# Multi-level Cell STT-RAM: Is It Realistic or Just a Dream?

Yaojun Zhang[1], Lu Zhang[1], Wujie Wen[1], Guangyu Sun[2], Yiran Chen[1]
[1]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA
[2]Center for Energy-efficient Computing and Applications School of EECS, Peking University, Beijing, China
{yaz24, luz14, wuw2, yic52}@pitt.edu, gsun@pku.edu.cn

***Abstract*—Spin-transfer torque random access memory (STT-RAM) is a promising nonvolatile memory technology aiming on-chip or embedded applications. In recent years, many researches have been conducted to improve the storage density and enhance the scalability of STT-RAM, such as reducing the write current and switching time of magnetic tunneling junction (MTJ) devices. In parallel with these efforts, the continuous increasing of tunnel magneto-resistance(TMR) ratio of the MTJ inspires the development of multi-level cell (MLC) STT-RAM, which allows multiple data bits be stored in a single memory cell. Two types of MLC STT-RAM cells, namely, parallel MLC and series MLC, were also proposed. The storage margin of a MLC STT-RAM cell, i.e., the distinction between the lowest and highest resistance states, is partitioned into multiple segments for multi-level data representation. As a result, the performance and reliability of MLC STT-RAM cells become more sensitive to the MOS and MTJ device variations and the thermal-induced randomness of MTJ switching. In this work, we systematically analyze the variation sources of MLC STT-RAM designs and their impacts on the reliability of the read and write operations. On top of that, we also discuss the optimal device parameters of the MLC MTJ for the minimization of the operation error rate of the MLC STT-RAM cells from statistical design perspective. Our simulation results show that under the current available technology, series MLC STT-RAM demonstrates overwhelming benefits in the read and write reliability compared to parallel MLC STT-RAM and could potentially satisfy the requirement of commercial practices.**

## I. Introduction

Spin-transfer torque random access memory (STT-RAM) is an emerging nonvolatile memory technology which aims embedded memory and on-chip cache applications. In an STT-RAM cell, data is stored as the two or more resistance states of a magnetic tunneling junction (MTJ) device. The resistance states of the MTJ is determined by the magnetization of the magnetic layers, which can be changed by passing through an electrical current with different polarizations. Such a unique storage mechanism offers STT-RAM many attractive characteristics, such as fast operation time, small memory cell size, radiation hardness, good CMOS process compatibility and scalability, etc. [14].

Compared to the mainstream on-chip memory technologies such as SRAM and embedded DRAM, an obvious drawback of STT-RAM is the high switching current of the MTJ, which incurs large write energy dissipation. Since the MTJ switching current in an STT-RAM cell is supplied by the MOS device, the magnitude of the switching current determines not only the write energy of STT-RAM but also its memory integration density. Hence, many new types of MTJ devices have been proposed to reduce the switching current, i.e., the MTJ with perpendicular magnetization [6] and the MTJ with dual tunneling barriers [11]. In parallel, multi-level cell (MLC) technology is also explored in STT-RAM designs to store more than one data bit in a single STT-RAM cell: following the improvement on the distinction between the lowest and the highest resistance states of the MTJ, it becomes possible to further divide the MTJ resistance into multiple levels to represent the combinations of multiple data bits [3], [7].

In general, the variability sources in STT-RAM designs include: 1) the device parametric deviations of MOS transistor and MTJ, such as the variations of the geometry sizes [1], the threshold voltage [19], and the magnetic materials [12]; and 2) the thermal fluctuations in the MTJ switching [13]. Compared to single-level cell (SLC) STT-RAM designs, the impacts of such design variabilities in MLC STT-RAM designs are even more prominent due to the scaled data storage margin. Although the impacts of these variations on the SLC STT-RAM designs have been well studied in many previous works [9], it is still unclear if MLC STT-RAM is a viable technology when all the design variabilities are taken into account. Also, the robustness of the different types of MLC MTJ designs require further investigations because obviously their resilience to the variations varies by the difference of device structures.

In this work, we systematically analyze the variability sources of two MLC STT-RAM designs, namely, parallel MLC and series MLC, and investigated the impacts of these variations on the memory performance and reliability. We hope our work can answer the questions that STT-RAM researchers have for a long time: *Can MLC STT-RAM designs be realized by using the existing technologies? And if not, how far away are we from it?* Luckily, our analysis shows that at least the series MLC STT-RAM may potentially be implemented by using the stacked MTJ structure [7] and achieve an acceptable reliability for some commercial applications. Based on our analysis, we also discuss the design optimization methods to minimize the operation error rate of MLC STT-RAM.

The rest of our paper is organized as follows: Section II gives the preliminary on MLC STT-RAM designs and the variability sources including process variations and thermal fluctuations; Section III analyzes the impacts of variations on the readability of MLC STT-RAM; Section IV discusses the impacts of variations on the writability of MLC STT-RAM; Section V concludes our works.

## II. Preliminary

### A. SLC MTJ and MLC MTJ

The data storage device in an STT-RAM cell is magnetic tunneling junction (MTJ), where a tunneling oxide layer is sandwiched between two ferromagnetic layers. The MTJ resistance is determined by the relative magnetization directions of the two ferromagnetic layers: when their magnetization directions are parallel (anti-parallel), the MTJ is in its low (high) resistance state, as shown in Fig. 1(a). The magnetization direction of the reference layer is fixed while that of the free layer can be flipped by passing a spin-polarized current [14]. A parameter called "tunneling magnetoresistance ratio ($TMR$)" is introduced to measure the distinction between the two resistance states of the MTJ as $(R_H - R_L)/R_L$. Here $R_H$ and $R_L$ denote the high- and the low-resistance states of the SLC MTJ, respectively



IEEE/ACM International Conference on Computer-Aided Design (ICCAD) 2012, November 5-8, 2012, San Jose, California, USA

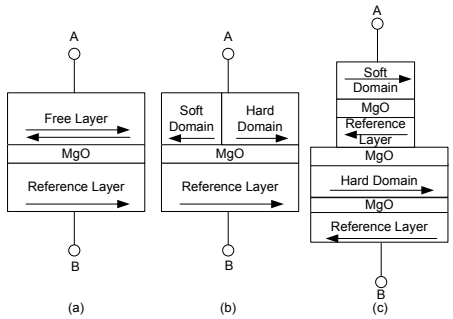


Fig. 1. (a) Conventional MTJ. (b) Parallel MLC MTJ. (c) Series MLC MTJ.

The multi-level cell (MLC) capability can be implemented by realizing four or more resistance levels in MTJ designs. At least two proposals of MLC MTJ structures have emerged [7], [10] so far, including parallel MLC MTJs and series MLC MTJs. Fig. 1(b) and (c) show the structures of a two-bit parallel MLC MTJ (b) and a two-bit series MLC MTJ (c), respectively.

In parallel MLC MTJs, the four resistance states – '00', '01', '10', and '11', are uniquely defined by the four combinations of the magnetic directions of the two magnetic domains in the free layer. The first and the second digits of the two-bit data refer to the resistance states of the hard domain and the soft domain, respectively, as shown in Fig. 2 [2]. In series MLC MTJs, the four resistance states are uniquely defined by the combinations of the relative magnetization of the two SLC MTJs. The minimal device size of a parallel MLC MTJ and the small SLC MTJ in a series MLC MTJ can be as the same as that of the normal SLC MTJ, which is defined by the required aspect ratio and the lithography limit. We note that the parallel MLC MTJ design is only applicable to in-plane MTJ technology because it requires the different aspect ratios of the two magnetic domains to achieve different switching current densities. The series MLC MTJ design, however, is compatible to the advanced MTJ technologies such as perpendicular MTJ etc. [15].

### B. Variability Sources in MLC STT-RAM Designs

The performance and reliability of MLC STT-RAM cells are seriously affected by mainly two types of variabilities, including a) the process variations of MOS and MTJ devices and b) the thermal fluctuations in MTJ switching process.

*1) Process Variations:* The major sources of MTJ device variations mainly include: 1) MTJ shape variations, i.e., the surface area variation; 2) MgO layer thickness variations; and 3) normally distributed localized fluctuation of magnetic anisotropy: $K = M_s \cdot H_k$.

The MTJ device variations affect the reliability of the two types of MLC MTJs in the different ways: In parallel MLC MTJs, the two parts of the MTJ in different magnetic domains (For simplicity, we also call them "two magnetic domains" in the rest of this paper) share the same free layer, reference layer and MgO layer. In such a small geometry size, we can assume the MgO layer thickness and the RA (resistance-area) of these two parts are fully correlated. Other

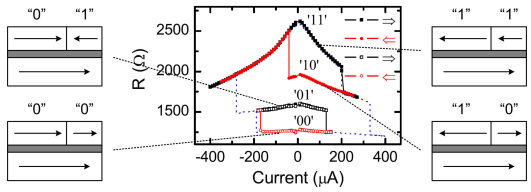


Fig. 2. Four resistance states of MTJ and R-I swap curve [2].

parameters, such as the MTJ surface areas, the magnetic anisotropy and the required switching current density can be very different for these two parts because they are determined by the magnetic domain partitioning. In series MLC MTJs, however, all these parameters of two SLC MTJs are close to each other and only partially correlated.

We note that the MOS device variations also impacts the robustness of MLC STT-RAM designs by causing the magnitude variations of the read and the write currents of the MTJ. In our reliability analysis of MLC STT-RAM, the parametric variability of MOS devices is represented by the variations of the current source output.

*2) Thermal Fluctuations:* The thermal fluctuations results in the randomness of the MTJ switching time so that the MTJ switching time becomes a distribution. A write failure occurs when the MTJ switching time is longer than the write pulse width. The impact of thermal fluctuations is an accumulative effects and determined by the length of the MTJ switching time. The reduction of switching current does not only prolong the MTJ switching time but also increases the ratio between the standard deviation and the mean value of the switching time [4], indicating a larger impact of thermal fluctuations. Hence, in MLC STT-RAM designs, the impacts of thermal fluctuations could be stronger than that in the SLC STT-RAM designs when the MTJ switching current density is lower than that of the SLC MTJ (e.g., during the soft-domain flipping in parallel MLC MTJs).

## III. Readability Analysis of MLC MTJs

### A. Nominal Analysis of the Readability of MLC MTJs

We assume that the resistances of the hard domain and the soft domain in a parallel MLC MTJ are $R_1$ and $R_2$, respectively. The corresponding the high and the low resistance states of the two domains are $R_{1H}$, $R_{1L}$, $R_{2H}$, and $R_{2L}$, respectively. The $TMR$ ratio of each domain is defined as: $\frac{R_{iH}-R_{iL}}{R_{iL}}, (i = 1, 2)$. As aforementioned in Section II-B1, the two magnetic domains share the same magnetic structure and MgO layer within a small proximity. Thus, we can safely assume the $RAs$ and the $TMRs$ of the two domains are the same, or $RA_{1j} = RA_{2j}, (j = H or L)$ and $\frac{R_{1H}}{R_{1L}} = \frac{R_{2H}}{R_{2L}}$. For the existing in-plane MTJ technology, the typical $TMR$ ratio is $1 \sim 1.2$ [7]. Because the size of the hard domain is larger than that of the soft domain, we have $R_{1H} < R_{2H}$ and $R_{1L} < R_{2L}$. In the simulations in our work, we assume the surface area of the parallel MLC MTJ is a $45nm \times 90nm$ ellipse, which is the minimum shape that satisfies the shape anisotropy requirement [5], [16] and is allowed by the lithography constraint of 45nm CMOS fabrications process.

Sense margin is one of the major concerns in MLC STT-RAM designs because the resistance state distinction of the MTJ is partitioned into multiple levels. Read errors happen when the distributions of the two adjacent resistance states (i.e., 00 vs. 01, 01 vs. 10, and 10 vs. 11) overlap with each other, or the distinction between the two resistance states is smaller than the sense amplifier resolution. The reading error rate can be reduced by maximizing the distinctions between every two adjacent states. Without considering the process variations, the goal of the nominal design method of MLC STT-RAM cell is to maximize the distinctions between the designed values of every two adjacent resistance states.

In the real implementation of parallel MLC MTJs, $R_{00} = R_{1L}||R_{2L}$ and $R_{11} = R_{1H}||R_{2H}$ are fixed by the MTJ designs. The changes of $R_{01}$ and $R_{10}$ are not independent and determined by the partitioning of the free layer. If we assume the surface area of the parallel MLC MTJ is $A$ and the surface area of the hard domain

is $A_1$, we have:

$$R_{1L}\cdot A_1 = R_{2L}\cdot(A - A_1) = R_{00}\cdot A, \quad (1)$$

$$R_{1H}\cdot A_1 = R_{2H}\cdot(A - A_1) = R_{11}\cdot A. \quad (2)$$

Here $A_1 > A/2$. The distinctions between every two adjacent resistance states can be calculated as:

$$D_{00-01} = R_{01} - R_{00} = \frac{TMR\cdot RA}{A}\cdot\frac{A-A_1}{A+A_1\cdot TMR} \quad (3)$$

$$D_{01-10} = R_{10} - R_{01} = \frac{[TMR\cdot(TMR+1)\cdot RA](2A_1-A)}{(A+TMR\cdot A_1)[TMR\cdot(A-A_1)+A]} \quad (4)$$

$$D_{10-11} = R_{11} - R_{10} = \frac{TMR\cdot(TMR+1)\cdot RA}{A}\cdot\frac{A-A_1}{TMR\cdot(A-A_1)+A} \quad (5)$$

We calculated the derivatives of $D_{00-01}$, $D_{01-10}$, and $D_{10-11}$ with respect to $A_1$ and have: $\frac{dD_{00-01}}{dA_1} < 0$, $\frac{dD_{10-11}}{dA_1} < 0$, and $\frac{dD_{01-10}}{dA_1} > 0$ when $A_1 \in [A/2, A]$. In other words, $D_{00-01}$ and $D_{10-11}$ monotonically decrease when $A_1$ increases from $A/2$ to $A$ and $D_{01-10}$ monotonically increases in the same range. Also, since $A - A_1 < A_1$ and $TMR \geq 1$, $D_{10-11}$ is always larger than $D_{00-01}$ based on Eq. (3) and (5). Therefore, the optimal design of parallel MLC MTJs happens when $D_{00-01} = D_{01-10}$ or:

$$(TMR+1)(\frac{R_{2L}}{R_{1L}})^2 - \frac{R_{2L}}{R_{1L}} = 2(TMR+1) \quad (6)$$

Here $R_{1L}||R_{2L} = R_{00}$.

In a series MLC MTJ, the optimal MTJ design happens when $D_{00-01} = D_{01-10} = D_{10-11}$, or:

$$R_{1L} = \frac{1}{2}R_{2L} \quad (7)$$

Here $R_{2L}$ is usually the low resistance state of the SLC MTJ with the minimum surface area (say, $A$). The optimal design parameters of a typical parallel MLC MTJ and a typical series MLC MTJ are: $RA = 20\Omega\mu A$, $TMR = 1.2$, The limitation sizes is 45nm×90nm.

### *B. Statistical Analysis of the Readability of MLC MTJs*

All the optimizations in Section III-A are based on the nominal values of the device parameters of MLC MTJs. In this section, we will analyze the impacts of process variations on the readability of MLC STT-RAM cells.

Fig. 3(a) and Fig. 3(b) shows the distributions of the four resistance states in a parallel MLC MTJ and a series MLC MTJ, respectively. Both MTJs are optimized by using the nominal optimization method presented in Section III-A. The standard deviations ($1\sigma$) of $RA$ and $TMR$ are 7% and 9%, respectively, based on the measurement data in [7]. In the nominal optimized parallel MLC MTJ, $\frac{R_1}{R_2} = 1.66$. In the nominal optimized series MLC MTJ, the surface area of the larger MTJ is 64nm×127nm, which corresponds to a low resistance state of $R_{2L} = 2500\Omega$. After the process variations are taken into account, the distributions of the resistance states overlap with each other, resulting in the read errors of the MLC MTJs. Because of the different deviations of every resistance state, the original nominal optimization that maximizes the distinctions between the nominal values of the adjacent resistance states is no longer able to guarantee the minimal overlaps between the adjacent resistance state distributions. A statistical optimization method is required for the minimization of the read error rate of MLC STT-RAM cells.

*1) Optimization of Parallel MLC MTJs:* In our design, we assume the size of the parallel MLC MTJs is the same as the minimum size of the SLC MTJ or 45nm×90nm. The resistances of the two magnetic domains can be adjusted by changing the partition of the free layer. The surface areas of the whole MTJ follows Gaussian distributions

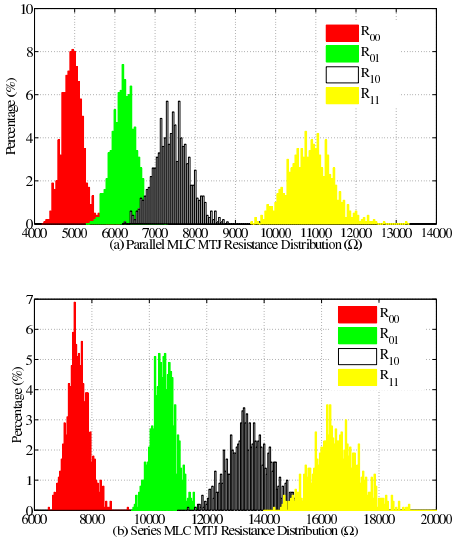


Fig. 3. Four state resistance distributions of (a) Parallel MLC MTJ and (b) Series MLC MTJ, optimized by nominal design method.

and the surface areas of the two magnetic domains follow a joint Gaussian distribution. To sense the four resistance states in a four-level parallel MLC MTJ, three reference resistances, i.e., $R_I$, $R_{II}$, $R_{III}$, are needed. The read error rates of reading $R_{00}$, $R_{01}$, $R_{10}$ and $R_{11}$ can be respectively expressed as:

$$\begin{aligned} P_{e00} &= P(R_{00} > R_I) \\ P_{e01} &= P(R_{01} < R_I) + P(R_{01} > R_{II}) \\ P_{e10} &= P(R_{10} < R_{II}) + P(R_{10} > R_{III}) \\ P_{e11} &= P(R_{11} < R_{III}) \end{aligned} \quad (8)$$

We note that the impacts of the read error rates of each resistance states are not accumulative in MLC STT-RAM designs: For a MLC STT-RAM cell, the highest read error rate is the maximum one of all resistance states, or, $P_e = Max(P_{e00}, P_{e01}, P_{e10}, P_{e11})$. To minimize the $P_{ei}, i = 00, 01, 10, 11$, the $R_I$, $R_{II}$, ideally, $R_{III}$ must be selected at the cross point of the two adjacent distributions. In memory designs, $P_e$ can be used to determine the required error tolerance capability. The read errors due to the MTJ resistance variations can be corrected or tolerated in the design practices by using error correction code (ECC) and design redundancy etc.

In Fig. 3(a), the overlaps of the resistance state distributions of the parallel MLC MTJ generate the read error rates of $P_{e00} = 0.73\%$, $P_{e01} = 6.44\%$, $P_{e10} = 6.05\%$ and $P_{e11} = 0.018\%$. High read error rates happen at $R_{00}$ and $R_{01}$, which are incurred by the large overlaps between these two resistance states.

If we assume that surface area of each magnetic domain follows Gaussian distribution as $A_i \sim N(A_i, \sigma_i)(i = 1, 2)$, the distribution of low resistance $R_{iL}(i = 1, 2)$ can be expressed as:

$$f_{iL}(x_i) = \frac{1}{\sqrt{2\pi}\sigma_i}\frac{RA}{x_i^2}e^{-\frac{(RA/x_i - R_{iL})^2}{2\sigma_i^2}} \quad (9)$$

[7] shows that the $TMR$ ratio also follows Gaussian distribution. We introduce a new variable $z = TMR+1$. Then in our simulations, $z \sim N(2.2, 9\% \times 1.2)$ and its distribution can be expressed as $f_0(z) = \frac{1}{0.108\sqrt{2\pi}}e^{-\frac{(z-2.2)^2}{0.023328}}$. $R_{iH} = z\cdot R_{iL}$.

The read error rate probability of every resistance state of the

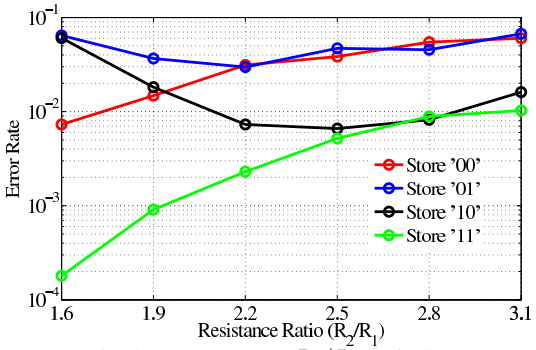


Fig. 4. Error Rate vs. $R_2/R_1$ Ratio Sweep.

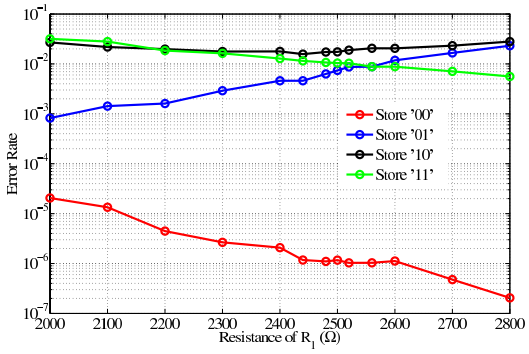


Fig. 5. Error Rate vs. Resistance of Hard Domain Sweep.

parallel MLC MTJ can be further derived from Eq.( 8) as:

$$\begin{aligned}
P(R_{00}) &= \frac{1}{\sqrt{2\pi}(\frac{\sigma_1+\sigma_2}{RA})}\frac{1}{R_{00}^2}\cdot e^{-\frac{(1/R_{00}-(R_{1L}+R_{2L})/RA)^2}{2(\frac{\sigma_1+\sigma_2}{RA})^2}} \\
P(R_{01}) &= \int_{-\infty}^{+\infty} f_{2L}(\frac{RA}{R_{01}}-x)\int_{-\infty}^{+\infty} f_{1L}(z)f_0(\frac{z}{x})\frac{z}{x^2}dzdx\frac{RA}{R_{01}^2} \\
P(R_{10}) &= \int_{-\infty}^{+\infty} f_{1L}(\frac{RA}{R_{10}}-x)\int_{-\infty}^{+\infty} f_{2L}(z)f_0(\frac{z}{x})\frac{z}{x^2}dzdx\frac{RA}{R_{10}^2} \\
P(R_{11}) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{1L}(x)f_0(\frac{x}{y})\frac{x}{y^2}dx \\
&\quad \cdot \int_{-\infty}^{+\infty} f_{2L}(x')f_0(\frac{x'}{R_{11}-y})\frac{x'}{(R_{11}-y)^2}dx'dy
\end{aligned} \tag{10}$$

Fig. 4 depicts the Monte-Carlo simulation results of the read error rate under the different ratios of the nominal resistances of the two magnetic domains ($R_2/R_1$) when the MTJ variations are considered. $P_{e11}$ is always lower than $P_{e00}$ due to the bigger distinction between $R_{10}$ and $R_{11}$ compared to the one between $R_{00}$ and $R_{01}$. Following the increase of $R_2/R_1$ from 1.6, both $P_{e00}$ and $P_{e11}$ increase, indicating the reduced distinction from the adjacent resistance states. However, the increase of $R_2/R_1$ decreases the $P_{e01}$ and $P_{e10}$ by raising the distinction between $R_{01}$ and $R_{10}$. When $R_2/R_1 = 2.2$, the parallel MLC MTJ achieves its lowest maximum read error rate as $P_{e00} = 3.31\%$, $P_{e01} = 2.97\%$, $P_{e10} = 0.73\%$ and $P_{e11} = 0.23\%$. The change of the optimal $R_2/R_1$ ratios in the nominal and statistical optimizations comes from the correlation between the standard deviation and the nominal values of the MTJ resistance state: the higher resistance is, the larger standard deviation of the resistance will be [17].

*2) Optimization of Series MLC MTJs:* In series MLC MTJ, the serially connected SLC MTJs are fabricated separately. The parameters of these two MTJs are partially correlated due to the spatial correlations. The two resistance states of the small SLC MTJ with the minimum size are $R_{2L} = 5000\Omega$ and $R_{2H} = 11000\Omega$, respectively. The distinctions between two adjacent resistance states can be adjusted by changing the surface area of the large SLC MTJ.

Similar to the analysis in Section III-B1, We have:

$$z = R_{iH}/R_{iL}, \tag{11}$$

and

$$g(R_{iH}, R_{iL}) = f_{iL}(R_{iL})\cdot f_0(\frac{R_{iH}}{R_{iL}})\cdot\frac{1}{R_{iL}}. \tag{12}$$

Here $g(R_{iH}, R_{iL})$ is the joint probability density function. Based on the integral of the variable $R_{iL}$, we can obtain the density function of $R_{iH}$ as:

$$\begin{aligned}
f_{iH}(x_i) &= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_i}\frac{RA}{R_{iL}^2}e^{-\frac{(RA/R_{iL}-R_{iL})^2}{2\sigma_i^2}} \\
&\quad \cdot\frac{1}{R_{iL}}\cdot\frac{1}{0.108\sqrt{2\pi}}e^{-\frac{(\frac{R_{iH}}{R_{iL}}-1.2)^2}{0.023328}}dR_{iL}
\end{aligned} \tag{13}$$

The read error rates of the resistance states of the series MLC MTJ are:

$$\begin{aligned}
P(R_{00}) &= \int_{-\infty}^{+\infty} f_{1L}(R_{00}-x)f_{2L}(x)dx \\
P(R_{01}) &= \int_{-\infty}^{+\infty} f_{1H}(R_{01}-x)f_{2L}(x)dx \\
P(R_{10}) &= \int_{-\infty}^{+\infty} f_{1L}(R_{10}-x)f_{2H}(x)dx \\
P(R_{11}) &= \int_{-\infty}^{+\infty} f_{1H}(R_{11}-x)f_{2H}(x)dx
\end{aligned} \tag{14}$$

Fig. 5 shows the read error rates of the four resistance states of the series MLC MTJ when the size of the large SLC MTJ changes. The variation of the large SLC MTJ size is represented by its low resistance state($R_{1L}$). The lowest maximum read error rate happens when $R_{1L} = 2440\Omega$, or the MTJ size is 64.5nm×129nm. It is very close to the result of the nominal optimization method – $R_{1L} = 2500\Omega$, or the MTJ size of 64nm×127nm. The corresponding read error rates of each resistance states are $P_{e00} = 0.000118\%$, $P_{e01} = 0.46\%$, $P_{e10} = 1.57\%$ and $P_{e11} = 1.15\%$. Compare to parallel MLC MTJs, series MLC MTJs demonstrated significantly lower read error rate under the same fabrication conditions. Although the read error rate has not achieved the commercial requirement yet, these results are still very encouraging.

## IV. Writability Analysis of MLC MTJs

In SLC MTJ designs, increasing the switching current density can effectively reduce the MTJ switching time and improve the write error rate of the SLC STT-RAM cell. In MLC MTJ designs, however, increasing the switching current when programming the MTJ to an intermediate resistance state may overwrite the MTJ to the next resistance level. The thermal fluctuations further complicate the situations of MLC MTJ programming by incurring the additional variability of MTJ switching time. In this section, we will discuss the impacts of these variations and the multi-level programming mechanisms on the writability of the MLC MTJs.

### A. Write Mechanism of MLC STT-RAM Cells

The write operation of a MLC STT-RAM cell is much more complex than that of a SLC STT-RAM cells – Both the polarizations and the amplitude of the switching current must be carefully tuned according to the current and the target resistance states.

The write scheme of parallel MLC MTJs has been discussed in [3]; In general, the soft domain can be switched by a small current (density) while the hard domain must be switched by a relatively large current (density). It means that the soft domain can be switched alone but the hard domain switching is always associated with the soft domain switching *if the original magnetization directions of the two domains are the same*. Hence, some resistance state transitions require two switching steps. For example, when a parallel MLC MTJs switches from $R_{00}$ to $R_{10}$, a large current is applied first to switch the MTJ from $R_{00}$ to $R_{11}$. Then a small current is applied to complete the transition from $R_{11}$ to $R_{10}$.

For easy analysis, we assume that the bits of a MLC MTJ from '00' to '11' follow the resistance value from low to high. As summarized in [2], the transitions of the MTJ resistance states can be classified into three types:

1) Soft transition (ST), which switches only the soft domain in a parallel MLC MTJ or the small SLC MTJ in a series MLC MTJ;
2) Hard transition (HT), which switches the both domains in a parallel MLC MTJ or both SLC MTJs in a series MLC MTJ to the same magnetization direction;
3) Two-step transition (TT), which utilizes two steps to switch the MLC MTJ to the target resistance states, i.e., one HT followed by one ST.

### B. Impacts of Thermal Fluctuations

We define the threshold switching current (density) as the minimal current (density) required to switching a MTJ within a switching time. The relationship between the magnetization switching time ($t_w$) and the nominal value of the threshold switching current density ($J_C$) can be divided in three working regions [14]. When $t_w < 10ns$, the reduction of $t_w$ requires the dramatic increase of the $J_C$. Also, due to the asymmetry of MTJ switching, the threshold switching current density of writing '1' is usually larger than that of writing '0' [18].

The thermal fluctuation demonstrates different impacts on the MTJ switching performance in the different working regions: For a low switching current density or a $T_w > 10ns$, the thermal fluctuation is dominated by the thermal component of internal energy; the MTJ switching time follows a Poisson distribution. For a high switching current density or a $T_w < 3ns$, the thermal fluctuation is dominated by the thermally active initial angle of procession; the MTJ switching time follows a Gaussian distribution [4]. The distribution of the MTJ switching time in the middle of these two regions follows a combination of the two distributions. In the write operations of MLC STT-RAM, the two parts of the MLC MTJs, i.e., the two magnetic domains in the parallel MLC MTJ or the two SLC MTJs in the series MLC MTJ, may experience different switching current densities, thermal fluctuations and even different threshold current densities (mainly exist in the parallel MLC MTJs). The MTJ switching could ends up with multiple possible resistance states with different probabilities, as we shall show in following sections.

### C. Write Operations of Parallel MLC MTJs

During the write operations of parallel MLC MTJs, the voltage ($V$) applied to the two terminals of the two magnetic domains are the same. For each domains, the switching current density has:

$$J_i = \frac{V}{R_i \cdot A_i} = \frac{V}{\frac{RA_i}{A_i} \cdot A_i} = \frac{V}{RA_i}, i = 1, 2. \quad (15)$$

It shows that after $V$ is fixed, the switching current density through each domain is uniquely determined by the $RA$ of the domain. Here $RA_i$ = $RA_L$ or $RA_L \cdot (TMR+1)$ for the low- or the high-resistance state, respectively. $RA_L$ is the $RA$ of the low resistance state. As we discussed in Section II-B1, the two magnetic domains of a parallel MLC MTJ have the exactly same RA when they are in the same resistance state. In such a case, the two magnetic domains have the the same current density. However, if the two domains are in the opposite resistance states, the current densities of them will be different.

Fig. 6(a) shows our simulation results of the relationships between the $T_w$ and $J_C$ for the two domains in a typical parallel MTJ. The MTJ parameters are scaled from the measured data of a 90×180nm elliptical MTJ device in [10]. Two domains demonstrate different $J_C$ even under the same $T_w$ due to the different shape anisotropy's etc.

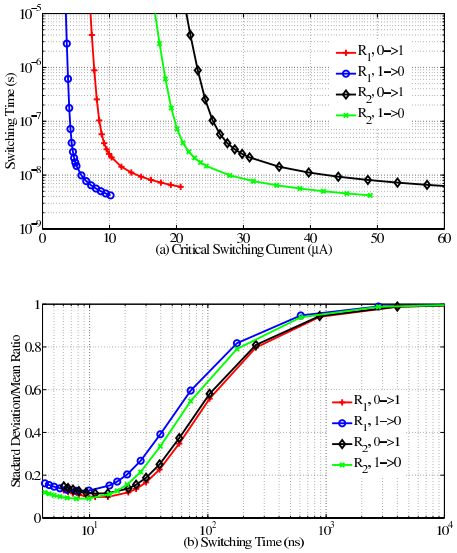


Fig. 6. Switching properties of the two domains for a parallel MLC MTJ. (a) switching time vs. switching current. (b) switching time standard deviation vs. switching current.

The write asymmetry is also observed in the result, i.e., the $J_C$ of '0'→'1' transition of the magnetic domain is always higher than that of '1'→'0' transition for the same $T_w$. The relative deviations of the $T_w$ of the two magnetic domains at the whole working region are shown in Fig. 6(b).

During the write operations of parallel MLC STT-RAM cells, the write current must be applied to switch only the domain(s) that need(s) to be flipped. However, the variability in the magnetization switching of the two domains can introduce write errors. Different from the SLC MTJ where the write error is only incurred by incomplete switching, the writing errors of the parallel MLC MTJ come from either the incomplete switching of the target domains (incomplete write) or overwriting the other domain to an undesired resistance state (overwrite). In a HT transition, only incomplete writes will happen because the write operations require either both domains flip together or only the hard domain flips if the soft domain has already been in the target resistance state. In such a case, increasing the switching current can effectively improve the switching performance of both domains and suppress the write error rate. In a ST transition, the situation can be divided into two scenarios: 1) If the destination resistance state is boundary state, i.e., $R_{00}$ and $R_{11}$, then only incomplete write failures are possible; 2) If the destination resistance state is intermediate state, i.e., $R_{01}$ and $R_{10}$, then both incomplete write and overwrite failures may occur. An appropriate switching current must be selected to achieve a low combined writing error rate. We denote the transitions in 2) as "dependent" transitions and the transitions in 1) and HT transitions as "independent" transitions.

Monte-Carlo simulations are conducted to evaluate the write error rates of the dependent transitions, i.e., 00 → 01 or 11 → 10, as shown in Fig. 7. Here we assume the MTJ switching current is supplied by an adjustable on-chip current source, whose output magnitude has an intrinsic standard deviation of 2% of the nominal value [8]. For a 10ns write pulse width, the optimal switching current for the transitions of '00'→'01' and '11'→'10' are $46.5\mu A$ and $49.9\mu A$, respectively. Fig. 7 also shows the changes of incomplete and overwrite errors over the whole simulated range. When the switching current decreases from the optimal value, the incomplete writes start to dominate the

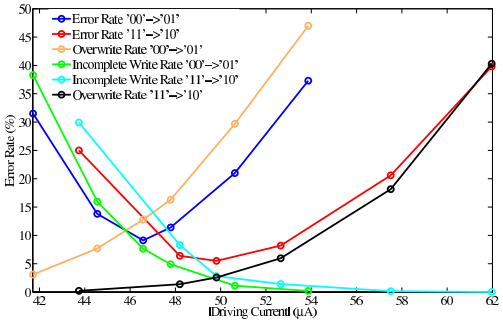


Fig. 7. Writing error rate in parallel MLC STT-RAM cell at $T_w$ = 10ns. Notes: The total error rate is not necessarily equal to the sum of incomplete error and overwrite error, which are the errors overwriting the hard domain or incurring the incomplete soft domain flipping, respectively.

write errors; When the switching current increases from the optimal value, the overwrite errors of the hard domain start to dominate the write errors. Nonetheless, the error rates of the two dependent transitions are still high ( 8.2%), indicating a large overlap area between the threshold switching current distributions of the hard domain and the soft domain.

Fig. 8 shows the write error rates of the dependent transitions of the parallel MLC MTJ at different switching currents when $T_w$ = 3ns, 10ns, and 100ns, respectively. The lowest write error rate is achieved at $T_w$ = 3ns. It is because that when $T_w$ reduces, the required MTJ switching current increases. The impact of the thermal fluctuations on the MTJ switching is suppressed and the distributions of the $T_w$ are compressed. This fact indicates that the parallel MLC MTJ better work at a fast working region to minimize the write error rate.

We can also map the uncertainties in the switching time of the parallel MLC MTJ under the fixed switching current into the distributions of the required switching currents for fixed switching time. Fig. 9(a) shows the distributions of the threshold switching current of the dependent transitions for the parallel MLC MTJ at a 10ns write pulse width. The distributions of the MTJ write current supplied by the on-chip current source are also depicted. Take the transition of '00'→'11' as an example, a write current is selected between the threshold current distributions of the transitions of '00'→'01' and '00'→'11'. The two types of write errors, including incomplete write and overwrite, are represented by the overlap between the distributions of the write current and the threshold switching current of '00'→'01' and the overlap between the distributions of the write current and the threshold switching current '00'→'11', respectively. Fig. 9(b) shows the distributions of the threshold switching current of the independent transitions for the parallel MLC MTJ at a 10ns write pulse width. Since only the target magnetic domain will flip

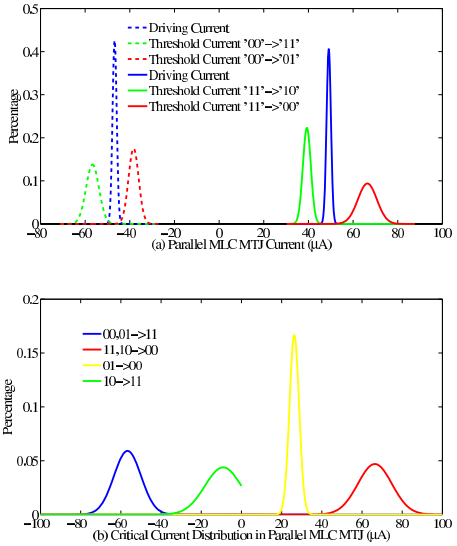


Fig. 9. Threshold current distributions of resistance state trasitions for the parallel MLC MTJ.(a) Dependent transitions. (b) Independent transitions.

during the independent transitions, a sufficiently large write current can be always applied to suppress the incomplete write errors without incurring any overwrite errors.

Similar to the distributions of the MTJ switching time, the distributions of the threshold switching current of the parallel MLC MTJ are also dependent on the working regions of the MTJ. After the distributions of the switching current of the resistance state transitions are obtained, the optimal write current can be derived as Fig. 9(a).

### *D. Write Operations of Series MLC MTJs*

In a series MLC MTJ, the magnitudes of the currents passing through the two SLC MTJs are the same. However, the applied current densities on the two SLC MTJs are different and determined by the different surface areas of them. In Section III-B2, the analysis on the read reliability of the series MLC MTJs shows that the optimal surface area ratio between the two MLC MTJs is around 2, or 45nm×90nm and 64.5nm×129nm at 45nm technology node. In our simulations, we also assume the two SLC MTJs maintain the same aspect ratios and were fabricated under the same conditions. Thus, they have the same switching properties, i.e., the same relationships between threshold switching current density and the switching time. Again, the switching current density on each SLC MTJ is controlled by the on-chip write current source.

Fig. 10 shows the write error rates of the dependent transitions of the series MLC MTJ under different switching currents for a 10ns write pulse width. The optimal switching current for the

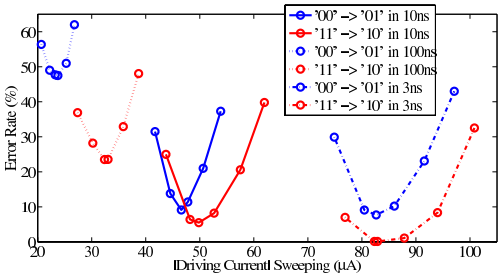


Fig. 8. Writing error rate in a parallel MLC STT-RAM cell at different $T_w$

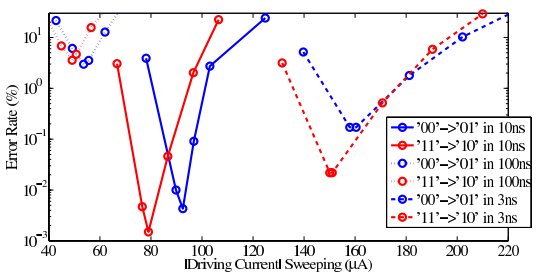


Fig. 10. Writing error rate in a series MLC STT-RAM cell at different $T_w$

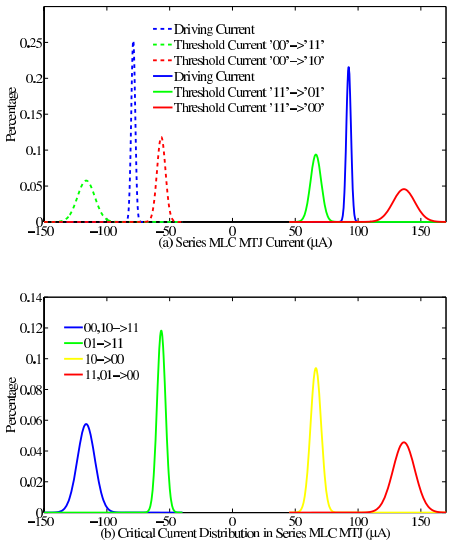


Fig. 11. Threshold current distributions of resistance state transitions for the series MLC MTJ.(a) Dependent transitions. (b) Independent transitions.

transitions of '00'→'10' and '11'→'01' are $79.0\mu A$ and $92.5\mu A$, respectively. Compared to parallel MLC MTJs, the write error rates of the dependent transitions are significantly reduced: the minimum write error rates of the transitions of '00'→'10' and '11'→'01' are only 0.0015% and 0.0043%, respectively. The improvement of the write reliability is because of the larger distinction between the threshold switching current distributions of the dependent transition and the adjacent resistance state transition, as shown in Fig. 11(a). For comparison purpose, the results of the independent resistance state transitions are shown in Fig. 11(b).

Fig. 10 also shows the write error rates of the dependent transitions of the serial MLC MTJ at different switching currents when $T_w$ = 3ns and 100ns, respectively. Similar dependency of the write error rate on the MTJ working region is observed. Interestingly, the minimum write error rate occurs when $T_w$ = 10ns, since the standard deviation/mean ratio reaches its minimum value (see Fig. 9(b)). Compared to parallel MLC MTJs, series MLC MTJs demonstrate much higher write reliability at the same technology node,, while requiring slightly larger switching current and higher write energy consumption.

## V. Conclusion

In this work, we quantitatively analyze the impacts of the process variations and the thermal fluctuations on the performance and reliability of both parallel and series multi-level cell (MLC) STT-RAM cell designs. Compared to conventional single-level cell (SLC) STT-RAM designs, the different storage mechanism of the MLC STT-RAM results in very unique operation failure models and reliability optimization concerns. Our results showed that the resistance states of both MLC STT-RAM cell structures must be optimized in the designs to minimize the read errors. The magnitude of the write current must be also carefully selected to suppress both incomplete write and overwrite failures.

Our simulation results show that series MLC STT-RAM demonstrates much better reliability in both write and read operations compared to the parallel MLC STT-RAM under the same fabrication conditions. Also, as expected, the readability is still the biggest concern in both MLC STT-RAM designs.

## Acknowledgment

This work is supported by NSF awards: CNS-1116171 and CCF-1217947.

## References

[1] Y. Chen, X. Wang, H. Li, H. Xi, Y. Yan, and W. Zhu. "Design Margin Exploration of Spin-Transfer Torque RAM (STT-RAM) in Scaled Technologies". *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 18(12):1724 –1734, Dec. 2010.

[2] Y. Chen, X. Wang, W. Zhu, H. Li, Z. Sun, G. Sun, and Y. Xie. "Access Scheme of Multi-Level Cell Spin-Transfer Torque Random Access Memory and Its Optimization". In *53rd IEEE International Midwest Symposium on Circuits and Systems*, pages 1109 –1112, Aug. 2010.

[3] Y. Chen, W.-F. Wong, H. Li, and C.-K. Koh. "Processor Caches Built Using Multi-Level Spin-Transfer Torque RAM Cells". In *International Symposium on Low Power Electronics and Design 2011*, pages 73 –78, Aug. 2011.

[4] Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L. Wang, and Y. Huai. "Spin-transfer Torque Switching in Magnetic Tunnel Junctions and sSpin-transfer Torque Random Access Memory". *Journal of Physics: Condensed Matter*, 19:165209, 2007.

[5] Y. Huai. "Spin-Transfer Torque MRAM (STT-MRAM): Challenges and Prospects". *AAPPS Bulletin*, 18(6):33–40, 2008.

[6] S. Ikeda, K. Miura, H. Yamamoto, K. Mizunuma, H. Gan, M. Endo, S. Kanai, J. Hayakawa, and H. Matsukura, F. Ohno. "A Perpendicular-anisotropy CoFeB-MgO Magnetic Tunnel Junction". *Nature Materials*, 9(9):721–724, 2010.

[7] T. Ishigaki, T. Kawahara, R. Takemura, K. Ono, K. Ito, H. Matsuoka, and H. Ohno. "A Multi-level-cell Spin-transfer Torque Memory with Series-stacked Magnetotunnel Junctions". In *Symposium on VLSI Technology*, pages 47 –48, Jun. 2010.

[8] M.-Y. Kim, H. Lee, and C. Kim. "PVT Variation Tolerant Current Source With On-Chip Digital Self-Calibration". *IEEE Transactions on Very Large Scale Integration Systems*, 20(4):737 –741, Apr. 2012.

[9] J. Li, H. Liu, S. Salahuddin, and K. Roy. "Variation-tolerant Spin-Torque Transfer (STT) MRAM Array for Yield Enhancement". In *IEEE Custom Integrated Circuits Conference*, pages 193 –196, Sep. 2008.

[10] X. Lou, Z. Gao, D. V. Dimitrov, and M. X. Tang. Demonstration of multilevel cell spin transfer switching in mgo magnetic tunnel junctions. *Applied Physics Letters*, 93(24):242502, 2008.

[11] N. Mojumder, C. Augustine, D. Nikonov, and K. Roy. "Effect of Quantum Confinement on Spin Transport and Magnetization Dynamics in Dual Barrier Spin Transfer Torque Magnetic Tunnel Junctions". *Journal of Applied Physics*, 108(10):104306–104306–12, Nov. 2010.

[12] K. Munira, W. Soffa, and A. Ghosh. "Comparative Material Issues for Fast Reliable Switching in STT-RAMs". In *11th IEEE Conference on Nanotechnology*, pages 1403 –1408, Aug. 2011.

[13] A. Nigam, C. W. Smullen, IV, V. Mohan, E. Chen, S. Gurumurthi, and M. R. Stan. "Delivering on the Promise of Universal Memory for Spin-Transfer Torque RAM (STT-RAM)". In *Proceedings of the 17th IEEE/ACM international symposium on Low-power electronics and design*, ISLPED '11, pages 121–126, Piscataway, NJ, USA, 2011. IEEE Press.

[14] A. Raychowdhury, D. Somasekhar, T. Karnik, and V. De. "Design Space and Scalability Exploration of 1T-1STT MTJ Memory Arrays in the Presence of Variability and Disturbances". In *IEEE International Conference on Electron Devices Meeting*, pages 1–4, Dec. 2009.

[15] R. Sbiaa, R. Law, S. Y. H. Lua, E. L. Tan, T. Tahmasebi, C. C. Wang, and S. N. Piramanayagam. "Spin Transfer Torque Switching for Multi-bit Per Cell Magnetic Memory with Perpendicular Anisotropy". *Applied Physics Letters*, 99(9):092506 –092506–3, Aug. 2011.

[16] J. Z. Sun. "Spin-current Interaction with A Monodomain Magnetic Body: A Model Study". *Phys. Rev. B*, 62:570–578, Jul 2000.

[17] Z. Sun, H. Li, Y. Chen, and X. Wang. "Variation Tolerant Sensing Scheme of Spin-Transfer Torque Memory for Yield Improvement". In *IEEE/ACM International Conference on Computer-Aided Design*, pages 432 –437, Nov. 2010.

[18] Y. Zhang, Y. Li, A.K.Jones, X. Wang, and Y. Chen. "Asymmetry of MTJ Switching and Its Implication to the STT-RAM Designs". *Design Automation and Test in Europe*, Mar. 2012.

[19] W. Zhao and Y. Cao. "New Generation of Predictive Technology Model for Sub-45 nm Early Design Exploration". *IEEE Transactions on Electron Devices*, 53(11):2816 –2823, Nov. 2006.