Chapter 5
Asymmetry in STT-RAM Cell Operations

Yaojun Zhang, Wujie Wen and Yiran Chen

Abstract Spin-transfer torque random access memory (STT-RAM) has emerged as a promising technology to replace SRAM and DRAM in embedded memory applications. In STT-RAM, the data are stored in a magnetic device (magnetic tunneling junction or MTJ) as different resistance states. The unique data storage mechanism of STT-RAM introduces the different design optimization concerns from conventional memories. As one important characteristic, programming "1" and "0" into an STT-RAM cell is very asymmetric in terms of performance, power, and reliability. In this chapter, we will review this asymmetry and analyze its sources. The impacts of this asymmetry on the STT-RAM cell optimization will be also discussed, followed by the introduction on a model to simulate the STT-RAM cell asymmetry.

5.1 Introduction

The conventional memory technologies such as SRAM, DRAM, and Flash have achieved a remarkable success in modern electronic designs. Following technology scaling, the shrunk feature size and the increased process variations impose serious concerns on the power and reliability of these conventional memory technologies. Many new memory technologies, including spin-transfer torque random access memory (STT-RAM), have emerged above the horizon. By leveraging a good combination of the non-volatility of Flash, the comparable cell density to DRAM, and the nanosecond programming time like SRAM, many applications of STT-RAM in

Y. Zhang
University of Pittsburgh, Pittsburgh, USA
e-mail: yaz24@pitt.edu

W. Wen
e-mail: wuw2@pitt.edu

Y. Chen (✉)
e-mail: yic52@pitt.edu

Y. Xie (ed.), *Emerging Memory Technologies*,
DOI: 10.1007/978-1-4419-9551-3_5,
© Springer Science+Business Media New York 2014

embedded memory and on-chip cache designs have been successfully demonstrated [16, 20, 23].

Conventional memory technologies rely on electrical charges to store the data. In STT-RAM, the data are represented as the resistance states of a magnetic tunneling junction (MTJ) device, which can be switched by applying a programming current with different polarizations. Compared to the charge-based storage mechanism, the dependency of this mechanism on the device volume is much less, introducing a better scalability. Nonetheless, STT-RAM still suffers from the process variation issue as the technology scales. Also, an intrinsic random process called thermal fluctuations may cause the intermittent errors in the read and write operations of STT-RAM.

As technology scales, the STT-RAM density and power consumption improve. However, process variations on STT-RAM cell designs, including MOS transistor device variations, MTJ geometry variations, and resistance variations, become prominent. Many researches have been conducted to simulate the impacts of process variations and thermal fluctuations on STT-RAM reliability. A common simulation flow is as follows: First, a macro-magnetic model runs extensively to characterize the MTJ switching behaviors under different MTJ device variations; after that, the derived statistical MTJ electrical properties, i.e., the resistance distributions, together with the CMOS device variations, are sent to SPICE simulations to get the MTJ programming current distributions. In some recent works, thermal fluctuations are handled as either the random magnetic field in macro-magnetic models or the resistance noise in SPICE simulations to obtain the MTJ switching time variation or the read disturbance.

In this chapter, the impacts of the device parametric variations in MTJ and transistors and intrinsic MTJ operating uncertainties on the performances and reliability of STT-RAM cells are systematically analyzed. A fast, scalable, and portable statistical STT-RAM reliability analysis methodology, namely "PS3-RAM", is introduced to simulate the impacts of multidimensional variations on the STT-RAM read and write operations. We reveal that the write mechanism of STT-RAM cells is highly asymmetric at different switching directions, i.e., '0'→'1' and '1'→'0'. Specifically, the switching of '0'→'1' takes longer time than '1'→'0' at the same switching current while suffering from the larger variations. This asymmetry is further aggravated by the different biasing conditions of the driving NMOS transistor at different switching directions and the device variations in both MTJ and NMOS transistors. An example of minimizing the design space pessimism that required to tolerate the asymmetry and variations in STT-RAM write performance is also shown.

5.2 STT-RAM Basics

Spin-transfer torque random access memory (STT-MRAM) uses MTJ device to store the information. An MTJ has two ferromagnetic layers (FL) and one oxide barrier layer (BL). The resistance of MTJ depends on the relative magnetization directions (MDs) of the two FLs. When their MDs are parallel or anti-parallel, the MTJ is in

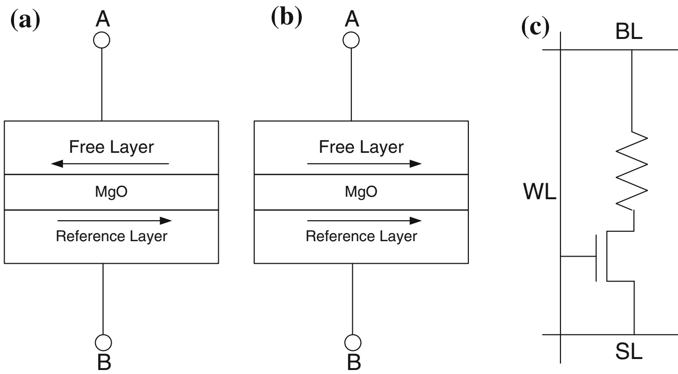

Fig. 5.1 MTJ structure. **a** Anti-parallel (high-resistance state). **b** Parallel (low-resistance state). **c** 1T1J STT-RAM cell structure

its low or high-resistance state, as illustrated in Fig. 5.1. R_h and R_l are usually used to denote the high and the low MTJ resistance, respectively. Tunneling magnetoresistance (TMR) is defined as $(R_h - R_l)/R_l$, which presents the distinction between the two resistance states.

In an MTJ, the MD of one FL (reference layer) is pinned, while the one of the other free layer (FL) can be flipped by applying a polarized write current through the MTJ. For example, the switching from low-resistance state ("0") to high-resistance state ("1") can be realized by applying a current from B to A, as shown in Fig. 5.1. A larger write current can shorten the MTJ switching time by paying the additional memory cell area overhead: In the popular "1T1J" (one-transistor-one-MTJ) cell structure (see Fig. 5.1c), the MTJ write current is supplied by the NMOS transistor. Increasing the write current requires a larger NMOS transistor. Also, the increased write current raises the breakdown possibility of the MTJ device.

5.3 Persistent and Non-persistent Variations

The persistent errors in an STT-RAM cell include the write errors incurred by the insufficient MTJ write current, the variation in MTJ critical switching current, and the insufficient read sense margin. They are mainly caused by the process variations in the NMOS transistors and the MTJ devices.

5.3.1 Persistent Errors

The persistent errors in an STT-RAM cell include the write errors incurred by the insufficient MTJ write current, the variation in MTJ critical switching current, and the insufficient read sense margin. They are mainly caused by the process variations in the NMOS transistors and the MTJ devices.

5.3.1.1 Transistor and MTJ Device Variation

The CMOS process variations that contribute to the variability of the driving strength of the NMOS transistor in an "1T1J" STT-RAM cell structure include random dopant fluctuations (RDFs), line-edge roughness (LER), shallow trench isolation (STI) stress, and the geometry variations in transistor channel length/width. Besides the geometry variations, most of the CMOS process variations are reflected as the threshold voltage deviations. The random variation of the threshold voltage is prominent in the scaled CMOS technology and can severely affect circuit stability and performance.

CMOS process variations affect not only the driving strength of the MOS transistor but also its equivalent resistance. The relative deviations of MOS transistor parameters are reduced when the transistor size increases.

The major sources of MTJ device variations include (1) MTJ shape variations; (2) MgO thickness variations; and (3) normally distributed localized fluctuation of magnetic anisotropy $K = M_s \cdot H_k$ [11]. The first two factors cause the variations in the MTJ resistance and the MTJ switching current by changing the bias conditions of the NMOS transistor. The third factor is the intrinsic variation in magnetic material that affects the MTJ switching threshold current density (Eq. 5.1) and the magnetization stability barrier height (Eq. 5.2) [11].

$$J_{C0} = \left(\frac{2e}{\hbar}\right)\left(\frac{\alpha}{\eta}\right)(t_F M_s)(H_k \pm H_{\text{ext}} + 2\pi M_s) \tag{5.1}$$

$$\Delta = \frac{K_u V}{k_B T} = \frac{M_s H_k V \cos^2(\theta)}{k_B T} \tag{5.2}$$

Here, the switching threshold current density J_{C0} is the minimal current density that causes the MTJ resistance flipping in the absence of any external magnetic field at 0K; e is the electron charge; α is the damping constant; M_s is the saturation magnetization; t_F is the thickness of the free layer; $\hbar$ is the reduced Planck's constant; H_k is the effective anisotropy field including magnetocrystalline anisotropy and shape anisotropy; H_{ext} is the external field; η is the spin-transfer efficiency; T is the working temperature; K_B is Boltzmann constant; and V is MTJ element volume.

Without considering any power rail fluctuations, during the write operation of an STT-RAM cell, the write current through the MTJ is mainly determined by the size of CMOS transistor and the MTJ resistance. The channel width, length, and threshold voltage of the NMOS transistor are the main parameters that affect the CMOS transistor performance. The standard variation in threshold voltage V_{th} decreases when the transistor size increases as $\sigma(V_{\text{th}}) \propto 1/\sqrt{WL}$. The MTJ resistance $R_{\text{MTJ}} \propto e^t/A$, where t is the tunneling oxide thickness and A is the MTJ surface area. The variations in both t and A follow Gaussian distribution [7]. Since $V_{\text{ds}} = V_{\text{dd}} - \Delta V_{\text{MTJ}}$, where $\Delta V_{\text{MTJ}} = I_{\text{MTJ}} \cdot R_{\text{MTJ}}$ is the voltage drop across the MTJ, V_{ds} should be a function of I_{MTJ}.

Normally, the variations in tunneling oxide thickness (t) and cross-sectional area (A) follow Gaussian distributions with a standard deviations of 2 or 5 % of their means, respectively [7]. Also, the MTJ variations are independent from the CMOS process variations since they are fabricated at different layers with different processes.

5.3.1.2 Fluctuation of Magnetic Anisotropy

Different from the CMOS process variations and the MTJ geometry variations that directly affecting the MTJ write current, the localized fluctuation of the MTJ magnetic anisotropy results in the variations in the switching threshold current density J_{C0}. In the concerned MTJ switching time range (a few ns to hundreds of ns), our magnetic model shows that the fluctuation of the MTJ magnetic anisotropy causes a standard deviation of the MTJ switching threshold current density that is about 2 % of the nominal value.

5.3.2 Non-persistent Errors

Device variations are introduced by the uncertainties during the manufacturing process. After the device is fabricated, the device parameters are fixed and their impacts on the circuit performance are deterministic. Besides the device variations of MOS transistor and MTJ, the MTJ switching performance is also affected by the intrinsic thermal fluctuations. The thermal-induced MTJ switching variation is a purely random process and cannot be deterministically repeated. It is the major source of the non-persistent errors in STT-RAM operations.

5.3.2.1 Thermal Fluctuation

In general, the impact of thermal fluctuations can be modeled by the thermal-induced random field h_{fluc} in stochastic Landau–Lifshitz–Gilbert (LLG) equation (Eq. 5.3) [1, 3, 5] as

$$\frac{\mathrm{d}\vec{m}}{\mathrm{d}t} = -\vec{m} \times (\vec{h}_{\text{eff}} + \vec{h}_{\text{fluc}}) + \alpha\vec{m} \times (\vec{m} \times (\vec{h}_{\text{eff}} + \vec{h}_{\text{fluc}})) + \frac{\vec{T}_{\text{norm}}}{M_s}. \tag{5.3}$$

Here, $\vec{m}$ is the normalized magnetization vector. Time t is normalized by γM_s. γ is the gyromagnetic ratio, and M_s is the magnetization saturation. $\vec{h}_{\text{eff}} = \frac{\vec{H}_{\text{eff}}}{M_s}$ is the normalized effective magnetic field. $\vec{h}_{\text{fluc}}$ is the normalized thermal agitation fluctuating field at finite temperature, which represents the thermal fluctuation. α is the LLG damping parameter. $\vec{T}_{\text{norm}} = \frac{\vec{T}}{M_s V}$ is the spin-torque term with units of magnetic field. The net spin torque $\vec{T}$ can be obtained through

microscopic quantum electronic spin transport model. Under the intrinsic thermal fluctuations, the MTJ switching time becomes unrepeatable and follows a distribution. As we shall show later, this distribution is also affected by the MTJ and NMOS transistor device variations and causes the asymmetric STT-RAM cell switching at two switching directions.

Based on the switching time, the switching process of an MTJ can be categorized into three working regions:

In a relatively long switching time ($>$ 10 ns), we have

$$I_{C1}(t_w) = I_{C0}(1 - (1/\Delta)\ln(t_w/\tau_0)). \tag{5.4}$$

Here, t_w is switching time; τ_0 is relaxation time; and T is the working temperature.

In a ultrashort switching time ($<$ 3 ns):

$$I_{C3}(t_w) = I_{C0} + C\ln(\pi/2\theta). \tag{5.5}$$

Here, C is a fitting parameter and θ is the initial angle between the magnetization vector and the easy axis.

When the MTJ switching time is in the middle (3 ns $< t_w <$ 10 ns), a dynamic reversal that combines the precessional and thermally activated switching occurs. Based on the simulation results of our macro-magnetic model, we derive a fitting function of the critical MTJ switching current I_{C2} as

$$I_{C2}(t_w) = 30(I_{C3}(3n) - I_{C1}(10n))/t_w + (10I_{C3}(3n) - 3I_{C1}(10n))/7. \tag{5.6}$$

Here, n is a fitting parameter.

Figure 5.2 shows the MTJ switching current versus the mean and the SDMR of the MTJ switching time. The device parameters are extracted from a 45 $\times$ 90 nm elliptical MTJ device, which have been calibrated with the measurement data of a real fabricated device from a leading magnetic recording company. The results of both switching directions (‘1’→‘0’ and ‘0’→‘1’) are depicted.

Figure 5.2a shows the means of MTJ switching current and switching time in both 1→0 (negative) and ‘0’→‘1’ (positive) switchings for the same MTJ. As Eq. 5.3, thermal fluctuation influences the magnetic process in MTJ switching and causes the variations in MTJ switching time. When MTJ working in a relatively long time region ($>$10 ns), the thermal fluctuation is dominated by the thermal component of internal energy; when MTJ working in a short time region ($<$10 ns), the thermal fluctuation is dominated by the thermally active initial angle of procession [21]. This uncertainty causes the unsuccessful writes if the MTJ does not switch before the write pulse width is removed or the read overwrite errors if the MTJ resistance switches before the read voltage/current is removed.

Figure 5.2b shows the distribution of MTJ switching time at both ‘1’→‘0’ and ‘0’→‘1’ switchings. When the mean of MTJ switching time decreases, its standard deviation decreases first and then increases. The minimal SDMR of the MTJ switching time occurs around $t_w = 10$ ns. Noted that when the MTJ is working at a small

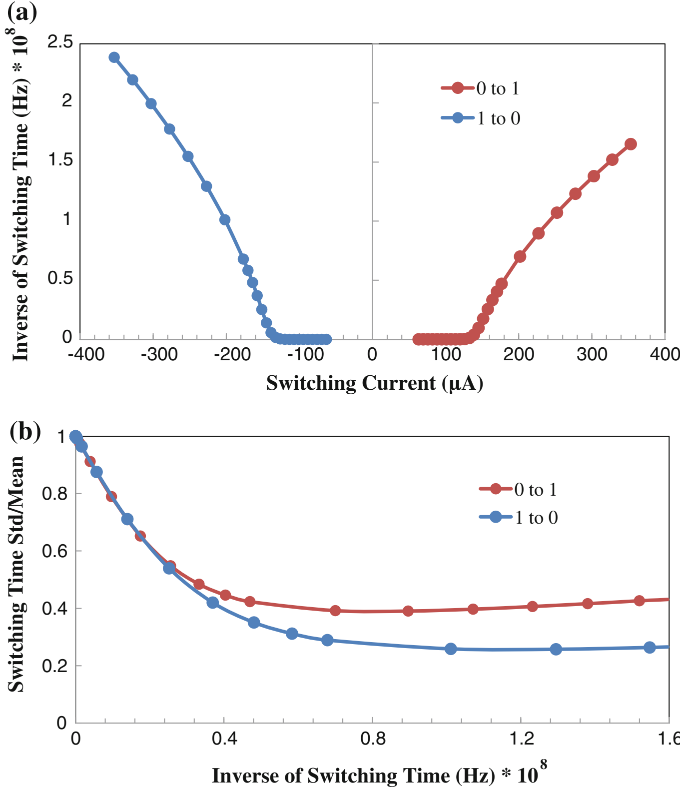

Fig. 5.2 **a** Switching current versus switching time mean. **b** Switching time mean versus switching time standard deviation/mean ratio (SDMR)

current region, the MTJ switching time follows Poisson's distribution; when the MTJ is working at a large current region, the MTJ switching time follows Gaussian distribution; and the MTJ switching time is a mixed distribution when it is in between the two above working regions.

The distinction between the means of the MTJ switching time of two switching directions under the same switching current can be explained as the asymmetric impacts of tunneling spin polarization P as

$$\frac{J_{C0}^{0\to 1}}{J_{C0}^{1\to 0}} = \frac{1+P^2}{1-P^2}. \tag{5.7}$$

Here, $J_{C0}^{0\to 1}$ and $J_{C0}^{1\to 0}$ denote the MTJ switching threshold current densities at the switchings of '0'→'1' and '1'→'0,' respectively.

The different standard deviations of the MTJ switching time at two switching directions, however, are caused by the asymmetric influences of the thermal agitation fluctuating field $\overrightarrow{h}_{\text{fluc}}$. Noticed that when the MTJ works at long

switching time (>40 ns) under a low switching current, the standard deviation of the MTJ switching time for both switching directions is almost the same. However, following the decrease in the MTJ switching time, the standard deviation difference of the MTJ switching time becomes prominent. It is due to the reduced thermal impacts and the increased asymmetry of the spin-torque term $\overrightarrow{T}_{\text{norm}}$ in MTJ switching under a high switching current. In general, MTJ switching time has a larger mean and a wider distribution in ‘0’→‘1’ switching than ‘1’→‘0’ switching under the same switching current.

5.3.2.2 Temperature Dependency

The switching performance of an MTJ can be improved by raising the working temperature. Higher temperature degrades the magnetization stability barrier height (Eq. 5.2) and reduces the critical MTJ switching current and/or the switching time. Figure 5.3 shows the critical MTJ switching currents versus switching time under different temperatures. The impacts of temperature variations are more significant in long working time region: The impact of thermal fluctuations on the MTJ switching performance is more prominent when the MTJ switching current is low, compared to the impact of spin torque.

The temperature sensitivity of the nominal switching time of the MTJ driven by the NMOS transistors with different sizes is shown in Fig. 5.4. The MTJ switching time increases when the temperature rises. It actually indicates that the improvement in MTJ magnetic switching cannot compensate the driving ability loss of the NMOS transistor when the working temperature rises up.

5.4 PS3-RAM Method

“PS3-RAM” is a fast, scalable, and portable statistical STT-RAM reliability analysis methodology. Figure 5.5 depicts the overview of “PS3-RAM” methods, including the sensitivity analysis for MTJ switching current (I), the I sample recovery,

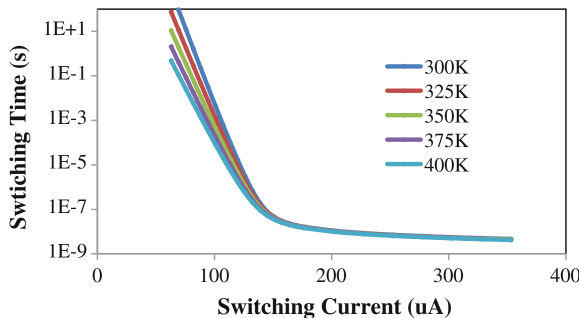

Fig. 5.3 MTJ Critical switching current versus switching time under varying temperature

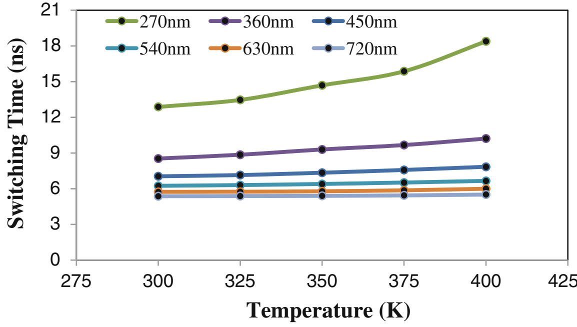

Fig. 5.4 Threshold switching time against temperature

and the statistical thermal analysis of STT-RAM. Array-level analysis and design optimizations can be also conducted using PS3-RAM.

5.4.1 Sensitivity Analysis on MTJ Switching

In this section, the sensitivity model is used for to characterize the MTJ switching current distribution. The contributions of different variation sources to the distribution of the MTJ switching current will be discussed. The definitions of the variables used in this section are summarized in Table 5.1.

5.4.1.1 Threshold Voltage Variation

The variations in channel length (L), width (W), and threshold voltage (V_{th} are three major factors inducing the variations in transistor driving ability. V_{th} variation mainly comes from RDF and LER, which is also the source of some geometry variations (i.e., L and W) [14, 21]. It is known that the V_{th} variation is also correlated with L and W and its variance decreases as the transistor size increases. The deviation of the V_{th} from the nominal value following the change in L (ΔL) can be modeled by [22]

Table 5.1 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	$\overline{V}_{th} = 0.466\,\text{V}$	compact model
Mgo thickness	$\overline{\tau} = 2.2\,\text{nm}$	$\sigma_\tau = 0.02\overline{\tau}$
MTJ surface area	$\overline{A} = 45 \times 90\,\text{nm}^2$	$\sigma_A = 0.05\overline{A}$
Resistance low	$R_L = 1000\,\Omega$	By calculation
Resistance high	$R_H = 2000\,\Omega$	By calculation

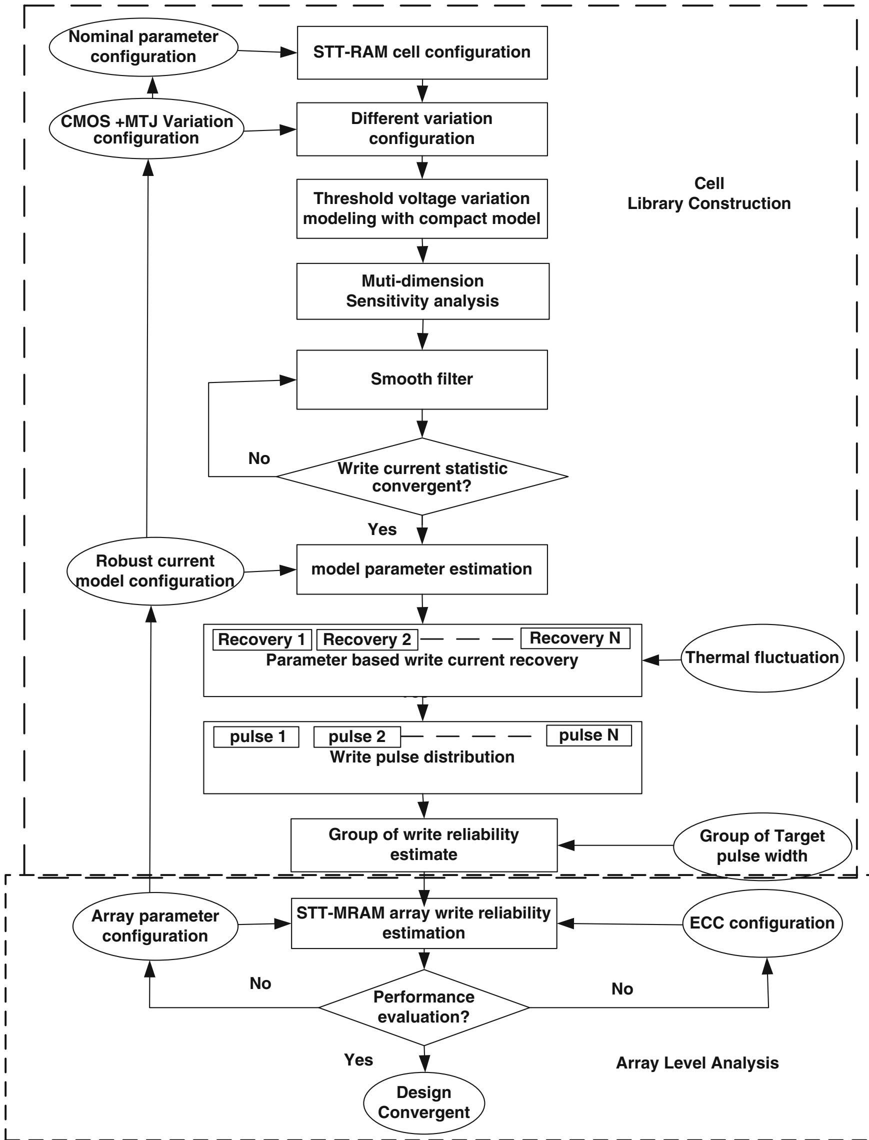

Fig. 5.5 Overview of PS3-RAM

$$\Delta V_{\text{th}} = \Delta V_{\text{th0}} + V_{\text{ds}}\exp(-\frac{L}{l'}) \cdot \frac{\Delta L}{l'}. \tag{5.8}$$

The standard deviation of V_{th} can be calculated as

$$\sigma_{V_{\rm th}}^2 = \frac{C_1}{WL} + \frac{C_2}{\exp\left(\mathrm{L}/l'\right)} \cdot \frac{W_c}{W} \cdot \sigma_L^2. \tag{5.9}$$

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 at the right side of Eq. (5.9) describes the RDF's contribution to $\sigma_{V_{\rm th}}$. The second term in Eq. (5.9) 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 in L and W.

5.4.1.2 Sensitivity Analysis on Variations

Although the contributions of MTJ and CMOS parameters 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 the loss of generality, the MTJ switching current I can be modeled by a function of W, L, $V_{\rm th}$, A, and τ. A and τ are the MTJ surface area and MgO layer thickness, respectively. The first-order Taylor expansion of I around the mean values of every parameter is

$$\begin{aligned} I\left(W, L, v_{\rm th}, A, \tau\right) &\approx I\left(\overline{W}, \bar{L}, \bar{V}_{\rm th}, \bar{A}, \bar{\tau}\right) + \tfrac{\partial I}{\partial W}\left(W - \overline{W}\right) \\ &+ \tfrac{\partial I}{\partial L}\left(L - \bar{L}\right) + \tfrac{\partial I}{\partial V_{\rm th}}\left(V_{\rm th} - \bar{V}_{\rm th}\right) + \tfrac{\partial I}{\partial A}\left(A - \bar{A}\right) + \tfrac{\partial I}{\partial \tau}\left(\tau - \bar{\tau}\right). \end{aligned} \tag{5.10}$$

As aforementioned, W, L, A, and τ generally follow Gaussian distribution [8]. $V_{\rm th}$ is correlated with W and L, as shown in Eq. (5.8) and (5.9). Because the MTJ resistance $R \propto \frac{e^{\tau}}{A}$ [8], we have

$$\frac{\partial I}{\partial A}\Delta A + \frac{\partial I}{\partial \tau}\Delta\tau = \frac{\partial I}{\partial R}\left(\frac{\partial R}{\partial A}\Delta A + \frac{\partial R}{\partial \tau}\Delta\tau\right) = \frac{\partial I}{\partial R}\Delta R. \tag{5.11}$$

It indicates that the combined contribution of A and τ is the same as the impact of MTJ resistance R. The difference between the actual I and its mathematical expectation μ_I can be calculated by

$$\begin{aligned} &I\left(W, L, V_{\rm th}, R\right) - E\left(I\left(\overline{W}, \bar{L}, \bar{V}_{\rm th}, \overline{R}\right)\right) \approx \\ &\frac{\partial I}{\partial W}\Delta W + \frac{\partial I}{\partial L}\Delta L + \frac{\partial I}{\partial V_{\rm th}}\Delta V_{\rm th} + \frac{\partial I}{\partial R}\Delta R. \end{aligned} \tag{5.12}$$

Here, $\mu_I \approx E\left(I\left(\overline{W}, \bar{L}, \bar{V}_{\rm th}, \overline{R}\right)\right) = I\left(\overline{W}, \bar{L}, \bar{V}_{\rm th}, \overline{R}\right)$ and the mean of MTJ resistance $\overline{R} \approx R\left(\bar{A}, \bar{\tau}\right)$. Combining Eqs. (5.8), (5.9), and (5.12), the standard deviation of I (σ_I) can be calculated as

$$\delta_I^2 = \left(\frac{\partial I}{\partial W}\right)^2 \sigma_W^2 + \left(\frac{\partial I}{\partial L}\right)^2 \sigma_L^2 + \left(\frac{\partial I}{\partial R}\right)^2 \sigma_R^2$$
$$+ \left(\frac{\partial I}{\partial V_{\text{th}}}\right)^2 \left(\frac{C_1}{WL} + \frac{C_2}{\exp\left(\text{L}/l'\right)} \cdot \frac{W_c}{W} \cdot \sigma_L^2\right)$$
$$+ 2\frac{\partial I}{\partial L}\frac{\partial I}{\partial v_{\text{th}}}\rho_1 \frac{1}{\sqrt{WL}}\sigma_L + 2\frac{\partial I}{\partial W}\frac{\partial I}{\partial V_{\text{th}}}\rho_2 \frac{1}{\sqrt{WL}}\sigma_W$$
$$+ 2\frac{\partial I}{\partial L}\frac{\partial I}{\partial V_{\text{th}}} V_{\text{ds}} \exp(-\frac{L}{l'})\frac{\sigma_L^2}{l'}. \tag{5.13}$$

Here, $\rho_1 = \frac{\text{cov}(V_{\text{th0}}, L)}{\sqrt{\sigma_{v_{\text{th0}}}^2 \sigma_L^2}}$ and $\rho_2 = \frac{\text{cov}(V_{\text{th0}}, W)}{\sqrt{\sigma_{V_{\text{th0}}}^2 \sigma_W^2}}$ are the correlation coefficients between V_{th0} and L or W, respectively [21]. $\sigma_{V_{\text{th0}}}^2 = \frac{C_1}{WL}$. The last three terms at the right side of Eq. (5.13) are significantly smaller than other terms and can be safely ignored in the simulations of normal STT-RAM operations.

The accuracy of the coefficient in front of the variances of every parameter at the right side of Eq. (5.13) can be improved by applying window-based smooth filtering. Take W as an example:

$$\left(\frac{\partial I}{\partial W}\right)_i = \frac{I\left(\overline{W} + i\Delta W, L, V_{\text{th}}, R\right) - I\left(\overline{W} - i\Delta W, L, V_{\text{th}}, R\right)}{2i\Delta W}, \tag{5.14}$$

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

$$\overline{\frac{\partial I}{\partial W}} = \sum_{i=1}^{K} \omega_i \left(\frac{\partial I}{\partial W}\right)_i. \tag{5.15}$$

Here, ω_i is the weight of sample i, which is determined by the window type, i.e., Hamming window or rectangular window [6].

5.4.1.3 Variation Contribution Analysis

The variations' contributions to I are mainly represented by the first four terms at the right side of Eq. (5.13) as

$$S_1 = \left(\frac{\partial I}{\partial W}\right)^2 \sigma_W^2, S_2 = \left(\frac{\partial I}{\partial L}\right)^2 \sigma_L^2, S_3 = \left(\frac{\partial I}{\partial R}\right)^2 \sigma_R^2$$
$$S_4 = \left(\frac{\partial I}{\partial V_{\text{th}}}\right)^2 \left(\frac{C_1}{WL} + \frac{C_2}{\exp\left(\text{L}/l'\right)} \cdot \frac{W_c}{W} \cdot \sigma_L^2\right). \tag{5.16}$$

As pointed out by many prior arts [22], 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, these phenomena can be well explained by using our sensitivity analysis.

As shown in Fig. 5.1, 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. $V_{\mathrm{gs}} = V_{\mathrm{dd}}$ and $V_{\mathrm{ds}} = V_{\mathrm{dd}} - IR$. The NMOS transistor is working in saturation region when the W is small or in triode region when W is large. Based on short-channel BSIM model, the MTJ switching current supplied by the NMOS transistor working in saturation region can be calculated by

$$I = \frac{\beta \cdot \left[(V_{\mathrm{dd}} - V_{\mathrm{th}}) (V_{\mathrm{dd}} - \mathrm{IR}) - \frac{a}{2} (V_{\mathrm{dd}} - \mathrm{IR})^2 \right]}{1 + \frac{1}{v_{\mathrm{sat}} L} (V_{\mathrm{dd}} - \mathrm{IR})}. \tag{5.17}$$

Here, $\beta = \frac{\mu_0 C_{ox}}{1 + U_0 (V_{\mathrm{dd}} - V_{\mathrm{th}})} \frac{W}{L}$. U_0 is the vertical field mobility reduction coefficient, μ_0 is 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 [10] and short-channel BSIM model [2, 13], 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 $\left(\frac{\partial I}{\partial V_{\mathrm{th}}}\right)^2$ as

$$\left(\frac{\partial I}{\partial W}\right)_0^2 \approx \frac{1}{(A_1 W + B_1)^4}, \quad \left(\frac{\partial I}{\partial L}\right)_0^2 \approx \frac{1}{\left(\frac{A_2}{W} + B_2 W + C\right)^2}$$
$$\left(\frac{\partial I}{\partial R}\right)_0^2 \approx \frac{1}{\left(\frac{A_3}{W} + B_3\right)^4}, \quad \left(\frac{\partial I}{\partial V_{\mathrm{th}}}\right)_0^2 \approx \frac{1}{\left(\frac{A_4}{\sqrt{W}} + B_4 \sqrt{W}\right)^4}. \tag{5.18}$$

Here, R is the high-resistance state of the MTJ, or R_H.

$$A_1 = \sqrt{\frac{\mu_0 C_{ox} V_{\mathrm{dd}} (V_{\mathrm{dd}} - V_{\mathrm{th}})}{L}} R$$
$$B_1 = \sqrt{\frac{L}{\mu_0 C_{ox} V_{\mathrm{dd}} (V_{\mathrm{dd}} - V_{\mathrm{th}})}}$$
$$A_2 = \frac{L^2}{\mu_0 C_{ox} V_{\mathrm{dd}} (V_{\mathrm{dd}} - V_{\mathrm{th}})}$$
$$B_2 = R^2 \mu_0 C_{ox} \frac{V_{\mathrm{dd}} - V_{\mathrm{th}}}{V_{\mathrm{dd}}}$$
$$A_3 = \frac{L}{\mu_0 C_{ox} \sqrt{V_{\mathrm{dd}}} (V_{\mathrm{dd}} - V_{\mathrm{th}})}, \; B_3 = \frac{R}{\sqrt{V_{\mathrm{dd}}}}, \; C = \frac{2LR}{V_{\mathrm{dd}}}$$

$$A_4 = \sqrt{\frac{L}{\mu_0 C_{ox} V_{\rm dd}}},\ B_4 = \sqrt{\frac{\mu_0 C_{ox}}{L V_{\rm dd}}} R\left(V_{\rm dd} - V_{\rm th}\right). \tag{5.19}$$

For a NMOS transistor working in triode region at '0'→'1' switching, the MTJ switching current becomes

$$I = \frac{\beta}{2a}\left[\left(V_{\rm dd} - IR - V_{\rm th}\right) - \frac{I}{W C_{ox} v_{\rm sat}^2}\right]^2. \tag{5.20}$$

where R is the low-resistance state of the MTJ, or R_L. We have

$$\left(\frac{\partial I}{\partial W}\right)_1^2 \approx \frac{1}{\left(A_5 W + B_5\right)^4},\quad \left(\frac{\partial I}{\partial L}\right)_1^2 \approx \frac{1}{\left(\frac{A_6}{W} + B_6\right)^2}$$
$$\left(\frac{\partial I}{\partial R}\right)_1^2 \approx \frac{1}{\left(\frac{A_7}{W} + B_7\right)^4},\quad \left(\frac{\partial I}{\partial V_{\rm th}}\right)_1^2 \approx \frac{1}{\left(\frac{A_8}{W} + B_8\right)^2}. \tag{5.21}$$

Here,

$$\begin{aligned}
A_5 &= \sqrt{\frac{2 C_{ox} v_{\rm sat} \mu_0}{La + \mu_0\left(V_{dd} - V_{\rm th}\right)}} R \\
B_5 &= \frac{\mu_0}{2 C_{ox} v_{\rm sat}\left[La + \mu_0\left(V_{dd} - V_{\rm th}\right)\right]} \\
A_6 &= \frac{\mu_0}{2a C_{ox} v_{\rm sat}^2},\ B_6 = \frac{R\mu_0}{a v_{\rm sat}} \\
A_7 &= \frac{1}{2 C_{ox} v_{sat}} \sqrt{\frac{\mu_0}{La v_{sat} + \mu_0\left(V_{\rm dd} - V_{\rm th}\right)}} \\
B_7 &= \sqrt{\frac{\mu_0}{La v_{\rm sat} + \mu_0\left(V_{\rm dd} - V_{\rm th}\right)}} R \\
A_8 &= \frac{1}{2 C_{ox} v_{\rm sat}},\ B_8 = R.
\end{aligned} \tag{5.22}$$

In general, a large S_i, $i = 1 \ldots 4$, corresponds to a large contribution to I variation. When W is approaching infinity, only S_3 is non-zero at '1'→'0' switching, while both S_2 and S_3 are non-zero at '0'→'1' switching. It indicates that the residual values of S_1–S_4 at '0'→'1' switching are 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.

5.4.1.4 Simulation Results of Sensitivity Analysis

Sensitivity analysis [4] 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 W, L, V_{th}, and R are defined as

$$P_i = \frac{S_i}{\sum_{i=1}^{4} S_i}, i = 1, 2, 3, 4. \tag{5.23}$$

Here,

$$S_1 = \left(\frac{\partial I}{\partial W}\right)^2 \sigma_W^2, S_2 = \left(\frac{\partial I}{\partial L}\right)^2 \sigma_L^2, S_3 = \left(\frac{\partial I}{\partial R}\right)^2 \sigma_R^2,$$
$$S_4 = \left(\frac{\partial I}{\partial V_{\text{th}}}\right)^2 \left(\frac{C_1}{WL} + \frac{C_2}{\exp\left(\mathrm{L}/l'\right)} \cdot \frac{W_c}{W} \cdot \sigma_L^2\right). \tag{5.24}$$

Figures 5.6 and 5.7 show the normalized contributions of every variation source at ‘0’→‘1’ and ‘1’→‘0’ switchings, respectively, at different transistor sizes. We can see that L and V_{th} are two major contributors to I variation at both switching directions when W is small. At ‘1’→‘0’ switching, the contribution of L ramps up 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 keep being the dominant factor over the simulated W range. At both switching directions, the contributions of R rise up when W increases.

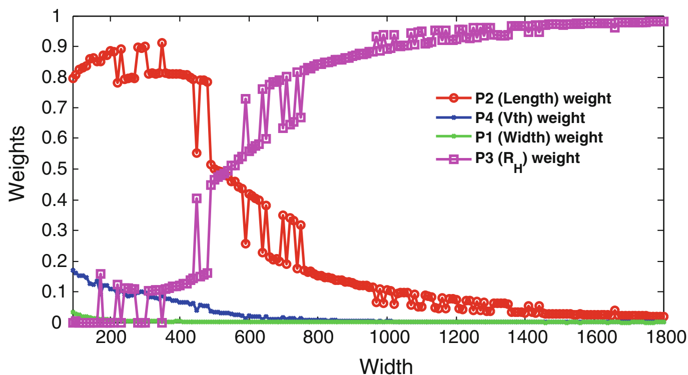

Fig. 5.6 The normalized contributions under different W at ‘1’→‘0’ switching

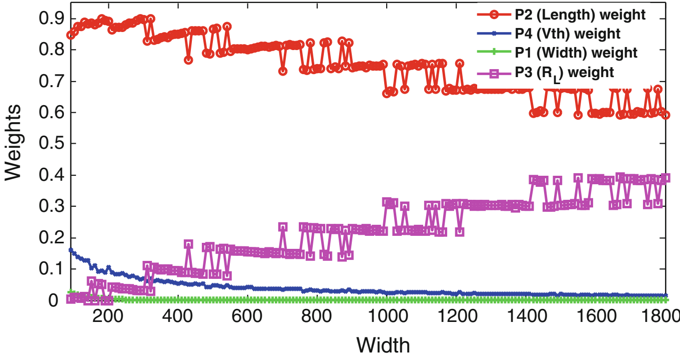

Fig. 5.7 The normalized contributions under different W at '0'→'1' switching

At '1'→'0' switching, the normalized contribution of R becomes almost 100 % when W is really large.

5.4.2 *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. It is found that dual-exponential function can provide an excellent accuracy of the typical distributions of I in modeling and recovering these distributions. The dual-exponential function for the I distributions is shown below:

$$f(I) = \begin{cases} a_1 e^{b_1(I-u)} & I \le u \\ a_2 e^{b_2(u-I)} & I > u. \end{cases} \tag{5.25}$$

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

$$\begin{aligned} &\int f(I)dI = 1 \\ &\int If(I)dI = E(I) \\ &\int I^2 f(I)dI = E(I)^2 + \sigma_I^2. \end{aligned} \tag{5.26}$$

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

The recovered I distribution can be used to generate the MTJ switching current samples, as shown in Fig. 5.8. 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. For example, if N samples are needed to generate within the confidence interval, at the point of $I = I_i$, a switching current sequence

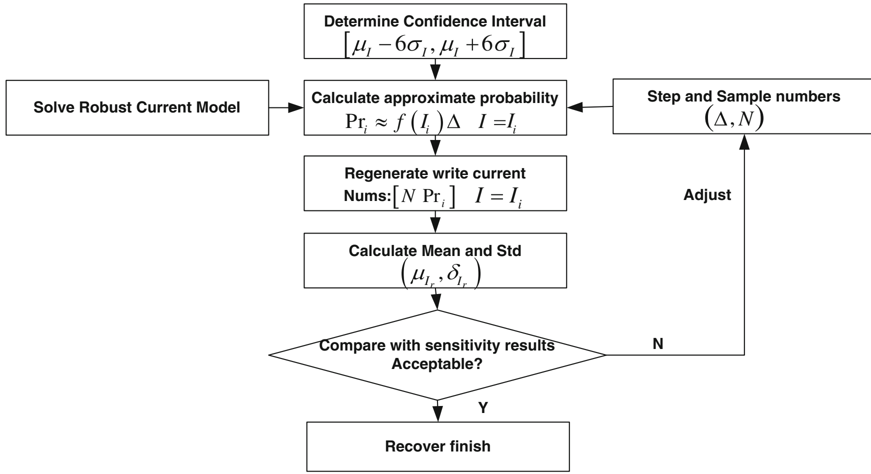

Fig. 5.8 Basic flow for MTJ switching current recovery

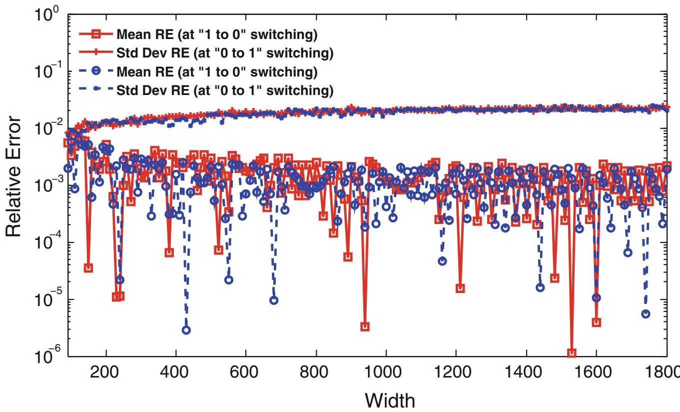

Fig. 5.9 Relative error of the recovered I w.r.t. the result from sensitivity analysis

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.

Figure 5.9 shows the relative error of the mean and the standard deviation of the recovered I distribution w.r.t. the results directly from the sensitivity analysis (see Eqs. (5.12) and (5.13)). The maximum relative error $<10^{-2}$, which proves the accuracy of our dual-exponential model.

Figures 5.10 and 5.11 compare the probability distribution functions (PDF's) of I from SPICE and Monte Carlo simulations and from the recovery process based on our sensitivity analysis at two switching directions. This method achieves good

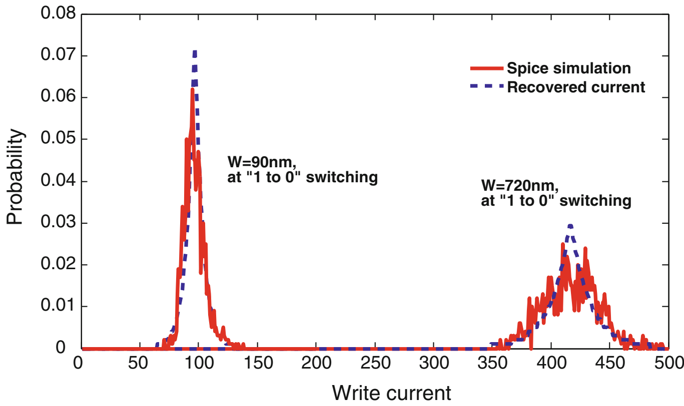

Fig. 5.10 Recovered I versus Monte Carlo result at '1'→'0'

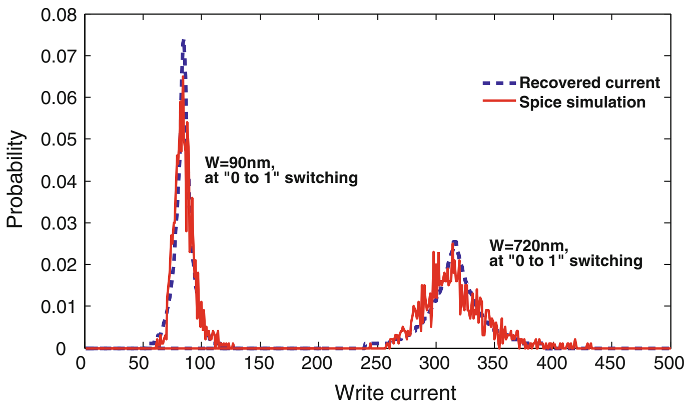

Fig. 5.11 Recovered I versus Monte Carlo result at '0'→'1'

accuracy at both simulated representative transistor channel widths ($W = 90\,\text{nm}$ or $= 720\,\text{nm}$).

5.4.3 Statistical Thermal Analysis

The variation in the MTJ switching time (τ_{th}) incurred by the thermal fluctuations follows Gaussian distribution when τ_{th} is below $10 \sim 20\,\text{ns}$, as Sect. 5.3.2 shows [22]. 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.

5.5 Write Reliability Analysis

In this section, the statistical analysis is conducted on the write reliability of STT-RAM cells by leveraging our PS3-RAM method. Both device variations and thermal fluctuations are considered.

5.5.1 Reliability Analysis of STT-RAM Cells

The write failure rate P_{WF} of an 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 $P_{WF} = P(\tau_{\text{th}} > T_w)$. τ_{th} is impacted by the MTJ switching current, MTJ and MOS device variations, MTJ switching direction, and thermal fluctuations. The simulation of P_{WF} can be conducted by PS3-RAM without incurring the costly Monte Carlo runs with hybrid SPICE and macro-magnetic modeling steps.

Figures 5.12 and 5.13 show the P_{WF}'s simulated with PS3-RAM for both switching directions at 300 K. The simulation environment is summarized in Table 5.1. For comparison purpose, the Monte Carlo simulation results are also presented. Different T_w's are selected at either switching directions due to the asymmetric MTJ switching performances [22]: ($T_w = 10, 15, 20$ ns at '0'→'1' and $T_w = 6, 8, 10, 12$ ns at '1'→'0'). The 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, in Fig. 5.14, the P_{WF} of different STT-RAM cell designs under different temperatures is also compared at this switching direction based on the result in Sect. 5.3.2. 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

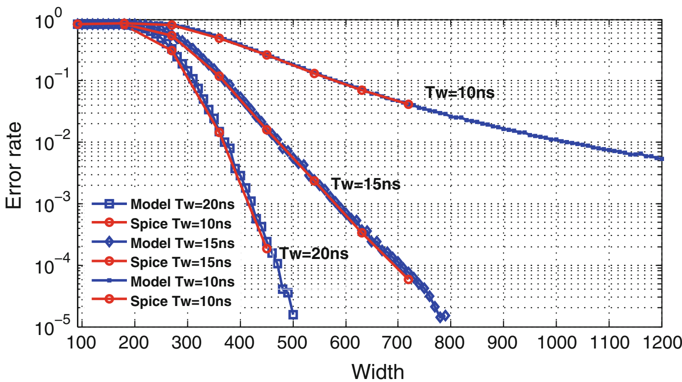

Fig. 5.12 Write failure rate at '0'→'1' when T = 300 K

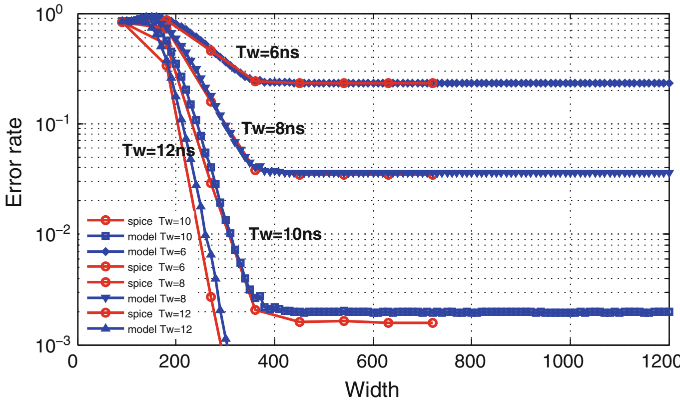

Fig. 5.13 Write failure rate at '1'→'0' when T = 300 K

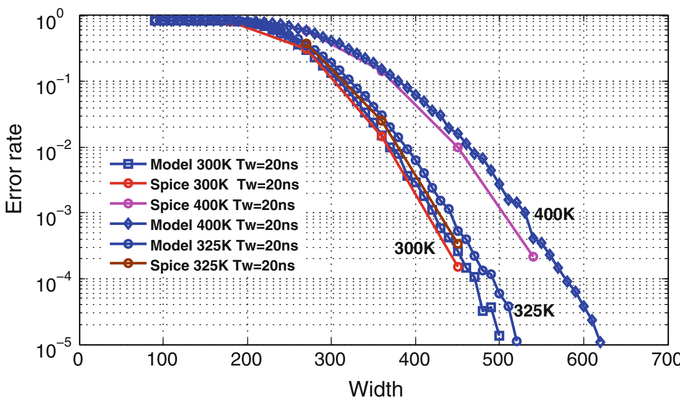

Fig. 5.14 P_{WF} under different temperatures at '0'→'1'

capture the small error rate change due to a little temperature shift (from T = 300 K to T = 325 K).

Figure 5.15 is one of the examples that PS3-RAM is used to explore the STT-RAM design space: The trade-off curves between P_{WF} and T_w are simulated at different W's. The corresponding trade-off between W and T_w can be easily identified on Fig. 5.15.

5.5.2 *Computation Complexity Evaluation*

We can also compare the computation complexity of our proposed PS3-RAM with the conventional simulation method. Assume the number of variation sources is M; for a statistical analysis of an STT-RAM design, the numbers of SPICE simulations

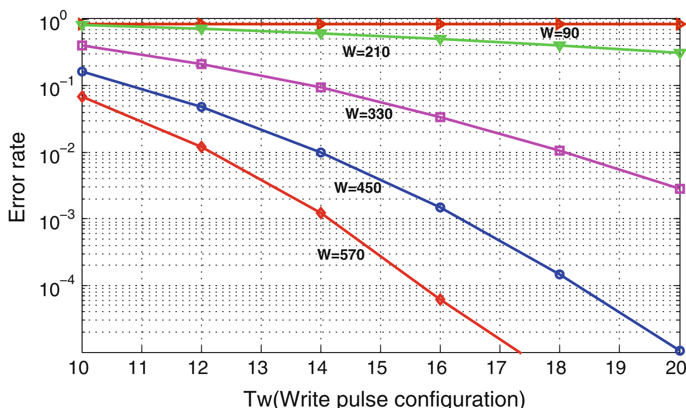

Fig. 5.15 STT-RAM design space exploration at '0'→'1'

required by conventional flow and PS3-RAM are $N_{\mathrm{std}} = {N_s}^M$ and $N_{PS3-\mathrm{RAM}} = 2KM + 1$, respectively. Here, K denotes the sample numbers for window-based smooth filter in sensitivity analysis. N_s is average sample numbers of every variation in the Monte Carlo simulations in conventional method and $K \ll N_s$. The speedup $X_{\mathrm{speedup}} \approx \frac{N_s^M}{2KM}$ 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 speedup is around 2.5×10^5.

5.6 Design Space Exploration

Device variations and thermal fluctuations on the write performance and reliability have been analyzed in the previous section. Based on the statistical analysis in Sect. 5.5, in this section, the outcomes of different design methodologies will be presented, followed by the exploration on the approaches to minimize the design pessimism for the STT-RAM write operations.

A corner design methodology is usually used to overcome the impacts of device variations. The design corner can be set up as the combinations of device parameters. In STT-RAM cell design, the design corner can be set up as follows:

Based on the impacts of the major sources of device variations, the worst corner happens when L, V_{th}, and τ show positive deviations from their nominal values and W shows a negative deviation from its nominal value. However, the worst corner of A is difficult to determined: A large MTJ surface raises the magnitude of MTJ switching threshold current, while it causes the reduction in MTJ resistance, which can improve the NMOS transistor driving ability and vice versa. Two sub-worst corners need to be created for both the positive and the negative deviations of A. Table 5.2 lists the

Table 5.2 3σ worst case parameters

W	L	V_{th}	τ	A
−15 %	+15 %	+15 %	+6 %	±15 %

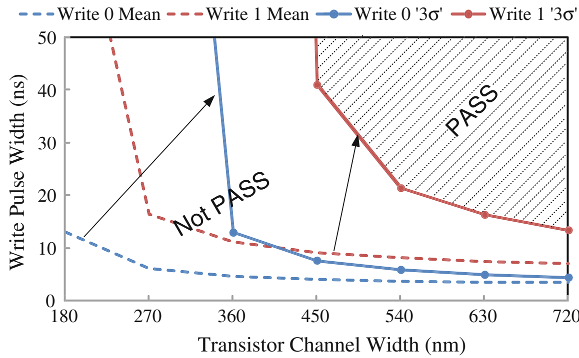

Fig. 5.16 Design space based on device parameters corner (Corner-P-I)

parameter deviations we used in the 3σ worst corner of both NMOS transistor and MTJ devices.

The simulated relationship between the NMOS transistor size and the required write pulse width is shown in Fig. 5.16. Here, only the device variations are considered and the thermal fluctuations are neglected. The required write pulse width value is calculated from the nominal relationship curve between the MTJ write current and the switching time (see Fig. 5.2a), while the MTJ write current is calculated based on the 3σ device parameter corner. The solid blue and red lines denote the results of '1'→'0' switching and '0'→'1' switching, respectively. The worst result is obtained when the MTJ surface area A is 15 % less than the nominal value. Simulation shows that '0'→'1' switching is the limiting switching direction, which requires larger transistor size and/or longer write pulse width. The pass region is constrained by the solid red line. This design method here is called as "Corner-P-I." For comparison purpose, the nominal design result that does not consider any device variations or thermal fluctuations is also plotted as the dash blue and red lines. A larger pass region is observed though it is an optimistic result.

5.6.1 Process Variation Aware Only Corner Design

There is another way to create design corner, i.e., directly using the 3σ value of the MTJ write current distribution to compute the required write pulse width. This method equals characterizing the MTJ write current corner by conventional statistical CMOS circuit design method and then deriving the MTJ switching time with the nominal MTJ switching curve. We refer to this design method as "Corner-P-II." The

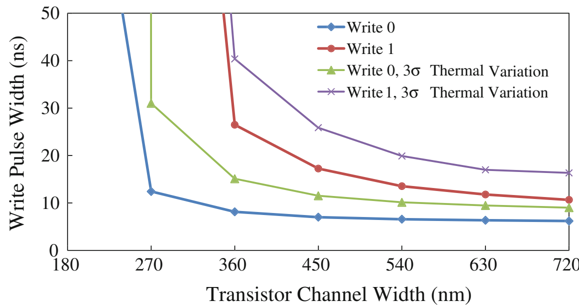

Fig. 5.17 Design space based on driving current corner (Corner-P-II & Corner-PT-II)

corresponding results are shown in Fig. 5.17. The pass region is relaxed from the "Corner-P-I" result by accurately estimating the 3σ corner value of the MTJ write current. However, this result may become optimistic as the thermal fluctuation is ignored.

5.6.2 Process Variation and Thermal Fluctuation Aware Corner Design

As aforementioned, thermal fluctuations cause the variation in MTJ switching time even the MTJ write current is fixed. If thermal fluctuations are considered, a corner representing MTJ switching time variation must be also created in the corner design of STT-RAM cell. For example, the distribution of MTJ switching time under certain MTJ write current can be obtained by macro-magnetic model. Then, the required MTJ write pulse width can be selected as the one corresponding to the $+3\sigma$ deviation of the MTJ switching time from its nominal value. The simulations results of the required write pulse width at different transistor size are also shown in Fig. 5.17. Compared to the result of "Corner-P-II," additional pessimism is added into the pass region because of the consideration on thermal fluctuations. Here, the same current corner of "Corner-P-II" is used, and this corner design is called as "Corner-PT-II."

5.6.3 Process Variation and Thermal Fluctuation Aware Statistical Design

It is well known that the combination of the worst corners of all device parameters may derive very pessimistic design since the worst cases seldom happen simultaneously. To reduce the design pessimism introduced by conventional corner designs, we established our macro-magnetic–SPICE design platform to simulate the statistical

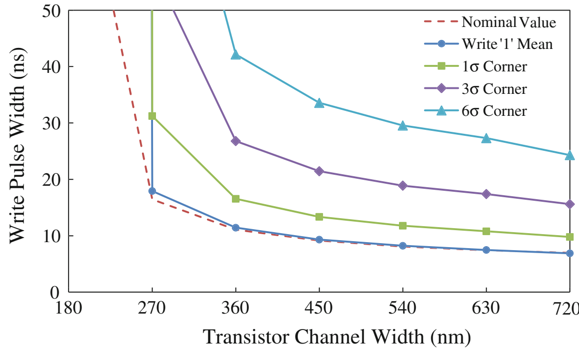

Fig. 5.18 Required write pulse width at various σ's in statistical design

property of STT-RAM cell operations. Monte Carlo simulations are run on both macro-magnetic MTJ model and SPICE transistor model to obtain the overall MTJ switching time distributions when both device variations and thermal fluctuations are considered.

Figure 5.18 shows the pass regions of the STT-RAM cell at different σ's of MTJ switching time. The pass region of the STT-RAM cell at $+3\sigma$ corner of MTJ switching time is between the results of "Corner-P-II" and "Corner-PT-II" designs, indicating the optimism of "Corner-P-II" and the pessimism of "Corner-PT-II." In any cases, '0'→'1' switching continues to be the limiting direction. It actually means that we should avoid the '0'→'1' switching in the operation of STT-RAM, as pointed out by many other studies [9].

5.7 Word-line Override Designs and Statistical Optimization Flow

5.7.1 Word-line Override Designs

As shown in Table 5.3, the effectiveness of increasing NMOS transistor size is degraded by the reduced V_{gs}. Word-line override, which compensates the loss of V_{gs} by adding additional voltage on WL for higher V_{gs}, was proposed to improve the driving ability of the NMOS transistor [17].

We assume that the WL voltage is raised to 1.1V rather than the normal 1V in '0'→'1' switching in WL override scheme. Table 5.3 compares the NMOS transistor driving abilities of the original design and the WL override design. Substantial improvement in MTJ write current is achieved in WL override design, while the incurred current variation is minimal. As a result, the total error rates reduce, as shown in Fig. 5.19.

Table 5.3 Driving current distribution with and without override voltage

Transistor size (nm)	Original design		Override design	
	Mean (μA)	Std. dev. (μA)	Mean (μA)	Std. dev. (μA)
180	148.28	14.35	169.07	14.39
270	194.75	18.11	222.21	18.19
360	230.18	20.68	262.80	20.85
450	258.18	22.76	294.89	23.01
540	280.79	24.51	320.83	24.87
630	299.91	26.15	342.69	26.62
720	315.41	27.31	360.49	27.88

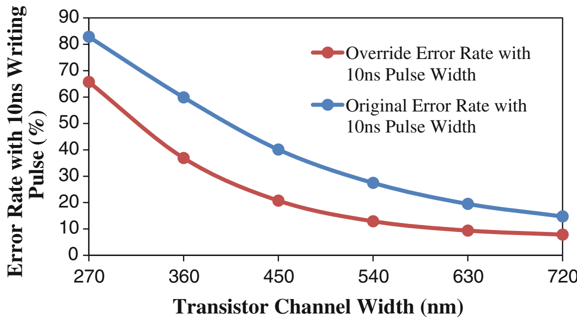

Fig. 5.19 Error rate for 10 ns writing pulse width

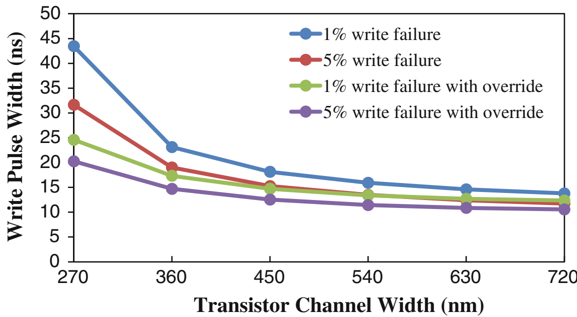

Fig. 5.20 1 and 5 % error rate for required writing pulse width

Figure 5.20 depicts the required write pulse widths by both original and WL override designs for certain total write error rates, in both ‘0’→‘1’ and ‘1’→‘0’ switches. Substantial reductions in the required writing pulse width are observed in WL override design. However, the effectiveness of WL override design is degraded by the reduced V_{gs} when the NMOS transistor size increases.

5.7.2 STT-RAM Cell Design Optimization Flow

Figure 5.21 illustrates an STT-RAM cell design optimization flow to estimate and minimize the operation errors. After the MTJ device parameters are given, the NMOS transistor sizes are calculated accordingly, based on the designed (nominal) values of both MTJ and CMOS parameters. Meanwhile, a reasonable operation pulse width will be calculated based on the nominal design. In the second step, the device parameter samples, including both the geometry and the material parameters, are generated based on the process variations in both NMOS transistor and MTJ. These samples are sent to the SPICE simulations to collect the write current samples through the MTJ. The third step takes into account the thermal fluctuation effects and the fluctuation of magnetic anisotropy under the given operation pulse width to calculate the distribution of the MTJ switching time and the write errors. Based on the specific write performance and the write error rate requirements, the optimal design points for both the NMOS transistor and the MTJ are found. If the result leads to a design

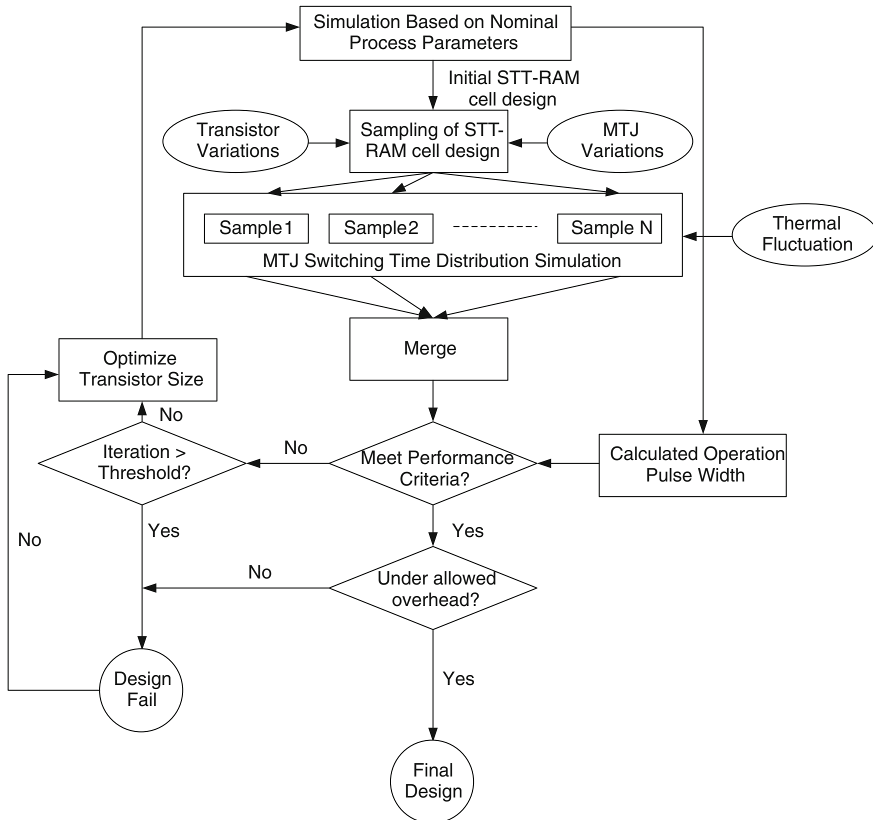

Fig. 5.21 Precess variation aware STT-RAM design flow

failure, then the override design will be applied to meet the performance. Similar design flow can be applied to the read error optimization or take it into the overall STT-RAM error rate optimization design flow.

5.8 Conclusion

In this chapter, we conduct a comprehensive discussion on the major variation sources in the STT-RAM designs and quantitatively analyze their impacts on the STT-RAM cell read and write operations. Both process variations and thermal fluctuations, which cause the significantly unbalanced write reliability at the switchings of '1'→'0' and '0'→'1', are considered in the analysis. After that, a fast and scalable statistical STT-RAM reliability analysis method named PS3-RAM is introduced. PS3-RAM is able to simulate the impact of the concerned variation sources on the statistical STT-RAM write performance, without running the costly Monte Carlo simulations on SPICE and macro-magnetic models. The effectiveness of different design methodologies are also evaluated, including nominal design, corner designs (with only device variations and with both device variations and thermal fluctuations) and full statistical design.

Acknowledgments This work was supported by National Science Foundation grants CNS-1116171 and CCF-1217947, and 49th Design Automation Conference A. Richard Newton Scholarship.

References

1. Berger, L. (Oct 1996). Emission of Spin waves by a magnetic multilayer traversed by a current. *Physical Review B*, *54*, 9353–9358.
2. BSIM.http://www-device.eecs.berkeley.edu/bsim3/. UC Berkeley.
3. Diao, Z., Li, Z., Wang, S., Ding, Y., Panchula, A., Chen, E., et al. (2007). Spin-transfer torque switching in magnetic tunnel junctions and Spin-transfer torque random access memory. *Journal of Physics: Condensed Matter*, *19*, 165209.
4. Doubilet, P., Begg, C. B., Weinstein, M. C., Braun, P., McNeil, B. J. (1985). A Practical approach: Probabilistic sensitivity analysis using Monte Carlo Simulation.
5. Gilbert, T. L. (1955). A lagrangian formulation of the gyromagnetic equation of the magnetization field. *Physics Review*, *100*(1243).
6. Harris, F. J. (Jan. 1978). On the use of windows for Harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, *66*(1), 51–83.
7. Li, J., Augustine, C., Salahuddin, S., Roy, K. (2008). Modeling of failure probability and statistical design of spin-torque transfer magnetic random access memory (STT MRAM) array for yield enhancement. In 45th Design Automation Conference, pp. 278–283, june 2008.
8. Li, J., Liu, H., Salahuddin, S., Roy, K. (2008). Variation-tolerant Spin-Torque transfer (STT) MRAM array for yield enhancement. In CICC, pp. 193–196, Sep. 2008.
9. Nigam, A., Smullen, C. W., Mohan, V., Chen, E., Gurumurthi, S., Stan, M. R. (2011). Delivering on the promise of universal memory for Spin-Transfer Torque RAM (STT-RAM), *International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 121–126, aug 2011.

10. Predictive Technology Model (PTM). http://www.eas.asu.edu/ptm/. ASU.
11. Raychowdhury, A., Somasekhar, D., Karnik, T., De V. (2009). Design space and scalability exploration of 1T–1STT MTJ memory arrays in the presence of variability and disturbances. In IEEE International Electron Devices Meeting (IEDM), pp. 1–4, dec. 2009.
12. Raychowdhury, A., Somasekhar, D., Karnik, T., De, V. (2009). Design space and scalability exploration of 1t–1stt MTJ memory arrays in the presence of variability and disturbances. In IEDM, pp. 1–4, Dec. 2009.
13. Sheu, B. J., Scharfetter, D. L., Ko, P.-K., & Jeng, M.-C. (Aug 1987). BSIM: Berkeley short-channel IGFET model for MOS transistors. *JSSC*, *22*(4), 558–566.
14. Singha, R., Balijepalli, A., Subramaniam, A., Liu, F., Nassif, S. (2007). Modeling and analysis of non-rectangular gate for post-lithography circuit simulation. In 44th DAC, pp. 823–828, June 2007.
15. Smullen, C. W., Nigam, A., Gurumurthi, S., Stan, M. R. (2011). The STeTSiMS STT-RAM simulation and modeling system. In ICCAD, pp. 318–325, Nov 2011.
16. Sun, G., Dong, X., Xie, Y., Li, J., Chen, Y. (2009). A novel architecture of the 3D stacked MRAM L2 cache for CMPs. In 15th HPCA, pp. 239–249. IEEE, 2009.
17. Wang, X., Zheng, Y., Xi, H., Dimitrov, D. (2008). Thermal fluctuation effects on Spin Torque induced switching: Mean and variations. *JAP*, *103*(3):034507–034507-4, Feb. 2008.
18. Xu, W., Chen, Y., Wang, X., Zhang, T. (2009). Improving STT MRAM storage density through smaller-than-worst-case transistor sizing. In 46th DAC, pp. 87–90, July 2009.
19. Xu, C., Niu, D., Zhu, X., Kang, H. S, Nowak, M., Yuan, X. (2011). Device architecture co-optimization of STT-RAM based memory for low power embedded systems. In ICCAD, p. 463–470, Nov 2011.
20. Xu, W., Sun, H., Chen, Y., Zhang, T. (2011). Design of last-level on-chip cache using Spin-Torque Transfer RAM (STT-RAM). In *IEEE Transactions on VLSI System*, pp. 483–493. IEEE, 2011.
21. Ye, Y., Liu, F., Nassif, S., Cao, Y. (2008). Statistical modeling and simulation of threshold variation under dopant fluctuations and line-edge roughness. In 45th DAC, pp. 900–905, June 2008.
22. Zhang, Y., Wang, X., Chen, Y. (2011). STT-RAM Cell Design Optimization for Persistent and Non-Persistent Error rate Reduction: A statistical Design View. In ICCAD, pp. 471–477, Nov. 2011.
23. Zhou, P., Zhao, B., Yang, J., Zhang, Y. (2009). Energy Reduction for STT-RAM Using Early Write Termination. In ICCAD, pp. 264–268. ACM, 2009.