we look at a simple invertible AND gate [25,48] operation, it is seen that devices with a staircaselike function like this are not suitable for performing as BSNs. This has been demonstrated experimentally in Refs. [49,50] where a stable MTJ was used as a bipolar resistor whose distribution was tuned by an external field. However, this issue could be resolved by introducing external or additional control parameters such as the external field shown in the same experiment.

C. Parameter dependence and design choices

Figure 3 is created with a fixed set of parameters for the resistor and coupled with a specific transistor technology, 14 nm FinFET models. In this section we explore how the transfer characteristics are affected by different parameters of the resistors and FET characteristics and how to choose the right combination of R and FET to be coupled.

1. Stochastic region

The stochastic region, which we define next, is a function of the resistance ratio n for nontunable resistors and biasing current I_0 for tunable resistors as shown in Fig. 5, which needs to be matched with the transistor characteristics.



Specifications:

R_p	17 kΩ
R_{ap}	35 kΩ
n	2
I_0	1 μA
I_{50}	15 A
V_{DD}	0.8 V
I_{Dsat}	15 μA

FIG. 3. Transfer characteristics: The BSN circuit is realized by coupling the fluctuating resistor that is the physical realization of the random variable r_i in the BSN equation to an NMOS that provides the tunability, and then to an inverter that thresholds the output. The four types of resistance are coupled to a 14 nm FinFET and the resistance parameters (based on experimental demonstrations of MTJs [47]) are chosen to match the transistor characteristics. All resistance types except for the bipolar nontunable are able to achieve the BSN operation following Eq. (2). To function as a BSN, the bipolar resistances need some means of tuning their probability distribution.

2. Effect of *n*

The resistance ratio $n = R_P/R_{AP}$ is directly related to the stochastic region Δv through the NMOS characteristics in the case of nontunable resistor designs. The edge of the stochastic region $v^{\pm}$ is defined by $V_i = V_{DD}/2 - [I^+R_P, I^-R_{AP}] \approx 0$, where the currents $I^{\pm}$ are determined by the NMOS, as shown in Fig. 6(c). For a desired $\Delta v = v^+ - v^-$ (stochastic region) and NMOS transistor, the required $n = R_{AP}/R_P$ should approximately equal I^+/I^-. Ideally, the minimum value of the resistance should be

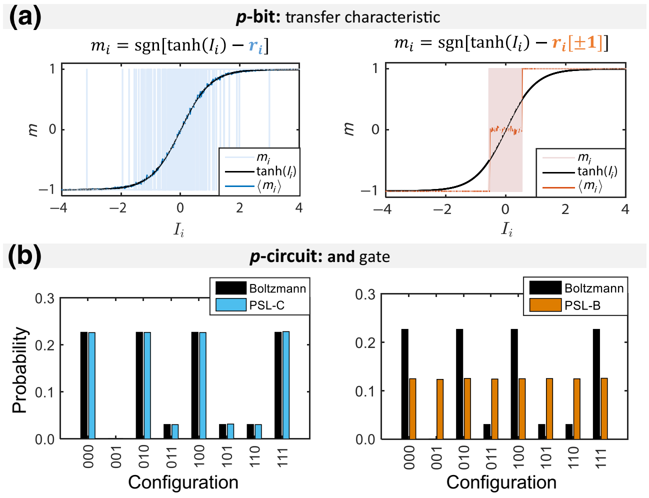


FIG. 4. Nontunable continuous versus bipolar resistance: (a) Transfer Characteristics shows that while the continuous resistor results in a sigmoidal output, the bipolar gives a stair-case like function. (b) The bipolar R is unable to follow the Boltzmann distribution of the invertible AND gate (description in Ref. [48]). All states remain equally probable.

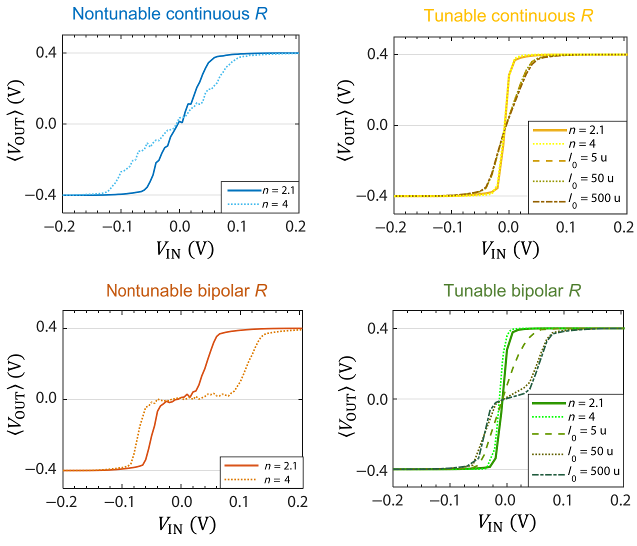


FIG. 5. Effect of n and I_0: The stochastic region of the nontunable resistances is determined by the resistance ratio $n = R_P/R_{AP}$, while the biasing current I_0 of tunable resistances controls the stochastic region. For large biasing currents, the tunable resistors behave effectively like nontunable resistances.

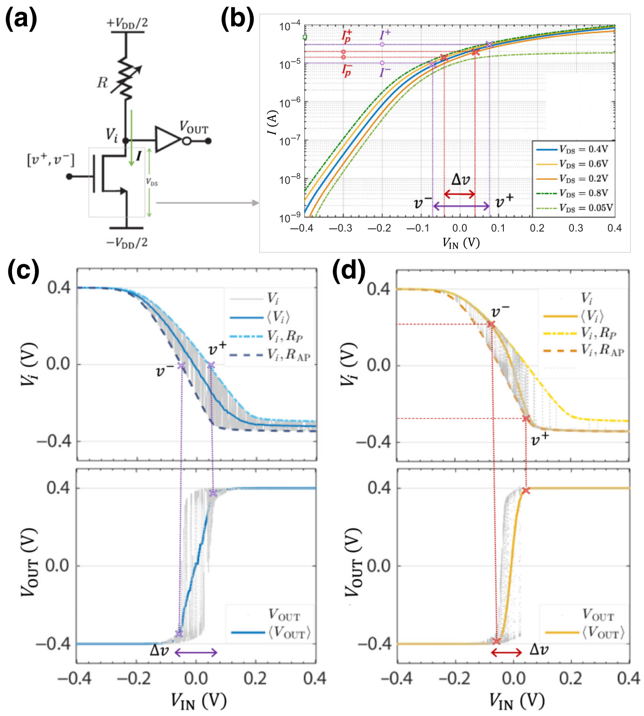


FIG. 6. Stochastic region boundaries: The stochastic region boundaries $[v^+, v^-]$ are set by different parameters for tunable and nontunable resistors. (a) The BSN circuit with (b) the current transfer characteristics of the 14 nm FinFET NMOS when $V_i \sim 0V$. (c) Nontunable R: in this case the boundaries are set when $V_i \approx 0$ and the resistance ratio $n = R_{AP}/R_P \approx I^+/I^-$. (d) Tunable R: the stochastic range is determined by pinning current I_P characteristics of the resistance. The transfer characteristics of each stage in (c) and (d) indicate the stochastic range v^+ and v^- and the relation to the NMOS characteristics in each case in (b).

$R_P = (V_{DD}/2)/I^+$ and to get full pinning, Δv should be less than V_{DD}. For a 14 nm FinFET, to get a stochastic region of $\Delta v = 50 - 200$ mV, the resistance ratio n should be around 2–50. The resistance ratio n is a measure of tunneling magnetoresistance (TMR) [$= (n-1) \times 100\%$] in the case of MTJs. For the nontunable case, TMR needs to be large enough to provide a voltage swing large enough to overcome the noise margins of the inverter [32], and it should be small enough so that output pinning is achieved within the given input range. Typically, MTJs have TMRs in the range 100%–300% [51] with a maximum reported TMR of 604% [52], so the resistance ratio of MTJs is well within the desired range, but the general requirements we outline should be applicable for other types of stochastic resistor as well.

3. Effect of I_0

In the case of tunable resistances, the stochastic region is independent of the resistance ratio and depends on the pinning current and thus the bias current ($I_P^{\pm} \propto I_0$) instead, as shown in Fig. 6(d). For large bias currents ($I_0 \gg I$), the tunable resistances act essentially like nontunable resistances. To get the full range of R, the NMOS needs to be able to supply the pinning current. If the pinning current is $3I_0$–$5I_0$, as shown in Fig. 2, then to get the full range of the resistance, $I_{P\max}^+$ needs to be around about $6I_0$–$10I_0$. In the case of 14 nm FinFETs, $I_{\max}^+$ is around about 40 μA, restricting I_0 to values less than 7 μA.

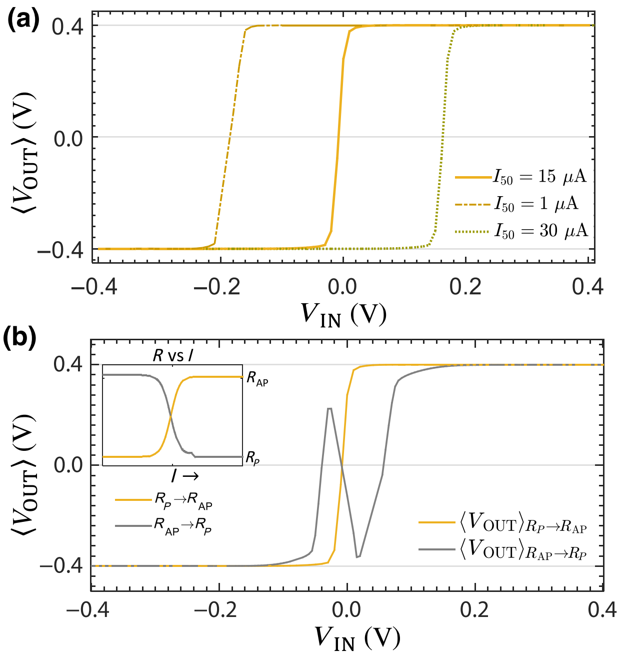


FIG. 7. (a) Choice of I_{50}: I_{50} is ideally a positive quantity matched with the $I_{D\text{sat}}$ of the transistor; changing I_{50} results in a lateral shift of the sigmoid. (b) The R versus I relationship: the output characteristics also depend on the nature of the resistance tunability with the circuit current I. If R decreases with I ($R_{AP} \rightarrow R_P$), the opposing characteristics of the transistor current and resistance change result in a nonmonotonic output.

4. Choice of I_{50}

Another parameter that is important for the operation of tunable resistors is I_{50}, which determines the midpoint of the sigmoid. It is the current at which the resistance on average spends equal time in the R_P and R_{AP} states [39]. As the circuit can only support positive current values, it needs to be a positive quantity and preferably matched with the saturation point ($V_{DS} = V_{GS}$) current $I_{D\text{sat}}$ of the NMOS transistor. Changing I_{50} shifts the transfer characteristics laterally, as shown in Fig. 7(a).

5. R versus I

One last requirement is that, for current tunable resistance with increasing current I, the resistance needs to increase from $R_P \rightarrow R_{AP}$. This can be understood intuitively: increasing I means that the NMOS transistor is becoming more conductive. If the MTJ concomitantly becomes more conductive as I increases, the transfer characteristics can show nonmonotonic behavior, as shown in Fig. 7(b). This requirement holds true irrespective of whether the circuit's R branch consists of a PMOS-1R or 1R-NMOS topology.

III. REALIZATION OF FLUCTUATING RESISTANCES WITH STOCHASTIC MAGNETIC TUNNEL JUNCTIONS

A MTJ whose free layer is a LBM could serve as a physical realization of fluctuating resistors. Depending on the nature and characteristics of the LBM magnetization fluctuations, we can get different types of R. Our previous analysis [32] was restricted to one type of LBM, the circular IMA with barrier less than k_BT; in this section we extend it to include all possible LBMs.

A general description of the energy associated with a magnet is given by [32]

$$E = \tfrac{1}{2}H_{kp}M_s\Omega(1-m_x^2) + \tfrac{1}{2}H_{ki}M_s\Omega(1-m_z^2) - \hat{H}_{\text{ext}}M_s\Omega\cdot\hat{m}, \qquad (3)$$

where $H_{kp} = 2K_s/t - 4\pi M_s$ is the perpendicular anisotropy field along the x axis, K_s is the surface anisotropy density, H_{ki} is the in-plane anisotropy along the z axis, H_{ext} is the external field, M_s is the saturation magnetization, and $\Omega = \pi(D/2)^2t$ is the volume of the magnet. By adjusting the thickness or the shape of the magnet, the magnetic anisotropy of the magnet can be scaled to behave like a low-barrier magnet [32,53]. Second-order magnetic anisotropy effects and in-plane components of demagnetization fields have not been considered here and left for a future investigation since the macroscopic model without it seems to be reasonably consistent with recent experimental results involving low-barrier magnets [39,54–56]. We use the stochastic Landau-Lifshitz-Gilbert (LLG) module fro our spintronics library [57] to simulate the LBM dynamics in HSPICE using its transient noise function. This model has been carefully benchmarked against general Fokker-Planck-based methods [30].

A. LBM fluctuation dynamics

By low-barrier magnets we refer to magnets whose barrier is less than $10k_BT$ or so, whose magnetization fluctuates randomly in the presence of thermal noise. Interestingly, the magnetization dynamics of low-barrier magnets with barrier less than k_BT are different from those with a slightly higher barrier [32,58]. The simple exponential dependence of the retention time of the magnetization state on the barrier height is not valid around or below k_BT [46].

Figure 8 shows the fluctuation dynamics, the magnetization distribution, and the auto-correlation time (τ_{CORR}) for low-barrier magnets. Magnetization fluctuations translate into resistance fluctuations in the MTJ, and we see that magnets with barrier less than k_BT act like continuous resistances, while slightly higher barrier magnets, which have more defined two states, give telegraphic fluctuations, and in both cases IMA magnets fluctuate orders of magnitude faster than their PMA counterparts due to a mechanism where the demagnetization field plays a central role [32,55,58–60].

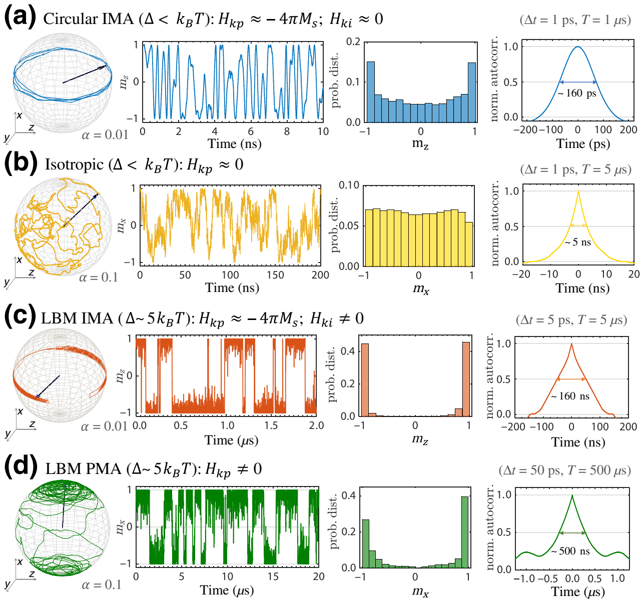


FIG. 8. Low-barrier magnet fluctuation dynamics: We use the benchmarked stochastic LLG module to simulate LBM dynamics. The saturation magnetization is considered to be $M_s = 1000$ emu/cc, $\Omega = 6.3\times10^{-19}$ cc, and H_k adjusted to get the indicated Δ. Each simulation is carried out with a time step at least 100 times smaller for a time duration 1000 times greater than characteristic timescales to avoid any simulation time dependencies; the exact parameters are indicated. Magnets with $\Delta < k_BT$ have more continuous fluctuations, with (b) having a more uniform distribution than (a), while slightly higher barrier magnets have a more telegraphic fluctuation. In both cases, the presence of high demagnetization fields cause faster fluctuations in IMA magnets.

B. Current response of LBMs

Magnetic fluctuations can be tuned by spin current. For high barrier magnets, the minimum current required to switch the magnetization is called the critical current [61]; in the case of low-barrier magnets, we refer to it as a biasing current, defined by the inverse of the derivative taken at $\langle m\rangle = 0$, mathematically expressed as $I_0 = (\langle m\rangle/I_S)^{-1}$ at low bias (I_S). The current required to pin the magnetization, similar to switching current in high-barrier magnets, is assumed to be about $3I_0$–$5I_0$, as indicated in Fig. 2. IMA magnets have a much larger pinning current than PMA magnets because of the large demagnetization field present due to their disk shape [32,61,62], meaning that transistors with much larger current ranges would be required for

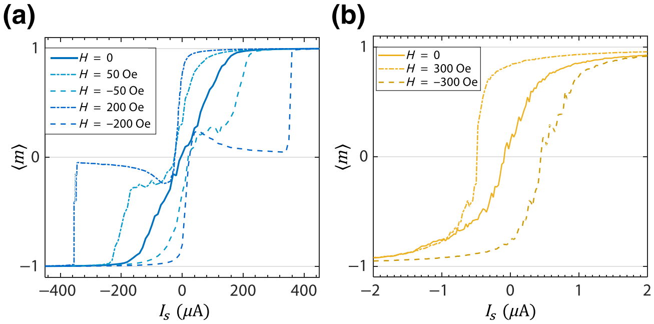


FIG. 9. Current response: LBM response to the spin current with and without external fields for a (a) circular IMA magnet ($H_{ki} \sim 0, H_{kp} \sim -H_D$) and (b) isotropic anisotropy magnet ($H_{kp} \sim 0$). Each point on the curve is a long-time ($T = 1$ μs, $\Delta t = 1$ ps) average magnetization from our benchmarked stochastic LLG module. The critical field for the IMA magnet is about 130 Oe and about 200 Oe for the isotropic magnet.

IMA magnet MTJs than PMA magnet MTJs for tunable resistors.

An important thing to note here is the current tunability in the presence of an external field that can arise, for example, due to the fixed, stable layer that acts as a reference to the free layer in the MTJ. In the case of high-barrier magnets, the spin-current-induced magnetic switching hysteresis loop just shifts for PMA magnets depending on the direction of the field, but for IMA magnets, the shape of the hysteresis and magnet dynamics are changed [61]. The large demagnetizing field present perpendicular to the magnetization plane in IMA magnets causes the magnetization to precess around it when spin current is applied in the opposite direction to the external field. The same is observed in low-barrier magnets, as shown in Fig. 9. The larger the external field, the more pronounced the effect. The uniform precessional motion kicks in at high field, when the current is close to the biasing current or higher, applied in the opposite direction to the field. Recently, this has been observed experimentally for low fields [55]. While this is an undesired effect in the case of our BSN operation, this can be useful in the context of oscillator-based networks [63].

This has important implications in terms of acting as a fluctuating resistance in a BSN circuit. IMA magnets with external fields (i.e., uncompensated dipolar fields in MTJ [64]) greater than their pinning field are not suited to function as tunable or nontunable resistors. IMA magnets with continuous magnetization coupled to a transistor with small saturation current (tens of microampere) compared to the biasing current of IMA (hundreds of microampere) can work as nontunable resistors, and as experimental observations in Ref. [55] suggest, can withstand small (compared to their pinning field) stray fields.

PMA magnet MTJs with their small biasing current (about a few to a few tens of microampere) when coupled to typical transistors act as tunable resistors in the BSN circuit. In this case the external bias field is actually preferred, since this enables positive I_{50} current [38].

Thus, if we coupled a MTJ with a 14 nm FinFET ($V_{DD} = 0.8$ and $I_{D\text{sat}} = 15$ μA) [31], the table in Fig. 10 summarizes the resistance mapping and the associated parameters.

IV. PERFORMANCE EVALUATION OF A MTJ-BASED BSN

In this section we compare the physical performance of these different SMTJs in a BSN.

A. Timescale of operation

The two relevant timescales of operation for a BSN are the correlation time τ_C, which is the average time it takes to

R type	MTJ free layer	τ_{CORR}	I_0	I_{50}	H_0
Nontunable continuous	$\Delta < k_B T$ Circular IMA	$\sqrt{8\ln(2)}\,\frac{1}{\gamma}\sqrt{\frac{M_s\Omega}{H_D k_B T}}$ ($\alpha < 0.1$) (sub-ns)	$\frac{2q}{\hbar}\sqrt{\frac{2}{\pi}}\sqrt{H_D M_s \Omega k_B T}$ ($\alpha < 0.1$) (0.1~1 mA)	(n/a)	$\frac{2k_BT}{M_s\Omega}$
Tunable continuous	$\Delta < k_B T$ Isotropic ‘PMA’	$\ln(2)\,\frac{1}{\gamma}\frac{M_s\Omega}{\alpha k_B T}$ ~10 ns	$\frac{6q}{\hbar}\alpha k_B T$ (0.4~4 μA)	$\frac{4q\alpha}{\hbar}\left(\frac{1}{2}H_{\text{ext}}M_s\Omega\right)$	$\frac{3k_BT}{M_s\Omega}$
Nontunable bipolar	$2k_BT < \Delta < 10k_BT$ IMA	$\propto \frac{e^{\Delta/k_BT}}{(1+H_D/2H_k)}$ 1 ns ~ 1 μs	$\frac{4q\alpha}{\hbar}\Delta\left(1+\frac{H_D}{2H_K}\right)$ (0.05~25 mA)	(n/a)	$\sim H_k$
Tunable bipolar	$2k_BT < \Delta < 10k_BT$ PMA	$\propto e^{\Delta/k_BT}$ 0.1~ 100 μs	$\frac{4q\alpha}{\hbar}\Delta$ (0.5~25 μA)	$\frac{4q\alpha}{\hbar}\left(\frac{1}{2}H_{\text{ext}}M_s\Omega\right)$	$\sim H_k$

FIG. 10. MTJ free layer and its corresponding *R* type along with corresponding characteristic parameters and their analytical expressions. The numbers in brackets indicate an approximate range of values for each parameter. The proportionality constant for the correlation time of magnets with $\Delta > k_B T$ is $\tau_0 \sim 0.1$–1 ns; see Ref. [46] for the exact equation.

produce a output at a given input, and the response time τ_N, which is defined as the average time it takes for the circuit to give a random output with the correct statistics as the input is changed [32]. Figure 11 shows the two timescales for the three types of fluctuating resistance for MTJs with two different timescales. For simplicity, we assume that the correlation time is the same for all types of magnet, but in reality they would follow the τ_{CORR} relations indicated in Fig. 10 [32,58].

Figure 11 shows that the response time τ_N for the nontunable resistor is independent of the fluctuation time of the resistance—it is rather proportional to the RC delay of the circuit—while for the tunable cases, the response time is related to the characteristic timescales of the resistor. But the time to give numbers or flip rate τ_C at $V_{\mathrm{IN}} = 0$ is entirely resistance fluctuation time dependent for all cases ($\tau_C \approx \tau_{\mathrm{CORR}}$). Thus, for the tunable case, the two said timescales of operation are likely to be similar as they are governed by the magnet fluctuation characteristics, while for the nontunable case, the response time that is RC dependent has the potential to be very short compared to the magnet-dependent correlation time. For most applications, this difference may not be of importance but, for some applications where the network is directed, like Bayesian inference, having two different timescales seems to be a requisite [65].

B. Power

Our SPICE simulations indicate that the average power consumed by the BSN circuit in its stochastic region is $\langle P\rangle \approx 2 \times V_{DD}I_{D\mathrm{sat}}$ [32]. The factor 2 is for the two branches: the MTJ branch and the inverter branch. This holds for all types of resistor. For a 14 nm FinFET with $V_{DD} = 0.8V$ and $I_{D\mathrm{sat}} \sim 1\ \mu$A, $\langle P\rangle \sim 20\ \mu$W. While the power is almost independent of TMR or the resistance ratio (n) for a set 50:50 point and technology for the MTJ branch, its joule heating increases with increasing TMR (approximately proportional to $\sqrt{n}$) in the positive pinning region as the NMOS resistance reduces. Thus, the lowest TMR that ensures a voltage swing V_i greater than the noise margin of the inverter is considered best suited for the BSN operation. The MTJ branch power could be reduced by operating in subthreshold region $I_{D\mathrm{sub}} \sim 1\ \mu$A, but this reduces the total power by 0.5 times while trading-off with a 10 times increase in the RC response time. Given the flexibility, it is preferable to design the MTJ to operate in the saturation region of the transistor. For the tunable case, this means matching $I_{50} \sim I_{D\mathrm{sat}}$; for the nontunable case, this means having $\langle R\rangle \approx (V_{DD}/2)/I_{D\mathrm{sat}}$.

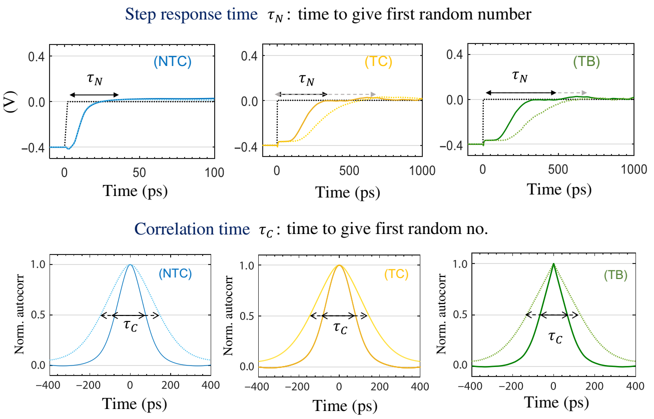


FIG. 11. Timescales of operation for each resistor type for two fluctuation times $\tau_C \sim [160\ \mathrm{ps}, 320\ \mathrm{ps}]$ are shown. The resistances are engineered to have similar characteristic timescales but different fluctuation behavior (tunable, nontunable, and continuous and bipolar fluctuations) for comparison purposes.

C. Energy

As there are two timescales associated with the BSN operation, we can define two energies; the energy required to give the first random number after the input has changed, $E_N \sim \tau_N\langle P\rangle$ and the energy required to produce new random numbers at a given input state, $E_C = \tau_C\langle P\rangle$. Fig. 12(a) shows the energy delay plot indicating the parameter ranges for each type of SMTJs. When describing the performance of a hardware BSN, we generally refer to the correlation time τ_C for delay and E_C for energy. The energy-delay numbers for a single BSN device operation can be used to project the system-level performance parameters for Ising processors built with them.

D. Hardware projections

Typically, the performance of an Ising hardware is measured in terms of the time and energy it takes to solve a specific problem. The time to solution depends not only

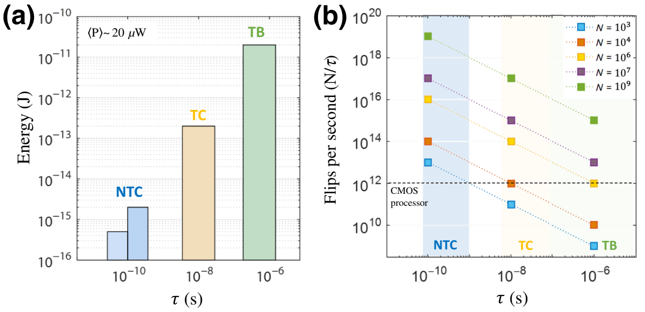
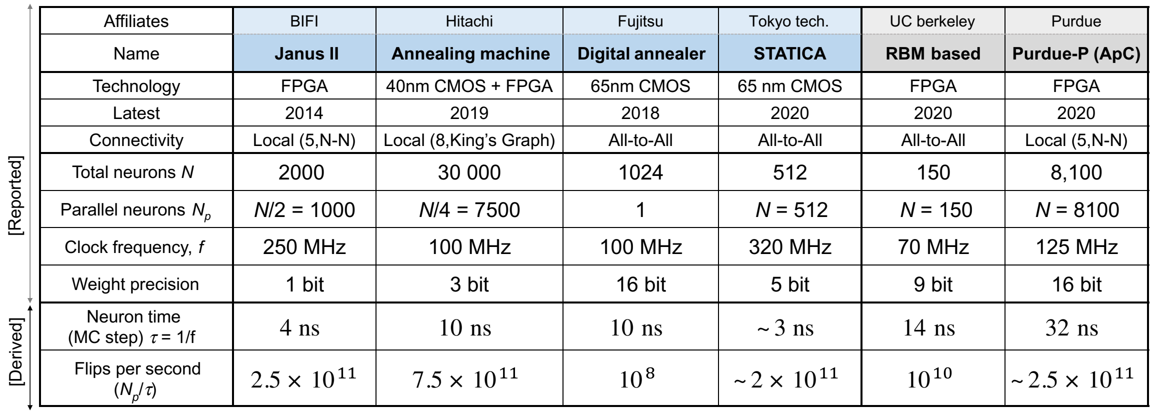


FIG. 12. (a) Energy delay of each type of MTJ-based BSN, assuming an average power of 20 μW and the timescales in Fig. 11. (b) Flips-per-second projections for different numbers of neurons for each type of MTJ. For these projections, only BSN performance numbers are used; the synapse would add to the power and thus energy per flip number.



	Affiliates	BIFI	Hitachi	Fujitsu	Tokyo tech.	UC berkeley	Purdue
[Reported]	Name	**Janus II**	**Annealing machine**	**Digital annealer**	**STATICA**	**RBM based**	**Purdue-P (ApC)**
	Technology	FPGA	40nm CMOS + FPGA	65nm CMOS	65 nm CMOS	FPGA	FPGA
	Latest	2014	2019	2018	2020	2020	2020
	Connectivity	Local (5,N-N)	Local (8,King's Graph)	All-to-All	All-to-All	All-to-All	Local (5,N-N)
	Total neurons N	2000	30 000	1024	512	150	8,100
	Parallel neurons N_p	$N/2$ = 1000	$N/4$ = 7500	1	N = 512	N = 150	N = 8100
	Clock frequency, f	250 MHz	100 MHz	100 MHz	320 MHz	70 MHz	125 MHz
	Weight precision	1 bit	3 bit	16 bit	5 bit	9 bit	16 bit
[Derived]	Neuron time (MC step) $\tau = 1/f$	4 ns	10 ns	10 ns	~3 ns	14 ns	32 ns
	Flips per second (N_p/τ)	2.5×10^{11}	7.5×10^{11}	10^{8}	$\sim 2 \times 10^{11}$	10^{10}	$\sim 2.5 \times 10^{11}$

FIG. 13. Flips per second (fps) is a substrate and algorithm-independent performance metric for simulated annealing processors much like the flops-per-second metric used for general purpose computers. It is a measure of how many flips, and hence spin configurations, the system can cycle through in a second. Flips per second can be derived from the reported performance metrics of the processors following Ref. [29]. The reported and derived quantities are as indicated. Current CMOS-based annealing processors perform at about 10^{12} fps. We project that MTJ-based hardware can increase by a few orders of magnitude.

on the physical hardware performance but also on the algorithm that is being implemented. Here, we emphasize measuring the hardware performance in terms of the purely hardware metric *flips per second* [19,29,66], which refers to the maximum number of spin configurations the hardware can cycle through per second. It depends on the number of spins in the system (N) and the time it takes for a spin to flip (τ), $f = N/\tau$.

For the digital annealers, the spin update time is usually determined by its clock period (τ_{clk}), which typically ranges in the tens of nanosecond range. To ensure fidelity, simultaneous updates of connected spins need to be avoided [67], forcing digital annealers that operate on the clock edge to update spins sequentially. Thus, in a network where all spins are connected, effectively only one spin can update per clock cycle [21]. But this needs not be the case if some spins are unconnected, (i.e., nearest-neighbor [1,19] or king-graph [20] connection, or if spins are parallelized by implementing special algorithms [22–24]. Based on the reported total spin number and clock speeds of digital annealing hardware today that have about 10 K neurons that can update about every 10 ns clock period, we derive an estimation of their performance at $f \sim 10^4/10^{-8} = 10^{12}$ flips per second [1,29], as shown in Fig. 13.

Compared to digital annealers the Ising spin hardware we present in this work can work autonomously, i.e, without a synchronizing clock or a sequencer [29,65,68]. In this mode, the speeds are governed by neuron (τ_{neu}) and synapse (τ_{syn}) times only, and to ensure fidelity and avoid simultaneous updates of connected BSNs, the synapse needs to update faster than the neuron ($\tau_{\text{syn}} < \tau_{\text{neu}}$). Sutton *et al.* [29] defined a metric $s = \tau_{\text{syn}}/\tau_{\text{neu}}$ and showed that to ensure the fidelity of operations s needs to be less than 1. The exact requirements are problem and architecture dependent. Memristive crossbar arrays paired with a fast summing amplifier synapse could operate very efficiently at as low as a few tens of picosecond speeds [26–28,69–71].

The digital annealers mimic the Ising spin using a combination of random-number generators (LFSR, Xoshiro, etc.), look-up tables, and comparators. The random number generator (RNG) unit is one of the most expensive elements in the design [72]. Even in the most optimized design, the RNG unit takes up about 11% of the total logic gate area [22]. The 3T-1MTJ design offers a drastic reduction in the area footprint, promising massive scalability, leveraging existing 1T-1MTJ magnetic RAM technology that already has 1 Gbit integrated cells [73,74].

Figure 12(b) projects the fps number considering $\tau \equiv \tau_{\text{neu}} \approx \tau_{\text{CORR}}$ for different numbers of spins, N. A MTJ realization with circular IMA, with about a nanosecond timescale can offer almost 2 orders of magnitude speedup with less than 10 k neurons. If spins are implemented in Gbit densities, all stochastic implementations seem to outperform the CMOS implementations. For such systems, the upper bound for N is ultimately determined by either the area or power budget of the chip. Note that the fps number does not reflect the connectivity of the spins or the algorithm implemented by the hardware. It also does not indicate the solution accuracy obtainable for specific problems [75]. What we highlight here is that, by using the natural physics of the MTJ, we can design a very compact realization of Eq. (2) compared to current state-of-the-art CMOS implementations, and despite being a magnetic circuit, low-barrier magnet implementations even offer an overall speed up due to their fast fluctuation rates.

V. CONCLUSION

In this paper, we present a comprehensive evaluation of naturally stochastic magnetic building blocks for implementing probabilistic algorithms compactly and efficiently. We generalize the proposed 3T-1MTJ design to a 3T-1SR design and present necessary design rules for the BSN operation that we hope will stimulate further interest in finding stochastic resistance with suitable properties. We extend the physical performance analysis of the 3T-1MTJ BSN design to include unstable MTJs with different low-barrier magnets as free layers. They are evaluated as physical realizations of the general stochastic resistor with respect to 14 nm FinFET transistors. IMA magnets with barrier less than or equal to k_BT prove to be the best option; low-barrier PMA magnets can function as current-tunable resistors as well. While careful optimization of the fixed layer to cancel the stray fields in the IMA MTJ is preferred, PMA can benefit from the presence of stray fields (can be a source of I_{50}). The most challenging set of working conditions are set for telegraphic IMA magnets; even if they are highly optimized and no stray fields are present in the circuit, they need to be coupled with high current transistors due to their high pinning currents because pairing them with low current transistors like 14 nm FinFET results in a staircaselike functional behavior that does not work as a p-bit.

These BSNs are an integral part of Ising machines that are often referred to as annealing processors. Using the 3T-MTJ BSN could speed up the operation of these processors by orders of magnitude. Another important application space for these BSNs is stochastic neural networks [68,76–78]. In fact, binary stochastic neurons are desired for deep learning networks, but are typically avoided because it is harder to generate random bits in the CMOS hardware [79]. Use of this compact neuron that relies on the natural physics of MTJs to provide stochastic binarization could accelerate computation in custom hardware [80,81] by faster evaluation of the BSN function [32] and also encourage algorithmic advancement using a BSN.

ACKNOWLEDGMENTS

This work is supported by the Center for Probabilistic Spin Logic for Low-Energy Boolean and Non-Boolean Computing (CAPSL), one of the Nanoelectronic Computing Research (nCORE) Centers as task 2759.005, a Semiconductor Research Corporation (SRC) program sponsored by the NSF through CCF 1739635.

APPENDIX: DERIVATION OF THE PINNING FIELD OF LOW-BARRIER MAGNETS

Magnets are generally used to store information putting the focus on the evaluating and predicting characteristics of stable high-barrier magnets. It is interesting to note that theoretical predictions and analytical derivations regarding low-barrier magnet ($\Delta \leq k_BT$) dynamics typically receive less attention as cases of "least practical interest" [82]. We document the analytical expressions associated with LBMs in Fig. 10. The expressions for the correlation time and biasing current can be found in Refs. [32,46,58,83]; in this appendix we derive the bias field.

We derive the expressions for the external magnetic field H_0 required to pin the magnetization of a LBM with $\Delta \leq k_BT$ here. We start from the energy expression for the magnet (E) and derive the expressions presented in Fig. 10 from the steady-state average magnetization defined by

$$\langle m \rangle = \frac{\int_{\theta=0}^{\theta=\pi} \int_{\phi=-\pi}^{\phi=\pi} \sin\theta \, d\phi \, d\theta \, m \exp(-E/k_BT)}{\int_{\theta=0}^{\theta=\pi/2} \int_{\phi=-\pi}^{\phi=\pi} \sin\theta \, d\phi \, d\theta \exp(-E/k_BT)}, \tag{A1}$$

where $(m_x, m_y, m_z) \equiv (\cos\theta, \sin\theta \sin\phi, \sin\theta \cos\phi)$.

1. Perpendicular magnetic anisotropy

In the case of a LBM with perpendicular magnetization, the anisotropy field along the x axis $H_{kp} \rightarrow 0$ and thus, for a field applied in the x direction, the energy expression, Eq. (1), is reduced to

$$E = -H_{\text{ext}} M_S \Omega m_x. \tag{A2}$$

Evaluation of Eq. (A1) with respect to this energy gives $\langle m_x \rangle = \coth(H_{\text{ext}} M_S \Omega / k_BT) - (H_{\text{ext}} M_S \Omega / k_BT) \approx \tanh(H_{\text{ext}} M_S \Omega / 3k_BT)$. Thus, to pin the magnetization to any of its states $\langle m_x \rangle = \pm 1$, the required external field for PMA magnets can be approximated by

$$|H_{\text{ext(PMA)}}| = \frac{3k_BT}{M_s\Omega}. \tag{A3}$$

2. In-plane magnetic anisotropy

For LBMs with in-plane magnets, the anisotropy field along the z axis $H_{ki} \rightarrow 0$ and a large demagnetization field H_D exists along the z axis that keeps the magnetization in plane. The energy expression, Eq. (1), in this case is

$$E = H_D M_S \Omega m_x^2 - H_{\text{ext}} M_S \Omega m_z. \tag{A4}$$

Once again evaluating Eq. (A1) with respect to this energy for very large demagnetizing fields ($H_D \rightarrow \infty$) can be simplified to $\langle m_z \rangle \approx H_{\text{ext}} M_S \Omega / 2k_BT$. Thus, to pin the magnetization to any of its states $\langle m_z \rangle = \pm 1$, the required external field for IMA magnets can be approximated by

$$|H_{\text{ext(IMA)}}| = \frac{2k_BT}{M_s\Omega}. \tag{A5}$$

The expression is independent of the demagnetization field. These empirical expressions match our SPICE simulation results quite well, as shown in Fig. 14.

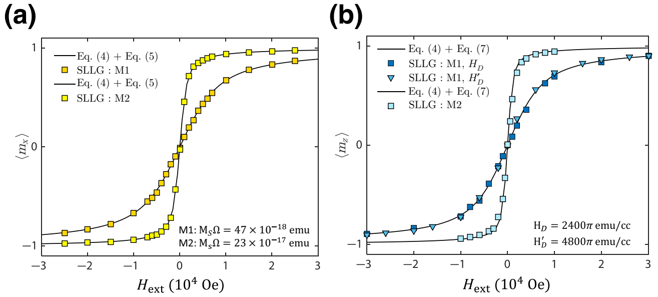


FIG. 14. Pinning field of low-barrier magnets: The numerical evaluations of the equations are compared to SPICE simulation results for (a) isotropic magnets and (b) circular IMA magnets that have $\Delta \leq k_B T$. The pinning fields are shown to be a function of $M_S\Omega$ only where $M_S = 600$ emu/cc and the volume of magnet Ω is varied. The pinning field values for IMA magnets indicate that they are independent of the large demagnetization field H_D. The precise correspondence between the analytical formulas and the numerical simulation also serves as a benchmark for our finite temperature (stochastic) LLG formulation.

[1] Masanao Yamaoka, Chihiro Yoshimura, Masato Hayashi, Takuya Okuyama, Hidetaka Aoki, and Hiroyuki Mizuno, in *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers* (IEEE, San Francisco, CA, USA, 2015), p. 1.

[2] Florian Neukart, Gabriele Compostella, Christian Seidel, David Von Dollen, Sheir Yarkoni, and Bob Parney, Traffic flow optimization using a quantum annealer, Front. ICT **4**, 29 (2017).

[3] Francisco Barahona, Martin Grötschel, Michael Jünger, and Gerhard Reinelt, An application of combinatorial optimization to statistical physics and circuit layout design, Oper. Res. **36**, 493 (1988).

[4] Chase Cook, Hengyang Zhao, Takashi Sato, Masayuki Hiromoto, and Sheldon X.-D. Tan, GPU based parallel Ising computing for combinatorial optimization problems in VLSI physical design, ArXiv:1807.10750 (2018).

[5] Gili Rosenberg, Poya Haghnegahdar, Phil Goddard, Peter Carr, Kesheng Wu, and Marcos López De Prado, Solving the optimal trading trajectory problem using a quantum annealer, IEEE J. Sel. Top. Signal Process. **10**, 1053 (2016).

[6] Hiromasa Sakaguchi, Koji Ogata, Tetsu Isomura, Shoko Utsunomiya, Yoshihisa Yamamoto, and Kazuyuki Aihara, Boltzmann sampling by degenerate optical parametric oscillator network for structure-based virtual screening, Entropy **18**, 365 (2016).

[7] Francisco Barahona, On the computational complexity of Ising spin glass models, J. Phys. A: Math. General **15**, 3241 (1982).

[8] Andrew Lucas, Ising formulations of many np problems, Front. Phys. **2**, 5 (2014).

[9] Brian Sutton, Kerem Yunus Camsari, Behtash Behin-Aein, and Supriyo Datta, Intrinsic optimization using stochastic nanomagnets, Sci. Rep. **7**, 44370 (2017).

[10] Binary stochastic neurons in tensorflow, https://r2rt.com/binary-stochastic-neurons-in-tensorflow.html.

[11] Mark W. Johnson, Mohammad H. S. Amin, Suzanne Gildert, Trevor Lanting, Firas Hamze, Neil Dickson, Richard Harris, Andrew J. Berkley, Jan Johansson, Paul Bunyk, *et al.*, Quantum annealing with manufactured spins, Nature **473**, 194 (2011).

[12] Peter L. McMahon, Alireza Marandi, Yoshitaka Haribara, Ryan Hamerly, Carsten Langrock, Shuhei Tamate, Takahiro Inagaki, Hiroki Takesue, Shoko Utsunomiya, Kazuyuki Aihara, *et al.*, A fully programmable 100-spin coherent Ising machine with all-to-all connections, Science **354**, 614 (2016).

[13] Sourav Dutta, Abhishek Khanna, Hanjong Paik, Darrell Schlom, Arijit Raychowdhury, Zoltan Toroczkai, and Suman Datta, Ising hamiltonian solver using stochastic phase-transition nano-oscillators, ArXiv:2007.12331 (2020).

[14] Hayato Goto, Kosuke Tatsumura, and Alexander R. Dixon, Combinatorial optimization by simulating adiabatic bifurcations in nonlinear hamiltonian systems, Sci. Adv. **5**, eaav2372 (2019).

[15] Tianshi Wang and Jaijeet Roychowdhury, in *International Conference on Unconventional Computation and Natural Computation* (Springer, Tokyo, Japan, 2019), p. 232.

[16] Ibrahim Ahmed, Po-Wei Chiu, and Chris H. Kim, in *2020 IEEE Symposium on VLSI Circuits* (IEEE, Honolulu, USA, 2020), p. 1.

[17] Jeffrey Chou, Suraj Bramhavar, Siddhartha Ghosh, and William Herzog, Analog coupled oscillator based weighted Ising machine, Sci. Rep. **9**, 1 (2019).

[18] Scott Kirkpatrick, C. Daniel Gelatt, and Mario P. Vecchi, Optimization by simulated annealing, Science **220**, 671 (1983).

[19] Marco Baity-Jesi, Rachel A. Baños, Andres Cruz, Luis Antonio Fernandez, José Miguel Gil-Narvión, Antonio Gordillo-Guerrero, David Iñiguez, Andrea Maiorano, Filippo Mantovani, Enzo Marinari, *et al.*, Janus II: A new generation application-driven computer for spin-system simulations, Comput. Phys. Commun. **185**, 550 (2014).

[20] Takashi Takemoto, Masato Hayashi, Chihiro Yoshimura, and Masanao Yamaoka, in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)* (IEEE, San Francisco, CA, USA, 2019), p. 52.

[21] Maliheh Aramon, Gili Rosenberg, Elisabetta Valiante, Toshiyuki Miyazawa, Hirotaka Tamura, and Helmut G. Katzgraber, Physics-inspired optimization for quadratic unconstrained problems using a digital annealer, Front. Phys. **7**, 48 (2019).

[22] Kasho Yamamoto, Kota Ando, Normann Mertig, Takashi Takemoto, Masanao Yamaoka, Hiroshi Teramoto, Akira Sakai, Shinya Takamaeda-Yamazaki, and Masato Motomura, in *2020 IEEE International Solid-State Circuits Conference-(ISSCC)* (IEEE, San Francisco, CA, USA, 2020), p. 138.

[23] Saavan Patel, Lili Chen, Philip Canoza, and Sayeef Salahuddin, Ising model optimization problems on a FPGA accelerated restricted Boltzmann machine, ArXiv:2008.04436 (2020).

[24] Saavan Patel, Philip Canoza, and Sayeef Salahuddin, Logically synthesized, hardware-accelerated, restricted Boltzmann machines for combinatorial optimization and integer factorization, ArXiv:2007.13489 (2020).

[25] Kerem Yunus Camsari, Sayeef Salahuddin, and Supriyo Datta, Implementing p-bits with embedded MTJ, IEEE Electron Device Lett. **38**, 1767 (2017).

[26] Lixue Xia, Peng Gu, Boxun Li, Tianqi Tang, Xiling Yin, Wenqin Huangfu, Shimeng Yu, Yu Cao, Yu Wang, and Huazhong Yang, Technological exploration of RRAM crossbar array for matrix-vector multiplication, J. Comput. Sci. Technol. **31**, 3 (2016).

[27] Fuxi Cai, Justin M. Correll, Seung Hwan Lee, Yong Lim, Vishishtha Bothra, Zhengya Zhang, Michael P. Flynn, and Wei D. Lu, A fully integrated reprogrammable memristor-CMOS system for efficient multiply–accumulate operations, Nat. Electron. **2**, 290 (2019).

[28] F. Merrikh Bayat, Mirko Prezioso, Bhaswar Chakrabarti, H. Nili, I. Kataeva, and D. Strukov, Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits, Nat. Commun. **9**, 1 (2018).

[29] Brian Sutton, Rafatul Faria, Lakshmi A. Ghantasala, Kerem Y. Camsari, and Supriyo Datta, Autonomous probabilistic coprocessing with petaflips per second, arXiv:1907.09664 (2019).

[30] Mustafa Mert Torunbalci, Pramey Upadhyaya, Sunil A. Bhave, and Kerem Y. Camsari, Modular compact modeling of MTJ devices, IEEE Trans. Electron Devices **65**, 4628 (2018).

[31] Predictive technology model (PTM), http://ptm.asu.edu/.

[32] Orchi Hassan, Rafatul Faria, Kerem Yunus Camsari, Jonathan Z. Sun, and Supriyo Datta, Low-barrier magnet design for efficient hardware binary stochastic neurons, IEEE Magn. Lett. **10**, 1 (2019).

[33] Matthew W. Daniels, Advait Madhavan, Philippe Talatchian, Alice Mizrahi, and Mark D. Stiles, Energy-Efficient Stochastic Computing with Superparamagnetic Tunnel Junctions, Phys. Rev. Appl. **13**, 034016 (2020).

[34] Bradley Parks, Mukund Bapna, Julianne Igbokwe, Hamid Almasi, Weigang Wang, and Sara A. Majetich, Superparamagnetic perpendicular magnetic tunnel junctions for true random number generators, AIP Adv. **8**, 055903 (2018).

[35] Julie Grollier, Damien Querlioz, K. Y. Camsari, Karin Everschor-Sitte, Shunsuke Fukami, and Mark D. Stiles, Neuromorphic spintronics, Nat. Electron. **3**, 1 (2020).

[36] Md Ahsanul Abeed and Supriyo Bandyopadhyay, Low energy barrier nanomagnet design for binary stochastic neurons: Design challenges for real nanomagnets with fabrication defects, IEEE Magn. Lett. **10**, 1 (2019).

[37] Justine L. Drobitch and Supriyo Bandyopadhyay, Reliability and scalability of p-bits implemented with low energy barrier nanomagnets, IEEE Magn. Lett. **10**, 1 (2019).

[38] William A. Borders, Ahmed Z. Pervaiz, Shunsuke Fukami, Kerem Y. Camsari, Hideo Ohno, and Supriyo Datta, Integer factorization using stochastic magnetic tunnel junctions, Nature **573**, 390 (2019).

[39] Brad Parks, Ahmed Abdelgawad, Thomas Wong, Richard F. L. Evans, and Sara A. Majetich, Magnetoresistance Dynamics in Superparamagnetic Co–Fe–B Nanodots, Phys. Rev. Appl. **13**, 014063 (2020).

[40] Suresh Cheemalavagu, Pinar Korkmaz, Krishna V. Palem, Bilge E. S. Akgul, and Lakshmi N. Chakrapani, in *IFIP International Conference on VLSI* (Princeton, New Jersey, USA, 2005), p. 535.

[41] Nikhil Shukla, Abhinav Parihar, Eugene Freeman, Hanjong Paik, Greg Stone, Vijaykrishnan Narayanan, Haidan Wen, Zhonghou Cai, Venkatraman Gopalan, Roman Engel-Herbert, *et al.*, Synchronized charge oscillations in correlated electron systems, Sci. Rep. **4**, 4964 (2014).

[42] Suhas Kumar, John Paul Strachan, and R. Stanley Williams, Chaotic dynamics in nanoscale $\mathrm{NbO_2}$ Mott memristors for analogue computing, Nature **548**, 318 (2017).

[43] Bernhard Stampfer, Feng Zhang, Yury Yuryevich Illarionov, Theresia Knobloch, Peng Wu, Michael Waltl, Alexander Grill, Joerg Appenzeller, and Tibor Grasser, Characterization of single defects in ultrascaled $\mathrm{MoS_2}$ field-effect transistors, ACS Nano **12**, 5368 (2018).

[44] Jialin Cai, Bin Fang, Like Zhang, Wenxing Lv, Baoshun Zhang, Tiejun Zhou, Giovanni Finocchio, and Zhongming Zeng, Voltage-Controlled Spintronic Stochastic Neuron Based on a Magnetic Tunnel Junction, Phys. Rev. Appl. **11**, 034015 (2019).

[45] Kerem Y. Camsari, Mustafa Mert Torunbalci, William A. Borders, Hideo Ohno, and Shunsuke Fukami, Double free-layer magnetic tunnel junctions for probabilistic bits, ArXiv:2012.06950 (2020).

[46] William T. Coffey and Yuri P. Kalmykov, Thermal fluctuations of magnetic nanoparticles: Fifty years after brown, J. Appl. Phys. **112**, 121301 (2012).

[47] C. J. Lin, S. H. Kang, Y. J. Wang, K Lee, X Zhu, W. C. Chen, X. Li, W. N. Hsu, Y. C. Kao, M. T. Liu, *et al.*, in *Electron Devices Meeting (IEDM), 2009 IEEE International* (IEEE, San Francisco, CA, USA, 2009), p. 1.

[48] Kerem Yunus Camsari, Rafatul Faria, Brian M. Sutton, and Supriyo Datta, Stochastic p-Bits for Invertible Logic, Phys. Rev. X **7**, 031014 (2017).

[49] Yang Lv, Robert P. Bloom, and Jian-Ping Wang, Experimental demonstration of probabilistic spin logic by magnetic tunnel junctions, IEEE Magn. Lett. **10**, 1 (2019).

[50] Brandon R. Zink, Yang Lv, and Jian-Ping Wang, Independent control of antiparallel-and parallel-state thermal stability factors in magnetic tunnel junctions for telegraphic signals with two degrees of tunability, IEEE Trans. Electron Devices **66**, 5353 (2019).

[51] Stuart S. P. Parkin, Christian Kaiser, Alex Panchula, Philip M. Rice, Brian Hughes, Mahesh Samant, and See-Hun Yang, Giant tunnelling magnetoresistance at room temperature with MgO (100) tunnel barriers, Nat. Mater. **3**, 862 (2004).

[52] S. Ikeda, J. Hayakawa, Y. Ashizawa, Y. M. Lee, K. Miura, H. Hasegawa, M. Tsunoda, F. Matsukura, and H. Ohno, Tunnel magnetoresistance of 604% at 300 K by suppression of ta diffusion in CoFeB/MgO/CoFeB pseudo-spin-valves annealed at high temperature, Appl. Phys. Lett. **93**, 082508 (2008).

[53] Punyashloka Debashis, Rafatul Faria, Kerem Y. Camsari, and Zhihong Chen, Design of stochastic nanomagnets for probabilistic spin logic, IEEE Magn. Lett. **9**, 1 (2018).

[54] Punyashloka Debashis, Rafatul Faria, Kerem Y. Camsari, Joerg Appenzeller, Supriyo Datta, and Zhihong Chen, in *Electron Devices Meeting (IEDM), 2016 IEEE International* (IEEE, San Francisco, CA, USA, 2016), p. 34.

[55] Christopher Safranski, Jan Kaiser, Philip Trouilloud, Pouya Hashemi, Guohan Hu, and Jonathan Z. Sun, Demonstration

of nanosecond operation in stochastic magnetic tunnel junctions, ArXiv:2010.14393 (2020).

[56] Chaoliang Zhang, Yutaro Takeuchi, Shunsuke Fukami, and Hideo Ohno, Field-free and sub-ns magnetization switching of magnetic tunnel junctions by combining spin-transfer torque and spin–orbit torque, Appl. Phys. Lett. **118**, 092406 (2021).

[57] nanohub.org: Modular approach to spintronics, https://nanohub.org/groups/spintronics.

[58] Jan Kaiser, Avinash Rustagi, Kerem Y. Camsari, Jonathan Z. Sun, Supriyo Datta, and Pramey Upadhyaya, Subnanosecond Fluctuations in Low-Barrier Nanomagnets, Phys. Rev. Appl. **12**, 054056 (2019).

[59] Matthew R. Pufall, William H. Rippard, Shehzaad Kaka, Steven E. Russek, Thomas J. Silva, Jordan Katine, and Matt Carey, Large-angle, gigahertz-rate random telegraph switching induced by spin-momentum transfer, Phys. Rev. B **69**, 214409 (2004).

[60] Rafatul Faria, Kerem Yunus Camsari, and Supriyo Datta, Low-barrier nanomagnets as p-bits for spin logic, IEEE Magn. Lett. **8**, 1 (2017).

[61] Jonathan Z. Sun, Spin-current interaction with a monodomain magnetic body: A model study, Phys. Rev. B **62**, 570 (2000).

[62] Rafatul Faria, Kerem Y. Camsari, and Supriyo Datta, Implementing Bayesian networks with embedded stochastic MRAM, AIP Adv. **8**, 045101 (2018).

[63] Miguel Romera, Philippe Talatchian, Sumito Tsunegi, Flavio Abreu Araujo, Vincent Cros, Paolo Bortolotti, Juan Trastoy, Kay Yakushiji, Akio Fukushima, Hitoshi Kubota, *et al.*, Vowel recognition with four coupled spin-torque nano-oscillators, Nature **563**, 230 (2018).

[64] Sarah Jenkins, Andrea Meo, Luke E. Elliott, Stephan K. Piotrowski, Mukund Bapna, Roy W. Chantrell, Sara A. Majetich, and Richard F. L. Evans, Magnetic stray fields in nanoscale magnetic tunnel junctions, J. Phys. D: Appl. Phys. **53**, 044001 (2019).

[65] Rafatul Faria, Jan Kaiser, Kerem Y. Camsari, and Supriyo Datta, Hardware design for autonomous Bayesian networks, ArXiv:2003.01767 (2020).

[66] Sergei V. Isakov, Ilia N. Zintchenko, Troels F. Rønnow, and Matthias Troyer, Optimised simulated annealing for Ising spin glasses, Comput. Phys. Commun. **192**, 265 (2015).

[67] Emile Aarts, Emile H. L. Aarts, and Jan Karel Lenstra, *Local Search in Combinatorial Optimization* (Princeton University Press, 2003).

[68] Jan Kaiser, Rafatul Faria, Kerem Y. Camsari, and Supriyo Datta, Probabilistic circuits for autonomous learning: A simulation study, Front. Comput. Neurosci. **14**, 1 (2020).

[69] Fuxi Cai, Suhas Kumar, Thomas Van Vaerenbergh, Rui Liu, Can Li, Shimeng Yu, Qiangfei Xia, J. Joshua Yang, Raymond Beausoleil, Wei Lu, *et al.*, Harnessing intrinsic noise in memristor hopfield neural networks for combinatorial optimization, ArXiv:1903.11194 (2019).

[70] Hao Huang, Johannes Heilmeyer, Markus Grözing, Manfred Berroth, Jochen Leibrich, and Werner Rosenkranz, An 8-bit 100-GS/s distributed DAC in 28-nm CMOS for optical communications, IEEE Trans. Microw. Theory Tech. **63**, 1211 (2015).

[71] Miao Hu, Catherine E. Graves, Can Li, Yunning Li, Ning Ge, Eric Montgomery, Noraica Davila, Hao Jiang, R. Stanley Williams, J. Joshua Yang, *et al.*, Memristor-based analog computation and neural network classification with a dot product engine, Adv. Mater. **30**, 1705914 (2018).

[72] Hidenori Gyoten, Masayuki Hiromoto, and Takashi Sato, Area efficient annealing processor for Ising model without random number generator, IEICE Trans. Inf. Syst. **101**, 314 (2018).

[73] S. Aggarwal, H. Almasi, M. DeHerrera, B. Hughes, S. Ikegawa, J. Janesky, H. K. Lee, H. Lu, F. B. Mancoff, K. Nagel, *et al.*, in *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, USA, 2019), p. 2.

[74] Everspin enters pilot production phase for the world's first 28 nm 1 Gb STT-MRAM component, Everspin Technology (2019), https://investor.everspin.com/news-releases/news-release-details/everspin-enters-pilot-production-phase-worlds-first-28-nm-1-gb.

[75] Xiangyu Zhang, Ramin Bashizade, Yicheng Wang, Cheng Lyu, Sayan Mukherjee, and Alvin R. Lebeck, Beyond application end-point results: Quantifying statistical robustness of MCMC accelerators, ArXiv:2003.04223 (2020).

[76] Shamma Nasrin, Justine L. Drobitch, Supriyo Bandyopadhyay, and Amit Ranjan Trivedi, Low power restricted Boltzmann machine using mixed-mode magneto-tunneling junctions, IEEE Electron Device Lett. **40**, 345 (2019).

[77] Catherine D. Schuman, Thomas E. Potok, Robert M. Patton, J. Douglas Birdwell, Mark E. Dean, Garrett S. Rose, and James S. Plank, A survey of neuromorphic computing and neural networks in hardware, arXiv:1705.06963 (2017).

[78] Geoffrey E. Hinton, Training products of experts by minimizing contrastive divergence, Neural Comput. **14**, 1771 (2002).

[79] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio, Binarized neural networks: Training neural networks with weights and activations constrained to +1 or −1, ArXiv:1602.02830 (2016).

[80] Chang-Hung Tsai, Wan-Ju Yu, Wing Hung Wong, and Chen-Yi Lee, A 41.3/26.7 pJ per neuron weight RBM processor supporting on-chip learning/inference for IoT applications, IEEE J. Solid-State Circuits **52**, 2601 (2017).

[81] Seongwook Park, Kyeongryeol Bong, Dongjoo Shin, Jinmook Lee, Sungpill Choi, and Hoi-Jun Yoo, in *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers* (IEEE, San Francisco, CA, USA, 2015), p. 1.

[82] William Fuller Brown Jr, Thermal fluctuations of a single-domain particle, Phys. Rev. **130**, 1677 (1963).

[83] Shehrin Sayed, Kerem Y. Camsari, Rafatul Faria, and Supriyo Datta, Rectification in Spin-Orbit Materials Using Low-Energy-Barrier Magnets, Phys. Rev. Appl. **11**, 054063 (2019).