A Statistical STT-RAM Retention Model for Fast Memory Subsystem Designs

Zihao Liu†, Wujie Wen†, Lei Jiang‡, Yier Jin∗, and Gang Quan†

†Department of ECE, Florida International University, Miami, FL 33174 USA (e-mail:{zliu021, wwen, gaquan}@fiu.edu)
‡Department of ISE, Indiana University Bloomington, Bloomington, IN 47405 USA (e-mail: jiang60@iu.edu)
∗Department of ECE, University of Central Florida, Orlando, FL 32816 USA (e-mail: yier.jin@eecs.ucf.edu).

***Abstract*—Spin-transfer torque random access memory (STT-RAM) is a promising nonvolatile memory (NVM) solution to implement on-chip caches and off-chip main memories for its high integration density and short access time, but it suffers from considerable write latency and energy overhead. Aggressively relaxing its non-volatility for write fast and write energy efficient memory subsystems has been quite debatable, due to the unclear retention behavior on a timescale of microseconds-to-seconds. Moreover, recent studies project that retention failure will eventually dominate the cell reliability as STT-RAM scales. As a result, a comprehensive understanding of the thermal noise induced STT-RAM retention mechanism has become a must. In this work, we develop a compact semi-analytical model for fast retention failure analysis. We then systematically analyze critical factors (e.g., initial angle, device dimension etc.) and their impacts on the STT-RAM retention behavior through our model. Our experimental results show that STT-RAM suffers from a soft-error style retention failure, which may happen instantly just after the last write finishes and is totally different from that of DRAM and Flash, i.e., the gradual charge loss process. Our model offers an excellent agreement with the results from golden macro-magnetic simulations in the region of interest without conducting expensive Monte-Carlo runs. At last, we demonstrate our model can enable architectural designers to rethink STT-RAM based memory designs by emphasizing its probabilistic retention property.**

I. Introduction

Moore's law continues by equipping a single chip with more and more cores enabling more applications to run concurrently on the same chip. To maintain scalable performance, modern computing systems require a scalable memory hierarchy to support the increasing working set size of these concurrent applications. However, traditional memory technologies, such as SRAM and DRAM, face a lot of severe challenges, e.g., significant leakage power, poor scalability and large process variation, at deep sub-micron technology nodes. Recent works identify spin-transfer torque random access memory (STT-RAM) as an universal memory solution to build on-chip caches [13], [11] and off-chip main memories [14], because of its short access time, high cell density and zero cell leakage. Today STT-RAM has become an off-the-shelf technology, i.e., after publishing several STT-RAM experimental chips ranging from 4Kb to 256Mb [7], [4], Everspin starts shipping 256Mb STT-RAM chips with DDR interfaces in April 2016 [3].

STT-RAMs use the low and high resistance states of magnetic tunneling junction (MTJ) to distinguish "0" and "1". The MTJ is able to maintain the data integrity for more than 10 years even without power supply. The MTJ resistance state can be changed normally as long as the energy field generated by a spin polarized current overcomes the MTJ energy barrier. Meanwhile, the resistance state of an idle MTJ is also subjected to change spontaneously, if the thermal noise induced magnetic field exceeds the energy barrier of MTJ. This is called a data retention failure. Unfortunately, the retention failure has emerged as one of the major reliability issues in the practical deployment of STT-RAM for two reasons: First, as process technology scales, the MTJ volume shrinking enlarges the thermal noise and decreases the MTJ energy barrier, leading to substantially exacerbated retention failures. A recent Intel study projects that retention failures will eventually surpass write / read errors and become the dominant cell level reliability factor at the $10nm$ technology node [5]. Second, relaxing STT-RAM retention is an effective method to facilitate the fast and energy-efficient STT-RAM cache designs. Based on the fact that the data in caches have much shorter lifetime (microseconds) than the retention time of STT-RAM (∼10 years), many researches accelerate write operations and lower write energy in STT-RAM based on-chip caches by aggressively sacrificing the retention time [13], [9]. Periodical refreshes are introduced to eliminate retention errors in caches. The effectiveness of the retention relaxing based scheme and the retention failure mitigation overhead heavily depend on the occurrence probability of STT-RAM retention errors. Therefore, it is critical to have an efficient method to qualify and quantify the STT-RAM retention failure.

The STT-RAM retention time estimation method in previous work [10] depending on the thermal activated model is originally used to analyze the MTJ flipping rate under a spin current. It can be summarized as: first, the thermal stability factor Δ is measured through the thermally activated model considering the effect of spin torque; second, the retention time is simply calculated as e^{Δ}. However, as retention errors are generated by random thermal noise, such a coarse-grained approach cannot provide full statistical information on retention time and is not applicable for the worst-case design in the region of interest (i.e. microseconds-to-seconds for cache). Moreover, since the thermal activated model adopted by the traditional retention estimation method [10] introduces a very small spin current in a large time window, it does not fully comply with the physical nature of retention error, i.e., a spontaneous bit flipping solely driven by the thermal noise. The measured results can be further distorted by the variations of injected currents and pulse widths.

In this paper, we propose a fast retention time analysis method which includes four integrated steps: "***Create-Transfer-Recover-Calculate***". Our contributions can be summarized as:

1) We calibrate the probabilistic STT-RAM retention time by carefully analyzing the distributions of the angle of the free layer magnetization at the interested time through solving the Landau-Lifshitz-Gilbert (LLG) equation. Compared to the previous method, our model is able to take both parametric and environmental variations into considerations, and deliver more accurate results without conducting time-consuming Monte-Carol simulations;
2) We translate all the critical parameters and their variabilities into an equivalent thermal noise value in our model for the fast and portable analysis of their impacts on STT-RAM's retention time. Experimental results show that the initial angle has little influence on the STT-RAM retention time;
3) Our results show that the estimation on STT-RAM retention made by previous work is too pessimistic. The lower-than-expected retention error rate enables the architectural designer to reduce the STT-RAM refresh frequency.

978-1-5090-1558-0/17/$31.00 ©2017 IEEE

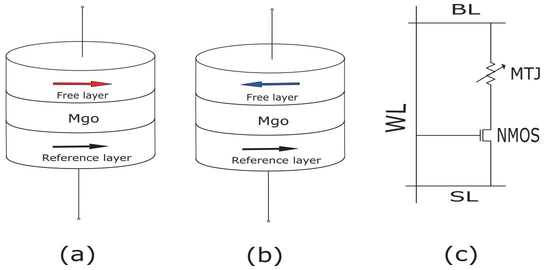


Fig. 1. STT-RAM basics. (a) Parallel (low resistance). (b) Anti-parallel (high resistance). (c) 1T1J cell structure.

II. Preliminary and Research Motivation

A. STT-RAM Basics

STT-RAM relies on the resistance state of Magnetic Tunnel Junction to store data. As Fig. 1(c) shows, the "1T1J" STT-RAM cell consists of one MTJ and one NMOS transistor. The typical MTJ device comprises a oxide barrier layer sandwiched between two ferromagnetic layers: one is reference layer with a fixed magnetization direction; the other is the free layer that can be switched by applying spin pulses with different polarizations. When the magnetization directions of two layers are parallel (anti-parallel), the MTJ has a low (high) resistance, indicating a '0' ('1'), as Fig. 1(a), (b) shows.

B. Retention Failure

The *retention failure* of STT-RAM is a phenomenon where an idle cell flips without applying any intentional excitation source. The *retention time* of a MTJ is a metric of the expected time until a data retention failure happens. Theoretically, within a sufficient long time, a retention failure can occur on a STT-RAM cell eventually because of the intrinsic thermal noise [16]. And the retention failure probability (rate) heavily depends on thermal noise, MTJ device setting and temperature etc. As STT-RAM scales aggressively into deep sub-micron nodes, significant variations appearing in MTJ dimensions and magnetic material parameters aggravate the retention errors. Furthermore, previous works [13], [9] deliberately shorten the MTJ retention time, so that STT-RAM writes can be improved in terms of latency and energy. More frequent retention errors are generated, demanding carefully designed refresh schemes. Hence it is critical to understand the occurrence probability of STT-RAM retention failures.

Previous studies [9], [10], [17] usually estimate the STT-RAM retention time (t_{ret}) by Eq.(1):

$$ln(t_{ret}) = \frac{A \cdot t_m \cdot H_k \cdot M_s}{2k_b \cdot T} \tag{1}$$

where A and t_m denote the MTJ area and its free layer thickness; M_s is the saturation magnetization; H_k indicates the uniaxial anisotropy factor; k_b is the Boltzmann's constant; and T is the operating temperature. However, this method suffers from several limitations: 1) The full statistical information of the retention time is not available; 2) The device and environmental variations, as well as the initial magnetic status, are not taken into considerations; 3) This equation is derived by assuming an excitation source–spin current is injected in MTJ, while an unintentional bit flipping is solely triggered by the thermal noise.

III. Thermal Induced Model and Validation

To address the looming retention crisis and overcome the limitation of the existing STT-RAM retention analysis, we introduce our systematic and accurate retention evaluation framework. Our macro-magnetic model can demonstrate and quantify precisely the stochastic retention behavior of STT-RAM.

A. Modeling

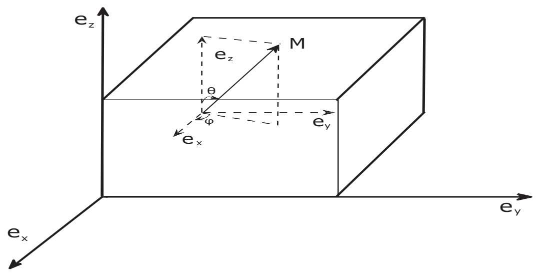


Fig. 2. MTJ model schematic diagram.

Fig. 2 illustrates the free layer magnetization ($\mathbf{M}$) of a MTJ cell. The magnetization switching process of a MTJ is determined by the direction changes of $\mathbf{M}$. The motion of $\mathbf{M}$ is usually depicted as a unit direction vector $\mathbf{n}_m = \mathbf{M}/|\mathbf{M}|$, since the magnitude of $\mathbf{M}$ remains as a constant value. As shown in Fig. 2, the coordinates of θ and φ, which are defined as the relative angles of $\mathbf{n}_m$ with respect to $\mathbf{e}_z$ and $\mathbf{e}_x$, can completely describe the dynamic behavior of $\mathbf{M}$ at any time. In general, the MTJ can be switched to a stable status, i.e. $\theta = 0$ $(\theta = \pi)$ for low (high) resistant state, by injecting an appropriate spin current with certain pulse width or the inherent thermal noise for a long time. To completely eliminate the impact of spin current and guarantee that thermal induced magnetic field is the only excitation source during data retention, the total energy of $\mathbf{M}$ is modeled as $U(\theta, \varphi) = U_K + U_p + U_l$. Here, the potential energy includes the uniaxial anisotropy energy $U_K = K\sin^2\theta$ and easy-plane anisotropy $U_p = K_P(\sin^2\theta\cos^2\varphi - 1)$, where $K = (1/2)M_sH_k$, $K_p = 2\pi {M_s}^2$, H_k and M_s are the Stoner-Wohlfarth switching field and saturation magnetization, respectively [12]. The thermal field energy U_l is the only excitation causing a possible data flipping and its corresponding Langevin random field is modeled as [6]:

$$h_{l,i} = h_l \mathbf{X}_i(t) = \sqrt{\frac{\alpha}{\alpha^2+1}\frac{2k_bT}{\gamma\mu_0\Delta t\Delta v M_s}}\mathbf{X}_i(t)\,(i = x, y, z) \tag{2}$$

where h_l indicates the noise intensity highly related to system temperature T and free layer volume Δv; α is Gilbert damping constant; k_b denotes Boltzmann constant; γ is Gyromagnetic ratio; μ_0 represents Permeability of free space; Δt is the time step for simulation. $\mathbf{X}_i(t)$ is a three dimensional Gaussian random noise with zero mean and unit variance in x, y and z axis.

The three elements in total energy U can generate three different torques [15]: the uniaxial anisotropy term:

$$\frac{\Gamma_1}{t_mK} = (2\sin\theta\cos\theta)[(\sin\varphi)e_x - (\cos\varphi)e_y] \tag{3}$$

The easy-plane anisotropy term:

$$\frac{\Gamma_2}{t_mK} = -2h_p[(\cos\theta\sin\theta\cos\varphi)e_y - (\cos\varphi\sin\varphi{\sin}^2\theta)e_z] \tag{4}$$

The Langevin random field term:

$$\begin{aligned}\tfrac{\Gamma_3}{t_mK} &= (h_{l,z}\sin\theta\sin\varphi - h_{l,y}\cos\theta)e_x + h_{l,x}\cos\theta e_y \\ &-h_{l,z}\sin\theta\cos\varphi e_y + (h_{l,y}\sin\theta\cos\varphi - h_{l,x}\sin\theta\sin\varphi)e_z\end{aligned} \tag{5}$$

Here t_m is the thickness of MTJ.

By substituting the aforementioned three torque terms into Landau-Lifshitz-Gilbert equation (LLG) [12], the motion change of the free layer magnetization vector with respect to time during data retention

TABLE I
DEVICE SETTINGS

Parameter	MTJ cell	Unit
H_k	100	m/As
ΔV	$200 \times 100 \times 2$	nm^3
M_s	1050	emu/cm^3
Initial θ	0.01	rad
Damping rate α	0.01	N/A

can be described as follows:

$$\begin{bmatrix} \theta' \\ \varphi' \end{bmatrix} = \sum_{i=1}^{3} \begin{bmatrix} \theta_i' \\ \varphi_i' \end{bmatrix} \tag{6}$$

Here the uniaxial anisotropy term is:

$$\begin{bmatrix} \theta_1' \\ \varphi_1' \end{bmatrix} = - \begin{bmatrix} \alpha \sin\theta \cos\theta \\ \cos\theta \end{bmatrix} \tag{7}$$

The easy-plane anisotropy term is:

$$\begin{bmatrix} \theta_2' \\ \varphi_2' \end{bmatrix} = -h_p \begin{bmatrix} (\sin\varphi + \alpha\cos\theta\cos\varphi)\sin\theta\cos\varphi \\ (\cos\varphi\cos\theta - \alpha\sin\varphi)\cos\varphi \end{bmatrix} \tag{8}$$

The Langevin random field term is:

$$\begin{bmatrix} \theta_3' \\ \varphi_3' \end{bmatrix} = - \begin{bmatrix} \alpha[h_{l,z}\sin\theta - \cos\theta(h_{l,x}\cos\varphi + h_{l,y}\sin\varphi)] \\ +(h_{l,y}\cos\varphi - h_{l,x}\sin\varphi) \\ [\alpha(h_{l,x}\sin\varphi - h_{l,y}\cos\varphi) + h_{l,z}\sin\theta \\ -\cos\theta(h_{l,x}\cos\varphi + h_{l,y}\sin\varphi)]/\sin\theta \end{bmatrix} \tag{9}$$

Substitute above three terms into Eq. (6), we can get an analytical model to simulate the motion change of the magnetization by using two ordinary differential equation at any time.

B. Model Validation

To validate our model, the fourth order Runge-Kutta algorithm is adopted to solve the aforementioned LLG ordinary differential equations. Monte-Carlo simulations are conducted to obtain the distributions of θ at simulated time. The device parameter settings are summarized in Table I [6]. The time step of our simulation is $\Delta t = 10$ps.

Fig. 3 shows our simulated probability density function (PDF) of θ at a 200μs simulation time. The results are characterized after running 10^4 Montes-Carlo simulations. For comparison purpose, the analytical PDF of θ (see Eq. (10)) is also presented [6]:

$$P(\theta) = \frac{2H_K M_s}{k_b T} \exp(-\frac{H_K M_s}{k_b T} \sin^2\theta) \tag{10}$$

The simulated PDF is generally very close to the one obtained from Eq. (10). θ has a symmetric PDF with highest frequency of occurrence around zero, which is well consistent with the results expected from random thermal noise.

We also demonstrate how the Langevin random field can suddenly flip the magnetization of free layer from one orientation to the other randomly within certain time. As an example, we assume the initial angle of θ is $\pi/2$ and conduct two different experiments by intentionally injecting or removing the thermal noise field. As Fig. 4(a) shows, θ is gradually reduced from $\theta = \pi/2$ to $\theta = 0$ and stabilizes at 0 without applying thermal noise. The unstable initial state of θ can only go towards 0 at a certain damping rate by the intrinsic anisotropy field. However, after applying the random thermal noise field, the stable state of θ becomes unpredictable: either $\theta = 0$ or $\theta = \pi$, as shown in Fig. 4(b). The perturbation and oscillation momentum introduced by thermal noise can increase the possibility of flipping the resistance state of the MTJ, i.e. making θ permanently cross $\pi/2$ and eventually stay at π.

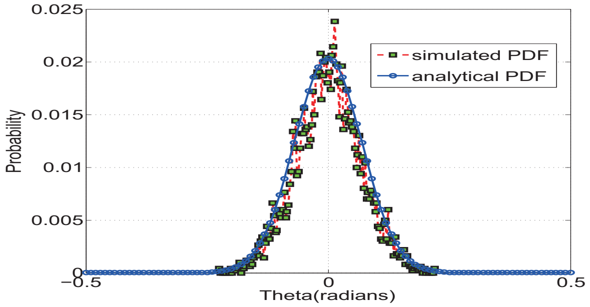


Fig. 3. Model validation of θ distribution.

IV. Methodology and Simulation

Similar as Flash memory [1], the retention time measurement of STT-RAM requires expensive Monte-Carlo simulations to capture a data retention failure that rarely happens within a sufficient long time window [5], i.e. one failure in 10^9 device hours continuous operation. Moreover, as Eq. (2) in section III shows, the key factors that can affect the stochastic retention behavior of an STT-RAM device may include the MTJ volume ΔV, system temperature T, the initial angle of θ etc. If we also take variations on such parameters (e.g., ΔV) into account, the simulation of data retention will become even difficult. In this section, we propose a fast and portable STT-RAM retention semi-analytical method, namely "***Create-Transfer-Recover-Calculate***", to characterize the full statistical information of data retention. A comprehensive analysis of the retention error behavior is also presented.

A. Methodology and Validations

The basic idea of our proposed "***Create-Transfer-Recover-Calculate***" is to extract the stochastic retention information from a fast-simulation cell, namely "highly scaled STT-RAM cell" (retention time–microseconds), and convert such information into that of the slow-simulation cell–"desired STT-RAM cell" (retention time–milliseconds to seconds or more), with a set of semi-analytical models. Fig. 5 presents the detailed framework of our proposed "***Create-Transfer-Recover-Calculate***", which mainly includes four important steps:

1) Step1: ***Create*** a "highly scaled STT-RAM cell" library for the fast retention time calibration;
2) Step2: ***Transfer*** the statistical information from the library to any "desired STT-RAM cell", as well as the variations information for noise intensity if the variation-aware mode is trigged;
3) Step3: ***Recover*** the retention time distributions of the "desired STT-RAM cell";
4) Step4: ***Calculate*** the retention failure rate within a given time for the normal mode or variation-aware mode.

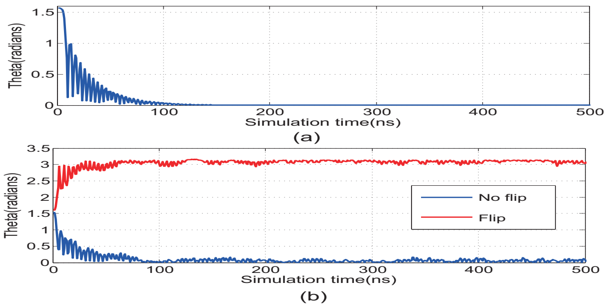


Fig. 4. Retention failure demonstration (a) without and (b) with thermal noise field.

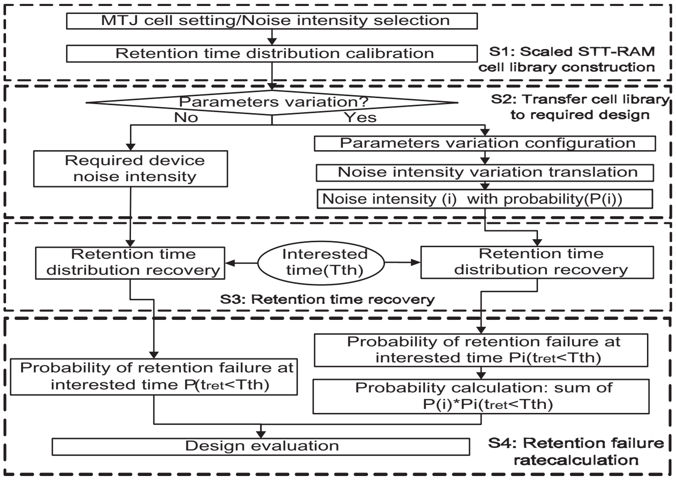

MTJ cell setting/Noise intensity selection
Retention time distribution calibration
S1: Scaled STT-RAM cell library construction
Parameters variation?
No
Yes
S2: Transfer cell library to required design
Parameters variation configuration
Required device noise intensity
Noise intensity variation translation
Noise intensity (i) with probability(P(i))
Retention time distribution recovery
Interested time(Tth)
Retention time distribution recovery
S3: Retention time recovery
Probability of retention failure at interested time Pi(tret<Tth)
Probability of retention failure at interested time P(tret<Tth)
Probability calculation: sum of P(i)*Pi(tret<Tth)
Design evaluation
S4: Retention failure ratecalculation

Fig. 5. Overview of our methodology.

The first step is to build a "highly scaled STT-RAM cell" library with much higher bit flipping probabilities by intentionally enhancing the noise intensity (h_l) of Langevin random field to significantly accelerate the simulation process. As Eq. (2) in section III indicates, any changes of the key parameters, can be successfully translated into a larger noise intensity, leading to a more prominent data retention failure within the same timing window. Consequently, the full statistical information and distribution of the retention time can be easily observed during the shortened simulated time. At the second step, the statistical information of a "desired STT-RAM cell", which usually has much longer retention time, will be estimated based on the "highly scaled STT-RAM cell" library from step one. Meanwhile, the variations from the input parameters, are transferred to the equivalent variation of noise intensity. At the third step, the retention time distribution of the "desired STT-RAM cell" is recovered. At last, the retention failure is calibrated based on the retention time distribution from the step three (normal mode).

As the following simulation and analysis focus on the statistical behavior of θ versus time based on the macro-magnetic model in section III, we make following parameter definitions and simulation settings: the retention time of a STT-RAM cell (t_{ret}) is defined as the time that a first bit flipping happens during data retention, i.e., from $\theta = 0$ to $|\theta| = \pi$ induced by Langevin random field. (μ, σ) are the mean and standard deviation of t_{ret}, respectively. The retention failure rate P_f denotes the probability that an idle cell flips odd times within a given time T_{th}. The parameters of MTJ cell in our simulations are also shown in Table I.

1) Create Distribution Model in Library: To minimize the overhead of the magnetic and Monte-Carlo simulations required to capture the statistical information of the retention time, our "highly scaled STT-RAM cell" library is derived from a set of noise-intensity enhanced cells, which cover 5 carefully selected levels of noise intensity (βh_l). Here, β is the noise level index normalized to the h_l, i.e., $\beta = 3.1, 3.2, 3.3, 3.4, 3.5$ for our MTJ cell model. For each selected level β, we perform extensive LLG macro-magnetic model-based Monte-carlo simulations for a given time T_{th} based on different β, i.e., running T_{th}=($10^4 \sim 5 * 10^6$)ns simulation for 10^5 times, to guarantee that the scaled cells can experience sufficient numbers of bit flippings at each T_{th}. The mean (μ) and standard deviation (σ) of the retention time t_{ret} can be accurately measured as long as the P_f is close to 1. Note that we explored different time steps here to make sure our simulation have guaranteed accuracy.

TABLE II
MAGNETIC SIMULATION RESULTS FOR SCALED CELLS

β	$T_{th}(ns)$	$P_f(T_{th})$	$\log \frac{-T_{th}}{\log(1-P_f)}$	$log(\mu)$
3.1	10^4	0.168	10.8834	N/A
	$5 * 10^4$	0.603	10.895	N/A
	10^5	0.857	10.847	N/A
	$2 * 10^6$	1	N/A	10.8733
3.5	10^4	0.874	8.472	N/A
	$2 * 10^4$	0.981	8.516	N/A
	$3 * 10^4$	0.996	8.491	N/A
	$4 * 10^4$	1	N/A	8.511

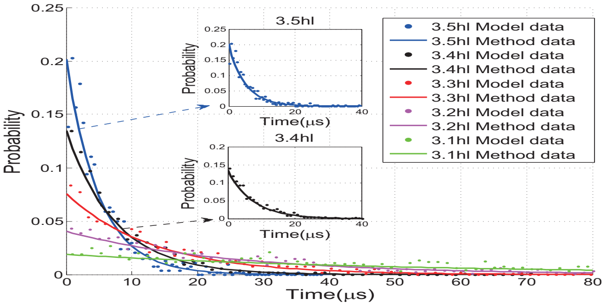


Fig. 6. Simulated retention time distribution from magnetic model v.s. exponential distribution from proposed method at different levels of noise intensity.

Table II shows the example magnetic simulation results of retention failure rate ($P_f(T_{th})$) w.r.t. different simulation time window (T_{th}) for the two "highly scaled STT-RAM cells" ($\beta = 3.1, 3.5$). The retention failure rate can be increased accordingly as the time elapses. We also found that the data in fourth column and last column can always satisfy the following equation at any given T_{th} for both cells:

$$\log \frac{-T_{th}}{\log(1-P_f)} = log(\mu) \tag{11}$$

Note that all the other scaled cells with different β also follow eq. (11). Consequently, the retention failure rate P_f for a given time T_{th} can be easily derived as:

$$P_f = P\left(t_{ret} \leq T_{th}\right) = 1 - e^{-T_{th}/\mu} \tag{12}$$

As expected, Eq. (12) also agrees well with the exponential function as predicted by the widely adopted Arrhenius-Neel theory for memory retention time analysis [1], [8].

Based on Eq. (12), we found that the exponential distribution may provide an excellent accuracy in modeling the probability distribution functions (PDFs) of t_{ret}. The employed exponential distribution can be expressed as:

$$P(t_{ret}) = \begin{cases} \frac{1}{\mu} e^{-\frac{1}{\mu} t_{ret}} & t_{ret} \geq 0, \\ 0 & t_{ret} < 0. \end{cases} \tag{13}$$

Here μ is the mean, as well as the standard deviation ($\sigma = \mu$) of the retention time.

Fig. 6 compares the PDFs of t_{ret} from the magnetic model and from the exponential distribution based on our method. The figure shows the proposed exponential distribution (as shown in Eq. (13)) achieve good accuracy at the simulated levels of noise intensity. Fig. 7 shows the relative ratio between the retention time's mean and standard deviation, which validates the correctness of our exponential distribution model.

2) Transfer Statistical Information: After aforementioned retention time distribution model is constructed, the next question becomes how to determine the statistical information of t_{ret} for any "desired STT-RAM cell", i.e. the actual value of the mean and standard deviations of t_{ret} for a long retention time cell, based on the information provided by the "highly scaled STT-RAM cell" library. As the noise intensity is a key parameter that links the MTJ cell and impacts the retention time, we propose to use a noise intensity to retention time transfer function (N-R function) to facilitate the associated information transfer.

Our further investigation suggests a quadratic function of the

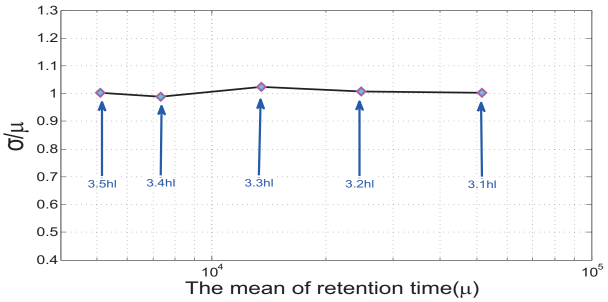


Fig. 7. The ratio between standard deviation and mean for retention time at different levels of noise intensity.

relationship between the noise level index β and logarithm of mean retention time $log(\mu)$, N-R function can be expressed by:

$$log(\mu) = a/\beta^2 + b \tag{14}$$

where a and b are device fitting parameters. Note that Eq.(1) and Eq. (2) suggest that the retention time can be approximately inversely proportional to the noise intensity: $\log(\mu) \propto 1/\beta^2$.

If the variability of the input parameters are to be considered, e.g. process variations of ΔV, we can easily calculate the distribution of h_l based on Eq. (2) in section III. As such, the histogram of the distorted h_l can be obtained for further variation-aware retention time analysis, i.e. $P_{h_l(i)}$ at each $h_l(i), i \le N$, here N is number of samples for h_l.

Fig. 8 shows the relationship between β and $log(\mu)$ based on our N-R function (method data in the figure). The results from our N-R function match the magnetic model results (the library data) very well. To verify the scalability of our N-R method, we also run extensive magnetic simulations at the noise index out of the constructed library, i.e. $\beta = 2.5, 4$ as the examples of the "desired STT-RAM cells". Our results (method data) can always align with the average value of $log(\mu)$ directly from costly magnetic model based monte-carlo simulations at enlarged or reduced noise index level. Note that mean retention time (μ) for the cells with a small noise intensity can be far exceeding those in the library, causing a significant simulation cost.

3) Recover Retention Time Distribution: The retention time distribution recovery for the "desired STT-RAM cell" can be easily conducted based on the predicted mean and the exponential distribution function. Note that only one PDF is required for retention failure calculation in a normal mode. However, multiple PDFs, corresponding to the number of samples of h_l in aforementioned section, are needed in a variation aware mode.

4) Calculate Retention Time Failure Rate: A retention failure happens if an expected bit flipping occurs within a specified time. Now the retention failure rate P_f of "desired STT-RAM cell" can be easily captured by combining Eq. (12) and Eq. (14) at any given time (T_{th}).

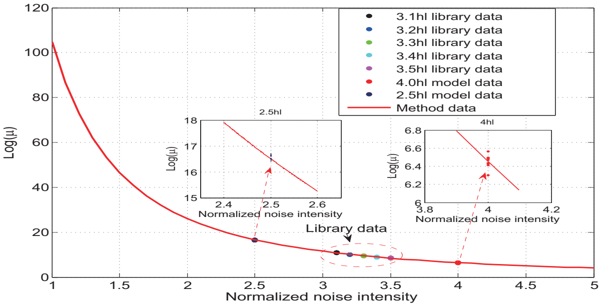


Fig. 8. The simulated relationship between β and $log(\mu)$ from magnetic model (library data) v.s. proposed N-R function (method).

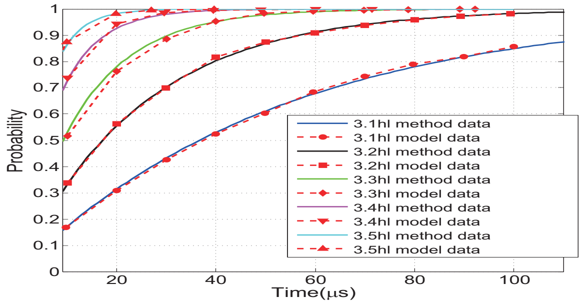


Fig. 9. CDF of the retention time from magnetic model v.s. the normal mode of proposed technique.

In the variation aware mode, the failure probability can be calculated by considering occurrence probability $P_{h_l(i)}$ of each $h_l(i)$, as well as the corresponding failure rate based on Eq. (12):

$$P_f{}^v(t) = P(t_{ret} \le T_{th}) = \sum_{i=1}^{N} P_{h_l(i)} * P_{f,h_l(i)}(T_{th}) \tag{15}$$

Here N is the number of samples for the distorted h_l.

Fig. 9 compare the cumulative distribution functions (CDFs) obtained from Eq. (12) based on our method and the magnetic monte-carlo simulations. As Fig. 9 show, our results (method data) can achieve very good accuracy at different levels of noise intensity and precisely describe the changing trend of the retention failure rate w.r.t. noise intensity. As the level of noise intensity increases, the retention time failure probability within a given time quickly increases, indicating a considerable average retention time reduction.

We also demonstrate the capability of our method in the presence of variations from any input parameters. As an example, we assume that both of MTJ area and thickness suffer from 2% variations. The noise intensity is assumed to be $3.2h_l$. Fig. 10 shows that the results from our method match well with the golden magnetic model if variations are existing. We also found that the results with variations are very close to those without variations, indicating that such small variations cannot change the statistical behavior of STT-RAM retention time. The main reason is that the distorted noise intensity in our simulation may either be increased or decreased by process variations, leading to the very limited CDF fluctuations.

B. Sensitivity Analysis of Initial Angle

As discussed in section IV-A4, the noise intensity h_l significantly impacts the retention time distribution. We now examine how the initial state of θ, namely initial angle θ_0 here, can affect the retention time by leveraging the power of our proposed "***Create-Transfer-Recover-Calculate***" technique. The initial state of φ is set as $\pi/2$ [12]. We adopt the same MTJ with the $3.5h_l$ noise intensity to accelerate the simulation process.

Fig. 11 depicts a dynamic changing of the θ's distribution w.r.t simulation time under a same initial state: $\theta_0 = 0.001$. As the

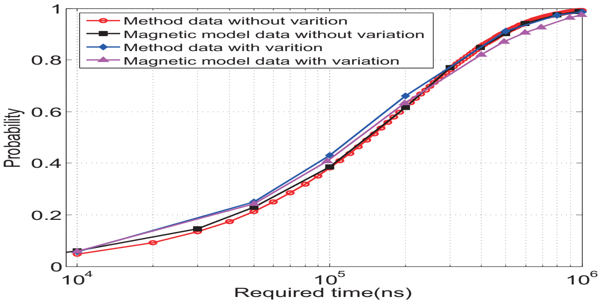


Fig. 10. CDF of the retention time from magnetic model v.s. the variation-aware mode of proposed technique.

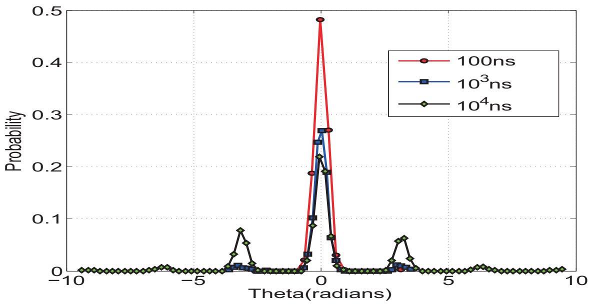


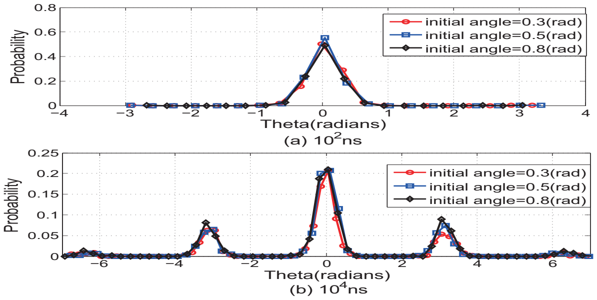
Fig. 11. The distributions of θ v.s. different simulation time for the same initial angle $\theta_0 = 0.001$.

simulation time increases, the occurrence probability of θ around 0 is reduced, while that of θ around $\pm\pi$ is increased significantly due to the noise intensity, showing a higher bit flipping rate. As the probability that θ can reach $\pm\pi$ is always existing at any time, a bit flipping of STT-RAM can happen instantly with no warning. Such a data retention failure mechanism is very different from that of DRAM and Flash, a gradual charge loss process.

We also conduct a set of simulations with three different initial angles of θ: $\theta_0 = 0.8, 0.5, 0.3$. As Fig. 12 shows, θ can always present the similar distributions at 10^2ns simulation time despite of the difference of the initial angle. As the simulation time grows to 10^4ns, the bit flipping rates among the three cases are all increased to a similar probability. As expected, our results clearly show that the initial angle does not impact the retention time distribution.

V. Architectural Indication

To exhibit the indication of our model on architecture designs, we show the retention error rate comparison between our model and the traditional method [9] in Table III. The retention error rate (P_t) v.s. a volume scaling factors (ΔV scaling) for a time period T_p derived from the traditional method is presented in the second column of Table III. As the noise intensity is proportional to the square of MTJ volume (see Eq. (2)), we further estimated the retention error rate (P_m) v.s. a equivalent noise intensity value (noise intensity) through our method. The simulations still follow the device parameter configurations in Table I. As Table III shows, the traditional method always produces a pessimistic result, i.e., P_t is always larger than P_m created by our model, indicating that previous works [9], [13] overestimate the retention error rate and adopt an unnecessarily high refresh frequency. Our observation is consistent with the conclusion from [2] that the actual retention time of MTJ device is longer than the value calculated by traditional method. The main reason is that during data retention mode, the STT-RAM's working mechanism is similar as the conventional magnetic random access memory (MRAM) with a special magnetic field (random thermal magnetic field) as the only excitation. As such, the associated energy barrier that a bit flipping needs to overcome will be much higher than that in the normal spin current driven mode. This observation also agrees well with the conclusion that the spin current based STT-RAM is easier to be written than the conventional MRAM. Consequently, our model is able to accurately estimate the STT-RAM retention error rate within a time frame of interest and lower the refresh frequency, so that the refresh energy can be reduced.



(a) 10^2ns

(b) 10^4ns

Fig. 12. The distribution of θ v.s. simulation time at different initial angle settings.

TABLE III
RETENTION ERROR RATE COMPARISON.

V scaling	P_t	**noise intensity**	P_m	T_p
$\Delta V/2^2$	0.1031	$2hl$	4×10^{-7}	$64ms$
$\Delta V/1.5^2$	5.5×10^{-5}	$1.5hl$	1.2×10^{-14}	$1s$
$\Delta V/3^2$	0.07	$3hl$	2.79×10^{-4}	$26.5us$

VI. Conclusion

STT-RAM greatly suffers from a unique data retention error introduced purely by the intrinsic thermal noise field. The accurate calibration of such a type of error is very difficult but crucial for advanced STT-RAM architectures at scaled technology nodes. In this paper, we propose a retention time acceleration analysis method, namely "***Create-Transfer-Recover-Calculate***", to capture the distribution and full statistical information of retention time at interested time. Our method can efficiently produce more precise results than previous solutions, providing a clear guidance for architectural designers. Moreover, our result shows that the initial angle has little impact on STT-RAM retention time. At last we reveal that the retention time estimated by previous works is too pessimistic.

VII. Acknowledgement

This work is supported in part by NSF under project CNS-1423137.

References

[1] Y. Cai, *et al.*, "Error Patterns in MLC NAND Flash Memory: Measurement, Characterization and Analysis," in *Date*, 2012.

[2] E. Chen, *et al.*, "Advances and Future Prospects of Spin-Transfer Torque Random Access Memory," *IEEE Transactions on Magnetics*, 46(6):1873–1878, June 2010.

[3] Everspin, "Everspin starts sampling 256Mb ST-MRAM chips, plans 1Gb chips by the end of 2016," http://goo.gl/0basVF.

[4] J. Janesky, *et al.*, "Device performance in a fully functional 800MHz DDR3 spin torque magnetic random access memory," in *IMW*, 2013.

[5] H. Naeimi, *et al.*, "Sttram Scaling And Retention Failure," *intel technology journal*, 17:54–75, 2013.

[6] A. Nigam, *et al.*, "Delivering on the Promise of Universal Memory for Spin-Transfer Torque RAM (STT-RAM)," in *ISLPED*, pages 121–126, IEEE/ACM, 2011.

[7] N. Rizzo, *et al.*, "A Fully Functional 64 Mb DDR3 ST-MRAM Built on 90 nm CMOS Technology," *IEEE Transactions on Magnetics*, 49(7), July 2013.

[8] N. D. Rizzo, *et al.*, "Thermally activated magnetization reversal in submicron," *APPLIED PHYSICS LETTERS*, 80:2335–2337, 2002.

[9] C. W. Smullen, *et al.*, "Relaxing non-volatility for fast and energy-efficient STT-RAM caches," in *HPCA*, 2011.

[10] C. W. Smullen, IV, *et al.*, "The STeTSiMS STT-RAM Simulation and Modeling System," in *ICCAD*, 2011.

[11] G. Sun, *et al.*, "A novel architecture of the 3D stacked MRAM L2 cache for CMPs," in *HPCA*, 2009.

[12] J. Z. Sun, "Spin-current interaction with a monodomain magnetic body: A model study," *The American Physical Society*, 62(1):570–578, 2000.

[13] Z. Sun, *et al.*, "Multi Retention Level STT-RAM Cache Designs with a Dynamic Refresh Scheme," in *MICRO*, pages 329–338, 2011.

[14] J. Wang, *et al.*, "Enabling High-performance LPDDRx-compatible MRAM," in *ISLPED*, 2014.

[15] P. Wang, *et al.*, "A Thermal and Process Variation Aware MTJ Switching Model and Its Applications in Soft Error Analysis," in *ICCAD*, pages 720–727, ACM, 2009.

[16] X. Wang, *et al.*, "Thermal Fluctuation Effects on Spin Torque Induced Switching: Mean and Variations," *JAP*, 103(3), Feb. 2008.

[17] W. Zhao, *et al.*, "Failure and reliability analysis of STT-MRAM," in *IMicroelectronics Reliability*, volume 52, pages 1848–852, 2012.