Reachability-Based Robustness Verification and Optimization of SRAM Dynamic Stability Under Process Variations

Yang Song, Hao Yu, *Senior Member, IEEE,* and Sai Manoj Pudukotai DinakarRao, *Student Member, IEEE*

***Abstract*—The dynamic stability margin of SRAM is largely suppressed at nanoscale due to not only dynamic noise but also process variation. This paper introduces an analog verification for SRAM dynamic stability under threshold-voltage variations. A zonotope-based reachability analysis by the backward Euler method is deployed for SRAM dynamic stability in state space with consideration of SRAM nonlinear dynamics. It can simultaneously consider multiple SRAM variation sources without multiple repeated computations. What is more, sensitivity analysis is developed for zonotope to optimize SRAM designs departing from unsafe regions by simultaneously tuning multiple SRAM device parameters. In addition, compared to the SRAM optimization by single-parameter small-signal sensitivity, the proposed method can converge faster with higher accuracy. As shown by numerical experiments, the proposed optimization method can achieve 600× speedup on average when compared to the repeated Monte Carlo simulations under the similar accuracy.**

***Index Terms*—Design for manufacturability, memory, mixed-mode, performance optimization, simulation, transistor-sizing.**

I. Introduction

ROBUSTNESS verification and optimization have become an emerging need for integrated circuit (IC) designs at nano-scale such as SRAMs. Static noise margin (SNM) [1], [2] is traditionally deployed for SRAM stability characterization because of its simple interpretation and measurement. As it may overestimate read-failure and underestimate write-failure, dynamic stability margin [3] is increasingly adopted by deploying critical word-line pulse-width that can produce a better estimation of failures. However, the verification of SRAM stability margin becomes even harder at nano-scale. Firstly, due to the nonlinear dynamics, the SRAM characteristic behavior becomes not digital but more analog. Moreover, process variations such as threshold-voltage variations [4]–[13] can further significantly suppress the SRAM stability margin, and hence result in higher failure rate during read/write operations.

The robustness verification and further optimization of SRAMs have become thereby necessary to provide designers a close scrutiny of potential hazards, such as threshold-voltage variations from all transistors. The primary challenges in traditional approaches for robustness verification and optimization are the complexity to deal with multiple dimensions of variation sources and device parameters. No matter the deterministic corner analysis or the statistical Monte Carlo analysis, the fundamental problem of complexity is from many repeated computations performed for each different condition of variation source and parameter. There is a need to efficiently report the status of failure with consideration of multiple variation sources just by one verification simultaneously. Moreover, a verification-driven optimization is also required, which can guide designs departing from unsafe region by further tuning multiple device parameters at the same time.

Many works are performed from statistical perspective [14], [15]. Based on dc characteristics of inverters, stability analysis has been performed in [14] by modeling failure with normal distribution even when failure occurs in tail of normal distribution. In [15], accurate estimation is achieved without the assumption of normal distribution of failure probability but is based on the most probable failure point searching. Moreover, importance sampling is utilized to avoid the prohibitive Monte Carlo simulation by only capturing relevant rare event in. A number of recent works have been performed from deterministic perspective as well [16]–[20]. For example, Euler–Newton curve-tracing [17] is utilized to find the boundary between the safe and unsafe regions without brute-force exploration. The work in [20] further formulates a dynamic stability margin to characterize the stability boundary, namely, the separatrix [18]. But, the search for boundary is limited to two parameters, and the computational cost can be prohibitive when considering parameter variations from all transistors. What is more, it is unclear how to perform parameter adjustment for SRAM robustness optimization that can help designs depart from unsafe regions.

Reachability analysis has been widely deployed in verification of system dynamics by exploring potential trajectories of operating points in state space. It can conveniently provide accurately predicted boundary of multiple trajectories under uncertain inputs and/or interval parameters by one-time computation, in contrast to simulate different trajectories one by one to explore. The reachability analysis has been deployed for a number of hard analog circuit verifications [21]–[26]. Starting from a set of uncertain inputs, the set of system

Manuscript received June 22, 2013; revised November 13, 2013 and January 17, 2014; accepted January 23, 2014. Date of current version March 17, 2014. The work was supported by the Singapore MOE Tier-1 funding RG 26/10. The preliminary result was published in ISPD'13. This paper was recommended by Associate Editor C. Sze.

The authors are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: haoyu@ntu.edu.sg).

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/TCAD.2014.2304704

trajectories in state space can be bounded by zonotope-based over-approximation [27], [28]. One can perform time-interval integrated reachability analysis with formed zonotope that can distinguish safe and unsafe regions at final set. As such, one can verify failure of system trajectory in state space without multiple repeated simulations. What is more, one can also formulate the set when adjusting multiple device parameters by zonotope approximation, and thereby further optimize the system trajectory to depart from the unsafe region to the safe region. However, the limitations of previous zonotope-based reachability analysis are mainly twofold. First, explicit time-interval integration is computationally expensive when considering nonlinearity during SRAM verifications. It is unknown how to develop a SPICE-like implicit time-interval integration of zonotope-based reachability analysis with consideration of linearization error update. Second, the zonotope-based sensitivity analysis is different from the traditional single-parameter small-signal sensitivity. One need to explore the set based sensitivity in term of distance to the safe/unsafe region inside a sequential of verifications based optimization.

In this paper, we introduce a zonotope-based reachability analysis for both verification and optimization of SRAM dynamic stability. The formulation of zonotope is based on over-approximation defined by a hypercube. Alternatively, sensitivity analysis of zonotope with respect to multiple device parameters is also derived to guide SRAM optimization that can depart from unsafe region. As a summary, there are two primary contributions of this paper. Firstly, to consider SRAM nonlinear dynamics, a backward Euler method is developed for the zonotope-based reachability analysis with linearization error control, which can efficiently consider multiple variation sources for the SRAM robustness verification. Secondly, a zonotope-based sensitivity analysis is introduced for safety distance, which can generate multiparameter large-signal sensitivity for the SRAM robustness optimization. The proposed verification and optimization procedures are both implemented in a SPICE-like simulator with nonlinear device model of transistors. Moreover, as multiple-parameter large-signal sensitivity is generated for safety distance, compared to the traditional single-parameter small-signal based sensitivity optimization, the proposed method can converge fast with high accuracy. Compared to Monte Carlo optimization, the proposed method can achieve speedups up to 600× with similar accuracy.

The rest of this paper is organized as follows. Section II reviews the SRAM failure mechanisms with consideration of threshold-voltage variations. Section III describes the zonotope-based based nonlinear reachability analysis, which is further deployed in the robustness optimization with safety distance sensitivity calculation for SRAM dynamic stability in Section IV. The proposed method is validated by experiments in Section V for different SRAM malfunctions including write and read failures. Conclusions are drawn in Section VI.

II. Problem Formulation of SRAM Failure Verification and Optimization

Similar to [16]–[20], the scope of this paper focuses on the transistor-level analytical approaches for the SRAM dynamic stability verification and optimization under threshold-voltage (V_{th}) variations. When statistical distribution of V_{th} variation is known, one can efficiently generate yield statistics from the transistor-level verification results. There exists serious concern of 6T-SRAM with V_{th} variation [5], [6]. The resulting mismatch among transistors can lead to SRAM functional failures during read and write operations. What is more, though transistor sizing may compensate the negative impact of V_{th} variations, it is unknown how to adjust transistor size for robustness optimization for the sake of SRAM dynamic stability.

In this section, we introduce the definition to quantitatively describe the robustness of SRAM dynamic stability.

Definition 1: Safety distance is the Euclidean distance $||p_{safe} - x||_2$ in the state space between operating point x and the safe state p_{safe}.

The operating point x or safe state p_{safe} depends on both the time instant t and input stimulus u. Hence, the safety distance $||p_{safe} - x||_2$ is a function of both time instant t and input stimulus u. The input corners of the state space are employed to ensure the safety for the possible input stimulus. The idea of zonotope-based reachability verification in the state space is based on the fact that the instability hazards can be visualized as unsafe regions and the Euclidean distance to safe region becomes a measure of system safety. From this perspective, it is the first time in literature to deploy Euclidean distance to describe the safety distance for verification and optimization by this paper.

Note that different from separatrix based approaches [18], [20], the safety distance provides indication on the optimization direction of trajectory. As such, it can be conveniently leveraged within reachability analysis to consider parameter and also input variations at the same time by performing one-time transient simulation.

A. Failure Mechanisms

In the following, we describe physical mechanisms of SRAM failures, including read and write failures in terms of safety distance in the state space. In addition, recall that there exist two convergent regions in the state space of SRAM [18]. Operating points on either region converge to the nearest equilibrium state.

1) *Write Failure:* A write failure refers to the inability to write data properly into the SRAM cell. During write operation, both access transistors should be strong enough to change the voltage level at internal nodes. As shown in Fig. 1, write operation can be described in the state space as the procedure of pulling the operating point from initial state (bottom right corner) to the target state (top left corner). Thus, the safety distance refers to the distance between one operating point and the target state at top-left corner. Given enough time, the operating point will converge to the nearest stable equilibrium state either at top-left or bottom-right corner. The write operation is aimed at pulling operating point into the same region with the target state and thus helping the safety distance converge to zero, as shown by point B in Fig. 1.

The Vth variations, however, may cause write failure. An increase in Vth can reduce the strength of the transistor. For

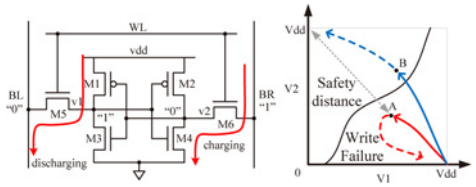

Fig. 1. Illustration of write failure by safety distance.

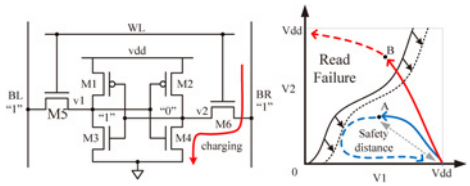

Fig. 2. Illustration of read failure by safety distance.

example, increase of Vth in M6 along with the decrease of Vth in M4 can make it more difficult to pull up v2. In the state space, the operating point moves slowly toward the target state under this condition. If operating point cannot reach other convergent region before access transistors are closed, it will move back to the initial state which means write failure happens. To resolve the failure, tuning width of M6 can be increased while narrowing M4 can help to reduce safety distance and hence can mitigate the side effect from Vth variations.

2) *Read Failure:* A read failure refers to the loss of previously stored data when SRAM flips to the other state during read operation. Access transistors need careful sizing such that their pull-up strengths are not strong enough to pull digital 0 to 1 or vice-versa during read operation. In the state space, one operating point of SRAM is inevitably perturbed and pulled toward the other convergent region. In this situation, the safety distance is from the operating point to its initial state. If read operation does not last too long, access transistors can be shut down before the operating point converges to the other region. The safety distance will converge to zero as the operating point returns to the initial state in the end, as shown by point A in Fig. 2.

The Vth variations may also cause read failure. For example, variations caused by mismatch between M4 and M6 can result in unbalanced pulling strengths and v2 can be pulled up more quickly. As a result, the operating point crosses to the other region before read operation ends with failure, as shown by point B in Fig. 2. To resolve the read failure, width of M6 needs to be scaled down to avoid excessive pulling strength. However, this may lead to write failure as illustrated in previous subsection. In addition, Vth variations in M1–4 affect the locations of converging regions on the state-variable plane. As the opposite converging region migrates closer to the initial state, it becomes more likely for read failure to happen.

Therefore, a full and fast verification considering Vth variations from all transistors are needed to include all potential hazards, which will be done by our reachability-based method. Another problem addressed in this paper is to find an appropriate combination of sizing from all transistors to optimize the robustness of SRAM dynamic stability by circumventing potential hazards caused by Vth variations. In addition, one needs to balance both read and write operations during the optimization for the SRAM dynamic stability.

B. SRAM Dynamics

1) *Nonlinearity:* One primary challenge for SRAM dynamic stability verification and optimization is its nonlinear dynamic behavior. The time-evolution of safety distance depends on the nonlinear dynamics of SRAMs, which can be described by the differential algebraic equation (DAE)

$$\frac{d}{dt}q(x(t),t)+f(x(t),t)+u(t)=0 \tag{1}$$

in which state variable vector $x(t)$ and input vector $u(t)$ are deployed. In this DAE equation, taking dynamics of SRAM as an example, $f(x(t),t)$ includes drain current of transistor; $q(x(t),t)$ is charge accumulated on the gate or parasitic capacitors; $u(t)$ is the input current as well as noise current in (6); and vector x includes node voltages and branch currents.

After Newton iteration is performed at a selected operating point (or nominal point) x^*, the SRAM nonlinear dynamics by $f(t)$ can be linearized as $\frac{\partial f}{\partial x}\Big|_{x=x^*}$. Based on the mean-value theorem, the dynamic equation at any neighbor operating point x can be expressed by

$$\begin{aligned}&\frac{d}{dt}q(x(t),t)+f(x^*,t)+u+\frac{\partial f}{\partial x}\Big|_{x=x^*}(x-x^*)\\&+\frac{1}{2}(x-x^*)^T\otimes\frac{\partial^2 f}{\partial x^2}\Big|_{x=\xi}\otimes(x-x^*)=0,\\&\xi\in\{x^*+\alpha(x-x^*)|0\le\alpha\le 1\}\end{aligned} \tag{2}$$

where x^* is the nominal point and x is one neighbor point near x^*; and $\otimes$ represents the tensor multiplication. The 2nd-order remainder in (2), i.e., the difference between nonlinear $f(t)$ and its linear approximation, is called as linearization error denoted by L.

The SRAM dynamic equation thereby can be depicted in a simplified form by

$$\frac{d}{dt}(q(x^*,t)+C\Delta x)+f(x^*,t)+u^*(t)+G\Delta x+\Delta u+L=0 \tag{3}$$

in which

$$\begin{aligned}&C=\frac{\partial q}{\partial x}\Big|_{x=x^*},\quad G=\frac{\partial f}{\partial x}\Big|_{x=x^*};\\&\Delta x=x-x^*,\quad \Delta u=u-u^*.\end{aligned} \tag{4}$$

Here, u is decomposed into u^* and Δu, in which u^* is the noiseless input and Δu is the input noise independent of x. Assume that $q(x,t)$ can be further decomposed into $q(x^*)$ and $C\Delta x$. Thus, one can obtain

$$\frac{d}{dt}q(x^*,t)+f(x^*,t)+u^*(t)=0 \tag{5a}$$

$$\frac{d}{dt}C\Delta x+G\Delta x+\Delta u+L=0 \tag{5b}$$

in which (5a) is the nonlinear differential equation for nominal point x^* and (5b) is the linear equation with Euclidean distance from nominal point x^* to the neighbor point x.

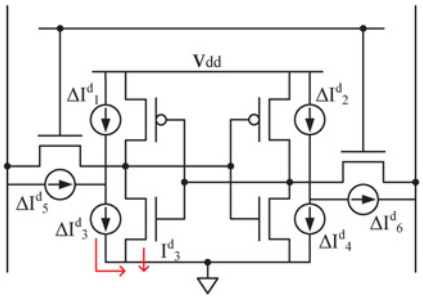

Fig. 3. SRAM with threshold-voltage variations modeled by additional current sources for all transistors.

On the basis of (5), reachability analysis can be deployed for SRAM dynamic stability verification and optimization in the state space as discussed later. Note that by using L, nonlinearity is considered and thus reachability analysis can be performed on nonlinear trajectories with high accuracy.

2) *Multiple Variation Sources and Multiple Device Parameters:* Moreover, we discuss how to introduce multiple variation sources and also multiple device parameters into the state equation. First, multiple threshold-voltage variation sources in SRAM can be introduced at the input u as additional noise current sources, which are added to the drain current of each transistor in SPICE as shown in Fig. 3.

Based on the first-order Taylor approximation, drain current for transistor operating in saturation region is presented in (6). For the simplicity of presentation, we only show the drain current equation in the saturation mode. In the experiment, we employ drain current equations that depend on transistor operation modes. The operation mode of each transistor is derived after Newton iterations at the present time instance similar to SPICE. Note that the threshold voltage variation is modeled as an *ad-hoc* noise drain current that is computed afterward based on the operation mode of the transistor. Operation mode transition can happen in the next time instance if the noise current value is large enough to cause the change

$$\begin{aligned} I^d + \Delta I^d &= \beta[V_{gs} - (V_{th} + \Delta V_{th})]^2 \\ \Delta I^d &\approx -\beta(V_{gs} - V_{th})\Delta V_{th}. \end{aligned} \tag{6}$$

The threshold-voltage variation of each transistor ΔI^d can be included by Δu of (5b)

$$\Delta u = u - u^* = [0, ..., \overbrace{\Delta I^d_j}^{\text{node } a}, ..., \overbrace{-\Delta I^d_j}^{\text{node } b}, ..., 0]^T \tag{7}$$

as an independent current source. Δu represents the jth variation of current source connected between nodes a and b. The other process variations can be also conveniently considered in the similar way.

What is more, perturbations of multiple device parameters can be considered as well. Suppose that each transistor in SRAM has width perturbation ΔW that affects transconductance g_m, namely, Δg_m. One can have

$$\Delta g_m = \frac{\partial g_m}{\partial W}\Delta W. \tag{8}$$

On basis of Δg_m, the device parameter perturbations can be included into conductance matrix by ΔG as follows:

$$\Delta G = \begin{pmatrix} \ddots & & \\ & \frac{\partial g_m}{\partial W} & -\frac{\partial g_m}{\partial W} \\ & -\frac{\partial g_m}{\partial W} & \frac{\partial g_m}{\partial W} \\ & & \ddots \end{pmatrix} \Delta W. \tag{9}$$

Based on the above discussions to include multiple variation sources and multiple device parameters, one can deploy zonotope to form a set of region for multiple variation sources and multiple device parameters. With the further development of linear multistep based integration for zonotope and its according sensitivity, one can develop reachability-based robustness verification and optimization as discussed in the later part.

C. Problem Formulation

Based on the aforementioned SRAM failure mechanisms and dynamic analysis, the objective of SRAM dynamic stability verification is to examine if the safety distance can be reduced at the final operating point, when considering interval values of threshold-voltage variations from all transistors.

If the safety distance fails to converge to zero, a robust optimization of SRAM dynamic stability is proposed in terms of safety distance with respect to the most adverse combination of V_{th} variations. We call this approach as a verification oriented robustness optimization.

SRAM Robustness Optimization: To ensure SRAM dynamic stability, one needs to minimize the safety distance measured at the final state of the system trajectory as follows:

$$\begin{aligned} \min_{w} \quad & F(w) \\ \text{s. t.:} \quad & W_{min} < w_i < W_{max}, i = 1, 2, ..., m. \end{aligned} \tag{10}$$

Here, w is the parameter or sizing vector for all transistors with a defined range $[W_{min}, W_{max}]$. Number of process variations is represented by m. In this paper, the weighted sum of safety distances for both read and write operations is deployed as the objective function by

$$F(w) = \begin{cases} D_w(w, t_w) + D_r(w, t_r), & \text{write and read failures} \\ D_w(w, t_w), & \text{write failure only} \\ D_r(w, t_r), & \text{read failure only} \end{cases} \tag{11}$$

where $D(w, t)$ is the safety distance and t is the pulse-width for read or write operation. Due to the symmetrical structure, three transistor pairs are used to represent the 6T-SRAM. Thus, the robustness optimization task to be performed reduces to a three-dimensioned parameter space, where a parameter-state point is denoted by $w \in \mathbf{R}^{3\times 1}$.

In the following, we will introduce a solution for the above problem by a zonotope-based reachability analysis. After the reachability analysis is performed for verification of safety distance, sensitivity of safety distance is also obtained to guide the optimization routine, which can eventually mitigate or even eliminate failures caused by V_{th} variations with improved SRAM dynamic stability.

TABLE I
PARAMETERS USED IN REACHABILITY-BASED VERIFICATION

Notation	Definition
$\mathcal{X}_0$, $\mathcal{X}_N$	Initial and final reachable sets in zonotope form
$\mathcal{M}$, $M^{(i)}$	Interval matrix in zonotope form and its matrix generator
$\mathcal{X}^{uni}$	Unidirectional zonotope
$g_k^{(i)}$	The ith generator at kth iteration
L_k	Linearization error at kth iteration
X_k^g	Generator matrix
$\mathcal{A}$	Jacobian matrix zonotope with variations
$\mathcal{C}$	Capacitance matrix zonotope with variations
U_k^g	Matrix with noise currents
g_m	Transistor transconductance
W	Transistor width
X_h^g	Homogeneous solution for (18)
X_i^g	Inhomogeneous solution for (18) due to input vector
X_e^g	Inhomogeneous solution for (18) due to linearization error

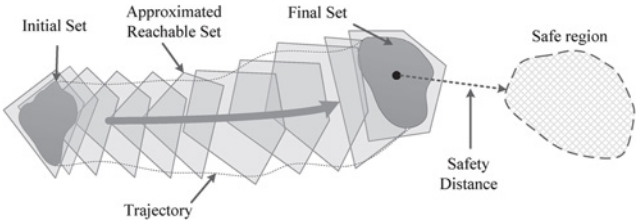

Fig. 4. System trajectory and safety distance with zonotopes.

III. ZONOTOPE-BASED REACHABILITY ANALYSIS FOR VERIFICATION

In this section, we show how to deploy zonotope-based reachability analysis for SRAM dynamic stability verification in terms of safety distance, and also discuss how to consider nonlinearity during SRAM dynamic stability verification.

Reachability analysis [23], [24], [27], [28] can efficiently determine a reachable region that one dynamic system evolves with a range of states. As such, one can perform one-time reachability analysis for all potential system trajectories that form the safe region with the safe distance determined from the final state set as shown in Fig. 4. The complete flow of reachability analysis for SRAM verification is shown in Algorithm 1. The notations used in this section and their definitions are listed in Table I. Here, $\mathcal{X}_0$ is the initial reachable set in zonotope form, and $\mathcal{X}_N$ is final reachable set used for calculation of safety distance and its sensitivity. With the linear multistep implementation for integration, the runtime cost or complexity of zonotope-based reachability analysis is similar with the transient analysis in SPICE. In the following, the details at each step are illustrated.

A. Reachable Set and Zonotope

Interval-value has been applied to model the uncertainties of state variables in [4], such as variation sources and device parameters. For example, if Δx_1, Δx_2 model uncertainties in two different dimensions of state variable x with c as interval center, the neighboring point including these variations can be modeled as: $x = c + [-1, 1]\Delta x_1 + [-1, 1]\Delta x_2$. However, there is no formal and efficient verification method developed to deal with multidimensional interval-value problem.

In this paper, we show that the multidimensional interval-value problem can be modeled by zonotope, which is a convex polytope to model multiple variation sources and multiple device parameters. Before defining zonotope, one important concept for reachability analysis is the reachable set.

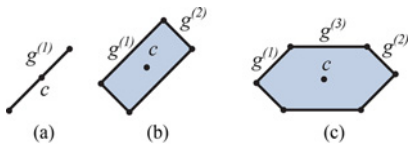

Fig. 5. Construction of zonotope. (a) $c + g^{(1)}$. (b) $c + g^{(1)} + g^{(2)}$. (c) $c + g^{(1)} + g^{(2)} + g^{(3)}$.

Definition 2: Reachable Set is the collection of all possible operating points or states in the state space that a system may visit, which can be approximated by an enclosing polytope.

One simple and symmetrical type of polytope, called zonotope [27] and is defined as follows.

Definition 3: Zonotope $\mathcal{X}$ is defined by

$$\mathcal{X} = \{x \in \mathbf{R}^{n\times 1} : x = c + \sum_{i=1}^{q}[-1, 1]g^{(i)}\}; \\ = (c, g^{(1)}, g^{(2)}, ...) \tag{12}$$

where $c \in \mathbf{R}^{n\times 1}$ is the zonotope center; and $g^{(i)} \in \mathbf{R}^{n\times 1}$ is called as zonotope generator.

As shown in (12), the so called zonotope is essentially a multidimensional interval in affine form or a hypercube with each generator as a variation in a different direction. Note that ellipsoid-modeled uncertainties are not considered in the reachability analysis in this paper, which will be addressed in future work.

Mathematically, the summation in zonotope needs to be interpreted as the Minkowski summation [28] of two finite sets such that the merged set can preserve convex property. Given two sets of zonotopes P and Q, Minkowski summation is performed by adding their zonotope centers and concatenating their generators as

$$P \oplus Q = \{p + q | p \in P, q \in Q\} \\ = (c_1 + c_2, g_1^{(1)}, ..., g_1^{(e)}, g_2^{(1)}, ..., g_2^{(u)}). \tag{13}$$

Here, c_1 and c_2 are the centers of zonotopes P, Q, respectively. Generators of P and Q are represented by $g_1^{(i)}$, $g_2^{(i)}$, respectively. A tight zonotope enclosing the convex hulls of two zonotopes $\mathbf{CH}(P, Q)$ can be found by $\overline{\mathbf{CH}}(P, Q)$ as follows:

$$\overline{\mathbf{CH}}(P, Q) = \frac{1}{2}(c_1 + c_2, g_1^{(1)} + g_2^{(1)}, ..., g_1^{(e)} + g_2^{(e)}, \\ c_1 - c_2, g_1^{(1)} - g_2^{(1)}, ..., g_1^{(e)} - g_2^{(e)}). \tag{14}$$

As all the generators are enclosed within in the zonotope, this forms a convex set. Note that the above property is applicable to the summations in (13) and (12).

One example of construction of zonotope with the addition of generator vectors is shown in Fig. 5. Here, c is the center of zonotope and generator vectors are represented as $g^{(1)}$, $g^{(2)}$ and $g^{(3)}$. We perform addition of zonotope vectors to preserve the convexity. In Fig. 5, initially a zonotope with a center c and generator $g^{(1)}$ is presented. Further to perform Minkowski summation, $g^{(2)}$ and its negative vector is added to $g^{(1)}$, which

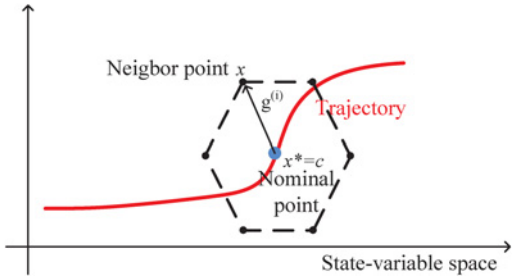

Fig. 6. Nonlinear SRAM dynamics and zonotope.

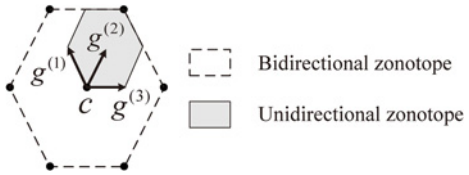

Fig. 7. Bidirectional zonotope and unidirectional zonotope.

results $g^{(2)}$ in two directions to form a convex zonotope. The same procedure is followed to perform Minkowski summation for additional generators.

Physically, a zonotope spans a polytope in the state space that covers all kinds of trajectories caused by uncertain initial sets as shown in Fig. 4. A zonotope indicating its center c and the generator vector $g^{(i)}$ is shown in Fig. 6, which indicates that the nominal point x^* can be relevant to the center of zonotope c; and the the deviation distance $(x - x^*)$ of the state variable x can be relevant to the generator g.

Note that scaling factors of generators are allowed to range from -1 to 1. Thus, the difference vector from the nominal point to other points within a reachable set varies in two directions. One can use bidirectional zonotopes to include all possible threshold variations during SRAM verification. However, for the calculation of sensitivity during robustness optimization, the scaling factor is defined within $[0, 1]$ to obtain a single-direction difference vector from the nominal point to any other neighbor point in the reachable set. We call the modified zonotope as the unidirectional zonotope (Fig. 7), which is determined by

$$\mathcal{X}^{uni} = \{x \in \mathbf{R}^{n\times 1} : x = x^* + \sum_{i=1}^{q}[0, 1]g^{(i)}\}. \tag{15}$$

What is more, similar to zonotope for state variable vector, one can model the interval values for state matrices by zonotope as well [28]. As such, the matrix zonotope is derived as

$$\mathcal{M} = \{M \in \mathbf{R}^{n\times n} : M \in M^{(0)} + \sum_{i=1}^{q}[0, 1]M^{(i)}\}. \tag{16}$$

Similar to zonotope of state variable vector, the matrix $M^{(0)}$ is the center matrix and the matrix $M^{(i)}$ is called the matrix generator, which contains the variation ranges of perturbed device parameters. Addition and multiplication rules for matrix zonotopes are similarly defined as vector zonotopes [28].

B. Reachability Analysis by Backward Euler Method

On the basis of nonlinear dynamics of SRAM discussed in (5), the zonotope-based reachability analysis is performed as follows. The detailed explanation of explicit integration can be found in [28], which is much more sophisticated and expensive than the proposed numerical integration method developed in this paper. In this paper, a SPICE-like zonotope evolution is developed based on backward Euler method [29].

First, note that (5) can be solved with discretized time-step h at kth-iteration by

$$\Delta x_k^{(i)} = A^{-1}(\frac{C}{h}\Delta x_{k-1}^{(i)} - \Delta u_k^{(j)} - L_k);$$
$$k = 1, ..., N; i = 1, ..., q; j = 1, ..., m \tag{17}$$

where $A = \frac{C}{h} + G$ is the Jacobian matrix, N represents the number of time steps, m represents the number of process variations and q is the number of zonotope generators.

Let the zonotope center be a nominal operating point x_k^*. Meanwhile, the zonotope generators $\Delta x_k^{(i)}$ are the Euclidean distances (4) from the nominal point x_k^* to neighbor points x_k. As such, one can have a zonotope of state variable vector by

$$\mathcal{X}_k = \left\{ x_k \in \mathbf{R}^{n\times 1} : x_k = x_k^* + \sum_{i=1}^{q}[-1, 1]\Delta x_k^{(i)} \right\}.$$

What is more, multiple threshold-voltage variation sources are included in form of zonotope generators based on (7) as

$$\mathcal{U}_k = \left\{ u_k \in \mathbf{R}^{n\times 1} : u_k = u_k^* + \sum_{j=1}^{m}[-1, 1]\Delta u_k^{(j)} \right\}.$$

The according iteration equation for zonotope-based verification is thereby built after substituting generator $\Delta x_k^{(i)}$ by generator matrix $X_k^g = [\Delta x_k^{(1)}, ..., \Delta x_k^{(q)}]$, $\Delta u_k^{(i)}$ by $U_k^g = [\Delta u_k^{(1)}, ..., \Delta u_k^{(m)}]$, Jacobian matrix A by matrix zonotope $\mathcal{A}$, and capacitance matrix C by matrix zonotope $\mathcal{C}$. As such

$$X_k^g = \mathcal{A}^{-1}\left(\frac{\mathcal{C}}{h}X_{k-1}^g - U_k^g - L_k\right), k = 1, ..., N. \tag{18}$$

What is more for robustness optimization, matrix zonotopes $\mathcal{A}$ and $\mathcal{C}$ can be built to consider perturbations from multiple device parameters, such as transistor width sizings in the case of SRAMs. In $\mathcal{A}$, interval conductance matrix ΔG can be computed using the interval values of transistor widths similar to (9). As such, the zonotope matrix can be further interpreted in terms of interval-valued matrices by

$$A \in \left[A^{(0)} - \sum_i |A^{(i)}|, A^{(0)} + \sum_i |A^{(i)}|\right]. \tag{19}$$

The matrix generator can be formed as follows:

$$A^{(i)} = \frac{\partial A^{(0)}}{\partial W} \cdot \Delta W^{(i)} = \Delta G^{(i)}. \tag{20}$$

Here, $A^{(0)}$ is the nominal state matrix without variations, and $A^{(i)}$ is the variation of state matrix caused by perturbation due to ith transistor width $\Delta W^{(i)}$.

In addition, note that in (18), the reciprocal of the matrix zonotope $\mathcal{A} = (A^{(0)}, ..., A^{(m)})$ can be evaluated in two steps

without explicit inversion. The first step is to calculate $(A^{(0)})^{-1}$ by LU decomposition

$$(A^{(0)})^{-1} = U^{-1}L^{-1}P^T I \tag{21}$$

where I is the identity matrix and P is the permutation matrix. The second step is the approximated expansion of $\mathcal{A}^{-1}$ by

$$\mathcal{A}^{-1} = ((A^{(0)})^{-1}, ..., (A^{(0)})^{-1}A^{(m)}(A^{(0)})^{-1}). \tag{22}$$

Recall that m represents the number of process variations. This approach enables an implementation of reachability analysis inside a SPICE-like simulator. However, such an approximated inversion may introduce additional source of error.

C. New Set Formulation by Minkowski Summation

Superposition principle allows to separate the solution of (18) into two parts: homogeneous solution X_h^g with respect to the initial state X_k^g when there is no input U_k^g; and inhomogeneous solution X_i^g accounting for the system input U_k^g supposed that the initial state X_k^g is the origin. Note that linearization error is also considered at the input during the update at each time step.

As such, the inhomogeneous solution can be further divided into the one due to input vector (X_i^g) and the other one for linearization error (X_e^g)

$$\begin{aligned} X_h^g &= \mathcal{A}^{-1}\frac{\mathcal{C}}{h}X_{k-1}^g \\ X_i^g &= -\mathcal{A}^{-1}U_k^g \\ X_e^g &= -\mathcal{A}^{-1}L_k. \end{aligned} \tag{23}$$

Given an initial set $\mathcal{X}_k$ at current time step, there are three sets of solutions computed. Multiplication of matrix zonotopes [28] can be used for solving (23). The number of generators may grow after zonotope multiplications. As such, an upper bound has to be set on the number of generators and generators with smallest magnitudes can be discarded in this process. The concatenation of these sets with convexity is performed by the aforementioned Minkowski summation [28] to form a convex zonotope.

Therefore, a new reachable set $\mathcal{X}_k$ is obtained by combining zonotope center x_k^* and generator X_k^g as

$$\begin{aligned} X_k^g &= X_h^g \oplus X_i^g \oplus X_e^g \\ \mathcal{X}_k &= (x_k^*, X_k^g) \end{aligned} \tag{24}$$

where $\oplus$ represents the Minkowski summation.

D. Linearization Error Control

Approximation of linearization error L_k in line 6 of Algorithm 1 is a critical step in each iteration cycle. L_k is a vector with interval values. It can be viewed as a zonotope with 0 as the center and the interval ranges as generators. Linearization error accounts for nonlinearity of SRAM dynamics. Here, nominal point x_k^* is the zonotope center for the current iteration x_k; and x_k varies within zonotope $\mathcal{X}_k$. As such, L_k cannot be exactly calculated but approximated for $x_k \in \mathcal{X}_k$. Detailed approximation for L_k can be found in [28] by

$$\begin{aligned} L_k = \frac{1}{2}(x_k - x_k^*)^T \otimes \left.\frac{\partial^2 f}{\partial x_k^2}\right|_{x_k=\xi} \otimes (x_k - x_k^*), \\ \xi \in \{x_k^* + \alpha(x_k - x_k^*) | 0 \le \alpha \le 1\}. \end{aligned} \tag{25}$$

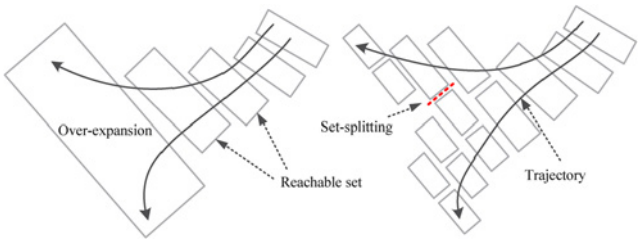

Fig. 8. Reachable sets with or without splitting.

By dynamic updating the approximated L_k, the convergence of zonotope-based reachability analysis can be guaranteed. What is more, one can further develop local-truncation-error control scheme similar to SPICE.

Next, the nonlinearity of SRAM is rather prominent in the transition area between two convergent regions where linearization error expands rapidly. Over-expanded reachable sets in Fig. 8 may be too rough to be meaningful. Based on (25), if the deviations $(x_k - x_k^*)$ between states are small, the second order derivatives are appropriate enough to approximate the linearization error. However, for strong nonlinearity, set-splitting needs to be performed to limit the deviations $(x_k - x_k^*)$ by cutting it into half and creating two nominal points, $x_k^* + |x_k - x_k^*|/2$ and $x_k^* - |x_k - x_k^*|/2$. After self-splitting, new zonotopes are formed but usually with overlap of each other. Avoiding the overlap can reduce unnecessary computations. One possible solution is to cancel the reachable sets that have already been reached, which can be performed by difference operation between $x_k^* + |x_k - x_k^*|/2$ and $x_k^* - |x_k - x_k^*|/2$.

A judgement condition for set splitting is shown as follows:

$$\mathbf{IH}(L_k) \subseteq [-\varepsilon, \varepsilon] \tag{26}$$

in which $\mathbf{IH}()$ is the interval hull operation that converts a zonotope to a multidimensional interval; and ε is an user-defined limit vector. After the current reachable set is divided into two subsets, along with a new trajectory being created, the zonotope-based reachability analysis is repeated at the current time point for the new subsets.

IV. Sensitivity of Safety Distance for Optimization

In this section, we first introduce the definition of safety distance under zonotope, and further discuss the according sensitivity calculation of safety distance, which is applied for SRAM dynamic stability optimization by tuning multiple SRAM device parameters simultaneously. Different notations and their definitions used in this section are listed in Table II.

A. Safety Distance

With the use of zonotope, safety distance in the state space can be obtained as follows. Assume that one safe state is located at p_{safe} in the state space. As for any zonotope in the form of (12), the safety distance for the reachable set can be expressed as

$$\mathcal{D} = \{d \in \mathbf{R}^{n\times 1} : d = p_{safe} - c - \sum_{i=1}^{q}[0, 1]g^{(i)}\}. \tag{27}$$

Algorithm 1: Reachability Analysis

Input: System equation, input vector $\mathcal{I}_{1,2,...N}$, initial set $\mathcal{X}_0$, simulation interval h, and the maximum number of time steps N.

Output: $\mathcal{X}_N$

1: **for** $(k = 1; k < N + 1; k++)$ **do**
2: $\mathcal{X}_{k-1} \rightarrow (x^*_{k-1}, X^g_{k-1})$
3: compute x_k and linearized matrices C_k, G_k
4: compute system matrix zonotopes $\mathcal{A}$ and $\mathcal{C}$
5: approximate linearization error L_k
6: **if** $\mathbf{IH}(L_k) \subseteq [-\varepsilon, \varepsilon]$ **then**
7: $X^g_h = \mathcal{A}^{-1}\frac{\mathcal{C}}{h}X^g_{k-1}$
8: $X^g_i = -\mathcal{A}^{-1}U_k$
9: $X^g_e = -\mathcal{A}^{-1}L_k$
10: $X^g_k = X^g_h \oplus X^g_i \oplus X^g_e$
11: $\mathcal{X}_k = (x^*_k, X^g_k)$
12: **else**
13: $\mathcal{X}_{k-1} = split(\mathcal{X}_{k-1})$
14: continue
15: **end if**
16: **end for**

TABLE II
PARAMETERS USED IN ROBUSTNESS OPTIMIZATION

Notation	Definition
$\mathcal{D}$	Safety distance for a reachable set
D	Safety distance for a specific point inside a reachable set
p_{safe}	Location of safe state in state space
x_{final}	Final reachable state
c_{final}	Center of zonotope in the final reachable set
g_{final}	Generator of zonotope in the final reachable set
$S(D, w)$, s	Large-signal sensitivity and small-signal sensitivity
w_k, Δw_k	Parameter vector and its increment
$F(w_k, t)$	Objective function
ρ_k	Gradient of objective function
β_k, γ	Empirical scaling factor

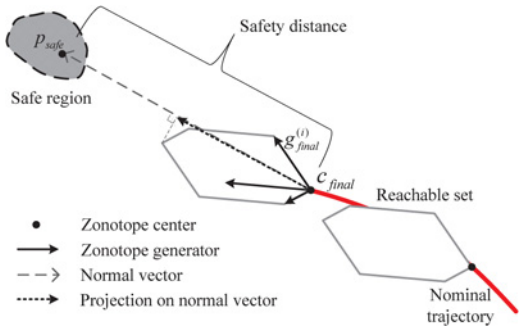

Fig. 9. Safety distance and its sensitivity in reachability analysis.

As shown in Fig. 9, for one specific point inside the reachable set, one certain safety distance can be determined as

$$D = ||p_{safe} - c - \sum_{i=1}^{q} \varepsilon^{(i)} g^{(i)}||_2, 0 \leq \varepsilon^{(i)} \leq 1 \qquad (28)$$

where $\varepsilon^{(i)}, i = 1, ...q$ is the coefficient of generators to determine the relative position of the point within the zonotope. Note that safety distance reduces to zero if zonotope settles in the safe region. As such, one can utilize it to verify the dynamic stability of SRAM.

B. Sensitivity of Safety Distance

With the use of reachability analysis by zonotope, trajectory of SRAM is obtained and the sensitivity of the safety distance D at the final reachable set

$$x_{final} = c_{final} + \sum_{i=1}^{q} [0, 1] g^{(i)}_{final}$$

can be calculated afterward. Note that the safety distance D for a reachable set can vary within a certain range as the perturbation of device parameters (20) can result in different operating points close-by.

The perturbation range of device parameters $[0, \Delta W]$ is in form of interval entries of the matrix zonotope (16) and is included in $\mathcal{A}$ in (23). By zonotope multiplication, the perturbation is transferred to generator Δx_k (namely, g_k) in X^g_k. After running reachability iterations, the generator of the final state (g_{final}) is used to derive the perturbation of the safety distance as follows:

$$\Delta D = \sum_{i=1}^{q} \frac{(p_{safe} - c_{final})^T}{||p_{safe} - c_{final}||_2} g^{(i)}_{final}. \qquad (29)$$

As shown in Fig. 9, the perturbation of the safety distance D at the final reachable set x_{final} is obtained by projecting zonotope generators $g^{(i)}_{final}$ to the normal vector $\frac{(p_{safe} - c_{final})^T}{||p_{safe} - c_{final}||_2}$, which is formed from the zonotope center c_{final} to the safe region p_{safe}.

As such, one can calculate the large-signal sensitivity $S(D, w)$ of the safety distance D with respect to device parameter w by

$$S(D, w) := \frac{\Delta D}{\Delta w} \qquad (30)$$

which becomes the ratio between their increment values ΔD and Δw for multiple device parameters simultaneously.

Note that different from the single-parameter small-signal sensitivity of state variable obtained by differentiating (2) with

$$s = \frac{\partial x_k}{\partial w} = -(\frac{C}{h} + G)^{-1} \frac{\partial G}{\partial w} (\frac{C}{h} + G)^{-1} (\frac{C}{h} x_{k-1} - u_k) \qquad (31)$$

in which the linearization error L_k of (2) is omitted and state variable x_{k-1} for the last time-step is assumed as constant. (For the simplicity of presentation, derivatives of capacitance matrices are omitted.) As such, one can observe that though the single-parameter small-signal sensitivity is easy to obtain, compared to the large-signal sensitivity $S(D, w)$ calculated from reachability analysis in (30), s may fail to measure the accumulated variation from the previous states by multiple parameters. What is more, without considering nonlinearity, the small-signal sensitivity may fail to provide accurate direction during the global optimization of system trajectory. In contrast, the proposed multiparameter large-signal sensitivity by safety distance in reachability analysis can be effectively utilized for SRAM dynamic stability optimization, which has faster convergence with higher accuracy as demonstrated by numerical experiment results.

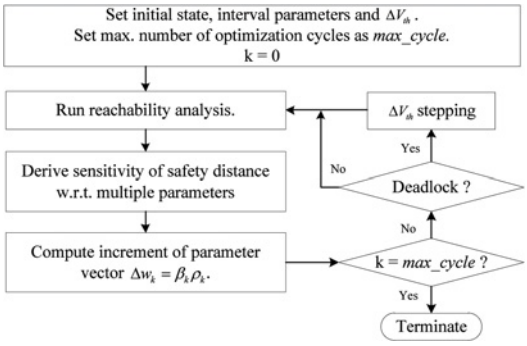

Fig. 10. Flow-chart of robust stability optimization of SRAMs.

C. Safety Distance Optimization by Sensitivity

The sensitivity of safety distance derived from the reachability analysis can guide the optimization direction that departs from unsafe region. It can effectively shorten the safety distance by tuning multiple device parameters. As such, one can embed it within any gradient-based optimization algorithm to achieve a robust SRAM design under process variations. Note that a dynamic stability optimization by sensitivity of safety distance is convenient and general; and hence can be applied for any circuits when safety distance is properly defined because there is no specific knowledge of circuit types required. In the case of robust stability optimization for SRAMs, transistor widths can be automatically sized to improve the dynamic stability. The complete flow of SRAM optimization is shown in Fig. 10 with following detailed steps.

First, in each searching step, the increment Δw_k of one parameter vector w_k is in the same direction with the sensitivity of the safety distance by

$$\Delta w_k = \beta_k \rho_k \tag{32}$$

where $\beta_k > 0$ is a scaling factor and ρ_k is the gradient of objective function, i.e., $S_k(F(w,t),w)$.

The parameter increment Δw_k for the next step is estimated by

$$F(w_k,t) + \Delta w_k^T \rho_k = 0. \tag{33}$$

As such, one can obtain

$$\beta_k = -\frac{F(w_k,t)}{\rho_k^T \rho_k} \tag{34}$$

after combining (32) with (33). Increment of parameter vector Δw (32) is obtained afterward.

Note that the objective function $F(w,t)$ changes nonlinearly in the parameter space but its gradient $\frac{\partial F(w,t)}{\partial w}$ becomes small in magnitude around the safe region. As such, an empirical scaling factor $\gamma < 1$ can be utilized to resize the estimated increment of parameter vector

$$\beta_k = -\gamma \frac{F(w_k,t)}{\rho_k^T \rho_k}, 0 < \gamma < 1 \tag{35}$$

such that the convergence of optimization can be improved. What is more, to further improve the convergence, the initial value stepping can be used when the searching is stuck in the deadlock or out of the feasible range of device parameters ($W_{min} < w_{1,2,3} < W_{max}$).

V. Experimental Results

With the use of zonotope-based reachability analysis, the robustness verification and optimization for SRAM dynamic stability are implemented inside a SPICE-like simulator by MATLAB. Manipulations of zonotopes are performed by a MATLAB toolbox named Multiparametric Toolbox (MPT) [30]. BSIM3 is used as the MOSFET transistor model. Threshold voltage variation in each transistor is introduced as a noise current source in (6). Its center value is 0 and variation is $|k\frac{W}{L}(V_{gs}-V_{th})\delta V_{th}|$, where δ is the variation range. Experiment data is collected on a desktop with Intel Core i5 3.2 GHz processor and 8 GB memory.

We first demonstrate zonotope-based reachability analysis upon SRAM dynamic stability verification under threshold-voltage variations. Then, we show robustness optimization on basis of zonotope-based sensitivity calculation. Further, we compare with Monte Carlo-based verification and also single-parameter small-signal sensitivity based optimization. For SRAM stability verification, we used 1000/2000 samples in order to show a comparison in reasonable runtime. For SRAM stability optimization, we used 100 000 samples when measuring the yield rate before and after optimization as shown in Fig. 16. But we cannot show 100 000 curves in one figure. What is more, for both of the verification and optimization, we set threshold voltage variation up to 30%. In addition note that during read operation, two charged external capacitors are connected to the outputs of SRAM. Data in SRAM is read after one of the external capacitors is discharged through SRAM. By comparison, during write operation, internal capacitors in SRAM are pulled down/up. Since internal capacitors are much smaller than the external capacitors for read operation. As such, write operation is observed faster than read operation in experiment results.

A. Dynamic Stability Verification Results

40 nm node is used in our experiment and 1V is chosen as the supply voltage. Moreover, the equilibrium state of SRAM usually does not settle at the exact v_{dd} or 0. Thus we start reachability analysis with an initial state set of $v_1 \in [0.98, 1.00]$ and $v_2 \in [0, 0.02]$.

1) *Verification of Write Operation:* The write operation is first verified by reachability analysis with consideration of threshold voltage variations. For comparison, Monte Carlo verification is performed to demonstrate the accuracy of reachability analysis. The duration of write signal is varied to exam SRAM behaviors under different conditions.

Verification results of write operation are shown in Fig. 11 with threshold-voltage variation range set to 10%. Larger variation range can be considered for verification when high-order noise model is available. The curves simulated by Monte Carlo verification are plotted in light purple and trajectories of reachability analysis are drawn in dark blue. Three different durations of write signal are tested, including 0.025 ns, 0.029 ns, and 0.050 ns.

In Fig. 11(a), write signal lasts for 0.025 ns. At the beginning, trajectories move toward the other corner of variable plane as data is being written into SRAM. Later, the turning

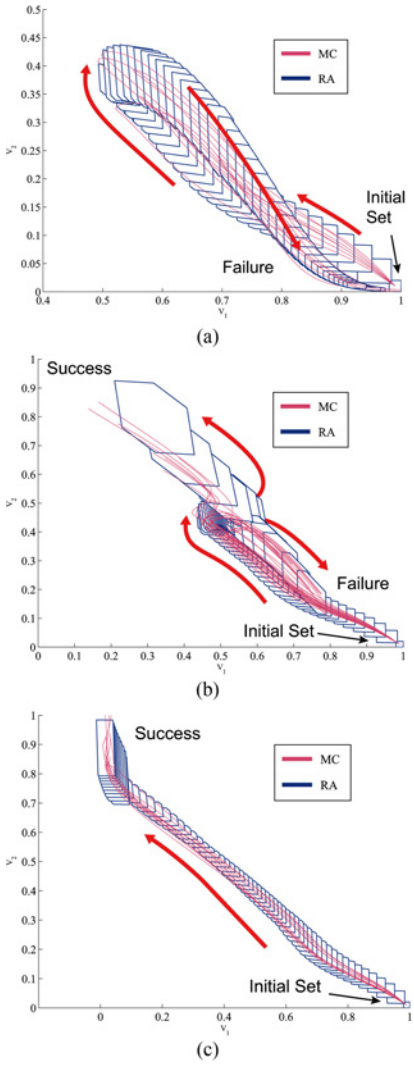

Fig. 11. Verification of write operation with threshold variation range of 10%. (a) Write operation fails with 0.025 ns writing pulse. (b) Write operation fails with 0.029 ns writing pulse. (c) Write operation succeeds with 0.050 ns writing pulse.

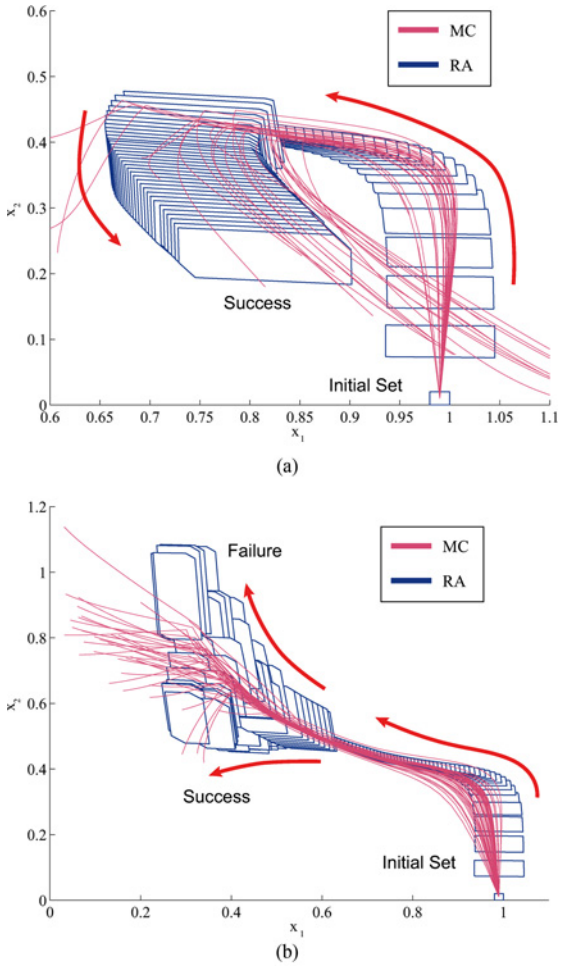

Fig. 12. Verification of read operation with threshold variation range of 30%. (a) Read operation succeeds with 6 ns pulse. (b) Read operation fails with 11 ns pulse.

point of trajectories is generated when the write signal flips to 0. Afterward, trajectories return to initial states. As such, the data fails to be written into the SRAM, which means that write failure happens.

When the write pulse increases to 0.029 ns in Fig. 11(b), trajectories of reachable sets split around the center of the state space. This happens when the write signal shuts down. Some of the new trajectories move back to initial states, which means that some states still fail in the write operation. To limit the computational cost, the number of trajectories needs to be constrained. For the simplicity of presentation, we show two trajectories in Fig. 11(b). Note that at the end of the trajectory departing from the failure region, the Monte Carlo curves do not settle with in reachable sets, which means the mismatch between Monte Carlo curves and reachable sets happens. This is because after some trajectories reachable sets are truncated, the rest trajectories may not cover all possible curves of Monte Carlo verification. Thus, the number of trajectories is a tradeoff between time and accuracy. An ideal set-splitting strategy can make the overlap between new reachable sets the smallest and thus the new trajectories can cover the most Monte Carlo curves.

Finally, when the duration increases to 0.050 ns in Fig. 11(c), all possible states finish write operation without failure. As shown in Fig. 11, curves of Monte Carlo verification remain within the reachable sets by reachability analysis under the similar accuracy. It indicates that reachability analysis can succeed in approximating the trajectories of SRAM for failure verification.

2) *Verification of Read Operation:* Next, the read operation can be also verified by reachability analysis. The verification result of read operation is compared with different durations of input signal while the *V*th variations are set as 30%. Duration of read signal is set to 6 ns and 11 ns.

As shown in Fig. 12, the Monte Carlo curves are plotted in light purple and the enclosing trajectories drawn by reachability analysis are in dark blue. When signal duration is 6 ns [Fig. 12(a)], all reachable sets recover back to the initial state after read operation finishes. But when the signal duration rises to 11 ns, most reachable sets head for the opposite state which means that read failure happens [Fig. 12(b)]. Due to the limited accuracy of the first-order noise current model in (6), the difference between Monte Carlo and the reachability

analysis can be observed. Yet reachability analysis is still able to catch most of the possible trajectories obtained by Monte Carlo simulations. Note that in the optimization experiment, the threshold voltage variation is predefined with no use of noise current equation in (6).

B. Stability Optimization Results

The setup of SRAM circuit in optimization is as follows. 40 nm CMOS is used as the technology node for our optimization experiment. Supply voltage of SRAM v_{dd} is set to $1V$. Initial states for SRAM are set as $v_1 = 1V$ and $v_2 = 0$, respectively. The transistors widths can change in the range of [100 nm, 600 nm] with step of 1 nm. Threshold variations of 30% are considered by verification and optimization. Note that interval values of threshold-voltage variations are considered by reachability analysis as input sources. For the robustness optimization, the interval values of transistor widths are further considered in zonotope matrix to derive sensitivity.

As mentioned in Section II-C, the dynamic stability of SRAM can be improved by shortening the safety distance and converging to the safe region. Although yield rate cannot be calculated based on safety distance, it can be optimized by improving the stability or reducing the failure rate of SRAMs. In this way, the safety distances of a number of SRAMs with different Vth deviations can be shortened as a whole. As such, less SRAMs end up outside safety region with yield rate $Y := 1 - \frac{N_{failure}}{N_{total}}$, which is increased as $N_{failure}$ reduces. As for write operation, strong pulling strength of M1, M4 and weak strength of the other transistors lead to high probability of write failure. Thus, the negative threshold-voltage variations are assumed for M1, M4, while positive threshold-voltage variations are assumed for the other transistors. For the same reason, negative threshold-voltage variations in M2, M3, M6, and positive threshold-voltage variations in M1, M4, M5 are used for read operation. The threshold-voltage variations are set as constant during optimization as the standard deviations.

1) *Optimization of Read or Write Failure:* To start with, we perform dynamic stability optimization for read operation only. Initial widths of three transistor pairs are $[W_1, W_3, W_5]$ =[200 nm, 300 nm, 300 nm] and pulse width is 9 ns. The process of stability optimization is shown in Fig. 13, in which trajectories are plotted in light purple; and reachable sets (i.e., zonotopes) due to parameter changes are drawn in dark blue. Unlike the situation in the previous section, zonotopes for SRAM optimization are quite small. This is because transistor widths are varied by step of 1 nm and thus the resulting variation range of trajectory is limited. Note that the sensitivity calculated here is multiparameter large-signal sensitivity with respect to one zonotope set, which is different from the classic single-parameter small-signal sensitivity in (31). As demonstrated later, the multiparameter large-signal sensitivity is more stable and accurate.

Three reachable sets are generated at each nominal point with different transistor widths. The final sets are used to derive large-signal sensitivities (Fig. 9) by sensitivity-based reachability analysis. The initial trajectory fails to converge to the safe region. After three iterations, the optimized trajectory recovers from read failure. The optimized widths are [148 nm, 343 nm ,217 nm].

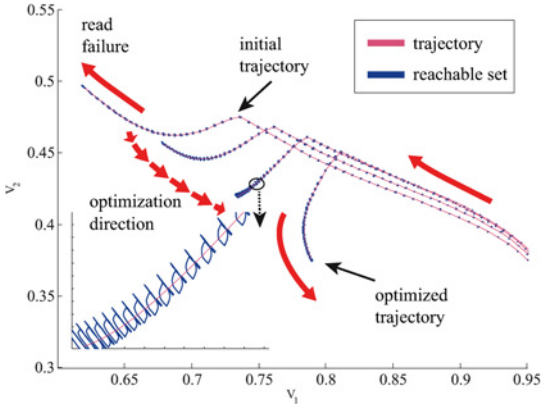

Fig. 13. Optimization of read operation only.

Then, we perform stability optimization for write operation only. We set the initial pair widths as $[W_1, W_3, W_5]$ =[400 nm, 500 nm, 350 nm] and reduce the pulse width to 0.050 ns. The stability optimization by large-signal sensitivity calculated from reachability analysis can certainly help guide the system trajectory to converge to the safe region within four iterations (Fig. 14). The optimized widths are [381 nm, 440 nm, 497 nm].

2) *Optimization of Read and Write Failure:* To optimize read and write failure simultaneously, initial transistor pair widths are randomly chosen as $W_1 = 200$ nm, $W_3 = 400$ nm and $W_5 = 400$ nm. Pulse width is 9 ns for read operation and is 0.024 ns for write operation. The process of stability optimization is shown in Fig. 15.

The optimization direction of trajectory for read operation and write operation are shown in Fig. 15(a) and (b), respectively. The trajectory after performing optimization to initial set of transistor widths is represented as initial. From Fig. 15(b), one can observe that at the beginning, write failure happens as the trajectory converges to the initial state. With the use of the proposed sensitivity-based reachability analysis for the dynamic stability optimization, the trajectory of read operation moves away from the wrongly converged region and finally moves to the target state after six iterations when tuning transistor pair sizes. Meanwhile, the read operation in Fig. 15(a) is considered, where read failure did not happen at the beginning. As the write operation is optimized, the trajectory for read operation deviates upward too. As such, the safety distance to the top-left corner (in this case) is decreased. In other words, the write operation is optimized at the expense of read operation to achieve a lower rate of failure for both cases.

The optimized transistor widths obtained by our approach are finally achieved as $W_1 = 192$ nm, $W_3 = 330$ nm and $W_5 = 586$ nm, respectively. Yield rate ($Y := 1 - \frac{N_{failure}}{N_{total}}$) considering both read and write functions is improved from 6.8% to 99.957%. Further improvement of yield rate can be achieved by introducing larger threshold variations during the optimization.

C. Comparisons

1) *SRAM Dynamic Stability Verification:* A detailed comparison between zonotope-based reachability analysis and

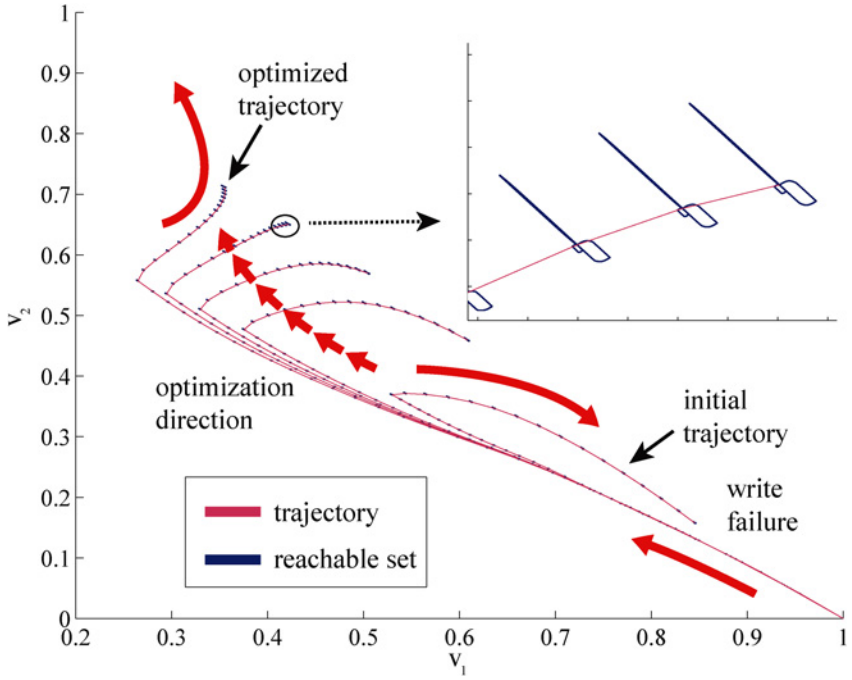

Fig. 14. Optimization of write operation only.

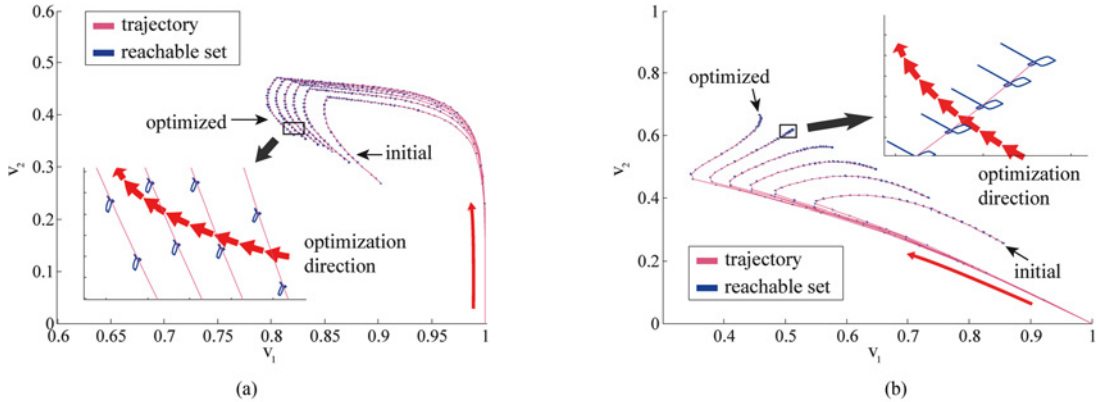

Fig. 15. Optimization procedure for SRAM dynamic stability. (a) Optimization of read operation. (b) Optimization of write operation.

Monte Carlo method is made upon the write operation. For the concern of huge time consumption by Monte Carlo method, we use 1000 samples that usually takes more than one hour for a single round of verification according to our experiment. Different durations of write signal are considered as well as different threshold-voltage variations in all transistors. Detailed experimental results are listed in Table III in which pulse refers to the duration of input signal; and acceleration is the ratio of time consumption of Monte Carlo to that of reachability analysis.

As shown in Table III, compared with Monte Carlo, reachability analysis can achieve speedup up to more than 800× for 1000 samples. When write signal duration is set to 0.025 ns [Fig. 11(a)] or 0.050 ns [Fig. 11(c)], only one trajectory is generated by reachability analysis. Linearization is performed around one nominal trajectory which takes up most of the simulation time. Thus the time consumption of reachability analysis is slightly higher than the simulation of one sample of Monte Carlo verification. As signal duration is set to 0.029 ns, reachable sets are split into different parts and two trajectories are generated. Therefore the runtime of reachability verification doubles and the speedup ratio reduces by half when the signal lasts 0.029 ns and 10% *V*th variations are introduced [Fig. 11(b)]. For all experiment cases listed in the Table III, the reachability analysis can achieve the similar accuracy as Monte Carlo method to report the failure region.

TABLE III
TIME CONSUMPTION OF SRAM VERIFICATION

Pulse (ns)	Threshold Variation	Reachability Analysis(s)	Monte Carlo(s)	Acceleration
0.025	1%	4.27	3682.89	861.97×
	5%	4.85	3679.84	759.41×
	10%	4.68	3725.21	795.35×
0.029	1%	4.09	3655.89	893.36×
	5%	6.14	3667.69	596.96×
	10%	9.10	3652.92	401.40×
0.050	1%	4.18	3651.56	872.76×
	5%	4.44	3652.77	822.22×
	10%	4.38	3649.81	833.31×

2) *SRAM Dynamic Stability Optimization:* The runtime of optimization at each iteration is listed in Table IV, where more than 600× runtime speedup can be achieved by our approach.

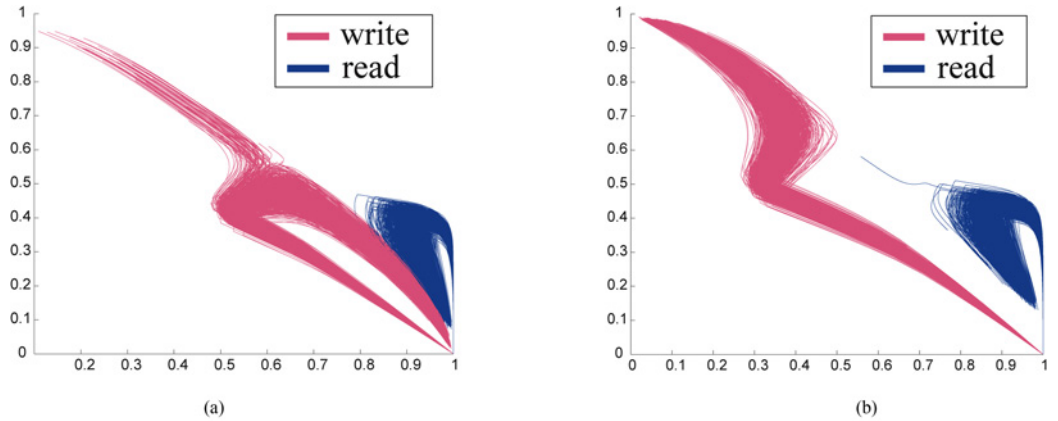

Fig. 16. Statistical yield calculation (a) before and (b) after optimization.

TABLE IV
RUNTIME COMPARISON OF SRAM STABILITY OPTIMIZATION

Iter.	Transistor Widths (nm)	Sensitivity based RA(s)	MC (s)	Speedup
1	[185, 371, 451]	9.37	5953.23	635.35×
2	[177, 359, 485]	9.69	5876.12	606.41×
3	[173, 349, 515]	9.53	5901.64	619.27×
4	[171, 340, 545]	9.34	5932.87	635.21×
5	[181, 329, 574]	9.58	5951.07	618.11×
6	[192, 330, 586]	9.51	5911.91	621.65×

The optimized transistor widths i.e. $[W_1, W_3, W_5]$ for read and write stability optimization are also represented in Table IV. Iteration number for optimization is represented as Iter and the time taken for optimization by using our proposed reachability based sensitivity analysis is listed under sensitivity-based RA column with its units in seconds, similarly time consumption for optimization by traditional Monte Carlo-based method is listed under MC column with time consumption in seconds and the corresponding speedup achieved by our proposed method is listed under speedup column. For example, for the first optimization step, the proposed optimization takes about 9 s while Monte Carlo-based method needs nearly 2 h. The time consumption of reachability analysis is roughly the same with one transient simulation, since most computation is used on the simulation of the nominal trajectory. Similarly, one can observe the variation in transistor widths at each iteration. As discussed previously the initial transistor widths are set to [200 nm, 400 nm, 400 nm], but the optimized set of transistor widths by our approach is [192 nm, 330 nm, 586 nm]. In our case, to derive large-signal sensitivity with respect to the three transistor pairs, reachability analysis is performed for three times.

Furthermore, we compare the our approach with another optimization routine by single-parameter small-signal sensitivity (31). For the same aforementioned test-case, the optimization result by small-signal sensitivity in shown in Fig. 17. Unlike in Fig. 15(b), the optimization routine by single-parameter small-signal sensitivity fails to find a feasible solution and results in negative width after three iterations. Transistor pair widths, i.e. $[W_1, W_3, W_5]$ are shown in Fig. 17. Note that W_5 fails to be tuned during optimization, because small-signal sensitivity with respect to W_5 is much smaller than the rest. Since the single-parameter small-signal sensitivity only depends on the location of the final state, the resulted gradient merely has

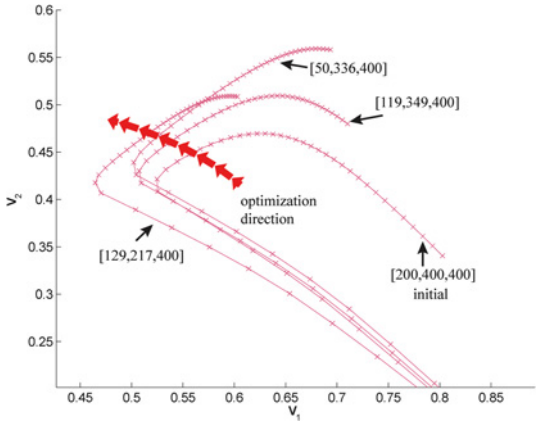

Fig. 17. Optimization of write operation by small-signal sensitivity.

local accuracy and changes irregularly as the trajectory moves. As a result the small-signal sensitivity does not lead to the convergence. The proposed large-signal sensitivity calculation in reachability analysis can achieve much higher accuracy for a faster converged SRAM optimization.

VI. CONCLUSION

In this paper, we are the first to develop the reachability analysis for the robustness verification and optimization of SRAM dynamic stability in the presence of multiple variation sources and device parameters from all transistors. By quantitatively describing SRAM robustness with a defined safety distance, our approach can efficiently provide not only stability verification but also optimization during the zonotope-based reachability analysis. By modeling variations as uncertain input currents added to the input range, the zonotope-based reachability analysis is deployed to provide the system performance boundary for the estimation of SRAM dynamic stability region. We are the first to develop the backward Euler-based zonotope evolution with linearization error update and control. Furthermore, the multiparameter large-signal sensitivity calculation is invented in term of zonotope, which is applied for the robustness optimization for SRAM dynamic stability. By simultaneously tuning multiple SRAM transistor widths, the resulted sensitivity of safety distance during reachability analysis can be deployed during the sequential optimizations to

guide SRAM design with operations departing from unsafe region and converge in safe region. In addition, compared to the traditional single-parameter small-signal based sensitivity optimization, our method can converge faster with higher accuracy. Compared to the Monte Carlo-based optimization, our method can achieve speedups up to 600× with similar accuracy.

REFERENCES

[1] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE J. Solid State Circuits*, vol. 22, no. 5, pp. 748–754, Oct. 1987.

[2] E. Grossar, M. Stucchi, K. Maex and W. Dehane, "Read stability and write-ability analysis of SRAM cells for nanometer technologies," *IEEE J. Solid State Circuits*, vol. 41, no. 11, pp. 2577–2588, Nov. 2006.

[3] S. O. Toh, Z. Guo, and B. Nikolić, "Dynamic SRAM stability characterization in 45 nm cmos," in *Proc. VLSIC*, Jun. 2010, pp. 35–36.

[4] A. Singhee, C. F. Yang, J. D. Ma, R. A. Rutenbar, "Probabilistic interval-valued computation: Toward a practical surrogate for statistics inside CAD tools," *IEEE Trans. Comput. Aided Design Integr. Circuits Syst.*, vol. 27, no. 12, pp. 2317–2330, Nov. 2008.

[5] S. Yaldiz, U. Arslan, X. Li and L. Pileggi, "Efficient statistical analysis of read timing failures in SRAM circuits," in *Proc. ISQED*, 2009, pp. 617, 621.

[6] C. Dong and X. Li, "Efficient SRAM failure rate prediction via Gibbs sampling," in *Proc. DAC*, Jun. 2011, pp. 200–205.

[7] H. Yu and S. X.-D. Tan, "Recent advance in computational prototyping for analysis of high-performance analog/RF ICs," in *Proc. ASICON*, Oct. 2009, pp. 760–764.

[8] F. Gong, H. Yu, Y. Shi, D. Kim, J. Ren and L. He, "Quickyield: An efficient global-search based parametric yield estimation with performance constraints," in *Proc. DAC*, Jun. 2010, pp. 392–397.

[9] F. Gong, H. Yu, and L. He, "Fast non-Monte-Carlo transient noise analysis for high-precision analog/RF circuits by stochastic orthogonal polynomials," in *Proc. DAC*, Jun. 2011, pp. 298–303.

[10] F. Gong, X. Liu, H. Yu, S. X-D. Tan, J. Ren and L. He, "A fast non-Monte-Carlo yield analysis and optimization by stochastic orthogonal polynomials," *ACM Trans. Design Autom. Electron. Syst.*, vol. 17, no. 1, pp. 10:1–10:23, Jan. 2012.

[11] H. Wang, H. Yu, and S. X.-D. Tan, "Fast timing analysis of clock networks considering environmental uncertainty," *VLSI J. Integr.*, vol. 45, no. 4, pp. 376–387, Sep. 2012.

[12] W. Wu, F. Gong, R. Krishnan, H. Yu, and L. He, "Exploiting parallelism by data dependency elimination: A case study of circuit simulation algorithms," *IEEE Design Test Comput.*, vol. 30, no. 1, pp. 26–35, Feb. 2013.

[13] F. Gong, S. B. Kazeruni, L. He and H. Yu, "Stochastic behavioral modeling analysis of analog/mixed-signal circuits," *IEEE Trans. Comput. Aided Design Integr. Circuits Syst.*, vol. 32, no. 1, pp. 24–33, Jan. 2013.

[14] K. Agarwal and S. Nassif, "Statistical analysis of SRAM cell stability," in *Proc. DAC*, 2006, pp. 57–62.

[15] D. E. Khalil, M. Khellah, N-S. Kim, Y. Ismail, T. Karnik and V. K. De, "Accurate estimation of SRAM dynamic stability," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 16, no. 12, pp. 1639–1647, Dec. 2008.

[16] B. Zhang, A. Arapostathis, S. Nassif and M. Orshansky, "Analytical modeling of SRAM dynamic stability," in *Proc. ICCAD*, Nov. 2006, pp. 315–322.

[17] S. Srivastava and J. Roychowdhury, "Rapid estimation of the probability of SRAM failure due to MOS threshold variations," in *Proc. CICC*, Sep. 2007, pp. 229–232.

[18] G. M. Huang, W. Dong, Y. Ho and P. Li, "Tracing SRAM separatrix for dynamic noise margin analysis under device mismatch," in *Proc. BMAS*, Sep. 2007, pp. 6–10.

[19] C. J. Gu and J. Roychowdhury, "An efficient, fully nonlinear, variability-aware non-Monte-Carlo yield estimation procedure with applications to SRAM cells and ring oscillators," in *Proc. ASPDAC*, Mar. 2008, pp. 754–761.

[20] W. Dong, P. Li, and G. M. Huang, "SRAM dynamic stability: Theory, variability and analysis," in *Proc. ICCAD*, Nov. 2008, pp. 378–385.

[21] S. Gupta, B. H. Korigh and R. A. Rutenbar, "Towards formal verification of analog designs," in *Proc. ICCAD*, Nov. 2004, pp. 210–217.

[22] G. Frehse, B. H. Krogh, and R. A. Rutenbar, "Verifying analog oscillator circuits using forward/backward abstraction refinement," in *Proc. DATE*, Mar. 2006, pp. 257–262.

[23] D. Walter, S. Little, C. Myers, N. Seegmiller, and T. Yoneda, "Verification of analog/mixed-signal circuits using symbolic methods," *IEEE Trans. Comput. Aided Design Integr. Circuits Syst.*, vol. 27, no. 12, pp. 2223–2235, Dec. 2008.

[24] M. Althoff, S. Yaldiz, A. Rajhans, X. Li, B. H. Krogh, and L. Pileggi, "Formal verification of phase-locked loops using reachability analysis and continuization," in *Proc. ICCAD*, Nov. 2011, pp. 659–666.

[25] Y. Song, H. Fu, H. Yu and G. Shi, "Stable backward reachability correction for PLL verification with consideration of environmental noise induced jitter," in *Proc. ASPDAC*, Jan. 2013, pp. 755–760.

[26] Y. Song, H. Yu, S. Manoj P. D. and G. Shi, "SRAM dynamic stability verification by reachability analysis with consideration of threshold voltage variations," in *Proc. ISPD*, 2013, pp. 43–49.

[27] A. Girard, "Reachability of uncertain linear systems using zonotopes," in *Proc. HSCC*, 2005, pp. 291–305.

[28] M. Althoff, "Reachability analysis and its application to the safety assessment of autonomous cars," Ph.D. dissertation, Dept. Electr. Eng., TUM, Munich, 2010.

[29] U. M. Ascher and L. R. Petzold, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Philadelphia, PA, USA: Society Ind. Appl. Math., 1998.

[30] M. Kvasnica, P. Grieder, and M. Baotić. (2013, Jul.). *Multi-parametric toolbox (MPT)*. MPT 2.6.3 [Online]. Available: http://control.ee.ethz.ch/∼mpt/

Yang Song received the B.S. and M.S. degrees in microelectronics from Shanghai Jiao Tong University, Shanghai, China, in 2006 and 2013, respectively, and is currently pursuing the Ph.D. degree from the University of California, San Diego, CA, USA.

From 2012 to 2013, he was a Project Officer with the VIRTUS IC Design Center of Excellence, Nanyang Technological University (NTU), Singapore, where he was a member of the NTU CMOS Emerging Technology group. His current research interests include applications of reachability analysis on circuit-level verification and optimization.

Hao Yu (M'06–SM'13) received the B.S. degree from Fudan University, Shanghai, China, in 1999 and the Ph.D. degree from the Electrical Engineering Department, University of California, San Diego, CA, USA, in 2007.

He was a Senior Research Staff with Berkeley Design Automation, Berkeley, CA, USA. Since 2009, he has been an Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His current research interests include 3-D-IC and RF-IC at nano-tera scale. He has authored 115 peer-reviewed IEEE/ACM publications.

Dr. Yu was a recipient of the Best Paper Award from the ACM TODAES'10, Best Paper Award nominations in DAC06, ICCAD'06, ASP-DAC'12, Best Student Paper (advisor) Finalist in SiRF'13, RFIC'13 and Inventor Award'08 from semiconductor research cooperation. He is an Associate Editor and Technical Program Committee Member for a number of IEEE/ACM journals and conferences.

Sai Manoj Pudukotai DinakarRao (S'13) received the M.Tech. degree from IIIT, Bangalore, India, in 2012, and is currently pursuing the Ph.D. degree from the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore.

His current research interests include 3-D-IC I/O modeling, thermal, and power management.

Mr. Manoj P. D. was a recipient of the A. Richard Newton Young Research Fellow Award in DAC 2013.