A Robustness Optimization of SRAM Dynamic Stability by Sensitivity-based Reachability Analysis

Yang Song, Sai Manoj P D and Hao Yu
School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore 639798
haoyu@ntu.edu.sg

***Abstract*—A robustness optimization of SRAM dynamic stability at nano-scale is developed in this paper by zonotope-based reachability analysis. A backward Euler method is developed to efficiently perform reachability analysis by zonotope to deal with multiple device parameters with tuning ranges. Moreover, a sensitivity calculation of zonotope is developed to optimize *safety distance* by simultaneously tuning multiple SRAM device parameters without multiple repeated computations. As such, sequential robustness optimizations can be performed such that the optimized SRAM designs can depart from unsafe region but converge into safe region. The proposed method is implemented inside a SPICE-like simulator. As shown by numerical experiments, the proposed method can achieve $600\times$ speedup on average compared to the traditional verification method by Monte-Carlo under the similar accuracy. In addition, compared to the traditional small-signal based sensitivity optimization, the proposed method can converge faster with high accuracy.**

I. Introduction

Stability verification and robustness optimization are emerging needs for integrated circuit (IC) designs at nano-scale. The stability challenge is pronounced for densely integrated SRAM circuits with minimum feature sizes. Note that static noise margin (SNM) [1] is traditionally deployed for SRAM stability characterization because of simple interpretation and measurement. As it may overestimate read-failure and underestimate write-failure, dynamic SRAM stability margin [2] is increasingly adopted by deploying critical word-line pulse-width that can produce a better estimation of failures. The verification of SRAM stability margin becomes harder at nano-scale. Firstly, due to the nonlinear dynamics, the SRAM characteristic behavior becomes less digital but more analog. Secondly, process variations such as threshold-voltage V_{th} [3], [4], [5], [6], [7] can further significantly suppress the SRAM stability margin, and result in higher failure rate during read/write operations.

Many recent works have been performed for SRAM stability characterization [8], [9], [10]. For example, Euler-Newton curve-tracing is utilized to find the boundary between safe and unsafe regions in the *parameter space* without brute-force exploration. But, this method is limited to considering two parameters, and the computational cost is prohibitive considering parameter variations from all transistors. The work in [10] formulates a dynamic stability margin to characterize the stability boundary, namely the *separatrix* [8] in the *state space*. The separatrix provides intuitive illustration of SRAM nonlinear behavior and is formed by combining two transient trajectories which start from the same equilibrium state on separatrix but move towards different directions. Yet, separatrix has to be re-computed every time when any parameter changes. What is more, it is unclear how to consider parameter tuning using separatrix for SRAM robustness optimization.

Note that reachability analysis has been widely deployed in stability verification of system dynamics by exploring potential trajectories of operating points in state space [11], [12]. It can accurately predict boundary of multiple trajectories with uncertain states by one-time simulation, in contrast to obtaining multiple trajectories in repeated simulations. The reachability analysis has been deployed for a number of hard analog circuit verifications [13], [14], [6]. A set of system trajectories in state space can be bounded by the zonotope-based over-approximation. A time-interval integrated reachability analysis with formed zonotope that can distinguish safe and unsafe regions is performed such that the failure in state space can be verified. However, it is unknown how to perform efficient reachability analysis that can be further applied for optimization with consideration of multiple device parameters.

In this paper, to consider multiple device parameters in one simulation, a zonotope-based reachability analysis is developed for robustness optimization of SRAM dynamic stability. Device parameters such as SRAM transistor widths are considered by zonotope matrices during reachability analysis. A corresponding sensitivity calculation is developed and deployed for the optimization of SRAM dynamic stability such that the SRAM designs can depart from unsafe regions. What is more, based on backward Euler method, zonotope-based verification and optimization procedures considering nonlinear device model of transistors are implemented in a SPICE-like simulator. Compared to the traditional Monte-Carlo based verification, our method achieves nearly up to $600\times$ speedup with similar accuracy. Moreover, as multiple-parameter large-signal sensitivity is generated for a safety distance, compared to the traditional single-parameter small-signal based sensitivity optimization, our method can converge faster with high accuracy.

II. Problem Formulation of SRAM Failure Analysis and Optimization

Similar to [8], [9], [10], [6], the scope of this paper focuses on the transistor-level analytical approaches for SRAM dynamic stability optimization under V_{th} variations. When statistical distribution of V_{th} variation is known, one can efficiently generate yield statistics from the transistor-level verification results. In a 6T-SRAM there exists serious stability failure concern with V_{th} variation [6], which can lead to SRAM functional failures during read and write operations. 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 following 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 one operating point x and the safe state p_{safe}.

Note that different from separatrix based approaches [8], [10], 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.

978-1-4799-2816-3/14/$31.00 ©2014 IEEE

A. Failure Mechanisms

Physical mechanisms of SRAM failures i.e. read and write failures in terms of safety distance in the state space are described here. In addition, there exist two convergent regions in the state-variable space of SRAM [8]. 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 on the state-variable plane 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 operating point and the target state. Given enough time, the operating point in any region will converge to the nearest stable equilibrium state. The write operation is aimed at pulling operating point to the target state and thus reducing the safety distance, as shown by point B in Fig.1.

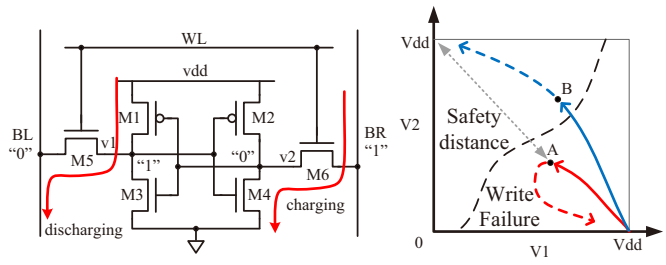


Fig. 1: Illustration of write failure.

V_{th} variations may cause write failure. An increase in V_{th} can reduce the strength of the transistor. For example, increase of V_{th} in M6 along with decrease of V_{th} in M4 can make it more difficult to pull up v_2. If operating point, which slowly moves towards target-state, cannot reach other convergent region before access transistors are closed, it will move back to the initial state implying a write failure. To resolve the failure, tuning width of M6 as increased while M4 as narrowed can help reduce safety distance and hence can mitigate the side effect from V_{th} variations.

2) Read Failure: A read-failure refers to the loss of the previously stored data in SRAM during read operation. Access transistors need careful sizing such that their pull-up strengths are not strong enough to pull digital "0" to "1" during read operation. On the state-variable plane, operating point of SRAM is inevitably perturbed and pulled towards 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 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.

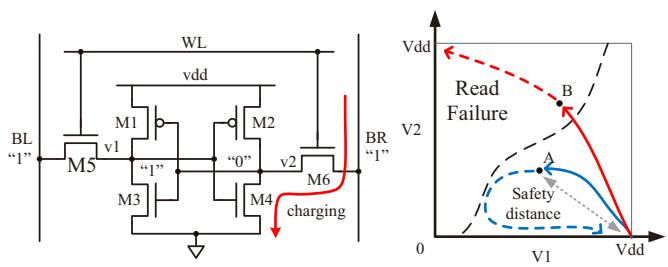


Fig. 2: Illustration of read failure.

V_{th} variations may also cause read failure. For example, variations caused by mismatch between M4 and M6 can result in unbalanced pulling strengths and v_2 can be pulled up more quickly. As a result, the operating point crosses to the other region resulting in read failure, as shown by point B in Fig.2. To resolve the failure, width of M6 needs to be scaled down to avoid excessive pulling strength, which may lead to write failure. In addition, V_{th} 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, the problem to solve 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 V_{th} variations.

B. SRAM Dynamics

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

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

where $x(t)$ is a state variable vector and $u(t)$ is input vector. Here $q(t)$, $u(t)$ contain charges and external sources; and $f(x,t)$ describes SRAM nonlinear dynamics. V_{th} variations of transistors can be introduced at input $u(t)$ as ad-hoc current sources [4].

After Newton iteration is performed at one selected operating point (or nominal point) x^*, $f(x(t),t)$ is linearized at this point as $\frac{\partial f}{\partial x}|_{x=x^*}$. Based on the mean-value theorem, the dynamic equation $f(x)$ at any neighbor operating point x can be expressed by a linear approximation with a 2nd-order residue i.e. the difference between nonlinear $f(t)$ and its linear approximation, called as *linearization error* denoted by L.

The SRAM dynamic equation (1) thereby can be depicted in a simplified form by

$$\begin{aligned} &\frac{d}{dt}q(x,t) + f(x^*,t) + u^*(t) + G(x-x^*) + L = 0; \\ &with \\ &L = \frac{1}{2}(x-x^*)^T \frac{\partial^2 f}{\partial x^2}|_{x=\xi}(x-x^*); \quad G = \left.\frac{\partial f}{\partial x}\right|_{x=x^*} \\ &\xi \in \{x^* + \alpha(x-x^*) | 0 \le \alpha \le 1\}. \end{aligned} \tag{2}$$

Assume that $q(x,t)$ can be further decomposed into $q(x^*)$ and $C\Delta x$. Thus, we have (3)

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

$$\frac{d}{dt}(C\Delta x) + G\Delta x + L = 0 \tag{3b}$$

where $C = \frac{\partial q}{\partial x}|_{x=x^*}$.

(3a) is the nonlinear differential equation for the nominal point x^* and (3b) is the linear equation with the Euclidean distance from x^* to the neighbor point x.

On the basis of (3), reachability analysis can be deployed for SRAM dynamic stability verification and optimization in the state space. Reachability analysis can be performed on nonlinear trajectories with high accuracy by considering L.

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

$$\Delta g_m = \frac{\partial g_m}{\partial W}\Delta W. \tag{4}$$

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

$$\Delta G = \begin{pmatrix} \ddots & & \\ & \frac{\partial g_m}{\partial W} & -\frac{\partial g_m}{\partial W} \\ & -\frac{\partial g_m}{\partial W} & \frac{\partial g_m}{\partial W} \\ & & \ddots \end{pmatrix} \Delta W. \tag{5}$$

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

C. Problem Formulation

Based on the aforementioned SRAM failure mechanisms and dynamics analysis, a robustness optimization for SRAM dynamic stability is proposed in terms of safety distance considering interval values of V_{th} variations from all transistors.

If the safety distance fails to converge to zero, a robust optimization of SRAM dynamic stability can be used to reduce safety distance, we call this as a verification oriented robustness optimization.

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

$$\begin{aligned} \min_{w} \quad & F(w,t) \\ \textit{subject to} \quad & W_{min} < w_i < W_{max}, i = 1, 2, ..., m. \end{aligned} \tag{6}$$

Here, w is the parameter or sizing vector for all m transistors with a defined range $[W_{min}, W_{max}]$. $F(w,t)$ is the objective function, weighted sum of safety distances for both read and write operations given by

$$F(w,t) = \begin{cases} D_w(w,t_w) + D_r(w,t_r), \textit{ write and read failures} \\ D_w(w,t_w), \textit{ write failure only} \\ D_r(w,t_r), \textit{ read failure only} \end{cases} \tag{7}$$

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 is performed in a three-dimensioned parameter space, where a parameter-state point is denoted by $w \in \mathbf{R}^{3\times 1}$. After the reachability analysis is performed, sensitivity of safety distance w.r.t. multiple device parameters can be obtained, which can guide the optimization routine to reduce or even eliminate failures caused by V_{th} variations with improved SRAM dynamic stability.

III. Zonotope-based Reachability Analysis for SRAM Dynamic Stability Verification

Reachability analysis [11], [12], [13], [6] can efficiently determine a reachable region that a dynamic system evolves within a range of states. As such, one can perform one-time reachability analysis to form safe region with safe distance determined from the final state set as shown in Fig.3.With the linear multi-step implementation, the runtime cost or complexity of zonotope-based reachability analysis is similar with transient analysis in SPICE.

Here, we formulate of SRAM robustness optimization by *safety distance* in the framework of reachability analysis.

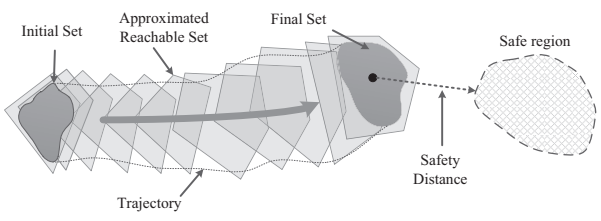


Fig. 3: System trajectory and safety distance with zonotopes.

A. Reachable Set and Zonotope

Interval-value analysis has been applied to model the uncertainties of state variables in [15], such as device parameters. For example, if Δx_1, Δx_2 model uncertainties in 2 different dimensions of state variable x with interval center c, then $x = c + [-1,1]\Delta x_1 + [-1,1]\Delta x_2$ is the neighboring point including these variations. However, there is no formal and efficient verification method developed to deal with multi-dimensional interval-value problem. Here, we model a multi-dimensional interval-value variable as a zonotope [16], [13], which is a convex polytope, to model multiple device parameters.

To start with, an important concept for reachability analysis is the reachable set.

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 hypercube.

One simple and symmetrical type of hypercube, called *zonotope* [11] is defined as follows.

Definition 3:

Zonotope $\mathcal{X}$ is defined by

$$\mathcal{X} = \{x \in \mathbf{R}^{n\times 1} : x = c + \sum_{i=1}^{q} [-1,1] g^{(i)}\} \tag{8}$$

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

One can observe from (8), a zonotope is essentially a multi-dimensional interval in affine form, with each generator $g^{(i)}$ in Fig.4 as a variation in a different direction, implying it can include device and parameter variations. Mathematically zonotope can also be expressed as *Minkowksi summation* [12] of two finite sets such that merged set preserves convexity. When reachability analysis for a nonlinear system is performed, the center of zonotope is utilized as the nominal point for the minimization of linearization error [16].

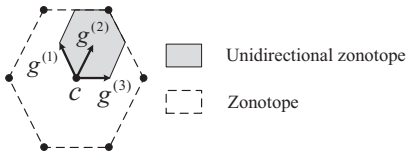


Fig. 4: Zonotope and unidirectional zonotope.

Note that in (8), range of scaling factors for generators is $[-1,1]$ implying the difference vector from nominal point within a reachable set varies in two directions. For the calculation of sensitivity during robustness optimization, the scaling factor is defined within $[0,1]$ to obtain a single-direction distance from the nominal point. We call this modified zonotope as *unidirectional zonotope* (Fig.4), determined by

$$\mathcal{X}^{uni} = \{x \in \mathbf{R}^{n\times 1} : x = c + \sum_{i=1}^{q} [0,1] g^{(i)}\}. \tag{9}$$

Similar to (9), interval values for state matrices can also be modeled as zonotope [12]. Such an interval matrix can be described as *matrix zonotopes* $\mathcal{M}$.

$$\mathcal{M} = \{M \in \mathbf{R}^{n\times n} : M = M^{(0)} + \sum_{i=1}^{q} [0,1] M^{(i)}\}. \tag{10}$$

Similar to zonotopes, the matrix $M^{(0)}$ is called *center matrix* and the matrix $M^{(i)}$ is called *generator matrix*, which contains the variation ranges of perturbed device parameters. Addition and multiplication rules for zonotopes and matrix zonotopes are defined in [12].

B. Reachability Analysis

To solve the dynamic equation of SRAM in (3a), a SPICE-like simulator is applied in this paper. Eq. (3b) with zonotope-based evolution can be solved by backward Euler method with discretized time-step h at kth-iteration by

$$\Delta x_k^{(i)} = A^{-1}(\frac{C}{h}\Delta x_{k-1}^{(i)} - L_k), k = 1, ..., K; i = 1, ..., q. \quad (11)$$

in which $A = \frac{C}{h} + G$ is Jacobian matrix, K and q represents number of time steps and zonotope generators respectively. These zonotope generators $\Delta x_k^{(i)}$ are the Euclidean distances in (2) from the nominal point (zonotope center) x^* to neighbor points x, with zonotopes formed as defined in (9).

The according iteration equation for reachability analysis is built after substituting $\Delta x_k^{(i)}$ by zonotope generator matrix $X_k = [\Delta x_k^{(1)}, ..., \Delta x_k^{(q)}]$, Jacobian matrix A by matrix zonotope $\mathcal{A}$, and capacitance matrix C by matrix zonotope $\mathcal{C}$. As such, one can have

$$X_k = \mathcal{A}^{-1}(\frac{\mathcal{C}}{h}X_{k-1} - L_k), k = 1, ..., K. \quad (12)$$

Here, $\mathcal{A}$ and $\mathcal{C}$ contain variations from multiple device parameters such as transistor width sizings in the case of SRAMs, whose solution is Minkowski summation [12] of individual matrices. In $\mathcal{A}$, interval conductance matrix ΔG can be computed using interval values of transistor width ΔW similar to (5). As such, the zonotope matrix in terms of interval-valued matrices given by

$$\begin{aligned} A &\in [A^{(0)} - \sum_i |A^{(i)}|, A^{(0)} + \sum_i |A^{(i)}|] \\ A^{(i)} &= \frac{\partial A^{(0)}}{\partial W}\Delta W^{(i)} = \Delta G^{(i)}. \end{aligned} \quad (13)$$

Here, $A^{(0)}$ is the nominal state matrix without variations, and $A^{(i)}$ is the variation of state matrix caused by perturbation due to the i-th transistor width $\Delta W^{(i)}$. Similarly other interval conductances including g_{ds} and g_{mb} and capacitances can be derived.

Moreover, in (12) the reciprocal of matrix zonotope $\mathcal{A}$ with m width variations $\mathcal{A} = (A^{(0)}, ..., A^{(m)})$ can be evaluated by two steps. Firstly, $(A^{(0)})^{-1}$ is calculated by LU decomposition. And then, one can have $\mathcal{A}^{-1}$ expanded as

$$\mathcal{A}^{-1} = ((A^{(0)})^{-1}, ..., (A^{(0)})^{-1}A^{(m)}(A^{(0)})^{-1}). \quad (14)$$

This approach leads to an implementation of reachability analysis by SPICE-like simulator.

IV. Robustness Optimization of SRAM Dynamic Stability

In this section, we first introduce safety distance under zonotope, and discuss the according sensitivity calculation of safety distance, which is applied for SRAM dynamic stability optimization by tuning multiple SRAM device parameters simultaneously.

A. Safety Distance

Assume that one safe state is located at p_{safe} in the state space. As for any zonotope in the form of (8), the safety distance for the reachable set can be expressed as

$$\mathcal{D} = \{d \in \mathbf{R}^{n\times 1} : d = p_{safe} - c - \sum_{i=1}^{q}[0, 1]g^{(i)}\}. \quad (15)$$

As shown in Fig.5, for one specific point inside a reachable set, safety distance D can be determined as

$$D = ||p_{safe} - c - \sum_{i=1}^{q}\varepsilon^{(i)}g^{(i)}||_2, 0 \le \varepsilon^{(i)} \le 1 \quad (16)$$

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

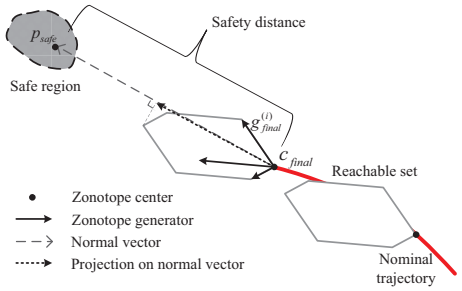


Fig. 5: Safety distance and its sensitivity in reachability analysis.

B. Sensitivity of Safety Distance

With the use of reachability analysis by zonotope, trajectory of SRAM is obtained at final reachable set x_{final} with center c_{final} as

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

Safety distance D for a reachable set can vary within a certain range as the perturbation of device parameters (13) can result in different operating points close-by.

Given the perturbation range of device parameters $[0, \Delta W]$ in form of interval entries of the matrix zonotope (10), the variation range of safety distance can be obtained from the final reachable set by

$$\Delta D = \sum_{i=1}^{q}\frac{(p_{safe} - c_{final})^T}{||p_{safe} - c_{final}||_2}g_{final}^{(i)}. \quad (17)$$

As shown in Fig.5, the perturbation of safety distance at the final reachable set is obtained by projecting zonotope generators $g_{final}^{(i)}$ to the normal vector $\frac{(p_{safe}-c_{final})^T}{||p_{safe}-c_{final}||_2}$, which is formed from zonotope center c_{final} to safe region p_{safe}.

As such, one can calculate the large-signal sensitivity $S(D, w)$ of the safety distance D w.r.t. device parameter w by

$$S(D, w) := \frac{\Delta D}{\Delta w}, \quad (18)$$

which becomes the ratio of 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 (11) with

$$s = \frac{\partial x_k}{\partial w} = -(\frac{C}{h} + G)^{-1}\frac{\partial G}{\partial w}(\frac{C}{h} + G)^{-1}(\frac{C}{h}x_{k-1} - u_k) \quad (19)$$

For the simplicity of presentation linearization error L_k and derivatives of capacitance matrices are omitted and state variable x_{k-1} is assumed as constant. Even though the single-parameter small-signal sensitivity is easy to obtain, compared to the large-signal sensitivity $S(D, w)$ in (18), s may fail to measure the accumulated variation from the previous states by multiple parameters. Moreover, small-signal sensitivity fails to provide accurate direction during optimization due to exclusion of nonlinearity. In contrast, the proposed multi-parameter large-signal sensitivity can be effectively utilized for SRAM dynamic stability optimization, which has faster convergence with higher accuracy as shown in experiment results.

C. Safety Distance Optimization by Sensitivity

The sensitivity of safety distance derived from reachability analysis can guide the optimization direction that departs from unsafe region and shorten safety distance by tuning multiple device parameters. So, it can be embedded in any gradient-based optimization algorithm to achieve a robust SRAM design under process variations. Moreover it can be applied for any circuits when safety distance is properly defined because no specific knowledge of circuit type is required. In case of robust stability optimization for SRAMs, transistor widths are

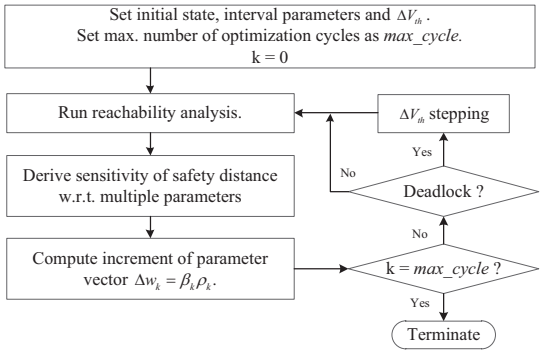


Fig. 6: Flow-chart of robust stability optimization of SRAMs.

sized to improve the dynamic stability. The complete flow of SRAM optimization is shown in Fig.6 with following detailed steps.

Firstly, in each searching step, increment Δw_k of parameter vector w_k is in the same direction with the sensitivity of safety distance.

$$\Delta w_k = \beta_k \rho_k \tag{20}$$

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

$$F(w_k,t) + \Delta w_k^T \rho_k = 0. \tag{21}$$

As such, one can obtain $\beta_k = -\frac{F(w_k,t)}{\rho_k^T \rho_k}$ after combining (20) with (21). Increment of parameter vector Δw (20) is obtained afterwards.

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

$$\beta_k = -\gamma \frac{F(w_k,t)}{\rho_k^T \rho_k}, 0 < \gamma < 1 \tag{22}$$

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

The proposed optimization technique is implemented in a SPICE-like simulator by MATLAB. Manipulations of zonotopes are performed by MATLAB toolbox named Multi-Parametric Toolbox (MPT) [17]. Experiment data is collected on a desktop with Intel Core i5 3.2GHz processor and 8GB memory.

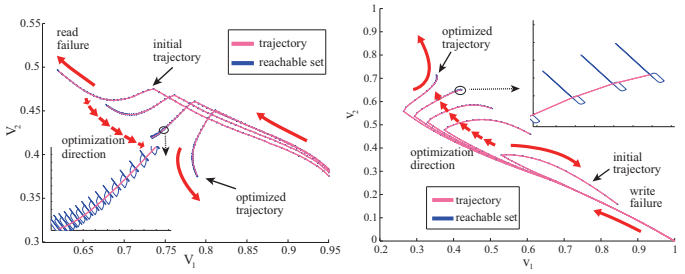
SRAM circuit with $40nm$ CMOS is used as the technology node. Supply voltage of SRAM v_{dd} is set to $1V$ with initial states set as $v_1 = 1V$ and $v_2 = 0$. The transistor widths are varied in the range of $[100, 600]nm$ with step of $1nm$. A budget of 30% threshold-voltage variation is considered for each transistor during the SRAM verification and optimization. For the robustness optimization, interval values of transistor widths are considered in zonotope matrix to derive sensitivity.

According to Section II, strong pulling strength of M1, M4 and other weak transistors lead to high probability of write failure. Thus the negative V_{th} variations are assumed for M1, M4 and positive variations for other transistors. Similarly for read operation, negative V_{th} variations in M2, M3, M6 and positive variations for other transistors are assumed. Variation magnitudes used for optimization are initially set as standard deviations.

A. Stability Optimization Results

1) Optimization of Read or Write Failure: Firstly, stability optimization for read operation only is performed with transistor widths $[W_1, W_3, W_5] = [200, 300, 300]nm$ and $9ns$ pulse width. The process of read stability optimization is shown in Fig.7(a), with trajectories plotted in light purple; and reachable sets (i.e. zonotopes) due to parameter changes in dark blue. Zonotopes are quite small due to small transistor width variation range $1nm$, resulting in limited variation range of trajectory. Multi-parameter large-signal sensitivity w.r.t. one zonotope set is calculated here, which is different from single-parameter small-signal sensitivity in (19).

At each nominal point 3 reachable sets are generated with different transistor widths and final sets are used to derive large-signal sensitivities (Fig.5). After 4 iterations, the optimized trajectory recovers from read failure. The optimized widths are $[148, 343, 217]nm$.

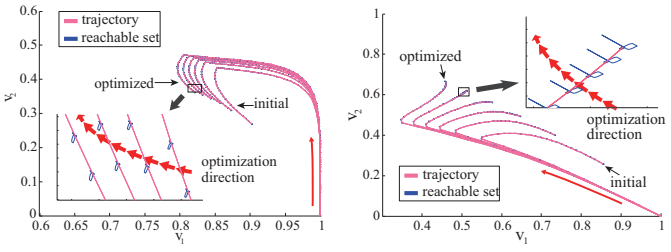


(a) optimization of read operation only (b) optimization of write operation only

Fig. 7: Optimization procedure for SRAM dynamic stability

Then, we perform stability optimization for write operation only with initial transistor widths $[400, 500, 350]nm$ and $0.050ns$ pulse width. The stability optimization by large-signal sensitivity calculated from reachability analysis can help the system trajectory to converge to safe region within 5 iterations (Fig.7(b)). The optimized transistor widths are $[381, 440, 497]nm$.

2) Optimization of Read and Write Failure: Initial transistor pair widths for read and write failure optimization are randomly chosen as $[200, 400, 400]nm$ with pulse width $9ns$ for read and $0.024ns$ for write operations.

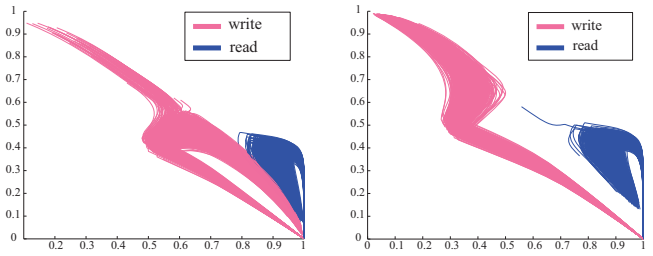


(a) optimization of read operation (b) optimization of write operation

Fig. 8: Optimization procedure for read and write operation

The optimization directions of trajectory for read and write operations are shown in Fig.8(a) and Fig.8(b) respectively. The trajectory after performing initial optimization is represented as *initial*. From Fig.8(b), we can observe that at the beginning, write failure happens as the trajectory converges to initial state. With the use of proposed optimization technique, the trajectory of write operation moves towards target state and converges after 6 iterations. Meanwhile, if the read operation in Fig.8(a) is considered, initially read failure did not happen. As the write operation optimizes, trajectory for read operation moves upward. 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 is $[W_1, W_3, W_5] = [192, 330, 586]nm$. Statistical analysis by Monte-Carlo with 1000 samples is performed for SRAM before and after



(a) before optimization (b) after optimization

Fig. 9: Statistical yield calculation before and after optimization

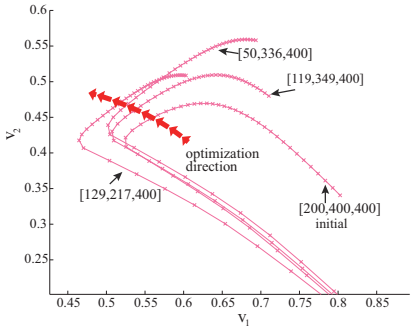


Fig. 10: Optimization of write operation by small-signal sensitivity

optimization (Fig.9). 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-voltage variations during the optimization.

B. Comparisons

Time consumption of our proposed method and traditional Monte-Carlo method for read and write failure optimization shown in Fig.8 are compared here. Monte-Carlo based optimization is performed with 2000 samples. Runtime of optimization and transistor widths at each iteration is listed in Table I, and more than 600× runtime speedup is achieved in our approach. *Iter* represents the iteration number; time consumed in *seconds* for optimization by proposed and traditional Monte-Carlo method are shown under *Sensitivity based RA* and *MC* columns respectively. Achieved speedup is listed under *speedup* column. For example, the proposed method takes about 9 seconds while Monte-Carlo based method needs nearly 2 hours for first iteration. 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. Variation of transistor widths at each iteration can also be observed. In our case, to derive large-signal sensitivity w.r.t. three transistor pairs, reachability analysis is performed 3 times.

TABLE I: 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×

Furthermore, we compare the our approach with another optimization routine by single-parameter small-signal sensitivity (19). For the same aforementioned test-case, the optimization result by small-signal sensitivity in shown in Fig.10. Unlike in Fig.8(b), the optimization routine by single-parameter small-signal sensitivity fails to find a feasible solution, and results in negative width after 3 iterations. Transistor pair widths, i.e. $[W_1, W_3, W_5]$ are shown in Fig.10. Note that W_5 fails to be tuned during optimization, because small-signal sensitivity w.r.t. W_5 is much smaller than the rest. Since the small-signal sensitivity only depends on the location of final state, the resulted gradient merely has local accuracy and changes irregularly as the trajectory moves and fails to converge. The proposed large-signal sensitivity by reachability analysis can achieve much higher accuracy for a faster converged SRAM optimization.

VI. Conclusions

To consider multiple device parameters during optimization, a zonotope-based reachability analysis is developed for robustness optimization of SRAM dynamic stability. Efficient backward Euler method is developed for zonotope-based reachability analysis for SRAM failure verification. Moreover, the proposed approach can generate sensitivity of zonotope by considering of multiple device parameters with large-signal changes. The resulted sensitivity of safety distance during reachability analysis can be deployed during sequential optimizations to guide SRAM designs departing from unsafe region. Compared to the traditional Monte-Carlo based verification, our method is 600× faster with similar accuracy. In addition, compared to the traditional single-parameter small-signal sensitivity based optimization, our method converges faster with high accuracy.

References

[1] E. Grossar and et.al. Read stability and write-ability analysis of SRAM cells for nanometer technologies. *IEEE JSSC*, 41(11):2577–2588, Nov 2006.

[2] S. O. Toh and et.al. Dynamic SRAM stability characterization in 45nm CMOS. In *IEEE Symp. on VLSI Circuits*, 2010.

[3] 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 *ACM/EDAC/IEEE DAC*, 2011.

[4] F. Gong and et.al. A fast non-monte-carlo yield analysis and optimization by stochastic orthogonal polynomials. *ACM TODAES*, 17(1):10:1–10:23, Jan 2012.

[5] H. Wang, H. Yu, and S. X-D. Tan. Fast timing analysis of clock networks considering environmental uncertainty. *Integration, the VLSI Journal*, 45(4):376 – 387, Sep 2012.

[6] Y. Song and et.al. SRAM dynamic stability verification by reachability analysis with consideration of threshold voltage variation. In *ACM ISPD*, 2013.

[7] S. B-Kazeruni and et.al. SPECO: Stochastic perturbation based clock tree optimization considering temperature uncertainty. *Elsevier Integration, the VLSI Journal*, 46(1):22 – 32, Jan 2013.

[8] G. M. Huang and et.al. Tracing SRAM separatrix for dynamic noise margin analysis under device mismatch. In *IEEE Int. BMAS Workshop*, 2007.

[9] 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 *IEEE/ACM ASP-DAC*, 2008.

[10] W. Dong and et.al. SRAM dynamic stability: theory, variability and analysis. In *ACM/IEEE ICCAD*, 2008.

[11] A. Girard. Reachability of uncertain linear systems using zonotopes. In *Int. conf. on Hybrid Systems: computation and control*. Springer, 2005.

[12] M. Althoff. Reachability analysis and its application to the safety assessment of autonomous cars. In *PhD Dissertation, TUM*, 2010.

[13] M. Althoff and et.al. Formal verification of phase-locked loops using reachability analysis and continuization. In *ACM/IEEE ICCAD*, 2011.

[14] Y. Song and et.al. Stable backward reachability correction for PLL verification with consideration of environmental noise induced jitter. In *IEEE/ACM ASP-DAC*, 2013.

[15] A. Singhee and et.al. Probabilistic interval-valued computation: Toward a practical surrogate for statistics inside CAD tools. *IEEE Trans. on CAD*, 27(12):2317–2330, Nov 2008.

[16] M. Althoff, O. Stursberg, and M. Buss. Reachability analysis of nonlinear systems with uncertain parameters using conservative linearization. In *IEEE Conf. on Decision and Control*, 2008.

[17] M. Kvasnica and et.al. Multi-parametric toolbox (mpt). MPT 2.6.3 is available at http://control.ee.ethz.ch/˜mpt/.