A Non-Volatile All-Spin Analog Matrix Multiplier: An Efficient Hardware Accelerator for Machine Learning

Rahnuma Rahman and Supriyo Bandyopadhyay, *Fellow, IEEE*

Abstract-- We propose and analyze a compact and *non-volatile* nanomagnetic (all-spin) analog matrix multiplier performing the multiply-and-accumulate (MAC) operation using two magnetic tunnel junctions – one activated by strain to act as the multiplier, and the other activated by spin-orbit torque pulses to act as a domain wall synapse that performs the operation of the accumulator. Each MAC operation can be performed in ~ 1 ns and the maximum energy dissipated per operation is ~ 100 aJ. This provides a very useful hardware accelerator for machine learning (e.g. training of deep neural networks), solving combinatorial optimization problems with Ising type machines, and other artificial intelligence tasks which often involve the multiplication of large matrices. The non-volatility 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—Matrix multiplication, magnetic tunnel junction, domain wall synapse, 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, etc.). Estimates by OpenAI predict an explosive growth of computational requirements in AI by a factor of $100\times$ every two years, which is a $50\times$ 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

(e.g. min-cut or max-cut problems), etc. 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 $x_i^{m+1} = f\left(\sum_j M_{ij}^m x_j^m\right)$, where f is a non-linear

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 paper, 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 *non-volatile* 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 magnetic tunnel junction (driven by spin-orbit torque) acting as a domain wall synapse [12]. Each MTJ can have a footprint of $\sim (100 \text{ nm})^2$, and with all the

This work was supported by the US National Science Foundation under grants CCF-2001255 and CCF-2006843.

The authors are with the Department of Electrical and Computer Engineering, Virginia Commonwealth University, Richmond, VA 23284, USA (email: rahmanr3@vcu.edu, sbandy@vcu.edu)

peripherals, the footprint of the entire device can be $< 1 \mu\text{m}^2$. The matrix multiplier can operate at clock rates of $\sim\text{GHz}$ and dissipate ~ 100 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” layer and a “soft” layer, separated by an intervening insulating spacer layer. Any residual dipole interaction between the hard and the soft layer creates an effective magnetic field H_d in the soft layer that is directed along the latter’s major axis (easy axis) in a direction opposite to the magnetization of the hard layer. The soft layer is magnetostrictive and placed in elastic contact with an underlying *poled* piezoelectric thin film deposited on a conducting substrate (this construct constitutes a 2-phase

axis (or the magnetization of the hard layer). The value of θ_{ss} depends on the applied strain and H_d . The hard layer’s magnetization remains unaffected. Since the resistance of the s-MTJ depends on the angle θ_{ss} between the magnetizations of the hard and the soft layers, the strain changes the MTJ resistance since the value of H_d is fixed. This is the operational basis of a “straintronic” MTJ (s-MTJ), whose basic function was 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.

Fig. 1: (a) A straintronic magnetic tunnel junction (s-MTJ) configured to produce a hyperbolic region in the transfer characteristic R_{MTJ} (magnetic tunnel junction resistance) versus V_G (gate voltage). (b) The transfer characteristic showing the hyperbolic region. (c) An analog multiplier implemented with a single s-MTJ. The two operands are encoded in V_{in1} and V_{in2} and the product of them is encoded in V_{out} or I_{out} . The s-MTJ is biased in the hyperbolic region of the transfer characteristic where the s-MTJ resistance is inversely proportional to $(V_G - \delta)$ with δ being a bias voltage.

multiferroic). Two electrically shorted electrodes, delineated on the piezoelectric film, flank the MTJ, while the back of the substrate is connected to 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) towards the minor axis (hard axis) because of the Villari effect. The rotation is arrested midway by the magnetic field H_d and hence the magnetization ultimately settles into a steady-state orientation that subtends some angle θ_{ss} with the major

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 θ_{ss} versus V_G relation. The s-MTJ resistance is given by

$R_{s-MTJ} = R_P + \frac{R_{AP} - R_P}{2} [1 - \cos \theta_{ss}]$, 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 θ_{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 proper choice of the s-MTJ parameters, we can produce a *linear* region in the G_{s-MTJ} vs. V_G characteristic where $1/R_{s-MTJ} = 1/R_{AP} + \kappa/(V_G - \delta) \Rightarrow G_{s-MTJ} = G_{AP} + \kappa/(V_G - \delta)$ [κ and δ are constants]. We show this analytically in the Appendix. In Fig. 2, we plot the θ_{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 ref. [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 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 $1/R_{s-MTJ} = 1/R_{AP} + \kappa/(V_G - \delta) \Rightarrow G_{s-MTJ} = G_{AP} + \kappa/(V_G - \delta)$. 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 elucidate this in the next subsection.

Fig. 2: Plots of (a) the steady-state value of the angle θ 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) The $1/R_{MTJ}$ 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 $R_P = 1$ k Ω and $R_{AP} = 2$ k Ω . The voltage δ and the constant κ obtained by fitting a straight line to this plot are shown in the figure. We get $\kappa = -0.4 \pm 0.045$ (k Ω -V) $^{-1}$ and $\delta = -0.26 \pm 0.013$ V. The various material parameters used to obtain these plots are given in Table I.

Table I: Parameters for the soft layer of the MTJ

Major axis dimension (L)	800 nm
Minor axis dimension (W)	700 nm
Thickness (d)	2.2 nm
Saturation magnetization (M_s)	8.5×10^5 A/m
Dipole coupling field (H_d)	1000 Oe
Gilbert damping constant (α)	0.1
Saturation magnetization (λ_s)	600 ppm
Young's modulus	120 GPa
Piezoelectric coefficient (d33)	1.5×10^{-9} C/N
Piezoelectric layer thickness	1 μ m

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

$$V_{out} = I_{out} R = \frac{R}{R + R_{s-MTJ}} V_{in2} \approx \frac{R}{R_{s-MTJ}} V_{in2} \approx R G_{s-MTJ} V_{in2} \quad (1)$$

if $R \ll R_{s-MTJ}$

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

$$V_{out} = R G_{AP} V_{in2} + R \kappa (V_{in1} \times V_{in2}) \approx R \kappa (V_{in1} \times V_{in2}) \quad \text{and} \quad (2)$$

$$I_{out} \approx \kappa (V_{in1} \times V_{in2}) \quad \text{since } R \ll R_{AP}$$

That 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 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 no more than 50 mV. We can, of course, increase the voltage range by redesigning with different parameters, but that 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 layer between the soft layer and the heavy metal, which will not

impede the operation of the accumulator. This configuration is shown in Fig. 3(a). The current pulses I_{out} pass through the heavy metal strip (which is the resistor R) and because of spin-orbit interaction in that strip, they inject spins into the soft layer of the p-MTJ (through the thin insulating and metallic layers). That causes domain wall motion in the latter during each pulse because of spin orbit torque due to the spin Hall effect [17-19]. The distance a domain wall moves over the duration of a pulse is approximately proportional to the amplitude of the pulse since the domain wall velocity is proportional to the current density. 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, a small fraction will be un-magnetized and will be the “domain wall” separating two domains, and the remainder of the soft layer will have its magnetization antiparallel to that of the hard layer. The fractions with parallel and anti-parallel magnetizations change with successive current pulses flowing through the heavy metal.

Fig. 3: (a) Schematic of the “accumulator” consisting of a p-MTJ integrated with a heavy metal strip. (b) Domain wall motion in the p-MTJ soft layer due to the flow of current through the heavy metal strip making up the resistor R in Fig. 1. (c) The conductance of the p-MTJ is the conductance of the parallel combination of three conductors associated with the anti-parallel configuration, domain wall interface, and parallel configuration.

This is the well-known basis of a domain wall synapse [12]. Here, we have used a p-MTJ in the spirit of ref. [12], but there is no reason why an MTJ with in-plane magnetic anisotropy cannot be used instead.

The conductance of the p-MTJ (measured between its hard and soft layers) is the conductance of the parallel combination of *three* conductors corresponding to the parallel configuration of the p-MTJ, the domain wall (DW) interface and the antiparallel configuration [12], as shown in Fig. 3(c). If the domain wall in the soft layer of the p-MTJ is located at a distance x from one edge and L is the length of the soft layer (excluding the domain wall width, which is w), then [12]

$$G_{p\text{-MTJ}}(x) = \underbrace{\frac{w}{L}G_{DW} + G_P}_{\text{constant } A} \left(1 - \frac{w}{L}\right) - \underbrace{\frac{(G_P - G_{AP})}{L}}_{\text{constant } B} x, \quad (3)$$

where G_P is the p-MTJ conductance in the parallel state, G_{AP} is the conductance in the antiparallel state and G_{DW} is the conductance associated with the domain wall in the soft layer.

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 Δt . The current $I_{out} \approx -\kappa(V_{in1} \times V_{in2})$ is a current pulse of amplitude proportional to $a \times b$ and width Δt . 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 therefore has an amplitude $(I_{out})_i \propto a_i \times b_i$.

The i -th current pulse will move the domain wall by an amount Δx_i where

$$\Delta x_i = v_i \Delta t, \quad (4)$$

and v_i is the domain wall velocity imparted by the i -th current pulse. The domain wall velocity is proportional to current density for low densities [18] and is hence proportional to the amplitude of the current pulse. Therefore, from Equation (4), we get

$$\Delta x_i \propto (I_{out})_i \propto a_i \times b_i. \quad (5)$$

The last equation is an important result showing that the amount by which the domain wall 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 Equations (3) and (5)

$$\begin{aligned} G_{p\text{-MTJ}} &= \frac{w}{L}G_{DW} + G_P \left(1 - \frac{w}{L}\right) - \frac{(G_P - G_{AP})}{L} \sum_i \Delta x_i \\ &= A - B \sum_i \Delta x_i = A - B \sum_k a_k \times b_k, \\ &= A - B \sum_m a_{im} \times b_{mj} = A - B c_{ij} \end{aligned} \quad (6)$$

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 Equation (6), we obtain

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 offers increased protection against cyber threats.

The extremely low energy dissipation also offers protection against hardware Trojans, which are disastrous for AI and are very hard to detect. Trojans, however, have an effect on the power consumption and therefore can be detected with a technique called side channel analysis [20], which searches for anomalies in power consumption. A *low power* matrix multiplier, which consumes very little power itself, will exacerbate power anomalies and facilitate Trojan detection.

V. APPENDIX

A.1: We consider the elliptical soft layer of a straintronic 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 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 θ shown in Fig. 5 is the angle between the magnetizations of the hard and soft layers of the s-MTJ.

Fig. 5: The axes designation used to simulate the magneto-dynamics of the soft layer of the straintronic MTJ employed in the multiplier.

Ref. [15] showed that the stochastic Landau-Lifshitz-Gilbert equation yields the following equations to describe the temporal evolution of the polar and azimuthal angles of the magnetization vector (θ , ϕ) in the soft layer in the presence of thermal noise:

$$\begin{aligned} (1 + \alpha^2) \frac{d\theta(t)}{dt} &= \frac{|\gamma|}{M_v} \left[A(t) \sin \theta(t) - \alpha B(t) \sin(2\theta(t)) + \alpha M_v H_d \sin \theta(t) + (\alpha P_{\theta(t)} + P_{\phi(t)}) \right] \\ (1 + \alpha^2) \frac{d\phi(t)}{dt} &= \frac{|\gamma|}{M_v} \left[\alpha A(t) + 2B(t) \cos \theta(t) - M_v H_d - \frac{P_{\theta(t)} - \alpha P_{\phi(t)}}{\sin \theta(t)} \right] \quad [\sin \theta(t) \neq 0] \end{aligned} \quad (\text{A1})$$

where

$$\begin{aligned} A(t) &= \frac{\mu_0}{2} M_s^2 \Omega (N_{d-xx} - N_{d-yy}) \sin(2\phi(t)) \\ B(t) &= \frac{\mu_0}{2} M_s^2 \Omega (N_{d-xx} \cos^2 \phi(t) + N_{d-yy} \sin^2 \phi(t) - N_{d-zz}) \\ &\quad + \frac{3}{2} \lambda_s \sigma \Omega \\ P_{\theta}(t) &= M_v [h_x(t) \cos \theta(t) \cos \phi(t) + h_y(t) \cos \theta(t) \sin \phi(t)] \\ &\quad - M_v h_z(t) \sin \theta(t) \\ P_{\phi}(t) &= M_v [h_y(t) \cos \phi(t) - h_x(t) \sin \phi(t)] \\ M_v &= \mu_0 M_s \Omega \\ \Omega &= (\pi/4) L_{\text{maj}} \times L_{\text{min}} \times d \\ N_{d-yy} &= \frac{\pi}{4} \left(\frac{d}{L_{\text{maj}}} \right) \left[1 + \frac{5}{4} \left(\frac{L_{\text{maj}} - L_{\text{min}}}{L_{\text{maj}}} \right) + \frac{21}{16} \left(\frac{L_{\text{maj}} - L_{\text{min}}}{L_{\text{maj}}} \right)^2 \right] \\ N_{d-zz} &= \frac{\pi}{4} \left(\frac{d}{L} \right) \left[1 - \frac{1}{4} \left(\frac{L_{\text{maj}} - L_{\text{min}}}{L_{\text{maj}}} \right) - \frac{3}{16} \left(\frac{L_{\text{maj}} - L_{\text{min}}}{L_{\text{maj}}} \right)^2 \right] \\ N_{d-xx} &= 1 - N_{d-yy} - N_{d-zz} \\ h_i(t) &= \sqrt{\frac{2\alpha kT}{|\gamma|(1 + \alpha^2) M_v \tau}} G_{(0,1)}^i(t) \quad (i = x, y, z) \\ \gamma &= 2\mu_B \mu_0 / \hbar = 2.21 \times 10^5 \text{ rad-m/Amp-s} \end{aligned} \quad (\text{A2})$$

Here M_s is the saturation magnetization of the soft layer, σ is the uniaxial stress generated in the soft layer along the major axis by the applied gate voltage V_G (we neglect the strain generated along the minor axis since it is much smaller), λ_s is the saturation magnetization of the soft layer's material, Ω is the soft layer's volume, L_{maj} is the length of the major axis, L_{min} is the length of the minor axis and d is the thickness of the soft layer. The quantity μ_0 is the permeability of free space and μ_B is the Bohr magneton. The field $h_i(t)$ [$i = x, y, z$] is the random magnetic field due to thermal noise and $G_{(0,1)}^i(t)$ are three statistically independent Gaussians of zero mean and unit standard deviation, while τ is the time-step used in the simulation (0.1 ps). The stress σ is related to the gate voltage

V_G according to the relation $\sigma = Y d_{33} \frac{V_G}{T}$, where Y is the

Young's modulus of the soft layer (Terfenol-D), d_{33} is the piezoelectric coefficient of the piezoelectric film (PMN-PT), and T is the thickness of the film.

Equations (A1) and (A2) are used with the parameters in Table I to find the steady state value of θ (i.e. θ_{ss}) as a function of V_G . This is shown in Fig. 2(a). Thermal noise introduces some randomness in the magnetization trajectory, and hence we find θ_{ss} as a function of V_G by averaging over 100 trajectories. This yields the plot in Fig. 2(a) and ultimately the plot in Fig. 2(b).

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

$R_{s-MTJ} = R_P + \frac{R_{AP} - R_P}{2} [1 - \cos \theta_{ss}]$, where θ_{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 ref. [15], we obtain that the magneto-static energy in the plane of the nanomagnet (i. e. when $\phi = 90^\circ$) for any magnetization orientation and at any given stress is

$$E = \left[\frac{\mu_0}{2} M_s^2 \Omega (N_{d-yy} - N_{d-zz}) + \frac{3}{2} \lambda_s \sigma \Omega \right] \sin^2 \theta + \frac{\mu_0}{2} M_s^2 \Omega N_{d-zz} - \frac{3}{2} \lambda_s \sigma \Omega + M_v H_d \cos \theta \quad (A3)$$

where 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 θ is that where the magneto-static energy is minimized.

Taking the derivative of Equation (A3) with respect to θ and setting it equal to zero, we find the angle where the energy is minimum, and it corresponds to the steady state value θ_{ss} . We get

$$\begin{aligned} \frac{\partial E}{\partial \theta} &= \left[\frac{\mu_0}{2} M_s^2 \Omega (N_{d-yy} - N_{d-zz}) + \frac{3}{2} \lambda_s \sigma \Omega \right] \sin(2\theta) - M_v H_d \sin \theta \\ &= \sin \theta \left[2 \left(\frac{\mu_0}{2} M_s^2 \Omega (N_{d-yy} - N_{d-zz}) + \frac{3}{2} \lambda_s \sigma \Omega \right) \cos \theta - M_v H_d \right] \\ &= 0 \end{aligned} \quad (A4)$$

Therefore,

$$\begin{aligned} \cos \theta_{ss} &= \frac{M_v H_d}{2 \left(\frac{\mu_0}{2} M_s^2 \Omega (N_{d-yy} - N_{d-zz}) + \frac{3}{2} \lambda_s \sigma \Omega \right)} \\ &= \frac{M_v H_d}{\left(\mu_0 M_s^2 \Omega (N_{d-yy} - N_{d-zz}) + 3 \lambda_s Y d_{33} \frac{V_G}{T} \right)} \\ &= \frac{M_v H_d}{3 \lambda_s Y d_{33} \Omega / T} \times \frac{1}{V_G + \mu_0 M_s^2 (N_{d-yy} - N_{d-zz}) T / (3 \lambda_s Y d_{33})} \\ &= \frac{\Gamma}{V_G - \gamma} \end{aligned}$$

(A5)

where $\Gamma = \frac{M_v H_d T}{3 \lambda_s Y d_{33} \Omega} = \frac{\mu_0 M_s H_d T}{3 \lambda_s Y d_{33}}$ and

$$\gamma = \frac{\mu_0 M_s^2 (N_{d-zz} - N_{d-yy}) T}{3 \lambda_s Y d_{33}}$$

It is easy to verify that the second derivative $\frac{\partial^2 E}{\partial \theta^2}$ is positive and hence this is indeed a minimum of the energy as opposed to a maximum.

Since a real solution of θ_{ss} is possible only if $|\cos \theta_{ss}| \leq 1$, it is obvious that $|V_G - \gamma| \geq \Gamma$. Using the values in Table I, 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 Equation (A5)

$$\begin{aligned} R_{s-MTJ} &= R_P + \frac{R_{AP} - R_P}{2} [1 - \cos \theta_{ss}] \\ &= \frac{R_{AP} + R_P}{2} - \frac{R_{AP} - R_P}{2} \frac{\Gamma}{V_G - \gamma} \end{aligned} \quad (A6)$$

and hence

$$\frac{1}{R_{s-MTJ}} = \frac{1}{\frac{R_{AP}}{2} \left(1 - \frac{\Gamma}{V_G - \gamma} \right) + \frac{R_P}{2} \left(1 + \frac{\Gamma}{V_G - \gamma} \right)} \quad (A7)$$

When $V_G - \gamma$ is close to $-\Gamma$, we can write $\frac{\Gamma}{V_G - \gamma} = -1 + \varepsilon$

where $|\varepsilon| \ll 1$. Hence from Equation (A7) we obtain

$$\begin{aligned} \frac{1}{R_{s-MTJ}} &= \frac{1}{R_{AP} (1 - \varepsilon/2) + R_P \varepsilon/2} \approx \frac{1}{R_{AP}} (1 + \varepsilon/2) \\ &= \frac{1}{R_{AP}} + \frac{1}{2 R_{AP}} \left(1 + \frac{\Gamma}{V_G - \gamma} \right) \\ &\approx \frac{1}{R_{AP}} - \frac{1}{2 R_{AP}} \left(\frac{V_G - \gamma + \Gamma}{\Gamma} \right) \quad [\text{since } V_G - \gamma \approx -\Gamma] \\ &= \frac{1}{R_{AP}} - \frac{1}{2 R_{AP} \Gamma} (V_G - [\gamma - \Gamma]) \end{aligned} \quad (A8)$$

Equation (A8) has the form $1/R_{s-MTJ} = 1/R_{AP} + \kappa(V_G - \delta)$ or

$$G_{s-MTJ} = G_{AP} + \kappa(V_G - \delta) \text{ where } \kappa = -\frac{1}{2 R_{AP} \Gamma} \text{ and } \delta = \gamma - \Gamma$$

. Thus, we have derived the existence of the linear region in the G_{s-MTJ} vs. 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 $R_{AP} = 2$ k Ω , we find that $\kappa = -0.96$ (k Ω -V) $^{-1}$ and $\delta = -0.261$ V. This value of δ shows excellent agreement with what we obtained in Fig. 2(b),

but κ 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 heavy metal strip was calculated as 50 μA . The corresponding current density through a 250 nm² cross-section is 2×10^{11} A/m². If v_d is the domain wall velocity at that current density (which is material dependent), then the maximum domain wall displacement caused by the maximum current pulse of duration 1 ns is $v_d \Delta t$ with $\Delta t = 1$ ns. Since the soft layer of the p-MTJ is 100 nm long, it can sustain $N = 100 \text{ nm}/v_d \Delta t$ current pulses before the domain wall moves completely through it. Hence the largest matrix size that can be handled is $N \times N$.

B. Digital (non-binary) multiplier

If we wish to use this device as a digital, but not just binary, multiplier (meaning its elements can have integral values that are not just 0 and 1), then we need to know what is the largest digit we can have as a matrix element. That depends on how small we can make the quantization step size when we digitize. The minimum step size is, say, twice the thermal noise voltage appearing at any input terminal and that is $2\sqrt{kT/C_{in}}$ where C_{in} is the input terminal capacitance [21]. We can reasonably assume that $C_{in} \sim 1\text{fF}$ when we factor in line capacitances. This makes the minimum step size ~ 4 mV at room temperature. Hence the largest digit that we can encode is $50 \text{ mV}/4 \text{ mV} = 12$. We can, of course, increase this number by using optimized design where the amplitude of the voltage pulses can exceed 50 mV. This would require decreasing κ . Here, however, we were interested in demonstrating just the basic principle and hence have not focused on design optimization.

We can also calculate the current density through the HM strip at the minimum step size of 4 mV. This is $\sim 4 \mu\text{A}$. The corresponding current density is $4 \mu\text{A}/250 \text{ nm}^2 = 1.6 \times 10^{10}$ A/m², which is more than enough to induce domain wall motion in many materials [22]. Hence, the smallest digit is 1 since the current pulse corresponding to this digit can also induce domain wall motion.

REFERENCES

- [1] AI and Compute. <https://openai.com/blog/ai-and-compute/>
- [2] R. Szeliski, *Computer Vision: Algorithms and Applications* (Springer, London, 2010).
- [3] J. Zhang, Z. Wang and N. A. Verma, "A matrix multiplying ADC implementing a machine learning classifier directly with data conversion", *IEEE International Solid State Circ. Conf.*, San Francisco, Digest of Technical Papers, 22-26 February 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*, 52, 261–271 (2016).
- [5] R. Yazdani, A. Segura, J.-M. Arnau and A. Gonzalez, "An ultra low-power hardware accelerator for automatic speech recognition", *Proc. IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Taipei, Taiwan, 15–19 October 2016; pp. 1–12.
- [6] N. Verma, A. Shueb, J. V. Gutttag and A. P. Chandrakasan, "A micro-power EEG acquisition SoC with integrated seizure detection processor for continuous patient monitoring", *Proc. IEEE Symposium on VLSI Circuits*, Kyoto, Japan, 16–18 June 2009; pp. 62–63.
- [7] K. H. Lee and N. A. Verma, "Low-Power Processor with Configurable Embedded Machine-Learning Accelerators for High-Order and Adaptive Analysis of Medical-Sensor Signals", *IEEE J. Solid-State Circuits*, 48, 1625–1637 (2013).
- [8] B. Randell, "Ludgate's analytical machine of 1909", *The Computer Journal*, 14, 317 (1971).
- [9] See, for example, S. Hong, H. Kang, J. Kim and K. Cho, "Low voltage time-based matrix multiplier-and-accumulator for neural computing system", *Electronics (MDPI)*, 9, 2138 (2020).
- [10] R. Hammerly, L. Bernstein, A. Sludds, M. Soljačić and D. Englund, "Large scale optical neural networks based on photoelectric multiplication", *Phys. Rev. X*, 9, 021032 (2019).
- [11] R. Hammerly, "The future of deep learning is photonic: Reducing the energy needs for neural networks might require computing with light", *IEEE Spectrum*, 58, 30 (2021).
- [12] A. Sengupta, Y. Shim and K. Roy, "Proposal for all-spin artificial neural network: Emulating neural and synaptic functionalities through domain wall motion in ferromagnets", *IEEE Trans. Biomed. Circ. Syst.*, 10, 1152 (2016).
- [13] J. Cui, J. L. Hockel, P. K. Nordeen, 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.* 103 232905 (2013).
- [14] Z. Y. Zhao, M. Jamali, N. D'Souza, D. Zhang, S. Bandyopadhyay, J. Atulasimha and J. P. Wang, "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.* 109 092403 (2016).
- [15] K. Roy, S. Bandyopadhyay and J. Atulasimha, "Energy dissipation and switching delay in stress-induced switching of multiferroic devices in the presence of thermal fluctuations", *J. Appl. Phys.* 112 023914 (2012).
- [16] S. Khasanvis, M. Li, M. Rahman, M. Salehi-Fashami, A. K. Biswas, J. Atulasimha, S. Bandyopadhyay and C. A. Moritz, "Self-similar magneto-electric nanocircuit technology for probabilistic inference engines", *IEEE Trans. Nanotechnol.*, 14, 980 (2015).
- [17] S. Emori, U. Bauer, S.-M. Ahn, E. Martinez and G. S. Beach, "Current-driven dynamics of chiral ferromagnetic domain walls", *Nature Mater.*, 12, 611 (2013).
- [18] K.-S. Ryu, L. Thomas, S.-H. Yang and S. S. P. Parkin, "Chiral spin torque at magnetic domain walls", *Nature Nanotechnol.*, 8, 527 (2013).
- [19] S. Emori, E. Martinez, K.-J. Lee, H.-W. Lee, U. Bauer, S.-M. Ahn, P. Agrawal, D. C. Bono and G. S. Beach, "Spin hall torque magnetometry of Dzyaloshinskii domain walls", *Phys. Rev. B*, 90, 184427 (2014).
- [20] S. Bhunia, M. S. Hsiao, M. Banga and S. Narasimhan, "Hardware Trojan attacks: Threat analysis and countermeasures", *Proc. IEEE*, 102, 1229 (2014).
- [21] L. B. Kish, "End of Moore's law: Thermal (noise) death of integration in micro and nano electronics", *Phys. Lett. A*, 305, 144.
- [22] D.-T. Ngo, K. Ikeda and H. Awano, "Direct observation of domain wall motion induced by low current density in TbFeCo wires", *Appl. Phys. Express*, 4, 093002 (2011).