NVSim-VX^s: An Improved NVSim for Variation Aware STT-RAM Simulation

Enes Eken¹, Linghao Song¹, Ismail Bayram¹, Cong Xu², Wujie Wen³, Yuan Xie⁴, Yiran Chen¹

¹Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA

² Hewlett Packard Labs, Palo Alto, CA 94304, USA

³Department of Electrical and Computer Engineering, Florida International University, Miami, FL, 33174, USA

⁴Department of Electrical and Computer Engineering, University of California at Santa Barbara, CA, 93106, USA

{ene4, lis75, isb11, yic52}@pitt.edu, cong.xu@hpe.com, wwen@fiu.edu, yuanxie@ece.ucsb.edu

Abstract—Spin-transfer torque random access memory (STT-RAM) recently received significant attentions for its promising characteristics in cache and memory applications. As an early-stage modeling tool, NVSim has been widely adopted for simulations of emerging nonvolatile memory technologies in computer architecture research, including STT-RAM, ReRAM, PCM, etc. In this work, we introduce a new member of NVSim family – NVSim-VX^s, which enables statistical simulation of STT-RAM for write performance, errors, and energy consumption. This enhanced model takes into account the impacts of parametric variabilities of CMOS and MTJ devices and the chip operating temperature. It is also calibrated with Monte-Carlo Simulations based on macro-magnetic and SPICE models, covering five technology nodes between 22nm and 90nm. NVSim-VX^s strongly supports the fast-growing needs of STT-RAM research on reliability analysis and enhancement, announcing the next important stage of NVSim development.

I. INTRODUCTION

“Post-silicon” devices have received increasing attentions in solid-state device and circuit society due to the concerns on continuous scaling of conventional CMOS technology. The high leakage power and significantly degraded reliability of mainstream memory technologies [12] inspired the popular research on emerging memory technologies: *spin-transfer torque random access memory* (STT-RAM), *resistive memory* (ReRAM), *phase-change memory* (PCM) [3], etc. In particular, STT-RAM demonstrates many characteristics that are of importance to on-chip cache and memory applications, such as high integration density, zero standby power, nanosecond access time, and excellent CMOS-compatibility [7].

It is known that write error is the major reliability issue in STT-RAM operations. Compared with conventional memory technologies, simulating a STT-RAM cell is very challenging because it requires understandings of both CMOS and magnetic devices. In [2], Chen *et al.* proposed the first combined magnetic and SPICE simulation framework to evaluate the write performance and energy of STT-RAM cells by considering the interaction between transistor and magnetic tunneling junction (MTJ) devices. Besides parametric variabilities that exist in conventional memory cells, thermal-induced switching randomness also significantly affects write operations of STT-RAM cells. Performing statistical analysis on the write reliability, hence, requires very costly and entangled Monte-Carlo simulations on both types of devices.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

DAC '16, June 05–09, 2016, Austin, TX, USA

Copyright © 2016 ACM. ISBN 978-1-4503-4236-0/16/06 \$15.00

DOI: <http://dx.doi.org/10.1145/2897937.2898053>

Block-level STT-RAM models have been also developed to fulfill the need in architectural analysis. Arcaro *et al.* integrated a STT-RAM model into CACTI [1] – a tool was originally used for conventional memory modeling and design [8]. Wu *et al.* presented an architecture-level simulation framework of the advanced perpendicular STT-RAM in [11]. Dong *et al.* released the most widely used STT-RAM block-level model, namely, NVSim [4]. However, none of the above models are able to simulate the impacts of CMOS or MTJ variations and consequently, the write errors of STT-RAM.

In this work, we introduce a new member of NVSim family – NVSim-VX^s, which enables statistical simulation of STT-RAM for write performance, errors, and energy consumption. Besides the parametric variabilities of both CMOS and MTJ devices, this enhanced model also takes into account chip operating temperature, which significantly affects the write reliability of STT-RAM. As the first systematic model to simulate the entangled relationships between different design parameters and metrics of STT-RAM at block-level, the major novelty we introduce to STT-RAM research can be summarized as follows:

- We derive statistical approximations of STT-RAM design metrics, e.g., switching time and energy consumption, and generate the corresponding compact models for fast statistical analysis;
- We implement the STT-RAM model based on the switching pattern of input bits, which have been proved as the major factor affecting the statistical STT-RAM design metrics;
- We develop the models of both perpendicular and in-plane STT-RAMs to support the scaling of STT-RAM technologies.

The first release of NVSim-VX^s supports 5 technology nodes {22, 32, 45, 65, 90}nm, and driving NMOS transistor size between 2–5× the minimum feature size. It also supports operating temperature between 300K to 375K. The model has been thoroughly calibrated with the Monte-Carlo simulations based on macro-magnetic and SPICE models to ensure the accuracy.

The rest of this paper is organized as follows: Section II presents the basics of STT-RAM and NVSim; Section III introduces the statistical compact models that are developed for NVSim-VX^s; Section IV discusses the block-level simulations of NVSim-VX^s and the impact of input bit switching patterns; Section V concludes our work.

II. PRELIMINARY

A. Basics of STT-RAM

In a STT-RAM cell, the data is stored as the resistance of a MTJ device, as shown in Fig. 1. The MTJ resistance is determined by the magnetization directions of the two ferromagnetic layers, i.e., in parallel (low-resistance) or anti-parallel (high-resistance). The magnetization of one of the ferromagnetic layers (reference layer) is fixed while that of the other ferromagnetic layer (free layer) can be switched by applying a write current with proper polarization. The

Fig. 1. STT-RAM basics. (a) In-plane MTJ. (b) Perpendicular MTJ. (c) 1T1J cell structure.

magnetization of the ferromagnetic layers can be either in parallel or perpendicular to the surface of the MTJ, namely, in-plane or perpendicular MTJ, as shown in Fig. 1(a) and (b), respectively. Fig. 1(c) shows a popular “1T1J” STT-RAM cell design where a NMOS transistor supplies the write and read current to the MTJ. The switching process of the MTJ is greatly affected by the amplitude of the current passing through it, which varies with the parametric variations of the MTJ and transistor. It is also affected by the thermal-induced fluctuations of magnetization precession. As pointed out by the prior art [13], this asymmetric structure results in a very unreliable ‘0(low)→‘1(high)’ switching, i.e., with much longer switching time and wider distribution than the other switching direction.

The magnetization switching of the MTJ under a write current can be modeled by *Landau-Lifshitz-Gilbert* (LLG) equation as [9]:

$$\frac{dm_f}{dt} + \alpha(m_f \times \frac{dm_f}{dt}) = \frac{1}{2}\gamma H_k \Sigma_{i=1}^4 \left(\frac{\Gamma_i}{l_t K} \right). \quad (1)$$

Here, M_s is free layer magnetization saturation and m_f represents the unit vector of the free layer. α is the Gilbert damping constant; γ is the gyromagnetic ratio; H_k is the Stoner-Wohlfarth switching field; l_t is the free layer thickness. The four torque (Γ) terms represent the factors that affecting the m_f dynamics – the uniaxial anisotropy, the easy-plane anisotropy, the Langevin random thermal field, and the spin torque term from the applied current. The parameters adopted in our macro-magnetic simulations are summarized in TABLE I [5].

B. Basics of NVSim

NVSim is a widely used open source simulation framework for circuit-level modeling of emerging nonvolatile memories like STT-RAM, ReRAM, PCM, etc [4]. NVSim can extract the memory design metrics, i.e., read/write latency, read/write energy consumption, area, etc. under the given design constraints, or optimize the design parameters.

To simulate an STT-RAM design in NVSim, users are expected to specify the write current value and switching time for both SET and RESET operations. However, obtaining correct values of these parameters requires running macro-magnetic models, which is not

TABLE I
DEVICE AND CIRCUIT SIMULATION PARAMETERS

Device Level	Parameter	Symbol	Value	Unit
	Mag. Saturation	M_s	230	kA/m
	Uniaxial Anisotropy	H_k	200	Oe
	Gilbert Constant	α	0.01	
	Free Layer Thickness	l_t	1	nm
Circuit Level	Parameter	Mean	Std. Dev	
	Channel length	$L = 22 \sim 90nm$	$\sigma_L = 0.05L$	
	Channel width	$W = 2L \sim 5L$	$\sigma_W = 0.05L$	
	Free layer volume	$V = L \times 2L \times 1nm^3$	$\sigma_V = 0.05V$	
	Resistance low	$R_L = 1000\Omega$	$\sigma_{R_L} = 0.05R_L$	
	Resistance high	$R_H = 2000\Omega$	$\sigma_{R_H} = 0.05R_H$

Fig. 2. Framework of the NVSim-VXs.

supported in the current version of NVSim based on pure circuit-level simulation. Moreover, the current version of NVSim supports neither the statistical analysis of STT-RAM, e.g., the variations of write performance (errors) and energy consumption, nor the impact of operating temperature.

III. NVSIM-VX^s FRAMEWORK

Fig. 2 presents the framework of our proposed NVSim-VX^s, which includes three important new features that are not supported by the existing deterministic STT-RAM simulators: the temperature-aware statistical switching time model, the statistical energy consumption model, and the write error rate model. Compared to the current version of NVSIM, NVSim-VX^s possesses a more flexible and user-friendly interface to facilitate its probabilistic design philosophy, i.e. allowing users to set circuit and architecture parameters as the inputs and obtain cell/block level statistical design metrics from the outputs. Furthermore, the important switching pattern (i.e. number of ‘0’→‘1’ or ‘1’→‘0’ flipping’s in write) dependent energy and reliability analysis is also enabled at block level.

A. Temperature-Aware Statistical STT-RAM Switching Time Model

The MTJ switching time variation is mainly generated from the following two torque terms: the Langevin random field and the spin torque, as suggested by Eq. (1). In specific, the randomness sources of the Langevin random field are the variations of MTJ surface area and the thickness of free layer while the spin torque is generated by the driving current, which is affected by process variations of both NMOS transistor and MTJ device [14]. Note that the above two

Fig. 3. Detailed flow for circuit-level and cell-level simulations in NVSim-VX^s.

Fig. 4. Simulated switching time distribution v.s. Log-normal distribution (45nm technology node, 90nm transistor width, 4 different temperatures): a) 300K b) 325K c) 350K d) 375K.

terms are also significantly affected by the fluctuation of operating temperature.

To minimize the costly hybrid CMOS-Magnetic simulations required to capture all the parametric variabilities, our temperature-aware statistical STT-RAM switching time model is derived and simplified from extensive LLG model-based Monte-Carlo simulations, which cover 2 MTJ resistance switching directions, 5 technology nodes, 7 different transistor widths, and 16 different temperature points, as illustrated in Fig. 3. In the first step, sensitivity analysis is conducted at different temperatures to characterize the driving current distributions [10]. Variations of the transistor channel length, the transistor width, and the MTJ resistance are also taken into considerations. Simultaneously considering all variability parameters can reduce the computation complexity from $O(N^k)$ to $O(N)$, where k is the number of variability parameters and N is the number of samples for each parameter. In the second step, we integrate both driving current distributions and the Langevin random field into LLG equation under different temperatures to obtain the temperature-aware switching time distributions. Finally, a fast and compact timing model that directly links the switching time variation (i.e., mean and standard deviation) to temperature and driving current can be achieved.

Based on the device parameters and simulation setup summarized in TABLE I, we performed Monte-Carlo simulations to obtain switching time distributions of '0'→'1' switching at 4 different temperatures for an STT-RAM cell with a $w = 2L$ NMOS transistor width at technology node $L = 45nm$, as depicted in Fig. 4. All the simulated switching time results are in excellent agreement with the Log-normal distributions at the concerned temperatures. As the temperature increases from 300K to 375K, the distribution of the MTJ switching time becomes broader, indicating the increased impact of temperature on MTJ switching and hence the STT-RAM cell write reliability. As we shall show next, the corresponding mean

Fig. 5. The standard deviation of switching time distributions for 7 transistor widths at 45nm technology node.

Fig. 6. The mean of switching time distributions for 7 transistor widths at 45nm technology node.

(μ) and standard deviation (σ) of the Log-normal MTJ switching time distribution can be directly linked with the driving current and temperature using our model.

Our further investigation suggests a linear approximation of the relationship between the μ/σ and temperature. This linear approximation of the temperature dependency of μ and σ of the MTJ switching time can be expressed by:

$$\begin{aligned}\sigma(w) &= m_\sigma(w) * T_n + \sigma_0(w), \\ \mu(w) &= m_\mu(w) * T_n + \mu_0(w).\end{aligned}\quad (2)$$

Here $m_\sigma(w)$ and $m_\mu(w)$ are the coefficient representing the temperature dependency of μ and σ at transistor width w . T_n is the normalized temperature. $\sigma_0(w)$ and $\mu_0(w)$ are the initial values of $\sigma(w)$ and $\mu(w)$, respectively, at $T_n = 0$.

Fig. 5 and 6 show the results of the linear approximation of σ and μ for 7 different transistor widths ($2L \sim 5L$, $L = 45nm$) at different temperatures (300K $\sim$ 375K), respectively. For comparison purpose, the results of Monte-Carlo simulations are also presented. It can be observed that our linear model provides very accurate approximation of the Monte-Carlo simulation results at the whole covered ranges of transistor widths and temperatures. As transistor width increases, both the temperature dependency and the initial values of σ and μ monotonically decreases, implying a less sensitivity to the temperature change and improved thermal robustness.

For a specific transistor width w , the temperature dependency and the initial values of σ and μ ($m_\sigma(w)$, $m_\mu(w)$) and ($\sigma_0(w)$, $\mu_0(w)$), are the functions of the MTJ driving current I as:

$$\begin{aligned}m_\sigma(w) &= a_{m_\sigma} * e^{b_{m_\sigma} I}, \\ m_\mu(w) &= a_{m_\mu} * e^{b_{m_\mu} I}.\end{aligned}\quad (3)$$

Fig. 7. The simulated results of our model and Monte-Carlo simulations for key parameters of σ and μ : (a) $m_\sigma(w)$ v.s. I ; (b) $\sigma_0(w)$ v.s. I ; (c) $m_\mu(w)$ v.s. I ; (d) $\mu_0(w)$ v.s. I .

Fig. 8. Overview of the statistical energy consumption characterization flow.

and

$$\begin{aligned}\sigma_0(w) &= a_{c\sigma} * e^{b_{c\sigma} I} + c_{c\sigma} * e^{d_{c\sigma} I}, \\ \mu_0(w) &= a_{c\mu} * e^{b_{c\mu} I} + c_{c\mu} * e^{d_{c\mu} I}.\end{aligned}\quad (4)$$

Here, the driving current I is determined by the NMOS transistor width w at different technology node. a_i, b_i, c_j and d_j are technology-dependent fitting parameters where $i = m\sigma, m\mu, c\sigma, c\mu$, $j = c\sigma, c\mu$. Fig. 7 depicts the simulated relationship between $m_\sigma(w)$, $m_\mu(w)$, $\sigma_0(w)$, $\mu_0(w)$ v.s. I based on our model in Eq. (3) and (4). The results includes the data at 7 different transistor widths (i.e., $2L \sim 5L$, $L = 45nm$). To validate our model, the Monte-Carlo simulation results are also included. The results show that our model matches the Monte-Carlo simulations very well in all the simulated cases.

By substituting Eq. (3) and (4) into Eq. (2), the MTJ switching time distributions can be expressed by:

$$\begin{aligned}\sigma(w) &= (a_{m\sigma} * e^{b_{m\sigma} I}) * T_n + (a_{c\sigma} * e^{b_{c\sigma} I} + c_{c\sigma} * e^{d_{c\sigma} I}), \\ \mu(w) &= (a_{m\mu} * e^{b_{m\mu} I}) * T_n + (a_{c\mu} * e^{b_{c\mu} I} + c_{c\mu} * e^{d_{c\mu} I}).\end{aligned}\quad (5)$$

Although the above illustrated examples are based on 45nm technology, our developed temperature-aware statistical STT-RAM switching time model is capable to capture the switching time variations for different transistor sizes ($2L \sim 5L$), different temperatures ($300K \sim 375K$) at different technology nodes ($22 \sim 90nm$), showing its adaptivity and scalability in advanced technology nodes.

B. STT-RAM Statistical Energy Modeling

In current version of NVSim, the write energy of STT-RAM is deterministically modeled without considering any fluctuations in write operations. The cell-level write energy is directly extracted from the given SET/RESET current, applied voltage, and write time. However, as previously discussed, the switching time of the MTJ in each STT-RAM cell varies with the parametric variations of the MTJ

Fig. 9. The write energy consumption distribution comparison of our model v.s. Monte-Carlo at 45nm technology node, 90nm transistor width, 60ns write pulse width.

and the NMOS transistor, and is influenced by thermal fluctuations. The MTJ resistance states, which affect the write current through the device, also follows some distributions. Hence, in NVSim-VX⁸, we characterize the statistical STT-RAM write energy consumption.

The energy consumption of an STT-RAM cell during a write operation, i.e., '0' → '1' switching, can be calculated using Joule's first law as:

$$E = (I_H * \tau_{sw} + I_L * (\tau_{wt} - \tau_{sw})) * V. \quad (6)$$

Here I_H is the initial high driving current at low resistance (R_L) state and I_L is the low post-switching current after the resistance state switches to high resistance (R_H). τ_{sw} is the actual MTJ switching time and τ_{wt} is the writing period (write time) for which the programming voltage (V) is applied. Note that we ignore the oscillation of the driving current generated by the magnetic precession. Also, both I_H and τ_{sw} are correlated, and subjected to the variations from CMOS/MTJ device and thermal fluctuations.

Fig. 8 depicts the overview of our proposed STT-RAM statistical write energy model, including the following five steps:

- 1) **Derive current information:** Obtain the driving current statistical information by conducting sensitivity analysis, as is discussed in statistical STT-RAM switching time model;
- 2) **Generate current sample:** Generate the driving current samples over the dual exponential current distribution and the statistical information [10];
- 3) **Obtain switching time distribution:** Send the driving current samples and the temperature to the temperature-aware statistical switching time model developed in Section III-A and generate different switching time distributions for each sample;
- 4) **Calculate statistical energy:** Calculate the energy by doing integral over user-specified write time at each driving current sample and switching time distribution pair as below:

$$E_i = \int_0^{\tau_{wt}} (I_{Hi} * t + I_L * (\tau_{wt} - t)) * V * f_i(t) * dt. \quad (7)$$

Here I_{Hi} denotes the i_{th} sample of the I_H obtained from dual exponential model, $f_i(t)$ is the probability density function of corresponding switching time distribution of the current sample. I_L is the low post-switching current.

- 5) **Dump energy distribution:** Calculate the mean (μ_E) and standard deviation (σ_E) of the write energy consumption as:

$$\mu_E = \frac{\sum_{i=1}^n E_i * f_i}{\sum_{i=1}^n f_i} \quad \text{and} \quad \sigma_E = \sqrt{\frac{\sum_{i=1}^n (E_i - \mu_E)^2 * f_i}{\sum_{i=1}^n f_i}}. \quad (8)$$

Here f_i is the number of occurrences of energy value E_i for the

Fig. 10. The mean of the energy consumption at 10ns write pulse width for 7 different transistor widths at 45nm technology node.

Fig. 11. The mean of energy consumption at different write pulse widths for $w = 2.5L$ transistor width ($L = 45nm$).

i current sample.

Our simulations show that the write energy consumption roughly follows a Gaussian distribution whose mean and standard deviation can be obtained from step 5. Fig. 9 shows the write energy distributions of both MTJ switching directions obtained by our model and Monte-Carlo simulations at $\tau_{wt} = 60ns$. The temperature is 350K and the transistor width $W = 2L$, $L = 45nm$. The results show that our model approximates the Monte-Carlo simulations very closely. Fig 10 compares the mean value of the write energy consumption of STT-RAM cell designs with various transistor widths ($2L \sim 5L$, $L = 45nm$) under different temperatures at ‘1’→‘0’ switching. Again, our model can always provide the results very close to that of the Monte-Carlo simulations with the simulated transistor sizes. As the temperature increases, energy consumption reduces almost linearly at large transistor widths (i.e. $3.5L \sim 5L$) because of the narrow distribution of the MTJ switching time (τ_{sw}). On the contrary, the changing rate of the energy consumption with temperature becomes nonlinear when the transistor width is small, indicating a high sensitivity to temperature change. It again proves that a large access transistor can help reducing the thermal-induced performance variations of STT-RAM. Fig. 11 shows the energy consumption of $W = 2.5L$ at three different τ_{wt} . Reducing the τ_{wt} slightly degrades the linearity of the temperature dependency of the energy consumption.

C. STT-RAM Write Error Rate

An STT-RAM write failure happens if the MTJ switching cannot complete within the applied write pulse width (or the write time τ_{wt}). Following technology scaling, write reliability emerges as one of the main challenges in STT-RAM designs.

Traditionally, calibrating the write error rate of an STT-RAM cell requires two runs of Monte-Carlo simulations and one sample/distribution processing: Firstly, circuit-level simulations are conducted to get the STT-RAM switching current distribution by considering all parametric variabilities; Secondly, the current distribution

Fig. 12. Simulated write error rate v.s. temperature from our model and Monte-Carlo for 22nm perpendicular MTJ and 45nm in-plane MTJ (10ns write pulse width, ‘0’→‘1’ switching).

Fig. 13. Simulated write error rate v.s. temperature from our model and Monte-Carlo at different write time (transistor width $w = 2.5L$, ‘0’→‘1’ switching).

is sent to a macro-magnetic model and the second-round Monte-Carlo simulations are performed to obtain the STT-RAM switching time distribution; Finally, the generated switching time samples or distribution must be compared with the given write time to calculate the write failure.

In NVSim-VX^s, the samples of the STT-RAM switching time can be obtained from the embedded temperature-aware statistical switching timing model without running the two costly Monte-Carlo simulations. However, the normal write error rate of a STT-RAM cell is so low that a large number of samples of the STT-RAM switching time still need to be generated to calculate the error rate. In our work, we introduce the mixture importance sampling technique in NVSim-VX^s to reduce the write error rate calculation cost as [6]:

$$E_{p(x)}[\theta] = E_{g(x)}[\theta * \frac{p(x)}{g(x)}]. \quad (9)$$

Here $p(x)$ denotes the switching time probability density function and $g(x)$ is the distorted sampling function defined as;

$$g(x) = \lambda_1 p(x) + \lambda_2 U(x) + (1 - \lambda_1 - \lambda_2) p(x - \mu_s), \quad (10)$$

where $0 \leq \lambda_1 + \lambda_2 < 1$. $U(x)$ is the uniform pdf; μ_s is the shifted center and is chosen experimentally.

Fig. 12 compares the results of write error rate of an STT-RAM obtained from NVSim-VX^s and Monte-Carlo simulations, respectively, for a 22nm perpendicular MTJ and an 45nm in-plane MTJ at ‘0’→‘1’ switching. Our model can achieve good accuracy at the 4 simulated transistor widths and precisely describe the changing trend of the write error rate with the temperature. Interestingly, Fig. 12 shows that the write error rate of the 22nm perpendicular MTJ always outperforms the one of 45nm in-plane MTJ at the similar relative transistor sizes ($2L \sim 3.5L$) and the same temperature. This result validates the conclusion that perpendicular STT-RAM is more promising at scaled technology node, i.e., below 45nm.

Fig. 13 shows the simulated write error rate over different temperatures for the 45nm in-plane MTJ with different write pulse widths. Increasing the write time can greatly reduce the write error rate at low temperatures; However, limited improvement is observed at higher temperatures.

IV. BLOCK LEVEL EXTENSION FOR NVSIM-VX^s

To make the NVSim-VX^s suitable for architecture-level simulation, we extended our cell level model to block level. NVSim-VX^s is capable to calculate the block level write energy or error rate more precisely by taking the switching pattern into consideration. To the best of our knowledge, this is the first time that such an important feature is integrated into nonvolatile memory simulators.

A. Block Level Energy Consumption

Write energy consumption of an STT-RAM cell is distinctive at two switching directions. The block-level energy estimation will be more accurate if the users can provide the switching patterns of the

Fig. 14. The mean and standard deviation of block-level energy consumption for 3 different flipping combinations at 10ns write time and $w = 3L$ transistor width, $L = 45nm$.

accessed block by considering the difference between the incoming data and the stored data. There are four possibilities of the bit switching: ‘0’→‘1’, ‘1’→‘0’, ‘0’→‘0’, and ‘1’→‘1’. For a switching of ‘ i ’→‘ j ’, $i, j = 0$ or 1, the mean and the standard deviation of total energy consumption spent on ‘ i ’→‘ j ’ switchings in a memory write can be expressed as:

$$\mu_{eT,ij} = N_{F,ij} * \mu_{e,ij} \quad \text{and} \quad \sigma_{eT,ij} = N_{F,ij} * \sigma_{e,ij}. \quad (11)$$

Here, $\mu_{e,ij}$ and $\sigma_{e,ij}$ are the mean and the standard deviation of the write energy of the STT-RAM cell, respectively, at ‘ i ’→‘ j ’ switching.

The distribution of the write energy consumption can be then described by:

$$\mu_{eT} = \sum_{ij} \mu_{eT,ij} \quad \text{and} \quad \sigma_{eT} = \sqrt{\sum_{ij} \sigma_{eT,ij}^2}. \quad (12)$$

When $i = j$, the stored data is actually overwritten by the same value. The write energy can be zeroed by applying a “read-before-write” technique to eliminate this redundant operation. “Read-before-write” is the default mode of NVSim-VX^s and the energy of one read operation is automatically included in the write energy calculation.

Fig. 14 shows an example of NVSim-VX^s results of the nominal write energy consumptions and their standard deviations of a 512-bit block with three different switching patterns.

B. Block Level Write Error Rate

Due to the asymmetry of the write error rates at both switching directions of STT-RAM cells, the block level write error rate is also highly related to the switching patterns of the array. The array-level write error rate can be easily extracted from the cell level results for a given array size as:

$$P = 1 - \prod_{i=0}^1 (1 - W_{ER_i})^{N_{F_i}} \quad (13)$$

Here, we assume the location information of bit switchings is known during programming. N_{F_i} and W_{ER_i} denote the number of flipping bits and the bit write error rate, respectively for the switching of

Fig. 15. The block level write error rate for 3 different flipping combinations at 10ns write time and $w = 3L$ transistor width, $L = 45nm$.

‘ i ’→‘ j ’. Fig. 15 shows an application example of NVSim-VX^s in simulating the write error rate of a 512-bit block with different switching patterns.

V. CONCLUSION

The existing deterministic nonvolatile memory simulators are incapable to capture the probabilistic behaviors of STT-RAM incurred by process variations and thermal fluctuations. In this work, we introduce NVSim-VX^s, an enhanced version of NVSim, to facility the need of statistical STT-RAM modeling for write performance, errors and energy consumption. The example simulation results show that increasing temperature dramatically degrades the STT-RAM’s performance and reliability. Moreover, NVSim-VX^s simulations quantitatively validate the advantages of perpendicular STT-RAM over in-plane STT-RAM, showing its promising capability to predict the scaling trend of STT-RAM.

VI. ACKNOWLEDGEMENT

This work was supported in part by NSF CNS-1202225, DOE DE-SC0013553, NSF 1461698/1500848/1533933.

REFERENCES

- [1] S. Arcaro, S. Di Carlo, M. Indaco, D. Pala, P. Prinetto, E. Vatajelu, et al. Integration of stt-mram model into cacti simulator. In *International Design & Test Symposium*, pages 67–72. IEEE, 2014.
- [2] 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) System*, 18(12):1724–1734, 2010.
- [3] Y.-C. Chen, H. Li, W. Zhang, and R. E. Pino. The 3-d stacking bipolar rram for high density. *IEEE Transactions on Nanotechnology*, 11(5):948–956, 2012.
- [4] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi. Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 31(7):994–1007, 2012.
- [5] E. Eken, Y. Zhang, W. Wen, R. Joshi, H. Li, and Y. Chen. A novel self-reference technique for stt-ram read and write reliability enhancement. *IEEE Transactions on Magnetics*, 50(11):1–4, 2014.
- [6] R. Kanj, R. Joshi, and S. Nassif. Mixture importance sampling and its application to the analysis of sram designs in the presence of rare failure events. In *Design Automation Conference*, pages 69–72. ACM, 2006.
- [7] M. Motoyoshi, I. Yamamura, W. Ohtsuka, M. Shouji, H. Yamagishi, M. Nakamura, H. Yamada, K. Tai, T. Kikutani, T. Sagara, et al. A study for 0.18μm high-density mram. In *Symposium on VLSI Technology*, pages 22–23, 2004.
- [8] P. Shivakumar and N. P. Jouppi. Cacti 3.0: An integrated cache timing, power, and area model. Technical report, Compaq Computer Corporation, 2001.
- [9] P. Wang, E. Eken, W. Zhang, R. Joshi, R. Kanj, and Y. Chen. A thermal and process variation aware mtj switching model and its applications in soft error analysis. In *More than Moore Technologies for Next Generation Computer Design*, pages 101–125. Springer, 2015.
- [10] W. Wen, Y. Zhang, Y. Chen, Y. Wang, and Y. Xie. Ps3-ram: a fast portable and scalable statistical stt-ram reliability analysis method. In *Design Automation Conference*, pages 1191–1196. ACM, 2012.
- [11] B. Wu, Y. Cheng, Y. Wang, A. Todri-Sanial, G. Sun, L. Torres, and W. Zhao. An architecture-level cache simulation framework supporting advanced pma stt-mram. In *Nanoscale Architectures (NANOARCH), 2015 IEEE/ACM International Symposium on*, pages 7–12. IEEE, 2015.
- [12] W. Xu, Y. Chen, X. Wang, and T. Zhang. Improving stt mram storage density through smaller-than-worst-case transistor sizing. In *Design Automation Conference*, pages 87–90. ACM, 2009.
- [13] Y. Zhang, I. Bayram, Y. Wang, H. Li, and Y. Chen. Adams: asymmetric differential stt-ram cell structure for reliable and high-performance applications. In *International Conference on Computer-Aided Design*, pages 9–16. IEEE Press, 2013.
- [14] Y. Zhang, X. Wang, and Y. Chen. Stt-ram cell design optimization for persistent and non-persistent error rate reduction: a statistical design view. In *International Conference on Computer-Aided Design*, pages 471–477. IEEE Press, 2011.