Special Session: STT-MRAMs: Technology, Design and Test

Anteneh Gebregiorgis[1], Lizhou Wu[1], Christopher Münch[2], Siddharth Rao[3], Mehdi B. Tahoori[2] and Said Hamdioui[1]

[1]Department of Quantum and Computer Engineering
Delft University of Technology
Delft, Netherlands

[2]Chair of Dependable Nano Computing
Karlsruhe Institute of Technology
Karlsruhe, Germany

[3]IMEC
Leuven, Belgium

2022 IEEE 40th VLSI Test Symposium (VTS) | 978-1-6654-1060-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/VTS52500.2021.9794278

***Abstract*—STT-MRAM has long been a promising non-volatile memory solution for the embedded application space owing to its attractive characteristics such as non-volatility, low leakage, high endurance, and scalability. However, the operating requirements for high-performance computing (HPC) and low power (LP) applications involve different challenges. This paper addresses different aspects of STT-MRAM; it will cover state-of-the-art, some new results and future challenges related to technology, design and test. While STT-MRAM devices have shown encouraging performance metrics at device-level, a key challenge has been achieving backend-of-line (BEOL) CMOS compatibility, while retaining the benefits of low power operation. Scaling demands to improve data densities have placed additional challenges in terms of addressing the impact of process-induced damage on device performance at CD $<$ 100 nm. In addition, the paper discusses the design of reliable read mechanism considering the variability effects. Moreover, the failure of traditional fault modeling and test approaches in model STT-MRAM unique defects for appropriate test solutions is demonstrated in this paper based on silicon data.**

***Index Terms*—STT-MRAM, Device-aware test, Reliability**

I. Introduction

The conventional embedded memories such as Static Random Access Memory (SRAM), embedded Dynamic Random Access Memory (eDRAM) and eFlash are now struggling to meet the increasing demands in terms of energy efficiency, reliability, scalability and manufacturing costs [1]. As one of the most promising non-volatile memory technologies, Spin-Transfer Torque Magnetic Random Access Memory (STT-MRAM) offers competitive write/read performance, endurance, density, retention, and power consumption benefits [2]. The tunability of these aspects makes it customizable as both embedded and discrete memory solutions for a variety of applications such as Internet-of-Things (IoT), automotive, aerospace, and last-level caches [3]. Therefore, STT-MRAM technology has received a large amount of attention for commercialization from major semiconductor companies such as Everspin, TSMC, Samsung, and Intel [3]. However, with its current technology stature, STT-MRAM cannot fully replace SRAM as it has relatively high write latency and energy, design and test challenges [4], [5]. The magnetic tunnel junction, which is the storage device of the STT-MRAM, needs high write current for a longer duration in order to switch its magnetic orientation (its content) resulting in high write energy. Moreover, the lack of proper and effective test strategy and design challenges are the main hindrance for the widespread applicability of STT-MRAM devices. Therefore, its imperative to address the technology, design and test challenges of STT-MRAM in order to make it a primary candidate for on-chip memories by replacing the struggling conventional memory technologies.

Addressing the fundamental challenges such as technological [6], [7], reliability [8] and test [5] challenges is crucial for the commercialization and widespread applicability of STT-MRAM [9]. Various industrial and academic research works have proposed to address these challenges separately [10], [11] and collectively [5], [9]. The works in [6], [7] investigated the future prospect and technological challenges of STT-MRAM for scaled technology nodes in order to enable the deployment of STT-MRAM for high performance cache memories. Similarly, the works in [8] investigated the read failure of STT-MRAM and developed mitigation techniques to tackle the issues. The work in [5] evaluated the failure mechanisms, developed proper fault models for device-aware testing of STT-MRAM. Thus, proper solutions addressing the technology, design and test challenges is crucial for the commercialization and adoption of STT-MRAM as the prime choice for embedded memories.

This paper addresses the technology, design and test aspects of STT-MRAM; it will cover the state-of-the-art, some new results and future challenges related to technology, design and test. While STT-MRAM devices have shown encouraging performance metrics at device-level, a key challenge has been achieving backend-of-line (BEOL) CMOS compatibility, while retaining the benefits of low power operation. Scaling demands to improve data densities have placed additional challenges in terms of addressing the impact of process-induced damage on device performance at CD $<$ 100 nm. additionally, the paper discusses the design of reliable read mechanism considering the variability effects. Moreover, the paper demonstrates, based on silicon data, how traditional fault modeling and test approaches fail to model STT-MRAM unique defects and hence fail in providing appropriate test solutions.

The reminder of this paper is organized as follows: Section II presents the technological and design challenges of STT-MRAM. Section III discusses the design of reliable sens-

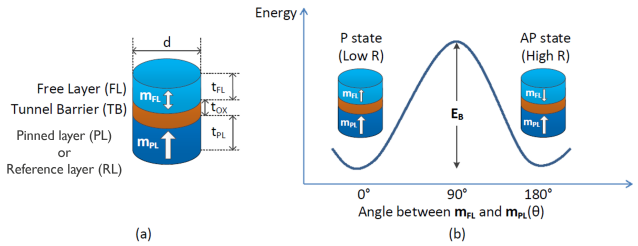


Fig. 1: Schematic representation of a pMTJ device and the energy barrier separating the stable binary states.

ing mechanisms for STT-MRAM followed by the discussion of device-aware testing technique for STT-MRAM devices in Section IV. Finally the paper is concluded in Section V.

II. STT-MRAM Technology for Embedded Applications

In this section, we will review the fundamentals and operating principles of an STT-MRAM device. This will be followed by an understanding of the manufacturing process flow of embedded STT-MRAMs, with a specific focus on the underlying challenges towards market adoption. Overcoming these challenges are critical to minimizing defects and maximizing yield, and we will review the current state-of-the-art within the framework of these challenges.

A. STT-MRAM Device Organization

An STT-MRAM device is composed of a magnetic tunnel junction (MTJ) driven by an access transistor. Data is hard coded into the MTJ by orienting the magnetization of one of its magnetic layers with respect to the other, thus enabling its non-volatile operation. Typically, the MTJ consists of a thin dielectric barrier (such as MgO) sandwiched by two ferromagnetic layers called the *free layer* (FL) and *reference layer* (RL) respectively. As the name suggests, the magnetization of the FL can be re-oriented by external means while that of the RL is robust to external influences.

1) Free layer (FL): The top layer is called free layer, which is typically made of CoFeB material (t_{FL}= 1.5nm) [12]. The magnetization (m_{FL}) in the FL is oriented along the easy axis (an energetically favorable direction), and can be switched to the opposite direction by STT. The saturation magnetization M_{s} and magnetic anisotropy field H_{k} are two key technology parameters that influence the write energy and the data retention against external thermal fluctuations (also known as thermal stability, Δ) of the FL. The easy axis lies in the thin film if the FL has in-plane magnetic anisotropy, whereas it points perpendicular to the free layer for perpendicular magnetic anisotropy (pMTJ). pMTJ STT-MRAM devices enable higher scalability and reduced write power consumption [2], and thus will form the main focus of this paper.

2) Tunnel barrier (TB): The MgO dielectric layer in the middle is called tunnel barrier. As the TB layer is ultrathin, typically $\sim 1\,\mathrm{nm}$ [12], electrons have chance to tunnel through it overcoming its potential barrier height $\bar{\varphi}$ [13]. This makes the device behave as a tunneling-like resistor. To compare the sheet resistivity of different MTJ designs, the *resistance-area* (RA) product [14] is used. This is a figure-of-merit which is commonly used in MRAM community, and it is independent on device size.

3) Reference Layer (RL) or Pinned layer (PL): The bottom ferromagnetic layer is referred to as pinned layer or the reference layer; typically, its thickness is $t_{\mathrm{RL}}{=}\sim 2.5\,\mathrm{nm}$ [12]. The magnetization (m_{RL}) of the RL is strongly pinned to a certain direction by an inner synthetic anti-ferromagnet (iSAF). With the fixed magnetization in RL as a reference, the magnetization in FL is either parallel (P state) or anti-parallel (AP state) to that of RL. As both terms refer to the same magnetic layer, we will use the term RL in the rest of this paper.

By sensing the magnetization of the FL with respect to the RL, the encoded data can be read out. Figure 1 schematically represents the MTJ unit, and the energy barrier (E_B) to be surmounted for a typical write operation.

B. Operating Principles, and Design Considerations of STT-MRAM Devices

Write operation: Information is written in MTJ devices through STT-induced magnetization reversal in the FL by means of a flowing electrical current through the MTJ device. To achieve this, a minimum energy greater than the energy barrier ($E > E_B$) between the two stables states must be supplied to the system. The energy supplied, and consequently the STT, can be correlated to the flowing current. Assuming a coherent model of magnetization reversal, the required write current (I_c) can be expressed as: $I_c = I_{c0}(1 - \frac{1}{\Delta}\ln(f_0 t_p))$, where $I_{c0} = \frac{8\alpha e M_s t}{\eta\hbar\pi d^2} H_k$ is the critical write current at zero temperature and is purely dependent on FL material properties such as the saturation magnetization (M_s), effective magnetic anisotropy field (H_k), damping (α) and spin polarization (η). The attempt frequency (f_0) is typically $\sim$1 ns, t_p refers to the operating pulse-width (inverse of frequency), and t, d refer to the physical dimensions of the MTJ device. Furthermore, the thermal stability (or data retention) can be written as $\Delta = (M_s H_k t d)/(2k_B T)$, where ($k_B T$) represents the ambient energy in the system due to random thermal fluctuations. For a data retention of 10 years at room temperature, the required thermal stability is $\sim$60–70, depending on the variation within the STT-MRAM array [14].

Read operation: A sense current (I_{rd}), is applied through the MTJ device to read the stored information. Depending on the relative orientations of the two magnetic layers, the measured resistance across the device can vary as shown in Figure 1b. This phenomenon is the well-known *tunnel magnetoresistance* (TMR) effect, and arises due to complex quantum mechanical tunneling phenomena [15]. The TMR is a figure-of-merit, expressed as a ratio, where $TMR{=}(R_{\mathrm{AP}}{-}R_{\mathrm{P}})/R_{\mathrm{P}}$, where R_{AP} and R_{P} are the resistances in the AP and P states respectively. A higher TMR ratio enables faster and more accurate read operations, thus leading to reduced read latencies. For commercially feasible STT-MRAM products, a TMR ratio of $\sim$150% is considered acceptable [2]. To avoid

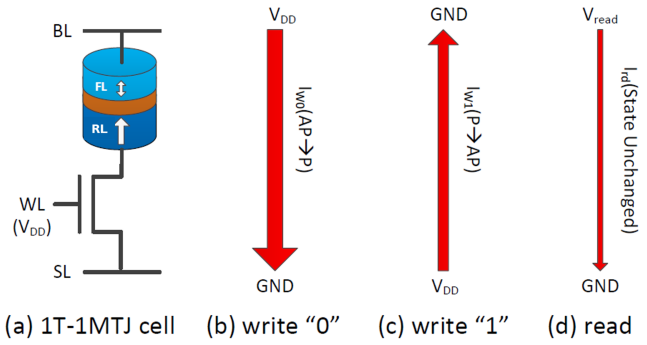


Fig. 2: STT-MRAM bit-cell architecture and operations.

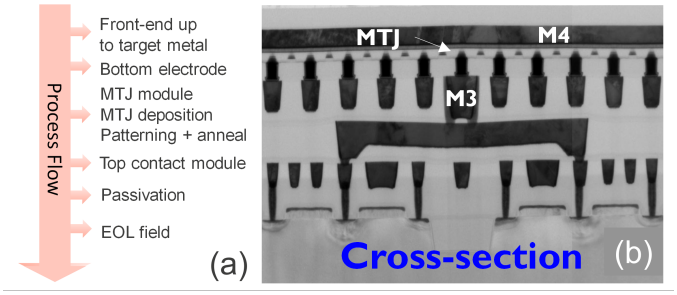


Fig. 3: (a) General manufacturing process of STT-MRAM (b) Cross-sectional TEM view of fully-integrated MRAM array with MTJ between M3 and M4 levels [20].

an inadvertent state change during read operations, known as read destructive fault [15], I_{rd} should be as small as possible; typically, I_{rd} is $< 50\%$ of the write currents for devices with a thermal stability $\Delta = 65$ [16].

For HPC applications such as last-level caches (LLC) that must operate at short write latencies (200 – 500 MHz), higher write current densities ($J_c \sim 5MA/cm^2$) can be accommodated to ensure 10-year data retention. On the other hand, LP applications require a strong reduction in the write current density ($J_c < 1MA/cm^2$) for similar retention, though at relaxed write speeds (~20 MHz). Thus, the tradeoff between write current (I_c) and thermal stability (Δ) becomes the key differentiator for STT-MRAM technology development. Enormous resources have been invested in the past decade at major foundries and tool suppliers to improve the write performance in STT-MRAMs, but the requirement of high TMR has so far limited the material development to interfacial CoFeB/MgO systems, presenting a unique challenge for further material exploration [17], [18].

STT-MRAM devices are commonly realized in a 1T-1MTJ bit-cell configuration, due to the overall area and performance benefits [19], [20]. Figure 2a shows a schematic representation of the device architecture, with an MTJ connected in series to a backend-of-the-line (BEOL) access transistor, typically NMOS. Row access is controlled with the word line (WL) connected to the NMOS gate, while the other two terminals are connected to a bit line (BL) and a source line (SL), respectively. They control write and read operations on the internal MTJ device depending on the magnitude and polarity of voltage applied across them, as shown in Figures 2(b-d).

C. Manufacturing Process and Challenges

STT-MRAM becomes an attractive option at advanced nodes by meeting stringent targets on chip density and scaled dimensions. Large scale integration enables area savings ~43% over existing SRAM memories at the 5nm node [7], which in turn requires a manufacturing process with high yield. Physical imperfections in the manufactured devices can result in defects, which negatively impact performance and yield. To mitigate these issues, it is necessary to understand the nature of the defects by testing and process innovation. Here, we review the standard process flow for the manufacturing of embedded STT-MRAM devices, as depicted in Figure 3a. The standard CMOS fabrication steps are retained for the access transistor, while the MTJ is integrated between the target metal levels, depending on the target application space.

As the processing up to the target metal remains standard, the challenges have been well-understood and addressed by foundries [7]. The following step involves the bottom electrode contact (BEC) deposition and patterning, which connects the MTJ stack to the bottom Cu metal lines. To ensure an ultra-smooth surface for the subsequent MTJ stack deposition, a chemical mechanical polishing (CMP) step is necessary [13]. High roughness on the BEC can lead to degraded TMR due to magnetic coupling effects across the MgO tunnel barrier such as orange peel coupling [21], and stray magnetic fields on the free layer from the pinning layers. The following steps are particularly critical for optimal device performance – MTJ stack deposition and patterning. State-of-the-art MTJ stacks are typically composed of ~15-20 metallic and insulating layers with polycrystalline interfaces. High data retention and TMR can be achieved by suitably engineering the crystallinity of CoFeB/MgO interfaces in the FL and RL. Through careful optimization of the layer thicknesses and annealing conditions (time and temperature), high perpendicular magnetic anisotropy (PMA) can be achieved [12], [22], which in turn enables a higher Δ and TMR at device-level. The patterning of MTJ stacks into nanopillars has been investigated by reactive-ion etch (RIE) and ion-beam etch (IBE) techniques [23], [24], with the latter demonstrating straighter pillar profiles, reduced footing and avoidance of magnetic layer corrosion. However, a significant challenge with IBE etch schemes is the strong sidewall redeposition encountered in the attempt to achieve straight sidewall profiles with some mitigation strategies available [23], [25]. The succeeding metallization steps are similar to standard BEOL processing of CMOS technology. The wafer with the patterned devices is then subjected to an end-of-line (EOL) magnetic field to set the magnetization direction of the RL in the MTJ devices. Figure 3b depicts a cross-sectional view of a 1 Mbit electrically-active STT-MRAM array with the MTJ integrated between M3 and M4 levels.

As efforts have ramped up to make STT-MRAM technology a manufacturable process, extensive research has been conducted into failure mechanisms due to processing defects at BEOL, FEOL and MTJ fabrication levels [25]. Significant advancements have been made in some areas, while challenges persist in others. We will briefly review the state-of-the-art of STT-MRAM technology today.

BEOL compatibility. During the typical dual damascene-based metallization processes, MTJ stacks are exposed to short bursts of annealing at $400\,^{\circ}\mathrm{C}$ depending on the metal level targeted. Therefore, higher thermal budget robustness is required in MTJ stacks, also referred to as 'BEOL compatibility'. Longer annealing times can lead to inter-diffusion within the MTJ stack layers, in turn causing PMA loss in the FL and RL. Through careful optimization of the layer thicknesses and introduction of diffusion-blocking layers, BEOL compatibility with TMR $\sim$200% has been demonstrated [20], [26]–[28].

Solder reflows. Standard packaging processes of SMT components requires a solder reflow process at $260\,^{\circ}\mathrm{C}$ for a duration of $\sim$3-5 minutes [3]. As packaging is carried out post-programming of the memory arrays, this requires a FL thermal stability of $\sim$40 at elevated temperatures to ensure the fidelity of the stored information. Recent reports at device-oriented conferences have demonstrated this capability by increasing the effective magnetic anisotropy of the FL, though this has been achieved in lieu of an increased write current [3], [29].

Write power consumption. To enable implementation at advanced logic nodes with lower drive capabilities, a reduction of the write current is still required. As detailed in the write operation of STT-MRAMs, this persistent challenge exists due to an inverse relationship between the write current and the targeted thermal stability. Due to the dependence on CoFeB/MgO interfaces for high TMR capabilities, most efforts till date have focused on optimization of the interface properties for write current reduction [7]. One approach to reduce the total write current by lowering the FL saturation magnetization (M_s) and optimizing film-level damping constant has shown promising results [17], [18], and may point to a re-thinking of the conventional FL design.

Scalability. Another approach to write current reduction is to reduce the CD of the patterned MTJ devices. As the FL volume reduces, the write current follows in a linear fashion. Such an approach requires not only tight MTJ pitch spacing ($<200\,\mathrm{nm}$), but also minimal damage of the patterned MTJs. A recent report demonstrated a 4 kbit array configuration of 40 nm nominal CD MTJs at a pitch spacing of 160 nm [30]. While an excellent demonstration, a large variability was observed in the measured resistance of the devices. This can be attributed to process-induced damage of the MTJs leading to an expected reduction in the device metrics. Thus, there is an urgent need for damage-minimizing patterning processes of scaled MTJs.

Patterning. Process-induced damage in scaled MTJs can occur due to the etch process itself, or the post-patterning oxygen treatment. These typically result in a reduction of the effective FL CD due to oxygen penetration in the MTJ along the CoFeB/MgO interfaces, also referred to as a 'bird's beak formation'. By optimizing the energy of the etching ions and controlling the amount of oxygen present on the sidewalls during the post-patterning treatment, the bird's beak formation can be significantly mitigated, as shown in Figure 4. Energy savings of $\sim$20% were reported at 1 Mbit-array level with the optimized etch process, in addition to reduced variability [20]. Modelling tools can further aid defect and damage-minimization strategies through cross-correlation with experiments. By correlating device metrics across an extensive design-of-experiments in the patterning space, damage-inducing processes can be optimized for improved yield and performance [31], [32].

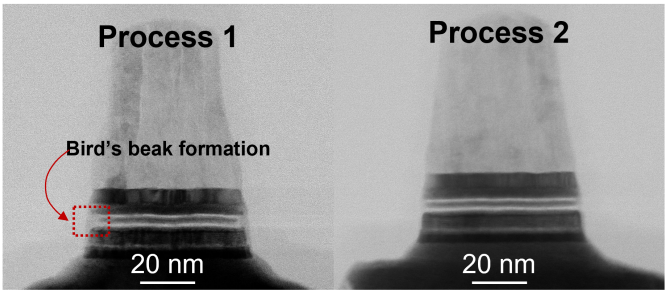


Fig. 4: Cross-sectional TEM images highlighting mitigation of bird's beak formation with etch optimization. pMTJ devices of 55nm CD were fabricated by the conventional etch process (Process 1) and an optimized etch process (Process 2) [20].

III. Design of Reliable Sensing Mechanisms in STT-MRAM

The read operation in resistive memories such as STT-MRAM is fundamentally different from conventional CMOS-based memories such as SRAM and DRAM. The read operation in STT-MRAM is typically performed by a *sense amplifier* (SA) and a reference value which is set in the middle of the two resistive states of the device. The SA amplifies the differential between the bit-line of the activated cell and another bit-line of a reference resistance, and determines whether a logical '1' or '0' was read.

However, all parts of the bit-cell array and read circuitry such as MTJs, MOS transistors and all circuit components are affected by *Process Variation* (PV). In addition to PV, thermal fluctuations have a significant impact on the electrical properties of MTJs [33], MOS transistors [34] and the reference circuitry. MOS transistors and MTJs have a different thermal behavior of their conductivity. MTJs are also affected differently depending on their magnetic state [35]. Therefore, the design and the structure of the reference circuitry has a significant impact on the read failure rate especially at high temperatures. In this section we evaluate different designs of the reference circuitry and their impacts on the reliability of the read operation in STT-MRAM.

A. STT-MRAM Read Operation

Figure 5 shows the concept of read circuitry in STT-MRAM with a generic reference. The read circuitry is usually placed

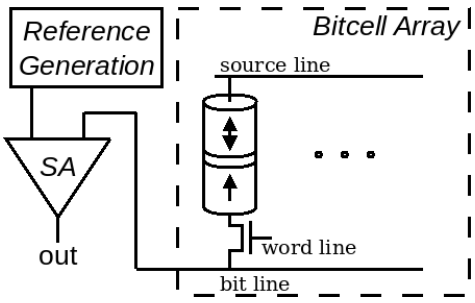
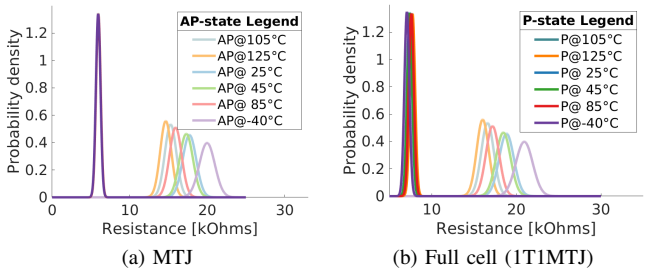


Fig. 5: STT-MRAM read concept. Multiple cells share the same sense amplifier (SA). A cell can be selected via the wordline. The selected cell is compared to a reference to read out the MTJ state.

column-wise and the resistive state of the selected cell in the row (corresponding to the read address in the wordline) is compared against the reference resistance. By passing small reading currents through reference and cell and compare them with the SA, the read decision can be made. The result of the comparison is the content of the read cell. The reference is usually placed in the middle of the two resistive states, i.e., $(R_P + R_{AP})/2$, however it should also track the temperature behavior of the bit-cells (including both MTJ and access transistor) and the rest of the read path, to ensure reliable read across all temperature ranges.

Figure 6 shows the distributions of MTJ and complete bit-cell (including the access transistor) variations under different temperatures. The effective resistance of MOS transistors shifts with temperature. The resistance of MTJs are nearly unaffected in P-state, however, their resistance in AP-state reduces with higher temperature. The combined effect results in nearly constant P-cell resistance, while AP-cell resistance reduces with temperature. Additionally, the temperature behavior of the cell effectively narrows down the sense margin between cell in P-state and cell in AP-state, and hence, makes reliable read much more challenging. Ideally, the reference should always be placed between the two distributions, otherwise a read fail can occur. A major challenge is that the thermal behavior of the reference could be completely different from the cell, depending on its design and structure.

In the following, we evaluate different reference resistance



(a) MTJ (b) Full cell (1T1MTJ)

Fig. 6: Temperature influence on MTJ and full cell (1T1MTJ) variations

Parameter	Value
VDD	0.8V
Nominal Temperature	27°C
MTJ radius	20 nm
Free/Oxide layer thickness	1.3/1.48 nm
RA	7.5 $\Omega\mu m^2$
TMR @ 0V	$>$ 150% for all temperatures

TABLE I: MTJ parameters and simulation setup

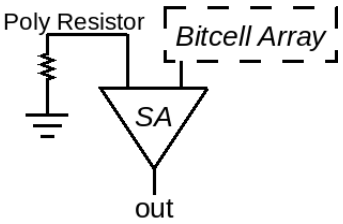


Fig. 7: Single Polysilicon reference resistance for the entire bitcell column.

designs, using Cadence Virtuoso. We have selected the GlobalFoundaries 22FDX CMOS process for the transistor models and use an MTJ model based on the description in [36]. The MTJ and the remaining simulation parameters can be found in Table I.

B. Poly-based Reference

Figure 7 shows the architecture of the reference scheme which is based on a single Polysilicon (Poly) resistor. The distribution of such Poly resistor under different temperatures is shown in Figure 8. As this figure shows, it is mostly temperature invariant and the resistance slightly increases with higher temperature. However, such temperature-invariant behavior becomes problematic when the cell resistance shifts over the temperature range. The other disadvantage of such a referencing scheme is that the area overhead of a Poly resistance is very high, so only one reference per SA can be used and it must be shared among multiple cells. It is possible to design a reference with with dummy reference cells consisting of a column of transistors and resistors [37]. This help to mitigate bitline parasitics during the sense operation.

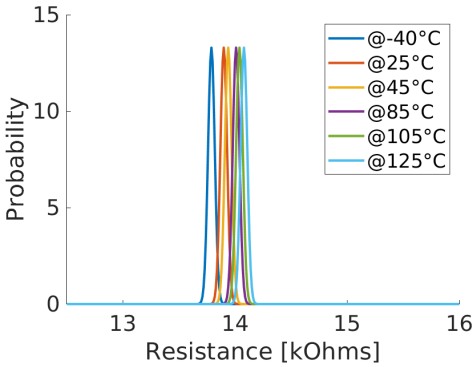


Fig. 8: Temperature behavior of Polysilicon resistor under process variation.

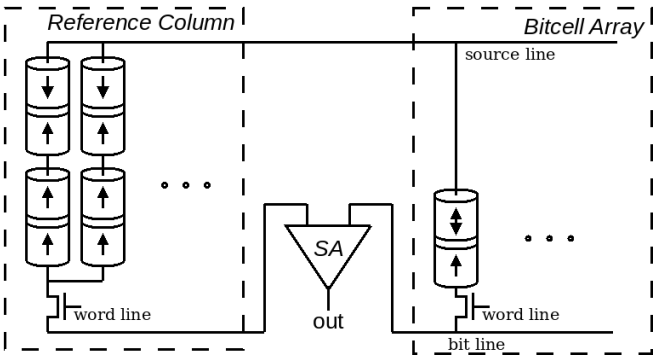


Fig. 9: 4-MTJ Compound reference architecture.

C. Compound MTJ Reference

Figure 9 shows the architecture of a referencing scheme which is made from an array of MTJs. A single reference uses two parallel connected substructures of two MTJs connected in series, a P-state MTJ and an AP-state MTJ, which can be enabled by the an access transistor. This access transistor is connected to the same wordline as the read cell. In practice, the SA can be shared among multiple bitlines with a multiplexer to amortize the overhead of the reference column. The resistance distribution of the 4MTJ compound reference under different temperatures is depicted in Figure 10. There is one reference per wordline/column of reference cells, each is built from 4 MTJs and an access transistor. This structure results in a resistance of $(R_{AP} + R_P)/2$ with two P-state MTJs and two AP-state MTJs. The effective resistance of this structure reduces with higher temperature. The temperature dependency of this reference is less than the 1T1MTJ cell, because of the P-state MTJs in the reference structure. In terms of area overhead, this structure is smaller than the single Poly resistor, and hence can easily be integrated in a dedicated reference bitline.

D. Hybrid Resistor-MTJ Reference

While the two architectures shows above have their own shortcomings, Figure 11 shows the architecture of a hybrid

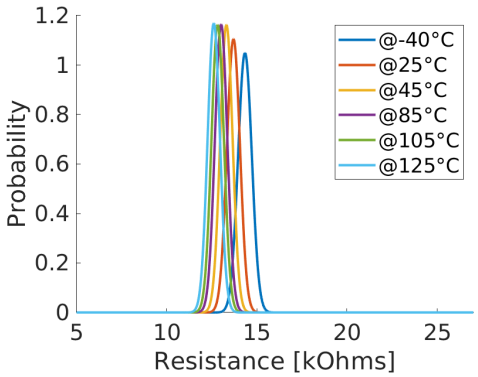


Fig. 10: Temperature behavior of 4-MTJ compound reference under process variation.

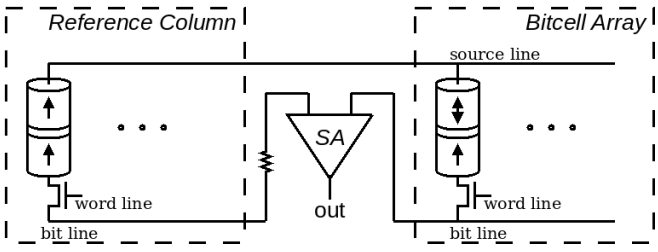


Fig. 11: Hybrid Poly-resistor + 1T1MTJ reference column architecture.

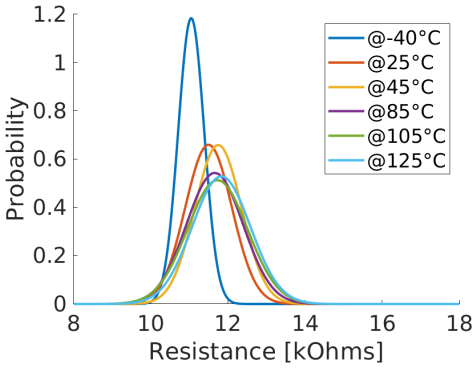


Fig. 12: Temperature behavior of hybrid poly resistor + 1T1MTJ reference under process variation.

referencing scheme which is made from a combination of a Poly resistor as well as a column of 1T1MTJ cells. As the 1T1MTJ cell is about 30% smaller than the compound cell, this offers a design choice between the good temperature behavior of the compound cell and the improved area usage of the hybrid reference.

The resistance distribution of one Poly resistor and one 1T1MTJ cell under different temperatures is shown in Figure 12. It is also possible to replace the poly resistor with a Negative Temperature Coefficient (NTC) resistor, which combines temperature-independent and temperature-dependent parts, while reusing the 1T1MTJ cell for reference [38].

E. Analysis and Comparison

Table II compares *Read Decision Failure (RDF)* rate across the temperature range for the different reference structures discussed above. In the Poly-only based reference, the Poly resistor is shared among all cells, which needs to be tuned based on all cells. The poly resistor does not follow the temperature behavior of bit-cells, therefore the RDF increases considerably with higher temperature. The compound reference is slightly larger than 1T1MTJ cell, and can better track the temperature

	-40°C	25°C	45°C	85°C	105°C	125°C
Poly	6.3e-13	4.5e-09	6.8e-08	1.7e-05	1.9e-04	1.8e-03
Compound	9.2e-07	9.8e-06	5.6e-06	5.6e-05	1.3e-04	2.6e-04
Hybrid	3.4e-10	1.1e-05	2.1e-06	2.0e-04	7.5e-04	1.5e-03

TABLE II: Read decision failure probability rates for different reference structures at different temperatures.

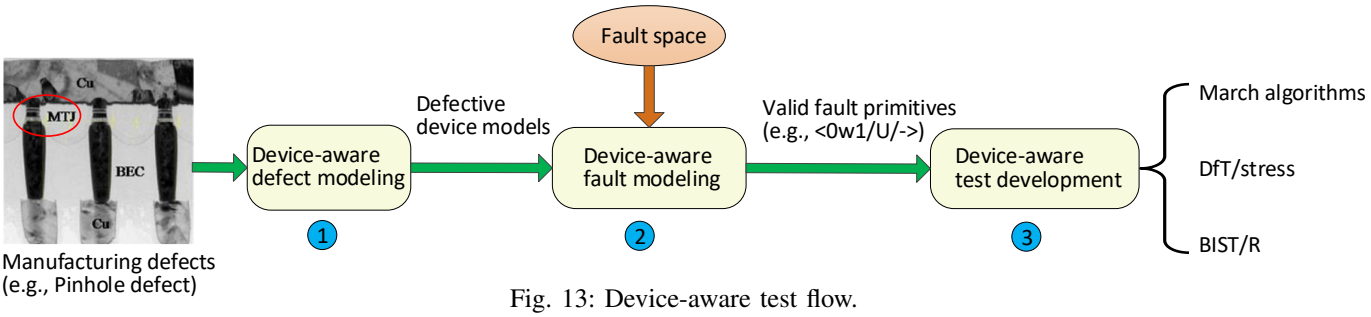


Fig. 13: Device-aware test flow.

behavior of the cell compared to the poly structure, thus only slowly degenerates with temperature. The Hybrid structure is similar to the Poly-only structure. The additional reference column helps to offset bitline parasitic, yet the temperature variation is not mitigated.

In direct comparison, the discussed references differ mainly in their ability to track the temperature variation, their area footprint and their possible handling of bitline parasitic. From a temperature perspective, the compound reference cell allows for the most consistent failure rate over the entire temperature range. However, the compound reference uses more MTJs per reference cell than the hybrid reference. This leads to an overhead of around 45% per reference cell, which is specifically impactful for larger arrays. Yet, both the compound and the hybrid reference increase their area based on the number of reference cells. On the other hand, this allows to mitigate bitline parasitics by enabling reference cells with similar positions with respect to the read cell and the sense amplifier. The single poly reference is not able to mitigate these parasitics on its own. A dummy bitline with only an access transistor per wordline can be used to address this, adding to the overall area of the solution.

IV. Device-Aware Test for STT-MRAM

In this section, we first discuss the limitation of the conventional test approach based on injection of linear resistors. Thereafter, we introduce the newly-proposed device-aware test approach. This new test approach is then applied to three key types of STT-MRAM devices defects: synthetic antiferromagnet flip, intermediate state, and pinhole defects.

A. Limitation of the Conventional Test Approach

The conventional test approach models any defect in an STT-MRAM device as a linear resistor (e.g., open and short), as can be found in the prior works [5], [13], [39]. The forming mechanism, location, occurrence probability are never taken into account and manifested as a difference in the defect model. In addition, emerging devices are typically nonlinear and have unique properties such as magnetic properties. In some recent works [40], [41], it has been demonstrated with both silicon measurements and circuit simulations that injecting linear resistors is not qualified to model defects in STT-MRAM and RRAM devices; the traditional test approach may even lead to non-existent fault models and vain test solutions.

B. Device-Aware Test Approach

We have proposed a new test approach: *Device-Aware Test* (DAT) to specifically address device-internal defects [41]. DAT targets DPPB-level test development and it is composed of three steps (see Figure 13) as follows.

1) Device-aware defect modeling. First, a device-internal defect needs to be physically analyzed and characterized to understand its forming mechanism, location, occurrence probability, and the key technology parameters that are affected. Thereafter, the effects of the defect are quantitatively incorporated into these technology parameters. Second, the defect-induced changes in the technology parameters are mapped into the device's electrical parameters. This allows us to obtain a parameterized defective device model. Third, the defective model can be further calibrated using silicon data if available.

2) Device-aware fault modeling. First, a complete fault space which describes all possible faults in emerging memories is defined. This is achieved by extending the conventional *Fault Primitive* (FP) notation $\langle S/F/R \rangle$ [42]. S denotes the operation sequence that sensitizes a fault, while R describes the readout result if the last operation in S is a read. F denotes the value in the faulty cell after S is applied. Apart from logic '0' and '1', F can also be 'U' (undefined), 'L' (extreme low) and 'H' (extreme high) states due to defects, as indicated by silicon data shown in [25]. Based on the extended FP definition, all memory faults are classified into two categories: *Easy-to-Detect* (EtD) faults and *Hard-to-Detect* (HtD) faults. EtD faults are those which can be detected by applying normal write and read operations, i.e., March tests, while HtD faults refer to those which cannot be guaranteed to be detected by March tests. Second, a systematic fault analysis based on circuit simulations for each targeted defect is conducted to derive realistic faults within the pre-defined fault space.

3) Device-aware test development. The accurate and realistic faults obtained from the previous step are used to develop test solutions at DPPB level. Specifically, EtD faults can simply be detected by March tests; HtD faults, however, need special DfT or stress tests. The clear mapping relations between physical defects and fault models enable us to not only reduce test escapes and time but also speed up yield learning and defect diagnosis [41].

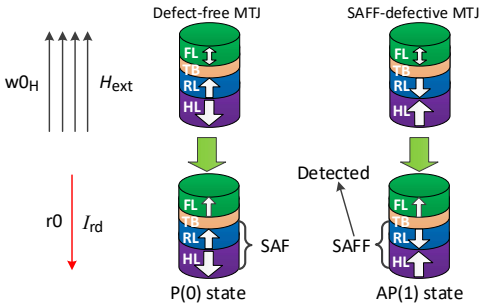


Fig. 14: Testing SAFF defects using a magnetic write '0' operation ($\mathrm{w0_H}$).

C. DAT for Synthetic Anti-Ferromagnet Flip (SAFF) Defects

SAFF defects mean that the SAF structure flips in MTJ devices, as illustrated in Figure 14. A probable cause of SAFF defects is an initial hard layer reversal due to inhomogeneities in coercivity arising during device fabrication [43].

1) Device-aware defect modeling: Based on comprehensive device characterization, we observed that SAFF defects lead to a flip of the intra-cell stray field $H_{\mathrm{s\_intra}}$ at the free layer of MTJ. Therefore, we physically modeled $H_{\mathrm{s\_intra}}$ [43]; In addition, we extended this model to inter-cell stray field $H_{\mathrm{s\_inter}}$ which is generated from neighboring cells. The effect of SAFF defects were then incorporated into $H_{\mathrm{s\_intra}}$ and $H_{\mathrm{s\_inter}}$. Thereafter, we mapped the SAFF-induced change in the stray fields to two key electrical parameters: I_{c} and t_{w}. By fitting to experimental data and model optimization, we obtained a defective MTJ model for SAFF defects.

2) Device-aware fault modeling: The SAFF-defective MTJ model was used to develop accurate and realistic fault models using SPICE-based circuit simulations. Figure 15 compares the fault modeling results using DAT and the conventional test approaches. DAT results in a HtD fault: intermittent passive neighborhood pattern sensitive fault ($\mathrm{PNPSF1_i}$), whereas the conventional test leads to four EtD faults as shown in the right circle. This indicates that these four faults are not qualified to cover SAFF defects in STT-MRAMs. Accordingly, the March tests targeting these four EtD faults obviously cannot guarantee the defection of SAFF defects.

3) Device-aware test development: To detect SAFF defects, we proposed a magnetic march test as follows:

$$\{\Updownarrow(\mathrm{w0_H}); \Updownarrow(\mathrm{r0})\} \text{ or } \{\Updownarrow(\mathrm{w1_H}); \Updownarrow(\mathrm{r1})\}.$$

Here, the first element $\mathrm{w0_H}$ ($\mathrm{w1_H}$) indicates a magnetic write '0' ('1') operation; i.e., an external field H_{ext} is applied to

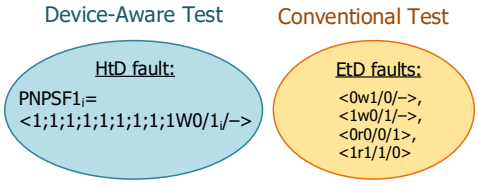


Fig. 15: Comparison of sensitized FPs due to SAFF defects: device-aware test vs. conventional test based on linear resistors.

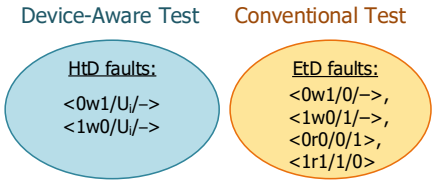


Fig. 16: Comparison of sensitized FPs due to IM state defects: device-aware test vs. conventional test based on linear resistors.

switch the MTJ state rather than driving an electric current through the MTJ device. Figure 14 illustrates the test process to guarantee the detection of SAFF defects.

D. DAT for Intermediate (IM) State Defects

IM state defects mean that a third abnormal state appears in MTJ devices, in addition to the bi-stable P and AP states; $R_{\mathrm{P}} < R_{\mathrm{IM}} < R_{\mathrm{AP}}$. The root causes can be attributed to multi-domain structure of the FL induced by the dipole field and large device size etc. [44].

1) Device-aware defect modeling: Our experimental results show that the occurrence of IM state defects is probabilistic depending on the switching direction, applied bias voltage, and device size; the retention time of IM state is larger than $10\,\mathrm{ms}$ after the removal of write pulses [44]. Therefore, we physically modeled an IM state defect by splitting the free layer into two regions: 1) P-state region and 2) AP-state region. Additionally, we modeled the probabilistic occurrence using Bernoulli distribution and the retention time using a statistic thermal stability model [44]. The physical modeling results were then mapped to key electrical parameters: R_{IM}, I_{c} and t_{w}. By fitting to experimental data and model optimization, we obtained a defective MTJ model with three resistive states.

2) Device-aware fault modeling: Figure 16 presents the fault modeling results. Two HtD faults were observed using our DAT approach; they are intermittent write transition faults: $\mathrm{W1TFU_i}$=⟨0w1/$\mathrm{U_i}$/–⟩ and $\mathrm{W0TFU_i}$=⟨1w0/$\mathrm{U_i}$/–⟩. Again, there is no overlap in the Venn diagram, meaning that IM state defects exhibit unique faulty behaviors which cannot be covered by linear resistor models.

3) Device-aware test development: To detect IM state defects, we proposed the following March algorithm with a weak

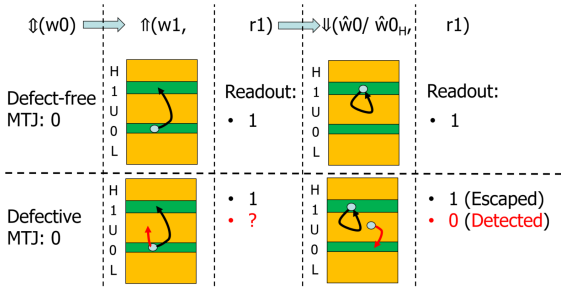


Fig. 17: Proposed March algorithm with a weak write operation $\hat{\mathrm{w}}0/\hat{\mathrm{w}}0_{\mathrm{H}}$.

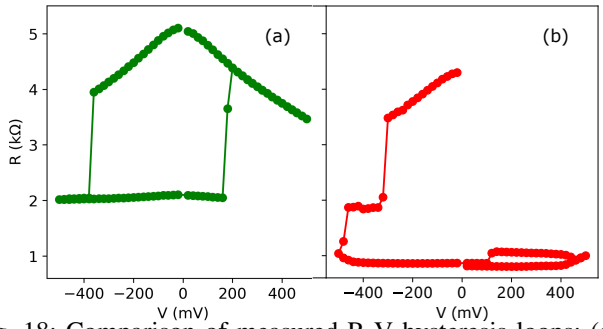


Fig. 18: Comparison of measured R-V hysteresis loops: (a) a good MTJ and (b) a defective MTJ.

write operation, as illustrated in Figure 17:

$$\{\Updownarrow\ (\mathrm{w0}); \Uparrow (\mathrm{w1}, \mathrm{r1}); \Downarrow (\widehat{\mathrm{w}}0/\widehat{\mathrm{w}}0_{\mathrm{H}}, \mathrm{r1})\}.$$

Here $\widehat{\mathrm{w}}0$ denotes a write '0' operation with a relatively weak current; it can be implemented by reducing current amplitude or duration. Similarly, $\widehat{\mathrm{w}}0_{\mathrm{H}}$ means a write '0' operation using a weak magnetic field. The weak write induces an IM→P transition while it is not strong enough to change AP state.

E. DAT for Pinhole Defects

Pinhole defects in the MgO tunnel barrier of MTJ take place during multi-layer deposition; they may form due to unoptimized deposition of thin films [40].

1) Device-aware defect modeling: Figure 18a shows the measured R-V loop of a good MTJ and Figure 18b shows the measurement data of a defective MTJ. Due to the non-linear behavior of MTJ, it is impossible to model the impact of a physical defect on the R-V loop simply by adding a linear resistor. For an MTJ device, its magnetic properties are as important as electrical ones; but linear resistors are unable to capture defect-induced changes in magnetic properties [40]. Hence, a new test approach is required to develop high-quality yet cost-efficient test solutions for device-internal defects. Our experimental results on real fabricated MTJ devices show that pinhole defects result in RA and TMR degradation [40]. With a small pinhole filled with CoFeB material from the above free layer, the tunneling current across the MgO barrier is shunted by a high-conductance path via the pinhole. In addition, small pinhole defects deteriorate over time because of Joule heating and/or an electric field across the pinhole circumference. Therefore, we modeled pinhole defects by incorporating their effects into the two technology parameters RA and TMR and

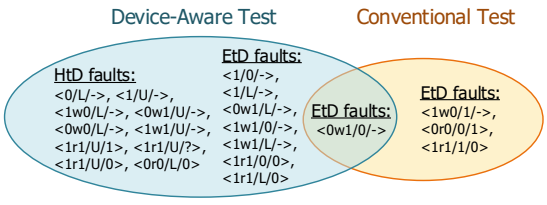


Fig. 19: Comparison of sensitized FPs due to pinhole defects: device-aware test vs. conventional test based on linear resistors.

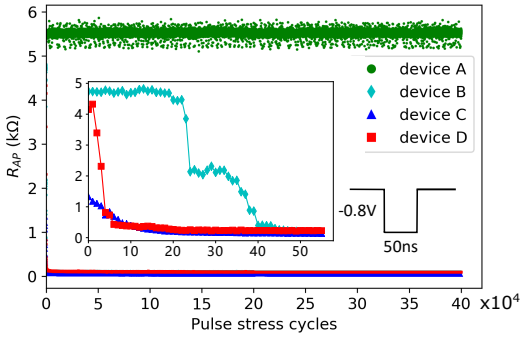


Fig. 20: Stress test for detecting small pinhole defects (HtD faults).

subsequently into the electrical parameters: R_{P}, R_{AP}, I_{c} and t_{w}. By fitting to experimental data and model optimization, we derived a pinhole-parameterized MTJ model.

2) Device-aware fault modeling: Figure 19 shows the fault modeling results for pinhole defects. It can be seen that our DAT approach results in 18 faults. Among these faults, 17 are not observed with resistor models while only a single EtD fault (W1TF0=⟨0w1/0/–⟩) is in overlap. Among the unique 17 faults generated by our DAT approach, 10 are HtD faults and the rest 7 are EtD faults.

3) Device-aware test development: EtD faults and HtD faults require different test solutions. March tests including the element ⇳(w1, r1) can guarantee the detection of all sensitized EtD faults. However, HtD faults require dedicated DfT or stress tests to detect them. For instance, a hammering write '1' stress test can be used as shwon in Figure 20 to intentionally enlarge the pinhole size and transform his faulty behavior from HtD faults to EtD ones, making its testing with just a March test feasible [25].

V. Conclusion

STT-MRAM has a promising potential to become the primary candidate for on-chip memories by replacing the struggling conventional memory technologies. However, various technology, reliability and test challenges of STT-MRAM need to be addressed for its widespread commercialization. This paper first discussed the manufacturing process, reliability and test challenges of STT-MRAM. Then, reliable sensing mechanism and device-aware testing techniques are presented to address those challenges to make STT-MRAM as a prime choice for embedded memory.

References

[1] International roadmap for devices and systems (irds). [Online]. Available: https://irds.ieee.org/

[2] D. Apalkov *et al.*, "Magnetoresistive random access memory," *Proceedings of the IEEE*, 2016.

[3] W. Gallagher *et al.*, "22nm stt-mram for reflow and automotive uses with high yield, reliability, and magnetic immunity and with performance and shielding options," in *IEDM*, 2019.

[4] S. Van Beek *et al.*, "Edge-induced reliability & performance degradation in stt-mram: an etch engineering solution," in *International Reliability Physics Symposium (IRPS)*, 2021.

[5] L. Wu *et al.*, "Survey on stt-mram testing: failure mechanisms, fault models, and tests," *arXiv preprint arXiv:2001.05463*, 2020.

[6] S. Yuasa *et al.*, "Future prospects of mram technologies," in *International Electron Devices Meeting*, 2013.

[7] S. Sakhare *et al.*, "Enablement of stt-mram as last level cache for the high performance computing domain at the 5nm node," in *International Electron Devices Meeting (IEDM)*, 2018.

[8] S. M. Nair *et al.*, "Mitigating read failures in stt-mram," in *VLSI Test Symposium (VTS)*, 2020.

[9] Y. Huai *et al.*, "Spin-transfer torque mram (stt-mram): Challenges and prospects," *AAPPS bulletin*, 2008.

[10] T. Na, S. H. Kang, and S.-O. Jung, "Stt-mram sensing: a review," *Transactions on Circuits and Systems II: Express Briefs*, 2020.

[11] S. Han *et al.*, "Reliability of stt-mram for various embedded applications," in *International Reliability Physics Symposium (IRPS)*, 2021.

[12] G. S. Kar *et al.*, "Co/Ni based p-MTJ stack for sub-20nm high density stand alone and high performance embedded memory application," in *IEDM*, 2014.

[13] L. Wu *et al.*, "Electrical modeling of STT-MRAM defects," in *ITC*, 2018.

[14] A. V. Khvalkovskiy *et al.*, "Basic principles of STT-MRAM cell operation in memory arrays," *J. Phys. D: Appl. Phys.*, 2013.

[15] W. Butler *et al.*, "Spin-dependent tunneling conductance of Fe — MgO — Fe sandwiches," *Physical Review B*, 2001.

[16] W. Zhao *et al.*, "Design considerations and strategies for high-reliable STT-MRAM," *Microelectronics Reliability*, 2011.

[17] J. M. Iwata-Harms *et al.*, "Ultrathin perpendicular magnetic anisotropy cofeb free layers for highly efficient, high speed writing in spin-transfer-torque magnetic random access memory," *Scientific reports*, 2019.

[18] T. S. Santos *et al.*, "Ultrathin perpendicular free layers for lowering the switching current in stt-mram," *Journal of Applied Physics*, 2020.

[19] Y. Lee *et al.*, "Highly scalable stt-mram with mtjs of top-pinned structure in 1t/1mtj cell," in *Symposium on VLSI Technology*, 2010.

[20] S. Rao *et al.*, "Stt-mram array performance improvement through optimization of ion beam etch and mtj for last-level cache application," in *International Memory Workshop*, 2021.

[21] E. I. Vatajelu *et al.*, "Challenges and solutions in emerging memory testing," *IEEE Transactions on Emerging Topics in Computing*, 2017.

[22] S. Ikeda *et al.*, "A perpendicular-anisotropy cofeb–mgo magnetic tunnel junction," *Nature materials*, 2010.

[23] T. Endoh *et al.*, "Etch process technology for high density stt-mram," in *International Magnetics Conference (INTERMAG)*, 2018.

[24] R. Islam *et al.*, "Dry etching strategy of spin-transfer-torque magnetic random access memory: A review," *Journal of Vacuum Science and Technology B*, 2020.

[25] L. Wu *et al.*, "Defect and fault modeling framework for STT-MRAM testing," *Trans. Emerg. Topics Comput.*, 2019.

[33] L. Zhang *et al.*, "Reliability and performance evaluation for stt-mram under temperature variation," in *International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems*, 2016.

[26] J. Swerts *et al.*, "Beol compatible high tunnel magneto resistance perpendicular magnetic tunnel junctions using a sacrificial mg layer as cofeb free layer cap," *Applied Physics Letters*, 2015.

[27] Y.-D. Chih *et al.*, "A 22nm 32mb embedded stt-mram with 10ns read speed, 1m cycle write endurance, 10 years retention at 150°c and high immunity to magnetic field interference," in *International Solid- State Circuits Conference - (ISSCC)*, 2020.

[28] Y. Song *et al.*, "Demonstration of highly manufacturable stt-mram embedded in 28nm logic," in *International Electron Devices Meeting (IEDM)*, 2018.

[29] V. B. Naik *et al.*, "Manufacturable 22nm fd-soi embedded mram technology for industrial-grade mcu and iot applications," in *International Electron Devices Meeting (IEDM)*, 2019.

[30] D. Edelstein *et al.*, "A 14 nm embedded stt-mram cmos technology," in *International Electron Devices Meeting (IEDM)*, 2020.

[31] J. Jeong and T. Endoh, "Novel oxygen showering process (osp) for extreme damage suppression of sub-20nm high density p-mtj array without ibe treatment," in *Symposium on VLSI Technology (VLSI Technology)*, 2015.

[32] N. Xu *et al.*, "Rare-failure oriented stt-mram technology optimization," in *Symposium on VLSI Technology*, 2018.

[34] F. Shoucair, "Design consideration in high temperature analog cmos integrated circuits," *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, 1986.

[35] X. Bi *et al.*, "STT-RAM Cell Design Considering CMOS and MTJ Temperature Dependence," *Transactions on Magnetics*, 2012.

[36] A. Mejdoubi *et al.*, "A compact model of precessional spin-transfer switching for MTJ with a perpendicular polarizer," in *MIEL*, 2012.

[37] A. Antonyan *et al.*, "Embedded MRAM Macro for eFlash Replacement," in *International Symposium on Circuits and Systems (ISCAS)*, 2018.

[38] E. M. Boujamaa *et al.*, "A 14.7Mb/mm2 28nm FDSOI STT-MRAM with Current Starved Read Path, 52Ω/Sigma Offset Voltage Sense Amplifier and Fully Trimmable CTAT Reference," in *Symposium on VLSI Circuits*, 2020.

[39] S. M. Nair, R. Bishnoi *et al.*, "Defect injection, fault modeling and test algorithm generation methodology for STT-MRAM," in *ITC*, 2018.

[40] L. Wu *et al.*, "Pinhole defect characterization and fault modeling for STT-MRAM testing," in *ETS*, 2019.

[41] M. Fieback *et al.*, "Device-aware test: A new test approach towards DPPB level," in *ITC*, 2019.

[42] S. Hamdioui and A. Van De Goor, "An experimental analysis of spot defects in SRAMs: realistic fault models and tests," in *ATS*, 2000.

[43] L. Wu *et al.*, "Characterization, modeling and test of synthetic antiferromagnet flip defect in STT-MRAMs," in *ITC*, 2020.

[44] L. Wu *et al.*, "Characterization and fault modeling of intermediate state defects in STT-MRAM," in *Design, Autom. & Test in Europe Conf. & Exhib.*, 2021.