Journal of Applied Physics

RESEARCH ARTICLE | JULY 01 2024

Thermal optimization of two-terminal SOT-MRAM

Haotian Su ; Heungdong Kwon ; William Hwang ; Fen Xue ; Çağıl Köroğlu ; Wilman Tsai; Mehdi Asheghi; Kenneth E. Goodson; Shan X. Wang ; Eric Pop

J. Appl. Phys. 136, 013901 (2024)
https://doi.org/10.1063/5.0211620

Journal of Applied Physics ARTICLE pubs.aip.org/aip/jap

Thermal optimization of two-terminal SOT-MRAM

Cite as: J. Appl. Phys. **136**, 013901 (2024); doi: 10.1063/5.0211620
Submitted: 1 April 2024 · Accepted: 14 May 2024 ·
Published Online: 1 July 2024

Haotian Su,[1] Heungdong Kwon,[2] William Hwang,[1] Fen Xue,[1] Çağıl Köroğlu,[1] Wilman Tsai,[3] Mehdi Asheghi,[2] Kenneth E. Goodson,[2,3] Shan X. Wang,[1,3] and Eric Pop[1,3,4,5,a)]

AFFILIATIONS

[1]Department of Electrical Engineering, Stanford University, Stanford, California 94305, USA
[2]Department of Mechanical Engineering, Stanford University, Stanford, California 94305, USA
[3]Department of Materials Science and Engineering, Stanford University, Stanford, California 94305, USA
[4]Department of Applied Physics, Stanford University, Stanford, California 94305, USA
[5]Precourt Institute for Energy, Stanford University, Stanford, California 94305, USA

[a)]**Author to whom correspondence should be addressed:** epop@stanford.edu

ABSTRACT

While magnetoresistive random-access memory (MRAM) stands out as a leading candidate for embedded nonvolatile memory and last-level cache applications, its endurance is compromised by substantial self-heating due to the high programming current density. The effect of self-heating on the endurance of the magnetic tunnel junction (MTJ) has primarily been studied in spin-transfer torque (STT)-MRAM. Here, we analyze the transient temperature response of two-terminal spin–orbit torque (SOT)-MRAM with a 1 ns switching current pulse using electro-thermal simulations. We estimate a peak temperature range of 350–450 °C in 40 nm diameter MTJs, underscoring the critical need for thermal management to improve endurance. We suggest several thermal engineering strategies to reduce the peak temperature by up to 120 °C in such devices, which could improve their endurance by at least a factor of 1000× at 0.75 V operating voltage. These results suggest that two-terminal SOT-MRAM could significantly outperform conventional STT-MRAM in terms of endurance, substantially benefiting from thermal engineering. These insights are pivotal for thermal optimization strategies in the development of MRAM technologies.

I. INTRODUCTION

Commercial spin-transfer torque magnetoresistive random-access memory (STT-MRAM) requires high write current density for fast switching, which makes it unsuitable for ultrafast operation in the sub-nanosecond regime with the near-infinite endurance desired for machine learning.[1–3] It is well established that the high write current in STT-MRAM through the MgO tunnel barrier accelerates the breakdown of the MgO barrier and, therefore, raises endurance concerns.[4–7] Spin–orbit torque (SOT)-MRAM, which relies on the torque generated by the in-plane current in a heavy metal, has the potential to overcome these limitations by decoupling the write and read current paths.[8–11] This alleviates some thermal challenges by lowering the current across the magnetic tunnel junction (MTJ). However, one major disadvantage of SOT-MRAM is that it is a three-terminal device, requiring two transistors per cell for the read and write, which increases the size of the cells, limiting their practical applications.

Two-terminal SOT-MRAM was developed to address the above issues.[12–14] These devices combine SOT with STT in a synergistic manner to achieve low write current, high density, ultrafast operation, thus being considered a next-generation MRAM technology. However, previous studies on SOT-MRAM showed that self-heating of the SOT line during the write process inevitably results in elevated MTJ temperature, causing its degradation and limiting endurance.[15–17] Thermal aspects must be considered to make MRAM a reliable technology, but few studies have examined the thermal physics of MRAM in order to address and improve such challenges.

In this work, we analyze the transient temperature response of two-terminal SOT-MRAM with 1 ns write pulses and show that the peak temperature can reach about 350–450 °C. We highlight the importance of including the appropriate interface physics (in the form of the thermal boundary conductance, TBC) of various material interfaces within MRAM and the reduced effective thermal

conductivity of the constituent thin films. We then propose practical solutions to reduce the peak temperature of two-terminal SOT-MRAM, by material engineering and SOT line design. We further discuss the effect of voltage and temperature reduction on the expected endurance of MRAM. Our findings show that by thermal engineering, the peak temperature of two-terminal SOT-MRAM can be reduced by up to 120 °C, which could improve the endurance of these devices by a factor of ∼10–100× at 1 V, or by over 1000× if the voltage were reduced to 0.75 V. Lastly, we show the thermal robustness of two-terminal SOT-MRAM in comparison with typical STT-MRAM. Our findings hold significant implications for the thermal optimization of MRAM.

II. METHODS

A. Two-terminal SOT-MRAM design

Two-terminal SOT-MRAM features an SOT line below the free layer of the MTJ, to improve switching speed and reduce the write current of such devices,[12,14] and retains the single-transistor cell design typical of STT-MRAM, as depicted in Fig. 1(a). The SOT line is connected in series with the MTJ, with both STT and SOT being generated with the injection of current. In conventional SOT materials, the SOT-generated spins are polarized along the in-plane *y*-orientation, and STT-generated spins are polarized coaxially to the magnetic anisotropy of MRAM. The spin polarization angle of SOT is nominally orthogonal to the magnetic anisotropy of MRAM such that maximum SOT is applied to the free layer when current is injected into the device.[12] As a result, switching begins as soon as the current is injected, and the orientation of the free layer can be switched between parallel (P) and antiparallel (AP) states with respect to the reference layer by reversing the current direction.

The MTJ stack considered in this study is β-W(7.2)/CoFeB(1)/MgO/CoFeB(1.1)/W(0.3)/Co(0.6)/Ru(0.85)/Pt(0.8)/$[Co(0.8)/Pt(0.86)]_6$/Ru(6), where the numbers in parentheses are thicknesses in nanometers (nm), and the subscript represents 6 repeated layers of Co/Pt. The SOT material is doped β-W with a spin Hall angle of 0.6 and electrical resistivity $\rho_{SOT} = 160\,\mu\Omega\,cm$.[18] The MTJ critical dimension (CD) is 40 nm for 14 nm CMOS technology,[19] the pitch is 80 nm, and the SOT length is 120 nm, as shown in Fig. 1(b). The free layer is 1 nm CoFeB because such an ultrathin film displays perpendicular magnetic anisotropy (PMA).[20,21] The synthetic antiferromagnet (SAF) made of $[Co(0.8)/Pt(0.86)]_6$ is used as part of the reference layer to generate PMA and pin the magnetization of the CoFeB reference layer.[22–24] The thickness of MgO is tailored to meet the operating voltage requirements.[25–27]

We carry out micromagnetic simulations of both two-terminal SOT-MRAM and typical STT-MRAM with 40 nm diameter MTJ, finding a critical switching current (defined as P → AP switching) of 264 vs 364 μA, with 1 ns write latency and $60k_BT$ thermal stability at 350 K ambient temperature, where k_B is the Boltzmann constant. More details on the micromagnetic simulation methodology can be found in Refs. 28 and 29, and the simulation parameters are listed in Fig. S1 within the supplementary material. In two-terminal SOT-MRAM, the STT current density is ∼72% of that in the typical STT-MRAM device, increasing endurance and reducing energy consumption at the same time.

According to our micromagnetic simulations, the required switching current density is $J_{SOT} = 92\,MA/cm^2$ and $J_{STT} = 21\,MA/cm^2$ for 1 ns switching of the above two-terminal SOT-MRAM. We consider the operating voltage to be 1 V in Secs. III A–III E. The

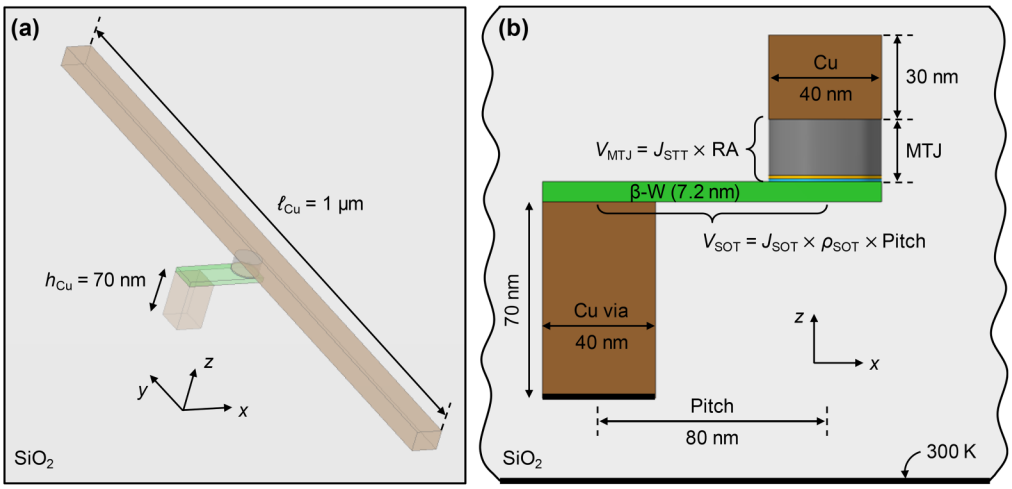

FIG. 1. (a) Two-terminal SOT-MRAM cell layout and setup of the finite-element method COMSOL model. Two-terminal SOT-MRAM is connected to a 1 μm long (ℓ_{Cu}) Cu metal bitline and a Cu via with 70 nm height (h_{Cu}), all fully encased by SiO_2 dielectric (gray region). (b) Cross section of the two-terminal SOT-MRAM, showing the dimensions of each segment. The device is encased in SiO_2, with the bottom of the Cu via and the bottom of the simulation domain set at 300 K. Pitch is defined as the distance between the center of the Cu via and that of the MTJ. The MTJ is connected in series with the β-W SOT line (green). The voltage is applied between the Cu via bottom and the top metal bitline. The voltage across the SOT line and MTJ can be calculated as shown in the figure.

voltage across the SOT line and across the MTJ can be approximated using equations in Fig. 1(b), i.e., $V_{SOT} = 0.12$ V and $V_{MTJ} = 0.88$ V, respectively. The required MTJ resistance-area product is RA $\approx 4.2\,\Omega\,\mu m^2$, which leads to a good tunneling magnetoresistance ratio (TMR) > 100%.[25–27]

B. Finite-element electro-thermal simulation model

Direct temperature measurements of nanoscale vertical devices like two-terminal SOT-MRAM pose significant challenges.[30,31] As a result, simulations become essential to estimate the peak temperature during the write pulse. In this study, we employed three-dimensional (3D) finite-element electro-thermal simulations using COMSOL Multiphysics® to model the transient temperature response of two-terminal SOT-MRAM. The device model is encased in SiO_2, as shown in Fig. 1, with the SiO_2 extending 5 μm in the x and y directions (not shown). The lower copper (Cu) via connecting to one end of the SOT line has a height of 70 nm and a width of 40 nm. The upper Cu metal bitline aligns with the MTJ, having a thickness of 30 nm and a length of 1 μm. The dimensions selected are consistent with experimental devices from Edelstein *et al.*[19] The upper metal bitline can connect to multiple two-terminal SOT-MRAM devices in practical MRAM chips. The bottom of the lower Cu via is connected to larger interconnects in MRAM chips[19,32] and it is set to thermal ground (~300 K), along with the bottom of the SiO_2 layer. We have verified that this location of thermal ground does not affect the simulated peak MTJ temperature. The simulation mesh is fine-tuned to ensure the accuracy of the results, with the mesh illustrated in Sec. S2 of the supplementary material. Considering the practical challenges in individually meshing each ultrathin layer (many under 1 nm), we lumped the reference multilayers together with an effective thermal conductivity listed in Sec. III C. This approach simplifies the mesh configuration while preserving the essential thermal characteristics. A comprehensive list of simulation parameters used can be found in Sec. S3 of the supplementary material, including the temperature dependence of thermal conductivity and the volumetric heat capacity of each material.

C. Two-terminal SOT-MRAM heating mechanism

In two-terminal SOT-MRAM, current flows through both the MTJ and the SOT line when voltage is applied, with both regions acting as heat sources. Given that the voltage across MgO is <1 V during the write process, it is expected that direct tunneling dominates over Fowler–Nordheim tunneling.[4] The tunneling process is ballistic and thus non-dissipative within the ultrathin (~1 nm) MgO; thus, heat is only generated in the receiving electrode, where injected hot electrons lose energy to the lattice after tunneling.[33–37] As a result, the temperature profile in the MTJ depends on the voltage polarity, and the heating mechanism does not lend itself to a simple analytical interpretation.

In this study, we use the same switching current density for both transitions (AP → P and P → AP), which means that the higher initial resistance in the AP → P transition requires a higher switching voltage, leading to more self-heating. (For example, the P → AP transition has ~50% of the AP → P temperature rise, for a TMR of ~100%.) Thus, we focus on the AP → P transition, wherein electrons tunnel from the reference layer to the free layer. The AP → P transition has also been identified[5,38–40] as the primary limiter of STT-MRAM endurance, leading to our decision to focus on it as the worst-case scenario.

After tunneling, we must consider the scattering of electrons in the ferromagnetic layer, which occurs with a very small inelastic mean free path (<1 nm), suggesting that heat is mainly generated in the immediate vicinity of the MgO tunnel barrier.[34,36] To simplify our simulations, we model the MTJ resistance-area product (RA) as the electrical contact resistance between MgO and the adjacent CoFeB layer at the receiving side, i.e., the CoFeB free layer. In practice, the actual temperature may be partly reduced due to the finite electron inelastic scattering mean free path in CoFeB, and future corrections can be implemented as new evidence becomes available.

III. RESULTS AND DISCUSSION

This section describes the transient temperature response of two-terminal SOT-MRAM in more detail. We highlight how the TBC between SOT-MRAM layers and the effective thermal conductivity of MTJ thin films affect the peak temperature prediction of SOT-MRAM devices. We demonstrate thermal engineering of two-terminal SOT-MRAM as well as the effect of reduced temperature on the expected endurance improvement of such devices. Lastly, we compare the thermal robustness of two-terminal SOT-MRAM with typical STT-MRAM and show advantages of the former in terms of endurance and power consumption.

A. Transient temperature response of two-terminal SOT-MRAM

We simulate the transient temperature response of two-terminal SOT-MRAM with the 1 ns write pulse described above, and the temperature distribution at the end of the pulse is illustrated in Fig. 2(a). We find that the peak temperature of the MgO layer can reach 457 °C (i.e., $\Delta T \approx 430$ K) during the write operation, with a temperature drop around 150 K at the top MgO interface. (Similarly, large temperature drops have also been noted at phase-change[41] and resistive[42] random-access memory interfaces.) The peak temperature is comparable to the temperature range for MRAM back end of line (BEOL) processing,[1,3,18,19] which highlights the need to reduce peak device temperature during operation to decelerate MgO degradation and possible Fe and oxygen diffusion.[15,16] Figure 2(b) shows the peak temperature rise in β-W, the CoFeB free layer, and MgO during a 1 ns write pulse.

The temperature rises linearly with time initially ($\propto t$) as the current pulse is applied due to rapid, nearly adiabatic, heating of the small device volume. At such short time scales, $\Delta T \approx E/C = Pt/C$, where E is the pulse energy, P is the input power (in W), and C is the device heat capacity (in J/K). After a short time, the device experiences some heat loss to its surrounding SiO_2 and the temperature rises proportionally to $t^{1/2}$, a similar temperature evolution previously observed in fast transistors.[43,44] After the heating pulse ends, we find a falling temperature time constant of $\tau_{fall} \approx 0.2$ ns. The peak temperatures of CoFeB and β-W are very close to each other due to high TBC between metal/metal interfaces. However, the peak temperature for the MgO layer is lower, partially due to a

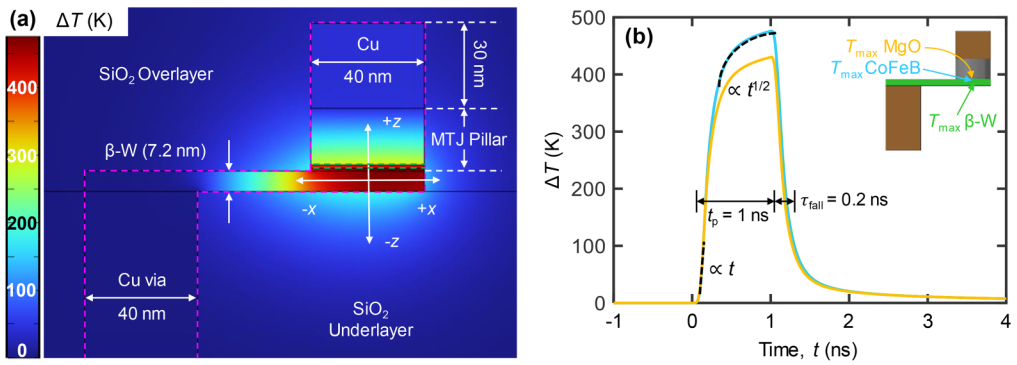

FIG. 2. (a) Temperature rise (ΔT, above ambient) of two-terminal SOT-MRAM at the end of a 1 ns write pulse, showing a peak MgO temperature of 457 °C ($\Delta T \approx 430$ K) with SiO_2 as the surrounding material. The purple dashed lines mark the interface between SOT-MRAM and surrounding materials, here with TBC = 50 MW m^{-2} K^{-1}. The green dashed lines mark the interfaces between CoFeB and MgO, with TBC values listed in Sec. S3 within the supplementary material. (b) Temperature rise (ΔT) of two-terminal SOT-MRAM during and after a write pulse, t_p = 1 ns. The dashed lines mark ΔT scaling with $\propto t$ and then with $\propto t^{1/2}$ during the applied current pulse. The temperature fall time constant is $\tau_{fall} \approx 0.2$ ns. The peak temperature rise of the β-W (green) and CoFeB (blue) are very similar and overlapping, due to high TBC between metal/metal interfaces.

lower TBC with CoFeB, and more details of this thermal interface are given in Sec. S3 of the supplementary material. The peak temperature occurs in the CoFeB layer instead of MgO because electrons tunnel (top-down) through the MgO and relax their energy in the CoFeB (see Sec. II C). However, due to reliability considerations, in the following discussion, we primarily monitor the peak temperature of the MgO layer for two-terminal SOT-MRAM.

B. Effect of thermal boundary conductance (TBC)

Many simulation studies of temperature in MRAM devices overlook the importance of thermal boundary conductance (TBC) with the surrounding materials, which can underestimate the peak device temperature during operation. The TBC and its inverse, the thermal boundary resistance (TBR = 1/TBC), cause a temperature difference across the boundary between two materials in the presence of a heat flux, $\Delta T = P''/\text{TBC}$, where P'' is the heat flux per unit area. A finite TBC (i.e., nonzero TBR) can cause temperature differences of tens or even hundreds of K across an interface.

In this section, we investigate the impact of TBCs between two-terminal SOT-MRAM and surrounding materials, represented by the purple dashed lines in Fig. 2(a). While the exact TBCs between β-W or the MTJ pillar with SiO_2 remain unidentified, general TBC values between metals and dielectrics usually lie in the range[35,45] of 50–300 MW m^{-2} K^{-1}. Furthermore, studies have shown that the TBCs between metals and insulators can vary depending on the fabrication process, interface quality, and bond strength.[45–48] Given this variability, our simulations address both extremes, TBC = 50 and 300 MW m^{-2} K^{-1}. Time-domain thermoreflectance (TDTR) measurements can be used to measure these unknown TBCs for more accurate temperature predictions of MRAM.[49]

Figure 3(a) shows the peak temperature rise during a nanosecond current pulse in two-terminal SOT-MRAM for varying TBC values: infinite TBC (i.e., TBR = 0, perfect heat flow across the boundary), TBC = 50 MW m^{-2} K^{-1} and TBC = 300 MW m^{-2} K^{-1}. A TBC of 50 MW m^{-2} K^{-1} can lead to an additional temperature rise up to ~118 K after a 1 ns pulse in the two-terminal SOT-MRAM geometry studied here. These results emphasize that numerous TBCs play a key role in determining MRAM device peak temperatures. In Secs. III C–III H, TBC refers to the TBCs between two-terminal SOT-MRAM and surrounding materials, unless otherwise specified.

C. Effect of thermal conductivity of MTJ thin films

For the MTJ stack, all the constituent materials are ultrathin films. In such films, it is known that the thermal conductivity is greatly suppressed compared to the bulk form due to increased electron and phonon boundary scattering.[45,48] For instance, one study revealed that a 180-repetition of Co(1.2 nm)/Cu(1.1 nm) multilayer had an effective cross-plane thermal conductivity[50] of ~5–7 W m^{-1} K^{-1}, while the thermal conductivities of bulk Co and Cu are ~100 and ~400 W m^{-1} K^{-1} at room temperature, respectively.[51] Therefore, in this work, we have carefully incorporated the thermal conductivity values of thin films for the consisting materials of the MTJ.

The effective cross-plane thermal conductivity of CoFeB/MgO/CoFeB has been measured to be ~0.5 W m^{-1} K^{-1}.[52] However, the effective thermal conductivity of the reference layer (k_{ref}) is unknown. Typically, the reference layer of MRAM with PMA consists of multiple layers of thin metals.[10,12,22–24] The reference layer in two-terminal SOT-MRAM considered in this study is CoFeB (1.1)/W(0.3)/Co(0.6)/Ru(0.85)/Pt(0.8)/$[Co(0.8)/Pt(0.86)]_6$/Ru(6), which consists of 16 sub-nm thin metal layers. These ultrathin metals show strongly suppressed thermal conductivities due to the increased density of interfaces and significant electron boundary scattering, because the layer thicknesses are comparable to or less than the electron mean free path.[45,48,53,54]

Hence, we consider various k_{ref} values in Fig. 3(b), and we find that the peak temperature of two-terminal SOT-MRAM could increase by up to ~202 K if k_{ref} is reduced from 15 to 5 W m^{-1} K^{-1}

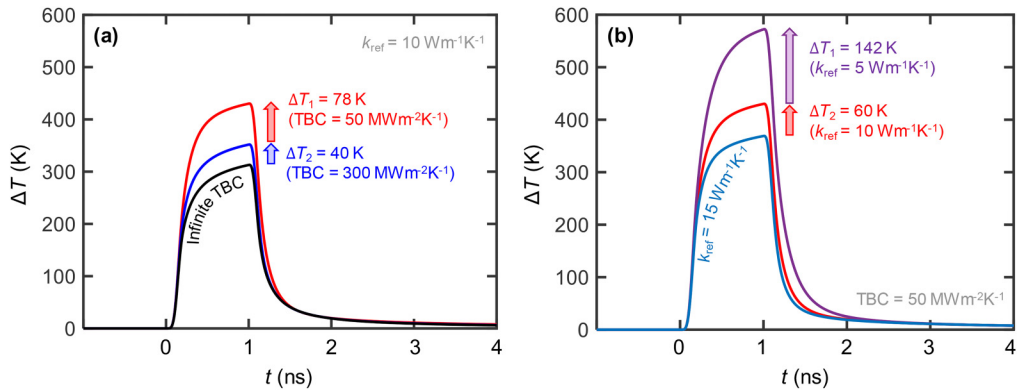

FIG. 3. (a) Transient temperature rise ΔT of two-terminal SOT-MRAM using different TBCs with surrounding materials, when k_{ref} = 10 Wm^{-1} K^{-1}. The peak ΔT rises by an additional 118 K when TBC = 50 MW m^{-2} K^{-1}, compared to when thermal interfaces are considered ideal (infinite TBC). (b) Transient ΔT of two-terminal SOT-MRAM with different k_{ref} when TBC = 50 MW m^{-2} K^{-1}. The peak ΔT could rise by an additional 202 K if k_{ref} is reduced from 15 to 5 Wm^{-1} K^{-1}.

while maintaining the TBC at 50 MW m^{-2} K^{-1}. This finding highlights the importance of incorporating accurate thermal conductivity values of reference layers when predicting the peak temperature of two-terminal SOT-MRAM. These results also demonstrate that the number of interfaces and the thickness of the reference layer have a great impact on the peak temperature of MRAM. In the following analysis, we set $k_{ref} \approx 10$ W m^{-1} K^{-1},[4,53] which provides a good estimation of the k_{ref} value. The thermal conductivity of β-W is estimated by the Wiedemann–Franz Law,[54–56] and more details can be found in Sec. S3 of the supplementary material.

D. Reducing peak temperature by material engineering

From Secs. III A–III C, we have shown the peak temperature of the MgO layer can be as high as 457 °C. Such high temperatures can accelerate the degradation of MgO (which also faces large electric fields), as well as accelerate oxygen and Fe diffusion, which can lead to MTJ failure.[15–17] Thus, the peak temperature must be reduced in order to improve endurance and make MRAM thermally robust. In this section, we examine the effect of changing the surrounding insulator from conventional SiO_2 (room temperature thermal conductivity ≈ 1.4 W m^{-1} K^{-1}) to other BEOL-compatible passivation materials with higher thermal conductivity, such as Si_3N_4[57,58] or AlN.[59,60] For example, BEOL-compatible AlN thin films can have a thermal conductivity of ∼18 W m^{-1} K^{-1} even at ∼100 nm thickness.[59] Here, we assume that the AlN thermal conductivity is isotropic, although the exact values may depend on the fabrication process, grain orientation, and defect density.[59]

We note that the dielectric constants of Si_3N_4 and AlN are higher[59,61] than for SiO_2, which may affect the ultrafast operation of MRAM through capacitance effects. Thus, here, we propose only

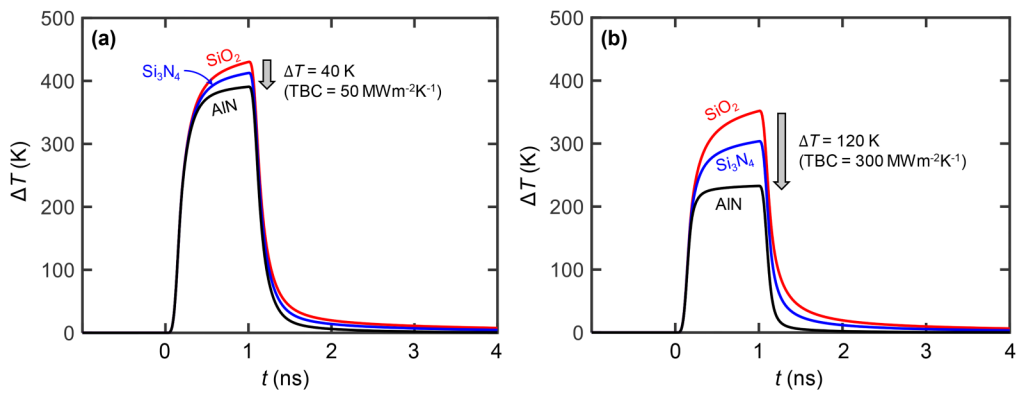

FIG. 4. Successive thermal improvement of two-terminal SOT-MRAM by replacing surrounding materials from SiO_2 to Si_3N_4 and to AlN, when the TBC between the device and surrounding materials is (a) 50 and (b) 300 MW m^{-2} K^{-1}, respectively. The higher thermal conductivity layers are more effective in reducing the peak temperature (ΔT = 120 K) when the TBC is also greater (300 MW m^{-2} K^{-1}). Here, k_{ref} = 10 Wm^{-1} K^{-1} and the current pulse width is 1 ns, similar to earlier figures unless otherwise stated.

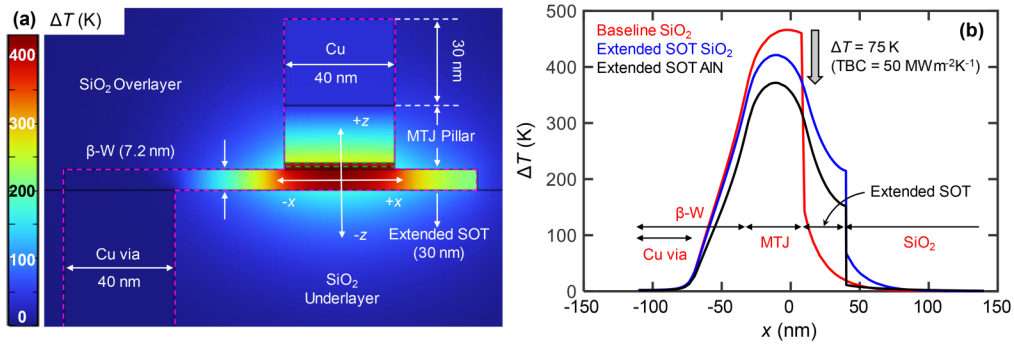

FIG. 5. (a) Temperature distribution of two-terminal SOT-MRAM with 30 nm extended SOT line, at the end of the 1 ns write pulse. Note the heat spreading into the SOT extension. (b) Temperature distribution along the SOT line x axis for three cases: the baseline with SiO_2 dielectric and no SOT extension (red), the case in (a) with SiO_2 dielectric and SOT extension (blue), and the case with AlN dielectric and SOT extension (black). TBC = 50 MW m^{-2} K^{-1} for all cases considered in this figure. The block arrow marks a 75 K temperature reduction for the case with AlN dielectric and 30 nm SOT extension. The large temperature drop where the SOT line ends (>100 K for all three cases) is due to the finite TBC with the dielectric.

changing a thin underlayer (40 nm) and overlayer (30 nm) below and above the β-W SOT line [see Fig. 1(b)] to achieve optimal heat dissipation with minimal impact on operation speed.

Figure 4(a) shows a 40 K reduction of peak MgO temperature when replacing SiO_2 with AlN, for the same TBC of 50 MW m^{-2} K^{-1}. This TBC between the SOT-MRAM device and surrounding materials has the effective thermal resistance of 28 nm SiO_2 (the so-called Kapitza length[45,48]). Because this is comparable to the device dimensions, the temperature reduction is relatively modest even when the device is surrounded by materials with higher thermal conductivity, like AlN. In contrast, Fig. 4(b) shows a 120 K reduction of peak temperature when the TBC is six times greater (300 MW m^{-2} K^{-1}) and we replace the surrounding SiO_2 with AlN. After the pulse is turned off ($t > 1$ ns), the temperature fall time constant also reduces from ~0.2 to ~0.12 ns. This indicates that the heat can dissipate faster when AlN is used as the surrounding material. We note that the TBC itself may change with the surrounding material (e.g., SiO_2 vs AlN), but the exact value is not known at the present time. Thus, our adoption of a fixed TBC here (e.g., 50 vs 300 MW m^{-2} K^{-1}) allows us to isolate the effect of the surrounding material's thermal conductivity.

These results suggest that the reduction of peak temperature will be more significant if the TBC between SOT-MRAM and its dielectric environment can be improved. Additionally, using Si_3N_4 and AlN as thin surrounding materials can effectively avoid oxygen diffusion, which has been one of the major degradation mechanisms in SOT-MRAM identified in the literature.[15,16]

E. Reducing peak temperature by SOT line extension

We introduce a novel method to lower the peak temperature of two-terminal SOT-MRAM during the write pulse. This is achieved by implementing an SOT line that extends beyond the MTJ [see Fig. 5(a)] and serves as a cooling fin, while preserving the original operating conditions. The thermal healing length (L_H) of the fin is the characteristic length of temperature decay along the fin,[62] and a fin much longer than L_H provides diminishing returns in cooling. The length of the SOT extension is thus chosen to be equal to the L_H along the β-W SOT line. This ensures optimal heat dissipation with minimal impact on areal density.

Assuming heat dissipates toward the underlying silicon substrate, the thermal healing length of the SOT line can be estimated as

$$L_H = \sqrt{\frac{k_{SOT}}{k_{ul}} d_{SOT}\left(d_{ul} + \frac{k_{ul}}{TBC}\right)},$$

where k_{SOT} and d_{SOT} are the thermal conductivity and thickness of the β-W SOT line, respectively, while k_{ul} and d_{ul} are the thermal conductivity and thickness of the underlayer beneath the SOT line, respectively. The second term in the parenthesis represents the Kapitza length,[45,48] which accounts for the TBC between the SOT and the underlayer. Typical L_H for our combination of materials and TBC range between 20 and 80 nm, as summarized in Sec. S4 of the supplementary material. We note the L_H estimated above represents an upper bound because it ignores three-dimensional heat spreading into the overlayer and the MTJ pillar. More accurate values of L_H can be obtained from numerical simulations by examining the temperature distribution along the β-W SOT line, e.g., in Fig. 2(a). However, the temperature drops (nearly) exponentially along the SOT line such that small variations in L_H are less significant. Based on these estimates, we employ a 30 nm extended SOT line in the following analysis.

Figure 5(a) illustrates the peak temperature rise of the device with an extended SOT line design, at the end of the current pulse. Using this approach, the peak temperature of two-terminal SOT-MRAM can be reduced from 457 to 422 °C, with the input power, material choices, and other thermal conditions described earlier. Remarkably, this reduction is comparable to changing the surrounding SiO_2 to AlN, as shown in Fig. 4(a), given the same TBC = 50 MW m^{-2} K^{-1}. If the extended SOT line design is *combined* with AlN (as surrounding dielectric), then the temperature reduction can be as high as 75 K. This case is shown in Fig. 5(b), which compares the temperature profile along the SOT line x axis, at the end of the current pulse, for the baseline case

(SiO_2 dielectric and no SOT extension), the case in Fig. 5(a) (SiO_2 dielectric and extended SOT), and the case with surrounding AlN (instead of SiO_2) and extended SOT.

We note that if TBC is larger, 300 MW m^{-2} K^{-1}, which enables better heat spreading into surrounding materials, the efficiency of the extended SOT line is less significant. This is particularly evident when AlN is used as the surrounding material, leading to a reduced L_H (from 73 to 25 nm, see Sec. S4 of the supplementary material) and only an ~8 K peak temperature drop. Consequently, the benefits of the extended SOT line are significant when heat spreading is limited by the thermal interfaces between SOT-MRAM and its surrounding materials.

F. Reducing peak temperature by voltage scaling

Further reduction of peak temperature can be achieved by lowering the operating voltage of the two-terminal SOT-MRAM. In Sec. II A, we showed that the critical switching current density is $J_{SOT} = 92$ MA/cm^2 and $J_{STT} = 21$ MA/cm^2 for 1 ns switching of two-terminal SOT-MRAM. Assuming the same current density, the operating voltage can be reduced by thinning down the MgO layer (i.e., reducing the MTJ RA value).[25–27] For example, reducing the MTJ RA value from ~4.2 to ~3 $\Omega \cdot \mu m^2$ lowers the operating voltage from 1 to 0.75 V, this reduction being entirely due to a lower voltage across the MTJ (from 0.88 to 0.63 V). The peak temperature of two-terminal SOT-MRAM exhibits an almost linear relationship with the total power input, decreasing from 457 °C at a 1 V operating voltage to 339 °C at 0.75 V, for the case with SiO_2 as the surrounding material and TBC of 50 MW m^{-2} K^{-1}. Lowering the operating voltage also reduces the stress across the MgO layer, reducing the odds of dielectric breakdown and improving MTJ endurance.[4–7] Several studies[63,64] also proposed reducing the MTJ RA to enhance the MTJ endurance.

We note that, while reducing the operating voltage can effectively reduce both the peak temperature and MTJ voltage stress, thereby improving endurance, the trade-off is a potential reduction of the TMR ratio. The reduced MTJ RA value of ~3 $\Omega \mu m^2$ assumed here is at the forefront of current MRAM technology,[25–27] but comes with a lower TMR ratio. The TMR ratio plays a crucial role in determining the on or off state of the device during the read operation. A reduced TMR ratio will complicate the readout process and corresponding circuits.[65,66] To address this trade-off, a novel MTJ process was developed by Ikegawa *et al.*[64] to decrease bit error rate while using a lower RA value for the MgO tunnel barrier.

G. Effect of temperature reduction on MTJ endurance improvement

In Secs. III D and III E, we showed that the peak MTJ temperature can be reduced by up to 75 K (120 K) using thermal engineering with a TBC of 50 MW m^{-2} K^{-1} (300 MW m^{-2} K^{-1}). This section explores how such temperature reduction could improve the endurance of MRAM devices. In this context, we refer to the study of Van Beek *et al.*,[17] who fabricated three-terminal SOT-MRAM with RA ≈ 4.5 $\Omega \mu m^2$, where heat generated by both STT and SOT current affected the MTJ temperature. We can draw parallels between this work and ours by thinking of the lower temperature achieved through a reduction of V_{SOT} by Van Beek *et al.*[17] having a similar effect as our thermal engineering.

For our baseline two-terminal SOT-MRAM, when TBC = 50 MW m^{-2} K^{-1}, the $V_{MTJ} = 0.88$ V with a peak temperature of 457 °C, this is equivalent to the case of $V_{STT} = 0.88$ V and $V_{SOT} \approx 0.4$ V in the study of Van Beek *et al.*,[17] with an endurance of ~10^4 cycles. With thermal engineering, we reduced the peak temperature to 382 °C, which corresponds to $V_{SOT} = \sim 0.3$ V in the work of Van Beek *et al.*,[17] or an endurance improvement to ~10^5–10^6 cycles. In other words, a temperature reduction of 75 K [such as in Fig. 5(b)] could lead to 10–100× endurance improvement. When the TBC is higher, 300 MW m^{-2} K^{-1}, and we thermally engineer the device with AlN surrounding material [as in Fig. 4(b)], the endurance improvement is expected to be a similar factor of 10–100×. This endurance improvement is partly limited by the high $V_{MTJ} = 0.88$ V.

However, when V_{MTJ} is lowered to 0.63 V (total voltage lowered to 0.75 V), as described in Sec. III F, the endurance improvement by thermal engineering is expected to be at least 1000×. This is due to the combined effects of lower V_{STT} stress[17] and temperature reduction on the endurance improvement. Overall, if a good TMR ratio can be obtained with low RA value, lowering the operating voltage from 1 to 0.75 V combined with thermal engineering could lead to significant endurance improvements.

Finally, we note that Van Beek *et al.*[17] used 10 ns stress pulses, while we expect that the 1 ns stress pulses used here would also improve endurance by at least an additional ~10× due to the reduced time the device spends at high temperature.[15] Combined these observations also suggest that additional studies are required to accurately quantify MRAM endurance improvements achievable through factors like voltage and thermal engineering by material properties, dimensions (i.e., thermal time constant), and geometry effects such as our extended SOT design.

H. Thermal robustness of SOT-MRAM relative to STT-MRAM

In Sec. II A, we have shown that, compared to typical STT-MRAM, the critical switching current can be reduced from 364 to 264 μA when two-terminal SOT-MRAM is implemented, under the same operational conditions. More details of micromagnetic simulations can be found in Sec. S1 of the supplementary material. Here, we take a closer look at the impact of this current reduction on thermal robustness and endurance improvement of two-terminal SOT-MRAM relative to STT-MRAM.

We also perform electro-thermal simulations for STT-MRAM. The required STT current density for typical STT-MRAM is ~29 MA/cm^2 for 1 ns switching with a 40 nm MTJ diameter. Here, we keep the MTJ RA value the same at 4.2 $\Omega \mu m^2$, and the operating voltage required is $V_{STT} = 1.22$ V. Typically, STT-MRAM is bottom-pinned, i.e., the reference layer is below the MgO tunnel barrier while the free layer is on top. To draw a fair comparison with two-terminal SOT-MRAM simulated in Secs. III A–III G, we simulate the same MTJ stacks and also assume a 7.2 nm capping layer above the free layer with 10 W m^{-1} K^{-1} thermal conductivity (the same as the β-W layer in two-terminal SOT-MRAM). The thickness of the capping layer on top of the free layer has a

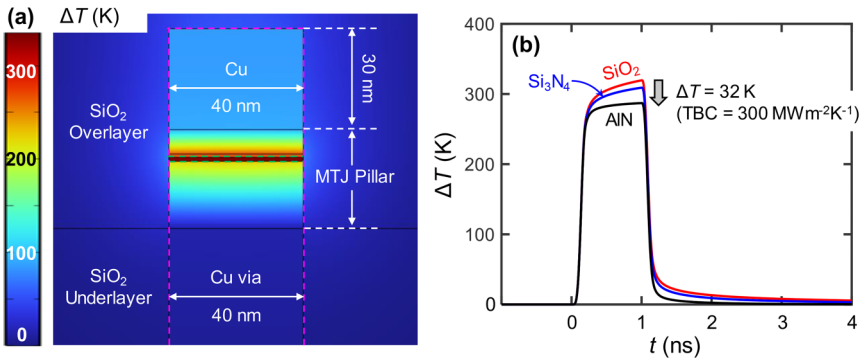

FIG. 6. (a) Temperature rise (ΔT, above ambient) distribution of STT-MRAM at the end of a 1 ns write pulse, showing a peak temperature of 346 °C with a SiO_2 surrounding material. The TBC between STT-MRAM with surrounding materials is 300 MW m^{-2} K^{-1} (interfaces marked by purple dashed lines). Horizontal green dashed lines mark the interface between CoFeB and MgO, with TBC values listed in Sec. S3 within the supplementary material. (b) Peak transient temperature rise (ΔT) of STT-MRAM during and after a 1 ns write pulse, when surrounding materials are SiO_2, Si_3N_4, or AlN, when TBC is 300 MW m^{-2} K^{-1}. Although the TBC itself may change with the surrounding material, this simulation allows us to isolate the effect of the thermal conductivity alone; here, we find this effect is small, no more than 32 K.

significant impact on the MTJ peak temperature, which agrees with previous results.[67]

Our simulations reveal that the peak temperature of STT-MRAM is 346 °C when the interface between STT-MRAM and surrounding materials has TBC = 300 MW m^{-2} K^{-1}. Our result here is comparable to the estimates of Van Beek *et al.*,[4] where the peak MgO temperature reached ∼280 °C at an ∼1.2 V operating voltage, for a 50 nm diameter MTJ with a higher RA value of 10 Ω μm^2. Quantitative comparisons across the literature present challenges due to variations in device dimensions, operating conditions, and the thermophysical properties of the MTJ layers. Although the input power is higher in STT-MRAM, the peak temperature of STT-MRAM is still lower compared to two-terminal SOT-MRAM due to direct heat dissipation through the electrodes, as shown in Fig. 6(a).

However, the high voltage stress across the MTJ (V_{STT}) leads to a much lower endurance in STT-MRAM, compared to two-terminal SOT-MRAM, where the voltage is shared between the MTJ and SOT line. Overall, the MTJ voltage stress is reduced from 1.22 V in STT-MRAM to 0.88 V in two-terminal SOT-MRAM. According to previous endurance and time-dependent dielectric breakdown (TDDB) results,[4–7] the MTJ endurance can be improved by more than a factor of 10^6 when the voltage stress reduces from ∼1.22 to ∼0.88 V. Van Beek *et al.*[15] report a significant improvement in endurance for SOT-MRAM when the voltage across the SOT is reduced to $V_{SOT} \approx 0.12$ V, suggesting a near-infinite endurance for the SOT line in our two-terminal SOT-MRAM.

Compared to two-terminal SOT-MRAM where heat spreads through both the SOT line and MTJ pillar, STT-MRAM has less area of heat dissipation due to the small area of the MTJ pillar. Thus, STT-MRAM benefits less from proposed thermal engineering solutions. Figure 6(b) shows only a 32 K reduction in peak temperature when replacing SiO_2 with AlN, for a good TBC of 300 MW m^{-2} K^{-1}. These results suggest that two-terminal SOT-MRAM is expected to have much higher endurance compared to typical STT-MRAM and will also benefit more from thermal engineering.

IV. CONCLUSION

Through simulations rooted in realistic considerations (e.g., thermal boundary conductance), we have shown that the peak temperature of two-terminal SOT-MRAM can reach 350–450 °C at the end of 1 ns write pulse (depending on thermal properties, including thermal conductivity and TBC). Such high temperatures will accelerate the MgO degradation and cause endurance issues, indicating the necessity of reducing peak temperature for endurance improvement.

We propose two methods to reduce the peak temperature during the write pulse. The first method involves using an extended SOT line as a heat spreader, while the second method replaces the surrounding dielectric with AlN, which has higher thermal conductivity. These work best when heat dissipation from two-terminal SOT-MRAM to its surrounding is limited by relatively low TBC and, combined, they can reduce the peak temperature by ∼75 K. If the TBC is relatively high, then using AlN as the surrounding material will reduce peak temperatures by ∼120 K. We further discuss the effect of reducing operating voltage on endurance improvement for two-terminal SOT-MRAM. Combined with our thermal engineering solution, reducing the operating voltage to 0.75 V could improve endurance by at least 1000×.

Lastly, we compare the thermal robustness and endurance of two-terminal SOT-MRAM with STT-MRAM and show that the former enables better control of temperature and significantly improved endurance due to lower voltage stress across the MTJ, partly enabled by voltage sharing with the SOT line. This study sheds light on thermal management methods for two-terminal SOT-MRAM devices and provides important guidelines for the thermal optimization of MRAM.

SUPPLEMENTARY MATERIAL

See the supplementary material for micromagnetic simulation parameters and results (Sec. S1), finite-element simulation mesh (Sec. S2), electro-thermal simulation parameters (Sec. S3), and thermal healing length calculations (Sec. S4).

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation and industry partners under Grant No. 2328804 in the Future of Semiconductors (FuSe) program, Grant No. 2345655 in the Partnerships for Innovation (PFI) program, Grant No. 2314591 in the Advanced Chip Engineering Design and Fabrication (ACED Fab) program, and in part by the Stanford Nonvolatile Memory Technology Research Initiative (NMTRI) Affiliates program. Ç.K. and E.P. also acknowledge partial support from SUPREME, one of seven JUMP 2.0 centers sponsored by the Semiconductor Research Corporation (SRC) and DARPA.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Haotian Su: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Funding acquisition (supporting); Investigation (lead); Methodology (lead); Software (lead); Validation (lead); Visualization (equal); Writing – original draft (lead); Writing – review & editing (equal). **Heungdong Kwon:** Conceptualization (supporting); Formal analysis (supporting); Methodology (supporting); Software (supporting); Validation (supporting); Visualization (supporting). **William Hwang:** Conceptualization (supporting); Formal analysis (supporting); Methodology (supporting); Software (supporting); Writing – review & editing (supporting). **Fen Xue:** Conceptualization (supporting); Formal analysis (supporting); Investigation (supporting); Visualization (supporting); Writing – review & editing (supporting). **Çağıl Köroğlu** Formal analysis (supporting); Methodology (supporting); Software (supporting); Writing – review & editing (supporting). **Wilman Tsai:** Conceptualization (supporting); Funding acquisition (supporting); Methodology (supporting); Project administration (supporting); Supervision (supporting). **Mehdi Asheghi:** Conceptualization (supporting); Formal analysis (supporting); Funding acquisition (supporting); Project administration (supporting); Supervision (supporting). **Kenneth E. Goodson:** Conceptualization (supporting); Funding acquisition (supporting); Project administration (supporting); Supervision (supporting). **Shan X. Wang:** Conceptualization (supporting); Formal analysis (supporting); Funding acquisition (equal); Investigation (supporting); Project administration (supporting); Supervision (supporting); Validation (supporting); Visualization (supporting); Writing – review & editing (supporting). **Eric Pop:** Conceptualization (supporting); Formal analysis (supporting); Funding acquisition (equal); Methodology (supporting); Project administration (supporting); Software (supporting); Supervision (lead); Validation (supporting); Visualization (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]D. C. Worledge, in *2022 14th IEEE International Memory Workshop (IMW)* (IEEE, 2022).
[2]O. Golonzka, J. G. Alzate, U. Arslan, M. Bohr, P. Bai, J. Brockman, B. Buford, C. Connor, N. Das, B. Doyle, T. Ghani, F. Hamzaoglu, P. Heil, P. Hentges, R. Jahan, D. Kencke, B. Lin, M. Lu, M. Mainuddin, M. Meterelliyoz, P. Nguyen, D. Nikonov, K. O'brien, J. ODonnell, K. Oguz, D. Ouellette, J. Park, J. Pellegren, C. Puls, P. Quintero, T. Rahman, A. Romang, M. Sekhar, A. Selarka, M. Seth, A. J. Smith, A. K. Smith, L. Wei, C. Wiegand, Z. Zhang, and K. Fischer, in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2018), pp. 18.1.1–18.1.4.
[3]G. Hu, M. G. Gottwald, Q. He, J. H. Park, G. Lauer, J. J. Nowak, S. L. Brown, B. Doris, D. Edelstein, E. R. Evarts, P. Hashemi, B. Khan, Y. H. Kim, C. Kothandaraman, N. Marchack, E. J. O'Sullivan, M. Reuter, R. P. Robertazzi, J. Z. Sun, T. Suwannasiri, P. L. Trouilloud, Y. Zhu, and D. C. Worledge, in *2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2017), pp. 38.3.1–38.3.4.
[4]S. Van Beek, B. J. O'Sullivan, P. J. Roussel, R. Degraeve, E. Bury, J. Swerts, S. Couet, L. Souriau, S. Kundu, S. Rao, W. Kim, F. Yasin, D. Crotti, D. Linten, and G. Kar, in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2018), pp. 25.2.1–25.2.4.
[5]X. Zhang, G. J. Zhang, L. J. Shen, P. P. Yu, and Y. F. Jiang, Solid State Electron. **173**, 107878 (2020).
[6]K. Hosotani, M. Nagamine, T. Ueda, H. Aikawa, S. Ikegawa, Y. Asao, H. Yoda, and A. Nitayama, Jpn. J. Appl. Phys. **49**, 04DD15 (2010).
[7]Y. Wang, H. Cai, L. A. D. Naviner, Y. Zhang, X. X. Zhao, E. Deng, J. O. Klein, and W. S. Zhao, IEEE Trans. Electron Devices **63**, 1762–1767 (2016).
[8]S. Couet, S. Rao, S. Van Beek, V. D. Nguyen, K. Garello, V. Kateel, G. Jayakumar, J. D. Costa, K. Cai, F. Yasin, D. Crotti, and G. S. Kar, in *2021 Symposium on VLSI Technology* (IEEE, 2021), pp. 1–2.
[9]K. Garello, F. Yasin, S. Couet, L. Souriau, J. Swerts, S. Rao, S. Van Beek, W. Kim, E. Liu, S. Kundu, D. Tsvetanova, N. Jossart, K. Croes, E. Grimaldi, M. Baumgartner, D. Crotti, A. Funiemont, P. Gambardella, and G. S. Kar, in *2018 IEEE Symposium on VLSI Circuits* (IEEE, 2018).
[10]E. Grimaldi, V. Krizakova, G. Sala, F. Yasin, S. Couet, G. S. Kar, K. Garello, and P. Gambardella, Nat. Nanotechnol. **15**, 111–117 (2020).
[11]J. Doevenspeck, K. Garello, B. Verhoef, R. Degraeve, S. Van Beek, D. Crotti, F. Yasin, S. Couet, G. Jayakumar, I. A. Papistas, P. Debacker, R. Lauwereins, W. Dehaene, G. S. Kar, S. Cosemans, A. Mallik, and D. Verkest, in *2020 IEEE Symposium on VLSI Technology* (IEEE, 2020).
[12]N. Sato, F. Xue, R. M. White, C. Bi, and S. X. Wang, Nat. Electron. **1**, 508–511 (2018).
[13]G. Q. Yu, Nat. Electron. **1**, 496–497 (2018).
[14]W. Hwang, F. Xue, W. Tsai, and S. X. Wang, in *2023 IEEE 34th Magnetic Recording Conference (TMRC)* (IEEE, 2023).
[15]S. Van Beek, K. Cai, S. Rao, G. Jayakumar, S. Couet, N. Jossart, A. Chasin, and G. S. Kar, in *2022 IEEE International Reliability Physics Symposium (IRPS)* (IEEE, 2022).
[16]S. Van Beek, K. M. Cai, K. Q. Fan, G. Talmelli, A. Trovato, N. Jossart, S. Rao, A. Chasin, and S. Couet, in *2023 IEEE International Reliability Physics Symposium (IRPS)* (IEEE, 2023).
[17]S. V. Beek, V. Kateel, K. Cai, N. Jossart, S. Rao, and S. Couet, in *2023 IEEE International Magnetic Conference—Short Papers (INTERMAG Short Papers)* (IEEE, 2023).

[18]M. Y. Song, C. M. Lee, S. Y. Yang, G. L. Chen, K. M. Chen, I. J. Wang, Y. C. Hsin, K. T. Chang, C. F. Hsu, S. H. Li, J. H. Wei, T. Y. Lee, M. F. Chang, X. Y. Bao, C. H. Diaz, and S. J. Lin, in *2022 IEEE Symposium on VLSI Technology and Circuits* (IEEE, 2022).
[19]D. Edelstein, M. Rizzolo, D. Sil, A. Dutta, J. DeBrosse, M. Wordeman, A. Arceo, I. C. Chu, J. Demarest, E. R. J. Edwards, E. R. Evarts, J. Fullam, A. Gasasira, G. Hu, M. Iwatake, R. Johnson, V. Katragadda, T. Levin, J. Li, Y. Liu, C. Long, T. Maffitt, S. McDermott, S. Mehta, V. Mehta, D. Metzler, J. Morillo, Y. Nakamura, S. Nguyen, P. Nieves, V. Pai, R. Patlolla, R. Pujari, R. Southwick, T. Standaert, O. van der Straten, H. Wu, C. C. Yang, D. Houssameddine, J. M. Slaughter, and D. C. Worledge, in *2020 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2020), pp. 11.5.1–11.5.4.
[20]D. C. Worledge, G. Hu, D. W. Abraham, J. Z. Sun, P. L. Trouilloud, J. Nowak, S. Brown, M. C. Gaidis, E. J. O'Sullivan, and R. P. Robertazzi, Appl. Phys. Lett. **98**, 022501 (2011).
[21]S. Ikeda, K. Miura, H. Yamamoto, K. Mizunuma, H. D. Gan, M. Endo, S. Kanai, J. Hayakawa, F. Matsukura, and H. Ohno, Nat. Mater. **9**, 721–724 (2010).
[22]M. X. Wang, W. L. Cai, D. Q. Zhu, Z. H. Wang, J. Kan, Z. Y. Zhao, K. H. Cao, Z. L. Wang, Y. G. Zhang, T. R. Zhang, C. Park, J. P. Wang, A. Fert, and W. S. Zhao, Nat. Electron. **1**, 582–588 (2018).
[23]S. Couet, T. Devolder, J. Swerts, S. Mertens, T. Lin, E. Liu, S. Van Elshocht, and G. S. Kar, Appl. Phys. Lett. **111**, 152406 (2017).
[24]S. E. Lee, T. H. Shim, and J. G. Park, NPG Asia Mater. **8**, e324 (2016).
[25]M. M. C. Souza, R. C. Sousa, C. Ducruet, S. Auffret, and B. Dieny, J. Appl. Phys. **107**, 09C702 (2010).
[26]K. Tsunekawa, D. D. Djayaprawira, M. Nagai, H. Maehara, S. Yamagata, N. Watanabe, S. Yuasa, Y. Suzuki, and K. Ando, Appl. Phys. Lett. **87**, 072503 (2005).
[27]S. Ikeda, J. Hayakawa, Y. M. Lee, R. Sasaki, T. Meguro, F. Matsukura, and H. Ohno, Jpn. J. Appl. Phys. **44**, L1442 (2005).
[28]W. Hwang, F. Xue, F. Zhang, M. Y. Song, C. M. Lee, E. Turgut, T. C. Chen, X. Y. Bao, W. L. Tsai, D. L. Fan, and S. X. Wang, IEEE Trans. Magn. **59**, 3400106 (2023).
[29]A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. Van Waeyenberge, AIP Adv. **4**, 107133 (2014).
[30]S. Deshmukh, M. M. Rojo, E. Yalon, S. Vaziri, C. Koroglu, R. Islam, R. A. Iglesias, K. Saraswat, and E. Pop, Sci. Adv. **8**, eabk1514 (2022).
[31]E. Yalon, S. Deshmukh, M. Muñoz Rojo, F. F. Lian, C. M. Neumann, F. Xiong, and E. Pop, Sci. Rep. **7**, 15360 (2017).
[32]P. H. Lee, C. F. Lee, Y. C. Shih, H. J. Lin, Y. A. Chang, C. H. Lu, Y. L. Chen, C. P. Lo, C. C. Chen, C. H. Kuo, T. L. Chou, C. Y. Wang, J. J. Wu, R. Wang, H. Chuang, Y. Wang, Y. D. Chih, and T. Y. J. Chang, in *2023 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2023), pp. 494–496.
[33]R. C. Sousa, I. L. Prejbeanu, D. Stanescu, B. Rodmacq, O. Redon, B. Dieny, J. G. Wang, and P. P. Freitas, J. Appl. Phys. **95**, 6783–6785 (2004).
[34]E. Gapihan, J. Hérault, R. C. Sousa, Y. Dahmane, B. Dieny, L. Vila, I. L. Prejbeanu, C. Ducruet, C. Portemont, K. Mackay, and J. P. Nozières, Appl. Phys. Lett. **100**, 2002410 (2012).
[35]E. Pop, Nano Res. **3**, 147–169 (2010).
[36]J. H. NamKoong and S. H. Lim, J. Phys. D: Appl. Phys. **42**, 225003 (2009).
[37]R. Lake and S. Datta, Phys. Rev. B **46**, 4757 (1992).
[38]J. J. Kan, C. Park, C. Ching, J. Ahn, Y. Xie, M. Pakala, and S. H. Kang, IEEE Trans. Electron Devices **64**, 3639 (2017).
[39]J. J. Kan, C. Park, C. Ching, J. Ahn, L. Xue, R. Wang, A. Kontos, S. Liang, M. Bangar, H. Chen, S. Hassan, S. Kim, M. Pakala, and S. H. Kang, in *2016 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2016), pp. 27.4.1–27.4.4.
[40]C. Yoshida and T. Sugii, in *2012 IEEE International Reliability Physics Symposium (IRPS)* (IEEE, 2012),
[41]A. I. Khan, A. Daus, R. Islam, K. M. Neilson, H. R. Lee, H. P. Wong, and E. Pop, Science **373**, 1243 (2021).
[42]A. Sarantopoulos, K. Lange, F. Rivadulla, S. Menzel, and R. Dittmann, arXiv:2402.07603 (2024).
[43]V. M. Dwyer, A. J. Franklin, and D. S. Campbell, Solid State Electron. **33**, 553–560 (1990).
[44]R. C. Joy and E. S. Schlig, IEEE Trans. Electron Devices **17**, 586–594 (1970).
[45]C. Monachon, L. Weber, and C. Dames, Annu. Rev. Mater. Res. **46**, 433–463 (2016).
[46]P. E. Hopkins, ISRN Mech. Eng. **2013**, 682586 (2013).
[47]C. Monachon, M. Hojeij, and L. Weber, Appl. Phys. Lett. **98**, 091905 (2011).
[48]D. G. Cahill, P. V. Braun, G. Chen, D. R. Clarke, S. H. Fan, K. E. Goodson, P. Keblinski, W. P. King, G. D. Mahan, A. Majumdar, H. J. Maris, S. R. Phillpot, E. Pop, and L. Shi, Appl. Phys. Rev. **1**, 011305 (2014).
[49]P. Q. Jiang, X. Qian, and R. G. Yang, J. Appl. Phys. **124**, 161103 (2018).
[50]M. A. M. Gijs, S. K. J. Lenczowski, R. J. M. van de Veerdonk, J. B. Giesbers, M. T. Johnson, and J. B. F. aan de Stegge, Phys. Rev. B **50**, 16733 (1994).
[51]C. Y. Ho, R. W. Powell, and P. E. Liley, J. Phys. Chem. Ref. Data **1**, 279–421 (1972).
[52]H. Jang, L. Marnitz, T. Huebner, J. Kimling, T. Kuschel, and D. G. Cahill, Phys. Rev. Appl. **13**, 024007 (2020).
[53]I. E. T. Iben, in *2016 38th Electrical Overstress/Electrostatic Discharge Symposium (EOS/ESD)* (IEEE, 2016),
[54]R. B. Wilson and D. G. Cahill, Phys. Rev. Lett. **108**, 255901 (2012).
[55]M. Jonson and G. D. Mahan, Phys. Rev. B **21**, 4223 (1980).
[56]G. D. Mahan and M. Bartkowiak, Appl. Phys. Lett. **74**, 953–954 (1999).
[57]H. Ftouni, C. Blanc, D. Tainoff, A. D. Fefferman, M. Defoort, K. J. Lulla, J. Richard, E. Collin, and O. Bourgeois, Phys. Rev. B **92**, 125439 (2015).
[58]P. Eriksson, J. Y. Andersson, and G. Stemme, J. Microelectromech. Syst. **6**, 55–61 (1997).
[59]C. Perez, A. J. Mcleod, M. E. Chen, S. I. Yi, S. Vaziri, R. Hood, S. T. Ueda, X. Y. Bao, M. Asheghi, W. Park, A. A. Talin, S. Kumar, E. Pop, A. C. Kummel, and K. E. Goodson, ACS Nano **17**, 21240–21250 (2023).
[60]R. L. Xu, M. Muñoz Rojo, S. M. Islam, A. Sood, B. Vareskic, A. Katre, N. Mingo, K. E. Goodson, H. G. Xing, D. Jena, and E. Pop, J. Appl. Phys. **126**, 185105 (2019).
[61]S. M. Hu, J. Electrochem. Soc. **113**, 693 (1966).
[62]M. Asheghi, M. N. Touzelbaev, K. E. Goodson, Y. K. Leung, and S. S. Wong, J. Heat Transfer **120**, 30–36 (1998).
[63]G. Jan, L. Thomas, S. Le, Y.-J. Lee, H. Liu, J. Zhu, J. Iwata-Harms, S. Patel, R.-Y. Tong, V. Sundar, S. Serrano-Guisan, D. Shen, R. He, J. Haq, Z. J. Teng, V. Lam, Y. Yang, Y.-J. Wang, T. Zhong, H. Fukuzawa, and P.-K. Wang, in *2018 IEEE Symposium on VLSI Technology* (IEEE, 2018).
[64]S. Ikegawa, K. Nagel, F. B. Mancoff, S. M. Alam, M. Arora, M. DeHerrera, H. K. Lee, S. Mukherjee, G. Shimon, J. J. Sun, I. Rahman, F. Neumeyer, H. Y. Chou, C. Tan, A. Shah, and S. Aggarwal, in *2022 International Electron Devices Meeting (IEDM)* (IEEE, 2022), pp. 10.4.1–10.4.4.
[65]T. Na, S. H. Kang, and S.-O. Jung, IEEE Trans. Circuits Syst. II: Exp. Briefs **68**, 12 (2021).
[66]J. DeBrosse, T. Maffitt, Y. Nakamura, G. Jan, and P. K. Wang, in *2015 IEEE Custom Integrated Circuits Conference (CICC)* (IEEE, 2015).
[67]B. Sikder and M. Z. Baten, in *2022 IEEE International Conference on Emerging Electronics (ICEE)* (IEEE, 2022).