Stochastic Switching in a Magnetic-Tunnel-Junction Neuron and a Bias-Dependent Néel-Arrhenius Model

Ming-Hung Wu,[1] I-Ting Wang,[1] Ming-Chun Hong,[1,2] Kuan-Ming Chen,[3] Yuan-Chieh Tseng,[3]
Jeng-Hua Wei,[2] and Tuo-Hung Hou[1,2,*]

[1] *Institute of Electronics, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan*
[2] *Electronic and Optoelectronic System Research Laboratories, Industrial Technology Research Institute, Hsinchu 310, Taiwan*
[3] *Department of Material Science and Engineering, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan*

(Received 14 May 2022; revised 17 August 2022; accepted 27 October 2022; published 13 December 2022)

Back hopping or telegraphic switching describes stochastic bistate oscillation in magnetic tunnel junctions (MTJ) at high bias, which significantly increases the write error rate of memory storage. Nevertheless, this unfavorable stochastic switching could be utilized to construct extremely compact spiking neuron circuits, where both the spike frequency and duty cycle are proportional to the applied bias voltage. This MTJ neuron is the fundamental building block of future all-spin neural networks. This paper analyzes the mechanism of stochastic switching in the MTJ neuron in detail. The self-heating effect at high bias is identified to induce thermal perturbation through a reduced energy barrier of a weakened perpendicular magnetic anisotropy. The high spike frequency is measured up to 10 MHz, which is mainly limited by the transition time between states and the multistate switching when approaching the Curie temperature of the ferromagnetic phase. Finally, to quantitatively describe the stochastic switching phenomenon, we establish a bias-dependent Néel-Arrhenius compact model, which calibrates well with experimental data. Based on this model, the potential and design guideline for realizing MTJ neurons with a spike frequency toward the gigahertz range are also discussed.

DOI: 10.1103/PhysRevApplied.18.064034

I. INTRODUCTION

The spiking neural network (SNN) inspired by the biological brain is known as the third-generation neural work for its potential for extremely high energy efficiency [1]. Compared with the prevalent deep neural network (DNN) that relies on high-precision data transmission and processing, the SNN transmits information through simple binary pulse trains in the time domain by using algorithms such as rate coding or rank order [2]. This brings two major advantages to hardware implementation using the emerging in-memory computing (IMC) architecture [3,4]. First, although IMC is praised for its minimal data movement to accelerate DNNs with low latency and energy consumption, the substantial energy and area overhead of essential peripheral circuits to support high-precision IMC operations, including high-precision analog-to-digital converters (ADC) and digital-to-analog converters (DAC), presents daunting challenges [5]. The requirement of processing only binary input and output information greatly reduces the complexity of peripheral circuits in SNN hardware. Second, time-series encoding is naturally more suitable for processing real-world time-series information, such as real-time object detection in video streaming, than steady-state DNNs. The spike neuron circuit is the fundamental building block of SNNs [6]. In the IMC architecture, the neuron integrates the time-varied weighted-sum current from the synaptic array of the previous neural layer and generates binary pulse trains as the output. Popular neuron types include the integrate-and-fire (I&F) neuron [7], leaky I&F neuron [7], and Poisson neuron [8]. Ideally, this output could be directly transmitted to the next neural layer as the input. Therefore, the neuron replaces the ADC and DAC circuits and the activation function in conventional DNNs. Furthermore, it simplifies unnecessary data conversion back and forward between the analog and digital domains to save energy and reduce latency.

Spike frequency, energy consumption per spike, and circuit area are key parameters of neuron circuits for implementing SNNs with low latency, high energy efficiency, and high density. The conventional CMOS neuron circuit is mature and has a high spike frequency, but its large area, including the integration capacitor, comparator, and reset circuit, makes it difficult to support highly parallel high-density IMC architectures. As a result, many studies have focused on emerging neuron circuits based

[*] thhou@nycu.edu.tw

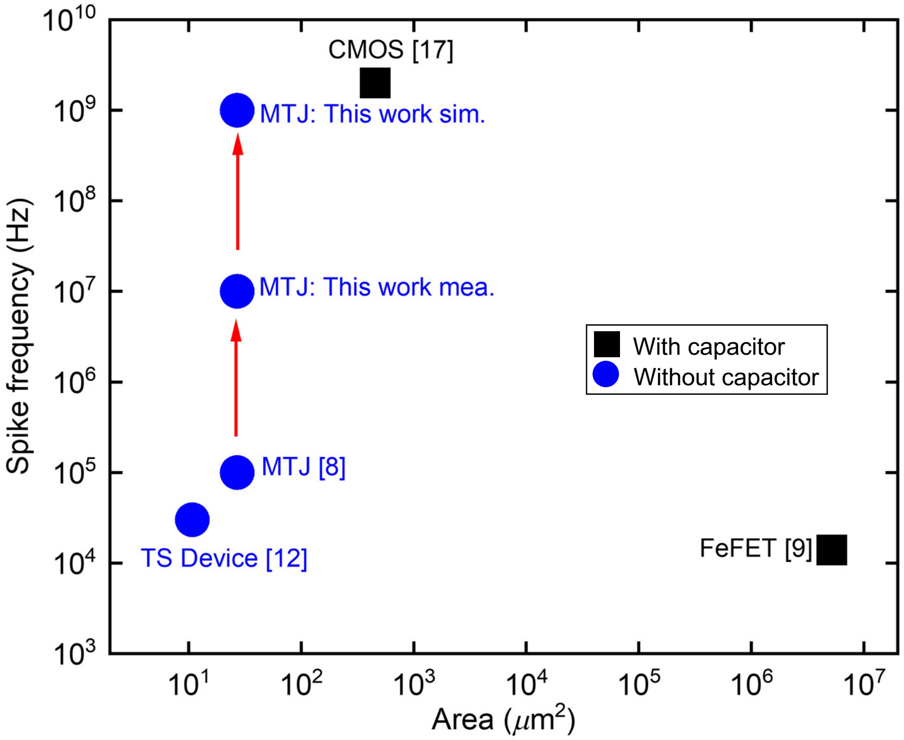

FIG. 1. Benchmark of CMOS-based and device-based neuron circuits implemented using 65-nm technology [16]. The spike frequency considers both the integration and firing periods. The area considers the complete neuron circuit including the necessary bias circuit, comparator, capacitor, and reset circuit.

on intrinsic device properties in ferroelectric memory [9–11], threshold-switching device (TS device) [12–14], magnetic random-access memory (MRAM), etc. [8,15]. A previous study quantitatively compared various neuron circuits implemented using 65-nm technology [16]. The spike frequency and circuit area of the respective technologies are summarized in Fig. 1 [8,9,12,17]. Among them, the MRAM neuron based on the magnetic tunnel junction (MTJ) structure shows stochastic oscillation between the low-resistance parallel (P) state and the high-resistance antiparallel (AP) state at high bias. The oscillation generates spike or pulse frequency and duty cycle depending on the voltage bias applied to the MRAM. Therefore, a single MRAM device acts as an extremely compact neuron circuit with no need for an additional integration capacitor, comparator, and reset circuit, presenting at least an order of magnitude area reduction compared with the conventional CMOS neuron. This MRAM neuron is different from the stochastic superparamagnetic tunnel junction (SMTJ) [18–20], which is intentionally designed with a low thermal stability factor, thus easily switched between the AP and P states by thermal fluctuation even at room temperature and low voltage bias. Although the SMTJ could potentially be utilized as a neuron, too, it could not be used to store the nonvolatile synaptic weight information in SNNs. The proposed MRAM device is designed with a standard high thermal stability factor for memory storage. It is thermally stable at normal read and retention conditions. The same device could be utilized as both the neuron (at high bias voltage) and synapse (at low bias voltage) for the easy integration of an all-spin SNN architecture [8]. Moreover, the proposed device is based on the commercially available perpendicular spin-transfer torque (STT) MRAM technology that has been successfully scaled down to 16 nm [21,22] and has reached a high density of 1 Gbit [23]. Its excellent scalability, compact area, and ease of large-scale integration make it extremely attractive for future SNN hardware.

The bistate oscillation of standard STT MRAM at high bias, also known as back hopping [24–29] or telegraphic switching [30,31], is not included in the typical design space for memory storage applications because it significantly increases the write error rate. In general, the oscillation can be understood as the thermal perturbation on the reduced energy barrier of a weakened perpendicular magnetic anisotropy (PMA) at high bias [8], but the relation between oscillation and spike frequency is truly vague. Additionally, although the spike frequency of MRAM neurons has been reported up to 100 kHz [15], which is among the highest in device-based compact neurons, it is slower than that of the CMOS neuron with gigahertz spike frequency by orders of magnitude, as shown in Fig. 1. A high spike frequency is beneficial for SNN applications because it accelerates computing and reduces processing latency. It also suppresses the energy consumption per spike by reducing the dc energy dissipation of the weighted-sum current flowing from the synaptic array into the neuron circuit [16]. Moreover, since the MRAM neuron is based on nonvolatile memory technology with finite endurance, a higher spike frequency also reduces the total stress time (t_{stress}) applied to the device for a given amount of processing load, thus improving the system reliability. The MRAM neuron operated at a frequency of 100 kHz consumes approximately 20 000 times more energy than the CMOS neuron operated at a frequency of 2 GHz [16]. Therefore, achieving high spike frequency in the MRAM neuron is critical for energy-efficient SNN applications.

In this paper, the comprehensive measurement and mechanism of telegraphic switching in the MRAM neurons at high bias voltage and high frequency are discussed in detail. An increased high spike frequency up to 10 MHz is demonstrated. We also successfully model the self-heating effect in MRAM neurons quantitatively using a bias-dependent Néel-Arrhenius model. This well-calibrated model allows us to explore the design guideline for realizing a highly competitive MRAM neuron with a high spike frequency toward the gigahertz range and the wide detection range required for future SNN applications. This article is organized as follows: In Sec. II, we describe the fabrication process of the proposed MRAM device. In Sec. III, we briefly review the operation principles of the MRAM-based I&F neuron [15] and Poisson neuron [8], which provide a general perspective connecting the device and the application. In Sec. IV, the comprehensive measurement of telegraphic switching and its relation with the strong self-heating effect are discussed. In Sec. V, the potential of high spike frequency in the MRAM neuron is discussed through both the experiment and a bias-dependent Néel-Arrhenius model.

II. FABRICATION PROCESS

A perpendicular magnetic tunnel junction (PMTJ), consisting of a Ta seed layer, a $[\text{Co/Pt}]_4/\text{Co/Ru}/[\text{Co/Pt}]_2/\text{Co}$ synthetic antiferromagnet (SAF) layer, a Co-Fe-B pinned layer, a MgO barrier layer, a Co-Fe-B/Ta/Co-Fe-B free layer, a MgO capping layer for enhancing PMA, and a Ta capping layer (from bottom to top), is prepared. All the ferromagnetic layers are deposited on the thermally oxidized silicon substrate at room temperature in an ultrahigh vacuum magnetron sputtering system and then annealed at 300 °C with an in-plane 1 T magnetic field for 2 h. The MgO barrier and capping layers are deposited using rf sputtering in an argon atmosphere. PMTJs are patterned using electron-beam lithography and reactive ion etching to complete a step etch-stop structure [32]. The diameter of the circular PMTJs is 80 nm. This fabrication process is identical to that for the standard PMTJ used for memory storage applications [32,33]. The device can be switched between the AP and P states by using bipolar write voltage/current pulses, and both the AP and P states are thermally stable at the read and retention conditions, unlike SMTJs with a low thermal stability factor [18–20]. Furthermore, the conventional PMTJ for memory storage applications is not intentionally designed to exhibit telegraphic switching without an external magnetic field. Two different PMTJ devices are fabricated with different SAF layer configurations to show both field-free and field-assisted telegraphic switching. The thickness of the Ru layer is 8 Å and 4.1 Å for the field-free and field-assisted samples, respectively, due to the difference in the stray field of the SAF layer. The conductance oscillation of the MTJ devices is measured using the Keysight B1530A waveform generator/fast measurement unit at different biases. For field-assisted switching, the external magnetic field is applied perpendicularly to the device through a custom electromagnet.

III. MTJ-BASED SPIKE NEURON AND ITS APPLICATION

Conventional CMOS-based I&F spike neurons consist of at least one integration capacitor, one comparator, and one reset circuit, as shown in Fig. 2(a). The operation is divided into the integration phase and the firing phase. In the integration phase, the membrane potential (V_m) increases when the weighted-sum current generated from a synaptic array accumulates in the integration capacitor. In the firing phase, the comparator generates a spike at the output potential (V_{out}) once V_m exceeds the preset and fixed threshold voltage (V_{th}). Then, the reset circuit resets the stored capacitive charges and initializes a new integration phase. As a result, the spike frequency increases proportionally with the weighted-sum current. The proposed compact MTJ neuron in the study overcomes the large circuit area of CMOS neurons by using the native stochastic switching characteristics of the MTJ at high bias voltage (V_b). On applying the weighted-sum current from synaptic arrays on the MTJ, the voltage drops on the low-resistance-state (LRS) MTJ, i.e., the P state, triggering finite switching probability. The cumulative switching probability (P_{sw}) increases with time, which is equivalent to the accumulation of V_m in the CMOS neuron, as shown in Fig. 2(b). With the increase of integration time, the MTJ eventually switches to a high resistance state (HRS), i.e., the AP state, and P_{sw} returns to zero. Similarly, P_{sw} starts to accumulate in the HRS MTJ at the same bias current, and the MTJ again switches back to the LRS with the increase of P_{sw}. V_{out} is larger at the HRS than that at the LRS, and, therefore, the consecutive MTJ switching processes generate repeated voltage

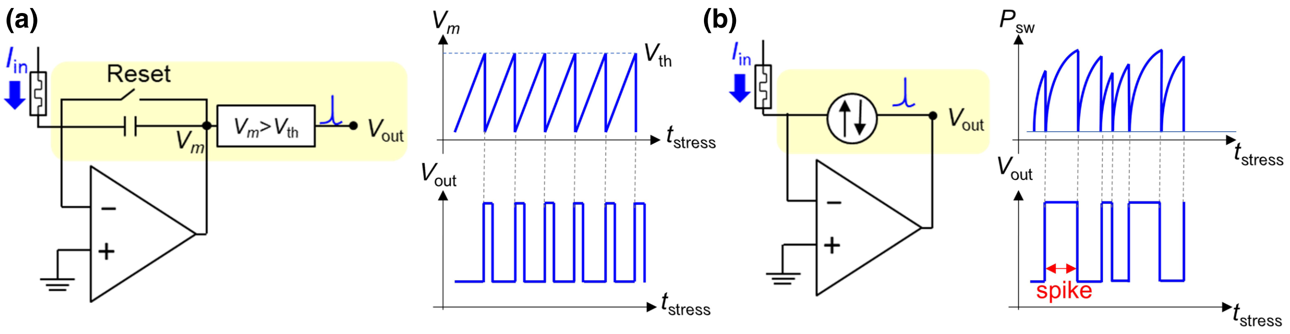

FIG. 2. Schematics of (a) the CMOS neuron and (b) the MTJ neuron. V_m of the CMOS neuron increases linearly as the weighted-sum current (I_{in}) accumulates on the integration capacitor and resets when V_m exceeds V_{th}. P_{sw}, the MTJ cumulative switching probability, also increases as I_{in} accumulates on the MTJ, but P_{sw} increases nonlinearly and the switching is stochastic between the AP and P states without the assistance of an external reset circuit.

spikes at V_{out} in a telegraphic switching process. Unlike the CMOS neuron where every spike is generated at a precise time only determined by the predesigned capacitance value and V_{th} for a given weighted-sum current, the MTJ spiking is stochastic. However, when the weighted-sum current increases, the overall spike frequency of MTJ neurons still increases as a result of a faster accumulation of P_{sw}, thus is suitable for various SNN applications. In our measurement, the MTJ spike frequency is proportional to V_b without applying an external magnetic field when designed properly, as shown in Fig. 3. Based on its I&F neuron characteristics, we demonstrated that the MTJ neuron could be used as a 4-bit time-domain ADC [15]. Utilizing a binary MTJ synapse together with the proposed MTJ neuron, the simulation of an all-spin neural network achieved 82% accuracy for the CIFAR-10 dataset in a challenging online training task by using only binary MTJ devices [15]. Although the relatively low accuracy is mainly limited by the nonideal gradient accumulation on the binary MTJ synapse, but not the MTJ neuron [4], this is a successful demonstration of an online-trained all-spin neural network.

In addition to the I&F neuron, the Poisson neuron transforms analog information into temporal signals by rate coding, for example, a Poisson spike train with a tunable duty cycle. The SNN with Poisson neurons allows all information to be processed in the analog domain without any energy-expensive analog-digital signal conversion. In our measurement, not only the spike frequency but also the duty cycle of the MTJ neuron is proportional to V_b, as shown in Fig. 3. Therefore, it could be used as both an I&F neuron and a Poisson neuron. We simulate the inference of the MNIST handwritten digit database using a SNN model converted from a DNN. The SNN inference considering the MTJ Poisson neuron achieves a high accuracy of 98.4% without converting information back and forward between the analog and digital domains. Interested readers could refer to [8] for the details on the algorithm and architecture of the all-spin SNN.

IV. STOCHASTIC SWITCHING MECHANISM OF THE MTJ NEURON

In this section, the physical mechanism behind the stochastic spiking or telegraphic switching characteristics of the MTJ neuron is discussed in detail using both electrical and magnetic measurements. The conductance-magnetic field (G-H) loop is measured at different V_b in Fig. 4(a). The switching magnetic field H_{sw} to the AP and P states at different V_b is extracted from the G-H loops and plotted in Fig. 4(b). H_{sw} is known to be dependent on STT, self-heating, and voltage-controlled magnetic anisotropy (VCMA) effects [34]. To decouple these effects, the offset field $H_{\text{offset}} = (H_{\text{sw}}^{\text{AP}\to\text{P}} + H_{\text{sw}}^{\text{P}\to\text{AP}})/2$ and the coercive field

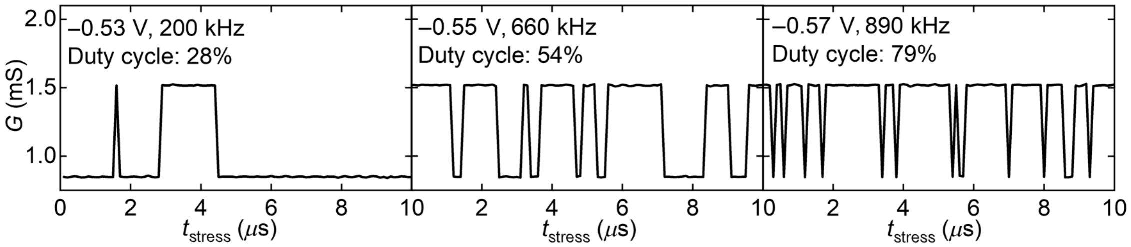

FIG. 3. Sampled MTJ conductance at different bias voltages without applying an external magnetic field. Both the duty cycle and spike frequency are proportional to the bias voltage.

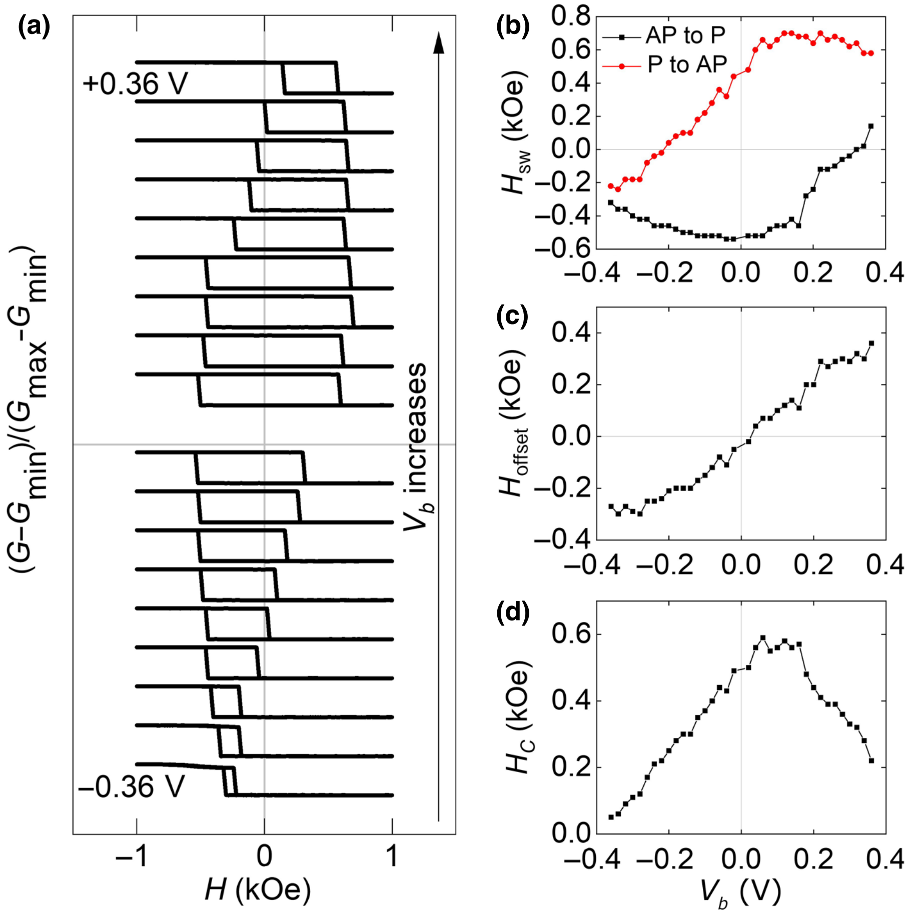

FIG. 4. (a) *G-H* loop at different V_b from −0.36 V to +0.36 V with a constant step, and extracted (b) H_{sw}, (c) H_{offset}, and (d) H_C values from (a).

$H_C = (H_{\mathrm{sw}}^{\mathrm{P}\to\mathrm{AP}} - H_{\mathrm{sw}}^{\mathrm{AP}\to\mathrm{P}})/2$ are extracted in Figs. 4(c) and 4(d), respectively. H_{offset}, which is dominated by the STT effect and the stray magnetic field of the device, is linearly proportional to V_b. H_C, which is dominated by the self-heating and VCMA effects, is a quadratic function of V_b. The reduced H_C at both bias polarities is mainly attributed to the self-heating effect induced by Joule heating when currents flow through the MgO barrier layer. The thermal energy is then directed to the free layer and weakens the PMA. The VCMA effect originates from the spin-orbit coupling effect when the charge accumulates at the free layer/MgO interface. The positive (negative) charges at the interface enhance (weaken) the PMA of the free layer and enlarge (shrink) H_C. Considering H_C is reduced at both polarities, the effect of self-heating is much stronger than that of VCMA, which only positively shifts the H_C vertex by 0.1 V in our device. To further analyze the self-heating effect, we apply various voltage biases with external magnetic fields and monitor the conductance of the same MTJ device, as shown in Figs. 5(a)–5(c). The external magnetic field is applied to compensate for the effect of nonzero H_{offset} at high V_b, which is induced by the stray magnetic field of the device and the STT effect. The nonzero H_{offset} leads to a preferred magnetization state in MTJ and suppresses oscillation at zero external magnetic field even though H_C is small enough due to the self-heating. If the magnitude of the external magnetic field is much larger than H_{offset}, magnetization follows the direction of the external magnetic field. Only when the external magnetic field is close to H_{offset} and the self-heating effect is strong does the unstable magnetization start to oscillate between the AP and P states. A duty cycle of 50% extracted from the telegraphic switching indicates that the applied external field compensates for the effect of nonzero H_{offset} exactly, and the switching is completely random depending on thermal perturbation [35]. For the MTJ neuron application, there are two favorable requirements. First, it is desired that the telegraphic switching occurs at zero external magnetic field. Thus, no external magnetic field is needed for neuron operations. This would require the opposite effects from the stray field and bias-dependent STT to cancel each other out. Therefore, the telegraphic switching always occurs at one instead of both polarities without an external field because the STT effect

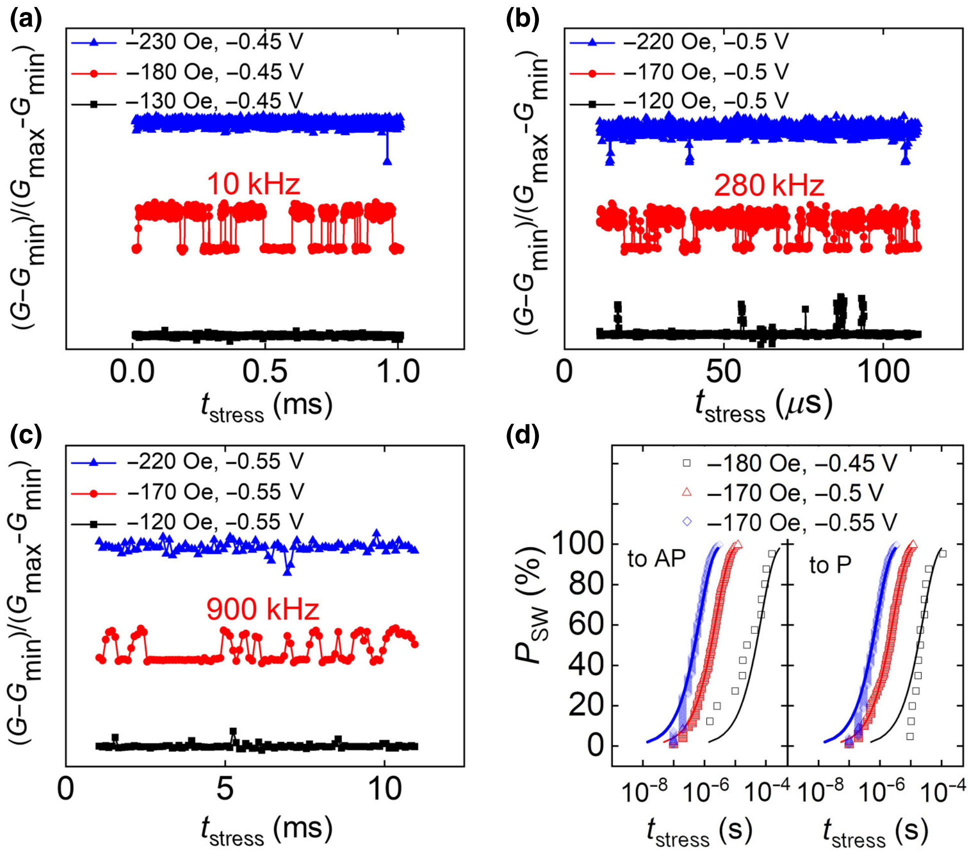

FIG. 5. (a)–(c) Sampled conductance of the same MTJ device at different bias voltages and different external magnetic fields. The red lines represent the conditions where the duty cycle is close to 50%. (d) P_{sw} vs t_{stress} to both the AP and P states extracted from the conditions with the duty cycle of 50% in (a)–(c). P_{sw} agrees well with the CDF of an exponential distribution statistically. The total sampling time is fixed at 1 ms for all conditions. The relatively poor fitting of the −180 Oe, −0.45 V condition with long switching times is likely due to the insufficient sampling points statistically.

is polarity-dependent [8]. The required stray field could be provided through the engineering of the SAF layer [36]. For example, two different samples measured in Fig. 3 (field-free) and Figs. 4 and 5 (field-assisted) have different SAF configurations. Both field-free and field-assisted telegraphic switching are possible. Second, the oscillation frequency of telegraphic switching should depend on V_b. Self-heating plays two important roles in telegraphic switching. It suppresses the PMA, thus lowering the energy barriers between the AP and P states. It also provides sufficient thermal energy for the magnetization to flip to the opposite state in a stochastic process. With the increase of V_b and self-heating, the frequency of telegraphic switching increases with the decrease of the PMA and the increase of thermal energy. Each switching time to the AP and P states in Figs. 5(a)–5(c) is extracted and P_{sw} to the AP and P states is plotted in Fig. 5(d), which follows a cumulative distribution function (CDF) of an exponential distribution [37] as,

$$P_{\text{sw}} = 1 - \exp\left(-\frac{t_{\text{stress}}}{\tau_{\text{sw}}}\right), \tag{1}$$

where τ_{sw} is the mean switching time to the AP and P states. The measurement result fits well with the CDF and suggests that the switching is random due to the thermal perturbation through a low energy barrier of a weakened PMA. The stochastic telegraphic switching of the MTJ is utilized to replace the integration capacitor, comparator, and reset circuit in the conventional CMOS neurons.

V. SPIKE FREQUENCY AND BIAS-DEPENDENT NÉEL-ARRHENIUS MODEL

To understand the dynamic of telegraphic switching in the MTJ, aiming for the potential of achieving high spike frequency, we extract the relation of τ_{sw} vs V_b in Fig. 6(a) and analyze the telegraphic switching process at three V_b in Fig. 6(b) by using high-speed (10 ns) and consecutive conductance readout. τ_{sw} could be further divided into the mean dwell time (τ_{dw}) and the mean transition time (τ_{tr}) [38]. During τ_{dw}, the MTJ conductance fluctuates at its respective LRS and HRS values. During τ_{tr}, the MTJ conductance is changed from one state to the other. τ_{dw} defines the incubation period when the magnetization direction vibrates due to thermal energy before flipping to

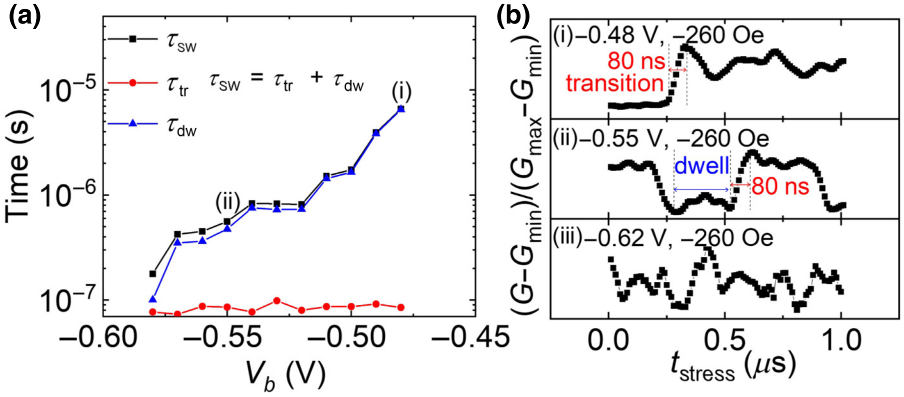

FIG. 6. (a) τ_{sw}, τ_{tr}, and τ_{dw} extracted from the sampling measurement in Fig. 5. (b) Enlarged waveforms of the sampling measurement and the definition of τ_{tr} and τ_{dw}. The switching in the bias regimes of (i),(ii) are bistate while that in the (iii) regime shows multistate conductance change.

the opposite state. The exact time of this incubation period in every switching event is random, but τ_{dw} is lowered as V_b increases. τ_{tr} defines the transition period of the magnetization flipping. The result shows that τ_{tr} is constant at different V_b, suggesting that τ_{tr} is dominated by the intrinsic damping of the free layer instead of the self-heating effect [39]. The damping constant is a material parameter related to the free layer interface [40,41]. Note that the MTJ spike frequency cannot keep increasing by simply increasing V_b. Although much faster than the previously reported value of 100 kHz [15], the measured spike frequency in our device saturates at approximately 10 MHz due to two reasons. First, the constant τ_{tr} of 80 ns eventually dominates τ_{sw} and limits the spike frequency. Second, the strong

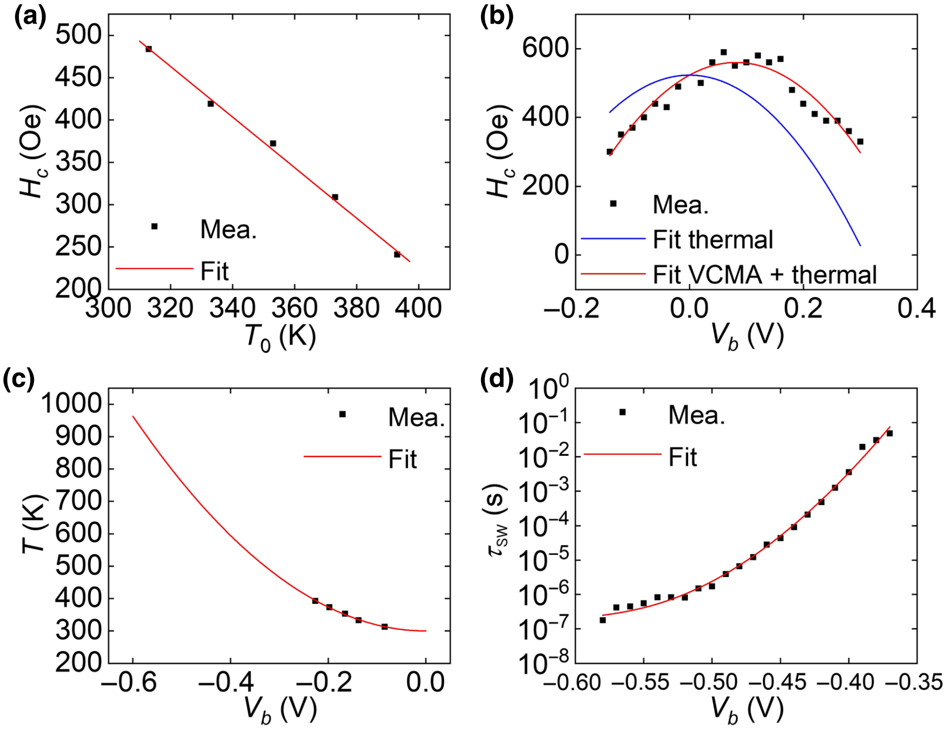

FIG. 7. Fitting of the bias-dependent Néel-Arrhenius model. (a) Linear fitting of H_C vs T, where T is the environmental temperature. The G-H curve is measured at V_b of 0.1 V to avoid Joule heating. (b) Fitting of H_C vs V_b at room temperature according to Eq. (5). H_C is affected by V_b through the VCMA and thermal effects. The contributions of these two effects are also simulated. (c) Fitting of T vs V_b according to Eq. (4). The implicit relation between T and V_b is obtained from H_C vs T and H_C vs V_b. (d) Fitting of τ_{sw} vs V_b according to Eqs. (2) and (3), and the T vs V_b relation in (c). A good agreement is obtained between the measurement and fitting. $\xi = 3$ and $K_{u0}V_F$= 1.98 eV are fitting parameters.

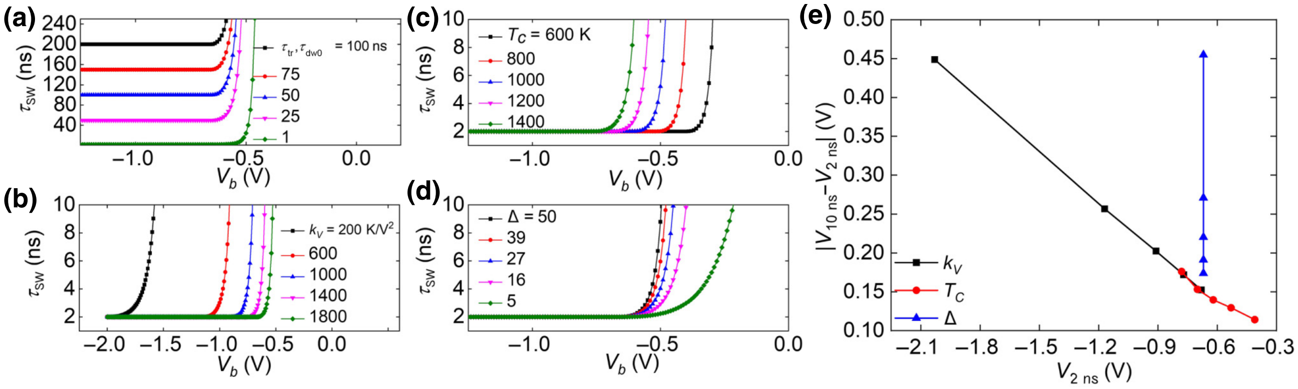

FIG. 8. τ_{sw} vs V_b simulated using the bias-dependent Néel-Arrhenius model by tuning (a) τ_{tr} and τ_{dw0}, (b) k_V, (c) T_C, and (d) Δ. k_V, T_C, and Δ obtained from the experimental fitting in Fig. 7 are used as default values while both τ_{tr} and τ_{dw0} of 1 ns are used as default values to explore the high-frequency capability toward the gigahertz range. When tuning one parameter, the other parameters are fixed at default values. (e) Sensing margin of V_b, i.e., $|V_{\text{10ns}}- V_{\text{2ns}}|$, vs V_{2ns} are extracted from (b)–(d). Large $|V_{\text{10ns}}- V_{\text{2ns}}|$ and small V_{2ns} are favorable for a wider detection range and robust reliability.

self-heating effect as V_b increases transforms the ferromagnetic phase of the free layer into the paramagnetic phase. Therefore, the MTJ starts to show multistate conductance change, for example, at $V_b = -0.62$ V in Fig. 6(b), instead of bistate telegraphic switching, thus is not suitable for spike neuron applications.

We establish an analytical compact model to explain the MTJ telegraphic switching quantitatively. We adopt the thermal Néel-Arrhenius model and calibrate it with the telegraphic switching measurement under voltage bias. The H_{offset} shifting due to the stray field and STT effects are compensated by an external magnetic field so that the thermal model is applicable to MTJ. τ_{sw} of MTJ following the Néel-Arrhenius law is expressed as [37]

$$\tau_{\text{sw}} = \tau_{\text{tr}} + \tau_{\text{dw0}} \exp\left(\frac{E_B}{k_B T}\right) = \tau_{\text{tr}} + \tau_{\text{dw0}} \exp\left(\frac{K_u V_F}{k_B T}\right), \tag{2}$$

where τ_{dw0} is the attempt period of incubation, E_B is the energy barrier of the PMA, k_B is the Boltzmann constant, T is the temperature, K_u is the energy density of the PMA, and V_F is the volume of the free layer. K_u includes the contribution from the perpendicular anisotropic field and saturation magnetization and is disturbed by the self-heating effect. By combining the Callen-Callen law and the Bloch law, K_u can be expressed as [42]

$$K_u = K_{u0}\left[1 - \left(\frac{T}{T_C}\right)^{\eta}\right]^{\xi}, \tag{3}$$

where K_{u0} is the K_u value at 0 K, T_C is the Curie temperature of the free layer, η is a fitting parameter obtained from the temperature-dependent measurement of saturation magnetization, and ξ is a power-law parameter fitted from Eqs. (2) and (3) [43]. The values of T_C of 1120 K and η of 1.64 are adopted from [44]. T is affected by the environmental temperature and the self-heating effect induced by Joule heating at the barrier layer when applying V_b. T follows the first law of thermodynamics and is simply described as [34]

$$T = T_0 + k_V V_b^2, \tag{4}$$

where k_V is a fitting parameter related to the MTJ resistance-area product (RA) value and the thermal capacitor. To extract k_V, we measure the relations of H_C vs T_0 and H_C vs V_b to obtain the implicit relation between T and V_b [45]. T_0 is varied by heating up the wafer substrate using a thermal chuck. First, the *G-H* loops are measured at 300 to 390 K using a small V_b of 0.1 V, and the H_C vs T relation is extracted in Fig. 7(a). Second, the *G-H* loops are measured at different V_b at room temperature, and then the H_C vs V_b relation is extracted in Fig. 7(b). The H_C vs V_b relation is affected by both the self-heating and VCMA effects and follows:

$$H_C = H_{C0} + k_{\text{VCMA}} V_b + k_{\text{th}} V_b^2, \tag{5}$$

where H_{C0} is H_C at room temperature without V_b, k_{VCMA} is the VCMA coefficient, and k_{th} is the thermal coefficient. The extracted k_{VCMA} is 901 Oe/V and k_{th} is 5505 Oe/V^2 in Fig. 7(b). k_{th} is much larger than k_{VCMA}, which is consistent with the telegraphic switching being mainly induced by self-heating instead of VCMA. Third, from the T vs V_b relation in Fig. 7(c), we extract k_V of 1841 K/V^2 by using Eq. (4). Note that T_C reported in the literature [44,46,47] is in the range of 800 to 1200 K, which corresponds to T at V_b of less than −0.5 V. In our measurement, the MTJ telegraphic switching transforms from bistate into multistate at $V_b = -0.62$ V, which corresponds to $T = 1000$ K. The results agree that the free layer is transformed from the ferromagnetic phase into the paramagnetic phase, i.e., E_B close to zero, above T_C. Finally, the measured τ_{sw} vs V_b is fit using the bias-dependent Néel-Arrhenius model from Eqs. (2)–(4), as shown in Fig. 7(d). When E_B is close to zero at large V_b, τ_{sw} is saturated to $\tau_{\mathrm{tr}}+\tau_{\mathrm{dw0}}$. The extracted τ_{tr} and τ_{dw0} of 80 and 100 ns limit the saturated spike frequency. Furthermore, it is worth noting that telegraphic switching is triggered by thermal fluctuation and thus the spike frequency is sensitive to T_0 [48]. A temperature compensation design that is commonly used in other analog circuits would be required for ensuring reliable operations. Because this compensation circuit could be used for multiple neurons across the chip, it will not significantly increase the area and power overhead.

To understand the design space and optimization direction of the MTJ neuron, the calibrated bias-dependent Néel-Arrhenius model obtained from the fitting of Fig. 7 is useful to predict the spike frequency and sensing margin of V_b as the functions of τ_{tr}, τ_{dw0}, k_V, T_C, and $\Delta(= K_u V_F/k_B T_0$ at room temperature). τ_{sw} of 2 to 10 ns is targeted for the desired high spike frequency, and the corresponding V_b is referred to as $V_{2\mathrm{ns}}$ and $V_{10\mathrm{ns}}$ for τ_{sw} of 2 and 10 ns, respectively. Because the V_b drop on the neuron device depends on the weight-summing current from the synaptic array, a larger sensing margin of V_b, i.e., $|V_{10\mathrm{ns}} - V_{2\mathrm{ns}}|$, allows a wider detection range of the weighted-sum current from synaptic arrays within a fixed neuron sensing time, e.g., 10 ns. If only a small range of synaptic current is detectable by the neuron, the functionality and accuracy of the SNN could be severely compromised. Therefore, the sensing margin is another important consideration for neuron applications. Some of the key findings are listed here. First, the saturation of τ_{sw} in Fig. 8(a) is determined by τ_{tr} and τ_{dw0} according to Eq. (2), which could be lowered by increasing the MTJ damping constant and adopting in-plane anisotropy MTJ [38,39]. A recent study demonstrated the feasibility of reducing τ_{tr} and τ_{dw0} down to 8 ns [47], highlighting the potential of the MTJ neuron for achieving gigahertz oscillation frequency. Second, k_V would affect T and E_B while T_C would affect E_B. The sensing margin of V_b increases by lowering k_V or increasing T_C, as shown in Figs. 8(b) and 8(c), respectively. However, the absolute values of V_b also increase. The upper bound of V_b is limited by the breakdown voltage of the MTJ, and thus it cannot be excessively high. In general, k_V is tunable by the MTJ RA value in a range of 200–1900 K/V^2 [34,45,49], while T_C is tunable by the thickness of the free layer in a range of $800-1200$ K [44,46,47]. Finally, lowering Δ increases the sensing margin of V_b without increasing the absolute value of V_b, as shown in Fig. 8(d). Δ could be reduced by appropriately scaling V_F or tuning the interfacial anisotropy of the free layer. However, the lower bound of Δ is set by the retention criterion if the same device is also used as a synapse. Figure 8(e) summarizes the effect of k_V, T_C, and Δ on the sensing margin of V_b. Overall, lowering τ_{tr}, τ_{dw0}, and Δ are most effective to increase the spike frequency and the sensing margin of V_b for achieving desired MTJ neuron properties.

VI. CONCLUSION

The MTJ neuron could be used as compact I&F and Poisson neurons for SNN applications with no need for additional integration capacitor, comparator, and reset circuit. Its telegraphic switching originates from the strong self-heating effect under external bias and achieves a high spike frequency of 10 MHz, one of the highest measured among the device-based spiking neurons. Moreover, a bias-dependent Néel-Arrhenius model is proposed and calibrated with the measured τ_{sw} vs V_b relation to quantitively describe the telegraphic switching. To further increase the spike frequency, aiming at lower latency, lower power consumption, and robust reliability, the design space for realizing an MTJ neuron with a CMOS neuron-compatible gigahertz spike frequency is explored by using the proposed model. For realizing compact and energy-efficient SNN hardware with high accuracy, future research should further address system-level concerns such as device and temperature variations in a large-scale SNN with numerous MTJ neurons, which is not discussed in this work. The offset cancelation and temperature compensation techniques that are commonly implemented in conventional CMOS-based analog circuits should be explored in designing a robust MTJ neuron.

ACKNOWLEDGMENT

This work was supported by the National Science and Technology Council, Taiwan under Grants No. 109-2221-E-009-020-MY3 and No. 111-2622-8-A49-018-SB, and also by the Industrial Technology Research Institute.

[1] W. Maass, Networks of spiking neurons: The third generation of neural network models, Neural Networks **10**, 1659 (1997).

[2] E. D. Adrian and Y. Zotterman, The impulses produced by sensory nerve endings, J. Physiol. **61**, 151 (1926).

[3] C.-C. Chang, J.-C. Liu, Y.-L. Shen, T. Chou, P.-C. Chen, I.-T. Wang, C.-C. Su, M.-H. Wu, B. Hudec, C.-C. Chang, *et al.*, in *Proceedings of the 2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2017), pp. 11.6.1–11.6.4.

[4] C.-C. Chang, M.-H. Wu, J.-W. Lin, C.-H. Li, V. Parmar, H.-Y. Lee, J.-H. Wei, S.-S. Sheu, M. Suri, T.-S. Chang, and T.-H. Hou, in *Proceedings of the 2019 56th ACM/IEEE Design Automation Conference (DAC)* (IEEE, Las Vegas, NV, 2019), pp. 1–6.

[5] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V.

Srikumar, ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars, ACM SIGARCH Comput. Archit. News **44,** 14 (2016).

[6] B. Yan, Q. Yang, W.-H. Chen, K.-T. Chang, J.-W. Su, C.-H. Hsu, S.-H. Li, H.-Y. Lee, S.-S. Sheu, M.-S. Ho, *et al.*, in *Proceedings of the 2019 Symposium on VLSI Technology* (IEEE, Kyoto, Japan, 2019), pp. T86–T87.

[7] M. Bouvier, A. Valentian, T. Mesquida, F. Rummens, M. Reyboz, E. Vianello, and E. Beigne, Spiking neural networks hardware implementations and challenges: A survey, ACM J. Emerg. Technol. Comput. Syst. **15,** 22 (2019).

[8] M.-H. Wu, M.-S. Huang, Z. Zhu, F.-X. Liang, M.-C. Hong, J. Deng, J.-H. Wei, S.-S. Sheu, C.-I. Wu, G. Liang, and T.-H. Hou, in *Proceedings of the 2020 Symposium on VLSI Technology* (IEEE, Honolulu, HI, 2020), pp. 1–2.

[9] Z. Wang, B. Crafton, J. Gomez, R. Xu, A. Luo, Z. Krivokapic, L. Martin, S. Datta, A. Raychowdhury, and A. I. Khan, in *Proceedings of the 2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2018), pp. 13.3.1–13.3.4.

[10] Z. Wang, S. Khandelwal, and A. I. Khan, Ferroelectric oscillators and their coupled networks, IEEE Electron Device Lett. **38,** 1614 (2017).

[11] Y. Fang, J. Gomez, Z. Wang, S. Datta, A. I. Khan, and A. Raychowdhury, Neuro-mimetic dynamics of a ferroelectric FET-based spiking neuron, IEEE Electron Device Lett. **40,** 1213 (2019).

[12] L. Gao, P.-Y. Chen, and S. Yu, NbO_x based oscillation neuron for neuromorphic computing, Appl. Phys. Lett. **111,** 103503 (2017).

[13] X. Zhang, W. Wang, Q. Liu, X. Zhao, J. Wei, R. Cao, Z. Yao, X. Zhu, F. Zhang, H. Lv, *et al*, An artificial neuron based on a threshold switching memristor, IEEE Electron Device Lett. **39,** 308 (2018).

[14] X. Zhang, Z. Wang, W. Song, R. Midya, Y. Zhuo, R. Wang, M. Rao, N. K. Upadhyay, Q. Xia, J. J. Yang, *et al.*, in *Proceedings of the 2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2019), pp. 6.7.1–6.7.4.

[15] M.-H. Wu, M.-C. Hong, C.-C. Chang, J.-H. Wei, H.-Y. Lee, S.-S. Sheu, and T.-H. Hou, in *Proceedings of the 2019 Symposium on VLSI Technology* (IEEE, Kyoto, Japan, 2019), pp. T34–T35.

[16] F.-X. Liang, I.-T. Wang, and T.-H. Hou, Progress and benchmark of spiking neuron devices and circuits, Adv. Intell. Syst. Comput. **3,** 2100007 (2021).

[17] C. Liu, B. Yan, C. Yang, L. Song, Z. Li, B. Liu, Y. Chen, and H. Li, in *Proceedings of the 2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)* (IEEE, San Francisco, CA, 2015), pp. 1–6.

[18] D. Vodenicarevic, N. Locatelli, A. Mizrahi, J. S. Friedman, A. F. Vincent, M. Romera, A. Fukushima, K. Yakushiji, H. Kubota, S. Yuasa, *et al.*, Low-Energy Truly Random Number Generation with Superparamagnetic Tunnel Junctions for Unconventional Computing, Phys. Rev. Appl. **8,** 054045 (2017).

[19] D. I. Suh, G. Y. Bae, H. S. Oh, and W. Park, Neural coding using telegraphic switching of magnetic tunnel junction, J. Appl. Phys. **117,** 17D714 (2015).

[20] C. M. Liyanagedera, A. Sengupta, A. Jaiswal, and K. Roy, Stochastic Spiking Neural Networks Enabled by Magnetic Tunnel Junctions: From Nontelegraphic to Telegraphic Switching Regimes, Phys. Rev. Appl. **8,** 064017 (2017).

[21] Y.-C. Shih, C.-F. Lee, Y.-A. Chang, P.-H. Lee, H.-J. Lin, Y.-L. Chen, C.-P. Lo, K.-F. Lin, T.-W. Chiang, Y.-J. Lee, *et al.*, in *Proceedings of the 2020 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2020), pp. 11.4.1–11.4.4.

[22] C.-H. Chen, C.-Y. Chang, C.-H. Weng, T.-H. Kuo, C.-Y. Wang, M.-C. Shih, T.-W. Chiang, Y.-J. Lee, R. Wang, K.-H. Shen, *et al.*, in *Proceedings of the 2021 Symposium on VLSI Technology* (IEEE, Kyoto, Japan, 2021), pp. 1–2.

[23] K. Lee, *et al.*, in *Proceedings of the 2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2019), pp. 2.2.1–2.2.4.

[24] T. Devolder, O. Bultynck, P. Bouquin, V. D. Nguyen, S. Rao, D. Wan, B. Sorée, I. P. Radu, G. S. Kar, and S. Couet, Back hopping in spin transfer torque switching of perpendicularly magnetized tunnel junctions, Phys. Rev. B **102,** 184406 (2020).

[25] K.-M. Chen, C.-W. Cheng, J.-H. Wei, Y.-C. Hsin, and Y.-C. Tseng, Effects of synthetic antiferromagnetic coupling on back-hopping of spin-transfer torque devices, Appl. Phys. Lett. **117,** 072405 (2020).

[26] J. Z. Sun, M. C. Gaidis, G. Hu, E. J. O'Sullivan, S. L. Brown, J. J. Nowak, P. L. Trouilloud, and D. C. Worledge, High-bias backhopping in nanosecond time-domain spin-torque switches of MgO-based magnetic tunnel junctions, J. Appl. Phys. **105,** 07D109 (2009).

[27] C. Abert, H. Sepehri-Amin, F. Bruckner, C. Vogler, M. Hayashi, and D. Suess, Back-Hopping in Spin-Transfer-Torque Devices: Possible Origin and Countermeasures, Phys. Rev. Appl. **9,** 054010 (2018).

[28] S.-H. Huang, K.-H. Shen, C.-W. Chien, S.-Y. Yang, J.-H. Shyu, D.-Y. Wang, K.-M. Kuo, T.-K. Ku, and D. Deng, in *Proceedings of the 2015 International Symposium on VLSI Technology, Systems and Applications* (IEEE, Hsinchu, Taiwan, 2015), pp. 1–2.

[29] T. Min, J. Z. Sun, R. Beach, D. Tang, and P. Wang, Back-hopping after spin torque transfer induced magnetization switching in magnetic tunneling junction cells, J. Appl. Phys. **105,** 07D126 (2009).

[30] B. R. Zink, Y. Lv, and J.-P. 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).

[31] B. R. Zink, Y. Lv, and J.-P. Wang, Telegraphic switching signals by magnet tunnel junctions for neural spiking signals with high information capacity, J. App. Phy. **124,** 152121 (2018).

[32] C.-W. Chien, D.-Y. Wang, S.-H. Huang, K.-H. Shen, S.-Y. Yang, J.-H. Shyu, C.-Y. Lo, K.-M. Kuo, Y.-S. Chen, Y.-H. Wang, *et al.*, Scaling properties of step-etch perpendicular magnetic tunnel junction with dual-CoFeB/MgO interfaces, IEEE Electron Device Lett. **35,** 738 (2014).

[33] M.-C. Hong, *et al.*, in *Proceedings of the 2022 Symposium on VLSI Technology* (IEEE, Kyoto, Japan, 2022), pp. 379–380.

[34] G. Mihajlović, N. Smith, T. Santos, J. Li, M. Tran, M. Carey, B. D. Terris, and J. A. Katine, Origin of the Resistance-Area-Product Dependence of Spin-Transfer-Torque Switching in Perpendicular Magnetic Random-Access Memory Cells, Phys. Rev. Appl. **13**, 024004 (2020).

[35] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevApplied.18.064034 for the method of extracting the mean switching time τ_{sw}.

[36] M. Yoshikawa, T. Kai, M. Amano, E. Kitagawa, T. Nagase, M. Nakayama, S. Takahashi, T. Ueda, T. Kishi, K. Tsuchida, *et al.*, Bit yield improvement by precise control of stray fields from SAF pinned layers for high-density MRAMs, J. Appl. Phys. **97**, 10P508 (2005).

[37] Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L.-C. Wang, and Y. Huai, Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory, J. Phys.: Condens. Matter **19**, 165209 (2007).

[38] S. Kanai, K. Hayakawa, H. Ohno, and S. Fukami, Theory of relaxation time of stochastic nanomagnets, Phys. Rev. B **103**, 094423 (2021).

[39] T. Taniguchi and H. Imamura, Thermal switching rate of a ferromagnetic material with uniaxial anisotropy, Phys. Rev. B **85**, 184403 (2012).

[40] H. Naganuma, S. Miura, H. Honjo, K. Nishioka, T. Watanabe, T. Nasuno, H. Inoue, T. V. A. Nguyen, Y. Endo, Y. Noguchi, *et al.*, in *Proceedings of the 2021 Symposium on VLSI Technology* (IEEE, Kyoto, Japan, 2021), pp. 1–2.

[41] S. Couet, T. Devolder, J. Swerts, S. Mertens, T. Lin, E. Liu, S. V. Elshocht, and G. S. Kar, Impact of Ta and W-based spacers in double MgO STT-MRAM free layers on perpendicular anisotropy and damping, Appl. Phys. Lett. **111**, 152406 (2017).

[42] H. B. Callen and E. Callen, The present status of the temperature dependence of magnetocrystalline anisotropy, and the l (l+ 1) 2 power law, J. Phys. Chem. Solids **27**, 1271 (1966).

[43] E. R. Callen and H. B. Callen, Anisotropic magnetization, J. Phys. Chem. Solids **16**, 310 (1960).

[44] J. G. Alzate, P. K. Amiri, G. Yu, P. Upadhyaya, J. A. Katine, J. Langer, B. Ocker, I. N. Krivorotov, and K. L. Wang, Temperature dependence of the voltage-controlled perpendicular anisotropy in nanoscale MgO|CoFeB|Ta magnetic tunnel junctions, Appl. Phys. Lett. **104**, 112410 (2014).

[45] N. Strelkov, A. Chavent, A. Timopheev, R. C. Sousa, I. L. Prejbeanu, L. D. Buda-Prejbeanu, and B. Dieny, Impact of Joule heating on the stability phase diagrams of perpendicular magnetic tunnel junctions, Phys. Rev. B **98**, 214410 (2018).

[46] K.-M. Lee, J. W. Choi, J. Sok, and B.-C. Min, Temperature dependence of the interfacial magnetic anisotropy in W/CoFeB/MgO, AIP Adv. **7**, 065107 (2017).

[47] K. Hayakawa, S. Kanai, T. Funatsu, J. Igarashi, B. Jinnai, W. A. Borders, H. Ohno, and S. Fukami, Nanosecond Random Telegraph Noise in In-Plane Magnetic Tunnel Junctions, Phys. Rev. Lett. **126**, 117202 (2021).

[48] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevApplied.18.064034 for the spiking frequency at different environmental temperatures T_0.

[49] A. Chavent, C. Ducruet, C. Portemont, L. Vila, J. Alvarez-Hérault, R. Sousa, I. L. Prejbeanu, and B. Dieny, Steady State and Dynamics of Joule Heating in Magnetic Tunnel Junctions Observed via the Temperature Dependence of RKKY Coupling, Phys. Rev. Appl. **6**, 034003 (2016).