Statistical Reliability/Energy Characterization in STT-RAM Cell Designs

Wujie Wen, Yaojun Zhang, and Yiran Chen

1 Introduction

Conventional memory technologies, i.e., SRAM, DRAM, and Flash, have achieved remarkable successes in modern computer industry. Following technology scaling, the shrunk feature size and the increased process variations impose serious power and reliability concerns on these technologies.

In recent years, many emerging nonvolatile memory technologies have emerged above the horizon. As one promising candidate, spin-transfer torque random access memory (STT-RAM) has demonstrated great potentials in embedded memory and on-chip cache designs [1–6] through a good combination of the non-volatility of Flash, the comparable cell density to DRAM, and the nanosecond programming time like SRAM.

In STT-RAM, the data is represented as the resistance state of a magnetic tunneling junction (MTJ) device. The MTJ resistance state can be programmed by applying a switching current with different polarizations. Compared to the charge-based storage mechanism of conventional memories, the magnetic storage mechanism of STT-RAM shows less dependency on the device volume and hence, better scalability. Nonetheless, despite of these advantages, the unreliable write operation and high write energy are to be the major issues in STT-RAM designs. And these design metrics are significantly impacted by the prominent statistical factors of STT-RAM, including CMOS/MTJ device process variations under scaled technology and the probabilistic MTJ switching behaviors [7, 8]. In particular, the randomness of MTJ switching process incurred by the thermal fluctuations may generate the intermittent write failures of STT-RAM cells.

W. Wen • Y. Zhang • Y. Chen (✉)
Department of Electrical & Computer Engineering, University of Pittsburgh,
Pittsburgh, PA 15213, USA
e-mail: YIC52@pitt.edu

© Springer International Publishing Switzerland 2015

W. Zhao, G. Prenat (eds.), *Spintronics-based Computing*,
DOI 10.1007/978-3-319-15180-9_7

Many studies were performed to evaluate the impacts of process variations and thermal fluctuations on STT-RAM reliability [9–11]. The general evaluation method is as follows: First, Monte-Carlo SPICE simulations are run extensively to characterize the distribution of the MTJ switching current I during the STT-RAM write operations, by considering the device variations of both MTJ and MOS transistor; Then I samples are sent into the macro-magnetic model to obtain the MTJ switching time (τ_{th}) distributions under thermal fluctuations; Finally, the τ_{th} distributions of all I samples are merged to generate the overall MTJ switching performance distribution. A write failure happens when the applied write pulse width is smaller than the needed τ_{th}. Nonetheless, the costly Monte-Carlo runs and the dependency on the macro-magnetic and SPICE simulations incur huge computation complexity [12–15], limiting the application of such a simulation method at the early stage of STT-RAM design and optimization. Meanwhile, the modeling of write energy in STT-RAM was also studied extensively [16]. However, many such works only assume that the write energy of STT-RAM is deterministic and cannot successfully take into account its statistical characteristic induced by process variations and thermal fluctuations.

In this chapter, we propose "PS3-RAM"—a fast, portable and scalable statistical STT-RAM reliability/energy analysis method. PS3-RAM includes three integrated steps: (1) characterizing the MTJ switching current distribution under both MTJ and CMOS device variations; (2) recovering MTJ switching current samples from the characterized distributions in MTJ switching performance evaluation; and (3) performing the simulation on the thermal-induced MTJ switching variations based on the recovered MTJ switching current samples. By introducing the sensitivity analysis technique to capture the statistical characteristics of the MTJ switching, and dual-exponential model to efficiently and accurately recover the MTJ switching current samples for statistical STT-RAM thermal analysis, PS3-RAM can achieve multiple orders-of-magnitude ($> 10^5$) run time cost reduction with marginal accuracy degradation under any variation configurations when compared to SPICE-based Monte-Carlo simulations. Finally, we released PS3-RAM from SPICE and macro-magnetic modeling and simulations, and extended its application into the array-level reliability analysis and the design space exploration of STT-RAM.

The structure of this chapter is organized as the follows: Section 2 gives the preliminary of STT-RAM; Section 3 presents the details of PS3-RAM method; Section 4 presents the application of our PS3-RAM on cell and array level reliability analysis and design space exploration; Section 5 shows the deterministic/statistical write energy analysis based on our PS3-RAM; Section 6 discusses the computation complexity; The last section-Appendix gives the detailed theoretical model deduction and its numerical validation for sensitivity analysis.

2 Preliminary

2.1 STT-RAM Basics

Figure 1c shows the popular "one-transistor-one-MTJ (1T1J)" STT-RAM cell structure, which includes a MTJ and a NMOS transistor connected in series. In the MTJ, an oxide barrier layer (e.g., MgO) is sandwiched between two ferromagnetic layers. '0' and '1' are stored as the different resistances of the MTJ, respectively. When the magnetization directions of two ferromagnetic layers are parallel (anti-parallel), the MTJ is in its low (high) resistance state. Figure 1a, b show the low and the high MTJ resistance states, which are denoted by R_L and R_H, respectively. The MTJ switches from '0' to '1' when the switching current drives from reference layer to free layer, or from '1' to '0' when the switching current drives in the opposite.

2.2 Process Variations and Programming Uncertainty of STT-RAM

2.2.1 Process Variations-Persistent Errors

The current through the MTJ is affected by the process variations of both transistor and MTJ. For example, the driving ability of the NMOS transistor is subject to the variations of transistor channel length (L), width (W), and threshold voltage (V_{th}). The MTJ resistance variation also affects the NMOS transistor driving ability by changing its bias condition. The degraded MTJ switching current leads to a longer MTJ switching time and consequently, results in an incomplete MTJ switching before the write pulse ends. This kind of errors is referred to as "persistent" errors, which are mainly incurred by only device parametric variations. Persistent errors can be measured and repeated after the chip is fabricated.

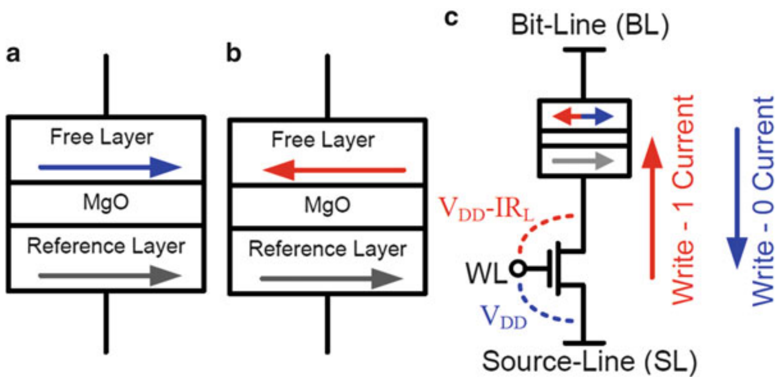

Fig. 1 STT-RAM basics. (**a**) Parallel (low resistance). (**b**) Anti-parallel (high resistance). (**c**) 1T1J cell structure

2.2.2 Thermal Fluctuation-Non-persistent Errors

Another kind of errors is called "non-persistent" errors, which happen intermittently and may not be repeated. The non-persistent errors of STT-RAM are mostly caused by the intrinsic thermal fluctuations during MTJ switching [17]. In general, the impact of thermal fluctuations can be modeled by the thermal induced random field h_{fluc} stochastic Landau-Lifshitz-Gilbert (LLG) equation (1) as [17]:

$$\frac{d\vec{m}}{dt} = -\vec{m} \times \left(\vec{h}_{eff} + \vec{h}_{fluc}\right) + \alpha\vec{m} \times \left(\vec{m} \times \left(\vec{h}_{eff} + \vec{h}_{fluc}\right)\right) + \frac{\vec{T}_{norm}}{M_s} \quad (1)$$

Where $\vec{m}$ is the normalized magnetization vector. Time t is normalized by γM_s; γ is the gyro-magnetic ratio and M_s is the magnetization saturation. $\vec{h}_{eff} = \frac{\vec{H}_{eff}}{M_s}$ is the normalized effective magnetic field. $\vec{h}_{fluc}$ is the normalized thermal agitation fluctuating field at finite temperature which represent the thermal fluctuation. α is the LLG damping parameter. $\vec{T}_{norm} = \frac{\vec{T}}{M_s V}$ is the spin torque term with units of magnetic field. And the net spin torque $\vec{T}$ can be obtained through microscopic quantum electronic spin transport model. Due to thermal fluctuations, the MTJ switching time will not be a constant value but rather a distribution even under a constant switching current.

3 PS3-RAM Method

Figure 2 depicts the overview of our proposed PS3-RAM method, mainly including the sensitivity analysis for MTJ switching current (I) characterization, the I sample recovery, and the statistical thermal analysis of STT-RAM. The first step is to configure the variation-aware cell library by inputting both the nominal design parameters and their corresponding variations, like the channel length/width/threshold voltage of NMOS transistor, as well as the thickness/area of MTJ device. Then a multi-dimension sensitivity analysis will be conducted to characterize the statistical properties of I, followed by an advanced filtering technology—smooth filter, to improve its accuracy. After that, the write current samples can be recovered based on the above characterized statistics and current distribution model. The write pulse distribution will be generated after mapping the switching current samples to the write pulse samples by considering the thermal fluctuations. Finally, the statistical write energy analysis and the STT-RAM cell write error rate can be performed based on the samples of the write current once the write pulse is determined. Array-level analysis and design optimizations can be also conducted by using PS3-RAM.

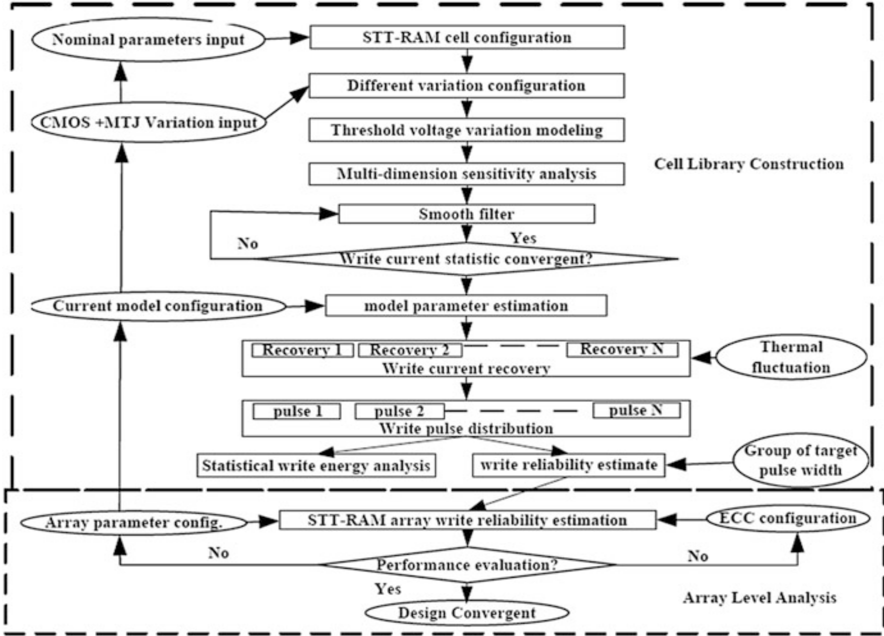

Fig. 2 Overview of PS3-RAM

3.1 *Sensitivity Analysis on MTJ Switching*

In this section, we present our sensitivity model used for the characterization of the MTJ switching current distribution. We then analyze the contributions of different variation sources to the distribution of the MTJ switching current in details. The definitions of the variables used in our analysis are summarized in Table 1.

3.1.1 Sensitivity Analysis on Variations

1) ***Threshold voltage variations:*** The variations of channel length, width and threshold voltage are three major factors causing the variations of transistor driving ability. V_{th} variation mainly comes from random dopant fluctuation (RDF) and line-edge roughness (LER), the latter of which is also the source of some geometry variations (i.e., L and W) [18, 19]. It is known that the V_{th} variation is also correlated with L and W and its variance decreases when the transistor size increases. The deviation of the V_{th} from the nominal value following the change of L (ΔL) can be modeled by [15]:

Table 1 Simulation parameters and environment setting

Parameters	Mean	Standard deviation
Channel length	$\overline{L} = 45\ \text{nm}$	$\sigma_L = 0.05\overline{L}$
Channel width	$\overline{W} = 90 \sim 1800\ \text{nm}$	$\sigma_W = 0.05\overline{L}$
Threshold voltage	$\overline{V}_{th} = 0.466\,\text{V}$	by calculation
MgO thickness	$\overline{T_{thick}} = 2.2\ \text{nm}$	$\sigma_{T_{thick}} = 0.02\overline{T_{thick}}$
MTJ surface area	$\overline{A} = 45 \times 90\ \text{nm}^2$	By calculation
Resistance low	$R_L = 1000\,\Omega$	By calculation
Resistance high	$R_H = 2000\,\Omega$	By calculation

$$\Delta V_{th} = \Delta V_{th0} + V_{ds}\ \exp\left(\frac{L}{l'}\right) \cdot \frac{\Delta L}{l'} \tag{2}$$

Then the standard deviation of V_{th} can be calculated as:

$$\sigma_{V_{th}}^2 = \frac{C_1}{WL} + \frac{C_2}{\exp\left(L/l'\right)} \cdot \frac{W_c}{W} \cdot \sigma_L^2 \tag{3}$$

Here W_c is the correlation length of non-rectangular gate (NRG) effect, which is caused by the randomness in sub-wavelength lithography. C_1, C_2 and l' are technology dependent coefficients. The first term in (3) describes the RDF's contribution to $\sigma_{V_{th}}$. The second term in (3) represents the contribution from NRG, which is heavily dependent on L and W. Following technology scaling, the contribution of this term becomes prominent due to the reduction of L and W.

2) ***Sensitivity analysis on variations:*** Although the contributions of MTJ and MOS transistor parametric variabilities to the MTJ switching current distribution cannot be explicitly expressed, it is still possible for us to conduct a sensitivity analysis to obtain the critical characteristics of the distribution. Without loss of generality, the MTJ switching current I can be modeled by a function of W, L, V_{th}, A and T_{thick}. A and T_{thick} are the MTJ surface area and MgO layer thickness, respectively. The 1st-order Taylor expansion of I around the mean values of every parameter is:

$$\begin{aligned} I(W, L, V_{th}, A, T_{thick}) \approx\ & I\left(\overline{W}, \overline{L}, \overline{V}_{th}, \overline{A}, \overline{T_{thick}}\right) + \frac{\partial I}{\partial W}\left(W - \overline{W}\right) + \frac{\partial I}{\partial L}\left(L - \overline{L}\right) \\ & + \frac{\partial I}{\partial V_{th}}\left(V_{th} - \overline{V}_{th}\right) + \frac{\partial I}{\partial A}\left(A - \overline{A}\right) + \frac{\partial I}{\partial T_{thick}}\left(T_{thick} - \overline{T_{thick}}\right) \end{aligned} \tag{4}$$

Here W, L and T_{thick} generally follow Gaussian distribution [9], A is the product of two independent Gaussian distributions, V_{th} is correlated with W and L, as shown in (2) and (3).

Because the MTJ resistance $R \propto \frac{e^{T_{thick}}}{A}$ [9], we have:

$$\frac{\partial I}{\partial A}\Delta A + \frac{\partial I}{\partial T_{thick}}\Delta T_{thick} = \frac{\partial I}{\partial R}\left(\frac{\partial R}{\partial A}\Delta A + \frac{\partial R}{\partial T_{thick}}\Delta T_{thick}\right) = \frac{\partial I}{\partial R}\Delta R \tag{5}$$

Equation (5) indicates that the combined contribution of A and T_{thick} is the same as the impact of MTJ resistance. The difference between the actual I and its mathematical expectation μ_I can be calculated by:

$$I(W, L, V_{th}, R) - E\left(I\left(\overline{W}, \overline{L}, \overline{V}_{th}, \overline{R}\right)\right) \approx \frac{\partial I}{\partial W}\Delta W + \frac{\partial I}{\partial L}\Delta L + \frac{\partial I}{\partial V_{th}}\Delta V_{th} + \frac{\partial I}{\partial R}\Delta R \tag{6}$$

Here we assume $\mu_I \approx E\left(I\left(\overline{W}, \overline{L}, \overline{V}_{th}, \overline{R}\right)\right) = I\left(\overline{W}, \overline{L}, \overline{V}_{th}, \overline{R}\right)$ and the mean of MTJ resistance $\overline{R} \approx R\left(\overline{A}, \overline{\tau}\right)$. Combining (2), (3), and (6), the standard deviation of I (σ_I) can be calculated as:

$$\begin{aligned}\sigma_I^2 = &\left(\frac{\partial I}{\partial W}\right)^2\sigma_W^2 + \left(\frac{\partial I}{\partial L}\right)^2\sigma_L^2 + \left(\frac{\partial I}{\partial R}\right)^2\sigma_R^2 \\ &+ \left(\frac{\partial I}{\partial V_{th}}\right)^2\left(\frac{C_1}{WL} + \frac{C_2}{\exp\left(L/l'\right)}\cdot\frac{W_c}{W}\cdot\sigma_L^2\right) + 2\frac{\partial I}{\partial L}\frac{\partial I}{\partial V_{th}}\rho_1\sqrt{\frac{C_1}{WL}}\sigma_L \\ &+ 2\frac{\partial I}{\partial W}\frac{\partial I}{\partial V_{th}}\rho_2\sqrt{\frac{C_1}{WL}}\sigma_W + 2\frac{\partial I}{\partial L}\frac{\partial I}{\partial V_{th}}V_{ds}\exp\left(-\frac{L}{l'}\right)\frac{\sigma_L^2}{l'}\end{aligned} \tag{7}$$

Here $\rho_1 = \frac{\mathrm{cov}(V_{th0}, \mathrm{L})}{\sqrt{\sigma_{V_{th0}}^2\sigma_L^2}}$ and $\rho_2 = \frac{\mathrm{cov}(V_{th0}, W)}{\sqrt{\sigma_{V_{th0}}^2\sigma_W^2}}$ are the correlation coefficients between V_{th0} and L or W, respectively [19]. $\sigma_{V_{th0}}^2 = \frac{C_1}{WL}$. Our further analysis shows that the last three terms at the right side of (7) are significantly smaller than other terms and can be safely ignored in the simulations of STT-RAM normal operations.

The accuracy of the coefficient in front of the variances of every parameter at the right side of (7) can be improved by applying window based smooth filtering. Take W as an example, we have:

$$\left(\frac{\partial I}{\partial W}\right)_i = \frac{I\left(\overline{W} + i\Delta W, L, V_{th}, R\right) - I\left(\overline{W} - i\Delta W, L, V_{th}, R\right)}{2i\Delta W} \tag{8}$$

where $i = 1, 2, \ldots K$. Different $\frac{\partial I}{\partial W}$ can be obtained at the different step i. K samples can be filtered out by a windows based smooth filter to balance the accuracy and the computation complexity as:

$$\overline{\frac{\partial I}{\partial W}} = \sum_{i=1}^{K}\omega_i\left(\frac{\partial I}{\partial W}\right)_i \tag{9}$$

Here ω_i is the weight of sample i, which is determined by the window type, i.e., Hamming window or Rectangular window [20].

3) ***Variation contribution analysis:*** The variations' contributions to I are mainly represented by the first four terms at the right side of (7) as:

$$
\begin{aligned}
S_1 &= \left(\frac{\partial I}{\partial W}\right)^2 \sigma_W^2, S_2 = \left(\frac{\partial I}{\partial L}\right)^2 \sigma_L^2, S_3 = \left(\frac{\partial I}{\partial R}\right)^2 \sigma_R^2 \\
S_4 &= \left(\frac{\partial I}{\partial V_{th}}\right)^2 \left(\frac{C_1}{WL} + \frac{C_2}{\exp(L/l')} \cdot \frac{W_c}{W} \cdot \sigma_L^2\right)
\end{aligned}
\tag{10}
$$

As pointed out by many prior-arts [21–24], an asymmetry exists in STT-RAM write operations: the switching time of '0' → '1' is longer than that of '1' → '0' and suffers from a larger variance. Also, the switching time variance of '0' → '1' is more sensitive to the transistor size changes than '1' → '0'. As we shall show later, this phenomena can be well explained by using our sensitivity analysis. To the best of our knowledge, this is the first time the asymmetric variations of STT-RAM write performance and their dependencies on the transistor size are explained and quantitatively analyzed.

As shown in Fig. 1, when writing '0', the word-line (WL) and bit-line (BL) are connected to V_{dd} while the source-line (SL) is connected to ground. $V_{gs} = V_{dd}$ and $V_{ds} = V_{dd} - IR$. The NMOS transistor is mainly working in triode region. Based on short-channel BSIM model, the MTJ switching current supplied by a NMOS transistor can be calculated by:

$$
I = \frac{\beta\left[(V_{dd} - V_{th})(V_{dd} - IR) - \frac{a}{2}(V_{dd} - IR)^2\right]}{1 + \frac{1}{v_{sat}L}(V_{dd} - IR)}
\tag{11}
$$

Here $\beta = \frac{\mu_0 C_{ox}}{1+U_0(V_{dd}-V_{th})}\frac{W}{L}$. U_0 is the vertical field mobility reduction coefficient, μ_0 is the electron mobility, C_{ox} is gate oxide capacitance per unit area, a is body-effect coefficient and v_{sat} is carrier velocity saturation. Based on short-channel PTM model [25] and BSIM model [26, 27], we derive $\left(\frac{\partial I}{\partial W}\right)^2$, $\left(\frac{\partial I}{\partial L}\right)^2$, $\left(\frac{\partial I}{\partial R}\right)^2$ and $\left(\frac{\partial I}{\partial V_{th}}\right)^2$ as:

$$
\begin{aligned}
&\left(\frac{\partial I}{\partial W}\right)_0^2 \approx \frac{1}{(A_1 W + B_1)^4}, \left(\frac{\partial I}{\partial L}\right)_0^2 \approx \frac{1}{\left(\frac{A_2}{W} + B_2 W + C\right)^4} \\
&\left(\frac{\partial I}{\partial R}\right)_0^2 \approx \frac{1}{\left(\frac{A_3}{W} + B_3\right)^4}, \left(\frac{\partial I}{\partial V_{th}}\right)_0^2 \approx \frac{1}{\left(\frac{A_4}{\sqrt{W}} + B_4\sqrt{W}\right)^4}
\end{aligned}
\tag{12}
$$

Our analytical deduction shows that the coefficients A_{1-4}, B_{1-4} and C are solely determined by W, L, V_{th} and R. The detailed expressions of coefficients

A_{1-4}, B_{1-4} and C can be found in the appendix. Here R is the high resistance state of the MTJ, or R_H. For a NMOS transistor at '0' → '1' switching, the MTJ switching current is:

$$I = \frac{\beta}{2a}\left[(V_{dd} - V_{th} - IR) - \frac{I}{WC_{ox}v_{sat}^2}\right]^2 \tag{13}$$

Here R is the low resistance state of the MTJ, or R_L. We have:

$$\begin{aligned}\left(\frac{\partial I}{\partial W}\right)_1^2 \approx \frac{1}{(A_5W + B_5)^4}, \left(\frac{\partial I}{\partial L}\right)_1^2 \approx \frac{1}{\left(\frac{A_6}{W} + B_6\right)^2} \\ \left(\frac{\partial I}{\partial R}\right)_1^2 \approx \frac{1}{\left(\frac{A_7}{W} + B_7\right)^4}, \left(\frac{\partial I}{\partial V_{th}}\right)_1^2 \approx \frac{1}{\left(\frac{A_8}{W} + B_8\right)^2}\end{aligned} \tag{14}$$

Again, A_{5-8} and B_{5-8} can be expressed as the function of W, L, V_{th} and R and the detailed expressions of those parameters can be found in the appendix in this chapter.

In general, a large S_i corresponds to a large contribution to I variation. When W is approaching infinity, only S_3 is nonzero at '1' → '0' switching while both S_2 and S_3 are nonzero at '0' → '1' switching. It indicates that the residual values of S_1–S_4 at '0' → '1' switching is larger than that at '1' → '0' switching when $W \to \infty$. In other words, '0' → '1' switching suffers from a larger MTJ switching current variation than '1' → '0' switching when NMOS transistor size is large.

4) ***Simulation results of sensitivity analysis:*** Sensitivity analysis [28] can be used to obtain the statistical parameters of MTJ switching current, i.e., the mean and the standard deviation, without running the costly SPICE and Monte-Carlo simulations. It can be also used to analyze the contributions of different variation sources to I variation in details. The normalized contributions (P_i) of variation resources, i.e., W, L, V_{th}, and R, are defined as:

$$P_i = \frac{S_i}{\sum_{i=1}^{4} S_i}, \; i = 1,2,3,4 \tag{15}$$

Figures 3 and 4 show the normalized contributions of every variation source at '0' → '1' and '1' → '0' switching's, respectively, at different transistor sizes. We can see that L and V_{th} are the first two major contributors to I variation at both switching directions when W is small. At '1' → '0' switching, the contribution of L raises until reaching its maximum value when W increases, and then quickly decreases when W further increases. At '0' → '1' switching, however, the contribution of L monotonically decreases, but keeps being the dominant factor over the simulated W range. At both switching directions, the contributions of R ramps up when W increases. At '1' → '0' switching, the normalized contribution of R becomes almost 100 % when W is really large.

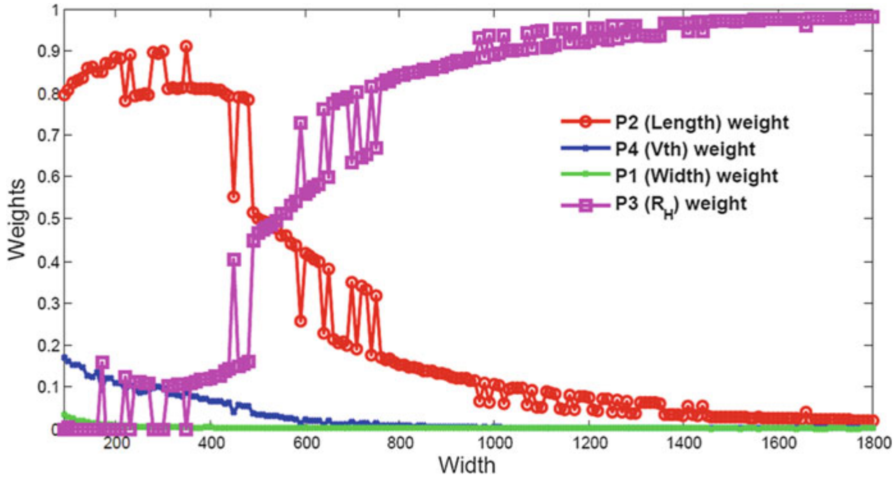

Fig. 3 The normalized contributions under different W at '1' → '0' switching

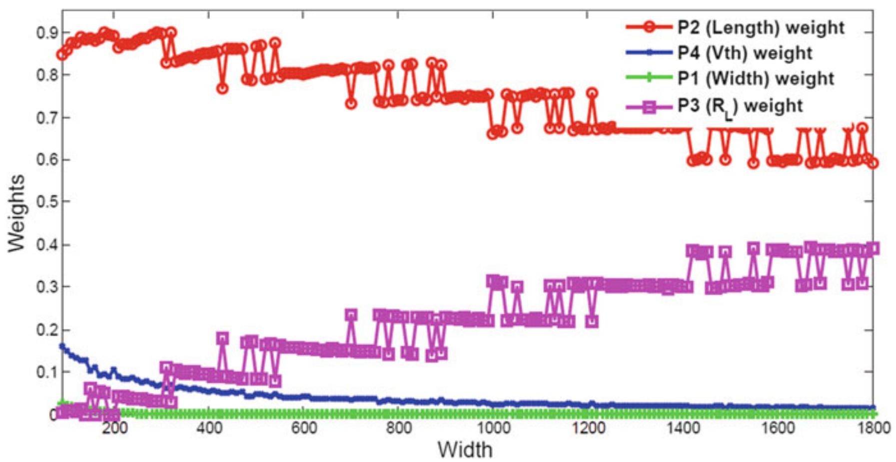

Fig. 4 The normalized contributions under different W at '0' → '1' switching

3.2 Write Current Distribution Recovery

After the I distribution is characterized by the sensitivity analysis, the next question becomes how to recover the distribution of I from the characterized information in the statistical analysis of STT-RAM reliability. We investigated the typical distributions of I in various STT-RAM cell designs and found that dual-exponential function can provide the excellent accuracy in modeling and recovering these

distributions. The dual-exponential function we used to recover the I distributions can be illustrated as:

$$f(I) = \begin{cases} a_1 e^{b_1(I-\mu)} & I \le \mu \\ a_2 e^{b_2(\mu-I)} & I > \mu \end{cases} \tag{16}$$

Here a_1, b_1, a_2, b_2 and μ are the fitting parameters, which can be calculated by matching the first and the second order momentums of the actual I distribution and the dual-exponential function as:

$$\begin{aligned} &\int f(I)dI = 1, \\ &\int If(I)dI = E(I), \\ &\int I^2 f(I)dI = E(I)^2 + \sigma_I^2 \end{aligned} \tag{17}$$

Here $E(I)$ and σ_I^2 are obtained from the sensitivity analysis.

The recovered I distribution can be used to generate the MTJ switching current samples, as shown in Fig. 5. At the beginning of the sample generation flow, the confidence interval for STT-RAM design is determined, e.g., $[\mu_I - 6\sigma_I, \mu_I + 6\sigma_I]$ for a six-sigma confidence interval. Assuming we need to generate N samples within the confidence interval, say, at the point of $I = I_i$, a switching current sequence of $[N\Pr_i]$ samples must be generated. Here $\Pr_i \approx f(I_i)\Delta$, Δ equals $\frac{12\sigma_I}{N}$, or the step of sampling generation. $f(I_i)$ is the dual-exponential function. Note that N determines both the analysis granularity and the level of the estimated error rate.

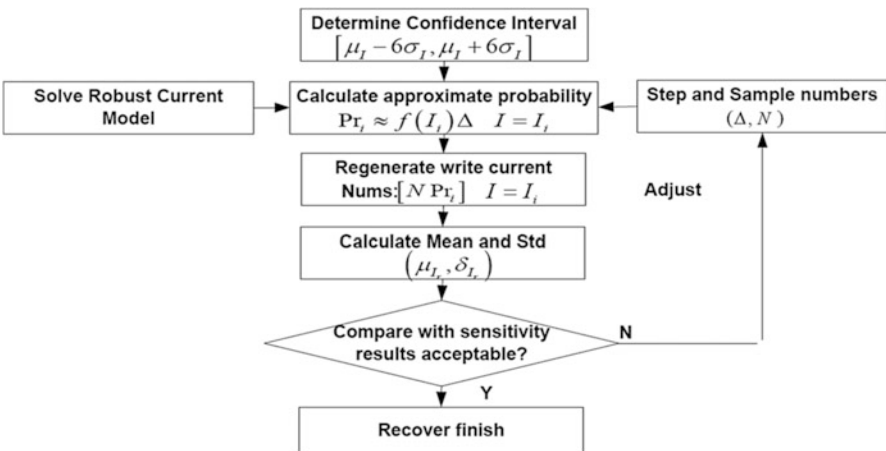

Fig. 5 Basic flow for MTJ switching current recovery

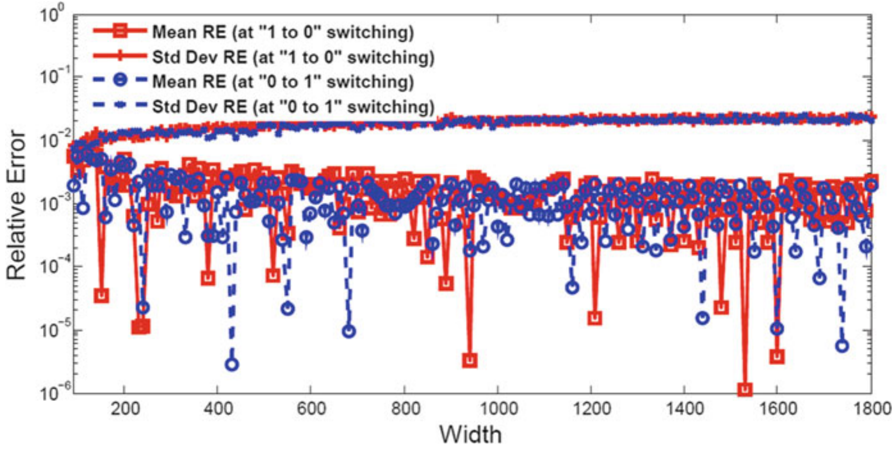

Fig. 6 Relative errors of the recovered I w.r.t. the results from sensitivity analysis

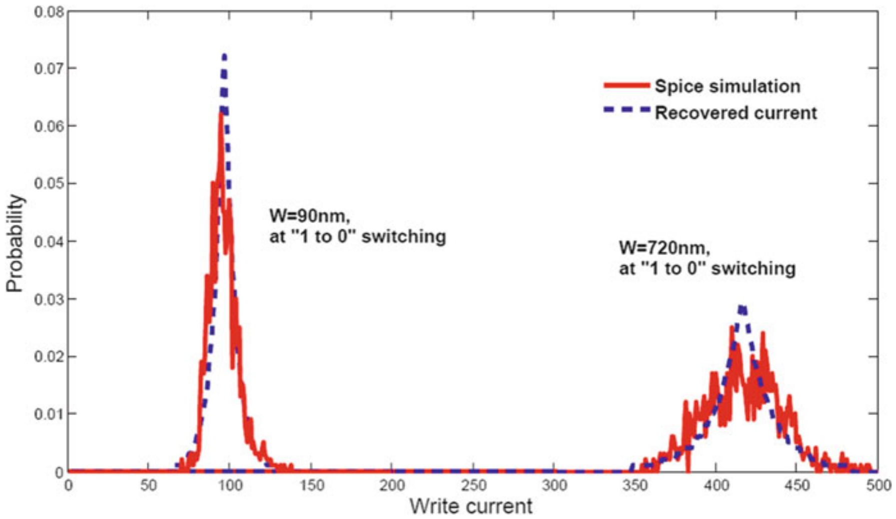

Fig. 7 Recovered I vs. Monte-Carlo result at '1' → '0'

Figure 6 shows the relative errors of the mean and the standard deviation of the recovered I distribution w.r.t. the results directly from the sensitivity analysis (see (6) and (7)). The maximum relative error $<10^{-2}$, which proves the accuracy of our dual-exponential model.

Figures 7 and 8 compare the probability distribution functions (PDF's) of I from the SPICE Monte-Carlo simulations and from the recovery process based on our sensitivity analysis at two switching directions. Our method achieves good accuracy at both representative transistor channel widths ($W = 720$ nm or $W = 720$ nm).

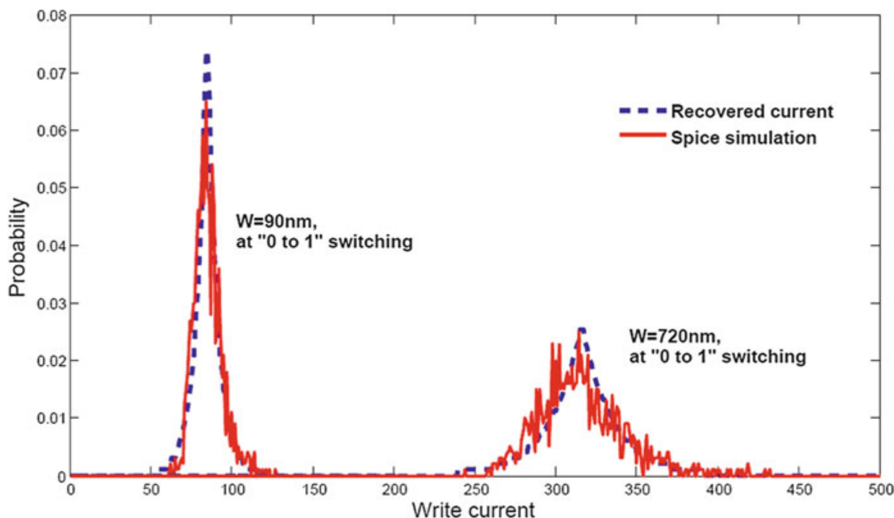

Fig. 8 Recovered I vs. Monte-Carlo result at '0' → '1'

3.3 Statistical Thermal Analysis

The variation of the MTJ switching time (τ_{th}) incurred by the thermal fluctuations follows Gaussian distribution when τ_{th} is below $10 \sim 20$ ns [21]. In this range, the distribution of τ_{th} can be easily constructed after the I is determined. The distribution of MTJ switching performance can be obtained by combining the τ_{th} distributions of all I samples.

4 Application 1: Write Reliability Analysis

In this section, we conduct the statistical analysis on the write reliability of STT-RAM cells by leveraging our PS3-RAM method. Both device variations and thermal fluctuations are considered in the analysis. We also extend our method into array-level evaluation and demonstrate its effectiveness in STT-RAM design optimizations.

4.1 Reliability Analysis of STT-RAM Cells

The write failure rate P_{WF} of a STT-RAM cell can be defined as the probability that the actual MTJ switching time τ_{th} is longer than the write pulse width T_w, or $P_{WF} = P(\tau_{th} > T_w)$. τ_{th} is affected by the MTJ switching current magnitude,

the MTJ and MOS device variations, the MTJ switching direction, and the thermal fluctuations. The conventional simulation of P_{WF} requires costly Monte-Carlo runs with hybrid SPICE and macro-magnetic modeling steps. Instead, we can use PS3-RAM to analyze the statistical STT-RAM write performance. The corresponding simulation environment is also summarized in Table 1.

Figures 9 and 10 depict the P_{WF}'s simulated by PS3-RAM for both switching directions at 300 K. For comparison purpose, the Monte-Carlo simulation results

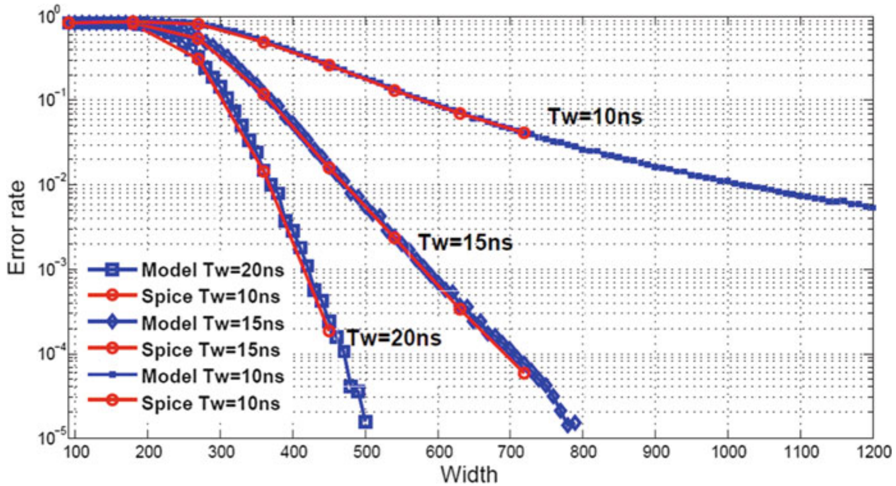

Fig. 9 Write failure rate at '0' → '1' when T = 300 K

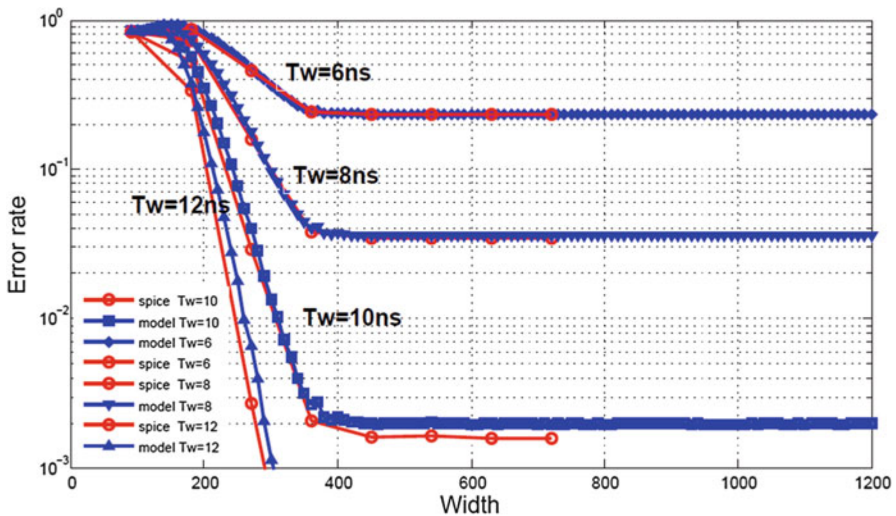

Fig. 10 Write failure rate at '1' → '0' when T = 300 K

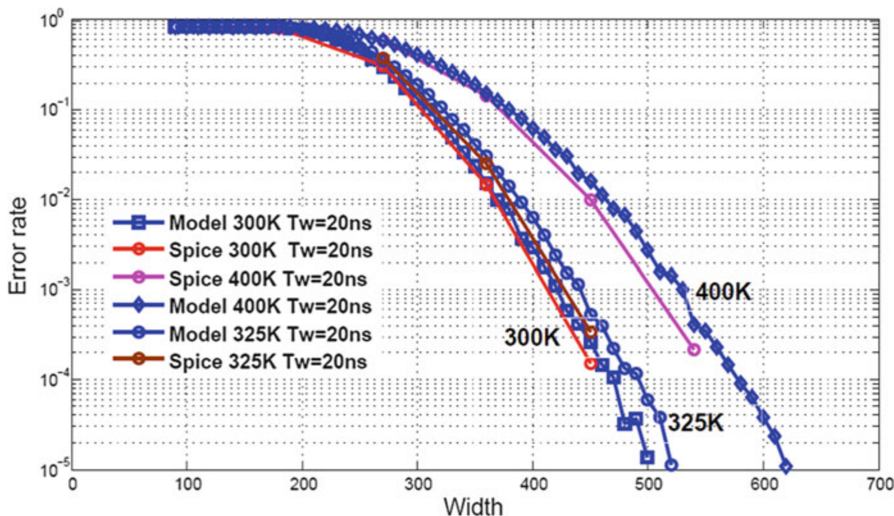

Fig. 11 P_{WF} under different temperatures at '0' → '1'

are also presented. Different T_w's are selected at either switching directions due to the asymmetric MTJ switching performances [21], i.e., $T_w = 10, 15, 20$ ns at '0' → '1' and $T_w = 6, 8, 10, 12$ ns at '1' → '0'. Our PS3-RAM results are in excellent agreement with the ones from Monte-Carlo simulations.

Since '0' → '1' is the limiting switching direction for STT-RAM reliability, we also compare the P_{WF}'s of different STTRAM cell designs under different temperatures at this switching direction in Fig. 11. The results show that PS3-RAM can provide very close but pessimistic results compared to those of the conventional simulations. PS3-RAM is also capable to precisely capture the small error rate change incurred by a moderate temperature shift (from T = 300 to 325 K).

It is known that prolonging the write pulse width and increasing the MTJ switching current (by sizing up the NMOS transistor) can reduce the P_{WF}. In Fig. 12, we demonstrate an example of using PS3-RAM to explore the STT-RAM design space: the tradeoff curves between P_{WF} and T_W are simulated at different W's. For a given P_{WF}, for example, the corresponding tradeoff between W and T_W can be easily identified on Fig. 12.

4.2 Array Level Analysis and Design Optimization

We use a 45 nm 256 Mb STT-RAM design [29] as the example to demonstrate how to extend our PS3-RAM into array-level analysis and design optimizations. The number of bits per memory block $N_{bit} = 256$ and the number of memory blocks $N_{word} = 1\text{M}$. To repair the operation errors of memory cells, circuit-level technique-ECC (error correction code) is usually applied [30]. Two types of ECC's with

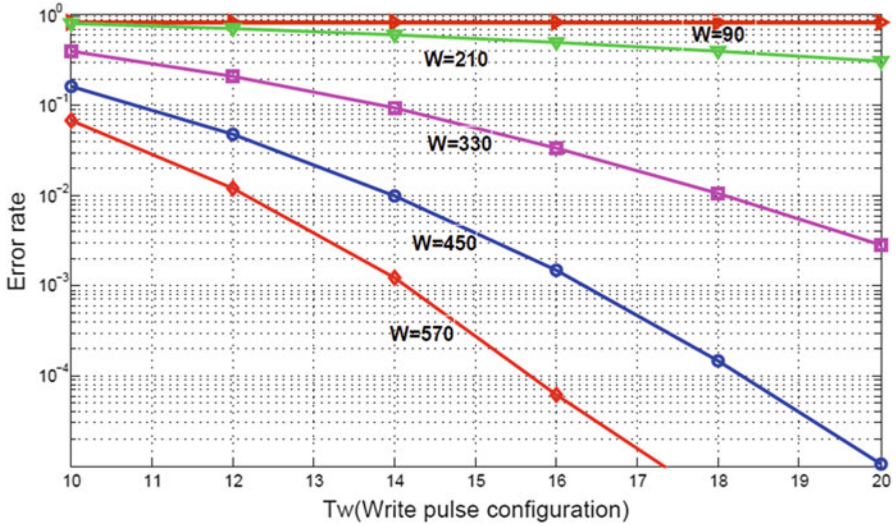

Fig. 12 STT-RAM design space exploration at '0' → '1'

different implementation costs are being considered, i.e., single-bit-correcting Hamming code and a set of multi-bits-correcting BCH codes. We use (n, k, t) to denote an ECC with n codeword length, k bit user bits being protected (256 bit here) and t bits being corrected. The ECC's corresponding to the error correction capability t from 1 to 5 are Hamming code (265; 256; 1) and four BCH codes–BCH1 (274; 256; 2), BCH2 (283; 256; 3), BCH3 (292; 256; 4) and BCH4 (301; 256; 5), respectively. The write yield of the memory array Y_{wr} can be defined as:

$$Y_{wr} = P(n_e \le t) = \sum_{i=0}^{t} C_n^i P_{WF}^i (1 - P_{WF})^{n-i} \tag{18}$$

Here, n_e denotes the total number of error bits in a write access. Y_{wr} indeed denotes the probability that the number of error bits in a write access is smaller than the error correction capability.

Figure 13 depicts the Y_{wr}'s under different combinations of ECC scheme and W when $T_W = 15$ ns at '0' → '1' switching. The ECC schemes required to satisfy $\sim 100\%$ Y_{wr} for different W are: (1) Hamming code for $W = 630$ nm; (2) BCH2 for $W = 540$ nm; and (3) BCH4 for $W = 480$ nm. The total memory array area can be estimated by using the STT-RAM cell size equation $\text{Area}_{\text{cell}} = 3(W/L + 1)(F^2)$ [31]. Calculation shows that combination (3) offers us the smallest STT-RAM array area, which is only 88 % and 95 % of the ones of (1) and (2), respectively. We note that PS3-RAM can be seamlessly embedded into the existing deterministic memory macro models [31] for the extended capability on the statistical reliability analysis and the multi-dimensional design optimizations on area, yield, performance and energy.

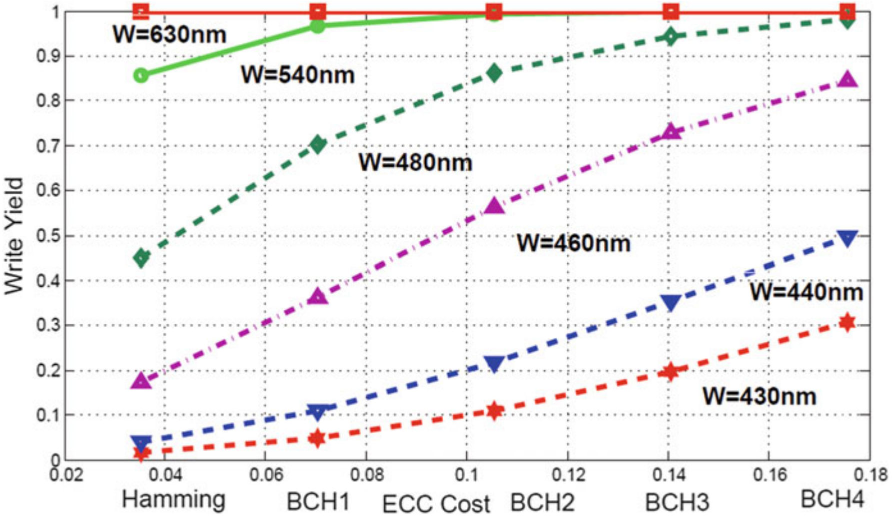

Fig. 13 Write yield with ECC's at '0' → '1', Tw = 15 ns

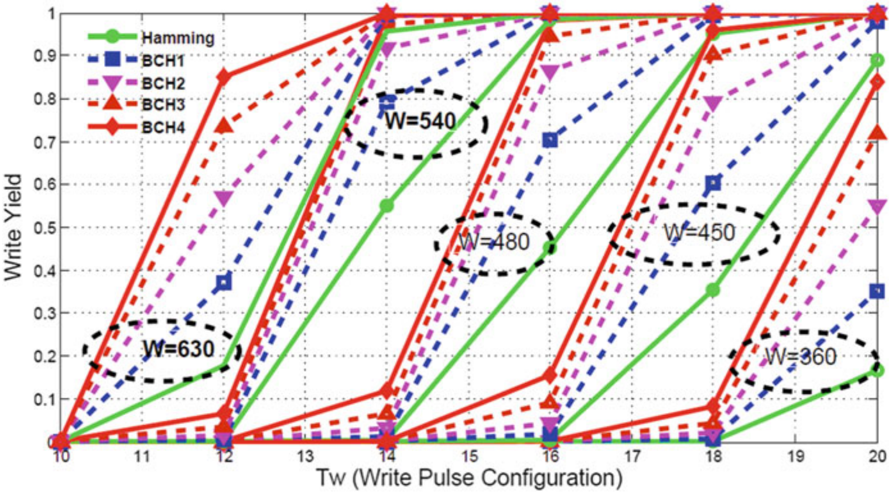

Fig. 14 Design space exploration at '0' → '1'

Figure 14 illustrates the STT-RAM design space in terms of the combinations of Y_{wr}, W, T_w and ECC scheme. After the pair of (Y_{wr}, T_w) is determined, the tradeoff between W and ECC can be found in the corresponding region on the figure. The result shows that PS3-RAM provides a fast and efficient method to perform the device/circuit/architecture co-optimization for STT-RAM designs.

5 Application 2: Write Energy Analysis

In addition to write reliability analysis, our PS3-RAM method can also precisely capture the write energy distributions influenced by the variations of device and working environment. In this section, we first prove that there is a sweet point of write pulse width for the minimum write energy without considering any variations. Then we introduce the concept of statistical write energy of STT-RAM cells considering both process variations and thermal fluctuations, and perform the statistical analysis on write energy using our PS3-RAM method.

5.1 Write Energy Without Variations

The write energy of a STT-RAM cell during each programming cycle without considering process and thermal variations is deterministic and can be modeled by (19) as:

$$E_{av} = I^2 R \tau_{th} \tag{19}$$

Here I denotes the switching current at either '0' $\rightarrow$ '1' or '1' $\rightarrow$ '0' switching, τ_{th} is the corresponding MTJ switching time and R is the MTJ resistance value, i.e., R_L (R_h) for '0' $\rightarrow$ '1'('1' $\rightarrow$ '0') switching. As discussed in prior art [21], the switching process of an STT-RAM cell can be divided into three working regions:

$$I = \begin{cases} I_{C_0}\left(1 - \dfrac{\ln(\tau_{th}/\tau_0)}{\Delta}\right), & \tau_{th} > 10 \text{ ns} \\ I_{C_0} + C\,\ln\left(\dfrac{\pi}{2\theta}\right)/\tau_{th}, & \tau_{th} < 3 \text{ ns} \\ \dfrac{P}{\tau_{th}} + Q, & 3 \leq \tau_{th} \leq 10 \text{ ns} \end{cases} \tag{20}$$

Here I_{C_0} is the critical switching current, Δ is thermal stability, $\tau_0 = 1$ ns is the relax time, θ is the initial angle between the magnetization vector and the easy axis, C, P, Q are fitting parameters.

For a relatively long switching time range ($\tau_{th} \approx 10 \sim 300$ ns), the undistorted write energy P_{av} can be calculated as:

$$E_{av} = I_{C_0}^2\left(1 - \frac{\ln(\tau_{th})}{\Delta}\right)^2 R\tau_{th} = \frac{I_{C_0}^2}{\Delta^2}(\Delta - \ln(\tau_{th}))^2 \tau_{th} \tag{21}$$

In the long switching time range, we have $\ln(\tau_{th}) < 0$. Thus, $(\Delta - \ln(\tau_{th}))^2$ or E_{av} monotonically raises as the write pulse τ_{th} increases and the minimized write energy E_{av} occurs at $\tau_{th} = 10$ ns.

In the ultra-short switching time range ($\tau_{th} < 3$ ns), E_{av} can be obtained as:

$$\begin{aligned} E_{av} &= \left(I_{C_0} + C\ln\left(\frac{\pi}{2\theta}\right)/\tau_{th}\right)^2 R\tau_{th} \\ &= 2I_{C_0}RC\ln\left(\frac{\pi}{2\theta}\right) + I_{C_0}^2 R\tau_{th} + \frac{C^2\ln^2(\pi/2\theta)R}{\tau_{th}} \\ &\geq 2I_{C_0}RC\ln\left(\frac{\pi}{2\theta}\right) + 2\sqrt{I_{C_0}^2 R^2 C^2 \ln^2(\pi/2\theta)} \\ &\geq 4I_{C_0}RC\ln\left(\frac{\pi}{2\theta}\right) \end{aligned} \tag{22}$$

As (22) shows, the minimum of E_{av} can be achieved when $\tau_{th} = C\ln\left(\frac{\pi}{2\theta}\right)/I_{C_0}$. However, for the ultra-short switching time range (usually $C\ln\left(\frac{\pi}{2\theta}\right)/I_{C_0} > 3\,\text{ns}$), E_{av} monotonically decreases as τ_{th} increases.

Similarly, in the middle switching time range ($3 \leq \tau_{th} \leq 10$ ns), E_{av} can be expressed as:

$$E_{av} = \left(\frac{P}{\tau_{th}} + Q\right)^2 R\tau_{th} = \left(\frac{P}{\sqrt{\tau_{th}}} + Q\sqrt{\tau_{th}}\right)^2 R \geq 4PQR \tag{23}$$

Again, the minimized E_{av} occurs at $\tau_{th} = \frac{P}{Q}$. Here $\frac{P}{Q} \geq 10$ ns based on our device parameters characterization [21]. Thus, the write energy E_{av} in this range monotonically decreases as τ_{th} grows.

According to the monotonicity of E_{av} in the three regions, the most energy-efficient switching point of E_{av} should be at $\tau_{th} = 10$ ns. To validate above theoretical deduction for the sweet point of E_{av}, the SPICE simulations are also conducted. Here the STT-RAM device model without considering process and thermal variations is also adopted from [21].

Figure 15 shows the simulated write energy E_{av} over different write pulse at '0' → '1' switching. As Fig. 15 shows, E_{av} monotonically decreases in the ultra-short switching range and continues decreasing in the middle range, but becomes monotonically increasing after entering the long switching time range. The sweet point of E_{av} occurs around $\tau_{th} = 10$ ns, which validates our theoretical analysis for the write energy without considering any variations.

We also present the simulated $E_{av} - \tau_{th}$ curve under different temperatures in Fig. 16. The trend and sweet point of $E_{av} - \tau_{th}$ curves remain almost the same when the temperature increases from T = 300 to 400 K. In fact, the write energy E_{av} decreases a little bit as the temperature increases. The reason is that the driving ability loss of the NMOS transistor (I) dominates E_{av} though the MTJ switching time (τ_{th}) increases when the working temperature raises.

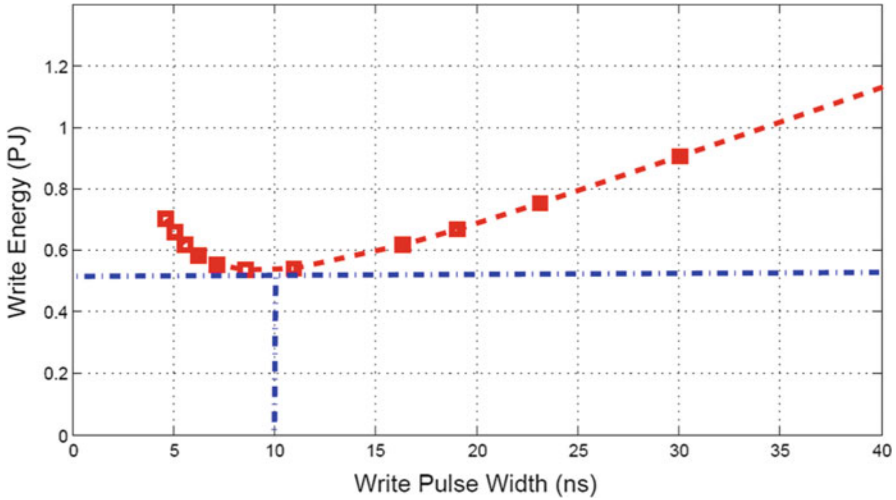

Fig. 15 Average Write Energy under different write pulse width when T = 300 K

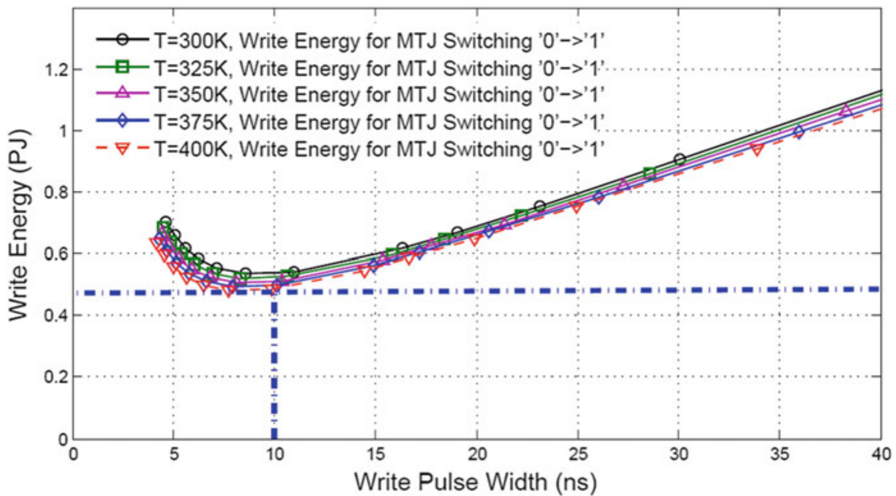

Fig. 16 Average Write Energy vs. write pulse width under different temperature

5.2 PS3-RAM for Statistical Write Energy

As discussed in previous section, the write energy of a STT-RAM cell can be deterministically optimized when all the variations are ignored. However, since the switching current I, the resistance R, and the switching time τ_{th} in (19) may be distorted by CMOS/MTJ process variations and thermal fluctuations, the

deterministic value will no longer be able to represent the statistic nature of the write energy of a STT-RAM cell. Accordingly, the optimized write energy at sweet point ($\tau_{th} = 10$ ns) shown in Fig. 15 should be expanded as a distribution.

Similar to the write failure analysis, we conduct the statistical write energy analysis using our PS3-RAM method. We choose the mean of NMOS transistor width $W = 540$ nm. The remained device parameters and variation configurations keep the same as Table 1.

Figures 17 and 18 show the simulated statistical write energy by PS3-RAM for both switching directions at 300 K. For comparison, the SPICE simulation results

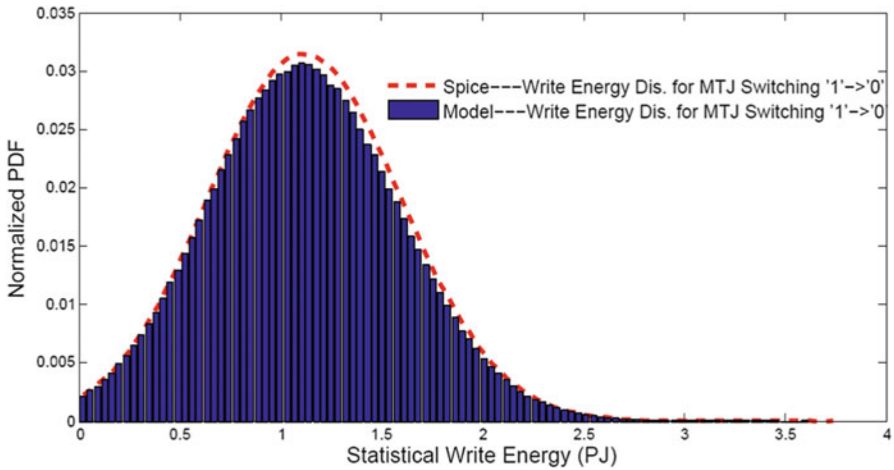

Fig. 17 Statistical Write Energy vs. write pulse width at '1' → '0'

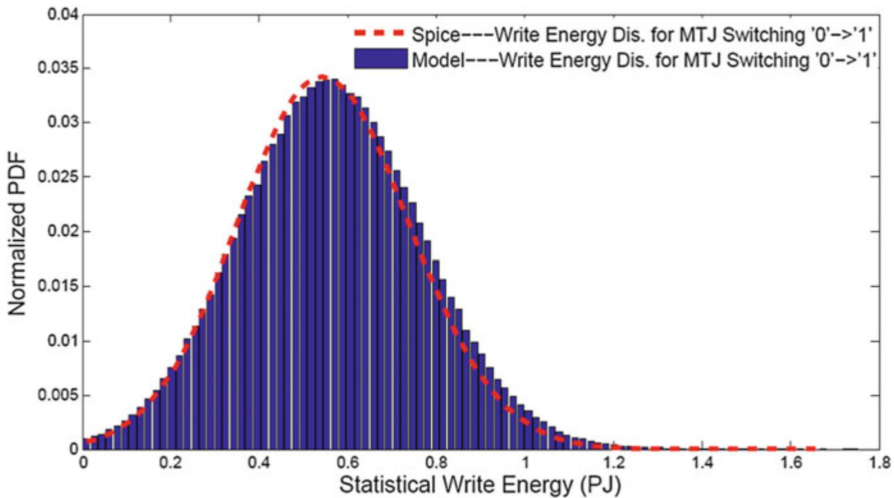

Fig. 18 Statistical Write Energy vs. write pulse width at '0' → '1'

are also presented. As shown in the figures, the distribution of write energy captured by our PS3-RAM method are in excellent agreement with the results from SPICE simulations at both ‘1’ $\rightarrow$ ‘0’ and ‘0’ $\rightarrow$ ‘1’ switching’s.

6 Computation Complexity Evaluation

We compared the computation complexity of our proposed PS3-RAM method with the conventional simulation method. Suppose the number of variation sources is M, for a statistical analysis of a STT-RAM cell design, the numbers of SPICE simulations required by conventional flow and PS3-RAM are $N_{std} = N_s^M$ and $N_{\text{PS3-RAM}} = 2KM + 1$, respectively. Here K denotes the sample numbers for window based smooth filter in sensitivity analysis, N_s is average sample number of every variation in the Monte-Carlo simulations in conventional method, $K \ll N_s$. Note that our switching current sample recovery flow does not require any extra Monte-Carlo simulations. The speedup $X_{\text{speedup}} \approx \frac{N_s^M}{2KM}$ can be up to multiple orders of magnitude: for example, if we set $N_s = 100$, $M = 4$, (note: V_{th} is not an independent variable) and $K = 50$, the speed up is around 2.5×10^5.

7 Conclusion

A fast and scalable statistical STT-RAM reliability/energy analysis method called PS3-RAM was developed in this chapter. PS3-RAM can simulate the impact of process variations and thermal fluctuations on the statistical STT-RAM write performance or write energy distributions, without running costly Monte-Carlo simulations on SPICE and macro-magnetic models. Simulation results show that PS3-RAM can achieve very high accuracy compared to the conventional simulation method, while achieving a speedup of multiple orders of magnitude. The great potentials of PS3-RAM in the application of the device/circuit/architecture co-optimization of STT-RAM designs are also demonstrated.

Appendix

In this appendix, the details on the model deduction in sensitivity analysis and the summary of the analytic results involved in the PS3-RAM development are given. Meanwhile, the validation of the analytic results based on Monte-Carlo simulations is also presented. Table 2 [26] summarizes some additional parameters used in this Appendix.

Sensitivity Analysis Model Deduction

The sensitivity analysis model is developed based on the electrical MTJ model and the simplified BSIM model [26, 27]. At '1' → '0' switching, the MTJ switching current supplied by an NMOS transistor working in the triode region is:

$$I = \frac{\beta\left[(V_{dd} - V_{th})(V_{dd} - IR) - \frac{a}{2}(V_{dd} - IR)^2\right]}{1 + \frac{1}{v_{sat}L}(V_{dd} - IR)} \tag{24}$$

Here $\beta = \frac{\mu_0 C_{ox}}{1 + U_0(V_{dd} - V_{th})}\frac{W}{L}$. U_0 is the vertical field mobility reduction coefficient, μ_0 is the electron mobility, C_{ox} is gate oxide capacitance per unit area, a is body-effect coefficient and v_{sat} is carrier velocity saturation. The MTJ is in its high resistance state, or $R = R_H$. Based on PTM [25] and BSIM [26], the partial derivatives in (6) can be calculated by ignoring the minor terms in the expansion of (24) as:

$$\begin{aligned}
&\left(\frac{\partial I}{\partial W}\right)_0^2 \approx \frac{1}{(A_1 W + B_1)^4}, \left(\frac{\partial I}{\partial L}\right)_0^2 \approx \frac{1}{\left(\frac{A_2}{W} + B_2 W + C\right)^4} \\
&\left(\frac{\partial I}{\partial R}\right)_0^2 \approx \frac{1}{\left(\frac{A_3}{W} + B_3\right)^4}, \left(\frac{\partial I}{\partial V_{th}}\right)_0^2 \approx \frac{1}{\left(\frac{A_4}{\sqrt{W}} + B_4\sqrt{W}\right)^4}
\end{aligned} \tag{25}$$

Table 2 Parameter definition

Variable	Definition
U_0	Vertical field mobility reduction coefficient
μ_0	Electron mobility
C_{ox}	Gate oxide capacitance per unit area
a	Body-effect coefficient
v_{sat}	Carrier velocity saturation

Here,

$$A_1 = \sqrt{\frac{\mu_0 C_{ox} V_{dd}(V_{dd} - V_{th})}{L}} R,$$

$$B_1 = \sqrt{\frac{L}{\mu_0 C_{ox} V_{dd}(V_{dd} - V_{th})}},$$

$$A_2 = \frac{L^2}{\mu_0 C_{ox} V_{dd}(V_{dd} - V_{th})}$$

$$B_2 = R^2 \mu_0 C_{ox} \frac{V_{dd} - V_{th}}{V_{dd}},$$

$$A_3 = \frac{L}{\mu_0 C_{ox} \sqrt{V_{dd}}(V_{dd} - V_{th})},$$

$$B_3 = \frac{R}{\sqrt{V_{dd}}}, C = \frac{2LR}{V_{dd}},$$

$$A_4 = \sqrt{\frac{L}{\mu_0 C_{ox} V_{dd}}},$$

$$B_4 = \sqrt{\frac{\mu_0 C_{ox}}{L V_{dd}}} R(V_{dd} - V_{th})$$

At '0' → '1' switching, the NMOS transistor is working in the saturation region. The current through the MTJ is:

$$I = \frac{\beta}{2a}\left[(V_{dd} - V_{th} - IR) - \frac{I}{W C_{ox} v_{sat}^2}\right]^2 \tag{26}$$

The MTJ is in its low resistance state, or $R = R_L$. The derivatives can be also calculated as:

$$\begin{aligned} \left(\frac{\partial I}{\partial W}\right)_1^2 \approx \frac{1}{(A_5 W + B_5)^4}, \left(\frac{\partial I}{\partial L}\right)_1^2 \approx \frac{1}{\left(\frac{A_6}{W} + B_6\right)^2} \\ \left(\frac{\partial I}{\partial R}\right)_1^2 \approx \frac{1}{\left(\frac{A_7}{W} + B_7\right)^4}, \left(\frac{\partial I}{\partial V_{th}}\right)_1^2 \approx \frac{1}{\left(\frac{A_8}{W} + B_8\right)^2} \end{aligned} \tag{27}$$

by ignoring the minor terms in the expansion of (24). Here,

$$A_5 = \sqrt{\frac{2\mu_0 C_{ox} v_{sat}}{La + \mu_0(V_{dd} - V_{th})}} R,$$
$$B_5 = \frac{\mu_0}{2C_{ox} v_{sat}[La + \mu_0(V_{dd} - V_{th})]},$$
$$A_6 = \frac{\mu_0}{2aC_{ox} v_{sat}^2},$$
$$B_6 = \frac{R\mu_0}{av_{sat}},$$
$$A_7 = \frac{1}{2C_{ox} v_{sat}} \sqrt{\frac{\mu_0}{Lav_{sat} + \mu_0(V_{dd} - V_{th})}},$$
$$B_7 = \sqrt{\frac{\mu_0}{Lav_{sat} + \mu_0(V_{dd} - V_{th})}} R,$$
$$A_8 = \frac{1}{2C_{ox} v_{sat}}, B_8 = R$$

The contributions of different variation sources to I are represented by:

$$S_1 = \left(\frac{\partial I}{\partial W}\right)^2 \sigma_W^2, S_2 = \left(\frac{\partial I}{\partial L}\right)^2 \sigma_L^2, S_3 = \left(\frac{\partial I}{\partial R}\right)^2 \sigma_R^2$$
$$S_4 = \left(\frac{\partial I}{\partial V_{th}}\right)^2 \left(\frac{C_1}{WL} + \frac{C_2}{\exp(L/l')} \cdot \frac{W_c}{W} \cdot \sigma_L^2\right) \quad (28)$$

Here S_1, S_2, S_3 and S_4 denote the variations induced by W, L, R (R_H or R_L) and V_{th}, respectively.

Analytic Results Summary

Table 3 shows the monotonicity and the upper or lower bounds of the variation contributions $S_1 - S_4$ as the transistor channel width W increases. Here, "↑", "↓" and "↗↘" denote monotonic increasing, monotonic decreasing and changing as a convex function. $K_1 = \frac{C_1}{L} + \frac{C_2 W_c \sigma_L^2}{\exp(L/l')}$. Table 3 also gives the maximum and minimum values of $S_1 - S_4$ and their corresponding W's.

Validation of Analytic Results

As (27) shows, $\left(\frac{\partial I}{\partial W}\right)^2$, $\left(\frac{\partial I}{\partial L}\right)^2$, and $\left(\frac{\partial I}{\partial R}\right)^2$ solely determine the trends of S_1, S_2, S_3, respectively, when W increases at both switching directions. The corresponding

Table 3 Summary of variation contribution

	Variation	Monoto	Bounds	$W \to \infty$
'0'	S_1	↑	$\min S_1 = 0$ $W = \infty$	$S_1 \to 0$
	S_2	↗↘	$\max S_2 = \left(\frac{V_{dd}}{4LR_H}\sigma_L\right)^2$ $W = \frac{L}{\mu_0 C_{ox}(V_{dd} - V_{th})R_H}$	$S_2 \to 0$
	S_3	↑	$\max S_3 = \left(\frac{V_{dd}}{R_H^2}\sigma_{R_H}\right)^2$ $W = \infty$	$\max S_3$
	S_4	↗↘	$\max S_4 = \frac{K_1 \mu_0 C_{ox} V_{dd}^2}{16LR_H(V_{dd} - V_{th})}$ $W = \frac{L}{\mu_0 C_{ox} R_H (V_{dd} - V_{th})}$	$S_4 \to 0$
'1'	S_1	↓	$\min S_1 = 0$ $W = \infty$	$S_1 \to 0$
	S_2	↑	$\max S_2 = \left(\frac{a v_{sat}}{R_L \mu_0}\sigma_L\right)^2$ $W = \infty$	$\max S_2$
	S_3	↑	$\max S_3 \approx \left(\frac{V_{dd} - V_{th}}{R_L^2}\sigma_{R_L}\right)^2$ $W = \infty$	$\max S_3$
	S_4	↗↘	$\max S_4 = \frac{C_{ox} v_{sat}}{2R_L} K_1$ $W = \frac{1}{2C_{ox} v_{sat} R_L}$	$S_4 \to 0$

Monte-Carlo simulation results of S_1, S_2, S_3 are shown in Figs. 19, 20, and 21, respectively.

Figure 19 shows S_1 monotonically decreases to zero as W increases to infinity at both switching directions. Its value at '1' → '0' switching is always greater than that at '0' → '1' switching because $A_1 < A_5$.

Figure 20 shows that the variation contribution of L at '0' → '1' switching is always larger than that at '1' → '0' switching. The gap between them reaches the maximum when $W \to \infty$.

Figure 21 shows that the contribution from MTJ resistance R becomes dominant in the MTJ switching current distribution when W is approaching infinity. Because $\left(\frac{V_{dd} - V_{th}}{R_L^2}\sigma_{R_L}\right)^2 < \left(\frac{V_{dd}}{R_H^2}\sigma_{R_H}\right)^2$, the normalized contribution of R is always larger at '1' → '0' switching than that at '0' → '1' switching. We note that the additional coefficient $\left(\frac{C_1}{WL} + \frac{C_2}{\exp(L/l')} \cdot \frac{W_c}{W} \cdot \sigma_L^2\right)$ at the right side of (28) after the $\left(\frac{\partial I}{\partial V_{th}}\right)^2$ results in the different features of $\left(\frac{\partial I}{\partial V_{th}}\right)^2$ from S_4 in our simulations.

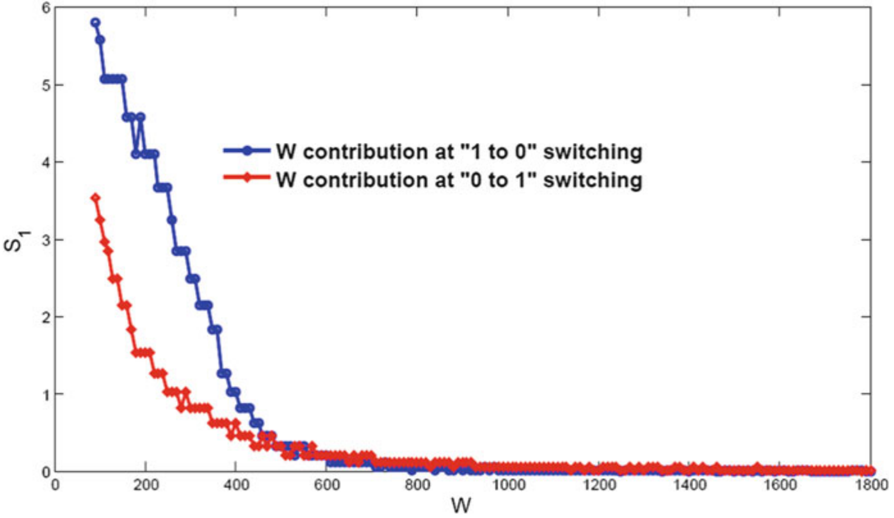

Fig. 19 Contributions from W

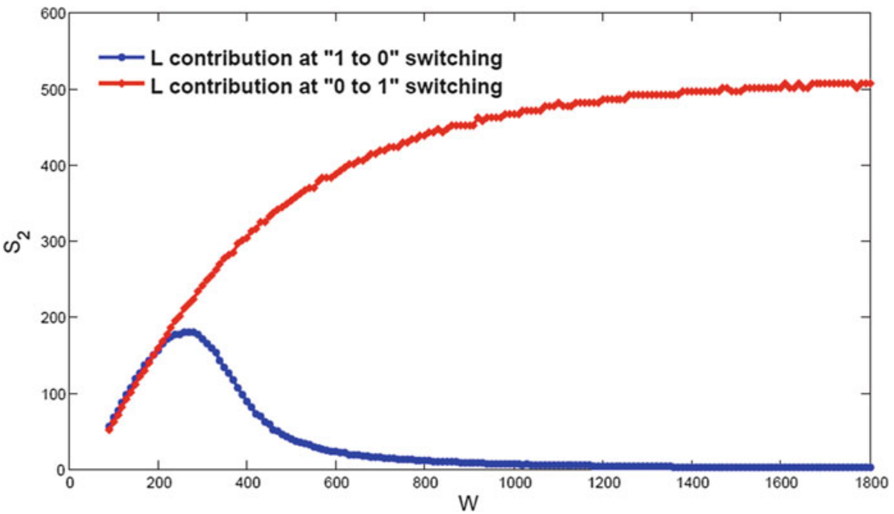

Fig. 20 Contributions from L

Figure 22 shows the values of $\left(\frac{\partial I}{\partial V_{th}}\right)^2$ at both switching directions. At '0' → '1' switching, $\left(\frac{\partial I}{\partial V_{th}}\right)^2$ increases monotonically when W grows. At '1' → '0' switching, $\left(\frac{\partial I}{\partial V_{th}}\right)^2$ increases first, then quickly decays to zero after reaching its maximum. These trends follow the expressions of $\left(\frac{\partial I}{\partial V_{th}}\right)^2$ at either switching directions very well.

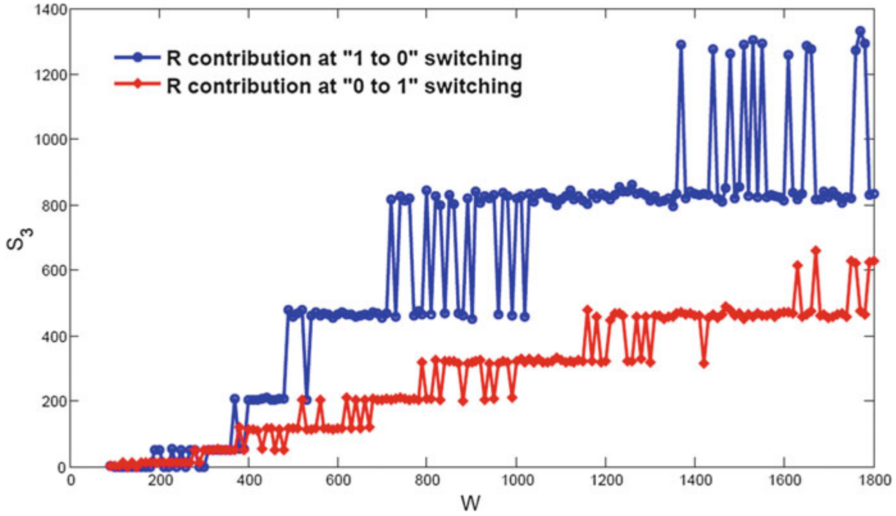

Fig. 21 Contributions from R

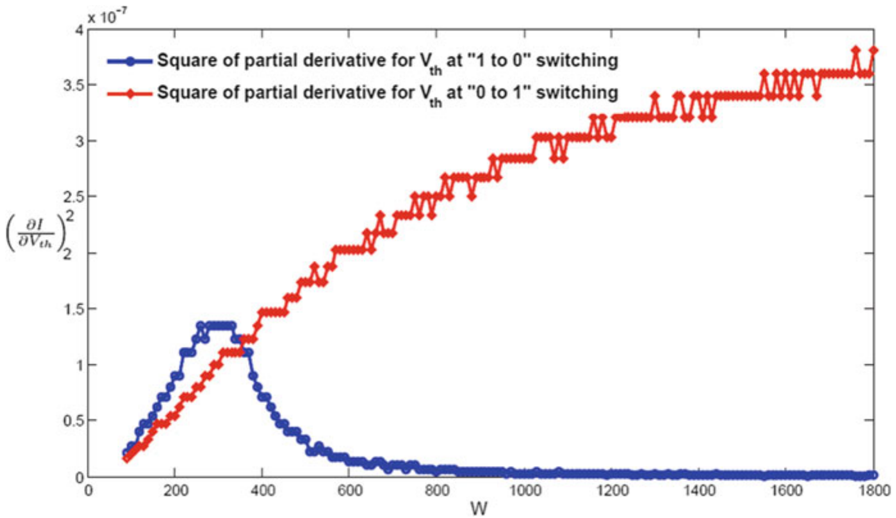

Fig. 22 Square partial derivatives for V_{th}

However, because of the additional coefficient on the top of $\left(\frac{\partial I}{\partial V_{th}}\right)^2$, S_4 does not follow the same trend of $\left(\frac{\partial I}{\partial V_{th}}\right)^2$ at either switching directions. Figure 23 shows that at '0' → '1' switching, S_4 increases first and then slowly decreases when W rises. At

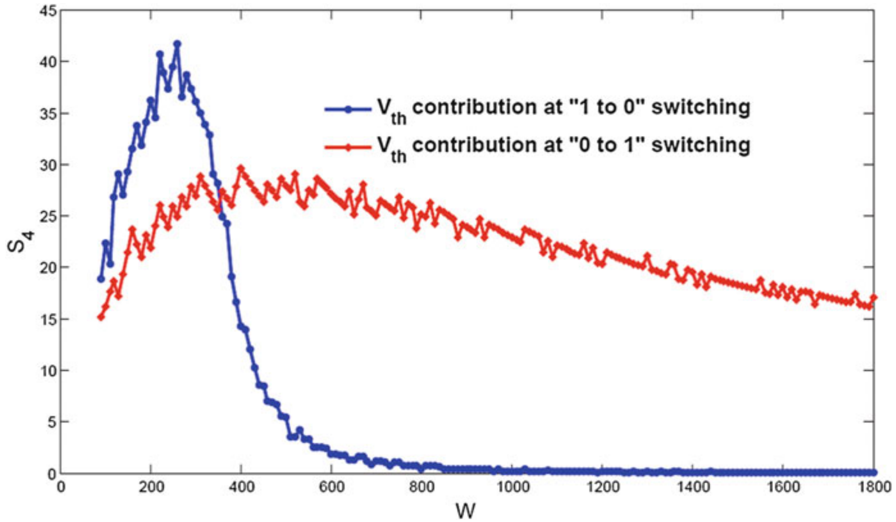

Fig. 23 Contributions from V_{th}

this switching direction, S_4 will become zero when $W \to \infty$ due to the existence of the additional coefficient $\left(\frac{C_1}{WL} + \frac{C_2}{\exp\left(L/l'\right)} \cdot \frac{W_c}{W} \cdot \sigma_L^2\right)$.

All these above results are well consistent with our analytic analysis in Table 3.

References

1. G. Sun, X. Dong, Y. Xie, J. Li, Y. Chen, in *A Novel Architecture of the 3D Stacked MRAM L2 Cache for CMPs*. 15th HPCA. IEEE, 2009, pp. 239–249
2. W. Xu, H. Sun, Y. Chen, T. Zhang, Design of last-level on-chip cache using spin-torque transfer ram (STT-RAM). IEEE Trans. VLSI Syst **19**, 483–493 (2011)
3. P. Zhou, B. Zhao, J. Yang, Y. Zhang, in *Energy Reduction for STTRAM Using Early Write Termination*. ICCAD. ACM, 2009, pp. 264–268.
4. C. Smullen, V. Mohan, A. Nigam, S. Gurumurthi, M. Stan, in *Relaxing Non-volatility for Fast and Energy-Efficient STT-RAM Caches*. High Performance Computer Architecture (HPCA), 2011 IEEE 17th International Symposium, 2011, pp. 50–61.
5. Y. Chen, W.-F. Wong, H. Li, C.-K. Koh, Y. Zhang, W. Wen, On chip caches built on multilevel spin-transfer torque RAM cells and its optimizations. J. Emerg. Technol. Comput. Syst. **9**(2), 16:1–16:22 (2013)
6. Z. Sun, X. Bi, H. Li, W. Wong, Z. Ong, X. Zhu, W. Wu, in *Multi Retention Level STT-RAM Cache Designs with a Dynamic Refresh Scheme*. Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture, 2011, pp. 329–338
7. Z. Sun, X. Bi, and H. Li, *Process Variation Aware Data Management for STT-RAM Cache Design*, in Proceedings of the 2012 ACM/IEEE International Symposium on Low Power Electronics and Design, 2012, pp. 179–184.
8. W. Wen, Y. Zhang, L. Zhang, Y. Chen, in *Loadsa: A yield-Driven Top-Down Design Method for STT-RAM Array*. Design Automation Conference (ASP-DAC), 2013 18th Asia and South Pacific, 2013, pp. 291–296

9. J. Li, H. Liu, S. Salahuddin, and K. Roy, in *Variation-Tolerant Spin-Torque Transfer (STT) MRAM Array for Yield Enhancement*, in CICC, 2008, pp. 193–196.
10. C. W. Smullen, A. Nigam, S. Gurumurthi, M. R. Stan, in *The STeTSiMS STT-RAM Simulation and Modeling System*. ICCAD, 2011, pp. 318–325
11. Y. Chen, X. Wang, H. Li, H. Xi, Y. Yan, W. Zhu, Design margin exploration of spin-transfer torque ram (stt-ram) in scaled technologies. IEEE Trans. Very Large Scale Integr. VLSI Syst. **18**(12), 1724–1734 (2010)
12. W. Zhao, E. Belhaire, V. Javerliac, Q. Mistral, E. Nicolle, C. Chappert, B. Dieny, in *Macro-Model of Spin-Transfer Torque Based Magnetic Tunnel Junction (MTJ) Device for Hybrid Magnetic/CMOS Design*. Proceeding Of IEEE International Behavioral Modeling and Simulation Conference (IEEE-BMAS), San Jose, 2006, pp. 40–44
13. M. El Baraji, V. Javerliac, W. Guo, G. Prenat, B. Dieny, Dynamic compact model of thermally-assisted switching magnetic tunnel junctions. J. Appl. Phys. **106**, 123906 (2009)
14. W. Guo, G. Prenat, V. Javerliac, M. El Baraji, N. de Mestier, C. Baraduc, B. Dieny, SPICE modelling of magnetic tunnel junctions written by spin transfer torque. J. Phys. D Appl. Phys. **43**, 215001 (2010)
15. G. Prenat, G. di Pendina, K. Torki, B. Dieny, J.-P. Nozières. in *Hybrid CMOS/Magnetic Process Design Kit and Application to the Design of High-Performances Non-Volatile Logic Circuits*. Proceedings of the International Conference on Computer-Aided Design, 2011, pp. 240–245
16. S. Chatterjee, M. Rasquinha, S. Yalamanchili, S. Mukhopadhyay, A scalable design methodology for energy minimization of STTRAM: a circuit and architecture perspective. IEEE Trans. Very Large Scale Integr. VLSI Syst. **19**(5), 809–817 (2011)
17. X. Wang, Y. Zheng, H. Xi, D. Dimitrov, Thermal fluctuation effects on spin torque induced switching: mean and variations. JAP **103**(3), 034507 (2008)
18. R. Singha, A. Balijepalli, A. Subramaniam, F. Liu, S. Nassif. in *Modeling and Analysis of Non-Rectangular Gate for Post-Lithography Circuit Simulation*. 44th DAC, 2007, pp. 823–828
19. Y. Ye, F. Liu, S. Nassif, Y. Cao, in *Statistical Modeling and Simulation of Threshold Variation under Dopant Fluctuations and Line-Edge Roughness*. 45th DAC, 2008, pp. 900–905.
20. F. Harris, On the use of windows for harmonic analysis with the discrete fourier transform. Proc. IEEE **66**(1), 51–83 (1978)
21. Y. Zhang, X. Wang, Y. Chen, in *STT-RAM Cell Design Optimization for Persistent and Non-Persistent Error rate Reduction: A statistical Design View*. ICCAD, 2011, pp. 471–477
22. Y. Zhang, W. Wen, Y. Chen, STT-RAM cell design considering MTJ asymmetric switching. *SPIN* **02**(03), 1240007 (2012)
23. W. Wen, Y. Zhang, Y. Chen, Y. Wang, Y. Xie, in *PS3-RAM: A Fast Portable and Scalable Statistical STT-RAM Reliability Analysis Method*. 49th DAC, 2012, 1187–1192
24. W. Wen, Y. Zhang, L. Zhang, Y. Chen, in *Loadsa: A Yield-Driven Top-Down Design Method for STT-RAM Array*. ASP-DAC, 2013, pp. 291-296
25. P. T. M. (PTM) ASU http://www.eas.asu.edu/ptm/
26. BSIM, UC Berkeley http://www-device.eecs.berkeley.edu/bsim3/
27. B. Sheu, D. Scharfetter, P.-K. Ko, M.-C. Jeng, BSIM: berkeley short-channel IGFET model for MOS transistors. JSSC **22**(4), 558–566 (1987)
28. P. Doubilet, C. Begg, M. Weinstein, P. Braun, B. McNeil, Probabilistic sensitivity analysis using monte carlo simulation. A practical approach. Med. Decis. Making **5**(2), 157–177 (1985)
29. W. Xu, Y. Chen, X. Wang, T. Zhang, in *Improving STT MRAM Storage Density Through Smaller-Than-Worst-Case Transistor Sizing*. 46th DAC, 2009, pp. 87–90
30. W. Kang, W. Zhao, Z. Wang, Y. Zhang, J.-O. Klein, Y. Zhang, C. Chappert, D. Ravelosona, A low-cost built-in error correction circuit design for STT-MRAM reliability improvement. Microelectron. Reliab. **53**, 1224–1229 (2013)
31. C. Xu, D. Niu, X. Zhu, H. S. Kang, M. Nowak, Y. Xie, in *Device Architecture Co-Optimization of STT-RAM Based Memory for Low Power Embedded Systems*. ICCAD, 2011, pp. 463–470.