EMACS: Efficient MBIST Architecture for Test and Characterization of STT-MRAM Arrays

Insik Yoon   Ashwin Chintaluri   Arijit Raychowdhury
School of Electrical & Computer Engineering,
Georgia Institute of Technology, Atlanta, Georgia
Email: { iyoon, achintaluri3 }@gatech.edu, arijit.raychowdhury@ece.gatech.edu

***Abstract*—Spin Transfer Torque Magnetic Random Access Memory (STT-MRAM) is an emerging memory technology which exhibits non-volatility, high density, high endurance and nanosecond read and write times. These attributes of STT-MRAM make it suitable for last level embedded caches. However, defect models, faults and test architectures for emerging memory technologies are relatively unexplored. This is further aggravated by the fact that STT-MRAM, like other post-CMOS technologies rely on novel physics of operation, which can result in unexplored read, write and retention fault models. In particular, the stochastic retention failure of STT-MRAM has a large impact on the test time. Conventional test schemes for retention of STT-MRAM need to be redesigned and optimized for testing large STT-MRAM arrays. This paper presents a comprehensive analysis of read, write and retention tests in STT-MRAM arrays. Resistive and capacitive defects and the corresponding faults are studied. A novel MBIST architecture and associated circuits are presented for measuring thermal stability (and hence retention times) in STT-MRAM bits for characterization and manufacturing tests, amidst variations and magnetic coupling. Trade-offs between fault localization, area overhead and test-times are presented.**

I. Introduction

In STT-MRAM, a Magnetic Tunnel Junction(MTJ) stores a single bit of information per cell. An MTJ comprises of a thin insulator (MgO) which is sandwiched between a ”fixed” ferromagnetic layer (CoFeB based) whose magnetic moment is pinned to one direction and a ”free” ferromagnetic layer whose moment changes direction based on applied external energy(field). When a spin-polarized current passes through a mono-domain ferromagnet, it attempts to polarize the current in its preferred direction of magnetic moment. As the ferromagnet absorbs some of the angular momentum of the electrons, it creates a torque that causes a flip in the direction of magnetization in the ferromagnet.The basic STT-MRAM cell comprises of an access transistor and a magnetic tunneling junction (MTJ) as shown in Fig. 1. The relative alignment of the ferromagnetic layers results in a high resistance path (Anti-parallel) or a low resistance (Parallel) path for the current, giving a notion of binary storage. Depending on the direction of current through the access transistor, the free layer magnetization flips from Anti-parallel to Parallel state or vice-versa resulting in change of bit from 1 to 0 or 0 to 1 respectively. The bias conditions applied for the write and read operations are shown in Fig. 1(a). As one can see, the write operation is bidirectional where the bit-line (BL) or source line (SL) are pulled high and the other pulled low. The read operation is unidirectional where a weak-current is passed through the cell and its resistive state is sensed using either a constant current scheme (Fig. 1)(b) or a BL discharge scheme [1]. The MTJ can either be an In-plane MTJ (I-MTJ) with magnetic anisotropy in plane due to shape anisotropy or a Perpendicular plane MTJ (P-MTJ) where magnetic anisotropy is aligned out of plane, independent of the shape of the free layer [2]. The relative merits and demerits of the two structures are being extensively studied [2], [3], [4] and in this paper, we will discuss both varieties of bitcells. The non-volatility of the MTJ is a key feature in STT-MRAMs and high thermal stability of the cell in scaled nodes is desired.

STT-MRAM arrays are expected to suffer from read and write failures which are induced by electrical defects and process variations. The role of variations in read and write have been extensively studied, including prior work by the authors [1]. However, the role of resistive and capacitive defects and coupling faults is relatively unexplored (except for preliminary work in [5]); and in the first part of this paper, we will explore (1) the types of defects, (2) their manifestation as faults, and (3) enhancements to memory test patterns to activate these faults. Apart from read and write faults, STT-MRAMs can also suffer from retention failures. The non-volatility (or retention characteristics) of the bit can be measured by the thermal stability factor Δ. [6][7] describe retention failure as a bit-flip in a cell caused by thermal noise. The thermal activation model of STT-MRAM in [6] suggests that a bit flip has a poisson distribution with time constant of $\tau.e^{\Delta}$ where $\tau \approx 1ns$. Conventional test methods for retention have very large number of test times. In the second part of this paper we explore worst case test patterns and propose a Memory Built In Self-Test (MBIST) architecture that can detect the retention failures along with read and write faults in a time-efficient manner. We propose EMACS as an efficient MBIST architecture that can perform in-situ read, write and retention (stochastic test) tests on STT-MRAM arrays. This work is based on a vertically-integrated, device to array modeling infrastructure that we have developed to analyze the physics of MTJ operation (amidst variations and thermal fluctuations) for various types of MTJ cells.

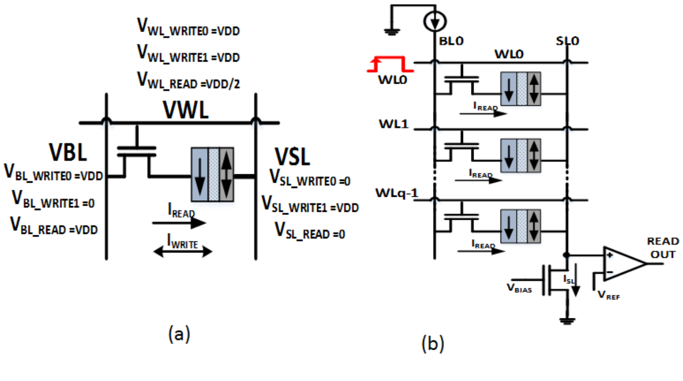


Figure 1: (a) Write operation and conventional voltage read operation of STT-MRAM, (b) Current read operation of STT-MRAM

978-1-4673-8773-6/16/$31.00 ©2016 IEEE

The rest of the paper is divided as follows. In section II, the modeling infrastructure is described. In Section III we discuss defects and corresponding fault models that can occur in an array. The role of magnetic coupling is discussed in section IV. In section V we discuss the challenges and necessary test patterns for retention testing. The MBIST architecture and circuits are discussed in section VI. Performance Analysis of EMACS is discussed in section VII. Some practical challenges and efficiency of EMACS is discussed in section VIII. Finally conclusions are drawn in section IX.

II. Modelling and Infrastructure Development

For the sake of brevity, we present a short description of the modeling infrastructure for array level analysis. Further details can be obtained in [1]. In the simplest form, STT-MRAM can be modeled as a 1 transistor, 1 resistor model in which the resistance changes instantly between a high state and a low state when the correct bias is applied. But such a static model doesn't capture the continuously varying resistance and the current through the device [8]. The magnetic dynamics under the influence of a current is described by the Landau Lifshitz and Gilbert (LLG) equation [8]. The free-layer magnetic-moment $m(t)$ evolves in the presence of a torque experienced because of uniaxial anisotropy field (T_U), easy plane anisotropy field (T_K), spin transfer torque from injected electrons (T_S) and thermally induced random torque component (T_{THERM}). The LLG under the total torque (T) is expressed in polar coordinates as:

$$\frac{1+\alpha^2}{\gamma H_\mathrm{k}}\begin{bmatrix}\frac{\partial\theta}{\partial t}\\ \frac{\partial\phi}{\partial t}\end{bmatrix} = \vec{T_\mathrm{U}} + \vec{T_\mathrm{K}} + \vec{T_\mathrm{S}} + \vec{T}_\mathrm{THERM} \tag{1}$$

where α is the LLG damping coefficient and γ is the gyromagnetic ratio. In a manner described in [14], the switching current density (J_{C0}) at T=0K can be described by:

$$J_\mathrm{c0} = \frac{\hbar}{2e}\frac{\alpha}{\eta}(tM_\mathrm{s}H_\mathrm{k})(1+\frac{2\pi M_\mathrm{s}}{H_\mathrm{k}}) \tag{2}$$

where e is the electronic charge, η is the polarization of the injected current, and t is the thickness of the free layer. LLG is implemented in a SPICE compatible environment [1] and it allows simulation of large arrays with control of material parameters as well as with statistical variations applied. For simulating the effect of thermal fields under long write times ($> 10ns$), we use a combination of SPICE simulations and model based approach [8], where the thermal stability under the application of a current density ($J_{applied}$) is modeled as $\Delta_{modified} = \Delta(1 - J_{applied}/J_{c0})$. The modeling infrastructure is combined with smart Monte Carlo techniques and allows parametric analysis under variations, defects and thermal fluctuations.

III. Electrical Defects and Fault Models

A. Read and Write Faults

There have been numerous studies on SRAM fault models and defects [9] and limited work on resistive bridging faults in ReRAMs and memristor [10][11]. These works have addressed static fault models and a limited number of dynamic fault models. In [1] the authors have identified both static and dynamic faults occurring both due to defects and due to variations in an STT-MRAM bit cell by using the full LLG solver model described in Section II. The faults during read and write are summarized in Table I. It should be noted that both I-MTJs and P-MTJs exhibit similar electrical characteristics, and hence the defect models and fault activations are also identical.

Fault Model	Affects	Key Cause
Transition Fault (TF)	WR	Relative Weak WR current due to stray resistive paths
Coupling Fault(CF)	WR	Neighboring cells switching
Stuck At Fault(SF)	WR	T0 , WL stuck at VDD or GND
Incorrect Read Fault (IRF)	RD	Current miscorrelation due to defects affecting WL,BL
Read Disturb Fault (RDF)	RD	Electrical disturbance at T0 node due to larger than normal RD current

Table I: Defects induced faults

B. Resistive Defects

It is observed that resistive shorts in the energy storage node or in the WL result in faults that occur when a particular pattern is written into adjacent bit cells. For example, when writing 0 to both cell0 and cell1, if there is a bridge between BL0 and SL1 this leads to weakening of BL0 leading to a TF0 (transition to 0 fault). The authors in [1] have explored resistive defects only. For the sake of completion we list all the possible resistive faults in a 2X2 array window and the conditions that activate these faults in Table II. We use these faults in the next section to propose enhancements to March C- test for complete coverage.

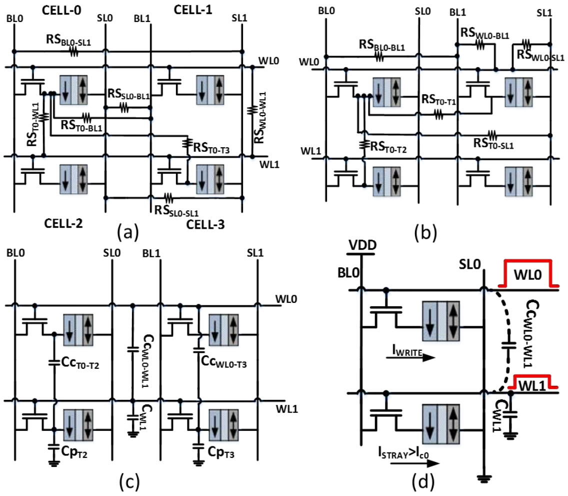


Figure 2: (a) and (b) Inter-cell Resistive defects (c) Most aggressive capacitive defects (d) Capacitive coupling between WLs that may cause CF

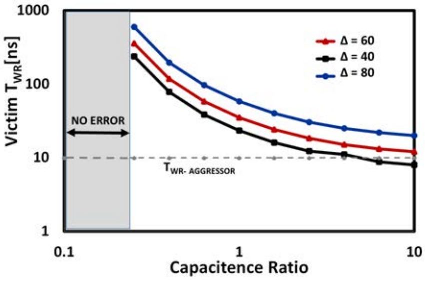


Figure 3: Victim Write time vs. Cap. ratio Cc/Cg

C. Capacitive defects

The role of capacitive coupling defects in STT-MRAM array, hitherto unexplored, can also lead to coupling faults. We introduce coupling capacitances between nodes of the cell (Fig. 2(c),(d)) and observe the effect of crosstalk on the write process of the cell. We have seen that the capacitive defects between WL and the internal node T0 are the most aggressive as they cause unintended writes to the neighboring cells. The ratio of the crosstalk capacitance (Cc) to the ground capacitance (Cg) is varied to observe the trend of how write time varies with strength of capacitive coupling. This quantifies the impact of the aggressor cell on the victim cell. It is seen that the capacitive coupling is relatively weak compared to resistive bridges; but if the victim is a weak cell, the effect is prominent. Of particular importance is WL coupling (Fig. 2(d)). Fig. 3 illustrates that a nominal aggressor cell (target $\Delta = 60$) causes a victim cell with $\Delta = 40$ (weak cell) to write inadvertently within the write window for coupling ratios of 1 and above. The complete list of capacitive defects, faults (if any) and their activation patterns is shown in Table III.

D. Fault Activation

Traditionally March C- has been known to give good coverage for the SAF, TF and CFs. To be able to cover the data-dependent CFs which are two cell dynamic functional fault models, Word Oriented March (WOM) test patterns are needed. From the tables above, we are able to determine the worst case test patterns for each defect which need to be exercised to activate the faults. According to [12], constructing a 2-bit WOM is done by concatenating a 2-bit inter-word oriented March test with the necessary 2-bit intra-word patterns that sensitize the faults. Using the patterns discovered in the above sections and constructing a WOM March test from March C- in a manner discussed in [12], we get:

$\{\Uparrow\!\Downarrow(w00); \Uparrow(r00, w11, r11); \Downarrow(r11, w00, r00);$
$\Downarrow(r00, w11, r11); \Downarrow(r11, w00); \Uparrow\!\Downarrow(r00);$
$\Uparrow(r00, w01, r01); \Uparrow(r01, w10, r10); \Downarrow(r01, w10, r10);$
$\Downarrow(r10, w11, r11); \Uparrow\!\Downarrow(r11);\}$

Location	Agressor Cell	Write Fault model	Write Data Pattern Victim	Write Data Pattern Aggressor	Read Fault model	Read Data Pattern Victim
BL0-SL1	1	TF0	xw0	xw0	IRF0	r0
SL0-SL1	1	TF0	xw0	xw1	No effect	NA
BL1-SL0	1	TF1	xw1	xw0	IRF0	r0
T0-WL1	1,2	SA1F,CF	xw0	Idle	IRF0	r0
T0-SL1	1	SA1F	xw0	xw0	IRF1	r1
T0-BL1	1	SA1F	xw0	xw1	IRF1	r1
T0-T1	1	SA1F,CF	xw0	xw1	RDF	r0
T0-T2	2	SA1F,CF	xw0	Idle/xw1	RDF	r0
T0-T3	3	SA1F,CF	xw0	Idle/xw1	IRF1	r1
BL0-BL1	1	TF1	xw1	xw0	IRF1	r1
WL0-BL1	1	SA1F	xw0	xw1	RDF	r0
WL0-SL1	1	SA1F	xw0	xw1	IRF0	r0
WL0-WL1	1,3	SA0F,SAF1 CF	xwx	Idle	IRF0	r0

Table II: Resistive defects, faults and activation patterns

Location	Agressor Cell	Write Fault model	Write Data Pattern Aggressor	Write Data Pattern Victim
BL0-SL1	1	No effect	xw0	xwx
SL0-SL1	1	No effect	wx0	xwx
BL0-SL0	1	No effect	xwx	xwx
WL0-T1	1,2	CF	xwx	1w0
T0-SL1	1	No effect	xw0	xw0
T0-BL1	1	No effect	xw1	xw0
T0-T1	1	CF	xw1	1w0
T0-T2	2	CF	Idle/xw1	1w0
T0-T3	3	CF	Idle/xw1	1w0
BL0-BL1	1	No effect	xw0	xw1
WL0-BL1	1	No effect	xw1	xw0
WL0-SL1	1	No effect	xw1	xw0
WL0-WL1	1,3,4	CF	xwx	xwx

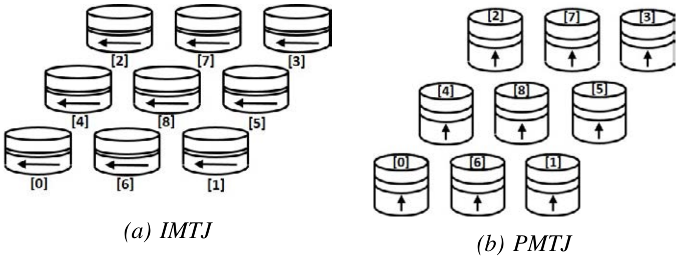
Table III: Capacitive defects, faults and activation patterns

$M0 = \Uparrow\!\Downarrow(w00); \quad M1 = \Uparrow(r00, w11, r11);$
$M2 = \Uparrow(r11, w00, r00); \quad M3 = \Downarrow(r00, w11, r11);$
$M4 = \Downarrow(r11, w00); \quad M5 = \Uparrow\!\Downarrow(r00);$
$M6 \Uparrow(r00, w01, r01); \quad M7 = \Uparrow(r01, w10, r10);$
$M8 = \Downarrow(r01, w10, r10); \quad M9 = \Downarrow(r10, w11, r11);$
$M10 = \Uparrow\!\Downarrow(r11);$

The SA1F and TF1 are sensitized and detected by M0 and M1 respectively. SA0F and TF0 are sensitized and detected by M1. RDF are detected if the last read operation in M1 and first read operation in M2 read differently. IRF are detected from M4 and M5. The various intra-word coupling faults are detected by M6, M7, M8 as depicted in the list above. M9 and M10 complete the test. The test patterns thus identified forms a part of the EMACS MBIST architecture that can identify read and write errors amidst electrical (resistive and capacitive) faults. In the next sections, we explore how EMACS can be extended for time-efficient retention testing.

IV. ROLE OF MAGNETIC COUPLING IN DENSE ARRAYS

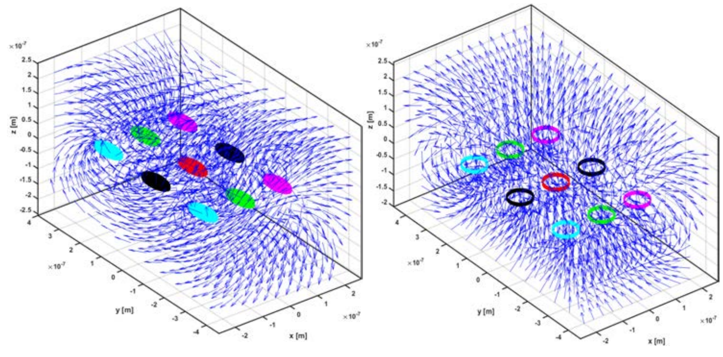
Since STT-MRAMs store information in nanomagnets, the energy stored in these nanomagnets can be affected by the data pattern in the neighboring cells. Just like electrical coupling, described above, we need to analyze the origin, magnitude and effects of magnetic coupling in STT-MRAM bitcells. In this



(a) IMTJ

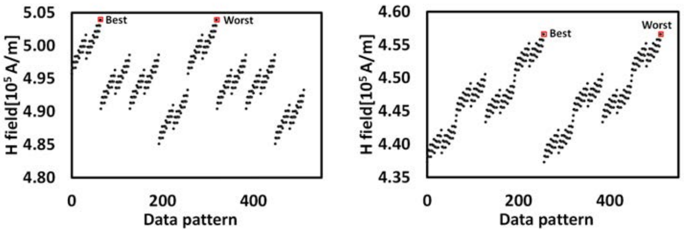
(b) PMTJ

Figure 4: Arrangement of MTJs in a 3X3 array



(a) IMTJ *(b) PMTJ*

Figure 5: Magnetic field visualization of IMTJ and PMTJ 3X3 arrays for the worst data pattern



(a) IMTJ *(b) PMTJ*

Figure 6: Residual H field vs. data pattern in IMTJ & PMTJ

section we provide a brief overview of our modeling infrastructure to determine the data pattern sensitivity on write times and stability. This falls under Neighborhood Pattern Sensitive Faults (NPSF) [13]. We model magnetic coupling effects on a 3x3 array and analyze the best case and worst case patterns with respect to write and retention and their overall effect on variability.

A. *Modeling Infrastructure for Magnetic coupling*

We model the MTJ as a solenoid which has imaginary current paths wrapped around itself to produce the same saturation magnetization(M_s). Since magnetic moment is derived from the volume and M_s of an MTJ ($M_s = \frac{Magnetic\ moment}{Volume\ of\ MTJ}$) and it is the product of current in the imaginary current path, cross section area of MTJ and number of coils, the amount of current needed to produce magnetic field can be readily calculated. From this relation, current is expressed as $\frac{M_s t}{\#\ coils}$, t is the thickness of a MTJ layer. The current loop around an MTJ is wrapped around in a direction to generate equal value and direction of M_s of a real MTJ and M_s direction is labeled in the figures. After current calculated from M_s to coils is applied, finite element method and Biot-Savart law are used to calculate magnetic field produced by MTJ at a specific position in space. Vector addition of the magnetic field produced by multiple aggressor magnets on the victim is used as a measure of the residual coupling field which is responsible for perturbing the magnetic dynamics of the victim cell. Fig. 4 shows the arrangement of the STT-MRAM bits in the 3X3 array (victim bit is cell-[8]). The saturation magnetization is set to 1.257e3A/m and physical dimensions of MTJs are set to $40nm \times 90nm$ and Δ = 60. The magnetic field pattern created by the IMTJ and PMTJ arrays on the victim cell (cell-[8]) is shown in Fig. 5.

B. *Impact of Magnetic Coupling on Write and Retention*

To visualize the best and worst case data patterns, we represent the information stored in the 3X3 array as a 9-bit number where each bit represents the data stored (0 for anti-parallel and 1 for parallel) in the i^{th} as shown in Fig. 4. The magnetic field in the figure is measured in the direction of the free layer magnetization. Therefore, the data pattern that yields the highest value of magnetic field when value is storing 0/1 is the best/worst case. Because of this encoding, data patterns 0 to 255 represent the victim storing a 0 and 256 to 511 represents the victim storing a 1. Fig. 6 show residual magnetic field strength from all the aggressors for all possible data arrangements. Residual field in the direction of the free layer's magnetization enhances stability and improves retention (thereby degrading writability) while residual fields in the opposite direction would tend to destabilize the magnet. We note that data pattern [000 111 111] and [100 111 111] are the best and worst case data patterns for thermal stability (or retention) for IMTJ. For PMTJ, best and worst data patterns are [011 111 111] and [111 111 111]. Due to the uni-axial anisotropy in two MTJ types, best and worst case data pattern are different for in-plane and perpendicular MTJs. The worst case patterns for the 3X3 block is shown in Fig. 7. We apply the residual magnetic field to solve LLG under an external field to determine the role of magnetic coupling on writability. The stored energy of the magnet is modified as $\Delta(H) = \Delta(H = 0)(1 \pm \frac{H}{H_k})^2$ under the presence of an external field (H) and shows significant impact on the average cell retention. Fig. 8 illustrates the role of magnetic coupling on average retention time and average write time on $15F^2$ cells at different P-MTJ technology nodes. We note (1) magnetic coupling has a weak effect on write, (2) due to the exponential dependence of retention time on Δ, average retention time has large variance between the best and worst case data patterns, and (3) as magnets are scaled while keeping Δ constant, the internal field H_k needs to be increased, which makes the cells less susceptible to external perturbations. Because of (2) we limit our discussion in this paper on magnetic coupling based data pattern dependence on retention testing only.

V. Test and Characterization of cell Retention

In STT-MRAM, retention time is defined as the time it takes for a cell to flip, a stochastic phenomenon, caused by thermal noise [14]. The average retention time is quantified as: $\tau = \tau_0 exp(\Delta)$ and $\Delta = \frac{K_u V}{k_B T} = \frac{H_k M_s V}{2k_B T}$ [14]. In order to ensure system reliability, each cell in an array must have enough thermal stability(Δ = 60 to guarantee 10 years of retention) against stochastic bit flip induced by thermal noise. With high Δ, a cell

1	0	1
1	1	1
1	0	1

(a) IMTJ block data pattern

1	1	1
1	1	1
1	1	1

(b) PMTJ block data pattern

Figure 7: Magnetic coupling induced worst-case data pattern for thermal stability

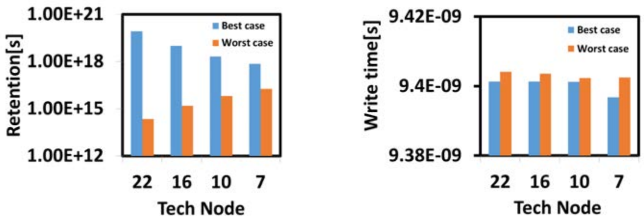


Figure 8: Magnetic coupling induced Data pattern dependence on (a) retention time (b) Write time in PMTJ cells

can have long retention but high Δ affects increase in write time and current. Due to this trade-off between power consumption and retention, [15][16][17] propose the use of quasi-stable cells with lower Δ to be used in caches. Whatever the design target may be, determining Δ in post-Silicon characterization and manufacturing tests is of utmost importance. However (1) the statistical nature of thermally activated bit-flips, (2) low failure probabilities, (3) large dependence on temperature and process parameters (M_S, H_K, t) and (4) exponential dependence of retention times and retention failure probability on Δ make it a challenging test problem, as has been noted in the Intel publication [6].

A. Challenges in Retention and Thermal Stability Tests

Very little work exists in published literature on test schemes and challenges from testing retention and thermal stability. While discussing the challenges in [6], Intel proposes a possible test methodology based on the thermal activation model.

$$P_{\text{sw}} = 1 - exp\left(-t/exp\left(\Delta\left(1 - \frac{I_{\text{WWR}}}{I_{\text{c0}}}\right)\right)\right) \quad (3)$$

P_{sw} is a switching probability of a cell and I_{WWR} is a Weak Write (WWR) current. The model is is used to obtain the values of I_{c0} and Δ by fitting bit-level experimental/test data [6][7][18][19]. From the thermal activation model for the case:

$$\frac{t_{\text{p}}}{\tau_0 exp(\Delta(1 - \frac{I_{\text{WWR}}}{I_{\text{c0}}}))} << 1 \quad (4)$$

using Taylor expansion and ignoring higher order terms [7][6]:

$$ln(P_{\text{sw}}) = ln(\frac{t_{\text{p}}}{t_0}) - \Delta(1 - \frac{I_{\text{WWR}}}{I_{\text{c0}}}) \quad (5)$$

where t_p is the pulse width for switching current. This model links P_{SW} and Δ under application of I_{WWR}. Since the thermal activation model is a stochastic model, a large number of successive tests is required to obtain statistically significant results. Also, the model is accurate when low switching current is applied during the long pulse width [6]. Experimental data from [7][18] suggests that switching current ratio of $\frac{I_{\text{WWR}}}{I_{\text{c0}}} \leq 0.8$ and switching pulse width of t_{p} = 100ns are the upper bounds of the thermal activation model for $P_{\text{sw}} \leq 1e-3$ [6] (Fig. 9). P_{sw} of $1e-1$ with ± 1 percent error margin and 99 percent confidence requires 5e+5 number of tests [20][6]. Based on this model, [6] proposes a test scheme where 100ns I_{WWR} pulses are applied and each bit read to determine a possible bit flip. After 5e+5 such tests with 10 different values of $\frac{I_{\text{WWR}}}{I_{\text{c0}}}$, generated by an embedded MBIST, we can obtain statistically significant test data to determine Δ through post-processing. Based on [6] the test algorithm is shown below:

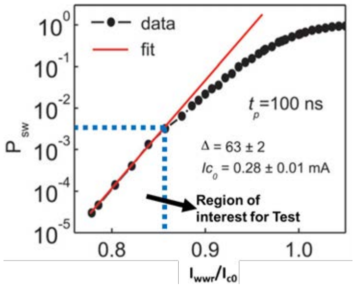


Figure 9: Experimental data of P_{SW} vs. Iwwr[7] showing the region of operation for test where the exponential thermal model is valid.

```
Result: Obtain $P_{sw}$ for every cells in an array
initialization;
$N_{row}$ = total number of rows;
Iwwr[N] = array that contains N number of $I_{WWR}$ value;
M = total number of experiments per each $I_{WWR}$;
for $i = 0; i < N_{row}; i{+}{+}$ do
    for $j = 0; j < N; j{+}{+}$ do
        Write test patterns into the line;
        for $k = 0; k < M; k{+}{+}$ do
            Apply current $I_{WWR}$[j] for $t_p$;
            Read the line value;
            if $value \neq test\ pattern$ then
                error counter of cells with error++;
                rewrite correct value to the row;
            end
        end
    end
end
```

Algorithm 1: Retention test algorithm with weak WR current

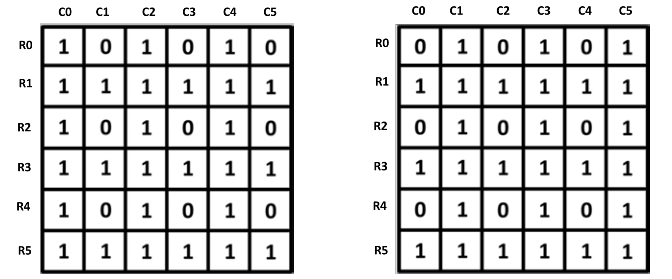
Using an MBIST the total test time is approximately 16mins to test two thousand lines of array when N is 5e5 with 10 I_{WWR}, t_{p} =100ns. Even though parallelism at a sub-array level can help to reduce retention test time, there is a clear limit in reducing the total retention test time. With increasing size of cell array, the retention test time with this MBIST is not feasible. Therefore, there is a strong need for efficient retention test algorithm which can reduce test time significantly. We address this issue in the next section.

B. Test patterns for retention test: Role of Magnetic Coupling

From the analysis of magnetic coupling, we identified the worst case data patterns for retention testing under magnetic coupling. Fig. 7 indicates the worst data patterns of IMTJ and PMTJ cells under which magnetic field coupling degrades the thermal stability the most. In order to consider magnetic coupling effect in retention test, we need to set the test pattern which has most impact on thermal stability. We first write the data pattern based on Fig. 7; and then perform retention test for cells under magnetic coupling. Fig. 10 indicates the two block data patterns for testing worst case stability in I-MTJ arrays. For P-MTJ the worst case pattern is all-ones.

VI. Proposed MBIST for Retention Testing

We extend EMACS to perform in-situ, statistical, retention testing of large STT-MRAM arrays. From the retention BIST algorithm [6], we apply a weak current and read the value of cells row by row to obtain P_{sw}. The principle drawbacks of the above scheme that we identify are:



	C0	C1	C2	C3	C4	C5
R0	1	0	1	0	1	0
R1	1	1	1	1	1	1
R2	1	0	1	0	1	0
R3	1	1	1	1	1	1
R4	1	0	1	0	1	0
R5	1	1	1	1	1	1

(a) IMTJ block data pattern A

	C0	C1	C2	C3	C4	C5
R0	0	1	0	1	0	1
R1	1	1	1	1	1	1
R2	0	1	0	1	0	1
R3	1	1	1	1	1	1
R4	0	1	0	1	0	1
R5	1	1	1	1	1	1

(b) IMTJ array data pattern B

Figure 10: IMTJ worst case data patterns for retention shown in a 5×5 *grid. For PMTJ the worst case pattern is all-ones.*

(1) The retention test time increases linearly when the row size of an array increases.

(2) The retention tests have to be carried out in an operating region where P_{SW} is very low. For example, for a cell with Δ = 60, applying $\frac{I_{\text{WWR}}}{I_{c0}}$ from 0.76 to 0.82 for t_p = 100ns sets P_{sw} to be 5.573e-5 to 0.002 based on Eqn 5. It indicates that for most of the iterations, a bit flip will not happen; which means most of the read operations after applying current are not necessary.

These two problems are main bottlenecks for improving speed of retention test. The retention test methodology and MBIST architecture that we propose focuses on how to overcome these two bottlenecks. If an error can be detected in an entire cell array with a fixed number of memory operations, we can decouple array size from the factors affecting retention time. Also, since the probability of occurrence of bit flip is low, rather than reading rows each time after applying current in search of a bit-flip, reading rows only after an error is detected will reduce retention test time. The proposed architecture reduces retention time significantly by:

1. Detecting errors column-wise
2. Avoiding search (reading rows) when error is not detected

By testing multiple rows in a column at the same time and searching for errors after error detection, retention time testing reduces significantly. The retention test is divided into two processes, (1) Error Detection (ED) and (2) Error Search (ES).

A. EMACS System Architecture for Statistical Retention Tests

Fig. 11 presents the top level system diagram of the proposed MBIST circuitry. Normal memory operation and test operation are distinguished by the test_en signal. For retention test, Error Detection (ED) and Error Search (ES) logic are parts of the control logic. Based on the outputs of the MBIST circuit, Error Detection logic asserts err_det signal and while err_det is asserted, Error Search is conducted. Error Search controls which rows to assert from a localization factor (to be described in the next subsection) and it outputs error location to the output of control logic as soon as it identifies error locations. Search_done signal is asserted if Error Search is over and it resets err_det signal. I_{WWR} bus is used to control voltage of bitline and word-line, which leads to different magnitudes of I_{WWR} current. Column of different resistors are used as a references for finding errors in blocks of rows and temperature sensor are located inside a sub array to monitor temperature inside a sub array. Each characterization test, which produces an experimental determination of Δ, is qualified by a temperature data. The proposed scheme allows massive parallelism in test and enables a fine trade-off between localization of weak cells and test time.

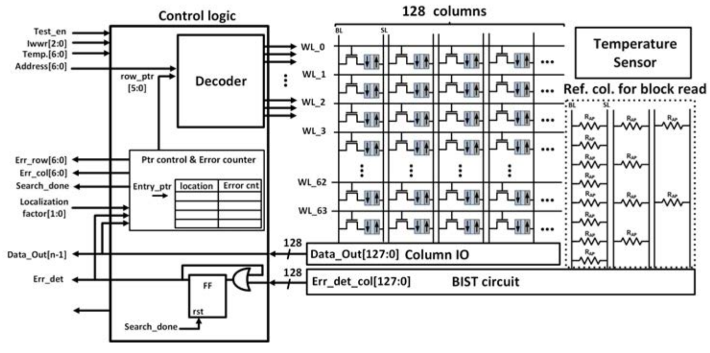


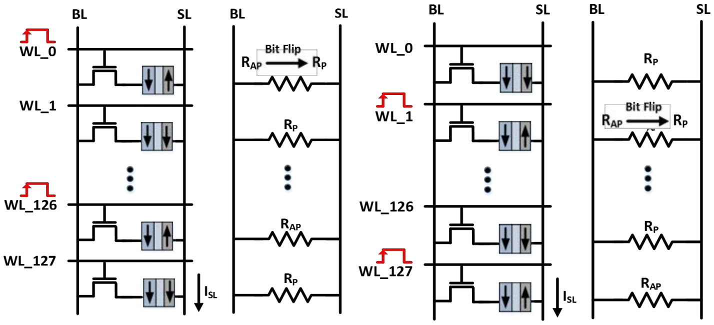
Figure 11: System architecture of EMACS MBIST applied to a 64×128 *array. EMACS is capable of read, write and statistical retention tests.*

B. Error Detection (ED)

The ED architecture is based on the MTJ property that any change in data (bit-flip) results in a change in resistance of the cell, which in-turn changes the current flowing through the cell. [21] uses this property to detect read disturb errors, by monitoring current difference (before and after the bit-flip) due to change in resistance.

In the proposed scheme (Fig. 12): (1) data patterns based on Fig. 10 are first written into the array, (2) retention test started by turning on multiple word-lines simultaneously, (3) I_{WWR} current injected through each cell which is storing a 1, (4) multiple read operations are conducted while passing I_{WWR} to check for a possible bit-flip, (5) next data pattern applied for full-coverage. For IMTJ, two block data patterns are identified in Fig. 10. To enable multiple simultaneous tests, odd numbered columns (C1,C3,..) are tested first Fig. 12a with block data pattern A, followed by testing of even numbered columns (C0, C2,...) using pattern B. Then the pattern is shifted vertically by one row and the process repeats. For PMTJ, the worst case pattern under magnetic coupling is all-ones, and hence all the columns can be tested simultaneously. Turning multiple word-lines in a column connects the MTJ resistance in parallel as shown in Fig. 12. The resistance of a MTJ is set to R_{ap} since cells store bit 1 in the figure. When I_{WWR} causes a bit flip in a cell, the resistance of a MTJ will change from R_{ap} to R_{p} as shown. The current flowing through source line of a column (I_{SL}) changes due to the resistance change. By detecting difference in I_{SL}, we can detect the existence of errors in a column.

However, due to low (150%[22]) TMR($= \frac{R_{\text{ap}} - R_{\text{p}}}{R_{\text{p}}}$) of the MTJ, the number of rows that can be simultaneously turned on and a bit-flip detected, is limited. With low TMR, the difference of total resistance of a column between a case with no errors and a case with a single error decreases and it affects difference in I_{SL}. Fig. 13 presents a trade-off between number of activated rows and the current difference of no error case and one error case with respect to different TMR values. It exhibits decreasing I_{SL} difference in percent as number of activated rows increases with different TMR. Due to process variation and temperature fluctuation, appropriate number of activated row must be set to gain enough margin in current difference. In this work, we limit



(a)

(b)

Figure 12: Multiple word-lines simultaneously turned ON to detect bit-flips according to worst case patterns identified in Fig.10 for IMTJ. (a) Simultaneous testing with block data pattern A applied to C1,C3,.., (b) Simultaneous testing with block data pattern B applied to C0,C2,... For PMTJ the worst case pattern is all-ones so all word-lines are turned on simultaneously.

the number of activated rows to 16, to distinguish between the "no error" and a "single error" in a column.

In the proposed test scheme, unlike the testing scheme from [6], we check errors while supplying I_{WWR} through 16 rows of cells. Since test scheme must apply $I_{\mathrm{WWR}} \leq 0.8 I_{\mathrm{c0}}$ for t_p and it is same as a strong read/weak write operation, we can detect errors within 16 rows in a column by monitoring I_{SL}. The read operation overhead after weak write is removed for the case when no error is detected during t_p. From this scheme, we can reduce t_{read} by $(1 - P_{\mathrm{sw}})N$ per I_{WWR} iteration, N is the number of test per I_{WWR} Fig. 14 shows the scheme for detecting a change in I_{SL} caused by a bit flip of a cell. The change in I_{SL} is amplified by current mirror and it is transferred to voltage difference and further amplified by multi stage common drain amplifier. Switched capacitors C1 and C2 sample the voltage at the common drain amplifier alternatively based on CLK and CLK_B signals. When bit-flip happens, the voltage difference between C1 and C2 is developed and maintained for a half clock cycle. Since the node voltages at C1 and C2 fluctuate when they are directly connected to the inputs of sense amplifier due to their small size, we implemented voltage keepers in between to avoid voltage fluctuation. By calibrating value of R1 and R2, in+ port is set to be always 10mV higher than in- to prevent metastability issue in sense amplifier. When sense amplifier enable is on, the sense amplifier fully differentiates the in+ and in- to VDD and GND. Fig. 15 presents waveform of switched capacitor control signals(CLK, CLK_B) and sense amplifier enable. Once WLs are asserted to supply I_{WWR} for t_p, CLK and CLK_B toggle to sample the voltage to C1 and C2. After capacitor C1 and C2 develop common mode voltage within t_{dev}, sense amplifier enable signal is asserted in the middle of every half clock cycle. Discharging of C1 or C2 must be finished before sense amplifier enable is asserted to apply maximum voltage difference in port in+ and in- of sense amplifier.

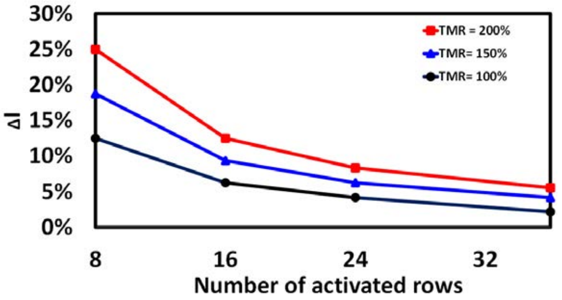


Figure 13: ΔI_{SL} vs. number of rows activated as a function of TMR

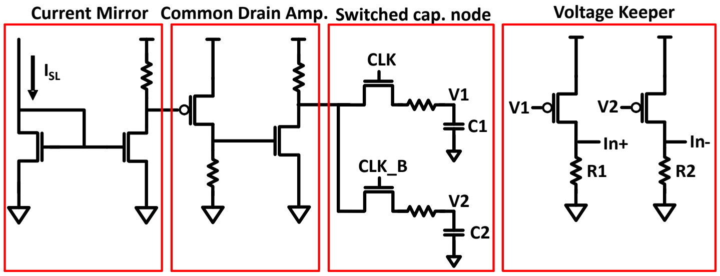


Figure 14: Error Detection circuit for a column with 16 rows

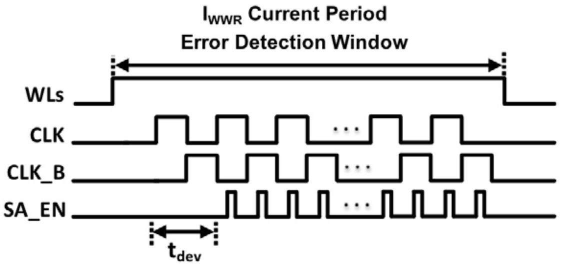
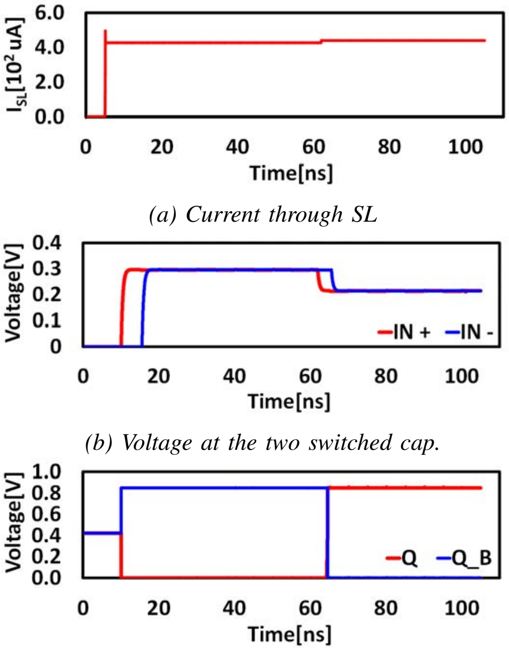


Figure 15: Timing Diagram illustrating the operation of the MBIST retention test

Fig. 16 shows the voltage across switched capacitors(C1, C2) and sense amplifier output when bit-flip happens. Around 60ns in Fig. 16a, current through SL is seen to increase due to the change in resistance ($R_{\mathrm{AP}} \rightarrow R_{\mathrm{P}}$) from a bit flip. Voltage difference across switched capacitor is maintained for half clock cycle in Fig. 16b and the sense amplifier resolves the voltage difference to VDD and GND when sense amp. enable is on. Fig. 17a summarizes the test-procedure and Fig. 17b



(a) Current through SL

(b) Voltage at the two switched cap.

(c) Sense Amp. Output

Figure 16: Transient analysis for error detection

shows the corresponding algorithm. Test flows A, B, C, D in yellow bounding boxes are presented in individual diagrams in Fig. 20,21. The main differences between proposed test scheme and [6] are error detection and search algorithms. In [6], the authors propose to apply I_{WWR} for t_p and read a row for every rows in an array. Instead, the proposed test scheme applies I_{WWR} to a block of rows and search for errors only when existence of error is identified by detecting a change in current through source line of a column. Fig. 18 illustrates a particular simulation run showing infrequent bit flips happening over time which are recorded in the current scheme. This allows estimation of P_{sw} and finally extrapolated to obtain Δ via Eqn. 3. The P_{sw} for a cluster of cells within an 8KB subarray is shown in Fig. 19. After ED, a search algorithm to localize the bit-flip is used and is discussed next.

C. Error Search and Localization

After detecting the existence of errors using the scheme above, searching the location of errors within the activated rows is necessary in order to obtain P_{sw} and thermal stability of cells. In this paper, we present three different error search schemes (exhaustive search, temporal locality search and search localization).

1) Exhaustive Search: The algorithm used after detecting the first error is exhaustive search. In exhaustive search, every row in a block of activated rows are read to locate errors. Once it obtains location of an error, the test scheme stores the location in a error table and re-writes original test pattern to a row with an error. When the last row in a block is read, it goes back to error detection flow algorithm. Error location stored in a table is used in a search which exploits temporal locality. Fig. 20 demonstrates each steps in exhaustive search.

2) Temporal locality search: Temporal locality search can reduce error search time when process variation on thermal stability of cells is large. The efficiency of search improves when performing manufacturing test, the test that identifies cells which do not meet target retention. Fig. 21 presents each steps in temporal locality search. Once error table is filled from exhaustive search, temporal locality search first reads rows in the table to locate errors. If the row specified in the table contains an error, it updates number of errors associated with the row in a table. When no error is found in the rows from the table, it switches to exhaustive search to find errors in other rows and add a row to a table when error is found in the row. After it finds an error, it reads the block of rows to ensure it corrected all errors.

3) Error Localization Search: Both exhaustive search and temporal locality search identify all locations of errors in the array. In terms of search time, however, both search scheme can be time consuming if the block size of activated rows for error detection is large. Instead of identifying which row contains errors for each column, we can set a block size in terms of row(N_{loc}) and search whether the block contains errors. For example, searching errors within 4 rows each time is 4 times faster than exhaustive row search. By reducing accuracy of error position, we reduces search time linearly as N_{loc} increases. Fig. 22 presents how search time varies based on the size of N_{loc}. The search time is compared with 5 different levels of localization. Table IV indicates how localization level maps to N_{loc}. Since search is conducted when error is detected, search time is a multiplication of error probability(P_{sw}), read time and $\frac{N_{act}}{N_{loc}}$. N_{act} is the number of rows activated for error detection. Search time in the Fig. 22 is calculated with the assumption that P_{sw} = 3e-3, number of I_{wwr} = 10 and number of experiments per I_{wwr} = 5e5. As we mentioned earlier, the search time decreases linearly when N_{loc} increases in the figure.

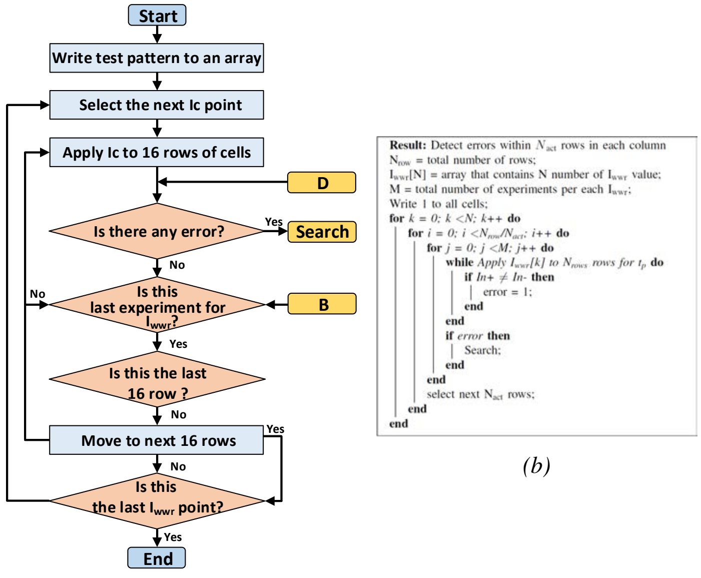
$$t_{\text{search}} = P_{\text{sw}} \times t_{\text{read}} \times \frac{N_{\text{act}}}{N_{\text{loc}}} \tag{6}$$

Localization level	Block size(row)
5	N_{act}
4	0.5 N_{act}
3	0.25 N_{act}
2	0.125 N_{act}
1	1

Table IV: Localization level in terms of no. of rows

D. Overhead of internally storing data

Retention test requires at least $8N_{cell}$ (total number of cells in an array) bits of memory to store number of bit flips per cell under assumptions that the maximum P_{sw} = 3e-3 and number of experiments per I_{wwr} is 5e5 to calculate thermal stability of each cells. Instead of storing error counts in the memory, test scheme can output row & column information when error is detected to calculate thermal stability outside the chip. However,

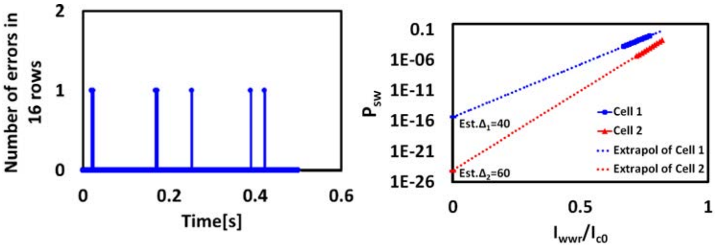

Start
Write test pattern to an array
Select the next Ic point
Apply Ic to 16 rows of cells
D
Is there any error?
Yes
Search
No
No
Is this last experiment for I_{wwr}?
B
Yes
Is this the last 16 row ?
No
Move to next 16 rows
Yes
No
Is this the last I_{wwr} point?
Yes
End

```
Result: Detect errors within $N_{act}$ rows in each column
$N_{row}$ = total number of rows;
$I_{wwr}[N]$ = array that contains N number of $I_{wwr}$ value;
M = total number of experiments per each $I_{wwr}$;
Write 1 to all cells;
for $k = 0; k < N; k{+}{+}$ do
    for $i = 0; i < N_{row}/N_{act}; i{+}{+}$ do
        for $j = 0; j < M; j{+}{+}$ do
            while Apply $I_{wwr}[k]$ to $N_{rows}$ rows for $t_p$ do
                if $In+ \neq In-$ then
                    error = 1;
                end
            end
            if error then
                Search;
            end
        end
        select next $N_{act}$ rows;
    end
end
```

(a)

(b)

Figure 17: (a) Flow chart (b) algorithm for bit-flip detection in a column



(a) Bit Errors as a measure of bit-flips over time

(b) P_{sw} vs.. I_{WWR}/I_{c0}

Figure 18: Estimating P_{sw} and Δ through EMACS

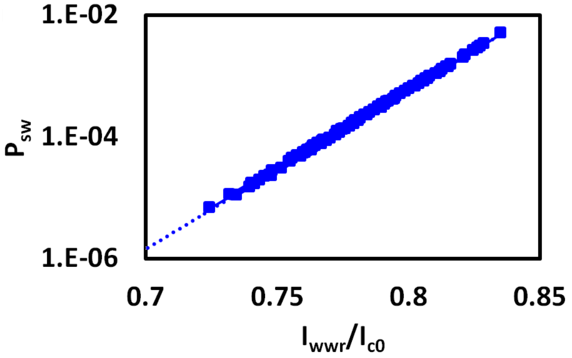


Figure 19: P_{sw} on a cluster of cells in an 8KB array showing a scatter which can be extrapolated to obtain Δ

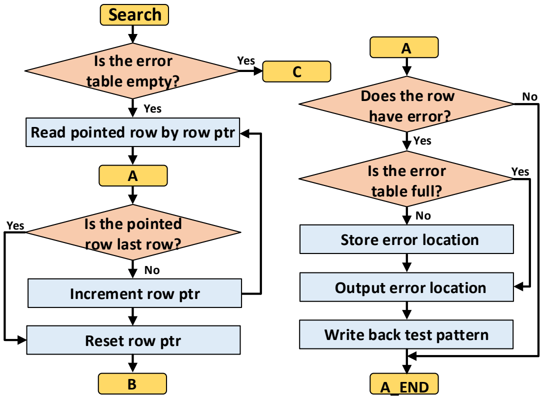


Figure 20: Flow chart for exhaustive error search

it adds complexity to test mode control logic and outputting error location is also time consuming. It should also be noted that block level identification of cell stability allows us to apply redundancy easily. Once a particular column is identified as having weak (low Δ) cells, we can swap it with a redundant column. So in manufacturing tests, localization at the granularity of a column is sufficient.

VII. Performance analysis

A. Test time Comparison

The retention test time of proposed test scheme can be calculated using the equation;

$$t_{\text{ret}} = [(t_{\text{p}} + t_{\text{search}}) \times \frac{N_{\text{row}}}{N_{\text{act}}}] \times M \times N_{\text{I}_{\text{wwr}}} \tag{7}$$

N_{row} is the total number of rows in an array, M is the number of experiments required for each I_{wwr} and $N_{I_{\text{wwr}}}$ is the total number of I_{wwr} needed to extrapolate P_{sw} vs. I_{wwr} to obtain cell retention. t_{search} is defined in equation 6. Fig. 23 presents the performance analysis in terms of time between [6] and EMACS. For testing retention for an array with 2000 rows, test scheme from [6] takes 16 mins to complete and proposed test scheme takes 1 min with N_{act}. If we increase N_{act} to be 32 and 64, the test time reduces to $\frac{1}{32}$, $\frac{1}{64}$ of the test time from [6].

B. Area overhead

Fig. 23b presents the area overhead of the proposed retention MBIST. For this analysis, we did not use any column mux techniques to reduce number of test circuit by half. Each column contains one set of test circuit including sense amp. described in Fig. 14 in the analysis. We assumed that each cell size in the array is $30F^2$ to calculate the area overhead of test circuit with respect to total array area. From Fig. 23b, we deduce that area overhead of test scheme decreases linearly with respect to number of rows in the array. With 512 rows, area overhead is 3.44% of the total cell array size and it reduces by half when number of rows doubles.

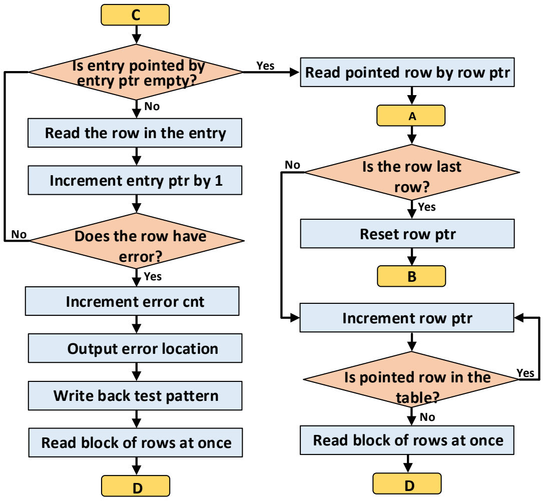


Figure 21: Flow chart for temporal locality search

VIII. Array Level Testing and Challenges

The proposed test-scheme, albeit a practical and faster test methodology, is still a statistical test enabled by an MBIST and suffers from measurement errors arising due to temperature changes and process variations. Since retention times are heavily dependent on temperature, we propose (1) to use embedded thermal sensors within the subarray to qualify each sub-array measurement with the corresponding temperature, or (2) insert idle states in between applying I_{wwr} and error detection process to maintain stable temperature. Another potential problem in the test-scheme is the process induced mismatches between cells. When a block of cells are written and read simultaneously, the I_{WWR} is not equally divided between the cells. This creates loss of accurate measurement of Δ and needs to be accounted for as a design guard-band. We carried out simulations of the EMACS test scheme by running tests under temperature and

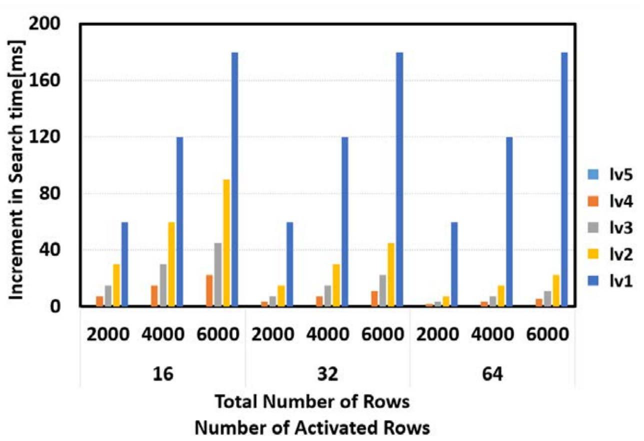
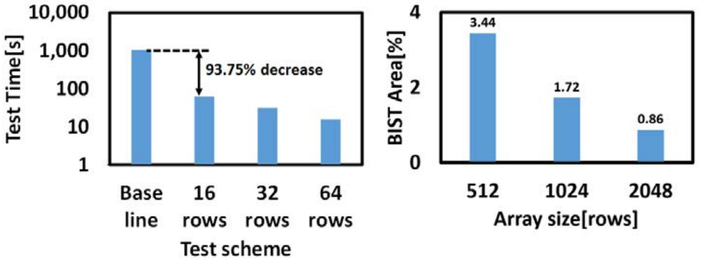
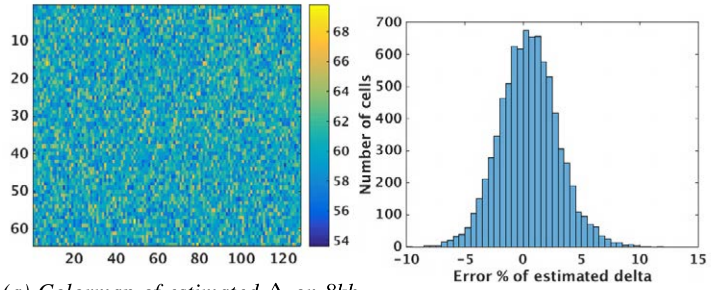


Figure 22: Search time increment w.r.t. localization level



(a) *(b)*

Figure 23: (a) Retention time vs. localization level, (b) Area overhead w.r.t. array size



(a) Colormap of estimated Δ on 8kb array from EMACS

(b) Error in estimation of Δ

Figure 24: Estimated Δ and Error of estimation from 8kb array using EMACS. The colormap represents cells in a 64×128 array.

process variations and trying to estimate Δ on an 8KB subarray amidst all the non-idealities. Fig. 24(a) presents the estimated thermal stability of 8kb cell subarray and Fig. 24b shows the accuracy of the test methodology for the collection of 8KB cells. It can be seen that the proposed scheme has bounded error of $< \pm 5\%$ and 93.75% decrease in test-time with respect to [6] and demonstrates the effectiveness of the proposed test methodology.

IX. Conclusion

This paper presents a comprehensive test methodology for STT-MRAM arrays. We identify electrical defects and magnetic coupling induced data pattern dependence on tests for read, write and retention. Finally, an MBIST architecture (EMACS) capable of collecting statistical data in an STT-MRAM subarray to estimate the thermal stability and retention is proposed. The proposed MBIST shows 93.75% improvement in test-time compared to a brute-force approach [6] with less that 5% estimation error.

Acknowledgment

The authors would like to thank Suriyaprakash Natarajan, Helia Naeimi and Abhinav Parihar for fruitful discussions. Insik Yoon, and Ashwin Chintaluri were supported by the Semiconductor Research Corporation (Task no. 2493.001).

References

[1] A. Chintaluri, H. Naeimi, S. Natarajan, and A. Raychowdhury, “Analysis of defects and variations in embedded spin transfer torque (stt) mram arrays,” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. PP, no. 99, pp. 1–11, 2016.

[2] S. A. Wolf, J. Lu, M. R. Stan, E. Chen, and D. M. Treger, “The promise of nanomagnetics and spintronics for future logic and universal memory,” *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2155–2168, 2010.

[3] G. Jan, L. Thomas, S. Le, Y. J. Lee, H. Liu, J. Zhu, R. Y. Tong, K. Pi, Y. J. Wang, D. Shen, R. He, J. Haq, J. Teng, V. Lam, R. Annapragada, T. Zhong, T. Torng, and P. K. Wang, “Demonstration of an MgO based anti-fuse OTP design integrated with a fully functional STT-MRAM at the Mbit level,” *Digest of Technical Papers - Symposium on VLSI Technology*, vol. 2015-Augus, pp. T164–T165, 2015.

[4] H. Yoda, S. Fujita, N. Shimomura, E. Kitagawa, K. Abe, K. Nomura, H. Noguchi, and J. Ito, “Progress of STT-MRAM technology and the effect on normally-off computing systems,” *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 259–262, 2012.

[5] A. Chintaluri, A. Parihar, S. Natarajan, H. Naeimi, and A. Raychowdhury, “A Model Study of Defects and Faults in Embedded Spin Transfer Torque (STT) MRAM Arrays,” *2015 IEEE 24th Asian Test Symposium (ATS)*, vol. 1, no. c, pp. 187–192, 2015.

[6] H. Naeimi, C. Augustine, A. Raychowdhury, S.-l. Lu, and J. Tschanz, *Intel Technology Journal, STTRAM Scaling and Retention Failure*, vol. 17. 2013.

[7] R. Heindl, W. H. Rippard, S. E. Russek, M. R. Pufall, and A. B. Kos, “Validity of the thermal activation model for spin-transfer torque switching in magnetic tunnel junctions,” *Journal of Applied Physics*, vol. 109, no. 7, 2011.

[8] J. Sun, “Spin-current interaction with a monodomain magnetic body: A model study,” *Physical Review B*, vol. 62, no. 1, pp. 570–578, 2000.

[9] R.-f. Huang, Y.-f. Chou, and C.-w. Wu, “Defect oriented fault analysis for SRAM,” *Proceedings of the 7th International Conference on Properties and Applications of Dielectric Materials ATS-03*, pp. 256–261, 2003.

[10] N. Z. Haron and S. Hamdioui, “On defect oriented testing for hybrid CMOS/memristor memory,” *Proceedings of the Asian Test Symposium*, pp. 353–358, 2011.

[11] Y. X. Chen and J. F. Li, “Fault modeling and testing of 1t1r memristor memories,” in *2015 IEEE 33rd VLSI Test Symposium (VTS)*, pp. 1–6, April 2015.

[12] A. J. V. D. Goor, I. B. S. Tlili, and S. Hamdioui, “Converting march tests for bit-oriented memories into tests for word-oriented memories,” pp. 46–52, Aug 1998.

[13] A. J. V. D. Goor, *Testing Semiconductor Memories: Theory and Practice*. John Wiley and Sons, 1998.

[14] A. V. Khvalkovskiy, D. Apalkov, S. Watts, R. Chepulskii, R. S. Beach, A. Ong, X. Tang, A. Driskill-Smith, W. H. Butler, P. B. Visscher, D. Lottis, E. Chen, V. Nikitin, and M. Krounbi, “Basic principles of STT-MRAM cell operation in memory arrays,” *Journal of Physics D: Applied Physics*, vol. 46, no. 13, p. 139601, 2013.

[15] A. Nigam, C. W. Smullen, V. Mohan, E. Chen, S. Gurumurthi, and M. R. Stan, “Delivering on the promise of universal memory for spin-transfer torque RAM (STT-RAM),” *Proceedings of the International Symposium on Low Power Electronics and Design*, vol. 1, pp. 121–126, 2011.

[16] A. Jog, A. K. Mishra, C. Xu, Y. Xie, V. Narayanan, R. Iyer, and C. R. Das, “Cache revive: Architecting volatile STT-RAM caches for enhanced performance in CMPs,” *Proceedings of the 49th Annual Design Automation Conference (DAC)*, pp. 243–252, 2012.

[17] Z. Sun, X. Bi, H. H. Li, W.-F. Wong, Z.-L. Ong, X. Zhu, and W. Wu, “Multi retention level STT-RAM cache designs with a dynamic refresh scheme,” *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 329–338, 2011.

[18] A. Driskill-Smith, S. Watts, D. Apalkov, D. Druist, X. Tang, Z. Diao, X. Luo, A. Ong, V. Nikitin, and E. Chen, “Non-volatile spin-transfer torque RAM (STT-RAM): An analysis of chip data, thermal stability and scalability,” *2010 IEEE International Memory Workshop, IMW 2010*, vol. 1, no. 408, pp. 5–7, 2010.

[19] M. Pakala, Y. Huai, T. Valet, Y. Ding, and Z. Diao, “Ciritical Current distribution in spin-transfer-switched magnetic tunnel junctions,” *Journal of Applied Physics 2005*, vol. 98, no. 5, 2005.

[20] D. Montgomery and G. Runger, *Applied Statistics and Probability for Engineers*. John Wiley and Sons, 2010.

[21] R. Bishnoi, F. Oboril, and M. B. Tahoori, “Read Disturb Fault Detection in STT-MRAM,” *International Test Conference*, pp. 1–7, 2014.

[22] G. Jan, L. Thomas, S. Le, Y.-j. Lee, H. Liu, J. Zhu, R.-y. Tong, K. Pi, Y.-J. Wang, D. Shen, R. He, J. Haq, J. Teng, V. Lam, K. Huang, T. Zhong, T. Torng, and P.-k. Wang, “Demonstration of fully functional 8Mb perpendicular STT-MRAM chips with sub-5ns writing for non-volatile embedded memories,” *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers*, vol. 093008, no. 2012, pp. 1–2, 2014.