Abstract—We propose and analyze a compact and nonvolatile nanomagnetic (all-spin) nonbinary matrix multiplier performing the multiply-and-accumulate (MAC) operation using two magnetic tunnel junctions (MTJs)—one activated by strain to act as the multiplier and the other activated by spin-orbit torque pulses to act as a domain wall (DW) synapse that performs the operation of the accumulator. Each MAC operation can be performed in \(\sim 5\) ns and the energy dissipated per operation is \(\sim 500\) aJ. This provides a very useful hardware accelerator for machine learning and artificial intelligence tasks that often involve the multiplication of large matrices. The nonvolatility allows the matrix multiplier to be embedded in powerful non-von-Neumann architectures. It also allows all computing to be done at the edge while reducing the need to access the cloud, thereby making artificial intelligence more resilient against cyberattacks.

Index Terms—Domain wall (DW) synapse, magnetic tunnel junction (MTJ), matrix multiplication, straintronics.

I. INTRODUCTION

Artificial intelligence (AI) is pervasive and ubiquitous in modern life (smart cities, smart appliances, autonomous self-driving vehicles, information processing, speech recognition, patient monitoring, and so on). Estimates by OpenAI predict an explosive growth of computational requirements in AI by a factor of 100× every two years, which is a 50× faster rate than Moore’s law governing the evolution of the chip industry [1]. Most AI applications leverage machine learning (or deep learning based on neural networks) to perform two primary functions—training and inference. Algorithms for these tasks require multiplication of large matrices, such as in updating the synaptic weight matrices in deep learning networks, which is an essential feature of training a neuronal circuit, solving combinatorial optimization problems with Ising machines (e.g., min-cut or max-cut problems) and so on. A deep neural network (DNN) is a sequence of layers, each connected to the next through a matrix multiplication \(\{x\} \rightarrow [M]\{x\}\) representing synaptic connections. The input to the \((m+1)\)th layer is related to the \(m\)th layer as \(x_{i}^{m+1} = f(\sum_{j} M_{ij} x_{j}^{m})\), where \(f\) is a nonlinear activation function. Hardware accelerators that can perform matrix multiplications rapidly and efficiently are therefore very attractive since they can speed up AI tasks immensely. They are particularly useful in computer vision [2], image and other classification tasks [3], approximate computing [4], speech recognition [5], patient monitoring [6], and biomedicine [7].

The earliest ideas for devising hardware-based matrix multipliers date back to 1909. Percy Ludgate conceived of a machine made of mechanical parts that was understandably unwieldy, slow, and unreliable [8]. Modern matrix multipliers employ electronic charge-based circuitry that are fast, convenient, and reliable [9] but also energy-hungry and volatile, i.e., they lose all information once powered off. Recently, matrix multipliers have been implemented with optical networks [10], [11], which can be extremely energy-efficient and fast, but their drawback is the large footprint. They too are usually volatile since they use capacitors. In this article, we present an all-magnetic (all-spin) implementation of a matrix multiplier, which is energy efficient, fast, and has a much smaller footprint than its optical counterparts. Its most important advantage is that it is nonvolatile, and hence, the matrix products can be stored indefinitely in the device after powering off.

Consider the matrix multiplication operation \(c_{ij} = \sum_{m} a_{im} b_{mj}\). This operation consists of multiplying pairs of numbers (one member of the pair picked from a row of one matrix and the other from a column of the other matrix) and then adding up the products of the pairs to produce an element of the product matrix. Thus, one would need: 1) a “multiplier” to multiply pairs of numbers and 2) an “accumulator” (which accumulates the individual products and adds them up). These are the two ingredients of a hardware accelerator for matrix multiplication. In this work, we implement the multiplier with a single straintronic magnetic tunnel junction (MTJ) and the accumulator with another MTJ (driven by spin-orbit torque) acting as a domain wall (DW) synapse [12]. Each MTJ can have a footprint of \(\sim 50\) 000 nm\(^2\), and with all the peripherals, the footprint of the entire device can be \(<2\) \(\mu\)m\(^2\). The matrix multiplier can operate at clock rates of \(\sim 200\) MHz and dissipate \(\sim 500\) aJ of energy per multiply-and-accumulate.
(MAC) operation. In the next two sections, we describe the multiplier and the accumulator.

II. MULTIPLIER

A schematic of the proposed multiplier is shown in Fig. 1. It consists of an elliptical MTJ that has a (magnetically) “hard” and a “soft” layer, separated by an intervening insulating spacer layer. Any residual dipole interaction between the hard and soft layers creates an effective magnetic field. The soft layer is magnetostrictive and placed in a two-phase multiferroic). Two electrically shorted electrodes, delineated on the piezoelectric film, flank the MTJ, while the back of the substrate is connected to the ground.

When a (gate) voltage $V_G$ is applied to the shorted electrode pair, it generates biaxial strain in the piezoelectric film pinched between the two electrodes, which is transferred to the elliptical soft layer. The strain is either compressive along the major axis and tensile along the minor axis of the soft layer or vice versa, depending on the voltage polarity [13]. With the right voltage polarity, these strains rotate the soft layer’s magnetization away from the major axis of the ellipse (the easy axis) toward the minor axis (hard axis) because of the Villari effect. The rotation is opposed by the magnetic field $H_d$, which would like to keep the magnetization pointing along the initial orientation. The interplay of these two effects ultimately makes the magnetization settle into an orientation that subtends some angle $\theta_{ss}$ with the major axis (or the magnetization of the hard layer). The value of $\theta_{ss}$ depends on the applied strain and $H_d$ (it corresponds to the location of the potential energy minimum in the presence of both strain and $H_d$). Because the hard layer’s magnetization remains unaffected by strain, $H_d$ is fixed and does not change. Therefore, as we change $V_G$ and the resulting strain, we will change $\theta_{ss}$ and, consequently, the MTJ resistance, which depends on $\theta_{ss}$. The working of the basic straintronic MTJ (s-MTJ) was experimentally demonstrated in [14].

To implement the multiplier, a constant current source $I_{bias}$ is connected between the hard and soft layers of the s-MTJ (terminals “1” and “2”), as shown in Fig. 1(a). This drives a current through the s-MTJ. The gate voltage $V_G$ is applied at terminal “3” to generate the strain in the soft layer, and a fourth terminal is connected to the hard layer (common with terminal “1’), which outputs a voltage $V_0$. Terminal 2, connected to the soft layer, is grounded, and hence, $V_0 = R_{s-MTJ}I_{bias}$, where $R_{s-MTJ}$ is the resistance of the s-MTJ that can be altered by the gate voltage $V_G$ generating strain, as explained before.

A. Rotation of the Soft Layer’s Magnetization Due to the Gate Voltage

We have modeled the rotation of the soft layer’s magnetization as a function of the gate voltage $V_G$ in the presence of $H_d$ and thermal noise using stochastic Landau–Lifshitz–Gilbert simulations [15]. This allows us to find the $\theta_{ss}$ versus $V_G$ relation. The s-MTJ resistance at any gate voltage is given by $R_{s-MTJ} = R_P + (R_{AP} - R_P/2)[1 - \cos \theta_{ss}(V_G)]$, where $R_P$ is the s-MTJ resistance when the magnetizations of the hard and soft layers are mutually parallel and $R_{AP}$ is the s-MTJ resistance when the magnetizations are antiparallel. From the $\theta_{ss}$ versus $V_G$ relation, we can therefore calculate the $1/R_{s-MTJ} (= G_{s-MTJ})$ versus $V_G$ characteristic, which we show qualitatively in Fig. 1(b). With the proper choice of the s-MTJ parameters, we can produce a linear region in the $G_{s-MTJ}$ versus $V_G$ characteristic where $1/R_{s-MTJ} = 1/R_{AP} + \kappa(V_G - \delta) \approx G_{s-MTJ} = G_{AP} + \kappa(V_G - \delta)$ [where $\kappa$ and $\delta$ are constants]. We show this analytically in the Appendix. In Fig. 2, we plot the $\theta_{ss}$ versus $V_G$ characteristics obtained from the stochastic Landau Lifshitz Gilbert simulation and the resulting $G_{s-MTJ}$ versus $V_G$ plot. The simulation procedure is described in [15] and the Appendix. The parameters for the elliptical soft layer of the s-MTJ used in the simulation are given in Table I. The soft layer is assumed to be made of Terfenol-D, which has large magnetostriction. The value of $H_d$ can be altered arbitrarily.
TABLE I
PARAMETERS FOR THE SOFT LAYER OF THE MTJ

| Parameter                              | Value |
|----------------------------------------|-------|
| Major axis dimension (L)               | 800 nm |
| Minor axis dimension (W)               | 700 nm |
| Thickness (d)                          | 2.2 nm |
| Saturation magnetization (Ms)          | 8.5 × 10^5 A/m |
| Dipole coupling field (Hd)             | 1000 Oe |
| Gilbert damping constant (α)           | 0.1   |
| Saturation magnetostriiction (Jc)     | 600 ppm|
| Young’s modulus                       | 120 GPa|
| Piezoelectric coefficient (d33)        | 1.5 × 10^-9 C/N |
| Piezoelectric layer thickness          | 1 μm   |

B. Operation of the Multiplier

To understand how the multiplier works, refer to Fig. 1(c) and note that \( V_{in1} = V_G - \delta \). Now, if \( V_G \) is within the linear region in Fig. 2(b), then \( G_{s-MTJ} = G_{AP} + \kappa (V_G - \delta) = G_{AP} + \kappa V_{in1} \). Also, note that the voltage dropped over the series resistor \( R \) is

\[
V_{out} = I_{out} R = \frac{R}{R + R_{s-MTJ}} V_{in2} \approx \frac{R}{R_{s-MTJ}} V_{in2} \approx R G_{s-MTJ} V_{in2} \quad \text{if} \quad R \ll R_{s-MTJ}.
\]

Replacing \( G_{s-MTJ} \) in (1) with \( G_{AP} + \kappa V_{in1} \), we get

\[
V_{out} = R G_{AP} V_{in2} + R \kappa (V_{in1} \times V_{in2}) \approx R \kappa (V_{in1} \times V_{in2}) \quad \text{and} \quad I_{out} \approx \kappa (V_{in1} \times V_{in2}) \quad \text{since} \quad R \ll R_{AP}.
\]

This implements a “multiplier” since the current \( I_{out} \) flowing through the s-MTJ (which is also the current through the series resistor \( R \)) is proportional to the product of the two input voltages \( V_{in1} \) and \( V_{in2} \). The voltage \( V_{out} \) is proportional to this current, and hence, it too is proportional to the product \( V_{in1} \times V_{in2} \). Similar ideas were used to design probability composer circuits for Bayesian inference engines in the past [16]. In our case, \( V_{in1} \) and \( V_{in2} \) are voltage “pulses” of fixed width and varying amplitude. Their amplitudes are proportional to the two matrix elements (multiplier and multiplicand) to be multiplied.

Note from Fig. 2(b) that the linear region in the plot extends over a voltage range of ∼100 mV. Therefore, for this choice of parameters, the amplitude of the \( V_{in1} \) pulse should be no more than ∼50 mV. Since we would like the two voltage pulses \( V_{in1} \) and \( V_{in2} \) to have similar limits on the amplitude, both should have an amplitude of no more than 50 mV. We can, of course, increase the voltage range by redesigning with different parameters, but this will increase the energy dissipation per MAC operation.

III. ACCUMULATOR

Next, imagine that the resistor \( R \) of Fig. 1(c) is a heavy metal (HM) strip, on top of which we place a p-MTJ (which is an MTJ whose ferromagnetic layers have perpendicular magnetic anisotropy) with the soft layer in contact with the HM strip. We can insert a thin insulating layer and a thin metallic strip. We can insert a thin insulating layer and a thin metallic strip and because of spin-orbit interaction in that strip, the distance a DW moves over the duration of a pulse is approximately proportional to the amplitude of the pulse, and we show this from micromagnetic simulations in the Appendix. The arrangement is shown in Fig. 3(b).

After any number of pulses, a fraction of the soft layer will have its magnetization parallel to that of the hard layer, and a small fraction will be unmagnetized and will be the “DW” separating two domains; the remainder of the soft layer will have its magnetization antiparallel to that of the hard layer.

![Figure 2](image-url)

**Fig. 2.** Plots of (a) steady-state value of the angle \( \theta \) between the magnetizations of the hard and soft layers of the MTJ as a function of the gate voltage \( V_G \) obtained from the stochastic Landau–Lifshitz–Gilbert simulation at room temperature (300 K). Because of thermal noise, which introduces randomness in the magnetization trajectory, this curve was obtained by averaging over 100 trajectories. (b) \( 1/R_{MTJ} \) versus \( V_G \) characteristic showing that there is a region (shaded in the figure) where the relation \( G_{s-MTJ} = G_{AP} + \kappa (V_G - \delta) \) holds approximately. For this plot, we assumed that \( R_P = 1 \) kΩ and \( R_{AP} = 2 \) kΩ. The voltage \( \delta \) and the constant \( \kappa \) obtained by fitting a straight line to this plot are shown in the figure. We get \( \kappa = -0.4 \pm 0.045 \) (kΩ·V)^{-1} and \( \delta = -0.26 \pm 0.013 \) V. The various material parameters used to obtain these plots are given in Table I.

By applying an external magnetic field aligned with the dipole coupling field. The piezoelectric film is assumed to be (001) PMN-PT, which has a large piezoelectric coefficient. The plot in Fig. 2(b) shows that there is indeed a region of \( V_G \) where the MTJ conductance varies linearly with gate voltage and obeys the relation given above. When the gate voltage \( V_G \) is chosen to be in that region, one can perform an analog multiplication of two input voltages \( V_{in1} \) and \( V_{in2} \) encoding the two matrix elements that are to be multiplied. We show this in the next subsection.
due to the flow of current through the HM strip making up the resistor with successive current pulses. This is the well-known basis of a DW synapse [12]. Here, we have used a p-MTJ in place of a non-volatile memory cell. The conductance of the p-MTJ is the conductance of the parallel combination of three conductors associated with the parallel configuration, DW interface, and parallel configuration.

The conductance of the p-MTJ (measured between its hard and soft layers) is the conductance of the parallel combination of three conductors associated with the parallel configuration, DW interface, and parallel configuration.

A. Operation of the Accumulator

To understand how the accumulator works, consider the fact that the amplitudes of the voltage pulses \( V_{in1} \) and \( V_{in2} \) are proportional to the two matrix elements \( a \) and \( b \) that are to be multiplied. The pulses all have a fixed width of \( \Delta t \). The current \( I_{out} \approx -k(V_{in1} \times V_{in2}) \) is a current pulse of amplitude proportional to \( a \times b \) and has a width \( \Delta t \). The factor \( a_i \) is encoded in the amplitude of the \( i \)th pulse of \( V_{in1} \) and \( b_i \) is encoded in the amplitude of the \( i \)th pulse of \( V_{in2} \). The \( i \)th current pulse flowing through the HM strip, therefore, has an amplitude \( (I_{out})_i \propto a_i \times b_i \).

The \( i \)th current pulse will move the DW by an amount \( \Delta x_i \) where

\[
\Delta x_i = v_i \Delta t
\]

and \( v_i \) is the DW velocity imparted by the \( i \)th current pulse. The DW velocity is proportional to the current density for low densities [18]. Consequently, the DW displacement \( \Delta x \) will be proportional to the amplitude of the current pulse since \( \Delta t \) is fixed. We show this to be approximately true based on simulations (see Appendix A3). Therefore, from (4), we get

\[
\Delta x_i \propto (I_{out})_i \propto a_i \times b_i.
\]

The last equation is an important result showing that the amount by which the DW moves after each pulse is proportional to the product of the two numbers \( a \) and \( b \). Since \( x = \sum_i \Delta x_i \), we get from (3) and (5)

\[
G_{p-MTJ} = x \frac{L}{G_{DW} + G_P \left( 1 - \frac{\omega}{L} \right)} - \frac{(G_P - G_{AP})}{L} \sum_i \Delta x_i
\]

\[
= A - B \sum_i \Delta x_i = A - B \sum_k a_k \times b_k
\]

\[
= A - B \sum_m a_{im} \times b_{mj} = A - B c_{ij}
\]

where \( a_{im} \) is the \((i,m)\)th element of matrix \( [a] \), \( b_{mj} \) is the \((m,j)\)th element of matrix \([b]\), and \( c_{ij} \) is the \((i,j)\)th element of the product matrix \([c] = [a] \times [b]\). The quantities \( A \) and \( B \) are constants. Finally, from (6), we obtain

\[
c_{ij} = \frac{A - G_{p-MTJ}}{B}.
\]

Fig. 4 shows the composite system that constitutes the all-spin matrix multiplier. In addition to the multiplier shown in Fig. 1(c) and the accumulator shown in Fig. 3(a), we use a voltage source \( V_s \) proportional to \( 1/B \), a conductor whose conductance is equal to \( A \), and another conductor whose conductance is \( G_0 \), where \( G_0 \gg A \). The current flowing through the last conductor is

\[
I_{G_0} \approx \frac{-V_s}{1/G_{p-MTJ} + 1/G_0} + \frac{V_s}{1/A + 1/G_0} \propto \frac{V_s(A - G_{p-MTJ})}{A - G_{p-MTJ}} \propto c_{ij}
\]
which is proportional to the \((i, j)\)th element of the product matrix. The voltage dropped over the last conductor is proportional to this current and hence proportional to the \((i, j)\)th element of the product matrix \(c_{ij}\). We just have to measure this voltage after the pulse sequence has ended (i.e., one row has been multiplied with one column) to obtain a voltage proportional to \(c_{ij}\), which is the result of multiplying the \(i\)th row of the first matrix with the \(j\)th column of the second. After obtaining \(c_{ij}\), the DW synapse is reset with a magnetic field or a reverse current pulse to make \(x = 0\), and then, the process is repeated to obtain the product of multiplying another row of the first matrix with another column of the second (which would be the next element of the product matrix).

### B. Energy Dissipation

The energy dissipation incurred during the rotation of a nanomagnet’s magnetization due to strain is very small—theoretically around 1 aJ at room temperature [15], while the energy dissipation associated with DW motion will be on the order of \(I^2 R \Delta t\), where \(I\) is the current inducing the DW motion, \(R\) is the resistance of the HM strip, and \(\Delta t\) is the pulsewidth. There is some additional dissipation in the passive resistors, but they can be made arbitrarily small by choosing the bias voltages to be small. We will neglect any other dissipation due to DW viscosity, which would be comparatively smaller. Therefore, the energy dissipated during each MAC operation is \(\sim I^2 R \Delta t\). We will assume that the HM strip has a width of 50 nm and thickness of 5 nm (cross-sectional area \(= 250 \text{ nm}^2\)) and length 1060 nm. Hence, its resistance is \(R = 424 \Omega\), if it is made of \(\beta\)-Ta whose resistivity is \(\sim 10^{-7} \Omega\cdot\text{m}\). From Fig. 1(c), we see that the current through the HM strip will have a maximum value of \(I_{\text{max}} = \frac{V_{\text{in2}}(\text{max})}{(R + R_P)} \approx V_{\text{in2}}(\text{max})/R_P\), which will have a maximum value of \(\sim 50 \mu\text{A}\) since \(V_{\text{in2}}(\text{max}) \sim 50 \text{ mV}\) and \(R_P\) can be 1 k\(\Omega\). Assuming a pulsewidth \(\Delta t = 0.5\) ns, the maximum energy dissipation per MAC operation is \(I^2 R \Delta t = (50 \mu\text{A})^2 \times 424 \Omega \times 0.5 \times 10^{-9} \text{ s} \sim 530 \text{ aJ}\). This is a small energy price to pay for the small footprint and the nonvolatility of this device.

### IV. Conclusion

We have shown how to implement a matrix multiplier with two MTJs, passive resistors and some bias sources. The energy dissipation per MAC operation is much smaller than what would be encountered in traditional electronic implementations, although not as small as in optical implementations [10]. Our matrix multiplier is also not as fast as optical implementations or even electronic implementations, but it is nonvolatile and will retain the result of the operation (i.e., the matrix element \(c_{ij}\) indefinitely after powering off. The nonvolatility is a major advantage since it will allow most or all computing to be performed at the edge without the need to access the cloud. This reduces the likelihood of hacking, data loss, intrusion, and eavesdropping. Cybersecurity is critical for artificial intelligence and the ability to perform all or most computing at the edge, with a nonvolatile hardware accelerator, offers increased protection against cyber threats.

The extremely low energy dissipation per MAC operation (\(\sim 500 \text{ aJ}\)) also offers protection against hardware Trojans, which are disastrous for AI and are very hard to detect. Trojans, however surreptitious, must consume some energy and hence can be detected with a technique called side-channel monitoring [20], which searches for anomalies in power consumption. A low-power matrix multiplier, which consumes very little power itself, will exacerbate power anomalies due to Trojans and facilitate Trojan detection.

### Appendix

A.1: We consider the elliptical soft layer of an s-MTJ, as shown in Fig. 5. This figure shows the axis designation with the \(z\)-axis along the major (easy) axis of the soft layer and the

---

**Fig. 4.** Composite matrix multiplier. \(V_s\) is a battery whose voltage output is inversely proportional to \(B\) and \(G_0\) is a conductor whose conductance is much larger than \(A\) and the conductance of the p-MTJ. The current flowing through the conductor after multiplication of one row with one column is complete is the corresponding element of the product matrix.
y-axis along the minor (hard) axis. We will assume that the hard layer’s magnetization is along its own easy axis and is pointing along the +z-direction. In that case, the polar angle \( \theta \) shown in Fig. 5 is the angle between the magnetizations of the hard and soft layers of the s-MTJ.

Roy et al. [15] showed how the stochastic Landau–Lifshitz–Gilbert equation yields the temporal evolution of the polar and azimuthal angles of the piezoelectric coefficient of the piezoelectric film (PMN-PT), and \( T \) is the thickness of the film.

In the simulation, we turn on \( V_G \) abruptly at time \( t = 0 \), and then, we follow the temporal evolution of the magnetization (and hence the angle \( \theta \)) until steady state is reached. Steady state is defined as the condition when \( \theta \) settles to a value and fluctuates slightly around it due to thermal noise. Because thermal noise can influence the switching trajectory (i.e., the temporal evolution of \( \theta \)) from the very start, the steady-state value is slightly different from run to run, and hence, we average over 1000 runs to find the steady-state value \( \theta_{ss} \).

**A2: Analytical Derivation of the Linear Region in the \( G_{s-MTJ} \) Versus \( V_G \) Characteristic**

Here, we show analytically that in our system, \( 1/R_{s-MTJ} = 1/R_{AP} + (\kappa(V_G - \delta)) \Rightarrow G_{s-MTJ} = G_{AP} + \kappa(V_G - \delta) \) in a specific region of gate voltage and derive what that region is.

The resistance of the s-MTJ as a function of the angle between the hard and soft layer’s magnetizations is given by \( R_{s-MTJ} = R_P + (R_{AP} - R_P)/2[1 - \cos(\theta_{ss})] \), where \( \theta_{ss} \) is the steady-state angle between the magnetizations of the hard and the soft layer at any given stress (or, equivalently, any given \( V_G \)). From [15], we obtain that the magnetostatic energy in the plane of the nanomagnet (i.e., when \( \phi = 90^\circ \)) for any magnetization orientation and at any given stress is

\[
E = \left[ \frac{\mu_0}{2} M_s^2 \Omega (N_{d-yy} - N_{d-zz}) + \frac{3}{2} \lambda_s \sigma \Omega \right] \sin^2 \theta
+ \left[ \frac{\mu_0}{2} M_s^2 \Omega N_{d-yy} - \frac{3}{2} \lambda_s \sigma \Omega + \mu_0 M_s \Omega H_d \cos \theta \right]
\]

where \( \mu_o \) is the permeability of free space, \( M_s \) is the saturation magnetization, \( \lambda_s \) is the saturation magnetostress, \( \sigma \) is the stress, \( N_{d-yy} \) and \( N_{d-zz} \) are the demagnetization factors along the minor and major axis (they depend on the nanomagnet’s dimensions), and \( \Omega \) is the nanomagnet’s volume. The quantity \( H_d \) is the effective magnetic field in the soft layer due to any residual dipole coupling with the hard layer. As mentioned earlier, this field is antiparallel to the magnetization of the hard layer. The strength of this field can be tailored by engineering the material composition of the hard layer, which is usually made of a synthetic antiferromagnet. It can also be adjusted with an external in-plane magnetic field, if needed.

The steady-state value of the angle \( \theta \) is that where the magnetostatic energy is minimized.

Taking the derivative of (A1) with respect to \( \theta \) and setting it equal to zero, we find the angle where the energy is minimum.

It corresponds to the steady-state value \( \theta_{ss} \). We get

\[
\frac{\partial E}{\partial \theta} = \left[ \frac{\mu_0}{2} M_s^2 \Omega (N_{d-yy} - N_{d-zz}) + \frac{3}{2} \lambda_s \sigma \Omega \right] \sin(2\theta) - \mu_0 M_s H_d \sin(\theta)
= \sin^3 \theta \left[ 2 \left( \frac{\mu_0}{2} M_s^2 \Omega (N_{d-yy} - N_{d-zz}) + \frac{3}{2} \lambda_s \sigma \Omega \right) \cos \theta - M_s H_d \right]
= 0.
\]  

(A2)

Here, \( M_s = \mu_0 M_s \). Solving for \( \cos \theta \) from (A2), we get

\[
\cos \theta_{ss} = \frac{M_s H_d}{2\left( \frac{\mu_0}{2} M_s^2 \Omega (N_{d-yy} - N_{d-zz}) + \frac{3}{2} \lambda_s \sigma \Omega \right) M_s H_d}
= \frac{\left( \mu_0 M_s^2 \Omega (N_{d-yy} - N_{d-zz}) + 3 \lambda_s \sigma Y_d M_s \frac{V_G}{\Omega} \right)}{3 \lambda_s \sigma Y_d \Omega / T}
= \frac{\Gamma}{V_G - \gamma}
\]

(A3)

where \( \Gamma = (M_s H_d T / (3 \lambda_s \sigma Y_d \Omega)) \) and \( \gamma = (\mu_0 M_s^2 (N_{d-yy} - N_{d-zz}) T / (3 \lambda_s \sigma Y_d \Omega)) \). It is easy to verify that the second derivative \( (\partial^2 E / \partial \theta^2) \) is positive, and hence, this is indeed a minimum of the energy, as opposed to a maximum.

Since a real solution of \( \theta_{ss} \) is possible only if \( |\cos \theta_{ss}| \leq 1 \), it is obvious that \( |V_G - \gamma| \geq \Gamma \). Using the values in Table 1, we obtain from the above expressions that \( \Gamma = 0.26 \) V and \( \gamma = -0.001 \) V. Hence, a steady-state solution for the angle between the magnetizations of the hard and soft layers (when they are not collinear) can be obtained only if \( |V_G + 0.001 \text{ V}| \geq 0.26 \) V and that is what we observe in Fig. 2(b) where the MTJ resistance begins to change only when \( V_G \leq -0.261 \) V.

Now, from (A5)

\[
R_{s-MTJ} = R_P + \frac{R_{AP} - R_P}{2}[1 - \cos(\theta_{ss})]
= \frac{R_P + R_{AP}}{2} - \frac{R_{AP} - R_P}{2} \frac{\Gamma}{V_G - \gamma}
\]

(A4)

and therefore,

\[
1 \quad R_{s-MTJ} = \frac{R_P}{2} \left( 1 - \frac{\Gamma}{V_G - \gamma} \right) + \frac{R_{AP}}{2} \left( 1 + \frac{\Gamma}{V_G - \gamma} \right).
\]

(A5)
When $V_G - \gamma$ is close to $-\Gamma$, we can write $\Gamma/(V_G - \gamma) = -1 + \epsilon$, where $|\epsilon| \ll 1$. Hence, from (A5), we obtain

$$\frac{1}{R_{s,MTJ}} = \frac{1}{R_{AP}(1 - \epsilon/2) + R_P e/2} \approx \frac{1}{R_{AP}}(1 + \epsilon/2)$$

$$= \frac{1}{R_{AP}} + \frac{1}{2R_{AP}} \left(1 + \frac{\Gamma}{V_G - \gamma}\right)$$

$$\approx \frac{1}{R_{AP}} - \frac{1}{2R_{AP}} \left(\frac{V_G - \gamma + \Gamma}{\Gamma}\right) [\text{since } V_G - \gamma \approx -\Gamma]$$

$$= \frac{1}{R_{AP}} - \frac{1}{2R_{AP}} \left(\frac{V_G}{\gamma - \Gamma}\right). \quad (A6)$$

Equation (A8) has the form $1/R_{s,MTJ} = 1/R_{AP} + \kappa(V_G - \delta)$ or $G_{s,MTJ} = G_{AP} + \kappa(V_G - \delta)$, where $\kappa = (1/2R_{AP}\Gamma)$ and $\delta = \gamma - \Gamma$. Thus, we have derived the existence of the linear region in the $G_{s,MTJ}$ versus $V_G$ characteristic analytically and found that it exists when $V_G - \gamma$ is close to $-\Gamma$.

Since $\Gamma = 0.26$ V and $\gamma = -0.001$ V, while $R_{AP} = 2$ kΩ in Fig. 2, we find that $\delta = -0.96$ (kΩ V)$^{-1}$ and $\delta = -0.261$ V. This value of $\delta$ shows excellent agreement with what we obtained in Fig. 2(b), but $\kappa$ is larger in magnitude by more than a factor of 2, which is still acceptable within the limits of the approximations used to derive this analytical result.

### A. Maximum Current Pulse Amplitude

The maximum current that flows through the HM strip was calculated as 50 μA. The corresponding current density through a 250-nm$^2$ cross section is $2 \times 10^{11}$ A/m$^2$. In Appendix A3, we show from room temperature micromagnetic simulations that the DW displacement at this current density is about 120 nm if we inject the current pulse for 0.5 ns and then allow a rest period of 4.5 ns for the DW to stabilize. Since the soft layer of the p-MTJ is 1060 nm long, it can sustain at least 1060/120 = 9 current pulses before the DW moves completely through it. Hence, the largest matrix size that can be handled is $9 \times 9$ with this construct. If we need to handle larger matrices, we have to make the length of the soft layer and HM strip larger, leading to larger HM resistance and correspondingly larger energy dissipation and larger device footprint.

### B. Nonbinary Multiplier

The construct described here is a nonbinary multiplier (meaning that its elements can have integral values that are not just 0 and 1). We will, of course, need to know the largest integer we can have as a matrix element. This depends on how small we can make the quantization step size when we digitize the input voltage pulses $V_{in1}$ and $V_{in2}$ representing the multiplier and multiplicand. The minimum step size is, say, twice the thermal noise voltage appearing at any input terminal, that is, $2(kT/C_{in})^{1/2}$, where $C_{in}$ is the input terminal capacitance [21]. We can reasonably assume that $C_{in} \sim 1$ fF when we factor in line capacitances. This makes the minimum step size $\sim 4$ mV at room temperature. Hence, the largest integer that we can encode is $50/4$ mV = 12. We can, of course, increase this number by using an optimized design where the amplitude of the voltage pulses can exceed 50 mV. This would require decreasing $\kappa$. Here, however, we were interested in demonstrating just the basic principle and hence have not focused on design optimization. Increasing the pulse amplitude will obviously lead to more energy dissipation as well.

We can also calculate the current through the HM strip at the minimum step size of 4 mV, which corresponds to the integer 1. This is $\sim 8$ μA. The corresponding current density is $8 \mu A/250 \text{ nm}^2 = 3.2 \times 10^{10}$ A/m$^2$, which is more than enough to induce DW motion in many materials via spin-orbit torque [22]. In fact, the results in the next subsection (Appendix A3) show that the DW displacement at this current density is about 10 nm. Hence, the smallest integer that we can have as a matrix element is 1 since the current pulse corresponding to this digit can induce sufficient DW motion. Thus, for this design, our integer range is 1–12.

#### A3: Room Temperature Micromagnetic Simulations of DW Motion in the Soft Layer of the Accumulator MTJ

It is well known that at room temperature, the DW motion is stochastic. After the current pulse inducing the DW motion subsides, the wall does not immediately stabilize but can move forward and backward—a phenomenon sometimes referred to as DW creep. It is very damaging for a DW synapse since it will hinder the DW displacement from being proportional to the current amplitude, which is critical to implementing the accumulator.

The solution is to make the edges of the soft layer grooved or notched, as shown in the insets of Fig. 6 [23], [24]. They stabilize the DW, mitigate the effect of edge roughness in the soft layer that can trap DWs [25], and prevent creep, but to ensure that the DW displacement is linearly proportional to the current amplitude (which is what we need) the pitch, depth, and width of the groove that will have to be chosen carefully. For this purpose, we carried out micromagnetic (MuMax3) simulations of domain well motion in the p-MTJ soft layer of dimensions $1060 \times 50 \times 1.5$ nm and assumed a spin Hall angle of 0.2, which is reasonable when the HM is β-Ta. The soft layer of the p-MTJ is assumed to be made of CoFeB. The notch dimensions and spacing are shown in the left inset of Fig. 6.

A current pulse was injected for 0.5 ns followed by a rest period of 4.5 ns within which the DW position stabilized. The simulations were carried out in the presence of random thermal noise at 300 K and the mean displacements and standard deviation (error bars) in the displacements of the DW are shown in Fig. 6 as a function of the current density injected into the HM strip. The mean and standard deviation were obtained from 100 runs of the MuMax3 simulations.

The best fit straight line is shown in this plot and the points representing the mean displacements do not stray too far from this line, showing that for this choice of groove parameters, the DW displacement is approximately proportional to the current density and, hence, the current amplitude. This is what is needed to implement the accumulator.

An interesting observation is that the standard deviation in the displacements is rather large and the question naturally arises if this is a consequence of the grooved structure or thermal noise. We have examined many different groove geometries and parameters. In all cases, we saw large...
standard deviations, and hence, it is likely that choosing a different groove geometry or pattern will not reduce the standard deviation significantly. It appears that the primary culprit is thermal noise, which introduces this large standard deviation.

**REFERENCES**

[1] Al and Compute. Accessed: May 16, 2018. [Online]. Available: https://openai.com/blog/ai-and-compute/

[2] R. Szeliski, *Computer Vision: Algorithms and Applications*. London, U.K.: Springer, 2010.

[3] J. Zhang, Z. Wang, and N. Verma, “A matrix-multiplying ADC implementing a machine-learning classifier directly with data conversion,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC)* Dig. Tech. Papers, Feb. 2015, pp. 332–333.

[4] E. H. Lee and S. S. Wong, “Analysis and design of a passive switched-capacitor matrix multiplier for approximate computing,” *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 261–271, Jan. 2017.

[5] R. Yazdani, A. Segura, J.-M. Arnaud, and A. Gonzalez, “An ultra low-power hardware accelerator for automatic speech recognition,” in *Proc. 49th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Taiwan, Oct. 2016, pp. 1–12.

[6] N. Verma, A. Shoeb, J. V. Guttag, and A. P. Chandrakasan, “A micro-power EEG acquisition SoC with integrated seizure detection processor for continuous patient monitoring,” in *Proc. IEEE Symp. VLSI Circuits*, Kyoto, Japan, Jun. 2009, pp. 62–63.

[7] K. H. Lee and N. Verma, “A low-power processor with configurable embedded machine-learning accelerators for high-order and adaptive analysis of medical-sensor signals,” *IEEE J. Solid-State Circuits*, vol. 48, no. 7, pp. 1625–1637, Jul. 2013, doi: 10.1109/JSSC.2013.2253226.

[8] B. Randell, “Ludgate’s analytical machine of 1909,” *Comput. J.*, vol. 14, no. 3, pp. 317–326, Mar. 1971, doi: 10.1093/comjnl/14.3.317.

[9] S. Hong, H. Kang, J. Kim, and K. Cho, “Low voltage time-based matrix multiplier-and-accumulator for neural computing system,” *Electronics*, vol. 9, no. 12, p. 2138, Dec. 2020, doi: 10.3390/electronics9122138.

[10] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D.Englund, “Large-scale optical neural networks based on photoelectric multiplication,” *Phys. Rev. X*, vol. 9, no. 2, May 2019, Art. no. 021032, doi: 10.1103/PhysRevX.9.021032.

[11] R. Hamerly, “The future of deep learning is photonic: Reducing the energy needs of neural networks might require computing with light,” *IEEE Spectr.*, vol. 58, no. 7, pp. 30–47, Jul. 2021, doi: 10.1109/MSPEC.2021.9475393.

[12] 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.*, vol. 10, no. 6, pp. 1152–1160, Dec. 2016, doi: 10.1109/TBCAS.2016.2525823.

[13] J. L. Hockel, P. K. Nordean, D. M. Pisani, C.-Y. Liang, G. P. Carman, and C. S. Lynch, “A method to control magnetism in individual strain-mediated magnetoelectric islands,” *Appl. Phys. Lett.*, vol. 103, no. 23, Dec. 2013, Art. no. 232905, doi: 10.1063/1.4838216.

[14] Z. Y. Zhao et al., “Giant voltage manipulation of MgO-based magnetic tunnel junctions via localized anisotropic strain: A potential pathway to ultra-energy-efficient memory technology,” *Appl. Phys. Lett.*, vol. 109, no. 9, 2016, Art. no. 092403, doi: 10.1063/1.4961670.

[15] K. Roy, S. Bandyopadhyay, and J. Atulasimha, “Energy dissipation and switching delay in stress-induced switching of multiferroic nanomagnets in the presence of thermal fluctuations,” *J. Appl. Phys.*, vol. 112, no. 2, Jul. 2012, Art. no. 023914, doi: 10.1063/1.4737792.

[16] S. Khasanvis et al., “Self-similar magneto-electric nanocircuit technology for probabilistic inference engines,” *IEEE Trans. Nanotechnol.*, vol. 14, no. 6, pp. 980–991, Nov. 2015, doi: 10.1109/TNANO.2015.2439618.

[17] S. Emori, U. Bauer, S.-M. Ahn, E. Martinez, and G. S. D. Beach, “Current-driven dynamics of chiral ferromagnetic domain walls,” *Nature Mater.*, vol. 12, pp. 611–616, Jun. 2013, doi: 10.1038/NMAT3675.

[18] K.-S. Ryu, L. Thomas, S.-H. Yang, and S. S. P. Parkin, “Chiral spin torque at magnetic domain walls,” *Nature Nanotechnol.*, vol. 8, no. 7, pp. 527–533, Jun. 2013, doi: 10.1038/nnano.2013.102.

[19] S. Emori et al., “Spin Hall torque magnetometry of Dzyaloshinskii domain walls,” *Phys. Rev. B, Condens. Matter*, vol. 90, Nov. 2014, Art. no. 184427, doi: 10.1103/PhysRevB.90.184427.

[20] B. Randell, “Ludgate’s analytical machine of 1909,” *Comput. J.*, vol. 14, no. 3, pp. 317–326, Mar. 1971, doi: 10.1093/comjnl/14.3.317.