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 exacerbate its stochastic switching behavior, leading to both test time failure 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, low 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 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 ($\Delta$) is an important parameter of the MTJ which is modeled as:

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

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), $\Delta$ 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
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 nondeterministic. 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, $R_{c}$, thickness of the FL/RL and oxide thickness. These variations in turn affect $\Delta$ 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 $\Delta$ on the radius is modeled as [9]:

$$I_{c}, \Delta \propto r^2$$

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]:

$$WP_{bit}(t) = \exp \left[ -\frac{t}{4\frac{I_{w}}{I_{c}}} \right]$$

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:

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

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]:

$$P_{RF} = 1 - \exp \left[ -\frac{t}{\tau_{\Delta}} \right]$$

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 ($\sigma V_{th}$) due to these random variations is given by the Pelgrom law [16]:

$$\sigma V_{th} \propto \frac{1}{\sqrt{W/L}}.$$
variation.

B. Permanent Fault Analysis

The variations in the parameters (Vth and r) beyond the nominal voltage and radius. Then for extreme variations, we increase the minimum supply voltage and temperature, and maximum threshold write operation, the worst-case operating conditions are mini-
simulations. The read/write margins are fixed based on the process variations.

Work, our focus is only on the permanent faults due to extreme at the tester. These faults are mainly caused due to extreme

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 Vth 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

making effective for yield improvement with minimum overheads.

oxide barrier breakdown of the MTJs. Hence a combination of current boosting and traditional techniques can be the most effective.

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.

Design, Automation And Test in Europe (DATE 2018) 267
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 Pelgrum 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 45nm 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 read 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 I, it can be observed that there are a high percentage of rows with large number of faults (74.35% for read and 63.58% for write). 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 faults. However, there is a limit to the amount of current boosting possible, due to Time Dependent Dielectric Breakdown (TDBD) 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 failures, 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.

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