A novel compound synapse using probabilistic spin-orbit-torque switching for MTJ based deep neural networks

Vaibhav Ostwal[1,2], Ramtin Zand[3], Ronald DeMara[4], Joerg Appenzeller[1,2]

[1,2] Electrical and Computer Engineering, Purdue University [3] Computer Science and Engineering, University of South Carolina [4] Electrical and Computer Engineering, University of Central Florida

Abstract: Analog electronic non-volatile memories mimicking synaptic operations are being explored for the implementation of neuromorphic computing systems. Compound synapses consisting of ensembles of stochastic binary elements are alternatives to analog memory synapses to achieve multilevel memory operation. Among existing binary memory technologies, magnetic tunneling junction (MTJ) based Magnetic Random Access Memory (MRAM) technology has matured to the point of commercialization. More importantly for this work, stochasticity is natural to the MTJ switching physics e.g devices referred as p-bits which mimic binary stochastic neurons. In this article, we experimentally demonstrate a novel compound synapse that uses stochastic spin-orbit torque (SOT) switching of an ensemble of nano-magnets that are located on one shared spin Hall effect (SHE) material channel, i.e. tantalum. By using a properly chosen pulse scheme, we are able to demonstrate linear potentiation and depression in the synapse, as required for many neuromorphic architectures. In addition to this experimental effort, we also performed circuit simulations on an SOT-MRAM based 784×200×10 deep belief network (DBN) consisting of p-bit based neurons and compound synapses. MNIST pattern recognition was used to evaluate the system performance, and our findings indicate that a significant reduction in recognition error rates can be achieved by improving the linearity of the potentiation and depression curves using an incremental pulse scheme.

I. Introduction

Analog electronic non-volatile memories (eNVMs) have attracted attention in the research community for their potential as synaptic elements [1]. The conductance of such an eNVM can increase or decrease in a continuous analog fashion, mimicking the potentiation or depression of a synapse. However, while Resistive Random Access Memory (RRAM) technology has shown the potential for achieving such analog conductance behavior, the reliable fabrication of analog RRAM devices has remained challenging [2]. Hence, compound synapses that consist of an ensemble of *binary* memory elements have been proposed. Employing the
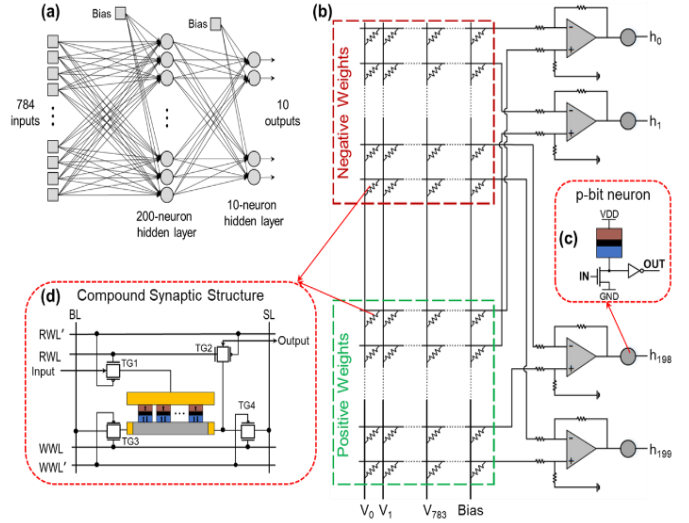

Fig. 1: (a) Graph representation of the 784×200×10 DBN (b) equivalent circuit for the first layer, (c) p-bit as neuron and (d) compound synapse implemented with MRAM cells.

probabilistic switching of individual memory elements, multilevel operation can be realized in a reproducible fashion. In fact, simulations of spiking neural networks (SNN)[3] using compound synapses from binary memory devices can elucidate the desired performance specifications. Moreover, experimental implementations based on an arrangement of parallel binary RRAM devices and simulations of convolutional neural networks (CNNs) [4] demonstrated multi-level operation of such compound synapse structures. However, RRAM technology is facing challenges in terms of current and voltage scaling and is prone to process variability and instabilities. On the other hand "MRAM has already found a niche market and is heading toward disruptive growth" according to Bhatti et al. [5]. Spin transfer torque (STT)-MRAM is close to foundry scale production [6][7] and wafer-scale manufacturability has been shown even for SOT-MRAM [8]. In fact, compound synapses based on a series arrangement of STT-MTJs have already been demonstrated [9][10].

Here we propose a new SOT-MRAM based compound synaptic structure as shown in fig 1(d) as part of a neural

network that utilizes yet another MTJ based element, i.e. a p-bit in fig 1(c) [11][12] as a binary stochastic neuron. (For details on the topic of p-bits see e.g. [11]). While the above mentioned STT-MTJ based synapse shares the same READ and WRITE path, implying that the resistance of the READ path affects the resistance of the WRITE path, our SOT synapse does not suffer from this problem, since writing occurs by means of the spin Hall effect (SHE), as discussed below. Moreover, in general SOT-MRAM is expected to perform better in terms of endurance, power consumption and speed [8]. In this article, we first experimentally demonstrate a novel spintronics compound SOT based synapse (fig 3a) and then utilize a modified version of this device (fig 1d) to simulate a spin-based based deep belief network (DBN). In particular, we have explored the accuracy of a 784×200×10 DBN as shown in fig 1 for MNIST pattern recognition, noting that the realization of a synaptic network with MRAM elements presents an attractive opportunity to build an all-MRAM based DBN. It is also worth noticing that the compound synaptic structure proposed here can readily be utilized in bio-inspired computing architectures such as oscillatory neural networks [13], reservoir computing [14], population coding [15] to just name a few, which all utilize the dynamics of stochastic or oscillatory MTJs to mimic neuronal functionality. Furthermore, multiple neuromorphic computing architectures have been proposed using stochastic memristive devices in [10][16][17][18].

For the experimental demonstration we utilize the intrinsic property of spin devices to exhibit thermally activated switching of their magnetization that is probabilistic in nature. We have built and characterized an array of nanomagnets with perpendicular shape anisotropy (PMA) located on a tantalum layer that acts as a spin Hall effect (SHE) channel. Probabilistic switching of the individual nanomagnets is used to realize our SOT synapse. Since an individual nanomagnet has a finite probability of switching for a properly chosen current pulse through the tantalum layer, the ensemble of nanomagnets shows a gradual increase (potentiation) or decrease (depression) in the total magnetization state similar to an analog memory element. While the observed potentiation and depression is non-linear with respect to the number of input pulses, a modified pulse scheme, as discussed below, can be used to alleviate this issue.

II. EXPERIMENTAL RESULTS

A   Ta(5nm)/CoFeB(1nm)/MgO(2nm)/Ta(2nm) magnetic stack with perpendicular magnetic anisotropy (PMA) was deposited on a Si/SiO$_2$ substrate using sputter deposition techniques. In the first e-beam lithography step, a Hall bar was patterned by etching through the entire Ta/CoFeB/MgO/Ta stack until the SiO$_2$ substrate was reached using Ar ion milling. Subsequently, either a single nanomagnet or an ensemble of nanomagnets are patterned on
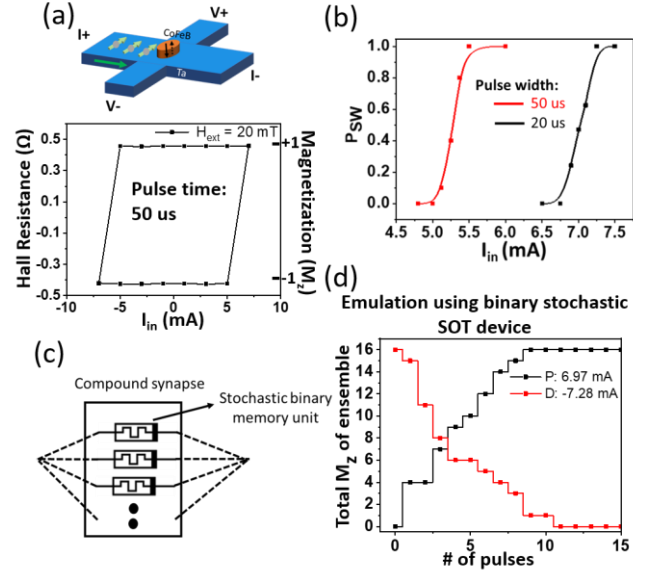

Fig. 2 (a) SOT switching of a single nanomagnet (binary memory unit) (b) switching probability (P_{sw}) curve for two different current pulse widths (c) compound synapse consisting of multiple stochastic binary units (d) experimental potentiation and depression curves emulating a 4-bit compound synapse using a SOT device with a single nanomagnet.

the Hall bars and defined by etching through the top part of the material stack until the bottom Ta layer was reached. Finally, Ti/Au metal pads were deposited using a standard lift-off process, enabling contacts to the devices. To detect the magnetization state of the system, Anomalous Hall Effect (AHE) measurements were performed using a lock-in scheme after quasi-static pulses of 20 to 50 us in width were applied to switch the nanomagnets. Fig 2(a) shows SOT switching of a fabricated Hall bar with a single nanomagnet using current pulses of 50 us width in the presence of an in-plane external field of 20 mT in the current direction. The Hall bar had a width of around 4 um and the nanomagnet on top is elliptical in shape with axes dimensions of 1 and 3 um respectively and the longer axis in the current direction. In our proof-of-concept devices, we are using an external in-plane field to break the symmetry during SOT switching of PMA magnets, similar to the approach used in previous experiments [19]. Note however that external field-free SOT switching of PMA magnets can be achieved with a modified device structure [20][21]. To explore the probabilistic nature of SOT switching, we first apply a high negative current (RESET) pulse [width: 50 us] to deterministically switch the nanomagnet to its -1 state, followed by a positive current pulse with varying amplitude (SET) to probabilistically switch the nanomagnet to its +1 state. Fig 2(b) depicts the average probability of switching. As expected, a higher SET current amplitude results in a higher probability of switching. In addition, measurements were also performed employing a pulse width of 20 us, which resulted in a similar trend but higher current requirements for the probabilistic switching,

consistent with the expectations for thermally activated spin torque switching. Note that thermally activated switching of nanomagnets is inherently probabilistic in nature and even pulses with widths of nanoseconds shows probabilistic switching similar to the observation in fig 2(b).

An n-bit compound synapse can be realized using 2^n such nanomagnets operating in the probabilistic switching regime and working in parallel or series (fig 2(c)). To emulate a 4-bit synapse with the "one nanomagnet device", the device was first set to its -1 state as described before. Next, 20 SET pulses at a current level of 6.97 mA with a pulse width of 20 us were applied to the sample. Note that this current level lies in the "finite switching probability" region of the nanomagnet in fig 2(b). Due to the probabilistic nature of switching, the nanomagnet can switch to its +1 state during any of the 20 SET pulses and will remain in this +1 state for all subsequent SET pulses, since the energetic barrier between the -1 and +1 state is of the order of $40k_BT$. Since a 4-bit compound synapse is an ensemble of such 16 devices, our experiment is repeated 16 times on the same device. In this way, instead of using an ensemble of 16 devices, we are able to use a single device to emulate the switching behavior of the ensemble. Fig 2(d) shows the outcome of this experiment. The number of nanomagnets switched is plotted versus the SET pulse number. The black line clearly illustrates the increasing probability of finding the one nanomagnet switched in all 16 repeated experiments in its +1 state. Such behavior is analogous to the potentiation (P) curve of an analog synapse where the conductance increases gradually with the number of input pulses. Similarly, depression (D) operation can be emulated with our "one nanomagnet device" when input current pulses of -7.28 mA are used – see red curve in fig 2(b). Note that by using an MTJ structure instead of our nanomagnet, the plot in fig 2(d) would result in an incremental change in conductance through the MTJ with the number of pulses instead of a change in anomalous Hall resistance, making the approach much more technology relevant and feasible.

Next, we have studied an actual 4-bit compound synapse. A cross shaped Hall bar structure from tantalum with 16 nanomagnets in the cross region was built (see SEM image of the device in fig 3(a)). Fig 3(b) shows the anomalous Hall resistance versus perpendicular external magnetic field of the 16 nanomagnet device with resistance steps that correspond to the switching of individual nanomagnets. Each nanomagnet's switching causes a resistance change of around 110 mOhm ≈ 1.8 Ohm / 16. Due to variations in the coercive fields of the 16 nanomagnets as well as the stochastic nature of the nanomagnets' switching steps occur at slightly different magnetic fields. As in the previous experimental approach, the 16 nanomagnet device is initially put in its -1 state, before being subjected to 20 us pulses of identical amplitude (pulse scheme I). Fig 3(d) shows the resulting P curves for three different current levels
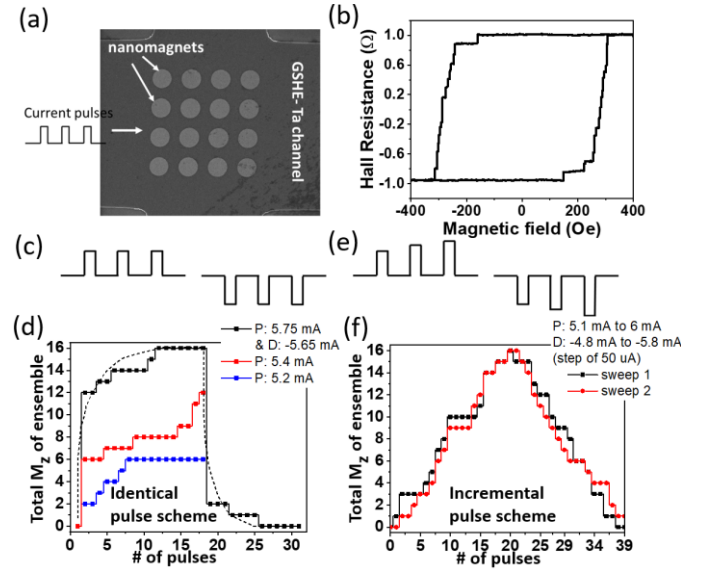

Fig 3 (a) Scanning electron microscope (SEM) image of our SOT based synapse with 16 nanomagnets (b) anomalous Hall effect vs. out-of-plane magnetic field for the device showing 110 mOhm steps, corresponding to individual nanomagnet switching (c) pulse scheme I: identical pulses (d) potentiation and depression curve of the device using pulse scheme I. 3 different potentiation curves showing that the rate of the potentiation can be controlled by the current amplitude (e) pulse scheme II: incremental pulses (f) potentiation and depression curve of the device with pulse scheme II.

in addition to one D curve, displaying very similar characteristics as fig 2(d). Similar to the emulated P and D curves (fig 2(d)), P and D curves of our compound synapse are nonlinear with respect to the pulse number and show a saturating behavior. Note that such non-linear behavior is also observed for analog type RRAM memristors due to the nature of filament formation and breaking during the SET and RESET process. For neural networks, a non-linear response of this type is in general undesirable, since it can have a detrimental impact on the accuracy of the network. To address this issue, researchers [22] have explored a modified pulse scheme in which, instead of using identical pulses, pulses with either incrementally increasing pulse-width or pulse-amplitude are used. Fig 3(f) shows the result of adopting this modified pulse scheme for our 4-bit compound synapse. A significantly improved linear characteristic is experimentally obtained when increasing the current amplitude from 5 mA to 6 mA (P-curve) and decreasing it from -4.8 mA to -5.8 mA (D-curve) in steps of 0.05 mA (pulse scheme II). Note that for two sweeps (red and black curve in fig 3(f)) the magnetization change is not identical as individual nanomagnets' switching is stochastic in nature. However, the net magnetization of the magnet ensemble shows an overall improved linearity in the characteristics.

III. APPLICATION-LEVEL SIMULATIONS

To analyze the effect of the above used switching pulse schemes on the application-level behavior of neuromorphic architectures, we have used the PIN-Sim framework [23] developed by the authors to realize a circuit-level implementation of a deep belief network (DBN) using our compound synaptic structure as weighted connection and MRAM-based p-bits as neurons, as shown in Fig 1. In particular, a MATLAB based module [24] is modified according to the characteristics of the neurons to perform an off-line training activity to tune the weights in a 784×200×10 DBN architecture. Once the network is trained, the obtained weights are mapped to various resistive levels that can be realized by our MRAM-based compound synapses, as illustrated by (1):

$$
\begin{aligned}
&\textbf{for } i = 1:(Q+1) \textbf{ do}\\
&\qquad\qquad R_{Synapse}(i) =\\
&\quad R_P R_{AP}/(R_{AP}(Q-(i-1)) + R_P(i-1))\\
&\textbf{end}
\end{aligned}
\tag{1}
$$

$$
R(\theta) = \begin{cases} R_P = R_{MTJ} & , \ \theta = 0° \\ R_{AP} = R_{MTJ}(1+TMR), & \theta = 180° \end{cases}
\tag{2}
$$

where the quantization factor (Q) defines the number of MTJs used in each compound synapse, and R_P and R_{AP} represent the resistance of the MTJ in its parallel and anti-parallel states, respectively. R_P and R_{AP} are obtained using (2), where TMR is the tunneling magnetoresistance and $R_{MTJ} = RA/Area$, in which RA is the MTJ's resistance-area product, and $Area$ is the surface of the MTJ. The accompanying Supplementary Materials provide detailed descriptions regarding the structure of the PIN-Sim, as well
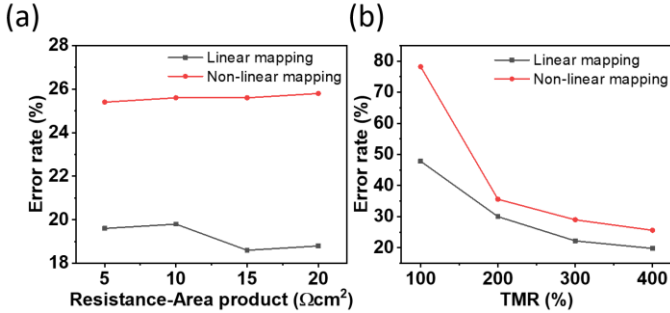
Fig. 4: (a) Error rate versus resistance-area, (b) Error rate versus tunneling magneto resistance (TMR) for a 784×200×10 DBN with 4-bit compound synapses

as the fundamentals of DBNs.

Fig 1(b) shows the structure of a 784×200×10 DBN circuit implemented by the modified PIN-Sim framework for MNIST pattern recognition application [25]. According to the experimental results shown in Fig 3, using the identical pulse scheme (scheme I) leads to simultaneous switching of multiple MTJs in the compound synapses. This results in randomly skipping some of the resistive levels that could be ideally realized by the synaptic structures, which we refer to

as non-linear mapping. In order to model the non-linear mapping behavior in the application-level simulations, we have used the average of 100 stochastic cases based on the mathematical model for the compound synapses that are switched by the identical pulse schemes to estimate the most probable resistive-levels that could be realized by the pulse scheme I. However, utilization of the incremental pulse scheme (scheme II) enables leveraging all the possible resistive levels that can be realized by a compound synapse, which we refer to as linear mapping.

The application-level simulation results obtained by the PIN-Sim framework for a 4-bit compound synapse (Q=16) are shown in Fig 4, which displays the relation between the DBN's error rate and two different device level parameters of the MTJs for nonlinear and linear mapping, respectively. In particular, fig 4(a) focuses on the relation between error rate and RA while the TMR value is assumed to be 400%, and Fig 4(b) shows the error rate versus TMR relation for RA= 10 $\Omega\mu m^2$. The obtained results show the significant effect of TMR on the accuracy of the network, while changing the RA value does not result in a major impact on the error rate. Moreover, it is shown in Fig 4 that increasing
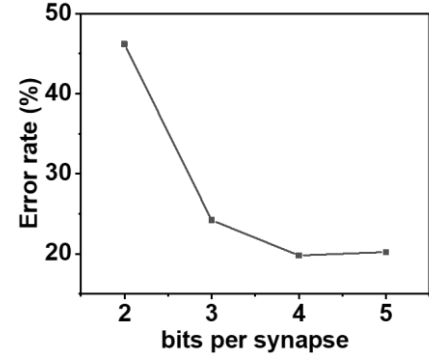
Fig. 5: Error rate versus bits per synapse.

the RA and TMR values can result in reduced power consumptions for the entire network. However, realization of higher TMR values would impose increased complexity in the fabrication process, of which a TMR of ~600% has been experimentally demonstrated at room temperature [26]. It is worth noting that although our application-level simulation results have not considered the stochasticity in the compound synaptic structures, which would result in cycle-to-cycle variations of the P and D curves as shown in fig 3(f), they help gaining insights into the dependence of network accuracy on device parameters such as TMR and RA products as well as understanding the relation between error rates and bit-precision as shown in fig 5. For instance, a bit-precision of more than 4-bits is not required within the trained network, since the error rates will saturate for higher bit-precision values. Such findings are particularly important

to guide future experiments, since the fabrication complexity of compound synaptic structures can significantly increase with larger bit-precision and TMR values. Finally, by comparing the non-linear and linear mapping realized by the identical (scheme I) and incremental (scheme II) pulse schemes, a significant decrease in the error rates for the linear mapping method compared to the non-linear mapping can be observed from fig 4. This highlights the effectiveness of the incremental pulse scheme approach investigated herein. It is worth noting that while cycle-to cycle (C2C) variations induced by the natural stochasticity of our synapses may change the absolute error rate values reported here, the relative accuracy plotted in fig 4 is important to understand the impact of the switching pulse schemes on the performance of the entire network. The obtained results highlight various opportunities for future work, such as: 1) investigating deeper neuromorphic architectures for large-scale applications, 2) leveraging the intrinsic stochasticity of the compound synapses to realize an energy-efficient training approach for quantized neural networks[3][4], or 3) applying multi-objective optimizations to tune the network parameters by including device models in the simulation loop.

To verify the energy improvements of our compound synaptic structure over a CMOS-based weighted connection, we have utilized the values reported in [27] and [28] to estimate the energy consumption for weighted sum operations in a 784×200×10 DBN using integer and floating-point (FP) multiplication and add (MAC) operations. As listed in Table I, the 4-bit compound synapse based DBN

TABLE I: ENERGY CONSUMPTION COMPARISON FOR WEIGHTED-SUM OPERATIONS IN A 784×200×10 DBN.

Weighted Connections	Energy Consumption			
	MAC Operations		Memory Access per Instruction	
	per Instruction	Total (nJ)	Cache (pJ)	DRAM (nJ)
8-bit Integer [27]	0.23 pJ	36	100	1.3-2.6
32-bit Integer [27]	3.2 pJ	510	100	1.3-2.6
16-bit FP [27]	1.5 pJ	240	100	1.3-2.6
32-bit FP [27]	4.8 pJ	730	100	1.3-2.6
4-bit Compound Synapse	NA[(1)]	0.21	Near-Zero	

(1) It is not available (NA) since all the MAC operations in the crossbar are performed simultaneously through the intrinsic physical characteristics of the compound synapses. We can only obtain the total energy consumption of the crossbar, which can later be used to estimate the energy per instruction, i.e. 0.0013 pJ.

circuit can realize approximately two orders-of-magnitude energy improvements over their CMOS-based counterparts. Note that increasing the bit-precision in a compound synapse from 4-bit to 16-bit or 32-bit can result in higher energy consumption, but this increase will not be beyond the significant energy reductions that are achieved compared to the integer and FP DBNs. Moreover, it was shown in [28]

that memory accesses normally consume more energy than arithmetic operations. Since our compound synapses indeed operate as memory cells, they enable in-memory computation, which decreases the memory access energy consumptions to near-zero values, as indicated in Table I. Finally, it should be noted that in this work we have only focused on the energy consumption which can be realized by our compound synapses, while authors have shown in [23] that the p-bit based neuron can also contribute to several orders of magnitudes energy reduction with respect to previously reported energy-efficient CMOS-based activation functions for DBN structures.

IV. CONCLUSION

In conclusion, we have experimentally demonstrated proof-of-concept 4-bit compound synapses using probabilistic spin-orbit torque switching. A modified incremental pulse scheme is shown to result in improved linearity of the synaptic behavior. Furthermore, circuit-level simulations of DBNs consisting of stochastic spin-torque devices, namely p-bits and compound synapses show that the linearity in the synaptic behavior and high TMR-values are crucial device parameters to achieve low error rates.

References

[1] Shimeng Yu, "Neuro-Inspired Computing With Emerging Nonvolatile Memory," *Proc. IEEE*, vol. 106, no. 2, 2018.

[2] G. C. Adam, A. Khiat, and T. Prodromakis, "Challenges hindering memristive neuromorphic hardware from going mainstream," *Nat. Commun.*, vol. 9, no. 1, pp. 2–5, 2018.

[3] J. Bill and R. Legenstein, "A compound memristive synapse model for statistical learning through STDP in spiking neural networks," *Front. Neurosci.*, vol. 8, no. DEC, pp. 1–18, 2014.

[4] E. Vianello *et al.*, "HfO$_2$-Based OxRAM Devices as Synapses for Convolutional Neural Networks," *IEEE Trans. Electron Devices*, vol. 62, no. 8, pp. 2494–2501, 2015.

[5] S. Bhatti, R. Sbiaa, A. Hirohata, H. Ohno, S. Fukami, and S. N. Piramanayagam, "Spintronics based random access memory: a review," *Mater. Today*, vol. 20, no. 9, pp. 530–548, 2017.

[6] G. H. Koh *et al.*, "Demonstration of Highly Manufacturable STT-MRAM Embedded in 28nm Logic," *IEEE Int. Electron Devices Meet.*, pp. 18–2, 2019.

[7] L. Wei *et al.*, "A 7Mb STT-MRAM in 22FFL FinFET Technology with 4ns Read Sensing Time at 0.9V Using Write-Verify-Write Scheme and Offset-Cancellation Sensing Technique," *IEEE Int. Solid-State Circuits Conf.*, pp. 480–481, 2019.

[8] M. Cubukcu *et al.*, "SOT-MRAM 300mm integration for low power and ultrafast embedded memories K.," *IEEE Trans. Magn.*, vol. 54, no. 4, pp. 81–82, 2018.

[9]   E. Raymenants *et al.*, "Chain of magnetic tunnel junctions as a spintronic memristor," *J. Appl. Phys.*, vol. 124, no. 15, 2018.

[10] D. Zhang *et al.*, "All Spin Artificial Neural Networks Based on Compound Spintronic Synapse and Neuron," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 4, pp. 828–836, 2016.

[11] K. Y. Camsari, R. Faria, B. M. Sutton, and S. Datta, "Stochastic p-bits for invertible logic," *Phys. Rev. X*, vol. 7, 2017.

[12] K. Y. Camsari, S. Salahuddin, and S. Datta, "Implementing p-bits with Embedded MTJ," *IEEE Electron Device Lett.*, vol. 38, no. 12, pp. 1767–1770, 2017.

[13] M. Romera *et al.*, "Vowel recognition with four coupled spin-torque nano-oscillators," *Nature*, vol. 563, no. 7730, pp. 230–234, 2018.

[14] J. Torrejon *et al.*, "Neuromorphic computing with nanoscale spintronic oscillators," *Nature*, vol. 547, no. 7664, pp. 428–431, 2017.

[15] A. Mizrahi *et al.*, "Neural-like computing with populations of superparamagnetic basis functions," *Nat. Commun.*, vol. 9, no. 1, pp. 1–11, 2018.

[16] C. Gamrat *et al.*, "Spin-Transfer Torque Magnetic Memory as a Stochastic Memristive Synapse for Neuromorphic Systems," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 2, pp. 166–174, 2015.

[17] D. Zhang, L. Zeng, and J.-O. Klein, "Stochastic Spintronic Device based Synapses and Spiking Neurons for Neuromorphic Computation," *IEEE/ACM Int. Symp. Nanoscale Archit.*, pp. 173–178, 2016.

[18] A. Mondal and A. Srivastava, "In-situ Stochastic Training of MTJ Crossbar based Neural Networks," *Proc. Int. Symp. Low Power Electron. Des.*, p. 51, 2018.

[19] L. Liu, C. Pai, Y. Li, H. W. Tseng, D. C. Ralph, and R. A. Buhrman, "Spin-Torque Switching with the Giant Spin Hall Effect of Tantalum," *Science (80-.).*, vol. 336, pp. 555–559, 2012.

[20] Y. W. Oh *et al.*, "Field-free switching of perpendicular magnetization through spin-orbit torque in antiferromagnet/ferromagnet/oxide structures," *Nat. Nanotechnol.*, vol. 11, no. 10, pp. 878–884, 2016.

[21] S. Fukami, C. Zhang, S. Duttagupta, A. Kurenkov, and H. Ohno, "Magnetization switching by spin-orbit torque in an antiferromagnet-ferromagnet bilayer system," *Nat. Mater.*, vol. 15, no. 5, pp. 535–541, 2016.

[22] M. Jerry *et al.*, "Ferroelectric FET Analog Synapse for Acceleration of Deep Neural Network Training," *IEEE Int. Electron Devices Meet.*, pp. 6–2, 2017.

[23] R. Zand, K. Y. Camsari, S. Datta, and R. F. DeMara, "Composable Probabilistic Inference Networks Using MRAM-based Stochastic Neurons," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 15, no. 2, p. 17, 2019.

[24] M. Tanaka and M. Okutomi, "A Novel Inference of a Restricted Boltzmann Machine," *Proc. - Int. Conf. Pattern Recognit.*, pp. 1526–1531, 2014.

[25] L. Y, B. L, B. Y, and H. P, "Gradient-Based Learning Applied to Document Recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[26] S. Ikeda *et al.*, "Tunnel magnetoresistance of 604% at 300 K by suppression of Ta diffusion in CoFeBMgOCoFeB pseudo-spin-valves annealed at high temperature," *Appl. Phys. Lett.*, vol. 93, no. 8, pp. 1–4, 2008.

[27] B. Zhuang, C. Shen, and I. Reid, "Training Compact Neural Networks with Binary Weights and Low Precision Activations," *J. Mach. Learn. Res. 18*, vol. 18, pp. 1–30, 2018.

[28] M. Horowitz, "Computing's Energy Problem: (and what we can do about it)," *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Pap.*, pp. 10–14, 2014.