Parametric Failure Modeling and Yield Analysis for STT-MRAM

Sarath Mohanachandran Nair, Rajendra Bishnoi and Mehdi B. Tahoori
Chair of Dependable Nano Computing (CDNC), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
Email: {sarath.nair, rajendra.bishnoi, mehdi.tahoori}@kit.edu

***Abstract*—The emerging *Spin Transfer Torque Magnetic Random Access Memory* (STT-MRAM) is a promising candidate to replace conventional on-chip memory technologies due to its advantages such as non-volatility, high density, scalability and unlimited endurance. However, as the technology scales, yield loss due to extreme parametric variations is becoming a major challenge for STT-MRAM because of its higher sensitivity to process variations as compared to CMOS memories. In addition, the parametric variations in STT-MRAM exacerbates its stochastic switching behavior, leading to both test time fails and reliability failures in the field. Since an STT-MRAM memory array consists of both CMOS and magnetic components, it is important to consider variations in both these components to obtain the failures at the system level. In this work, we model the parametric failures of STT-MRAM at the system level considering the correlation among bit-cells as well as the impact of peripheral components. The proposed approach provides realistic fault distribution maps and equips the designer to investigate the efficacy of different combinations of defect tolerance techniques for an effective design-for-yield exploration.**

I. Introduction

Increased leakage power has become a major factor affecting the scalability of conventional CMOS memories such as SRAM and DRAM in advanced technology nodes [1]. As a solution, various emerging non-volatile memories are in consideration to replace CMOS memories, at least for a subset of the on-chip memory hierarchy. Among these, *Spin Transfer Torque Magnetic Random Access memory* (STT-MRAM) is the most promising candidate due to its advantages such as scalability, high endurance, long retention and fast accesses [2].

A typical STT-MRAM bit-cell consists of a Magnetic Tunnel Junction (MTJ), which is the storage element, and an NMOS access transistor. However, as technology scales down, STT-MRAM is affected by manufacturing variations in the magnetic fabrication process as well as the CMOS process. In addition, the impact of process variation on the magnetic devices exacerbates the stochastic switching of the MTJ. The CMOS device variations are primarily due to Random Dopant Fluctuation (RDF), Line-Edge Roughness (LER) and Shallow-Trench Isolation (STI) stress [3]. The combined effect of magnetic and CMOS variations on the bit-cell and peripheral circuitry result in both reliability failures in the field and permanent faults at the tester in STT-MRAM based memories.

A good understanding of the failure behavior and failure map can help the designers to incorporate the right combination of defect tolerance techniques to overcome the yield loss. The existing fault models for conventional CMOS memory technologies cannot be directly applied to STT-MRAM because of the the fundamental difference in the operation [4]. In addition, due to non-volatility and stochasticity, some of the failure mechanisms (such as read disturb and retention failures) are unique to STT-MRAM. The yield analysis framework should also consider the entire memory system including the bit-cell array and peripherals which can guide the designer to employ appropriate design-for-yield schemes.

There are a few works which analyze the transient (reliability) and permanent faults in STT-MRAM and propose solutions to mitigate these faults [4–6]. However, most of these works primarily focus on the bit-cell level modeling and also do not consider both reliability failures and permanent faults. The existing works do not perform a system-level fault modeling considering the bit-cell and the periphery. Furthermore, the previous works do not consider the correlation among the parameters of neighboring bit-cells.

In this work, we consider parametric variations in the bit-cells and peripherals as well as the correlation among neighboring cells to get the fault distribution map of the memory array, due to both permanent faults and reliability failures. The framework can be used for a design-for-yield exploration combining various defect tolerance techniques (like ECC and redundancy) to mitigate permanent and reliability failures. We observe that unique yield improvement techniques specific to STT-MRAM are far more effective than conventional techniques (such as redundant rows/columns and ECC).

The rest of this paper is organized as follows. In Section II, the basics of STT-MRAM is introduced. Next, in Section III, we discuss the parametric failures affecting STT-MRAM. Section IV explains the methodology employed for developing the fault modeling framework, followed by the results which are demonstrated in Section V. Finally, Section VI concludes the paper.

II. Background

The storage element in STT-MRAM is the *Magnetic Tunnel Junction* (MTJ) (see Fig. 1(a)), comprising of two ferromagnetic layers (e.g., CoFeB) separated by a thin oxide layer (e.g., MgO). The magnetic orientation of one of the layers is fixed, known as the *Reference Layer* (RL), whereas that of the other layer, known as the *Free Layer* (FL), can be rotated freely by passing a spin polarized current. When the magnetic orientation of the FL is parallel (anti-parallel) to that of the RL, the MTJ cell is in low (high) resistance state. To switch the MTJ from the anti-parallel (parallel) to the parallel (anti-parallel) state, the current has to flow from the FL (RL) to the RL (FL). On the other hand, to read a value, a low unidirectional current has to flow through the MTJ which is sensed using a sense amplifier.

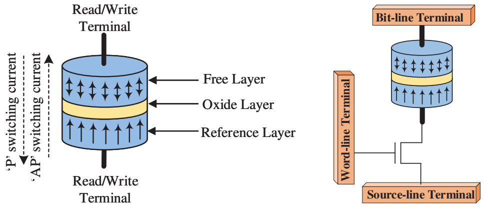
The *Thermal Stability Factor* (Δ) is an important parameter of the MTJ which is modeled as:

$$\Delta = \frac{V \cdot H_k \cdot M_s}{2 \cdot K_B \cdot T}, \tag{1}$$

where V, M_s, K_B, T and H_k are the volume of the free layer, the saturation magnetization, the Boltzmann constant, the temperature in Kelvin and the effective field anisotropy respectively. As shown in Eq. (1), Δ is proportional to the volume, and hence is affected by manufacturing variations.

A typical STT-MRAM bit-cell structure is shown in Fig. 1(b). It consists of one MTJ and an NMOS access transistor. A typical memory system consists of banks, mats and

978-3-9819263-0-9/DATE18/©2018 EDAA



(a) MTJ (b) Bit-cell

Fig. 1. Magnetic Tunnel Junction (MTJ) and STT-MRAM bit-cell.

subarrays [7]. The subarray is the basic building block of the memory system. It consists of the bit-cell array and the CMOS based peripheral components which are used to select and drive the appropriate memory block for read and write operations.

The faults in STT-MRAM can be classified into reliability (transient) faults and permanent faults. The reliability faults, which occur mainly due to stochastic switching of the MTJ, manifest during lifetime operation of the chip and are non-deterministic. On the other hand, permanent or persistent faults are deterministic and can be repeated after the chip fabrication, resulting in fails at the tester. Extreme process variations and spot defects (opens and shorts) are the primary cause of these faults.

Error Correcting Code (ECC), Redundancy Repair (RR) and Fault Masking (FM) techniques are typically employed to mitigate faults in logic and memory chips [5]. ECCs are typically used to detect and correct reliability faults whereas RR and FM are typically used to repair permanent faults.

An ECC scheme for correcting e errors in k data bits is represented as ECC(n, k, e) where n is the word size and $n-k$ is the number of check bits. The storage overhead of ECC ($\frac{n-k}{n}$) increases as the number of errors e in the data increases. In RR techniques, a faulty row or column is replaced with a spare one. Hence these techniques result in a large overhead, since an entire row/column is required to repair even a single fault. This can be overcome by FM, but at the cost of more complex addressing and accessing schemes. There are also some solutions to improve the redundancy efficiency by combining the ECC and RR techniques [8]. A failure distribution map of the memory array can help in choosing the right combination of the above defect tolerance techniques.

III. Parametric variations in STT-MRAM

The fabrication of STT-MRAM requires two different fabrication processes, a magnetic-process for the MTJs and a CMOS-process for the access transistors and peripherals. Variations in either process can affect the characteristics of the memory.

A. Random Variations

1) Variations in MTJ: Imperfections in the magnetic manufacturing process cause variations in the MTJ parameters such as radius, M_s, H_k, thickness of the FL/RL and oxide thickness. These variations in turn affect Δ and the critical current (I_c). In this work, we lump all the parameter variations of the MTJ into radius variations. We have not considered individual variations of the other parameters due to the limitations of our MTJ model as well as the tractability of the statistical simulations. However this can easily be added to the analysis if it is supported by the used MTJ model. The dependence of I_c and Δ on the radius is modeled as [9]:

$$I_c, \Delta \propto r^2 \tag{2}$$

The variations in r alters the switching threshold current (critical current), resistance values and the TMR ratio [10, 11], resulting in read, write and retention failures. The write probability of a bit-cell can be modeled by the following equation [12]:

$$WP_{bit}(t) = exp\left[\frac{-\pi^2(I-1)\triangle}{4(Ie^{C(I-1)t}-1)}\right], \; I = \frac{I_w}{I_c}, \tag{3}$$

where t is the write period, I_w is the write current, I_c is the critical current and C is a constant determined by the material and technology parameters. The WER is given by:

$$WER_{bit}(t) = 1 - WP_{bit}(t). \tag{4}$$

A retention failure in STT-MRAM happens when the magnetic orientation of the MTJ spontaneously flips due to thermal noise, causing the bit-cell to lose its content. The retention failure probability (P_{RF}) for a given time period (t) can be computed as [13]:

$$P_{RF} = 1 - exp[-\frac{t}{\tau \cdot e^{\Delta}}] \tag{5}$$

2) Variations in access transistor: The STT-MRAM bit-cell is also influenced by variations in the CMOS access transistor. The variations in the CMOS fabrication process causes variations in the device threshold voltage (V_{th}) primarily due to RDF, LER and STI stress [14, 15]. The standard deviation of the threshold voltage (σV_{th}) due to these random variations is given by the Pelgrom law [16]:

$$\sigma V_{th} \propto \frac{1}{\sqrt{WL}}, \tag{6}$$

where L and W are the effective length and width of the transistors respectively.

3) Variations in peripheral circuitry: The variations in the threshold voltage as per Eq. 6 affects the on-current of the CMOS based peripheral components causing variations in the read/write current of the bit-cell. Hence the combined effect of bit-cell and peripheral variations significantly impact the overall access latency of the memory system. Furthermore, extreme parametric variations can cause the latencies to extend beyond the design margins, resulting in permanent faults.

B. Systematic Variations

In addition to the random variations described in Section III-A, STT-MRAM is also affected by systematic variations. These variations show strong spatial correlations, which means that the variations among neighboring cells are much smaller compared to cells that are far apart from each other. In this work, we model systematic variations in the radius (r) of the MTJ and the threshold voltage (V_{th}) of the access transistor. Since the MTJ manufacturing process is compatible with the CMOS process, we assume that the MTJ radius variation is similar to that of CMOS variations. We use the VARIUS tool [17] to obtain the correlation map for these parameters.

IV. Yield analysis Framework

The overall yield analysis flow is presented in Fig. 2. The correlation map of the parameters (r and V_{th}) are obtained assuming a gaussian distribution. For each of these correlation maps, we get the parametric failures (both permanent and reliability failures) by performing Monte-Carlo simulations for the entire memory system including the bit-cell and peripheral components. The yield is then obtained by performing Monte-Carlo over multiple maps (corresponding to different chip instances). We then explore the right combination and efficacy of different defect tolerance techniques to obtain a target yield.

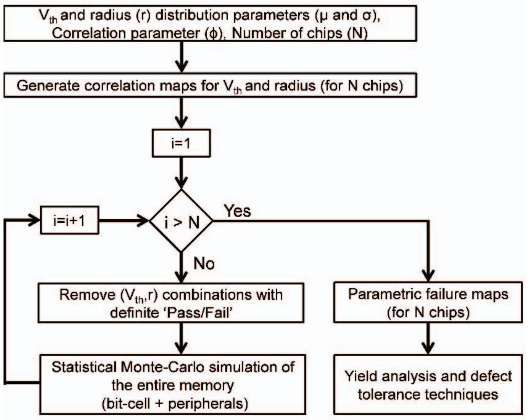

V_th and radius (r) distribution parameters (μ and σ), Correlation parameter (ϕ), Number of chips (N)
Generate correlation maps for V_th and radius (for N chips)
i=1
i=i+1
i > N
Yes
No
Remove (V_th,r) combinations with definite 'Pass/Fail'
Statistical Monte-Carlo simulation of the entire memory (bit-cell + peripherals)
Parametric failure maps (for N chips)
Yield analysis and defect tolerance techniques

Fig. 2. Proposed yield analysis flow.

A. Obtaining Correlation maps

The correlation map for the bit-cell parameters are obtained from VARIUS [17]. In this tool, the systematic variation is modeled using a multivariate normal distribution with a spherical spatial correlation structure as given in Eq. 7:

$$\rho(x) = \begin{cases} 1 - \frac{3x}{2\phi} + \frac{x^3}{3\phi^3}, & (x \leq \phi) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

In the above equation, $\rho(x)$ is the correlation function for two points separated by a distance x and ϕ is the correlation parameter, which specifies the range over which two points are correlated, expressed as a fraction of the chip's width. Two cells which are at a distance less than ϕ are assumed to be correlated while those with distance greater than ϕ have no correlation.

The input to the VARIUS tool are the mean (μ) and standard deviation (σ) of the parameter under consideration and also the correlation parameter (ϕ). We run Monte-Carlo simulations using VARIUS [17] to generate independent spatial correlation map of the parameters (V_{th} and r).

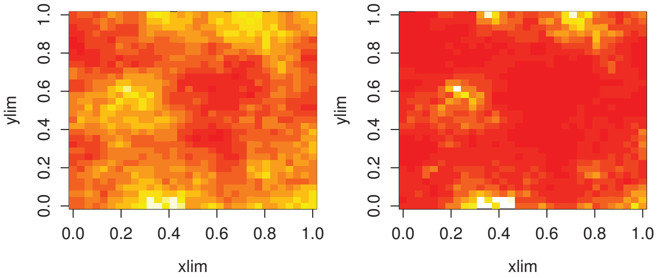
B. Permanent Fault Analysis

The permanent faults are deterministic and can be repeated at the tester. These faults are mainly caused due to extreme process variations or spot defects (opens and shorts). In this work, our focus is only on the permanent faults due to extreme process variations.

The standard 1T-1MTJ bit-cell structure is used for our simulations. The read/write margins are fixed based on the worst-case operating conditions of the cell. For instance, for the write operation, the worst-case operating conditions are minimum supply voltage and temperature, and maximum threshold voltage and radius. Then for extreme variations, we increase the variations in the parameters (V_{th} and r) beyond the nominal variation.

The first step is to get the correlation maps of V_{th} and r as explained in Section IV-A. Next, the latency distributions of the peripheral components are obtained using a hierarchical and hybrid Monte-Carlo approach as proposed in [18]. Then, for each bit-cell, depending on the specific V_{th} and r values for the bit-cell as well as the periphery path, SPICE simulations are performed to determine whether the cell is functional or not (based on the provided margins). Then, using the Monte-Carlo method, the above process is repeated for all the bit-cells in a memory array to obtain the fault distribution map.

The fault map for read and write thus obtained for one of the Monte-Carlo runs for a 32×32 memory array is shown in Fig. 3. The analysis of the fault map from different Monte-Carlo runs gives the number of failures and their distribution in the memory array, which can provide insights into the defect tolerance techniques required for yield improvement.



(a) Read failure distribution map (b) Write failure distribution map

Fig. 3. Permanent read and write failure distribution map for a 32×32 memory array for radius (μ = 20 nm, σ = 6%), V_{th} (μ = 397.9 mV, σ = 3.76%) and ϕ = 0.5. Lighter colors indicate potential failure points.

C. Reliability Fault Analysis

The reliability (transient) faults are non-deterministic faults occurring primarily due the stochastic switching of the MTJ, and are typically expressed by respective error rates. It can be seen that these failures primarily depend on the thermal stability factor Δ and the write current, which in turn depends on the radius (r) and the threshold voltage (V_{th}). Hence different bit-cells have different failure rates, according to their process points (r, V_{th}).

If e_i is the failure probability of the i^{th} bit-cell and n is the word size, then the failure probability (error rate) of the entire word, E is given by:

$$E = 1 - \prod_{i=1}^{n}(1 - e_i) \tag{8}$$

The word error rate E specifies the number of reliability faults per memory access. Since these faults happen in the field, the error rates should be kept to a minimum. For instance, the target values of WER for a memory array should be around 10^{-9} or lower [2].

D. Yield Exploration

Yield exploration can be done from the failure maps of different Monte-Carlo runs corresponding to different chip instances and by analyzing the number of faults in a row or column. If there are large number of faults per row or column, i.e., the faults are clustered, then RR is a good technique to mitigate these faults. On the other hand, for a small number of faults per row or column, i.e., when faults are more uniformly distributed, ECC might be a good option. In case of more number of single isolated faults, advanced techniques such as those proposed in [5] could be optimal for yield improvement.

Besides the conventional yield improvement techniques, we also explore some of the techniques specific to STT-MRAM. Since the switching probability and the latency of STT-MRAM is highly sensitive to the write current, current boosting can significantly decrease the write latency resulting in reduced write failures. This current boosting can be achieved by increasing the transistor sizing of write drivers. However, the amount of current boosting is limited to ensure that it does not lead to oxide barrier breakdown of the MTJs. Hence a combination of current boosting and traditional techniques can be the most effective for yield improvement with minimum overheads.

TABLE I. % OF CHIPS WITH THEIR FAULT TYPES FOR A 512×512 MEMORY ARRAY

Fault type		Maximum number of faults per line				
		0	1	2	3	> 3
Permanent	% of chips with write fault	72%	3%	2%	1%	22%
	% of chips with read fault	79%	4%	2%	1%	14%
Reliability	% of chips with retention fault	22%	2%	5%	6%	65%

TABLE II. LINE FAULT DISTRIBUTION FOR A 512×512 MEMORY ARRAY

Fault type	1-fault line	2-faults line	3-faults line	> 3 faults line
Write Fault	11.67%	7.19%	6.78%	74.35%
Read Fault	17.53%	10.21%	8.67%	63.58%
Retention Fault	11.38%	8.05%	6.04%	74.52%

V. RESULTS

A. Experimental Setup

We have employed the TSMC SPICE models for the CMOS access transistor and the Perpendicular Magnetic Anisotropy (PMA) MTJ model from [19]. The MTJ radius variation is assumed to be 5% whereas the threshold voltage variations in the CMOS components (bit-cell and periphery) are assumed to follow the Pelgrom law [16]. For extreme variations, we consider 20% extra variations compared to the nominal variations. We have done our analysis on a 512×512 memory array at 45 nm technology node.

B. Results

The fault distribution in the memory array is given in Table I. The percentage of chips with no permanent faults is 72% for write fault and 79% for read fault. Hence, with no defect tolerance techniques employed, the baseline yield is 72% considering read faults and 79% considering write faults. From Table II, it can be observed that there are a high percentage of rows with large number of faults (74.35% for write and 63.58% for read). This clustering of faults is because of the high correlation among the parameters of the neighboring cells.

Table III shows the yield improvement with different defect tolerance techniques and the respective area overhead costs. The storage area overhead for ECC is calculated from the number of ECC bits required to correct e errors and detect $e+1$ errors as $10e+1$ [20]. For the current boosting technique, the area and energy overhead for the additional circuitry are around 5.38% and 0.65% respectively, which is obtained from NVSim [7].

The results show that under the same area constraint, the current boosting technique is the most effective technique to mitigate write failures. However, there is a limit to the amount of current boosting possible, due to Time Dependent Dielectric Breakdown (TDDB) of the MTJ. Hence, we limit ourselves to around 10% current boost. Current boosting is not very effective to mitigate read faults, since an increase in current, although reduces the read decision faults, increases the probability of read disturb. It can also be seen that the ECC technique has the least effectiveness for yield improvement of permanent faults. The best combination for yield improvement is based on current boosting and modest RR.

In Table I, we also show the yield for one of the reliability failures, namely the retention failure. The results show that the yield is around 22%, which means that 78% of the chips are likely to have reliability failures due to short retention time in the field. The yield can be improved as shown in Table III, however even with RR of 10%, the yield improves to only around 61%. This observation is in line with those reported in other works such as [20], where retention failures are seen as a major reliability concern for STT-MRAM in advanced technology nodes.

TABLE III. YIELD IMPROVEMENT USING DEFECT TOLERANCE TECHNIQUES

Defect Tolerance Technique	Yield (without current boost)				Yield (10% current boost)	
	Write	Read	Retention	Area Overhead	Write	Area overhead
None	72%	79%	22%	0	95%	5.38%
ECC-1	75%	83%	24%	2.15%	95%	7.53%
ECC-2	77%	85%	29%	4.10%	95%	9.48%
ECC-3	78%	86%	35%	6.05%	96%	11.43%
RR (2%)	83%	86%	34%	2%	98%	7.38%
RR (4%)	85%	91%	44%	4%	99%	9.38%
RR (6%)	90%	95%	48%	6%	100%	11.38%
RR (8%)	95%	97%	54%	8%	100%	13.38%
RR (10%)	97%	97%	61%	10%	100%	15.38%

VI. CONCLUSIONS

In this work, we propose a framework for yield analysis of STT-MRAM memory arrays considering reliability and permanent faults due to parametric variations. We have considered the variations in the bit-cell and the peripheral components as well as the spatial correlation among the bit-cells in our analysis. The results show that the above considerations lead to a different fault distribution map as compared to previous works. Our framework also allows the designer to perform a design-for-yield exploration and investigate the efficacy of different defect tolerance techniques.

VII. ACKNOWLEDGEMENT

This work was supported in part by the European Commission under Horizon-2020 Program as part of the GREAT project (http://www.great-research.eu/) and in part by ANR/DFG as part of the MASTA project.

REFERENCES

[1] N. S. Kim *et al.*, “Leakage current: Moore’s law meets static power,” *computer*, vol. 36, no. 12, pp. 68–75, 2003.

[2] D. Apalkov *et al.*, “Spin-transfer torque magnetic random access memory (STT-MRAM),” *JETC*, vol. 9, no. 2, p. 13, 2013.

[3] Y. Ye *et al.*, “Statistical modeling and simulation of threshold variation under dopant fluctuations and line-edge roughness,” in *DAC*, pp. 900–905, ACM, 2008.

[4] A. Chintaluri *et al.*, “Analysis of defects and variations in embedded spin transfer torque (stt) mram arrays,” *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 6, no. 3, pp. 319–329, 2016.

[5] W. Kang *et al.*, “Yield and reliability improvement techniques for emerging non-volatile STT-MRAM,” *IEEE Trans. Emerg. Sel. Topics Circuits Syst*, vol. 5, no. 1, pp. 28–39, 2015.

[6] Y. Zhang *et al.*, “Persistent and nonpersistent error optimization for STT-RAM cell design,” *TCAD*, vol. 36, no. 7, pp. 1181–1192, 2017.

[7] X. Dong *et al.*, “NVSim: A circuit-level performance, energy, and area model for emerging non-volatile memory,” in *Emerging Memory Technologies*, pp. 15–50, Springer, 2014.

[8] C.-L. Su *et al.*, “An integrated ECC and redundancy repair scheme for memory reliability enhancement,” in *DFT*, pp. 81–89, 2005.

[9] W. Kang *et al.*, “Reconfigurable codesign of STT-MRAM under process variations in deeply scaled technology,” *TED*, vol. 62, no. 6, pp. 1769–1777, 2015.

[10] J. Li *et al.*, “Modeling of failure probability and statistical design of spin-torque transfer magnetic random access memory (STT MRAM) array for yield enhancement,” in *DAC*, pp. 278–283, 2008.

[11] W. Zhao *et al.*, “Failure analysis in magnetic tunnel junction nanopillar with interfacial perpendicular magnetic anisotropy,” *Materials*, vol. 9, no. 1, p. 41, 2016.

[12] K. Munira *et al.*, “A quasi-analytical model for energy-delay-reliability tradeoff studies during write operations in a perpendicular STT-RAM cell,” *TED*, vol. 59, no. 8, pp. 2221–2226, 2012.

[13] C. W. Smullen *et al.*, “Relaxing non-volatility for fast and energy-efficient STT-RAM caches,” in *HPCA*, pp. 50–61, 2011.

[14] Y. Ye *et al.*, “Statistical modeling and simulation of threshold variation under random dopant fluctuations and line-edge roughness,” *TVLSI*, vol. 19, no. 6, pp. 987–996, 2011.

[15] S. Borkar *et al.*, “Parameter variations and impact on circuits and microarchitecture,” in *DAC*, pp. 338–342, 2003.

[16] M. J. Pelgrom *et al.*, “Matching properties of mos transistors,” *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, 1989.

[17] S. R. Sarangi *et al.*, “VARIUS: A model of process variation and resulting timing errors for microarchitects,” *IEEE Trans. Semicond. Manuf.*, vol. 21, no. 1, pp. 3–13, 2008.

[18] S. M. Nair *et al.*, “VAET-STT: A variation aware estimator tool for STT-MRAM based memories,” in *DATE*, pp. 1456–1461, 2017.

[19] W. Guo *et al.*, “Spice modelling of magnetic tunnel junctions written by spin-transfer torque,” *Journal of Physics D: Applied Physics*, vol. 43, no. 21, p. 215001, 2010.

[20] H. Naeimi *et al.*, “STTRAM scaling and retention failure.,” *Intel Technology Journal*, vol. 17, no. 1, pp. 54–75, 2013.