Nano Research

Open Access

Research Article

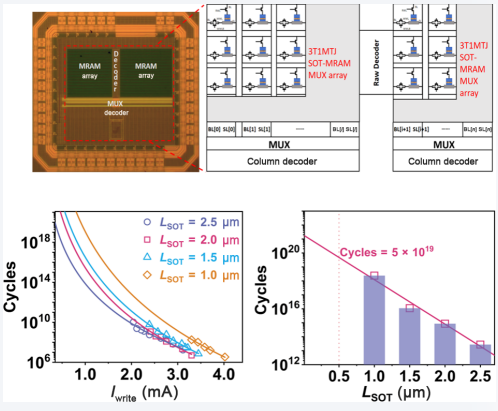
Unveiling the endurance limit of SOT-MRAM with CoFeB/MgO/CoFeB junctions: From elemental interdiffusion to array-level reliability

Chuanpeng Jiang[1,2,§], Jinhao Li[1,2,§], Chao Wang[4], Shiyang Lu[1], Xingyu Yao[1], Wenlong Cai[1], Danrong Xiong[3], Xiaofei Fan[3], Hong-Xi Liu[3] ✉, Gefei Wang[3], He Zhang[1], Kaihua Cao[1,2] ✉, Zhaohao Wang[1,4] ✉, and Weisheng Zhao[1]

[1] *School of Integrated Circuit Science and Engineering, Beihang University, Beijing 10083, China*
[2] *Integrated Circuit and Intelligent Instruments Innovation Center, Qingdao Research Institute, Beihang University, Qingdao 266100, China*
[3] *Truth Memory Corporation, Beijing 10089, China*
[4] *State Key Laboratory of Spintronics, Hangzhou International Innovation Institute, Beihang University, Hangzhou 311115, China*
[§] *Chuanpeng Jiang and Jinhao Li contributed equally to this work.*

Cite this article: *Nano Research*, **2026**, *19*, 94908447. https://doi.org/10.26599/NR.2026.94908447

ABSTRACT: Spin-orbit torque magnetic random access memory (SOT-MRAM), which is theoretically predicted with "unlimited" endurance, has been regarded as a promising alternative to conventional static random access memory (SRAM)-based L1 to L3 caches in advanced complementary metal-oxide-semiconductor (CMOS) technology nodes. However, contrary to the predicted "unlimited" endurance, the experimental test reveals that practical SOT-MRAM fails upon reaching an endurance limit, and the underlying mechanism remains unclear. We also lack information on how these limits depend on the materials or structures used, particularly at the array level. In this work, we conducted a thorough investigation of the endurance performance of SOT-MRAM arrays through cycling tests under various voltages and thermal stress conditions. Experimental results show that SOT-MRAM failures mainly happen due to thermal interdiffusion in the magnetic tunnel junctions (MTJs). This process is largely caused by Joule heating over 180 K in the SOT channel during write operations. This heat generation creates a fundamental limitation on achieving "unlimited" endurance in practical implementations. To enhance the endurance of SOT-MRAM, we proposed several optimization strategies from the perspectives of material and structure engineering, including optimizing the SOT channel material and reducing the SOT channel length. Through these optimization measures, we successfully improved the endurance of SOT-MRAM to over 10^{18} cycles at 125 °C, meeting the requirements for "unlimited" operation (≥ 10^{17} cycles) in edge computing applications. This provides design guidelines for realizing SOT-MRAM's potential as a candidate to SRAM in high-performance cache hierarchies.



KEYWORDS: SOT-MRAM array, endurance, accelerated stress test, elemental interdiffusion, power law model

1 Introduction

Artificial intelligence (AI) is increasingly influencing a wide range of application domains. This trend necessitates enhanced computational capabilities at the edge of the network to process vast amounts of data with reduced latency [1, 2]. The transition from centralized cloud-based algorithm training to edge-based inference demands significant improvements in computational performance, particularly in terms of energy efficiency, latency, and endurance [3]. Among existing silicon-based complementary metal-oxide-semiconductor (Si-CMOS) technologies and emerging devices [4–7], magnetic random access memory (MRAM) emerges as a promising candidate to meet the stringent requirements of multiple application scenarios [8, 9], as shown in Fig. 1(a). Spin-transfer torque-magnetic random access memory (STT-MRAM) offers

Received: October 15, 2025; **Revised:** January 14, 2026
Accepted: January 15, 2026

✉Address correspondence to Zhaohao Wang, zhaohao.wang@buaa.edu.cn; Kaihua Cao, kaihua.cao@buaa.edu.cn; Hong-Xi Liu, hongxi_liu@tmc-bj.cn

several advantages, including fast write and read speeds on the order of tens of nanoseconds [10], low power consumption of 10 pJ/bit [11], and compatibility with advanced CMOS nodes below 28 nm [12]. These characteristics have facilitated the integration of STT-MRAM into system-on-chip (SoC) designs [13, 14] and its deployment in various wearable devices and automotive applications [15, 16]. However, the endurance of STT-MRAM presents a critical concern. The high write current associated with its fast write speed can lead to fatigue in the MgO tunneling barrier of the two-terminal magnetic tunnel junctions (MTJs) [17, 18]. This fatigue limits the write cycles to a range between 10^6 and 10^9, restricting the use of STT-MRAM as working memory and last level cache [19]. To tackle this challenge, three-terminal spin-orbit torque (SOT) MRAM has been proposed [20, 21]. Figure 1(b) shows the write current flows through the SOT channel rather than MTJ, which could result in unlimited endurance in principle [22, 23]. This innovation holds promise for significantly extending the operational life of MRAM, enhancing its feasibility for use as working memory in edge computing. Semiconductor foundries and research institutes like TSMC [24, 25], IMEC [26, 27], Tohoku University [28, 29], and Beihang University [30, 31] have allocated resources to development in SOT-MRAM.

However, some studies have challenged the theoretical assumption of "unlimited" endurance in SOT-MRAM. Initial investigations by Simon et al. [32–34] showed that SOT-MTJs exhibit a maximum endurance of 10^8 cycles under 100 ns write pulses. Failure analysis revealed that oxygen diffusion from the substrate to the SOT-MTJs, driven by Joule heating in the SOT channel rather than electric field-induced dielectric breakdown, is the dominant degradation mechanism. The work by Shiokawa et al. [35] on 8 × 8 SOT-MRAM arrays identified electromigration-induced void formation in SOT channels under high-current cycling conditions. However, several issues remain unclear in the above-mentioned studies. Firstly, there is a controversy about the failure mechanism of SOT-MRAM. Different studies attribute the failure of SOT-MRAM to thermal activation diffusion [32–34] and electromigration [35], but the specific mechanism still needs to be further investigated. Secondly, the failure model lacks mathematical verification. Currently, the failure time fitting model of SOT-MRAM is only derived empirically and has not been fully verified by experimental data. Thirdly, the parameter dependence of SOT-MRAM endurance has not been quantified. The impacts of factors such as the SOT channel materials, the SOT channel length, and the SOT-MRAM structure on the endurance have still not been quantitatively analyzed. Fourthly, the existing study primarily focuses on the isolated SOT-MTJs. The study on endurance of large-scale SOT-MRAM arrays remains lacking. Additionally, for scenarios where SOT-MRAM has potential applications, such as in the automotive and aerospace industries, its endurance performance under high-temperature environments has not been studied yet.

In this work, we present a comprehensive array-level investigation into the endurance of 2 Kb SOT-MRAM arrays under combined voltage and thermal stress, and propose corresponding device- and array-level optimization strategies. By integrating accelerated stress tests with multiphysics simulations, we identify that the dominant failure mechanism in SOT-MRAM arrays is thermally-induced elemental interdiffusion within the MTJ stack, driven by Joule heating in the SOT channel, a mechanism that differs from previously reported oxygen diffusion or electromigration. Through the mechanism-based statistical modeling, we further establish that the power law model, rather than conventional *E*-model, provides a more accurate description and extrapolation of endurance failure, predicting an endurance of ~ 10^{14} cycles for standard devices at 125 °C. Moreover, we demonstrate that the material engineering combined with structural scaling can enhance the endurance by four orders of magnitude, achieving over 10^{18} cycles at 125 °C, a performance level that meets the "unlimited" endurance target for edge computing caches. This work clarifies the high temperature reliability limits of SOT-MRAM and provides actionable design guidelines for its implementation in industrial memory hierarchies.

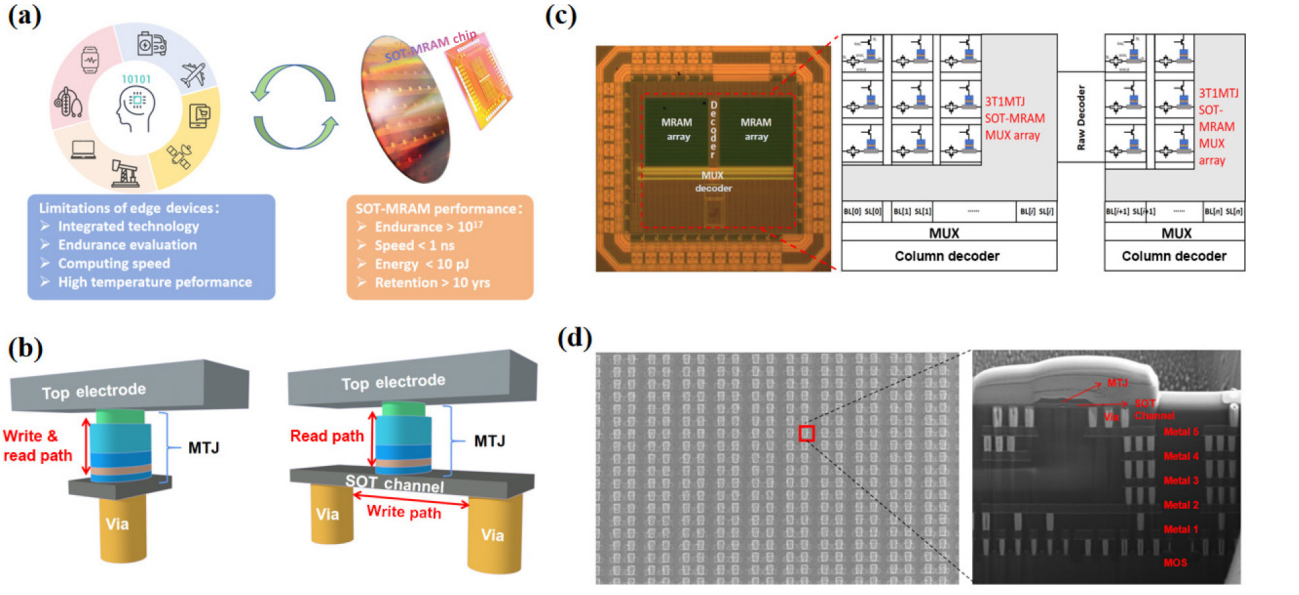


Figure 1 (a) Various application fields of AI and the excellent performance required by SOT-MRAM as an edge computing device. (b) Typical structures of STT-MRAM and SOT-MRAM. In STT-MRAM, both the read and write currents pass through the MTJ, whereas in SOT-MRAM, the read and write paths are separated. (c) Depicting a MUX-array diagram of the SOT-MRAM, comprising MTJ array, row/column decoders, and MUX circuit. (d) The scanning electron microscopy image of SOT array and TEM image of SOT-MTJ.

2 Experimental

The manufacturing of SOT-MRAM arrays was implemented based on the 8-inch process platform [36–38]. Figure 1(c) illustrates schematic diagram of the 2 Kb SOT-MRAM multiplexer (MUX) array. The MTJ served as the data storage element, featuring a designed critical dimension (CD) of 350 nm × 1050 nm, which was patterned into an elliptical shape. The process-of-record (POR) SOT channel was etched to a dumbbell shape with a width of 1200 nm and a length of 2500 nm. The transmission electron microscopy (TEM) image of the SOT unit is shown in Fig. 1(d), where MTJ is deposited between Metal 5 and the deposited layer. The detailed descriptions of SOT-MRAM array are listed in Section S1 in the Electronic Supplementary Material (ESM).

The electrical properties of SOT-MRAM arrays over a wide temperature range from −40 to 125 °C were first characterized, where the temperature was controlled and stabilized using a thermally-regulated probe station chuck. The electrical properties provide an experimental basis for determining the voltage parameters in accelerated stress tests. Figure 2(a) shows the cumulative distribution function (CDF) plot of switching voltage (V_S) over a wide temperature range from −40 to 125 °C. The V_S values were the average switching voltages for the P-to-AP and AP-to-P directions of each SOT-MTJ, where P and AP corresponded to low and high resistance states (R_P and R_{AP}), respectively. The CDF plot was normalized to the median V_S value, referred to as the critical switching voltage (V_C) at room temperature (RT). There was a change of 30% in V_C across the temperature ranges. Accordingly, the test voltage (V_t) was set to 1.8–2.5 V to ensure that V_t always remained higher than the operating voltage ($V_{OP} = V_C + 3\sigma$), even under extreme temperature conditions at 125 °C.

Next, the metrology of endurance tests was described. Figure 2(b) illustrates the constant voltage stress (CVS) pattern schematically used in the tests. For the test waveform, there are two distinct sets of operations: the write operation and the read operation. Each write cycle consists of two pulses: One is in the positive direction to write the MTJ from AP to P state; and the other is in the reverse direction, re-writing the SOT-MTJs back to the original AP state. The write pulse width is denoted as t_{write}, while the delay between two pulses is referred to as t_{delay}. One complete

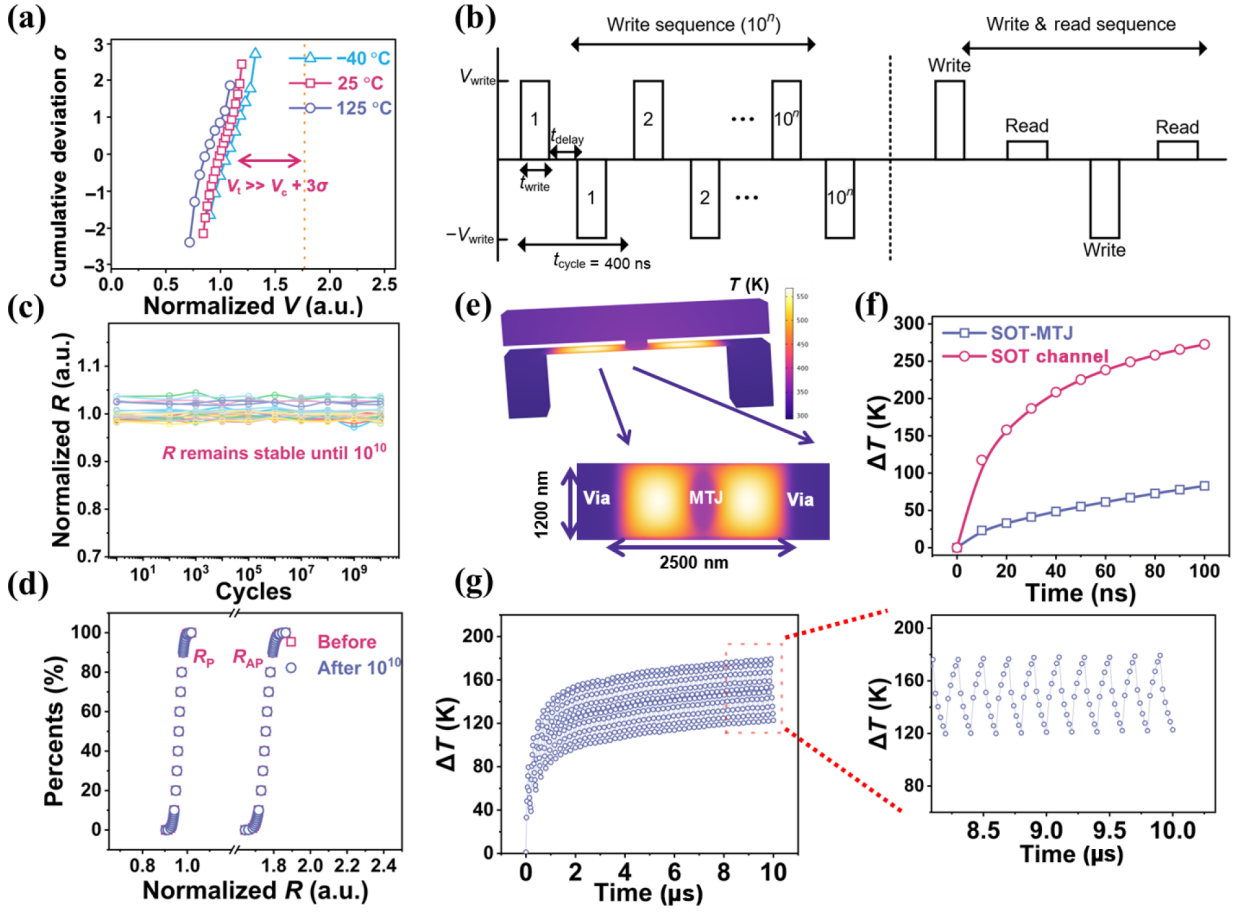


Figure 2 (a) Distribution of V_S in the SOT-MRAM array at various temperatures. The voltage is normalized by the value of the V_C at RT. The dashed line suggests the stress voltage V_t applied for the cycling tests and $V_t > V_C + 3\sigma$ at RT. (b) Waveform diagram applied in the cycling tests. Each cycle consists of a set of positive and negative pulses. After each 10^6 write cycles, the resistance of the SOT-MTJs is recorded under a 100 mV read bias across the SOT-MTJs. (c) R_P as a function of writing cycling counts at RT. Each color line-dot represents one SOT-MTJ of the total 50 measured bits. R_P is normalized by the median value. (d) The cumulative deviations of R_P and R_{AP} were measured at the initial state and after 10^{10} write cycles. R_P and R_{AP} are normalized by the median value of R_P at the initial state. (e) The thermal profile of the SOT-MTJ under V_t with a 100-ns-width pulse. The color bar represents the temperature ranging from 293 to 560 K. The ΔT at a region 300 nm away from the SOT-MTJ exceeds 270 K, while the ΔT near the SOT-MTJ is approximately 80 K. (f) The temporal temperature evolution near the MTJ. The squares and circles are simulation results, and the lines are fitting curves. The simulation results reveal a rapid temperature rise during the initial 20 ns, followed by a more gradual increase from 20 to 100 ns. (g) Simulation of the temperature evolution curve reveals that the SOT-MTJ temperature exhibits a cumulative increase with successive pulse applications, reaching thermal equilibrium at approximately 10 μs with a ΔT of ~ 180 K.

write cycle lasts for 400 ns. The t_{write} and t_{delay} were both set to 100 ns with a duty ratio of 50% under POR test conditions. After reaching each 10^6 cycles, analog read operations were performed to check the resistance of SOT-MTJs. The failure criterion was set when the tunneling magnetoresistance (TMR = $(R_{AP} - R_P)/R_P \times 100\%$) changed by 20% of the initial value.

3 Results and discussion

3.1 Endurance under voltage stress at room temperature

We first describe the cycling test results at room temperature. Figure 2(c) shows R_P measured for 50 bits in the MUX array. Each line represents an individual SOT-MTJ. Notably, no bits failed even after 10^{10} write cycles under V_t of 2.5 V. Figure 2(d) shows the cumulative deviation of both R_P and R_{AP} at the initial state and after 10^{10} cycles. The measured resistances at three different cycle counts all follow normal distributions, with the distributions of R_P and R_{AP} showing no significant resistance drift. The cycling results at RT confirm remarkable endurance even under voltage stress after 10^{10} cycles. The test was terminated at 10^{10} cycles due to time constraints. We also achieved a high endurance of up to 10^{12} cycles for our fabricated SOT-MTJs (Section S2 in the ESM).

However, these results differ from the reported ones [32–34], in which SOT-MTJ failure was observed, attributed to the oxygen diffusion from the substrate to the MgO layer due to the accumulated Joule heating. To understand the differences, we systematically investigate the thermal characteristics of SOT-MRAM through simulation. We employed the heat conduction equation to simulate the thermal dynamics during the write operations [39, 40].

$$\rho C_p \frac{\partial T}{\partial t} = \nabla \cdot (k \nabla T) + Q \quad (1)$$

where ρ is the material density, C_p is the heat capacity, k is the thermal conductivity, T is the temperature, and Q is the heat generation rate per unit volume. The transient temperature evolution of SOT-MRAM is simulated using a 100-ns-width write pulse. Section S3 in the ESM describes the detailed structure and parameters of the simulation.

Figure 2(e) shows the temperature distribution along the SOT channel at the end of the write pulse. The SOT channel exhibits the highest temperature rise ΔT of approximately 270 K at a region of 300 nm away from the SOT-MTJ. This significant temperature rise can be attributed to the accumulated Joule heating during the write operation. Because the SOT channel is surrounded by either SiO_x or SiN_x layers with low thermal conductivity, the heat cannot be easily dissipated. However, the temperature at the MTJ location shows a relatively moderate increase of 80 K because the Joule heating can be dissipated vertically along the MTJ pillars. The temporal temperature evolution for the SOT-MTJ in Fig. 2(f) shows a rapid initial increase within the first 20 ns to 30 K, driven by adiabatic heating. Subsequently, it increases in proportion to the square root of time to about 80 K at 100 ns, because the thermal energy begins to dissipate into the surroundings. The simulated temperature rise ΔT for the SOT channel follows a similar trend to that of the SOT-MTJ region but with higher ΔT values. The relaxation time between write pulses during the cycling tests also needs to be considered in the simulations. As shown in Fig. 2(g), in the case of cycling tests, even though there is a relaxation time between the write pulses, the complete heat dissipation is not achieved within the 100 ns relaxation period. This leads to a cumulative temperature rise over repeated write cycles, which is higher than the case for the single write operation. The device will finally reach a thermal equilibrium state, where the temperature rise and heat dissipation are in balance, causing the temperature to saturate.

In our study, the ΔT of the SOT-MTJ differs from that in Ref. [34]. This variation can be attributed to differences in the current density, resistivity, and physical dimensions, as shown in Section S4 in the ESM. Specifically, in our SOT-MTJ array, the current density is one-third of that in Ref. [34], and the width of the SOT channel is an order of magnitude larger, enabling faster heat dissipation. These factors collectively explain the higher ΔT observed in Ref. [34]. In addition, in Section S5 in the ESM, we demonstrate that the simulation results are consistent with the test results.

In a summary of the cycling tests and simulation results at RT, even though the V_t applied was twice the normal V_{OP}, we did not observe any degradation of the SOT-MTJs. Compared with other studies, the high endurance stability achieved in the SOT-MRAM MUX array may originate from relatively low-temperature rise during the cycling at RT which is insufficient to activate elemental diffusion. However, some application scenarios require SOT-MRAM to operate reliably at high temperatures, which imposes strict reliability constraints. It is necessary to examine the endurance performance at high temperatures such as 125 °C.

3.2 Endurance under both voltage and temperature stress with failure model analysis

To verify the hypothesis, we conducted endurance tests using the same methodology at RT, but under an additional thermal stress (T_S) of 125 °C. As the temperature rises from RT to 125 °C, the initial TMR decreases from 80% to 50% before cycling, consistent with the typical temperature-dependent behavior [41]. When conducting the cycling test, the temperature of SOT-MTJ is the sum of the T_S and the temperature generated by the write current of the SOT channel. We observe that with the increase in the number of write cycles, the resistance value of MTJ changes differently from that at RT. Figures 3(a) and 3(b) show the TMR and R_{AP} as functions of the number of write cycles. During testing, the TMR remains stable at approximately 50% up to 10^6 cycles and then undergoes significant degradation beyond 10^7 cycles. The R_{AP} exhibited slightly different behavior, showing a minor increase (4%) up to 10^7 cycles, after which it displayed a noticeable downward trend. These results indicate substantial deterioration of the SOT-MTJs due to structural or material changes induced by the combined effects of the V_t and T_S, which are markedly different from the RT.

The cycling failure data were analyzed using the Weibull distribution and the log-normal distribution [42, 43]. As shown in Fig. 3(c), the failure time distribution of the 100 devices deviates from the 95% confidence interval (CI) of the Weibull distribution but aligns well with the log-normal distribution. This is particularly evident in the tail region of the data. For example, the one percentile (1%) failure cycle value of 6.4×10^6 falls within the 95% CI of the log-normal model (5.2×10^6 to 7.2×10^6 cycles), whereas this range lies outside the corresponding confidence interval of the Weibull model. These results indicate that the log-normal distribution provides statistically higher accuracy than the Weibull distribution when modeling the failure cycles of SOT devices. Furthermore, it implies that the failure mechanism of SOT-MTJs

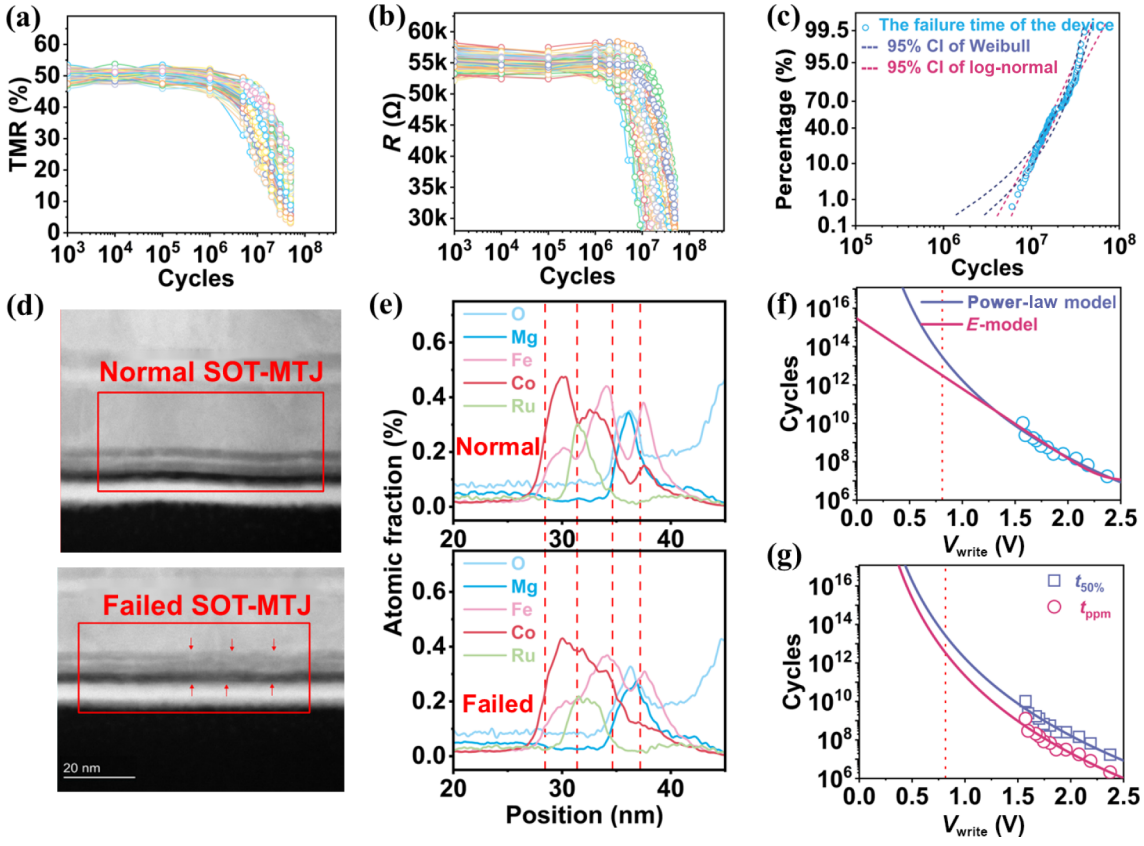


Figure 3 The change of (a) TMR and (b) R_{AP} as a function of writing cycling counts. Each curve represents a single SOT-MTJ device, with a total of over 100 devices characterized. The TMR and R_{AP} remain stable when the cycling counts are below 10^6. However, as the cycling counts increase, both the TMR and R_{AP} exhibit a significant degradation trend. (c) The statistical distributions of write failure cycles are presented using Weibull and log-normal fitting. The curves represent the 95% confidence intervals. The tail data of the Weibull distribution deviate from the confidence interval, while the tail data of the log-normal distribution are well-contained within the confidence interval. (d) TEM images comparing a normal SOT-MTJ and a failed SOT-MTJ. The normal SOT-MTJ exhibits sharp and continuous interfaces in the stack, whereas the failed SOT-MTJ shows intermixing at the CoFeB/MgO interface. (e) The EDS profile of the failed device reveals a broadened Mg peak and a reduced Mg/O intensity ratio. Additionally, the signals of Co, Fe, and Ru exhibit significant smearing and corresponding peak broadening. (f) The $t_{50\%}$ of SOT-MTJs are extrapolated from the acceleration voltage to the operating voltage by the *E*-model and the power-law model. The power-law model is more suitable for fitting the acceleration voltage. (g) The failure time is extrapolated from $t_{50\%}$ to t_{ppm}, and endurance cycles are reduced from 2.75×10^{13} to 3×10^{12} at the operating voltage.

may differ from the MgO dielectric breakdown observed in two-terminal STT-MTJs under voltage stress [17, 44], where the failure data follow the Weibull distribution.

The high-resolution scanning TEM (STEM) was performed on the failed bits after cycling to confirm the failure mechanism. Figure 3(d) compares a cross-sectional STEM image of the failed bit with that of the normal bit as a reference. The normal MTJ maintains sharp and well-defined interfaces between each layer (Fig. 3(d), top image). In contrast, the failed MTJ displays distinct contrast variations within the film stack (Fig. 3(d), bottom image), probably resulting from the element inter-diffusion. Complementary to the STEM analysis, Fig. 3(e) presents the energy-dispersive X-ray spectroscopy (EDS) profiles extracted from Fig. 3(d), quantifying the spatial distribution of key elements. The EDS profile of the failed device shows two critical abnormal phenomena compared with those of normal ones: (1) broadening of the Mg peak and a relatively lower ratio of Mg/O intensity, and (2) significant smearing of Co, Fe, and Ru signals and corresponding broadened peaks. This comparative elemental mapping conclusively shows that the device failure after cycling tests under V_t and T_S is correlated with element interdiffusion. The inter-diffusion likely induces two concurrent degradation mechanisms: (1) the formation of defects through stoichiometric imbalance within the MgO tunneling barrier, and (2) the modification of interface spin polarization characteristics, both of which contribute to the observed R_{AP} and TMR deterioration. These findings are phenomenologically consistent with Simon's other observation [32], even though more severe element diffusion has been found in our study, probably due to more accumulated Joule heating when the devices are tested at a T_S of 125 °C. Moreover, the STEM analysis explained why the statistical distribution of the cycles-to-failure of the SOT-MTJs in this study follows a log-normal behavior. This is because the log-normal distribution typically characterizes failure dominated by elemental diffusion, whereas the Weibull distribution is more suitable for describing failure governed by dielectric breakdown. Through experimental verification, we confirm that the failure mechanism of the SOT-MTJ originates from elemental diffusion and the failure distribution characteristics should be accurately described using a log-normal distribution.

Due to subtle process variations across the array, the inter-device non-uniformity may arise, leading to the non-uniform local Joule heating when the current flows through the SOT channel. This, in turn, results in a distributed rate of the thermally activated interdiffusion process. Therefore, we apply a log-normal distribution to statistically analyze the endurance cycles of devices within the array and extract the cycle number corresponding to a

50% failure probability ($t_{50\%}$) as the characteristic lifetime of the entire array under specified voltage and temperature conditions. Figure 3(f) shows the correlation between write voltage and $t_{50\%}$ at 125 °C. The endurance under V_{OP} is predicted utilizing the *E*-model and the power-law model, which are commonly employed for accelerated stress extrapolation analysis for thin oxides. The *E*-model, formulated as $TF \propto \exp(-\gamma E)$, is physically grounded in electric-field-driven dielectric breakdown mechanisms, where the applied field directly lowers the activation energy for bond breakage in the oxide [42, 43]. This model is well-suited for two-terminal MTJs in STT-MRAM, where the write current passes through the tunnel barrier. In contrast, the failure of three-terminal SOT-MRAM is initiated by Joule heating in the SOT channel, which promotes the thermally-activated elemental interdiffusion in the adjacent MTJ stack (Figs. 3(d) and 3(e)). This temperature rise scales with the Joule heating power ($P \propto V/R$) in the SOT channel [34, 43]. Consequently, the endurance exhibits a power-law dependence on the applied write voltage.

Both models capture the inverse relationship between write voltage and endurance cycles, although they diverge in their predictions at low voltage. The quantitative comparison using the Akaike and Bayesian information criteria (AIC/BIC) demonstrates that the power-law model provides a superior fit to acceleration data (see Section S6 in the ESM). Consequently, we employ the power-law model for extrapolation, as it more accurately represents the voltage-acceleration trend of the thermally-driven degradation observed in SOT-MRAM devices. We mathematically provide a forecast of ppm-level endurance (i.e., t_{1ppm}) based on the log-normal distribution. Figure 3(g) shows the endurance extrapolated from $t_{50\%}$ in Fig. 3(f) to t_{1ppm}. The extrapolation results reveal an exceptional $t_{50\%}$ endurance of ~ 10^{14} cycles under V_{OP} at 125 °C, demonstrating remarkably high-temperature reliability. However, for more stringent ppm-level reliability requirement, the endurance can still reach ~ 10^{12} cycles. Although this performance is already suitable for Mb-level array applications, further improvement in cycling endurance remains essential to meet the growing demands of high-density memory applications.

In summary, the endurance characterization demonstrates a cycling capability approaching 10^{14} cycles at 125 °C. The material characterization and structural analysis indicate that the primary failure mechanism in SOT-MTJs arises from the elemental interdiffusion at the interfaces. These findings highlight the critical need to enhance SOT-MRAM endurance performance through thermal management strategies that minimize Joule heating generation and optimize heat dissipation pathways.

3.3 Factors influencing endurance

Under the experimental conditions, the SOT-MRAM memory array demonstrates a cycling endurance up to 10^{14} cycles. However, when extrapolated to a 1 ppm failure rate criterion, the cycling endurance degrades to only 10^{12} cycles, significantly lower than the expected target of 1×10^{17} cycles [9]. To address this issue, we systematically evaluated a variety of potential methods to enhance the endurance of the array. These methods cover multiple professional dimensions such as structural design and materials engineering. According to the basic principles of thermodynamics, reducing the operating temperature of the SOT-MTJ is the key to improving the cycling endurance of the SOT-MRAM. Therefore, effectively reducing the Joule heating effect and simultaneously optimizing the heat dissipation mechanism are important ways to extend the service life of the device. In this section, we will conduct detailed discussions on the two key aspects of reducing heat generation and optimizing heat dissipation.

First, we discuss the key factors influencing heat generation. This can be achieved by decreasing the critical switching current and reducing the resistance of the SOT channel. Reducing the critical switching current is one of the core objectives in optimizing SOT-MRAM devices, as it not only significantly reduces power consumption but also effectively mitigates thermal effects during device operation. To reduce the critical switching current, we enhanced the spin Hall angle (SHA) of the SOT channel through material doping engineering and introduced a spacer layer in the MTJ structure, effectively suppressing current shunting effects [37]. Experimental results show that these synergistic optimization strategies significantly improve the magnetic moment switching efficiency of SOT, achieving a reduction of 25% in write current compared to the pre-optimized conditions. Based on the quadratic relationship between Joule heating and current, the heat generation of the device can be reduced by 43.75%.

Decreasing the resistance of the SOT channel (R_{SOT}) can also effectively reduce heat generation. This can be achieved by shortening the SOT channel length (L_{SOT}). To evaluate the impact of L_{SOT} on the endurance of SOT-MRAM, we successfully fabricated SOT-MRAM arrays with L_{SOT} scaled down from 2.5 to 1 μm, as shown in Fig. 4(a). Then, we validated their endurance performance based on the cycling tests under V_t and T_s of 125 °C mentioned above. Figure 4(b) presents the measured failed cycles of SOT-MRAM arrays with different values of L_{SOT}, along with the extrapolated values using the power-law model. As shown in Fig. 4(b), the endurance increases from 2.75×10^{13} to 2.38×10^{18} cycles as the values of L_{SOT} are reduced from 2.5 to 1 μm. Thermal simulations reveal a reduction in ΔT from 80 (L_{SOT} = 2.5 μm) to 40 K (L_{SOT} = 1 μm) at a 100-ns-width pulse, representing a decrease of 50% in thermal loading. Therefore, reducing L_{SOT} is a feasible and effective approach to enhance endurance. In this study, the SOT-MTJ's CD is 350 nm × 1050 nm. Even after processing optimization to reduce the L_{SOT} length to 1 μm, the SOT channel retains an approximate 500 nm redundancy (L_{SOT} = MTJ width + 200 nm process margin). As illustrated in Fig. 4(c), even under a conservative linear fitting estimation, the endurance of SOT-MTJs is projected to reach 1×10^{20} cycles when L_{SOT} is scaled down to 500 nm. These findings underscore the critical role of L_{SOT} scaling in achieving high-performance and reliable SOT-MRAM devices.

Besides reducing heat generation in the SOT channel, enhancing the heat dissipation capability of SOT-MRAM is also a viable strategy to improve its endurance. NAND-like devices can achieve high-density storage and highly reliable selective writing through mechanisms such as toggle spin torques (TST) or voltage-controlled magnetic anisotropy (VCMA) [45–47], making them one of the key development directions in SOT-MRAM. We propose an innovative high-reliability SOT-MRAM structure that integrates low-resistivity metal vias beneath the SOT channel between adjacent MTJs in a conventional NAND-like architecture (where four MTJs share one SOT channel), forming a composite thermal-electrical synergistic optimization network. As shown in Fig. 4(d), the metal vias employ high-conductivity materials, significantly higher than the SOT channel material (e.g., β-W thermal conductivity = 10 W/(m·K)), enabling the current to preferentially shunt through the vias during write operations, thereby reducing the effective resistance of the SOT channel and minimizing the Joule heat generation. Simultaneously, the vias act

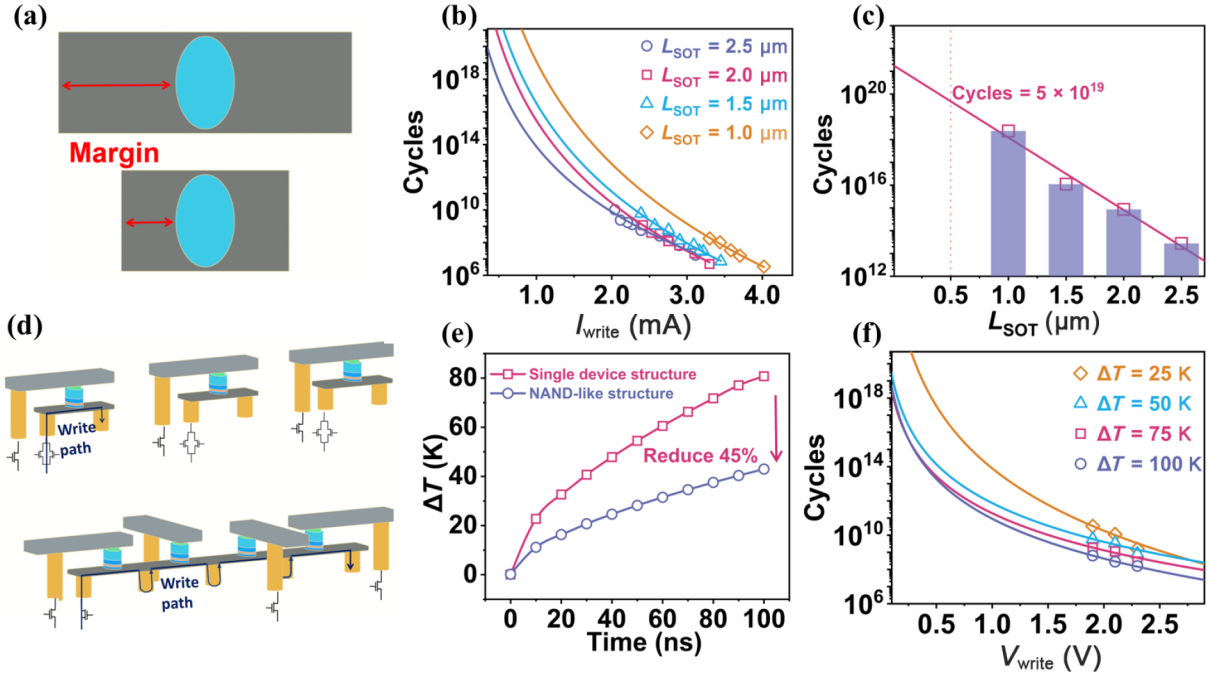


Figure 4 (a) Reducing the length of the SOT channel can decrease resistance and thereby lower the heat generation. (b) As L_{SOT} decreases from 2.5 to 1 μm, a significant improvement in endurance cycles is observed. Specifically, when L_{SOT} is 2.5 μm, the cycling counts reach 2.75×10^{13}. However, when L_{SOT} is reduced to 1 μm, the cycling counts significantly increase to over 1×10^{18}. (c) Based on the extrapolation of failure cycles at critical switching currents from the experimental data, L_{SOT} down to 500 nm could significantly enhance the endurance, reaching approximately 5×10^{19} cycles. (d) Schematic diagram of the NAND-like structure. Four MTJs share one SOT channel to deposit metal vias among the MTJs. (e) The comparison of SOT-MTJ temperature rise between the two structures after a 100-ns-width write pulse shows that the temperature rise is reduced by 45%. (f) The endurance of SOT-MTJs exhibits a clear dependence on ΔT, with a pronounced trend of increasing endurance cycles as ΔT decreases. This significant thermal improvement is projected to enhance endurance by two orders of magnitude.

as vertical thermal conduction paths (thermal conductivity > 400 W/(m·K)), working in conjunction with the top electrode heat dissipation path to form a multi-stage thermal management system, with simulations demonstrating a temperature reduction from 80 to 40 K (Fig. 4(e)). During write operations, the shunting effect of the vias lowers the current through the high-resistance SOT channel while maintaining sufficient spin-orbit torque for free layer switching. For read operations, a small vertical current passes through the target MTJ with independent MOSFET gating for access, while the vias' high thermal conductivity rapidly dissipates read-induced heat to prevent thermal crosstalk. This design enhances the endurance through dual mechanisms: Suppressing high-temperature-induced MTJ interface diffusion and SOT channel lattice distortion via multi-via integration, while the current shunting effect reduces operational current in the SOT channel, significantly mitigating failure risks. As shown in Fig. 4(f), we investigated the temperature-dependent endurance characteristics of SOT-MRAM by varying the experimental temperature. The results demonstrate a negative correlation between the device's endurance performance and the temperature variation under a fixed write voltage. When ΔT decreases from 100 to 25 K, the endurance cycles improve by 2 to 3 orders of magnitude.

This section presents a co-design centered on thermal management, which reduces the effective temperature rise in the MTJ region through minimizing heat generation at the SOT channel and actively dissipating heat. The approach includes optimizing the SOT channel material to enhance the spin Hall angle, while shortening the channel length from 2.5 to 1 μm, thereby reducing both the switching current and Joule heating. Additionally, low-resistivity and high-thermal-conductivity metal vias are innovatively integrated into a NAND-like array architecture. These vias not only shunt write current away from the high-resistance SOT channel to reduce heat generation but also act as efficient vertical thermal conduits to rapidly dissipate accumulated heat from the MTJ region. This holistic thermal-optimization design not only achieves a four-order-of-magnitude improvement in endurance but also provides a viable technical pathway toward achieving "unlimited" endurance in high-density, practical SOT-MRAM arrays.

4 Conclusions

In conclusion, this study provides the first array-level endurance analysis of SOT-MRAM under high-temperature operating conditions, revealing that the dominant failure mechanism is thermally activated elemental interdiffusion within the MTJ stack. By employing a combined log-normal–power-law failure model and implementing the channel-length scaling along with the structural thermal management, we demonstrate endurance exceeding 10^{18} cycles at 125 °C, a significant advance over prior reports. These findings offer practical design guidelines for realizing SOT-MRAM in high-performance and high-reliability applications.

Electronic Supplementary Material: Supplementary material (CMOS-compatible SOT-MRAM array, endurance test of isolated devices, simulation structure and parameters, comparison between this study and prior results, verification of MTJ temperature changes during the writing process, failure model analysis, and analysis of only applying temperature stress conditions) is available in the online version of this article at https://doi.org/10.26599/NR.2026.94908447.

Data availability

All data needed to support the conclusions in the paper are presented in the manuscript and the Electronic Supplementary Material. Additional data related to this paper may be requested fromthe corresponding author upon request

Acknowledgements

This work was supported by the National Program for Support of Top-notch Young Professionals, the Fundamental Research Funds for the Central Universities, the National Key Research and Development Program of China (Nos. 2022YFB4400200, 2021YFB3601303, and 2021YFB3601304), and the National Natural Science Foundation of China (Nos. 62001014, 92164206, 62401031, and 62171013).

Declaration of competing interest

All the contributing authors report no conflict of interests in this work.

Author contribution statement

C. P. J.: Data curation, validation, project administration, writing manuscript, experimental design. J. H. L.: Data curation, writing manuscript, project administration. C. W.: Validation, experimental design. S. Y. L.: Project administration, writing manuscript. X. Y. Y.: Writing manuscript. W. L. C.: Data curation. D. R. X.: Writing manuscript. X. F. F.: Data curation. H.-X. L.: Project administration, writing manuscript. G. F. W.: Project administration. H. Z.: Project administration. K. H. C.: Project administration. Z. H. W.: Project administration, writing manuscript, funding acquisition. W. S. Z.: Project administration, funding acquisition. All the authors have approved the final manuscript.

Use of AI statement

None.

References

[1] Meuser, T.; Lovén, L.; Bhuyan, M.; Patil, S. G.; Dustdar, S.; Aral, A.; Bayhan, S.; Becker, C.; Lara, E. D.; Ding, A. Y. et al. Revisiting edge AI: Opportunities and challenges. *IEEE Internet Comput.* **2024**, *28*, 49–59.

[2] Zheng, Y. Q.; Xu, H.; Li, Z. X.; Li, L. L.; Yu, Y. C.; Jiang, P. F.; Shi, Y. M.; Zhang, J.; Huang, Y. Q.; Luo, Q. et al. Artificial intelligence-driven approaches in semiconductor research. *Adv. Mater.* **2025**, *37*, 2504378.

[3] Ale, L.; Zhang, N.; King, S. A.; Chen, D. J. Empowering generative AI through mobile edge computing. *Nat. Rev. Electr. Eng.* **2024**, *1*, 478–486.

[4] Yoo, H. J. 1.2 Intelligence on silicon: From deep-neural-network accelerators to brain mimicking AI-SoCs. In *2019 IEEE International Solid-State Circuits Conference*, San Francisco, USA, 2019, pp 20–26.

[5] Lu, A. N.; Lee, J.; Kim, T. H.; Karim, M. A. U.; Park, R. S.; Simka, H.; Yu, S. M. High-speed emerging memories for AI hardware accelerators. *Nat. Rev. Electr. Eng.* **2024**, *1*, 24–34.

[6] Li, Z. X.; Xu, H.; Zheng, Y. Q.; Liu, L. C.; Li, L. L.; Lou, Z.; Wang, L. L. A reconfigurable heterostructure transistor array for monocular 3D parallax reconstruction. *Nat. Electron.* **2025**, *8*, 46–55.

[7] Zhong, B. W.; Qin, X. K.; Xu, H.; Deng, F.; Wang, H. L.; Li, L. L.; Li, Z. X.; Zhang, W. X.; Lou, Z.; Wang, L. L. Monolithic cell-on-memristor architecture enables wafer-scale integration of oscillatory chemoreceptors for bio-realistic gustatory chips. *Nat. Mater.*, in press, https://doi.org/10.1038/s41563-025-02436-y.

[8] Shao, Q. M.; Wang, Z. R.; Yang, J. J. Efficient AI with MRAM. *Nat. Electron.* **2022**, *5*, 67–68.

[9] Worledge, D. C.; Hu, G. H. Spin-transfer torque magnetoresistive random access memory technology status and future directions. *Nat. Rev. Electr. Eng.* **2024**, *1*, 730–747.

[10] Dieny, B.; Prejbeanu, I. L.; Garello, K.; Gambardella, P.; Freitas, P.; Lehndorff, R.; Raberg, W.; Ebels, U.; Demokritov, S. O.; Akerman, J. et al. Opportunities and challenges for spintronics in the microelectronics industry. *Nat. Electron.* **2020**, *3*, 446–459.

[11] Lee, T. Y.; Lee, J. M.; Oh, Y. W.; Kim, H. H.; Kwon, B. S.; Yamane, K.; Kim, M. K.; Jang, P. H.; Lee, D. K.; Gwak, H. J. et al. World-most energy-efficient 14 nm automotive eMRAM technology for high-endurance applications. In *2024 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, USA, 2024, pp 1–4.

[12] Worledge, D. C. Spin-transfer-torque MRAM: The next revolution in memory. In *2022 IEEE International Memory Workshop (IMW)*, Dresden, Germany, 2022, pp 1–4.

[13] Oh, J.; Park, J.; Suh, K.; Lee, K.; Hwang, S.; Bak, M.; Kim, H.; Kwon, B.; Lee, D.; Kim, M. et al. 14 nm FinFET node embedded MRAM technology for automotive non-volatile RAM applications with endurance over 1E12-cycles. In *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Honolulu, USA, 2024, pp 1–2.

[14] Clarke, P. *TSMC Embedded MRAM is Key to Gyrfalcon AI Chip* [Online]. Technology News 2018, https://www.eenewseurope.com/ (accessed Sep 22, 2018).

[15] Lee, P. H.; Lee, C. F.; Shih, Y. C.; Lin, H. J.; Chang, Y. A.; Lu, C. H.; Chen, Y. L.; Lo, C. P.; Chen, C. C.; Kuo, C. H. et al. 33.1 A 16 nm 32 Mb embedded STT-MRAM with a 6 ns read-access time, a 1M-cycle write endurance, 20-year retention at 150 °C and MTJ-OTP solutions for magnetic immunity. In *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, USA, 2023, pp 494–496.

[16] Ko, S.; Shim, J.; Park, J. H.; Lim, W.; Jung, H.; Bak, J. H.; Jeong, D.; Lee, J. W.; Whang, H.; Eom, M. et al. Key technologies of scaling embedded MRAM to 8 nm logic and beyond for automotive application. In *2024 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, USA, 2024, pp 1–4.

[17] Tan, J.; Lim, J. H.; Sikder, B.; Baten, M. Z.; Kwon, J. H.; Yamane, K.; Naik, V. B.; Raghavan, N.; Pey, K. L. Impact of voltage polarity on time-dependent dielectric breakdown of 1-nm MgO-based STT-MRAM with self-heating correction. *IEEE Trans. Electron Dev.* **2023**, *70*, 76–82.

[18] Wei, Z.; Kim, W.; Wang, Z.; Hu, L.; Jung, D.; Zhang, L.; Huai, Y. Accurate and fast STT-MRAM endurance evaluation using a novel metric for asymmetric bipolar stress and deep learning. In *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Honolulu, USA, 2022, pp 373–374.

[19] Hu, G. H. Spin-transfer torque MRAM-status and outlook. In *2023 IEEE International Magnetic Conference-Short Papers (INTERMAG Short Papers)*, Sendai, Japan, 2023, pp 1–2.

[20] Mihai Miron, I.; Gaudin, G.; Auffret, S.; Rodmacq, B.; Schuhl, A.; Pizzini, S.; Vogel, J.; Gambardella, P. Current-driven spin torque induced by the Rashba effect in a ferromagnetic metal layer. *Nat. Mater.* **2010**, *9*, 230–234.

[21] Miron, I. M.; Garello, K.; Gaudin, G.; Zermatten, P. J.; Costache, M. V.; Auffret, S.; Bandiera, S.; Rodmacq, B.; Schuhl, A.; Gambardella, P. Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection. *Nature* **2011**, *476*, 189–193.

[22] Shao, Q. M.; Li, P.; Liu, L. Q.; Yang, H.; Fukami, S.; Razavi, A.; Wu, H.; Wang, K.; Freimuth, F.; Mokrousov, Y. et al. Roadmap of spin-orbit torques. *IEEE Trans. Magn.* **2021**, *57*, 800439.

[23] Nguyen, V. D.; Rao, S.; Wostyn, K.; Couet, S. Recent progress in spin-

orbit torque magnetic random-access memory. *npj Spintron.* **2024**, *2*, 48.

[24] Song, M. Y.; Chen, K. L.; Chen, K. M.; Chang, K. T.; Wang, I. J.; Hsin, Y. C., Lin, C. Y.; Ambrosi, E.; Khwa, W. S.; Lu, Y. L. et al. High RA dual-MTJ SOT-MRAM devices for high speed (10 ns) compute-in-memory applications. In *2023 International Electron Devices Meeting (IEDM)*, San Francisco, USA, 2023, pp 1–4.

[25] Li, X.; Lin, S. J.; Dc, M.; Liao, Y. C.; Yao, C. Y.; Naeemi, A.; Tsai, W.; Wang, S. X. Materials requirements of high-speed and low-power spin-orbit-torque magnetic random-access memory. *IEEE J. Electron Dev. Soc.* **2020**, *8*, 674–680.

[26] Yasin, F.; Palomino, A.; Kumar, A.; Pica, V.; Van Beek, S.; Talmelli, G.; Nguyen, V. D.; Cosemans, S.; Crotti, D.; Wostyn, K. et al. Extremely scaled perpendicular SOT-MRAM array integration on 300 mm wafer. In *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Honolulu, USA, 2024, pp 1–2.

[27] Rao, S.; Cai, K.; Talmelli, G.; Franchina-Vergel, N.; Janssens, W.; Hody, H.; Yasin, F.; Wostyn, K.; Couet, S. Spin-orbit torque MRAM for ultrafast cache and neuromorphic computing applications. In *2023 IEEE International Memory Workshop (IMW)*, Monterey, USA, 2023, pp 1–4.

[28] Naganuma, H.; Miura, S.; Honjo, H.; Nishioka, K.; Watanabe, T.; Nasuno, T.; Inoue, H.; Nguyen, T. V. A.; Endo, T.; Noguchi, Y. et al. Advanced 18 nm Quad-MTJ technology overcomes dilemma of retention and endurance under scaling beyond 2X nm. In *2021 Symposium on VLSI Technology*, Kyoto, Japan, 2021, pp 1–2.

[29] Jinnai, B.; Igarashi, J.; Watanabe, K.; Funatsu, T.; Sato, H.; Fukami, S.; Ohno, H. High-performance shape-anisotropy magnetic tunnel junctions down to 2.3 nm. In *2020 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, USA, 2020, pp 24.6.1–24.6.4.

[30] Du, A.; Zhu, D. Q.; Cao, K. H.; Zhang, Z. Z.; Guo, Z. X.; Shi, K. W.; Xiong, D. R.; Xiao, R.; Cai, W. L.; Yin, J. L. et al. Electrical manipulation and detection of antiferromagnetism in magnetic tunnel junctions. *Nat. Electron.* **2023**, *6*, 425–433.

[31] Peng, S. Z.; Zhu, D. Q.; Li, W. X.; Wu, H.; Grutter, A. J.; Gilbert, D. A.; Lu, J. Q.; Xiong, D. R.; Cai, W. L.; Shafer, P. et al. Exchange bias switching in an antiferromagnet/ferromagnet bilayer driven by spin-orbit torque. *Nat. Electron.* **2020**, *3*, 757–764.

[32] Van Beek, S.; Cai, K. M.; Rao, S.; Jayakumar, G.; Couet, S.; Jossart, N.; Chasin, A.; Kar, G. S. MTJ degradation in SOT-MRAM by self-heating-induced diffusion. In *2022 IEEE International Reliability Physics Symposium (IRPS)*, Dallas, USA, 2022, pp 4A.2-1–4A.2-4.

[33] Van Beek, S.; Cai, K. M.; Fan, K. Q.; Talmelli, G.; Trovato, A.; Jossart, N.; Rao, S.; Chasin, A.; Couet, S. MTJ degradation in multi-pillar SOT-MRAM with selective writing. In *2023 IEEE International Reliability Physics Symposium*, Monterey, USA, 2023, pp 1–7.

[34] Van Beek, S.; Kateel, V.; Cai, K. M.; Jossart, N.; Rao, S.; Couet, S. Impact of SOT & STT stress on MTJ degradation in SOT-MRAM. In *2023 IEEE International Magnetic Conference-Short Papers (INTERMAG Short Papers)*, Sendai, Japan, 2023, pp 1–2.

[35] Shiokawa, Y.; Komura, E.; Ishitani, Y.; Tsumita, A.; Suda, K.; Kakinuma, Y.; Sasaki, T. High write endurance up to 10^{12} cycles in a spin current-type magnetic memory array. *AIP Adv.* **2019**, *9*, 035236.

[36] Zhang, H. C.; Ma, X. Y.; Jiang, C. P.; Yin, J. L.; Lyu, S.; Lu, S. Y.; Shang, X. T.; Man, B. W.; Zhang, C.; Li, D. D. et al. Integration of high-performance spin-orbit torque MRAM devices by 200-mm-wafer manufacturing platform. *J. Semicond.* **2022**, *43*, 102501.

[37] Jiang, C. P.; Lu, S. Y.; Zhang, Z. K.; Fan, X. F.; Xiong, D. R.; Li, J. H.; Liu, H. X.; Wang, G. F.; Zhang, H.; Jin, H. et al. Demonstration of 128 Kb SOT-MRAM chip with 5 ns write and 15 ns read speed, high endurance over 10^{10} and low ECC-on bit error rate. In *2024 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, USA, 2024, pp 1–4.

[38] Xiong, D. R.; Fan, X. F.; Jiang, C. P.; Wang, G. F.; Liu, H. X.; Zhang, H. C.; Lu, S. Y.; Lin, J. H.; Zhang, H.; Cao, K. H. et al. First CMOS-integrated 128 Kb antiferromagnet-based MRAM with immunity to 3 T magnetic fields. In *2024 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, USA, 2024, pp 1–4.

[39] Hadámek, T.; Jørstad, N. P.; de Orio, R. L.; Goes, W.; Selberherr, S.; Sverdlov, V. A comprehensive study of temperature and its effects in SOT-MRAM devices. *Micromachines* **2023**, *14*, 1581.

[40] Su, H. T.; Kwon, H.; Hwang, W.; Xue, F.; Köroğlu, Ç.; Tsai, W.; Asheghi, M.; Goodson, K. E.; Wang, S. X.; Pop, E. Thermal optimization of two-terminal SOT-MRAM. *J. Appl. Phys.* **2024**, *136*, 013901.

[41] Jiang, C. P.; Li, J. H.; Zhang, H. C.; Lu, S. Y.; Li, P. B.; Wang, C.; Zhang, Z. K.; Hou, Z. Y.; Liu, X.; Feng, J. G. et al. Demonstration of a manufacturable SOT-MRAM multiplexer array towards industrial applications. *J. Semicond.* **2023**, *44*, 122501.

[42] Padovani, A.; La Torraca, P.; Strand, J.; Shluger, A.; Milo, V.; Larcher, L. Towards a universal model of dielectric breakdown. In *2023 IEEE International Reliability Physics Symposium*, Monterey, USA, 2023, pp 1–8.

[43] Padovani, A.; La Torraca, P.; Strand, J.; Larcher, L.; Shluger, A. L. Dielectric breakdown of oxide films in electronic devices. *Nat. Rev. Mater.* **2024**, *9*, 607–627.

[44] Sato, H.; Shin, H. M.; Jung, H.; Lee, S. W.; Bae, H.; Kwon, H.; Ryu, K. H.; Lim, W. C.; Han, Y. S.; Jeong, J. H. et al. Comprehensive study on prediction of endurance properties from breakdown voltage in high-reliable STT-MRAM. In *2023 IEEE International Reliability Physics Symposium (IRPS)*, Monterey, USA, 2023, pp 1–5.

[45] Wang, Z. H.; Zhang, L.; Wang, M. X.; Wang, Z. L.; Zhu, D. Q.; Zhang, Y. G.; Zhao, W. S. High-density NAND-like spin transfer torque memory with spin orbit torque erase operation. *IEEE Electron Device Lett.* **2018**, *39*, 343–346.

[46] Wang, Z. H., Zhou, H. X.; Wang, M. L.; Cai, W. L.; Zhu, D. Q.; Klein, J. O.; Zhao, W. S. Proposal of toggle spin torques magnetic RAM for ultrafast computing. *IEEE Electron Device Lett.* **2019**, *40*, 726–729.

[47] Zhao, Y. L.; Yang, J. L.; Li, B.; Cheng, X. Z.; Ye, X. C.; Wang, X. Y.; Jia, X. T.; Wang, Z. H.; Zhang, Y. G.; Zhao, W. S. NAND-SPIN-based processing-in-MRAM architecture for convolutional neural network acceleration. *Sci. China Inf. Sci.* **2023**, *66*, 142401.