NANYANG TECHNOLOGICAL UNIVERSITY

ASP-DAC 2014

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

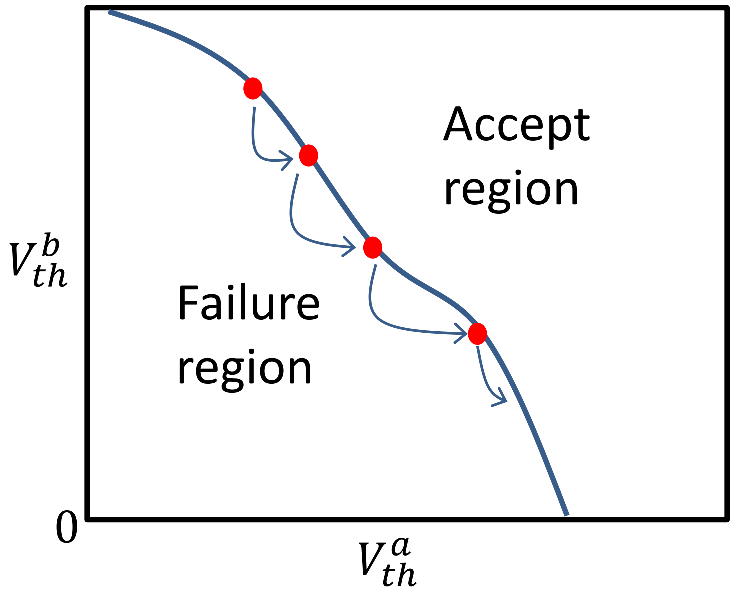
Outline

- **SRAM Robustness Optimization Problem**
- **Reachability Analysis by Zonotope**
- **Safety Distance and Large-signal Sensitivity**
- **SRAM Reliability Optimization**
- **Experimental Results**
- **Summary**

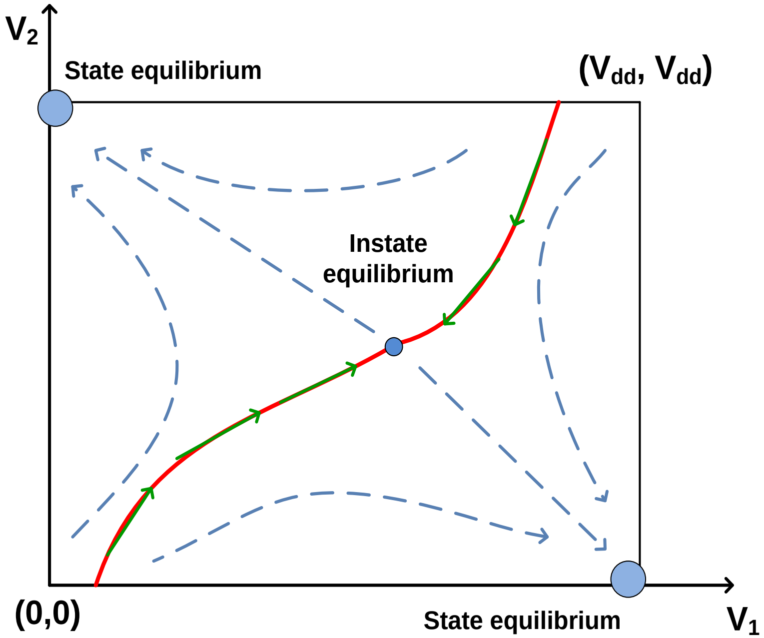
SRAM Robustness Optimization

- **Stability verification and robustness optimization become hard for SRAM circuits**
  - Process variations, mismatch among transistors cause failures at advanced nodes
- **Static noise margin (SNM):** overestimates read failure and underestimates write failure
- **Dynamic stability margin** is adopted by deploying critical word-line pulse-width
  - How to verify and optimize?

Previous Work of SRAM Verification



- Search for points on boundary of failure region in parameter space*1.



- Separatrix: boundary separating two stable regions in parameter space*2

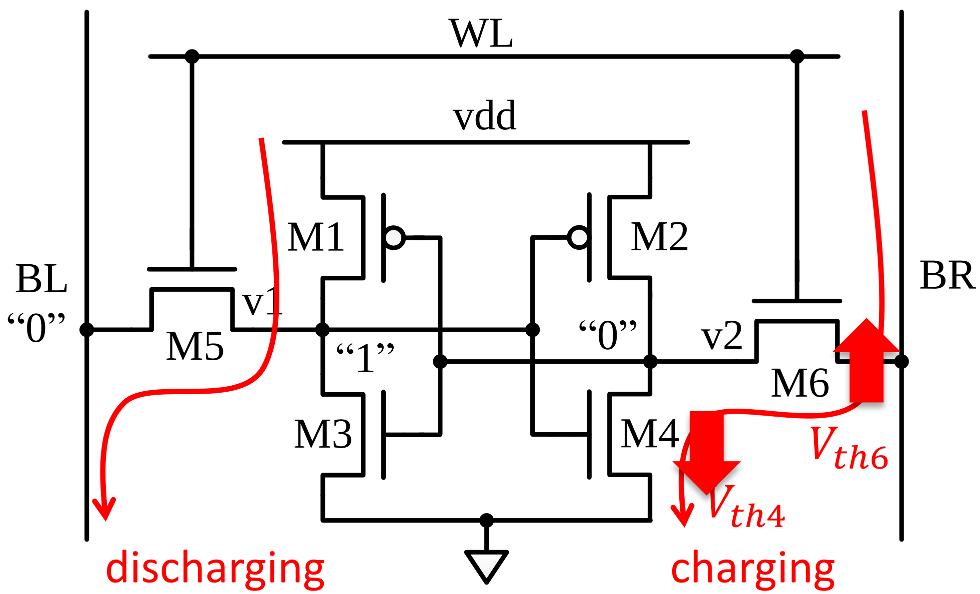
Confined in 2-D space, i.e. only two parameters considered.

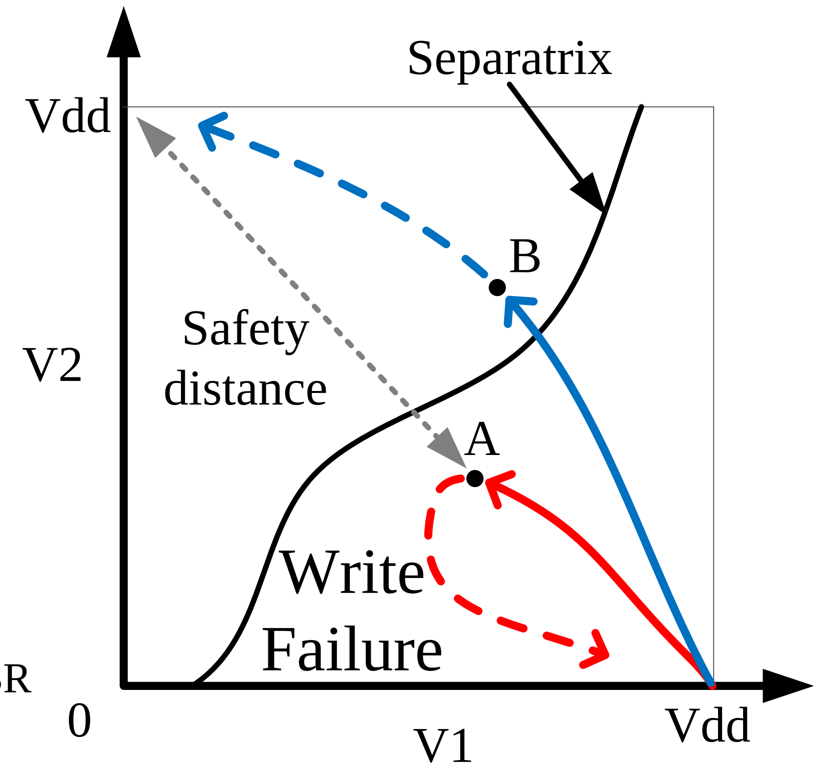
*1 W. Dong and et.al. *ICCAD*, 2008

*2 G M Huanag and et.al. IEEE Int. BMAS Workshop, 2007

Write Failure Analysis by Safety Distance

- Initial state (v1,v2) = (vdd, 0)
- Target state (v1, v2) = (0, vdd)





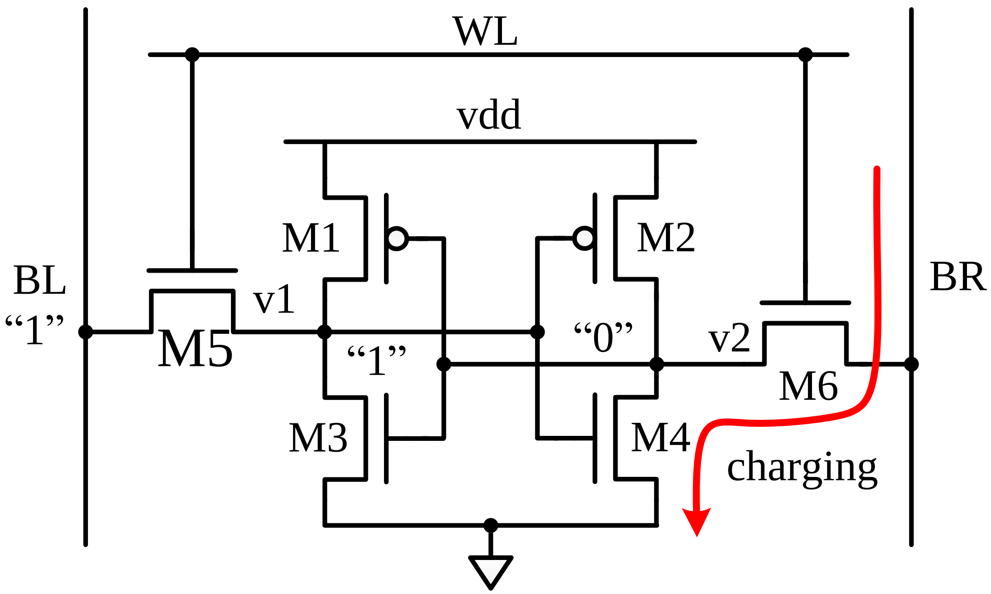
- Threshold voltage variation causes difficulty to move state point to the target state.

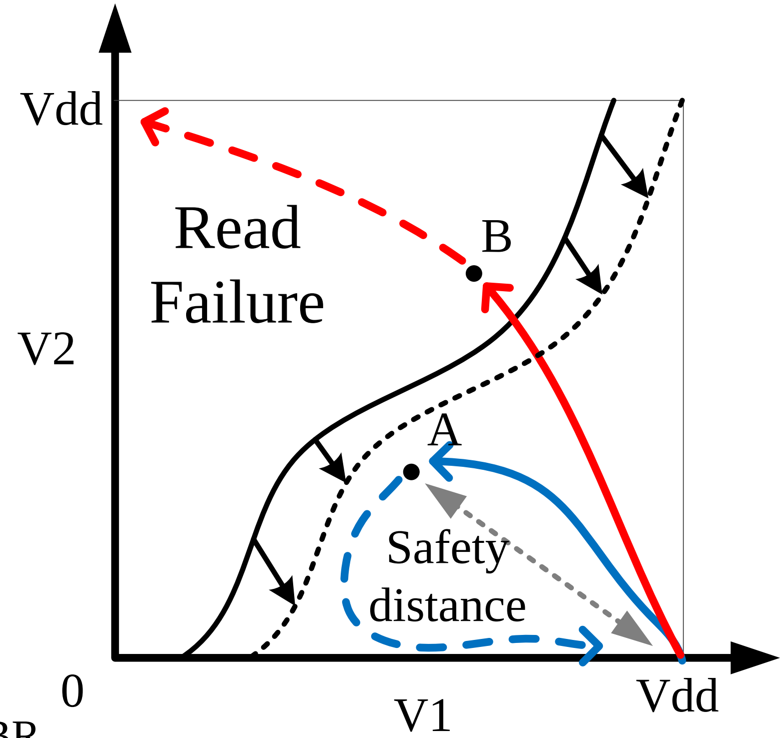
Safety Distance *is the Euclidean distance in the state space between the operating and the safe state region.*

Read Failure Analysis by Safety Distance

- Initial state (v1,v2) = (vdd, 0)
- Target state (v1, v2) = (Vdd,0)

- Internal state is aimed to maintain regardless perturbation during read operation.





- Mismatch between M4 and M6
- Mismatch among M1-4

SRAM Nonlinear Dynamics

- Nonlinear dynamics of SRAM can be defined as

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

Charges
State vector
Input

- Based on mean-value theorem $f(x)$ at neighborhood of the nominal poi[illegible]n b[illegible]ear approximation with 2^{nd} order residue as

$$\frac{d}{dt}q(x,t) + f(x^*,t) + u^*(t) + G(x - x^*) + \boxed{\frac{1}{2}(x - x^*)^T \cdot \frac{\partial^2 f}{\partial x^2}|_{x=\xi} \cdot (x - x^*)} = 0.$$

Linearization error

- Assuming q(x,t) can be decomposed

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

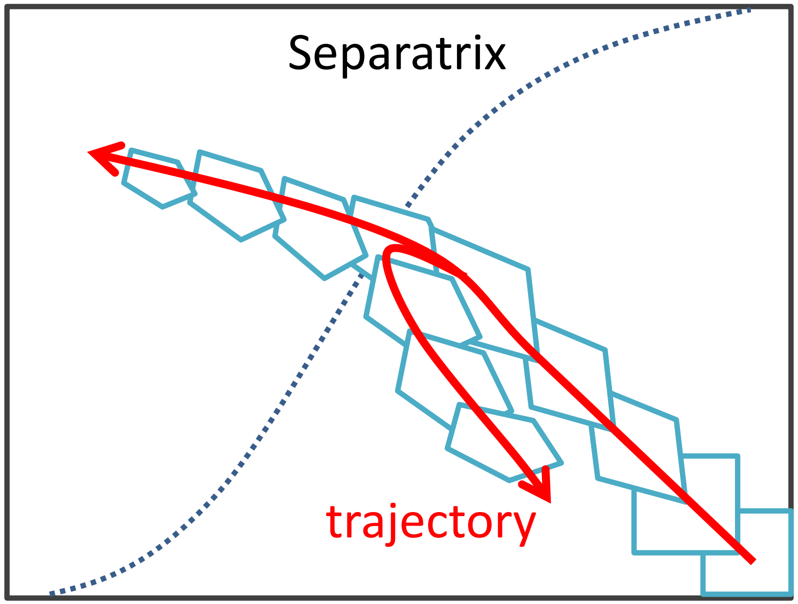
Nonlinear dynamics at nominal point

Linear dynamics at nominal point

$$C = \frac{\partial q}{\partial x}|_{x=x^*} \qquad G = \frac{\partial f}{\partial x}|_{x=x^*}$$

SRAM Verification by Reachability Analysis

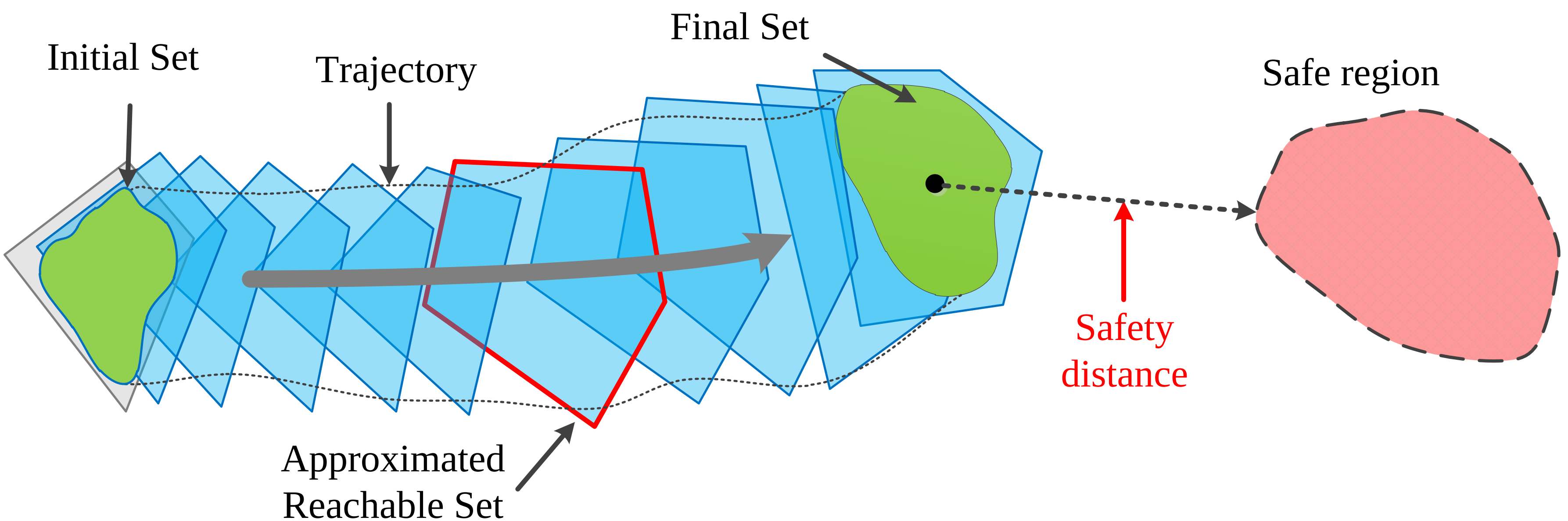
State-variable space



- Fast verification of SRAM nonlinear dynamics by reachability analysis
  - Variations from multiple sources considered at the same time
- For example, transconductance of multiple transistors considering variation in their widths can be added as follows

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

Reachability Analysis: Unidirectional Zonotope

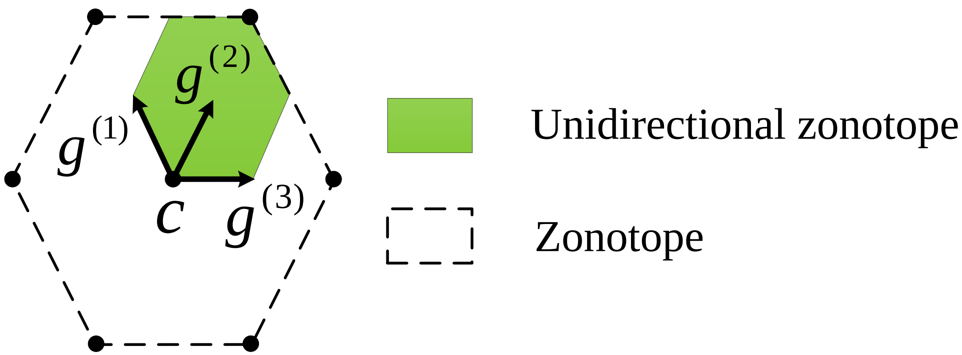


Zonotope

Set of points in n-dimensional polygon with generator $g^{(i)}$

Unidirectional zonotope

$$z^{uni} = c + \sum_{i=1}^{e} \beta_i g^{(i)}, 0 \leq \beta_i \leq 1$$



Reachability Analysis with Uncertain Parameters

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

SPICE-like simulator

Backward Euler method

- Backward Euler method with discretized time-step h at k-th iteration by

$$\Delta {x_k}^{(i)} = A^{-1}(\frac{C}{h} x^{(i)}{}_{k-1} - L_k)$$

Euclidean distance

$$A = \frac{C}{h} + G$$

- Considering all parameter variations as zonotope, linear Multi-step integration for reachability analysis

$$X_k = \mathcal{A}^{-1}(\frac{\mathcal{C}}{h} X_{k-1} - L_k)$$

$$X_k = \left[\Delta {x_k}^{(1)}, \dots, \Delta {x_k}^{(m)}\right]$$

Zonotope generator matrix

Reachability Analysis with Uncertain Parameters

- Parameter variations can be considered by the interval matrix A which is represented by a matrix zonotope.
- Zonotope matrix represented in terms of interval-valued matrices as

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

in which each matrix zonotope generator $A^{(i)}$ contains the variation range of a parameter.

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

- The inverse of $\mathcal{A}$ width variations is approximated as follows

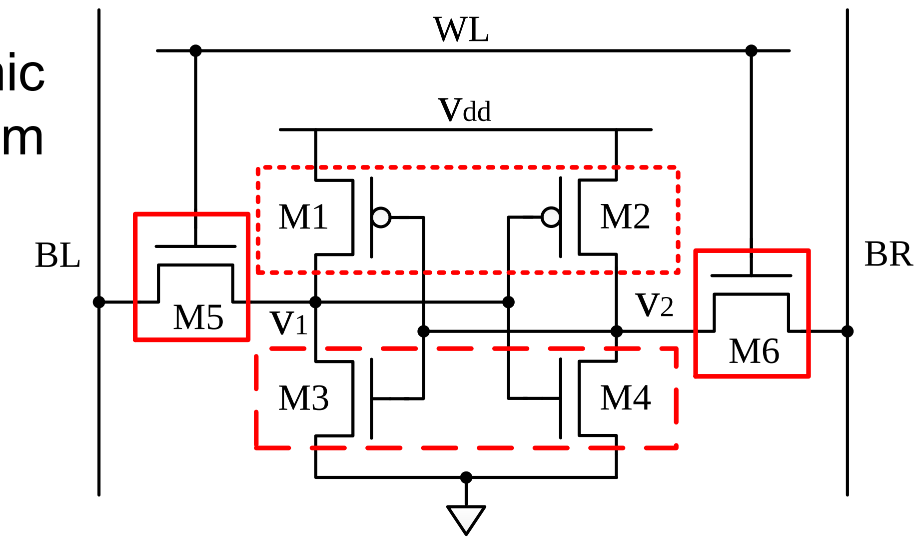
$$\mathcal{A}^{-1} = \left(\left(\mathcal{A}^{(0)}\right)^{-1}, \ldots, \left(\mathcal{A}^{(0)}\right)^{-1} \mathcal{A}^{(m)} \left(\mathcal{A}^{(0)}\right)^{-1} \right)$$

Inverse of $\mathcal{A}^{(0)}$ is computed by LU decomposition

SRAM Robustness Optimization Problem

- Optimization of SRAM dynamic stability is modeled as a minimum value problem.

$$\min \; F(\vec{w})$$
$$S.T. W_{min} < w_i < W_{max}$$



Due to symmetric structure, $\vec{w} \in R^{3\times1}$

Safety distance for write operation

Pulse width

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

Safety distance for read operation

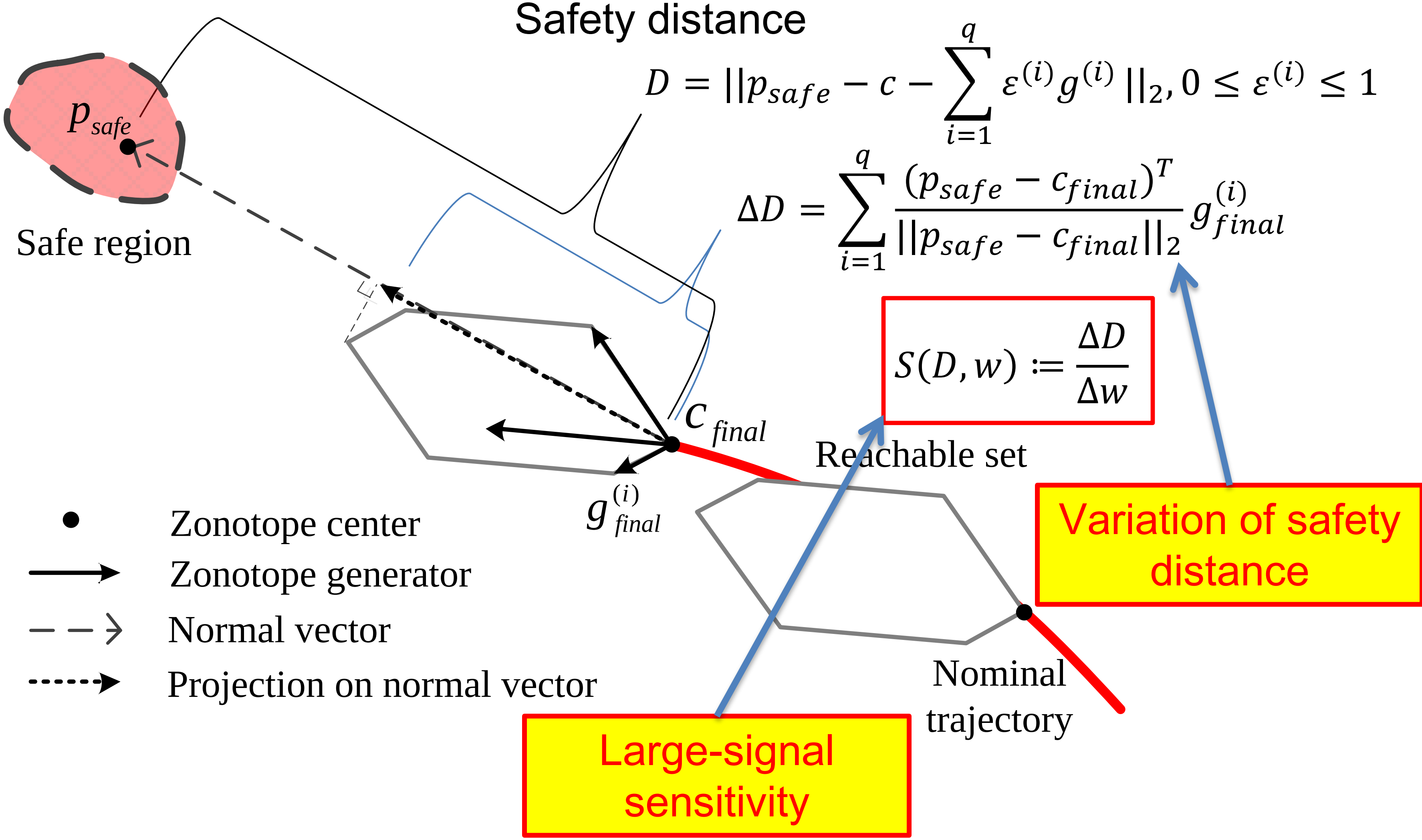
Large-signal Sensitivity of Safety Distance

Safety distance

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

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

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



Reliability Optimization with Sensitivity of Safety Distance

- Based on the calculated sensitivity, optimization is performed
- Increment of parameter vector in direction of optimization is

$$\Delta w_k = \beta_k \underline{\rho_k}$$ the gradient of obj. function

- Suppose gradient is constant in the state-variable space:

$$F(w_k, t) + {\Delta w_{k+1}}^T \rho_k = 0$$

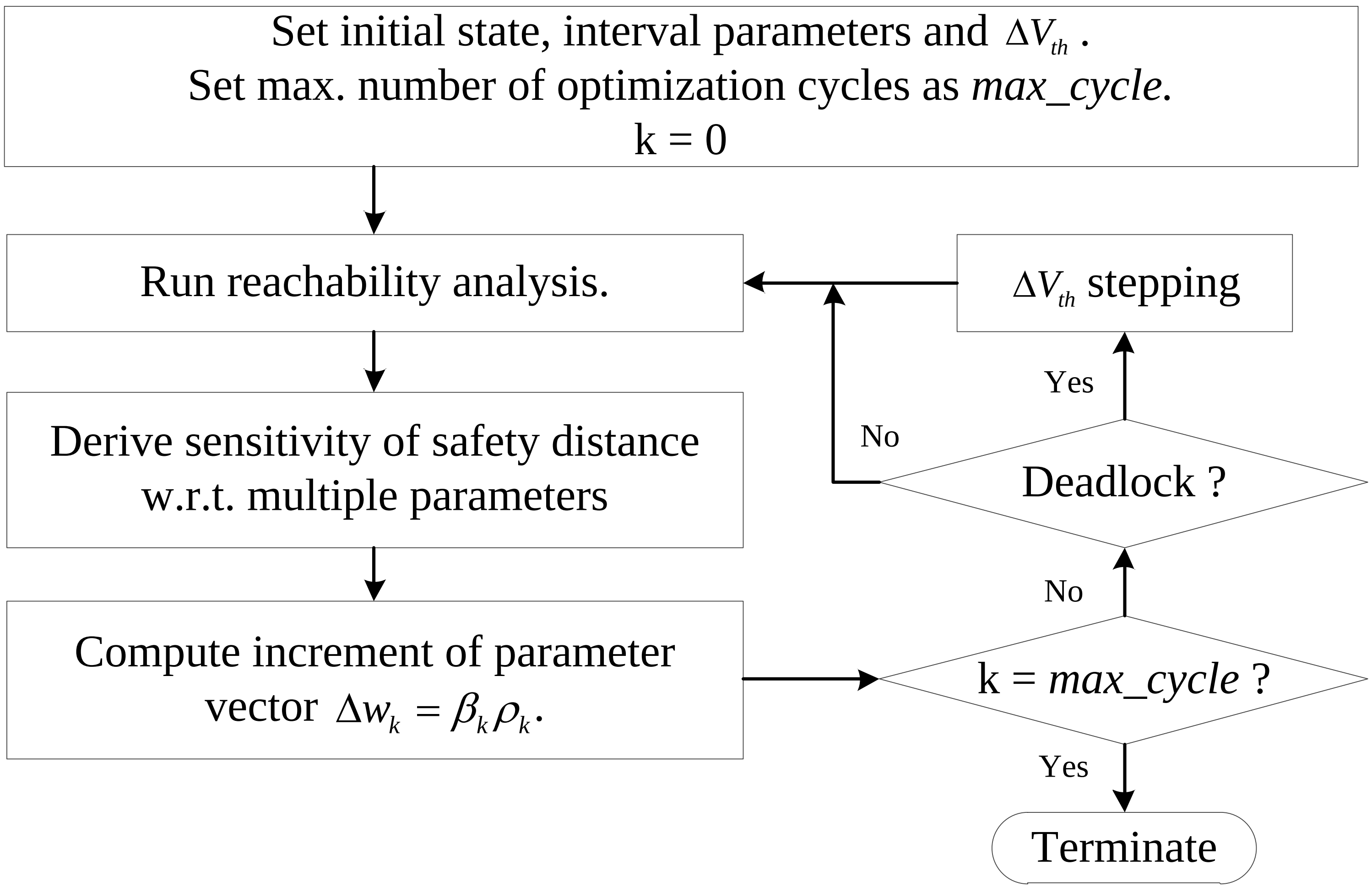
$$\beta_{k+1} = -\frac{F(w_k, t)}{{\rho_k}^T \rho_k}$$

- Virtually gradient decreases as safety distance becomes smaller. In other words, gradient should be smaller in the next search step.

$$\beta_{k+1} = -\gamma \frac{F(w_k,t)}{{\rho_k}^T \rho_k}, \; 0 < \gamma < 1$$

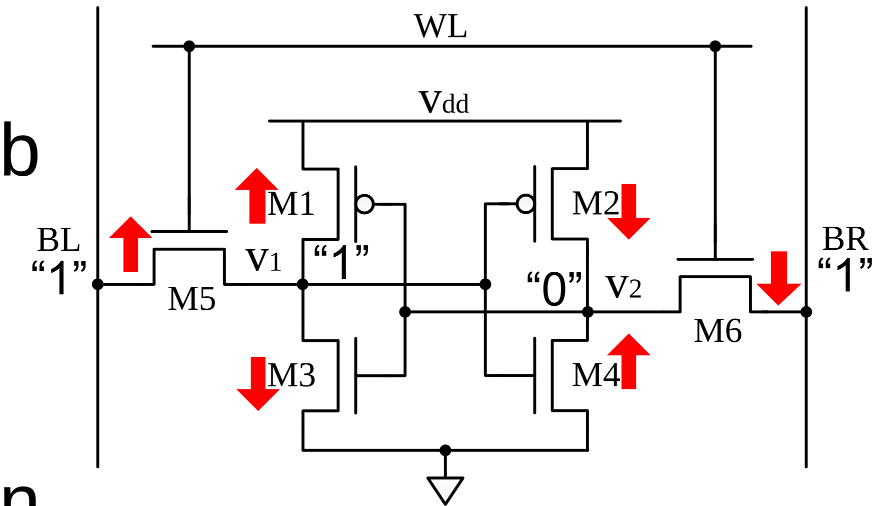
Use empirical factor γ to modulate step size.

SRAM Robustness Optimization Flow

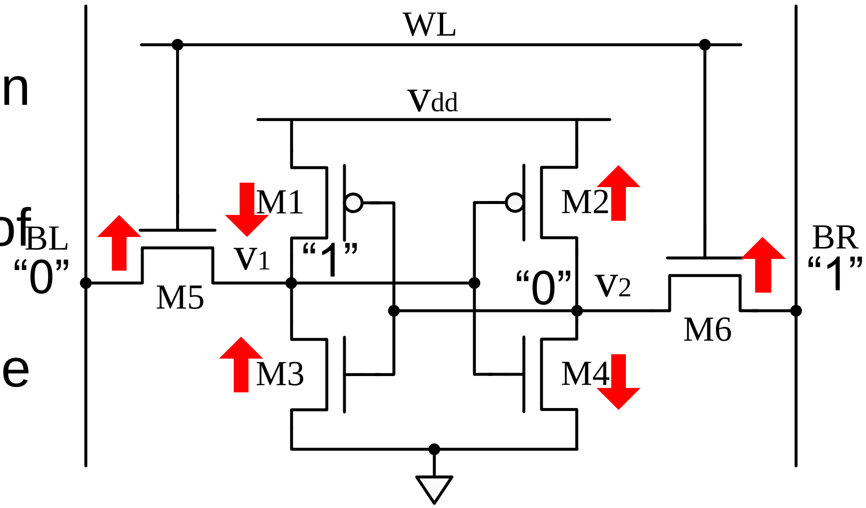


Experimental Results

- Implemented in Matlab Platform
  - Core i5 3.2GHz processor
  - 8GB memory
- Preset simulation parameters
  - Transistor width varies within $[100,600]nm$
  - Relative standard deviation of $V_{th} = 10\%$.
  - Optimization is performed for the most adverse situations.

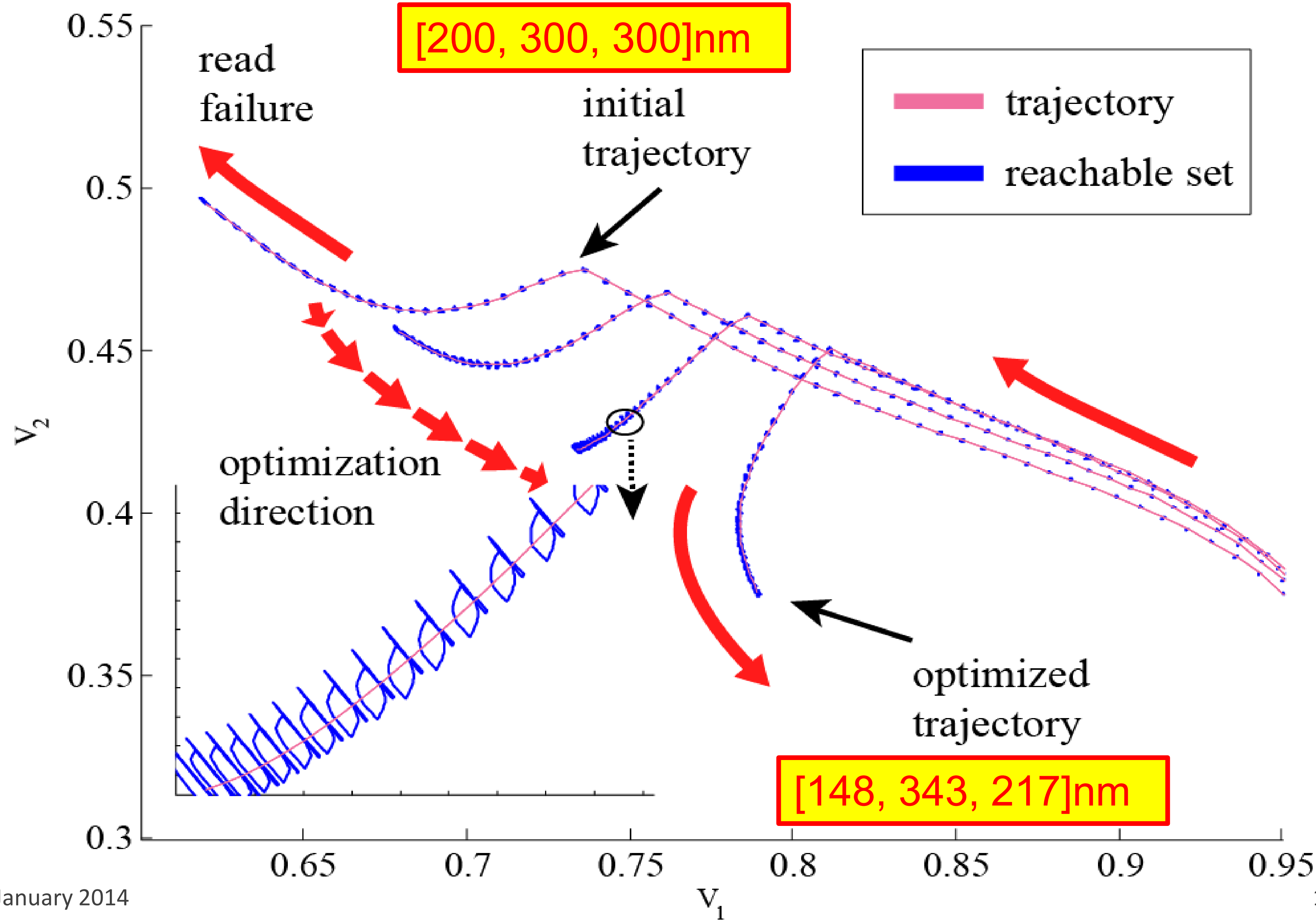


Threshold variations for read operation

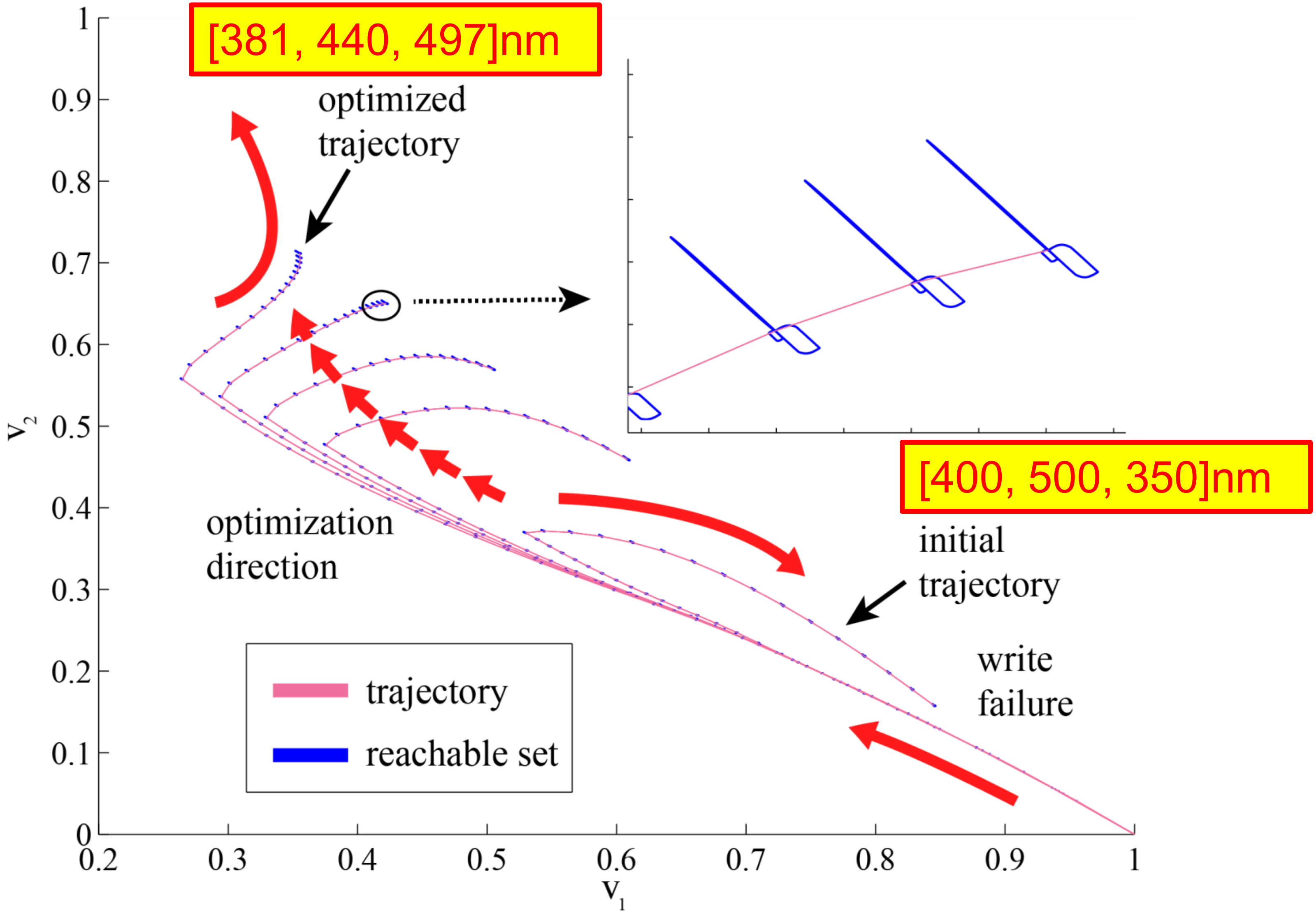


Threshold variations for write operation

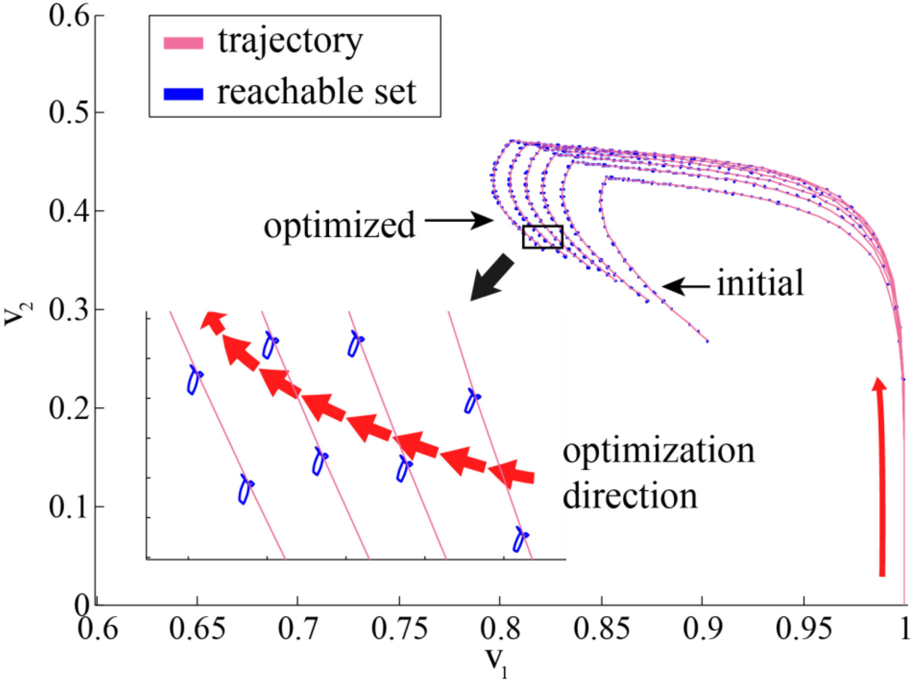
Optimization of Read Failure



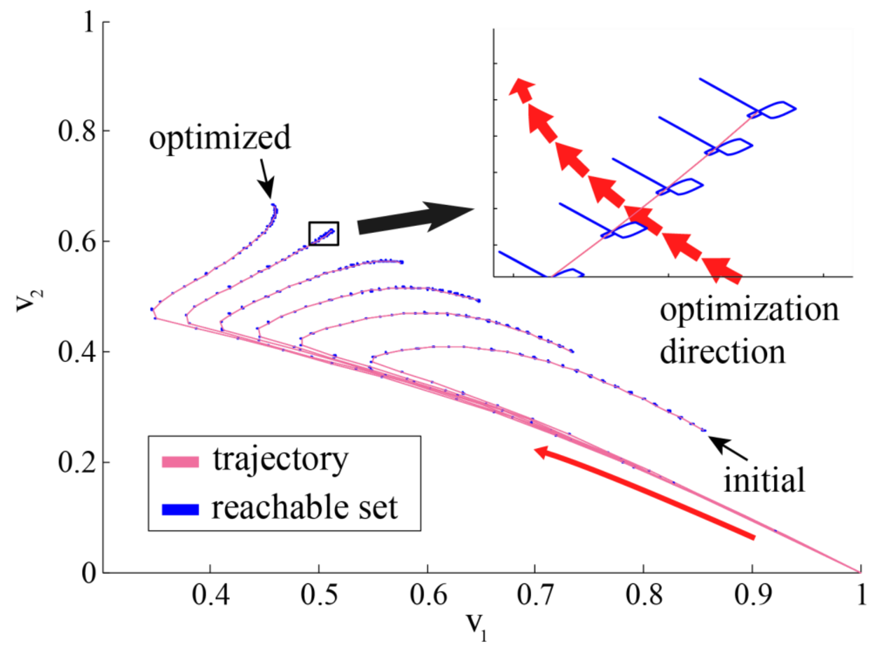
Optimization of Write Failure



Optimization of Read + Write Failure

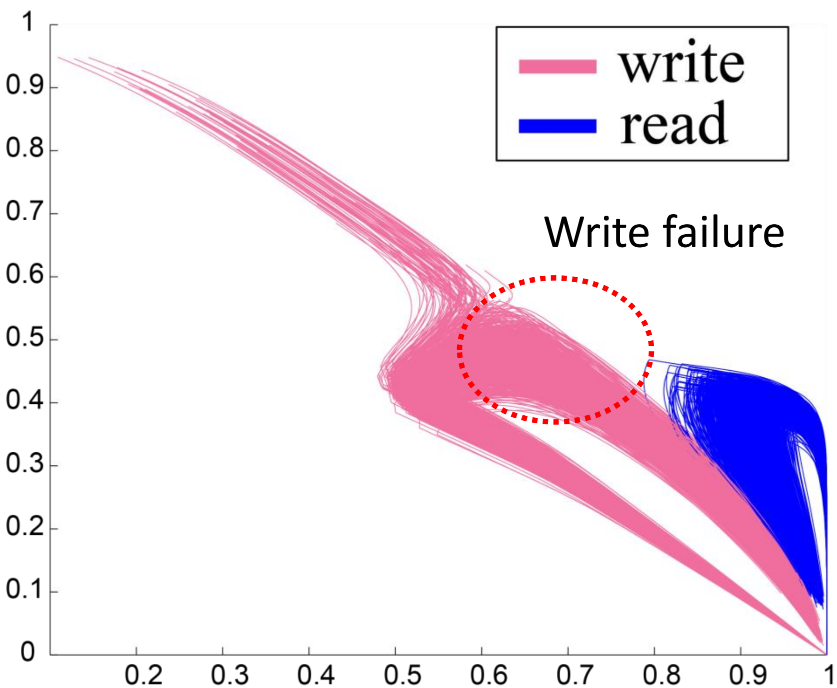


- Optimization result: $[W_1, W_3, W_5] = [192{,}330{,}586]\text{nm}$



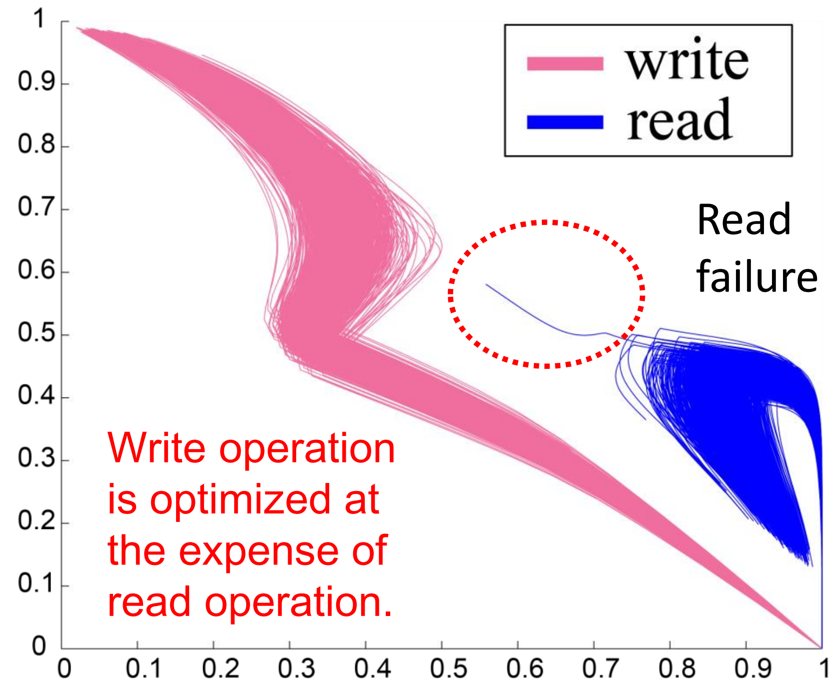
- Transistor widths : $[W_1, W_3, W_5] = [200{,}400{,}400]\text{nm}$
- Read pulse width is 9ns.
- Write pulse width is 0.024ns.

Yield Before and After Optimization



$$Yield\ rate \equiv 1 - \frac{N_{failure}}{N_{total}}$$

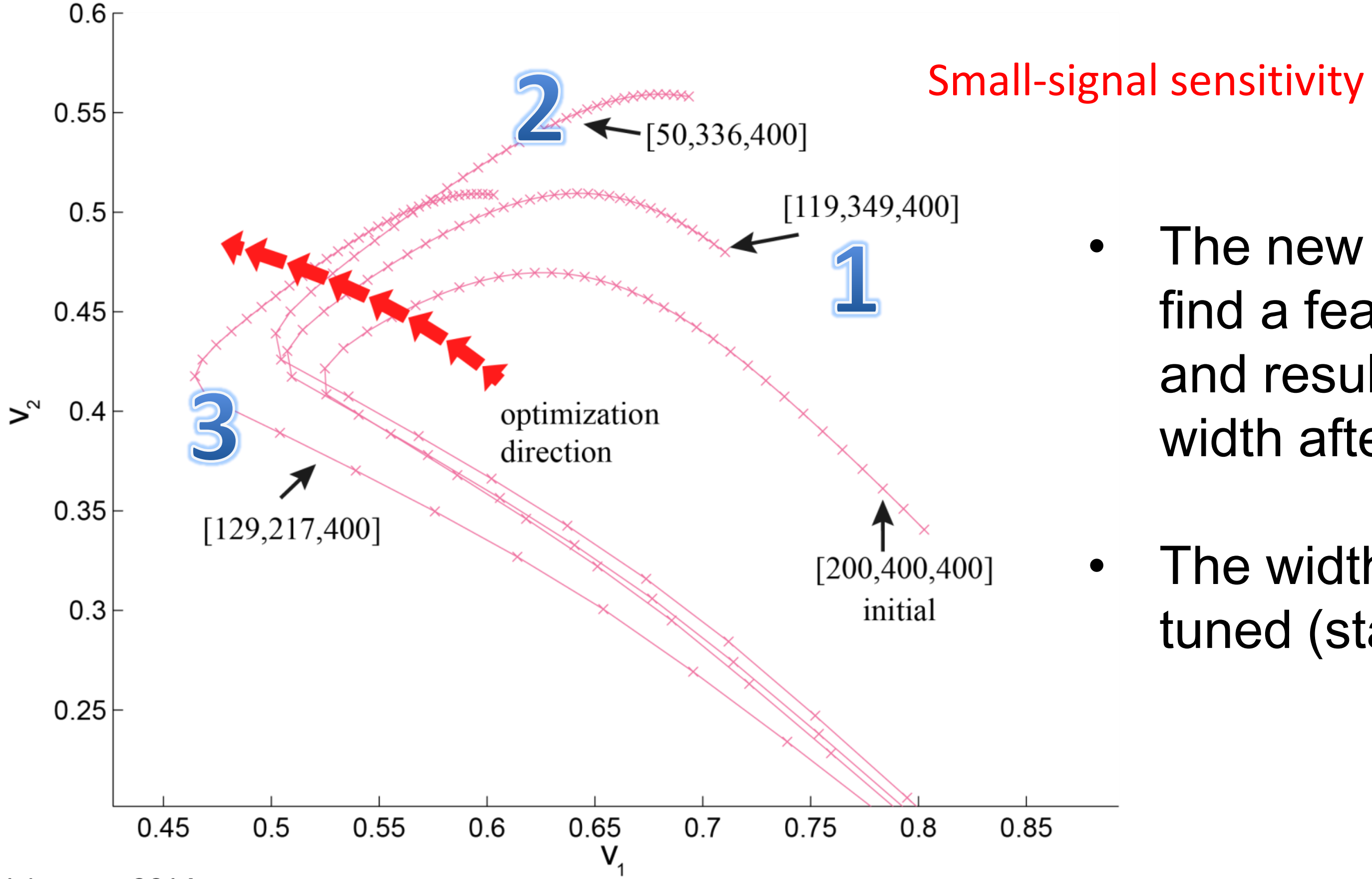
- Yield rate is improved from 6.8% to 99.957%.



- Verify yield rate before and after optimization by Monte Carlo with 1000 samples.

Optimization with Small-signal Sensitivity

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



- The new routine fails to find a feasible solution and results in negative width after 3 iterations.
- The width M5 fails to be tuned (stays at 400nm).

Runtime Comparison

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

Summary

- **Formulated SRAM dynamic stability verification and optimization problem**
- **Proposed large-signal sensitivity of safety distance in the state space by reachability analysis**
- **Significantly improved SRAM yield rate in presence of parameter variations**

Thank You!

Please send comments to haoyu@ntu.edu.sg

http://www.ntucmosetgp.net

NANYANG TECHNOLOGICAL UNIVERSITY