Temporal Memory with Magnetic Racetracks

Hamed Vakili*, Mohammad Nazmus Sakib†, Samiran Ganguly†, Mircea Stan†,
Matthew W. Daniels‡, Advait Madhavan‡§, Mark D. Stiles‡ and Avik W. Ghosh*
*Department of Physics, University of Virginia, Charlottesville, VA 22903 USA
†Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903 USA
‡Alternative Computing Group, National Institute of Standards and Technology, Gaithersburg, MD
§Institute for Research in Physics and Applied Electronics, University of Maryland, College Park, MD

***Abstract*—Race logic is a relative timing code that represents information in a wavefront of digital edges on a set of wires in order to accelerate dynamic programming and machine learning algorithms. Skyrmions, bubbles, and domain walls are mobile magnetic configurations (solitons) with applications for Boolean data storage. We propose to use current-induced displacement of these solitons on magnetic racetracks as a native temporal memory for race logic computing. Locally synchronized racetracks can spatially store relative timings of digital edges and provide non-destructive read-out. The linear kinematics of skyrmion motion, the tunability and low-voltage asynchronous operation of the proposed device, and the elimination of any need for constant skyrmion nucleation and annihilation make these magnetic racetracks a natural memory for low-power, high-throughput race logic applications.**

***Index Terms*—Skyrmions, Domain Walls, Racetrack, Race Logic, Temporal Memory.**

I. Displacement based Magnetic Memories and Arrival Time Codes

When energy efficiency becomes the predominant metric in computing systems, the choice of information representation becomes increasingly important. A recently proposed temporal coding scheme known as race logic [1]–[3] promises orders-of-magnitude energy improvement over classical approaches. In race logic, information is encoded in the relative timing between digital rising edges on different wires, with respect to some temporal reference. This allows each wire to encode multiple bits of information depending on when the rising edge arrives. These rising edges, though encoding multiple bits, incur only a single transition per wire during the course of a computation, keeping activity factors low. This allows conventional Boolean primitives to perform non-traditional operations at a very low energy cost. For example, a single OR gate allows the first arriving edge to pass, effectively performing a two input MIN function,[1] as shown in figure 1(a). The rest of the fundamental race logic primitives including MAX, INHIBIT and ADD-BY-CONSTANT gates are shown in figure 1. Previous work has mathematically proven the universality of these primitives, constrained by the the physics of causality [4], [5]: any temporally invariant and causal function can be implemented with these primitives.

[1]Note that a multi-bit MIN operation performed with a Boolean magnitude comparator would incur a much higher area and power cost than a single two-input OR gate.

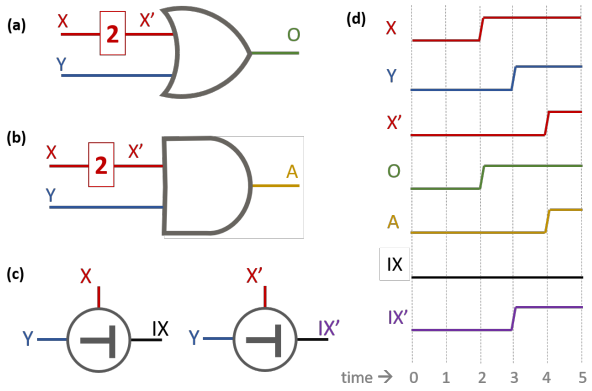

Fig. 1. Information encoding and primitives of race logic: Panel (a) shows how OR gates and delay elements can be used to perform the temporal MIN and addition by a constant operations. Panel (b) shows the same with an AND gate performing the MAX operation. Panel (c) shows temporal inhibit operation where the inhibiting signals (X and X') inhibit the incoming signal (Y) depending on their relative arrival times. Panel (d) shows the information representation through timing diagrams for the signals used in (a-c).

Such an approach to computation differs from the conventional Boolean approach and results in vastly differing architectures. Computations are generally performed by setting up the problem in a spatially arranged network of operators. Digital temporal wavefronts are presented to the inputs of such an array and the way the wavefront navigates through the network performs the computation. Hence, such an approach maps well to dynamic programming problems, where the spatial data structure of the problem (like graphs or trees) can be mapped to supporting spatial architectures. Previous work has demonstrated sizeable improvements in performance and energy efficiency for such constrained problems such as decision trees for machine learning (with 4- to 6-bit precision) [2] or directed-acyclic-graph-based genetic sequencing (4-bit precision) [3].

Though a promising approach, one major impediment in implementing race-logic-based temporal computing systems is the need for a memory that can easily store such temporally coded information. Such storage would enable more complicated processing than can be done with simple logic gates. Previously, counter and current starved inverter based circuits have been proposed as tunable delays, but their energy cost, volatility and limited reconfigurability have caused researchers

to look at novel approaches to store temporal information. Resistive approaches proposed in [6] suffer from non-linearity between the read and write processes, a problem that is avoided in the magnetic counterparts explored here.

Magnetic devices play key roles in data storage from magnetic tapes to hard disk drives to tunnel junction memories. A racetrack memory [7] is similar to a magnetic disk but without physically moving parts. Translating the magnetic configuration along the track plays the role of moving a magnetic tape. Racetracks with skyrmions [8], a particular magnetic configuration discussed in Sec. II can be used to store Boolean information with the presence of a skyrmion at a particular location indicating a one and the absence a zero, for example. This information can be subsequently read by translating the skyrmions past a detector that reads the changing magnetic state through a resistance change. Translation can be achieved by several mechanisms including passing a current through a heavy metal layer underlying the track. This current injects a spin current into the magnetic layer creating a spin-orbit torque [9], [10] that rotates the magnetization. The subsequent local rotations of the magnetization give rise to an effective translation of the magnetization pattern.

In this paper, we present a design (Fig. 2) of a temporal memory cell that linearly converts information from the time domain to a displacement domain. This is done by using current pulses of varying lengths in time to translate skyrmions corresponding distances in space. The linear relationship between pulse length and displacement, thereby encodes the arrival times of the pulses. We present details of this memory, the non-destructive readout, a way to reset memory, and a comparison with conventional Boolean approaches in the next few sections. Section II provides background on the magnetic technology that is used. Section III describes how complementary metal-oxide-semiconductor (CMOS) circuits can be interfaced with such technology to perform read and write operations. Section IV describes our simulation results for ideal skyrmion racetracks followed by a discussion of non-idealities from imperfections in material properties and how they affect the the operation of this memory.

II. Domain Walls and Skyrmions for Memory

Magnetic memories generally store information in the orientations of the microscopic magnetic domains that reside within the magnetic material. Magnetizations tend to have their energy minimized when the moments point in either direction along a preferred axis. Such a binary configuration naturally lends itself to binary information encoding in which one direction corresponds to 0 and the other to 1. This encoding is used in magnetic tapes, hard disk drives, and magnetic random access memory. In continuous media, the magnetization tends to form domains with the magnetization in two neighboring regions roughly uniform and separated by a narrow region where the magnetization rotates from one domain to the other.

When magnetic materials are fabricated in a 2D geometry, energetic considerations typically prefer that the magnetization lie in the plane. However, it is possible to tune the anisotropies of the material such that the magnetization tends to point parallel to the interface normal, which we label $\hat{z}$. We choose such a material in this work. In the uniform configuration, the magnetization points in the same direction everywhere in a sample, either along $\hat{z}$ or $-\hat{z}$. However, other configurations can exist in a metastable state [11]. One such configuration consists of two stable regions (one along $\hat{z}$ and one along $-\hat{z}$) separated by a localized 180° domain wall. Localized domain walls are generated between oppositely oriented magnetic regions through the balance of exchange and anisotropy forces, the latter of which can arise due to intrinsic magnetocrystalline anisotropy or from magnetostatic interactions [12], [13].

This domain wall divides two regions of uniform magnetization. A skyrmion can be imagined by making the domain wall circular so that it surrounds a region of uniform magnetization (inside the circle) from an region with the opposite magnetization (outside the circle). Both domain walls and skyrmions are topologically protected, meaning that there is a significant barrier to eliminate them, giving them the necessary lifetimes for storage [14]. For skyrmions, the domain wall separating the two regions twists in a particular way, which is described by a topological index known as a winding number. Using ultra-thin ferromagnets (FM) on top of heavy metal (HM) layers breaks inversion symmetry and generates an additional energy term, the Dzyaloshinskii-Moriya interaction (DMI) [15], which stabilizes skyrmions.

Both domain wall and skyrmions have been proposed as fundamental elements of future magnetic memories. For domain walls, ones and zeros are encoded in different magnetization directions; for skyrmions, it is their presence or absence which encodes a bit. In the particular class of memory devices called racetracks [7], [8] that we consider here, the sample geometry is a long thin wire. The information stored in the racetrack can be translated along the wire by passing a current through the wire. Here, this translation is facilitated by the inclusion of a heavy metal layer next to the magnetic layer. An electric current driven through the heavy metal layer injects a spin current, through a spin Hall effect, perpendicular to the FM/HM interface into the magnetic layer, thereby applying a spin-orbit torque to the ferromagnetic layer [16]–[19]. This torque can drive skyrmions and domain walls in a variety of ultra-thin magnetic systems at speeds as high as 500 m/s to 1 km/s for relatively low current densities. Possible systems include ferromagnets, antiferromagnets (AFMs), heavy-metal heterostructures, and the two systems of particular interest here: synthetic antiferromagnets [20]–[22], and nearly compensated ferrimagnets. For skyrmions to have sufficiently long lifetimes, their energy barriers should be above 30 kT to 40 kT. Such an energy barrier should be achievable by skyrmions of $\approx 20\ nm$ diameter. It is also possible to use narrower racetracks, smaller than 100 nm, to have a even more efficient device. The parameters we have used, ≈ 50 nm skyrmion diameter and 200 nm racetrack width are to show that the device can be more efficient compared to a CMOS based alternative even without using the state of the art parameters. In our simulations, we consider a racetrack

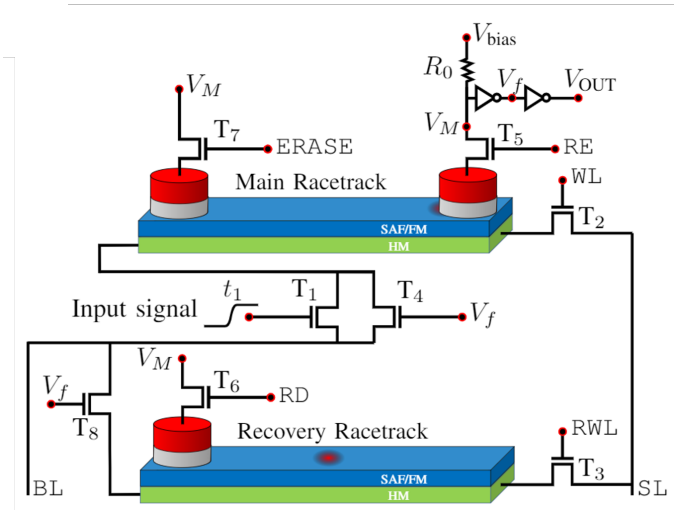

Fig. 2. Concept and circuit illustration: The general arrangement and detailed description of the circuits of a single column of $N = 1$ cell which can be used to store and play temporal wavefronts. The racetracks consist of a bilayer of synthetic antiferromagnet (SAF) or a ferromagnet (FM) with magnetic tunnel junctions (MTJs) to detect when a skyrmion appears under them. Control lines consist of the bit line (BL), source line (SL), write line (WL), read enable (RE), recovery line (RD), recovery write line (RWL), and erase line (ERASE).

that is either a two-sublattice near-compensated ferrimagnet (FiM) like CoGd or a synthetic antiferromagnet, where the low saturation magnetization gives small-sized high-speed skyrmions that propagate along the track without deflection to the side as discussed in Sec. IV.

The dynamics of domain walls and skyrmions under the effect of an electronic current is well studied [23], [24], and can be modeled using classical mechanics. In this work we make use of these classical kinemetic behaviors of domain walls and skyrmions by exploiting the simple linear response $x = vt$ between displacement x and time t, mediated by a constant velocity v. To model its motion in synthetic antiferromagnets, we use a collective coordinate description of a rigid skyrmion, that is, we assume the skyrmion texture has only translational degrees of freedom and derive equations for these from the governing Landau-Lifshitz-Gilbert equations. We assume that the only driving forces on the skyrmions come from spin orbit torques arising from the spin Hall effect (spin Hall angle Θ_{sh}) in the heavy metal layer.

The resulting Thiele equation gives an instantaneous speed for the skyrmion [23], [25]

$$v = \frac{\pi\gamma\hbar}{2e} \frac{I_d \Delta}{\sqrt{(4\pi)^2\langle N\rangle^2 + \alpha^2 \mathcal{D}_{xx}^2}} \frac{\Theta_{\text{sh}}}{\Sigma_i t_i M_{s_i}} j, \qquad (1)$$

where j is the electrical current density in the heavy metal layer, t is the thickness of that layer, Δ is the characteristic domain wall length, M_s is the saturation magnetization in each layer of the synthetic antiferromagnet, α is the Gilbert damping, and γ is the gyromagnetic ratio. Two characteristic properties of the skyrmion texture are the (integer) winding number $\langle N\rangle = (4\pi)^{-1} \int \mathbf{m} \cdot (\partial_x \mathbf{m} \times \partial_y \mathbf{m})\, \mathrm{d}x\, \mathrm{d}y$ and the longitudinal component of the dissipation tensor $\mathcal{D}_{xx} = \int (\partial_x \boldsymbol{m})^2\, \mathrm{d}x\, \mathrm{d}y$, which provides a kinetic friction force for the moving skyrmion. The direction of the skyrmion motion is determined by the ratio of $4\pi\langle N\rangle$ to $\alpha\mathcal{D}_{xx}$. Because of the off-diagonal response of a skyrmion's velocity, $\alpha\mathcal{D}_{xx}$ ends up being principally responsible for the longitudinal skyrmion motion along the track while the winding number $\langle N\rangle$ gives a transverse Magnus force. Finally, the factor I_d accounts for the spatially varying response of the skyrmion due to its spin texture; this term can be approximated as $I_d \approx e^{-r/\Delta} + \frac{\pi r}{\Delta}$ in rigid skyrmions of radius r, where r is defined by the $m_z = 0$ contour [25]. Note that, contrary to ferromagnetic skyrmions, skyrmions in synthetic antiferromagnets are not expected to experience a Magnus force, due to cancellation between the two oppositely magnetized layers [26]; their net topological charge is $\langle N\rangle = 0$. We develop a circuit module that integrates Eq. (1) over time to obtain the instantaneous skyrmion location on the racetracks, which is used to capture the effect of micromagnetic simulations as shown in Fig. 3.

The detection scheme is based on the tunneling magnetoresistance (TMR) found in magnetic tunnel junctions. In a magnetic tunnel junction, two magnetic layers are brought close to each other and separated by an insulating layer. The resistance through this material stack depends on the relative orientations of the magnetizations on either side of the insulator. This effect allows the information stored in these magnetic domains to be electrically read out. In this system, the insulating layer is deposited directly on top of the racetrack layer and an additional ferromagnetic layer with a fixed magnetization is deposited on top of that. The resistance measured by this detector changes when there is a skyrmion in the racetrack below the tunnel junction.

III. Temporal Memory Circuits and Simulations

A. Temporal-to-Spatial Mapping Scheme

A cell of the proposed magnetic skyrmion based temporal memory, comprising of two parallel racetracks, is shown in Fig. 2. The full memory consists of N such cells interfaced with an N channel data bus on which the input wavefronts arrive. The linearity of skyrmion displacement with the drive current pulse length is central to storing temporal data into a linearly mapped spatial skyrmion arrangement on a racetrack. The operating principle of the memory is simple. After a write operation, the temporal location of the incoming edge in the computation window should be mirrored by the spatial location of the skyrmion in the racetrack as shown in Fig. 3. Hence when the memory is read, the spatial displacement of the skyrmion leads to a corresponding temporal signal, faithfully reproducing the relative times of the incoming edges. This is done in the following way.

The temporal window of computation is defined as τ, and given a fixed current density of operation, this determines the length of the track $L = \tau v$, where v is the velocity of the skyrmion at the given current density. We assume an erased array to begin with, and a pre-nucleated skyrmion at its origin to avoid skyrmion nucleation costs which are generally understood to be expensive [16], although recent

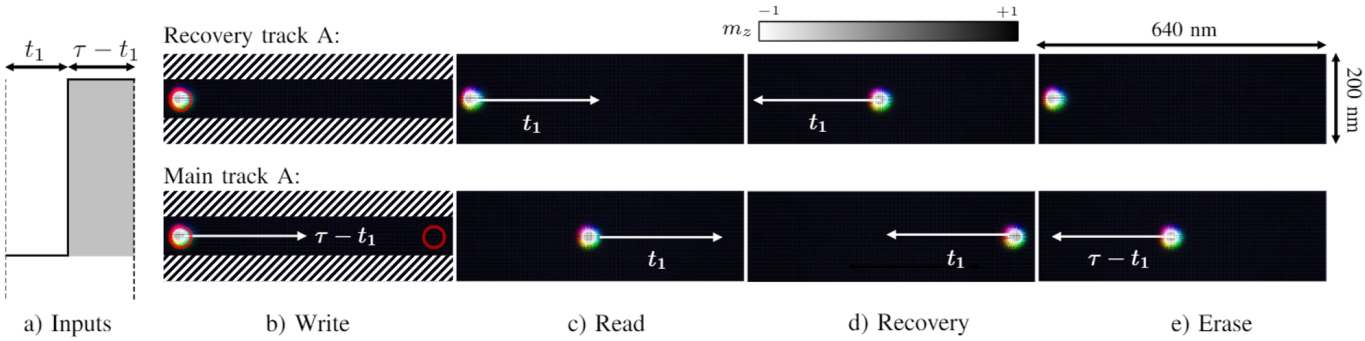

Fig. 3. (a) An example of a waveform presented as input to be stored by the memory cell. (b-e) Four operational phases of the racetrack temporal memory from a micromagentic simulation. The figures show a background of up-spin magnetic material (black) with a down-spin skyrmion (white); color indicates the in-plane spin orientation. In each phase, the skyrmion location is shown at the beginning of that phase while the arrows point to the location of the skyrmion at the end of the phase. The red circles in the first column indicate the magnetic tunnel junction (MTJ) placement, and the patterned area of the racetracks represents an energy barrier. Amplified repulsion from edges helps keep the skyrmions on their path to the read MTJ locations while preventing them from annihilation or pinning at the edges.

work [27] suggests this may not be prohibitive. During the write process, an incoming rising edge triggers the movement of the skyrmion along the track for a time corresponding to $\tau - t_1$. During the read process, the same current density, hence velocity of the skyrmion, causes a rising edge at the output at time t_1. Though conceptually straightforward, there are some engineering challenges that need to be addressed.

First, readout of such a memory changes the skyrmion position, leading to loss of the state, resulting in a destructive read. We fix this problem by adding a recovery track which saves the skyrmion state. This adds a process known as the recovery operation that restores the previously written value back to the racetrack. The use of a recovery track doubles the complexity of the memory. There are almost certainly other solutions to making a non-destructive memory and some might be less complex but we have not identified one. Second, the supporting circuitry must be developed that allows reliable operation of such a memory, which depends on being able to sense the arrival of a skyrmion with magnetic tunnel junction based readout circuits. These circuits are shown next. Third, such sensing circuits can be designed such that they are shared, hence improving the area and energy efficiency of such a memory. Last, such a memory should be able to work in the presence of non-idealities of the magnetic material. The next section discusses the non-idealities and their effects in detail.

B. Write Operation

The circuit for the write operation is shown in Fig. 2. As described before, we assume pre-nucleated skyrmions in the starting position of the main racetrack and the recovery racetrack as the initial condition for the write operation. Note that the write operation involves only the primary/main racetrack and not the recovery racetrack, hence only transistors T_1 and T_2 are turned on. The write operation corresponds to shifting the skyrmion in the primary racetrack by the application of the write current (Fig. 3b.), resulting in the duty cycle dependence of the final position of the skyrmion. This is done as follows.

A high ($V_{DD} = 1$ V for 45 nm CMOS technology node) voltage is applied to the bit line (BL) and ground to the source line (SL) while the write line (WL) transistor, T_2, is turned on to establish a current path. When the temporally coded incoming rising edge arrives at the gate of the transistor T_1, the write path through the heavy metal layer is turned on and the skyrmion begins its motion along the racetrack. When the incoming signal falls to ground, the transistor T_1 turns off, causing the skyrmion to stop moving and isolating the main racetrack from the BL, successfully storing temporal information spatially. The control signals and corresponding skyrmion motion as a function of time are shown in Fig. 4a.

C. Read Operation

Once the input signal is captured in the racetrack by displacement of the skyrmion to the appropriate position, it can be read. As can be seen in Fig. 3c, this read operation changes the skyrmion position and hence is destructive, leading us to employ a corresponding recovery track as shown in Fig. 2. During the read operation, the skyrmion in the main racetrack reaches the end of the racetrack, sending off a rising edge, while the recovery track is written with the corresponding t_1 value. This keeps a constant relative displacement between the read and recovery tracks, allowing easy recovery as described in the next subsection.

For the read operation, the circuits are arranged as follows. The BL is biased to high ($V_{DD} = 1$ V for 45 nm CMOS technology node) and SL is biased to ground, similar to the write operation. Transistor T_5 is enabled for readout while the transistor T_1 is turned off. The node voltage V_M is low and V_f is high at the beginning of the read operation due to the low resistance value of the readout MTJ in the absence of skyrmion. This high V_f turns on the identically sized transistors T_4 and T_8 allowing the same current to flow through both the primary and recovery racetrack. As soon as the rising edge that initiates the read operation appears at the inputs to transistors T_2 and T_3, the skyrmion on both racetracks see the same current in the HM layer and exhibit identical

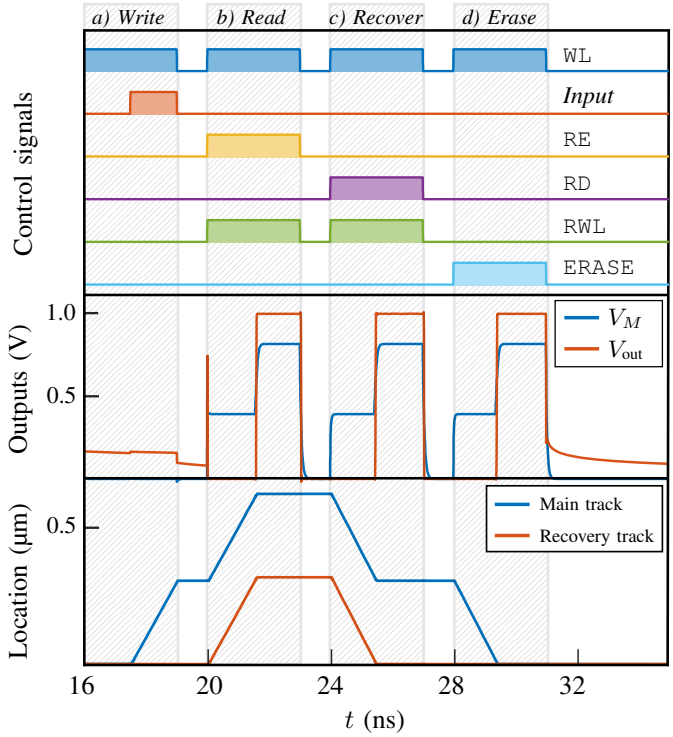

Fig. 4. Control signals, output signals, and skyrmion positions on the two tracks through four phases memory cell operation. Top: control bits for the write line (WL), read enable (RE), recovery line (RD), recovery write line (RWL), erase line (ERASE), and the incoming signal. The middle panel shows the synchronizing signal V_{OUT} and the node voltage V_M. The bottom panel shows the instantaneous skyrmion positions on the tracks. The four operational phases are highlighted with gray boxes. a) WL drives the skyrmion along the main racetrack, once the input arrives. b) WL and RWL drive the main and recovery skyrmions, respectively, until the main track skyrmion reaches the MTJ read stack, which is enabled by RE. The resulting resistance change in the MTJ, detected via the feedback voltage V_f (inverse of V_{OUT}), halts the drive currents through the tracks. c) WL and RWL drive the main and recovery skyrmions as in the read operation, but with the relative polarity of source and bit lines inverted to enable a reverse drive current. Motion stops when the recovery track skyrmion is detected by the RD-enabled MTJ stack, restoring the original position of the main track skyrmion before the read operation. d) The ERASE signal drives the main skyrmion back to the origin point until it is detected and the drive current is turned off. V_{bias} is held constant at 1.2 V during the read, recovery, erase operations.

dynamics, making them move the same distance along the racetracks as shown in Fig. 3c. When the skyrmion along the main racetrack reaches the end of the racetrack, V_M sees an increase in voltage, causing V_f to fall to ground turning off transistors T_4, and T_8 ending the read process. The recovery skyrmion is now frozen at the location corresponding to $t = t_1$ as shown in Fig. 3c.

In order to correctly detect the skyrmion, we use MTJ readout circuitry based on the difference in resistance between there being a skyrmion beneath the detector, R_{sk}, or not, R_{P}. This difference can be characterized by an effective TMR, $\beta = (R_{\text{sk}} - R_{\text{P}})/R_{\text{P}}$. A bias voltage V_{bias} with a reference resistor R_0 produces a voltage swing that depends on the effective TMR β at node V_M in the presence or absence of a skyrmion. The voltage at the node V_M is then amplified by an inverter to feed high or low feedback voltage V_f to the gates of the nMOS transistors T_4 and T_8. Note that the readout circuitry is biased in such a pattern that the feedback voltage V_f turns on the transistors T_4 and T_8 during the read, recovery, and erase operations only.

The reference resistance should provide the maximum contrast between the skyrmion and no-skyrmion states under the MTJ. To ensure functionality, R_0 must be between R_{0-} and R_{0+}. In general $R_{0\pm} = \tilde{R} - u\rho \pm \rho\sqrt{1 + u^2 - 2\tilde{R}u/\rho}$, where $\tilde{R} = R_T + (R_s + R_p)/2$, $u = V_{\text{bias}}/V_{\text{min}}$, and $\rho = (R_s - R_p)/2$. V_{min} is the minimum voltage difference between the skyrmion and no-skyrmion states needed at node V_M to ensure switching behavior at V_f, and V_{bias} is the bias voltage. The maximum possible voltage swing is given when $R_0 = \sqrt{R_{\text{sk}} R_{\text{P}}}$. Higher resistances increase efficiency and can be achieved by using a thicker MgO layer in the MTJ. It is essential that the effective TMR be large and the filling of the MTJs by the skyrmions be almost complete because the read-out circuitry begins to function poorly when the effective TMR is below 50 %. Using domain walls instead of skyrmions would give a larger read-out signal at the potential cost of more complicated motion.

Note that though V_M is connected to three different MTJs, only one of them is active in each operating phase. This node and its associated readout circuitry can therefore be shared by these MTJs. This is achieved by controlling inputs to transistors $\mathrm{T}_5, \mathrm{T}_6$ and T_7.

D. Recovery Operation

The recovery operation compensates for the destructive read by restoring the primary and recovery skyrmions to their pre-written values as is shown in Fig. 3d. This operation is the opposite of the read operation and is performed almost identically, except with a reversed applied current. Hence, the BL is biased to ground while the SL is biased to high. Using transistor T_6 with the MTJ, the shared readout circuit is triggered when the skyrmion returns to its original location under the MTJ in the recovery track (Fig. 4c.) This detects the end of the operation and turns off transistors T_8 and T_4, hence shutting off the current path. After this operation is complete, the skyrmion in the recovery track is restored to its default origin position, while the skyrmion in the primary track is restored to its value before the read. The cell is now ready for another read as described before, or an erase operation.

E. Erase Operation

The final operation is the erase operation to return the cell to its configuration before any write operations, i.e., the configuration with the skyrmions on both racetracks at the origin. The erase operation is similar to the write operation as only the skyrmion in the main track is moved, but the direction of the drive current and hence skyrmion motion is reversed (Fig. 3e.). The ending of this phase is determined by the MTJ at the beginning of the main track. Note that in this phase, the other two MTJs are not required, causing transistors T_5 and T_6 to be disabled, while transistor T_7 can be enabled. As soon as the node voltage V_f of the readout circuitry turns low, the

transistor T_4 turns off, signaling the end of the erase operation (Fig. 4d.). The cells are now ready to write a new state.

F. Robustness to parameter variations

The memory is designed in such a way that during the read and recovery operations, only the travel times of the main and recovery track skyrmions matter, not the distances covered by them; and the travel time is controlled by the feedback voltage V_f and as long as the travel time of both the main and recovery track skyrmions is the same, the different velocities in the racetracks will have no impact. For example, if the skyrmion velocities are different in the main and recovery tracks, both the skyrmions will move until the main racetrack skyrmion reaches under the read MTJ stack. Now, during the recovery operation, both the skyrmions will retrace the distances they traveled during the read operation, resulting in the restoration of the skyrmion position (i.e., at the beginning for the recovery track and at the same place where it was positioned after the write operation for the main track).

IV. Results and Discussion

A. Performance and Comparison with CMOS

In order to understand and quantify the performance of this wavefront memory cell, we perform detailed circuit simulations using a modular approach. In particular, we construct a complete circuit model using the 45 nm and 16 nm technology node obtained from the Predictive Technology Model [28] for the driving transistors and the module discussed in Sec. II for the dynamics of the skyrmions in the magnetic racetracks. We compare our results to conventional CMOS implementations constrained by the requirements of the application. Race logic has been successfully demonstrated to work well for 4 to 6 bit precision for a variety of applications such as decision trees and DNA sequence alignment [2], [3]. We limit ourselves to a similar precision in order to make a fair comparison.

To describe the magnetic dynamics, our simulations use a Gilbert damping constant $\alpha = 0.1$, saturation magnetization $M_s = 3 \times 10^5$ A/m, magnetic anisotropy $K = 135$ kJ/m^3, Dzyaloshinskii-Moriya interaction strength $D = 1.2$ mJ/m^2, and exchange stiffness $A_{ex} = 7.5 \times 10^{-12}$ J/m. Consequently, the domain wall width is $\Delta = \pi D/4K \approx 7.0$ nm and the skyrmion radius is approximately 24 nm [29]. We use these parameters to model a synthetic antiferromagnet with thickness of $t_{\text{FM}} = 2$ nm for each layer, length of 640 nm, and width of 200 nm. The heavy metal thickness is assumed to be 5 nm. The MTJ diameter is 40 nm with a TMR of ≈ 400 %, based on $R_P = 6.67$ kΩ and $R_{AP} = 33.3$ kΩ. The circuit behaves acceptably provided the bare TMR is greater than 100 %, or an effective TMR of 50 %. With the chosen parameters, the skyrmion size is comparable to that of the MTJ. From the structure of the skyrmion, we compute a 50 % reduction in the TMR to an effective TMR close to 200 %. The reference resistance is taken to be $R_0 = 15$ kΩ, giving a swing voltage of ≈ 425 mV.

The saddle point barrier to skyrmion annihilation is $E_{\text{b}} \approx 23 A_{\text{ex}} t_{\text{FM}} - E_{\text{skyr}} \approx 100\ kT$, where k is the Boltzmann constant and T is the temperature. Assuming a mean lifetime to annihilation of [30] $\tau = f_0 e^{E_{\text{b}}/k_B T}$ with $f_0 = 10^{-10}$, the lifetime of these skyrmions is in the range of years at room temperature. While the Magnus force is small in the compensated synthetic antiferromagnet, the wide track requires skyrmion injection right down the middle to hit the MTJ. This constraint can be avoided with narrower racetracks with repulsive edges.

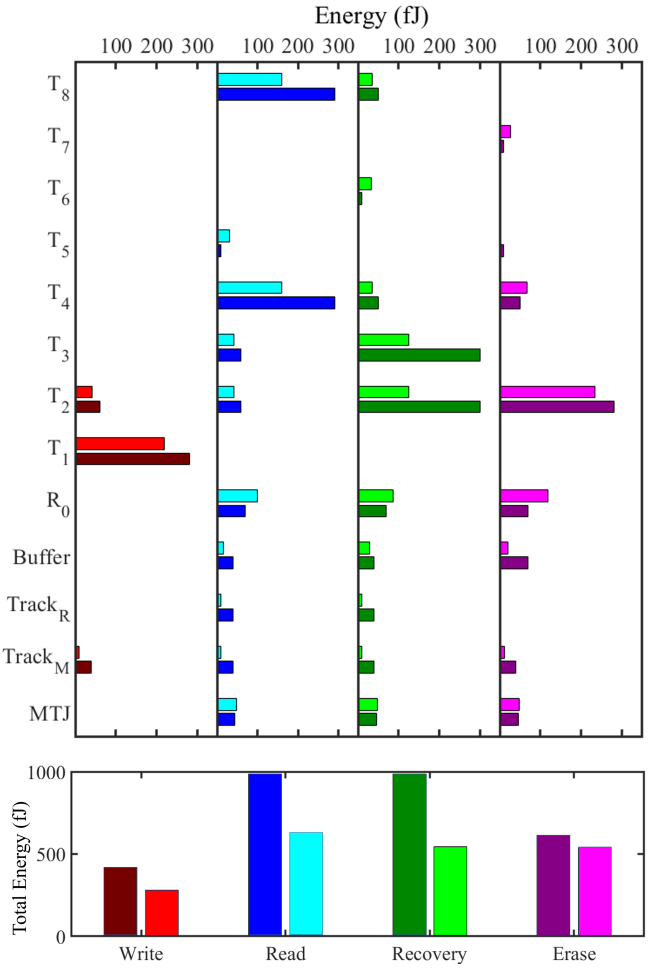

Fig. 5. Total energy, and energy breakdown per circuit element, for an average case of wavefront capture on a set of 640 nm racetracks. The read and recovery processes consume the most energy because both tracks are active in these operations. In all operations, the largest energy dissipation arises from high overdrive voltage of transistor, with only minimal contributions from Joule heating in the main (Track$_M$) and recovery (Track$_R$) racetracks. Darker and brighter color bars show the results done with 45 nm and 16 nm CMOS transistors, respectively.

Fig. 5 gives the energies consumed for these four operations in each component, for an average case in which the skyrmion is moved to the middle of the racetrack. The bottom panel of Fig. 5 shows that the energy consumption is highest in the read and recovery operations, as it involves driving both racetracks, while the write and erase operations, single track operations, consume relatively less energy. For the chosen parameters for the racetrack and the racetrack drive currents (≈ 244 μA), the total energy consumption of the complete cell is ≈ 2.8 pJ (45

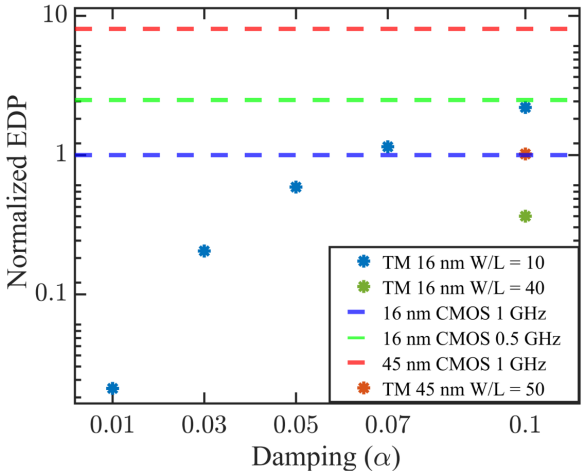

Fig. 6. Normalized energy delay product (EDP) of 4-bit CMOS counters and four temporal memory cells (TM) with different damping constant with 16 nm and 45 nm transistors (normalized to EDP of 1 GHz 16 nm counter-based CMOS temporal memory).

nm CMOS technology) for the full cycle of the four memory operations, at a 50 % duty cycle—that is, the case where the write process puts the skyrmion midway in the racetrack. This is the average case ($t = \tau/2$) of all possible temporal data recordings in the range $[0, \tau]$. The top panel of Fig. 5 shows that the energy consumption in the racetracks themselves (due to Joule heating) is only a minor component of the total energy consumption and is on the order of ≈ 45 fJ.

Such a memory cell is only useful if it has advantages over alternatives based solely on CMOS. Here we make a comparison with an equivalent conventional 4 bit CMOS temporal memory by using an up-counter coupled with a latch or an SRAM cell. The up-counter counts 'clock ticks' and can thereby digitize a clock delay. An up-counter is built out of multiple positive edge-triggered T-flip flops (TFF) and combinational circuits. Textbook implementations of one such TFF requires 20 MOSFETs. Our racetracks store 'analog' temporal information, while a counter stores quantized information. For example, a 4 stage counter, which has a $2^4 = 16$ step quantization of temporal data, will require 94 transistors. Coupling them with simple S-R latches to store the memory requires another 32 transistors, yielding a total of 126 transistors. In addition, this scheme stores the data logarithmically and its readout requires either a Boolean decoder circuit increasing the number of transistors (an estimate for a 4×16 decoder requires 172 transistors) or a clock generator driven by latch readout (component count dependent on implementation scheme).

The proposed design can also replay back the temporal data during the recovery operation making the recovery operation another read operation. For low damping ($\alpha = 0.01$), the EDP of the temporal memory is smaller by factors of 32 and 81 compared with the CMOS counterpart working at 1 GHz and 500 MHz clock frequencies, respectively. Moreover, EDP advantage will be even greater for higher numbers of read operations per write operation. The above comparison was made for a full cycle operation where write, read, and erase energies are counted once. When the temporal data is read multiple times per write, everything but the read energy will be amortized. Thus, for 10 and 100 reads, the EDP improves by factors of 74 and 84 respectively. At high damping ($\alpha = 0.1$), the CMOS counterpart has better EDP only when the driving transistor resistance is low for the temporal memory. A higher W/L ratio of the driving transistors decreases the EDP at the cost of increased area; the EDP improves by factors of 3, 11, and 16 when the temporal data is read 1 time, 10 times, and 100 times, respectively, compared to the CMOS counterpart at 1 GHz clock frequency. An additional advantage is the built-in non-volatility of the skyrmion racetrack; the proposed cell can be powered off without losing information for significant lengths of time, compared to the volatile nature of a pure CMOS design. For systems where long term memory is required the proposed memory cell would consume significantly less energy than the counter-based CMOS-only design.

B. Non-Idealities of Racetracks and Possible Solutions

Our simulations are based on pristine racetracks at zero temperature. In experimental realizations, there will be a number of complications. The most significant constraints on the operation of this memory are non-idealities in the tracks and device variation from track to track. To reliably store and recall temporal signals, skyrmion velocities need to be predictable and narrowly distributed among all tracks. Today, tracks cannot be fabricated with adequate reproducibility to achieve these margins. If applications of skyrmion racetracks become commercially appealing, we expect that such interest would lead to substantial improvement in fabrication quality and make the manufacture of our device realistic. The velocity distribution ultimately limits the precision of the memory and the number of significant bits it can store and recall. We argue below that even four bits of precision could be useful.

Finite temperature and pinning: At finite temperature, skyrmions in close to ideal racetracks undergo rapid diffusion and Brownian motion [31] that would make the racetracks unusable. At the other extreme, both intentional defects such as notches [32] or unintentional ones such as material non-uniformities give rise to pinning centers that can keep skyrmions pinned, preventing thermal diffusion but also preventing the current induced motion necessary for the operation of the racetrack. For the proposed racetrack to be viable, this pinning needs to be strong enough that the skymion remains localized with high probability for a sufficient time. If we require that a skymion remain localized to within 40 nm with probability $1 - 10^{-8}$ for 1 h, the diffusion constant must be less than 10^{-22} m^2/s.

On the other hand, the presence of pinning centers leads to a minimum drive current for predictable motion. Below that drive, the motion is in the creep regime and consists of repeated stochastic thermally assisted depinning and pinning. Reliable racetrack memory requires that the pinning be weak enough to allow low drive current predictable motion but

strong enough that Brownian motion is minimized during storage. The study of pinning is still in its infancy with reported diffusion constants ranging from $> 10^{-11}$ m^2/s in optimized material stacks [31] to $< 10^{-20}$ m^2/s [33]. The latter system exhibits predictable current-induced motion [33] for similar current density to what is used in this paper, suggesting that there is an experimental regime in which the racetrack described here functions as reported. As material fabrication capabilities improve, it may be desirable to control the motion with lithographically designed pinning cites. The memory would then be discrete rather than analog, but such memory is sufficient for a variety of race-logic applications.

These arguments hold for domain walls, but skymions and domain walls are more strongly affected by different types of disorder. Since skyrmions are localized to the center of the racetrack, they are less susceptible to pinning by edge roughness. Domain walls are more extended so they are less susceptible to pinning by the anisotropic grains that can exist in a racetrack [17].

Skyrmion readability: The read-out of skyrmion motion is through the TMR effect of the MTJ as mentioned before. The relative area of the skyrmion under the MTJ is a critical factor for obtaining a sufficiently large resistance swing to control the output swing of signal V_M in Fig. 3. Matching the area can be achieved by using large skyrmions to fill a significant area under the MTJ reader. Using domain walls removes this problem as they can be large enough to provide nearly full MTJ cross-section coverage. Improving the TMR of MTJs would allow smaller skyrmions to be read with sufficient contrast. Another solution is to use additional transistors at the read stage to amplify the smaller voltage contrast provided by a low read TMR, but with a higher area and energy cost.

Parameter Variations: Meaningful comparison of times stored on different write tracks requires a tight distribution of skyrmion velocities on the write tracks. In addition, all of the tracks must be long enough that the fastest skyrmions from the distributions of speeds would not reach the end of a track over the maximum write duration and get annihilated. On the other hand, the requirements for the recovery tracks are less stringent except for constraints on their length. During the read and recovery operations, only the travel times of the main and recovery track skyrmions matter, not the distances they cover. The travel time is controlled by the feedback voltage V_f so that as long as the travel time of both the main and recovery track skyrmions is the same, the different velocities will have no impact. However, to ensure that the recovery skyrmion never gets driven off the end of the track, the recovery tracks need to be long enough that the fastest skyrmion will not reach the end in the time that it takes the slowest skyrmion to travel between the two write track MTJs.

The much faster time scales of CMOS control circuitry compared to that of the current pulses used to move the skyrmion ensures the fidelity of the read process. The considered skyrmion speed is 200 m/s while the transistor switches in picoseconds. The propagation delay of the inverter used in the shared read circuitry is 4 ps (for 45 nm CMOS) during which the skyrmion moves only 0.8 nm. If we incorporate parasitic effects, variations, and the MTJ switching time (time for the V_M to go high), the skyrmion movement will increase. If the isolation time of the racetrack during the read operation becomes 40 ps after incorporating all the delays, the skyrmion will only move 8 nm, smaller than the size of the skyrmion.

Edge annihilation: Skyrmions are susceptible to annihilation by the edges of the racetrack. This can be partially ameliorated by use of wide racetracks, whereby we can avoid the possibility of skyrmion drifting to the edges and getting annihilated due to Magnus effect. This however can worsen the skyrmion readability as this scales up the size of the read MTJs, or makes skyrmions miss the MTJ as discussed above. While the use of synthetic antiferromagnet racetracks with compensation for Magnus force can considerably reduce this issue, use of reflecting edges through anisotropy engineering, such as using ion beam irradiation, or geometry engineering by using thicker edges can also help with this issue [34], [35]. By using reflective edges, the skyrmion will more reliably reach the MTJ position, as the repulsive force from the edges will confine the skyrmion to the middle of racetrack.

All of the issues listed above are common for all magnetic racetracks, an area of active research, and this device will benefit from any future improvements through better fabrication capabilities and novel extrinsic circuit techniques.

C. Applications of race logic

Race logic performs well against conventional approaches on tasks that involve searching through a large space and that are amenable to a dynamic programming approach. It effectively implements the principle from dynamic programming of optimality, which constrains solutions of sub-computations to be (partial) solutions of the global computation, in turn allowing chaining of temporal computations. Examples of useful computational problems obeying the principle of optimality are shortest paths of graphs and machine classification with decision trees. These applications require decisions to be made at each node in a graph or tree about what computational path needs to be taken next. By mapping the input into temporal edges and by effectively utilizing tunable delay elements, such decisions can be made using simple OR and AND gates on single wires.The precision required of the delay elements by such applications is limited to between 4 to 6 bits. With this limited precision, Ref. [1] shows a 200× energy advantage over conventional limited precision digital approaches. Ref. [3] describes a fabricated chip that performs a similar task at a power level of 70 mW. Ref. [2] performs classification with decision trees at an energy delay product 4 orders of magnitude superior to state of the art.

V. Conclusion

In this work, we utilize a novel technology based on skyrmions in magnetic racetracks to provide energy-efficient memory for temporal computing. Our proposal uses the displacement of pre-nucleated skyrmions to store temporal information in magnetic racetracks, avoiding the energy cost

required to nucleate them. The readout of this memory is made non-destructive by doubling the number of racetracks to store the information in the second racetrack during the read process. While this increases the energy consumption and area of the memory, we limit the impact by sharing readout and control circuits for the different operations involved in using the memory. The linearity of the stored information with respect to the input limits the circuitry required for translation between the time and displacement domains. By changing the current used to drive the skyrmions, and hence their velocity, these memory cells can be tuned so their time scale matches the range of input information. Such an efficient memory will greatly expand the problem domain that can be efficiently addressed by timing based computing.

VI. Acknowledgments

This work is funded in part by the Defense Advanced Research Projects Agency (DARPA) Topological Excitations in Electronics (TEE) program (grant D18AP00009). A.M. acknowledges support under the Cooperative Research Agreement Award No. 70NANB14H209, through the University of Maryland. We would like to thank Andrew Kent, Joe Poon, Geoffrey Beach, Kai Litzius, and Brian Hoskins for insightful discussions.

References

[1] A. Madhavan, T. Sherwood, and D. Strukov, "Race Logic: A hardware acceleration for dynamic programming algorithms," in 2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA), Jun. 2014, pp. 517–528, iSSN: 1063-6897.

[2] G. Tzimpragos, A. Madhavan, D. Vasudevan, D. Strukov, and T. Sherwood, "Boosted race trees for low energy classification," in Proceedings of the Twenty-Forth International Conference on Architectural Support for Programming Languages and Operating Systems, ser. ASPLOS '19, April 2019.

[3] A. Madhavan, T. Sherwood, and D.Strukov, "A 4-mm 2 180-nm-cmos 15-giga-cell-updates-per-second dna sequence alignment engine based on asynchronous race conditions," in 2017 IEEE Custom Integrated Circuits Conference (CICC). IEEE, 2017, pp. 1–4.

[4] J. Smith, "Space-time algebra: A model for neocortical computation," in 2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA). IEEE, 2018, pp. 289–300.

[5] A. Madhavan, G. Tzimpragos, M. Stiles, and T. Sherwood, "A truth-matrix view into unary computing."

[6] A. Madhavan and M.D.Stiles, "Storing and retrieving wavefronts with resistivetemporal memory," in 2020 IEEE International Symposium on Circuits and Systems (ISCAS), submitted. IEEE, 2020, pp. 1–4.

[7] S. S. P. Parkin, M. Hayashi, and L. Thomas, "Magnetic Domain-Wall Racetrack Memory," Science, vol. 320, no. 5873, pp. 190–194, Apr. 2008.

[8] R. Tomasello, E. Martinez, R. Zivieri, L. Torres, M. Carpentieri, and G. Finocchio, "A strategy for the design of skyrmion racetrack memories," Sci Rep, vol. 4, no. 1, p. 6784, May 2015.

[9] C. Hanneken, F. Otte, A. Kubetzka, B. Dupé, N. Romming, K. von Bergmann, R. Wiesendanger, and S. Heinze, "Electrical detection of magnetic skyrmions by tunnelling non-collinear magnetoresistance," Nature Nanotechnology, vol. 10, no. 12, pp. 1039–1042, Dec. 2015.

[10] K. Y. Camsari, R. Faria, B. M. Sutton, and S. Datta, "Stochastic p -Bits for Invertible Logic," Phys. Rev. X, vol. 7, no. 3, p. 031014, Jul. 2017.

[11] A. Belavin and A. Polyakov, "Metastable states of two-dimensional isotropic ferromagnets," JETP lett, vol. 22, no. 10, pp. 245–248, 1975.

[12] A. Thiaville, J. M. García, and J. Miltat, "Domain wall dynamics in nanowires," Journal of Magnetism and Magnetic Materials, vol. 242-245, pp. 1061–1063, Apr. 2002.

[13] G. S. D. Beach, C. Nistor, C. Knutson, M. Tsoi, and J. L. Erskine, "Dynamics of field-driven domain-wall propagation in ferromagnetic nanowires," Nature Mater, vol. 4, no. 10, pp. 741–744, Oct. 2005.

[14] A. N. Bogdanov and U. K. Rößler, "Chiral Symmetry Breaking in Magnetic Thin Films and Multilayers," Phys. Rev. Lett., vol. 87, no. 3, p. 037203, Jun. 2001.

[15] I. Dzyaloshinsky, "A thermodynamic theory of "weak" ferromagnetism of antiferromagnetics," Journal of Physics and Chemistry of Solids, vol. 4, no. 4, Jan. 241–255.

[16] J. Sampaio, V. Cros, S. Rohart, A. Thiaville, and A. Fert, "Nucleation, stability and current-induced motion of isolated magnetic skyrmions in nanostructures," Nature nanotechnology, vol. 8, no. 11, pp. 839–844, 2013.

[17] R. Tomasello, V. Puliafito, E. Martinez, A. Manchon, M. Ricci, M. Carpentieri, and G. Finocchio, "Performance of synthetic antiferromagnetic racetrack memory: domain wall versus skyrmion," J. Phys. D: Appl. Phys., vol. 50, no. 32, p. 325302, Aug. 2017.

[18] X. Zhang, Y. Zhou, and M. Ezawa, "Magnetic bilayer-skyrmions without skyrmion Hall effect," Nat Commun, vol. 7, no. 1, p. 10293, Apr. 2016.

[19] H. Vakili, Y. Xie, and A. W. Ghosh, "Self focusing hybrid skyrmions in spatially varying canted ferromagnetic systems." [Online]. Available: http://arxiv.org/abs/2001.00729

[20] W. Legrand, D. Maccariello, F. Ajejas, S. Collin, A. Vecchiola, K. Bouzehouane, N. Reyren, V. Cros, and A. Fert, "Room-temperature stabilization of antiferromagnetic skyrmions in synthetic antiferromagnets," Nat. Mater., Sep. 2019.

[21] T. Dohi, S. DuttaGupta, S. Fukami, and H. Ohno, "Formation and current-induced motion of synthetic antiferromagnetic skyrmion bubbles," Nat Commun, vol. 10, no. 1, p. 5153, Dec. 2019.

[22] Y. Xie, J. Ma, H. Vakilitaleghani, Y. Tan, and A. W. Ghosh, "Computational search for ultrasmall and fast skyrmions in the Inverse Heusler family," arXiv:1901.09446 [cond-mat], Jan. 2019. [Online]. Available: http://arxiv.org/abs/1901.09446

[23] A. A. Thiele, "Steady-State Motion of Magnetic Domains," Phys. Rev. Lett., vol. 30, no. 6, pp. 230–233, Feb. 1973.

[24] E. G. Tveten, A. Qaiumzadeh, O. A. Tretiakov, and A. Brataas, "Staggered dynamics in antiferromagnets by collective coordinates," Phys. Rev. Lett., vol. 110, p. 127208, Mar 2013. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.110.127208

[25] F. Büttner, I. Lemesh, and G. S. D. Beach, "Theory of isolated magnetic skyrmions: From fundamentals to room temperature applications," Sci Rep, vol. 8, no. 1, p. 4464, Dec. 2018.

[26] J. Barker and O. A. Tretiakov, "Static and dynamical properties of antiferromagnetic skyrmions in the presence of applied current and temperature," Phys. Rev. Lett., vol. 116, p. 147203, Apr 2016. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.116.147203

[27] M. G. Mankalale, Z. Zhao, J. Wang, and S. S. Sapatnekar, "Skylogic—a proposal for a skyrmion-based logic device," IEEE Transactions on Electron Devices, vol. 66, no. 4, pp. 1990–1996, 2019.

[28] "Predictive technology model (ptm)." [Online]. Available: http://ptm.asu.edu/

[29] X. S. Wang, H. Y. Yuan, and X. R. Wang, "A theory on skyrmion size," Communications Physics, vol. 1, no. 1, p. 31, Jul. 2018.

[30] P. F. Bessarab, G. P. Müller, I. S. Lobanov, F. N. Rybakov, N. S. Kiselev, H. Jónsson, V. M. Uzdin, S. Blügel, L. Bergqvist, and A. Delin, "Lifetime of racetrack skyrmions," vol. 8, no. 1, p. 3433. [Online]. Available: https://doi.org/10.1038/s41598-018-21623-3

[31] L. Zhao, Z. Wang, X. Zhang, X. Liang, J. Xia, K. Wu, H.-A. Zhou, Y. Dong, G. Yu, K. L. Wang, X. Liu, Y. Zhou, and W. Jiang, "Topology-dependent brownian gyromotion of a single skyrmion," Phys. Rev. Lett., vol. 125, p. 027206, Jul 2020. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.125.027206

[32] D. Suess, C. Vogler, F. Bruckner, P. Heistracher, F. Slanovc, and C. Abert, "Spin torque efficiency and analytic error rate estimates of skyrmion racetrack memory," Scientific Reports, vol. 9, no. 1, p. 4827, 2019.

[33] G. Beach, private communication.

[34] H. T. Fook, W. L. Gan, I. Purnama, and W. S. Lew, "Mitigation of magnus force in current-induced skyrmion dynamics," vol. 51, no. 11, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/document/7108038/

[35] K. Fallon, S. Hughes, K. Zeissler, W. Legrand, F. Ajejas, D. Maccariello, S. McFadzean, W. Smith, D. McGrouther, S. Collin, N. Reyren, V. Cros, C. H. Marrows, and S. McVitie, "Controlled individual skyrmion nucleation at artificial defects formed by ion irradiation," p. 1907450. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/smll.201907450