

Evaluation of Wafer-Scale SOT-MRAM for Analog Crossbar Array Applications

Samuel Liu^{1,2}, Chen-Yu Hu³, Ming-Yuan Song³, Xinyu Bao³, and Jean Anne C. Incorvia^{1,2}

¹ Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA

² Microelectronics Research Center, The University of Texas at Austin, Austin, TX, USA

² Corporate Research, Taiwan Semiconductor Manufacturing Corporation, Hsinchu, Taiwan

ABSTRACT

Analog crossbar arrays consisting of emerging memory devices can greatly alleviate the computational strain required by vector matrix multiplications for neural network applications. The ability to produce spin orbit torque-magnetic random-access memory (SOT-MRAM) at wafer-scale positions SOT-MRAM as a strong memory candidate. In this work, we fabricate and measure 300 mm-compatible SOT-MRAM with 150% tunnel magnetoresistance ratio, fast (2 ns) and low voltage (<1 V) operation, low energy dissipation (350 fJ), low write noise (0.1%), and low device-to-device variation of 10%. Through 2-bit quantization aware training and noisy training as mitigation techniques, the measured SOT-MRAM devices attain 95% on MNIST. The bi-stable anisotropy and stochastic switching of SOT-MRAM can additionally be leveraged for stochastic training of binary neural networks, able to reach ideal accuracy for a single device. Lastly, the devices were evaluated on implementation of probabilistic graph modeling and the interplay of tunnel magnetoresistance ratio, probability curve distribution, and conductance noise was shown to reduce potential errors in implementation. Through these results, SOT-MRAM is shown to be a uniquely effective candidate for implementation of crossbar accelerators in memory- and energy-limited applications, able to take advantage of stochastic operation and bi-stability to beneficial results in neural network applications.

BODY

As the widespread use of machine learning algorithms demands ever greater computational weight, unconventional computing architectures are becoming increasingly important to address this demand¹.

Among them, analog crossbar array circuits leveraging non-volatile memories have been proposed to accelerate vector matrix multiplications, one of the most costly and ubiquitous mathematical operations in machine learning applications^{2,3}. A wide variety of memories have been explored to build these crossbar arrays, ranging from more traditional SRAM^{4,5}, DRAM⁶, and flash⁷ memories to less traditional memories such as phase change memory (PCM)⁸⁻¹¹, resistive random access memory (RRAM)¹²⁻¹⁶, and magnetic random access memory (MRAM)^{17,18}, to more experimental memories such as 2D materials¹⁹⁻²² or ionic material-based electrochemical random access memory (ECRAM)²³⁻²⁸. Among these, MRAM has the advantages of relatively fast operation of a few ns via ferromagnets to potential ps-fs via antiferromagnets^{29,30}, low voltage operation < 1 V, high write endurance up to 10^{15} cycles, and the existence of semiconductor processes to produce reliable wafer-scale arrays³¹⁻³⁵. MRAM is based around the magnetic tunnel junction (MTJ), where two ferromagnetic layers sandwich an insulating tunnel barrier and the device is in a low (high) resistance state when the relative magnetization of the two ferromagnetic layers is parallel (anti-parallel). While previous works have explored implementations of wafer scale spin transfer torque-MRAM (STT-MRAM) in analog crossbar arrays for neuromorphic computing applications¹⁷, spin orbit torque-MRAM (SOT-MRAM) has only recently been applied to wafer scale memory applications^{18,33,35,36}. While SOT-MRAM is a three-terminal memory in contrast to STT-MRAM, a two-terminal memory, leading to increased area cost, SOT-driven magnetization switching has advantages of higher speed (down to ps timescales)³⁷, up to 10 times better energy efficiency³⁸, and up to 10 times lower switching voltage³⁸ due to the higher comparative efficiency of SOT switching vs. STT switching (*e.g.* ~0.58 charge-to-spin conversion in CoFeB spin valves³⁹ vs. 0.9 in Pt-based SOT structures⁴⁰). Additionally, the physical separation of the higher current write channel through the heavy metal SOT layer, and the lower current read channel through the tunnel barrier facilitates potentially higher write endurance. This is because the main cause of device degradation in MTJ-based devices is degradation of the insulating tunnel barrier due to Joule heating⁴¹. However, for neuromorphic applications of synaptic weight representation, SOT-MRAM retains the same drawback of STT-MRAM where a scaled MTJ can usually only represent two states due to shape, magnetocrystalline, or interfacial anisotropy effects. While much previous research

has explored how more states can be introduced to magnetic devices through multi-domain nucleation^{42,43}, domain walls^{44–50}, skyrmions^{51–54}, and magneto-ionics^{55–60}, we focus on applications that can benefit from the highly nonlinear switching effect induced by magnetic anisotropy. In this work, we measured SOT-MRAM devices sampled from wafer-scale fabrication of 4Kb memory arrays and analyze their projected performance on data-driven applications of neural network inference acceleration on 2-bit quantized networks, stochastic training of binary neural networks, and probabilistic graph model representation. We identify implementation strategies and areas of improvement and establish that SOT-MRAM is well-suited for limited-resource edge computing applications due to high speed, high endurance, high energy efficiency, and distinct switching characteristics derived from magnetic physics. These results show there is an important role in neuromorphic computing for SOT-MRAM non-volatile memories that have stochastic switching dynamics coupled with two stable resultant states.

To evaluate the expected performance of the devices at the state-of-the-art level, a 4Kb array of SOT-MRAM devices were fabricated using the same methodology described in Ref. ¹⁸. Figure 1a is a depiction of the typical structure of the in-plane magnetic anisotropy SOT-MTJ device along with the electrical operation of the bitcell. Figure 1b depicts a scanning electron microscope capture of the cross-section of the devices. The films were grown using sputter deposition followed by 400 °C annealing for 30 minutes. The resultant film had a RA of $10 \Omega^*\mu\text{m}^2$ and an average tunnel magnetoresistance (TMR) of 170% as measured by current in-plane tunneling. The heavy metal layer is composite W, which was measured to have a spin Hall angle of 0.6. The devices were then fabricated by patterning with electron-beam lithography followed by etching using a hard mask reactive ion etch followed by an Ar inductively couple plasma etch, which was done to avoid damage to the sidewalls of the MTJ. Lastly, an ion beam etch was performed at low energy to eliminate redeposition on the MTJ sidewalls and preserve the SOT underlayer. The MTJ devices were patterned into ellipses of dimension 75 nm by 230 nm. Figure 1c shows 100 field loops of a single SOT-MTJ device, showing sharp, symmetric switching current loops with high effective TMR at 150% and high thermal stability 152. When operated with a pulse width of 10 ns, the

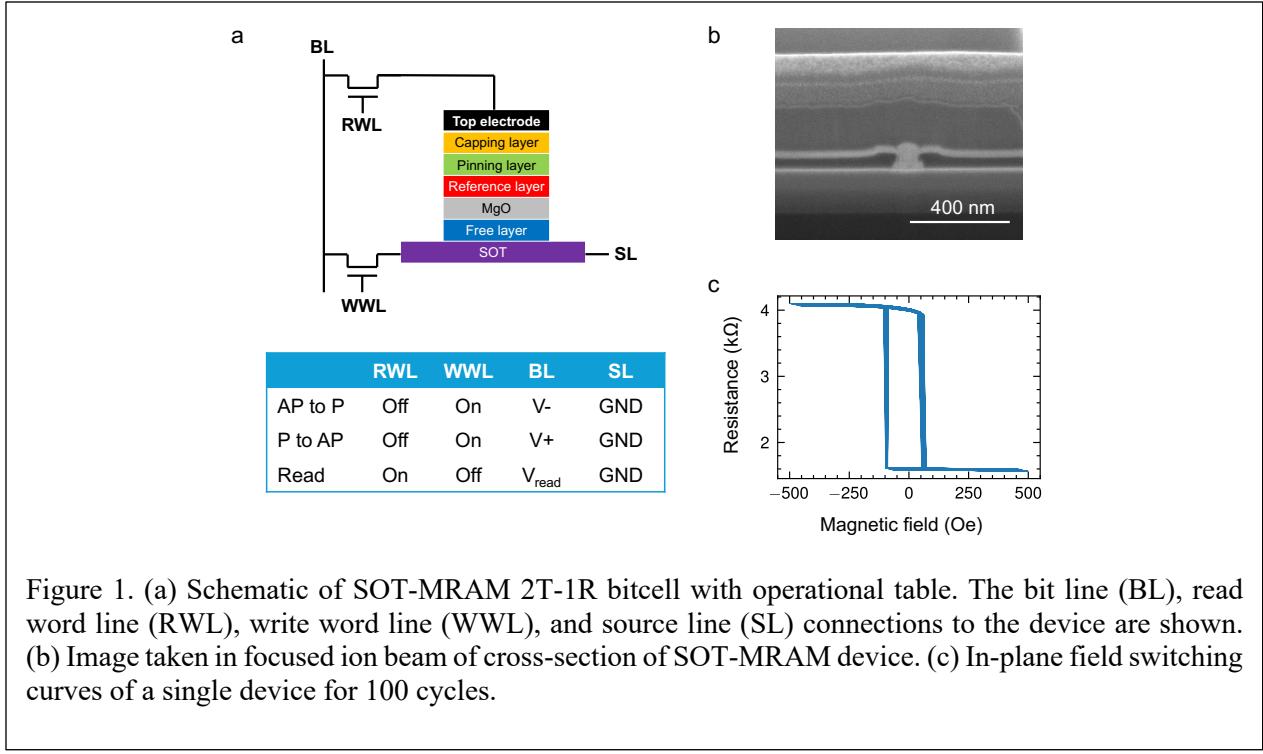


Figure 1. (a) Schematic of SOT-MRAM 2T-1R bitcell with operational table. The bit line (BL), read word line (RWL), write word line (WWL), and source line (SL) connections to the device are shown. (b) Image taken in focused ion beam of cross-section of SOT-MRAM device. (c) In-plane field switching curves of a single device for 100 cycles.

resulting energy dissipation is an average of 350 fJ per write event. As previously shown¹⁸, devices show no degradation at $>10^{12}$ write pulses (followed by read pulses), showing high endurance effective for a wide variety of memory and in-memory computing applications. These characteristics are well-suited for analog neural network inference acceleration applications, where retention and conductance stability have been shown to be two of the more important factors when evaluating inference accelerators⁶¹. Additionally, low write energy and fast writes are also important as the size of the models can exceed the amount of crossbar memory available, leading to necessary writes to perform matrix multiplication acceleration⁶². While the lack of analog states can be a drawback for inference applications, the bi-stability of MTJs can be leveraged using quantization aware training.

Due to the stable two states, we propose the use of SOT-MRAM devices for acceleration of 2-bit quantized neural networks, meaning that weights are constrained to values of -1, 0, and 1, represented by two SOT-MRAM devices per cell. While this limits the effective resolution of the crossbar array elements, very little digital memory storage is required to maintain a backup of the weights, synergizing well with the application to resource-limited systems⁶³. Networks that are quantized to 2 bits post-training traditionally suffer severe accuracy penalties, but this is alleviated through quantization aware training (QAT)⁶⁴. LeNet-

5 architecture convolutional neural networks (CNNs) shown in Fig. 2a were trained using Keras⁶⁵ on the MNIST handwritten digits dataset⁶⁶ and evaluated in CrossSim⁶⁷, a crossbar simulator that samples from measured data, with ideal differential weights at varying quantization level post-training, shown in Fig. 2b. The conventionally trained model (blue curve) shows a noticeable inference accuracy drop when quantized to 3 bits, with a severe drop-off at 2-bit quantization. However, when the CNN is trained using QAT (green curve), this accuracy is largely recovered, with a maximum validation accuracy of 98.2% compared to a maximum validation accuracy in the full precision model (assumed to be operated at 8 bits) at 99.2%, a small drop of ~1% with a 94% reduction in model memory usage compared to 32-bit floating precision.

To predict the performance of the network accurately, the high resistance state (HRS) and low resistance state (LRS) of 100 different devices on the wafer were measured, resulting in the distribution shown in Fig. 2c. The devices have an average TMR of 166%. The spread of the conductance values is approximately proportional, where the HRS has a conductance spread of $k_{\sigma,HRS} = \frac{\sigma_{G,HRS}}{\mu_{G,HRS}} = 9.4\%$ and the LRS has a conductance spread of $k_{\sigma,LRS} = \frac{\sigma_{G,LRS}}{\mu_{G,LRS}} = 10.2\%$, where σ_G and μ_G are respectively the standard

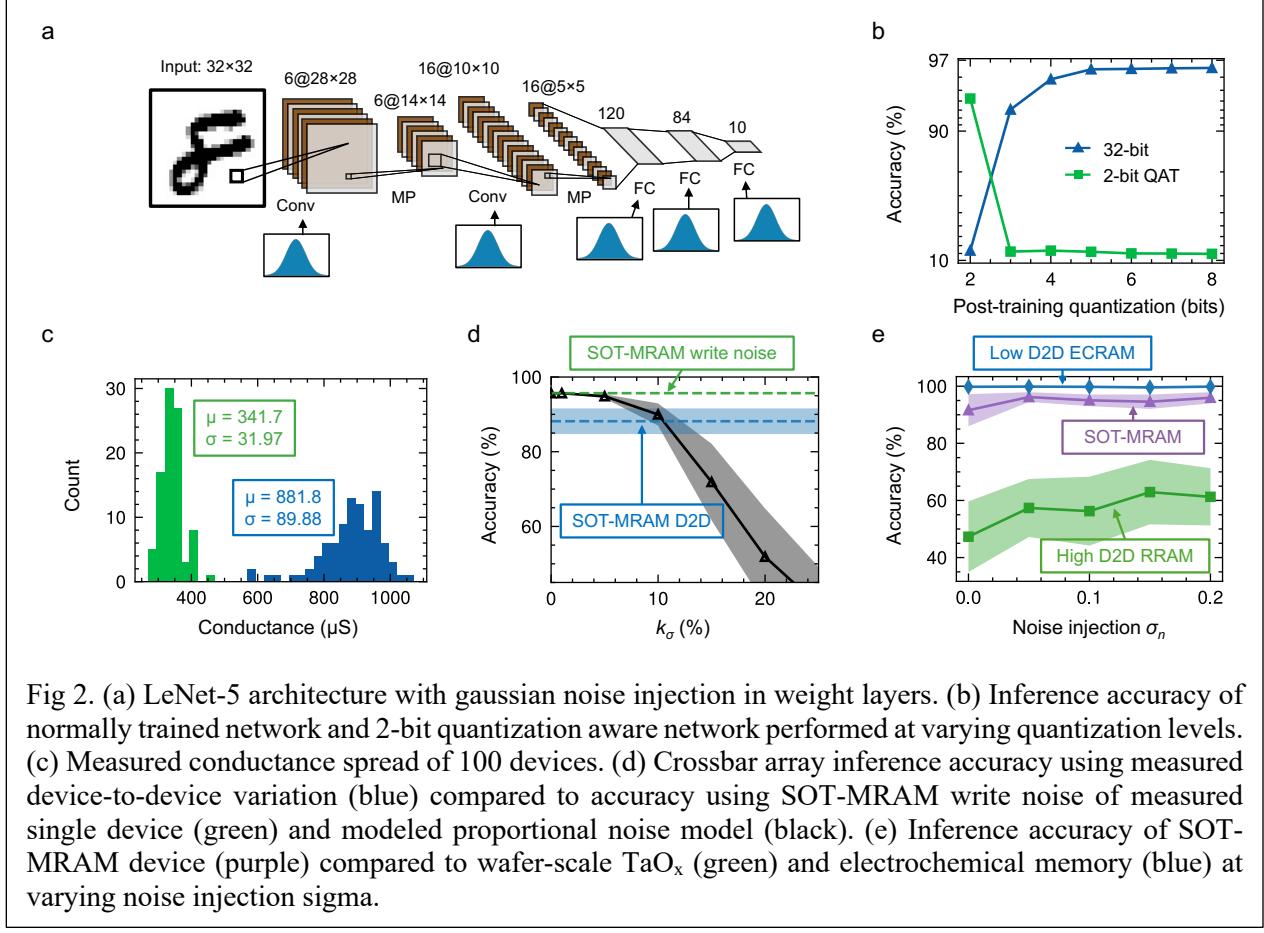


Fig 2. (a) LeNet-5 architecture with gaussian noise injection in weight layers. (b) Inference accuracy of normally trained network and 2-bit quantization aware network performed at varying quantization levels. (c) Measured conductance spread of 100 devices. (d) Crossbar array inference accuracy using measured device-to-device variation (blue) compared to accuracy using SOT-MRAM write noise of measured single device (green) and modeled proportional noise model (black). (e) Inference accuracy of SOT-MRAM device (purple) compared to wafer-scale TaO_x (green) and electrochemical memory (blue) at varying noise injection sigma.

deviation and mean of the conductance state. This variation is then applied into CrossSim by generating crossbars with device conductance variations matching the data and the 2-bit QAT CNN is modeled on the array of representative devices. Figure 2d shows the inference accuracy of the array over 20 runs with different seeds (blue), with an average accuracy of 88% with a standard deviation of 3.2%, shown by the shading. This is also compared with an array that is evaluated with only the write noise of a single device (green), with an average accuracy of 95.1% and a standard deviation of 0.2%. The write noise of a single device, at 0.107%, was calculated by analyzing the HRS and LRS of the device in Fig. 1c at 0 field. This is a projection of the limit of the maximum accuracy achievable with a SOT-MRAM array if device-to-device variation was minimized. The black series depicts a generic SOT-MRAM device with weight-proportional noise. This data is meant to represent wafer-scale SOT-MRAM of various device-to-device variations. The intercept of the weight-proportional devices and the SOT-MRAM array is at $k_\sigma = 10.4\%$, where k_σ is the weight proportional noise level, showing agreement between the experimental data and the weight-

proportional model. From the results in Fig. 2d, the device-to-device variation results in a significant reduction in inference accuracy compared to the single device case.

While tightening the device-to-device variation is necessary to increase the inference performance, at variations of approximately 5%, there is still a noticeable accuracy drop. One strategy to alleviate this by using noise injection during training, which is applied in each weight layer as shown in Fig. 2a. Noise injection has been shown to aid convergence during training to better error minima⁶⁸ and result in higher noise tolerance during inference^{69,70}. Figure 2e shows inference accuracy for varying training noise injection levels (arb. units) for the characterized SOT-MRAM devices along with an example device with relatively high write noise and device-to-device variation at 23.1% (TaO_x RRAM⁷¹) and example relatively low write noise and device-to-device variation at 1.1% (ECRAM²³) for comparison. The results indicate that at moderate noise injection levels, the performance of the SOT-MRAM array increases by 6% at the maximum improvement, matching expectations that the network becomes more noise resilient. However, this trend is not as clear with the characterized RRAM devices with significantly worse variation, indicating that the variation of the SOT-MRAM devices falls within an accepted variation where noise injection can result in increased inference accuracy for the system. Supplementary Fig. S1 shows the validation accuracy of the trained networks, showing that the noise applied to the network training eventually becomes large enough to prevent the network from reaching a better error minimum, resulting in worse accuracy. The interplay between getting a better error minimum and training the network to be noise resilient results in an effective sweet spot for noise injection during training, that the measured SOT-MRAM array meets.

Due to the measured high speed and high endurance of the SOT-MRAM devices, online neural network training acceleration is a promising application for the devices. The binary weight nature of the SOT-MRAM can be leveraged to accelerate binary neural networks (BNNs), shown in Fig. 3a, which constrain weight and activation values to only the two values of -1 and +1, in contrast to the 2-bit quantized network shown previously in Fig. 2 with -1, 0, and +1 as weights. The architecture of this network was chosen to be a multilayer perceptron (MLP), with 784 input units, 10 output units, and 2 hidden layers of

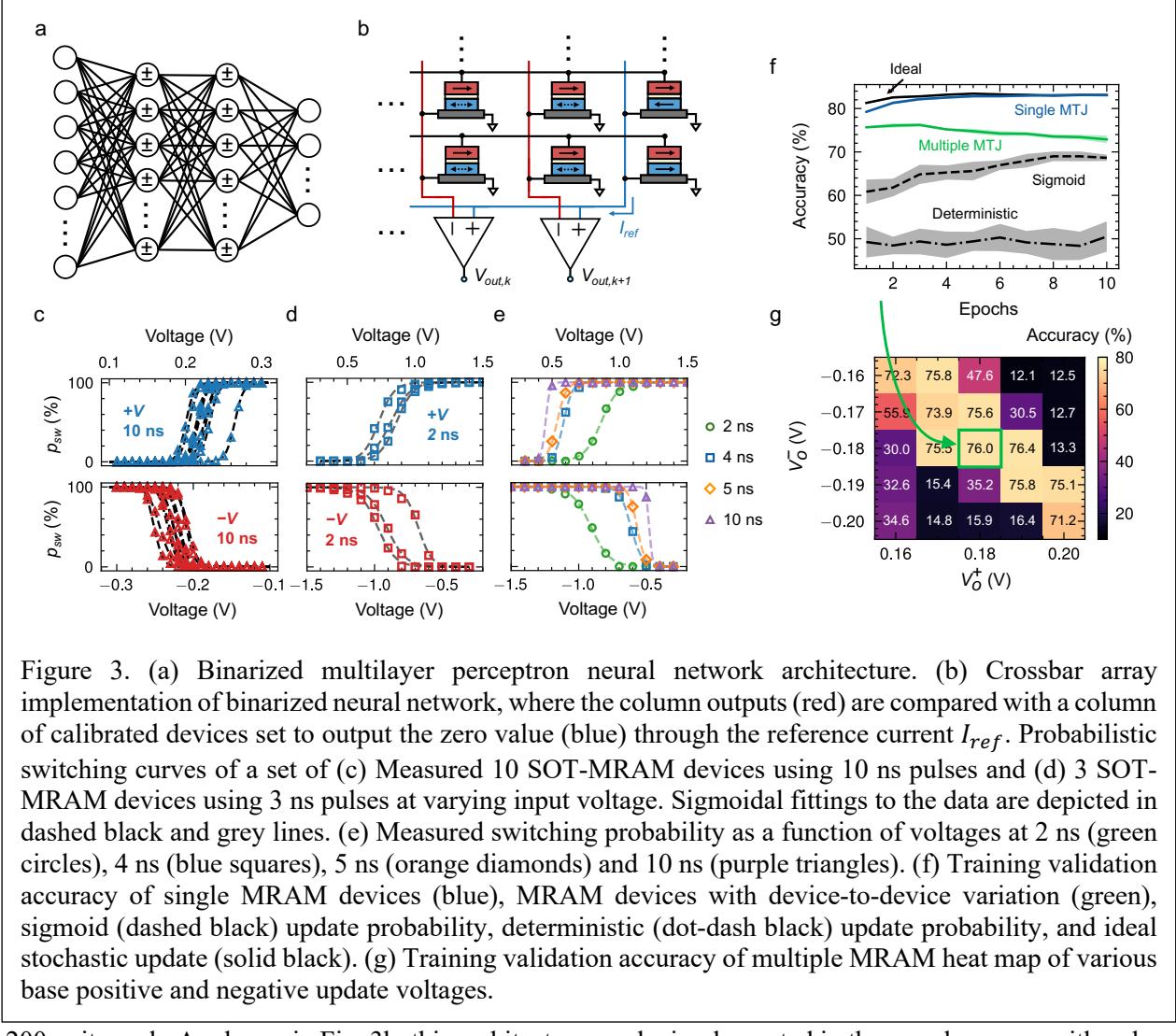


Figure 3. (a) Binarized multilayer perceptron neural network architecture. (b) Crossbar array implementation of binarized neural network, where the column outputs (red) are compared with a column of calibrated devices set to output the zero value (blue) through the reference current I_{ref} . Probabilistic switching curves of a set of (c) Measured 10 SOT-MRAM devices using 10 ns pulses and (d) 3 SOT-MRAM devices using 3 ns pulses at varying input voltage. Sigmoidal fittings to the data are depicted in dashed black and grey lines. (e) Measured switching probability as a function of voltages at 2 ns (green circles), 4 ns (blue squares), 5 ns (orange diamonds) and 10 ns (purple triangles). (f) Training validation accuracy of single MRAM devices (blue), MRAM devices with device-to-device variation (green), sigmoid (dashed black) update probability, deterministic (dot-dash black) update probability, and ideal stochastic update (solid black). (g) Training validation accuracy of multiple MRAM heat map of various base positive and negative update voltages.

200 units each. As shown in Fig. 3b, this architecture can be implemented in the crossbar array with only a single device per weight, which is then subtracted from a reference column. The output of the array is also simplified, allowing a single comparator to be used to determine the activation value as opposed to an analog-to-digital converter, often the costliest component in a crossbar array in terms of energy dissipation⁷². Training of this network was implemented in PyTorch⁷³ by modifying the Adam optimizer⁷⁴ with the lookup tables for 10 ns pulse duration shown in Fig. 3c. The training dataset was chosen to be Fashion-MNIST⁷⁵, split into 60,000 training images and 10,000 validation images. Though we perform the simulations with the 10 ns lookup tables due to having the highest number of measured devices, measurements for a shorter pulse duration of 2 ns are shown in Fig. 3d. Here, the voltage window for probabilistic switching is much

wider than the case for 10 ns, at approximately 0.5 V between 0% and 100% compared to 0.05 V for the 10 ns case, indicating that faster sampling is preferable to reduce the precision requirements of control circuitry. This is corroborated by Fig. 3e, where probability curves of varying pulse duration for a single device are shown.

Figure 3f shows the stochastic training performance in terms of predicted validation accuracy of the array of measured SOT-MRAM devices. When operating using a single MTJ, the network reaches 83% accuracy on average (blue curve), which is close to ideal for this network and task (black curve). When accounting for the device-to-device spread in threshold voltage shown in Fig. 3c, there is a noticeable drop in validation accuracy, maximizing at 76% on average (multiple MTJ green curve). The results are compared with the use of a deterministic update rule and a sigmoid update rule, both of which train worse than the SOT-MRAM network accounting for device-to-device variations. The deterministic update is calculated by taking the sign of the updated weight. The sigmoidal update is obtained by mapping the weights onto the probabilistic function:

$$p = \frac{1}{1 + e^{-\frac{w}{\sigma_s}}}$$

where the probability of switching p is dependent on the post-update weight level w and the width of the sigmoid σ_s . The deterministic update performs worse because small updates are immediately lost because of the rounding effect of taking the sign of the updated weight. As a result, only a sufficiently large update can change the weight, leading to quick saturation of learning. For the sigmoidal update, there is a small chance for weights at the extreme values of -1 and +1 to flip to the opposite value. This leads to an effective ambient noise during training that is detrimental to accuracy. This contrasts with the SOT-MRAM devices, which can only have a chance to switch to +1 (-1) during positive (negative) updates, leading to behaviors that are more similar to backpropagation-based learning rules. The ideal stochastic update follows these same characteristics and is described in Ref. ⁷⁶. While updates to other types of stochastic devices follow the same principle, the inherent variability of SOT-MRAM devices remain low because of magnetic anisotropy, which always forces the resistance state of the device to be either high or low, comparable to

the read noise of the device, unlike other devices such as RRAM or PCRAM, which can have resistances in the intermediate states. Additionally, since conductance drift in SOT-MRAM devices is solely due to degradation of the tunnel barrier, SOT-MRAM can achieve greater stability and predictability over time and cycles compared to RRAM and PCRAM, where the devices experience drift due to a host of causes such as ion diffusion, grain boundary differences, and generally high stress switching^{9,13,14}.

While the biases for each device of the SOT-MRAM array could be saved to prevent device-to-device variation impact by forcing all the probability curves to lie up, using the same biasing voltage for all devices is much more simplistic to implement. Figure 3g shows an evaluation of validation accuracy using different bias voltage values for the positive and negative updates for the measured distributions shown in Fig. 3c. The network performs the best when the biasing voltages are above ± 0.17 V. This is likely because if the biasing voltage is too low, then small updates would result in no change in the weight, a detriment to the network performance similar to that of the deterministic update. When the bias voltages are ± 0.20 V, there is also a loss in accuracy. This is because for a subset of devices, this bias voltage is enough to induce a large chance of switching the device at every update voltage. The result is similar to using a sigmoidal update, where there is an introduction of extra update noise that is not beneficial for training. The low accuracies in the lower left and upper right quadrants of the graph can be explained by the impact of asymmetry on neural network performance, where an asymmetric update severely reduces the performance of networks^{49,77}. From the results, we can conclude that while using SOT-MRAM devices to effectively implement crossbar arrays for binary neural network applications is feasible, the biasing of the update rules for the weights must be carefully chosen to preserve the accuracy of the implemented network.

Due to the tunability of the probabilistic switching characteristics of MTJs, various types of stochastic MTJs from superparamagnetic tunnel junctions^{78–80} to probabilistic-write STT-MRAM^{81,82} have been characterized and engineered to accelerate Ising and simulated annealing algorithms to solve optimization problems^{83–89}. Most of these approaches involve a single device used to generate a random bitstream and peripheral circuitry to process the analog signal. Here, the characterization of wafer-scale

SOT-MRAM allows a prediction of performance on applications involving many devices. One such example is the use of many small crossbar arrays to accelerate probabilistic graph model (PGM) calculations. PGMs can be applied to optimization problems such as steady-state estimation^{90,91} and the PageRank algorithm⁹², widely used for determining web-page relatedness⁹³. The relatedness of two webpage nodes can be represented by the proportion of pass vs. no pass that is sampled at the output. A simple PGM with three nodes 0, 1, and 2 is shown in Fig. 4a. The directional weights of the PGM describe a strength of connection that is represented by probability. Here, there are two possible paths for Node 0 to be connected to Node 2. There is a direct connection (orange) with a probability of 0.2 and an indirect connection passing through Node 1 (purple) with probabilities of 0.6 and 0.5. The crossbar array implementation of the graph is shown in Fig. 4b, where the two possible paths for node connection are shown (left, middle), along with the case where neither path is on (right). These small crossbars are sampled N times until a satisfactory distribution can be attained. In practice, many small crossbars can be fabricated to perform the sampling in parallel. Here, the device crossbars were generated and analyzed through Ngspice⁹⁴ through the PySpice package⁹⁵. The results of sampling these crossbars constructed from the measured device probability curves (p-curves) shown in Fig. 3c are shown in Fig. 4c. To sample multiple devices, the average of all p-curves was used to construct a lookup table to identify the necessary voltages to attain the targeted probability. The left graph of Fig. 4c depicts the outputs while only using the measured write noise of a single device, while the right graph depicts the outputs with the full device-to-device variation of the conductance. A threshold voltage to determine whether a sample is a pass or a no-pass is shown in the dashed black line, at 11.5 μ A, and the input voltage for sampling was set to be 10 mV.

From the right graph of Fig. 4c, it is evident that several of the samples cross the threshold, resulting in an unexpected reading compared to the actual configuration of the crossbar. This ratio of number of

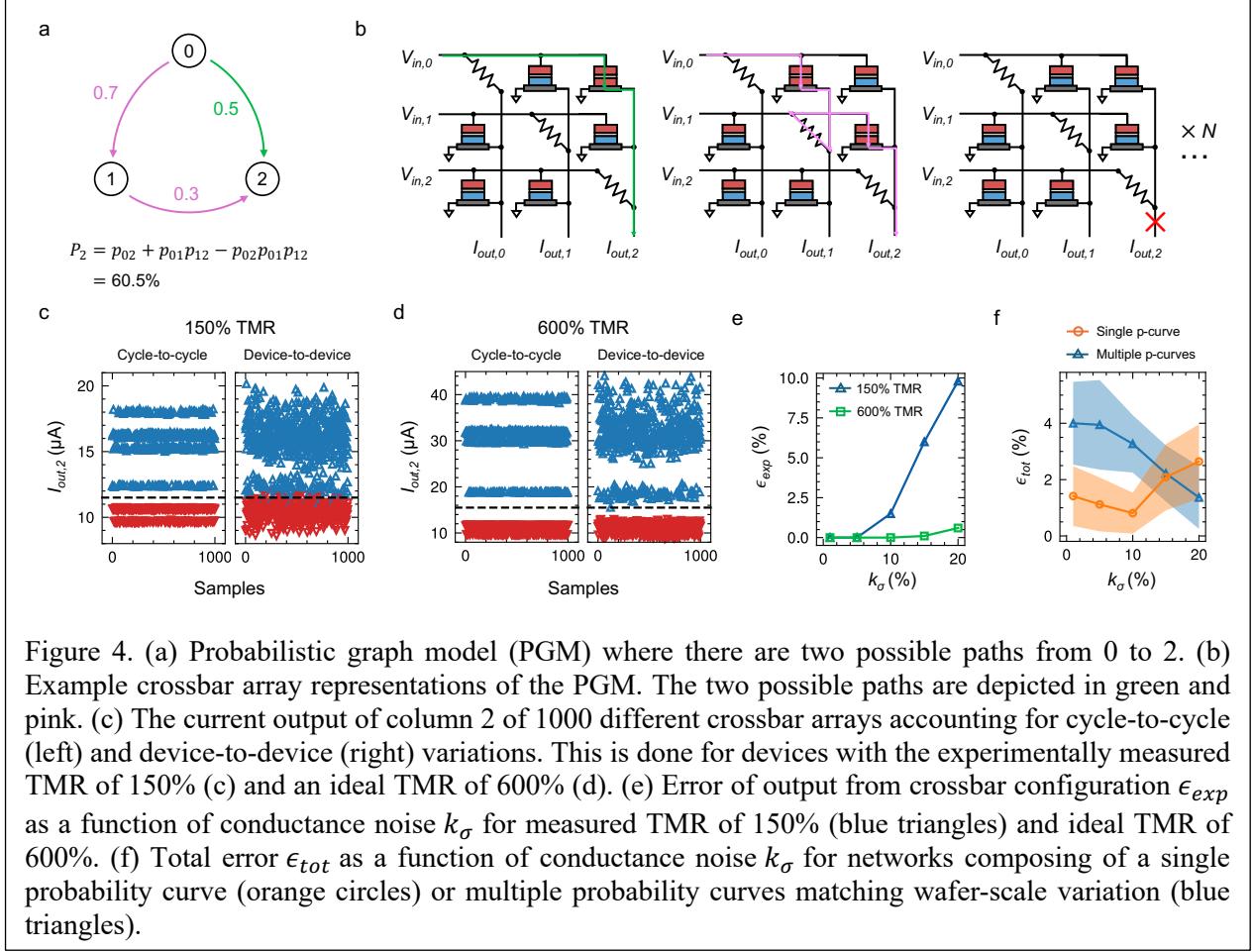


Figure 4. (a) Probabilistic graph model (PGM) where there are two possible paths from 0 to 2. (b) Example crossbar array representations of the PGM. The two possible paths are depicted in green and pink. (c) The current output of column 2 of 1000 different crossbar arrays accounting for cycle-to-cycle (left) and device-to-device (right) variations. This is done for devices with the experimentally measured TMR of 150% (c) and an ideal TMR of 600% (d). (e) Error of output from crossbar configuration ϵ_{exp} as a function of conductance noise k_σ for measured TMR of 150% (blue triangles) and ideal TMR of 600%. (f) Total error ϵ_{tot} as a function of conductance noise k_σ for networks composing of a single probability curve (orange circles) or multiple probability curves matching wafer-scale variation (blue triangles).

expectation errors to the total number of samples is described as ϵ_{exp} . The window where a threshold can be set is small due to the limited on-off ratio of the devices coupled with the device-to-device variation. In Fig. 4d, the same sampling was done again with a projection of the TMR to an idealized value of 600%. Here, even when accounting for the measured device-to-device variations of the wafer, there is a sufficient output current window to set a solid threshold. Figure 4e compares the ϵ_{exp} for the measured 150% TMR devices and the projected 600% TMR at varying levels of device-to-device variation k_σ . The results indicate that ϵ_{exp} remains low and rises much more slowly for the 600% TMR devices compared to the 150% TMR devices, indicating that a design metric of higher on/off ratio can mitigate errors introduced from device-to-device variation of the conductance, which becomes more important as the complexity and dimension of the PGM increases.

While the error of expectation ϵ_{exp} might be large because of many samples crossing the threshold due to device-to-device variation, if the threshold is set appropriately between the lowest pass level and highest no-pass level, the statistical result of the actual error can be close to 0% if the distribution of false pass results and false no-pass results is roughly equal. Here, we define the total error to be $\epsilon_{tot} = |P_{2,out} - P_{2,ideal}|$ where $P_{2,out}$ is the sampled probability of a pass and $P_{2,ideal}$ is the ideal probability of a pass. As seen in Fig. 3c, the measured devices have a significant spread in the threshold voltages for the p-curves. In Fig. 4f, the difference in ϵ_{tot} between wafer-scale device-to-device variation in p-curve spread (blue triangles) and the ideal situation of having a single sampled p-curve (orange squares) is analyzed as a function of device-to-device conductance noise k_σ . At k_σ under 15%, the error matches intuitive results, where the PGMs sampling from a single p-curve performs with lower error than the PGMs sampled from all the measured p-curves. Of note however, at k_σ equal to and greater than 15%, this trend reverses. This is most likely because the multiple device lookup table for switching voltages was constructed from the average of all devices, leading to a skewed result for the output at low noise. This is not the case for the single p-curve PGMs, where the lookup table was calibrated perfectly. As a result of the accurate calibration of the single p-curve, deviations due to conductance noise k_σ increase ϵ_{tot} . In contrast, for the multiple p-curve PGMs, the increase conductance noise k_σ smears the error function for the output probability, in this specific case leading to an increased overlap between the expected output probability $P_{2,ideal}$ and the output probability of the PGMs $P_{2,out}$, a beneficial outcome of the interplay between the error introduced by threshold voltage noise and conductance noise. These results show the most important improvement area for SOT-MRAM for PGM representation is to increase TMR, leading to a larger threshold window and larger simulated PGMs, while errors resulting from device-to-device variation can be mitigated. The interplay between threshold voltage variations and conductance variations can be engineered to reduce errors for PGM sampling of non-ideal devices.

Device	Speed	Energy	Endurance	On/off ratio	Write noise	Number of states	Device-to-device
SOT-MRAM (this work)	2 ns	350 fJ	$>10^{12}$	2.5	0.107%	2 states	~10%
STT-MRAM ¹⁷	90 ns	27 pJ (cell)	$>10^8$	2	Not available	2 states	Not available
RRAM ⁷¹	10 ns	100 pJ	Not available	13	~11.5%	8 states	~23.1%
RRAM ¹⁶	5 ms	787.5 pJ	$>10^5$	7	~1%	16 states	Not available
RRAM ¹⁵	80 μ s	400 pJ	$>10^5$	~10	~1%	16 states	~29%
ECRAM ²³	100 ns	3.5 μ J	$>10^7$	5	<1%	~90 states	~1.1%
PCRAM ¹¹	50 ns	1 pJ	$>10^{11}$	~100	~5%	12 states	~9%
PCRAM ⁸	<100 ns	~2 nJ	Not available	~100	~8%	~12 states	Not available

Table 1. Comparison of synaptic devices with wafer-scale production ability.

Overall, the advantages of the SOT-MRAM in this work are fast speed, low energy dissipation, and high endurance compared to other memories at similar maturity, shown in Table 1. The applications analyzed in this work show how the application can play to those strengths. Memory-efficient 2-bit neural networks with QAT mitigate limited resistance states and device-to-device variation, and binary neural networks and PGM representation leverage the advantages of magnetic anisotropy and probabilistic switching. Additionally, the demonstrated high endurance and low write energy dissipation allowing flexibility even if many write cycles are necessary.

In conclusion, we have shown that SOT-MRAM is a stand-out memory for edge computing applications. Measurements of wafer-scale fabricated SOT-MRAM devices from a 4Kb memory array are used to evaluate SOT-MRAM's effectiveness on several edge-specific applications, including 2-bit quantized inference, stochastic training of binary neural networks, and PGM representation. We identify implementation strategies for the outlined applications and identify increased TMR and reduced device-to-device variation as avenues of improvement for SOT-MRAM, but through comparisons with other technologies such as RRAM and ECRAM, show that SOT-MRAM is a leading candidate for edge computing.

Methods

The SOT-MRAM devices were fabricated by patterning with electron-beam lithography followed by etching using a hard mask reactive ion etch followed by an Ar inductively couple plasma etch, which was done to avoid damage to the sidewalls of the MTJ. Lastly, an ion beam etch was performed at low energy to eliminate redeposition on the MTJ sidewalls and preserve the SOT underlayer.

References

1. G. Finocchio, J. A. C. Incorvia, J. S. Friedman, Q. Yang, A. Giordano, J. Grollier, H. Yang, F. Ciubotaru, A. V Chumak, A. J. Naeemi, S. D. Cotofana, R. Tomasello, C. Panagopoulos, M. Carpentieri, P. Lin, G. Pan, J. J. Yang, A. Todri-Sanial, G. Boschetto, *et al.* Roadmap for unconventional computing with nanotechnology. *Nano Futures* **8**, 012001 (2024).
2. I. Chakraborty, M. Ali, A. Ankit, S. Jain, S. Roy, S. Sridharan, A. Agrawal, A. Raghunathan & K. Roy. Resistive Crossbars as Approximate Hardware Building Blocks for Machine Learning: Opportunities and Challenges. *Proceedings of the IEEE* **108**, 2276–2310 (2020).
3. Q. Xia & J. J. Yang. Memristive crossbar arrays for brain-inspired computing. *Nat Mater* **18**, 309–323 (2019).
4. P. Merolla, J. Arthur, F. Akopyan, N. Imam, R. Manohar & D. S. Modha. A digital neurosynaptic core using embedded crossbar memory with 45pJ per spike in 45nm. in *2011 IEEE Custom Integrated Circuits Conference (CICC)* 1–4 (IEEE, 2011). doi:10.1109/CICC.2011.6055294
5. Y. Han, T. Li, X. Cheng, L. Wang, J. Han, Y. Zhao & X. Zeng. Radiation Hardened 12T SRAM With Crossbar-Based Peripheral Circuit in 28nm CMOS Technology. *IEEE Transactions on Circuits and Systems I: Regular Papers* **68**, 2962–2975 (2021).
6. F. L. Traversa, F. Bonani, Y. V Pershin & M. Di Ventra. Dynamic computing random access memory. *Nanotechnology* **25**, 285201 (2014).
7. T. P. Xiao, B. Feinberg, C. H. Bennett, V. Agrawal, P. Saxena, V. Prabhakar, K. Ramkumar, H. Medu, V. Raghavan, R. Chettuvetty, S. Agarwal & M. J. Marinella. An Accurate, Error-Tolerant, and Energy-Efficient Neural Network Inference Engine Based on SONOS Analog Memory. *IEEE Transactions on Circuits and Systems I: Regular Papers* **69**, 1480–1493 (2022).
8. M. Le Gallo, R. Khaddam-Aljameh, M. Stanisavljevic, A. Vasilopoulos, B. Kersting, M. Dazzi, G. Karunaratne, M. Brändli, A. Singh, S. M. Müller, J. Büchel, X. Timoneda, V. Joshi, M. J. Rasch, U. Egger, A. Garofalo, A. Petropoulos, T. Antonakopoulos, K. Brew, *et al.* A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference. *Nat Electron* **6**, 680–693 (2023).
9. M. Le Gallo & A. Sebastian. An overview of phase-change memory device physics. *J Phys D Appl Phys* **53**, (2020).
10. G. W. Burr, M. J. BrightSky, A. Sebastian, H.-Y. Cheng, J.-Y. Wu, S. Kim, N. E. Sosa, N. Papandreou, H.-L. Lung, H. Pozidis, E. Eleftheriou & C. H. Lam. Recent Progress in Phase-Change <?Pub _newline ?> Memory Technology. *IEEE J Emerg Sel Top Circuits Syst* **6**, 146–162 (2016).
11. Z. Song, D. Cai, Y. Cheng, L. Wang, S. Lv, T. Xin & G. Feng. 12-state multi-level cell storage implemented in a 128 Mb phase change memory chip. *Nanoscale* **13**, 10455–10461 (2021).
12. W. Wan, R. Kubendran, C. Schaefer, S. B. Eryilmaz, W. Zhang, D. Wu, S. Deiss, P. Raina, H. Qian, B. Gao, S. Joshi, H. Wu, H.-S. P. Wong & G. Cauwenberghs. A compute-in-memory chip based on resistive random-access memory. *Nature* **608**, 504–512 (2022).

13. F. Pan, S. Gao, C. Chen, C. Song & F. Zeng. Recent progress in resistive random access memories: Materials, switching mechanisms, and performance. *Materials Science and Engineering R: Reports* **83**, 1–59 (2014).
14. H. Wang & X. Yan. Overview of Resistive Random Access Memory (RRAM): Materials, Filament Mechanisms, Performance Optimization, and Prospects. *physica status solidi (RRL) – Rapid Research Letters* **13**, (2019).
15. S. Choi, S. S. Bezugam, T. Bhattacharya, D. Kwon & D. B. Strukov. Wafer-scale fabrication of memristive passive crossbar circuits for brain-scale neuromorphic computing. *Nat Commun* **16**, 8757 (2025).
16. J. Park, A. Kumar, Y. Zhou, S. Oh, J.-H. Kim, Y. Shi, S. Jain, G. Hota, E. Qiu, A. L. Nagle, I. K. Schuller, C. D. Schuman, G. Cauwenberghs & D. Kuzum. Multi-level, forming and filament free, bulk switching trilayer RRAM for neuromorphic computing at the edge. *Nat Commun* **15**, 3492 (2024).
17. S. Jung, H. Lee, S. Myung, H. Kim, S. K. Yoon, S.-W. Kwon, Y. Ju, M. Kim, W. Yi, S. Han, B. Kwon, B. Seo, K. Lee, G.-H. Koh, K. Lee, Y. Song, C. Choi, D. Ham & S. J. Kim. A crossbar array of magnetoresistive memory devices for in-memory computing. *Nature* **601**, 211–216 (2022).
18. M. Y. Song, C. M. Lee, S. Y. Yang, G. L. Chen, K. M. Chen, I. J. Wang, Y. C. Hsin, K. T. Chang, C. F. Hsu, S. H. Li, J. H. Wei, T. Y. Lee, M. F. Chang, X. Y. Bao, C. H. Diaz & S. J. Lin. High speed (1ns) and low voltage (1.5V) demonstration of 8Kb SOT-MRAM array. in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* 377–378 (IEEE, 2022). doi:10.1109/VLSITechnologyandCir46769.2022.9830149
19. D. Kireev, S. Liu, H. Jin, T. Patrick Xiao, C. H. Bennett, D. Akinwande & J. A. C. Incorvia. Metaplastic and energy-efficient biocompatible graphene artificial synaptic transistors for enhanced accuracy neuromorphic computing. *Nat Commun* **13**, 4386 (2022).
20. Y. Zheng, S. Ghosh & S. Das. A Butterfly-Inspired Multisensory Neuromorphic Platform for Integration of Visual and Chemical Cues. *Advanced Materials* **36**, (2024).
21. L. Qin, P. Guan, J. Shao, Y. Xiao, Y. Yu, J. Su, C. Zhang, Y. Li, S. Liu, P. Li, D. Ouyang, W. He, F. Liu, K. Zhu, K. Liu, Z. Yao, J. Wu, Y. Zhao, H. Li, et al. Molecular crystal memristors. *Nat Nanotechnol* (2025). doi:10.1038/s41565-025-02013-z
22. Y. Shen, K. Zhu, Y. Xiao, D. Waldhör, A. H. Basher, T. Knobloch, S. Pazos, X. Liang, W. Zheng, Y. Yuan, J. B. Roldan, U. Schwingenschlögl, H. Tian, H. Wu, T. F. Schranghamer, N. Trainor, J. M. Redwing, S. Das, T. Grasser, et al. Two-dimensional-materials-based transistors using hexagonal boron nitride dielectrics and metal gate electrodes with high cohesive energy. *Nat Electron* **7**, 856–867 (2024).
23. P. Chen, F. Liu, P. Lin, P. Li, Y. Xiao, B. Zhang & G. Pan. Open-loop analog programmable electrochemical memory array. *Nat Commun* **14**, 6184 (2023).
24. X. Yao, K. Klyukin, W. Lu, M. Onen, S. Ryu, D. Kim, N. Emond, I. Waluyo, A. Hunt, J. A. del Alamo, J. Li & B. Yildiz. Protomic solid-state electrochemical synapse for physical neural networks. *Nat Commun* **11**, 3134 (2020).

25. Y. Li, T. P. Xiao, C. H. Bennett, E. Isele, A. Melianas, H. Tao, M. J. Marinella, A. Salleo, E. J. Fuller & A. A. Talin. In situ Parallel Training of Analog Neural Network Using Electrochemical Random-Access Memory. *Front Neurosci* **15**, (2021).
26. Y. van de Burgt, E. Lubberman, E. J. Fuller, S. T. Keene, G. C. Faria, S. Agarwal, M. J. Marinella, A. Alec Talin & A. Salleo. A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing. *Nat Mater* **16**, 414–418 (2017).
27. H. M. Resalat Faruque, M. A. Islam, M. J. Voegtle, N. Holtzman, I.-T. Bae, S. S. Swain, K. Beom, R. C. Dempsey, C. J. Smith, T. P. Xiao, C. H. Bennett, S. Agarwal, M. N. Kozicki, A. A. Talin & M. J. Marinella. Dual-ion ECRAm as a stable and accurate analog synapse. *Device* **3**, 100928 (2025).
28. A. A. T. Talin. Harnessing ion tuning mechanisms for neuromorphic computing: from artificial synapses to dynamically reconfigurable architectures. in *Low-Dimensional Materials and Devices 2024* (eds. Kobayashi, N. P., Talin, A. A., Davydov, A. V. & Islam, M. S.) 8 (SPIE, 2024). doi:10.1117/12.3029400
29. D.-F. Shao & E. Y. Tsymbal. Antiferromagnetic tunnel junctions for spintronics. *npj Spintronics* **2**, 13 (2024).
30. P. Qin, H. Yan, X. Wang, H. Chen, Z. Meng, J. Dong, M. Zhu, J. Cai, Z. Feng, X. Zhou, L. Liu, T. Zhang, Z. Zeng, J. Zhang, C. Jiang & Z. Liu. Room-temperature magnetoresistance in an all-antiferromagnetic tunnel junction. *Nature* **613**, 485–489 (2023).
31. Y. C. Ong, Y. H. Chen, H. H. Wang, J. Q. Liang, Y. S. Chen, J. Huang, T. W. Chiang, J. C. Huang, C. H. Weng, C. Y. Wang, B. P. H. Lee, A. Y. J. Wang, K. C. Huang & H. Chuang. Design-Technology-Reliability Co-Optimization for MRAM-OTP Integration — A Methodological Approach. in *2025 IEEE International Reliability Physics Symposium (IRPS) 01–06* (IEEE, 2025). doi:10.1109/IRPS48204.2025.10982744
32. M. G. Gottwald, G. Hu, P. L. Trouilloud, L. Rehm, C. Safranski, G. Kim, S. L. Brown, J. Bruley, C. P. D'Emic, O. Gunawan, H. Jung, C. Lavoie, J. Lee, J. Liang, M. Robbins, J. Z. Sun, P. Hashemi & D. C. Worledge. First Demonstration of High Retention Energy Barriers and 2 ns Switching, Using Magnetic Ordered-Alloy-Based STT MRAM Devices. in *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits) 1–2* (IEEE, 2024). doi:10.1109/VLSITechnologyandCir46783.2024.10631319
33. E. Liu, W. Yang, K. Zhou, Y. Gao, Z. Ji, D. Zeng, M. Wang, Q. Li, Y. Xi, D. Yang, G. Chen, H. Zhou, Y. Sun, Z. Zheng, Q. Guo, Q. Dai, F. Meng & S. He. A Novel Channel-Less SOT-MRAM with 115% TMR, 2 ns Switching, and High Bit Yield (>99.9%). in *2024 IEEE International Electron Devices Meeting (IEDM) 1–4* (IEEE, 2024). doi:10.1109/IEDM50854.2024.10873500
34. S. Ko, J. Shim, J. H. Park, W. Lim, H. Jung, J. H. Bak, D. Jeong, J. W. Lee, H. Whang, M. Eom, D. Shin, J. Lee, S. Noh, J. Yang, J.-H. Park, Y. Kim, C. Kim, J. H. Kim, T. Y. Lee, et al. Key Technologies of Scaling Embedded MRAM to 8nm Logic and Beyond for Automotive Application. in *2024 IEEE International Electron Devices Meeting (IEDM) 1–4* (IEEE, 2024). doi:10.1109/IEDM50854.2024.10873495

35. F. Yasin, A. Palomino, A. Kumar, V. Pica, S. Van Beek, G. Talmelli, V. D. Nguyen, S. Cosemans, D. Crotti, K. Wostyn, G. S. Kar & S. Couet. Extremely Scaled Perpendicular SOT-MRAM Array Integration on 300mm Wafer. in *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits) 1–2* (IEEE, 2024).
doi:10.1109/VLSITechnologyandCir46783.2024.10631340
36. H. Zhang, X. Ma, C. Jiang, J. Yin, S. Lyu, S. Lu, X. Shang, B. Man, C. Zhang, D. Li, S. Li, W. Chen, H. Liu, G. Wang, K. Cao, Z. Wang & W. Zhao. Integration of high-performance spin-orbit torque MRAM devices by 200-mm-wafer manufacturing platform. *Journal of Semiconductors* **43**, 102501 (2022).
37. K. Jhuria, J. Hohlfeld, A. Pattabi, E. Martin, A. Y. Arriola Córdova, X. Shi, R. Lo Conte, S. Petit-Watelot, J. C. Rojas-Sánchez, G. Malinowski, S. Mangin, A. Lemaître, M. Hehn, J. Bokor, R. B. Wilson & J. Gorchon. Spin-orbit torque switching of a ferromagnet with picosecond electrical pulses. *Nat Electron* **3**, 680–686 (2020).
38. P. Kumar & A. Naeemi. Benchmarking of spin-orbit torque vs spin-transfer torque devices. *Appl Phys Lett* **121**, (2022).
39. G. Jan, Y.-J. Wang, T. Moriyama, Y.-J. Lee, M. Lin, T. Zhong, R.-Y. Tong, T. Torng & P.-K. Wang. High Spin Torque Efficiency of Magnetic Tunnel Junctions with MgO/CoFeB/MgO Free Layer. *Applied Physics Express* **5**, 093008 (2012).
40. C.-Y. Hu, Y.-F. Chiu, C.-C. Tsai, C.-C. Huang, K.-H. Chen, C.-W. Peng, C.-M. Lee, M.-Y. Song, Y.-L. Huang, S.-J. Lin & C.-F. Pai. Toward 100% Spin-Orbit Torque Efficiency with High Spin-Orbital Hall Conductivity Pt-Cr Alloys. *ACS Appl Electron Mater* **4**, 1099–1108 (2022).
41. V. D. Nguyen, S. Rao, K. Wostyn & S. Couet. Recent progress in spin-orbit torque magnetic random-access memory. *npj Spintronics* **2**, 48 (2024).
42. J. Jeong, Y. Jang, M. Kang, S. Hwang, J. Park & B. Park. Spintronic Artificial Synapses Using Voltage-Controlled Multilevel Magnetic States. *Adv Electron Mater* **10**, (2024).
43. J. Liu, T. Xu, H. Feng, L. Zhao, J. Tang, L. Fang & W. Jiang. Compensated Ferrimagnet Based Artificial Synapse and Neuron for Ultrafast Neuromorphic Computing. *Adv Funct Mater* **32**, (2022).
44. D. Koh, D.-J. Kim, T. Kim, M. Kang, D. D. Viet, H. Hong, J.-R. Jeong, J. Park & B.-G. Park. Multilevel Nanoarray Spin-Orbit Torque Device for Process-in-Memory Applications. *Nano Lett* (2025).
doi:10.1021/acs.nanolett.5c02634
45. C. Cui, S. Liu, J. Kwon & J. A. C. Incovia. Spintronic Artificial Neurons Showing Integrate-and-Fire Behavior with Reliable Cycling Operation. *Nano Lett* **25**, 361–367 (2025).
46. T. Leonard, N. Zogbi, S. Liu, W. S. Rogers, C. H. Bennett & J. A. C. Incovia. Shape Anisotropy-Dependent Leaking in Magnetic Neurons for Bio-Mimetic Neuromorphic Computing. *ACS Nano* **19**, 3470–3477 (2025).

47. S. A. Siddiqui, S. Dutta, A. Tang, L. Liu, C. A. Ross & M. A. Baldo. Magnetic Domain Wall Based Synaptic and Activation Function Generator for Neuromorphic Accelerators. *Nano Lett* **20**, 1033–1040 (2020).
48. M. S. Alam, W. al Misba & J. Atulasimha. Quantized non-volatile nanomagnetic domain wall synapse based autoencoder for efficient unsupervised network anomaly detection. *Neuromorphic Comput and Eng* **4**, 024012 (2024).
49. T. Leonard, S. Liu, M. Alamdar, H. Jin, C. Cui, O. G. Akinola, L. Xue, T. P. Xiao, J. S. Friedman, M. J. Marinella, C. H. Bennett & J. A. C. Incorvia. Shape-Dependent Multi-Weight Magnetic Artificial Synapses for Neuromorphic Computing. *Adv Electron Mater* **8**, 2200563 (2022).
50. S. Liu, T. P. Xiao, C. Cui, J. A. C. Incorvia, C. H. Bennett & M. J. Marinella. A domain wall-magnetic tunnel junction artificial synapse with notched geometry for accurate and efficient training of deep neural networks. *Appl Phys Lett* **118**, (2021).
51. K. M. Song, J. S. Jeong, B. Pan, X. Zhang, J. Xia, S. Cha, T. E. Park, K. Kim, S. Finizio, J. Raabe, J. Chang, Y. Zhou, W. Zhao, W. Kang, H. Ju & S. Woo. Skyrmion-based artificial synapses for neuromorphic computing. *Nat Electron* **3**, 148–155 (2020).
52. Y. Sun, T. Lin, N. Lei, X. Chen, W. Kang, Z. Zhao, D. Wei, C. Chen, S. Pang, L. Hu, L. Yang, E. Dong, L. Zhao, L. Liu, Z. Yuan, A. Ullrich, C. H. Back, J. Zhang, D. Pan, *et al.* Experimental demonstration of a skyrmion-enhanced strain-mediated physical reservoir computing system. *Nat Commun* **14**, 3434 (2023).
53. T. da Câmara Santa Clara Gomes, Y. Sassi, D. Sanz-Hernández, S. Krishnia, S. Collin, M.-B. Martin, P. Seneor, V. Cros, J. Grollier & N. Reyren. Neuromorphic weighted sums with magnetic skyrmions. *Nat Electron* **8**, 204–214 (2025).
54. S. Chen, J. Lourembam, P. Ho, A. K. J. Toh, J. Huang, X. Chen, H. K. Tan, S. L. K. Yap, R. J. J. Lim, H. R. Tan, T. S. Suraj, M. I. Sim, Y. T. Toh, I. Lim, N. C. B. Lim, J. Zhou, H. J. Chung, S. Ter Lim & A. Soumyanarayanan. All-electrical skyrmionic magnetic tunnel junction. *Nature* **627**, 522–527 (2024).
55. G. Bernard, K. Cottart, M.-A. Syskaki, V. Porée, A. Resta, A. Nicolaou, A. Durnez, S. Ono, A. Mora Hernandez, J. Langer, D. Querlioz & L. Herrera Diez. Dynamic Control of Weight-Update Linearity in Magneto-Ionic Synapses. *Nano Lett* **25**, 1443–1450 (2025).
56. A. J. Tan, M. Huang, C. O. Avci, F. Büttner, M. Mann, W. Hu, C. Mazzoli, S. Wilkins, H. L. Tuller & G. S. D. Beach. Magneto-ionic control of magnetism using a solid-state proton pump. *Nat Mater* **18**, 35–41 (2019).
57. S. Liu, T. P. Xiao, J. Kwon, B. J. Debusschere, S. Agarwal, J. A. C. Incorvia & C. H. Bennett. Bayesian neural networks using magnetic tunnel junction-based probabilistic in-memory computing. *Frontiers in Nanotechnology* **4**, 1021943 (2022).
58. S. Das, R. Mansell, L. Flajšman, M.-A. Syskaki, J. Langer & S. van Dijken. Magneto-ionic synapse for reservoir computing. *Phys Rev Appl* **23**, 054043 (2025).

59. P. Monalisha, Z. Ma, E. Pellicer, E. Menéndez & J. Sort. A Multilevel Magnetic Synapse Based on Voltage-Tuneable Magnetism by Nitrogen Ion Migration. *Adv Electron Mater* **9**, (2023).
60. P. Monalisha, M. Ameziane, I. Spasojevic, E. Pellicer, R. Mansell, E. Menéndez, S. van Dijken & J. Sort. Magnetoionics for Synaptic Devices and Neuromorphic Computing: Recent Advances, Challenges, and Future Perspectives. *Small Science* (2024). doi:10.1002/smss.202400133
61. X. Sun & S. Yu. Impact of Non-Ideal Characteristics of Resistive Synaptic Devices on Implementing Convolutional Neural Networks. *IEEE J Emerg Sel Top Circuits Syst* **9**, 570–579 (2019).
62. J. Xu, H. Liu, X. Peng, Z. Duan, X. Liao & H. Jin. A Cascaded ReRAM-based Crossbar Architecture for Transformer Neural Network Acceleration. *ACM Transact Des Autom Electron Syst* **30**, 1–23 (2025).
63. L. Wei, Z. Ma, C. Yang & Q. Yao. Advances in the Neural Network Quantization: A Comprehensive Review. *Applied Sciences* **14**, 7445 (2024).
64. Y. Nahshan, B. Chmiel, C. Baskin, E. Zheltonozhskii, R. Banner, A. M. Bronstein & A. Mendelson. Loss aware post-training quantization. *Mach Learn* **110**, 3245–3262 (2021).
65. F. and others Chollet. Keras. <https://keras.io> (2015).
66. Li Deng. The MNIST Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Process Mag* **29**, 141–142 (2012).
67. Ben Feinberg, T. Patrick Xiao, Curtis J. Brinker, Christopher H. Bennett, Matthew J. Marinella & Sapan Agarwal. CrossSim: accuracy simulation of analog in-memory computing. <https://github.com/sandialabs/cross-sim>
68. G. An. The Effects of Adding Noise During Backpropagation Training on a Generalization Performance. *Neural Comput* **8**, 643–674 (1996).
69. T. Leonard, S. Liu, H. Jin & J. A. C. Incorvia. Stochastic domain wall-magnetic tunnel junction artificial neurons for noise-resilient spiking neural networks. *Appl Phys Lett* **122**, 262406 (2023).
70. H. Noh, T. You, J. Mun & B. Han. Regularizing Deep Neural Networks by Noise: Its Interpretation and Optimization. *Adv Neural Inf Process Syst* 5101–5118 (2017).
71. C. H. Bennett, D. Garland, R. B. Jacobs-Gedrim, S. Agarwal & M. J. Marinella. Wafer-Scale TaO_x Device Variability and Implications for Neuromorphic Computing Applications. in *2019 IEEE International Reliability Physics Symposium (IRPS)* 1–4 (IEEE, 2019). doi:10.1109/IRPS.2019.8720596
72. F. Aguirre, A. Sebastian, M. Le Gallo, W. Song, T. Wang, J. J. Yang, W. Lu, M.-F. Chang, D. Ielmini, Y. Yang, A. Mehonic, A. Kenyon, M. A. Villena, J. B. Roldán, Y. Wu, H.-H. Hsu, N. Raghavan, J. Suñé, E. Miranda, *et al.* Hardware implementation of memristor-based artificial neural networks. *Nat Commun* **15**, 1974 (2024).
73. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances*

in Neural Information Processing Systems (eds. Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. & Garnett, R.) **32**, (Curran Associates, Inc., 2019).

74. D. P. Kingma & J. Ba. Adam: A Method for Stochastic Optimization. (2014).
75. H. Xiao, K. Rasul & R. Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747* (2017).
76. T. Hirtzlin, B. Penkovsky, M. Bocquet, J.-O. Klein, J.-M. Portal & D. Querlioz. Stochastic Computing for Hardware Implementation of Binarized Neural Networks. *IEEE Access* **7**, 76394–76403 (2019).
77. A. Laborieux, M. Ernoult, T. Hirtzlin & D. Querlioz. Synaptic metaplasticity in binarized neural networks. *Nat Commun* **12**, 2549 (2021).
78. D. Vodenicarevic, N. Locatelli, A. Mizrahi, J. S. Friedman, A. F. Vincent, M. Romera, A. Fukushima, K. Yakushiji, H. Kubota, S. Yuasa, S. Tiwari, J. Grollier & D. Querlioz. Low-Energy Truly Random Number Generation with Superparamagnetic Tunnel Junctions for Unconventional Computing. *Phys Rev Appl* **8**, 1–9 (2017).
79. W. A. Borders, A. Z. Pervaiz, S. Fukami, K. Y. Camsari, H. Ohno & S. Datta. Integer factorization using stochastic magnetic tunnel junctions. *Nature* **573**, 390–393 (2019).
80. C. Safranski, J. Kaiser, P. Trouilloud, P. Hashemi, G. Hu & J. Z. Sun. Demonstration of Nanosecond Operation in Stochastic Magnetic Tunnel Junctions. *Nano Lett* **21**, 2040–2045 (2021).
81. A. Dubovski, T. Criss, A. S. El Valli, L. Rehm, A. D. Kent & A. Haas. One Trillion True Random Bits Generated With a Field-Programmable Gate Array Actuated Magnetic Tunnel Junction. *IEEE Magn Lett* **15**, 1–4 (2024).
82. A. Sidi El Valli, M. Tsao, J. D. Smith, S. Misra & A. D. Kent. High-speed tunable generation of random number distributions using actuated perpendicular magnetic tunnel junctions. *Appl Phys Lett* **126**, (2025).
83. S. Liu, J. Kwon, P. W. Bessler, S. G. Cardwell, C. Schuman, J. D. Smith, J. B. Aimone, S. Misra & J. A. C. Incorvia. Random Bitstream Generation Using Voltage-Controlled Magnetic Anisotropy and Spin Orbit Torque Magnetic Tunnel Junctions. *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* **8**, 194–202 (2022).
84. E. Raimondo, E. Garzón, Y. Shao, A. Grimaldi, S. Chiappini, R. Tomasello, N. Davila-Melendez, J. A. Katine, M. Carpentieri, M. Chiappini, M. Lanuzza, P. Khalili Amiri & G. Finocchio. High-Performance and Reliable Probabilistic Ising Machine Based on Simulated Quantum Annealing. *Phys Rev X* **15**, 041001 (2025).
85. C. Duffee, J. Athas, Y. Shao, N. D. Melendez, E. Raimondo, J. A. Katine, K. Y. Camsari, G. Finocchio & P. Khalili Amiri. An integrated-circuit-based probabilistic computer that uses voltage-controlled magnetic tunnel junctions as its entropy source. *Nat Electron* **8**, 784–793 (2025).
86. A. Maicke, J. Arzate, S. Liu, J. Kwon, J. D. Smith, J. B. Aimone, S. Misra, C. Schuman, S. G. Cardwell & J. A. C. Incorvia. Magnetic tunnel junction random number generators applied to dynamically tuned probability trees driven by spin orbit torque. *Nanotechnology* **35**, 275204 (2024).

87. A. Grimaldi, L. Mazza, E. Raimondo, P. Tullo, D. Rodrigues, K. Y. Camsari, V. Crupi, M. Carpentieri, V. Puliafito & G. Finocchio. Evaluating Spintronics-Compatible Implementations of Ising Machines. *Phys Rev Appl* **20**, 024005 (2023).
88. C. Delacour, M. M. H. Sajeeb, J. P. Hespanha & K. Y. Camsari. Two-dimensional parallel tempering for constrained optimization. *Phys Rev E* **112**, L023301 (2025).
89. M. M. H. Sajeeb, N. A. Aadit, S. Chowdhury, T. Wu, C. Smith, D. Chinmay, A. Raut, K. Y. Camsari, C. Delacour & T. Srimani. Scalable connectivity for Ising machines: Dense to sparse. *Phys Rev Appl* **24**, 014005 (2025).
90. A. Carpinone, M. Giorgio, R. Langella & A. Testa. Markov chain modeling for very-short-term wind power forecasting. *Electric Power Systems Research* **122**, 152–158 (2015).
91. L. Carrillo, J. A. Escobar, J. B. Clempner & A. S. Poznyak. Optimization problems in chemical reactions using continuous-time Markov chains. *J Math Chem* **54**, 1233–1254 (2016).
92. P. Berkhin. A Survey on PageRank Computing. *Internet Math* **2**, 73–120 (2005).
93. Y. H. Jang, S. H. Lee, J. Han, S. Cheong, S. K. Shim, J. Han, S. K. Ryoo & C. S. Hwang. Memristive Crossbar Array-Based Probabilistic Graph Modeling. *Advanced Materials* **36**, (2024).
94. Vera Albrecht, Phil Barker, Steven J. Borley, Stuart Brorson, Glaeser S. Dezai, Matt Flax, Daniele Foci, Alan Gillespie, Chris Inbody, Stefan Jones, Laurent Lemaitre, Paolo Nenzi, Arno W. Peters, Serban-Mihai Popescu, Georg Post, Emmanuel Rouat, Lionel Sainte Claque, Hitoshi Tanaka, Stephan Tiel, et al. Ngspice. <https://ngspice.sourceforge.io/index.html> (2008).
95. Fabrice Salvaire. PySpice. <https://pyspice.fabrice-salvaire.fr> (2021).