Title
Nanoscale SRAM Variability and Optimization

Author
Toh, Seng Oon

Publication Date
2011

Peer reviewed|Thesis/dissertation

Nanoscale SRAM Variability and Optimization

by

Seng Oon Toh

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Borivoje Nikolić, Chair
Professor Tsu-Jae King Liu
Professor Paul Wright

Fall 2011

Nanoscale SRAM Variability and Optimization

Abstract

Nanoscale SRAM Variability and Optimization

by

Seng Oon Toh

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Borivoje Nikolić, Chair

Robust SRAM design is one of the key challenges of process technology scaling. The steady pace of process technology scaling allows doubling memory array sizes, which requires thorough characterization of the impact of sources of process variability on SRAM operation. SRAM arrays are traditionally designed using static noise margins which are known to be optimistic in writeability and pessimistic in read stability. This work presents techniques for characterizing SRAM using dynamic stability metrics, which better represent actual SRAM operating conditions. Quantitative relationships between static and dynamic stability metrics are explored through statistical circuit simulations. Nano-scale SRAM design is traditionally complicated by sources of variability related to physical variability in the structure of the transistors, such as random dopant distribution. This work identifies temporal sources of variability in transistor intrinsic parameters, caused by random telegraph signaling (RTS) noise, which is directly correlated with fluctuation in SRAM performance. A large-scale dynamic stability characterization architecture is introduced and implemented in an early commercial low-power 45 nm CMOS process. This is used to experimentally verify the expected correlations between static and dynamic stability metrics. Outliers of up to 100X which are not correlated between static and dynamic stability metrics were observed and identified to be due to enhanced sensitivity of dynamic stability metrics to variability. Measurement techniques for characterizing temporal sources of variability caused by RTS noise, with particular emphasis on the large-signal bias change response typically encountered in SRAM operation, are used to collect large-scale statistics and to estimate the statistical impact of RTS noise on large SRAM arrays.

An importance sampling algorithm adapted for dynamic stability metrics is developed in this work. This algorithm is used to estimate improvements in SRAM robustness expected from new process technology options such as FDSOI, different bitcell designs such as the 8T-SRAM, as well as several read-assist and write-assist techniques. An optimization framework enabled by this importance sampling algorithm is used to design SRAM arrays with maximum array efficiency through joint-optimization between process technology, bitcell

design, and array organization. In conclusion, this dissertation identifies important sources of variability in nano-scale SRAM and also introduces the relevant optimization tools for performing variability-aware SRAM design.

To my wife, Jessica.

Contents

List of Figures **v**

List of Tables **xii**

1 Static and Dynamic SRAM Stability **1**
1.1 Introduction . . . 1
1.2 Background . . . 2
1.2.1 SRAM Margins . . . 3
1.2.2 SRAM Array . . . 4
1.3 Read Access Metrics . . . 5
1.3.1 Static Read Current (I_{read}) . . . 5
1.3.2 Read Access Time (T_{access}) . . . 6
1.4 Read Stability Metrics . . . 6
1.4.1 Static Read Stability Margins . . . 6
1.4.2 Critical Read Stability (T_{read}) . . . 7
1.5 Writeability Metrics . . . 7
1.5.1 Static Writeability Margins . . . 7
1.5.2 Critical Writeability (T_{write}) . . . 8

2 Dynamic Stability Characterization **11**
2.1 Dynamic Stability Characterization Architecture . . . 12
2.1.1 Programmable Pulse Generator . . . 12
2.1.2 Word-line Sampler . . . 14
2.1.3 Non-Destructive Read-Back . . . 15
2.1.4 Built-In-Self-Test . . . 16
2.2 45 nm CMOS Implementation . . . 16
2.3 Measurement Results . . . 17
2.3.1 Pulse Generator . . . 18
2.3.2 Read Access Time . . . 18
2.3.3 Critical Writeability . . . 20
2.3.4 Critical Read Stability . . . 21

2.4 Impact of Assist Techniques . . . 23
2.4.1 Write Assist . . . 23
2.4.2 Read Assist . . . 24

3 Random Telegraph Signal and SRAM Variability 25
3.1 Introduction . . . 25
3.2 Dynamics of Random Telegraph Signal . . . 26
3.2.1 Background . . . 26
3.2.2 Bias and Temperature Dependence . . . 30
3.2.3 Dynamic Behavior Response under Large-Bias Changes . . . 35
3.2.4 Alternating-Bias Technique . . . 39
3.3 Amplitudes of Random Telegraph Signaling Noise . . . 41
3.3.1 Background . . . 41
3.3.2 Empirical Bias-Dependent Model . . . 46
3.4 Impact of Random Telegraph Signal on Static SRAM margins . . . 48
3.4.1 45 nm SRAM Transistors . . . 50
3.4.2 Static Write Margin . . . 51
3.5 Impact of Random Telegraph Signal on SRAM Dynamic Stability . . . 57
3.5.1 RTS in pull-down NMOS transistor 1 . . . 61
3.5.2 RTS in pull-up PMOS transistor 2 . . . 62
3.5.3 RTS in pull-up PMOS transistor 1 . . . 63
3.5.4 RTS in pass-gate NMOS transistors . . . 64
3.5.5 Statistical Distributions and Implications to SRAM Robustness . . . 65
3.6 Correlation Between RTS and NBTI . . . 68

4 Statistical Estimation of SRAM Dynamic Stability 70
4.1 Introduction . . . 70
4.2 Importance Sampling and Most Probable Failure Point Search . . . 72
4.3 Importance Sampling and Dynamic Stability . . . 77
4.3.1 Dynamic Read Stability . . . 77
4.3.2 Dynamic Writeability . . . 80
4.3.3 Dynamic Read Access . . . 83
4.4 Importance Sampling with non-Gaussian Distributions . . . 85
4.4.1 Lognormal Distribution . . . 85
4.4.2 Random Telegraph Signal and Dynamic Read Stability . . . 86

5 Stochastic Optimization of SRAM Bitcell and Array 92
5.1 Introduction . . . 92
5.2 Bitcell Optimization . . . 92
5.2.1 Global Process Variation . . . 92
5.2.2 V_{DD} Scaling . . . 94

5.2.3 Bitcell Variety 95
5.2.4 8T SRAM 96
5.2.5 Process Technology 99
5.3 Array Optimization 101
5.3.1 Read and Write Assist 102
5.3.2 Bit-line Capacitance 103
5.3.3 Array Segmentation 104

6 Conclusion **111**
6.1 Key Contributions 111
6.2 Future Work 113

Bibliography **115**

List of Figures

1.1 SRAM array VDD reported in ISSCC and VLSI (2004-2010) and ITRS predictions. . . . 2
1.2 Schematic of an SRAM bitcell. . . . 2
1.3 (a) Schematic of bitcell for SRAM margin measurement. (b) Static noise margin (SNM) extraction. (c) Write margin (I_{write}) extraction. . . . 3
1.4 Schematic of an $m \times n$ SRAM array. . . . 4
1.5 (a) Schematic of a 6-T SRAM cell storing a 0 on the left internal node. (b) Simulated waveforms corresponding to failed read access with pulse-width, T_A. Output of the sense-amplifier ($Data$) resolves to the incorrect value. (c) Simulated waveforms corresponding to successful read access with a longer pulse-width, T_B. Output of the sense-amplifier ($Data$) resolves to the correct value. . . . 6
1.6 (a) Schematic of a 6-T SRAM cell under read stress. (b) Simulated waveforms corresponding to read stable access with pulse-width, T_A. The state of the bitcell is retained after read operation. (c) Simulated waveforms corresponding to read upset with a longer pulse-width, T_B. The state of the bitcell is accidently flipped by the read operation. . . . 8
1.7 Simulated scatter plot showing the correlation between critical read stability (T_{read}) and negative static read margin (SRRV). . . . 8
1.8 Simulated waveforms corresponding to an SRAM bitcell under read-after-read access. . . . 9
1.9 (a) Schematic of a 6-T SRAM cell under write access. (b) Simulated waveforms corresponding to failed write access with pulse-width, T_A. The bitcell retains original state. (c) Simulated waveforms corresponding to successful write access with a longer pulse-width, T_B. A new value is written into the bitcell. . . . 10
1.10 Simulated scatter plot comparing critical writeability (T_{write}) and static write margin (BWTV) obtained from Monte-Carlo simulations. . . . 10

2.1 SRAM array organization for static and dynamic stability characterization. . 12
2.2 Frequency mixing programmable pulse generator with corresponding waveforms. 13

2.3 Statistics at the output of a flip-flop sub-sampling ϕ_1 using ϕ_0 with jitter. . . 13
2.4 Architecture of averaging *sync* generator. 14
2.5 Capacitive summing phase comparator and simulated waveforms before and after calibration. 15
2.6 Built-In-Self-Test state machine for finding the critical word-line pulse width. 16
2.7 (a) Die photo of the 45 nm CMOS test chip. (b) Die photo of active area. BLS : bitline switches; WLS : word-line samplers; LS+WLD : level shifters and word-line drivers; CIO : column I/O circuitry 17
2.8 SRAM writeability and read stability fail bit count measured from a 45nm CMOS SRAM. 18
2.9 Plots of (a) multiple sub-sampled word-line waveforms and (b) codeword to pulse width transfer function and measured error. 19
2.10 (a) Histogram of measured read access time. (b) Scatter plot showing correlation between read access time and static read current after normalization with sense-amplifier offset voltage and bit-line capacitance. 19
2.11 Critical writeability vs. static write margin of (a) same side and (b) opposite side of SRAM cell measured at $V_{DD,low}$. 20
2.12 Critical read stability versus static read margin. 21
2.13 Statistical distributions of critical read stability under single read and read-after-read access with 20 ns clock period. 22
2.14 Critical read stability of a selected bitcell as a function of the number of read-after-read cycles. The different curves correspond to the period of the read-after-read cycles. 22
2.15 Survival function of T_{write} under different bias conditions: (a) Word-line boosting; (b) V_{CELL} under-drive; (c) PMOS reverse body-bias; (d) Negative bit-line. 23
2.16 Survival function of T_{read} without assist techniques, with -100 mV word-line bias offset and with +100 mV V_{CELL} offset. 24

3.1 Drain current measured from a PMOS SRAM transistor showing RTS. . . . 27
3.2 Exponential fit of extracted time constants corresponding to (a) time until capture (τ_c) and (b) time until emission (τ_e). 27
3.3 Power spectral density of RTS drain current measurement. 28
3.4 Energy band diagram of oxide trap with tunneling probabilities. 28
3.5 Drain current of PMOS SRAM transistor under different gate biases. 30
3.6 Dependence of τ_c and τ_e on gate bias. 31
3.7 Drain current of PMOS SRAM transistor at different temperatures. 31
3.8 Measured mark-space ratio at different gate bias and temperatures with extrapolation. Lines indicate exponential fits. 32
3.9 Surface contour of average trap occupancy (α) as a function of gate bias and temperature, extrapolated using Equation 3.2 with $K = 26$ and $\frac{z}{T_{ox}} = 0.35$. . 32

3.10 (a) V_{GS} dependence of τ_c (filled circles) and τ_e (open squares) and (b) mark-space ratio of a type II trap. Reference [53] 33
3.11 Band diagram of type II oxide trap with tunneling probabilities. 34
3.12 Drain current of PMOS SRAM transistor, with type III, under different gate biases. 35
3.13 Type III trap dependence of τ_c and τ_e on gate bias. 35
3.14 Cross-section of transistor with oxide trap as well as associated waveforms for characterizing the response of RTS to large-bias changes. 36
3.15 (a) Single trace and average response of a trap to large-bias change (V_{init} = 1.0V, $V_{measure}$ = 0.7V) (b) Large-bias response of trap at different temperatures. 37
3.16 (a) Single trace and average response of a trap to large-bias change (V_{init} = 0.0V, $V_{measure}$ = 0.7V) (b) Large-bias response of trap at different temperatures. 37
3.17 (a) Average response of a trap to large-bias change with different values of T_{init} (V_{init} = 0.0V, $V_{measure}$ = 0.7V) (b) Estimated average response of the trap to large-bias change with V_{init} = 0.7V and $V_{measure}$ = 0.0V. 38
3.18 (a) Single trace and average response of a type-II trap to large-bias change (V_{init} = 1.0V, $V_{measure}$ = 0.7V) (b) (V_{init} = 0.0V, $V_{measure}$ = 0.7V) 39
3.19 Measured drain current of SRAM PU PMOS transistor demonstrating RTS with long time constant. 40
3.20 Waveforms of the alternating-bias technique. 40
3.21 Drain current of the same transistor measured using conventional and alternating-bias techniques. 41
3.22 Gumbel plots of RTS drain current fluctuations measured using conventional and alternating-bias techniques with constant time. 42
3.23 Histogram and lognormal distribution fits of RTS I_{ds} fluctuations measured using conventional and alternating-bias techniques with constant time. . . . 42
3.24 Bias dependence of RTS amplitude extracted from 45 nm SRAM (a) NMOS and (b) PMOS transistors. Solid lines correspond to number fluctuation model fitted to data at V_{DS} = 0.1V. 44
3.25 45 nm NMOS RTS amplitude fitted to correlated model. (a) Extracted scattering co-efficients as a function of electron density and fitted models. (b) Measured RTS amplitude across different biases and fitted models. Solid lines and dashed lines correspond to $1/\sqrt{N}$ and $log(N)$ models for scattering co-efficient, respectively. 46
3.26 V_{DS} dependence of ΔV_{th} for a trap measured from PMOS and NMOS transistors at constant gate bias ($|V_{GS}|$ = 0.8 V). 47
3.27 Residual ΔV_{th} after normalization by linear fit of V_{DS} dependence, as a function of electron density. 47
3.28 Equivalent ΔV_{th} fluctuation due to RTS across gate and drain biases extracted from (a) NMOS and (b) PMOS SRAM transistors. Solid lines represent the empirical model fitted to the data. 49

3.29 RTS amplitude for (a) NMOS and (b) PMOS transistors estimate using SPICE models extended with empirical RTS amplitude model. 49
3.30 Gumbel plots of normalized RTS fluctuations in drain currents, measured from transistors in the padded-out SRAM cells. 50
3.31 Schematic of I_{write} measurement. The N-curve is the sum of the currents flowing out of the internal node. 52
3.32 Currents contributing to I_{write}. 52
3.33 Currents and voltages measured from padded-out SRAM cell with RTS. RTS characteristics of the N-curve are influenced by RTS in both PU5 and PG3. Voltage fluctuation at node CL is minimal due to the low impedance of this node. 53
3.34 N-curves measured from multiple sweeps demonstrating the technique for extracting (a) nominal I_{write} and (b) RTS fluctuation in I_{write} due to RTS. . 54
3.35 (a) Histogram of nominal I_{write}. (b) Gumbel plots of I_{write} RTS fluctuation. The Maximum RTS amplitude normalized to σI_{write} at each operating voltage does not change significantly, although the shape of the distribution changes. 54
3.36 (a) Quantile-quantile plot of I_{write} with normal distribution fit. (b) Quantile-quantile plot of RTS fluctuation in I_{write} with hybrid distribution fit. 55
3.37 Scatter plots of measured nominal I_{write} and RTS fluctuation at nominal and low operating voltages. 55
3.38 Joint probability density function of nominal I_{write} and RTS fluctuation at nominal V_{dd}. 56
3.39 Joint probability density function of nominal I_{write} and RTS fluctuation at low V_{dd}. 56
3.40 FBR of SRAM at different voltages. 58
3.41 Measured fail bit count of a 64kb SRAM array. 58
3.42 (a) High-speed N-curve sweeps of SRAM write margin at two temperatures. (b) Zoomed-in view of region defining static SRAM write margin. 59
3.43 SRAM access patterns for evaluating the impact of RTS on dynamic read and write ability. 60
3.44 (a) SRAM schematic for read access with internal node CL storing a "0". (b) SRAM schematic for writing a "0" into node CH. 60
3.45 (a) Large-bias change response occupancy of a trap in $PD1$. (b) Statistical distributions of T_{access} for single-read and read-after-write. 61
3.46 Dependence of T_{access} fluctuation on delay since last write access (T_{relax}). . . 62
3.47 (a) Large-bias change occupancy of a type-II trap in $PD1$. (b) Statistical distributions of T_{access} for single-read and read-after-write. 62
3.48 (a) Large-bias change occupancy of a trap in $PU2$. (b) Statistical distributions of T_{write} for single-write and write-after-write. 63
3.49 (a) Large-bias change occupancy of a type-II trap in $PU2$. (b) Statistical distributions of T_{write} for single-write and write-after-write. 63

3.50 (a) Large-bias change occupancy of a trap in $PU1$. (b) Statistical distributions of T_{write} for single-write and write-after-write. 64
3.51 Simulated trap occupancy in $PG1$ under successive read-after-read access. (a) Zoomed-in view (b) Expanded view 65
3.52 (a) Large-bias change occupancy of a trap in $PG1$. (b) Statistical distributions of T_{access} with N read-after-read cycles, saturting at N=128. 65
3.53 Scatter plot of ΔT_{access} due to RTS vs. nominal T_{access}. 67
3.54 Scatter plot of ΔT_{write} due to RTS vs. nominal T_{write}. 67
3.55 (a) Measured V_{th} degradation due to NBTI. (b) RTS observed in the same transistor at 25^{o}C super-imposed on top of V_{th} degradation observed during NBTI characterization at 125^{o}C. 69

4.1 Schematics of SRAM for: (a) Read static noise margin analysis. (b) Write static noise margin analysis. 73
4.2 Sensitivity of read static noise margin (RSNM) to V_{th} variability centered at: (a) $\boldsymbol{x} = \mathbf{0}$. (b) $\boldsymbol{x} = MPFP$. 73
4.3 Sensitivity of bit-line write margin (BLWM) to V_{th} variability centered at: (a) $\boldsymbol{x} = \mathbf{0}$. (b) $\boldsymbol{x} = MPFP$. 74
4.4 MPFP search with locally evaluated gradients. 75
4.5 RSNM histogram of 5000 samples generated from original Gaussian distributions shifted by the MPFP vector. 76
4.6 Evolution of (a) fail bit rate ($\hat{p_{IS}}$) and (b) convergence metric (ρ) as a function of run number. 76
4.7 General algorithm for estimating statistical dynamic read stability reliability of SRAM. 77
4.8 Sensitivity of dynamic read stability (T_{read}) to V_{th} variability. 78
4.9 Evolution of (a) fail bit rate ($\hat{p_{IS}}$) and (b) convergence metric (ρ) as a function of run number, corresponding to T_{read} importance sampling. 80
4.10 General algorithm for estimating statistical dynamic writeability reliability of SRAM. 81
4.11 Sensitivity of dynamic writeability (T_{write}) to V_{th} variability. 81
4.12 Evolution of (a) fail bit rate ($\hat{p_{IS}}$) and (b) convergence metric (ρ) as a function of run number, corresponding to T_{write} importance sampling. 83
4.13 Sensitivity of final bit-line voltage at T_{access} to V_{th} variability. 84
4.14 Evolution of (a) fail bit rate ($\hat{p_{IS}}$) and (b) convergence metric (ρ) as a function of run number, corresponding to dynamic read access importance sampling. . 85
4.15 Simulated lognormal distributions of V_{th} fluctuation due to RTS in NMOS and PMOS transistors. 88
4.16 Conceptual plot demonstrating shift in dependencies between cell sigma and V_{DD} with worst-case RTS combination, best-case RTS combination, and no RTS. 88

4.17 Evolution of (a) fail bit rate ($\hat{p_{IS}}$) and (b) convergence metric (ρ) as a function of run number, corresponding to T_{read} with lognormal RTS distributions evaluated using importance sampling and conventional Monte Carlo simulations. 90
4.18 Fail bit rate corresponding to T_{read} as a function of V_{DD} without RTS and with worst-case RTS. $T_{read} = 1$ ns. 90

5.1 NMOS and PMOS global process variation space annotated with failure contours for read and write operation, as well as process corners [86]. 93
5.2 Dependence of fail bit rate on V_{DD} for static and dynamic writeability, read stability, and dynamic read access. 95
5.3 Fail bit rate as a function of V_{DD} for a high performance and high density bitcell from a particular process technology. The flattening of fail bit rate at higher V_{DD} is due to the weaker sensitivity of dynamic read stability to V_{DD} compared to other failure modes. 96
5.4 Schematic of 8T SRAM bitcell. 98
5.5 Fail bit rate of 8T and 6T bitcells as a function of V_{DD}. 98
5.6 Dynamic stability of 8T bitcell with and without V_{WWL} boost. 99
5.7 Dynamic access (1 ns pulse) fail bit rate of 8T bitcell with and without V_{RWL} boost. 100
5.8 Comparison of fail bit rate for a similar SRAM bitcell implemented either in bulk CMOS or FDSOI, demonstrating more than 200 mV reduction in V_{min} 101
5.9 Impact of word-line voltage offset on SRAM stability across different process corners. 102
5.10 (a) Fail bit rate degradation with increasing bit-line capacitance. (b) Fail bit rate degradation as a function of the number of rows in a bit-line with worst-case fail bit rate for dynamic read access and dynamic read stability combined into a single curve. 104
5.11 64 kb SRAM array segmented into (a) 256 rows x 256 columns, and (b) 64 rows x 1024 columns. 105
5.12 (a) Dependence of SRAM reliability (Z-value) corresponding to dynamic writeability as a function of V_{DD} across different access pulse-widths (ns), and (b) contour plot of SRAM reliability corresponding to dynamic writeability. 106
5.13 SRAM reliability corresponding to dynamic read stability as a function of V_{DD} across different bit-line capacitance (1 ns pulse-width). 107
5.14 (a) Dependence of SRAM reliability (Z-value) corresponding to dynamic read access as a function of V_{DD} across different (a) bit-line capacitance, (b) sense-amplifier offset voltages, and (c) access pulse-widths. 107
5.15 (a) Minimum cache area as a function of different cache sizes compiled using two different bitcells, (b) column height corresponding to the minimum-area solution for each cache size. 109

5.16 Optimum area of a 1MB cache memory optimized at different operating V_{DD} conditions. The corresponding optimum column height (# rows) is also plotted on the same graph. 110

List of Tables

1.1 Sensitivity analysis of writeability to the respective transistor V_{th} variation. . 10

4.1 Progress of MPFP search algorithm applied to RSNM 75
4.2 Progress of feasible T_{read} search algorithm. $T_{read,target} = 1$ ns 79
4.3 Progress of T_{read} MPFP search algorithm. $T_{read,target} = 1$ ns 80
4.4 Progress of T_{write} MPFP search algorithm. $T_{write,target} = 1$ ns 82
4.5 Progress of T_{access} MPFP search algorithm. $\Delta V_{BL,target} = 0.1$ V 84
4.6 Progress of T_{read} MPFP search algorithm with lognormal RTS distributions in $PU1$ and $PU2$. $T_{read,target} = 1.0$ ns 89

Acknowledgments

First and foremost, I would like thank my advisor, Professor Borivoje Nikolić for guiding me on the various technical aspects related to this dissertation. It was from his insight into pressing issues impacting the scalability of SRAM, that this thesis was born, and it was due to his comprehensive knowledge of various disciplines that this work was finally completed. I am in great appreciation of his patience in working with me over countless iterations of drafts and presentations to ensure impeccable delivery of the technical content. I would also like to thank him for other kind words of advice related to life, family, and my career.

I would also like to acknowledge Professor Tsu-Jae King Liu who served as chair of my qualifying exam committee and participated as a member of my dissertation committee. She also played an active role as a co-advisor on various research projects. I would also like to thank her students, Dr. Changhwan Shin, Nattapol Damrongplasit, and Min Hee Cho for their help in testchip tapeout and many discussions on device physics.

I am also grateful to Professor Paul Wright for participating as a member of my qualifying exam committee and dissertation committee and helping me place my research within the context of interest to society. I am also thankful to Professors Andy Neureuther and Costas Spanos for their feedback provided during countless variability group meetings. I would also like to thank one of Professor Neureuther's students, Dr. Lynn Wang, for her collaboration on interpreting variability results within the context of lithography. I am also indebted to Professor Martin Wainwright who provided advice on importance sampling and optimization. I would also like to thank Professor Jan Rabaey for being a member of my dissertation committee and his valuable advice earlier on in my Ph. D. career.

My research has been supported by the Center for Circuit & System Solutions (C2S2) Focus Center. The final year of my research was particularly made possible through IBM's generous Ph. D. fellowship. Chip fabrication was donated by STMicroelectronics.

I am appreciative of the help and support from the staff at the EECS graduate office and the Berkeley Wireless Research Center (BWRC). In particular, Ruth Gjerde, Tom Boot, Brenda Farrell, Pierce Chua, and Olivia Nolan for their administrative support; Kevin Zimmerman and Brian Richards for maintaining a top notch computing environment equipped with up-to-date tools and design kits; Susan Mellers and Chang Chun for maintaining the BWRC lab and co-ordinating availability of lab equipment; and Gary Kelson for directing the BWRC. I would also like to acknowledge various students of the department - Dr. Zheng Guo, Dr. Liang-Teck Pang, and Dr. Andrew Carlson for passing down to me a rich set of circuits, scripts, and techniques for variability characterization and overseeing my research; Dr. Bastien Giraud for his help on testchip debugging; Lauren Jones and Jason Tsai for their help on a testchip tapeout; Patrick Bennett for de-processing the testchips for failure analysis; Sriramkumar Venugopalan for helping me on RTS noise amplitude modeling; Katerina Papadopoulou and Brian Zimmer for their assistance on FDSOI testchip tapeouts; and Andrew Mairena for performing measurements.

This thesis would not have been possible and would lack relevance to the industry if not

for the mentorship of a few industrial domain experts. I do not have enough words to thank Dr. Yasumasa Tsukamoto for bringing me up to speed on state of the art SRAM statistical analysis, introducing me to new sources of SRAM variability, as well as being a friend. I would also like to thank Dr. Stephen Kosonocky, Dr. Olivier Thomas, and Dr. Sani Nassiff for providing constant feedback on my research.

Finally, I would like to thank my wife and son - Jessica and Ethan Toh - for standing by me and graciously giving me time to complete this work. I am also grateful to my parents - Professor Seong Chong Toh and How Chan Wong - for supporting me through this long journey.

Chapter 1

Static and Dynamic SRAM Stability

1.1 Introduction

SRAM scaling has been identified as one of the bottlenecks for supply voltage (V_{DD}) reduction in current and future technology nodes. Minimum SRAM operating voltage (V_{MIN}) is a function of the magnitude of process-induced variability as well as the array size. Aggressive SRAM bitcell scaling, coupled with continued increase in SRAM array sizes, has resulted in stagnation in SRAM V_{DD} scaling. This trend is observed in reported values of SRAM array V_{DD} and is recognized in the latest edition of the International Technology Roadmap for Semiconductors (ITRS) (Figure 1.1) [30]. V_{MIN} is traditionally estimated using various static noise margins (SNM) [82, 61]. These metrics are known to be optimistic in writeability and pessimistic in read stability primarily due to the fact that SRAM access is a dynamic operation and occurs within a finite duration versus the main assumption of infinite access time in static noise margins [87]. Dynamic stability metrics, derived from the SRAM under dynamic access, have been proposed to provide a better estimate of SRAM V_{MIN} [32, 37, 62]. While these metrics have been studied extensively through simulations, results based on large-scale silicon characterization of both read and write stability have only been reported in this dissertation. Similarly, a quantitative relationship between the static and dynamic read and write margins is studied here. The sensitivity of dynamic stability to non-idealities such as random telegraph signaling (RTS) noise and aging is also studied in this work.

This chapter first reviews conventional static and dynamic 6 transistor SRAM metrics as well as their expected correlations. Monte-Carlo simulations, introducing Gaussian distributions of V_{th} to the 6 SRAM transistors, are presented in this section to illustrate expected correlations between the metrics.

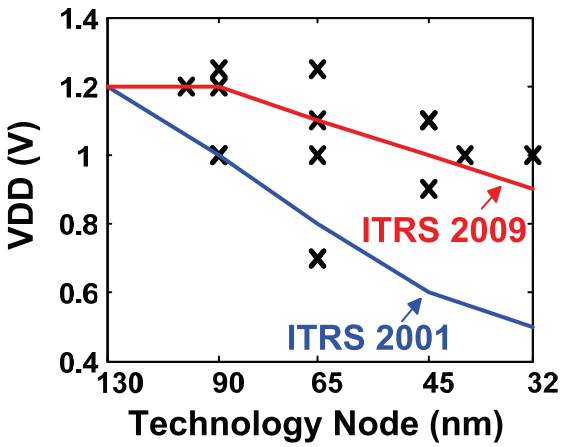


Figure 1.1: SRAM array VDD reported in ISSCC and VLSI (2004-2010) and ITRS predictions.

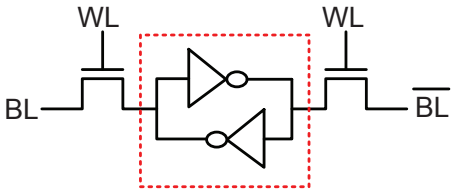


Figure 1.2: Schematic of an SRAM bitcell.

1.2 Background

Memory represents one of the fundamental building blocks of any computation system as it is required to retrieve variables for the computation as well as to store results of the calculation. Memory can be divided into two classes: volatile and non-volatile. Volatile memory loses the stored information when the power supply is removed while non-volatile memory typically retains the data for a few years. Static Random Access Memory (SRAM), which is the main subject of this work, is classified as volatile memory because it relies on a power supply to statically retain data. It is further classified as a random access memory because it allows access to arbitrary locations of data without any restrictions.

Each bit of data within an SRAM is stored in SRAM bitcells. Figure 1.2 illustrates a schematic representation of a bitcell circuit. The core of the SRAM bitcell (enclosed by dotted lines) consists of cross-coupled inverting gain elements and is used to store the state of the bitcell. Assuming that one node in the core is storing an electrical representation of a logic-0. This node drives the input of an inverting gain element to logic-1 which then causes the next inverting gain element to drive the original node back to a logic-0. This

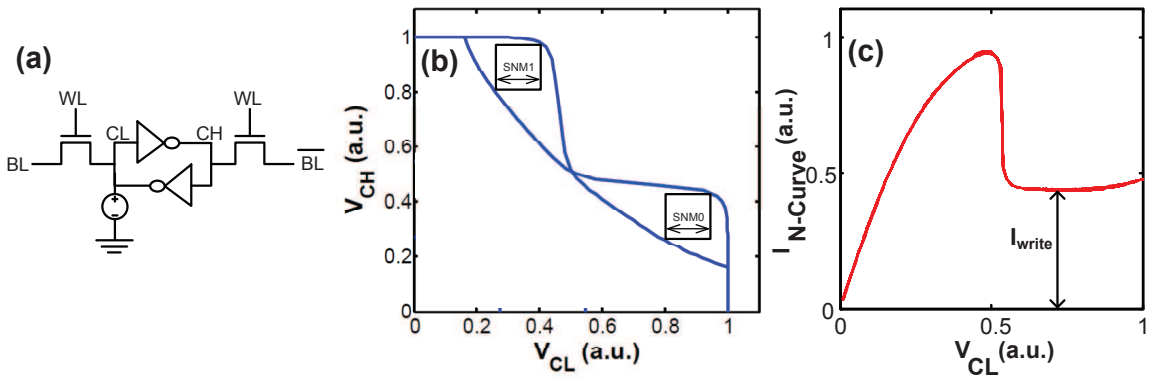


Figure 1.3: (a) Schematic of bitcell for SRAM margin measurement. (b) Static noise margin (SNM) extraction. (c) Write margin (I_{write}) extraction.

positive feedback configuration increases robustness of the circuit to noise. The bitcell core is accessed using a pair of access transistors that are used to read the contents of the bitcell or to write new values. These transistors connect the core of the bitcell to a pair of bit-lines (BL and $\overline{BL}$) and are enabled by pulsing on the word-line signal (WL). These access transistors are designed to be strong enough to overcome the positive feedback in the bitcell core and overwrite the contents of the bitcell. On the other hand, they also need to be weak enough that the content of the bitcell is not accidentally disturbed during a read operation. Herein lies one of the fundamental constraints in SRAM bitcell design - the tradeoff between read and write margin. The cross-coupled inverting gain elements in the core are usually implemented as two static CMOS inverters with each static CMOS inverter consisting of a PMOS and NMOS transistor. This SRAM bitcell design is usually called a 6T bitcell due to the fact that there are 6 transistors in a bitcell.

1.2.1 SRAM Margins

SRAM margins are used to quantify the robustness of a read and write operation. 1.3 (a) illustrates the schematic of a bitcell set up for SRAM static margin extraction. A voltage source is connected to one of the internal nodes (CL). This node voltage is swept while measuring voltages at other nodes or the current flowing out of this voltage source ($I_{N-Curve}$). Figure 1.3 (b) plots the technique for characterizing the static noise margin (SNM) of a bitcell [61]. Both bit-lines and the word-line (BL, $\overline{BL}$, WL) are connected to V_{DD}. The first curve is measured by sweeping the voltage of node CL while monitoring the voltage at node CH. This essentially traces the switching characteristic of one of the cross-coupled inverters. The other curve is measured by sweeping node CH while monitoring the voltage at node CL. The static noise margin for storing a 0 or 1 ($SNM0$ and $SNM1$) corresponds to the side of the largest square that fits into the respective lobes. This static

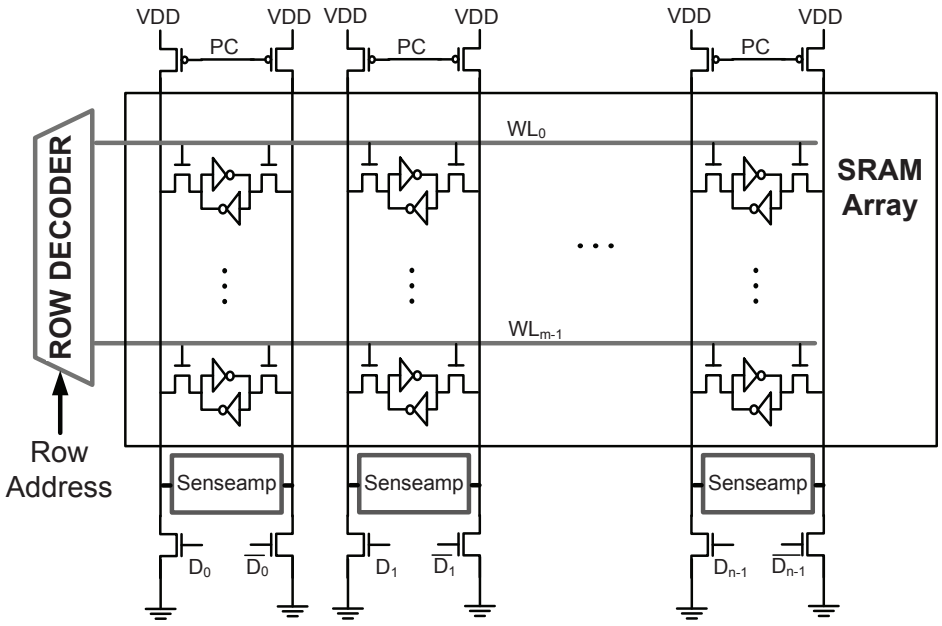


Figure 1.4: Schematic of an $m \times n$ SRAM array.

margin essentially characterizes the largest voltage perturbation that can be sustained in the internal nodes of the bitcell before the bitcell looses the ability to store two states. The bias conditions applied on the bitcell for write margin characterization depend on the data that is being written into the bitcell. To write a logic-0 into node CL, node BL is grounded while nodes $\overline{BL}$ and WL are connected to V_{DD}. The voltage at node CL is then swept from V_{DD} to ground while the current flowing through the voltage source ($I_{N-Curve}$) is monitored. This emulates the bias conditions applied to the transistors as a logic-0 is written to node CL. $I_{N-Curve}$ observed during this voltage sweep is plotted on Figure 1.3 (c). The static write margin(I_{write}) is defined as the minimum current observed at the right side of the plot [13]. The I_{write} write margin is preferred over an alternative write noise margin proposed by Seevinck *et al.* because of easier extraction from experimental results and better correlation with SRAM write V_{min} across a wide range of voltages [24].

1.2.2 SRAM Array

Figure 1.4 illustrates an array of SRAM bitcells representing an organization typically encountered in SRAM array implementations. The bitcells are organized into m rows and n columns. Additional circuits are located at the periphery of the array. This peripheral circuitry is used to support random read and write access to the data stored in the SRAM bitcells. All bitcells in a column share a common pair of bit-lines and column peripheral circuitry while all bitcells in a row share a common word-line, driven by the row decoder.

The pre-charge headers at the top of the array are used to pre-charge the bit-lines up to V_{DD} prior to a read or write operation. During a read operation, the row decoder enables one of the word-lines, depending on the row address that is selected. The enabled word-line turns on the access transistors which connects the bitcell core to the bit-lines and transfers the contents of the bitcell onto the bit-lines. A sense-amplifier is then used to read the values stored on the bit-lines and sent to the output of the array. During a write operation, one of the bit-lines in a pair is selectively dis-charged using the transistors at the bottom of the array, depending on the new data that is to be written into the bitcells. The word-line corresponding to the desired row where new data is to be stored is then enabled. This transfers the state of the bit-lines into the core of the bitcells.

The amount of capacitance on the bit-lines plays a critical role in SRAM performance and stability. The bit-line capacitance is dominated mainly by diffusion capacitance of all the access transistors sharing the same bit-line. Coupling capacitance of the bit-line to neighboring wires also plays a role. From an array efficiency perspective, it is highly desirable to have long bit-lines tied to many bitcells, sharing a single sense-amplifier. This however results in large bit-line capacitance which reduces SRAM read performance due to the extra time required to discharge the higher capacitance. Energy efficiency of the array is also degraded as this large capacitance needs to be charged and dis-charged frequently. More recently, hierarchical bit-lines have been introduced to minimize bit-line capacitance. These array organization involves segmenting the bit-lines into many local bit-lines with a small number of bitcells sharing the bit-line and logically combining the results from all the bit-lines. Although this requires more area overhead, this scheme could potentially allow the use of smaller less robust bitcells with weaker read current, ultimately resulting in a smaller SRAM array with lower energy consumption.

1.3 Read Access Metrics

1.3.1 Static Read Current (I_{read})

Static read current (I_{read}) corresponds to the current that is being sourced from the bit-line into the SRAM node storing a 0. Under SRAM read operation, this current is responsible for discharging the pre-charged bit-line capacitances (C_{BL}) enough to overcome the offset voltage (V_{offset}) of the sense-amplifier to result in a correct value being latched. It is expected to correlate with actual read access time (T_{access}) according to Equation 1.1

$$T_{access} \propto \frac{C_{BL} \times V_{offset}}{I_{read}} \tag{1.1}$$

Actual read access time might deviate from this linear relationship due to leakage currents from inactive bitcells sharing the bit-line as well as the fact that C_{BL} is a distributed RC network spanning the entire column of the SRAM array. Degradation in I_{read} and read access

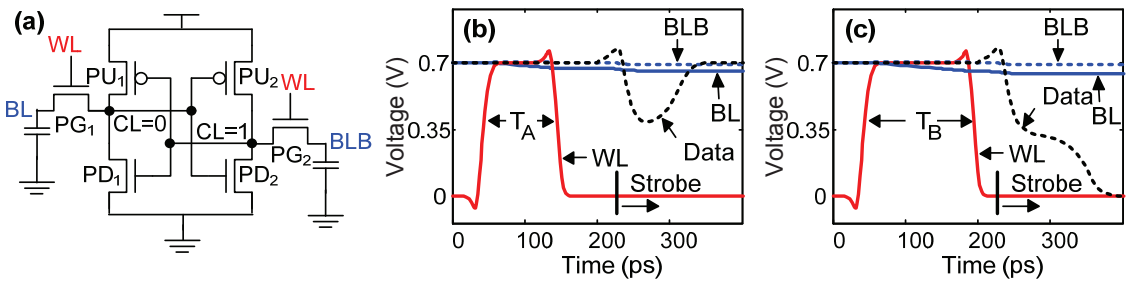


Figure 1.5: (a) Schematic of a 6-T SRAM cell storing a 0 on the left internal node. (b) Simulated waveforms corresponding to failed read access with pulse-width, T_A. Output of the sense-amplifier ($Data$) resolves to the incorrect value. (c) Simulated waveforms corresponding to successful read access with a longer pulse-width, T_B. Output of the sense-amplifier ($Data$) resolves to the correct value.

time due to RTS also contributes to this discrepancy, as will be shown in Section 3.5.

1.3.2 Read Access Time (T_{access})

Figure 1.5 illustrates an SRAM bitcell undergoing read access with pulse-widths T_A and T_B. Pulse-width T_A is too short to sufficiently discharge the bit-line capacitance to overcome offset in the sense-amplifier. There exists a pulse-width, T_{access} ($T_A < T_{access} < T_B$), where the sense-amplifier is on the threshold of a successful read access that is defined as the read access time. This is similar to the dynamic access failure criteria defined in [37]. This definition of read access time isolates out variability in the read access operation due to variability of the SRAM bitcell and ignores other delays such as word-line driver delay and sense-amplifier delay.

1.4 Read Stability Metrics

1.4.1 Static Read Stability Margins

Conventional stability metrics, such as SNM and N-curves [82, 61], require sweeping internal nodes in order to obtain the voltage transfer curves, which is not practical for evaluating large arrays. Supply read retention voltage (SRRV), which does not require access to the internal nodes, is used in this chapter to characterize static read stability [24]. This metric characterizes read stability margins by decreasing the supply voltage of the bitcell core while monitoring the currents flowing through the bit-lines which are held at V_{DD}. A flip in data stored by the bitcell is correlated to a sharp change in the bitcell currents. The SRRV margin is defined as the additional bitcell core voltage reduction that can be tolerated before the

bitcell loses the data. A direct correlation between this and other stability metrics has already been established in [24].

1.4.2 Critical Read Stability (T_{read})

Figure 1.6 illustrates an SRAM bitcell undergoing read stress with pulse-widths T_A and T_B. Pulse-width T_A is short enough that the internal nodes (CH and CL) return back to their original levels after the word-line pulse. The longer pulse-width T_B subjects the bitcell to too much read stress, causing the cell to flip to an opposite state after the word-line pulse. There exists a pulse-width, T_{read} ($T_A < T_{read} < T_B$), where the bitcell is on the threshold of a read upset, that is defined as the critical read stability. This is similar to the dynamic read failure criteria defined in [37]. This metric does not require access to the internal nodes of the SRAM cell. A subsequent read operation (read-back) is used to verify the contents of the bitcell and to determine if a read disturb had occurred. This read-back operation however needs to be performed under operating conditions which guarantee a correct read operation. This is usually done by raising the supply voltage of the bitcell to improve read stability.

A bitcell with positive static read margin will have infinite T_{read} while a bitcell with zero or negative static read margin will have a finite value of T_{read}. With the SRRV margin, it is possible to characterize a negative static read margin for a particular bitcell by measuring how much additional bitcell V_{DD} (V_{CELL}), above the nominal voltage, is required to maintain the stored state of the bitcell. Figure 1.7 plots the positive correlation observed between SRRV and T_{read} extracted from Monte-Carlo simulations. Although T_{read} is observed to be exponentially dependent on static read margin, it is impossible to accurately estimate exact values of critical read stability from a voltage screen test at elevated V_{CELL} due to the large dispersion (up to 10x) observed in T_{read} at a particular SRRV.

SRAM access with read-after-read operation presents the worst-case condition for critical read stability [37, 62]. Figure 1.8 illustrates the waveforms corresponding to an SRAM bitcell with read-after-read access. The SRAM bitcell is stable after the first word-line pulse but is subsequently corrupted by the second pulse. It is therefore important to characterize T_{read} as a function of the number of read-after-read pulses as well as the access frequency.

1.5 Writeability Metrics

1.5.1 Static Writeability Margins

Margins such as write noise margin (WNM) and write N-curve require sweeping internal nodes in order to obtain the voltage transfer curves [13, 8]. In this chapter, bit-line write trip voltage (BWTV) is characterized instead of other static metrics because BWTV can be measured by sweeping the bit-line voltages of the SRAM bitcell while monitoring the currents

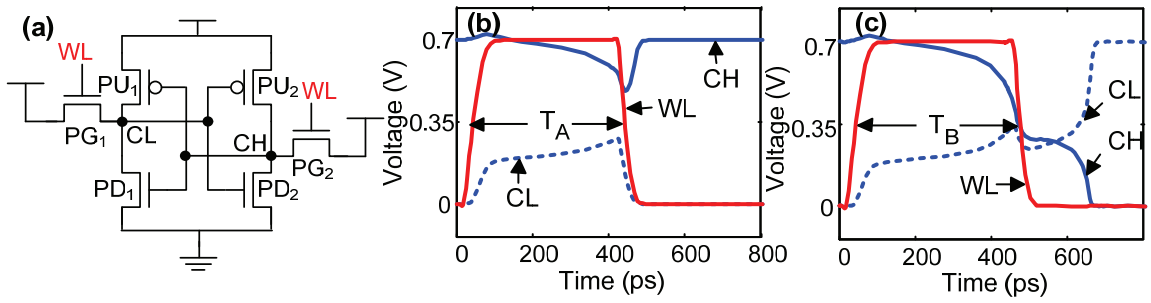


Figure 1.6: (a) Schematic of a 6-T SRAM cell under read stress. (b) Simulated waveforms corresponding to read stable access with pulse-width, T_A. The state of the bitcell is retained after read operation. (c) Simulated waveforms corresponding to read upset with a longer pulse-width, T_B. The state of the bitcell is accidently flipped by the read operation.

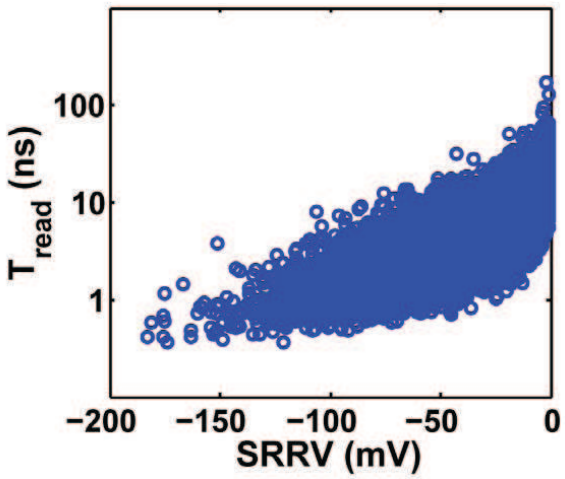


Figure 1.7: Simulated scatter plot showing the correlation between critical read stability (T_{read}) and negative static read margin (SRRV).

flowing through the bit-lines. BWTV write margin corresponds to the voltage applied on the word-line which is required to successfully write a new value into the bitcell. This is observed in the bit-line currents as a sharp change in the magnitude and direction of current flowing through the bit-lines. Correlations established with this margin can be extended to other static margins based on previously established relationships [24].

1.5.2 Critical Writeability (T_{write})

Figure 1.9 illustrates write operation to a SRAM bitcell with pulse-widths T_A and T_B. Pulse-width T_A is too short to overwrite the contents of the SRAM cell while pulse-width T_B

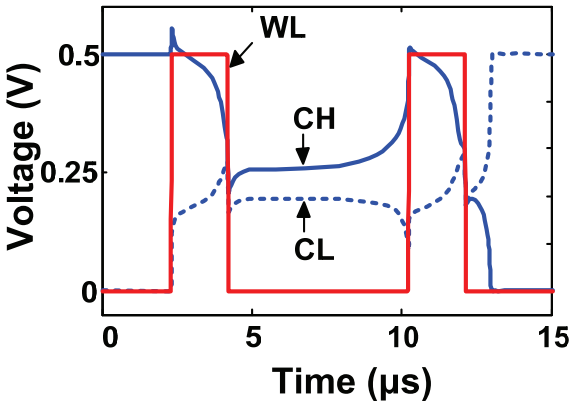


Figure 1.8: Simulated waveforms corresponding to an SRAM bitcell under read-after-read access.

is sufficient to complete the write operation. There exists a critical pulse-width, T_{write} ($T_A < T_{write} < T_B$), where the bitcell is on the threshold of a successful write access that is defined as the critical writeability. This is similar to the dynamic write failure criteria defined in [37]. This metric does not require access to the internal nodes of the SRAM cell. The challenge, however, is to reliably evaluate the contents of the bitcell after the test, without accidently disrupting the stored state.

Figure 1.10 plots the expected correlation between T_{write} and static write margin, based on Monte-Carlo simulations. Bitcells with poor static write margin (smaller values) are expected to be correlated with poor T_{write} (larger values). The dispersion between T_{write} and BWTV is small, especially at lower static margins, implying the possibility of using voltage screening either by reducing V_{CELL} or word-line bias to identify cells with poor T_{write}. Table 1.1 tabulates the sensitivities between the respective write margins to V_{th} variability in the 6 transistors of an SRAM bitcell under write operation as illustrated in Figure 1.9(a). The sensitivities in Table 1.1 reflect the negative correlation between BWTV and T_{write}. Both margins have similar magnitude of sensitivities except for the pull-up transistors, as T_{write} is correlated with variability in transistor $PU1$ while BWTV is independent, and $PU1$ is positively correlated with poor T_{write} while $PU2$ is negatively correlated with poor T_{write}. This suggests that T_{write} is more susceptible to cell asymmetry than estimated previously. Read-before-write or read-after-write does not need to be considered because the read operation only helps to upset the cell and complete the write operation [37]. T_{write} under write-after-write access, however, needs to be characterized to evaluate the impact of RTS on T_{write}.

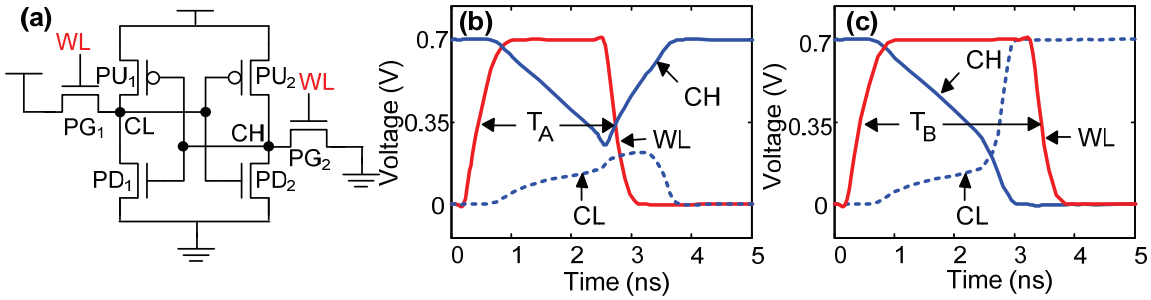


Figure 1.9: (a) Schematic of a 6-T SRAM cell under write access. (b) Simulated waveforms corresponding to failed write access with pulse-width, T_A. The bitcell retains original state. (c) Simulated waveforms corresponding to successful write access with a longer pulse-width, T_B. A new value is written into the bitcell.

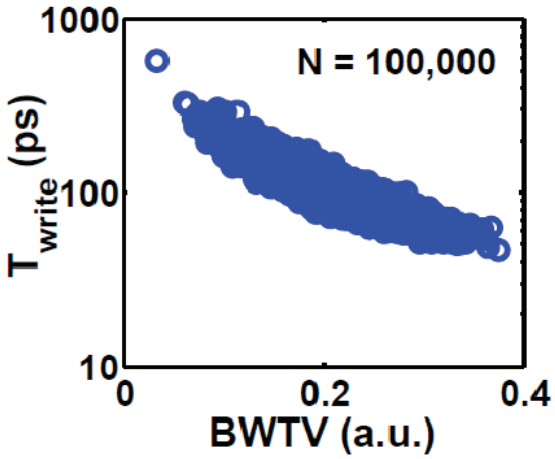


Figure 1.10: Simulated scatter plot comparing critical writeability (T_{write}) and static write margin (BWTV) obtained from Monte-Carlo simulations.

Write Metric	PD1	PG1	PU1	PD2	PG2	PU2
BWTV (V/V)	0.3	-0.3	0	0	-0.8	0.6
T_{write} (ns/V)	-0.08	1.0	1.3	0.02	3.9	-0.7

Table 1.1: Sensitivity analysis of writeability to the respective transistor V_{th} variation.

Chapter 2

Dynamic Stability Characterization

This chapter presents a characterization architecture for measuring dynamic SRAM stability through pulsed word-lines calibrated up to 10 ps accuracy [71, 72]. Measuring word-line pulse-widths calibrates out any timing uncertainty introduced by SRAM peripheral circuits, thus allowing characterization of the fundamental variability of the SRAM bitcells. This characterization methodology is validated in a commercial low-power 45nm CMOS process. The test chip also provides a means of correlation with static read and write metrics via direct bit-line measurements [24]. This method is used to identify new sources of variability in dynamic stability by observing deviations from expected correlations between dynamic stability and static margins. All voltage margins in the text are normalized to the supply voltage. Studied margins typically exhibit proportionality to the supply voltage, and normalizing them allows for comparison with prior studies (e.g. [24, 8]).

2.1 Dynamic Stability Characterization Architecture

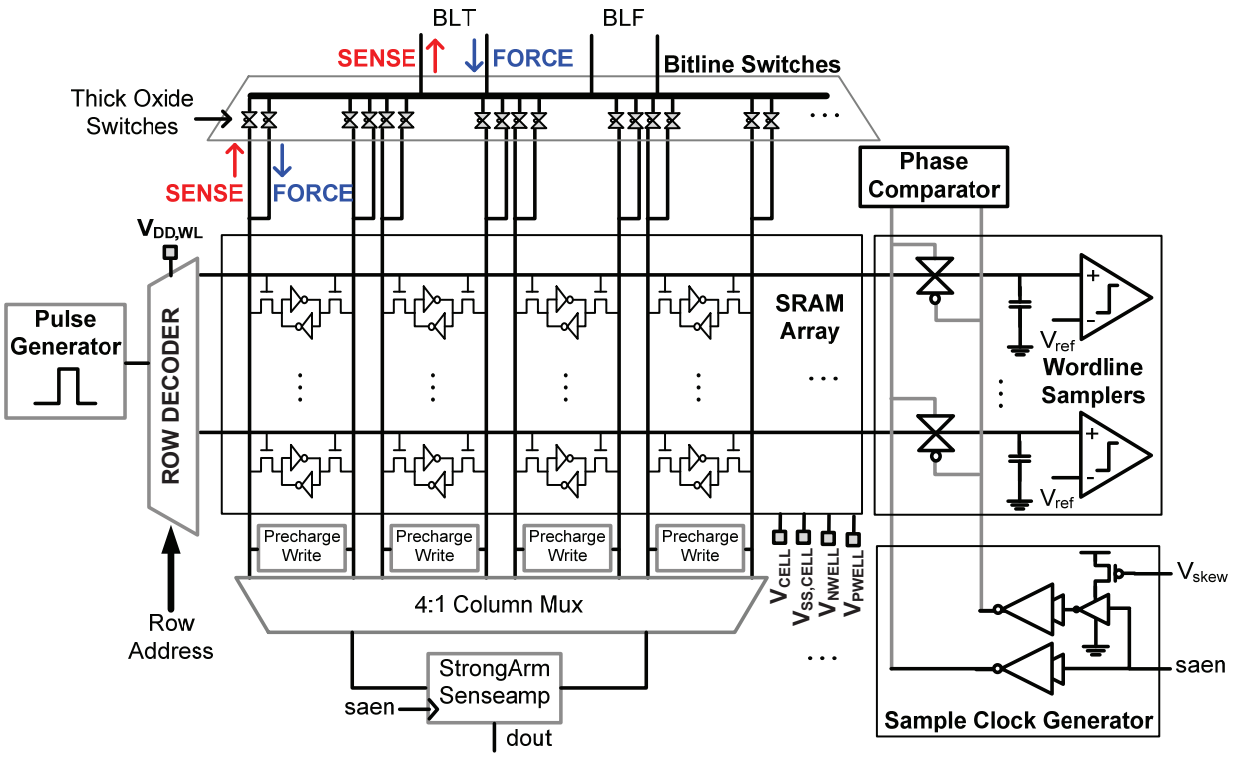


Figure 2.1: SRAM array organization for static and dynamic stability characterization.

Figure 2.1 presents the SRAM array configuration for the characterization of dynamic metrics. It also shows the necessary infrastructure for collecting static metrics for the purpose of establishing correlations with dynamic metrics. The SRAM bitcells under test are organized into a conventional SRAM array. Various array bias voltages ($V_{DD,WL}$, V_{CELL}, $V_{SS,CELL}$, V_{NWELL}, and V_{PWELL}) are connected to pads to characterize the SRAM under different read/write assist modes. A programmable pulse is generated on-chip and delivered to a single word-line at a time using existing row decoders. This architecture makes extensive use of simple circuits and calibration to ensure ease of implementation while providing measurements with high fidelity even in highly-scaled process technologies.

2.1.1 Programmable Pulse Generator

A programmable pulse is generated by simply mixing together two clocks, ϕ_0 and ϕ_1, that have a slight offset in clock period (ΔT) (Figure 2.2). This generates a pulse train with a difference in pulse-width of ΔT between successive pulses. A counter is then used to pass the

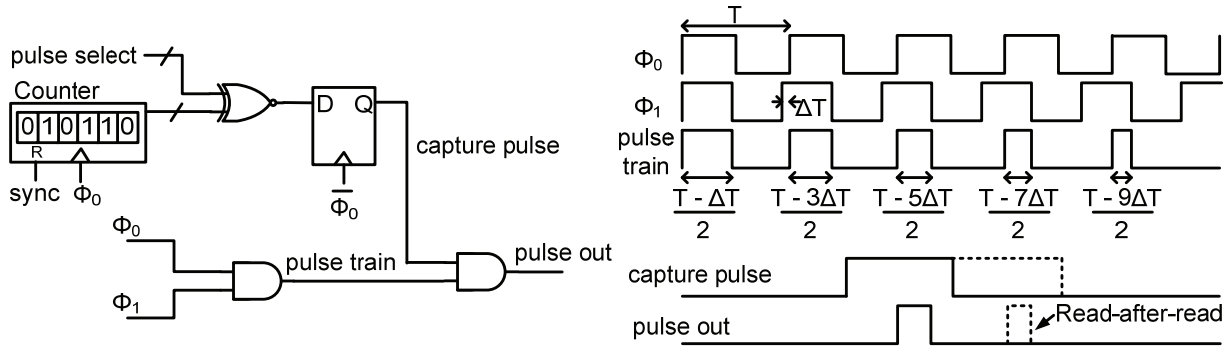


Figure 2.2: Frequency mixing programmable pulse generator with corresponding waveforms.

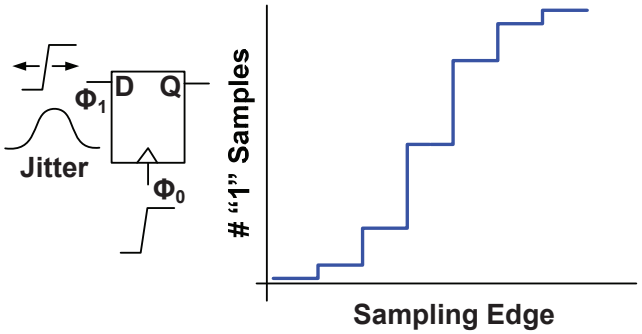


Figure 2.3: Statistics at the output of a flip-flop sub-sampling ϕ_1 using ϕ_0 with jitter.

desired pulse based on a programmed codeword. This pass signal can also be programmed to be held for multiple clock cycles to generate multiple pulses, simulating read-after-read access. The sync signal used to reset the counter is generated digitally on-chip based on statistics of the beat frequency between ϕ_0 and ϕ_1 ,averaged over 128 samples to minimize the impact of clock jitter. This signal can also be programmed to be held for multiple clock cycles to generate multiple pulses, simulating read-after-read access. This scheme allows generation of pulses with wide dynamic ranges that are immune to global process corners by tuning T and ΔT of the externally generated clocks. It also generates high bandwidth pulses (up to 100 ps) using circuits operating at much lower frequencies (10 MHz). This property is highly desirable when building characterization circuits in new process technologies with unknown performance. The lower limit of the pulse-width that can be generated is 50 ps. This lower limit is due to the limited slew rates of the word-line drivers.

A *sync* signal resets the pulse-generator counter when ϕ_0 and ϕ_1 are at a fixed phase relationship. This is generated by sub-sampling ϕ_1 using ϕ_0, producing a clock signal (A) at the beat frequency ($\phi_1 - \phi_0$). Figure 2.3 illustrates the flip-flop clocked by ϕ_0 that is

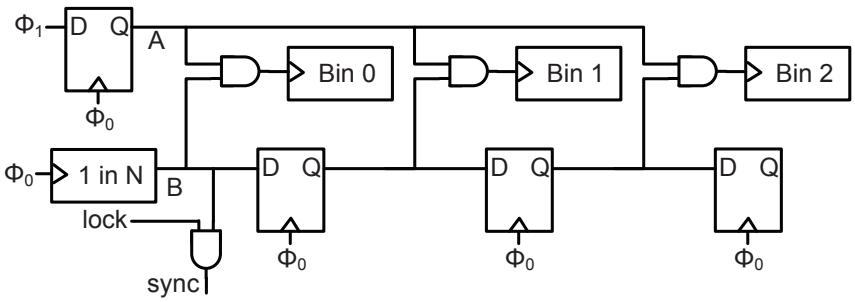


Figure 2.4: Architecture of averaging *sync* generator.

used to sub-sample ϕ_1. Jitter present in both clock signals introduce uncertainty in the *sync* signal. The impact of jitter on the *sync* generation scheme is simplified by input-referring all jitter on ϕ_0 to the input of the flip-flop and assuming that ϕ_0 is an ideal clock signal. The plot in Figure 2.3 illustrates statistics of the output of this flip-flop observed at a few sampling edges around the expected transition of the beat-signal from "0" to "1". The ideal transition of the beat-signal from "0" to "1" occurs when the probability of observing a 1 at the output at the corresponding sampling edge exceeds 50%. Figure 2.4 illustrates the schematic of the averaging *sync* generator used to obtain a highly stable *sync* signal with better than 10 ps accuracy even with up to 50 ps jitter on the clocks. A counter is configured to generate a signal every N clock cycles (N corresponds to $T_{\phi_1}/\Delta T$). This signal (B) is used to collect 3 successive samples of the sub-sampled clock (A) using 3 counters (Bin 0, Bin 1, and Bin 2). These statistics are collected for a programmable number of cycles, to average out the impact of clock jitter, before the contents of the counters are checked. This circuit effectively estimates the cumulative density function (CDF) of signal A at 3 points, spaced by ΔT. The 1 in N counter is then advanced or delayed by a single count until the mean of the sampled CDF falls into Bin 1 at which time the lock signal is asserted and the sync signal is produced. An alternative to this all-digital approach is to apply an analog low-pass filter on signal A however this requires passive components for accurately setting the cutoff frequency of the filter and lacks reconfigurability.

2.1.2 Word-line Sampler

To avoid process-induced uncertainties, the exact pulse width is measured by word-line samplers located on every word-line (Figure 2.1). This contrasts to prior work in which a small subset of the word-lines is sampled [51, 31]. The sampler consists of small transmission gates sampling the word-line pulse on a parasitic capacitance. Charge injection by the sampling clock, non-linearity of the transmission gates, and offset voltages of the comparators are calibrated out by tuning the reference voltage of the comparators. The differential clock driving the transmission gates is calibrated using a phase comparator to minimize aperture

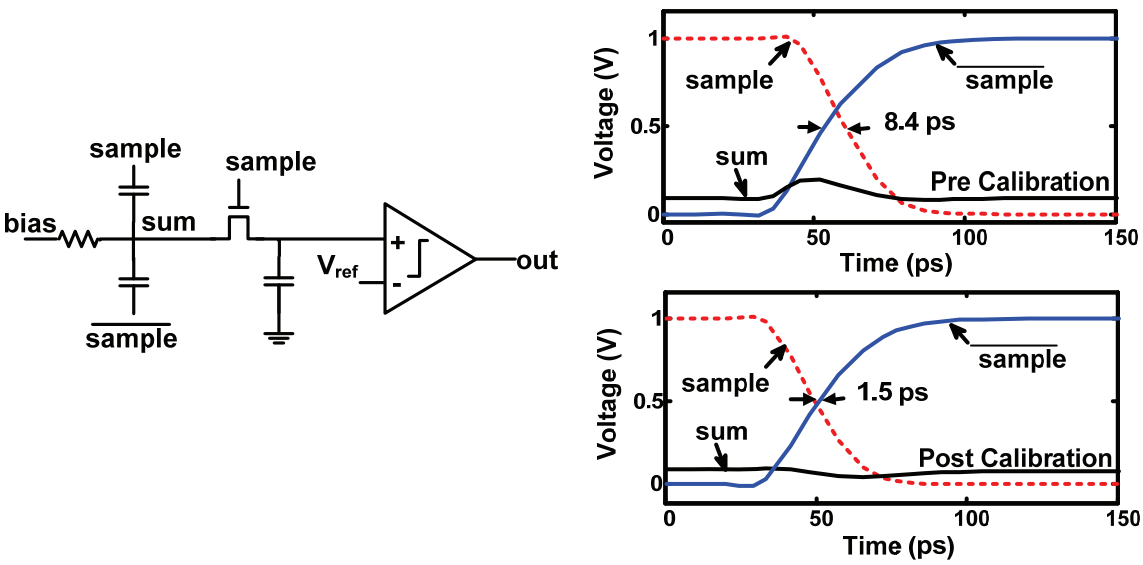


Figure 2.5: Capacitive summing phase comparator and simulated waveforms before and after calibration.

uncertainty in sampling the rising and falling edges of the word-line pulse (Figure 2.5). An ideal differential clock should have no common mode component. This phase comparator takes advantage of this fact and detects the common mode component by summing these two signals using capacitors. The calibration scheme then proceeds to skew the edges of the clock until the glitch on the sum node is minimized. A Monte-Carlo simulation of this scheme reveals that it reduces the phase offset of respective edges to less than 3 ps. The word-line pulse-width is finally measured by skewing the externally generated *saen* signal with respect to ϕ_0 with 1 ps resolution. This word-line sampling scheme produces finer resolution compared to delay-line samplers [31].

2.1.3 Non-Destructive Read-Back

Read-back refers to the operation of verifying the contents of the bitcell after a critical write or read operation that is being characterized. In writeability characterization, the purpose of the read-back operation is to check if a new value has been written into the bitcell. In read stability characterization, read-back is used to check if the stored data is lost. It is extremely important to make sure that this read-back process does not disturb the contents of the bitcell to ensure that the process that is being characterized is the SRAM operation prior to the read-back. Non-destructive read-back of the SRAM bitcells is accomplished using multiple minimum width read pulses. This allows the bitcell to gradually discharge the bitline capacitance without excessive read stress. Alternatively, V_{CELL} is raised to the

nominal voltage prior to read-back, especially when characterizing bitcells at low voltages.

2.1.4 Built-In-Self-Test

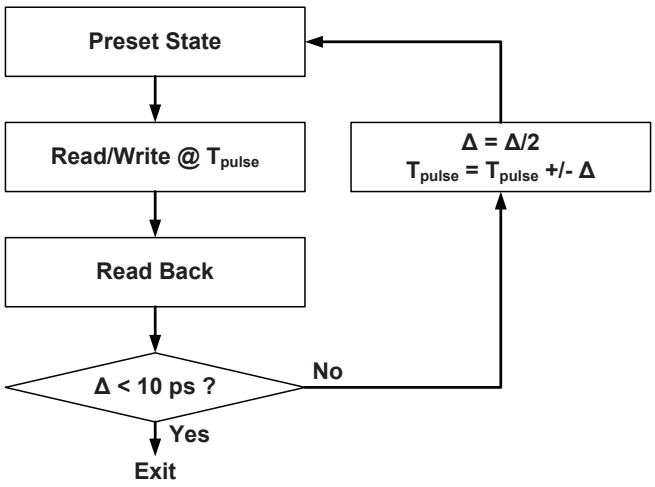


Figure 2.6: Built-In-Self-Test state machine for finding the critical word-line pulse width.

Dynamic stability characterization involves applying different word-line pulse widths to the SRAM bitcell until the critical word-line pulse width for writeability and read stability is found. This word-line pulse width search is implemented on-chip as a 3-phase state machine for binary search of the critical word-line pulse width (Figure 2.6). The first phase presets the SRAM bitcell to a known state. The second phase performs a read or write operation at the pulse width being tested and finally the contents of the cell are read back. This sequence is then repeated again with a shorter or longer pulse width, depending on the result of the read back operation. The built-in-self-test (BIST) can also be configured to take multiple samples at each pulse width to average out the impact of jitter on the instantaneous pulse width. The BIST is implemented using synthesized digital logic, placed, and routed using automated tools. Custom hand-designed logic is not necessary in the BIST implementation of this characterization architecture because all the high speed components of the design are limited to the pulse generator.

2.2 45 nm CMOS Implementation

A 1.55 mm x 1.55 mm test chip [71], [73], [80] is implemented (Figure 2.7) in a low-power strained-Si 45 nm CMOS process [33] with Poly-Si/SiO_xN_y gate stack and 7 metal layers. Experimental, high density 0.252 μm^2 6T SRAM bitcells that are smaller than ITRS

requirements for the 45 nm technology node are characterized to observe a larger impact of process-induced variability on SRAM performance and also to predict variability in future scaled transistors. The test chip consists of two 64x256 arrays and two 128x256 arrays with full static and dynamic stability characterization coverage. The narrower array (64 columns) has reduced word-line parasitics and is used to characterize dynamic stability at high speeds with strict requirements of rise- and fall-transition times. The word-line samplers contribute to a 16% array area overhead. The level-shifters and bit-line switches incur a larger area penalty and are required solely for static margin characterization.

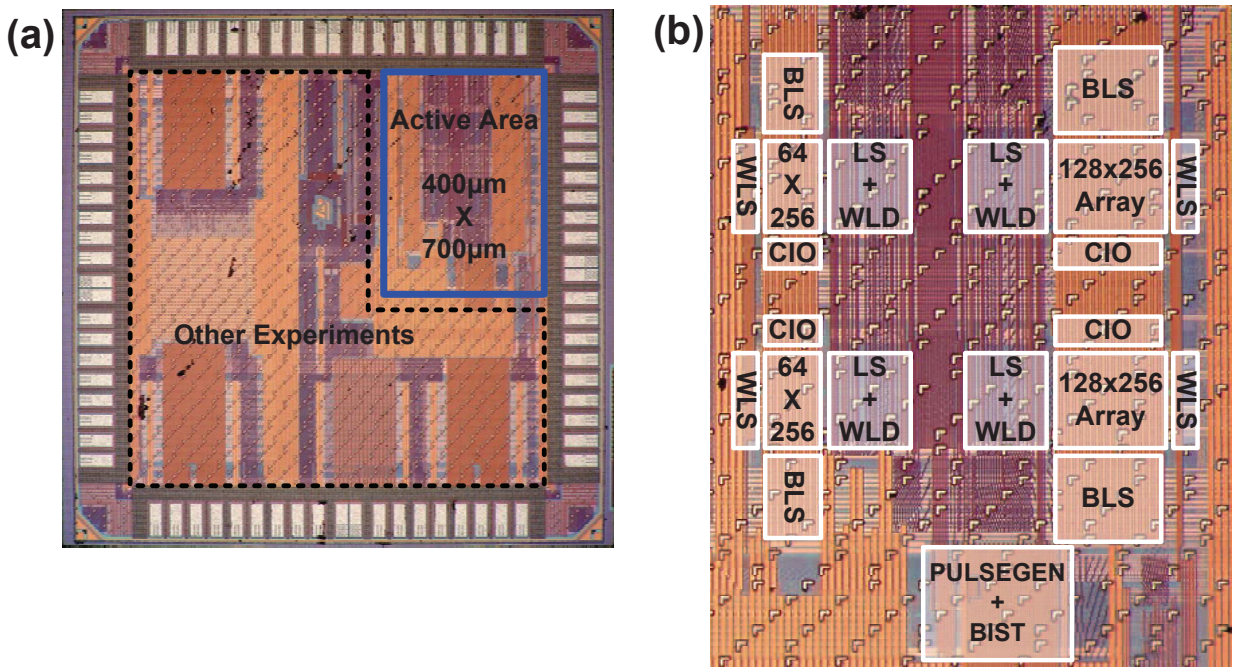


Figure 2.7: (a) Die photo of the 45 nm CMOS test chip. (b) Die photo of active area. BLS : bitline switches; WLS : word-line samplers; LS+WLD : level shifters and word-line drivers; CIO : column I/O circuitry

2.3 Measurement Results

Figure 2.8 illustrates fail bit count measured from the test chip, indicating 10-100X discrepancy between quasi-static (>1 s with bit-lines driven) and dynamic access. Static access fail bit counts are optimistic for writeability and pessimistic for read stability, compared to those for dynamic access. More than 10 write failures were observed at nominal V_{DD} when the bitcells were accessed with 1 ns pulses even though no write failures occurred when the bitcells were accessed quasi-statically. No read upset failures occurred when the

bitcells were accessed with 20 ns pulses even though tens of failed bits were observed when the same bitcells were accessed quasi-statically.

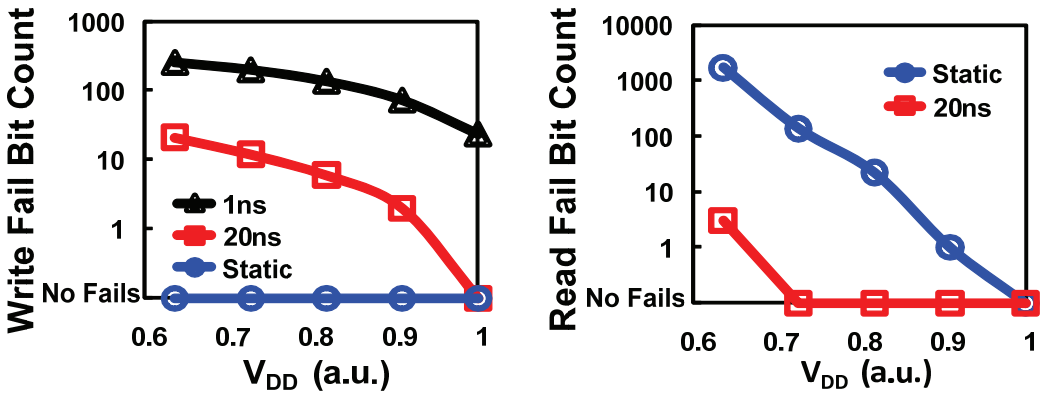


Figure 2.8: SRAM writeability and read stability fail bit count measured from a 45nm CMOS SRAM.

2.3.1 Pulse Generator

Multiple complete waveforms of word-line pulses were sub-sampled and plotted in real time in Figure 2.9(a). Good rise and fall transition times of 75 ps and 30 ps were observed. Note that the rise and fall transitions account for a significant portion of narrow pulses (less than 100 ps) and effectively limit the correlation between static and dynamic margins. The pulse-width, corresponding to the delay between the 50% voltage level of the rise and fall transitions, was measured across different codewords. The transfer function and the measured linearity error are plotted in Figure 2.9(b). Up to 100 ps of non-linearity was observed in the transfer function. This error is believed to be caused by voltage droop in the power supply grid as the pulse is being distributed across the chip. These non-idealities demonstrate the importance of calibrating word-line pulse-widths at every word-line in order to calibrate out this source of uncertainty from actual variability in the bitcells. All dynamic SRAM measurements presented are based on word-lines calibrated to 10 ps resolution using low-jitter signal generators and averaging.

2.3.2 Read Access Time

Figure 2.10(a) plots the statistical distribution of T_{access} measured from 1024 bitcells at 0.8X nominal V_{DD}. These 1024 bitcells were spread across multiple columns in the SRAM array with sense-amplifiers having different offset voltages on different columns. The distribution is observed to be multi-modal, a superposition of multiple Gaussian distributions. The multi-modal nature of this distribution is due to the strong dependence of read access

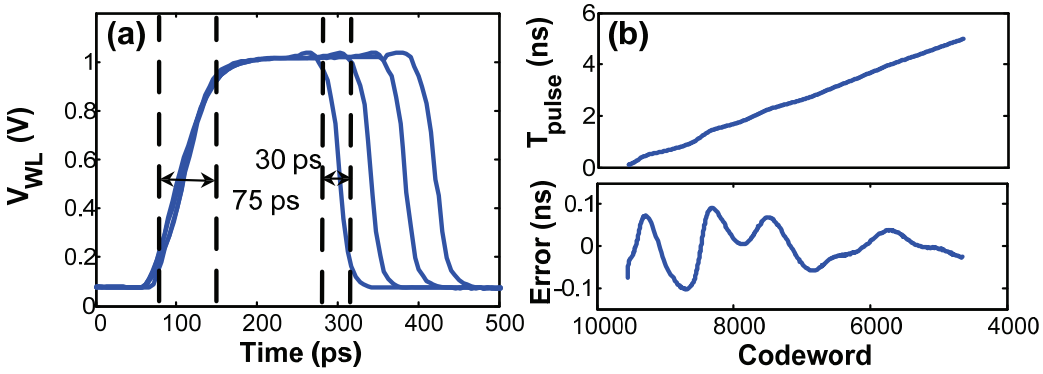


Figure 2.9: Plots of (a) multiple sub-sampled word-line waveforms and (b) codeword to pulse width transfer function and measured error.

time on sense-amplifier offset voltage, as predicted in Equation 1.1. The offset voltage of each sense-amplifier was characterized separately using the bit-line switches and external measurement equipment to normalize out this factor in T_{access}. Measurements of T_{access}, normalized with separately characterized sense-amplifier offset voltages and estimated bit-line capacitance, was observed to correlate ($R^2 = 0.69$) with static read current (Figure 2.10(b)). The remaining dispersion in the data is due to the inherent difference between I_{read} statically measured out of the bitcell at a fixed bit-line voltage and the transient bitcell current as the bit-line is being discharged.

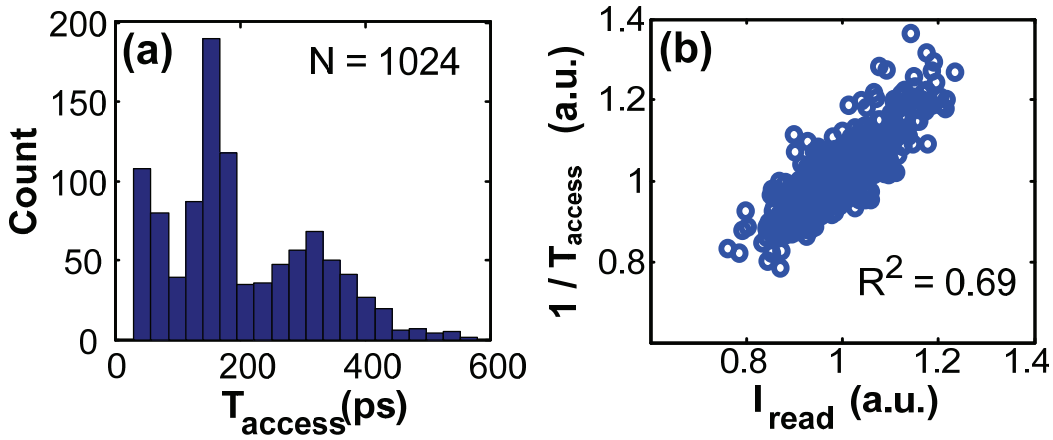


Figure 2.10: (a) Histogram of measured read access time. (b) Scatter plot showing correlation between read access time and static read current after normalization with sense-amplifier offset voltage and bit-line capacitance.

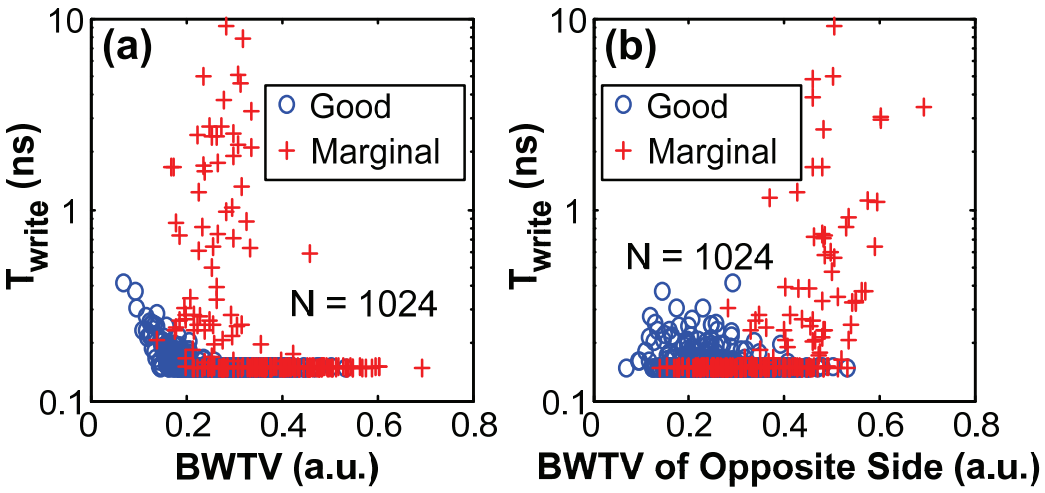


Figure 2.11: Critical writeability vs. static write margin of (a) same side and (b) opposite side of SRAM cell measured at $V_{DD,low}$.

2.3.3 Critical Writeability

Figure 2.11(a) plots measurements of critical writeability vs., the static write margin for writing the same data value to the same bitcell. Each data-point of T_{write} corresponds to an average of 128 measurements. Expected correlation between poor BWTV and T_{write} is observed in Figure 2.11(a), however, the uncorrelated outliers exceed the correlated data-points by more than ten times. These outliers are observed to appear exclusively in bitcells that have large static write margin on the opposite side of the cell (Figure 2.11(b)). Further analysis of individual transistor characteristics using Direct Bit-line Transistor Access (DBTA) (ref. [18]) revealed that a large number of bitcells sampled had large drain series resistance in one of the PMOS transistors. These marginal transistors were found to be on the side opposite to the half-cell being written to (PU1 in Figure 1.9(a)), causing a significant degradation in the speed of the bitcell for pulling the storage node up to V_{DD}. The remaining bitcells showed good correlation between T_{write} and BWTV metrics, after the marginal cells were screened out (Figure 2.11(a)). These marginal transistors did not degrade static write margin due to the negligible sensitivity of the margin to variability in $PU1$ (Table 1.1).

Voltage screen tests such as described in [7] are commonly used to screen out defects and early failures in SRAM arrays. Such tests are usually carried out in-line at wafer sort using testers running at lower frequencies than actual operating frequencies. The lack of correlation between the outliers in critical writeability and static write margin invalidates results obtained from such tests because the bitcells screened by these tests are not the bitcells that fail first at normal operating frequencies.

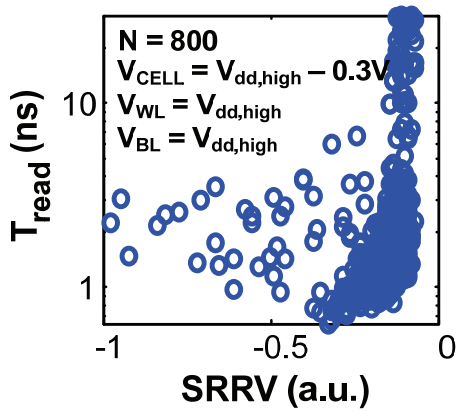


Figure 2.12: Critical read stability versus static read margin.

2.3.4 Critical Read Stability

Figure 2.12 plots measurements of critical read stability against the negative static read margin. These measurements were obtained by lowering V_{CELL} by 300 mV relative to word-line and bit-line pre-charge voltage levels, to increase the probability of observing cells that are unstable under static access. The expected correlation between T_{read} and negative SRRV (ref. Figure 1.7) was observed in measurements. Bitcells with marginally negative static read margin (approximately 0.1 a.u.) were observed to have a large dispersion in T_{read} ranging from 1 ns to 1 μs. This dispersion reduces as the bitcell SRRV becomes more negative. The minimum T_{read} observed was 630 ps, indicating that this SRAM bitcell can be accessed with pulse-widths shorter than 630 ps without read upsets even with 300 mV of V_{CELL} droop. Outliers with extremely poor SRRV that are not correlated with smaller values of T_{read} were observed. Such outliers were not observed in Monte-Carlo simulations of a large 100,000 sample set (Figure 1.7).

Figure 2.13 plots statistical distributions of critical read stability under single read and read-after-read access with 20 ns period (T_{cycle}). As expected, T_{read} degrades under read-after-read conditions [37]. Bitcells with small values of T_{read} (less than 2 ns) were observed to shift only by a small amount, while bitcells with larger T_{read} were observed to degrade by up to 1 ns. Susceptibility of a bitcell to read-after-read upset depends on the proximity of the internal node voltages to the rails when the next read pulse arrives. Bitcells with smaller values of T_{read} are less susceptible to read-after-read upsets, compared to bitcells with larger T_{read} accessed with the same T_{cycle}, because these bitcells have longer recovery periods to settle at the rail voltages. Figure 2.14 plots T_{read} of a single bitcell as a function of the number of read-after-read pulses across decreasing T_{cycle}. The degradation in T_{read}, due to read-after-read, increases as T_{cycle} is decreased. This degradation saturates eventually after 6 cycles in direct agreement with [37]. Evidence of slight T_{read} degradation even with a relatively slow T_{cycle} of 67 ns suggests that the recovery period of this bitcell is more than

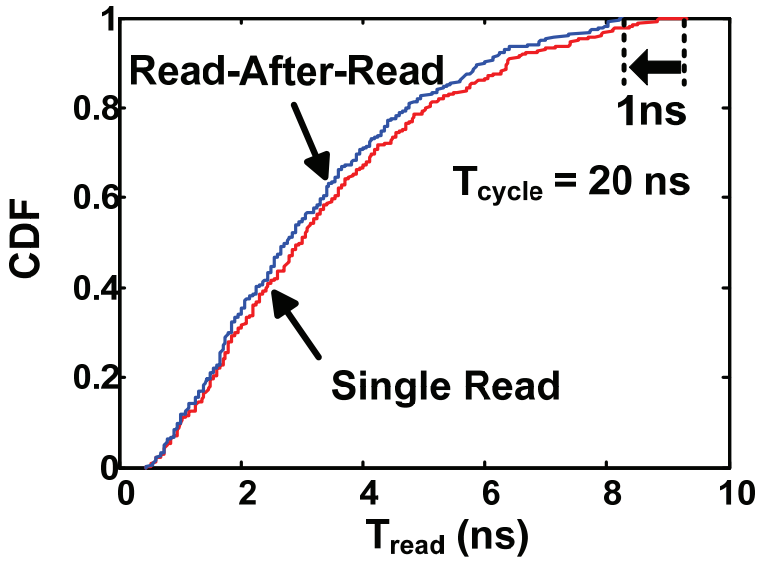


Figure 2.13: Statistical distributions of critical read stability under single read and read-after-read access with 20 ns clock period.

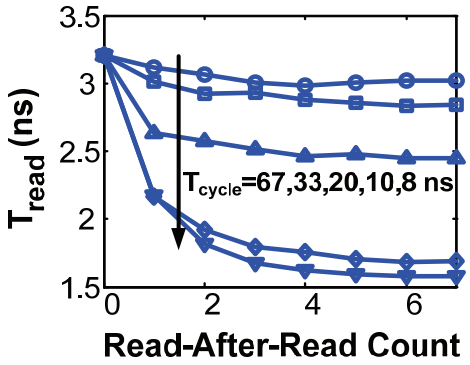
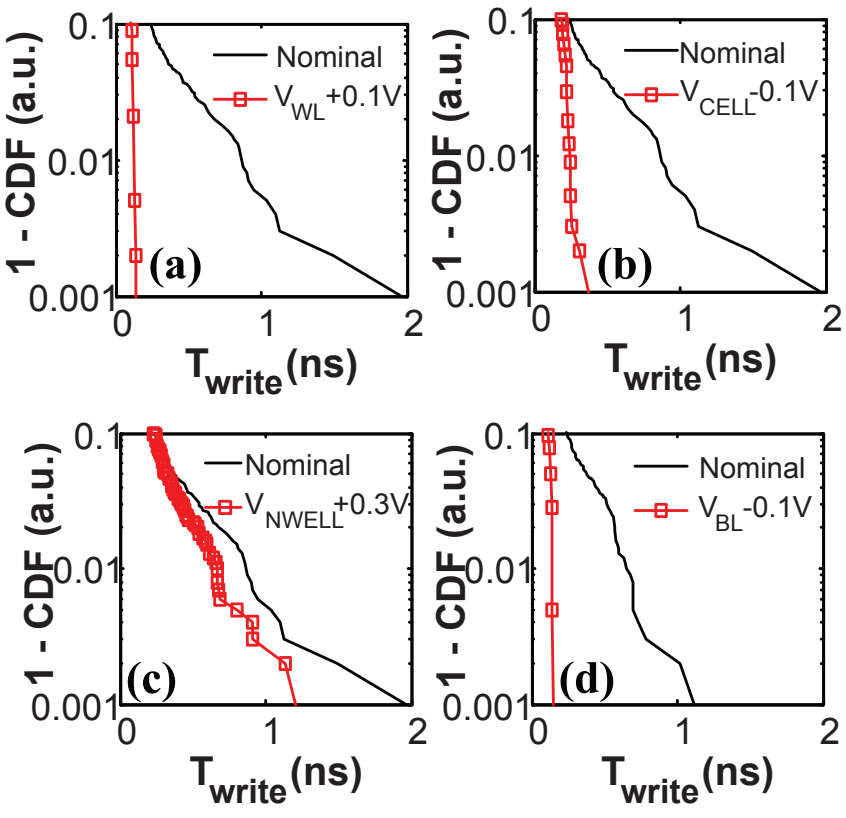


Figure 2.14: Critical read stability of a selected bitcell as a function of the number of read-after-read cycles. The different curves correspond to the period of the read-after-read cycles.

67 ns, which is greater than 20 times the single-read T_{read} of this bitcell (3.2 ns).



[H]

Figure 2.15: Survival function of T_{write} under different bias conditions: (a) Word-line boosting; (b) V_{CELL} under-drive; (c) PMOS reverse body-bias; (d) Negative bit-line.

2.4 Impact of Assist Techniques

2.4.1 Write Assist

Figure 2.15 compares the impact of different assist techniques on T_{write}. Word-line voltage (V_{WL}) boosting and V_{CELL} under-drive resulted in significant speed-up of T_{write} [88]. V_{WL} boost was slightly more effective than V_{CELL} under-drive because it increases the strength of the pass-gate transistors which have the strongest impact on T_{write}. Figure 2.15(c) plots the statistical distributions of T_{write} under 300 mV of PMOS reverse body-bias (RBB) [85]. Not much improvement in T_{write} was observed even with 300 mV of RBB due to the small body-effect coefficient for this 45 nm CMOS process. RBB might even have a detrimental effect on T_{write}, due to the opposite sensitivities of T_{write} to variability in $PU1$ and $PU2$. Figure 2.15(d) investigates write assist using negative voltage levels on the bit-lines [63]. A 100 mV negative bit-line bias results in a significant improvement in T_{write}.

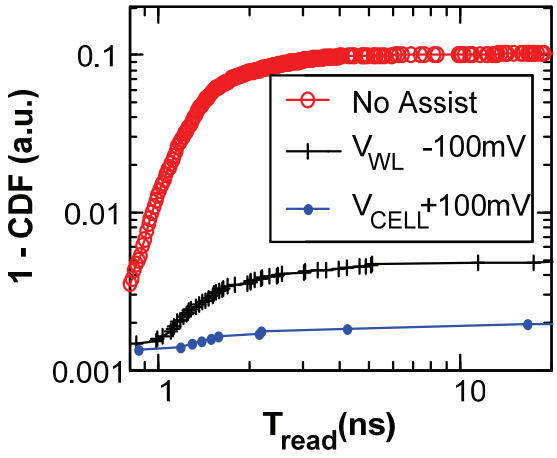


Figure 2.16: Survival function of T_{read} without assist techniques, with -100 mV word-line bias offset and with +100 mV V_{CELL} offset.

2.4.2 Read Assist

Figure 2.16 demonstrates the effectiveness of V_{CELL} boosting and V_{WL} under-drive for read assist [54]. V_{CELL} boosting was found to provide a larger improvement in critical read stability compared to V_{WL} under-drive. SRAM design using assist techniques involves a delicate balance of bias voltages in order to balance out the improvement in one margin with the degradation in the other. The strong sensitivity of read and write stability to V_{WL} and V_{CELL} biasing suggests the possibility of using these two voltage tuning knobs to increase the overall reliability of the SRAM array. Results in this section however demonstrate that this technique needs to be used with caution as slight offsets in V_{WL} will affect T_{write} and T_{read} exponentially. Because of this, any uncertainty or noise in setting V_{WL} could potentially result in large write or read stability failures. Evidence of this is observed in simulated results presented in Figure 5.6 where strong opposite sensitivities of read and write failures to word-line voltage boost is observed.

Chapter 3

Random Telegraph Signal and SRAM Variability

3.1 Introduction

Random telegraph signal (RTS) noise corresponds to a low-frequency noise phenomena that modulates intrinsic transistor parameters [59, 38]. This phenomena is caused by trapping and de-trapping of electrons or holes at traps within the gate oxide stack of the transistor. Theoretically, noise amplitude due to RTS scales with the inverse of channel area (L x W) and the square of effective gate oxide thickness ($t_{ox,e}^2$) [47]. The inverse relationship to channel area is due to the fact that RTS is related to a fluctuation in the number of carriers in the channel, which is primarily determined by the channel area. Although RTS amplitude should remain constant with conventional technology scaling (L, W, and $t_{ox,e}$ are scaled by the same factor), the stagnation of $t_{ox,e}$ scaling due to increase in gate leakage currents has caused an increase in RTS amplitude observed in highly scaled technologies. RTS amplitude is further exacerbated by the discreteness of dopants in the channel, creating percolation paths of carriers in the channel that are strongly attenuated by the occupation of traps in the oxide located within proximity of these paths [6]. It has been demonstrated that ΔV_{th} due to RTS exceeds ΔV_{th} due to random dopant fluctuation (RDF) at 3 sigma levels at the 22 nm technology node [66]. RTS is therefore likely to impact SRAM robustness due to the small channel area used in SRAM bitcells and also the large number of functional bitcells required in typical designs (greater than 6 sigma).

This chapter first examines the dynamics of RTS with emphasis on the bias and temperature dependence of these dynamics and the change in these dynamics when large bias swings are applied on the transistor. This is followed by a review of various RTS amplitude models and proposal of an empirical model that is able to capture the bias dependence of RTS amplitude. The impact of RTS on SRAM variability is first examined by evaluating the statistical distributions of RTS amplitude in the pull-up, pull-down, and

pass-gate transistors. The impact of RTS in the individual transistors on SRAM write margins is then evaluated using a statistical model, calibrated with 45 nm SRAM bitcells, and also verified experimentally through large-scale measurements of SRAM V_{min}. Finally, the impact of low-frequency RTS on SRAM bitcells operating at high frequencies is presented. All experimental data presented in this chapter are based on measurements from transistors fabricated in a pre-production 45 nm process with Poly-Si/SiO_xN_y gate stack.

3.2 Dynamics of Random Telegraph Signal

3.2.1 Background

Figure 3.1 plots the drain current measured from a PMOS SRAM transistor exhibiting a random telegraph signaling noise. The drain current fluctuates between two discrete levels with the higher drain current corresponding to an empty trap and the lower drain current level corresponding to the trap being occupied. Also present in this measurement is a secondary RTS noise with much smaller amplitude. The dynamics of RTS are characterized using time constants that characterize the time till capture of a carrier (τ_c) and the time till emission of a carrier (τ_e) [59]. Figure 3.2 plots the histogram of τ_c and τ_e extracted from multiple RTS events. τ_c and τ_e are observed to fit exponential distributions with their respective time constants that correspond to the average time constants [38]. Figure 3.3 plots the frequency characteristics of the RTS trace. This RTS noise in the time domain translates to a Lorentzian spectrum in the frequency domain with a corner frequency and $1/f^2$ characteristic decay. It has been speculated that the combination of multiple RTS noise sources in a transistor with uniformly separated corner frequencies sum up to $1/f$ noise [38]. In SRAM transistors, it is not feasible to observe such a characteristic because the small transistor geometries effectively limit the number of traps in the oxide to a small value that is insufficient to sum up to a general $1/f$ noise characteristic.

There is a large interest in modeling the physical origin of RTS dynamics because an accurate RTS model would allow inference of the quality of the gate oxide stack just by monitoring the dynamics of RTS. Several theories exist for explaining the time constants of these traps. The McWhorter model [45] suggests that the drain current fluctuations are due to elastic tunneling of carriers between the channel and gate oxide defects. Based on this model, the time constants of RTS are governed by elastic tunneling events that are related to the depth of the trap in the gate oxide. The McWhorter elastic tunneling model has largely been dismissed in recent studies due to its inability to account for temperature activation of the observed noise spectrum. More recently, it has been demonstrated that the time constants predicted using the McWhorter model are greater than 3 orders of magnitude faster than what is being observed in modern devices [11].

Kirton and Uren proposed that the time constants of charge trapping and de-trapping are governed by lattice relaxation [38]. A trap exists within a lattice of Si-O_2. Trapping and de-

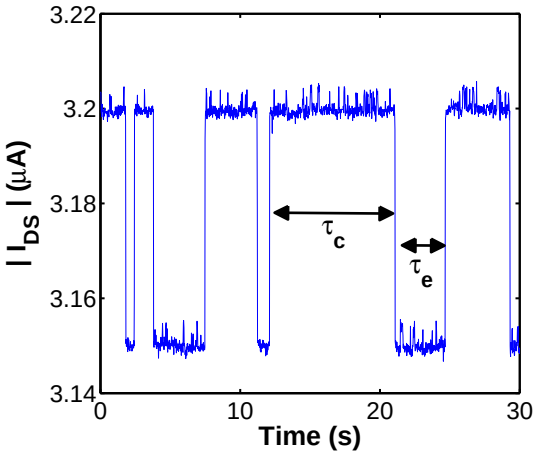


Figure 3.1: Drain current measured from a PMOS SRAM transistor showing RTS.

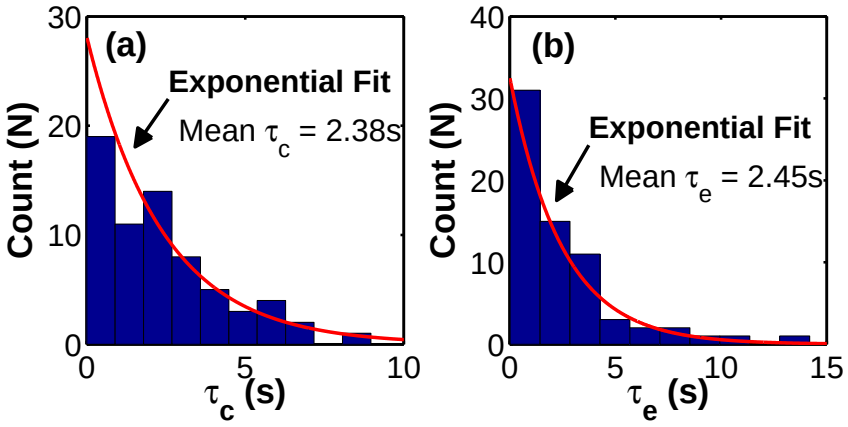


Figure 3.2: Exponential fit of extracted time constants corresponding to (a) time until capture (τ_c) and (b) time until emission (τ_e).

trapping of a charge involves overcoming the energy barrier associated with re-organization of the lattice configuration around the defect. This re-organization process therefore determines the time constants of these traps. This model however fails to capture the strong bias dependence of τ_e, a fact that is acknowledged by the authors. More advanced models, based on lattice relaxation, have been proposed and results look promising [43, 23]. The current state is that, after more than 5 decades of research, there is still no consensus within the modeling community on the physical origin of RTS dynamics. This thesis does not intend to contribute to the ongoing debate. As such, it will rely on actual RTS characterization for most conclusions and use physical models for extrapolation when qualitative results are

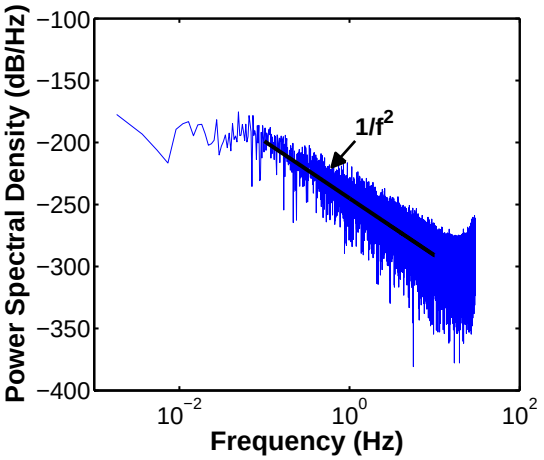


Figure 3.3: Power spectral density of RTS drain current measurement.

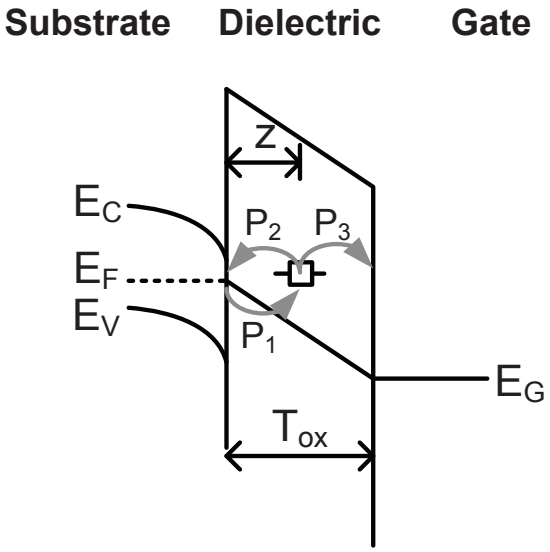


Figure 3.4: Energy band diagram of oxide trap with tunneling probabilities.

sufficient to prove a point.

$$\frac{\bar{\tau_c}}{\bar{\tau_e}} = \frac{1 - f_T}{f_T} = g\ e^{(\frac{E_T - E_F}{k_B T})} \tag{3.1}$$

$$ln\frac{\bar{\tau_c}}{\bar{\tau_e}} = K - \frac{q}{k_B T}\Big[\Big(1 - \frac{z}{T_{ox}}\Big)\Psi_s + \frac{z}{T_{ox}}V_g\Big] \tag{3.2}$$

$$\pi_i P_{ij} = \pi_j P_{ji} \tag{3.3}$$

The mark-space ratio ($\frac{\bar{\tau_c}}{\bar{\tau_e}}$) is yet another dynamic RTS metric. It is particularly attractive because it is easily extracted from experimental data where only a single trap is active in the measurement window by measuring the duty cycle of the signal [68, 53]. Equation 3.1 is obtained by detailed balance (equal numbers of up and down jumps), where f_T is the trap occupancy, g is the trap degeneracy, E_T is the trap energy level while E_F is the Fermi level, k_B is Botzmann's constant and T is the temperature [59]. For easier comparison with measurements, this equation can be written as Equation 3.2, where q is the charge of an electron, $\frac{z}{T_{ox}}$ is the relative depth of the trap from the channel, Ψ_s is the surface potential, V_g is the gate voltage, and K is a constant [29]. Ralls *et al.* reported that oxide trap depth can be inferred from measuring the response of the mark-space ratio to gate bias and fitting the data to the first derivative (with respect to V_g) of Equation 3.2. Special attention needs to be made to model the dependence of Ψ_s to V_g in order to arrive at a sound estimate of $\frac{z}{T_{ox}}$, especially when the gate bias is close to the threshold voltage [38]. This technique has been used to investigate oxide traps in high-κ metal gate stacks and also to collect large-scale statistics of trap depth and activation energy [53, 42, 78]. Equation 3.1 is derived based on detailed balance and makes no assumption of the tunneling mechanism. As such, it can be applied even if there exists uncertainty in the actual tunneling mechanism.

Next, let us consider the assumptions required for applying detailed balance. Figure 3.4 illustrates the energy band diagram of an NMOS transistor with an oxide trap. P_1 indicates the tunneling probability of a carrier from the channel into the oxide trap while P_2 and P_3 indicate the tunneling probabilities for de-trapping back into the channel or out into the gate, respectively. Detailed balance requires that Equation 3.3 holds true. This means that the carrier originates and ends at the same state which is only true if P_3 is 0. While this is true for most cases (traps that are important to transistor performance are usually close to the channel), cases where de-trapping occurs out through the gate have been observed [42]. Detailed balance might not hold also in cases where trapping/de-trapping occurs through intermediary states. Estimations of RTS dynamics based on detailed balance are therefore effective for most cases but needs to be used with caution to account for the various cases where the assumptions are not valid.

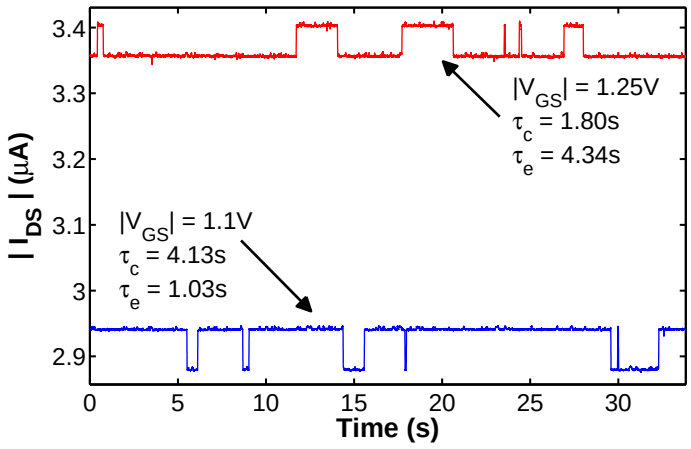


Figure 3.5: Drain current of PMOS SRAM transistor under different gate biases.

3.2.2 Bias and Temperature Dependence

Figure 3.5 plots transient measurements of drain current from a PMOS transistor at two different bias conditions, with an obvious RTS noise, demonstrating the strong dependence of RTS dynamics on bias. At $|V_{GS}|$ of 1.1 V, the trap is mostly unoccupied ($\frac{\bar{\tau_c}}{\bar{\tau_e}} > 1.0$). At the higher bias condition, the trap stays mostly occupied ($\frac{\bar{\tau_c}}{\bar{\tau_e}} < 1.0$). This negative dependence of $\frac{\bar{\tau_c}}{\bar{\tau_e}}$ on gate bias matches Equation 3.2, suggesting that this is a trap where the carrier jumps from the channel into the oxide trap and detraps back into the channel ($P_3 = 0$ in Figure 3.4). In this instance, both τ_c and τ_e respond strongly to gate bias, however this is not a general trend across the entire bias range. Figure 3.6 tracks the evolution of the time constants across different gate biases demonstrating that at low voltages, τ_c is extremely sensitive to slight changes in gate bias while τ_e is weakly dependent on gate bias. The opposite sensitivities are observed at higher voltages. Figure 3.7 plots transient measurements of drain current from the same PMOS transistor at different temperatures. Increasing temperature speeds up both the capture and emission time constants, indicating that temperature plays an important role in the transport of a carrier into and out of the trap.

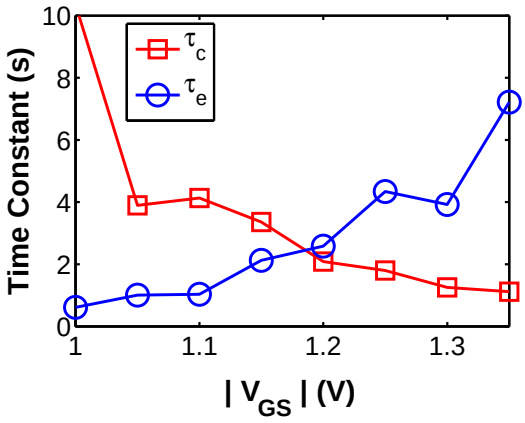


Figure 3.6: Dependence of τ_c and τ_e on gate bias.

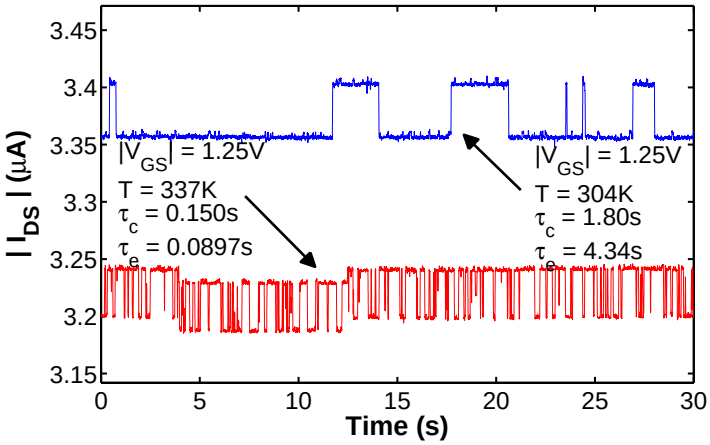


Figure 3.7: Drain current of PMOS SRAM transistor at different temperatures.

Figure 3.8 summarizes measurements of RTS dynamics extracted from the same trap, across different gate bias and temperatures. Parameters (K and $\frac{z}{T_{ox}}$ for Equation 3.2 were extracted based on exponential fits to the data. These parameters are used to model the average trap occupancy across bias and temperature in Figure 3.9. Average trap occupancy, or α, is related to the respective RTS time constants according to Equation 3.4. K and Ψ_s were assumed to be constant in this analysis. While a lot of assumptions were made in the extrapolation used to produce Figure 3.9, it serves the purpose to demonstrate that increasing $|V_{GS}|$ causes a trap that is initially always empty ($\alpha = 0$) to become fully occupied ($\alpha = 1$) while increasing temperature speeds up the time constants of the trap but only has a slight impact on average trap occupancy.

$$\alpha = f_t = \frac{1}{1 + \frac{\tau_c}{\tau_e}} \tag{3.4}$$

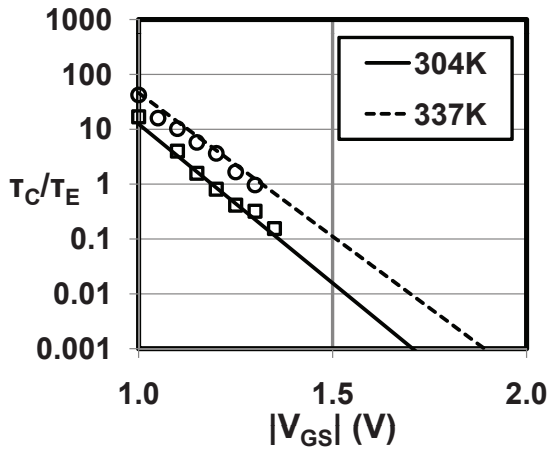


Figure 3.8: Measured mark-space ratio at different gate bias and temperatures with extrapolation. Lines indicate exponential fits.

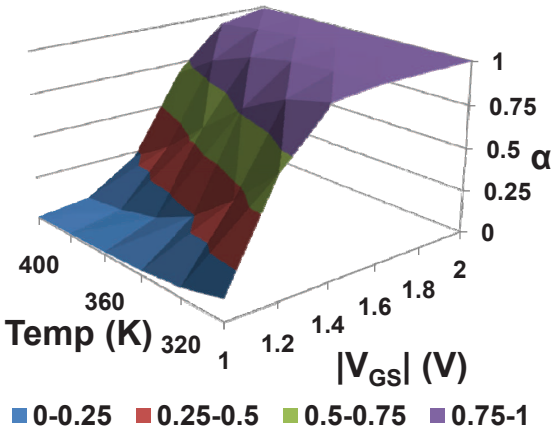


Figure 3.9: Surface contour of average trap occupancy (α) as a function of gate bias and temperature, extrapolated using Equation 3.2 with $K = 26$ and $\frac{z}{T_{ox}} = 0.35$.

The oxide traps analyzed so far exhibit a negative dependency of the mark-space ratio on gate bias i.e. increasing gate bias increases the probability of trap occupancy. There exists a different trap variety that exhibits a positive dependency of the mark-space ratio on gate bias [53, 59]. Figure 3.10 plots measurements of the dynamics of these traps, referred

to as type II traps by Nagumo . τ_c increases with gate bias while τ_e decreases with gate bias. The net effect is that increasing gate bias decreases the probability of trap occupancy. This trap is clearly active under normal transistor operating conditions ($V_{DD} < 1V$) and will result in performance degradation especially in SRAM. Ralls *et al.* characterized these traps as a positive scattering center that seems to be neutralized by capture of a carrier [59]. Nagumo *et al.* speculates that type II traps are due to exchange of carriers between the gate electrode and the oxide trap. Figure 3.11 illustrates the band diagram corresponding to a type II trap. P_1 and P_2 in this figure corresponds to the tunneling probability of a carrier from the gate electrode into the trap and from the trap back into the gate electrode. This differs from the band diagram corresponding to type I traps (ref. Figure 3.4) where the trap interacts with a carrier tunneling from the channel. Equation 3.2 is modified to reflect this assumption and the trap depth and energies are extracted, assuming detailed balance. Analysis of these parameters, based on statistics collected from multiple traps, indicate that these type II traps are physically closer to the gate electrode and are within the proximity of the gate work function (E_G), as depicted in Figure 3.11. While this model for type II traps sounds plausible, it still needs to be verified by measuring the gate current and correlating fluctuations in gate current with fluctuations in drain current, which is a challenging measurement due to the small amplitude of gate current fluctuation caused by a single carrier.

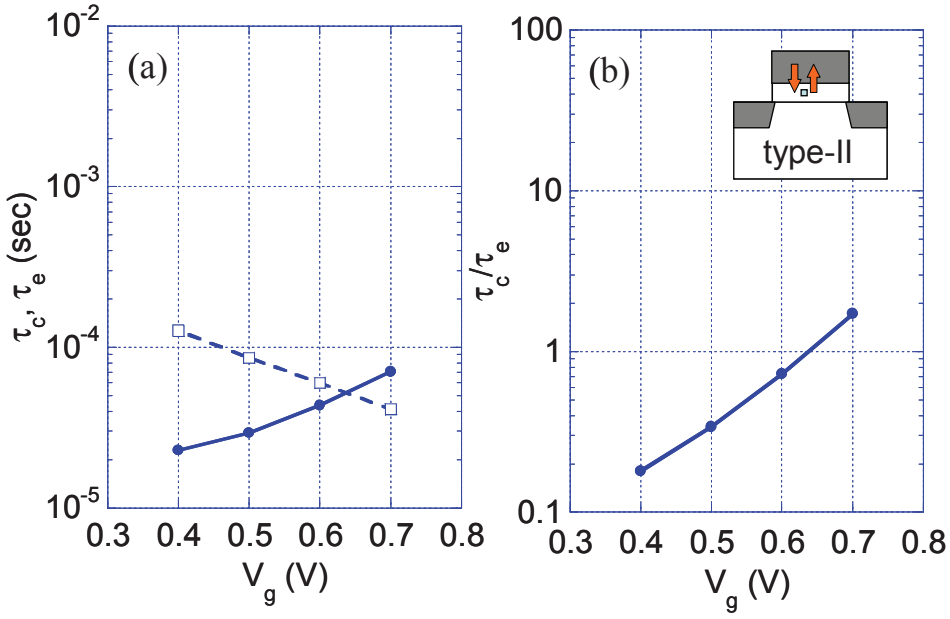


Figure 3.10: (a) V_{GS} dependence of τ_c (filled circles) and τ_e (open squares) and (b) mark-space ratio of a type II trap. Reference [53]

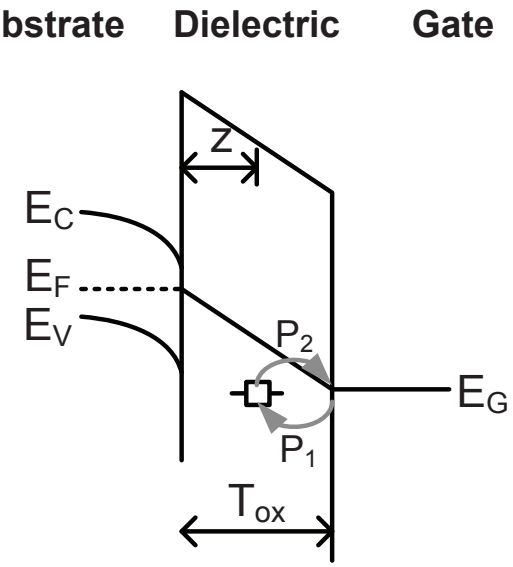


Figure 3.11: Band diagram of type II oxide trap with tunneling probabilities.

Other trap characteristics have also been observed. Figures 3.12 and 3.13 plot the characteristics of a type III trap. Both τ_c and τ_e corresponding to this trap decrease with increasing gate bias. Traps with this characteristics have also been reported by others and correlated directly with fluctuation in gate current [42]. These traps are usually active under high gate voltage and are most probably caused by trapping from channel to oxide trap and de-trapping out through the gate (finite P_3 in Figure 3.4). Type III traps are not that critical under normal operation because they are only active at high voltages. These traps were not observed under nominal bias conditions of less than 1 V. They were only observed when the transistor was biased with high gate bias voltages for bias-temperature instability (BTI) characterization. These traps appear as RTS noise during BTI characterization and need to be factored out in order to differentiate between transistor degradation caused by BTI and RTS.

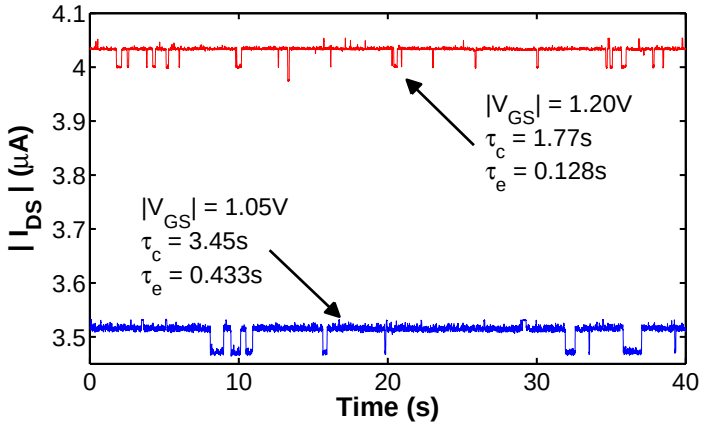


Figure 3.12: Drain current of PMOS SRAM transistor, with type III, under different gate biases.

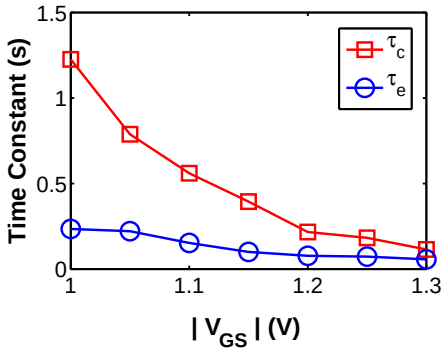


Figure 3.13: Type III trap dependence of τ_c and τ_e on gate bias.

3.2.3 Dynamic Behavior Response under Large-Bias Changes

The discussion of RTS dynamics so far has focussed on time constants of the trap under fixed-bias conditions. These characteristics might be sufficient for circuits where transistors are operating at a fixed-bias condition and are only subject to small signal stimulus, such as amplifiers. Other classes of circuits such as SRAM, comparators, oscillators, phase locked loops, and mixers undergo large-bias changes as part of the circuit operation. RTS dynamics under large-bias changes need to be well understood in order to properly estimate the impact of RTS on circuit operation. It is also possible to take advantage of this large-bias response behavior for minimizing the total low-frequency noise in the system [76].

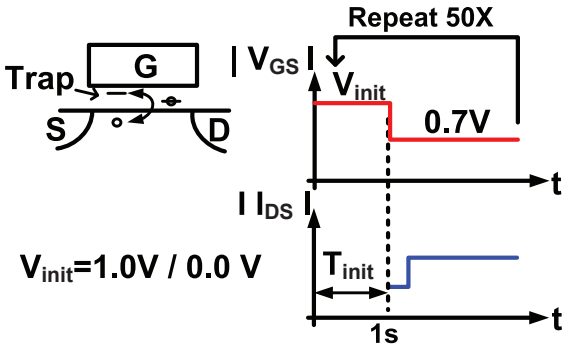


Figure 3.14: Cross-section of transistor with oxide trap as well as associated waveforms for characterizing the response of RTS to large-bias changes.

Figure 3.14 illustrates the waveforms for characterizing the large-bias response of oxide traps. The transistor is initially biased to a V_{init} voltage of either 1.0V or 0.0V for a duration of T_{init}. This initial bias condition forces trapping or de-trapping of the carriers in the oxide, depending on the trap type and V_{init}. The bias on the gate of the transistor is then changed to $V_{measure}$ (0.7V) and the drain current is monitored. The values of V_{init} are chosen to characterize the response of these traps corresponding to scenarios where the transistor is initially off (V_{init} = 0.0V) or on (V_{init} = 1.0V) prior to it being utilized in the circuit.

Figure 3.15(a) plots the single-measurement trace and average drain current measured from an SRAM *PD* NMOS transistor under this scheme. The single trace indicates a low current level state at the start of the measurement, corresponding to a trap being filled. The drain current eventually transitions up to a high level state corresponding to a carrier leaving an oxide trap. Assuming that these two current levels are caused by a single carrier occupying or leaving a trap, we can assign the low level to α of 100%, corresponding to the trap being filled, and the high current level to α of 0%, corresponding to the trap getting emptied. Analyzing the 50 run average, with α calibrated to the low and high current levels, reveals that the trap is always occupied immediately after the transition from V_{init} to $V_{measure}$ for more than 10 ms before the average trap occupancy decays exponentially to a steady-state level. Note that the final steady-state level is non-zero, indicating that the trap is active under $V_{measure}$ bias conditions, as is also obvious in the single trace. Figure 3.16 plots the single-measurement and average drain current extracted with V_{init} = 0.0V, measured from a different transistor. The V_{init} bias condition in this case forces an empty trap state (α = 0%) initially after the large-bias change. The trap eventually transitions to a new steady-state trap occupancy of 95% after more than 0.5 s.

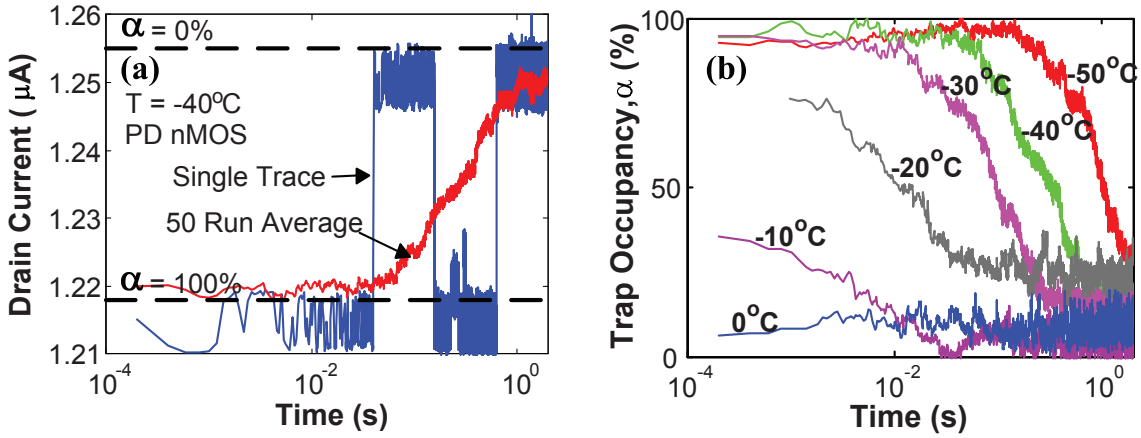


Figure 3.15: (a) Single trace and average response of a trap to large-bias change (V_{init} = 1.0V, $V_{measure}$ = 0.7V) (b) Large-bias response of trap at different temperatures.

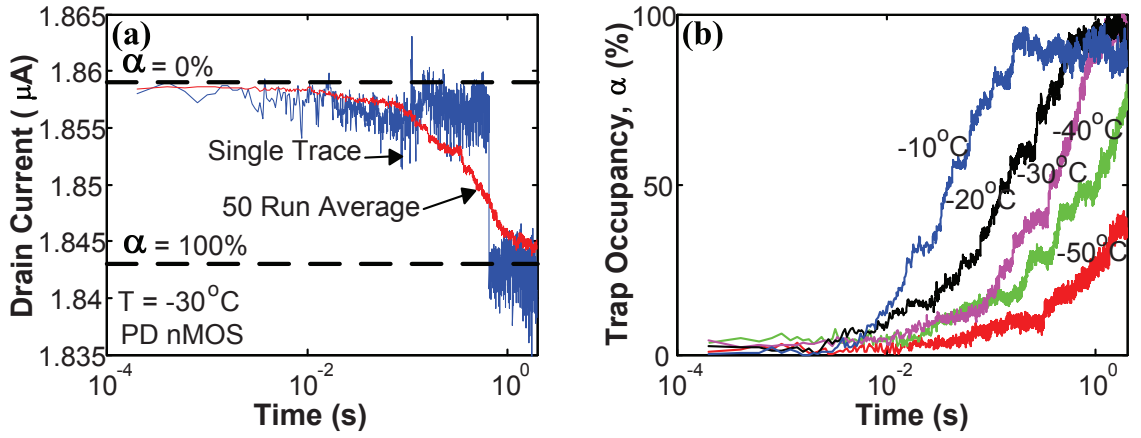


Figure 3.16: (a) Single trace and average response of a trap to large-bias change (V_{init} = 0.0V, $V_{measure}$ = 0.7V) (b) Large-bias response of trap at different temperatures.

Figures 3.15(b) and 3.16(b) illustrate the estimated trap occupancy corresponding to the respective large-bias changes, across different temperatures. The large-bias response time constant (τ_r) of the trap is strongly temperature dependent and is observed to decrease by 3 orders of magnitude as temperature is increased by 40°C (Figure 3.15(b). The trap occupancy appears to be constant at 0°C but it is likely that τ_r is too fast at this temperature for the measurement equipment to capture.

These results demonstrate that trap occupancy does not change instantaneously with bias changes [40]. The occupancy of these particular traps at initial and steady-state conditions

matches bias dependence of conventional traps (Figure 3.8). The time constant of the exponential decay of trap occupancy has been demonstrated to be related to the RTS time constants at $V_{measure}$ bias conditions, together with the initial and average trap occupancy, according to Equation 3.5 [26]. Time constants of RTS dynamics (τ_c and τ_e) can be extracted from large-bias response characteristics by measuring τ_r and $\alpha_{steady-state}$ and solving for the time constants using Equations 3.5 and 3.4.

$$\begin{aligned} \alpha(t) &= \alpha_{steady-state} + (\alpha_{t=0} - \alpha_{steady-state})e^{-\frac{t}{\tau_r}} \\ \frac{1}{\tau_r} &= \frac{1}{\tau_c} + \frac{1}{\tau_e} \end{aligned} \tag{3.5}$$

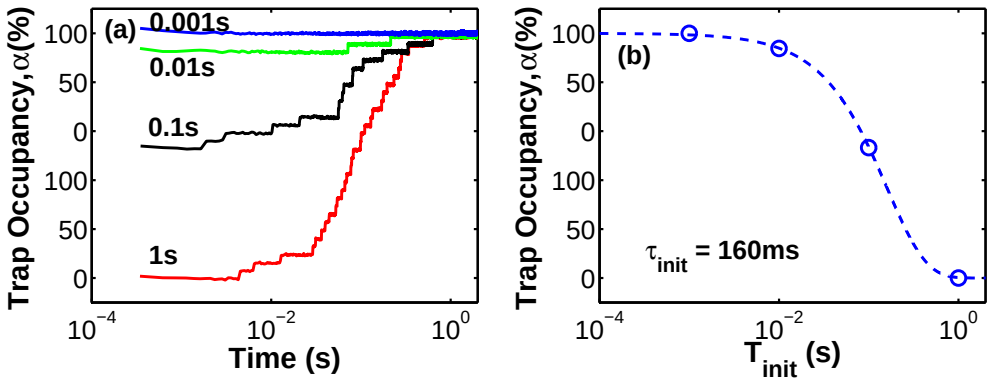


Figure 3.17: (a) Average response of a trap to large-bias change with different values of T_{init} ($V_{init} = 0.0$V, $V_{measure} = 0.7$V) (b) Estimated average response of the trap to large-bias change with $V_{init} = 0.7$V and $V_{measure} = 0.0$V.

Figure 3.17(a) plots the large-bias response of a trap measured with different initial bias durations (T_{init} ref. Figure 3.14). At a T_{init} of 1ms, the short pulse to 0.0V is insufficient to force emptying of the oxide trap. The average trap occupancy therefore appears to remain at 100% throughout the measurement period. The initial trap occupancy gradually decreases as T_{init} is increased and eventually saturates at 1 s. The dependence of the initial trap occupancy on T_{init} has important consequences when performing large-bias characterization of traps *i.e.* T_{init} needs to be long enough for the trap to converge to the steady-state condition corresponding to the V_{init} bias condition in order to observe the most significant large-bias trap response. This trap response characteristic has actually been exploited to estimate RTS dynamics at off-state bias conditions where it is infeasible to measure the drain current of the device [26, 46]. The first drain current measurement at $V_{measure}$ effectively samples the final trap occupancy state prior to the large-bias change, assuming that the delay till the first measurement is much shorter than the trap time constants. To perform this characterization, the large-bias characterization procedure is repeated at different T_{init}

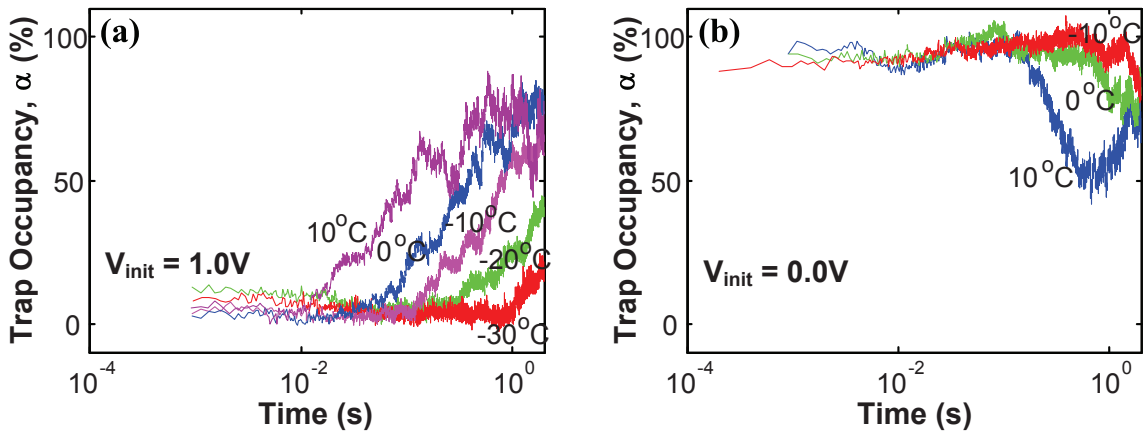


Figure 3.18: (a) Single trace and average response of a type-II trap to large-bias change (V_{init} = 1.0V, $V_{measure}$ = 0.7V) (b) (V_{init} = 0.0V, $V_{measure}$ = 0.7V)

conditions and the initial trap occupancy for each setting is recorded. Figure 3.17(b) plots the result of this procedure as well as a best-fit of the data-points to an exponentially decaying function. The time constant of this function is annotated on the plot.

Figure 3.18 plots the large-bias response of a type-II trap under different values of V_{init} and temperature. Contrary to conventional traps, the V_{init} bias of 1.0V forces de-trapping of the oxide trap. This is observed in the 0% initial trap occupancy observed in Figure 3.18(a). The trap occupancy eventually converges to a steady-state trap occupancy level with a characteristic time constant, corresponding to exponential decay. The opposite trend is observed under V_{init} = 0.0V bias conditions where the initial trap occupancy is at 100%.

3.2.4 Alternating-Bias Technique

RTS amplitude measurements are conventionally performed by measuring the drain current (I_{ds}) of the transistor under a constant gate bias [38, 53, 66]. Long measurement periods are required to observe RTS-related fluctuations caused by deep traps with long time constants. Figure 3.19 plots the drain current of a transistor exhibiting such a trap with a long time constant. In this example, the transistor current needs to be measured for more than 10 minutes before the trap that results in the worst-case RTS fluctuation in the drain current is observed. This makes it prohibitive to analyze a large population of transistors, to obtain statistics necessary for estimation of the properties of large SRAM arrays.

A measurement technique is therefore introduced to accelerate the oxide trapping and de-trapping processes by pulsing the gate bias to a high stress voltage (V_{stress}) prior to sampling the current, and subsequently to a negative voltage ($V_{accumulation}$) prior to sampling the current (Figure 3.20). This technique takes advantage of the large-bias trap response characteristics presented in this section. Gate biases of V_{stress} and $V_{accumulation}$ force trapping

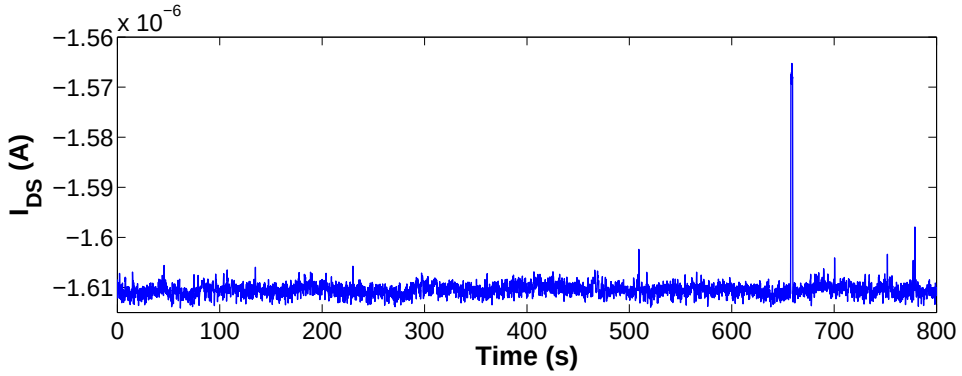


Figure 3.19: Measured drain current of SRAM *PU* PMOS transistor demonstrating RTS with long time constant.

and de-trapping of the oxide trap. The trap remains occupied/empty even though the stress voltage is removed because instantaneous trap occupancy converges to a new steady-state value as a decaying exponential [40]. Figure 3.21 plots the drain currents measured using this technique, demonstrating the 10x decrease in measurement time required to observe similar RTS fluctuation. It takes 32s for a trap to be occupied in the conventional technique. By applying an initial 0.5ms gate stress before measurement, the trap is occupied from the start of the experiment.

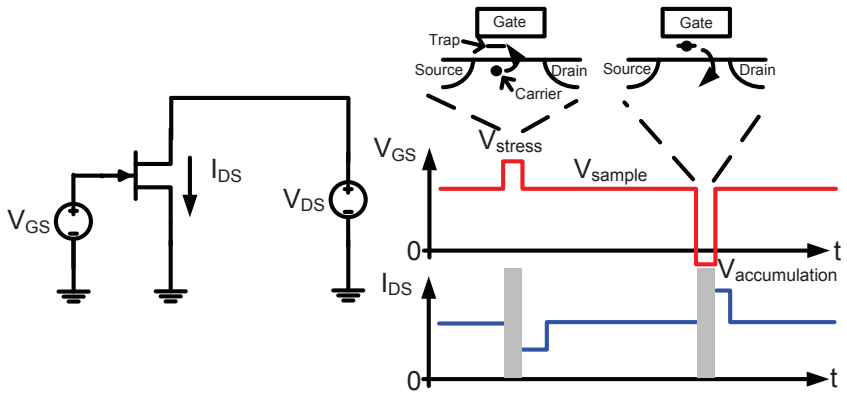


Figure 3.20: Waveforms of the alternating-bias technique.

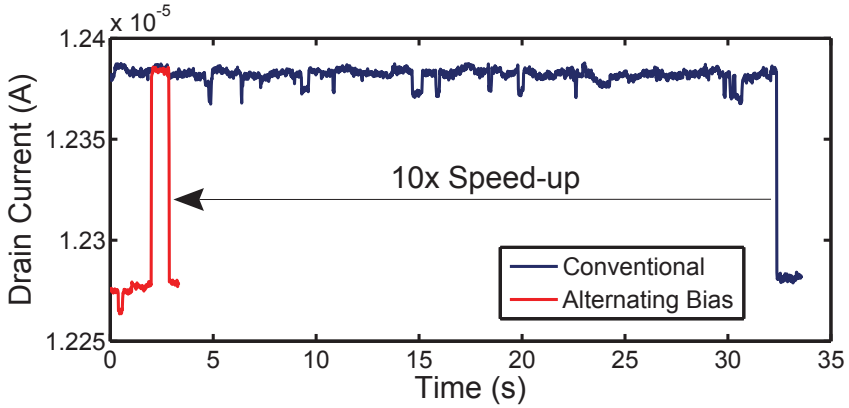


Figure 3.21: Drain current of the same transistor measured using conventional and alternating-bias techniques.

Figure 3.22 plots the cumulative density function (CDF) corresponding to the magnitude of the drain current fluctuations caused by RTS as extracted using the conventional measurement technique of applying a constant bias as well as the alternating-bias technique. The drain current of each transistor was measured for a fixed duration of 60s in each case to provide a fair comparison between both techniques. The CDF corresponding to the alternating-bias technique is clearly shifted towards the right, indicating that this technique observes larger magnitudes of drain current fluctuation from the same set of transistors, compared to the conventional technique. The largest difference in the statistical distribution is observed between the 50^{th} and 99^{th} percentile where a shift of up to 50% (100nA to 150nA is observed). This range of percentiles corresponds to the bulk of the distribution. Figure 3.23 plots the histograms corresponding to these two measurement techniques, together with fitted lognormal distributions, clearly demonstrating the differences in the bulk of the distributions. The bulk of a distribution plays an important role in estimating the statistical impact of RTS on circuit performance because it plays a larger role in determining the median of the final failure distribution due to the large population of samples present in the bulk. The conventional technique underestimates the magnitude of RTS fluctuations in the bulk of the distribution due to insufficient measurement period. Alternately, measurement period of the conventional technique can be increased while sacrificing the sample population that can be collected within a reasonable amount of time.

3.3 Amplitudes of Random Telegraph Signaling Noise

3.3.1 Background

RTS amplitudes have an equally important impact, relative to RTS dynamics, because this needs to be well-characterized in order to margin for this source of variability in circuit

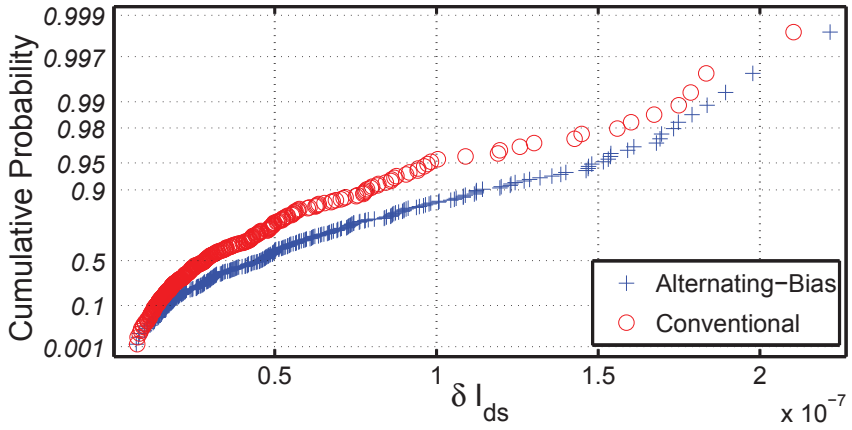


Figure 3.22: Gumbel plots of RTS drain current fluctuations measured using conventional and alternating-bias techniques with constant time.

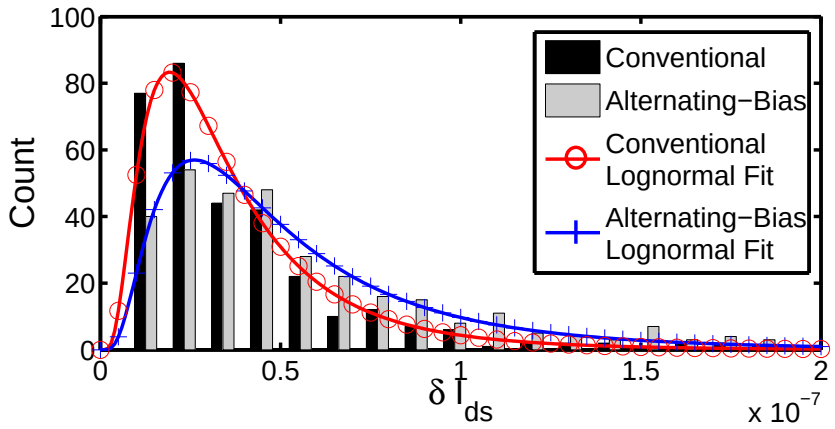


Figure 3.23: Histogram and lognormal distribution fits of RTS I_{ds} fluctuations measured using conventional and alternating-bias techniques with constant time.

design. This is particularly relevant in circuit design using highly scaled transistors (such as SRAM) because RTS amplitude is exacerbated by transistor scaling. More than 50% fluctuation in transistor drain current has been observed in highly scaled CMOS transistors [11]. Observation of RTS amplitude, fitted to a physics-based model, is also interesting because it can be correlated with RTS dynamics to extract trap parameters and can also be used as an in-situ probe of carrier transport characteristics [29]. Over the past five decades, several physics-based models have been proposed to model RTS amplitude. These models either attribute fluctuations in conductivity of the channel to fluctuations of the total number of carriers (ΔN) or mobility fluctuation ($\Delta\mu$) [45, 27]. More recently, unified

models combining these two theories have been proposed to reconcile the discrepancy between predictions and actual observation of RTS amplitude [29].

The classical number fluctuation theory [45] is based on the principle that the change in the conductivity of the channel, observed as a fluctuation in the drain current, is caused by carriers leaving the inversion layer to occupy traps, therefore resulting in a net reduction in current. Generally, charge in a trap is screened by charges on the gate, interface states, the inversion layer, and the depletion region. A change in the charge state of the trap therefore needs to be balanced by these other charges. The change in the inversion layer charge (Q_n) due to a change in the charge of a trap (Q_t) is modeled by Equation 3.6 [38]. ϕ_s is the surface potential while C_{ox}, C_{it}, and C_D are the capacitance per unit area of the oxide, interface states, and depletion layer respectively. In the strong inversion regime, the denominator is dominated by $\frac{\delta Q_n}{\delta \phi_s}$. $\frac{\delta Q_n}{\delta Q_t}$ is therefore -1 and this model predicts that the fluctuation in normalized drain current will be equal to the inverse of the number of carriers in the channel ($\frac{\Delta I_D}{I_D} = -\frac{1}{N}$). This however does not hold in the weak inversion regime.

$$\frac{\delta Q_n}{\delta Q_t} = \frac{\frac{\delta Q_n}{\delta \phi_s}}{C_{ox} + C_{it} + C_D - \frac{\delta Q_n}{\delta \phi_s}} \tag{3.6}$$

Ghibaudo *et al.*, equated the local change in the oxide charge due to trapping to a flatband voltage fluctuation, which was then used to derive Equation 3.7 which is a more generalized equation that is applicable to all regions of transistor operation (strong inversion, weak inversion, saturation regime, and linear regime). g_m is the transconductance while $\frac{d}{t_{ox}}$ is the relative depth of the trap into the oxide. It is interesting to note that representing RTS amplitude as an equivalent ΔV_{th}, by dividing the measured ΔI_D with the transconductance at the bias point [66, 53], is similar to the number fluctuation model. The extracted ΔV_{th} is equivalent to $\frac{q}{WLC_{ox}}\left(1 - \frac{d}{t_{ox}}\right)$ in Equation 3.7

$$\frac{\Delta I_D}{I_D} = \frac{g_m}{I_D}\frac{q}{WLC_{ox}}\left(1 - \frac{d}{t_{ox}}\right) \tag{3.7}$$

Validity of this model is evaluated by comparing actual RTS amplitudes extracted from 45 nm SRAM transistors with predictions of this model. Figure 3.24 plots normalized fluctuation in drain current due to RTS, extracted from NMOS and PMOS transistors as a function of different drain and gate biases. Equation 3.7 is fitted to the data at $|V_{DS}| = 0.1$ V to extract the coefficients, α_1 and α_2, for the NMOS and PMOS transistors respectively. The transconductance (g_m) and drain current (I_D) are characterized from the actual device at the respective bias conditions. This model is then used to estimate the RTS amplitude of these particular traps at $|V_{DS}| = 1.0$ V. Even though there exists a slight discrepancy between the model and measured data, the number fluctuation model appears to provide a good fit to the data regardless of whether the transistor is operating in the linear or saturation regime, especially when the transistor is in strong inversion. A large discrepancy however is observed when the transistor enters the weak inversion region. The number fluctuation

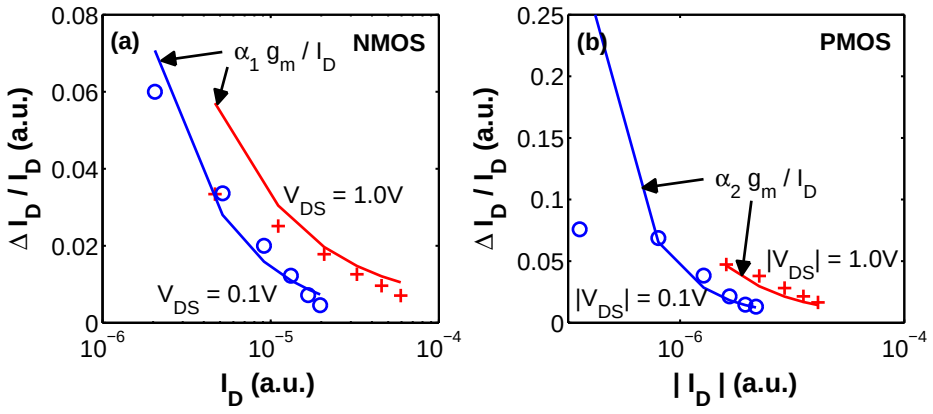


Figure 3.24: Bias dependence of RTS amplitude extracted from 45 nm SRAM (a) NMOS and (b) PMOS transistors. Solid lines correspond to number fluctuation model fitted to data at $V_{DS} = 0.1$V.

model over-estimates the magnitude of the actual RTS fluctuation in the weak inversion region. Furthermore, this model fails to capture the cross-over between the data-points corresponding to $V_{DS} = 0.1$ V and 1.0 V that is observed in Figure 3.24(a).

$$\left\langle \left(\frac{\Delta G}{G} \right)^2 \right\rangle = \frac{\alpha}{N} \frac{\Delta f}{f} \tag{3.8}$$

Mobility fluctuation models ($\Delta\mu$) are based on the empirical relationship in Equation 3.8, observed by Hooge [27]. G is the conductance, N the total number of mobile charge carriers in the sample and α is a dimensionless constant with the value $2\text{x}10^{-3}$. This empirical relationship was derived based on analysis of $1/f$ noise in homogeneous samples. This model has been applied to MOS transistors operating in the linear region, which is the closest approximation to a homogeneous sample, with good correlation observed in PMOS transistors with buried channels and poor correlation observed in NMOS transistors with surface channels [2]. These promising results however do not apply to current surface channel silicon PMOS transistors with p^+ polysilicon gates or even metal gates. This model also predicts a $1/N$ or $1/I_D$ dependence which is not observed in Figure 3.24 where the normalized RTS amplitude appears to be saturating to a constant value at weak inversion.

Correlated models have also been introduced in an attempt to provide a physical model for RTS amplitude that is applicable to both PMOS and NMOS transistors. The correlated model or unified model is based on the principle that capture and emission of a single carrier by an oxide trap will induce correlated fluctuations in channel carrier number and mobility [29]. The correlated mobility fluctuation is caused by Coulomb scattering of carriers in the channel by the oxide trap states. Normalized RTS amplitude caused by these two correlated sources is modeled using Equation 3.9. The first term corresponds to the number fluctuation

component while the second term corresponds to the correlated mobility fluctuation. The sign of the second component depends on whether the scattering center corresponding to the oxide trap is active when the trap is occupied (positive) or is neutralized by trap occupation (negative). The partial derivative $\frac{\delta N}{\delta N_t}$ is equal to the change in the number of carriers in the inversion layer due to a change in the number of trapped charges and is related to Equation 3.6. The partial derivative $\frac{\delta \mu}{\delta N_t}$ which models the dependence of the effective mobility on the number of traps is evaluated using Matthiesen's rule for effective mobility. The correlated model can be written as Equation 3.10 by evaluating the partial derivatives and substituting the $\frac{g_m}{I_D}$ form of the number fluctuation model, as proposed by Ghibaudo *et al* [22]. The α parameter is a Coulomb scattering coefficient that is also a function of carrier density.

$$\frac{\Delta I_D}{I_D} = \frac{1}{WL}\Big(\frac{1}{N}\frac{\delta N}{\delta N_t} \pm \frac{1}{\mu}\frac{\delta \mu}{\delta N_t}\Big) \tag{3.9}$$

$$\frac{\Delta I_D}{I_D} = \frac{1}{WL}\Big[\frac{g_m}{I_D}\frac{q}{C_{ox}}\Big(1 - \frac{d}{t_{ox}}\Big) \pm \alpha N_t\Big] \tag{3.10}$$

Figure 3.25 plots RTS amplitude measurements extracted from a 45 nm NMOS transistor, fitted to the correlated model. The scattering coefficient (Figure 3.25(a)) was extracted by normalizing the measured RTS amplitude with $\frac{g_m}{I_D}$ extracted from the actual transistors while N and μ were estimated from a SPICE model, calibrated to the actual transistor. The various data points were measured across an entire range of drain and gate biases ranging from 0.1 V to 1.1 V. As expected, α is strongly dependent on electron density, however, actual values of α observed are more than an order of magnitude larger than that is expected based on theoretical estimations. At low inversion, theory predicts that α saturates to approximately 3 x 10^{-15} Vs due to majority carrier screening [77]. These larger than expected values might be due to percolation paths in the channel of the transistor, due to discrete doping profiles [6]. Nevertheless, there is clearly a need to model the carrier density dependence of α to obtain a good estimate of RTS amplitude, using the correlated model. Hung *et al.* proposes modeling this dependence as $\alpha_0 + \alpha_1 log(N)$ which was justified based on good fitting observed by the authors within a small range of N [29]. Figure 3.25(a) plots the best-fit of this model, as well as a $1/\sqrt{N}$ model proposed by Vandamme *et al.* The $log(N)$ model appears to best match the general trend observed in the data but there exists a large dispersion in the data-points, centered around the $log(N)$ trend. This dispersion could possibly be caused by phonon and surface roughness scattering dominating over Coulomb scattering at bias conditions with high vertical electric fields, which is not being modeled [44].

Figure 3.25(b) plots the RTS amplitude estimated using these two models of α across different gate biases and two drain biases. The correlated model using the $log(N)$ model for α appears to fit the data corresponding to $V_{DS} = 0.1V$ relatively well but grossly over-estimates the amplitude corresponding to the $V_{DS} = 1.0V$ bias condition. A better fit can actually be obtained by fitting α to an empirical $1/N^2$ function but this makes the model empirical resulting in the inability to extract physical parameters from RTS amplitude measurements.

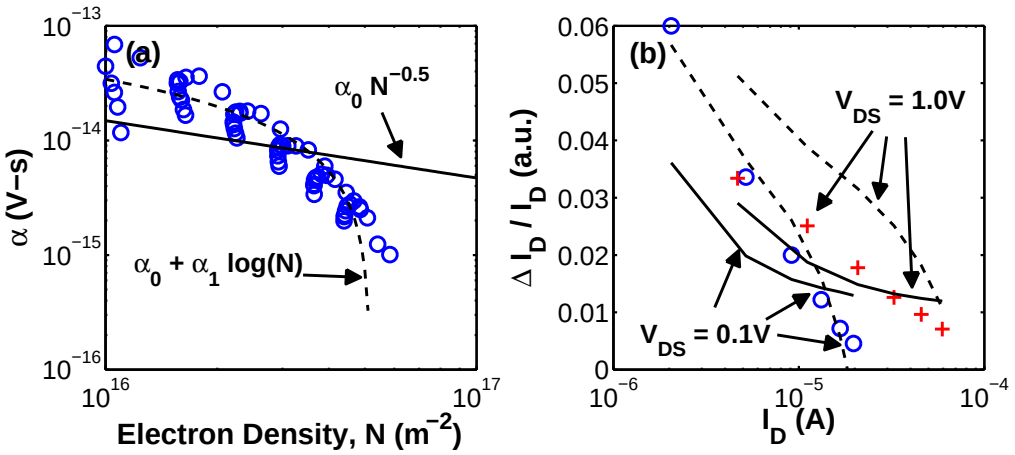


Figure 3.25: 45 nm NMOS RTS amplitude fitted to correlated model. (a) Extracted scattering co-efficients as a function of electron density and fitted models. (b) Measured RTS amplitude across different biases and fitted models. Solid lines and dashed lines correspond to $1/\sqrt{N}$ and $log(N)$ models for scattering co-efficient, respectively.

3.3.2 Empirical Bias-Dependent Model

Physical models of RTS amplitude such as the number fluctuation model and correlated model, are at most only able to fit actual data under limited bias conditions. In this section, an empirical model is proposed that is able to capture the bias dependence of RTS amplitude across a wider range of bias conditions. This model is based on assigning each trap an equivalent ΔV_{th} that is bias-dependent. Once the bias-dependence of ΔV_{th} is modeled, a statistical model is derived by assigning a statistical distribution to ΔV_{th} based on RTS amplitude distributions sampled from multiple transistors.

Figure 3.26 plots measured ΔV_{th} due to RTS measured from PMOS and NMOS transistor at a constant gate bias. ΔV_{th} was extracted from measurements of drain current fluctuation (ΔI_D) by dividing measured ΔI_D with the g_m characterized from the actual transistor. Note that the magnitude of ΔV_{th} extracted using this technique is an order of magnitude larger than what is expected from $\frac{q}{WLC_{ox}}$, derived for a continuously doped channel, which is attributed to the impact of discrete doping in the channel creating percolation paths of carriers that are more susceptible to Coulomb scattering due to trap states [6]. A slight dependence on V_{DS} is observed in the data. This might be due to the drain terminal modulating the channel pinch-off region relative to the lateral location of the trap and therefore changing the influence of the trap state on channel conductivity. This V_{DS} dependence is modeled in this empirical model using a linear function for both PMOS and NMOS transistors.

Figure 3.27 plots the ΔV_{th} measured from the NMOS transistor, normalized by a linear fit to the V_{DS} dependence of the data, as a function of electron density (N). Data-points

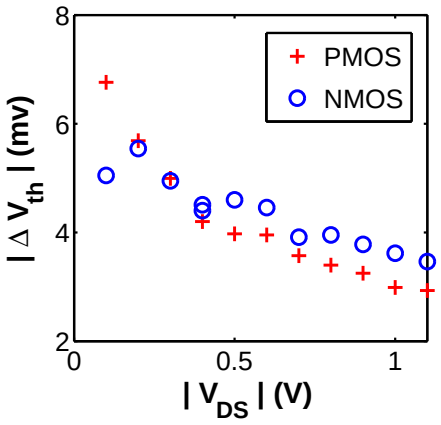


Figure 3.26: V_{DS} dependence of ΔV_{th} for a trap measured from PMOS and NMOS transistors at constant gate bias ($|V_{GS}| = 0.8$ V).

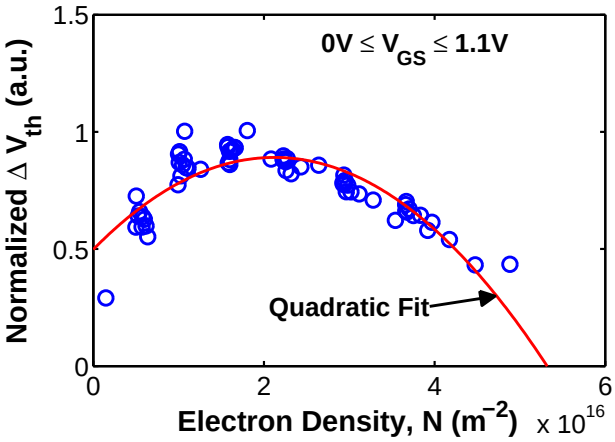


Figure 3.27: Residual ΔV_{th} after normalization by linear fit of V_{DS} dependence, as a function of electron density.

corresponding to all V_{GS} biases are plotted on the graph. The normalized ΔV_{th} has a strong dependence on electron density in the channel. This dependence could be due to different scattering mechanisms coming into play across the bias conditions [44]. It could also be due to the bias-dependence of the transconductance of the transistor (g_m) since all RTS amplitude measurements are normalized by g_m. A quadratic fit to the data is also plotted in Figure 3.27, showing a good fit to the data. The carrier density dependence of the residual V_{th} is therefore modeled as a quadratic function in the empirical model.

Equation 3.11 summarizes the proposed bias-dependent empirical model that is valid for both NMOS and PMOS transistors (taking the absolute value of the variables). The

parameters, p_0 and p_1, specify the quadratic dependence of ΔV_{th} on carrier density and need to be calibrated separately for NMOS and PMOS transistors. ΔV_{th0} specifies the nominal magnitude of the trap and is dependent on the location of the trap with respect to the dopant configuration of each instance of a transistor, due to the percolation path effect. Statistical distributions of ΔV_{th} due to RTS, extracted at small $|V_{DS}|$ and $|V_{GS}| \approx V_{DD}/2$ such as reported in [46, 53, 66], can be used directly as this parameter for statistical simulation. A fixed V_{DS} slope of 1.1×10^{-3} is used as this corresponds to the V_{DS} dependence observed in most of the RTS samples.

$$\Delta V_{th} = \Big(\Delta V_{th0} - 1.1 \times 10^{-3} V_{DS}\Big) \times \Big(p_0 N^2 + p_1 N + 1\Big) V \tag{3.11}$$

Figure 3.28 illustrates the fitting of the empirical model to RTS amplitude in NMOS and PMOS samples. The empirical model is able to fit the measurements across a broad range of bias conditions, spanning strong inversion to weak inversion and also linear mode to saturation mode of operation. Although some points do not fall exactly on the solid lines, the solid lines still track the general trend of the measured results much better than any of the physical models such as number fluctuation models or correlated models. The results in Figure 3.28 only cover a limited range of drain currents down to a few μA. This small range is due to the limited capability of the measurement equipment. To verify that this empirical model is self-consistent across the full range of transistor operation, RTS amplitudes are simulated using the empirical model, based on transistor parameters (g_m, I_D, N) obtained from SPICE models (Figure 3.28). Figure 3.29 plots results of this analysis. Normalized fluctuation in drain current saturates to fixed levels at small I_D which matches the characteristics observed through 3D atomistic simulation of RTS amplitude [6]. Furthermore, the cross-over of the lines corresponding to different drain biases, a characteristic that is observed in measured data (Figure 3.24), is also observed in the simulated RTS amplitudes.

3.4 Impact of Random Telegraph Signal on Static SRAM margins

This chapter has thus far focussed on the dynamics and amplitude of a single trap. While this gives much insight into the physics of this phenomena, the main reason RTS is presently a big concern in circuit design is due to the large spread in the statistical distribution of RTS noise amplitude measured from transistors with similar dimensions. This section presents statistical distributions of RTS amplitude measured from transistors used in a 45 nm SRAM bitcell. It then goes on further to study the impact of RTS in the 6 transistors used in a conventional SRAM bitcell on overall SRAM margins.

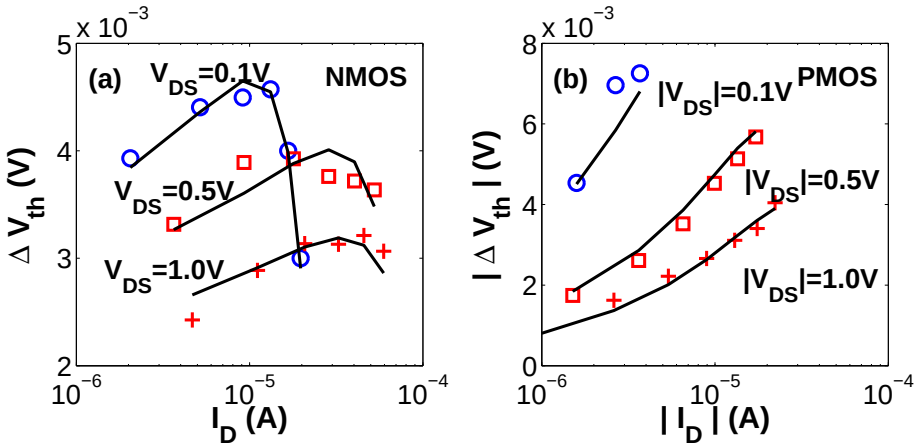


Figure 3.28: Equivalent ΔV_{th} fluctuation due to RTS across gate and drain biases extracted from (a) NMOS and (b) PMOS SRAM transistors. Solid lines represent the empirical model fitted to the data.

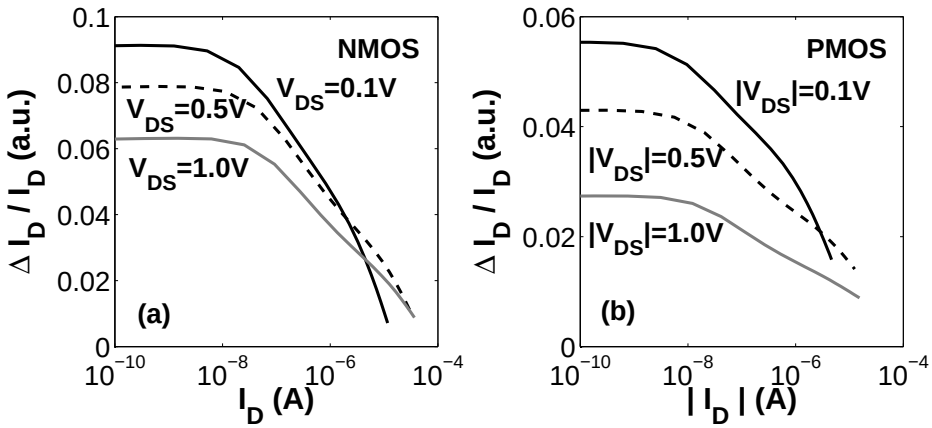


Figure 3.29: RTS amplitude for (a) NMOS and (b) PMOS transistors estimate using SPICE models extended with empirical RTS amplitude model.

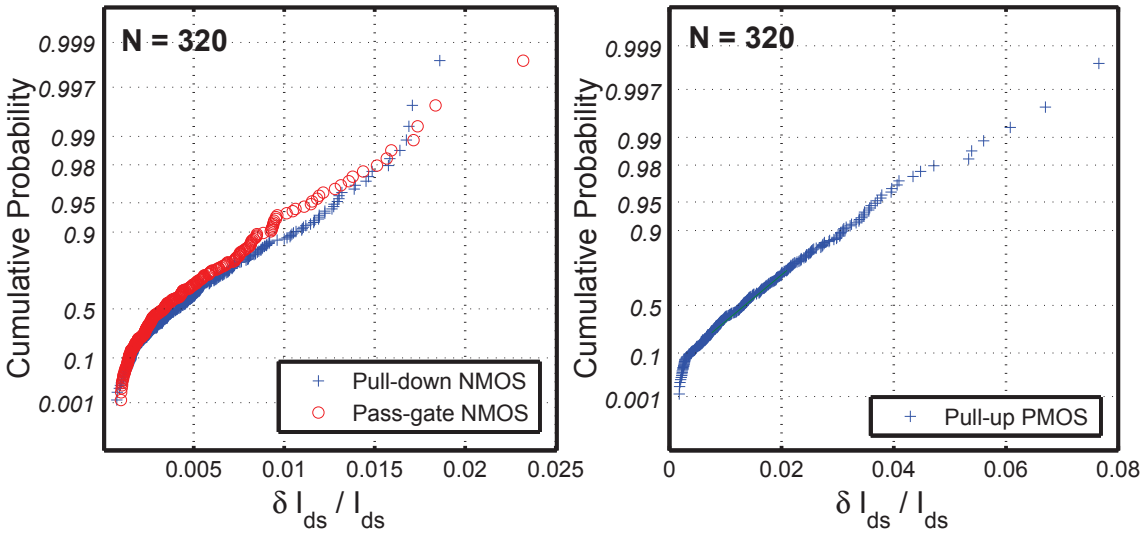


Figure 3.30: Gumbel plots of normalized RTS fluctuations in drain currents, measured from transistors in the padded-out SRAM cells.

3.4.1 45 nm SRAM Transistors

Figure 3.30 plots statistical distributions of RTS amplitudes that are measured from the pull-up (PU) PMOS, pull-down (PD) NMOS, and pass-gate (PG) NMOS transistor in an SRAM bitcell. These distributions were characterized at $|V_{GS}| = 0.7$ V and $|V_{DS}| = 50$ mV using the alternating-bias technique. This gate bias condition was chosen because it corresponds to critical bias conditions in an SRAM bitcell that determines the read and write margins. The transistors were biased into the linear region ($|V_{DS}| = 50$ mV) because this results in largest drain current fluctuation, which improves the signal to noise ratio of the measurement. SRAM bitcells are usually optimized for both read and write margins, resulting in the pulll-up transistor having the smallest channel area, followed by the pass-gate and the pull-down transistor. Figure 3.30(a) demonstrates that the pull-down transistors exhibit slightly larger RTS amplitudes compared to the pass-gate transistors even though the larger effective area transistor should have a smaller fluctuation. A similar trend is also reported in [65]. This discrepancy is partially due to a 6% higher mobility in the pass-gate devices compared to the pull-down devices, due to layout dependent effects. This higher mobility reduces the relative contribution of a trap in a pass-gate transistor as is predicted using the correlated model (Equation 3.9).

The $\geq$ 4x difference in RTS amplitudes between PMOS and NMOS devices can only be partially explained by the smaller channel area of the PMOS device, relative to the NMOS devices. Boutchacha *et al.* observed larger scattering coefficients in PMOS transistors compared to NMOS transistors which can result in a larger correlated mobility fluctuation

in PMOS transistors[9]. The transition to P^+ poly-silicon gate electrodes for PMOS and then to High-κ metal gates with work-function tuned specifically for PMOS transistors, brings the channel inversion-layer centroid closer to the surface which increases the Coulomb scattering of trapped charge on the channel carriers. Historically, buried channel PMOS transistors exhibited less RTS amplitude relative to surface channel NMOS transistors of similar dimensions, leading to PMOS transistors being preferred for critical transistors in circuit topologies where reduction of low-frequency noise is critical [25]. These results indicate that NMOS transistors should be used in circuits utilizing nanoscale transistors where minimizing low-frequency noise is the main priority.

3.4.2 Static Write Margin

The I_{write} metric, derived from write N-curves [13] is used to characterize write margin fluctuation of the SRAM cells caused by RTS. Figure 3.31 illustrates the schematic for I_{write} characterization while Figure 3.32 plots the various currents flowing through the respective transistors that contribute to the write margin. This static write margin is characterized by sweeping the internal node, corresponding to the half of the SRAM bitcell that is being written to, using a voltage source while monitoring the current flowing through this voltage source (labelled N-curve in Figure 3.31). The process of sweeping this internal node from V_{DD} to 0 V emulates the write operation, as a new value is being written into the bitcell. The I_{write} metric corresponds to the minimum point of the N-curve after the peak. It measures the relative strengths of PG3 and PD1 compared to PU5 during the critical phase of the write operation. An SRAM cell requires a positive I_{write} to be writeable. This current-based metric is favored over voltage-based metrics, such as wordline and bitline write metrics [24], because the measured margin fluctuations can be easily correlated with drain current fluctuations observed in the transistors. Figure 3.33 plots the RTS in I_{write} and its constituents, measured from one cell. Although $PU5$ contributes a small amount to the N-curve current, its RTS is reflected in I_{write} because RTS amplitudes are much larger for the PMOS transistors than for the NMOS transistors (Figure 3.30). Furthermore, $|V_{DS}|$ of $PU5$ during the critical region where I_{write} is defined is smaller than V_{DS} of $PG3$. Smaller $|V_{DS}|$ bias causes a larger fractional change in current due to RTS ($\Delta I_D/I_D$) that also increases the RTS noise contribution from $PU5$. These results indicate that RTS in SRAM write margin is dependent on both bias and RTS in multiple transistors, and requires a more accurate model than fixed shifts in V_{th} [66, 53] or single transistor RTS.

Fail Bit Rate (FBR) Estimation

A statistical model of SRAM write failure is developed in order to estimate the impact of RTS on write margin in large arrays. This model is calibrated based on characterization of I_{write} from padded out SRAM bitcells implemented in a 45 nm testchip [24]. Worst-case fluctuations in I_{write} are extracted by applying stress voltages to the pass-gate and pull-up

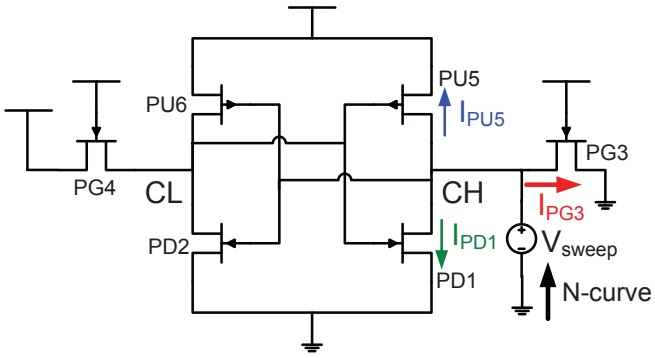


Figure 3.31: Schematic of I_{write} measurement. The N-curve is the sum of the currents flowing out of the internal node.

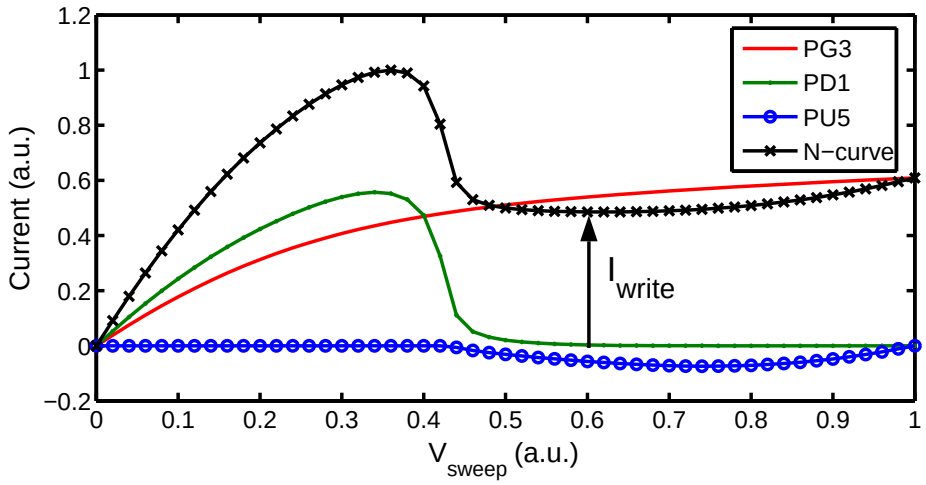


Figure 3.32: Currents contributing to I_{write}.

transistors before measurement. Multiple N-curve sweeps were measured from each SRAM bitcell to characterize the nominal value of I_{write} as well as δI_{write} as illustrated in Figure 3.34. Figure 3.35 plots the statistical distributions of nominal I_{write} and RTS measured at two operating voltages. Measured nominal I_{write} data is fitted to a normal distribution while RTS fluctuation data is fitted to a hybrid distribution (Figure 3.36). This hybrid distribution, which consists of a lognormal distribution for the bulk of the distribution and a generalized pareto distribution (GPD) for the tail of the distribution, allows accurate modeling of the entire distribution. The probability density function (f_{hybrid}) corresponding to this hybrid distribution is listed in Equation 3.12. A threshold (x_0) is defined where points less than this threshold are modeled using the lognormal distribution ($f_{lognornal}$) with parameters

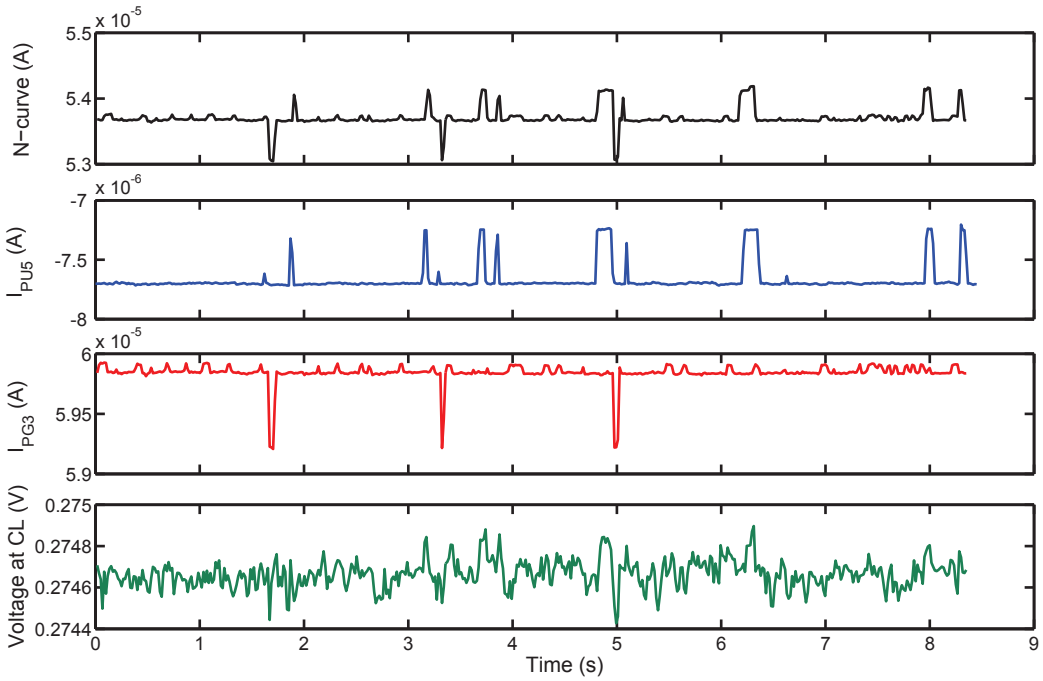


Figure 3.33: Currents and voltages measured from padded-out SRAM cell with RTS. RTS characteristics of the N-curve are influenced by RTS in both PU5 and PG3. Voltage fluctuation at node CL is minimal due to the low impedance of this node.

$\mu_{logn}, \sigma_{logn}$, while points larger or equal to the threshold, are modelled as a generalized pareto distribution (f_{GPD}) with parameters $\mu_{GPD}, \sigma_{GPD}, \xi_{GPD}$. f_{GPD} is normalized by the survivor function $(1 - F_{lognormal}(x_0))$ to adjust for the fact that the GPD distribution is only modeling the tail of the distribution above the threshold x_0.

$$
\begin{aligned}
&f_{hybrid}(x; \mu_{logn}, \sigma_{logn}, \mu_{logn}, \sigma_{logn}, x_0) \\
&= \begin{cases} f_{lognormal}(x; \mu_{logn}, \sigma_{logn}) & \text{if } x < x_0 \\ f_{GPD}(x; \mu_{GPD}, \sigma_{GPD}, \xi_{GPD})(1 - F_{lognormal}(x_0; \mu_{logn}, \sigma_{logn})) & \text{if } x \geq x_0 \end{cases} \\
&f_{lognormal}(x; \mu_{logn}, \sigma_{logn}) = \frac{1}{x\sqrt{2\pi\sigma_{logn}^2}} e^{-\frac{(\ln x - \mu_{logn})^2}{2\sigma_{logn}^2}} \\
&F_{lognormal}(x_0; \mu_{logn}, \sigma_{logn}) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left[\frac{\ln x - \mu_{logn}}{\sqrt{2\sigma_{logn}^2}}\right] \\
&f_{GPD}(x; \mu_{GPD}, \sigma_{GPD}, \xi_{GPD}) = \frac{1}{\sigma_{GPD}}(1 + \xi_{GPD}(\frac{x - \mu_{GPD}}{\sigma_{GPD}}))^{-(1/\xi_{GPD}+1)}
\end{aligned}
\tag{3.12}
$$

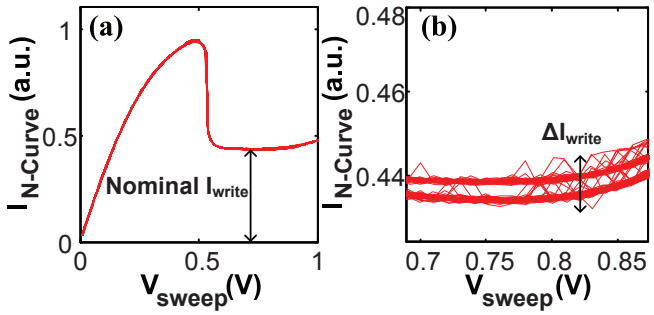


Figure 3.34: N-curves measured from multiple sweeps demonstrating the technique for extracting (a) nominal I_{write} and (b) RTS fluctuation in I_{write} due to RTS.

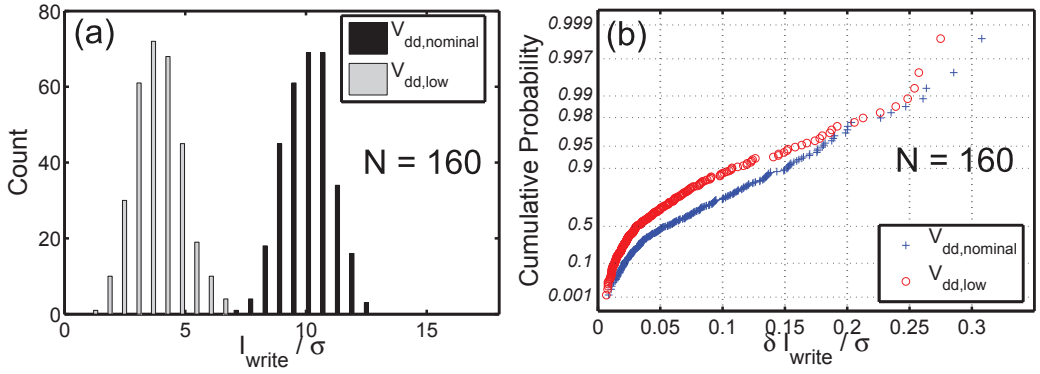


Figure 3.35: (a) Histogram of nominal I_{write}. (b) Gumbel plots of I_{write} RTS fluctuation. The Maximum RTS amplitude normalized to σI_{write} at each operating voltage does not change significantly, although the shape of the distribution changes.

Figure 3.37 plots the correlation between the nominal I_{write} and RTS fluctuation, measured at two voltages. The RTS amplitude appears to be uncorrelated with nominal I_{write} values which indicates that nominal I_{write} and δI_{write} can be modeled using two independent distributions. The joint probability density function (PDF) is then obtained from these two distributions by simple multiplication. Figure 3.38 plots the numerical joint PDF obtained from these two distributions, represented as contours of equal probability. Write failure occurs when $I_{write} - \delta I_{write} \leq 0$. This failure region is demarcated by a write failure contour ($I_{write} = 0$), illustrated in Figure 3.38. The most probable failure point (MPFP) corresponds to the point in the design space where the first failure is likely to occur. This point is numerically evaluated by tracing the write failure contour and finding the point with the highest probability. The FBR of the SRAM is calculated by integrating the joint PDF up to the boundary defined by the probability corresponding to the MPFP.

Figures 3.38 and 3.39 plot the joint PDFs, estimated based on measured I_{write} at nominal

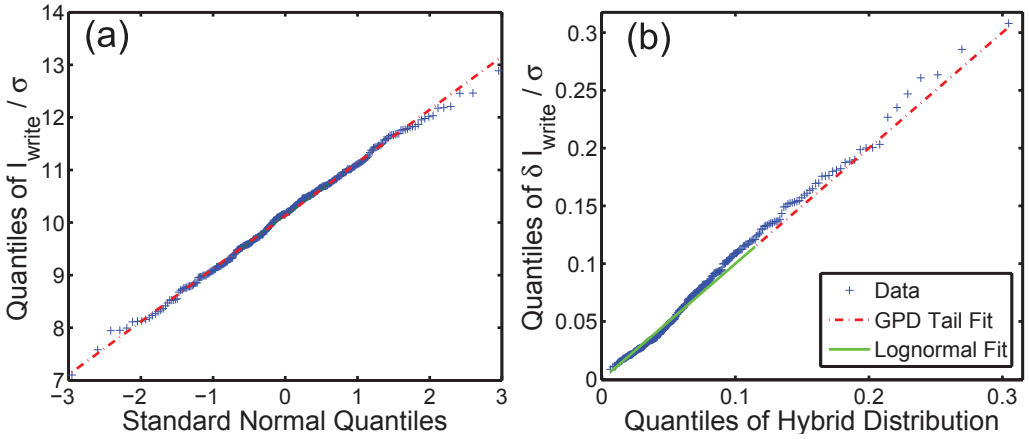


Figure 3.36: (a) Quantile-quantile plot of I_{write} with normal distribution fit. (b) Quantile-quantile plot of RTS fluctuation in I_{write} with hybrid distribution fit.

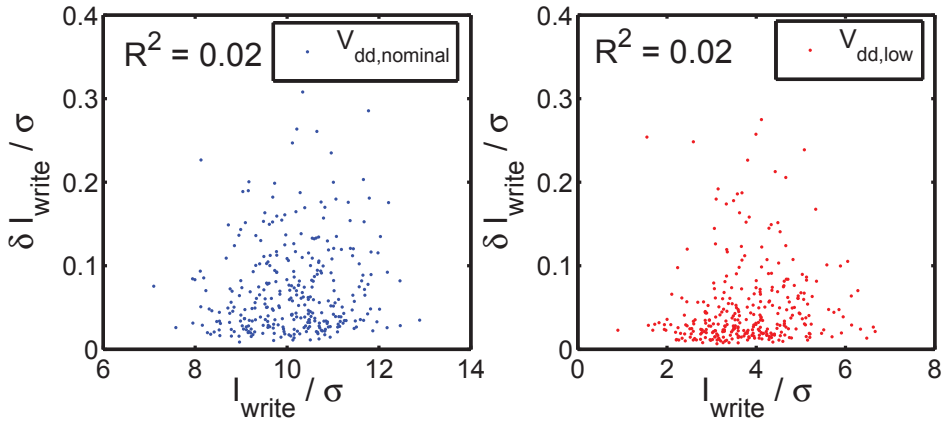


Figure 3.37: Scatter plots of measured nominal I_{write} and RTS fluctuation at nominal and low operating voltages.

and low operating voltages. The MPFP of both operating conditions occurs within the same region, characterized by low nominal I_{write} and small RTS fluctuation despite the different shapes of the joint PDFs. The proximity of the MPFP to small values of RTS fluctuation highlights the importance of modeling an accurate RTS fluctuation distribution in the bulk. The alternating-bias technique provides this accurate model.

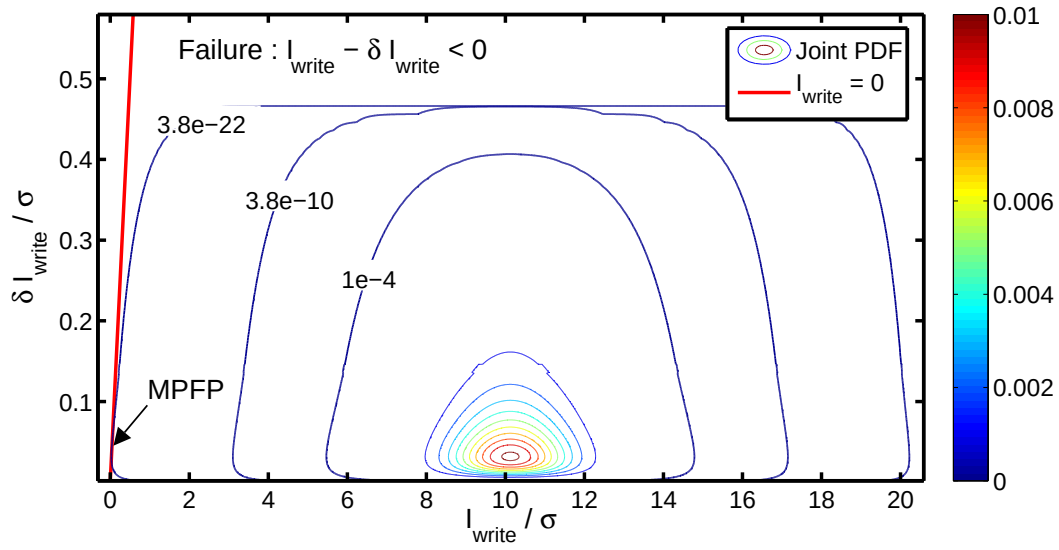


Figure 3.38: Joint probability density function of nominal I_{write} and RTS fluctuation at nominal V_{dd}.

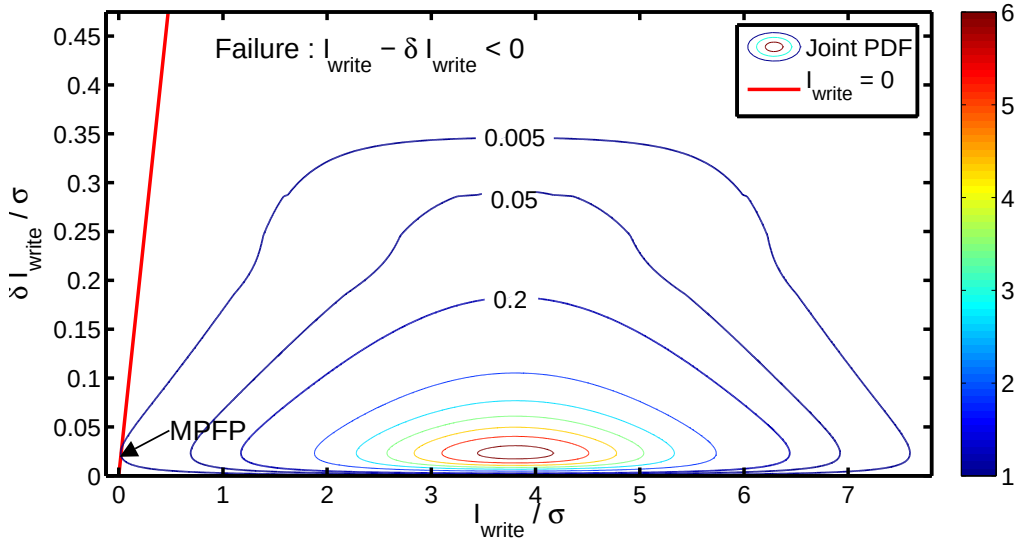


Figure 3.39: Joint probability density function of nominal I_{write} and RTS fluctuation at low V_{dd}.

FBR and V_{min} Degradation

The FBR of the SRAM bitcell is estimated at different voltages based on statistical distributions fitted to measured data (Figure 3.40). FBR increases by 0.6σ at low voltages and 0.2σ at nominal operating voltage. V_{min} is defined as the minimum cell supply voltage for performing a particular SRAM operation. V_{min} degradation for write operation is estimated from Fig. 3.40 by observing the increase in V_{dd} required to maintain a similar FBR. V_{min} is degraded by less than 50mV for small arrays (4σ) and is not significant in large arrays. The FBR degradation by RTS at the nominal voltage is minimal even in large SRAM arrays (10σ) because the MPFP is dominated by the bulk of the RTS fluctuation distributions rather than the tail.

The estimated V_{min} degradation corresponding to a 64kb SRAM array is experimentally verified by measuring V_{min} fluctuation in an actual array. V_{min} of the entire 64kb array is first sampled. 100 bitcells with worst V_{min} are then identified from this initial sample. V_{min} of these 100 bitcells is then measured continuously to identify the worst case V_{min} fluctuation. The minimum (nominal) and maximum V_{min} values of this subset is measured and illustrated in Fig. 3.41 as fail bit count. Fail bit count is defined as the number of cells with V_{min} greater than a corresponding voltage. Extrapolation of fail bit count to 10^0 estimates the largest V_{min} in the array. Assuming that these fluctuations in V_{min} characterized from the SRAM array are due to RTS, the measured results (0.04 a.u.) matches V_{min} degradation estimated numerically. Mismatch in the absolute V_{min} values between these two figures is caused by layout-induced differences between the 64kb SRAM arrays and the SRAM macros used for I_{write} measurements [24]. Statistical analysis confirmed by measurements therefore indicates that V_{min} degradation due to RTS is less than 50 mV, even in the presence of large RTS fluctuations.

3.5 Impact of Random Telegraph Signal on SRAM Dynamic Stability

SRAM arrays are typically accessed at high frequencies ranging from 10's of MHz to a few GHz. This section analyzes the impact of low frequency RTS noise on SRAM circuits operating at high speed [69]. Figure 3.42 presents measurements of multiple N-curve write margin sweeps at two different temperatures. Each sweep is performed as fast as possible, which is limited by the sampling speed of semiconductor parametric analyzers (typically within the range of a few kHz). The fluctuation in measured $I_{N-curve}$ observed at 30^{o}C indicates the presence of a trap with time constants faster than the measurement period of each N-curve sweep. The measurement is repeated at a colder temperature (-40^{o}C) to slow down the time constants of this trap and to emulate the expected behavior of the bitcell in the case where the N-curve sweeps are actually faster than the time constants of the trap (as is the case under normal SRAM operating frequencies). In this case, the trap stays in a

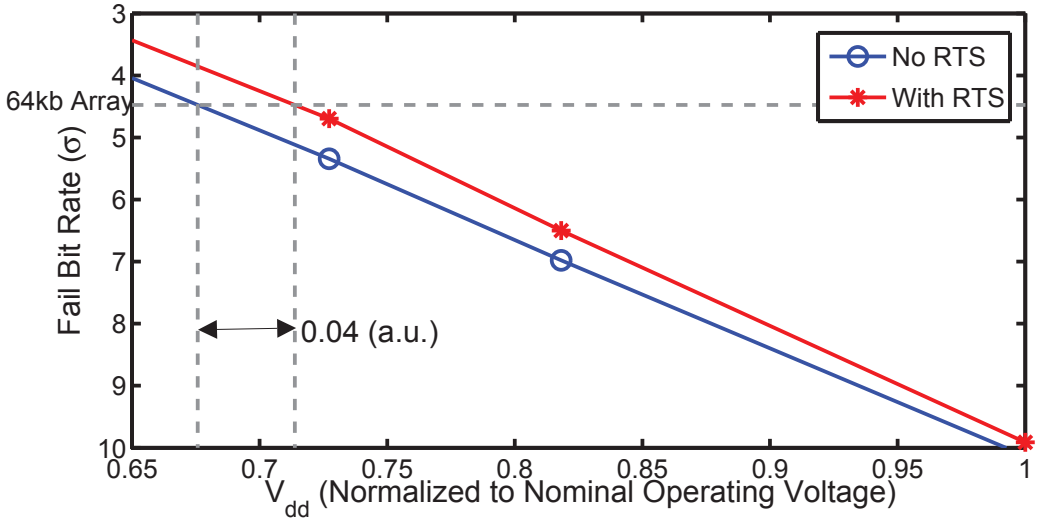


Figure 3.40: FBR of SRAM at different voltages.

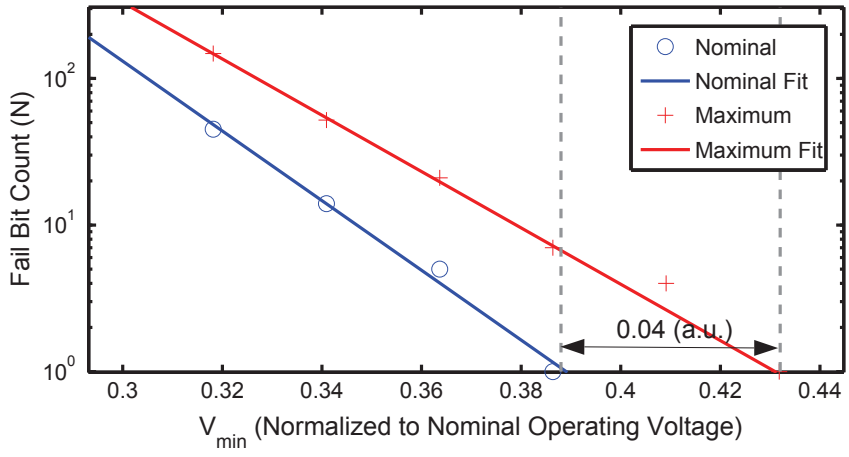


Figure 3.41: Measured fail bit count of a 64kb SRAM array.

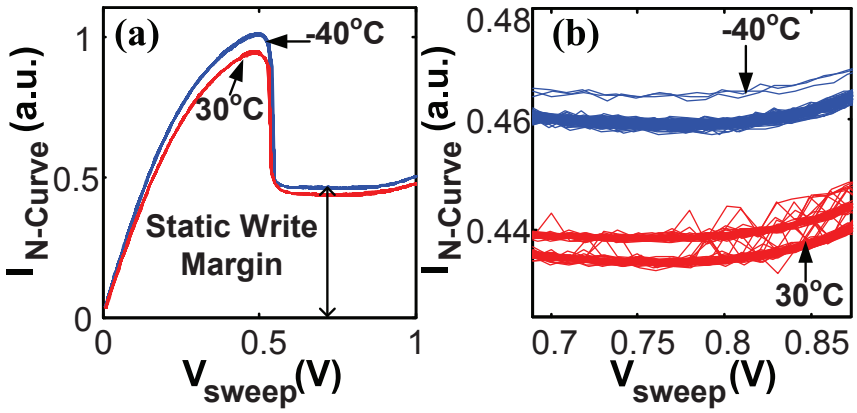


Figure 3.42: (a) High-speed N-curve sweeps of SRAM write margin at two temperatures. (b) Zoomed-in view of region defining static SRAM write margin.

fixed state throughout the duration of the measurement. RTS is therefore expected to result in hysteretic behavior in SRAM dynamic stability.

RTS dynamics are strongly dependent on bias conditions. Furthermore, the instantaneous trap occupancy is dependent on previous bias conditions. The dynamic stability of SRAM bitcells is analyzed using different access patterns typically encountered under normal SRAM access (Figure 3.43) to apply different combinations of large-bias changes on the SRAM transistors. The initial preset phase of dynamic stability characterization is followed by a relatively long period of at least 100 ms to let the traps in the respective transistors settle to their new steady-state condition. T_{write} and T_{access} are characterized immediately after this hold period in the case of single-write and single-read. For write-after-write or read-after-write access patterns, the first write operation is performed, followed by a relaxation period of T_{relax} before the final write or read operation is performed. T_{relax} is minimized (down to 200 ns) to prevent the traps from converging to a new steady-state condition before T_{write} or T_{access} corresponding to the final operation is characterized. Even though it might be illogical for a computer to perform a write-after-write operation back to back on the same SRAM bitcell, it is reasonable to expect these two operations to be spaced apart by 200 ns which corresponds to a few hundred clock cycles in modern microprocessors. The read-after-read access pattern performs N read access operations on the bitcell, spaced apart by T_{cycle}, prior to the final read operation where T_{access} is characterized.

Specific bitcells, with large RTS identified only in a single transistor, were selected to study the impact of RTS in a specific transistor on dynamic stability. The impact of RTS in these specific transistors on T_{access} and T_{write} will be presented next with reference to the schematics in Figure 3.44.

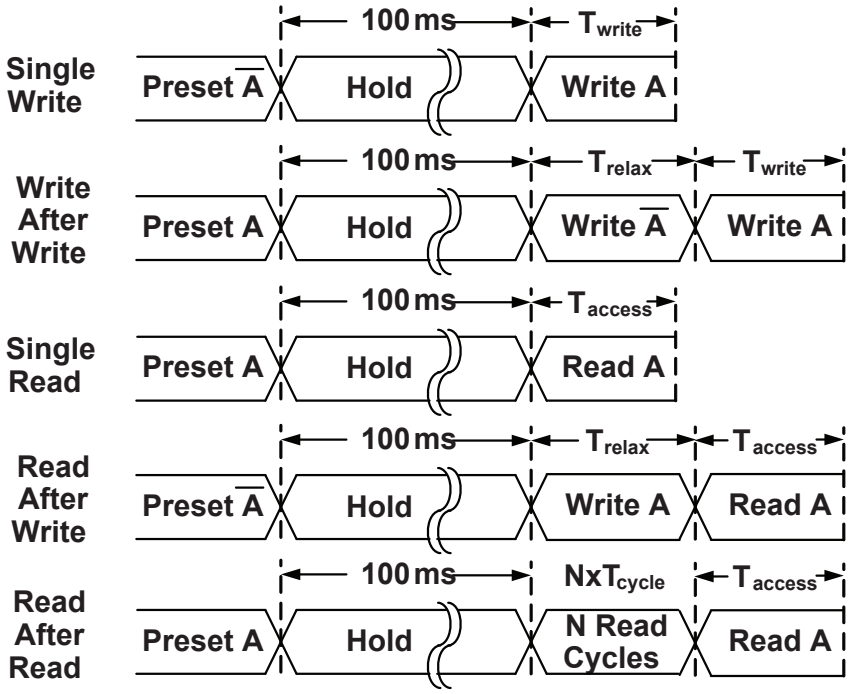


Figure 3.43: SRAM access patterns for evaluating the impact of RTS on dynamic read and write ability.

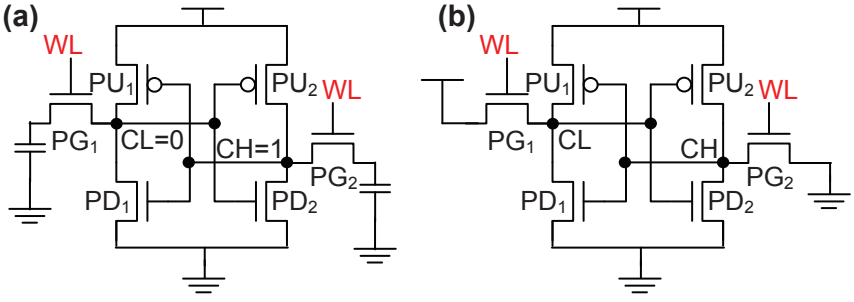


Figure 3.44: (a) SRAM schematic for read access with internal node CL storing a "0". (b) SRAM schematic for writing a "0" into node CH.

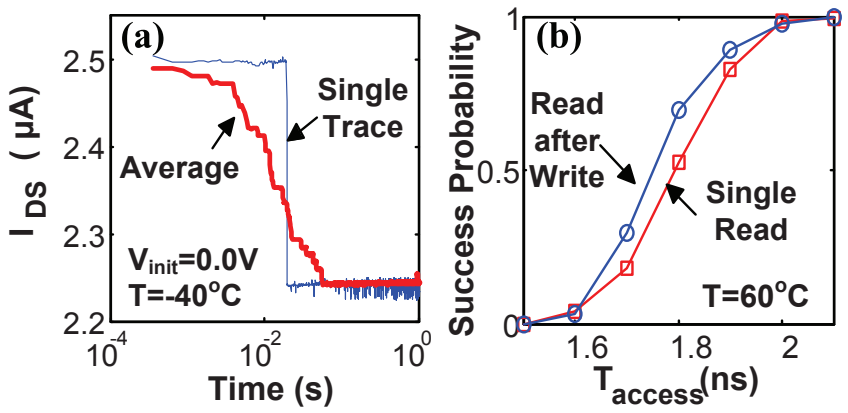


Figure 3.45: (a) Large-bias change response occupancy of a trap in $PD1$. (b) Statistical distributions of T_{access} for single-read and read-after-write.

3.5.1 RTS in pull-down NMOS transistor 1

Figure 3.45 shows the trap occupancy response behavior to a large-bias change for transistor $PD1$ and its effect on SRAM read access time. Each measurement in Figure 3.45(b) corresponds to the probability of observing correct read operations at the corresponding T_{access} pulse width, measured based on 128 samples. The bias conditions prior to read access (ref. Figure 3.44(a)) force trap occupancy in $PD1$ which degrades T_{access} for single-read access. Read-after-write access, on the other hand, only applies the detrimental bias condition of $V_{GS} = V_{DD}$ on $PD1$ for T_{relax} (200 ns in this case) which is not sufficient for the trap to arrive at a full trap occupancy steady-state condition. Subsequently, transistor characteristics of $PD1$ are not degraded in the case of read-after-write, resulting in T_{access} that is faster than the single-read case. The hysteresis observed in T_{access} is dependent on T_{relax}. Figure 3.46 demonstrates that the hysteresis disappears when T_{relax} is greater than 5 μs. Measuring the hysteresis window as a function of T_{relax} for read-after-write operation can be used as an indirect method for measuring the large-bias change response occupancy of a trap in $PD1$. The 5 μs time constant observed in this case is much faster than what is observable using direct transistor measurements.

The opposite trend is observed in a different cell with a type-II trap in $PD1$, due to the opposite gate-bias dependence of type-II traps (Figure 3.47). Read-after-write access in this case results in degraded T_{access} relative to single-read access because the bias conditions held for more than 100 ms prior to the first operation in read-after-write access forces trap occupancy in $PD1$ which degrades T_{access}.

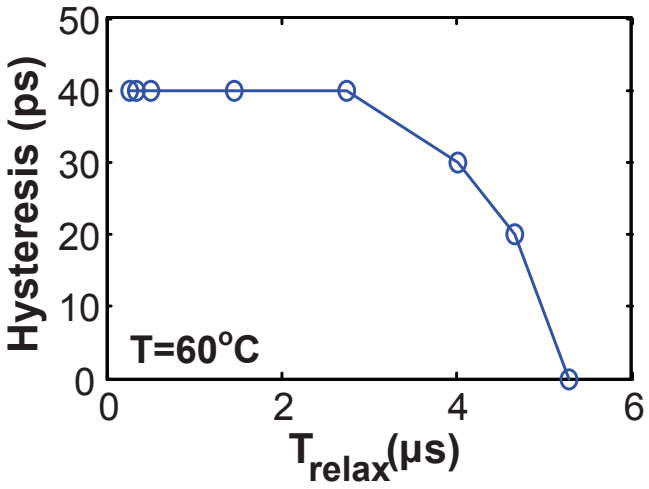


Figure 3.46: Dependence of T_{access} fluctuation on delay since last write access (T_{relax}).

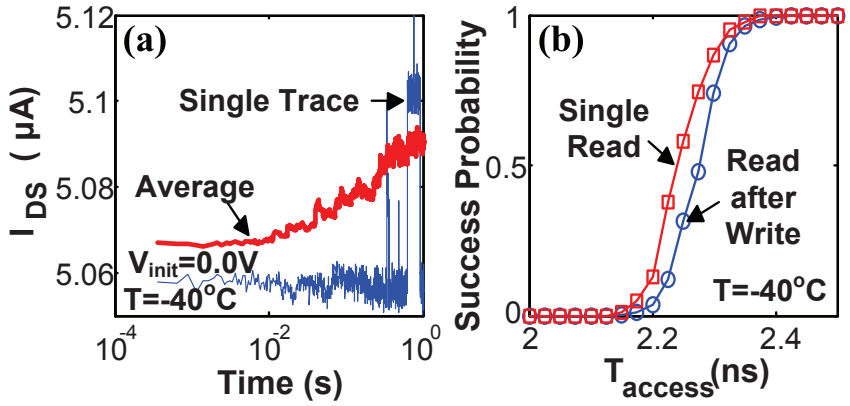


Figure 3.47: (a) Large-bias change occupancy of a type-II trap in $PD1$. (b) Statistical distributions of T_{access} for single-read and read-after-write.

3.5.2 RTS in pull-up PMOS transistor 2

Figure 3.48 shows the trap occupancy response behavior to a large-bias change for transistor $PU2$ and its effect on SRAM dynamic write ability. The $|V_{GS}| = V_{DD}$ bias condition on $PU2$ prior to the final write operation, forces trap occupancy of the RTS identified in Figure 3.48(a), which degrades the characteristics of $PU2$, resulting in improvement of the dynamic writeability of the SRAM bitcell. Write-after-write access on the other hand only subjects $PU2$ to this bias condition for a short period prior to the final write operation which is much faster than the large-bias response time constant of this trap. Write-after-write access therefore does not get the benefit of RTS in $PU2$ improving T_{write} and is therefore slower

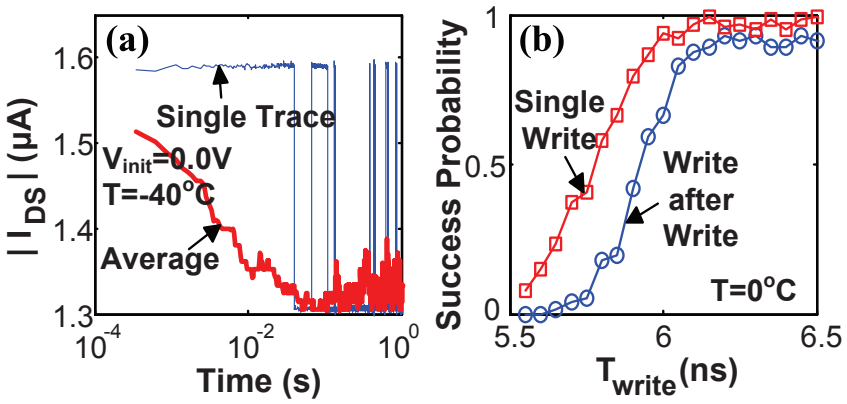


Figure 3.48: (a) Large-bias change occupancy of a trap in $PU2$. (b) Statistical distributions of T_{write} for single-write and write-after-write.

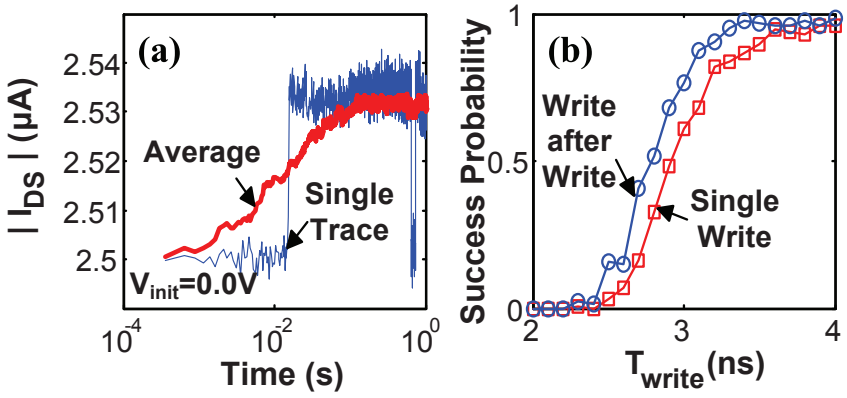


Figure 3.49: (a) Large-bias change occupancy of a type-II trap in $PU2$. (b) Statistical distributions of T_{write} for single-write and write-after-write.

compared to single-write. The opposite trend is observed, where single-write is slower than write-after-write, in the case of a type-II trap present in $PU2$, due to the opposite bias dependence of the trap (Figure 3.49).

3.5.3 RTS in pull-up PMOS transistor 1

Figure 3.50 shows the trap occupancy response behavior to a large-bias change for transistor $PU1$ and its effect on T_{write}. Write-after-write access in this case is slower than single-write. The $|V_{GS}| = V_{DD}$ bias condition on $PU1$ prior to the first write operation in a write-after-write access, forces trap occupancy of the RTS identified in Figure 3.50(a), which degrades the characteristics of $PU1$. An SRAM bitcell is dependent on $PU1$ to complete the write

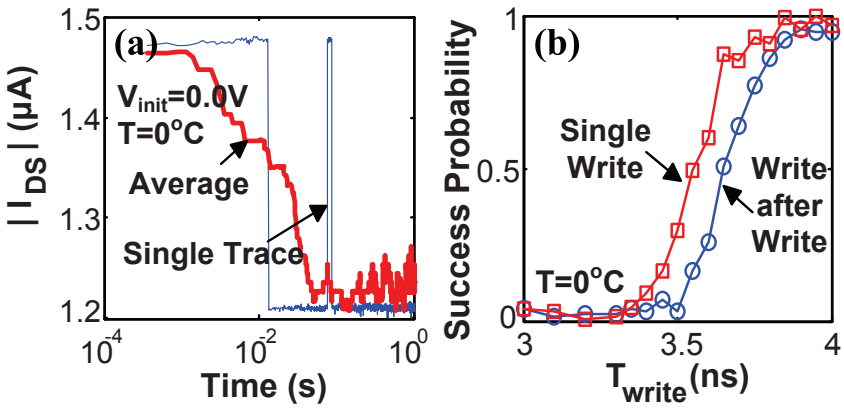


Figure 3.50: (a) Large-bias change occupancy of a trap in $PU1$. (b) Statistical distributions of T_{write} for single-write and write-after-write.

operation by pulling the node up to V_{DD}. Degradation of $PU1$ due to an occupied trap in this transistor therefore degrades T_{write}. Single write access on the other hand does not apply this detrimental bias condition on $PU1$ and is therefore faster than write-after-write access. An opposite trend is expected if a type-II trap is present in $PU1$.

3.5.4 RTS in pass-gate NMOS transistors

The impact of RTS on the pass-gate transistors is slightly different from the other transistors located in the cross-coupled inverters of an SRAM bitcell because the pass-gate transistors are only subject to short gate pulses under normal SRAM operation. Even though the pass-gate transistors are not subject to a long $V_{GS} = V_{DD}$ bias prior to read or write access that forces occupancy of conventional traps in these transistors, it is still possible for repeated access on the same bitcell to cause trap occupancy and therefore degrade characteristics of the pass-gate transistors. Figure 3.51 illustrates the simulated occupancy of a trap in a transistor that is subject to word-line pulses with 50% duty cycle, with a cycle time of 1 ns. The large-bias response time constant for a bias switch corresponding to the word-line turning on ($\tau_{WL,On}$) is 10 ns while the time constant for a bias switch corresponding to the word-line turning off ($\tau_{WL,Off}$) is 15 ns. Figure 3.51(b) illustrates the evolution of average trap occupancy over 100 cycles, which eventually saturates at 60%. This trap therefore has a 60% probability of being occupied after repeated access.

To verify this experimentally, Figure 3.52(a) shows the trap occupancy response behavior to a large-bias change for transistor $PG1$ while Figure 3.52(b) plots the dependence of T_{access} on the number (N) of read-after-read cycles. T_{cycle} is 10 ns in this case. T_{access} degrades as the number of repeated read-after-read cycles before the final read operation increases. This degradation eventually saturates after 128 cycles. An opposite dependence on N is expected

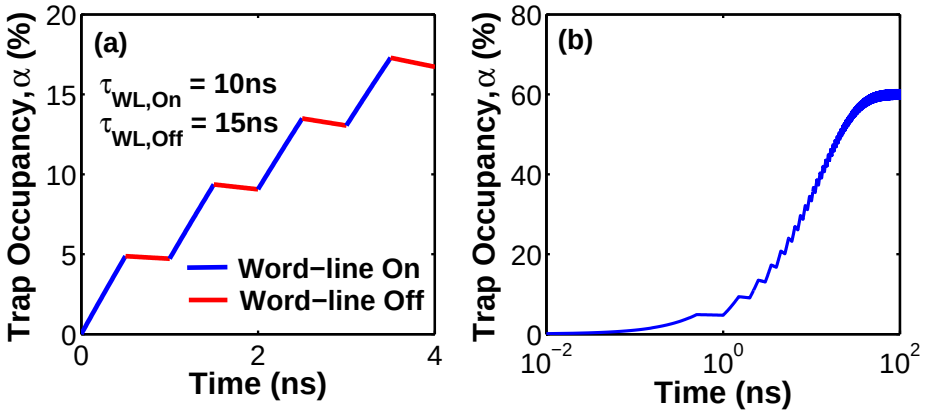


Figure 3.51: Simulated trap occupancy in $PG1$ under successive read-after-read access. (a) Zoomed-in view (b) Expanded view

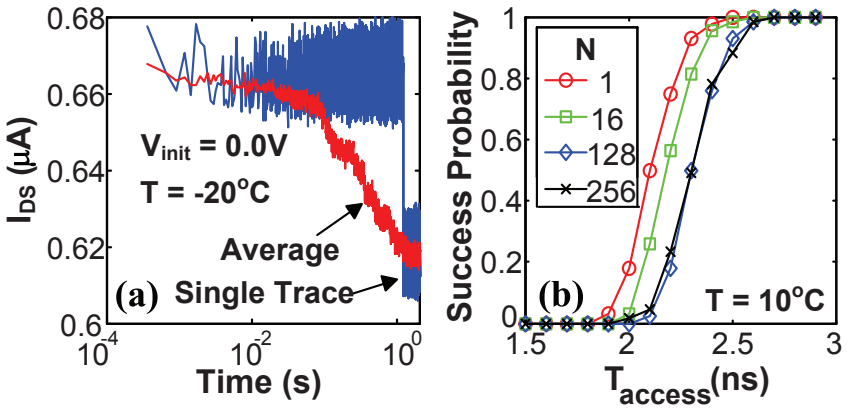


Figure 3.52: (a) Large-bias change occupancy of a trap in $PG1$. (b) Statistical distributions of T_{access} with N read-after-read cycles, saturting at N=128.

for a type-II trap present in $PG1$. Presence of a conventional trap in $PG2$ is expected to degrade T_{write} in situations where multiple access operations are performed on the same bitcell prior to the final write operation.

3.5.5 Statistical Distributions and Implications to SRAM Robustness

Figures 3.53 and 3.54 plot the correlation between the fluctuation observed in the respective metrics with their nominal value. The fluctuation is characterized by measuring T_{access} or T_{write} once under single operation, and another time under read-after-write or write-

after-write access. This two point measurement provides an efficient method for measuring the largest fluctuation in the respective metrics due to RTS. The worst-case fluctuation of T_{access} observed is 29%. The positive value of ΔT_{access} due to $T_{single-read}$ being greater than $T_{read-after-write}$ is indicative of a conventional trap. This is confirmed by directly observing the trap in $PD1$. For write operation, the worst-case fluctuation observed is 45%. The negative value of ΔT_{write}, due to $T_{single-write}$ being smaller than $T_{write-after-write}$, is indicative of a conventional trap. This is also confirmed by directly characterizing the trap in $PU2$.

Analysis on the impact of conventional traps in the respective transistors, discussed previously, indicate that single-read T_{access} is always slower than read-after-write (ref. Figure 3.45) while single-write T_{write} is always slower than write-after-write (ref. Figures 3.48 and 3.50). Observation of negative values of ΔT_{access} in Figure 3.53 and positive values of ΔT_{write} in Figure 3.54 indicate that type-II traps are quite commonly found in the transistors and degrade transistor performance significantly. Type-II traps, in addition to conventional traps, therefore need to be accounted for in SRAM reliability characterization and statistical design.

These results also suggest that RTS time constants can be ignored when estimating the impact of RTS on dynamic stability. This is due to the fact that the time constants of RTS are orders of magnitude slower than SRAM access times, resulting in the traps being either occupied or empty when the bitcell is being accessed. This greatly simplifies simulation of SRAM operation because the transient characteristics of RTS can be omitted. The impact of RTS is simply estimated by applying the worst-case and best-case trap occupancy states in the transistors, depending on the actual access operation, while performing a transient simulation to evaluate the respective margin. One possible exception to this might be the pass-gate transistors which are typically pulsed. Assigning a full trap occupancy to these transistor might be too pessimistic as the gate pulses might not be sufficient to cause full trap occupancy, as demonstrated in Figure 3.51. However, SRAM reliability needs to be guaranteed to extremely low levels of probability due to the large number of accesses a bitcell is subject to in its lifetime. Therefore, assigning worst-case degradation due to full trap occupancy in the pass-gate transistors is justified.

In conclusion, although RTS can cause up to 45% fluctuation in the margins, these large fluctuations are correlated with SRAM bitcells with nominal performance. Weak bitcells (large T_{access} and large T_{write}) located at the outliers of the distribution are correlated with smaller fluctuation in the metrics. This is the result of convolving a long-tailed statistical distribution, characteristic of RTS amplitude, with a Gaussian distribution, characteristic of traditional sources of variability such as RDF. The impact of RTS on dynamic stability therefore needs to be evaluated statistically to determine the expected degradation in nominally weak cells that are most likely to fail due to RTS fluctuation.

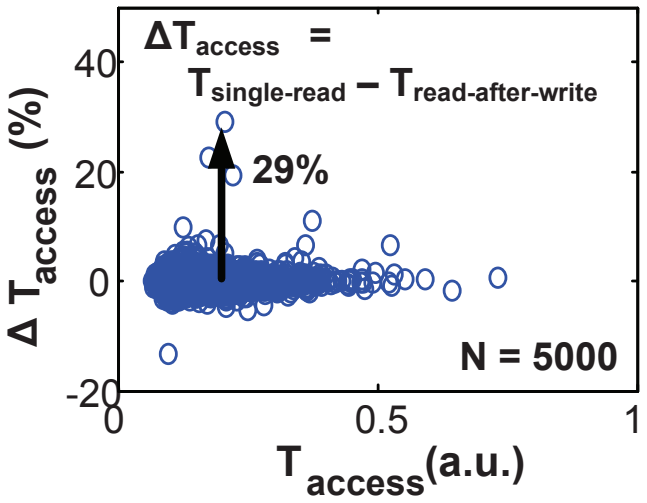


Figure 3.53: Scatter plot of ΔT_{access} due to RTS vs. nominal T_{access}.

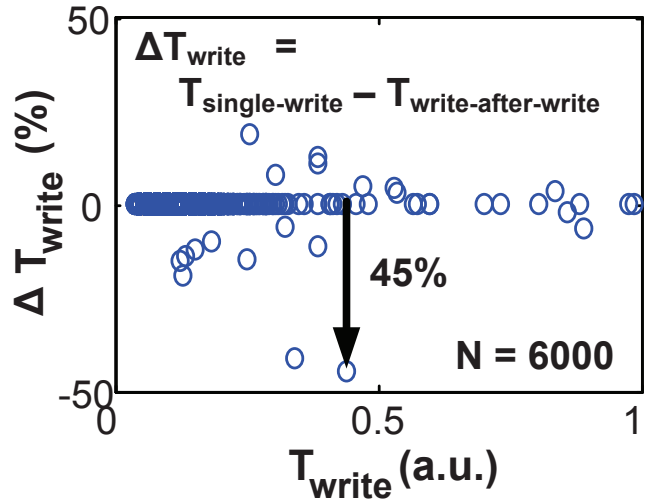


Figure 3.54: Scatter plot of ΔT_{write} due to RTS vs. nominal T_{write}.

3.6 Correlation Between RTS and NBTI

Negative bias temperature instability (NBTI) is another source of temporal variability in PMOS transistor performance. NBTI is observed as degradation in transistor V_{th} with time which occurs when a negative bias is applied on the gate of PMOS transistors (turning them on) and is exacerbated by elevated temperatures. Figure 3.55 (a) plots the V_{th} shift observed in an SRAM PMOS pull-up transistor after 10,000 seconds of -1.5 V stress at 125^{o}C. The threshold voltage of the transistor appears to continue degrading as a function of stress time. V_{th} degradation due to NBTI is observed to occur in a step-wise manner, especially in extremely scaled transistors such as those found in SRAM bitcells. This degradation is believed to be caused by either $Si - H$ bonds at the gate oxide interface breaking, leaving behind charged dangling bonds, or hole traps within the gate oxide capturing or releasing holes [23, 75, 46]. Due to the fact that these hole traps also cause RTS noise, it is believed that degradation in V_{th} due to RTS is a component of the degradation observed in NBTI. Figure 3.55 (b) presents experimental evidence supporting this. This graph plots a zoomed-in view of the V_{th} degradation observed in Figure 3.55 (a). RTS noise of the same transistor was sampled under different bias and temperature conditions (-1.0 V 25^{o}C) prior to the NBTI experiment and is super-imposed on Figure 3.55 (b). The magnitude of the rapidly fluctuating RTS noise observed at -1.0V, 25^{o}C is correlated with a step-wise transition in V_{th} observed at -1.5V, 125^{o}C. This change in trap dynamic behavior due to bias and temperature is characteristic of type-I traps (ref. Figure 3.9).

These results have important implications on accounting for the impact of RTS and NBTI on SRAM operation. RTS and NBTI have mostly been considered as unrelated sources of variability in SRAM operation [70, 1, 28, 67, 3]. The results in this thesis demonstrate that since transistor degradation due to RTS noise is included in NBTI margin, considering these two components separately will result in double-counting the impact on SRAM.

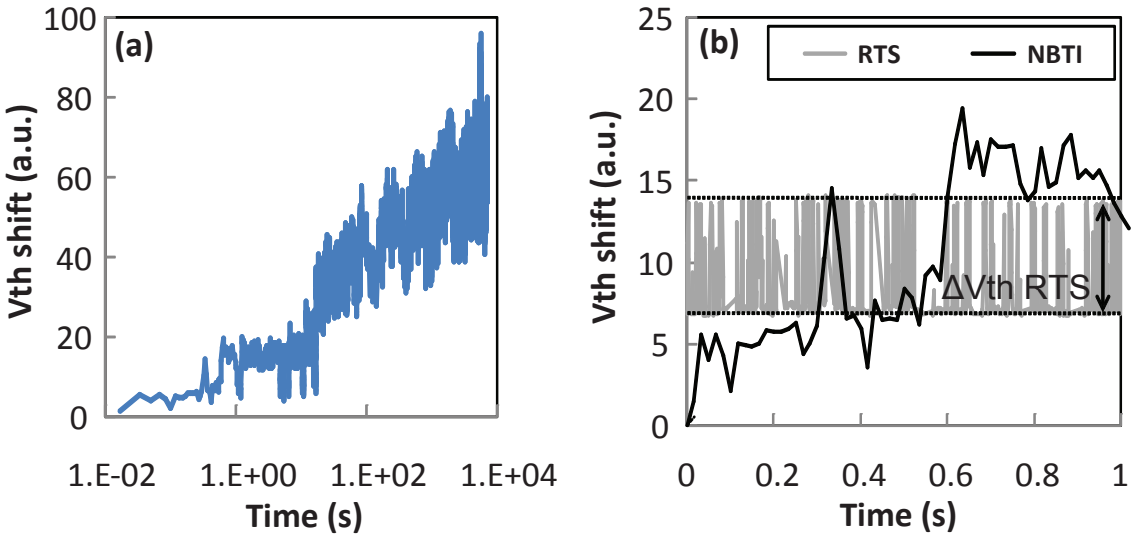


Figure 3.55: (a) Measured V_{th} degradation due to NBTI. (b) RTS observed in the same transistor at 25^{o}C super-imposed on top of V_{th} degradation observed during NBTI characterization at 125^{o}C.

Chapter 4

Statistical Estimation of SRAM Dynamic Stability

4.1 Introduction

Robust SRAM design requires guaranteeing correct functionality of dynamic read stability, writeability, and read access in every single SRAM bitcell, across process margins, operating voltages, and temperature. This is becoming challenging due to increasing process and operating uncertainties on the one hand and also increasing demand for larger SRAM arrays. Various statistical methods for estimating SRAM reliability have been proposed and have been effectively used at various levels of SRAM design.

Process technologist frequently report the μ and σ of a particular SRAM bitcell for read stability and writeability to quickly compare the robustness of various bitcell choices and expected robustness at different operating voltages [4, 19, 8]. The robustness of the SRAM bitcell is then estimated by assuming a Gaussian distribution and extrapolating the read and write margins out to the failure point. This method is attractive because μ and σ can be easily estimated from measurements of less than 100 SRAM bitcells which can be generated from device simulations or from actual fabricated testchips. This technique however assumes that the distributions of read and write margins are Gaussian up to multiple sigma. Static margins have been shown to follow a Gaussian distribution up to multiple σ but they deviate from the Gaussian distribution at low V_{DD} [24]. Furthermore these techniques do not apply to dynamic stability as it has been demonstrated that all distributions of dynamic stability are non-Gaussian [37].

Sensitivity analysis has also been frequently used to estimate sensitivities of a metric f, such as read or write margin, to perturbations in various parameters x_i, such as transistor V_{th}. Sensitivity analysis can be further extended to estimate SRAM reliability by formulating the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \text{cell sigma} = \sqrt{\boldsymbol{x}^T \boldsymbol{x}} \\ \text{subject to} \quad & f_{\boldsymbol{x}=\mathbf{0}} + \sum_i \frac{\delta f}{\delta x_i} x_i = 0,\ i = 1, \ldots, n. \end{aligned} \tag{4.1}$$

where x_i denotes the respective V_{th} variation for transistor i normalized with respect to its standard deviation $\sigma_{V_{th}}$. The goal of this optimization is to find the vector of parameters $\boldsymbol{x}$ with minimum geometrical distance from the origin such that the estimated metric f is 0. The reliability of the SRAM bitcell is parameterized by cell sigma as defined in Equation 4.1, assuming that the distributions of each of the parameters x_i are independent Gaussian random variables and by applying the central limit theorem. The optimal cell sigma obtained from this optimization effectively characterizes the largest amount of variability the bitcell is able to sustain before a failure occurs. This statistical estimation method is equivalent to the Most Probable Failure Point (MPFP) and worst-case vector method defined in other works [74, 12, 37]. The merit of this method lies in the fact that no assumption is made on the statistical distribution of f. This technique is therefore able to handle the non-Gaussian distributions found in dynamic stability metrics [37]. It can also be extended to handle non-linearities in the sensitivities by iteratively evaluating the sensitivities at each search point [74]. The main limitation of the MPFP method lies in the large number of iterations required to converge to a solution, especially when f is a non-linear function of $\boldsymbol{x}$. There's also the risk of converging to a local optimum but it has been demonstrated that the solution obtained through the MPFP method is usually the global optimum [12].

A straightforward approach for estimating reliability of SRAM bitcells is by performing Monte Carlo simulations. In this scheme, a large population of samples is generated from the original distributions of parameters x_i. The margin, f, is then evaluated from each sample set. The probably of failure is then estimated as follows, where A is the set of points in the sample space that meet the failure conditions.

$$\hat{p_{MC}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\boldsymbol{x_i} \in A) \tag{4.2}$$

Monte Carlo simulations suffer from unattractive long runtimes in order to reach the low failure rates required for multi-megabit SRAMs. It has been reported that analysis of a cell failing at a rate of 10^{-5} to 10^{-6} required 40 million samples that took two months of runtime [20]. Importance sampling, which comes under the class of variance reduction methods [20], offers to speed up the runtime of Monte Carlo simulations by generating sample sets from a different distribution, $\hat{f}(\boldsymbol{x})$ such that the probability of observing samples in the failure region A is increased. Let $f(\boldsymbol{x})$ be the probability density of $\boldsymbol{x}$ and $\hat{f}(\boldsymbol{x})$ be the probability density of the importance sampling distribution. The probability of failure ($\hat{p_{IS}}$) is estimated using Equation 4.3 [20]. A figure of merit, $\rho(\hat{p_{IS}})$, is defined in Equation 4.4 and is used as

an indicator of when to stop the simulation. A value of 0.1 corresponds to an accuracy in the estimate of $\hat{p_{IS}}$ of at least 90% with a confidence of 90%.

$$\begin{aligned} w(\boldsymbol{x}) &= \frac{f(\boldsymbol{x})}{\hat{f}(\boldsymbol{x})} \\ \hat{p_{IS}} &= \frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{x_i})\mathbf{1}(\boldsymbol{x_i} \in A) \end{aligned} \tag{4.3}$$

$$\rho(\hat{p_{IS}}) = \frac{\sqrt{VAR(\hat{p_{IS}})}}{\hat{p_{IS}}} \tag{4.4}$$

The main challenge with importance sampling is to find the distribution $\hat{f}(\boldsymbol{x})$ such that the runtime of importance sampling, compared to conventional Monte Carlo simulations, is reduced. The most common approach is to shift the original distribution by a shift vector, $\boldsymbol{x_s}$, at the boundary of the region of failure ($\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x} - \boldsymbol{x_s})$). [35] finds a shift vector by uniformly sampling the entire space and setting the shift vector at the region where the most failures are sampled. The importance sampling distribution is also augmented with a uniform distribution covering the entire space, in addition to the shifted original distribution, to ensure that all regions are sampled. [20] demonstrates that based on the classical Large Deviation Theory, the optimal shift vector is the vector that has the minimum $\mathscr{L}^2$-norm and lies on the failure contour surrounding A. Up to 10000X reduction in runtime, compared to conventional Monte Carlo simulations, was observed when this technique was used to estimate SRAM reliability using static noise margins. The optimal shift vector proposed in [20] is similar to the MPFP or Worst-Case Vector described earlier. We can therefore apply techniques for finding the MPFP to efficiently find a shift vector. Furthermore, the large number of iterations required to converge to a MPFP can be relaxed if applied to shift vector search because the final estimate of the reliability is based on a sampling of multiple points surrounding the MPFP, instead of the actual MPFP.

This chapter formally defines the importance sampling and MPFP search algorithm using static margins. Various adaptions are then introduced to adapt this algorithm to dynamic stability metrics which tend to be non-linear compared to static noise margins. Finally, this algorithm is adapted to handle non-Gaussian distributions.

4.2 Importance Sampling and Most Probable Failure Point Search

The optimization problem for finding the MPFP is defined in Equation 4.1. One major assumption made in this problem formulation is that the metric, f, is a linear function of the parameters $\boldsymbol{x}$. Figures 4.2 and 4.3 plot the deviation in the respective margins for static read margin and static write margin as a function of variability in each transistor (Figure

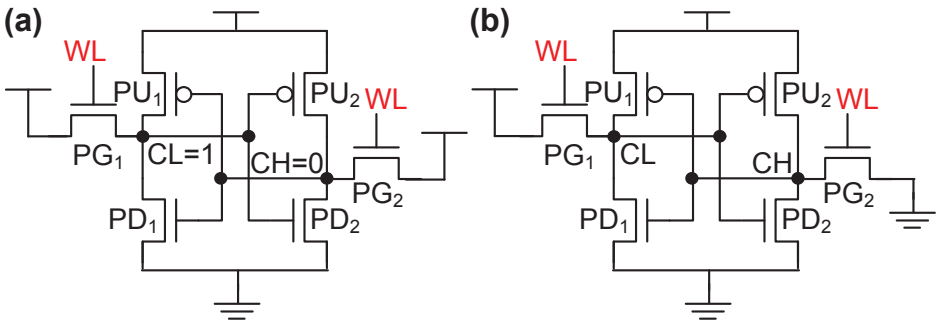


Figure 4.1: Schematics of SRAM for: (a) Read static noise margin analysis. (b) Write static noise margin analysis.

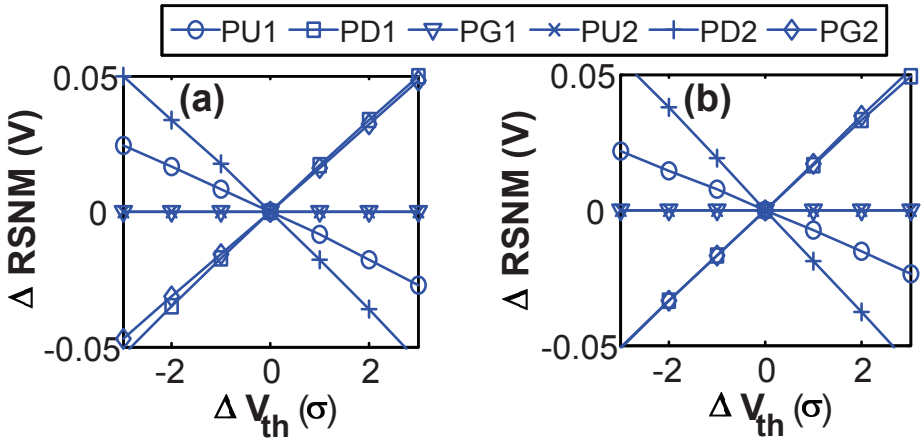


Figure 4.2: Sensitivity of read static noise margin (RSNM) to V_{th} variability centered at: (a) $\boldsymbol{x} = \mathbf{0}$. (b) $\boldsymbol{x} = MPFP$.

4.1). These sensitivities were collected at two different points: the origin ($\boldsymbol{x} = \mathbf{0}$) and at the MPFP. The margins do indeed appear to be linearly dependent on ΔV_{th} up to 3 σ, however slight deviations from linearity are observable in the plots. These non-linearities result in errors in estimating the MPFP [74]. Furthermore, the sensitivity of f to the parameters change as $\boldsymbol{x}$ changes. For example, the sensitivity of RSNM to $PD2$ increases at the MPFP compared to the origin (Figure 4.2).

These non-linearities as well as changes in sensitivity can be accounted for by iteratively solving for the MPFP using locally evaluated gradients. Figure 4.4 illustrates this procedure visually in 2 dimensions of variability. The algorithm initially starts at the origin. The local gradients $(\frac{\delta f}{\delta V_{th,x}}, \frac{\delta f}{\delta V_{th,y}})$ are then evaluated, using multiple calls to a SPICE simulator while varying each parameter. A modified version of Equation 4.1 is then solved. Equation 4.5

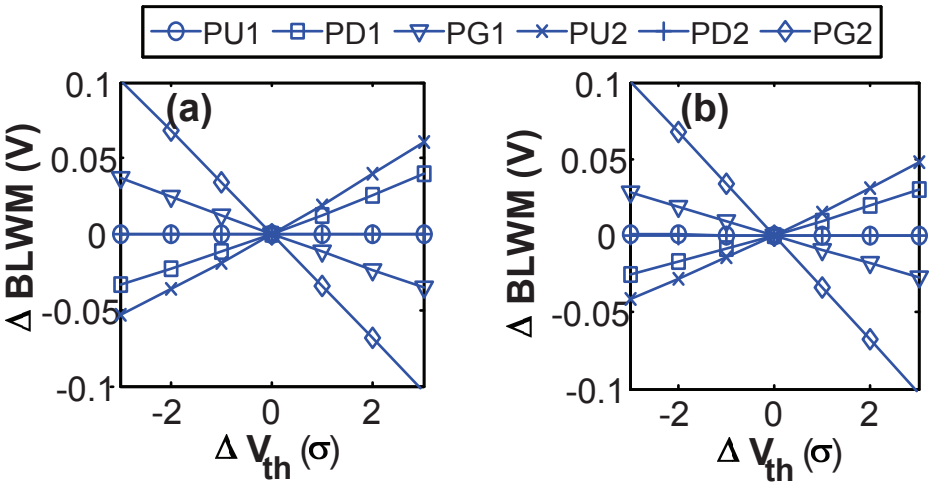


Figure 4.3: Sensitivity of bit-line write margin (BLWM) to V_{th} variability centered at: (a) $\boldsymbol{x} = \mathbf{0}$. (b) $\boldsymbol{x} = MPFP$.

takes into account that the gradients are evaluated locally at some point, x_{ci}, instead of the origin, while a constant is added to the vectors for evaluating cell sigma. The objective function still remains convex despite of this modification and therefore can be solved for the global optimum. Locally evaluated gradients provide a better approximation of the surface at the current location that's being evaluated and promises to converge faster. The disadvantage is that the SPICE simulator needs to be called at every iteration to compute the new gradients. This optimization problem is iteratively executed, each time using gradients evaluated at the current solution, until the algorithm converges at the failure contour (Figure 4.4). Carlson proposes a different method where the optimizer first finds a point on the failure contour before optimizing for the MPFP [12]. The optimization problem proposed in Equation 4.5 optimizes for both conditions concurrently by specifying the most-probable (lowest cell-sigma) point as the objective and the failure point as a constraint. This method will therefore arrive at the MPFP with less iterations.

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \text{cell sigma} = \sqrt{(\boldsymbol{x}+\boldsymbol{x_{ci}})^T(\boldsymbol{x}+\boldsymbol{x_{ci}})} \\ \text{subject to} \quad & f_{\boldsymbol{x_{ci}}} + \sum_i \frac{\delta f}{\delta x_i} x_i = 0,\ i = 1, \ldots, n. \end{aligned} \tag{4.5}$$

Table 4.1 tabulates the points in the 6 parameter V_{th} space evaluated by the proposed MPFP search algorithm applied towards finding the MPFP corresponding to static read margin (RSNM). The algorithm initially finds a failure point with large cell-sigma but eventually arrives at other failure points with lower cell-sigma and hence are more probable. As expected, $PG1$ and $PU2$ do not contribute to cell-sigma because of the negligible

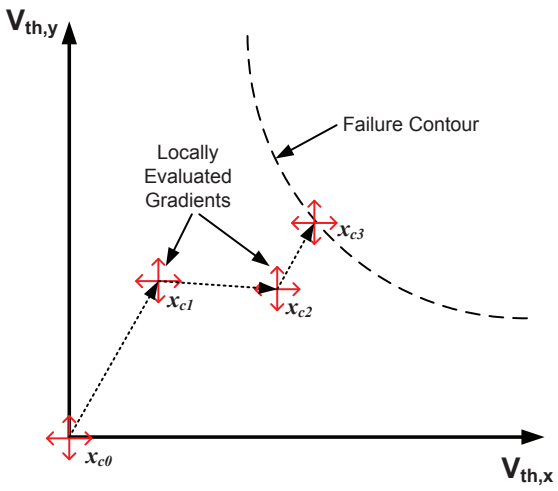


Figure 4.4: MPFP search with locally evaluated gradients.

Iteration	PU1	PD1	PG1	PU2	PD2	PG2	Cell Sigma
1	2.04	-3.31	0.00	0.00	1.35	-3.05	5.12
2	1.38	-2.69	0.00	0.00	2.80	-2.91	4.99
3	1.28	-2.12	0.00	0.00	2.70	-3.24	4.86
4	1.53	-2.49	0.00	0.00	2.92	-2.68	4.92
5	1.23	-2.15	0.00	0.00	2.77	-3.16	4.88
6	1.33	-2.55	0.00	0.00	3.02	-2.61	4.92
7	1.30	-2.94	0.00	0.00	2.74	-2.48	4.90
8	0.77	-2.03	0.00	0.00	3.59	-2.21	4.74
9	2.10	-2.30	0.00	0.00	2.59	-2.43	4.72
10	1.36	-2.40	0.00	0.00	3.14	-2.50	4.87

Table 4.1: Progress of MPFP search algorithm applied to RSNM

sensitivity of RSNM to variability in these transistors. Note that this algorithm did not converge to a stable point after 10 iterations. While convergence is absolutely necessary for SRAM reliability analysis that relies on cell-sigma as the metric for SRAM reliability, importance sampling with MPFP search only needs an approximate point, close to the actual MPFP.

The final SRAM reliability estimate is derived based on importance sampling using the final vector ($\boldsymbol{x_s}$) obtained from the MPFP search algorithm. Original Gaussian distributions corresponding to each V_{th} parameter is shifted by the corresponding element in $\boldsymbol{x_s}$ while the sigmas are kept constant. Figure 4.5 plots the RSNM histogram of 5000 samples generated

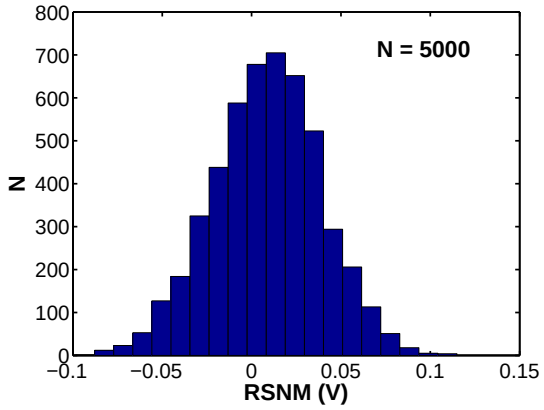


Figure 4.5: RSNM histogram of 5000 samples generated from original Gaussian distributions shifted by the MPFP vector.

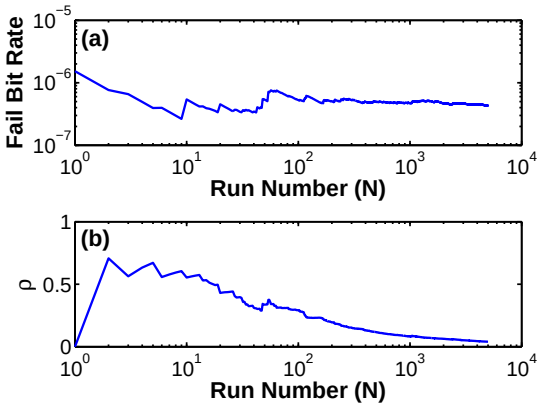


Figure 4.6: Evolution of (a) fail bit rate ($p\hat{_{IS}}$) and (b) convergence metric (ρ) as a function of run number.

from this shifted distribution, confirming that approximately half of the samples generated from these shifted distribution fall in the failure region ($RSNM < 0$). The actual mean of the histogram is approximately 10 mV from the ideal boundary of 0 mV due to the fact that the shift vector was not exactly the actual MPFP. This slight deviation is inconsequential as the importance sampling algorithm was still able to converge to a result quickly. Figure 4.6 plots the evolution of the importance sampling algorithm as a function of the run number. The algorithm converged to a good estimate ($\rho(p\hat{_{IS}}) < 0.1$) within 1000 simulations.

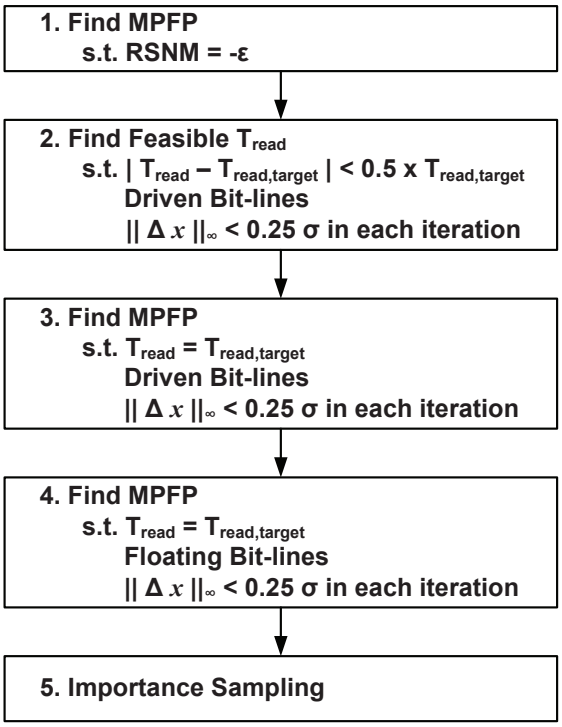


Figure 4.7: General algorithm for estimating statistical dynamic read stability reliability of SRAM.

4.3 Importance Sampling and Dynamic Stability

In this subsection, the importance sampling with MPFP search algorithm are applied towards estimating SRAM reliability using dynamic stability metrics based on critical word-line pulse-widths. Optimizing for dynamic stability is more complex than static margins because the margins are much more sensitive to slight perturbations in transistor parameters, compared to static margins. Techniques will be presented for improving convergence of the MPFP search algorithm.

4.3.1 Dynamic Read Stability

Searching for the MPFP with dynamic read stability is challenging because part of the search space is undefined. These points correspond to SRAM bitcells with non-zero static read margins and therefore infinite T_{read}. The MPFP algorithm defined for static margins therefore needs to be modified to account for this non-ideality. Figure 4.7 presents a general description of the algorithm used to find the MPFP of SRAM using dynamic read stability. The first step in this algorithm involves finding the MPFP using a static read margin

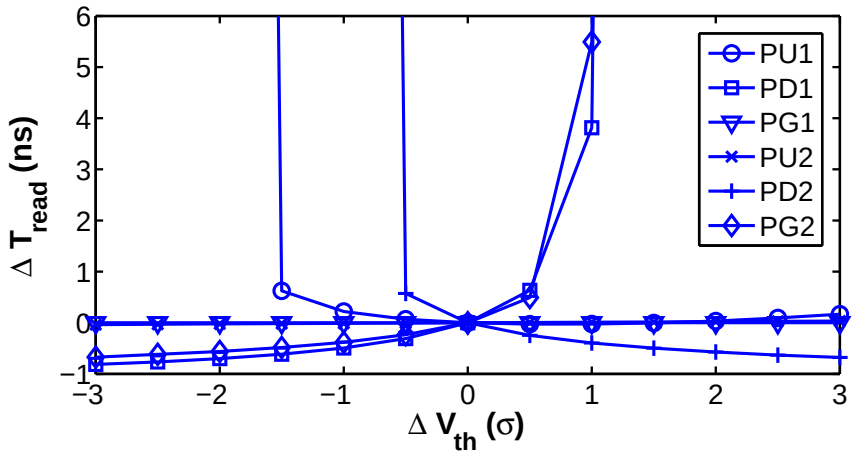


Figure 4.8: Sensitivity of dynamic read stability (T_{read}) to V_{th} variability.

(RSNM). RSNM is used initially instead of T_{read} because this margin is defined across the entire search space. The failure criteria is set at some slightly negative value $-\epsilon$ instead of 0 as this slight negative value guarantees that the algorithm converges to a point where T_{read} is finite. Even though RSNM is used instead of T_{read}, the solution obtained from this step should be within proximity of the MPFP with dynamic read stability, due to the correlation between these two metrics which was demonstrated in Chapter 1.

Figure 4.8 plots the sensitivity of T_{read} to V_{th} variability in the respective transistors (Figure 4.1). T_{read} appears to be more non-linearly dependent on V_{th} compared to static margins. The large jumps in the plot correspond to conditions where the SRAM bitcell becomes statically stable, resulting in T_{read} becoming infinite. This highly non-linear function together with the existence of discontinuities make it challenging to apply the MPFP search algorithm presented previously on dynamic read stability. Instead of directly solving for the MPFP, step 2 of the algorithm (Figure 4.7) first finds a point in the design space where T_{read} is within a certain range from the T_{read} constraint. This ensures that the MPFP optimization problem with the strict constraint that $T_{read} = T_{read,target}$, carried out in step 3, is feasible. Furthermore, the largest change in any dimension of the design point during each iteration is limited to 0.25 σ. This is to keep the search within a local region where the sensitivities can still be approximated as linear functions. Equation 4.6 is solved for x, which is the vector of shifts in the parameter space, in the least-squares sense. Table 4.2 lists the progress of the feasibility search algorithm with the respective transistor parameters and cell sigma at each iteration. The algorithm found a feasible point (within 2x of $T_{read,target}$) after 4 iterations. Note that the cell sigma increased during each iteration because no explicit objective function for minimizing cell sigma was specified.

Once a feasible point is found, the algorithm proceeds to step 3 where it looks for the MPFP with the constraint that $T_{read} = T_{read,target}$. The dependence of ΔT_{read} on the

Iteration	PU1	PD1	PG1	PU2	PD2	PG2	Cell Sigma	T_{read} (ns)
1	1.36	-2.32	0.00	0.00	3.01	-2.41	4.70	7.63
2	1.36	-2.57	0.00	0.00	3.01	-2.41	4.83	2.73
3	1.61	-2.57	0.00	0.00	3.01	-2.41	4.91	2.34
4	1.61	-2.57	0.00	0.00	3.26	-2.41	5.06	1.77

Table 4.2: Progress of feasible T_{read} search algorithm. $T_{read,target} = 1$ ns

parameters is modeled as a quadratic equation to model the non-linearity in the sensitivities (Figure 4.8), resulting in faster and more accurate convergence. Equation 4.7 defines the optimization problem for finding the MPFP of T_{read} with quadratic constraints. In order to keep the optimization problem convex, the equality in the constraint was replaced with an inequality [10]. This optimization problem is convex as long as $\frac{\delta^2 T_{read}}{\delta x_i^2}$ is greater than or equal 0. This is generally true as is observed in Figure 4.8. An equality is not required in this case because the objective function of finding the MPFP only activates the upper bound inequality due to the fact that maximizing probability of failure automatically increases T_{read}. Table 4.3 tabulates the progress of the MPFP search, demonstrating quick convergence to the target of 1 ns. An optional step, where the bit-line capacitances are pre-charged and left floating, is introduced before running importance sampling on the MPFP for instances where SRAM reliability under more realistic mode of operation is desired. Introducing this as a sequential step after MPFP search with driven bit-lines improves convergence of the algorithm. For cases where SRAM reliability under different bit-line capacitances is required, the solution corresponding to the driven bit-lines can be saved and reused when solving for the MPFP corresponding to each bit-line capacitance configuration.

$$\begin{aligned} Ax &= T_{read,target} - T_{read} \\ A &= \left[\frac{\delta T_{read}}{\delta x_1} \quad \frac{\delta T_{read}}{\delta x_2} \quad \dots \quad \frac{\delta T_{read}}{\delta x_n}\right] \end{aligned} \tag{4.6}$$

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \text{cell sigma} = \sqrt{(\boldsymbol{x}+\boldsymbol{x_{ci}})^T(\boldsymbol{x}+\boldsymbol{x_{ci}})} \\ \text{subject to} \quad & T_{read,}\boldsymbol{x_{ci}} + \sum_i \frac{\delta^2 T_{read}}{\delta x_i^2} x_i^2 + \sum_i \frac{\delta T_{read}}{\delta x_i} x_i \le T_{read,target},\ i = 1, \dots, n. \end{aligned} \tag{4.7}$$

Importance sampling, using the shift vectors obtained from MPFP search, is used to obtain an estimate of the fail bit rate of the SRAM bitcell. The SPICE simulation used for importance sampling basically performs a read operation on an SRAM bitcell, at the specified $T_{read,target}$ and checks if data corruption occurred. This is much faster than finding the critical word-line pulse-width for dynamic read stability that is required for MPFP search. Figure 4.9 plots the evolution of the estimate of the fail bit rate ($\hat{p_{IS}}$) a well as the figure

Iteration	PU1	PD1	PG1	PU2	PD2	PG2	Cell Sigma	T_{read} (ns)
1	1.61	-2.57	0.00	0.00	3.26	-2.41	5.06	1.77
2	1.44	-2.82	0.00	-0.05	3.37	-2.59	5.29	1.18
3	1.19	-2.98	0.00	-0.05	3.40	-2.70	5.40	0.97
4	0.94	-3.02	0.00	-0.06	3.25	-2.84	5.40	0.97
5	0.69	-3.04	0.00	-0.05	3.33	-2.73	5.32	0.99
6	0.44	-2.93	0.00	-0.03	3.21	-2.92	5.26	1.07
7	0.35	-2.90	0.00	-0.03	3.30	-2.89	5.27	1.06
8	0.55	-3.07	0.00	-0.05	3.16	-2.87	5.29	1.01

Table 4.3: Progress of T_{read} MPFP search algorithm. $T_{read,target} = 1$ ns

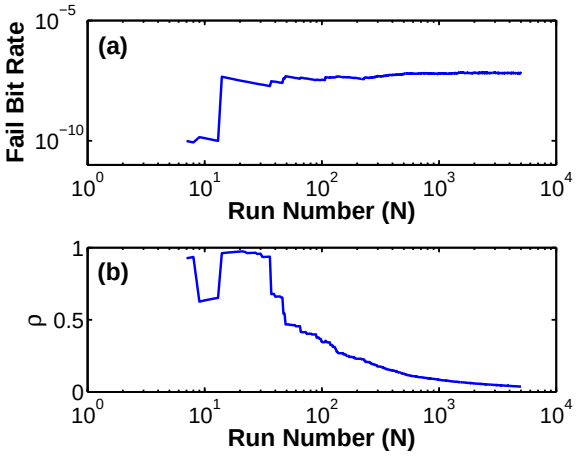


Figure 4.9: Evolution of (a) fail bit rate ($\hat{p_{IS}}$) and (b) convergence metric (ρ) as a function of run number, corresponding to T_{read} importance sampling.

of merit (ρ), demonstrating convergence after 1000 samples. The joint probability of the MPFP was also compared against the joint probability of other failures in the dataset of 5000 samples and was observed to be much larger, confirming the effectiveness of the MPFP search algorithm proposed in this section.

4.3.2 Dynamic Writeability

Figure 4.10 illustrates the general algorithm for estimating SRAM robustness using dynamic writeability as the write margin. Just as in the case of dynamic read stability, the algorithm first finds the MPFP using a static write margin (BLWM) before switching to dynamic writeability (T_{write}). The failure point in this case is defined as some small value (ϵ) above

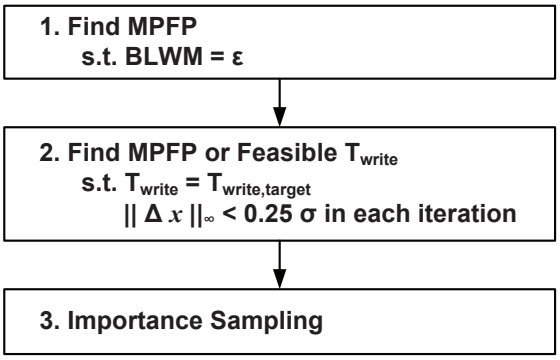


Figure 4.10: General algorithm for estimating statistical dynamic writeability reliability of SRAM.

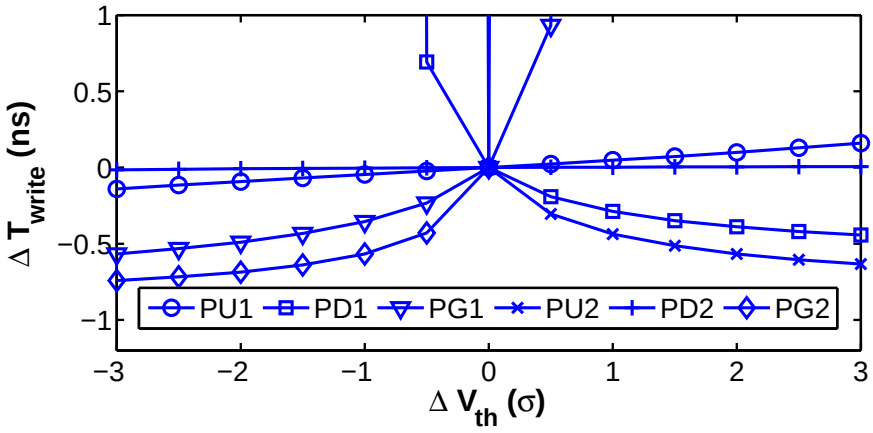


Figure 4.11: Sensitivity of dynamic writeability (T_{write}) to V_{th} variability.

zero so that the vector obtained from this step is on the verge of write failure. The algorithm takes advantage of the fact that static write margin and dynamic writeability are correlated and hence the MPFP corresponding to a static write margin provides a good starting point to search for the MPFP corresponding to dynamic writeability. Figure 4.11 plots the sensitivities of T_{write} to V_{th} variability in the respective transistors. The main advantage of using BLWM instead of T_{write} initially is that sensitivities of BLWM to transistor parameters (Figure 4.3) are more linear compared to T_{write}. The MPFP search algorithm is therefore able to converge to a solution with less number of iterations, due to the better modeled sensitivities.

Iteration	PU1	PD1	PG1	PU2	PD2	PG2	Cell Sigma	T_{write} (ns)
1	0.00	-1.21	1.64	-2.42	0.00	4.27	5.32	1.69
2	0.01	-1.37	1.39	-2.17	0.00	4.02	4.97	0.60
3	0.01	-1.37	1.39	-2.17	0.00	4.27	5.18	0.88
4	0.07	-1.12	1.54	-1.92	0.01	4.40	5.17	0.85
5	0.11	-1.37	1.29	-2.17	0.01	4.32	5.19	0.92
6	0.04	-1.12	1.04	-1.92	0.00	4.46	5.09	0.70
7	0.27	-1.37	1.29	-2.17	0.02	4.42	5.28	1.30
8	0.02	-1.12	1.04	-2.31	0.00	4.17	5.01	0.63

Table 4.4: Progress of T_{write} MPFP search algorithm. $T_{write,target} = 1$ ns

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \text{cell sigma} = \sqrt{(\boldsymbol{x}+\boldsymbol{x_{ci}})^T(\boldsymbol{x}+\boldsymbol{x_{ci}})} \\ \text{subject to} \quad & T_{write,\boldsymbol{x_{ci}}} + \sum_i \frac{\delta T_{write}}{\delta x_i} x_i = T_{write,target} \qquad i = 1,\ldots,n. \end{aligned} \tag{4.8}$$

Equation 4.8 describes the optimization problem for finding the MPFP corresponding to T_{write}. The sensitivities of T_{write} to the parameters are modeled as linear equations even though the sensitivities are highly non-linear (Figure 4.11) in order to keep the problem convex. The technique of using an inequality constraint used for dynamic read stability (Equation 4.7) does not apply to dynamic writeability because the upper bound inequality is not active in the optimization. Maximizing failure probability automatically decreases T_{write}. Placing a lower bound inequality in the constraint, with positive coefficients of $\frac{\delta^2 T_{write}}{\delta x_i^2}$ observed in Figure 4.11, would result in a non-convex problem [10]. The MPFP search algorithm for T_{write} is therefore expected to perform poorly due to the poor fitting of the sensitivities to a linear model. Furthermore, T_{write} is really sensitive to variability in some transistors. For example, a slight variability in $PG2$ and $PU2$ resulted in T_{write} of the SRAM bitcell going to infinity, corresponding to an unwritable bitcell (Figure 4.11). Due to this strong sensitivity, the MPFP search algorithm (Figure 4.10) is augmented with a fallback in situations where the optimization is infeasible. The fallback is to search for a feasible design point, similar to the second step in the dynamic read stability algorithm.

Table 4.4 lists the progress of the MPFP search algorithm. Each iteration appears to center around the target T_{write} of 1 ns but the algorithm never converges to this value even after 8 iterations. There also appears to be some oscillation between each iteration which is caused by the linear fit of the sensitivities that underestimates the expected ΔT_{write} for a specific ΔV_{th} especially when the dependence is quadratic. Figure 4.12 plots the evolution of the importance sampler as a function of run number. The algorithm converges to a good estimate ($\rho < 0.1$) of fail bit rate within less than 1000 runs. It is surprising to note that even

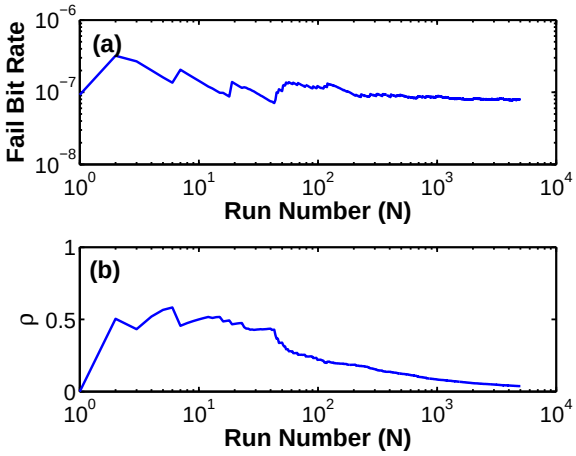


Figure 4.12: Evolution of (a) fail bit rate ($\hat{p_{IS}}$) and (b) convergence metric (ρ) as a function of run number, corresponding to T_{write} importance sampling.

though the MPFP search algorithm did not converge, the final search point still provided a good shift vector for the importance sampler. This demonstrates that using MPFP search algorithms in combination with importance sampling provides a good combination for estimating the reliability of SRAM. MPFP provides a good shift vector for the importance sampler while the importance sampling process relaxes convergence requirements of the MPFP search algorithm.

4.3.3 Dynamic Read Access

Dynamic read access is an equally important SRAM metric as dynamic read and writeability. The word-line pulse-width (T_{access}) together with the bit-line capacitance (C_{BL}) and SRAM bitcell read current determines how much bit-line voltage (V_{BL}) differential is generated. The goal is to create sufficient V_{BL} differential, greater than $V_{BL,target}$, to compensate for sense-amplifier offset voltages. The MPFP search algorithm for dynamic read access is defined in Equation 4.9. This formulation fixes T_{access} at $T_{access,target}$ and uses sensitivities of V_{BL} to V_{th} in the constraint. Sensitivities of V_{BL} at fixed T_{access} are used instead of sensitivities of T_{access} at fixed V_{BL} to simplify the SPICE simulations used for generating the sensitivities. Finding V_{BL} involves simulating a single SRAM access at $T_{access,target}$ while finding T_{access} of an SRAM bitcell for a target V_{BL} involves multiple simulations to find the critical word-line pulse-width. Linear sensitivities of V_{BL} to transistor parameters are sufficient to model the search space as the sensitivities appear to be mostly linear (Figure 4.13).

Table 4.5 tabulates the progress of the MPFP search algorithm. The algorithm converges to the target V_{BL} after 3 iterations. Cell sigma gradually decreases from 4.81 to 4.61 in the next few iterations as the algorithm searches for the most-probable point. The vector from

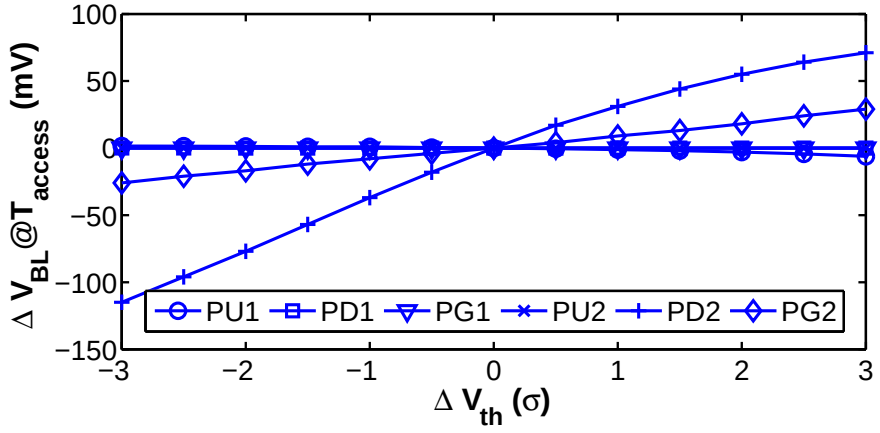


Figure 4.13: Sensitivity of final bit-line voltage at T_{access} to V_{th} variability.

Iteration	PU1	PD1	PG1	PU2	PD2	PG2	Cell Sigma	ΔV_{BL} (V)
1	-1.00	-1.00	0.00	0.00	3.00	3.00	4.47	0.17
2	-0.02	-0.00	0.00	0.00	3.75	3.12	4.88	0.12
3	-0.07	-0.00	0.00	0.00	4.16	2.41	4.81	0.10
4	-0.10	-0.00	0.00	-0.00	4.33	1.81	4.69	0.10
5	-0.11	-0.00	0.00	-0.00	4.39	1.48	4.64	0.10
6	-0.08	-0.00	0.00	-0.00	4.45	1.20	4.61	0.10
7	-0.11	-0.00	0.00	-0.00	4.43	1.27	4.61	0.10
8	-0.14	-0.00	0.00	-0.00	4.48	1.07	4.61	0.10

Table 4.5: Progress of T_{access} MPFP search algorithm. $\Delta V_{BL,target} = 0.1$ V

the last iteration is used as a shift vector for importance sampling to estimate the fail bit rate corresponding to dynamic read access. Figure 4.14 plots the evolution of the estimate of fail bit rate as a function of run number. The algorithm converges to a good estimate ($\rho < 0.1$) of fail bit rate within less than 1000 runs.

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \text{cell sigma} = \sqrt{(\boldsymbol{x}+\boldsymbol{x_{ci}})^T(\boldsymbol{x}+\boldsymbol{x_{ci}})} \\ \text{subject to} \quad & V_{BL,\boldsymbol{x_{ci}}} + \sum_i \frac{\delta V_{BL}}{\delta x_i} x_i = V_{BL,target} \qquad i = 1,\ldots,n \\ & T_{access} = T_{access,target} \\ & C_{BL} = C_{BL,target} \end{aligned} \tag{4.9}$$

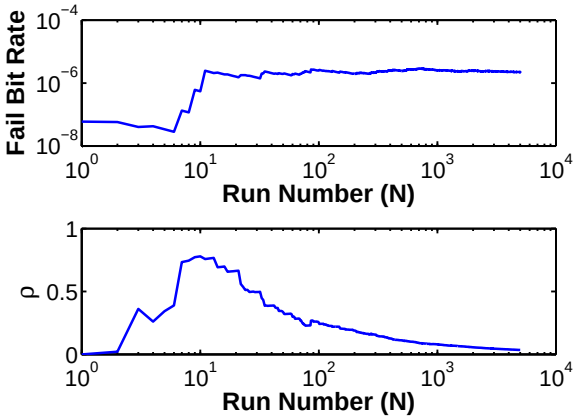


Figure 4.14: Evolution of (a) fail bit rate ($\hat{p_{IS}}$) and (b) convergence metric (ρ) as a function of run number, corresponding to dynamic read access importance sampling.

4.4 Importance Sampling with non-Gaussian Distributions

Most statistical analyses of SRAM model the distribution of the random variables as Gaussian distributions [12, 20, 32, 36, 74]. While these works have demonstrated good correlation between statistical estimations and actual measurements, it is uncertain whether this will continue to hold true in future technologies. It has been demonstrated that V_{th} variability due to random dopant fluctuation diverges from a Gaussian distribution 3 σ away from the mean in 13 nm technology devices [60]. This is based on 3-D atomistic simulation of 140000 devices where only the dopant locations in the channel are varied and all other sources of variability such as line-edge roughness are eliminated. This deviation is significant especially in SRAM design, where the failing bitcells typically sustain variability in the transistors of up to 4 or 5 σ (ref. Table 4.5). Furthermore, other sources of variability that are becoming increasingly important, such as RTS and NBTI, are known to be non-Gaussian (ref. Figure 3.30) [28]. This section attempts to take an initial step in introducing non-Gaussian distributions to importance sampling by considering lognormal distributions that can be used to model RTS and NBTI distributions.

4.4.1 Lognormal Distribution

Equation 4.10 lists the probability density function for a lognormal distribution. The name of the distribution originates from the fact that taking the logarithm of random variables derived from a lognormal distribution results in normally (Gaussian) distributed random variables. The algorithm developed in this chapter for analyzing reliability of SRAM involves

first finding the MPFP using an iterative search algorithm that was demonstrated to converge well for Gaussian distributions. To investigate whether this still holds true for situations where Gaussian distributions are combined with lognormal distributions, let us first consider the objective function of the MPFP search algorithm, listed in Equation 4.11.

By definition, the MPFP is a point where the joint probability density function (jpdf) is maximized. Taking the natural logarithm and negating the objective function (Equation 4.12), results in Equation 4.13. Variables x_i correspond to the Gaussian random variables while x_j correspond to the lognormal random variables. Substituting Equation 4.14 into Equation 4.13 results in Equation 4.15 which is a quadratic equation. The objective function is also convex because the coefficients of the quadratic terms are positive. The MPFP can therefore be solved for the global optimum, within a short run-time, even when lognormal random variables are mixed together with Gaussian random variables.

$$p(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(lnx-\mu)^2}{2\sigma^2}} \tag{4.10}$$

$$\underset{\boldsymbol{x}}{\text{maximize}} \quad jpdf(\boldsymbol{x}) \tag{4.11}$$

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad -ln(jpdf(\boldsymbol{x})) \tag{4.12}$$

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \sum_i \frac{(x_i-\mu_i)^2}{2\sigma_i^2} + \sum_j ln(x_j) + \frac{(ln(x_j)-\mu_j)^2}{2\sigma_j^2} \qquad i=1,\ldots,m \qquad j=1,\ldots,n. \tag{4.13}$$

$$y_j = ln(x_j) \tag{4.14}$$

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \sum_i \frac{(x_i-\mu_i)^2}{2\sigma_i^2} + \sum_j y_j + \frac{(y_j-\mu_j)^2}{2\sigma_j^2} \qquad i=1,\ldots,m \qquad j=1,\ldots,n. \tag{4.15}$$

4.4.2 Random Telegraph Signal and Dynamic Read Stability

The impact of random telegraph signal on dynamic read stability is evaluated using the importance sampling methodology adapted for non-Gaussian distributions. RTS introduces uncertainty in SRAM performance that is time-dependent. Chapter 3 demonstrates that RTS is governed by stochastic processes that determine the time until capture as well as time until emission of a carrier interacting with a particular trap. Furthermore, the magnitude of fluctuation in transistor intrinsic performance, such as I_{DS} or V_{th}, is also stochastic and is dependent on specific physical properties of each trap. Takeuchi and Aadithya *et al.* place

equal emphasis on the amplitude and dynamics of these traps in their evaluation of the impact of RTS on SRAM stability [64, 1]. These algorithms try to predict the probability of SRAM failures at specific times during the lifetime of a particular SRAM bitcell. Although the detail of their analysis is commendable, the extensive silicon characterization required to calibrate the models limit their practicality. Their proposed algorithms require extensive characterization of many traps in a large sample of transistors in order to collect sufficient statistics for modeling the RTS amplitudes and time constants. Emphasis on accurately modeling the time constants of the traps prohibits the use of the alternating-bias technique proposed in Section 3.2.4 which allows sampling of worst-case RTS amplitudes with relatively smaller sampling periods. Furthermore, estimating the probability of SRAM failure at a particular instance in time requires making assumptions of prior access patterns due to the influence of large-signal bias changes encountered during SRAM operations on trap probability (ref. Section 3.2.3). This section simplifies the problem by exploring the upper-bound and evaluating the probability of SRAM failure due to RTS throughout the lifetime of the SRAM bitcell. This simplification eliminates the need to model statistical distributions of RTS time constants and allows the use of alternating-bias RTS characterization for statistical RTS amplitude characterization.

Statistical distributions of RTS amplitude are modeled as lognormal distributions of threshold voltage degradation (ΔV_{th}). Figure 4.15 illustrates the probability density functions corresponding to NMOS and PMOS SRAM transistors. These lognormal distributions were fitted to measured fluctuation in drain currents (ΔI_{DS}) extracted from 45 nm CMOS SRAM transistors and normalized by the transconductance (g_m) of the sampled transistors to obtain ΔV_{th}. RTS amplitude is modeled as a simple shift in V_{th} instead of a more accurate empirical bias-dependent model (ref. Section 3.3.2) because compiled SPICE models with the empirical model are not available. This will likely impact accuracy of the results especially at higher bias voltages. Alternatively the empirical bias-dependent model can be implemented as a Verilog-A model with a slight increase in simulation runtime compared to a compiled SPICE model. Both pull-down and pass-gate NMOS transistors were modeled using similar distributions for simplicity, even though it has been demonstrated that these transistors have slightly different distributions (ref. Figure 3.30).

RTS introduces degradation in any of the 6 transistors of a 6T SRAM bitcell when electrons or holes occupy the traps in the respective transistors. Trap occupancy in any of these 6 transistors can either improve or degrade the dynamic read stability of the bitcell depending on the sensitivity of the read stability metric to degradation in the respective transistors. For example, based on the sensitivities illustrated in Figure 4.8, occupied traps in transistors $PU1$ and $PD2$ result in degradation of dynamic read stability (T_{read}) due to increased transistor V_{th}. On the other hand, occupied traps in transistors $PG2$ and $PD1$ result in improvement of T_{read}. Figure 4.16 illustrates the conceptual dependence of fail bit rate (cell sigma) on V_{DD} under different cases. The "Worst RTS" plot corresponds to the condition where all traps in transistors $PU1$ and $PD2$ are occupied, resulting in degraded fail bit rate corresponding to dynamic read stability. The "Best RTS" plot corresponds to

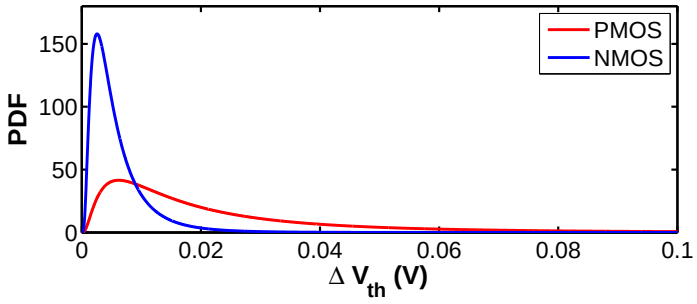


Figure 4.15: Simulated lognormal distributions of V_{th} fluctuation due to RTS in NMOS and PMOS transistors.

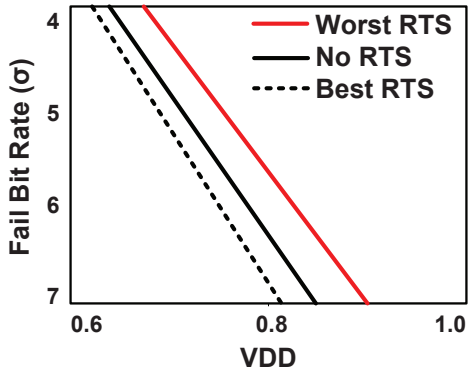


Figure 4.16: Conceptual plot demonstrating shift in dependencies between cell sigma and V_{DD} with worst-case RTS combination, best-case RTS combination, and no RTS.

the best-case combination where traps in transistors $PG2$ and $PD1$ are occupied, which improves SRAM reliability corresponding to dynamic read stability. The plot corresponding to "No RTS" represents the cell sigma one would observe from this SRAM bitcell if all traps in the 6 transistors of the bitcell were unoccupied. In this importance sampling analysis, lognormal distributions of ΔV_{th} due to RTS are only introduced to transistors $PU1$ and $PD2$ to evaluate the worst-case reliability degradation due to RTS.

The importance sampling algorithm for estimating dynamic read stability fail bit rates in the presence of non-Gaussian RTS distributions proceeds by first finding the MPFP using static read margin (RSNM). This is to assist in the convergence of the algorithm (ref. Section 4.3.1). Only Gaussian distributions corresponding to V_{th} variability are introduced during this phase to simplify the optimization problem and to minimize runtime. This does not impact quality of the final result as this initial phase is only meant for finding a good starting point for the next phase of the algorithm. The algorithm then proceeds to find the MPFP using T_{read} as the metric while introducing the lognormal distributions corresponding to

Iter	**PU1** (σ)	**PD1** (σ)	**PG1** (σ)	**PU2** (σ)	**PD2** (σ)	**PG2** (σ)	**RTS PU1** **(mV)**	**RTS PD2** **(mV)**	**JPDF**	T_{read} **(ns)**
1	1.27	-2.47	0.00	0.00	2.81	-2.45	15.9	4.49	12098	0.832
2	1.02	-2.72	0.00	-0.05	2.56	-2.2	12.3	3.49	44163	0.986
3	0.77	-2.77	0.00	-0.02	2.51	-2.09	9.62	3.07	83685	1.072
4	0.61	-2.52	0.00	-0.01	2.56	-2.34	7.86	3.03	99736	1.017

Table 4.6: Progress of T_{read} MPFP search algorithm with lognormal RTS distributions in $PU1$ and $PU2$. $T_{read,target} = 1.0$ ns

RTS in transistors $PU1$ and $PD2$ in addition to the Gaussian distributions corresponding to V_{th} variability. The optimization problem is formulated as a quadratic program with linear constraints. The objective function is simply the quadratic equation presented in Equation 4.15 which has been demonstrated to be a convex function. The linear constraints model the dependence of T_{read} on variability in the 6 SRAM transistors using locally evaluated gradients. A Taylor approximation of Equation 4.14 is used to translate the variables y_j in Equation 4.15 into x_j which is then multiplied by the locally evaluated gradients.

Table 4.6 lists four iterations of the MPFP search algorithm during the second phase where T_{read} is used as the read stability margin and lognormal RTS distributions are considered. Variables corresponding to Gaussian distributions of V_{th} variability in the 6 transistors are listed in terms of σ corresponding to the respective distributions, while the two variables corresponding to lognormal RTS distributions in $PU1$ and $PD2$ are listed in units of absolute mV. The joint probability density function (JPDF) of each iteration is calculated by multiplying the PDF of each variable which is estimated from their respective statistical distributions. Results in Table 4.6 indicate convergence to the target 1 ns T_{read} as well as an increase of the objective function (JPDF) in each iteration, which is the desired outcome of the optimization (ref. Equation 4.11). The MPFP estimated using this algorithm was verified through a 1,000,000 run Monte Carlo simulation. Out of the 1,000,000 simulations, only 22 runs had T_{read} less than 1 ns, which corresponds to dynamic read stability failures. The maximum JPDF from these 22 failing runs was only 27,754 which is less than the 99,736 obtained using the MPFP search algorithm. It is interesting to observe that the MPFP vector has relatively small RTS components in $PU1$ and $PD2$ which demonstrates the importance of accurate modeling of the bulk of RTS distributions.

Importance sampling using the original Gaussian and lognormal distributions, shifted by the MPFP vector, is then used to estimate the fail bit rate. Figure 4.17 plots the fail bit rate and figure of merit (ρ) obtained from the importance sampling algorithm, as a function of the number of samples (N). Also plotted on the same plots is the fail bit rate and figure of merit corresponding to conventional Monte Carlo simulations. The importance sampling algorithm quickly converges to the final fail bit rate within 5,000 samples which is indicated

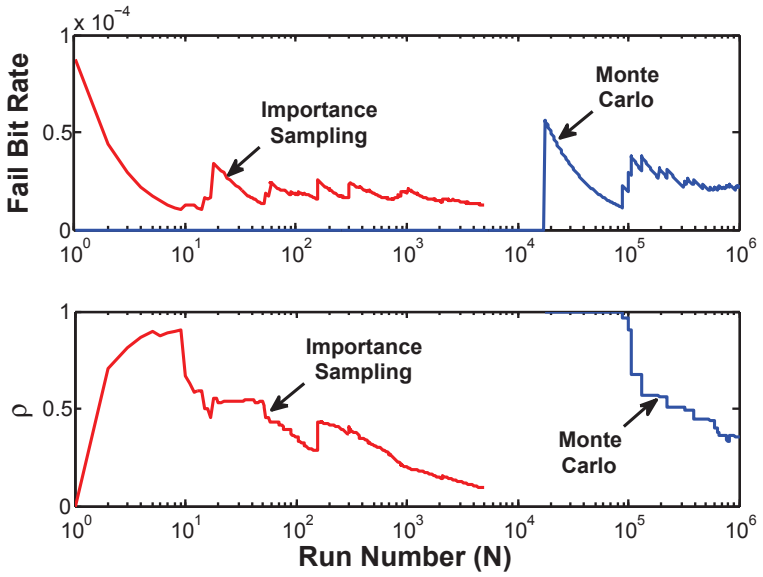


Figure 4.17: Evolution of (a) fail bit rate ($p\hat{}_{IS}$) and (b) convergence metric (ρ) as a function of run number, corresponding to T_{read} with lognormal RTS distributions evaluated using importance sampling and conventional Monte Carlo simulations.

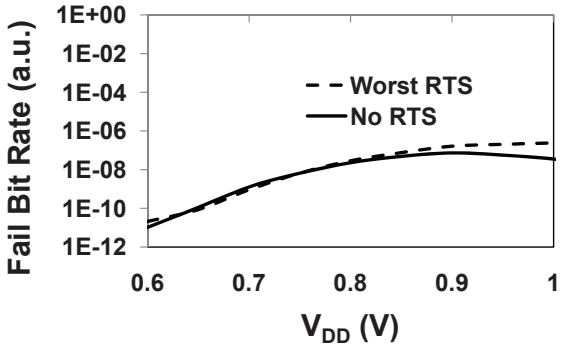


Figure 4.18: Fail bit rate corresponding to T_{read} as a function of V_{DD} without RTS and with worst-case RTS. $T_{read} = 1$ ns.

by the figure of merit (ρ) falling within the desired bound (0.1). Conventional Monte Carlo simulations however only encountered the first bit failure after 10,000 samples and requires more than 1,000,000 samples to converge to the same value as the fail bit rate predicted using importance sampling. The importance sampling algorithm therefore demonstrates a 1,000X reduction in runtime.

Figure 4.18 plots the fail bit rate corresponding to dynamic read stability as a function of V_{DD} estimated using importance sampling. The "No RTS" plot was estimated with only

Gaussian distributions of V_{th} variability introduced in the 6 transistors while the "Worst RTS" plot was estimated by introducing additional lognormal RTS ΔV_{th} distributions in transistors $PU1$ and $PD2$, which further degrades dynamic read stability. These results indicate that RTS causes significant degradation in SRAM fail bit rate, corresponding to dynamic read stability, at higher operating voltages. This degradation tapers off as V_{DD} is reduced. The estimation of dynamic read stability degradation due to RTS estimated in this section still needs to be verified through measurements. One possible source of discrepancy between these estimations and actual measurements might be due to the fact that RTS was modeled as fixed shifts in V_{th} instead of more accurate bias-dependent models. This results in over-estimation at higher voltages and under-estimation at lower voltages. These results however do demonstrate that non-Gaussian lognormal distributions can be introduced into importance sampling without significantly increasing convergence or runtime of the algorithm.

Chapter 5

Stochastic Optimization of SRAM Bitcell and Array

5.1 Introduction

Recognizing the dynamic nature of SRAM stability and access opens up a new horizon of SRAM optimization that is not available using conventional static margins. This chapter will explore this optimization space using importance sampling to evaluate the SRAM reliability resulting from each design choice. The first section explores design choices that can be made at the bitcell level, such as V_{DD}, bitcell variety, bitcell topology, and even the optimal process technology. The next section explores tradeoffs that can be made at a higher level when organizing the bitcells into arrays.

5.2 Bitcell Optimization

5.2.1 Global Process Variation

Robust SRAM design not only needs to account for local transistor mismatch, it also needs to account for within-die variation, within-wafer variation, as well as lot-to-lot variation [81, 57, 55]. Within the same die, an SRAM bitcell located close to the edge of the array will have different characteristics compared to a bitcell located at the center of an array. This can be caused by variation in stress due to proximity to shallow trench isolation (STI) and also thermal gradients encountered by the transistors during processing, due to different layout context [24, 48]. At the wafer level, variations in pattern density might cause systematic variations across the wafer during the chemical-mechanical polishing (CMP) step causing transistors in the center of the wafer to perform differently from transistors located at the boundaries. All these sources of variation can be summarized into two parameters that specify NMOS and PMOS performance. Foundries typically guarantee that the average

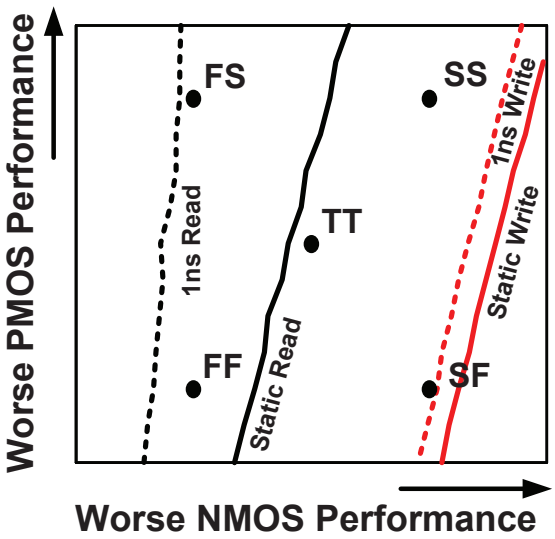


Figure 5.1: NMOS and PMOS global process variation space annotated with failure contours for read and write operation, as well as process corners [86].

NMOS and PMOS performance of each die will fall within defined fast and slow limits. Figure 5.1 illustrates the four process corners corresponding to both NMOS and PMOS at the fast corner (FF), both NMOS and PMOS at the slow corner (SS), NMOS at the fast corner but PMOS at the slow corner (FS), and NMOS at the slow corner but PMOS at the fast corner (SF). TT corresponds to the typical performance expected from both NMOS and PMOS transistors in the process technology.

Figure 5.1 is annotated with lines corresponding to failure boundaries for dynamic read stability (1 ns Read), static read stability (Static Read), dynamic writeability (1 ns Write), and static writeability (Static Write) [86]. The region to the right of the lines for read stability defines points in the space of global process variation where read stability is guaranteed to a certain probability (for example, 6σ). Even though the average NMOS and PMOS performance on a die might be at a particular point in the figure, local variation defines a cloud of points surrounding this particular point. SRAM reliability is guaranteed up to a specific probability level even in the presence of local variation. These results indicate that this particular bitcell, under the assumptions of the operating condition and target reliability levels made in this analysis, is not static read stable across the entire process window. This however does not reflect the true reliability of the bitcell as the FS and FF corners fit within the read stability region when evaluated using dynamic read stability with 1 ns pulse-width. Points to the left of the lines for writeability define global process regions where writeability is guaranteed up to a particular reliability level, according to the respective margins. The curve shifts to the left, with respect to static writeability, when dynamic writeability is used to evaluate the true write reliability of the SRAM bitcell but still bounds the SF corner.

These results demonstrate the importance of using dynamic stability to evaluate the reliability of SRAM bitcells across global process corners. Drawing conclusions just on static margins result in an overly pessimistic estimation of read stability while being slightly optimistic for writeability. Note that the FS corner presents the worst-case corner for read stability because the contour corresponding to dynamic read stability comes closest to this point while the SF corner presents the worst-case corner for write stability for similar reasons. Dynamic read access is critical at the slow NMOS corner and is independent of PMOS performance. A simple approach to guarantee SRAM reliability across global process variation is to verify read stability at the FS corner and writeability at the SF corner [54].

5.2.2 V_{DD} Scaling

This section explores the dependence of various SRAM stability metrics on V_{DD} scaling. Figure **??** plots estimated fail bit rates as a function of V_{DD}. All voltage applied on the bitcell (V_{WL}, $V_{precharge}$, V_{CELL}) are set to the corresponding V_{DD} value in this analysis. The bit-lines are assumed to be loaded with capacitance corresponding to 128 bitcells. Fail bit rate corresponding to static read stability failures increases exponentially with decreasing V_{DD}. This is expected as scaling V_{DD} directly impacts the eye opening in the butterfly curves corresponding to RSNM. The fail bit rate corresponding to dynamic read stability failures under 1 ns pulse-width access also increases exponentially initially with V_{DD} scaling. The fail bit rates corresponding to dynamic read stability are more than 4 orders of magnitude lower than the rates corresponding to static read stability in this case. Dynamic read stability failures actually start decreasing as V_{DD} is scaled past a certain threshold. The reduction in word-line voltage applied on the pass-gate transistors increases the on resistance of this transistor and therefore increases the time constant associated with disruption of the internal node during read access, relative to the word-line pulse-width of 1 ns.

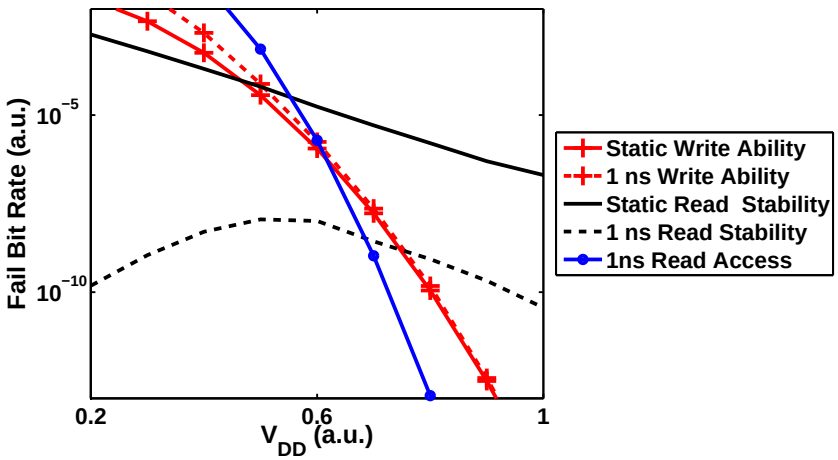


Figure 5.2: Dependence of fail bit rate on V_{DD} for static and dynamic writeability, read stability, and dynamic read access.

Successful read operation requires the state of the bitcell to be preserved as well as sufficient bit-line voltage differential created before enabling the sense amplifiers. Although the fail bit rate corresponding to dynamic read stability decreases with V_{DD} after a certain point, read failures are now dominated by read access failures which are observed to increase quite drastically with V_{DD} scaling. Figure 5.2 also indicates the fail bit rates corresponding to static writes and writes with 1 ns pulse-width. Fail bit rates increase when evaluating writeability of the bitcell under 1 ns pulse-width activation compared to static write margins. The divergence between these two lines however is minimal at high V_{DD} and only increases slightly at lower V_{DD}.

The fail bit rate of a bitcell is therefore dominated by different SRAM access operations at any particular value of V_{DD}. At higher voltages, dynamic read stability becomes the limiting factor of fail bit rate. Fail bit rates corresponding to writeability and read access however are highly sensitive to V_{DD} scaling and quickly become the limiting factor as V_{DD} is reduced.

5.2.3 Bitcell Variety

Multiple varieties of SRAM bitcells are typically offered in a certain process technology [5, 33]. These bitcells are sized differently and are typically targeted towards either high density or high performance (high read current) applications. Register files and first level caches might require high performance bitcells in order to keep up with operating frequencies of execution units while last level caches typically utilize high density bitcells to obtain large cache memories. High performance bitcells are usually designed with larger pass-gate and pull-down transistors to improve read access and writeability simultaneously at the expense

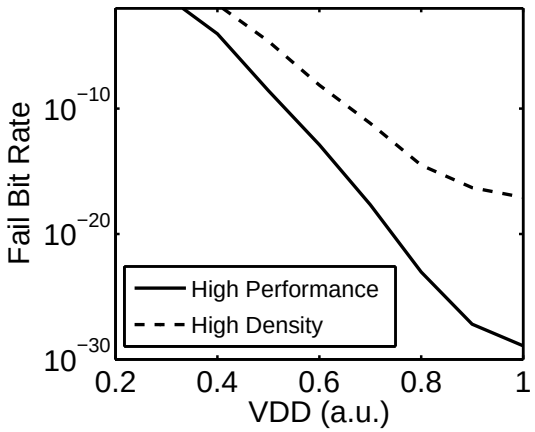


Figure 5.3: Fail bit rate as a function of V_{DD} for a high performance and high density bitcell from a particular process technology. The flattening of fail bit rate at higher V_{DD} is due to the weaker sensitivity of dynamic read stability to V_{DD} compared to other failure modes.

of larger bitcell area, relative to high density bitcells. The larger transistors also have the added benefit of smaller local variation, resulting in a lower V_{min}. Even if the difference in local variation between the high density and high performance bitcells is ignored, the high density bitcell might be less reliable from a manufacturing perspective because the process technology is typically pushed to the limits in order to achieve these small area bitcells.

Figure 5.3 compares the fail bit rates corresponding to a larger high performance bitcell and a smaller high density bitcell. Both of these bitcells are compared at similar bit-line capacitance loading and word-line pulse-width. Each of the curves in the plot correspond to the fail bit rates of the respective bitcells with dynamic read stability, dynamic read access, and dynamic writeability factored in. The high performance bitcell which incurs a 30% area penalty compared to the high density bitcell has much lower fail bit rates across the entire V_{DD} range. The high performance bitcell therefore has a lower V_{min} compared to the high density bitcell. Although the high performance bitcell is clearly more robust than the high density bitcell, the larger bitcell size might not be acceptable in large memory arrays, such as L2 and L3 caches, where area is one of the most important metrics.

5.2.4 8T SRAM

6T SRAM bitcells need to be optimized for both read stability and writeability simultaneously, both of which have conflicting requirements. These opposing requirements, together with increasing mismatch between the transistors, make it increasingly difficult to scale the size of bitcells in each technology node while trying to minimize V_{DD}. 8T SRAM bitcells, which share the same schematic as 6T SRAM except that one of the internal nodes is buffered

out through a separate read path, have been proposed to alleviate this problem [16]. This bitcell is robust enough to be utilized both in high performance microprocessors and sub-threshold SRAM [34, 79]. Figure 5.4 illustrates the schematic of an 8T SRAM bitcell. Read and write operations are performed separately through different ports. The write operation is performed through the 6T portion of the bitcell using the two bit-lines, BL and BLB, while activating the pass-gates by raising the write word-line (WWL). The read operation is performed by pre-charging the read bit-line (RBL) before enabling the read word-line (RWL).

The 6T portion of the 8T bitcell can be optimized for writeability by increasing the widths of the pass-gates while decreasing the widths of the pull-down transistors [16]. This not only increases the ratio between the pass-gate and the pull-down transistors, it also increases the trip point of the inverters, thus speeding up write completion. Pass-gate and pull-down transistors in optimized 8T bitcells are typically of the same width. This removes the diffusion area notch and thus eliminates the problem of corner rounding, resulting in less mismatch between transistors [56]. Figure 5.5 plots the fail bit rates of a 6T bitcell as well as an 8T bitcell, as a function of V_{DD}. The 8T bitcell was derived from the 6T bitcell by widening the pass-gates to match the widths of the pull-down transistors while adding two extra transistors. The 8T bitcell has almost four orders of magnitude improvement in fail bit rate at higher voltages. This improvement however tapers off as V_{DD} is scaled. Chang *et al.* noted that the extension of node CH to the gate of transistor RPD introduces extra capacitance on this node resulting in asymmetry in write performance [16]. In this case, a difference in fail bit rates of not more than 2X is observed when comparing write operations to node CL and node CH.

Figure 5.5 also plots the fail bit rate corresponding to dynamic read stability for both bitcells. Note that even though the read stability of the 8T bitcell is supposed to be worse than a 6T bitcell, due to the larger pass-gate transistors, the difference in fail bit rates between these two bitcell varieties is minimal when evaluated using dynamic read stability. Furthermore the 8T bitcell appears to have lower fail bit rates at low V_{DD} due to the larger pass-gate transistors having less variability. Fail bit rate at higher V_{DD} is clearly dominated by read stability failures. Even though the 8T bitcell has a dedicated read port, it is possible to subject the bitcells to a similar condition as read stability if column multiplexing is used and some bitcells in a row are half-selected. Column multiplexing therefore needs to be avoided in order to realize the maximum benefit of 8T bitcells at high V_{DD}. Other work has also suggested a write-back scheme to reinforce the stored data state every time a bitcell is half-selected [50]. Placing all bits corresponding to a word side-by-side within an array (as is the case when column multiplexing is not employed) poses an increased risk of multiple bit getting corrupted due to soft errors. This requires more error correction bits in order to detect and correct these errors [16]. The results in Figure 5.5 indicate that at $V_{DD} = 0.7$ a.u. and below, column multiplexing can still be used without a concern for half-select failures because the failures are dominated by write failures.

Dynamic writeability is clearly limiting the fail bit rate of an 8T bitcell at low V_{DD}.

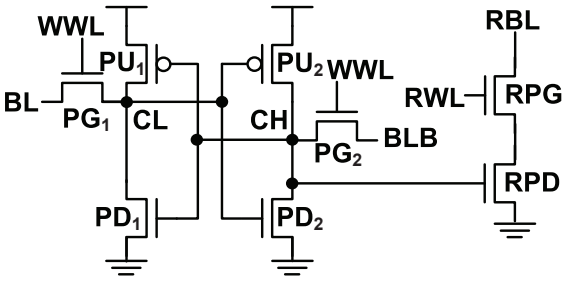


Figure 5.4: Schematic of 8T SRAM bitcell.

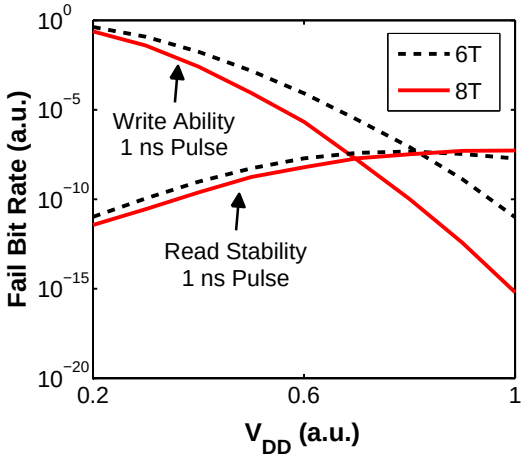


Figure 5.5: Fail bit rate of 8T and 6T bitcells as a function of V_{DD}.

Morita *et al.* demonstrated a functional 64-Mb 8T SRAM array at 0.42 V by keeping both the write word-line voltage (V_{WWL}) and read word-line voltage (V_{RWL}) at a fixed higher voltage while the array voltage was lowered [50]. Figure 5.6 and 5.7 plots the impact of this assist technique on SRAM fail bit rates under dynamic access. Keeping the word-line voltage at 1.0 a.u. while all other voltages such as the core array voltage and bit-line precharge voltage are scaled results in significant reduction in fail bit rate corresponding to the write operation. The stronger pass-gates however applies additional read stress on the bitcell, resulting in increased failures corresponding to dynamic read stability. Figure 5.6 demonstrates that it is possible to perform reliable write operations on an 8T bitcell at relatively short pulse-widths (1 ns) even at low values of supply voltage. Operating in this mode however requires avoiding half-selects at all cost in order to avoid read stability failures. Figure 5.7 also demonstrates significant reduction in dynamic read access failures when V_{RWL} is kept constant while V_{DD} is scaled. These plots are generated based on the assumption that 8 bitcells are sharing a read bit-line, which is typical in 8T array designs [16, 34]. Read access failures, however, do still depend strongly on V_{DD} because the gate

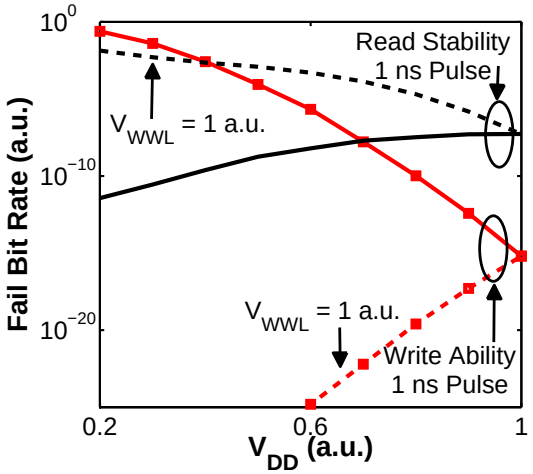


Figure 5.6: Dynamic stability of 8T bitcell with and without V_{WWL} boost.

voltage of the pull-down device in the read stack (RPG) is scaled together with V_{DD}. These failures can be reduced by making this device wider, or reducing the sense amplifier offset voltage margin. Alternatively, the read word-line can also be extended at low V_{DD} which can be done without concern for possible read stability failures, as is the case for the 6T bitcell.

The 8T bitcell therefore provides a robust alternative to the conventional 6T bitcell. This however comes at the expense of increased bitcell area. The smallest area overhead reported is 30% [15], however it is more realistic to expect area overheads of 50% to 100%. This is in addition to the extra peripheral area overhead required for short read bit-lines, dual word-line drivers, and no column multiplexing. These area costs might be acceptable for a register file or L1 cache where power and performance is more important but is most likely prohibitive in last level caches where memory density is the most important requirement.

5.2.5 Process Technology

Transistor variability due to random dopant fluctuation is clearly one of the main challenges of reliable SRAM design. V_{th} variability due to RDF has been demonstrated to be dependent on oxide thickness (T_{ox}), dielectric constant of the gate oxide (ϵ_{ox}), number of dopants in the channel (N), effective channel width (W_{eff}), and (L_{eff}) according to Equation 5.1 [49].

$$\sigma_{V_{th,RDF}} \propto \frac{T_{ox}}{\epsilon_{ox}} \frac{\sqrt[4]{N}}{\sqrt{W_{eff} L_{eff}}} \tag{5.1}$$

Historically, T_{ox} has been scaled at every technology node to control the impact of RDF as device dimensions ($W_{eff} L_{eff}$) were scaled. This trend however was slowed down when

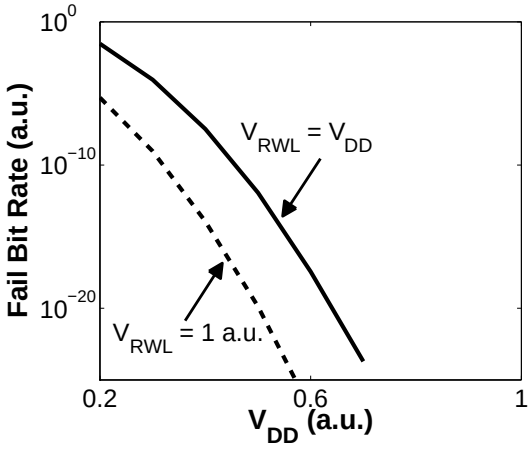


Figure 5.7: Dynamic access (1 ns pulse) fail bit rate of 8T bitcell with and without V_{RWL} boost.

gate leakage concerns limited T_{ox} scaling with conventional gate oxides at the 65 nm process technology node [41]. Introduction of high-K dielectrics to replace conventional gate oxides has effectively recovered this optimization knob by enabling further scaling of the effective gate oxide thickness, while keeping gate leakage in check.

Matching properties of transistors in various process technologies are usually specified by means of the Pelgrom number (A_{VT})[58]. This parameter is extracted experimentally by extracting the σ of the difference in V_{th} of two matched transistors for an ensemble of devices with different widths and lengths. A_{VT} corresponds to the best-fit slope to the results when plotted with $1/\sqrt{WL}$ on the x-axis and $\sigma\Delta V_{th}$ on the y-axis. Arnaud *et al.* recently reported A_{VT} of 2 mV μm in a 28 nm bulk CMOS process technology with high-K metal gates [5]. Even though SRAM V_{min} of 0.75 V has been reported in this process technology, it is still highly desirable to reduce V_{min} further especially in low power applications.

Intrinsic channel devices, such as fully depleted silicon-on-insulator (FDSOI) and FinFET, promise to reduce A_{VT} further by eliminating the problem of RDF. Instead of doping the channel to set the threshold voltage of the device, the channel is left at the intrinsic doping level. V_{th} is determined by the work-function of the metal gate and doping of the back-gate in FDSOI devices [21]. A_{VT} of 0.95 mV μm has been reported in an FDSOI device with 25 nm gate length [83], which is approximately a 2X improvement from bulk CMOS. These devices also boast steeper sub-threshold slope compared to bulk CMOS technologies which allows reducing V_{th} while keeping leakage currents at an acceptable level. This results in better gate overdrive ($V_{GS} - V_{th}$) thus allowing further V_{DD} scaling, without too much loss in performance.

In order to evaluate the impact of a 2X reduction in A_{VT} on SRAM reliability, the fail bit rate of a bitcell implemented either in bulk CMOS or FDSOI was evaluated. The bitcell

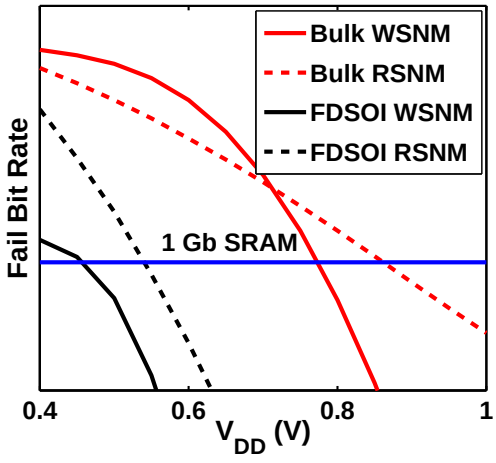


Figure 5.8: Comparison of fail bit rate for a similar SRAM bitcell implemented either in bulk CMOS or FDSOI, demonstrating more than 200 mV reduction in V_{min}

used in this analysis is optimized for the bulk CMOS process. To evaluate the V_{min} reduction obtained using an FDSOI process technology, the bulk transistor models were substituted with FDSOI transistor models. The FDSOI devices used in this analysis correspond to ultra thin-body devices with counter-doped ground planes (NMOS ground plane is doped P-type while PMOS ground plane is doped N-type) which offer the best short-channel control [21]. A common metal gate stack with a mid-gap work function is shared between NMOS and PMOS devices to reduce process complexity. No further V_{th} tuning of the FDSOI devices was performed. Figure 5.8 presents the results of this analysis. Static read and write margins (RSNM and WSNM) were used in this analysis, instead of dynamic stability metrics, to minimize runtime of the simulations. Based on these results, a 200 mV reduction in V_{min} is expected even if a bitcell is blindly ported to an FDSOI process technology, without any further V_{th} tuning or device sizing. This improvement in bitcell reliability comes primarily from the reduction in σV_{th}. Better short channel control in FDSOI technology also reduces drain-induced barrier lowering (DIBL) which results in improved read margins. The magnitude of the V_{min} reduction matches results obtained from statistical 3D atomistic simulations [17].

5.3 Array Optimization

This chapter has thus far focussed on techniques for optimizing a single SRAM bitcell to achieve the best performance in the presence of variability. In reality, an SRAM bitcell is usually arrayed and accessed using row and column peripheral circuitry such as word-line drivers and sense amplifiers. The array configuration and peripheral circuitry provide

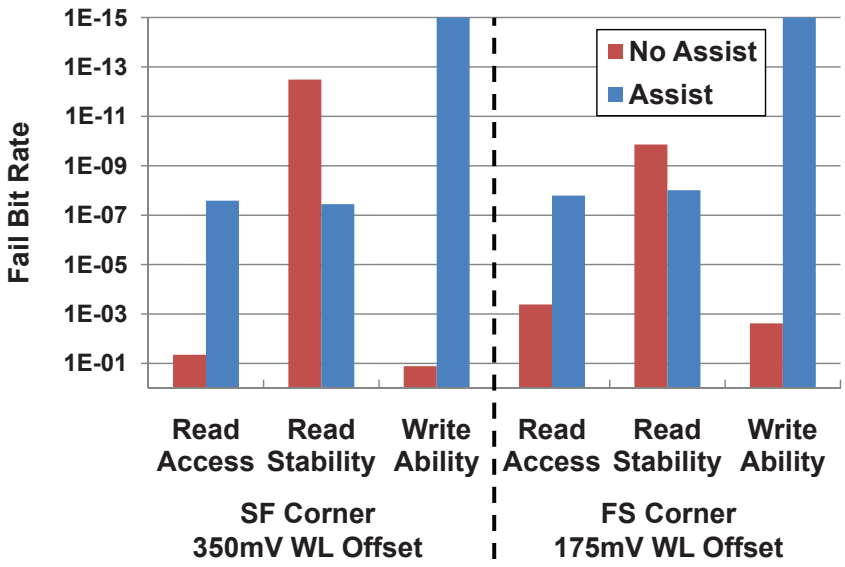


Figure 5.9: Impact of word-line voltage offset on SRAM stability across different process corners.

additional tuning knobs that can be used to optimize bitcell performance further.

5.3.1 Read and Write Assist

Section 5.2.1 demonstrated the need to satisfy both read and write requirements across global process corners when optimizing an SRAM bitcell. This is to ensure that the bitcell remains functional, within the desired specifications, regardless of location of the bitcell in the die, wafer, or the process conditions of the lot. This results in over-design of SRAM bitcells which translates to extra V_{MIN} guard-banding. Read and write assist techniques have been introduced as a technique to compensate for process corners and therefore eliminate wasteful guard bands [56, 14, 39].

Figure 5.9 plots the estimated fail bit rates corresponding to read access, read stability, and writeability at the *SF* and *FS* global process corners. This analysis was done at a lower operating voltage (both cell-supply and word-line voltage) to emulate failure conditions encountered at low V_{MIN} operating conditions. SRAM failures at low operating voltages are dominated by read access and writeability failures while dynamic read stability appears to be more reliable by up to 6 orders of magnitude. This discrepancy is strongly dependent on the process corner. For example, the *SF* global process corner degrades writeability and read access performance further, while improving read stability, resulting in an even larger difference between the fail bit rates of each operation.

The circuit design of a 6T SRAM bitcell allows trading off read stability for writeability

and read access performance. One way of realizing this is by adjusting the word-line voltage applied to a bitcell that is being accessed. Figure 5.9 plots the impact of applying a 350 mV word-line voltage offset (or boost) on the bitcell operating at the SF corner as well as the impact of applying a 175 mV word-line voltage offset on the bitcell operating at the FS corner. The exact value of word-line offset in each case was selected for optimized trade-off between read stability and writeability, as well as read access performance. Dynamic writeability appears to have a stronger sensitivity to word-line voltage offset compared to the other margins. This is observed in the larger reduction in writeability failure rates compared to read access, for the same amount of word-line offset. The write assist optimization in this case therefore simplifies into a tradeoff between read access performance and read stability. Sensors designed for detecting the process conditions of an array [39, 14] and automatically generating an assist bias can therefore be simplified to detect only 2 types of failures instead of all 3. This optimization however only applies to the bitcell design used in this analysis, where a large sensitivity of writeability to word-line voltage is observed.

The optimal word-line offset voltages in both global process corners improved the reliability of the bitcells from an unusable level to fail bit rates better than 1.0^{-7}. These levels are sufficient for low power embedded applications requiring small cache memories. The optimal word-line offset voltage in each process corner differs quite significantly. More word-line offset voltage (350 mV vs. 175 mV) is required in the SF corner to compensate for the poor performance of the NMOS pass-gate transistors, compared to the FS global process corner. In this case, applying a blanket 350 mV word-line offset voltage across all global process corners results in too much degradation of read stability in the FS process corner resulting in a net increase in SRAM failures. The optimal word-line offset voltage for achieving the best SRAM reliability is not only dependent on global process corners, it is also dependent on temperature and transistor aging. This dependence on multiple factors motivates the need for embedded sensors which detect the actual SRAM margins, from a statistically significant sample set, and apply optimal compensation [39, 14].

5.3.2 Bit-line Capacitance

Bit-line capacitance alters significantly the fail bit rates corresponding to dynamic read access and dynamic read stability. Figure 5.10(a) plots multiple curves corresponding to dynamic read stability and dynamic read access with increasing bit-line capacitance. Extra bit-line capacitance reduces the voltage droop on the bit-line that is being discharged by the bitcell, and therefore subjects the bitcell to more read stress due to larger V_{DS} on the pass-gate transistors. Failures corresponding to dynamic read access, on the other hand, also increases as a result of increased bit-line capacitance. The reduced bit-line droop due to the larger capacitance increases the number of read failures because not enough bit-line discharge voltage is accumulated to meet the sense amplifier offset voltage margin. The red curve in Figure 5.10 which corresponds to dynamic writeability does not shift with increasing bit-line capacitance because this analysis assumes that the bit-lines are statically driven either to

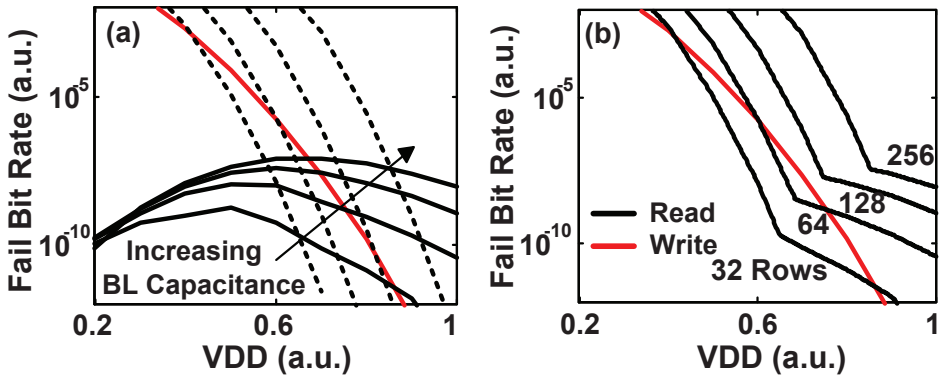


Figure 5.10: (a) Fail bit rate degradation with increasing bit-line capacitance. (b) Fail bit rate degradation as a function of the number of rows in a bit-line with worst-case fail bit rate for dynamic read access and dynamic read stability combined into a single curve.

V_{DD} or ground during write access.

The main array optimization parameter that determines the bit-line capacitance is the number of rows in in a sub-array. Other sources of bit-line capacitance such as capacitance of the column multiplexors and sense amplifiers are relatively smaller but can also be taken into account in array optimization. Figure 5.10(b) combines the curves corresponding to read stability and read access according to the worst-case fail bit rate and is grouped according to the number of rows in a sub-array. As expected, increasing the number of rows, increases the bit-line capacitance and therefore degrades the fail bit rate of the array. At high V_{DD}, where failures are usually limited by read stability failures, the lowest fail bit rate is obtained by minimizing the number of rows in a sub-array, as demonstrated in [84]. Assigning a smaller number of rows to a sub-array however incurs a larger area penalty and reduces the effective memory density. In a different scenario, where a certain fail bit rate level is tolerable, the array is optimized by choosing the number of rows such that V_{DD} can be reduced as much as possible without dynamic read stability or dynamic read access failures being the limiting factor. For example, if a fail bit rate of 10^{-7} is required, assigning 64 rows to a sub-array is sufficient to maximize V_{DD} reduction.

5.3.3 Array Segmentation

This section analyzes the impact of SRAM array segmentation on area, performance, and reliability of an SRAM array. Array segmentation refers to the organization of the total number of bitcells into various SRAM banks and sub-arrays. For example, Figure 5.11 illustrates two different configurations for a 64 kb SRAM array. Figure 5.11 has 256 bitcells sharing the same bit-line while Figure 5.11 (b) has only 64 bitcells sharing the same bitline. Section 5.3.2 has demonstrated the close relationship between bit-line capacitance and

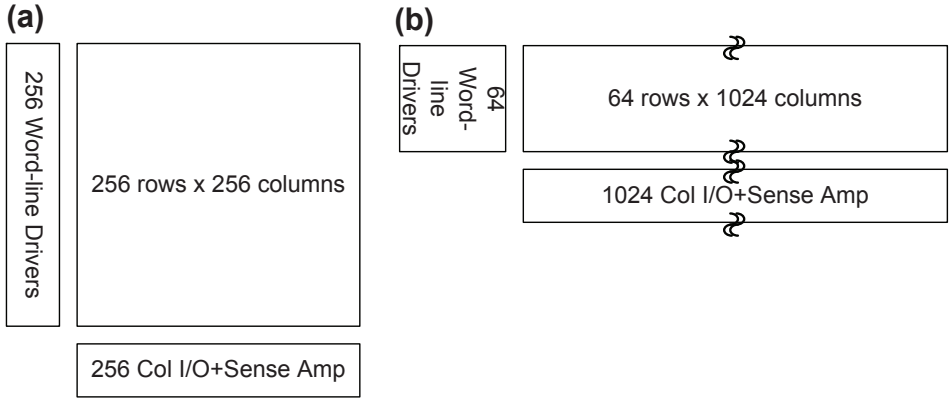


Figure 5.11: 64 kb SRAM array segmented into (a) 256 rows x 256 columns, and (b) 64 rows x 1024 columns.

dynamic stability of the SRAM bitcell. Since bit-line capacitance is determined primarily by the number of SRAM bitcells sharing the same bit-line in an SRAM array, array segmentation is expected to play a major role in determining the performance, and reliability of the memory.

Array segmentation to improve SRAM reliability however often comes at the price of increased area, which is one of the critical parameters in SRAM array design. For example, segmenting a 64 kb array into 64 rows and 1024 columns (ref. Figure 5.11) results in higher peripheral device overhead, compared to the other scheme with 256 rows, because every 64 bits requires a dedicated column I/O and sense-amplifier. These column peripheral devices are shared over 256 bits in the same column, in Figure 5.11 (a) resulting in lower peripheral device overhead. Segmenting the array with 64 rows however allows the same word-line driver to be shared between 1024 bits in a row, instead of 256 bits, resulting in a slightly lower row periphery device overhead. This reduction however is slightly offset by an increase in the size of the word-line drivers, necessary to drive the increased word-line capacitance of the additional columns.

Optimizing SRAM arrays using array segmentation is best handled using memory compilers which are equipped with power, area, and performance scaling characteristics of all the blocks that constitute an SRAM array. CACTI [52] is an open-source SRAM compiler that is frequently used by computer architects for evaluating the impact of cache configurations on power, performance, area, and compute efficiency. CACTI is equipped with fairly accurate analytical power, performance, and area models of interconnect and peripheral circuitry (sense-amplifiers, word-line drivers, column multiplexors, and decoders) that are calibrated to guidelines published in the ITRS. CACTI exhaustively searches all possible SRAM array segmentations to find the optimal configuration for minimizing a specified objective function which can be a weighted average of area, power, and performance. CACTI however treats an SRAM bitcell as a blackbox and is unaware of the impact of SRAM array

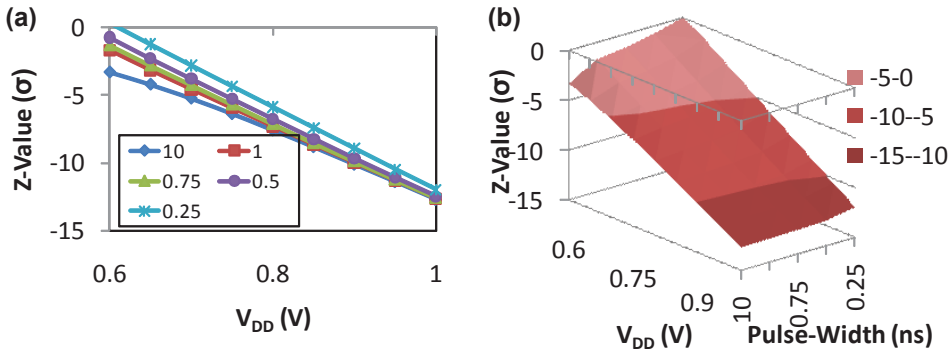


Figure 5.12: (a) Dependence of SRAM reliability (Z-value) corresponding to dynamic writeability as a function of V_{DD} across different access pulse-widths (ns), and (b) contour plot of SRAM reliability corresponding to dynamic writeability.

parameters such as column height, sense-amplifier offset, and access time on SRAM dynamic read stability, writeability, and read access. This results in CACTI reporting optimal array configurations with array heights of more than 1024 cells, which is not encountered in actual production-like SRAM arrays manufactured in modern process technologies [84, 39, 81].

The quality of results obtained from CACTI correlates better with actual SRAM bitcell performance by introducing a statistical SRAM bitcell dynamic stability model into the memory compiler. Dynamic stability is used to model the read stability, writeability, and read access performance of the bitcell in order to establish the relationship between SRAM array segmentation and bitcell performance. A statistical model (instead of a nominal-case or worst-case model) allows the memory compiler to provide a reliability estimate of the entire array and also use SRAM array reliability as one of the constraints in array optimization. This capability in CACTI is enabled by introducing lookup tables populated with pre-characterized performance of SRAM as a function of access time, V_{DD}, and bit-line capacitance. Figure 5.12 (a) plots an example of a lookup table corresponding to dynamic writeability. This lookup table models statistical SRAM writeability (Z-values) as a function of V_{DD} and access time. The bit-lines are assumed to be statically driven to V_{DD} and ground in this case, to eliminate the dependence on bit-line capacitance and simplify the lookup table by one dimension. Linear interpolation between points defined in the lookup table is used to obtain Z-values for other design points which provides the memory compiler the flexibility of choosing design points from a continuous range. Figure 5.12 (b) visualizes the surface obtained using this interpolation procedure. The lookup table for writeability is generated at the NMOS slow, PMOS fast corner, to capture the worst-case global process corner for writeability.

Dynamic read stability is modeled using a lookup table of Z-values that are dependent on V_{DD}, bit-line capacitance, and access time. Figure 5.13 plots a single cross-section of this 3-dimensional lookup table, keeping access time constant at 1 ns. This figure plots SRAM

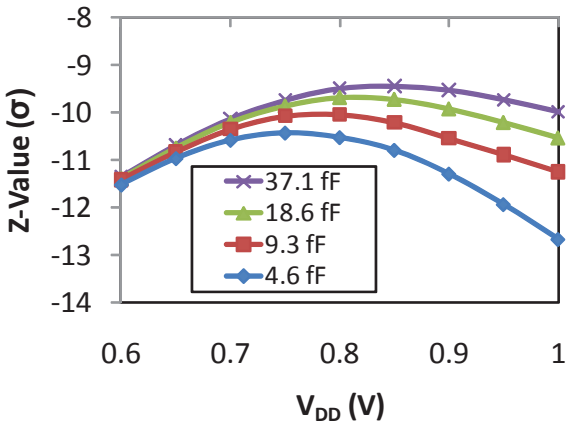


Figure 5.13: SRAM reliability corresponding to dynamic read stability as a function of V_{DD} across different bit-line capacitance (1 ns pulse-width).

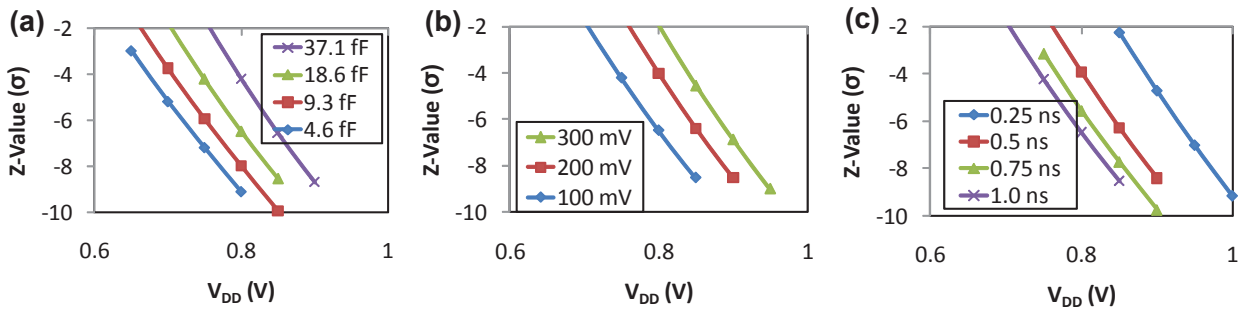


Figure 5.14: (a) Dependence of SRAM reliability (Z-value) corresponding to dynamic read access as a function of V_{DD} across different (a) bit-line capacitance, (b) sense-amplifier offset voltages, and (c) access pulse-widths.

reliability (Z-values) corresponding to dynamic read stability as a function of V_{DD} at different bit-line capacitances. The lookup table for dynamic read stability is generated at the NMOS fast, PMOS slow corner, to capture the worst-case process corner for read stability. The strong dependence of SRAM read stability on bit-line capacitance especially at higher V_{DD} demonstrates the importance of this design variable in SRAM array optimization. Degraded Z-values with larger bit-line capacitances enforces an upper limit to the number of SRAM rows that the memory compiler can select while still maintaining the required reliability of the cache memory. This avoids the non-ideality observed in memory compilers such as CACTI where the optimized array has an infeasible large number of rows.

Dynamic read access is modeled using a lookup table of Z-values that are dependent on V_{DD}, bit-line capacitance, access time, and sense-amplifier offset voltage. Figure 5.14 (a) plots the dependence of SRAM read access reliability on V_{DD} across different bit-line

capacitances, extracted from the 4-dimensional lookup table at a fixed sense-amplifier offset and access time. Smaller bit-line capacitance (shorter rows) results in better SRAM read access reliability at a given value of V_{DD}, as expected. Figure 5.14 (b) plots a snapshot taken from the lookup table at constant bit-line capacitance and access time, showing the dependence of read access reliability on sense-amplifier offset voltage. Sense-amplifier offset voltages of 100 mV correspond to small-signal differential sensing schemes while the larger offset voltages are needed for large-signal single-ended sensing schemes. While smaller offset voltage clearly results in better dynamic read access reliability, the differential-sensing sense-amplifiers with small offset voltage do incur additional peripheral area overhead, compared to much simpler single-ended inverters which have larger sense-amplifier offset voltages. Modeling the dependence of read access on sense-amplifier offset voltage provides the flexibility to the memory compiler for considering the tradeoffs in sense-amplifier selection on SRAM reliability. Figure 5.14 (c) shows the dependence of read access reliability across different access times. The lookup table for read access is generated at the NMOS slow, PMOS slow corner to capture the worst-case global process corner for read access.

These statistical models of SRAM dynamic stability were integrated into CACTI as a module that gets evaluated for every SRAM array segmentation considered by the memory compiler. A binary value is generated by this statistical module that indicates whether the array configuration meets the required reliability level or fails to meet this criteria. Array configurations that satisfy the reliability criteria are saved by the memory compiler in a list of feasible design points. This set of feasible design points is then used by the memory compiler to determine the optimal design point, based on a weighted objective function of area, power, and performance. Figure 5.15 (a) plots the minimum area for cache memories of different sizes implemented using two different bitcells from the same process technology. All designs were optimized for the same clock frequency and V_{DD}. Bitcell A is approximately 20% smaller than bitcell B in this technology and has weaker read and write stability as well as read access performance. One would expect that arrays implemented using bitcell A would be smaller than bitcell B however Figure 5.15 (a) indicates that the cache area of these two bitcell options are almost similar up to 16 MB arrays. At 64 MB, the cache memory implemented using the larger bitcell (B) ends up being smaller than the array implemented using the smaller bitcell (A). This unexpected results is explained using Figure 5.15 (b) which plots the column height of the SRAM sub-arrays, constituting the entire cache memory. Bitcell A was configured with 128 bitcells in a column to maintain dynamic read stability and to meet dynamic read access speed requirements. Bitcell B, on the other hand, is robust enough to be arranged in 256 height columns, for all the different cache configurations. The larger column height results in better array efficiency which makes the cache area corresponding to the larger bitcell competitive with the cache constructed out of smaller bitcells with poorer array efficiency. In the 64 MB cache configuration, the memory compiler resorts to segmenting bitcell A sub-arrays with 64 bitcell heights in order to meet the required reliability of a 64 MB cache memory. This further degrades array efficiency, resulting in the area of this cache memory being larger than the cache implemented using

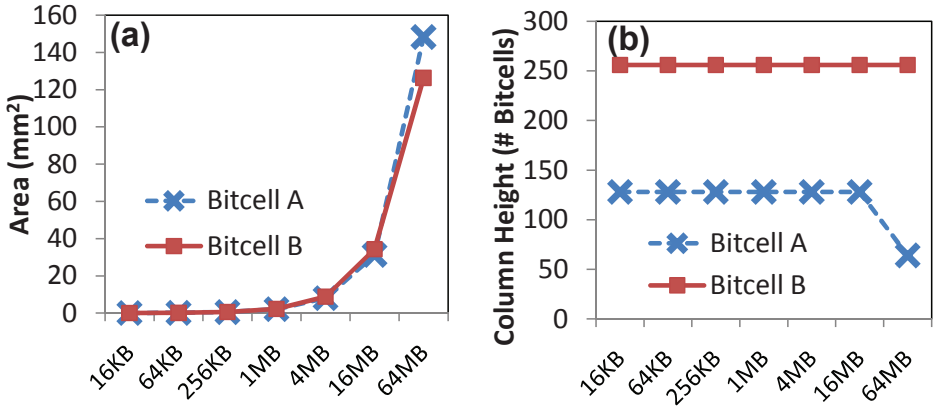


Figure 5.15: (a) Minimum cache area as a function of different cache sizes compiled using two different bitcells, (b) column height corresponding to the minimum-area solution for each cache size.

bitcell B.

Figure 5.16 evaluates the optimal cache memory configuration for a 1 MB array implemented using bitcell B, as a function of V_{DD}. At 1.0 V, the optimum cache configuration with minimum area consists of sub-arrays with 256 rows or 256 bitcells sharing the same bit-line. This bitcell is most likely able to function reliably with more than 256 bitcells per bit-line but this limit was imposed on the memory compiler to avoid unrealistically long bit-lines. Bitcell performance is degraded as V_{DD} is scaled down to 0.85 V. At this operating voltage, an SRAM array with 256 rows is most likely still functional in the nominal case but is not able to guarantee functionality of all bits in the 1 MB cache memory. The memory compiler is able to arrive at a solution that is still functional at this voltage by reducing the number of rows in each sub-array down to 128 bitcells, which improves dynamic read stability as well as read access performance. This incurs a slight increase in cache memory area due to degraded area efficiency of the shorter sub-arrays. The memory configuration can be further optimized for lower V_{DD} down to 0.73 V where the memory compiler is not able to find a feasible solution that meets the reliability requirements of a 1 MB array. These results demonstrate the importance of optimizing the array segmentation at the target minimum operating voltage in order to ensure proper functionality at all operating voltages. Designing SRAM arrays for low voltage operation however does incur a large area penalty. In this example the cache memory area increased by almost 2X as the operating voltage is decreased from 1 V down to 0.75 V.

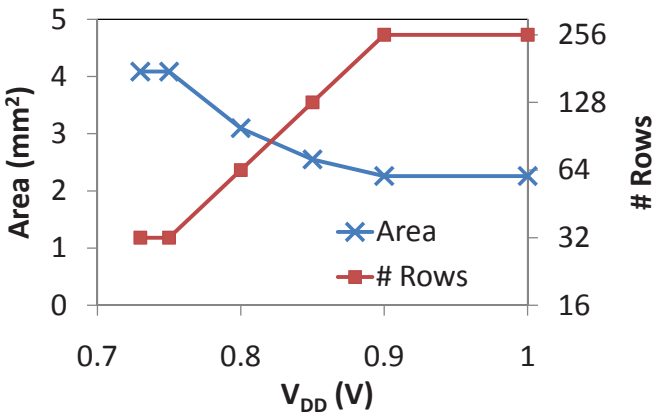


Figure 5.16: Optimum area of a 1MB cache memory optimized at different operating V_{DD} conditions. The corresponding optimum column height (# rows) is also plotted on the same graph.

Chapter 6

Conclusion

Nano-scale SRAM design is traditionally complicated by sources of variability related to physical variability in the structure of the transistors, such as random dopant distribution. This work identifies temporal sources of variability in transistor intrinsic parameters, caused by random telegraph signaling (RTS) noise, which is directly correlated with fluctuation in SRAM performance. A large-scale dynamic stability characterization architecture implemented in an early commercial low-power 45 nm CMOS process is used to experimentally verify the expected correlations between static and dynamic stability metrics. An optimization framework enabled by an importance sampling algorithm is used to design SRAM arrays with maximum array efficiency through joint-optimization between process technology, bitcell design, and array organization.

6.1 Key Contributions

The key contributions of this work are as follows:

- The correlations between static and dynamic stability metrics were analyzed through sensitivity analysis and Monte Carlo simulations. Theoretical expectations were verified experimentally on a 45 nm CMOS testchip with 10 ps accuracy.

Dynamic stability metrics for characterizing SRAM are known to provide better predictions of SRAM bit failures compared to traditional static noise margins. The relationship between these two types of metrics were quantitatively analyzed using sensitivity analyses and Monte Carlo simulations. It was demonstrated that while static and dynamic metrics do exhibit correlations in read access, read stability, and writeability, there exists differences in the sensitivities of the metrics to variability in the SRAM transistors which results in more than 10X differences between SRAM failure rates estimated using static and dynamic metrics. Furthermore, dynamic writeability exhibited sensitivity to degradation in more transistors compared to static write margins indicating increased impact of

process variability on actual SRAM operation. A characterization architecture capable of characterizing the dynamic stability of bitcells on a large scale, in-situ within a product-like SRAM array, was designed and validated in a 45 nm CMOS testchip. This characterization methodology makes extensive use of on-chip calibration circuitry to remove any uncertainty in the measurements and is well-suited for deployment in relatively immature process technologies for early technology exploration.

- Dynamics of RTS noise were investigated within the context of large-signal bias changes typically encountered in SRAM operation. SRAM access patterns to maximize the impact of traps on SRAM read stability and writeability were also developed and used to collect large-scale statistics efficiently.

RTS noise is traditionally analyzed under fixed bias conditions. Recognizing the fact that transistors within an SRAM bitcell are actually exposed to large-signal bias changes as read and write operations are applied on the bitcell, the large-signal bias response of traps were analyzed in detail. Traps contributing to RTS noise were observed to respond only after a few mili-seconds after a bias change was applied on the transistor. SRAM access patterns consisting of isolated and successive repeated operations were developed to take advantage of this fact and to speed up observation of the impact of these traps on SRAM bitcell failures, especially at the tails of the statistical distributions. This is a significant improvement from traditional sampling methods that rely on taking a large number of samples on the same bitcell to determine the impact of RTS on SRAM operation.

- The statistical impact of RTS noise on SRAM V_{MIN} was evaluated using a statistical model, calibrated to measured results from a 45 nm CMOS testchip. It was predicted and verified that even though RTS results in large degradation in the respective transistors, the actual degradation to V_{MIN} of a large array is only 50 mV.

RTS has been identified as one of the major sources of variability in nanoscale SRAM. Large temporal fluctuations of threshold voltage in highly scaled CMOS transistors have been reported and attributed to RTS. The statistical distributions of RTS noise amplitude were considered in this work, together with other existing sources of variability, within the context of the circuit operation of a 6T SRAM bitcell. A measurement technique for sampling worst-case threshold voltage variation due to RTS was used to sample statistical distributions of RTS noise amplitude. A numerical method was then used to convolve these empirical distributions to estimate the impact of RTS on SRAM V_{MIN}. Results of this analysis demonstrated that even though large RTS noise amplitudes might be present in nanoscale SRAM transistors, the interaction between the long-tailed RTS amplitude distributions and other Gaussian-like distributions results in a higher likelihood that the outlier bitcell limiting V_{MIN} of an array will have a smaller component of degradation caused by RTS.

- The importance sampling algorithm for estimating the robustness of SRAM was adapted for dynamic stability metrics demonstrating more than 4 orders of magnitude reduction in run-time compared to traditional Monte Carlo analysis.

Nanoscale SRAM design not only involves optimizing the nominal performance of a bitcell but also performance of bitcells at the tails of the statistical distributions which often determines the worst-case bounds of operation, such as V_{MIN}. Traditional statistical analysis methods either require making assumptions of the shape of the statistical distribution (such as Gaussian distribution) or generating a large number of samples (such as Monte Carlo analysis). This work develops the algorithms required to perform importance sampling using the more accurate dynamic stability metrics introduced in this work. These algorithms were demonstrated to converge to a solution within less than 10,000 samples. This technique was further extended to handle non-Gaussian distributions such as RTS noise amplitude.

- A methodology for performing joint-optimization between process technology, bitcell design, and array segmentation was developed, resulting in optimal design of large SRAM arrays.

SRAM design is an optimization space of many variables and is not only limited to just bitcell design. Choices that are made at different levels of design such as selection of process technology options (bulk vs. FDSOI), bitcell architecture (6T vs. 8T), circuit assist techniques (word-line voltage tuning), target operating V_{DD}, as well as array segmentation can vastly affect the area, power, and performance of the SRAM array. Recognizing the dynamic nature of SRAM stability and access opens up a new horizon of SRAM optimization that is not available using conventional static margins. The importance sampling algorithms developed in this work provide the means for quantifying the impact of each design choice while dynamic stability metrics provides the link between array segmentation and bitcell performance. This optimization tool exhaustively searches the entire optimization space using pre-characterized models of SRAM performance to select an optimal SRAM array design.

6.2 Future Work

Nanoscale SRAM design is becoming increasingly challenging due to the additional sources of variability introduced by advanced process technology options which are exacerbated by decreasing transistor sizes. This work has identified RTS as an increasingly significant source of variability and also analyzed its impact on SRAM reliability. Although this work has demonstrated that transistor variability due to random dopant fluctuation still ultimately contributes the most to degradation in SRAM V_{MIN}, it is interesting to investigate whether this still holds true in intrinsic channel devices such as FDSOI and FinFET where random dopant fluctuation is less of a problem. The potential dominance of RTS in transistor

variability makes it even more important to develop physical or empirical bias-dependent models which accurately models RTS amplitude and dynamics. Bias temperature instability (BTI) is yet another source of transistor variability due to aging which is becoming more important. Recently, it has been identified that RTS is related to BTI through the oxide traps which are inadvertently switched when BTI stress is applied [75, 23]. Further work is required to develop techniques to account for SRAM degradation due to RTS and BTI without accidentally double-counting any of these components.

The dynamic stability characterization architecture developed in this work has been proven on a 45 nm CMOS testchip and demonstrated the expected correlations (as well as unexpected outliers) between dynamic and static metrics. Although this architecture is capable of characterizing the performance of thousands of SRAM bitcells automatically, this methodology can be further extended to characterize millions of bitcells. The main limitation of the current architecture is due to the extra circuits and characterization steps required to achieve better than 10 ps timing resolution as well as simultaneous dynamic and static characterization capability on the same set of bitcells. Relaxing this strict timing requirement and removing the need for static characterization capability avoids the need for area-inefficient bit-line switches and word-line samplers and allows integration of more SRAM bitcells within a similar area. Integrated voltage regulators can also be integrated together with a more advanced BIST state machine to switch bias voltages applied to the array and allow more advanced read/write assist characterization while reducing test time.

This work has demonstrated the feasibility of introducing non-Gaussian statistical distributions of transistor variability into importance sampling for estimating SRAM reliability. In this work, RTS amplitude distributions were modeled as non-Gaussian distributions while threshold voltage fluctuation due to random dopant fluctuation was still assumed to be Gaussian. Recent 3D atomistic simulation studies have indicated that this will no longer hold true in the 13 nm gate length transistors [60]. Statistical tools for estimating SRAM reliability with non-Gaussian distributions will be needed to design large SRAM arrays in future technology nodes. Further work is therefore required to introduce arbitrary non-Gaussian distributions into an importance sampling framework while ensuring optimality of the results.

Bibliography

[1] K. Aadithya, S. Venogopalan, A. Demir, and J. Roychowdhury. MUSTARD: a coupled, stochastic/deterministic, discrete/continuous technique for predicting the impact of random telegraph noise on SRAMs and DRAMs. In *Design Automation Conference (DAC), 2011 48th ACM/EDAC/IEEE*, pages 292–297. IEEE, 2011.

[2] A.A. Abidi and C.R. Viswanathan. Flicker noise in CMOS transistors from subthreshold to strong inversion at various temperatures. *IEEE Transactions on Electron Devices*, 41(11):1965–1971, 1994.

[3] M. Agostinelli, J. Hicks, J. Xu, B. Woolery, K. Mistry, K. Zhang, S. Jacobs, J. Jopling, W. Yang, B. Lee, and Others. Erratic fluctuations of SRAM cache Vmin at the 90nm process technology node. In *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, volume 00, pages 655–658. IEEE, 2006.

[4] F Andrieu, O Weber, J Mazurier, O Thomas, J-P Noel, C. Fenouillet-Beranger, J-p Mazellier, P Perreau, T Poiroux, Y Morand, and Others. Low leakage and low variability Ultra-Thin Body and Buried Oxide (UT2B) SOI technology for 20nm low power CMOS and beyond. In *VLSI Technology (VLSIT), 2010 Symposium on*, pages 57–58. IEEE, 2010.

[5] F. Arnaud, A. Thean, M. Eller, M. Lipinski, YW Teh, M. Ostermayr, K. Kang, NS Kim, K. Ohuchi, J.P. Han, and Others. Competitive and cost effective high-k based 28nm CMOS technology for low power applications. In *Electron Devices Meeting (IEDM), 2009 IEEE International*, pages 1–4. IEEE, 2010.

[6] A. Asenov, R. Balasubramaniam, A.R. Brown, and J.H. Davies. RTS amplitudes in decananometer MOSFETs: 3-D simulation study. *Electron Devices, IEEE Transactions on*, 50(3):839–845, March 2003.

[7] M. Ball, J. Rosal, R. McKee, Wk. Loh, T. Houston, R. Garcia, J. Raval, D. Li, R. Hollingsworth, R. Gury, and Others. A Screening Methodology for VMIN Drift in SRAM Arrays with Application to Sub-65nm Nodes. In *Electron Devices Meeting, 2006. IEDM'06. International*, number 972, pages 1–4. IEEE, 2007.

[8] A. Bhavnagarwala, S. Kosonocky, C. Radens, K. Stawiasz, R. Mann, Q. Ye, and K. Chin. Fluctuation limits & scaling opportunities for CMOS SRAM cells. In *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, volume 00, pages 659–662. IEEE, 2006.

[9] T. Boutchacha, G. Ghibaudo, and B. Belmekki. Study of low frequency noise in the 0.18 μm silicon CMOS transistors. *ICMTS 1999. Proceedings of 1999 International Conference on Microelectronic Test Structures (Cat. No.99CH36307)*, 12(March):84–88, 1999.

[10] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[11] J.P. Campbell, J. Qin, K.P. Cheung, L.C. Yu, J.S. Suehle, A. Oates, and K. Sheng. Random telegraph noise in highly scaled nMOSFETs. *2009 IEEE International Reliability Physics Symposium*, pages 382–388, 2009.

[12] A. Carlson. *Device and Circuit Techniques for Reducing Variation in Nanoscale SRAM Copyright by Andrew Evert Carlson*. PhD thesis, University of California, Berkeley, 2008.

[13] A. Carlson, Z. Guo, S. Balasubramanian, L.-t. Pang, K. Liu, and B. Nikolic. FinFET SRAM with enhanced read/write margins. In *International SOI Conference, 2006 IEEE*, volume 10, pages 105–106. IEEE, October 2007.

[14] A. Carlson, Z. Guo, L.T. Pang, T.J.K. Liu, and B. Nikolic. Compensation of systematic variations through optimal biasing of SRAM wordlines. In *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, number Cicc, pages 411–414. IEEE, September 2008.

[15] L. Chang, D.M. Fried, J. Hergenrother, J.W. Sleight, R.H. Dennard, R.K. Montoye, L. Sekaric, S.J. McNab, A.W. Topol, C.D. Adams, and Others. Stable SRAM cell design for the 32 nm node and beyond. In *VLSI Technology, 2005. Digest of Technical Papers. 2005 Symposium on*, pages 128–129. IEEE, 2005.

[16] L. Chang, R. K. Montoye, Y. Nakamura, K. A. Batson, R. J. Eickemeyer, R. H. Dennard, W. Haensch, and D. Jamsek. An 8T-SRAM for Variability Tolerance and Low-Voltage Operation in High-Performance Caches. *IEEE Journal of Solid-State Circuits*, 43(4):956–963, April 2008.

[17] S. Changhwan, M. H. Cho, Y. Tsukamoto, B.-Y. Nguyen, C. Mazure, B. Nikolic, and T.-J. K. Liu. Performance and Area Scaling Benefits of FD-SOI Technology for 6-T SRAM Cells at the 22-nm Node. *IEEE Transactions on Electron Devices*, 57(6):1301–1309, June 2010.

[18] X. Deng, W.K. Loh, B. Pious, T.W. Houston, L. Liu, B. Khan, and D. Corum. Characterization of bit transistors in a functional SRAM. In *VLSI Circuits, 2008 IEEE Symposium on*, pages 44–45. IEEE, June 2008.

[19] A Dixit, KG Anil, E Baravelli, P Roussel, A Mercha, C Gustin, M Bamal, E. Grossar, R. Rooyackers, E. Augendre, and Others. Impact of stochastic mismatch on measured SRAM performance of FinFETs with resist/spacer-defined fins: Role of line-edge-roughness. In *Electron Devices Meeting, 2006. IEDM'06. International*, volume 3001, pages 1–4. IEEE, 2007.

[20] L. Dolecek, M. Qazi, and D. Shah. Breaking the simulation barrier: SRAM evaluation through norm minimization. *2008 IEEE/ACM International Conference on Computer-Aided Design*, pages 322–329, November 2008.

[21] C Fenouillet-Beranger, O Thomas, and P Perreau. Efficient Multi-VT FDSOI technology with UTBOX for low power circuit design. In *VLSI Technology (VLSIT), 2010 Symposium on*, pages 65–66, 2010.

[22] G. Ghibaudo, O. Roux, Ch. Nguyen-Duc, F. Balestra, and J. Brini. Improved Analysis of Low Frequency Noise in Field-Effect MOS Transistors. *Physica Status Solidi (a)*, 124(2):571–581, April 1991.

[23] T. Grasser, H. Reisinger, W. Goes, T. Aichinger, P. Hehenberger, P.-J. Wagner, M. Nelhiebel, J. Franco, and B. Kaczer. Switching oxide traps as the missing link between negative bias temperature instability and random telegraph noise. In *Electron Devices Meeting (IEDM), 2009 IEEE International*, pages 1–4. IEEE, December 2010.

[24] Z. Guo. *Large-Scale Variability Characterization and Robust Design Techniques for Nanoscale SRAM*. PhD thesis, University of California, Berkeley, 2009.

[25] R.R. Harrison and C. Charles. A low-power low-noise cmos for amplifier neural recording applications. *IEEE Journal of Solid-State Circuits*, 38(6):958–965, June 2003.

[26] E. Hoekstra. Large signal excitation measurement techniques for RTS noise in MOSFETs. In *EUROCON 2005 - The International Conference on "Computer as a Tool"*, volume 2, pages 1863–1866. Ieee, 2006.

[27] F Hooge. Discussion of recent experiments on 1/ noise. *Physica*, 60(1):130–144, July 1972.

[28] V Huard, C Parthasarathy, C Guerin, T Valentin, E Pion, M Mammasse, N Planes, and L Camus. NBTI degradation: From transistor to SRAM arrays. In *Reliability Physics Symposium, 2008. IRPS 2008. IEEE International*, pages 289–300. IEEE, 2008.

[29] KK Hung, PK Ko, C. Hu, and YC Cheng. Random Telegraph Noise of Deep-Submicrometer MOSFET's. *Electron Device Letters, IEEE*, 11(2):90–92, 1990.

[30] ITRS. International Technology Roadmap for Semiconductors 2009 Update System Drivers, 2009.

[31] R. Joshi, R. Houle, K. Batson, D. Rodko, P. Patel, W. Huott, R. Franch, Y. Chan, D. Plass, S. Wilson, and Others. 6.6+ GHz Low Vmin, read and half select disturb-free 1.2 Mb SRAM. In *VLSI Circuits, 2007 IEEE Symposium on*, pages 250–251. IEEE, 2007.

[32] R.V. Joshi, S. Mukhopadhyay, D.W. Plass, Y.H. Chan, and A. Devgan. Variability analysis for Sub-100 nm PD/SOI CMOS SRAM Cell. *Proceedings of the 30th European Solid-State Circuits Conference*, pages 211–214.

[33] E. Josse, S. Parihar, O. Callen, P. Ferreira, C. Monget, a. Farcy, M. Zaleski, D. Villanueva, R. Ranica, M. Bidaud, D. Barge, C. Laviron, N. Auriac, C. Le Cam, S. Harrison, S. Warrick, F. Leverd, P. Gouraud, S. Zoll, F. Guyader, E. Perrin, E. Baylac, J. Belledent, B. Icard, B. Minghetti, S. Manakli, L. Pain, V. Huard, G. Ribes, K. Rochereau, S. Bordez, C. Blanc, a. Margain, D. Delille, R. Pantel, K. Barla, N. Cave, and M. Haond. A Cost-Effective Low Power Platform for the 45-nm Technology Node. *2006 International Electron Devices Meeting*, pages 1–4, December 2006.

[34] R. Jotwani, S. Sundaram, S. Kosonocky, A. Schaefer, V. Andrade, G. Constant, A. Novak, and S. Naffziger. An x86-64 Core Implemented in 32nm SOI CMOS. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, number February 2009, pages 106–107. IEEE, 2010.

[35] R. Kanj, R. Joshi, and S. Nassif. Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events. *2006 43rd ACM/IEEE Design Automation Conference*, (x):69–72, 2006.

[36] D. E. Khalil, M. Khellah, N.-S. Kim, Y. Ismail, T. Karnik, and V. De. SRAM dynamic stability estimation using MPFP and its applications. *Microelectronics Journal*, 40(11):1523–1530, November 2009.

[37] D.E. Khalil, M. Khellah, N.S. Kim, Y. Ismail, T. Karnik, and V.K. De. Accurate Estimation of SRAM Dynamic Stability. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 16(12):1639–1647, 2008.

[38] M.J. Kirton and M.J. Uren. Noise in solid-state microstructures: A new perspective on individual defects, interface states and low-frequency (1/) noise. *Advances in Physics*, 38(4):367–468, January 1989.

[39] P. Kolar, E. Karl, U. Bhattacharya, F. Hamzaoglu, H. Nho, Y.G. Ng, Y. Wang, and K. Zhang. A 32 nm high-k metal gate SRAM with adaptive dynamic stability enhancement for low-voltage operation. *Solid-State Circuits, IEEE Journal of*, 46(1):76–84, 2011.

[40] J S Kolhatkar, E Hoekstra, C Salm, Van Der Wel, and H Wallinga. RTS Noise in MOSFETs under Steady-State and Large-Signal Excitation. In *2004 IEEE International Electron Devices Meeting (IEDM)*, pages 759–762, 2004.

[41] K.J. Kuhn. Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS. In *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pages 471–474. IEEE, 2008.

[42] S. Lee, H.-J. Cho, Y. Son, D. S. Lee, and H. Shin. Characterization of oxide traps leading to RTN in high-k and metal gate MOSFETs. In *Electron Devices Meeting (IEDM), 2009 IEEE International*, pages 1–4. IEEE, December 2010.

[43] A.J. Lelis, T.R. Oldham, H.E. Boesch, and F.B. McLean. The nature of the trapped hole annealing process. *IEEE Transactions on Nuclear Science*, 36(6):1808–1815, 1989.

[44] S.T. Martin, G.P. Li, E. Worley, and J. White. The gate bias and geometry dependence of random telegraph signal amplitudes [MOSFET]. *Electron Device Letters, IEEE*, 18(9):444–446, 1997.

[45] A L McWhorter. 1/f noise and germanium surface properites. *Semiconductor Surface Physics*, pages 207–228, 1957.

[46] H. Miki, N. Tega, Z. Ren, C. P. D. Emic, Y. Zhu, D. J Frank, M. A. Guillorn, D. Park, W. Haensch, and K. Torii. Hysteretic Drain-Current Behavior Due to Random Telegraph Noise in Scaled-Down FETs with High-K/Metal-Gate Stacks. In *2010 IEEE International Electron Devices Meeting (IEDM)*, pages 620–623, 2010.

[47] H. Mikoshiba. 1/f noise in n-channel silicon-gate MOS transistors. *Electron Devices, IEEE Transactions on*, 29(6):965–970, June 2005.

[48] T Miyashita, T Kubo, YS Kim, M Nishikawa, Y Tamura, J Mitani, M Okuno, T Tanaka, H Suzuki, T Sakata, and Others. A study on millisecond annealing (MSA) induced layout dependence for flash lamp annealing (FLA) and laser spike annealing (LSA) in multiple MSA scheme with 45 nm high-performance technology. In *Electron Devices Meeting (IEDM), 2009 IEEE International*, pages 1–4. IEEE, 2010.

[49] T. Mizuno and J. Okumtura. Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's. *Electron Devices, IEEE*, 41(11):2216–2221, 2002.

[50] Y. Morita, H. Fujiwara, H. Noguchi, Y. Iguchi, K. Nii, H. Kawaguchi, and M. Yoshimoto. An Area-Conscious Low-Voltage-Oriented 8T-SRAM Design under DVS Environment. In *VLSI Circuits, 2007 IEEE Symposium on*, volume 18, pages 256–257. IEEE, 2007.

[51] Y. Morita, K. Nose, K. Noguchi, S. Takami, K. Goto, Y. Aimoto, A. Kimura, and M. Mizuno. Small-defect detection in sub-100nm SRAM cells using a WL-pulse timing-margin measurement scheme. In *VLSI Circuits (VLSIC), 2010 IEEE Symposium on*, pages 37–38. IEEE, 2010.

[52] N. Muralimanohar, R. Balasubramonian, and N. P. Jouppi. CACTI 6 . 0 : A Tool to Model Large Caches CACTI 6 . 0 : A Tool to Model Large Caches, 2009.

[53] T. Nagumo, K. Takeuchi, T. Hase, and Y. Hayashi. Statistical Characterization of Trap Position , Energy , Amplitude and Time Constants by RTN Measurement of Multiple Individual Traps. In *2010 IEEE International Electron Devices Meeting (IEDM)*, pages 628–631, 2010.

[54] K. Nii, M. Yabuuchi, Y. Tsukamoto, S. Ohbayashi, S. Imaoka, H. Makino, Y. Yamagami, S. Ishikura, T. Terano, T. Oashi, and Others. A 45-nm bulk CMOS embedded SRAM with improved immunity against process and temperature variations. *Solid-State Circuits, IEEE Journal of*, 43(1):180–191, January 2008.

[55] B. Nikolic, Bastien Giraud, Zheng Guo, L.T. Pang, J.H. Park, and S.O. Toh. Technology variability from a design perspective. In *Custom Integrated Circuits Conference (CICC), 2010 IEEE*, pages 1–8. IEEE, 2010.

[56] S. Ohbayashi, M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Imaoka, Y. Oda, T. Yoshihara, M. Igarashi, M. Takeuchi, H. Kawashima, and Others. A 65-nm SoC embedded 6T-SRAM designed for manufacturability with read and write operation stabilizing circuits. *Solid-State Circuits, IEEE Journal of*, 42(4):820–829, 2007.

[57] L.T. Pang, K. Qian, C.J. Spanos, and B. Nikolic. Measurement and analysis of variability in 45 nm strained-Si CMOS technology. *Solid-State Circuits, IEEE Journal of*, 44(8):2233–2243, 2009.

[58] M.J.M. Pelgrom, a.C.J. Duinmaijer, and a.P.G. Welbers. Matching properties of MOS transistors. *IEEE Journal of Solid-State Circuits*, 24(5):1433–1439, October 1989.

[59] K. S. Ralls, W. J. Skocpol, L. D. Jackel, R. E. Howard, L. A. Fetter, R. W. Epworth, and D. M. Tennant. Discrete Resistance Switching in Submicrometer Silicon Inversion Layers: Individual Interface Traps and Low-Frequency (1/f) Noise. *Phys. Rev. Lett.*, 52(3):228–231, 1984.

[60] D. Reid, C. Millar, G. Roy, S. Roy, and A. Asenov. Analysis of Threshold Voltage Distribution Due to Random Dopants: A 100 000-Sample 3-D Simulation Study. *Electron Devices, IEEE Transactions on*, 56(10):2255–2263, October 2009.

[61] E. Seevinck, FJ List, and J. Lohstroh. Static-Noise Margin Analysis of MOS SRAM Cells. *Solid-State Circuits, IEEE Journal of*, 22(5):748–754, 2002.

[62] M. Sharifkhani and M. Sachdev. SRAM cell stability: A dynamic perspective. *Solid-State Circuits, IEEE Journal of*, 44(2):609–619, February 2009.

[63] N. Shibata, H. Kiya, S. Kurita, H. Okamoto, M. Tan'no, and T. Douseki. A 0.5-V 25-MHz 1-mW 256-kb MTCMOS/SOI SRAM for solar-power-operated portable personal digital equipment-sure write operation by using step-down negatively overdriven bitline scheme. *Solid-State Circuits, IEEE Journal of*, 41(3):728–742, 2006.

[64] K. Takeuchi, T. Nagumo, and T. Hase. Comprehensive SRAM design methodology for RTN reliability. In *VLSI Technology (VLSIT), 2011 Symposium on*, pages 130–131. IEEE, 2011.

[65] N. Tega, H. Miki, F. Pagette, DJ Frank, A. Ray, MJ Rooks, W. Haensch, and K. Torii. Increasing Threshold Voltage Variation due to Random Telegraph Noise in FETs as Gate Lengths Scale to 20 nm. In *VLSI Technology, 2009 Symposium on*, pages 50–51. IEEE, 2009.

[66] N. Tega, H. Miki, Z. Ren, C. P. D'Emic, Y. Zhu, D. J. Frank, J. Cai, M. a. Guillorn, D.-G. Park, W. Haensch, and K. Torii. Reduction of random telegraph noise in High-/ metal-gate stacks for 22 nm generation FETs. *2009 IEEE International Electron Devices Meeting (IEDM)*, pages 1–4, December 2009.

[67] N. Tega, H. Miki, M. Yamaoka, H. Kume, T. Mine, T. Ishida, Y. Mori, and R. Yamada. Impact of threshold voltage fluctuation due to random telegraph noise on scaled-down SRAM. *2008 IEEE International Reliability Physics Symposium*, pages 541–546, April 2008.

[68] A Teramoto, T Fujisawa, K Abe, S Sugawa, and T Ohmi. Statistical evaluation for trap energy level of RTS characteristics. In *VLSI Technology (VLSIT), 2010 Symposium on*, volume 38, pages 99–100. IEEE, 2010.

[69] S. O. Toh, T.-J. K. Liu, and B. Nikolic. Impact of Random Telegraph Signaling Noise on SRAM Stability. In *VLSI Technology (VLSIT), 2011 Symposium on*, number 510, pages 204–205, 2011.

[70] S. O. Toh, Y. Tsukamoto, Z. Guo, L. Jones, T.-J.K. Liu, and B. Nikolic. Impact of Random Telegraph Signals on Vmin in 45nm SRAM. In *IEEE International Electron Devices Meeting IEDM, Tech Dig*, pages 768–770, 2009.

[71] S.O. Toh, Z. Guo, and B. Nikolic. Dynamic SRAM stability characterization in 45nm CMOS. In *VLSI Circuits (VLSIC), 2010 IEEE Symposium on*, number 510, pages 35–36. IEEE, 2010.

[72] S.O. Toh, Zheng Guo, T.-J.K. Liu, and B. Nikolic. Characterization of Dynamic SRAM Stability in 45 nm CMOS. *Solid-State Circuits, IEEE Journal of*, 46(99):1–1, 2011.

[73] J. Tsai, S.O. Toh, Z. Guo, L.T. Pang, T.-J.K. Liu, and B. Nikolic. SRAM stability characterization using tunable ring oscillators in 45nm CMOS. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, volume 41, pages 354–355. IEEE, February 2010.

[74] Y. Tsukamoto, K. Nii, S. Imaoka, Y. Oda, S. Ohbayashi, T. Yoshizawa, H. Makino, K. Ishibashi, and H. Shinohara. Worst-case analysis to obtain stable read/write DC margin of high density 6T-SRAM-array with local Vth variability. In *IEEE/ACM International Conference on Computer-Aided Design, 2005.*, volume 1, pages 398–405. IEEE, 2005.

[75] Y. Tsukamoto, S.O. Toh, C. Shin, A. Mairena, T.-J.K. Liu, and B. Nikolic. Analysis of the relationship between random telegraph signal and negative bias temperature instability. In *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pages 1117–1121. IEEE, 2010.

[76] A. P. Van der Wel, Eric A. M. K., J.S. Kolhatkar, E. Hoekstra, M. F. Snoeij, C. Salm, H. Wallinga, and B. Nauta. Low-Frequency Noise Phenomena in Switched MOSFETs. *IEEE Journal of Solid-State Circuits*, 42(3):540–550, March 2007.

[77] E.P. Vandamme and L.K.J. Vandamme. Critical discussion on unified 1/f noise models for MOSFETs. *Electron Devices, IEEE Transactions on*, 47(11):2146–2152, 2002.

[78] D. Veksler, G. Bersuker, H. Park, C. Young, K. Y. Lim, W. Taylor, S. Lee, and H. Shin. The critical role of the defect structural relaxation for interpretation of noise measurements in MOSFETs. In *2009 IEEE International Integrated Reliability Workshop Final Report*, volume 1, pages 102–105. Ieee, October 2009.

[79] N. Verma and A.P. Chandrakasan. A 256 kb 65 nm 8T subthreshold SRAM employing sense-amplifier redundancy. *Solid-State Circuits, IEEE Journal of*, 43(1):141–149, 2008.

[80] L.T.N. Wang, N Xu, S Toh, AR Neureuther, T.-J.K. Liu, and B Nikolic. Parameter-specific ring oscillator for process monitoring at the 45 nm node. In *Custom Integrated Circuits Conference (CICC), 2010 IEEE*, pages 1–4. IEEE, 2010.

[81] Y. Wang, U. Bhattacharya, F. Hamzaoglu, P. Kolar, Y.G. Ng, L. Wei, Y. Zhang, K. Zhang, and M. Bohr. A 4.0 GHz 291 Mb Voltage-Scalable SRAM Design in a

32 nm High-k + Metal-Gate CMOS Technology With Integrated Power Management. *Solid-State Circuits, IEEE Journal of*, 45(1):103–110, 2009.

[82] C. Wann, R. Wong, D.J. Frank, R. Mann, P. Croce, D. Lea, D. Hoyniak, J. Toomey, M. Weybright, and J. Sudijono. SRAM Cell Design for Stability Methodology. *IEEE VLSI-TSA International Symposium on VLSI Technology, 2005. (VLSI-TSA-Tech).*, (April):21–22, 2001.

[83] O. Weber, O. Faynot, F. Andrieu, C. Buj-Dufournet, F. Allain, P. Scheiblin, J. Foucher, N. Daval, D. Lafond, L. Tosti, and Others. High immunity to threshold voltage variability in undoped ultra-thin FDSOI MOSFETs and its physical understanding. In *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pages 1–4. IEEE, December 2009.

[84] D. Weiss, M. Dreesen, M. Ciraula, C. Henrion, C. Helt, R. Freese, T. Miles, A. Karegar, R. Schreiber, B. Schneller, and J. Wuu. An 8MB Level-3 Cache in 32nm SOI with Column-Select Aliasing. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, pages 258–259, 2011.

[85] M. Yamaoka, N. Maeda, Y. Shimazaki, and K. Osada. A 65nm Low-Power High-Density SRAM Operable at 1.0V under 3σ Systematic Variation Using Separate Vth Monitoring and Body Bias for NMOS and PMOS. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2008 IEEE International*, pages 384–386, 2008.

[86] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. Shimazaki, K. Nii, S. Shimada, K. Yanagisawa, and T. Kawahara. Low-Power Embedded SRAM Modules with Expanded Margins for Writing. In *2005 IEEE International Solid-State Circuits Conference, 2005. Digest of Technical Papers. ISSCC*, pages 480–611, 2005.

[87] M. Yamaoka, K. Osada, and T. Kawahara. A cell-activation-time controlled SRAM for low-voltage operation in DVFS SoCs using dynamic stability analysis. *ESSCIRC 2008 - 34th European Solid-State Circuits Conference*, pages 286–289, September 2008.

[88] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr. A 3-GHz 70-Mb SRAM in 65-nm CMOS technology with integrated column-based dynamic power supply. *Solid-State Circuits, IEEE Journal of*, 41(1):146–151, 2006.