Design Considerations for Energy-Efficient and Variation-Tolerant Nonvolatile Logic

Jinghua Yang, *Member, IEEE*, Aykut Dengi, and Sarma Vrudhula, *Fellow, IEEE*

***Abstract*— Systems powered by harvested energy must consume very low power and withstand frequent interruptions in power. Nonvolatile logic (NVL) addresses the latter by saving the system state in flipflops enhanced with spin-transfer torque magnetic tunnel junctions (STT-MTJs) as the nonvolatile storage devices. Manufacturing variations in the STT-MTJs and in CMOS transistors significantly reduce yield, leading to overdesign and high-energy consumption. A detailed analysis of the design tradeoffs in the driver circuitry for performing backup and restore, and a novel method to design the energy optimal driver for a given yield is presented. Next, efficient designs of two nonvolatile flip-flop (NVFF) circuits are presented, in which the backup time is determined on a per-chip basis, resulting in minimizing the energy wastage and satisfying the yield constraint. To achieve a yield of 98%, the conventional approach would have to expend nearly 5× more energy than the minimum required, whereas the proposed tunable approach expends only 26% more energy than the minimum. Also included are the energy consumption of the proposed NVFF designs when used in two larger function blocks. Experimental results were based on a commercial 40-nm process design kit, and HSPICE simulations with foundry supplied statistical models and data.**

***Index Terms*— Energy harvesting, flip-flop, Internet of Things (IoT), low power, magnetic tunnel junction (MTJ), nonvolatile logic (NVL), nonvolatile memory (NVM), resistive random access memory.**

I. Introduction

MICROELECTRONIC circuits that obtain their energy from ambient energy sources (AESs) such as solar, piezoelectric, vibration, airflow, and thermoelectric [1] are expected to become essential for the burgeoning field of the Internet of Things (IoT). Although there are substantial differences among them in power density (ranging from tens of μW to tens of mW), as well as variations in the delivered energy over time, it is the intermittent nature of the delivered energy by AES that poses the most difficult challenge for microelectronic systems as they are generally architected for continuous operation. Hence, quickly predicting an impending power disruption, and saving the state in some form of nonvolatile storage is critical for all but the simplest devices. The emergence of CMOS-compatible nonvolatile memory (NVM) technologies (e.g., MRAM, RRAM, PCRAM, CBRAM, FeRAM, and STT-RAM) over the past decade has opened the way for new circuit architectures for near instantaneous and energy-efficient backup and recovery.

NVM for backup and restoration during a power disruption can be implemented in one of two ways. One option is to have a NVM array (NVMA) that is separate from the local (volatile) registers where the intermediate computation results are stored [2]. Before the power failure, the data in all the registers would be saved serially in the NVMA and later serially restored. The other option (e.g., [3]–[11]), which is the focus of this paper, is to have each register be a nonvolatile flip-flop (NVFF), which operates like a regular flip-flop in normal mode, but has the added capability of storing its state in a local nonvolatile device before power failure.

The nonvolatile devices that are most often employed in the various NVFF designs have a common characteristic, namely, that they require a critical current to be delivered for some minimum duration in order to switch their state. Process variations, including both within die and die-to-die variations pose a major challenge in circuits with NV devices. These, along with variations in the CMOS circuits that drive the NV device, result in statistical variations in the actual current being delivered. Designing with such variations in mind requires quantifying the ensuing tradeoffs between reliability (probability of successful backup), area of the driver circuits, backup and restoration time, and power consumption. Optimal driver design of a NVFF considering process variations and examination of the tradeoffs have not been well explored in the existing literature. Ignoring variations in the transistors and MTJ devices will result in poor functional yield. However, the traditional worst case-corners approach results in significant wastage of energy during backup.

II. Overview of This Paper

This paper presents an architecture and method for variation-tolerant, energy-optimal design of a NVFF, that uses a spin-transfer torque magnetic tunnel junction (STT-MTJ) as the nonvolatile device. The first part of this paper (Section III) explores the tradeoffs in the design of the backup driver, independent of the NVFF circuit architecture. In the absence of process variations, the driver size (i.e., transistor widths) that minimizes the total backup energy is obtained, under the

Manuscript received August 30, 2017; revised December 29, 2017; accepted February 18, 2018. Date of publication March 22, 2018; date of current version November 30, 2018. This work was supported in part by NSF under Grant 1230401, Grant 1237856, and Grant 1701241 and in part by the NSF IUCRC Center for Embedded Systems. *(Corresponding author: Sarma Vrudhula.)*

The authors are with the School of CIDSE, Arizona State University, Tempe, AZ 85281 USA (e-mail: Jinghua.Yang@asu.edu; Aykut.Dengi@asu.edu; vrudhula@asu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2018.2812700

constraint of equal time to back up a 1 or a 0 (Section III-B). However in the presence of process variations, the specified driver size represents a nominal or mean value around which statistical perturbations occur, i.e, each nominal value represents a population of circuits. A method is presented that identifies the nominal value of the driver size, that minimizes the average energy over the corresponding population, subject to satisfying a given yield (Section III-C).

In practice, the value of the backup time so obtained (i.e., from models) will be be an upper bound on the actual backup time required to satisfy the given yield. Using the upper bound for the backup time wastes a substantial amount energy because many chips may require far less time to successfully backup the data. This motivates the need for adjusting the backup time on a per-chip basis after fabrication. Toward this end, the paper presents two designs of a NVFF with *scan*, denoted as nonvolatile scan flip-flop (NVSFF)-DM (Section IV-D) and NVSFF-MS (Section IV-E). These designs have the same nonvolatile storage unit (NVSU) (Section IV-C), but differ in the design of the *volatile* flip-flop component. The volatile component of NVSFF-DM is sense-amp-based flip-flop whose outputs are differential, and hence can be directly connected to the NVSU. Alternatively, the volatile component of the NVSFF-MS is a conventional master-slave flip-flop whose design is modified to properly interface with the NVSU.

The proposed designs have several advantages over prior work [3]–[5]. In contrast to existing designs, the control circuitry in the NVSU is much simpler, and allows for near-instant backup and restoration, allowing a computation to be interrupted in midstream and resumed where it was suspended, with minimum hardware overhead for the control unit. In addition, using a scan mechanism, both designs allow for the actual backup time to be determined on a per-chip basis, which turns out to be much smaller with the optimally sized driver.

The evaluations of the backup time and energy consumption of optimally sized NVSFF-DM and NVSFF-MS flip-flops, and the energy savings when they are used in larger circuits are presented in Section V. Section VI summarizes the prior work and Section VII presents the conclusions.

III. Nonvolatile Flip-Flop Design Tradeoffs

A common required component for storing and restoring data into and from the NV devices is the *backup driver* [5], [7]. Fig. 1(a) shows the key components of such a circuit, without any of the control logic. It consists of two inverters in series with an STT-MTJ device. A brief, high-level description of the behavior of an STT-MTJ, sufficient to explain the design and optimization of the backup driver circuit, follows.

A. STT-MTJ Cell

An STT-MTJ cell consists of two ferromagnetic layers separated by an oxide insulation layer (usually M_gO) [see Fig. 1(b)]. The magnetization of the reference layer is fixed, whereas that of the free layer can be switched. When the spin orientations in the two layers are parallel (antiparallel), the STT-MTJ cell has a low (high) resistance, denoted by

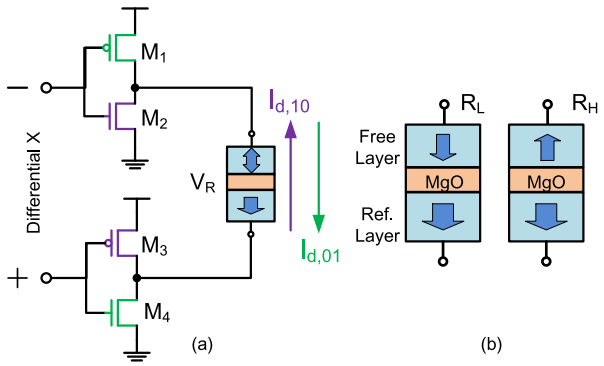


Fig. 1. (a) Simplified driver circuit providing bidirectional current to switch STT-MTJ cell. (b) Structure of STT-MTJ.

R_L (R_H), which represent the logic 0 and 1, respectively. TMR $= (R_H - R_L)/R_L$ represents the relative separation between the two resistance values, with typical values between 50% to 200%, and can be as high as 600% [12]. It is assumed that R_H and R_L are constants, independent of the voltage across the device, and that the change in resistance between R_L and R_H is abrupt. The switching time τ is the time at which the abrupt change takes place. Due to thermal fluctuations, the STT-MTJ switching is a stochastic [13]–[15]. However, deterministic switching is assumed when the device current I_d exceeds a critical value I_c.

Applying $X = 1$ in the backup driver will cause a current $I_{d,01}$ to flow through M_1, the STT-MTJ, and M_4. This must exceed a critical current $I_{c,01}$ for a duration of τ_{01} in order for the STT-MTJ to switch from R_L to R_H. Similarly, $X = 0$ will cause a current $I_{d,10}$ to flow in the reverse direction through M_3, the STT-MTJ, and M_2. This current must exceed a critical current $I_{c,10}$ for a minimum duration of τ_{10}, in order for the device to switch from R_H to R_L. Thus, the four critical parameters associated with an MTJ are R_L, R_H, I_c, and τ.

There has been extensive work on the development of compact models of STT-MTJ devices [15]–[18]. For feature sizes below 40 nm, the model described in [15] (also in [18]) is used here, as it integrates a number of physical models, enabling the analysis of static, dynamic and stochastic behavior, and reports results that show good agreement with experiments. The following simplified expressions for R_L and R_H and the switching time τ of an STT-MTJ taken from [18] are utilized in the methodology followed in this paper. The parameters α, β, and κ include multiple physical parameters that are explained in [18]. For the purposes herein, they are technology constants

$$R_L = \alpha t_{\rm ox} e^{\beta t_{\rm ox}} \tag{1}$$

$$R_H = (1 + \text{TMR}) \cdot R_L \tag{2}$$

$$\tau = \kappa \frac{1}{| I_d - I_c |}. \tag{3}$$

R_L and R_H are comparable to the on-channel resistances of the CMOS transistors in the driver. Therefore, the voltage drop across the MTJ during switching, combined with the fixed power supply V_{dd}, limits the maximum current that a driver can deliver. That driver current depends on the transistor dimensions together with R_L and R_H, which are

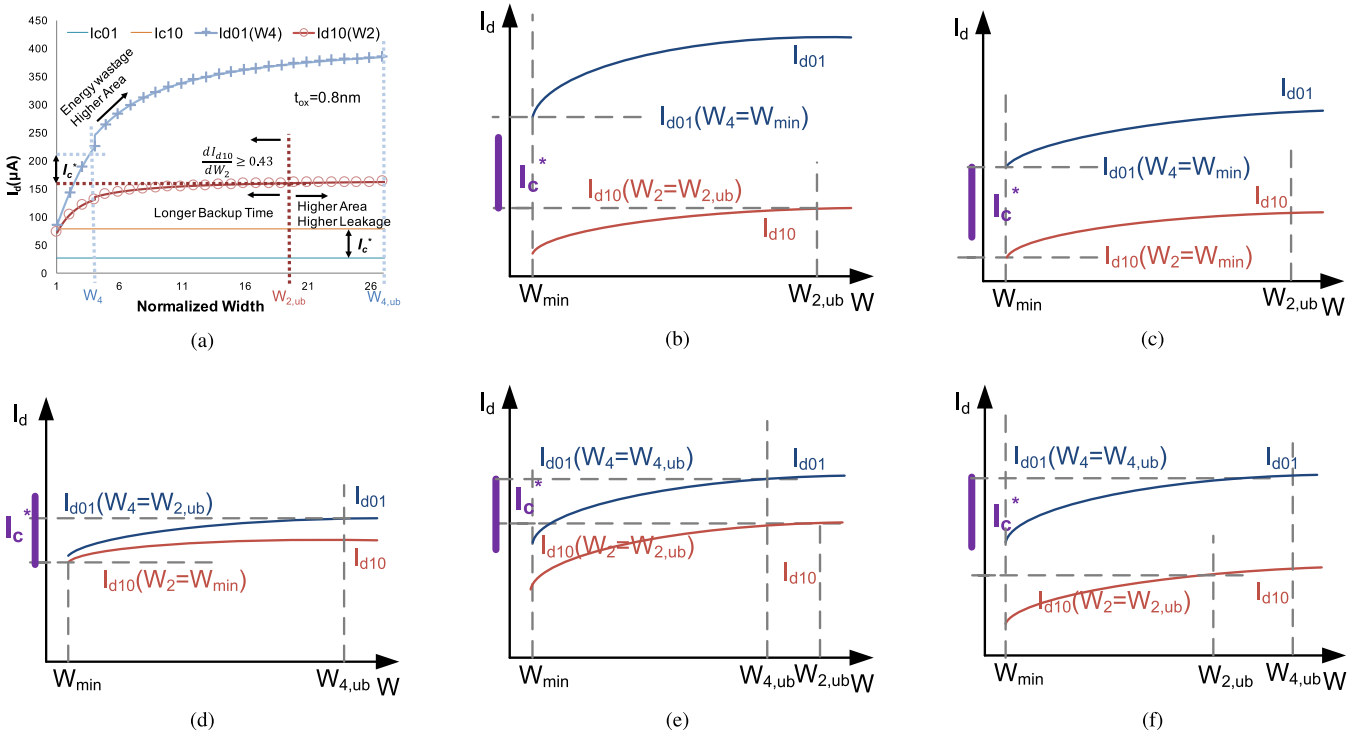


Fig. 2. Possible cases of driver current versus transistor width. (a) $I_{d,01}$ and $I_{d,10}$ versus normalized width. $\bar{t}_{\rm ox} = 0.8$ nm. (b) Case I. (c) Case II. (d) Case III. (e) Case IV. (f) Case V.

in turn related to $t_{\rm ox}$ of the MTJ (1) and (2). Local and global process variations in transistors and MTJs make the driver current a statistically varying quantity among different devices on the same die and among the same devices on different dices. However, before considering process variations, it will be instructive to examine the factors that affect the transistor sizes in the driver, and how those sizes might be determined.

B. Driver Sizes Ignoring Process Variations

$I_{d,01}$, and $I_{d,10}$ are functions of R_L, R_H, and the transistor widths W_4 and W_2, where R_L and R_H are determined by $t_{\rm ox}$ [see (1)]. Writing a 1 (0) in the MTJ will require $I_{d,01}(t_{\rm ox}, W_4) > I_{c,01}$ ($I_{d,10}(t_{\rm ox}, W_2) > I_{c,10}$), and the corresponding switching time τ_{01} (τ_{10}) will be inversely proportional to the excess current (3).

Let $\gamma = W_1/W_4 = W_3/W_2$ denote the ratio of the width of pFET M_1 (M_3) to the width of nFET M_4 (M_2), and assume that γ is fixed. Fig. 2(a) shows HSPICE generated plots of $I_{d,01}$ and $I_{d,10}$ as a function of the width of the corresponding nFETs W_4 and W_2, respectively, for a specific value of $t_{\rm ox}$.

From Fig. 2(a), it is seen that any pair of values for W_4 and W_2 are feasible as long as the corresponding $I_{d,01}(W_4) > I_{c,01}$ and $I_{10}(W_2) > I_{c,10}$. The objective is to choose values that minimize the total energy $E_{\rm total}$ required to store a 0 and 1. $E_{\rm total} = V_{dd}(\tau_{01} I_{d,01}(W_4) + \tau_{10} I_{d,10}(W_2))$. Let $\tau = \max\{\tau_{01}, \tau_{10}\}$ be single time to backup a 0 or a 1. Then,

$$E_{\rm total} = V_{dd}[\tau_{01} I_{d,01}(W_4) + (\tau - \tau_{01}) I^*_{d,01}(W_4) + \tau_{10} I_{d,10}(W_2) + (\tau - \tau_{10}) I^*_{d,10}(W_2)]. \quad (4)$$

$I^*_{d,01}(W_4)$ and $I^*_{d,10}(W_2)$ are the currents after the state transitions have completed. They are different from $I_{d,01}(W_4)$ and $I_{d,10}(W_2)$ because of the change in the device resistances. $E_{\rm total}$ is at least $V_{dd}(\tau_{01} I_{d,01}(W_4) + \tau_{10} I_{d,10}(W_2))$. Hence, the minimum of the average or total energy with a single backup time would require that $\tau = \tau_{01} = \tau_{10}$. Then, using (3), $I_{d,01}(W_4) - I_{c,01} = I_{d,10}(W_2) - I_{c,10}$, or equivalently, $I_{d,01}(W_4) - I_{d,10}(W_2) = I_{c,01} - I_{c,10} = I^*_c$, where I^*_c is independent of W. Therefore, the basic constraint that needs to be satisfied when determining the driver size is

$$I_{d,01}(W_4) = I_{d,10}(W_2) + I^*_c. \quad (5)$$

If (5) is satisfied, then the total energy is $E_{\rm total} = V_{dd}\tau(2I_{d,10}(W_2) + I^*_c)$. Now $\tau = \tau_{10} = \kappa/(I_{d,10}(W_2) - I_{c,10})$, and $E_{\rm total}$ can be written as

$$E_{\rm total} = V_{dd}\kappa \left(\frac{2I_{d,10}(W_2) + I^*_c}{I_{d,10}(W_2) - I_{c,10}} \right). \quad (6)$$

Equation (6) shows that with equal switching times for storing a 0 and 1, minimizing the total energy is equivalent to maximizing $I_{d,10}(W_2)$. This fact can be used to determine W_2 and $I_{d,10}(W_2)$. W_4 is determined by solving (5).

Fig. 2(a) shows plots of $I_{d,10}(W_2)$ (lower curve) and $I_{d,01}(W_4)$ as a function of driver transistor width,[1] which are enumerated in discrete increments. $W_{\rm min}$ is the minimum possible width. $W_{2,ub}$ and $W_{4,ub}$ denote widths at which the currents $I_{d,10}$ and $I_{d,01}$ have saturated, i.e., for some small $\epsilon > 0$, $W_{2,ub} = \min\{W \mid dI_{10}/dW \le \epsilon\}$, and $W_{4,ub} =$

[1]Transistor width is normalized to the minimum width allowed in the technology. Therefore, $W_{\rm min} = 1$

$\min\{W \mid dI_{01}/dW \leq \epsilon\}$. Choosing a value larger than $W_{2,ub}$ or $W_{4,ub}$ will not increase the current appreciably, but increases area. As E_{total} decreases with I_d, and I_d is monotonic with respect to W, the width W_2 that maximizes I_d can be determined by examining the boundary conditions.

Case I $I_{d,01}(W_4 = W_{\min}) > I_{d,10}(W_2 = W_{2,ub}) + I_c^*$*:*

This is shown in Fig. 2(b), and corresponds to the situation where $R_L \ll R_H$. (The low and high resistances are widely separated.) Even choosing $W_2 = W_{2,ub}$, there is no corresponding value of W_4 for which $I_{d,10}(W_{2,ub}) + I_c^* = I_{d,01}(W_4)$, i.e., equal backup times is not possible, and (5) cannot be satisfied. Therefore, the only choice is $W_4 = W_{\min}$. Choosing a larger value for W_4 makes writing a logic 1 even faster and wastes energy and area, because the actual backup time is determined by the time required to write a logic 0. Choosing a smaller value for W_2 makes writing a logic 0 even slower.

Note that with $R_L \ll R_H$, the process of *reading* is more robust, at the expense of increased energy for writing. This is opposite to the general conclusion on NVM design that wide R_L and R_H separation is always desired. In an AES powered nonvolatile logic (NVL) design, devices with widely separated resistance states like an RRAM cell require more energy for writing data than MTJs, while providing greater robustness when reading data.

Case II $I_{d,01}(W_4 = W_{\min}) > I_{d,10}(W_2 = W_{\min}) + I_c^*$*:*

This is depicted Fig. 2(c). Since I_d is monotonically increasing, $I_{d,10}(W_2 = W_{2,ub}) > I_{d,10}(W_2 = W_{\min})$. Therefore, Case I implies this Case. Hence if Case I fails, and this Case is true, then

$$\begin{aligned} I_{d,10}(W_2 = W_{2,ub}) &> I_{d,01}(W_4 = W_{\min}) - I_c^* \\ &> I_{d,10}(W_2 = W_{\min}). \end{aligned}$$

Equation (5) has a solution with $W_2 = W_{2,ub}$, and $W_4 = I_{d,01}^{-1}(I_{d,10}(W_2 = W_{2,ub}) + I_c^*)$. Note that choosing $W_4 = W_{4,ub}$ will not satisfy (5).

Case III $I_{d,01}(W_4 = W_{4,ub}) < I_{d,10}(W_2 = W_{\min}) + I_c^*$*:*

This is shown in Fig. 2(d), and corresponds to the situation when R_L and R_H are very close and their magnitudes are high, resulting in lower and flatter I_d curves. Higher resistances might be desired so as to reduce the possibility of a *read disturb* and improve thermal stability. In this situation, (5) has no solution, and the only option is $W_4 = W_{4,ub}$, and $W_2 = W_{\min}$. This speeds up the writing of a logic 1, and slows the writing of a logic 0, when compared to both transistors being of minimum size.

Case IV $I_{d,01}(W_4 = W_{4,ub}) < I_{d,10}(W_2 = W_{2,ub}) + I_c^*$*:*

This is shown in Fig. 2(b). Since $I_{d,10}(W_2 = W_{\min}) < I_{d,10}(W_2 = W_{2,ub})$, Case III implies this Case. Hence, if Case III fails, and this Case holds, then

$$\begin{aligned} I_{d,10}(W_2 = W_{2,ub}) &> I_{d,01}(W_4 = W_{4,ub}) - I_c^* \\ &> I_{d,10}(W_2 = W_{\min}). \end{aligned}$$

Equation (5) has a solution, which is $W_4 = W_{4,ub}$ and $W_2 = I_{d,10}^{-1}(I_{d,01}(W_4 = W_{4,ub}) - I_c^*)$.

Case V $I_{d,01}(W_4 = W_{4,ub}) > I_{d,10}(W_2 = W_{2,ub}) + I_c^*$*:*

Algorithm 1 Computes Optimal Transistors Sizes W_2, W_4

EOPTDRIVERSIZE($W_{min}, W_{2,ub}, W_{4,ub}$);
output: Energy optimal values of W_2, W_4
/* case I */
if $I_{d,01}(W_4 = W_{min}) > I_{d,10}(W_2 = W_{2,ub}) + I_c^*$ **then**
  $W_2 = W_{2,ub}$;
  $W_4 = W_{min}$;
endif
/* case II */
else if $I_{d,10}(W_2 = W_{2,ub}) > I_{d,01}(W_4 = W_{min}) - I_c^* > I_{d,10}(W_2 = W_{min})$ **then**
  $W_2 = W_{2,ub}$;
  $W_4 = I_{d,01}^{-1}(I_{d,10}(W_2 = W_{2,ub}) + I_c^*)$;
endif
/* case III */
else if $I_{d,01}(W_4 = W_{4,ub}) < I_{d,10}(W_2 = W_{min}) + I_c^*$ **then**
  $W_2 = W_{min}$;
  $W_4 = W_{4,ub}$;
endif
/* case IV */
else if $I_{d,10}(W_2 = W_{min}) < I_{d,01}(W_4 = W_{4,ub}) - I_c^* < I_{d,10}(W_2 = W_{2,ub})$ **then**
  $W_4 = W_{4,ub}$;
  $W_2 = W_2 = I_{d,10}^{-1}(I_{d,01}(W_4 = W_{4,ub}) - I_c^*)$;
endif
/* case V */
else
  $W_2 = W_{2,ub}$;
  $W_4 = I_{d,01}^{-1}(I_{d,10}(W_2 = W_{2,ub}) + I_c^*)$;
endif

From Fig. 2(f), it is apparent that there is solution to (5), given by $W_2 = W_{2,ub}$ and $W_4 = I_{d,01}^{-1}(I_{d,10}(W_2 = W_{2,ub}) + I_c^*)$. Once again, note that choosing $W_4 = W_{4,ub}$ first, does not lead to a solution.

These five cases are summarized in Procedure EOPTDRIVERSIZE shown in Algorithm 1.

C. Driver Sizes Considering Process Variations

Algorithm 1 is now adapted for the case where the parameters of the transistors in the driver and the MTJ device are subject to manufacturing variations. For an MTJ device, the primary design parameter is its dimension and for the driver circuit, they are the dimensions of the transistors M_1–M_4. There are several secondary nondesign parameters associated with the MTJ, such as localized fluctuation of magnetic anisotropy, thermally activated initial procession angle, and thermal component of internal energy [19], whose variations are not modeled in this paper.

For an MTJ device, it has been shown that variations in t_{ox} have the most significant impact on energy consumption [13]. This is due to fact that R_H and R_L have an exponential dependence on t_{ox} [see (1) and (2)]. During fabrication, the oxide is grown over the entire die, and consequently, it is

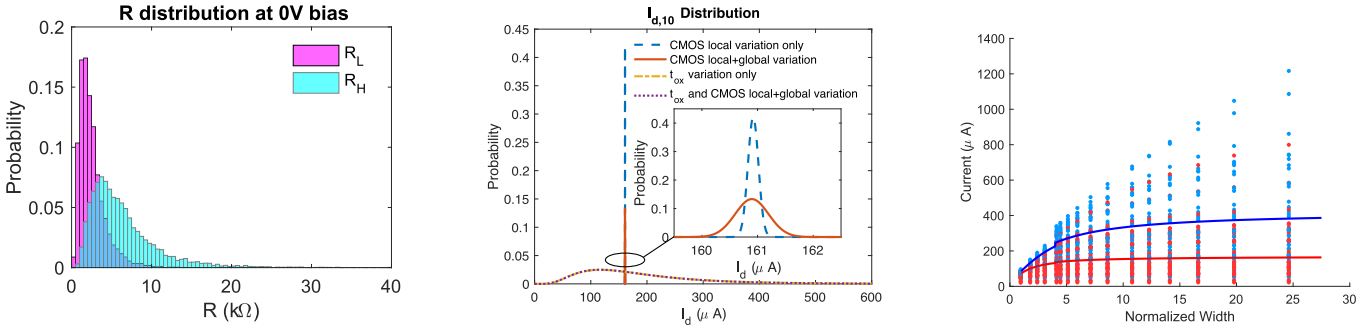


Fig. 3. Effect of process variations on parameters that impact driver design. Data for CMOS is from a 40-nm commercial library with foundry supplied parameters and HSPICE models. Data for MTJ variations was generated assuming $\overline{t}_{\rm ox} = 0.8$ nm and $\sigma_{\rm tox} = 0.1\overline{t}_{\rm ox}$, and using models in [15] and [18]. Note: to avoid clutter, in (c) only a subset of widths are plotted. (a) Frequency histograms of R_L and R_H using 10K MonteCarlo samples. (b) Frequency histogram of I_d using 10K Monte Carlo samples. (c) I_d current versus driver width. Blue dotted line is $I_{d,10}$, Red dotted line is $I_{d,01}$. Lines are with no process variation, dots are currents with $t_{\rm ox}$ and CMOS (local and global) variation.

assumed that the variation in its thickness is the same for all devices. Hence, following [13], [20], $t_{\rm ox}$ variation is assumed to global. Consequently, the length $L_{\rm MTJ}$ and width $W_{\rm MTJ}$ of the MTJ can be assumed to be fixed at the minimum feature size of the technology, and that the deviations in $t_{\rm ox}$ among different MTJs on a given die will be the same. On the other hand, the dimensions of the CMOS transistors in the driver are assumed to be subject to both local and global variations. Thus, the widths W_2 and W_4 are modeled as independent random variables centered around their respective nominal values $\overline{W}_2$ and $\overline{W}_4$, which are to be specified as part of the design.

Variations in $t_{\rm ox}$ result in variations in R_L and R_H [see Fig. 3(a)], and variations in $t_{\rm ox}$, W_2, and W_4 will result in corresponding variations in the driver currents. Fig. 3(b) shows frequency histograms of I_d in a driver, assuming different sources of variations. The inset plot shows the histogram of I_d considering local and global variations only in the driver transistors, and the outer plot includes variations in the transistor dimensions and $t_{\rm ox}$ of the MTJ. The plots indicate that variations in $t_{\rm ox}$ overwhelm the effect of variations in the transistors' dimensions. However, in the interest of generality and applicability to scaled geometries, the currents $I_{d,01}$ and $I_{d,10}$ are modeled as a function of a collection of random variables over the parameter space $(W_2, W_4, t_{\rm ox})$.

Fig. 3(c) shows plots of I_d as a function of the (normalized) widths of the driver's transistors. The red ($I_{d,10}$) and blue ($I_{d,01}$) solid curves correspond to the case where no variations are considered in the transistor dimensions nor in the $t_{\rm ox}$ of the MTJ. These plots are similar to those shown in Fig. 2(a). The plots also show individual populations (10K) of the $I_{d,10}$ and $I_{d,01}$ values generated by Monte Carlo simulations, by varying $(W_2, W_4, t_{\rm ox})$ around their nominal values $[\overline{W}_{2,i}, \overline{W}_{4,j}, \overline{t}_{\rm ox}]$, for $(i, j) \in [1, n]$. Let $S(\overline{W}_2, \overline{W}_4, \overline{t}_{\rm ox})$ denote the population of samples centered at $(\overline{W}_2, \overline{W}_4, \overline{t}_{\rm ox})$.

The problem to determine the energy-optimal driver size in the presence of process variations is to identify the population (i.e., the nominal values $\overline{W}_2, \overline{W}_4$) that have at least $y\%$ (y being the yield) of the samples resulting in a successful backup and restore, and have minimum average energy. Yield and energy are related. To see how to compute energy as a function of yield, consider samples of I_d shown in Fig. 3(c). Each pair of data points (red and blue dots) within a population has an associated backup time τ_{01} and τ_{10}, that can computed using (3). The corresponding total energy would be calculated by (4) where $\tau = \tau_y$. This energy is computed for all the samples in a given population whose backup times fall within the y percentile, for a given yield y.

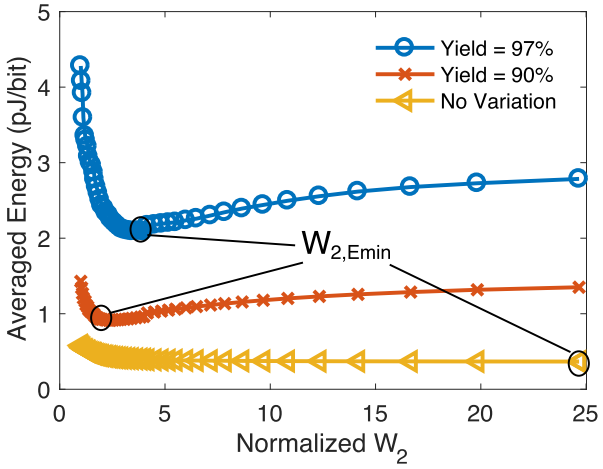


Fig. 4. Average total energy versus driver width, for different yields, accounting for process variations. Minimum energy is achieved with $W_{2,Emin} = W_{2,ub}$ when no variation is included. In the presence of process variations, yield constrained minimum energy can be achieved with smaller $W_{2,Emin}$, whose value depends on the target yield.

Fig. 4 shows plots of the average energy versus the driver width, for several values of the yield y. It is clear that unlike the deterministic case [see Fig. 2(a)], the minimum of the average energy does not necessarily correspond to the largest value of the transistor width (i.e., maximum current) but instead to some intermediate value. The smaller W_2 implies lower current and longer backup time.

Algorithm 2 Procedure to Compute Optimal Sizes of Driver Transistors W_1, W_4, W_3, W_2 Considering Process Variations

1 EOPTDRIVERSIZEWPR($[W_{min}, W_{ub}], t_{ox}, y$) ;
output: Energy optimal values of $\overline{W}_2$, $\overline{W}_4$ and τ_y
2 $i = 1$;
3 $\overline{W}_{2,0} = \overline{W}_{2,ub}$;
4 $\overline{W}_{4,0} = \overline{W}_{4,ub}$;
5 $E_{avg,0} = \infty$;
6 **while** $\overline{W}_{min} \leq \overline{W}_i \leq \overline{W}_{ub}$ **do**
7 $[\overline{W}_{2,i}, \overline{W}_{4,i}]$ =
8 EOPTDRIVERSIZE($\overline{W}_{min}$, $\overline{W}_{2,i-1}$, $\overline{W}_{4,i-1}$) ;
/* Generate N MonteCarlo samples */
9 $S_j = (W_{2,i,j}, W_{4,i,j}, t_{ox,j})$ =MC($\overline{W}_{2,i}$, $\overline{W}_{4,i}$, $\overline{t}_{ox}$);
10 **for** *j=1:N* **do**
/* Find driving current by HSPICE simulation */
11 ($I_{d,01,j}$, $I_{d,10,j}$) = HSPICE(S_j) ;
12 ($\tau_{01,j}$, $\tau_{10,j}$) = Eqn 3 ($I_{d,01,j}$, $I_{d,10,j}$) ;
13 τ_j = max($\tau_{01,j}$, $\tau_{10,j}$) ;
14 **end**
/* y% of switching times $\leq \tau_y$ */
15 τ_y : Prob($\tau \leq \tau_y$) = y;
16 **for** *j=1:N* **do**
17 **if** $\tau_j \leq \tau_y$ **then**
18 E_j= Eqn 4 (τ_y,$\tau_{01,j}$, $\tau_{10,j}$,$I_{d,01,j}$, $I_{d,10,j}$) ;
19 **endif**
20 **end**
21 $E_{avg,i} = (E_1 + E_2 + \cdots + E_N)/(yN)$;
22 **if** $E_{avg,i} > E_{avg,i-1}$ **then**
23 **return** $\overline{W}_{2,i-1} + \Delta W$, $\overline{W}_{4,i-1} + \Delta W$, τ_y;
24 **endif**
25 $\overline{W}_{2,i} = \overline{W}_{2,i} - \Delta W$;
26 $\overline{W}_{4,i} = \overline{W}_{4,i} - \Delta W$;
27 $i = i + 1$;
28 **end**

The procedure to determine the nominal widths of the driver transistors in the presence of process variations is shown in Algorithm (2). The objective is to identify the nominal values $(\overline{W}_2, \overline{W}_4)$ that define a population $S(\overline{W}_2, \overline{W}_4, \overline{t}_{ox})$ whose ensemble average energy computed over all those outcomes whose backup times fall below τ_y (the y percentile value of the backup time) is minimum. The procedure is a nonparametric or data-driven approach, using the empirical distribution of currents generated by Monte Carlo simulations to compute averages. As the set of transistor widths form a discrete set, the procedure starts with setting the nominal values to their respective upper bounds (lines 3 and 4), and iterates over the discrete set (line 6). Procedure EOPTDRIVERSIZE is used to determine the next nominal value around which to generate the sample population (lines 7–9), and then the backup times and currents are computed for each sample point (lines 10–14). The average of the samples whose backup times are within the y percentile value is computed (lines 15–21). The minimum average energy value is retained, and the procedure terminates as soon the average starts to increase (lines 22 and 23).

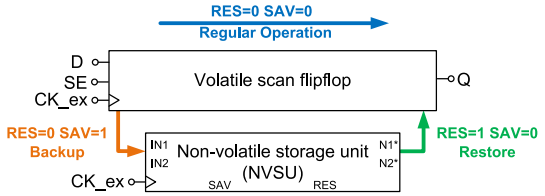


Fig. 5. Basic structure of NVSFF.

IV. NONVOLATILE FLIPFLOPS WITH SCAN

A. Yield Versus Energy Consumption

Fig. 4 shows that higher yield requires higher energy expenditure. One way to reduce backup energy is to boost the voltage [21]. However, this is not practical for the type of low-voltage, low-power application-specified integrated circuit (ASICs) employing energy harvesting that are the target of this paper. Techniques for improving the energy efficiency by balancing the backup times used in NVM as described in [14], [21], and [22] are not applicable for NVFFs. For this reason, the method described in Section III, minimizes the average energy under a yield constraint by sizing the drivers separately. Other techniques that improve the write margin by increasing the driver size (to increase I_d) and the backup time, result in high-energy consumption [14], [22]. Device engineering as in [23] can also be done to trade retention time with write energy. However that is outside the scope of this paper.

The backup time τ_y determined by procedure EOPTDRIVERSIZEWPR ensures that, with a high probability, y% of the dice will succeed in backup of a "1" and "0." However, the conservative choice of τ_y results in wasted energy for most of the dice. This motivates the adaptive approach of determining the backup time on a per-chip basis. This section presents the architecture of a NVFF equipped with a scan mechanism, which allows for dynamically testing and adjusting the backup time to minimize the backup energy. This scan mechanism is compatible with the normal scan available on traditional flipflops, and hence has minimum hardware cost.

B. NVSFF Basic Structure

The general structure of a nonvolatile scan flip-flop (NVSFF) is shown in Fig. 5. A NVSU is attached to a volatile flip-flop. This NVSFF has five modes of operation. In the normal mode (regular operation) and normal scan mode, it acts like an edge-triggered scan flip-flop. In these modes RES $=$ 0 and SAV $=$ 0, which together disconnect the path between the NVSU and the volatile flip-flop. During the backup mode, the flip-flop state is stored into the NVSU. After the backup mode is completed, the system can be safely powered off without losing the intermediate results. During the restore mode, the previous stored state is read out and presented on the flip-flop output Q. The nonvolatile test mode is a combination of the normal scan mode, the backup mode and the restore mode. This operation mode is mainly

for performing the nonvolatile device test and determining the backup time. Details of the circuit operation and design considerations are presented in the following.

C. Nonvolatile Storage Unit

The architecture of the NVSU is shown in Fig. 6. It takes two differential signals IN1 and IN2, and produces two differential outputs $N1^*$ and $N2^*$. RES and SAV control the operation mode. Two STT-MTJ devices are included in the NVSU. The one labeled STT_data store the state during backup mode. The one labeled STT_ref serves as a reference, used during the restore mode.

1) Normal Mode and Normal Scan Mode: The NVSU is inactive during the normal mode, and is turned off to save power. The input and output transistors are sized small to reduce the parasitics on the normal signal path.

2) Backup Mode: RES $= 0$ and SAV $= 1$ sets the NVSU to the backup mode. The unit labeled as state sense amplifier is inactive in this mode. Current will flow through write tri-state buffers TB1 and TB2 and set the state of STT_data. The current direction is determined by IN1 and IN2. Compared to the driver shown in Fig. 1, TB1 and TB2 consist of one pFET and two nFETs in a stack. The extra SAV driven pFET eliminates a false path to MTJ during restore mode.

The SAV signal is independent of the clock, and as long as SAV $= 1$ and inputs are differential, TB1 and TB2 will provide the necessary current to store the data. Note that consistent with other works on NVM and NVL [2], [24], [25], it is assumed that there exists a mechanism that will predict an impending power system failure and will initiate the backup by setting SAV $= 1$ during the period with CK $= 1$. A method to predict such a failure can be found in [25].

3) Restore Mode: When the power is reestablished, the state of the flip-flop can be restored by setting SAV $= 0$ and RES $= 1$. The two tristate buffers TB1 and TB2 are disabled. When $Rd = 0$, $N1^* = 1$ and $N2^* = 1$. When $Rd : 0 \rightarrow 1$, $M14$ and $M15$ in Fig. 6 become active, creating discharge paths to ground for both $N1^*$ and $N2^*$. Assuming STT_data $= R_L \ll R_{\text{ref}}$, the positive feedback in the state sense amplifier will sense the conductance difference between two discharging paths and set $N2^* = 0$ and $N1^* = 1$, which drive a regular flip-flop and set its output to $Q = 0$.

A *read disturb* occurs when the stored state in an STT-MTJ is flipped on a read operation. The probability of a read disturb in the NVSU can be reduced by using smaller transistors or lowering the power supply voltage for the state sense amplifier, at the cost of a longer restoration time. Unlike NVM implementations in which the stored data would be read more than once, in the NVFF with backup and restore, the stored data would only be restored to the datapath once. When the next power interrupt occurs, new data would be backed up. Therefore, the read disturb is not the primary concern in NVFF design.

4) Nonvolatile Test Mode: This mode is applied to test the functionality of other two modes as well as determine an optimal backup time. Unlike the other operation modes, this involves a sequence of operations. It starts in the normal scan mode (SE $= 1$, SAV $= 0$, and RES $= 0$) that scans in the test data, resulting in the data appearing at each output Q. After the data have been scanned in, the NVSFF is switched to the backup mode and restore mode. After a backup and restore step, the previous test data will be present at the output, if both steps completed successfully. Then, the output data are scanned out for verification by switching to the normal scan mode. The backup time is the duration when SAV $= 1$. The control signal sequence is shown in Fig. 7.

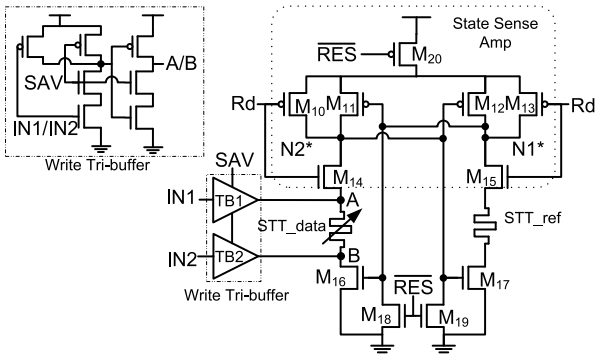


Fig. 6. Schematic of NVSU. The NVSU includes a write buffer, two STT-MTJ devices, and a state sense amplifier.

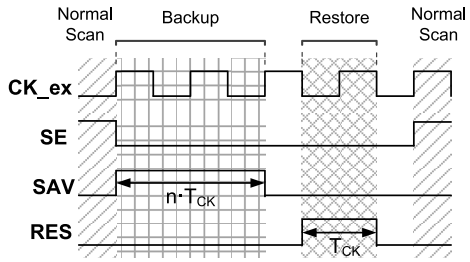


Fig. 7. Control signal sequence during nonvolatile test mode.

5) Timing of Control Signals: RES is synchronized with the falling edge of the clock, and therefore can be easily generated by a negative edge triggered flip-flop. Rd is generated by both RES and CK, which feeds into state sense amplifier. SAV controls TB1 and TB2. When input signals IN1 and IN2 are stable, the duration of SAV determines backup time τ. Although SAV can be synchronous or asynchronous, a synchronous signal is preferred as it can easily be generated by a counter followed by a flip-flop, and the total backup time would simply be $\lceil \tau / T \rceil \times T$, where T is the clock period. An asynchronous SAV can be generated by a separate pulse generation circuit, where τ is controlled by the pulsewidth. In an energy-area-constrained digital system, a synchronous SAV would be preferred because control circuitry would be smaller and consume less power than an on-chip pulse generator. The one disadvantage of using a synchronous SAV is that granularity with which τ can be adjusted is one clock period. Therefore, if the clock period is large, an asynchronous SAV may actually result in lower energy expenditure.

6) Timing of Input-Output Signals: During backup mode, IN1 and IN2 should be differential and stable. No current

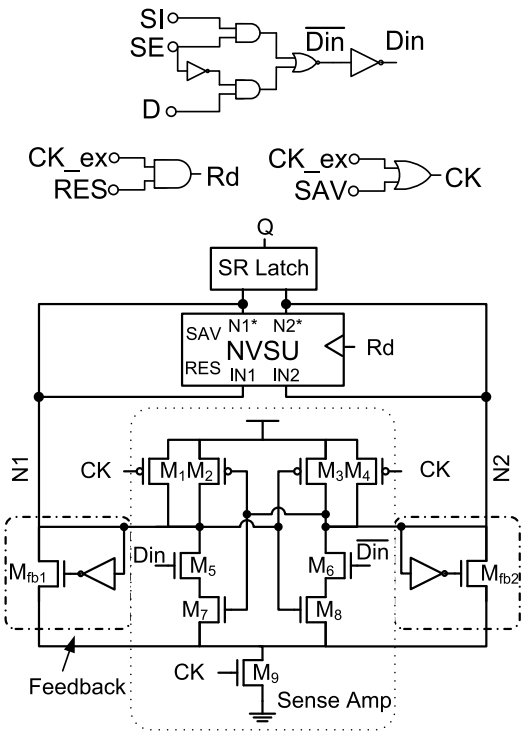


Fig. 8. Schematic of the NVSFF-DM. The tristate buffers between NVSU and SR latch are not shown.

would flow through STT_data if IN1 $= IN2$. If both signals flip, the current direction would change. During the restore mode, the outputs $N1^*$ and $N2^*$ will become differential after the state sense amplifier evaluates, when CK $= 1$ and $Rd = 1$. When CK $= 0$ and $Rd = 0$, both $N1^*$ and $N2^*$ are reset to 1. A latch is required to maintain the evaluation results on the NVSFF outputs when CK is low.

D. Nonvolatile Scan Differential Flip-Flop (NVSFF-DM)

The NVSU takes a pair of differential inputs during the backup mode, and produces a pair of differential outputs during the restore mode. Therefore, the simplest type of flip-flop to interface with the NVSU would be a differential or sense-amp-based flip-flop. Fig. 8 shows such a modified version of the flip-flop [26] interfaced with the NVSU. The combined unit is referred as NVSFF-DM. The circuit includes a differential sense amplifier with its output $N1$ and $N2$ connected to both the SR-latch and the NVSU. The inputs to the SR-latch can be switched from either the sense amplifier or the NVSU outputs. The tristate buffers between SR-latch and the two sources are not shown. In the normal mode, when CK $= 0$, it is easy to verify that $(N1, N2) = (1, 1)$. When CK : $0 \rightarrow 1$, $(N1, N2) = (0, 1)$ or $(N1, N2) = (1, 0)$, depending on the input D. $(N1, N2)$ set the output of SR-latch accordingly. The two feedback loops in Fig. 8 are there to eliminate potential floating nodes that are present in conventional differential flipflops [27]. $(N1, N2)$ become differential and stable after evaluation is completed.

The internal CK is gated by SAV and Rd is gated by RES. SAV ensures that CK remains at 1 during the backup mode, and RES ensures that Rd follows the external clock CK_ex only during restore mode. CK-gating makes sure that $(N1, N2)$ change from $(1, 1)$ to $(0, 1)$ or $(1, 0)$ only once. Rd gating ensures that the state sense amplifier will operate and consume power only during the restore mode. The SR-latch latches the output either from the sense amplifier or the NVSU. The requirements imposed by the NVSU on its inputs and outputs are satisfied with these settings of the NVSFF-DM.

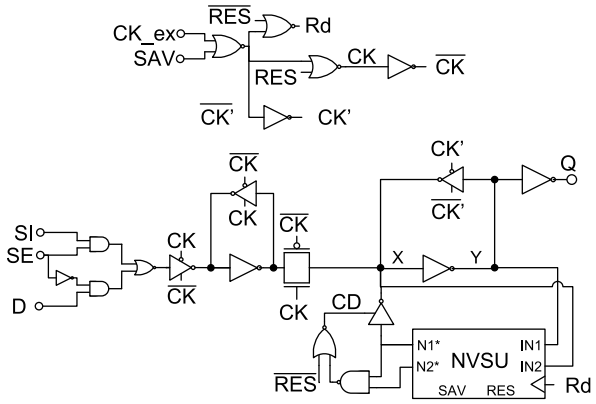


Fig. 9. Schematic of NVSFF-MS.

E. Nonvolatile Scan Master-Slave Flip-Flop (NVSFF-MS)

With some modification, the NVSU can also be combined with a conventional master-slave flip-flop to form a nonvolatile scan master-slave flip-flop (NVSFF-MS). This is shown in Fig. 9. The scan mechanism is the same as in a conventional D-flip-flop. However, the NVSU needs to be properly interfaced with the master and slave latches. The NVSU receives inputs (X and Y) from the slave latch during backup mode and sends its output back to the same node (X) during restore mode. To prevent the NVSU from interfering with the slave latch during normal mode and backup mode, a tristate buffer is used to buffer the output of NVSU. This buffer should be turned on only when NVSU is in the restore mode and its outputs are ready. Since the outputs of NVSU would become differential only when they are ready, a completion detection signal CD is derived from $N1^*$ and $N2^*$ to drive the tristate buffer. Unlike the NVSFF-DM, the slave latch and transmission gate between master and slave latch in NVSFF-MS are driven by different derived clocks derived from the master clock CK_ex. During the restore mode, the transmission gate should be turned off to block the signal from master latch. After the state is restored into the slave latch, the slave latch should be able to latch the data when external clock goes to 0.

The schematic of NVSFF-MS is shown in Fig. 9. In normal mode and normal scan mode, both SAV and RES are 0, NVSFF-MS operates the same as normal scan flip-flop. The internal clock signals CK, CK$'$, $\overline{\text{CK}}$ and $\overline{\text{CK}'}$ follow the external CK_ex under different conditions. CK follows CK_ex when both SAV $= 0$ and RES $= 0$, and CK$'$ follows CK_ex when SAV $= 0$.

Nodes X and Y are fed into NVSU as differential inputs IN1 and IN2. In the backup mode, SAV $= 1$, RES $= 0$.

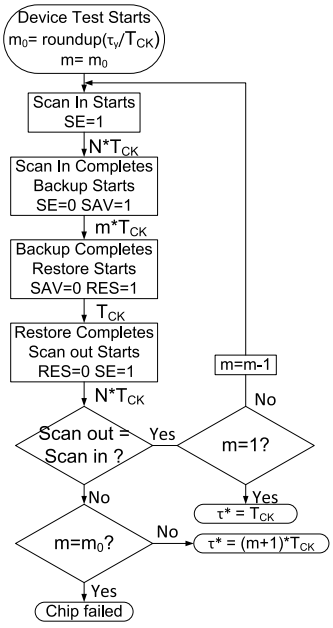


Fig. 10. Nonvolatile scan test procedure. N is number of flipflops in design.

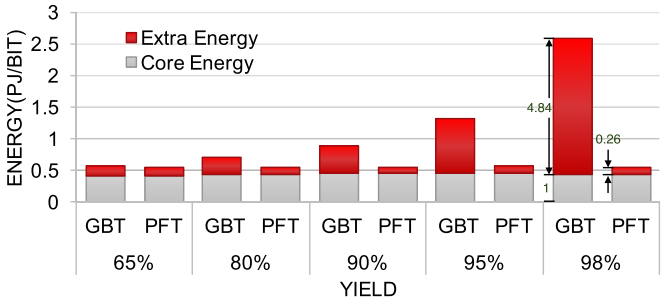


Fig. 11. GBT: Single, global backup time. PFT: postfabrication tuning. Core energy is the same for GBT and PFT. For achieving high yield, the energy wastage with PFT is much less than GBT.

Then, $\mathrm{CK} = \mathrm{CK}' = 1$ and $\overline{\mathrm{CK}} = \overline{\mathrm{CK}'} = 0$. This disconnects the master from its inputs and the slave, so that the value of the master can be saved in the NVSU. $\mathrm{RES} = 0$, $\mathrm{CD} = 0$ blocks $N1^*$ to node X. It ensures that X and Y are kept differential and stable during entire backup mode.

During restore mode, $\mathrm{RES} = 1$, $\mathrm{CK} = 0$, and $\overline{\mathrm{CK}} = 1$. The transmission gate between master and slave latches is blocked. In the meantime, Rd, CK' and $\overline{\mathrm{CK}'}$ follow CK_ex. When $\mathrm{CK}\_ex = 0$, $N1^* = N2^* = 1$, and $\mathrm{CD} = 0$. The slave latches its previous state. When $\mathrm{CK}\_ex : 0 \rightarrow 1$, the state sense amplifier in NVSU sets $N1^*$ and $N2^*$ into opposite values. These two differential signals set $\mathrm{CD} = 1$, which enables the tristate buffer between the NVSU and the slave. The value of $\overline{N1}^*$ is therefore sent to the slave latch to set its output Q.

F. Extension of Scan for Nonvolatile Test

The conventional scan mechanism can be extended to include the test of nonvolatile devices in each NVFF. The test procedure shown in Fig. 10 allows determining the *actual* or *chip-specific* backup time, after fabrication. Procedure EOPTDRIVERSIZEWPR described in Section III returns the nominal driver size that minimizes the average energy, and τ_y, which is the backup time for y percentile of the corresponding population. By definition, setting the backup time for all chips to τ_y would, with high probability, result in $y\%$ of the chips being successfully backed up. However, each specific chip might be successfully backed up with a smaller backup time. This smaller backup time, denoted by τ^*, is computed by using the scan mechanism on each chip. Once it is computed, it can be saved and used for backup whenever required. The energy savings using τ^* versus using τ_y can be substantial.

Fig. 10 shows an outline of the scan procedure to determine τ^*. If τ is the backup time computed by procedure EOPTDRIVERSIZEWPR, then the least number of clock cycles whose total duration exceeds τ is $m(\tau) = roundup(\tau/T_{CK})$. In Fig. 10, this is initialized to $m = m_0 = m(\tau_y)$. Then, data are scanned into all the NVSFFs, and the backup mode is made active (i.e., $\mathrm{SAV} = 1$) for m cycles. Next, a restore is performed, and the data are scanned out. If there are no differences between the data scanned in and scanned out, then m cycles were sufficient. Otherwise m is decremented, the procedure is repeated. If on some iteration, the scanned out values differ from the scanned in values, then the number of cycles was not sufficient. If this happens on the first iteration, where $m = m_0$, then this chip is considered to have not met the yield criterion and deemed to have failed. On the other hand if the error appears on some value of m other than the first, then the previous iteration succeeded, and the minimum backup time is $\tau^* = (m+1)T_{CK}$.

Fig. 11 shows the energy expenditure using the two different backup times: one using τ_y, which is termed as global backup time (GBT), and the other using τ^*, which is termed as postfabrication tuning (PFT). The savings in energy using PFT for a yield of 98% is nearly 80% compared to using GBT. Note that τ^* was computed using procedure EOPTDRIVERSIZEWPR with τ_y in line 18 being replaced by τ_j, and updating E_j only if $\tau_j \leq \tau_y$.

G. Robustness of the Restore Operation

The focus of this paper has been on the energy efficiency and robustness of the write or backup operation, because it results in greater power consumption than the read or restore operation. To read the state of a NVFF, the data have to be sensed and compared with a reference. Hence, process variations can result in a failure of this operation as well. In a NVFF the read circuitry is independent of the driver circuit used for the write operation. Consequently, techniques used to improve the robustness of the NVFF read operation by device parameter optimization [22] or by introducing redundancy [11] can be directly applied to the proposed NVFF design. Note that the proposed postfabrication tuning method shown in Fig. 10 verifies that both the backup and restore operations are successful as it searches from the smallest backup time.

All the components excluding backup and restore circuits are standard CMOS logic blocks, and hence are subject to

TABLE I
STT-MTJ PARAMETERS

Parameter	Value
MgO thickness(μ)	0.8 nm, 0.85nm
Free layer thickness	1.3 nm
Area	$40\ nm \times 40\ nm$
Resistance area product)	$5\ \Omega \cdot \mu m^2$
TMR at zero bias	150 %
STD of variation(σ)	3%, 5%, 10 % [19]
MonteCarlo cases	10000

process variations. Their yields are generally orders of magnitude higher than the STT-MTJ and other emerging devices. Consequently, reduction of yield in the CMOS blocks due to process variations was not considered.

V. EXPERIMENTAL RESULTS

This section contains simulation based evaluations of the proposed NVFF circuits as well as the results on a larger design incorporating the NVSFFs. The circuits were designed using a commercial process design kit for 40-nm GP process. Other standard cells in 40 nm were used in circuit automated synthesis. The power and delay values were obtained using HSPICE.

A. STT-MTJ Cell

The device simulations are based on the models in [13] and [18]. The STT-MTJ has a square shape top view with both width and length equal to 40 nm. Other parameters are shown in Table I. As $t_{\rm ox}$ is the most significant factor on energy consumption, to simplify the analysis, perturbations in $t_{\rm ox}$ are assumed to be Gaussian. To study the impact of the variations in $t_{\rm ox}$ on the resistances of the MTJ, 10 000 Monte Carlo simulations were performed with the mean $\mu_{\rm tox}$ and sigma $\sigma_{\rm tox}$ of $t_{\rm ox}$ set to .8 nm and 10% of mean [19]. Other physical parameters remained constant. Fig. 3(a) shows the distribution of $R_L(0)$ and $R_H(0)$. The mean and sigma of the resistances are summarized in Table II. $I_{c,01}$ is $78.71\mu A$ and $I_{c,10}$ is $27.77\mu A$. If a single power supply is used in NVSFF design, the maximum voltage drop cross MTJ could not exceed than its V_{dd}, which is 0.9 V in used 40-nm technology. Therefore, the maximum resistance can be calculated as

$$R_{H,\max} = V_{dd}/I_{c,10} = 32.4\ k\Omega$$
$$R_{L,\max} = V_{dd}/I_{c,01} = 11.43\ k\Omega.$$

Table II shows the mean and standard deviation of resistances for two different mean values of $t_{\rm ox}$. A smaller $t_{\rm ox}$ is preferred to ensure that the 3σ of R_L and R_H are below the maximum resistances dictated by the power supply. Based on Table II, $\mu_{\rm tox} = 0.8$ nm and $\sigma_{\rm tox}/\mu_{\rm tox} = 0.1$ is assumed.

B. Performance Evaluation of Proposed NVSFFs

Table III shows the delay and the energy delay product of the two NVSFF as well as a volatile master-slave scan flip-flop (SFF-MS) designs. The setup time ($T_{\rm setup}$) of the NVSFF-DM is negative, in contrast to the positive setup time

TABLE II
MEAN AND STANDARD DEVIATIONS OF STT-MTJ RESISTANCES VERSUS $t_{\rm ox}$. THE MEAN OF RANDOM VARIABLE $t_{\rm ox}$ IS SET TO TWO VALUES, .85 NM AND .8 NM, WITH SIGMA EQUAL TO 3%, 5%, AND 10% OF $\overline{t}_{\rm ox}$

μ_{tox} (nm)	σ_{tox} (%)	μ_{RH} ($k\Omega$)	σ_{RH} ($k\Omega$)	μ_{RL} ($k\Omega$)	σ_{RL} ($k\Omega$)
0.85	10%	9.59	6.91	3.84	2.76
	5%	8.23	2.74	3.29	1.09
	3%	7.96	1.57	3.18	0.62
0.8	10%	6.39	4.33	2.56	1.73
	5%	5.57	1.75	2.23	0.70
	3%	5.41	1.01	2.16	0.40

TABLE III
PERFORMANCE OF NVSFF-MS, NVSFF-DM, AND SFF-MS. THE AVERAGE ENERGY IS BASED ON 30% INPUT SWITCHING ACTIVITY. SIMULATION CONDITIONS ARE: 25 °C, 0.9 V, TT CORNER, AND OUTPUT LOAD OF 3 fF

	T_{C2Q} (ps)	T_{setup} (ps)	T_{total} (ps)	Energy (fJ/cyc)	EDP (fJ·ps)
NVSFF-MS	60.28	6.90	67.18	4.10	275.56
NVSFF-DM	46.99	-2.99	44.00	5.99	263.51
SFF-MS	38.08	16.74	54.82	2.218	121.59

TABLE IV
DELAY AND ENERGY OF RESTORE FROM NVSU TO OUTPUT

	Recover '0'		Recover '1'	
	Delay (ps)	Energy (fJ/bit)	Delay (ps)	Energy (fJ/bit)
NVSFF-MS	107.7	15.92	142.3	13.5
NVSFF-DM	83.87	17.75	82.84	19.41

of the NVSFF-MS. Hence, the total delay of NVSFF-DM is less than that of a NVSFF-MS. Compared to the NVSFF-MS, the average energy consumption (measured with 30% input switching activity) is higher in NVSFF-DM, but the EDP is similar due to the lower total delay of the NVSFF-DM. The total delay of SFF-MS is between the two NVSFFs, but its energy and EDP are much less than both NVSFFs. The area overhead of the NVSU in the NVSFF-DM and NVSFF-MS, makes their size about twice that of the SFF-MS. However, this does not translate to a similar increase in area of a whole circuit with either of the NVSFF cells (see Table VII).

A reference MTJ (STT_ref) is required in the state sense amplifier (see Fig. 6). The resistance of STT_ref is between R_H and R_L. Since the state recovery is implemented by the sensing current flow, $R_{\rm ref}$ is set to be harmonic mean of R_H and R_L. $1/R_{\rm ref} = 2(1/R_H + 1/R_L)$. The resistance of STT_ref is achieved by changing the dimension of the MTJ to 55 nm $\times$ 50 nm, and $R_{\rm ref}$ is 3.09 $k\Omega$. The recovery time of two designs are shown in Table IV. In this paper, global perturbations in $t_{\rm ox}$ are the most significant source of variations in the device resistances. Therefore, relative differences between $R_{\rm ref}$ and R_H/R_L would remain constant on a die.

Table V shows a comparison of NVSFF-MS and NVSFF-DM with published data on two other designs. The setup time and delay of the sense amplifier based NVFF (SA-MFF) in [7] are similar to the NVSFF-DM. Although the forward body bias (FBB) feature of fully depleted silicon-on-insulator (FDSOI) can improve the energy delay product,

TABLE V
COMPARISON OF NONVOLATILE FLIP-FLOP WITH PRIOR REPORTED DATA

	NVSFF-MS	NVSFF-DM	Ref. [7]	Ref. [5]
Tech	40nm	40 nm	28 nm FDSOI	45 nm
T_{setup}	6.9 ps	-3.0 ps	-4.9ps	75.2 ps
T_{C2Q}	60.3 ps	47.0 ps	50.1	203.3 ps
Backup time	Tunable		Fixed	Fixed
Backup energy	504fJ/bit		N/A	N/A
Restore time	142.3ps	83.9ps	N/A	2.01 ns
Restore energy	15.92 fJ/bit	19.41 fJ/bit	N/A	170.9 fJ/bit

TABLE VI
COMPARISON OF BACKUP SCHEMES. (A) AND (B) USE SINGLE BACKUP TIME FOR ALL DICE, AND (C) REFERS TO CHIP-SPECIFIC BACKUP TIME. (B) AND (C) INCLUDE VARIATIONS IN BOTH CMOS AND MTJ

	Yield	Driver Size	τ (ns)	Energy (pJ/bit)
(A) No Variation	100%	107.5	2.17	0.367
(B) Global Backup Time	97%	20.9	14.6	1.811
(C) Post Fab. Tuning	97%	32.8	1.96-12.84	0.504

the SA-MFF uses a fixed write pulse for backup, which has a significantly high failure rate (24.6%) due to MTJ variations. The NVFF in [5] has a large positive setup time, and exhibits a dc current during a read operation.

Table VI shows the energy consumption of NVSU during the backup mode. Three driver sizes were examined to evaluate their effects on the energy consumption. The driver sizes were determined based on method described in Section III. Ignoring variations, the minimum energy is achieved with the largest driver size (107.5). When both CMOS and MTJ variations are included, the single global backup time $\tau_{97} = 14.6$ ns, whereas the chip-specific backup times ranged from 1.96 to 12.84 ns (over 10K samples). However, the energy expenditure of the former was more than $3.5X$ than the latter. Moreover the sizing and PFT approach results in an energy expenditure that is close to the ideal case with no variations.

C. Performance Evaluation of Circuits

Both NVSFFs are characterized using a standard characterization tool. To demonstrate the performance impact of NVSFFs on larger circuits, two circuits, an 8-b multiply-and-accumulate (MAC) unit, and a 32-b adder were synthesized using the two different NVSFFs and a SFF-MS.

1) MAC Unit: The circuit structure is shown in Fig. 12. The MAC unit was synthesized using Genus from Cadence, with two different combinations of standard cells: 1) standard logic with NVSFF-MSs and 2) standard logic with NVSFF-DMs. Note that the total number of flipflops (16 input and output) in both designs is the same, and both were synthesized for the same target clock period of 1.835 ns.

Table VII shows of the results of the synthesis. The column *Cell Count* indicates the total number of standard cells. The designs with NVSFF-DMs have 11.6% fewer cell counts and 16% less area compared with the one with NVSFF-MSs.

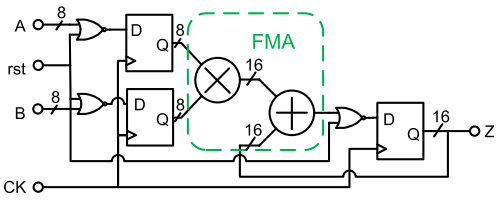


Fig. 12. 8-b MAC unit. It includes input and output flipflops, a synchronous reset, and fused multiply-add (FMA) unit.

TABLE VII
COMPARISON OF LOGIC CELL COUNT AND AREA USING DIFFERENT FLIPFLOPS IN MAC AND ADDER

	MAC unit		32-bit Adder	
Flipflop Type	Cell Count	Area (μm^2)	Cell Count	Area (μm^2)
NVSFF-MS	603	3040	482	2517
NVSFF-DM	533	2555	465	2342
SFF-MS	580	2795	477	2098

Even though NVSFF-DM consumes greater power, its smaller (negative) setup time allows the synthesis tool to reduce the logic cone driving the flip-flop to a greater degree than in the case of the NVSFF-MS.

Power estimation was done by PTPX from Synopsys, using the library characterized data. Input sequences with 10%, 20%, and 30% switching activities were supplied to the circuit. The average energy was measured by averaging the energy consumption across more than 100 cycles. Fig. 13(a) shows the total energy consumption of the two circuits versus input switching activity. The NV-MAC unit with NVSFF-DM consumed about 18.7%, 18.9%, and 19% less energy than the NV-MAC unit with NVSFF-MS. As with delay (see Table III), both area and energy consumption of the MAC with SFF-MS are between those with NVSFF-DM and NVSFF-MS.

2) 32-b Adder: Two 32-b adders are designed and synthesized in the same way as the MAC unit. There are 97 flipflops in the design. The synthesized results are shown in Table VII. The design with NVSFF-DMs has only 3.5% fewer cells and 7% smaller area than the one with NVSFF-MSs. The energy consumption with three switching activities are also very close, about 0.9%, 5.8%, and 7.2% fewer on NVSFF-DMs, shown in Fig. 13(b). Compared with the MAC unit, the 32-b adder has fewer logic cells and more flipflops. The NVSFF-DM has lower total delay (setup plus clock-to-Q) but slightly higher power consumption than the NVSFF-MS. The reduced delay allows synthesis tools to absorb the extra slack by reducing the size of the logic cone driving the flip-flop. Note that for the 32-b adder, the reduction in the size of its logic cones when using NVSFF-DM was not sufficient to compensate for it larger power consumption due to its greater number of flipflops. Since SFF-MS is smaller than NVSFFs, the total area of the adder with SFF-MS is 10.4% and 16.6% smaller than the one with NVSFF-DMs and NVSFF-MSs, respectively.

VI. LITERATURE REVIEW

Several different circuit architectures for NVFFs appear in [3]–[5], [7]–[9], and [11]. The earlier efforts [3], [4] using FeRAMs reported substantial penalties in area

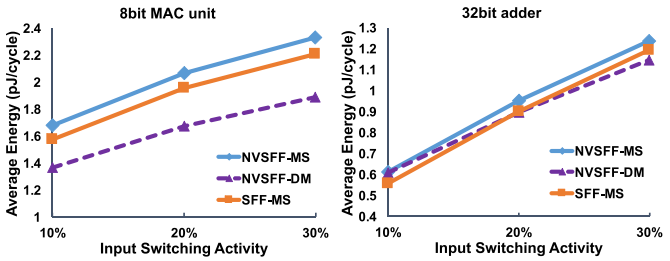


Fig. 13. Total energy versus input switching activity under normal operation. The simulation is done by PTPX under 25 °C typical corner.

(10X larger than regular flip-flop), performance (delays in $\sim\mu$s) and energy. The emerging spin-based MTJ devices such as STT, spin orbit torque (SOT), or programmable metallization cell (PMC) with high density, low switching energy, and fast switching times, are promising candidates for NVM and NVL.

References [5], [7], [9], [11] describe the design of a NVFF with STT-MTJ. Ryu *et al.* [5] focuses on the design of the write circuit to provide higher driving current thereby reducing the backup time. Bishnoi *et al.* [11] uses redundancy in the cell design to tolerate a single MTJ fault. It improves the robustness of the restore operation at the cost of doubling backup energy. The focus is on tolerating single failures, without considering the design of the driver circuitry, which would have a significant impact on the energy, performance and yield. Mahalanabis *et al.* [8] describes a PMC-based NVFF. The high ratio of the high-resistance to low-resistance states of the CBRAM improve robustness of the restore operation at low voltages.

A sense-amplifier-based NVFF using STT-MTJ and FDSOI technology is described in [7]. The FBB provided by FDSOI improves both speed and energy consumption. The robustness of the backup and restore operations are improved by the increasing supply voltage(VCMA effect). Kang *et al.* [9] shows that this can reduce the backup energy by 98.4% and 74.6% when compared to a NVFFs with STT-MTJ and SHE devices.

The discovery of SOT switching provides a more efficient way to reverse magnetization [28]. Compared with an STT-MTJ, SOT switching is faster [29], and the three terminal structure allows for separate optimization of the write and read paths. Kwon *et al.* [6] and Bishnoi *et al.* [10] describe SOT-MTJ-based NVFF designs, showing that they have the potential for higher speed, lower energy, and higher reliability than STT-MTJ devices. The design optimization and circuit architectures presented herein can easily be adapted to SOT-MTJ devices.

Techniques such as those in [30]–[32] address robust design of NVM in the presence of process variations. Zhao *et al.* [33] and Cai *et al.* [34] focus on the logic in NVM design. These techniques are generally not well-suited for the design of NVFF due to the complexity of circuits and, in some cases, the use of analog components [35].

VII. Conclusion

The key components in NVL are the flipflops that represent the state of the system at any given time. For near instantaneous backup and restoration of the state, it is best to enhance the flipflops with NV storage. The optimal design of the driver circuit to save the state in a NV device is critically important for energy efficiency and robustness due to process variations. This paper presented a systematic approach to the energy optimal design of the backup driver and the determination of the corresponding backup subject to satisfying a yield constraint. To further reduce energy wastage, a novel method is presented that adjusts the backup time on a per-chip basis, after fabrication. This substantially reduces the energy wasted when compared to using a single backup time for all chips. Also included is the design of NVFFs that enables the postfabrication tuning of the backup time through the use of a scan mechanism. Significant energy reduction with postfabrication tuning is demonstrated both in theory and in two circuit implementations: a 32-b adder and a 8-b MAC unit. The proposed methodology allows conversion of any ASIC design to one that is completely nonvolatile using commercial synthesis flows.

References

[1] S. Priya and D. J. Inman, *Energy Harvesting Technologies*, 1st ed. New York, NY, USA: Springer, 2008.

[2] S. Khanna, S. C. Bartling, M. Clinton, S. Summerfelt, J. A. Rodriguez, and H. P. McAdams, "An FRAM-based nonvolatile logic MCU SoC exhibiting 100% digital state retention at VDD = 0V achieving zero leakage with <400-ns wakeup time for ULP applications," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 95–106, Jan. 2014.

[3] M. Koga *et al.*, "First prototype of a genuine power-gatable reconfigurable logic chip with FeRAM cells," in *Proc. Int. Conf. Field Program. Logic Appl.*, Aug./Sep. 2010, pp. 298–303.

[4] Y. Wang *et al.*, "A 3us wake-up time nonvolatile processor based on ferroelectric flip-flops," in *Proc. ESSCIRC (ESSCIRC)*, Sep. 2012, pp. 149–152.

[5] K. Ryu, J. Kim, J. Jung, J. P. Kim, S. H. Kang, and S.-O. Jung, "A magnetic tunnel junction based zero standby leakage current retention flip-flop," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 11, pp. 2044–2053, Nov. 2012.

[6] K.-W. Kwon, S. H. Choday, Y. Kim, X. Fong, S. P. Park, and K. Roy, "SHE-NVFF: Spin Hall effect-based nonvolatile flip-flop for power gating architecture," *IEEE Electron Device Lett.*, vol. 35, no. 4, pp. 488–490, Apr. 2014.

[7] H. Cai, Y. Wang, W. Zhao, and L. A. B. Naviner, "Multiplexing sense-amplifier-based magnetic flip-flop in a 28-nm FDSOI technology," *IEEE Trans. Nanotechnol.*, vol. 14, no. 4, pp. 761–767, Jul. 2015.

[8] D. Mahalanabis, V. Bharadwaj, H. J. Barnaby, S. Vrudhula, and M. N. Kozicki, "A nonvolatile sense amplifier flip-flop using programmable metallization cells," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 5, no. 2, pp. 205–213, Jun. 2015.

[9] W. Kang, Y. Ran, W. Lv, Y. Zhang, and W. Zhao, "High-speed, low-power, magnetic non-volatile flip-flop with voltage-controlled, magnetic anisotropy assistance," *IEEE Magn. Lett.*, vol. 7, pp. 1–5, 2016.

[10] R. Bishnoi, F. Oboril, and M. B. Tahoori, "Non-volatile non-shadow flip-flop using spin orbit torque for efficient normally-off computing," in *Proc. 21st Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2016, pp. 769–774.

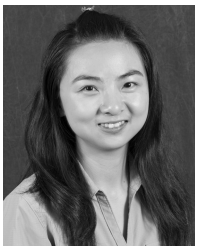
[11] R. Bishnoi, F. Oboril, and M. B. Tahoori, "Design of defect and fault-tolerant nonvolatile spintronic flip-flops," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 4, pp. 1421–1432, Feb. 2017.

[12] Y. Zhang *et al.*, "Compact modeling of perpendicular-anisotropy CoFeB/MgO magnetic tunnel junctions," *IEEE Trans. Electron Devices*, vol. 59, no. 3, pp. 819–826, Mar. 2012.

[13] K. Munira, W. H. Butler, and A. W. Ghosh, "A quasi-analytical model for energy-delay-reliability tradeoff studies during write operations in a perpendicular STT-RAM cell," *IEEE Trans. Electron Devices*, vol. 59, no. 8, pp. 2221–2226, Aug. 2012.

[14] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. B. Tahoori, "Improving write performance for STT-MRAM," *IEEE Trans. Magn.*, vol. 52, no. 8, Aug. 2016, Art. no. 3401611.

[15] Y. Wang, Y. Zhang, E. Y. Deng, J. O. Klein, L. A. B. Naviner, and W. S. Zhao, "Compact model of magnetic tunnel junction with stochastic spin transfer torque switching for reliability analyses," *Microelectron. Rel.*, vol. 54, nos. 9–10, pp. 1774–1778, Sep./Oct. 2014.
[16] J. Z. Sun *et al.*, "Sullivan, W. J. Gallagher, and D. C. Worledge, "Effect of subvolume excitation and spin-torque efficiency on magnetic switching," *Phys. Rev. B, Condens. Matter*, vol. 84, no. 6, p. 064413, 2011. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.84.064413
[17] N. Xu *et al.*, "Physics-based compact modeling framework for state-of-the-art and emerging STT-MRAM technology," in *IEDM Tech. Dig.*, Dec. 2015, pp. 28.5.1–28.5.4.
[18] Y. Zhang *et al.*, "Compact model of subvolume MTJ and its design application at nanoscale technology nodes," *IEEE Trans. Electron Devices*, vol. 62, no. 6, pp. 2048–2055, Jun. 2015.
[19] 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 *IEDM Tech. Dig.*, Dec. 2009, pp. 1–4.
[20] M. Wirnshofer, *Variation-Aware Adaptive Voltage Scaling for Digital CMOS Circuits* (Springer Series in Advanced Microelectronics). Amsterdam, The Netherlands: Springer, 2013, ch. 2.
[21] S. Motaman, S. Ghosh, and N. Rathi, "Impact of process-variations in STTRAM and adaptive boosting for robustness," in *Proc. Design, Autom. Test Europe Conf. Exhib. (DATE)*, Mar. 2015, pp. 1431–1436.
[22] Y. Zhang, X. Wang, H. Li, and Y. Chen, "STT-RAM cell optimization considering MTJ and CMOS variations," *IEEE Trans. Magn.*, vol. 47, no. 10, pp. 2962–2965, Oct. 2011.
[23] Y. Halawani, B. Mohammad, D. Homouz, M. Al-Qutayri, and H. Saleh, "Modeling and optimization of memristor and STT-RAM-based memory for low-power applications," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 3, pp. 1003–1014, Mar. 2016.
[24] K. Ma *et al.*, "Architecture exploration for ambient energy harvesting nonvolatile processors," in *Proc. IEEE 21st Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2015, pp. 526–537.
[25] D. Balsamo, A. S. Weddell, G. V. Merrett, B. M. Al-Hashimi, D. Brunelli, and L. Benini, "Hibernus: Sustaining computation during intermittent supply for energy-harvesting systems," *IEEE Embedded Syst. Lett.*, vol. 7, no. 1, pp. 15–18, Mar. 2015.
[26] J. Yang, N. Kulkarni, J. Davis, and S. Vrudhula, "Fast and robust differential flipflops and their extension to multi-input threshold gates," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2015, pp. 822–825.
[27] N. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*. London, U.K.: Pearson Education, 2005.
[28] I. M. Miron *et al.*, "Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection," *Nature*, vol. 476, no. 7359, pp. 189–193, Aug. 2011.
[29] K. Garello *et al.*, "Ultrafast magnetization switching by spin-orbit torques," *Appl. Phys. Lett.*, vol. 105, no. 21, p. 212402, Nov. 2014.
[30] H. Yu and Y. Wang, *Design Exploration of Emerging Nano-scale Non-volatile Memory*, 1st ed. New York, NY, USA: Springer-Verlag, 2014.
[31] W. Kang, L. Zhang, J.-O. Klein, Y. Zhang, D. Ravelosona, and W. Zhao, "Reconfigurable codesign of STT-MRAM under process variations in deeply scaled technology," *IEEE Trans. Electron Devices*, vol. 62, no. 6, pp. 1769–1777, Jun. 2015.
[32] S. Wang, H. Lee, F. Ebrahimi, P. K. Amiri, K. L. Wang, and P. Gupta, "Comparative evaluation of spin-transfer-torque and magnetoelectric random access memory," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 6, no. 2, pp. 134–145, Jun. 2016.
[33] W. Zhao *et al.*, "Synchronous non-volatile logic gate design based on resistive switching memories," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 61, no. 2, pp. 443–454, Feb. 2014.
[34] H. Cai, Y. Wang, L. A. De Barros Naviner, and W. Zhao, "Robust ultra-low power non-volatile logic-in-memory circuits in FD-SOI technology," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 4, pp. 847–857, Apr. 2017.
[35] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. B. Tahoori, "Read disturb fault detection in STT-MRAM," in *Proc. Int. Test Conf.*, Oct. 2014, pp. 1–7.
[36] N. Kulkarni, J. Yang, J.-S. Seo, and S. Vrudhula, "Reducing power, leakage, and area of standard-cell asics using threshold logic flip-flops," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 9, pp. 2873–2886, Sep. 2016.
[37] J. Yang, N. Kulkarni, S. Yu, and S. Vrudhula, "Integration of threshold logic gates with RRAM devices for energy efficient and robust operation," in *Proc. IEEE/ACM Int. Symp. Nanosc. Archit. (NANOARCH)*, Jul. 2014, pp. 39–44.



Jinghua Yang (M'15) received the B.S.E.E. degree from Harbin Institute of Technology, Harbin, China, in 2007 and the M.S.E.E. degree from Xi'an Jiaotong University, Xian, China, in 2010. She is currently working toward the Ph.D. degree at the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ, USA.

From 2008 to 2010, she was with the Institute of Microelectronics of Chinese Academy of Sciences, Beijing, China, focusing on analog IC design. Her current research interests include high-performance low-power digital circuits, threshold logic circuit and algorithms, emerging device technology, and memory devices.

Aykut Dengi received the B.S. degree in electrical and computer engineering from Bilkent University, Ankara, Turkey, in 1992 and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 1997.

He was the Founder of SystemIC, Phoenix, AZ, USA, where he researched and developed an analysis and design tool for communication systems. He led the Advanced Research and Development Group, Cadence Design Systems Inc., where he was involved in electronic design automation (EDA) algorithms and software, in particular for RF, analog, and mixed-signal integrated circuits (ICs). He was a Principle Investigator for the UltraSYN project under the Defense Advanced Research Projects Agency NeoCAD program, where he led the development of a complete schematic-to-layout synthesis and verification system for RF ICs. From 1994 to 2001, he was with Motorola Inc., Austin, TX, USA, focused on interconnect and passive device modeling and simulation, signal integrity verification for high-performance digital and mixed-signal ICs, electromagnetic simulation, and design flows for RF ICs. He is currently an Associate Research Professor at Arizona State University, Tempe, AZ, USA, and the Director of the Internet of Things Collaboratory. His current research interests include EDA, low-power high-performance digital designs, analog, RF and mixed-signal design, and computational electromagnetics.

Sarma Vrudhula (M'85–SM'02–F'16) received the B.Math. degree from the University of Waterloo, Waterloo, ON, Canada and the M.S.E.E. and Ph.D. degrees in electrical and computer engineering from the University of Southern California, Los Angeles, CA, USA.

He was a Professor at the ECE Department, University of Arizona, Tucson, AZ, USA, and was on the faculty of the EE-Systems Department at the University of Southern California. He was also the Founding Director of the NSF Center for Low Power Electronics at the University of Arizona.

He is a Professor of Computer Science and Engineering with Arizona State University, Tempe, AZ, USA, and the Director of the NSF I/UCRC Center for Embedded Systems. His research interests include design automation and computer aided design for digital integrated circuit and systems; low-power circuit design; energy management of circuits and systems; energy optimization of battery powered computing systems, including smartphones, wireless sensor networks, and Internet of Things systems that relie energy harvesting; system level dynamic power and thermal management of multicore processors and system-on-chip (SoC); statistical methods for the analysis of process variations; statistical optimization of performance, power, and leakage; a new circuit architectures of threshold logic circuits for the design of ASICs and field-programmable gate arrays; nonconventional methods for implementing logic, including technology mapping with threshold logic circuits; the implementation of threshold logic using resistive memory devices; and the design and optimization of nonvolatile logic.