Novel Magnetic Burn-In for Retention Testing of STTRAM

Mohammad Nasim Imtiaz Khan, Anirudh S Iyengar and Swaroop Ghosh
School of Electrical Engineering and Computer Science
Pennsylvania State University, University Park, PA-16801, USA
{muk392, asi7 and szg212}@psu.edu

Abstract*—Spin-Transfer Torque RAM (STTRAM) is an emerging Non-Volatile Memory (NVM) technology that has drawn significant attention due to complete elimination of bitcell leakage. However, it brings new challenges in characterizing the retention time of the array during test. Significant shift of retention time under static (process variation (PV)) and dynamic (voltage, temperature fluctuation) variability furthers this issue. In this paper, we propose a novel magnetic burn-in (MBI) test which can be implemented with minimal changes in the existing test flow to enable STTRAM retention testing at short test time. The magnetic burn-in is also combined with thermal burn-in (MBI+BI) for further compression of retention and test time. Simulation results indicate MBI with 220Oe (at 25C) can improve the test time by $3.71x10^{13}$ X while MBI+BI with 220Oe at 125C can improve the test time by $1.97x10^{14}$X.

I. INTRODUCTION

At the end of Silicon roadmap, several emerging non-Silicon memory technologies such as Ferroelectric RAM (FeRAM), STTRAM, and Resistive RAM (ReRAM) have surfaced [1-4]. STTRAM in particular offers high density, performance and endurance compared to the other memory technologies. Although attractive, PV and temperature fluctuations affect the bit-to-bit retention time of the STTRAM significantly. The variation become much worse at lower retention time and large LLC. Established volatile memory technologies such as, Dynamic RAM (DRAM) and embedded DRAM (eDRAM) undergo special retention tests to certify the refresh rate [2]. Although SRAM does not require retention tests, it does encounter random failures due to soft errors [5]. In matured NVMs such as, NAND-flash and NOR-flash, the retention is attributed to their manufacture specifications, and hence are not subjected to aggressive retention testing.

Conventional test flow lacks retention test methodology for NVMs in the design. The eDRAM retention tests cannot be extended to STTRAM due to test time overhead. Fig. 1 (a) describes the traditional test flow for Integrated Circuit (IC) testing. After manufacturing, the ICs go through wafer-level test where the defective ICs are identified using quiescent current test (I_{DDQ}) and discarded. Next, the wafers go through burn-in test where the temperature and voltage is elevated, few basic tests are run and the passing chips are diced out and full functional and structural test is performed on them. Finally, the chips are packaged after which another round of test is performed to screen out faulty chips. For eDRAM, the retention test is carried out after burn-in in order to only test the passing chips [5]. However, the same test flow will degrade the test time for STTRAM significantly, since the retention is compressed from a few years to a few seconds. Therefore, retention test time compression is desirable to ensure a small test time and a faster time to market.

Another challenge is stochastic nature of the retention since the retention time of the same bit can fluctuate randomly depending on the environmental noise. Other issues include dependency of retention time to manufacturing process variations, temperature and disturb current. Traditional test techniques fail to capture these bitcell-specific behaviors and is too time consuming to be practical. A retention test methodology for STTRAM is suggested in [6] however, the details of implementation is not described. A Design-for-Test (DFT) technique is proposed in [7] to compress the retention time from few secs to few µs by injecting constant DC disturb current during test. If the bitcells meet the compressed retention time then it indirectly confirms the target retention time. The usage of temperature (during burn-in) is also proposed to compress the test time. However, the method is power hungry and requires major DFT changes.

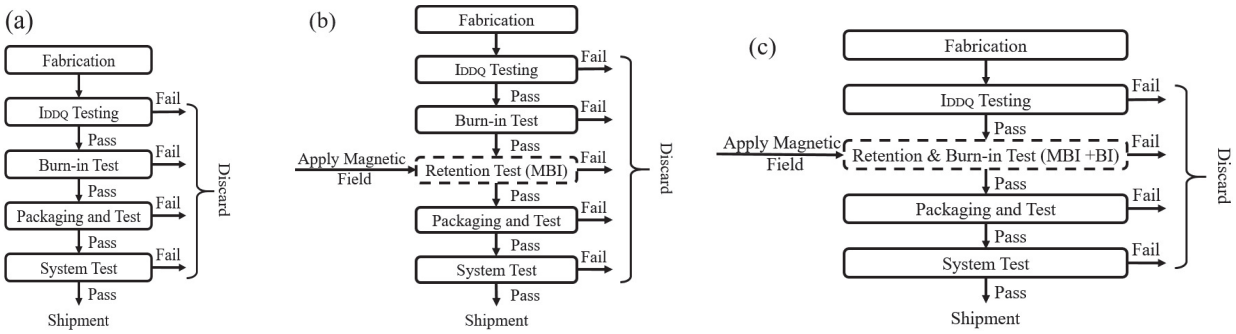

Fig. 1(a) Traditional test flow; (b) proposed magnetic burn-in (MBI); and (c) proposed magnetic and thermal burn-in (MBI+BI).

978-3-9815370-8-6/17/$31.00 ©2017 IEEE

In this paper, we propose to introduce a magnetic burn-in in the test flow, which is much similar to conventional burn-in, will compress the retention time of the bits for reduced test time. Fig. 1(b) shows the modified test flow where Magnetic Burn-In (MBI) is introduced after the thermal burn-in (BI) to characterize the retention time. A magnetic chamber with features to run read/write operations on memory is needed to conduct the MBI. Fig. 1(c) shows an alternative test flow where MBI is combined with conventional BI (MBI+BI). The advantage is higher compression of retention time due to both magnetic field and temperature which further improves the test time. Although attractive, MBI+BI will require a chamber equipped with magnetic field. Since conventional BI can only run some basic tests, the proposed MBI+BI will require enhancement of BI flow to incorporate retention test. *To best of our knowledge, this is the first attempt to study the STTRAM retention test time reduction using magnetic field burn-in*. We make following contributions in this paper:

- Compress the test time using MBI.
- Improve the test time further by combining MBI with BI.
- Does not incur DFT overhead, maintains low test power
- Maintain minimal changes in the existing test flow.

II. STTRAM Basics and Retention Time

A. Basics of STTRAM: Fig. 2(a) shows the STTRAM cell schematic with Magnetic Tunnel Junction (MTJ) as the storage element [8] [9]. The MTJ contains a free and a pinned magnetic layer. The resistance of the MTJ stack is high (low) if free layer magnetic orientation is anti-parallel (parallel) compared to the fixed layer. The MTJ can be toggled from parallel (data '0') to anti-parallel (data '1') (or vice versa) using current induced Spin Torque Transfer by passing the appropriate write current from source-line to bitline (or vice versa). For successful write, the write current must be greater than the critical current (I_{co}). The data in MTJ is stored in the form of magnetization. The two magnetization states of the free layer form the stable states separated by an energy barrier 'Δ' (Fig. 2(b)). To switch the magnetization from one state to another, the free layer is excited with enough energy (by passing current) to overcome this barrier. From (1)-(2) we note that Δ_H is a function of effective magnetic field and temperature [10]. Therefore, these parameters can be varied to lower the energy barrier and compress retention time during test mode.

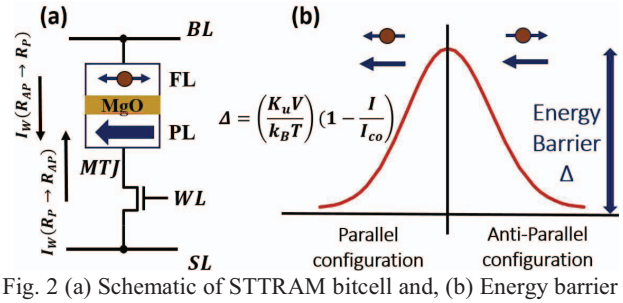

Fig. 2 (a) Schematic of STTRAM bitcell and, (b) Energy barrier separating the two MTJ states that determines the retention time.

B. Static Retention Modeling: The STTRAM retention time (t_{ret}) is given by $t_{ret} = t_0 \exp[\Delta]$ [11], where, t_0 is the attempt time (~1ns). Δ is calculated from (1), and the influence of an external field on Δ i.e. Δ_H, is obtained from (2). It is noted that the retention time is exponentially dependent on STTRAM dimensions, external field and ambient temperature. Fig. 3 shows the cumulative distribution function of retention time with applied magnetic field. A field of 220Oe, applied opposite to the free layer magnetic orientation, reduces the retention time from 1 year to <1μs. In this work, the dependency of Δ on external field is exploited for test time reduction and certify the bits under magnetic stress.

C. Stochastic Retention Modeling: Retention time of the

$$\Delta = E_0/k_BT = H_k M_s V/2k_BT \quad (1)$$

$$\Delta_H = \Delta\left(1 - \frac{H_{eff}}{H_k}\right)^2 \quad (2)$$

$$\overrightarrow{H_{th}} = \vec{\xi}\sqrt{2k_bT/\mu_o M_s V \Delta t} \quad (3)$$

$$\overrightarrow{H_{eff}} = \overrightarrow{H_{ext}} + \overrightarrow{H_{ant}} + \overrightarrow{H_{demag}} + \overrightarrow{H_{th}} \quad (4)$$

where, Δ_H= retention energy barrier in presence of applied external magnetic field, E_0= energy barrier, H_k= anisotropy magnetic field, M_s= saturation magnetization, V = the volume of the free layer, k_B= Boltzmann's constant, T= temperature, ξ= a standard Gaussian random variable in 3D space, μ_0= permeability off free space, Δ_t= constant time step used in the numerical simulation, H_{eff}= total effective magnetic field.

STTRAM becomes stochastic in presence of thermal noise. Therefore, the retention time fluctuates dynamically and certifying the bitcell retention requires multiple tests to capture the worst-case behavior. The combined effect of large number of tests and long test time makes the overall retention characterization a time-consuming process. Thermal excitation (noise) causes the magnetic moment of the STTRAM to precess about its easy axis leading to a variation of the initial angle (θ) [12,13]. In order to account for stochastic variation due to thermal noise, a corresponding field term (H_{th}) with zero mean is incorporated into the LLG ((3)-(4)). (4) is used in equation (2) to account for the thermal noise induced retention variation. In this work, all simulations are performed for MTJ of volume 1.0413 x10^{-23} m^3 with 1yr retention time. The MTJ was modelled on [14].

III. Retention Time Analysis

A. Impact of External Magnetic Field: From (1)-(2) we observe that retention time greatly depends on external magnetic field. To further analyze the impact of magnetic field, a 1-million-point Monte-Carlo analysis is conducted with a 3σ of 5% for MTJ dimensions, and with a mean retention time of 1 year. The magnetic field is applied opposite to the magnetic orientation of the free layer. Fig. 3 shows the cumulative dis-

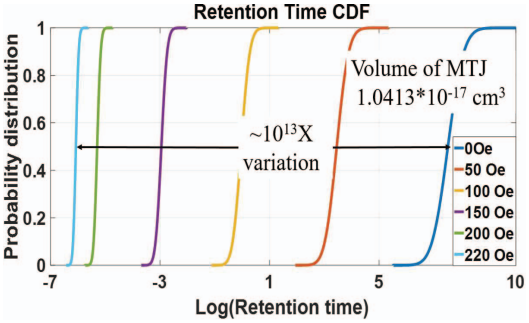

Fig. 3 Impact of magnetic field on retention time.

K_u=150150 erg/cc
Ms=790Oe.
Hk=380Oe Δ_0=37.99
Analysis was done with 3σ of 5% for MTJ dimensions.

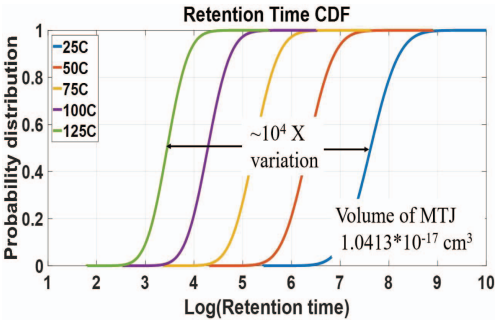

Fig. 4 Impact of temperature on retention time.

tribution function of the retention time under different magnitude of magnetic field (30Oe to 220Oe). We observe that by increasing the external magnetic field (H_{eff}), the retention time decreases exponentially. Without any external magnetic field, the retention time is ~1 year, however, the mean retention is compressed by ~1.15x10^{4}X, ~3.55x10^{7}X and ~3.89x10^{13}X for 50Oe, 100Oe and 220Oe respectively.

B. Impact of Temperature: From (1)-(2) we also note that retention time is depentent on the temperature. Fig. 4 shows the cumulative distribution function of the retention times under different temperatures (25°C to 125°C) using same setup as above. It is can be noted that under high temperatures, the retention time is drastically reduced (by ~10^{4}X @ 125°C), which can therefore be exploited to reduce test time.

C. Impact of Other Factors: Temperature has a two-fold impact on the retention time of a STTRAM array: (i) it lowers the energy barrier; and, (ii) it shifts the magnetization from its easy axis due to noise. Any displacement of the initial angle from the easy axis reduces the retention time. Therefore, the same bit can exhibit multiple retention times when tested multiple times. The worst-case retention time must be identified to guarantee functional correctness of the memory. Fig. 5 shows the stochastic retention time distribution of the base design, with MBI and with MBI+BI. It is evident that the distribution remains similar. Therefore, the worst-case retention time found using MBI+BI and/or MBI can be mapped to the corresponding worst case of the base case.

IV. Retention Time Compression and Testing

In this section, we discuss the proposed MBI and MBI+BI based test time compression. We then describe the retention time identification using a linear and binary search scheme on the compressed retention time.

A. *Compression using MBI:* The high-level view of the test process in MBI is as follows:

- First, a pre-defined data pattern ('1' or '0') is written to the entire memory array.
- A known magnetic field is applied in the direction opposite to the magnetic orientation of the free layer of the memory array. This compresses the retention time of the bits by a known factor, let's say by '*Y*' times.
- The retention time is identified by employing either the linear or binary search process (discussed in Section IV.C).
- Finally, the retention time is decompressed by multiplying the measured retention time by '*Y*'.

This technique is a simple and effective way to reduce the retention time. However, an additional step in the test flow is required (Fig. 1(b)).

B. Compression using MBI+BI: In this approach (with same steps as above), we exploit both MBI and BI to compress the retention time of the STTRAM reducing the overall test time further. By exposing the 'hot' chips during burn-in to an external magnetic field, a cumulative effect in retention time compression is achieved. This method poses two benefits over the previous approach: (i) no need for an additional test step since magnetic field could be applied in the conven-

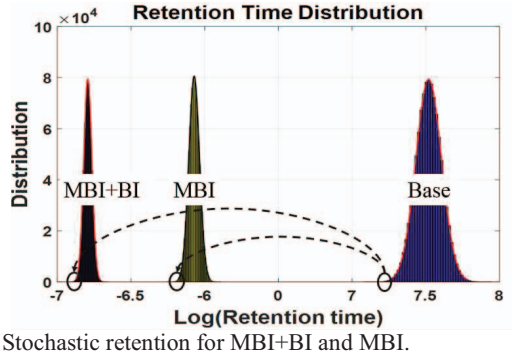

Fig. 5 Stochastic retention for MBI+BI and MBI.

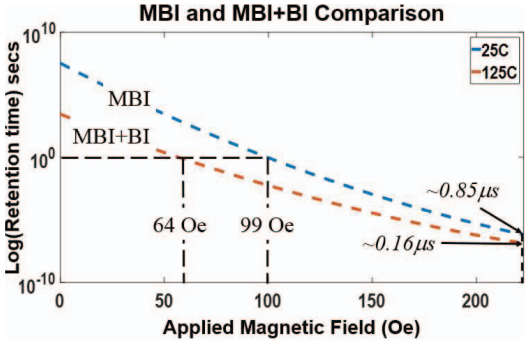

Fig. 6 Comparison between MBI with MBI+BI. It can be observed that at 0Oe, MBI+BI (@125°C) can compress the retention time by ~1.17x10^{4}X, at 100Oe can compress by ~5.7x10^{9}X and at 220Oe can compress by ~2.05x10^{14}X.

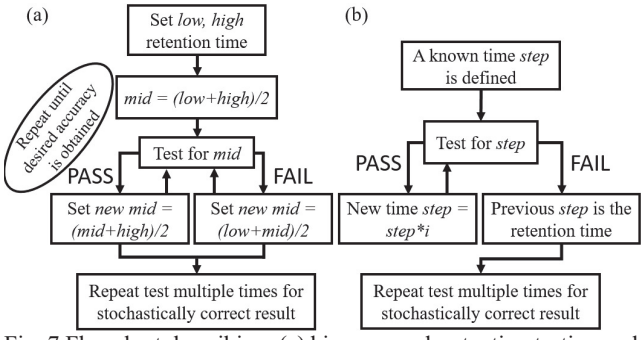

Fig. 7 Flowchart describing: (a) binary search retention testing and, (b) linear search retention testing.

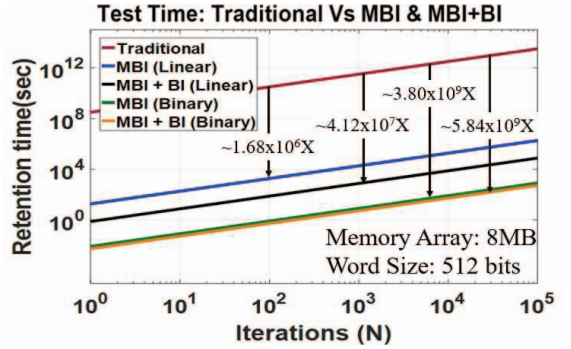

Fig. 8 Retention test time comparison using LS & BS.

tional burn-in chamber; and, (ii) by combining the compression obtained by both temperature and magnetic field, a much lower magnetic field will be required to achieve the same degree of compression. Thus, reducing test power. Fig. 6 shows a comparison between MBI and MBI+BI. It can be noted that, for same degree of compression, MBI+BI requires only 64Oe magnetic field as compared to 99Oe for BI. It can also be noted that MBI with 220Oe (@25°C) can reduce the retention time from 1 year to ~0.85µs ($3.71x10^{13}$X compression) while MBI+BI with 220Oe @125°C can further reduce the retention time to ~0.16µs ($1.97x10^{14}$X compression). Therefore, MBI+BI (5.3X better with 220Oe @125°C) is a more efficient approach for retention time testing.

Regardless of the retention compression approach employed (MBI or MBI+BI), multiple iterations of retention testing due to stochastic variations, is necessary to obtain statistically correct (worst case) retention time.

C. Linear/Binary Retention Time Search: We employ two methods of retention search namely, (a) Binary Search (BS) and, (b) Linear Search (LS) [7]. The search method is altered to capture the field assisted retention time compression. The BS routine (Fig. 7 (a)), utilizes the mean of an upper limit ('high') and a lower limit ('low') of the possible compressed retention time. The retention time of the array is compared with the 'mid', and if the worst-case retention time exceeds this 'mid' then the 'new mid' value is the average of 'mid' and 'high', and if the worst-case retention is below this 'mid' value, then the 'new mid' is the average of the 'low' and 'mid'. The process is repeated for multiple iterations in order to obtain a reasonably high level of accuracy.

In LS algorithm (Fig. 7(b)) a known step is chosen, following which the array is tested for retention in multiples of step until a failure is observed. The process is repeated for the rest of the memory elements. Finally, the worst-case retention time equal to the lowest step for failure is observed. The step value is the resolution of the retention test i.e. the retention is tested in multiples of this time step. The process is repeated until all the bits are written into and read periodically. It is evident that a low time step will provide a high level of accuracy and vice versa. Both LS and BS is repeated multiple times to ensure that a stochastically worst case retention time is obtained. Fig. 8 shows the retention test time comparison using LS and BS routine for 8MB of memory array with 512 bits of word size. We note that, for LS (BS), MBI improves test time by ~$1.68x10^6$X (~$3.80x10^9$X) and MBI+BI improves test time by ~$4.12x10^7$X (~$5.84x10^9$X) as compared to traditional method.

V. Conclusion

In this paper, we proposed to exploit susceptibility of STTRAM to magnetic field and temperature for test time reduction. We introduced two techniques namely, magnetic burn-in and combination of magnetic and thermal burn-in for retention time compression during test. The proposed techniques improve the test time by several orders of magnitude.

Acknowledgment

We acknowledge the support of NSF grants CNS- 1441757 and SRC grant 2442.001.

References

[1] Kryder, Mark H., et al. "After hard drives—what comes next?."*Magnetics, IEEE Transactions on* 45, (2009).
[2] Hamzaoglu, Fatih, et al. "13.1 A 1Gb 2GHz embedded DRAM in 22nm tri-gate CMOS technology." *ISSCC, 2014*
[3] Swaroop Ghosh, et al, "Security and Privacy Threats to On-Chip Non-Volatile Memories and Countermeasures?." *ICCAD,* 2016.
[4] Haron, Nor Zaidi, et al. "DfT schemes for resistive open defects in RRAMs." DATE, 2012.
[5] S. Mittal, et al, "A Survey Of Architectural Approaches for Managing Embedded DRAM and Non-Volatile On-Chip Caches," *TPDS*, 2015
[6] Naeimi, Helia, et al. "STTRAM scaling and retention failure." Intel Technology Journal 17, 2013.
[7] A. Iyengar, et al, "Retention Testing Methodology for STTRAM," D&T, 2016.
[8] J. Jang et al. "Self-Correcting STTRAM Under Magnetic Field Attack", DAC, 2015
[9] S. Ghosh, "Spintronics and Security: Prospects, Vulnerabilities, Attack Models, and Preventions," in *Proceedings of the IEEE*.
[10] https://arxiv.org/ftp/arxiv/papers/1107/1107.5007.pdf
[11] Raychowdhury, Arijit, et al. "Design space and scalability exploration of 1T-1STT MTJ memory arrays in the presence of variability and disturbances." *IEDM,* 2009.
[12] Sun, J. Z. "Spin-current interaction with a monodomain magnetic body: A model study." Physical Review B 62, 2000.
[13] Brown Jr, et al. "Thermal fluctuations of a single□domain particle." *JAP*,1963.
[14] S. Srinivasan, "All spin logic: Modeling multi-magnet networks interacting via spin currents". PhD diss., Purdue University, 2012.