A domain wall-magnetic tunnel junction artificial synapse with notched geometry for accurate and efficient training of deep neural networks

Cite as: Appl. Phys. Lett. **118**, 202405 (2021); https://doi.org/10.1063/5.0046032
Submitted: 31 January 2021 . Accepted: 01 May 2021 . Published Online: 21 May 2021

Samuel Liu, T. Patrick Xiao, Can Cui, Jean Anne C. Incorvia, Christopher H. Bennett, and Matthew J. Marinella

Applied Physics Letters ARTICLE scitation.org/journal/apl

A domain wall-magnetic tunnel junction artificial synapse with notched geometry for accurate and efficient training of deep neural networks

Cite as: Appl. Phys. Lett. **118**, 202405 (2021); doi: 10.1063/5.0046032
Submitted: 31 January 2021 · Accepted: 1 May 2021 ·
Published Online: 21 May 2021

Samuel Liu,[1,a)] T. Patrick Xiao,[2] Can Cui,[1] Jean Anne C. Incorvia,[1] Christopher H. Bennett,[2] and Matthew J. Marinella[2]

AFFILIATIONS

[1]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78712, USA
[2]Sandia National Laboratories, Albuquerque, New Mexico 87123, USA

Note: This paper is part of the APL Special Collection on Mesoscopic Magnetic Systems: From Fundamental Properties to Devices.
[a)]**Author to whom correspondence should be addressed:** liukts@utexas.edu

ABSTRACT

Inspired by the parallelism and efficiency of the brain, several candidates for artificial synapse devices have been developed for neuromorphic computing, yet a nonlinear and asymmetric synaptic response curve precludes their use for backpropagation, the foundation of modern supervised learning. Spintronic devices—which benefit from high endurance, low power consumption, low latency, and CMOS compatibility—are a promising technology for memory, and domain-wall magnetic tunnel junction (DW-MTJ) devices have been shown to implement synaptic functions such as long-term potentiation and spike-timing dependent plasticity. In this work, we propose a notched DW-MTJ synapse as a candidate for supervised learning. Using micromagnetic simulations at room temperature, we show that notched synapses ensure the non-volatility of the synaptic weight and allow for highly linear, symmetric, and reproducible weight updates using either spin transfer torque (STT) or spin–orbit torque (SOT) mechanisms of DW propagation. We use lookup tables constructed from micromagnetics simulations to model the training of neural networks built with DW-MTJ synapses on both the MNIST and Fashion-MNIST image classification tasks. Accounting for thermal noise and realistic process variations, the DW-MTJ devices achieve classification accuracy close to ideal floating-point updates using both STT and SOT devices at room temperature and at 400 K. Our work establishes the basis for a magnetic artificial synapse that can eventually lead to hardware neural networks with fully spintronic matrix operations implementing machine learning.

In-memory computing overcomes the memory wall in von Neumann architectures, where the data-intensive computation is frequently bottlenecked by slow and expensive memory accesses.[1] Foremost among these data-driven applications is the processing of artificial neural networks, modeled after biological neurons interconnected by tunable synapses. Massively parallel analog computation within a resistive memory array, where the memory devices serve as synapses, is a promising approach to lower the energy consumption of training and deploying neural networks.[2] Nonvolatile memory devices such as resistive random access memory (ReRAM),[3,4] phase change memory (PCM),[5,6] conductive bridge RAM (CBRAM),[7] and electrochemical or polymer-based memory[8–10] have all been demonstrated to implement multi-level synaptic functionality and, in many cases, adequate cycle-to-cycle variability. However, many of these devices exhibit nonlinear and/or asymmetric responses to programing pulses, making it difficult to accurately implement the ubiquitous backpropagation algorithm for neural network training.[11,12] These drawbacks, along with high write voltages or currents,[5] diminish the energy benefits of nonvolatile memory-based training accelerators and limit their generalizability to complex machine learning problems.

Spintronic memory has attracted interest for its high write endurance, low power consumption, and small size. For neuromorphic applications, domain wall-magnetic tunnel junction (DW-MTJ) devices[13,14] have previously been shown to emulate leaky integrate and fire (LIF) neuron functionality[15–17] as well as long-term potentiation (LTP) and spike-timing dependent plasticity (STDP) synaptic behaviors.[18] In contrast to two-terminal MTJs, the three-terminal DW-MTJ enables isolation of the read and write paths, contributing to reduced

wear on the MgO tunnel barrier. The DW-MTJ device, shown in Fig. 1, consists of a perpendicularly magnetized ferromagnetic (FM) track containing a DW, separated from a fixed FM layer by a thin MgO barrier. Current applied between the L and R terminals propagates the DW through spin transfer torque (STT). With the inclusion of a heavy metal (HM) layer underneath the FM track, DW motion can also be induced by spin–orbit torque (SOT) at a lower current density. The conductance of the MTJ stack, which represents the synaptic weight, is determined by the DW position and can be read by a small vertical current between OUT and either L or R that does not displace the DW. To date, most work on DW synaptic neuromorphic systems have focused on small-scale implementations of bio-inspired learning rules, e.g., STDP or STDP-like rules,[19–21] rather than on state-of-the-art deep neural networks that can be applied to complex, high-dimensional machine learning tasks.[22] Supervised learning systems proposed so far with DW synapses[23] have not considered the implications of room-temperature drift, stochasticity, and process variations on the feasibility of these systems.

In this Letter, we show using micromagnetic simulations that a DW-MTJ device can function effectively as a synapse with a high intrinsic linearity and symmetry. Notches are added to the track to stabilize the DW position and to improve the linearity and repeatability of synaptic updates. We build lookup-table models of device behavior, based on micromagnetic simulations of both STT- and SOT-driven DW motion, which capture thermally induced cycle-to-cycle variability of DW updates as well as process-induced device-to-device variability in the MTJ conductance. Next, we use these lookup tables to simulate the training of DW-MTJ synapses on the MNIST[24] and Fashion-MNIST[25] image classification datasets and evaluate their accuracy.

The magnetization dynamics of DW-MTJ artificial synapses are modeled using MuMax3, a micromagnetics simulator that solves the Landau–Lifshitz–Gilbert (LLG) equation.[26] The free FM layer is a rectangular wire that is 1050 nm long, 50 nm wide, and 1.5 nm thick for STT (1750 nm long for SOT). The wire is bounded on both ends by 30 nm long regions of fixed magnetization (50 nm for SOT). The free layer is assumed to be CoFeB (exchange stiffness $A_{ex} = 1.3 \times 10^{-11}$ J/m, saturation magnetization $M_{sat} = 0.8 \times 10^{6}$ A/m, magnetocrystalline anisotropy $\hat{z}K_u = 5 \times 10^{5}$ J/m^3, Gilbert damping factor $\alpha = 0.05$, non-adiabaticity factor $\xi = 0.05$, and spin polarization $P = 0.7$), while the HM layer for the SOT device is Ta [spin Hall angle $\theta_H = 0.2$ and

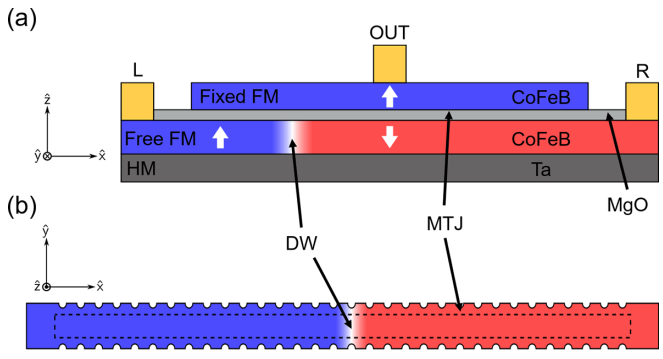

FIG. 1. (a) Side and (b) top profile of DW-MTJ artificial synapse. The dotted rectangle depicts the placement of the MTJ over the FM track.

interfacial Dzyaloshinskii–Moriya interaction (DMI) $D_{ind} = -0.5 \times 10^{-3}$ J/m^2]. The HM layer has the same thickness and resistivity as the FM, so that half of the injected current acts on the DW by STT and the other half by SOT. The DW position is approximated using the average magnetization of the wire along its length.

In a perfectly smooth FM wire and in the absence of a driving current or field, the DW tends to drift toward the center to minimize its interaction energy with the pinned magnetization regions. This is illustrated by the blue curve in Fig. 2(a), where the DW is initialized at the right edge ($x \approx 960$ nm) and drifts by approximately 200 nm over 50 ns. By adding semi-circular notches to the edges of the track that act as pinning sites, shown in Fig. 1(b), the DW position can be made nonvolatile. A further benefit of the notches is shown in Fig. 2(b), where the DW is driven by a sequence of 1 ns long 50 μA pulses separated by 4 ns relaxation periods. Without notches, the DW updates are nonlinear, controlled both by the applied current and its position along the track, which determines the rate of drift. The notches ensure linear updates independent of position and enable lithographic control of synaptic weight values. Other methods for DW pinning include using interlayer exchange coupling of multiple MTJs or by introducing defects in the shape or anisotropy of the free layer.[23,28] However, notches were chosen as the preferred pinning method because their controllable positions guarantee update linearity, which cannot readily be obtained using randomly distributed defects. In addition, notches are less complex to lithographically define than interlayer exchange and can be more easily scaled.

At non-zero ambient temperatures, thermal fluctuations can cause spontaneous depinning of the DW. As a result, the notch must provide a sufficiently deep energy well to ensure synaptic nonvolatility. The notch spacing must also be greater than the DW width to prevent drift. This is particularly important for SOT-driven devices, since the tilting of the DW due to DMI during and after current injection can cause uncontrolled movement between notches. Figure 2(c) illustrates this with snapshots of the $\hat{z}$ magnetization of a section of the track. Here, a 0.5 ns long 27 μA pulse is applied to an SOT device with a 30 nm notch spacing. With this spacing, the DW interacts with both adjacent notches and experiences tilting long after the current stimulus has ceased, eventually settling unpredictably to one of the notches. For the SOT device, notches with a 5 nm radius-spaced 50 nm apart are necessary to suppress the effect of thermal fluctuations, while for the STT device, a notch spacing of 30 nm is sufficient. These represent lower bounds on notch spacing that provide nonvolatility, predictable updates, and the availability of many states along the track. These notch spacings are attainable using electron beam lithography and ion mill etching; well-controlled nanomagnetic feature sizes down to 10 nm spacing have been demonstrated, and MTJs have been fabricated with diameters below 50 nm.[29,30]

The synaptic functionality of DW-MTJ devices is demonstrated using both STT- and SOT-driven devices with 32 equally spaced notches (32 weight levels). Synaptic updates are characterized using a sequence of positive current pulses followed by negative pulses, which ramp the DW position along the track. To shift the DW-MTJ by one weight level, the pulse duration is fixed to 1 ns for STT devices (0.5 ns for SOT), and the amplitude is set to 50 μA for STT (27 μA for SOT). The ramp is repeated 30 times to quantify the cycle-to-cycle variability in the update amount, which can be induced by thermal noise.

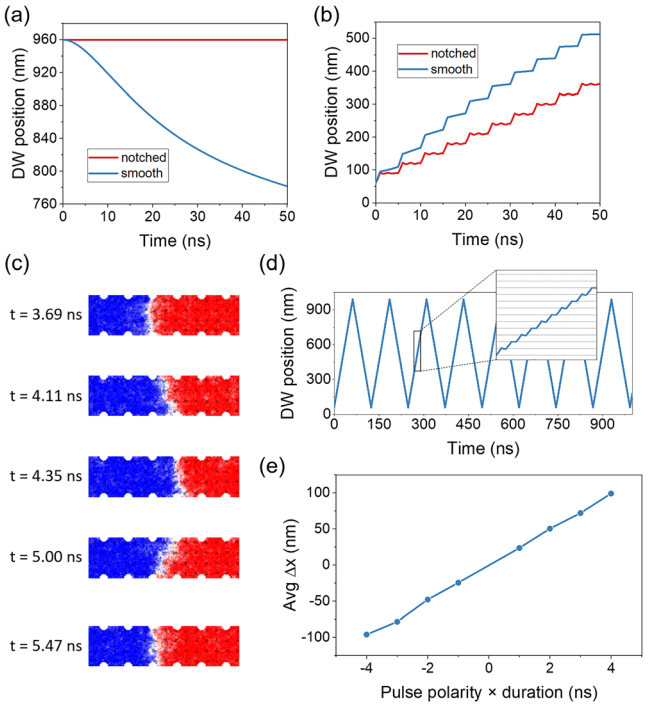

FIG. 2. (a) DW drift in smooth and notched DW-MTJ devices with no applied current. (b) DW response to a current pulse train, for smooth and notched DW-MTJ devices. (c) DW tilting in an SOT-driven DW-MTJ device. (d) Ramping of the DW position through 32 discrete levels using periodic positive and negative current pulses. The horizontal lines in the inset denote notch locations. (e) The change in the DW position Δx, averaged over 32 notch positions, is a linear function of pulse duration for both pulse polarities.

Figure 2(d) shows the ramp of DW position at 0 K for the STT device. In the inset, the notch positions depicted by the horizontal lines show the lithographically defined levels to which the DW can settle. The linear slope of the ramp in both directions indicates that synaptic weight changes are both highly linear (minimal state-dependence) and highly symmetric (same response in both directions). This suggests that DW-MTJ synapses can implement backpropagation with high fidelity.

The size of a weight update Δx can be linearly modulated using the magnitude or duration of the applied current pulse. This follows from the relationship between DW velocity and current density described by Beach *et al.*:[31]

$$\bar{v}_J = \frac{g\mu_B P}{2eM_{sat}} j, \tag{1}$$

where g is the Landé factor, μ_B is the Bohr magneton, P is the spin polarization, and e is the electron charge. To validate this property, positive and negative pulses of varying duration were applied to the DW at each of the 32 notches, and the average Δx for each duration is computed. The result, shown in Fig. 2(e), confirms that Δx varies linearly with pulse duration and that this response is symmetric for the two update polarities.

The DW position x is converted to a synapse conductance by treating the MTJ as two resistors in parallel: one over the region where the free and fixed FM layers have parallel magnetizations and one over the region where they are anti-parallel. The DW positions collected from 30 ramps are used to construct a probabilistic lookup table of the change in conductance ΔG for each initial conductance G. Figure 3(a) shows the lookup tables for STT and SOT devices at 0 K and 300 K (at 0 K, SOT behaves similarly to STT). Temperature introduces stochasticity to the updates, and this is significantly more pronounced in the SOT devices where, due to DMI, the DW prefers a Néel geometry.

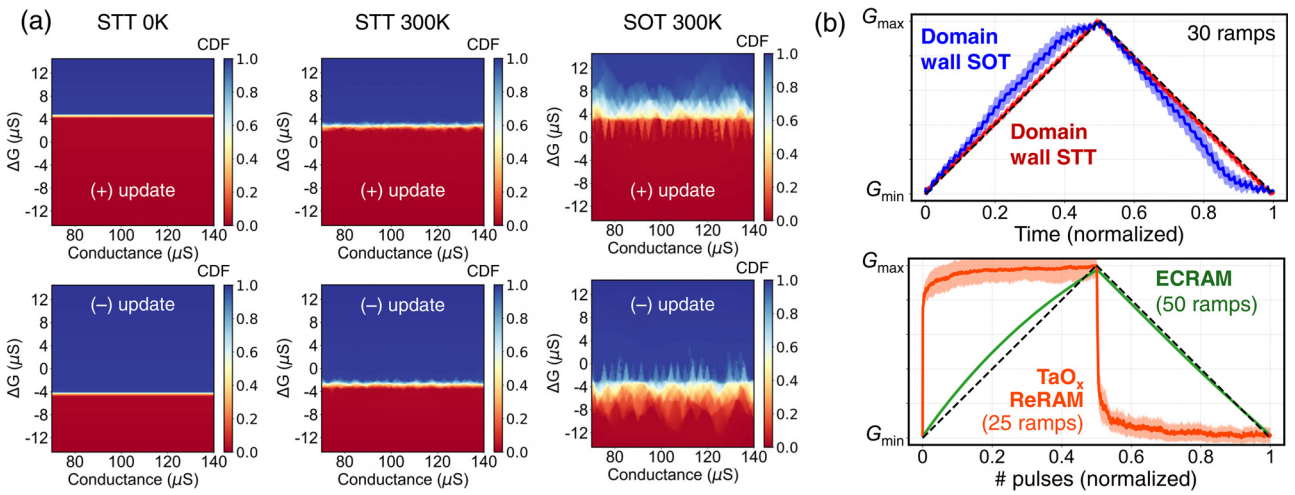

FIG. 3. (a) Conductance G vs ΔG statistics of the DW-MTJ at 0 K and 300 K for STT and SOT propagation. Each heatmap is constructed from 930 data points from micromagnetic simulations, and the color represents the cumulative distribution function (CDF) of ΔG at each conductance, where the CDF is the probability that a given conductance update is less than or equal to the plotted value of ΔG. In training simulations, the CDF is randomly sampled for each synapse to obtain the value of an update. (b) Comparison of update linearity and symmetry with experimental data from two other synaptic devices: ECRAM[10] and TaO_x ReRAM.[27] The ECRAM data are freely available on Materials Commons, and the TaO_x ramp data were taken from a larger database of TaO_x device ramp data, whose neuromorphic implications are summarized in Bennett *et al.*[27] The axes are normalized to allow a visual comparison. The shaded regions indicate the amount of cycle-to-cycle variability (two standard deviations) at each position within a ramp, taken over the indicated number of ramps.

Thermal fluctuations in the DW magnetization, together with its interaction with the notches, can randomly cause the DW to either not be displaced or to propagate multiple levels in order to maintain a Néel configuration. Nonetheless, when averaged over the noise, all of the lookup tables show highly linear updates, indicated by an expected value of ΔG that is nearly independent of G.

Figure 3(b) compares the simulated ramp response of DW devices at 300 K with experimentally measured ramp data from two previously published devices: electrochemical RAM (ECRAM),[10] which is highly linear and reproducible, and TaO_x ReRAM,[27] which is highly nonlinear and asymmetric. Table I compares the linearity, symmetry, and stochasticity of several published devices, where the parameters are extracted as described in Ref. 12. The DW synapse exhibits greater write noise than ECRAM, but using STT motion, the device has excellent linearity and symmetry in comparison to the best demonstrated synaptic devices.

We sample the generated lookup tables to simulate the training of DW-MTJ synapse arrays using CrossSim.[33] To model device-to-device variation within the array induced by process variations, each device is assigned a different perturbed lookup table. The perturbations are added as random variations in the MTJ parallel resistance R_p (12.8%) and the tunnel magnetoresistance ratio TMR (6.9%). We assume normally distributed variations with magnitudes obtained from Ref. 34 at a 45 nm critical dimension, which is slightly less than the track width. For computational tractability, we generate 20 perturbed lookup tables for each combination of technology (STT/SOT) and temperature and assign them randomly to devices in the array.[27]

As shown in Fig. 4(a), the matrix–vector multiplication can be executed on the DW-MTJ array during forward propagation by reading the MTJ resistance (OUT terminal). Weight updates are performed using the L and R terminals of the track, as shown in Fig. 4(b). Using backpropagation with stochastic gradient descent (SGD), each update to the weight matrix is an outer product of two vectors. A parallel outer product update can be efficiently executed in a DW-MTJ array by simultaneously driving the L terminals (rows) and the R terminals (columns) with time-coded and voltage-coded pulses, respectively, to obtain a multiplicative effect.[35]

The DW-MTJ synapse is evaluated on two image classification tasks—MNIST handwritten digits [see Fig. 5(a)] and the more difficult Fashion-MNIST clothing items [see Fig. 5(b)]—using the same two-layer multilayer perceptron topology with 300 hidden neurons. Each network consists of a 785×300 and a 301×10 weight matrix (including bias). A sigmoid and a softmax activation, after the first and second layers, respectively, are computed digitally at floating-point precision. Signed weights are implemented using the difference in conductance of two DW-MTJ devices: $W_{i,j} = G^+_{i,j} - G^-_{i,j}$. The complementary weight components are placed in two separate arrays, and updates are always applied to both halves of a synapse to prevent conductance saturation.[11] SGD is used with a fixed learning rate schedule for all simulations: the learning rate begins as α and is reduced to $\alpha/2$, $\alpha/3$, $\alpha/4$, and $\alpha/5$ after the third, fifth, eighth, and tenth training epochs, respectively.

Figures 5(a) and 5(b) show the training performance of STT DW-MTJ synapse arrays at 300 K compared to training with ideal numeric updates. We have used the MNIST and Fashion-MNIST test sets for validation. For each series, three networks are trained with random initial seeds; the data points show the average, while the colored

TABLE I. Comparison of simulated update properties of DW devices at 300 K with several published synaptic devices. The nonlinearity parameter is defined as in Ref. 12. For an ideal symmetric response, both the sign and magnitude of the nonlinearity are equal.

Synapse device	Nonlinearity (+/− updates)	Cycle-to-cycle variation
Domain wall STT	+0.07/−0.15	0.77%
Domain wall SOT	+0.80/−0.81	3.23%
ECRAM[10]	+0.70/−0.12	0.023%
TaO_x/HfO_x ReRAM[32] (analysis by Ref. 12)	+0.04/−0.63	3.70%
TaO_x ReRAM[27]	+668/−51.7	11.2%

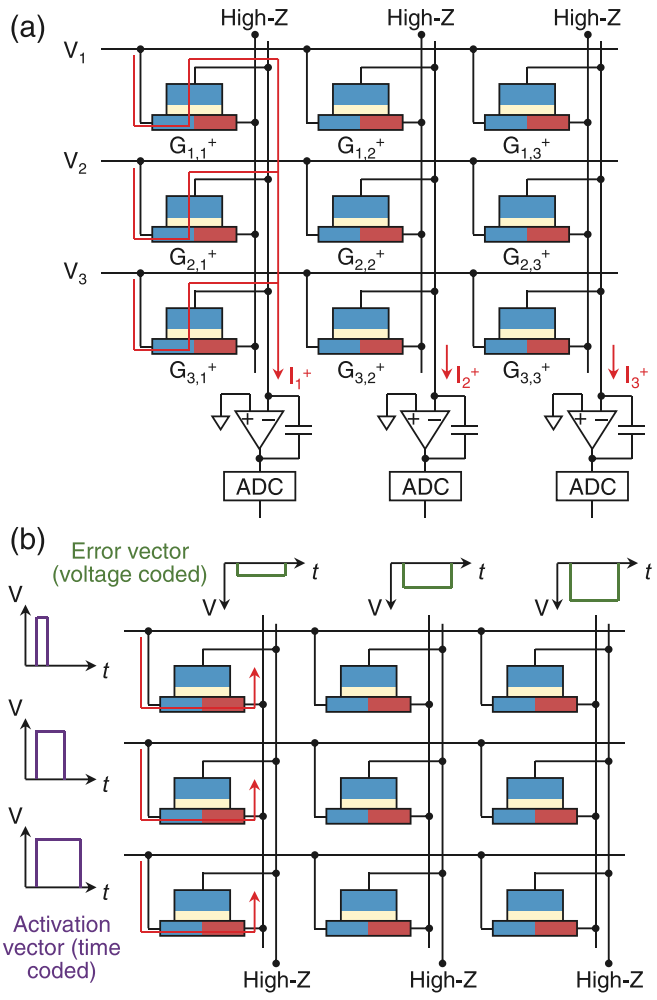

FIG. 4. Crossbar array of DW-MTJ synapses used in (a) matrix–vector multiplication mode and (b) outer product update mode. In (a), summed currents are integrated on a capacitor and digitized by an analog-to-digital converter (ADC), as in Ref. 36.

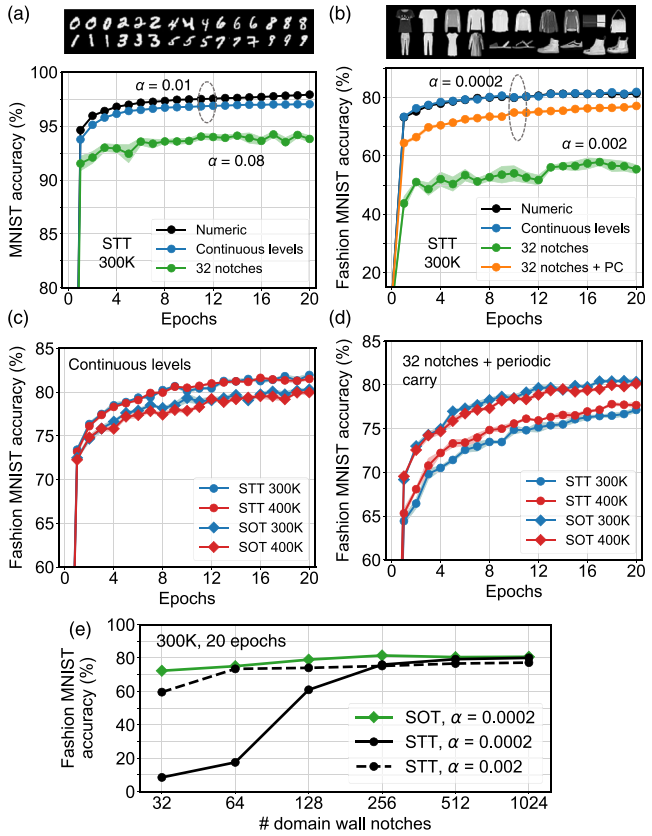

FIG. 5. Validation accuracy of a two-layer multilayer perceptron (MLP) (785 × 300, 301 × 10) of STT DW-MTJ synapses at 300 K for (a) MNIST and (b) Fashion-MNIST over 20 training epochs. PC indicates periodic carry. α is the initial learning rate. (c) Fashion-MNIST accuracy for different update mechanisms (SOT/STT) and temperatures assuming continuous weight levels, and (d) assuming 32 notches with periodic carry. (e) Fashion-MNIST accuracy (best of 20 epochs) vs the number of synaptic levels in STT and SOT DW-MTJ devices at 300 K without periodic carry.

areas signify standard deviation. If the DW-MTJ is idealized to have continuous levels without drift, the performance is very close to ideal even with cycle-to-cycle and device-to-device variations introduced by temperature and MTJ variations, respectively. The resilience to MTJ process variation arises from the high linearity of the devices: even if the same conductance maps to different DW positions in different devices, the update strength will be the same since it is largely independent of the starting state.

The geometry with 32 notches (green) has a discretizing effect on DW position: the updated conductance is rounded to the closest discrete level. For this case, the learning rate α is increased to prevent a large number of small updates from being reduced to zero; however, this results in inferior convergence relative to the continuous case. In both classification tasks, the notched synapses suffer a significant performance loss even with an optimized learning rate, with a greater loss (>20%) for the Fashion-MNIST task. The accuracy can be partially recovered using periodic carry, which splits the high and low significance bits of each weight into two devices with 32 notches each,[33] increasing the effective weight resolution of each nanosynapse. With increasing dataset and neural network complexity, more weight levels (notches) are needed to obtain ideal numeric accuracy. Indeed, back-propagation using SGD typically has a clear lower bound on allowable bit resolution.[36,37]

Figures 5(c) and 5(d) compare the training performance on Fashion-MNIST of STT and SOT devices and different temperatures, assuming ideal continuous and notched synapses with periodic carry, respectively. In the continuous case, the superior accuracy of STT devices in Fig. 5(c) results from their smaller cycle-to-cycle variability in ΔG. For both device types, the accuracy is roughly the same for 300 K and 400 K, which reflects the similarity in their lookup tables. On the other hand, for the notched devices in Fig. 5(d), a higher accuracy is attained with SOT than with STT. This arises from the fact that when the desired synapse update ΔG is small, the more stochastic SOT device is more likely to yield a non-zero conductance update than the STT device, where many of the updates will be too small to move the DW.

To reduce the accuracy loss caused by discretization, an alternative to periodic carry is to use a longer FM wire with more notches to increase the available number of weight levels. Figure 5(e) shows the effect of the number of notches on Fashion-MNIST accuracy using STT and SOT devices at 300 K. Surprisingly, the SOT device attains a much higher accuracy than the STT device when the number of notches is small; with just 32 notches, an accuracy of 72% is achieved on Fashion-MNIST compared to 10% using STT at the same learning rate. We attribute this to the greater stochasticity of the SOT mechanism, which allows small device updates that would otherwise fail to move the DW to the next notch to occasionally produce a change in conductance. As with the stochastic rounding technique used in software,[36] when this effect is averaged over thousands of updates, an effectively higher resolution is achieved for the weight updates than the actual number of notches present. Noisy updates have also been shown to reduce overfitting.[38] This benefit can be approximated using STT with a higher learning rate (dashed curve), but with a lower accuracy at both a small and large number of notches relative to SOT.

Material parameters may also influence the neuromorphic performance and efficiency of the DW-MTJ synapse. Assuming a free layer with perpendicular anisotropy, both the DW velocity and the DW width depend on the track's magnetic properties. Based on Eq. (1), an increase in P or a decrease in M_{sat} would increase the DW velocity for a given applied current, allowing the same weight update to be performed with a lower energy. For small notch spacings close to half the DW width, an increase in DW width can increase the stochasticity of a weight update. The expression for DW width $\delta = \pi\sqrt{A/K}$ indicates that choosing a material with increased exchange stiffness A or reduced perpendicular anisotropy K can lead to more stochastic updates with the same device geometry. Additionally, in choosing the material for the HM layer, the torque contribution from the spin Hall effect described in Ref. 39 can be used to deduce the relevant material parameters,

$$\tau_1^{SHE} = \frac{\hbar\theta_H j}{2eM_{sat}}. \quad (2)$$

Choosing a material with larger spin Hall angle θ_H leads to an increased DW velocity, increasing the energy efficiency of a weight update. The HM layer also mediates the magnitude of DMI, which is a large contributor to the stochasticity found in SOT devices. By choosing a material that induces stronger or weaker DMI, the stochasticity

of a weight update can be augmented or reduced. In addition to increasing the effective weight update resolution, tunable stochasticity can also enable efficient implementations of probabilistic learning algorithms.

In summary, our micromagnetics-based modeling of DW-MTJ nanosynapses with a notched geometry demonstrates their suitability for on-chip learning using backpropagation. Well-engineered notches eliminate DW drift in both STT and SOT DW-MTJs, bestowing synaptic non-volatility. Device lookup tables constructed from micromagnetics simulations of 32-level devices display highly linear and symmetric synaptic response, leading to classification accuracies approaching ideal numeric performance on the MNIST task. When taking into account the pinning of DWs to discrete notches, there is an accuracy penalty for more complex tasks such as Fashion-MNIST. This penalty could be alleviated by increasing the weight resolution (adding more notches), using multiple devices to represent the synapse bits (periodic carry), or by exploiting the stochasticity that is inherent to the physics of SOT devices. Since our results imply that discretization due to notches is the major roadblock to software-equivalent neural network performance, the effect of stochastic rounding will be investigated in future work to mitigate this drawback while retaining the increased linearity of a notched geometry. Overall, our physics-rich neural network simulations may be a foundational step in the realization of analog spintronic neuromorphic computation.

The authors acknowledge the support from Sandia's Laboratory-Directed Research and Development program, funding from the National Science Foundation CAREER under the Award No. 1940788, and computing resources from the Texas Advanced Computing Center (TACC) at the University of Texas at Austin (http://www.tacc.utexas.edu). This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in this paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by NTESS, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under Contract No. DE-NA0003525.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]W. A. Wulf and S. A. McKee, "Hitting the memory wall," ACM SIGARCH Comput. Archit. News **23**, 20–24 (1995).
[2]T. P. Xiao, C. H. Bennett, B. Feinberg, S. Agarwal, and M. J. Marinella, "Analog architectures for neural network acceleration based on non-volatile memory," Appl. Phys. Rev. **7**, 031301 (2020).
[3]H. Akinaga and H. Shima, "Resistive random access memory (ReRAM) based on metal oxides," Proc. IEEE **98**, 2237–2251 (2010).
[4]F. Pan, S. Gao, C. Chen, C. Song, and F. Zeng, "Recent progress in resistive random access memories: Materials, switching mechanisms, and performance," Mater. Sci. Eng., R **83**, 1–59 (2014).
[5]H. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Phase change memory," Proc. IEEE **98**, 2201–2227 (2010).
[6]G. W. Burr, M. J. Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson, B. Kurdi, C. Lam, L. A. Lastras, A. Padilla, B. Rajendran, S. Raoux, and R. S. Shenoy, "Phase change memory technology," J. Vac. Sci. Technol. B **28**, 223–262 (2010).
[7]M. Kund, G. Beitel, C. Pinnow, T. Rohr, J. Schumann, R. Symanczyk, K. Ufert, and G. Muller, "Conductive bridging RAM (CBRAM): An emerging non-volatile memory technology scalable to sub 20 nm," in IEEE International Electron Devices Meeting, IEDM Technical Digest (2005), pp. 754–757.
[8]Y. Van De Burgt, E. Lubberman, E. J. Fuller, S. T. Keene, G. C. Faria, S. Agarwal, M. J. Marinella, A. Alec Talin, and A. Salleo, "A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing," Nat. Mater. **16**, 414–418 (2017).
[9]P. Gkoupidenis, N. Schaefer, B. Garlan, and G. G. Malliaras, "Neuromorphic functions in PEDOT:PSS organic electrochemical transistors," Adv. Mater. **27**, 7176–7180 (2015).
[10]Y. Li, T. P. Xiao, C. H. Bennett, E. Isele, A. Melianas, H. Tao, M. J. Marinella, A. Salleo, E. J. Fuller, and A. A. Talin, "*In situ* parallel training of analog neural network using electrochemical random-access memory," Front. Neurosci. **15**, 323 (2021).
[11]S. Agarwal, S. J. Plimpton, D. R. Hughart, A. H. Hsia, I. Richter, J. A. Cox, C. D. James, and M. J. Marinella, "Resistive memory device requirements for a neural algorithm accelerator," in *International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2016), pp. 929–938.
[12]X. Sun and S. Yu, "Impact of non-ideal characteristics of resistive synaptic devices on implementing convolutional neural networks," IEEE J. Emerging Sel. Top. Circuits Syst. **9**, 570–579 (2019).
[13]J. A. Currivan, Y. Jang, M. D. Mascaro, M. A. Baldo, and C. A. Ross, "Low energy magnetic domain wall logic in short, narrow, ferromagnetic wires," IEEE Magn. Lett. **3**, 3000104 (2012).
[14]J. A. Currivan-Incorvia, S. Siddiqui, S. Dutta, E. R. Evarts, J. Zhang, D. Bono, C. A. Ross, and M. A. Baldo, "Logic circuit prototypes for three-terminal magnetic tunnel junctions with mobile domain walls," Nat. Commun. **7**, 10275 (2016).
[15]C. Cui, O. G. Akinola, N. Hassan, C. H. Bennett, M. J. Marinella, J. S. Friedman, and J. A. C. Incorvia, "Maximized lateral inhibition in paired magnetic domain wall racetracks for neuromorphic computing," arXiv:1912.04505 (2019).
[16]N. Hassan, X. Hu, L. Jiang-Wei, W. H. Brigner, O. G. Akinola, F. Garcia-Sanchez, M. Pasquale, C. H. Bennett, J. A. C. Incorvia, and J. S. Friedman, "Magnetic domain wall neuron with lateral inhibition," J. Appl. Phys. **124**, 152127 (2018).
[17]A. Sengupta, Y. Shim, and K. Roy, "Proposal for an all-spin artificial neural network: Emulating neural and synaptic functionalities through domain wall motion in ferromagnets," IEEE Trans. Biomed. Circuits Syst. **10**, 1152–1160 (2016).
[18]O. Akinola, X. Hu, C. H. Bennett, M. Marinella, J. S. Friedman, and J. A. C. Incorvia, "Three-terminal magnetic tunnel junction synapse circuits showing spike-timing-dependent plasticity," J. Phys. D **52**, 49LT01 (2019).
[19]U. Sahu, A. Pandey, K. Goyal, and D. Bhowmik, "Spike time dependent plasticity (STDP) enabled learning in spiking neural networks using domain wall based synapses and neurons," AIP Adv. **9**, 125339 (2019).
[20]K. Yue, Y. Liu, R. K. Lake, and A. C. Parker, "A brain-plausible neuromorphic on-the-fly learning system implemented with magnetic domain wall analog memristors," Sci. Adv. **5**, eaau8170 (2019).
[21]C. H. Bennett, N. Hassan, X. Hu, J. A. C. Incornvia, J. S. Friedman, and M. J. Marinella, "Semi-supervised learning and inference in domain-wall magnetic tunnel junction (DW-MTJ) neural networks," in *Spintronics XII* (International Society for Optics and Photonics, 2019), Vol. 11090, p. 110903I.
[22]V. Joshi, M. L. Gallo, S. Haefeli, I. Boybat, S. R. Nandakumar, C. Piveteau, M. Dazzi, B. Rajendran, A. Sebastian, and E. Eleftheriou, "Accurate deep neural network inference using computational phase-change memory," Nat. Commun. **11**, 2473 (2020).
[23]D. Kaushik, J. Sharda, and D. Bhowmik, "Synapse cell optimization and back-propagation algorithm implementation in a domain wall synapse based crossbar neural network for scalable on-chip learning," Nanotechnology **31**, 364004 (2020).
[24]Y. Lecun, C. Cortes, and C. J. C. Burges, see http://yann.lecun.com/exdb/mnist/ for "The MNIST Database of Handwritten Digits."
[25]H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," arXiv:1708.07747 (2017).
[26]A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. Van Waeyenberge, "The design and verification of MuMax3," AIP Adv. **4**, 107133 (2014).

[27]C. H. Bennett, D. Garland, R. B. Jacobs-Gedrim, S. Agarwal, and M. J. Marinella, "Wafer-scale TaOx device variability and implications for neuromorphic computing applications," in *IEEE International Reliability Physics Symposium (IRPS)* (IEEE, 2019), pp. 1–4.
[28]S. A. Siddiqui, S. Dutta, A. Tang, L. Liu, C. A. Ross, and M. A. Baldo, "Magnetic domain wall based synaptic and activation function generator for neuromorphic accelerators," Nano Lett. **20**, 1033–1040 (2020).
[29]A. Lyle, J. Harms, T. Klein, A. Lentsch, A. Klemm, D. Martens, and J. P. Wang, "Integration of spintronic interface for nanomagnetic arrays," AIP Adv. **1**, 042177 (2011).
[30]B. Jinnai, K. Watanabe, S. Fukami, and H. Ohno, "Scaling magnetic tunnel junction down to single-digit nanometers—Challenges and prospects," Appl. Phys. Lett. **116**, 160501 (2020).
[31]G. S. Beach, M. Tsoi, and J. L. Erskine, "Current-induced domain wall motion," J. Magn. Magn. Mater. **320**, 1272–1281 (2008).
[32]W. Wu, H. Wu, B. Gao, P. Yao, X. Zhang, X. Peng, S. Yu, and H. Qian, "A methodology to improve linearity of analog RRAM for neuromorphic computing," in IEEE Symposium on VLSI Technology (2018), pp. 103–104.
[33]S. Agarwal, R. B. Jacobs Gedrim, A. H. Hsia, D. R. Hughart, E. J. Fuller, A. A. Talin, C. D. James, S. J. Plimpton, and M. J. Marinella, "Achieving ideal accuracies in analog neuromorphic computing using periodic carry," in Symposium on VLSI Technology (2017), pp. T174–T175.
[34]L. Xue, A. Kontos, C. Lazik, S. Liang, and M. Pakala, "Scalability of magnetic tunnel junctions patterned by a novel plasma ribbon beam etching process on 300 mm wafers," IEEE Trans. Magn. **51**, 4401503 (2015).
[35]M. J. Marinella, S. Agarwal, A. Hsia, I. Richter, R. Jacobs-Gedrim, J. Niroula, S. J. Plimpton, E. Ipek, and C. D. James, "Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator," IEEE J. Emerging Sel. Top. Circuits Syst. **8**, 86–101 (2018).
[36]S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning* (ICML'15) (JMLR.org, 2015), Vol. 37, p. 1737–1746.
[37]P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh *et al.*, "Mixed precision training," arXiv:1710.03740 (2017).
[38]H. Noh, T. You, J. Mun, and B. Han, "Regularizing deep neural networks by noise: Its interpretation and optimization," in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), Vol. 30, pp. 5109–5118.
[39]A. V. Khvalkovskiy, V. Cros, D. Apalkov, V. Nikitin, M. Krounbi, K. A. Zvezdin, A. Anane, J. Grollier, and A. Fert, "Matching domain-wall configuration and spin-orbit torques for efficient domain-wall motion," Phys. Rev. B **87**, 020402 (2013).