

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Large-Scale Variability Characterization and Robust Design Techniques for Nanoscale SRAM

Permalink

<https://escholarship.org/uc/item/9jk6x3wc>

Author

Guo, Zheng

Publication Date

2009

Peer reviewed|Thesis/dissertation

**Large-Scale Variability Characterization and Robust Design Techniques for
Nanoscale SRAM**

by

Zheng Guo

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 Robert C. Leachman

Fall 2009

The dissertation of Zheng Guo, titled Large-Scale Variability Characterization and Robust Design Techniques for Nanoscale SRAM, is approved:

Chair

Date

Date

Date

University of California, Berkeley

**Large-Scale Variability Characterization and Robust Design Techniques for
Nanoscale SRAM**

Copyright © 2009
by
Zheng Guo

Abstract

Large-Scale Variability Characterization and Robust Design Techniques for Nanoscale SRAM

by

Zheng Guo

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Borivoje Nikolić, Chair

Continued increase in the process variability is perceived to be a major roadblock for future technology scaling. Its impact is particularly pronounced in large memory arrays due to both the utilization of minimum sized transistors and their extremely large data capacity. In order to enable the continued scaling of the next-generation embedded static random access memory (SRAM), the ability to monitor and characterize, on-chip, the variations in SRAM functionality and performance becomes critical for both gaining a deeper understanding of the sources of variability and for developing more robust circuits and topologies. This work presents a methodology to characterize, directly, the impact of process variability on the functionality of large SRAM-based cache memories - capable of collecting massive silicon data at little hardware and/or design overhead. In addition, a thorough investigation of various SRAM read stability and writeability metrics, including the proposed large-scale design metrics, is conducted to further understand the utility of each metric for SRAM yield prediction. The large-scale characterization methodology is validated on two different test chips, fabricated in an early commercial low-power 45nm CMOS process. This method can be easily extended to capture more than 6 standard deviations of parameter variations by increasing the SRAM array size, and therefore can serve as a valuable addition to the next-generation SRAM development vehicle.

The enablement of future SRAM scaling will require technology and circuit co-design. The FinFET technology is particularly attractive for nanoscale SRAM design not only for its reduced σ_{VTH} and better control of the short channel effects (SCE), but also for the architectural flexibility enabled by its unique independently-gated (IG) operation. New bitcell designs are presented to take advantage of this IG operation in the form of a dynamic pass-gate feedback (PGFB). It is shown that the IG FinFET design using dynamic PGFB can both dramatically enhance the read stability of a 6-T SRAM cell and enable the practical design of a 4-T SRAM cell. While increased variability presents a formidable challenge for future SRAM scaling, the presented methodologies, both in testing and design, can facilitate its continuation.

Contents

List of Figures	iii
List of Tables	xiv
1 Introduction	1
1.1 Technology and SRAM Scaling Trends	1
1.2 Limitations to SRAM Scaling	3
1.3 Combating Variability: Contemporary Work	8
1.4 Research Goal	10
1.5 Dissertation Outline	11
2 Characterizing SRAM Read Stability and Writeability	12
2.1 Introduction	12
2.2 Conventional SRAM Design Metrics	12
2.3 Large-Scale SRAM Design Metrics	33
3 Variability Characterization Test Chip	46
3.1 Introduction	46
3.2 Implementation for Characterizing the Conventional SRAM Design Metrics .	46
3.3 Implementation for Characterizing the Large-Scale SRAM Design Metrics .	52
3.4 45nm Low-Power CMOS Test-Chips	60
4 Analysis of Measured Variability	62
4.1 Introduction	62
4.2 Read Stability and Writeability Measurements	62
4.3 Measurements and Estimation of the SRAM Minimum Operating Voltage .	74
4.4 Read Current Measurements	91
4.5 Impact of Systematic Variability on SRAM Cell Stability	95
4.6 Enhancement of SRAM Cell Stability using Assist Circuits	102
4.7 Summary	105
5 Robust SRAM Design using FinFETs	107
5.1 Introduction	107
5.2 FinFET Technology for SRAM Design	109
5.3 6-T FinFET based SRAM design	113
5.4 4-T FinFET based SRAM design	122

5.5	Summary	126
6	Conclusion	130
6.1	Key Contributions	130
6.2	Future Work	132
6.3	Final Words	133
	Bibliography	134

List of Figures

1.1	Continued aggressive scaling for the transistor gate length (L_G) by $0.7\times$ every 2 years. ITRS data prior to 2001 are extracted from [120]; ITRS data after 2001 are provided from ITRS (2001-2007 editions) [73] scaling specifications. Intel data are provided from [12, 91, 99, 120] and indicate a significant slowdown in the scaling of the transistor L_G	2
1.2	Recent trend in SRAM scaling, indicating a $\sim 0.5\times$ scaling factor of the cell area per technology generation. The industry data is collected from various conference publications; the SRAM cell areas reported by Intel's technology development are provided from [12, 24, 91, 99, 120]; the available bit densities are provided from [158]. The ITRS scaling specifications [73] indicate smaller cell areas than the industry publications, but they correspond to the years of production (for different technology nodes), which typically lack behind industry publications by approximately 2 years (since the 90nm node).	3
1.3	Recent trend in SRAM V_{DD} scaling [158] with a few high-density and high-performance designs (in 90nm, 65nm, and 45nm nodes) [12, 67, 148, 149, 162, 163] shown as examples.	4
1.4	Evolution of wavelength used in optical lithography [120], illustrating a shift to subwavelength lithography for the recent technology nodes. Currently, 193nm (ArF) immersion lithography is used for 45nm production and 32nm/22nm development. It is still uncertain, at this time, when extreme ultraviolet lithography (EUVL) will be available for high-volume manufacturing, to close the gap between the wavelength of the light source and the minimum feature size.	5
1.5	(a) The average number of dopant atoms in the channel decreases with technology scaling; resulting in (b) increased σ_{VTH} due to RDF. The figures are adapted, with estimated data points, from [85].	5
1.6	Recent advancements in strain engineering have led to higher improvements in the PMOS I_{DSAT} over the NMOS I_{DSAT} - data estimated from [99]. For a minimum-sized SRAM bitcell, writeability is degraded due to a decreasing cell α -ratio.	7
1.7	(a) Per-cell leakage current increases at a much higher rate than the per-cell read current (I_{READ}) with technology scaling. (b) As a result, a large fraction of the bit-line signal may be lost for higher column heights. Figures adapted, with estimated data points, from [114].	8

1.8	Measured and simulated σ_{VTH} due to RDF, under equivalent doping conditions, for (a) a $65nm$ process and (b) a $45nm$ process. The figures are adapted, with estimated data points, from [85].	9
2.1	Schematic of a 6-T SRAM bitcell.	13
2.2	(a) Definition of HSNM from simulated butterfly-curve. (b) Definition of RSNM from simulated butterfly-curve.	14
2.3	(a) Read butterfly-curve rotated by 45° . (b) Length of the side of each embedded square within the rotated butterfly-curves.	15
2.4	(a) N-curve for sweeping V_{CL} during the standby cycle. (b) I_D of pull-down, pass-gate, and pull-up transistors while sweeping V_{CL} during the standby N-curve characterization.	16
2.5	(a) N-curve for sweeping V_{CL} during the read cycle. (b) I_D of pull-down, pass-gate, and pull-up transistors while sweeping V_{CL} during the read N-curve characterization.	16
2.6	(a) Butterfly-curve for SNM extraction during the read operation. (b) N-curve for sweeping V_{CL} during the read operation (x- and y-axis reversed for comparison). (c) N-curve for sweeping V_{CH} during the read operation.	17
2.7	Scatter plots for SVNM versus RSNM obtained from 3k-sample MC simulations using a commercial low-power $45nm$ CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$	18
2.8	(a) VTC pair for an SRAM cell with a negative read margin. (b) N-curve while sweeping V_{CL} for an SRAM cell with a negative read margin (x- and y-axis reversed for comparison). (c) N-curve while sweeping V_{CH} for an SRAM cell with a negative read margin.	19
2.9	Scatter plots for SINM versus RSNM obtained from 3k-sample MC simulations using a commercial low-power $45nm$ CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$	19
2.10	(a) I_D of pull-down, pass-gate, and pull-up transistors while sweeping V_{CH} for extracting the SINM near corner B . (b) I_D of pull-down, pass-gate, and pull-up transistors as a function of V_{CH} while sweeping V_{CL} for the characterization of corner A in the RSNM extraction.	20
2.11	Distribution densities of the pass-gate transistor and the pull-down transistor current contributions to SINM at $V_{DD} = 1.1V$, $0.9V$, and $0.6V$	21
2.12	Distribution densities of (a) RSNM and (b) SINM at $V_{DD} = 1.1V$, $0.9V$, and $0.6V$	22
2.13	Scatter plots for SPNM versus RSNM obtained from 3k-sample MC simulations using a commercial low-power $45nm$ CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$	23
2.14	Fail probability as a function of V_{DD} - used to illustrate the definition of V_{MIN}	23
2.15	Scatter plots for RSNM versus $V_{MIN,RD}$ obtained from 3k-sample MC simulations using a commercial low-power $45nm$ CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$	24

2.16 Scatter plots for SVNM versus $V_{MIN,RD}$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$	24
2.17 Scatter plots for SINM versus $V_{MIN,RD}$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$	25
2.18 Scatter plots for SPNM versus $V_{MIN,RD}$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$	25
2.19 Definition of WNM from simulated VTC-pair for (a) writing a '0' into CL (or '1' into CH) and (b) writing a '1' into CH (or '0' into CL)	27
2.20 The VTC pair, rotated by 45°, for (a) writing a '0' into CL (or '1' into CH) and (b) writing a '0' into CH (or '1' into CL). (c) Length of the side of each embedded square within the rotated VTC pair.	28
2.21 Definition of I_W from measured N-curve for (a) writing a '0' into CL (or '1' into CH) and (b) writing a '1' into CH (or '0' into CL). (c) I_D of pull-down, pass-gate, and pull-up transistors while sweeping V_{CH} during N-curve characterization for writing a '1' into CH	29
2.22 Definition of WTV and WTI from measured N-curve.	30
2.23 Scatter plots for I_W versus WNM obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 0.9V$ and (b) $V_{DD} = 0.6V$	31
2.24 (a) I_D of pull-down, pass-gate, and pull-up transistors while sweeping V_{CH} for extracting the I_W near corner B . (b) I_D of pull-down, pass-gate, and pull-up transistors as a function of V_{CH} while sweeping V_{CL} for the characterization of corner A in the WNM extraction.	31
2.25 Scatter plots for WNM versus $V_{MIN,WRT}$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 0.9V$ and (b) $V_{DD} = 0.6V$	32
2.26 Scatter plots for I_W versus $V_{MIN,WRT}$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 0.9V$ and (b) $V_{DD} = 0.6V$	32
2.27 (a) Measurement setup for characterizing SRRV. (b) Definition of SRRV from simulated transfer curve.	34
2.28 (a) SRRV transfer curves for storing a '0' at the less read-stable CH node; all transfer curves exhibit sharp fall off in $I_{MEAS,BLC}$. (b) SRRV transfer curves for storing a '0' at the more read-stable CL node; only the transfer curve corresponding to a marginal intrinsic mismatch exhibit a sharp fall-off in $I_{MEAS,BL}$, while the transfer curve corresponding to a high intrinsic mismatch shows a smooth bend in $I_{MEAS,BL}$. (c) SRRV transfer curves for storing a '0' at the more read-stable CH node; only the transfer curve corresponding to a marginal intrinsic mismatch exhibit a sharp fall-off in $I_{MEAS,BLC}$, while the transfer curve corresponding to a high intrinsic mismatch shows a smooth bend in $I_{MEAS,BLC}$. (c) SRRV transfer curves for storing a '0' at the less read-stable CL node; all transfer curves exhibit sharp fall off in $I_{MEAS,BL}$	35

2.29	(a) Measurement setup for characterizing WRRV. (b) Definition of WRRV from simulated transfer curve.	37
2.30	(a) WRRV transfer curves for storing a '0' at the less read-stable CH node; all transfer curves exhibit sharp fall off in $I_{MEAS,BLC}$. (b) WRRV transfer curves for storing a '0' at the more read-stable CL node; only the transfer curve corresponding to a marginal intrinsic mismatch exhibit a sharp fall-off in $I_{MEAS,BL}$, while the transfer curve corresponding to a high intrinsic mismatch shows a smooth $I_{MEAS,BL}$. (c) WRRV transfer curves for storing a '0' at the more read-stable CH node; only the transfer curve corresponding to a marginal intrinsic mismatch exhibit a sharp fall-off in $I_{MEAS,BLC}$, while the transfer curve corresponding to a high intrinsic mismatch shows a smooth $I_{MEAS,BLC}$. (c) WRRV transfer curves for storing a '0' at the less read-stable CL node; all transfer curves exhibit sharp fall off in $I_{MEAS,BL}$	38
2.31	Scatter plots for (a) SRRV versus WRRV, (b) SRRV versus RSNM, and (c) WRRV versus RSNM at $V_{DD} = 0.6V$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process.	39
2.32	Scatter plots for (a) SRRV versus $V_{MIN,RD}$ and (b) WRRV versus $V_{MIN,RD}$ at $V_{DD} = 0.6V$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process.	39
2.33	(a) Measurement setup for characterizing BWTV. (b) Definition of BWTV from simulated transfer curve.	41
2.34	(a) Measurement setup for characterizing WWTV. (b) Definition of WWTV from simulated transfer curve.	41
2.35	Scatter plots for (a) BWTV versus WWTV, (b) BWTV versus WNM, and (c) WWTV versus WNM at $V_{DD} = 0.6V$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process.	42
2.36	Scatter plots for (a) BWTV versus I_W and (b) WWTV versus I_W at $V_{DD} = 0.6V$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process.	42
2.37	Scatter plots for (a) BWTV versus $V_{MIN,WRT}$ and (b) WWTV versus $V_{MIN,WRT}$ at $V_{DD} = 0.6V$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process.	43
2.38	(a) Flow chart for $V_{MIN,RD}$ characterization and (b) bit-line currents at different V_{DD} points for $V_{MIN,RD}$ extraction.	44
2.39	(a) Flow chart for $V_{MIN,WRT}$ characterization and (b) bit-line currents at different V_{DD} points for $V_{MIN,WRT}$ extraction.	45
3.1	Circuit diagram of the all-internal-node access characterization scheme for the SRAM macros.	47
3.2	(a) Schematic of the thick-oxide CMOS transmission gate used in the switch network for both the all-internal-node access scheme and the direct bit-line access scheme. Simulated V_{OUT} versus I_D of the switch for (b) $V_{IN} = 1.1V$ and (c) $V_{IN} = 0V$	49
3.3	Layout view of an SRAM macro constructed for a 20-row by 40-column array.	50
3.4	Layout cartoon for a $0.374 \mu\text{m}^2$ bitcell with all 10 internal nodes wired out..	51

3.5	Layout cartoon for a $0.299 \mu\text{m}^2$ bitcell with all 10 internal nodes wired out	51
3.6	Layout cartoon for a $0.252 \mu\text{m}^2$ bitcell with all 10 internal nodes wired out. The SRAM CUT is outlined by the dotted line.	52
3.7	Circuit diagram of the direct bit-line characterization scheme for the functional SRAM arrays.	53
3.8	Layout view (up to the M2 layer) showing the construction of the first level of the bit-line switch hierarchy within the column pitch of an SRAM sub-array using the $0.374 \mu\text{m}^2$ bitcell design.	54
3.9	Layout view (up to the M2 layer) showing the construction of the first level of the bit-line switch hierarchy folded to fit within $2\times$ the column pitch of an SRAM sub-array using the $0.299 \mu\text{m}^2$ bitcell design.	54
3.10	(a) Layout view showing the M2-M3 connection of V_{CELL} outside the 128×256 mini-array. (b) Layout view showing the M3-M4 connection of $V_{SS,CELL}$ inside the 128×256 mini-array.	55
3.11	Simulated waveforms during (a) the read cycle and (b) the write cycle showing correct functionality.	56
3.12	Circuit diagram of a 3LSB-3MSB segmented DAC implemented to perform on-chip word-line sweep.	57
3.13	Circuit diagram showing how the output of the 6-bit DAC is multiplexed to drive the word-line.	57
3.14	(a) Simulated gain and phase frequency response waveforms for the folded cascode operation amplifier used in the unity gain buffer. (b) Simulated transient waveforms showing the characterization of WWTW using the 6-bit DAC. Note: the BL current is scaled and does not reflect actual current values from the simulation.	59
3.15	Die photo of the first low-power 45nm CMOS test chip.	60
3.16	Die photo of the second low-power 45nm CMOS test chip. This test chip allows the characterization of 3 different SRAM bitcell designs, yielding cell areas of $0.374 \mu\text{m}^2$, $0.299 \mu\text{m}^2$, and $0.252 \mu\text{m}^2$	61
4.1	Measured (a) butterfly-curves for RSNM extraction, (b) N-curves for SVNM and SINM (as well as SPNM) extraction, (c) VTC pairs for WNM extraction, and (d) N-curves for I_W extraction from SRAM macros using all-internal-node access.	63
4.2	Scatter plots for (a) SVNM versus RSNM, (b) SINM versus RSNM, and (c) SPNM versus RSNM measured at $V_{DD} = 0.6V$ from the same SRAM macros using all-internal-node access.	64
4.3	Distribution densities of (a) RSNM/SVNM, (b) SINM, and (c) SPNM measured at $V_{DD} = 0.6V$ from the same SRAM macros using all-internal-node access. Note: the y-axis scale differs from Figure 2.12 because each metric is normalized to its σ value.	64
4.4	Scatter plot for I_W versus WNM measured at $V_{DD} = 0.6V$ from the same SRAM macros using all-internal-node access.	65

4.5	Measured transfer curves for (a) SRRV extraction, (b) WRRV extraction, (c) BWTV extraction, and (d) WWTW extraction from functional SRAM arrays using direct bit-line access.	66
4.6	(a) Measured SRRV transfer curves for storing a '0' at the less read-stable <i>CH</i> node; all transfer curves exhibit sharp fall off in $I_{MEAS,BLC}$. (b) Measured SRRV transfer curves for storing a '0' at the more read-stable <i>CL</i> node; only some transfer curves exhibit a sharp fall-off in $I_{MEAS,BL}$ while other transfer curves do not. (c) Measured WRRV transfer curves for storing a '0' at the less read-stable <i>CH</i> node; all transfer curves exhibit sharp fall off in $I_{MEAS,BLC}$. (d) Measured WRRV transfer curves for storing a '0' at the more read-stable <i>CL</i> node; only some transfer curves exhibit a sharp fall-off in $I_{MEAS,BL}$ while other transfer curves do not.	67
4.7	(a) Scatter plot for WRRV versus RSNM measured from the same SRAM macro using all-internal-node access at $V_{DD} = 0.8V$ and $0.5V$. (b) Scatter plot for SRRV versus WRRV measured from the same functional SRAM array using direct bit-line access at $V_{DD} = 0.8V$ and $0.5V$	68
4.8	(a) Scatter plot for WWTW versus I_W measured from the same SRAM macro using all-internal-node access at $V_{DD} = 0.8V$ and at $V_{DD} = 0.5V$ with $V_{NW} = 0.2V$. (b) Scatter plot for BWTV versus WWTW measured from the same functional SRAM array using direct bit-line access at $V_{DD} = 0.8V$ and $0.5V$	68
4.9	Semi-log plots for (a) the measured read metric distributions using RSNM, SRRV, and WRRV at $V_{DD} = 0.7V$; and (b) the measured write metric distributions using I_W , BWTV, and WTTW at $V_{DD} = 0.7V$	69
4.10	Normal probability plots for (a) SRRV, (b) WRRV, and (c) RSNM measured at $V_{DD} = 0.7V$	70
4.11	Normal probability plots for (a) BWTV, (b) WWTW, and (c) I_W measured at $V_{DD} = 0.7V$	70
4.12	Measured (a) μ , (b) σ , and (c) μ/σ of SRRV and WRRV as a function of V_{DD}	71
4.13	The bit-line current sensitivity to the <i>WL</i> overdrive is reduced due to a rise in the '0' storage node voltage and is more pronounced as V_{DD} is increased.	72
4.14	Measured (a) μ , (b) σ , and (c) μ/σ of BWTV and WWTW as a function of V_{DD}	73
4.15	Distributions of (a) $V_{MIN,RD}$ and (b) $V_{MIN,WRT}$ measured in a 64kb functional SRAM sub-array.	74
4.16	Scatter plots for (a) SRRV versus $V_{MIN,RD}$ and (b) WWTW versus $V_{MIN,WRT}$ measured in a 64kb functional SRAM sub-array demonstrating excellent correlation near failure. SRRV and WWTW are measured at $V_{DD} = 0.6V$; a 100mV word-line weak write is applied during $V_{MIN,WRT}$ characterization.	75
4.17	Sensitivity of (a) RSNM and (b) WNM to differential and common-mode variations in the pull-down, pass-gate, and pull-up transistor pairs within an SRAM bitcell [32].	77
4.18	(a) Scatter plot of RSNM1 versus RSNM2, along with a linear fit, showing a negative correlation between the read stability of the two data polarities. (b) Scatter plot of WNM1 versus WNM2, along with a linear fit, showing a positive correlation between the writeability for the two data polarities.	78

4.19	Semi-log plot of the distribution density of the actual RSNM - taken as the minimum of two RSNMs - extracted from a 3k-sample MC simulation fitted in (a) using the normal <i>PDF</i> and the <i>PDF</i> defined by equations 4.3-4.5; and in (b) using the <i>PDF</i> defined by equations 4.3-4.5 with $\rho = 0$ and with the extracted ρ value. The worst-case tail matches nicely to the <i>PDF</i> defined by equations 4.3-4.5 with and without modeling the ρ . (c) Fitted <i>PDFs</i> using equations 4.3-4.5 for three different values of V_{DD} . The probability of read stability failure at each value of V_{DD} is equal to the area under the <i>PDF</i> and to the left of the line $y = 0$.	79
4.20	The simulated (a) μ and (b) σ for RSNM and SRRV, using 3k-sample MC simulations, as a function of V_{DD} along with the corresponding polynomial fit and the norm of the residuals.	81
4.21	(a) The scatter plot for RSNM2 versus RSNM1, along with a linear fit, showing a negative correlation. Here, $V_{DD} = 0.8V$ is selected without any particular reason. (b) The coefficient of correlation, ρ , between RSNM1 and RSNM2 as a function of V_{DD} along with the quadratic fit.	82
4.22	(a) The scatter plot for SRRV2 versus SRRV1 with a linear fit, for SRAM cells allowing the characterization of SRRV for both data polarities, showing a positive correlation. (b) Semi-log plot of the distribution density of SRRV extracted from a 5k-sample MC simulation fitted using the normal <i>PDF</i> and the <i>PDFs</i> defined by equations 4.3-4.5 and by equation 4.8. The worst-case tail matches nicely to the <i>PDFs</i> defined by either equations 4.3-4.5 or equation 4.8. A 5k-sample MC simulation is used to extract adequate samples of both SRRV1 and SRRV2 for accurate ρ extraction.	83
4.23	Semi-log plot for the read fail probability as a function of V_{DD} , both extracted from a 100k-sample MC simulation and estimated (a) using RSNM and SRRV, and (b) using SVNM. (c) $V_{MIN,RD}$ as a function of the number of functional SRAM cells, in units of σ , extracted from a 100k-sample MC simulation and estimated using RSNM, SRRV, and SVNM. The estimations using RSNM and SRRV matches very well against the results from MC, whereas the estimation using SVNM does not. The estimation using SVNM is done twice - with μ and σ fitted using (1) a full range of V_{DD} values and (2) only higher V_{DD} values.	84
4.24	Distribution densities, at three different supply voltages, of (a) RSNM, (b) SRRV, (c) SVNM, and (d) SINM extracted from 3k-sample MC simulations for a single data polarity. The distributions of both SVNM and SINM become non-Gaussian as V_{DD} is reduced.	85
4.25	Distribution densities of (a) WNM, (b) WWTW, and (c) I_W extracted from 3k-sample MC simulations for a single data polarity. The distribution of I_W become non-Gaussian as V_{DD} is reduced.	86
4.26	(a) Semi-log plot for the read fail probability as a function of V_{DD} ; and (b) $V_{MIN,RD}$ as a function of the number of functional SRAM cells, in units of σ - both extracted from a 100k-sample MC simulation and estimated using RSNM, with μ fitted either linearly or to a 3 rd order polynomial.	87

4.27 (a) Semi-log plot for the read fail probability as a function of V_{DD} ; and (b) $V_{MIN,RD}$ as a function of the number of functional SRAM cells, in units of σ - both extracted from a 100k-sample MC simulation and estimated using RSNM, with its <i>PDF</i> modeled by either equations 4.3-4.5 or equation 4.8.	87
4.28 $V_{MIN,RD}$ as a function of the number of functional SRAM cells, in units of σ , both extracted from a 100k-sample MC simulation and estimated using - (a) RSNM with its <i>PDF</i> modeled by either equation 4.8 or equation 4.1; and (b) SRRV with its <i>PDF</i> modeled by either equation 4.8 or equation 4.1.	88
4.29 $V_{MIN,RD}$ as a function of the number of functional SRAM cells, in units of σ , both extracted from a 100k-sample MC simulation and estimated using RSNM, with its <i>PDF</i> modeled by either equations 4.3-4.5 or equation 4.8. A systematic mismatch is introduced to the bitcell through a differential adjustment in the L_G of the pull-down transistor pair - producing a $\sim 8\%$ shift in μ and a $\sim 6\%$ shift in σ between SRRV1 and SRRV2.	88
4.30 (a) Semi-log plot for the read fail probability as a function of V_{DD} ; and (b) $V_{MIN,RD}$ as a function of the number of functional SRAM cells, in units of σ - both extracted from a 100k-sample MC simulation and estimated using SRRV, with its <i>PDF</i> modeled by either equation 4.8 or equation 4.1. A systematic mismatch is introduced to the bitcell through a differential adjustment in the L_G of the pull-down transistor pair - producing a $\sim 8\%$ shift in μ and a $\sim 6\%$ shift in σ between SRRV1 and SRRV2.	89
4.31 The (a) μ and (b) σ of SRRV and WWTW, measured for 8k bitcells from a 64kb SRAM sub-array, as a function of V_{DD} along with the corresponding linear fit and the norm of the residuals.	90
4.32 Semi-log plot of the distribution densities of (a) SRRV, modeled using equation 4.8; and (b) WWTW, modeled using equations 4.3-4.5. Both SRRV and WWTW, in this example, are measured for a 64kb SRAM sub-array. The worst-case tail matches well in both cases. The scatter plots of SRRV2 versus SRRV1 and WWTW2 versus WWTW1 are included to show the positive correlation from measurement.	91
4.33 Semi-log plot for (a) the read fail probability as a function of V_{DD} , measured for a 64kb SRAM sub-array and estimated using SRRV; and (b) the write fail probability as a function of V_{DD} , measured for a 64kb SRAM sub-array and estimated using WWTW. (c) $V_{MIN,RD}/V_{MIN,WRT}$ as a function of the number of functional SRAM cells, in units of σ - both measured and estimated. To reduce writeability and expose higher $V_{MIN,WRT}$ values, NMOS RBB and PMOS FBB are applied during $V_{MIN,WRT}$ and WWTW measurements - word-line weak write is not applied because WWTW characterization requires direct word-line control. The data in (c) are so normalized to display both $V_{MIN,RD}$ and $V_{MIN,WRT}$ results in the same graph. Results indicate good matching between the estimated values and the measurement data.	92
4.34 $\sigma_{IREAD}/\mu_{IREAD}$, measured from four separate 64kb SRAM sub-arrays, as a function of $1/\sqrt{W} \times L$ for three different SRAM bitcell designs with cell areas of $0.374 \mu\text{m}^2$, $0.299 \mu\text{m}^2$, and $0.252 \mu\text{m}^2$	93

4.35 Normal probability plots for (a) I_{READ} measured at $V_{DD} = 1.1V$ and (b) I_{READ} measured at $V_{DD} = 0.7V$ and $V_{DD} = 0.5V$. The data in (b) is normalized to the mean of I_{READ} measured at $V_{DD} = 1.1V$	94
4.36 (a) The distribution density of the measured minimum I_{READ} over a 1kb SRAM block showing a long tail to the left. (b) Gumbel probability plot for the lower tail of the distribution for the measured minimum I_{READ} over a 1kb SRAM block.	94
4.37 (a) Layout view for a 20×40 SRAM array, with gate-poly in the vertical direction, wired for all-internal-node access. Each array inside the test macro is surrounded by wide regions of STI in all directions. (b) μ of the measured I_{DSAT} for pull-down, pass-gate, and pull-up transistors as a function of the distance from the edge of the array (normalized to the average distance). (c) μ of the measured $V_{TH,LIN}$ for pull-down, pass-gate, and pull-up transistors as a function of the distance from the edge of the array. (d) μ of RSNM and I_W as a function of the distance from the edge of the array. All measurements are taken from SRAM macros via all-internal-node access.	96
4.38 Measured (a) SRRV, (b) WWTW, and (c) I_{READ} as a function of row and column position within a 256×256 (64kb) functional SRAM sub-array. . . .	97
4.39 (a) A 4-cell cluster in an SRAM array showing the 4 cell orientations; the storage nodes of orientations C and D are reversed in the drawing for clarification. (b) Wafer map identifying the measured chips. Measured (c) read disturb frequency, (d) μ of WWTW, and (e) μ of I_{READ} for two test chips on the same wafer as a function of the cell storage node and the cell orientation.	99
4.40 Layout cartoon of an SRAM cell showing the corner rounding of the PMOS diffusions and the NMOS diffusions.	100
4.41 (a) Locations of the measured test chips on two different wafers. (b) Fail bit count as a function of V_{DD} during a static read operation, and (c) fail bit count as a function of V_{DD} during a static write operation, measured for 64kb SRAM sub-arrays on three test chips from two different wafers.	101
4.42 Distribution densities of (a) measured read margins (using either SRRV or WRRV) without RAC, and with BCV and SWL; and (b) measured write margins (using WWTW) without WAC, and with CVD and NBL. Measurements are taken for 2k-samples from a 64kb SRAM sub-array at $V_{DD} = 0.7V$	102
4.43 (a) Simplified schematic of the boosted cell V_{DD} (BCV) and the cell V_{DD} down (CVD) scheme for read and write assist. (b) Fail bit count as a function of V_{DD} during a read operation, measured for the same 64kb SRAM sub-array with no read assist circuits (RAC) and with a 100mV BCV. (b) Fail bit count as a function of V_{DD} during a write operation measured for the same 64kb SRAM sub-array with no write assist circuits (WAC) and with a 100mV CVD.	103
4.44 (a) Simplified schematic of the suppressed word-line (SWL) and the negative bit-line (NBL) scheme for read and write assist. (b) Fail bit count as a function of V_{DD} during a read operation measured for the same 64kb SRAM sub-array with no read assist circuits (RAC) and with a 100mV SWL. (b) Fail bit count as a function of V_{DD} during a write operation measured for the same 64kb SRAM sub-array with no write assist circuits (WAC) and with a 100mV NBL.	104

4.45	Circuit diagram for a column based biasing scheme, implemented in [162], to independently achieve high read stability and writeability.	105
5.1	Schematic of a (a) 8-T dual-port SRAM bitcell, (b) 10-T dual-port cross-point SRAM bitcell, and (c) 8-T single-port cross-point SRAM bitcell.	109
5.2	(a) Cross-sectional schematic of the FinFET structure. The gates of the FinFET can either (b) swing together in double-gated (DG) operation or (c) swing independently in independently-gated (IG) operation.	110
5.3	(a) Thin-cell layout for a conventional 6-T bulk-Si based SRAM cell with β -ratio = 1.5. The dark outline indicates the area of one memory cell. (b) Read butterfly-curves for a conventional 6-T bulk-Si based SRAM cell with β -ratio = 1.5 and β -ratio = 2.0. (c) Impact of cell β -ratio on the cell read-and write-margins (RSNM is used as the read metric and BWTV is used as the write metric). β -ratio is adjusted in (b) and (c) by changing the channel widths of the pull-down transistors.	113
5.4	Thin-cell layout for a conventional double-gated (DG) FinFET based 6-T SRAM cell with β -ratio = 1. The dark outline indicates the area of one memory cell.	114
5.5	Thin-cell layout for a conventional double-gated (DG) FinFET based 6-T SRAM cell with (a) 2 fins in the pull-down transistors and (b) $L_{G,pass\text{-}gate} = 2 \times L_{G,pull\text{-}down}$. The dark outline indicates the area of one memory cell.	114
5.6	Read butterfly-curves for a conventional DG FinFET based 6-T SRAM cell with (a) 1 fin and 2 fins in the pull-down transistors, and (b) $L_G = 22nm$ and $L_G = 44nm$ for the pass-gate transistors. (c) Impact of cell β -ratio, determined by the number of pull-down transistor fins, on the cell read- and write-margins (RSNM is used as the read metric and BWTV is used as the write metric).	115
5.7	(a) Thin-cell layout for a DG FinFET based 6-T SRAM cell with fin-rotation to increase the effective cell β -ratio. The outline indicates the area of one memory cell. (b) Read butterfly-curves for a DG FinFET based 6-T SRAM cell with fin-rotation, showing improved RSNM.	116
5.8	(a) Schematic for an IG FinFET based 6-T SRAM cell with dynamic PGFB. (b) Read butterfly-curves for an IG FinFET based 6-T SRAM cell with dynamic PGFB, showing significantly improved RSNM. (c) Thin-cell layout for an IG FinFET based 6-T SRAM cell with dynamic PGFB, indicating zero area penalty compared to the conventional DG 6-T design. The dark outline indicates the area of one memory cell. Note the use of BG-FinFET NMOS pass-gate transistors involves gate separation, as indicated in the layout by the dark region over their fins.	117
5.9	(a) Iso-writeability comparison of RSNM and (b) iso-RSNM comparison of I_{READ} , over a wide range of V_{DD} , between the conventional DG 6-T bitcell and the IG PGFB 6-T bitcell. Writeability and read stability are equalized at each supply voltage, between the two designs, using Φ_m tuning.	118

5.10 RSNM and I_{READ} , of an IG 6-T SRAM bitcell with dynamic PGFB, as a function of Φ_m . The values of I_{READ} are normalized to that of a conventional DG 6-T design with $\Phi_m = 4.75eV$	119
5.11 (a) RSNM and BWTV, of an IG 6-T SRAM bitcell with dynamic PGFB, as a function of Φ_m . (b) HSNM and BWTV, of an IG 6-T SRAM bitcell with dynamic PGFB, as a function of V_{CELL}	120
5.12 (a) Bit-line voltage simulations for the conventional DG 6-T SRAM design and the IG PGFB 6-T SRAM design, with 128 bitcells per column. (b) Impact of dynamic PGFB on sensing speed - where $\Delta T/T$ is the normalized difference in the bit-line discharging time between a conventional 6-T design and a IG PGFB 6-T designs; and the sense amplifier offset is the tolerable offset voltage at the inputs of the sense amplifier. Less than 5% impact on sensing speed is incurred when using sense amplifiers with less than 100mV offset voltage.	121
5.13 (a) Schematic for a conventional loadless 4-T SRAM bitcell. (b) RSNM and HSNM of a conventional loadless bulk-Si based 4-T SRAM bitcell, with β -ratio= 2, as a function of the difference in NMOS to PMOS V_{TH}	122
5.14 (a) Schematic and (b) layout for an IG FinFET based loadless 4-T SRAM bitcell with dynamic PGFB. Here, β -ratio= 1 is assumed. (c) Read and standby butterfly-curves for an IG FinFET based 4-T SRAM cell with dynamic PGFB (β -ratio= 1 and $\Phi_m = 4.65eV$), showing significantly improved RSNM and HSNM. (d) PMOS pass-gate current as a function of the opposing storage node voltage, illustrating the selective injection of $I_{RETENTION}$ when the storage node holds a '1'.	123
5.15 Write cycle simulation for the IG PGFB 4-T design (with $\Phi_m = 4.65eV$) illustrating (a) the successful write operation for an accessed bitcell and (b) the successful data retention for a neighboring bitcell (in the same column). $V_{WL} = -200mV$ is applied during the write cycle.	124
5.16 Bit-line voltage simulations for the IG PGFB 4-T SRAM design with varying column heights.	125
5.17 (a) Schematic for a gated- V_{SS} leakage reduction scheme. (b) The impact of leakage reduction on the HSNM of a IG PGFB 4-T SRAM bitcell.	126
5.18 Impact of process variations on the cell RSNM for various bulk-Si and FinFET SRAM bitcells. The Monte Carlo (MC) simulations are run in mixed-mode using Taurus. Geometric variations in L_G and T_{Si} (with $3\sigma(L_G) = 3\sigma(T_{Si}) = 10\%L_G$) are considered for FinFETs, whereas only RDF is considered for bulk-Si MOSFETs [63]. The RSNM extracted for FinFET based designs show much tighter distributions (i.e. smaller σ) than that of the bulk-Si based design.	128

List of Tables

1.1	Random and systematic sources of process variability [110].	6
5.1	Expected RDF-induced V_{TH} variation, expressed as a percent of the 90nm node value (for $W/L = 2$), following the ITRS scaling specifications.	108
5.2	Transistor parameters used for Taurus simulations.	111
5.3	45nm node general logic design rules used for SRAM bitcell layouts.	112
5.4	Simulated HSNM and per-cell standby leakage currents for the IG PGFB 4-T design.	126
5.5	Cell area, RSNM (and HSNM), and per-cell standby leakage currents for various bitcell designs.	127

Acknowledgments

I would like to begin by expressing my sincere gratitude to my thesis advisor, Prof. Borivoje Nikolić, for his patient, yet focused, guidance throughout the course of this thesis. His broad, yet profound, knowledge of the issues and challenges faced by the designs of both analog and digital integrated circuits has paved ways for the breadths of projects conducted within our research group. I am particularly impressed with the level of fairness and integrity he brings to his work and hope to take that with me as I enter the industry.

I would also like to thank Prof. Tsu-Jae King Liu, for her gracious support as my qualifying exam committee chair and for participating in my dissertation committee. I also greatly appreciate her careful revisions during the preparations of a number of conference and journal publications. In addition, I am pleased to acknowledge the successful collaborations with her students, in particular Dr. Sriram Balasubramanian and Dr. Andrew E. Carlson, in several projects. I would also like to thank Changhwan Shin, one of Prof. Tsu-Jae King Liu's current students, for his help with chip measurements.

My research has been supported by the National Defense Science and Engineering Graduate (NDSEG) Fellowship, the National Science Foundation Infrastructure Grant No. 0403427, and also through the Center for Circuit & System Solutions (C2S2) Focus Center. The chip fabrication donations were provided by STMicroelectronics - I would like to extend my sincere appreciation to Ernesto Perea and the engineers at STMicroelectronics, Crolles for their support.

I am genuinely thankful for Prof. Robert C. Leachman, for his kind participation in my qualifying exam, as well as dissertation, committee. I also appreciate the valuable inputs from Prof. Andy Neureuther, Prof. Costas Spanos, and their students during the various variability group meetings. In addition, I would like to thank Prof. Andrei Vladimirescu for his help in optimizing the simulation environment for my research.

I am grateful for all the help and support from the staff at the Berkeley Wireless Research Center (BWRC). In particular, I would like to thank Tom Boot, as well as Brenda Farrell, for their patient and gracious help on many administrative matters; Brian Richards for his help on the installation of many design kits, as well as his support during several tape out processes; Kevin Zimmerman and Ken Tang for their technical support; Susan H. Mellers for her help and support with the lab equipments; and Gary Kelson for his dependable reminders for seminars, retreats, as well as technical reports. I would also like to acknowledge the many current and formal students at BWRC - Dr. Liang-Teck Pang, for helpful technical discussions and his support during the first $45nm$ CMOS tape out; Prof. Zhengya Zhang, for being a great friend, as well as roommate during the many retreats, and also for his comedy relief; Seng Oon Toh and Lauren Jones for their help with the $45nm$ CMOS tape outs, as well as chip measurements; Kenneth T. Duong and Jason Tsai for their help during several $45nm$ CMOS tape outs; Dr. Bastien Giraud for his careful review for a journal submission; and all others for various technical, as well as non-technical, interactions. I also thank Dr. Yasumasa Tsukamoto for many very insightful and helpful technical discussions during his stay at Berkeley as a visiting scholar from Renesas Technology. In addition, I want to extend my sincere gratitude to Dr. Sanu K. Mathew and Dr. Azeez J. Bhavnagarwala for their helpful mentoring during my internships at Intel and IBM, respectively. I would like to particularly thank Dr. Sanu K. Mathew for his support for several fellowship applications

and Dr. Azeez J. Bhavnagarwala for his gracious job recommendations.

Finally, and certainly not the least, I thank God for the gift of faith and for His steadfast love. I would like to also express my heartfelt appreciation for my parents, my younger brother, my lovely girlfriend - Angela S. Park - and her entire family, and everyone at KCPC for their constant prayer and support throughout the hardest and darkest durations of my dissertation work.

Chapter 1

Introduction

Due to a combination of the different challenges associated with the scaling of advanced CMOS technology and the broad range in the memory performance requirements of today's system-on-chip (SoC) and microprocessor (μP) designs, the once-well-understood art of embedded memory design has evolved into one of the most exciting, yet least comprehended, areas of semiconductor research. Embedded memories already occupy well over 50% of the total silicon area today, and this number is projected [2] to reach over 90% by the year 2014. Static random access memory (SRAM) represents one of the most prevalent forms of embedded memory, accounting for over 90% of the total transistor count in some of today's high-end μP designs [1]. Presently, continued increase in the process variability is perceived to be the biggest roadblock to the scaling of multi-megabit SRAM circuits. This dissertation addresses the impending barrier, fostered by process variability, to SRAM scaling through three different means - by proposing a methodology to characterize, directly, the impact of process variability on the functionality of large SRAM-based cache memories - capable of collecting massive silicon data at little hardware and/or design overhead; by building a better understanding for the usage of various SRAM read stability and writeability metrics; and by proposing new SRAM bitcell designs, through the application of technology and circuit co-design, in a thin-body double-gated (DG) FinFET process.

1.1 Technology and SRAM Scaling Trends

Despite the emergence of recent barriers to the scaling of the CMOS technology, the semiconductor industry continues to enjoy an exponential growth in accordance to Moore's law [92, 93]; in fact, its growth has accelerated recently [102]. Figure 1.1 illustrates the continued scaling of transistor dimensions by $0.7\times$ every 2 years. While a $0.7\times$ scaling of the transistor dimensions achieve a $2\times$ reduction in the required silicon area for a given functionality, the μP die size has not shrunk dramatically over the years¹; rather, the scaling of transistor dimensions has manifested itself in a higher level of integration - of both functionality and memory [155]. In particular, embedded memory is considered to be a key performance enabler for high-end multi-core μPs as it can provide fast on-chip data

¹In fact, the μP die size has experienced a growth of about 25% per technology generation [25] until the late 90's; this trend has stopped to limit the power consumption.

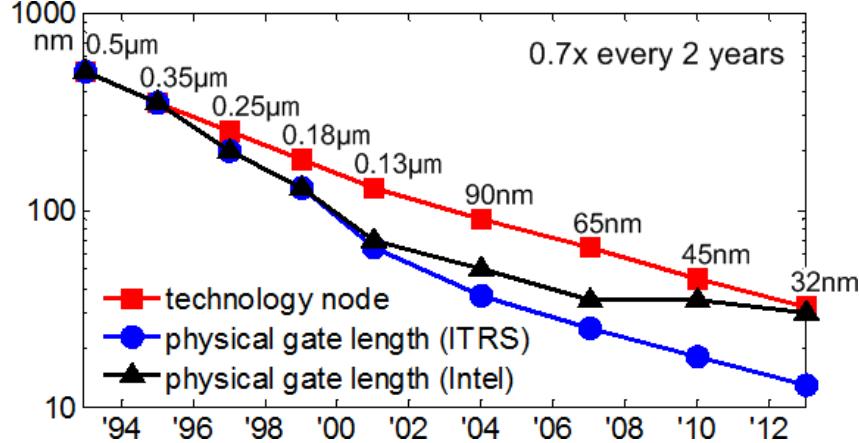


Figure 1.1: Continued aggressive scaling for the transistor gate length (L_G) by $0.7 \times$ every 2 years. ITRS data prior to 2001 are extracted from [120]; ITRS data after 2001 are provided from ITRS (2001-2007 editions) [73] scaling specifications. Intel data are provided from [12, 91, 99, 120] and indicate a significant slow-down in the scaling of the transistor L_G .

communication with the central processing unit (CPU). Most modern μP designs adopt a multi-level cache memory hierarchy, where the lowest level cache (L1 or level one) contains a small amount memory cells allowing the fastest data access, and the highest level cache (commonly L2 for personal computing and L3 for enterprise servers) contains a large amount of slower memory cells allowing the highest bit density. With aggressive scaling, the ever-increasing L2/L3 cache sizes have now become popular specifications in many of today's consumer and enterprise server μP products. The 6-transistor (6-T) SRAM cell, due to its fast random access performance, is by far the most dominant form of cache memory element in today's μPs ; although recent advancements in high-performance embedded dynamic random access memory (eDRAM) [16, 145] indicate achievable performances comparable to the slower SRAM bitcells used for the highest level (L2/L3) cache memory, while offering higher bit densities, and therefore suggest a possible replacement. However, although demonstrated to be relatively inexpensive ($\sim 7\%$ cost overhead) for the silicon-on-insulator (SOI) technology [145], the integration of logic and eDRAM process remains expensive in the bulk-Si CMOS technology due to the necessity of a thick collar to suppress the parasitic leakage. In addition, the scalability of high performance eDRAM, which already suffers from a short data retention time, may be challenging due to an ever-shrinking V_{TH} requirement - where the lowest V_{TH} must meet the retention specifications and the highest V_{TH} must meet the performance requirements [84]. Therefore, the scaling of embedded SRAM continues to set the pace for the development of the next-generation high-performance SoC.

Figure 1.2 shows that the recent scaling of embedded SRAM is able to keep up with Moore's law - indicating a $\sim 0.5 \times$ scaling factor of the cell area per technology generation. However, the margins (between the smallest achievable bitcells and the $0.5 \times$ scaling requirement) have been shrinking since the 65nm technology node. Note that while the ITRS scaling specifications indicate smaller cell areas than the industry publications, they correspond to the years of production (for different technology nodes), which typically lack behind industry publications by approximately 2 years (since the 90nm node).

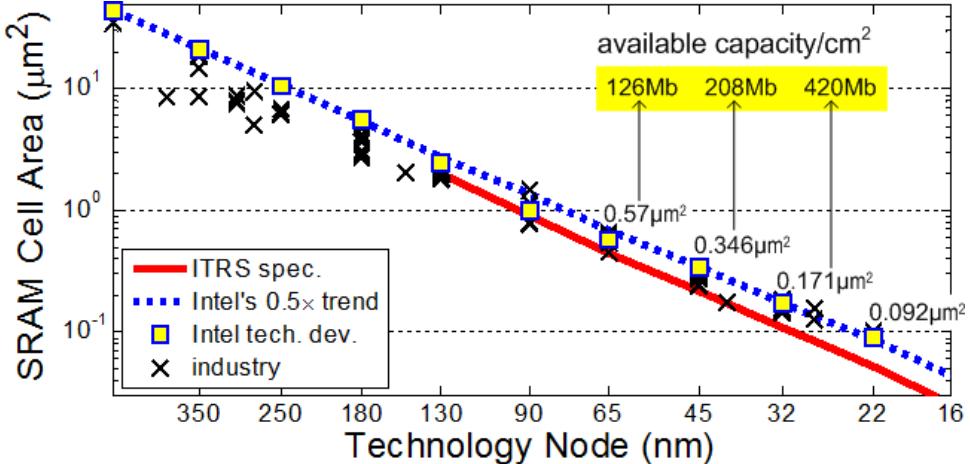


Figure 1.2: Recent trend in SRAM scaling, indicating a $\sim 0.5\times$ scaling factor of the cell area per technology generation. The industry data is collected from various conference publications; the SRAM cell areas reported by Intel’s technology development are provided from [12, 24, 91, 99, 120]; the available bit densities are provided from [158]. The ITRS scaling specifications [73] indicate smaller cell areas than the industry publications, but they correspond to the years of production (for different technology nodes), which typically lack behind industry publications by approximately 2 years (since the 90nm node).

The SRAM operating voltage (V_{DD}) scaling trend (Figure 1.3), however, indicates a significant slow-down in the recent years and a saturation at around $V_{DD} = 1.0V$ [158]. The ITRS projections have also dramatically shifted. This is primarily a result of the inhibited scaling of V_{TH} due to the exponentially dependent subthreshold leakage currents of modern MOSFETs. In addition, an increasing σ_{VTH} , dominated by random dopant fluctuation (RDF), has emerged as a second barrier to SRAM V_{DD} scaling in recent process nodes - by degrading the SRAM stability.

1.2 Limitations to SRAM Scaling

1.2.1 Process Variability

The existence of process variability [121] and attempts to characterize it [6] have long been documented in the semiconductor literature. However, the control of process variability has not kept pace with the aggressively scaled transistor dimensions [98]. This is particularly problematic as transistor scaling approaches atomic-scale dimensions. As a result, although the SRAM cell area has kept pace with technology scaling (at least for now), the ratios of the standard deviations over the means, of the SRAM stability margins, continue to increase. Concurrently, high-end μP have been increasing the amount of on-die cache to improve the performance. This simultaneous increase of both the memory size and the variability, therefore, presents one of the greatest obstacles in semiconductor research today.

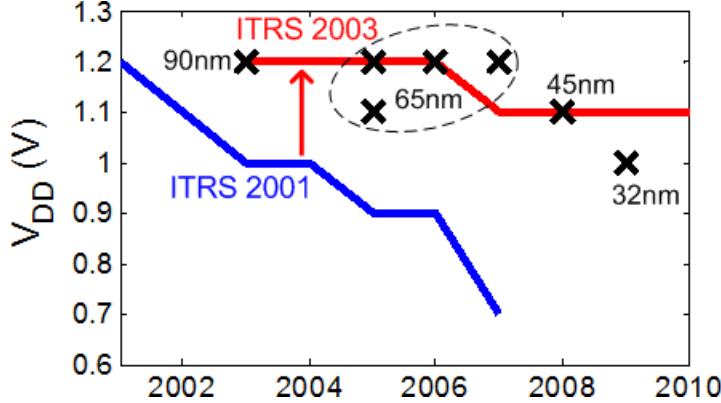


Figure 1.3: Recent trend in SRAM V_{DD} scaling [158] with a few high-density and high-performance designs (in 90nm, 65nm, and 45nm nodes) [12, 67, 148, 149, 162, 163] shown as examples.

Subwavelength Lithography

As the critical dimensions (CD) continue to shrink rapidly, the evolution of the light source used in optical lithography falls behind. Figure 1.4 illustrates a shift to subwavelength lithography (approximately) since the 180nm technology node. To print ever-shrinking patterns using a longer wavelength, compensation schemes such as optical proximity correction (OPC) and phase shift masks (PSM) have been adopted. While such compensation schemes enable the fabrication of subwavelength features, CD control remains difficult as the gap between the wavelength and the minimum feature size continues to grow [77, 110]. Currently, 193nm (ArF) immersion lithography is used for 45nm production² and 32nm/22nm development. It is still uncertain, at this time, when extreme ultraviolet lithography (EUVL) will be available for high-volume manufacturing, to close the gap between the wavelength of the light source and the minimum feature size.

In addition to the lithography process, the etch process also contributes to CD variations, since the typical physical gate length (L_G) is significantly smaller than the printed linewidth [73]. This, and other sources of variability related to the issues of manufacturing control, can be classified as extrinsic sources of variability [107]. To limit the impact of CD variability, recent trend from the semiconductor industry [12] reveals a longer physical L_G than specified by the ITRS. In addition, the scaling of the physical L_G , in the semiconductor industry, has slowed significantly³ - nearly coming to a complete halt [91, 99], in contrast to the ITRS scaling specifications (Figure 1.1). If CD control does not improve, the scaling of embedded SRAM may be severely limited as both the array size and the ratios of the standard deviations over the means, of the bitcell stability margins, increase simultaneously.

Intrinsic Atomic-Scale Random Variations

As transistor scaling approaches atomic-scale dimensions, the effects of random device parameter fluctuations become important [107]. In this regime, small fluctuations in the

²Immersion lithography is not used for Intel's 45nm node.

³Scaling of the physical L_G is also limited by other effects, such as LER and RDF.

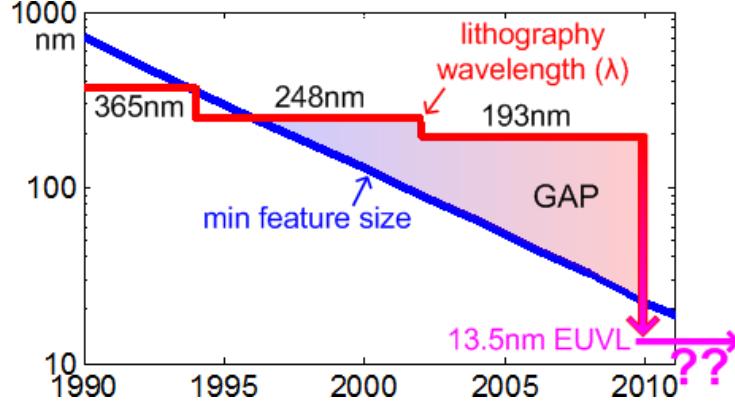


Figure 1.4: Evolution of wavelength used in optical lithography [120], illustrating a shift to subwavelength lithography for the recent technology nodes. Currently, 193nm (ArF) immersion lithography is used for 45nm production and 32nm/22nm development. It is still uncertain, at this time, when extreme ultraviolet lithography (EUVL) will be available for high-volume manufacturing, to close the gap between the wavelength of the light source and the minimum feature size.

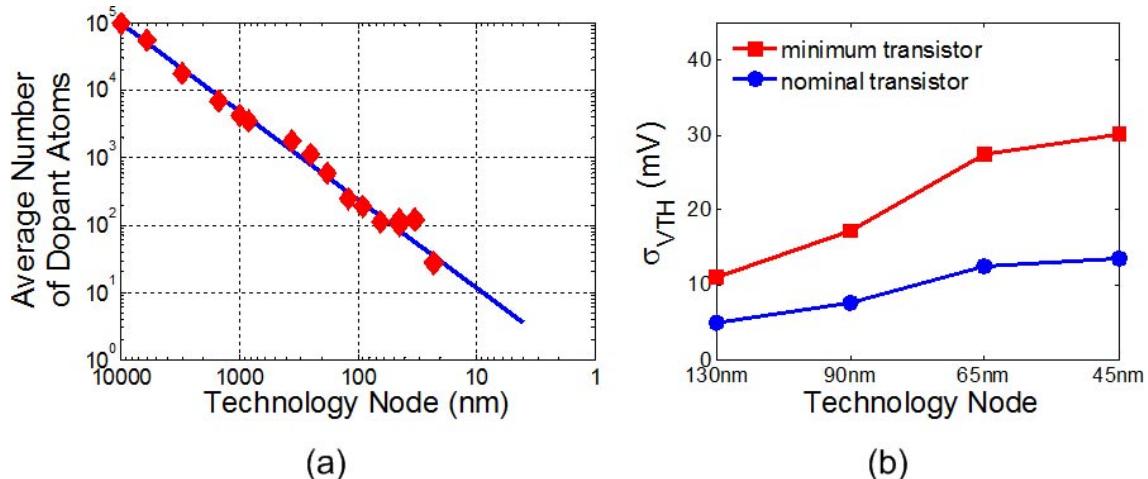
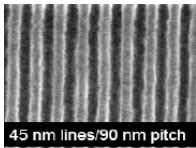
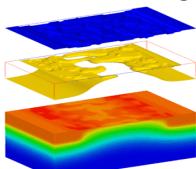
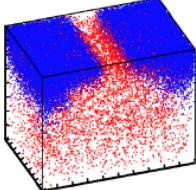


Figure 1.5: (a) The average number of dopant atoms in the channel decreases with technology scaling; resulting in (b) increased σ_{VTH} due to RDF. The figures are adapted, with estimated data points, from [85].

number and/or location of atoms may result in significant parameter variations as the total count becomes small. In particular, random dopant fluctuation (RDF) is perceived as the primary supplier of σ_{VTH} [9,135] in modern bulk-Si MOSFETs and will continue to be so, at least, until $L_G < 20\text{nm}$ [10]. This RDF-induced σ_{VTH} is expected to increase as the average number of dopants in the channel decreases with scaling, in accordance with Pelgrom's model [113] - i.e. $\sigma_{VTH} \propto 1/\sqrt{W_{EFF} \times L_{EFF}}$. Figure 1.5 graphically illustrates a decrease in the average number of dopant atoms in the channel and an increase in RDF-induced σ_{VTH} with technology scaling.

A second significant source of random V_{TH} variations is line-edge roughness (LER) -

Table 1.1: Random and systematic sources of process variability [110].

Parameter	Random	Systematic
Gate Length (L_G)	Line-edge roughness (LER) [105] 	Lithography and etching: proximity effects, orientation [106]
Gate Oxide Thickness (T_{OX})	Si/SiO_2 and $SiO_2/Poly-Si$ interface roughness [11] 	Non-uniformity in oxide growth
Channel Dopant Concentration (N_{CH})	RDF-induced σ_{VTH} [55] 	Non-uniformity in dopant implantation, dosage, diffusion
Threshold Voltage (V_{TH}) (non- N_{CH} related)	Random anneal temperature and strain effects	Non-uniform annealing temperatures [117] (metal coverage over gate); biaxial strain
Mobility (μ)	Random strain distributions	Systematic variation of strain in the Si due to STI, S/D area, contacts, gate density, etc.

defined as the random variation in the transistor L_G along its width [107]. Factors contributing to LER include statistical variation in the photon count during lithographic exposure, the aerial image contrast, and the absorption rate and composition of the photoresist [19, 49]. Effects of LER are expected to become significant as transistor dimensions shrink below $50nm$, and severe at below $32nm$ [107]. In particular, the σ_{VTH} due to LER is expected to dominate over RDF at $L_G < 20nm$ [10], if the present LER value of approximately $4nm$ does not scale. Both LER and RDF also likely contribute to the significant slow-down in the scaling of the transistor L_G (as mentioned previously and shown in Figure 1.1) in the recent technology nodes [91, 99].

Additional sources of variability may include gate oxide interface roughness, strain-induced mobility (μ) variations, etc. A summary of the different sources of random and systematic variations [110] is provided, for completeness, in Table 1.1.

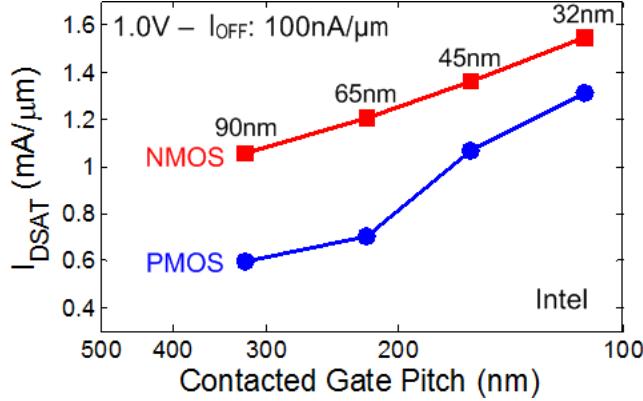


Figure 1.6: Recent advancements in strain engineering have led to higher improvements in the PMOS I_{DSAT} over the NMOS I_{DSAT} - data estimated from [99]. For a minimum-sized SRAM bitcell, writeability is degraded due to a decreasing cell α -ratio.

1.2.2 Transistor Characteristics

In addition to increases in the standard deviations of the SRAM stability margins due to process variability, technology scaling also affects the means. In modern MOSFETs, short channel behaviors such as drain-induced barrier lowering (DIBL), channel length modulation, velocity saturation, etc. [152] impact the transistor on-state characteristics. DIBL is particularly destructive to the read stability of a 6-T SRAM bitcell (at higher supply voltages), as it degrades the transistor output impedance and thus the inverter gain in the read voltage transfer curves (VTC) - see Section 2.2.1.

Additionally, recent advancements in strain engineering have enhanced the hole mobility (μ_p) at a faster rate than the electron mobility (μ_n), leading to a faster improvement in the PMOS on-current compared to the NMOS on-current (Figure 1.6) - thus closing the gap between NMOS and PMOS performance. While this may benefit logic circuits, through a reduction in the gate logical effort [133] - by allowing equal pull-up versus pull-down strengths with smaller PMOS-to-NMOS ratios; it can be detrimental to the writeability of a minimum-sized SRAM bitcell, through a reduction in the cell α -ratio (Section 2.2.2).

Furthermore, increased subthreshold and gate leakage currents may impact the SRAM array segmentation - thus limiting the array efficiency, and may also place an upper bound to the array size due to a power constraint. Figure 1.7a shows that the per-cell leakage current increases at a much higher rate than the per-cell read current (I_{READ}) with technology scaling. In the worst-case, the bit-line leakage currents from all un-accessed bitcells in the same column compete against the I_{READ} of the cell under read access (Section 5.3.4). As a result, a large fraction of the bit-line signal may be lost for tall columns in a scaled process (Figure 1.7b). On top of the signal loss, read access time requirements may further limit the column height [3]. In addition, as the transistor gate oxide thickness continues to shrink, not only does the gate leakage increase, the gate oxide reliability also becomes compromised [4, 114, 122].

Finally, while the recent trend indicates a decrease in the per-cell sensitivity to soft errors (or single event upsets) - as the scaling of the collection area compensates the reduction

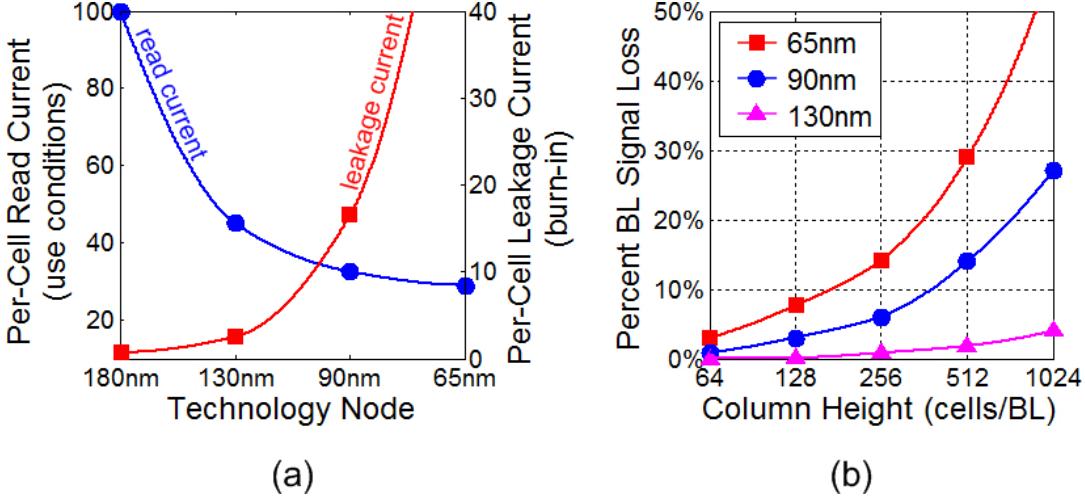


Figure 1.7: (a) Per-cell leakage current increases at a much higher rate than the per-cell read current (I_{READ}) with technology scaling. (b) As a result, a large fraction of the bit-line signal may be lost for higher column heights. Figures adapted, with estimated data points, from [114].

in the cell critical charge (Q_{CRIT}) - with technology scaling [56, 115], the per-chip soft error rate (SER) has been steadily increasing due to a $\sim 2\times$ increase in the bit capacity per generation.

Therefore, with continued technology scaling, embedded SRAM simultaneously suffers from the ever-increasing standard deviations over means of the stability margins (due to process variability), increased leakage, reduced gate oxide reliability, and reduced SER reliability. While each issue presents a formidable challenge to the continuation of SRAM scaling, this work focuses primarily on the impact of process variability on SRAM functionality.

1.3 Combating Variability: Contemporary Work

The impact of process variability on SRAM functionality has been an active area of research. To better investigate the failure mechanisms of SRAM bitcells, new metrics to numerically quantify the cell read stability [61, 119, 150] and the cell writeability [20, 21, 30, 61, 119, 134, 150] have been proposed to complement the classical static noise margin (SNM) [123]. In addition, methods to analytically model the SRAM read margin [22, 61, 70] and write margin [61] have been developed. While these techniques offer fast yield analysis and/or design optimization, their accuracies are compromised by the approximations made in their formulations. As transistor models become more and more complex with scaling, the error from these approximations will inevitably grow.

Another common approach is to estimate the yield using SPICE- and/or TCAD-based Monte Carlo simulations. While this approach can be time consuming, statistical methods can be applied to quickly estimate the probability of failure [27, 46, 78, 128, 147]. However, the accuracies of these methods depend on the device models. As process becomes increasingly complex and harder to control, designers can no longer rely on model accuracy

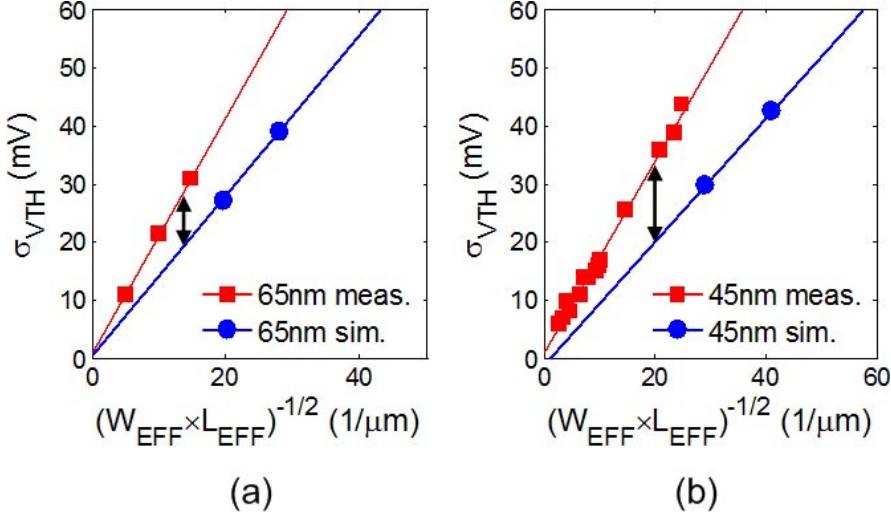


Figure 1.8: Measured and simulated σ_{VTH} due to RDF, under equivalent doping conditions, for (a) a 65nm process and (b) a 45nm process. The figures are adapted, with estimated data points, from [85].

to fully capture the random effects in large cache memories. As an example, measured and simulated σ_{VTH} due to RDF are plotted in Figure 1.8 for 65nm and 45nm process nodes, illustrating large gaps between silicon measurements and simulations.

Recently, methods have been developed to characterize SRAM variability through measuring DC read/write margins in small SRAM macros with wired-out storage nodes [20, 21]. This significantly enhances the accuracy of SRAM failure analysis over both analytical methods and simulations, but requires the removal of upper metal layers and the insertion of large switch networks to access all internal storage nodes. As a result, this approach is limited to delivering smaller data volume that may be unsuitable for failure analysis of large cache memory. Thus, SRAM designers continue to rely on collecting distributions of bit-line read currents (I_{READ}) [53] to gauge the performance and minimum operating voltage (V_{MIN}) [4, 14] to gauge the SRAM read stability and writeability in large functional SRAM arrays. However, direct correlation between measured SRAM read/write margins and V_{MIN} in large functional SRAM arrays has not been established.

In addition to characterizing the impact of process variability on SRAM functionality, circuit techniques have been adopted to maintain functionality in the presence of process variations. The simplest form of such techniques involves the optimization of transistor sizing to either shift the means of the read/write margin distributions - through adjusting the cell *beta-* and *alpha*-ratios, or decrease the standard deviations of the read/write margin distributions - through collectively increasing the transistor dimensions, or both. Alternatively, bitcell designs implemented with extra (i.e. more than six) transistors have been proposed to enhance the cell margins [35, 37, 154]. However, these techniques inevitably result in larger cell areas, undermining the fundamental drive to increase density. To increase the array robustness of smaller bitcell designs, assist techniques [41, 100, 104, 116, 125, 144, 156, 162] can be implemented to widen the SRAM design margins by shifting the SRAM operating point away from failure (Section 4.6). However, such techniques degrade the array efficiency and,

thus, the array density.

1.4 Research Goal

With aggressive technology scaling, the construction of a large memory array now presents an extreme example of variability-aware design. To satisfy the functionality of hundreds of millions of SRAM cells in current on-die cache memories, the design has to provide more than 6 standard deviations of margin to parameter variations. This is becoming increasingly challenging to satisfy, and presents a major problem for continued scaling of memory density.

In addition to technological advances, the enablement of future SRAM scaling will depend on the ability to monitor and characterize, on-chip, the variations in SRAM functionality and performance; this is critical for both gaining a deeper understanding of the sources of variability and for developing more robust circuits and topologies. This dissertation facilitates the design of embedded SRAM in the presence of process variability in the following ways:

- ◊ **Developing a methodology to characterize, directly, the impact of process variability on the functionality of large SRAM-based cache memories.**

SRAM read stability and writeability are, conventionally, measured from standalone SRAM macros with wired-out storage nodes (Section 1.3). This often requires a very large area overhead that is associated with the switch network, and therefore, is typically limited to delivering smaller data volume. A method to directly measure the SRAM read stability and writeability from functional arrays is developed, using direct bit-line access. This method is capable of collecting massive silicon data at little hardware and/or design overhead (compared to the conventional method) and is validated on two test chips, fabricated in an early commercial low-power 45nm CMOS process.

- ◊ **Building a better understanding for the usage of various SRAM read stability and writeability metrics.**

Recently, several read stability and writeability metrics have been proposed (Section 1.3). While each metric has been shown to provide a good indication for the SRAM read/write functionality, detailed examinations reveal discrepancies among the various metrics under different characterization conditions. To further assess the different metrics, their respective suitabilities for yield prediction are also compared. Having a deeper understanding of the various design metrics can help designers to more effectively use them for SRAM yield analysis.

- ◊ **Designing new SRAM bitcells using thin-body double-gated (DG) FinFETs.**

Due to the ever-increasing RDF-induced σ_{VTH} in planar bulk-Si MOSFETs, SRAM design using a thin-body DG FinFET process is investigated as an alternative. The FinFET technology not only offers reduced σ_{VTH} , due to the elimination of RDF, and better control of the short channel effects (SCE), but also provides an opportunity for better stability trade-offs through its independently-gated (IG) operation.

1.5 Dissertation Outline

Chapter 2 reviews and compares the various conventional SRAM stability metrics in detail. The limitations of these conventional metrics are highlighted, and the large-scale SRAM read stability and writeability metrics are introduced as solutions. In addition, a method for per-cell minimum operating voltage, V_{MIN} , characterization using direct bit-line measurements is also described. The correlations between the various conventional and large-scale metrics, as well as per-cell V_{MIN} , are studied, in detail, through Monte Carlo simulations; and speculations for the utilities of the different metrics for V_{MIN} estimation are made.

Chapter 3 details the implementations of two variability characterization test chips in a commercial low-power strained-Si 45nm CMOS process. These test chips allow measurements of both the conventional SRAM design metrics and the large-scale SRAM design metrics for three different bitcell designs - achieving cell areas of 0.374 μm^2 , 0.299 μm^2 , and 0.252 μm^2 .

Chapter 4 presents the measurement results, where direct correlations between the conventional SRAM design metrics and the large-scale SRAM design metrics, and between the large-scale SRAM design metrics and the per-cell V_{MIN} , are established. The large-scale characterization of SRAM variability is attractive for early stages of SRAM development due to its ability to capture massive statistical data at a very low design and area overhead, compared to the conventional method. In addition, a method to estimate the V_{MIN} of a functional SRAM array using the large-scale read/write margin measurements is described. Sources of systematic variations and their impacts on the SRAM cell stability are studied. Finally, the impact of several read and write assist circuits on the SRAM cell stability is investigated.

Chapter 5 evaluates FinFET based SRAM as an alternative in nanoscale memory design. Both 6-T and 4-T SRAM bitcells are analyzed using mixed-mode Taurus simulations. New bitcell designs are proposed to take advantage of the unique independently-gated (IG) operation of the FinFET technology. It is shown that the IG FinFET design using dynamic pass-gate feedback (PGFB) can both dramatically enhance the read stability of a 6-T SRAM cell and enable the practical design of a 4-T SRAM cell.

Finally, Chapter 6 presents a summary of this dissertation - highlighting the key contributions of this work, along with future research directions.

Chapter 2

Characterizing SRAM Read Stability and Writeability

2.1 Introduction

The most important properties of an SRAM array, in addition to power and performance targets, are its density and its yield, which is limited by the impact of process variability on per-cell functionality. Yield can be guaranteed for large SRAM arrays by providing sufficient design margins for functionality, which are determined by transistor sizing (W and L_G), the selection of transistor threshold voltages (V_{TH}), and the SRAM cell supply voltage (V_{CELL}); and/or by implementing assist techniques [41, 100, 104, 125, 144, 156, 162]. In order to investigate the impact of process variability on the functionality of an SRAM cell, the metrics for characterizing SRAM read stability and writeability must first be understood.

Section 2.2 explores in detail the conventional read stability and writeability metrics used in recent studies [20, 21, 30, 61, 119, 123, 150] for SRAM yield estimation. Section 2.3 highlights the limitations of the conventional metrics and introduces the large-scale SRAM read stability and writeability metrics as solutions. In addition, a method for per-cell minimum operating voltage, V_{MIN} , characterization using direct bit-line measurements is also described.

2.2 Conventional SRAM Design Metrics

2.2.1 Standby and Read Stability Metrics

Figure 2.1 shows the schematic of a 6-T SRAM bitcell. It consists of two cross coupled inverters ($P_L - N_L$ and $P_R - N_R$) for data retention and two pass-gate transistors (N_{AXL} and N_{AXR}) for read/write access. In the standby mode, assuming a prevalently used precharge-high bit-line scheme, the word-line (WL) is driven low and both bit-lines (BL and BLC) are precharged to the operating voltage (V_{DD}). During this mode, the PMOS pull-up transistor (P_R) must compensate for all leakage paths to V_{SS} at the '1' storage node (CH). This is satisfied by keeping the cell supply voltage (V_{CELL}) sufficiently high. However, the degradation of I_{ON}/I_{OFF} ratio and a significant increase in the gate leakage in recent

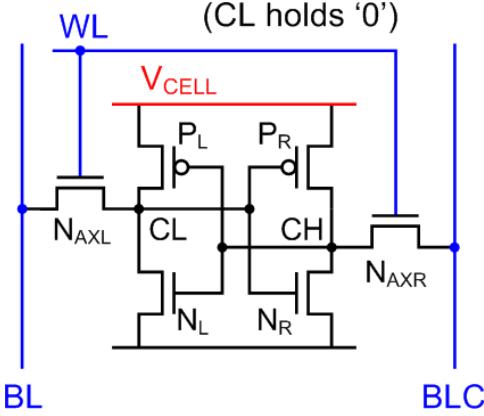


Figure 2.1: Schematic of a 6-T SRAM bitcell.

technology nodes [114], coupled with the recent trend of reducing V_{CELL} during standby to limit SRAM static power consumption [118], make data retention in large memory arrays a progressively more difficult task.

During the read operation, the word-line (WL) is driven high and both bit-lines (BL and BLC) float around the operating voltage (V_{DD}). Due to the activation of the pass-gate transistor N_{AXL} , the storage node voltage V_{CL} rises above $0V$, to a voltage determined by the resistive voltage divider set up by the pass-gate transistor N_{AXL} and the pull-down transistor N_L between BL and the storage node CL . If V_{CL} exceeds the trip point of inverter $P_R - N_R$ during the read cycle, the cell bit will flip, causing a read upset. Similar to data retention during standby, data retention during the read cycle can be guaranteed by keeping V_{CELL} sufficiently high, as the cost of increased power consumption during the read cycle. Alternately, read stability can also be satisfied by increasing the strength of the pull-down transistor relative to the pass-gate transistor - i.e. by increasing the SRAM cell β -ratio, which is defined as the strength ratio of the pull-down transistor to the pass-gate transistor. Since SRAM cells are commonly implemented using minimum geometry transistors to maximize the density, this is typically achieved by either increasing the pull-down transistor channel width (W) or the pass-gate transistor channel length (L_G), thus trading off cell compactness for enhanced cell stability.

Hold and Read Static Noise Margin

The most common metric for characterizing SRAM data stability is the static noise margin (SNM). SNM can be extracted from the voltage transfer characteristics (VTC) generated for the two halves of an SRAM cell [123]. The VTC during the standby mode can be measured by sweeping the voltage at the storage node CH (or CL) with both bit-lines (BL and BLC) biased at V_{DD} and the word-line (WL) biased at V_{SS} while monitoring the node voltage at CL (or CH). Figure 2.2a plots the resulting VTC pair, more commonly referred to as the butterfly-curve, simulated in a commercial low-power 45nm CMOS process for the standby mode. During the read operation, pass-gate transistors turn ON and the VTC can be measured by sweeping the voltage at the storage node CH (or CL) with both bit-lines (BL and BLC) and the word-line (WL) biased at V_{DD} while monitoring the node voltage

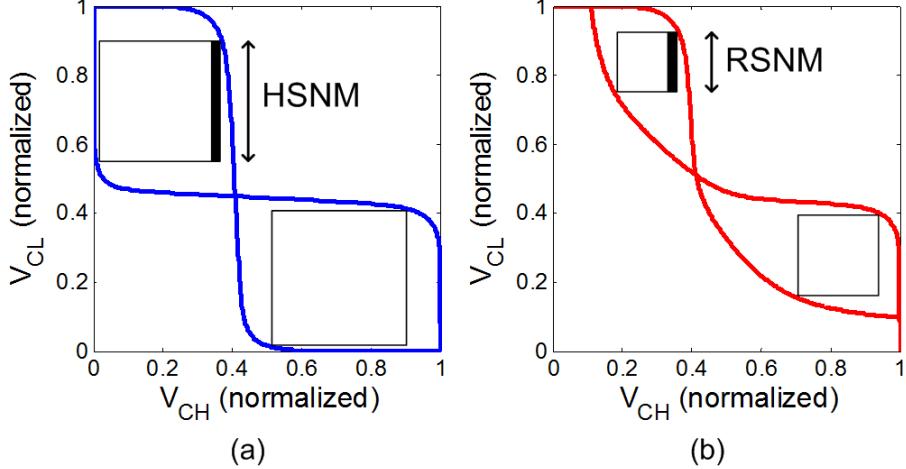


Figure 2.2: (a) Definition of HSNM from simulated butterfly-curve. (b) Definition of RSNM from simulated butterfly-curve.

at CL (or CH). The butterfly-curve simulated for the read cycle is plotted in Figure 2.2b. (V_{CELL} is biased at V_{DD} for both measurements.) Both butterfly-curves (during standby and read cycles) illustrate a bistable circuit operation with 2 stable points - at low V_{CH} (high V_{CL}) and at high V_{CH} (low V_{CL}), and 1 metastable point - at moderate V_{CH} and V_{CL} . The SNM of a bitcell for storing a certain data polarity can be quantified by the side of the largest square embedded within the corresponding opening in the butterfly-curve - i.e. the stability of storing a '0' at CH ('1' at CL) is gauged by the side of the largest square embedded within the upper-left opening of the butterfly-curve; likewise, the stability of storing a '1' at CH ('0' at CL) is gauged by the side of the largest square embedded within the lower-right opening of the butterfly-curve. Therefore, the SNM of an SRAM cell is equal to the side of the smaller maximum-square¹ and it represents the maximum tolerable DC noise voltage simultaneously added to storage nodes CH and CL before corrupting its data.

The value of the SNM can be analytically extracted from the butterfly-curve by rotating both the x-axis and the y-axis by 45° [123]. The vertical distance between the two resulting curves corresponds to the diagonal of each square that can be embedded within the butterfly-curve. Figure 2.3 plots (a) the rotated butterfly-curve simulated for the read cycle and (b) the length of the side of each embedded square. The shorter peak in Figure 2.3b represents the SNM of the SRAM cell during the read cycle. The SNM captured during the standby mode is commonly referred as the hold static noise margin (HSNM) and the SNM captured during the read cycle is commonly referred as the read static noise margin (RSNM). Since N_{AXL} operates in parallel with P_L during the read cycle and keeps V_{CL} from ever reaching 0V, the output-low voltage (V_{OL}) of the inverter $P_L - N_{AXL} - N_L$ will be non-zero and the gain in the inverter VTC will decrease [22], causing a reduction in the SNM. Thus, under the same operating conditions (i.e. V_{DD} , etc.), the SRAM cell is more susceptible to noise during the read cycle.

¹The side of the square is taken instead of the diagonal because SNM measures the amount of noise voltage added to V_{CH} and V_{CL} (i.e. x- and y-axis in Figure 2.2) in order to corrupt the cell data.

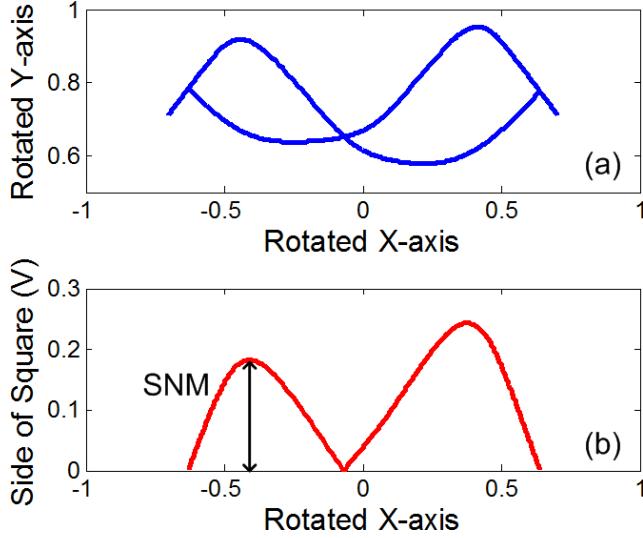


Figure 2.3: (a) Read butterfly-curve rotated by 45° . (b) Length of the side of each embedded square within the rotated butterfly-curves.

N-Curve

Alternately, the SRAM read stability can be characterized using the N-curve [150], which simplifies the analytical extraction (compared to the SNM). The N-curve can be measured by sweeping the voltage at the storage node CH (or CL) while monitoring the current externally sourced into the CH (or CL) node. For N-curve characterization during the standby mode, BL and BLC are biased at V_{DD} while WL is biased at V_{SS} ; for N-curve characterization during the read cycle, BL , BLC , and WL are all biased at V_{DD} . (V_{CELL} is biased at V_{DD} for both measurements.) Figure 2.4a plots the N-curve for an SRAM cell during the standby operation and Figure 2.4b graphically illustrates the relative current contributions of the pull-down, the pass-gate, and the pull-up transistors to the N-curve. Figure 2.5 presents similar plots for the read operation. As expected, the current contribution of the pass-gate transistor is approximately zero during the standby operation. During the read operation, the pass-gate current contributes negatively to the N-curve and pulls it down.

The butterfly-curve, for SNM extraction, is compared against the N-curve in Figure 2.6 for the read operation. Figure 2.6a plots the butterfly-curve with V_{CH} on the y-axis and V_{CL} on the x-axis. The N-curve captured while sweeping V_{CL} is plotted in Figure 2.6b and the N-curve captured while sweeping V_{CH} is plotted in Figure 2.6c. The 3 zero-crossings of each N-curve corresponds to the 3 intersection points on each butterfly-curve - the first and last zero-crossings of the N-curve correspond to the 2 stable roots of the butterfly-curve, while the middle zero-crossing of the N-curve corresponds to the meta-stable root of the butterfly-curve. The locations of the 3 zero-crossings are in agreement with the Kirchhoff's current law (KCL) since the N-curve measures the amount of external current required at each storage node such that the voltage conditions are satisfied while sweeping that storage node - because KCL is satisfied at each intersection point of the butterfly-curve without any external current, the N-curve should cross zero at all such intersection points. The static

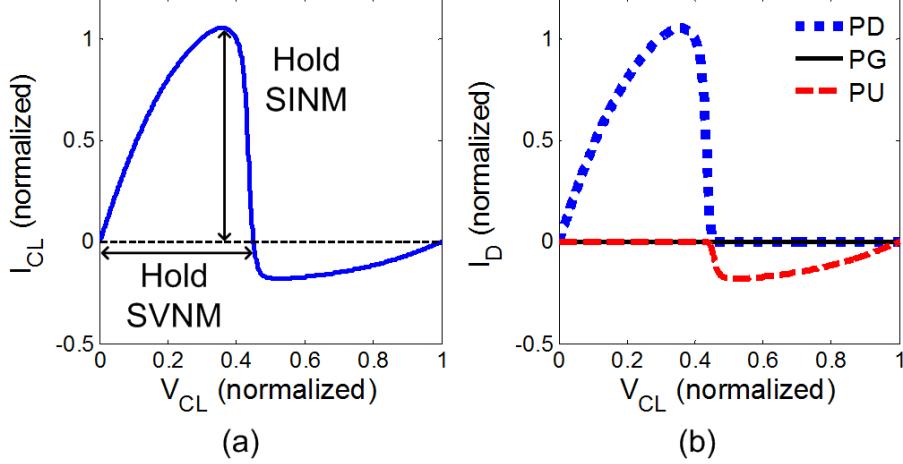


Figure 2.4: (a) N-curve for sweeping V_{CL} during the standby cycle. (b) I_D of pull-down, pass-gate, and pull-up transistors while sweeping V_{CL} during the standby N-curve characterization.

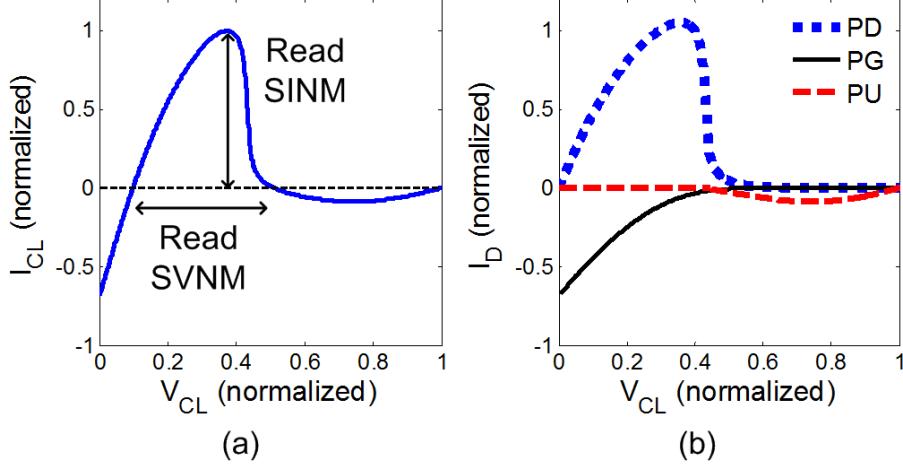


Figure 2.5: (a) N-curve for sweeping V_{CL} during the read cycle. (b) I_D of pull-down, pass-gate, and pull-up transistors while sweeping V_{CL} during the read N-curve characterization.

voltage noise margin (SVNM) [61], also referred to as the critical voltage (V_{CRIT}) [150], is defined as the voltage difference between the first 2 zero-crossings of the N-curve and represents the maximum tolerable DC noise voltage added to the sweeping storage node (i.e. the storage node corresponding to the x-axis) before data corruption. The static current noise margin (SINM) [61], also referred to as the critical current (I_{CRIT}) [150], is defined as the first peak current in the N-curve, located near corner B in Figure 2.6a, and represents the maximum tolerable DC noise current injected into the sweeping storage node (i.e. the storage node corresponding to the x-axis) before data corruption. The SINM (or I_{CRIT}) effectively measures the pull-down transistor current minus the pass-gate transistor current. A third stability metric, the static power noise margin (SPNM) [61] or the critical power (P_{CRIT}) [150], includes both voltage and current information and is defined as the area underneath

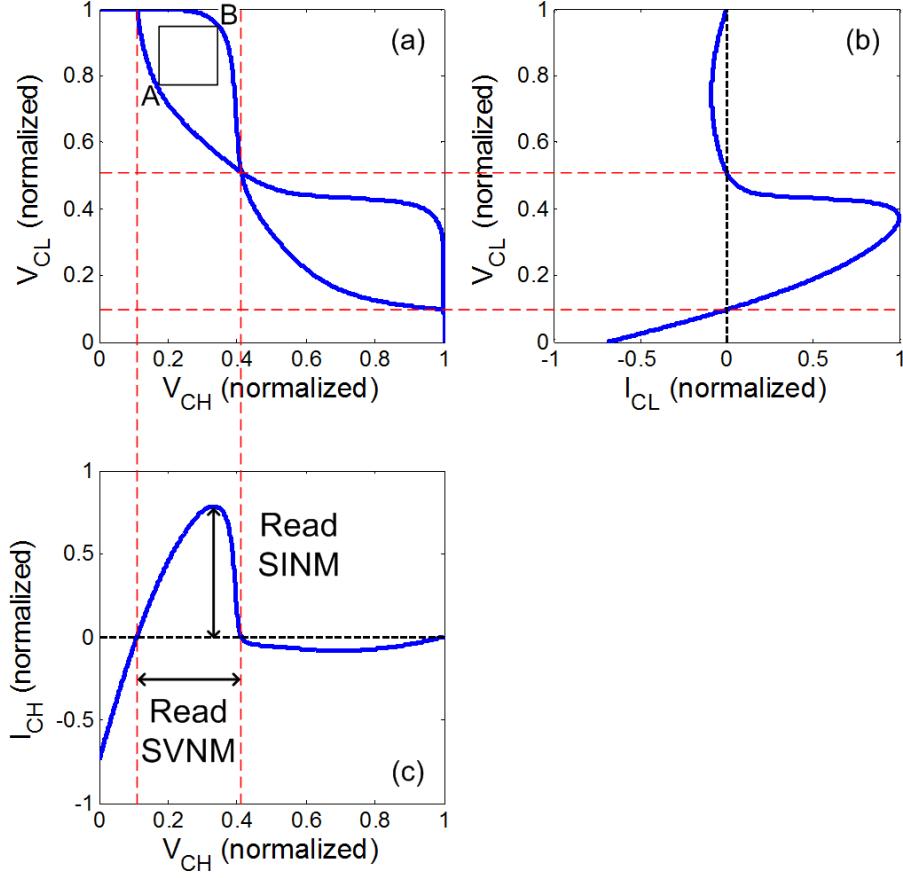


Figure 2.6: (a) Butterfly-curve for SNM extraction during the read operation. (b) N-curve for sweeping V_{CL} during the read operation (x- and y-axis reversed for comparison). (c) N-curve for sweeping V_{CH} during the read operation.

the N-curve (and above the line $y = 0$). Similar to the SNM, the SVNM/SINM/SPNM can be extracted for two data polarities - the SVNM/SINM/SPNM for storing a '0' at CH ('1' at CL) can be found by sweeping V_{CH} and the SVNM/SINM/SPNM for storing a '0' at CL ('1' at CH) can be found by sweeping V_{CL} .

Butterfly-curve versus N-Curve

Since the N-curve is a relatively recent concept, it is important to investigate its accuracy in estimating SRAM read stability against the well-studied SNM. Figure 2.7, Figure 2.9, and Figure 2.13 show the scatter plots of the three N-curve read stability metrics - SVNM, SINM, and SPNM - versus RSNM obtained from 3k-sample Monte Carlo (MC) simulations, with common-mode global variations in L_G , W , T_{OX} , and V_{TH} as well as random mismatch in V_{TH} for all transistors, using a commercial low-power 45nm CMOS process. The MC simulations are done at $V_{DD} = 1.1V$, $0.9V$, and $0.6V$ to examine the correlations between the N-curve metrics and RSNM under higher operating voltages, with high read stability, as well as at lower operating voltages, where the bitcells approach read stability failure. Figure 2.7a-b show a reasonable linear correlation, with some dispersion, between SVNM and RSNM

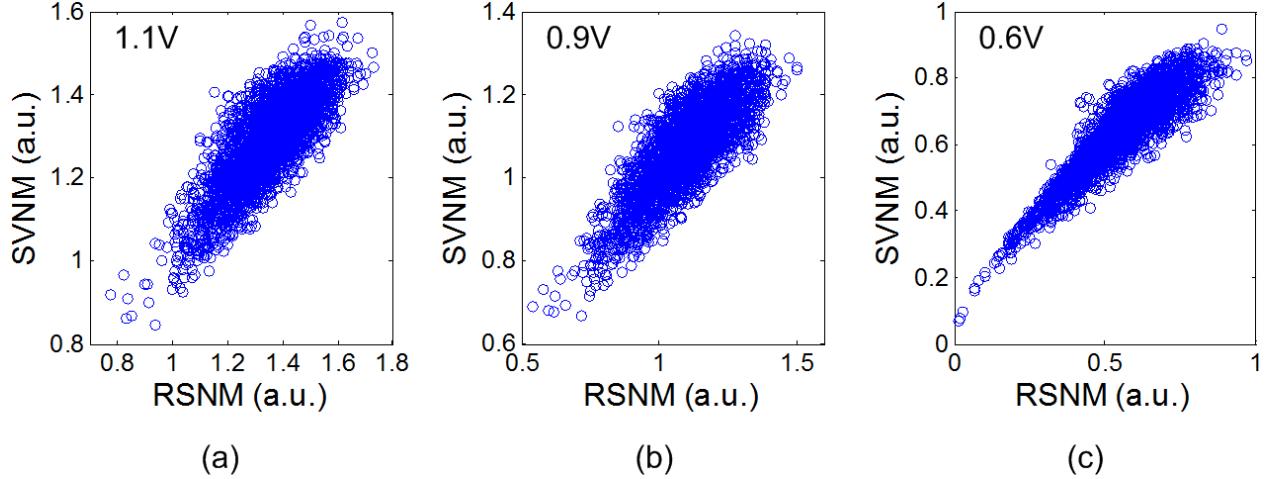


Figure 2.7: Scatter plots for SVNM versus RSNM obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$.

at higher operating voltages - $V_{DD} = 1.1V$, and $0.9V$. The dispersion can be attributed to the fact that while both RSNM and SVNM attempt to measure the size of the eye-opening in the butterfly-curve, the SVNM only captures the intersection points on the butterfly-curve whereas the RSNM tracks the actual VTC and measures the maximum separation in the eye-opening. As V_{DD} is reduced to $0.6V$, the correlation between SVNM and RSNM is significantly improved and a near 1-to-1 mapping is established between the two metrics near read stability failure, which is identified by the origin of Figure 2.7c. This indicates that both RSNM and SVNM share a common point of failure, which is in agreement with theory since $SVNM = 0$ implies a single intersection point at the eye-opening of the butterfly-curve and translates to a zero RSNM. However, while results indicate that the correlation between SVNM and RSNM is very good down to the zero crossing, it is important to note that a negative SVNM cannot be quantized, as the N-curve does not cross the $y = 0$ line, whereas a negative RSNM can be quantized as the side of the minimum square embedded between the VTC pair² at the region where the eye-opening no longer exists - this is similar to the write noise margin (WNM) definition discussed in Section 2.2.2. This scenario is described in Figure 2.8.

Figure 2.9 illustrates a marginal correlation between SINM and RSNM at higher operating voltages - $V_{DD} = 1.1V$, and $0.9V$. This correlation does not improve significantly with a decreasing V_{DD} until very close to the point of read stability failure - corresponding to the lower left corner of Figure 2.9c. The poor correlation between SINM and RSNM at higher operating voltages can be partially attributed to a difference in the sensitivities of a current noise margin (SINM) and a voltage noise margin (RSNM). In addition, a closer examination of Figure 2.6a reveals a difference in the bias conditions of SINM versus RSNM extraction - while the SINM is extracted near corner B of Figure 2.6a, the RSNM is extracted using both corners A and B [22]. More specifically, from Figure 2.6, the SINM is extracted

²The VTC pair can no longer be described as the butterfly-curve as the unstable data polarity does not produce an eye-opening.

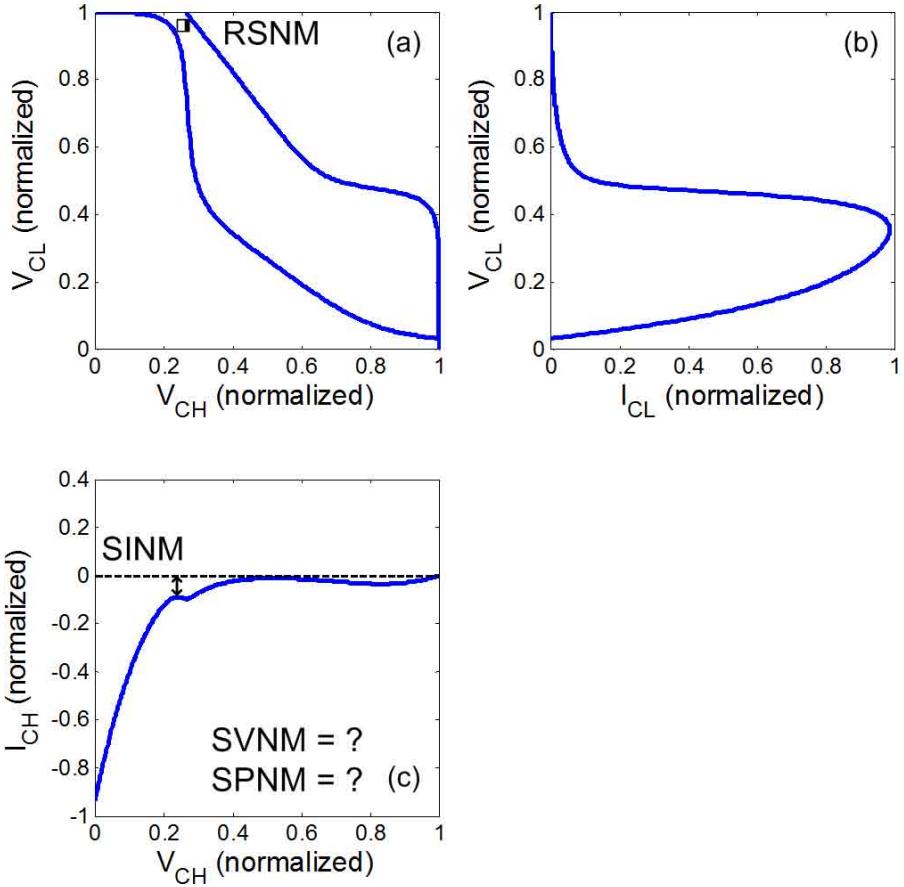


Figure 2.8: (a) VTC pair for an SRAM cell with a negative read margin. (b) N-curve while sweeping V_{CL} for an SRAM cell with a negative read margin (x- and y-axis reversed for comparison). (c) N-curve while sweeping V_{CH} for an SRAM cell with a negative read margin.

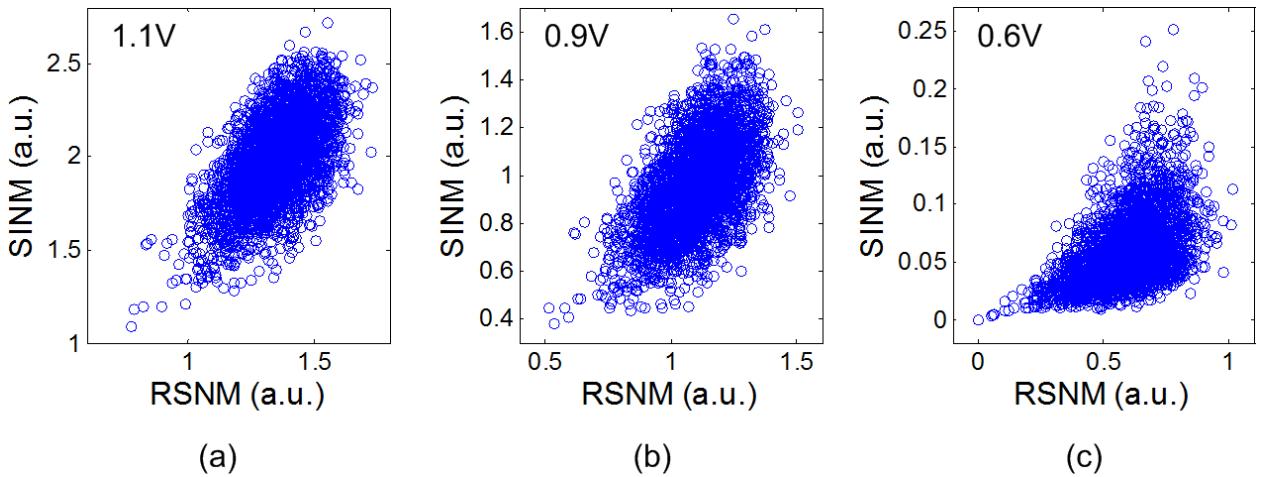


Figure 2.9: Scatter plots for SINM versus RSNM obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$.

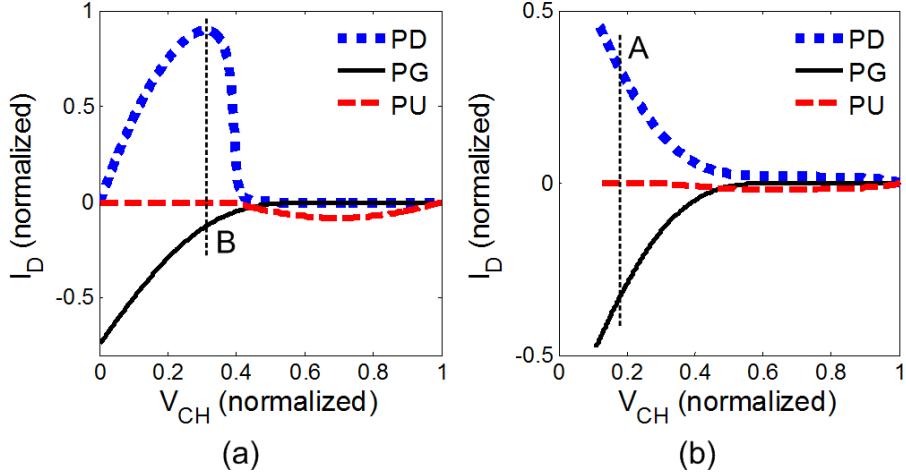


Figure 2.10: (a) I_D of pull-down, pass-gate, and pull-up transistors while sweeping V_{CH} for extracting the SINM near corner B . (b) I_D of pull-down, pass-gate, and pull-up transistors as a function of V_{CH} while sweeping V_{CL} for the characterization of corner A in the RSNM extraction.

as the current difference between N_R and N_{AXR} (refer to Figure 2.1) with both V_{CL} and V_{CH} determined near corner B of Figure 2.6a. The RSNM, on the other hand, is extracted as the maximum distance between the $V_{CL}(V_{CH})^3$ curve near corner B - using transistors P_L , N_L , and N_{AXL} - and the $V_{CH}(V_{CL})$ curve near corner A - using transistors P_R , N_R , and N_{AXR} . Therefore, transistors N_R and N_{AXR} are biased differently for extracting the RSNM (near corner A) and for extracting the SINM (near corner B) - this difference appears in both V_{CL} and V_{CH} and is approximately equal to the RSNM of the SRAM bitcell (i.e. the difference between A and B along the x- and y-axis). Figure 2.10 graphically illustrates the impact of this difference in bias on the drain currents of transistors N_R and N_{AXR} during SINM and RSNM extraction; where the pull-down transistor (N_R) is biased by both V_{CL} and V_{CH} and the pass-gate transistor is biased by V_{CH} . Due to the high current sensitivities of both the pull-down and the pass-gate transistors near corner A (for the extraction of the RSNM) and corner B (where the SINM is measured), this difference in bias helps to increase the dispersion between the extracted SINM and RSNM values.

Since the bias difference between the SINM and the RSNM extraction is approximately equal to the RSNM of the SRAM bitcell, the dispersion between the extracted SINM and RSNM values is expected to decrease significantly when V_{DD} is low (e.g at 0.6V). While this is true, a different phenomenon limits the correlation between SINM and RSNM when V_{DD} is decreased. This can be understood by examining Figure 2.5b, which reveals that, at the V_{CL} value where the N-curve peaks (i.e. where the SINM is measured), the pass-gate transistor approaches weak-inversion even at higher operating voltages - this is possible as high V_{TH} transistors are used and the V_{TH} of the pass-gate transistor is further increased by a reverse body bias (RBB) effect because its source node is biased at a higher potential than the P-well, which is biased at V_{SS} . As a result, the distribution of the pass-gate transistor current contribution is expected to deviate from that of a normal random variable. As the

³ $V_{CL}(V_{CH})$ corresponds to V_{CL} as a function of V_{CH} .

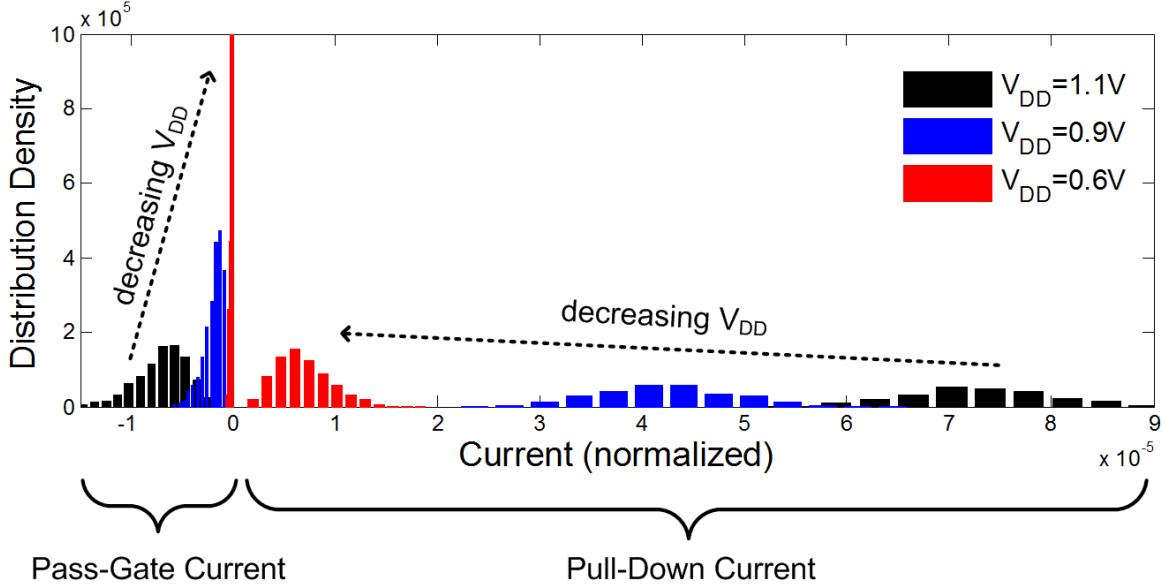


Figure 2.11: Distribution densities of the pass-gate transistor and the pull-down transistor current contributions to SINM at $V_{DD} = 1.1V$, $0.9V$, and $0.6V$.

operating voltage is further reduced, the pass-gate transistor approaches the subthreshold region, where the drain current varies exponentially with the transistor V_{TH} , and the distribution of the pass-gate transistor current contribution becomes log-normal. Figure 2.11 plots the distribution densities of the pass-gate transistor and the pull-down transistor current contributions to the SINM at $V_{DD} = 1.1V$, $0.9V$, and $0.6V$. As expected, the pass-gate transistor current distribution deviates from that of a normal random variable as V_{DD} is decreased from $1.1V$ and becomes log-normal as V_{DD} approaches $0.6V$. Figure 2.11 also shows that the pull-down transistor current contribution is normally distributed at $V_{DD} = 1.1V$ and $0.9V$. However, at $V_{DD} = 0.6V$, the pull-down transistor current distribution deviates from that of a normal random variable as the pull-down transistor approaches weak-inversion due to a high V_{TH} and a low V_{DD} . Since the SINM effectively measures the pull-down transistor current minus the pass-gate transistor current at the N-curve peak, the SINM distribution is also expected to deviate from that of a normal random variable as V_{DD} is reduced. Figure 2.12 plots the distribution densities of the RSNM and the SINM at $V_{DD} = 1.1V$, $0.9V$, and $0.6V$. As expected, the SINM data deviates from a normal distribution as V_{DD} is reduced to $0.6V$ ⁴ whereas the RSNM data continues to show a normal distribution. Consequently, elevated dispersion exists between the extracted SINM and RSNM values in Figure 2.9c at $V_{DD} = 0.6V$. However, Figure 2.9c does show a significant improvement in the correlation between SINM and RSNM near the point of read stability failure, indicating that the two metrics can track each other near failure and share a common zero-crossing - this is in agreement with theory since $SINM = 0$ coincides with $SVNM = 0$, which translates to a zero RSNM as the eye-opening in the butterfly-curve closes up completely. Figure 2.9c also

⁴The SINM data fits well to a normal distribution at $V_{DD} = 1.1V$ and $0.9V$ despite the deviation of the pass-gate transistor current from a normal distribution because the pull-down transistor current dominates in determining the SINM value.

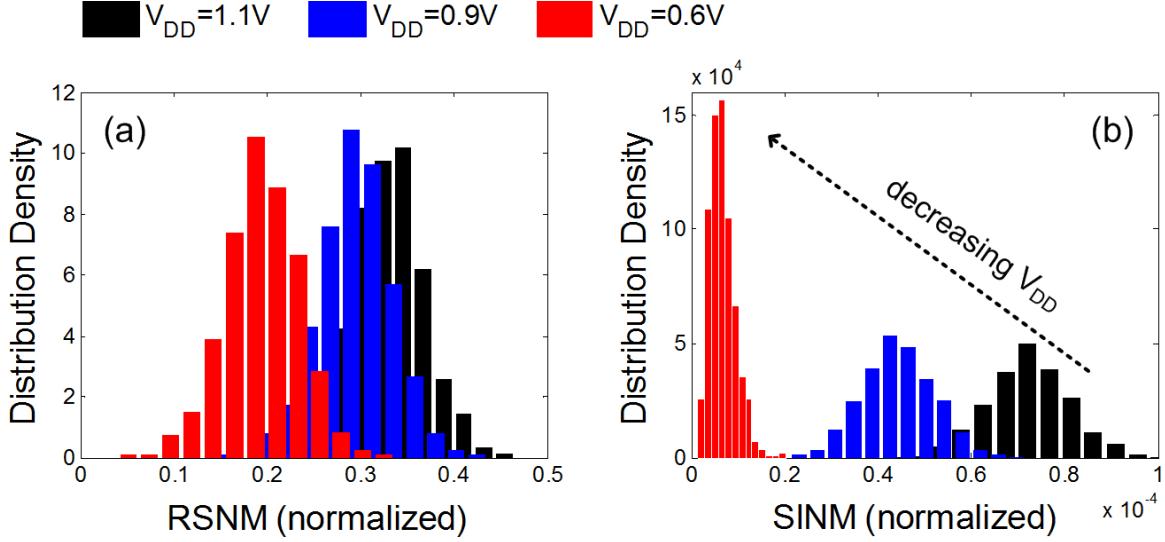


Figure 2.12: Distribution densities of (a) RSNM and (b) SINM at $V_{DD} = 1.1V$, $0.9V$, and $0.6V$.

indicates a reduction in the sensitivity of the SINM metric as compared to the RSNM near read stability failure - this is expected from the behavior of the log-normal distribution for the SINM at $V_{DD} = 0.6V$ (Figure 2.12). However, whereas a negative SVNM cannot be quantized, a negative SIVM can still be quantized as the first peak in the N-curve even if the peak is negative (Figure 2.8).

Figure 2.13 exhibits excellent correlation between SPNM and RSNM - even better than the correlation between SVNM and RSNM - at $V_{DD} = 1.1V$. However, the correlation between SPNM and RSNM degrades as the operating voltage is reduced - this can be attributed to the deviation of the extracted SINM from a normal distribution that exacerbates the correlation between SINM and RSNM at lower operating voltages, since the SPNM contains both voltage and current information. Nevertheless, results indicate that SPNM does track RSNM reasonably well down to $V_{DD} = 0.6V$. At $V_{DD} = 0.6V$, although the correlation degrades at higher SPNM and RSNM values, the lower tail in the scatter plot does show a significant improvement in the correlation. Similar to the SINM versus RSNM scatter plot in Figure 2.9c, Figure 2.13c bends in the direction of the origin, indicating that SPNM and RSNM also share a common zero-crossing. This is in agreement with theory since $SPNM = 0$ coincides with both $SINM = 0$ and $SVNM = 0$, which translates to a zero RSNM as the eye-opening in the butterfly-curve closes up completely. Similar to the SVNM, however, a negative SPNM cannot be quantized as the N-curve does not cross the $y = 0$ line (Figure 2.8).

The authors in [61] claim that the N-curve offers a more complete and proper definition of the SRAM read stability criteria over the butterfly-curve by providing both voltage and current information. In this claim, the authors highlight the comparison of the read stability between two SRAM bitcells with similar transistor ratios, but with the widths of all transistors in one bitcell sized up by $2\times$. The author shows that while the two bitcells exhibit the same RSNM and SVNM, the doubly-sized bitcell achieves nearly a $2\times$ increase

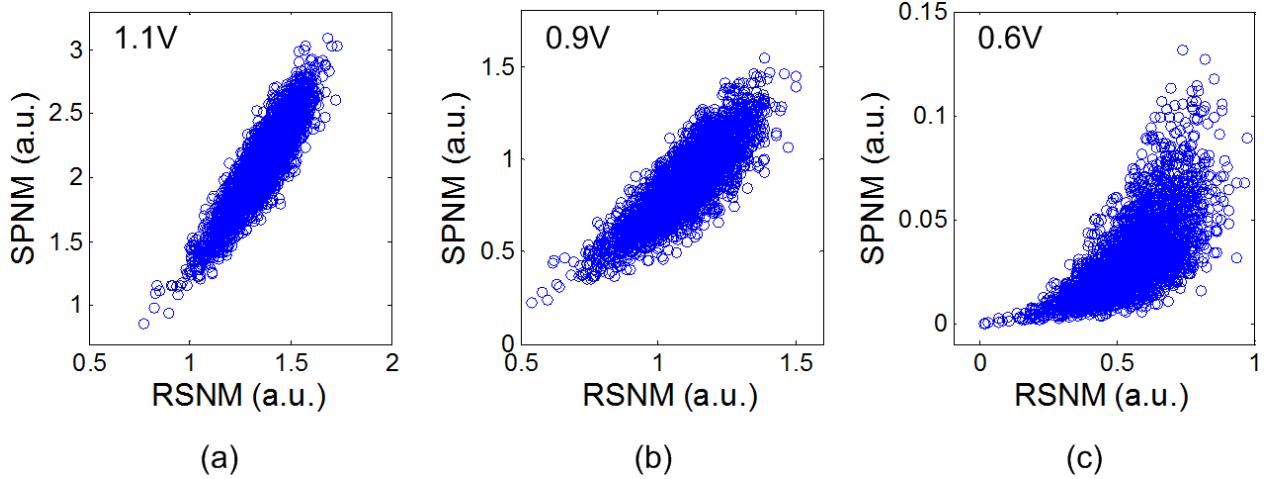


Figure 2.13: Scatter plots for SPNM versus RSNM obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$.

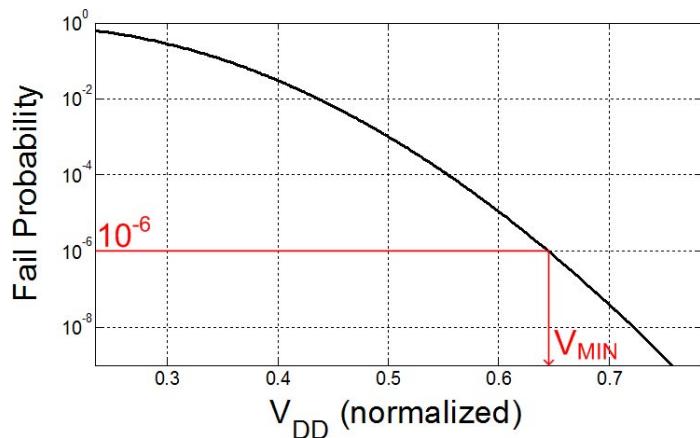


Figure 2.14: Fail probability as a function of V_{DD} - used to illustrate the definition of V_{MIN} .

in the SIVM - suggesting that this SRAM bitcell design can withstand up to $2\times$ more DC noise current at its storage nodes before data corruption and therefore should have higher read stability. However, to more properly compare the different read stability metrics, the SRAM yield should be considered.

Minimum Operating Voltage (V_{MIN})

A common method to quantify the yield of large SRAM arrays has been the SRAM array minimum operating voltage (V_{MIN}) [4, 14] - this quantification may be a partial consequence of embedded SRAM design becoming more and more power limited [160]. Figure 2.14 plots the SRAM fail probability as a function of V_{DD} . The SRAM array V_{MIN} can be defined as the highest V_{DD} able to keep the fail probability below a specified threshold. This threshold depends on the memory array size and the number of correctable failures through

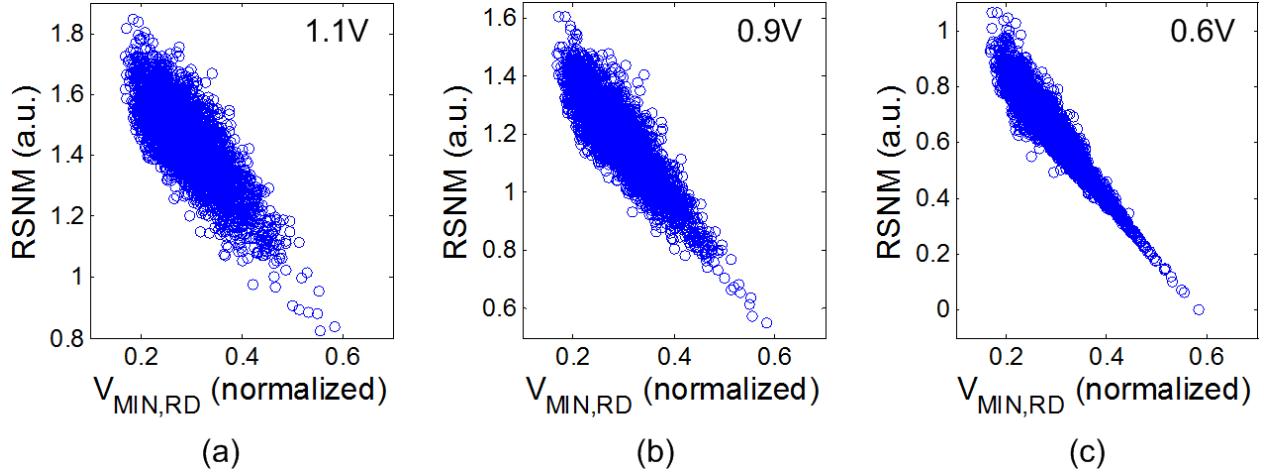


Figure 2.15: Scatter plots for RSNM versus $V_{MIN,RD}$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$.

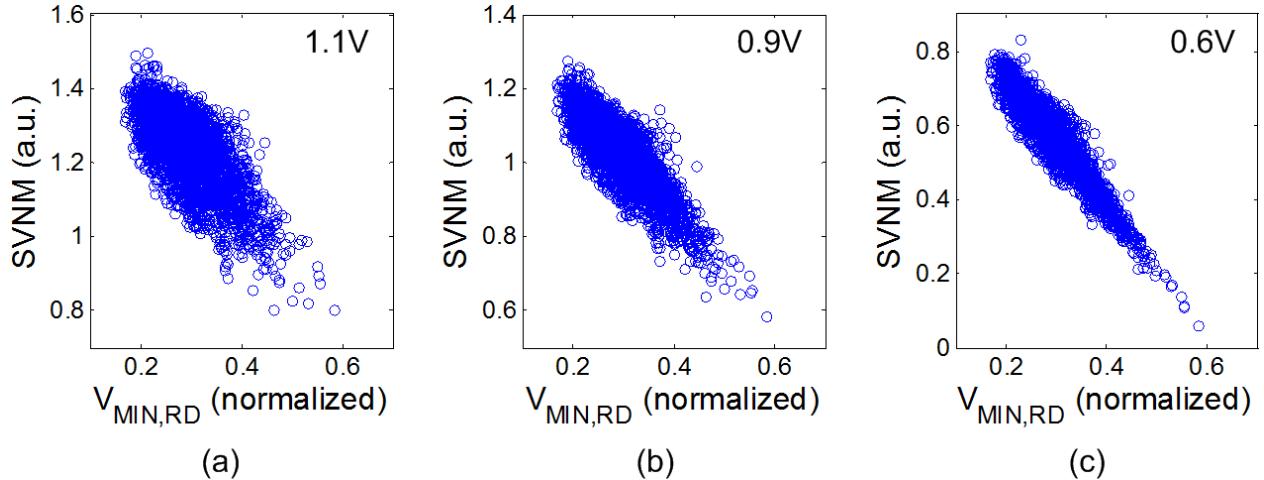


Figure 2.16: Scatter plots for SVNM versus $V_{MIN,RD}$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$.

error-correction codes (ECC) and/or redundancy. For example, consider a memory array of 1 million bitcells with no ECC/redundancy; as illustrated in Figure 2.14, its V_{MIN} is equal to the V_{DD} corresponding to a fail probability of 10^{-6} .

In the case of a static read operation, the fail probability corresponds to the probability for which read margin < 0 . To better assess the different read stability metrics, a per-cell V_{MIN} can be characterized during the read cycle (i.e. $V_{MIN,RD}$), along with the various metrics, as the V_{DD} for which read retention is no longer satisfied - i.e. by decreasing V_{DD} until a pre-initialized state is flipped.

Figures 2.15-2.18 show the scatter plots of all four aforementioned read stability metrics - RSNM, SVNM, SINM, and SPNM - versus $V_{MIN,RD}$ obtained from 3k-sample MC

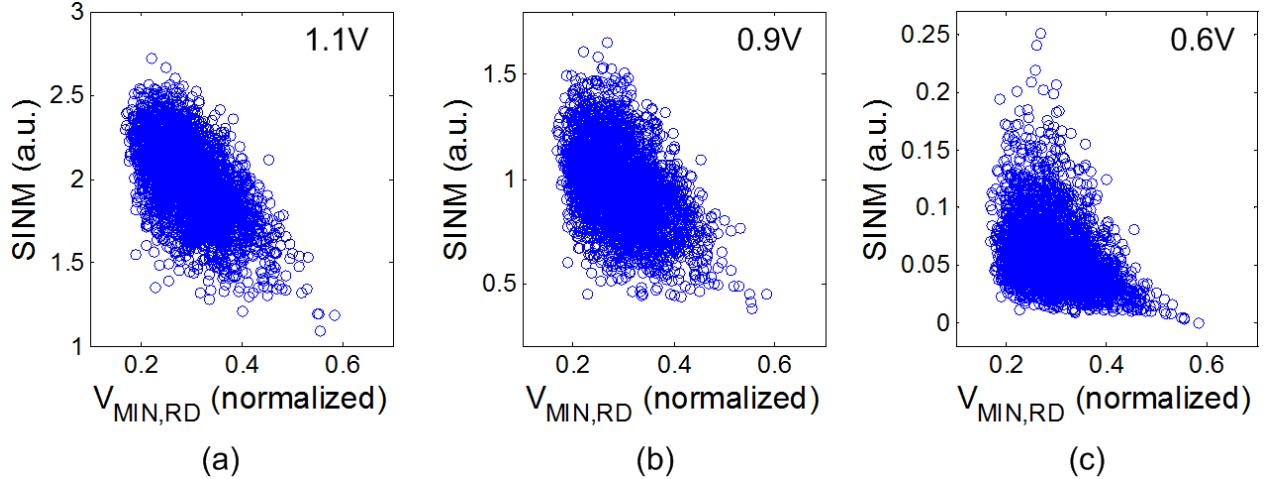


Figure 2.17: Scatter plots for SINM versus $V_{MIN,RD}$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$.

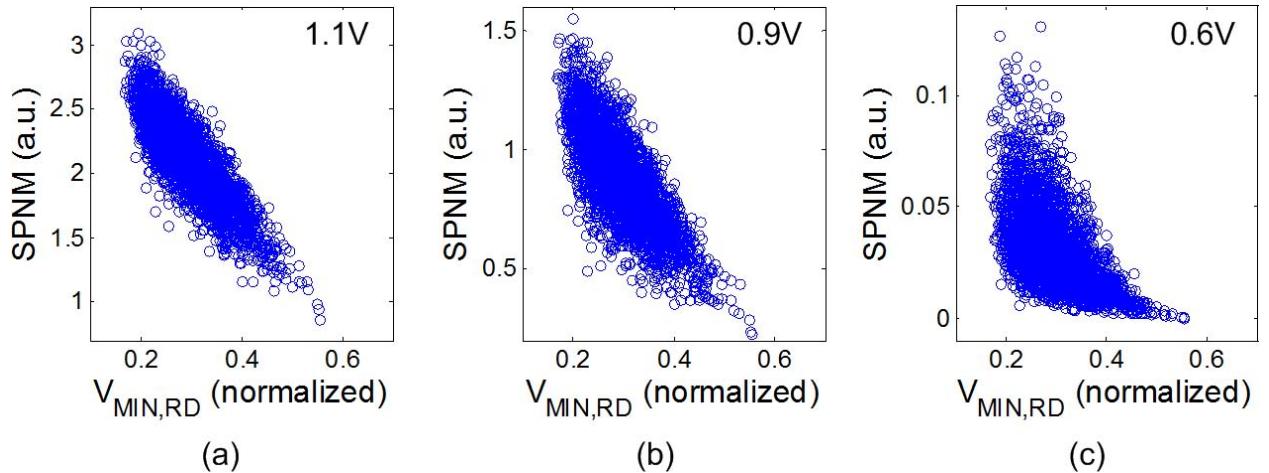


Figure 2.18: Scatter plots for SPNM versus $V_{MIN,RD}$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 1.1V$, (b) $V_{DD} = 0.9V$, and (c) $V_{DD} = 0.6V$.

simulations using a commercial low-power 45nm CMOS process. The MC simulations for the read stability metrics are done at $V_{DD} = 1.1V$, $0.9V$, and $0.6V$ to examine their correlations with $V_{MIN,RD}$ under higher operating voltages, with larger read margins, as well as at lower operating voltages, where the read margins approach zero. Figures 2.15 and 2.16 indicate good correlations between RS NM/SVN M and $V_{MIN,RD}$ at all operating voltages. The correlations between RS NM/SVN M and $V_{MIN,RD}$ improve as the operating voltage for read margin extraction is reduced from $1.1V$ to $0.6V$. At $V_{DD} = 0.6V$, the correlations are excellent and a near 1-to-1 mapping between RS NM/SVN M and $V_{MIN,RD}$ is established, indicating that both RS NM and SV NM can effectively track $V_{MIN,RD}$. Conversely, Figure 2.17 shows that the SINM does not correlate very well with $V_{MIN,RD}$. However, this cor-

relation improves significantly at $V_{DD} = 0.6V$ for SRAM cells near read stability failure - i.e. in the region of high $V_{MIN,RD}$ values and low SINM values. Figure 2.18 shows that SPNM can achieve a better correlation with $V_{MIN,RD}$ than SINM. However, the correlation between SPNM and $V_{MIN,RD}$ degrades as the operating voltage is reduced. Nevertheless, the lower-right tails of the SPNM- $V_{MIN,RD}$ scatter plots in Figure 2.18 do show decent correlations, especially at $V_{DD} = 0.6V$, indicating that the SPNM can reasonably track $V_{MIN,RD}$ for SRAM cells with low read stability. This investigation concludes that while the N-curve does offer both voltage noise margin and current (and also power) noise margin metrics, only the voltage margin metrics - RSNM and SVNM - can most effectively track the SRAM $V_{MIN,RD}$. However, Section 4.3.2 shows that the accuracy of $V_{MIN,RD}$ estimation using SVNM may be questionable due its inability to quantify a negative read margin. In addition, although SINM and SPNM demonstrate good correlation with $V_{MIN,RD}$ near read stability failure, their non-Gaussian distributions limit their utility in $V_{MIN,RD}$ estimation (Section 4.3.2).

2.2.2 Writeability Metrics

During the write cycle, the bit-lines (BL and BLC) are driven differentially by the data input and WL is asserted. Assuming that the storage node CH holds a '1' as in Figure 2.1 and BLC is driven low, the WL assertion forms a resistive voltage divider between the falling BLC and the storage node CH through transistors N_{AXR} and P_R . If the voltage divider pulls V_{CH} below the trip point of inverter $P_L - N_L$, a successful write operation takes place. Since the write operation depends on a voltage division between the pass-gate transistor (N_{AXR}) and the pull-up transistor (P_R), the writeability of an SRAM cell can be guaranteed by increasing the strength of the pass-gate transistor relative to the pull-up transistor - i.e. by increasing the SRAM cell α -ratio, which is defined as the strength ratio of the pass-gate transistor to the pull-up transistor. This is typically achieved by either increasing the pass-gate transistor channel width (W) or the pull-up transistor channel length (L_G) while keeping its W small - note that the requirements for writeability partially conflicts with the requirements for read stability presented in Section 2.2.1; for this reason, achieving balanced read stability and writeability remains as one of the most critical challenges in nanoscale SRAM design. Alternately, the SRAM cell α -ratio can be adjusted through the selection of the NMOS and PMOS transistor threshold voltages (V_{TH}) and/or by optimizing the electron- versus hole-mobility via channel strain.

Write Noise Margin

The SRAM writeability can be gauged by the write noise margin (WNM) [20, 21], which can be extracted from the voltage transfer characteristics (VTC) generated for the two halves of an SRAM cell biased for the write operation. Figure 2.19 graphically illustrates the two VTC pairs generated to characterize the SRAM writeability for both data polarities - i.e. writing a '0' into CL ('1' into CH) and writing a '0' into CH ('1' into CL). During the write operation, the bit-lines are differentially driven to either V_{DD} or V_{SS} corresponding to the data input. The ability to write a '0' into storage node CL ('1' into storage node CH) is assessed by a write-VTC - captured by sweeping V_{CH} while monitoring V_{CL} with WL and

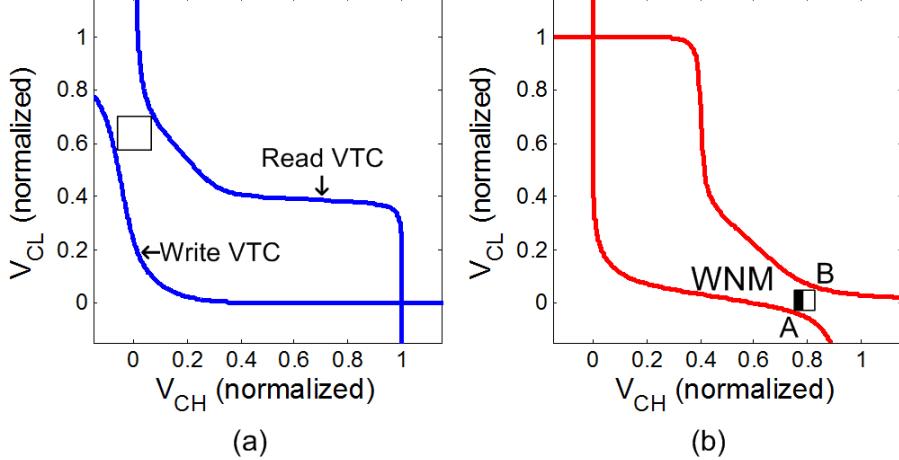


Figure 2.19: Definition of WNM from simulated VTC-pair for (a) writing a '0' into CL (or '1' into CH) and (b) writing a '1' into CH (or '0' into CL).

BLC biased at V_{DD} and BL biased at V_{SS} - and a read-VTC⁵ - captured by sweeping V_{CL} while monitoring V_{CH} with WL , BL , and BLC all biased at V_{DD} ; this VTC pair is plotted in Figure 2.19a. Similarly, the ability to write a '0' into storage node CH ('1' into storage node CL) is assessed by a read-VTC - captured by sweeping V_{CH} while monitoring V_{CL} with WL , BL , and BLC all biased at V_{DD} - and a write-VTC - captured by sweeping V_{CL} while monitoring V_{CH} with WL and BL biased at V_{DD} and BLC biased at V_{SS} ; this VTC pair is plotted in Figure 2.19b. (V_{CELL} is biased at V_{DD} for all above cases.) A 200mV sweep margin is added at the beginning and the end of the sweeping step - i.e. a $-0.2V$ to $1.1V$ storage node sweep is done for $V_{DD} = 0.9V$ (which requires only a $0V$ to $0.9V$ sweep). This is done to expose the convexity of the write-VTC for bitcells with higher writeability. Figure 2.19 illustrate a monostable circuit operation with a single intersection point corresponding to the written data polarity. The WNM of a bitcell for writing either data polarity is quantified by the side of the smallest square embedded between the corresponding read- and write-VTC pair located on the opposite half of the transfer curves - i.e. away from the stable point. The SRAM cell WNM is equal to the side of the smaller minimum-square (near corners A and B in Figure 2.19b). When WNM falls below zero, the write-VTC intersects the read-VTC, indicating a positive retention margin even when the bit-lines are differentially driven (while WL is asserted), thus suggesting an inability to write.

The value of the WNM can be analytically extracted from the VTC-pair by rotating both the x-axis and the y-axis by 45° , similar to the SNM extraction. The vertical distance between the two resulting curves correspond to the diagonal of each square that can be embedded within the VTC-pair. Figure 2.20a-b plots the rotated VTC-pair simulated for writing both data polarities into the SRAM cell. Figure 2.20c plots the length of the side of each embedded square. This length is plotted only for the region of interest - e.g. for writing a '0' into CL (Figure 2.20a), the length is plotted only for $x < 0$, corresponding

⁵This is referred to as the read-VTC because the transfer curves are the same as the VTC captured during the read SNM extraction. The measurement setups are similar as well - i.e. WL , BL and BLC all biased at V_{DD} while sweeping the storage node.

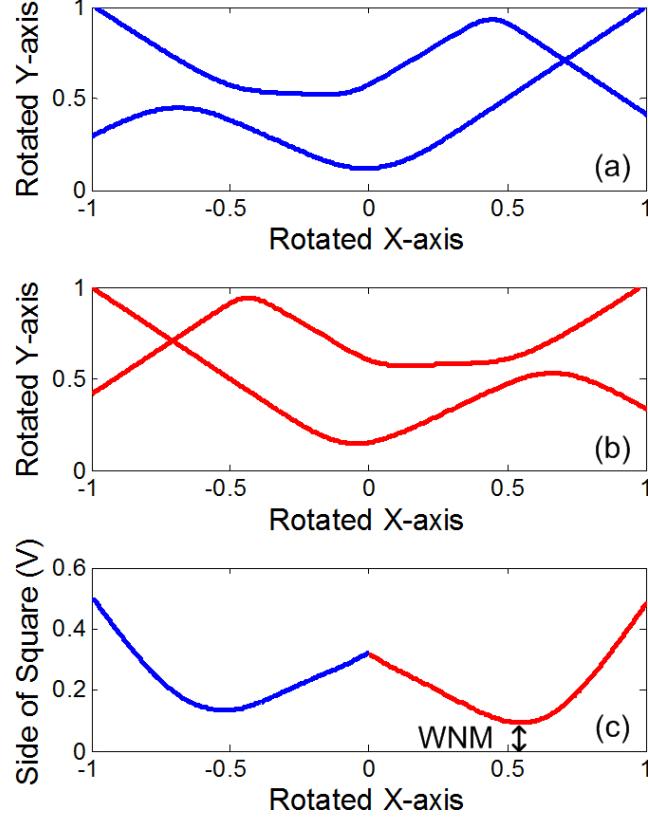


Figure 2.20: The VTC pair, rotated by 45° , for (a) writing a '0' into CL (or '1' into CH) and (b) writing a '0' into CH (or '1' into CL). (c) Length of the side of each embedded square within the rotated VTC pair.

to the upper-left half in Figure 2.19a; this is because $V_{CL} = V_{OL}$ and $V_{CH} = V_{OH}$ (where V_{OH} denotes the output-high voltage) should be the only stable root of the VTC-pair - this method looks for a second root away from this stable root, above the trip point of the inverter $P_R - N_{AXR} - N_R$ (i.e. for which $V_{CH} = V_{OL}$). The shorter valley in Figure 2.20c represents the WNM of the SRAM cell during the write cycle.

N-curve - Writeability Current

Alternate to WNM, the SRAM writeability can be characterized using the N-curve, similar to the case for read stability. Unlike the N-curve setup to characterize the read stability, however, the N-curve for writeability characterization, for writing a '0' into CL ('1' into CH), is measured by sweeping the voltage at the storage node CL with WL and BLC biased at V_{DD} and BL biased at V_{SS} while monitoring the current externally sourced into the CL node. Figure 2.21a plots the N-curve for writing a '0' into the CL node of an SRAM cell. Likewise, to characterize writeability for writing a '0' into CH ('1' into CL), V_{CH} is swept with WL and BL biased at V_{DD} and BLC biased at V_{SS} while monitoring the current externally sourced into the CH node. Figure 2.21b plots the N-curve for writing a '0' into the CH node of an SRAM cell. (V_{CELL} is biased at V_{DD} for both measurements.) The

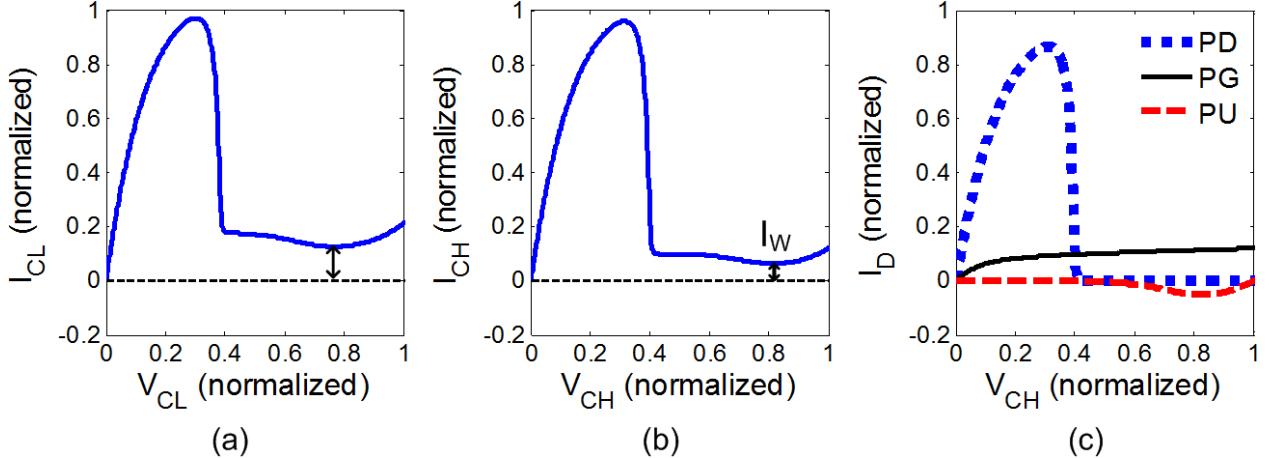


Figure 2.21: Definition of I_W from measured N-curve for (a) writing a '0' into CL (or '1' into CH) and (b) writing a '1' into CH (or '0' into CL). (c) I_D of pull-down, pass-gate, and pull-up transistors while sweeping V_{CH} during N-curve characterization for writing a '1' into CH .

relative current contributions of the pull-down, the pass-gate, and the pull-up transistors to the N-curve (for writing a '0' into CH) is illustrated in Figure 2.21c. The writeability current [30, 119], I_W , is defined as the current valley of the N-curve, located near corner B in Figure 2.19b, and effectively measures the pass-gate transistor current minus the pull-up transistor current as the current contribution from the pull-down transistor near the current valley is approximately zero. A larger I_W corresponds to a more writeable bitcell, while $I_W < 0$ represents a write failure. It should be noted that if $I_W \leq 0$, the write-VTC will intersect the read-VTC (in Figure 2.19b) at the same V_{CH} point(s) where the N-curve current I_{CH} intersects the line $y = 0$ (in Figure 2.21b), resulting in zero or negative WNM.

It has been reported in literature that the writeability metrics can be extracted using the exact same N-curve as the read stability metrics [61]. The write trip voltage (WTV) is defined as the voltage difference between the last 2 zero-crossings on the N-curve and represents the voltage drop needed to flip the data polarity of the SRAM bitcell when both bit-lines are biased at V_{DD} . The write trip current (WTI) is defined as the current valley of the N-curve and represents the current needed to flip the data polarity of the SRAM bitcell when both bit-lines are biased at V_{DD} . Figure 2.22 graphically illustrates the definitions of WTV and WTI on the N-curve. One important property of the WTV and the WTI is that they are measured with both bit-lines biased at V_{DD} and, therefore, do not reflect actual transistor operation during a write cycle, when the bit-lines are driven differentially. The most notable difference is in the pass-gate transistor operation. During WTV and WTI measurements, the source node⁶ of the pass-gate transistor at the sweeping storage node is tied to the bit-line and is biased at V_{DD} . As a result, a reverse body bias (RBB) is exerted on the pass-gate transistor. Conversely, during a write operation, the source node of the same pass-gate transistor is biased at V_{SS} and therefore exerts no RBB. In addition, the

⁶The source node corresponds to the drain/source terminal of the pass-gate transistor with a lower potential during a write operation.

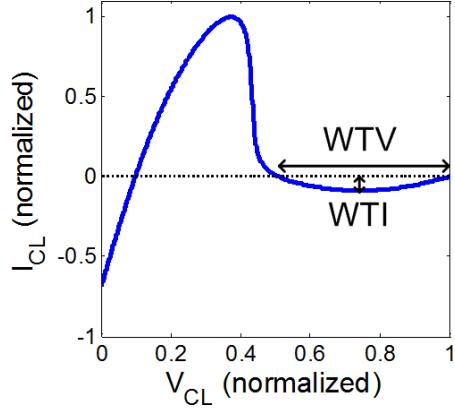


Figure 2.22: Definition of WTV and WTI from measured N-curve.

V_{DS} of the pass-gate transistor also differ in the two cases near the region where WTV and WTI are defined, thus exerting different degrees of drain induced barrier lowering (DIBL). Consequently, I_W , measured when the SRAM bitcell is biased under a write operation, is better suited for writeability characterization than WTV and WTI.

Write Noise Margin versus Writeability Current

Similar to the case for the read stability metrics, the correlation between I_W and WNM is examined. Figure 2.23 shows the scatter plot of I_W versus WNM obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process. The MC simulations are done at $V_{DD} = 0.9V$ and $0.6V$ to examine the correlations between I_W and WNM under higher operating voltage, with high writeability, as well as at lower operating voltages, where the bitcells approach writeability failure. At $V_{DD} = 0.9V$, the WNM values saturate as the write-VTC for SRAM cells with higher writeability fail to expose convexity even with a 200mV sweep margin. As a result, the correlation between I_W and WNM is exacerbated due to an error in the extraction process. In addition, the correlation between I_W and WNM also suffers due to similar reasons as the correlation between SINM and RSNM (Section 2.2.1); however, due to much lower sensitivities of the pass-gate transistor and the pull-up transistor currents near corners A (for the WNM extraction) and B (where the I_W is measured) in Figure 2.24, the correlation between I_W and WNM does not suffer as much from a difference in the bias point of the I_W versus the WNM extraction. As V_{DD} is reduced to $0.6V$, convexity is successfully exposed in the write-VTC for all or most SRAM cells and the correlation between I_W and WNM is significantly improved, especially near writeability failure. Elevated dispersion still exists in the region of high writeability as both the pull-up transistor and the pass-gate transistor enter the subthreshold region, resulting in a log-normal distribution for the extracted I_W - this is similar to the case for the SINM distribution (Section 2.2.1). However, better correlation is observed between I_W and WNM due to a lower sensitivity to the bias point in I_W versus WNM extraction. Moreover, Figure 2.23b reveals that I_W and WNM share the same failure point as the scatter plot crosses the origin.

To further investigate the accuracy of I_W and WNM in estimating the SRAM write-

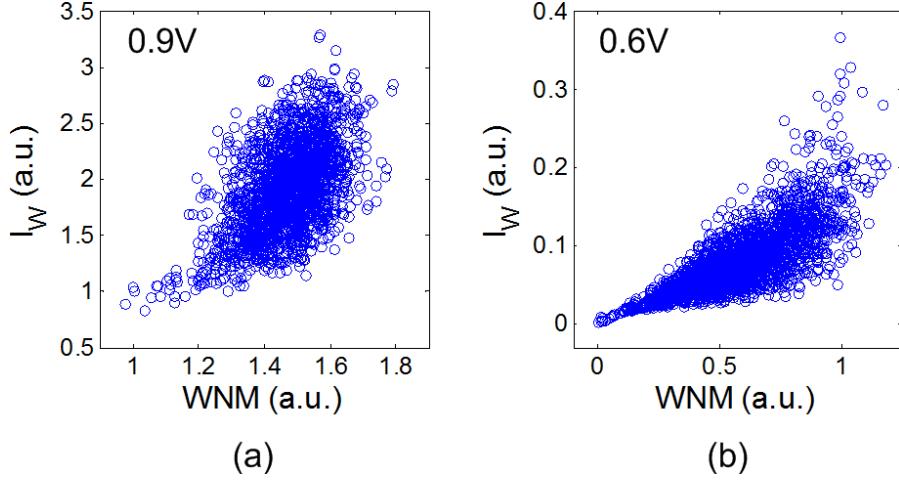


Figure 2.23: Scatter plots for I_W versus WNM obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 0.9V$ and (b) $V_{DD} = 0.6V$.

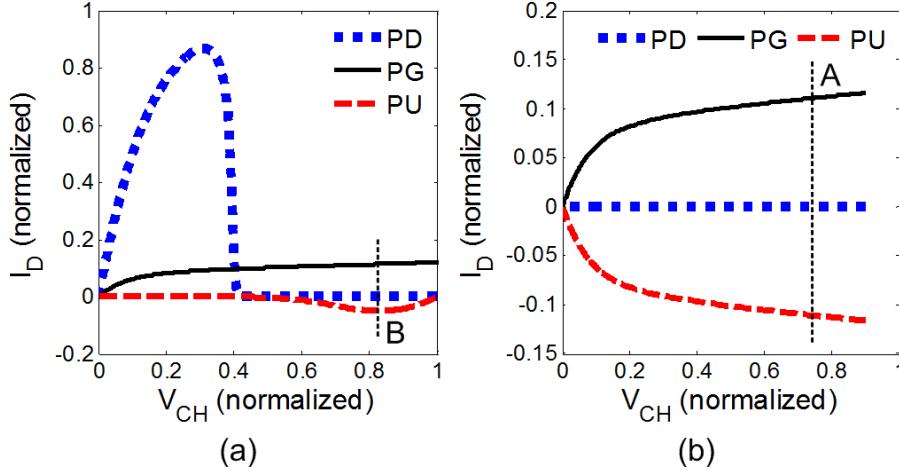


Figure 2.24: (a) I_D of pull-down, pass-gate, and pull-up transistors while sweeping V_{CH} for extracting the I_W near corner B. (b) I_D of pull-down, pass-gate, and pull-up transistors as a function of V_{CH} while sweeping V_{CL} for the characterization of corner A in the WNM extraction.

ability, their correlation with the SRAM cell V_{MIN} characterized for the write cycle - $V_{MIN,WRT}$ - is examined. The per-cell $V_{MIN,WRT}$ can be characterized as the V_{DD} for which a successful write operation can no longer take place - i.e. by decreasing V_{DD} until a pre-initialized state can no longer be flipped during a write operation. In the case of a static write operation, the fail probability (in Figure 2.14) corresponds to the probability for which write margin < 0.

Figures 2.25-2.26 show the scatter plots of WNM and I_W versus $V_{MIN,WRT}$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process. The MC simulations for WNM and I_W are done at $V_{DD} = 0.9V$ and $0.6V$ to examine their correlations with $V_{MIN,WRT}$ under higher operating voltages, with larger write margins, as

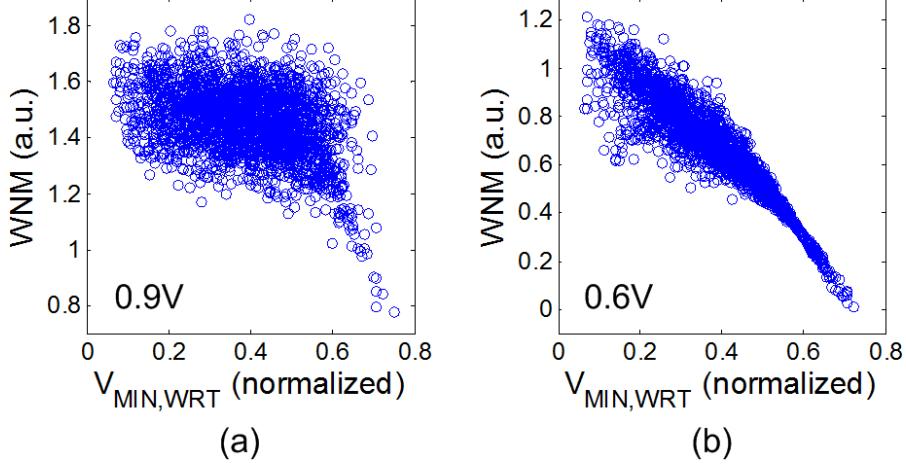


Figure 2.25: Scatter plots for WNM versus $V_{MIN,WRT}$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 0.9V$ and (b) $V_{DD} = 0.6V$

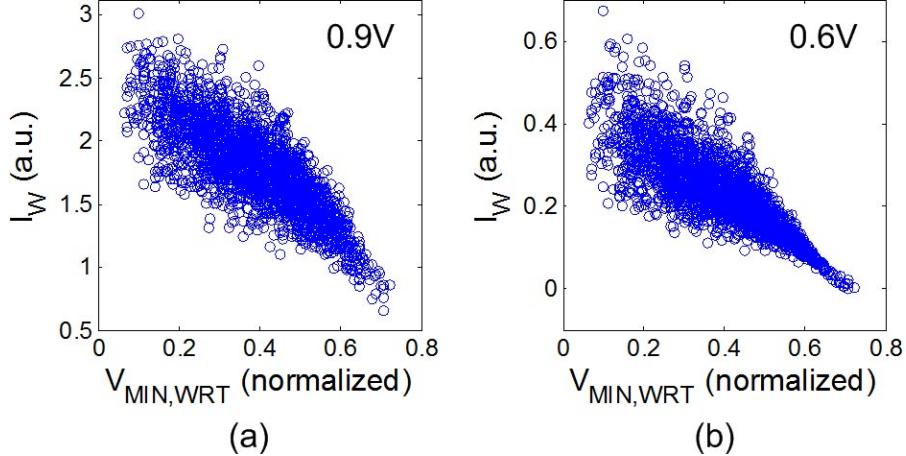


Figure 2.26: Scatter plots for I_W versus $V_{MIN,WRT}$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process at (a) $V_{DD} = 0.9V$ and (b) $V_{DD} = 0.6V$.

well as at lower operating voltages, where the write margins approach zero. Figure 2.25a reveals that the correlation between WNM and $V_{MIN,WRT}$ at $V_{DD} = 0.9V$ is limited due to a saturation in the WNM data near high write margin values. As mentioned, this is caused by an inability to expose convexity in the write-VTC for bitcells with high writeability. At $V_{DD} = 0.6V$ (Figure 2.25b), convexity is successfully exposed in the write-VTC and excellent correlation is established between WNM and $V_{MIN,WRT}$, achieving a near 1-to-1 mapping against $V_{MIN,WRT}$. Figure 2.26a exhibit good correlation between I_W and $V_{MIN,WRT}$ at $V_{DD} = 0.9V$. When V_{DD} is reduced to 0.6V (Figure 2.26b), the correlation between I_W and $V_{MIN,WRT}$ is significantly improved, especially in the region of high $V_{MIN,WRT}$ and low I_W values; although a larger dispersion is observed in the scatter plot for I_W versus $V_{MIN,WRT}$, especially in the region of high writeability. This investigation concludes that both WNM

and I_W can effectively track the SRAM $V_{MIN,WRT}$. However, the non-Gaussian distribution of the extracted I_W limits its utility in $V_{MIN,WRT}$ estimation (Section 4.3.2).

2.3 Large-Scale SRAM Design Metrics

2.3.1 Limitations of the Conventional Metrics

While the conventional DC read stability and writeability metrics presented in the previous section can effectively track the SRAM V_{MIN} during the read/write cycles, their measurements require access to the internal storage nodes. As a result, the major drawback associated with the conventional metrics is the inability to measure them in dense functional SRAM arrays because of the metal spacing constraints for routing out internal storage nodes and the significant area overhead associated with the switch array. This results in an insufficient number of data points for failure analysis of large cache memories. In addition, the removal of upper metal layers is necessary for routing out the internal storage nodes, which may require changes in the layout of the SRAM bitcell down to as low as the M1 layer. This, coupled with the fact that the conventional metrics are typically extracted from standalone test macros, may result in discrepancies between the measured conventional metrics and the actual read/write margins of SRAM cells in a functional array - i.e. conventional metrics measured from standalone test macros may not reflect the actual functionality of SRAM cells in a dense array. Furthermore, a direct correlation between the conventional metrics and the V_{MIN} in large functional SRAM arrays cannot be easily established on silicon⁷ as the conventional metrics are typically extracted from standalone test macros.

To increase the sample size and to allow direct correlation with per-cell V_{MIN} , the SRAM array must stay intact; in this case, the SRAM read stability and writeability must be characterized by accessing only the bit-lines, the word-line, and the cell supply voltages. As an example, bit-line access has been previously applied to detect and isolate faulty SRAM cells in memory arrays [153]. Similarly, large-scale performance of the SRAM cells has been characterized through distributions of the per-cell read currents (I_{READ}) [53] and the per-cell V_{MIN} [4, 14]. However, direct correlation between measured SRAM read/write margins and V_{MIN} in large functional SRAM arrays has not been established. However, direct correlation between measured SRAM read/write margins and V_{MIN} in large functional SRAM arrays has not been established. In this section, a method for characterizing the SRAM cell read stability and writeability in functional SRAM arrays [64, 65] by taking advantage of direct bit-line measurements while adjusting the bit-line, the word-line, and the cell supply voltages is presented. Furthermore, a method to characterize the per-cell V_{MIN} during standby, read, and write cycles using direct bit-line measurements is also described.

⁷Direct correlation between the conventional metrics and the SRAM V_{MIN} is established only in simulation in Section 2.2. The simulated SRAM cell V_{MIN} may not reflect the per-cell V_{MIN} measured from functional arrays.

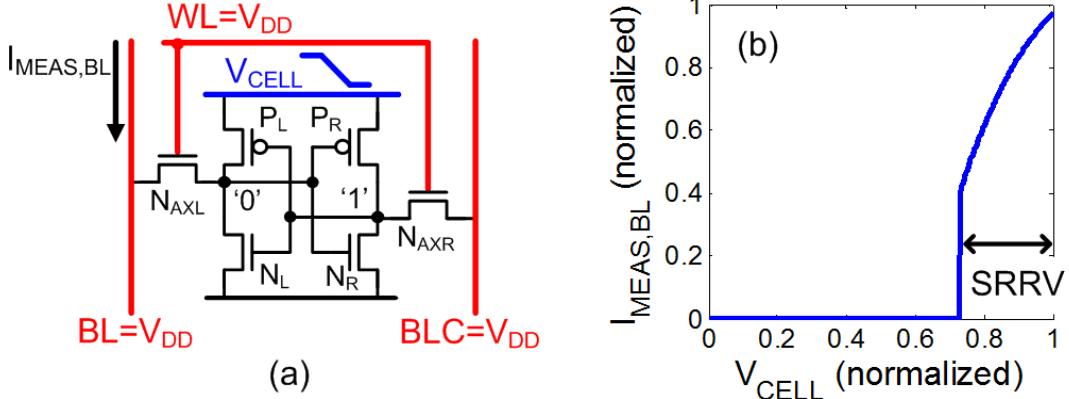


Figure 2.27: (a) Measurement setup for characterizing SRRV. (b) Definition of SRRV from simulated transfer curve.

2.3.2 Large-Scale Read Stability Metrics

Supply Read Retention Voltage

During the read cycle, both bit-lines (initially) float around V_{DD} while the word-line is driven high, and the cell state is retained by keeping the cell supply sufficiently high. The SRAM read stability in functional SRAM arrays can be measured as the lowest cell supply voltage for data retention during a read cycle - this is equivalent to finding the SRAM data retention voltage under a read operation and is denoted as the supply read retention voltage (SRRV) [64, 65]. Figure 2.27a graphically illustrates the measurement setup for characterizing SRRV. After the SRAM cell is first initialized to a known state, both BL and BLC are precharged at V_{DD} and WL is activated to emulate a read operation. The BL current at the '0' storage side, $I_{MEAS,BL}$ ⁸, is monitored while ramping down the SRAM cell supply voltage, V_{CELL} ⁹. When V_{CELL} is dropped sufficiently low, the SRAM cell loses its ability for data retention when N_{AXL} dominates N_L so that CL , originally holding a '0', rises above the trip point of inverter $P_R - N_R$. At that point, the cell state flips and a read upset occurs, signified by a sudden drop in $I_{MEAS,BL}$. The simulated transfer curve, of $I_{MEAS,BL}$ as a function of V_{CELL} , is plotted in Figure 2.27b. The difference between V_{DD} and the value of V_{CELL} causing $I_{MEAS,BL}$ to suddenly drop quantifies the SRRV of the SRAM cell. When $SRRV = 0$, the SRAM cell is biased for a nominal read operation with WL , BL , BLC and V_{CELL} all biased at V_{DD} . $SRRV > 0$ indicates that V_{CELL} can be dropped below V_{DD} , which decreases the gate-source voltage (V_{GS}) of the pull-down transistor (at the '0' storage side), without disturbing the stored data. Therefore, the SRRV effectively measures the maximum tolerable reduction in the cell β -ratio - through a reduction of the pull-down transistor V_{GS} , while maintaining the operating condition of the pass-gate transistor - before causing data corruption during the read cycle.

Intrinsic mismatch of transistors within an SRAM cell typically causes the bitcell to favor one data polarity over the other, resulting in an asymmetry in the cell robustness to

⁸If '0' is stored at CH , $I_{MEAS,BLC}$ is monitored instead.

⁹ V_{CELL} should be ramped from above V_{DD} when characterizing SRAM cells with negative read margin.

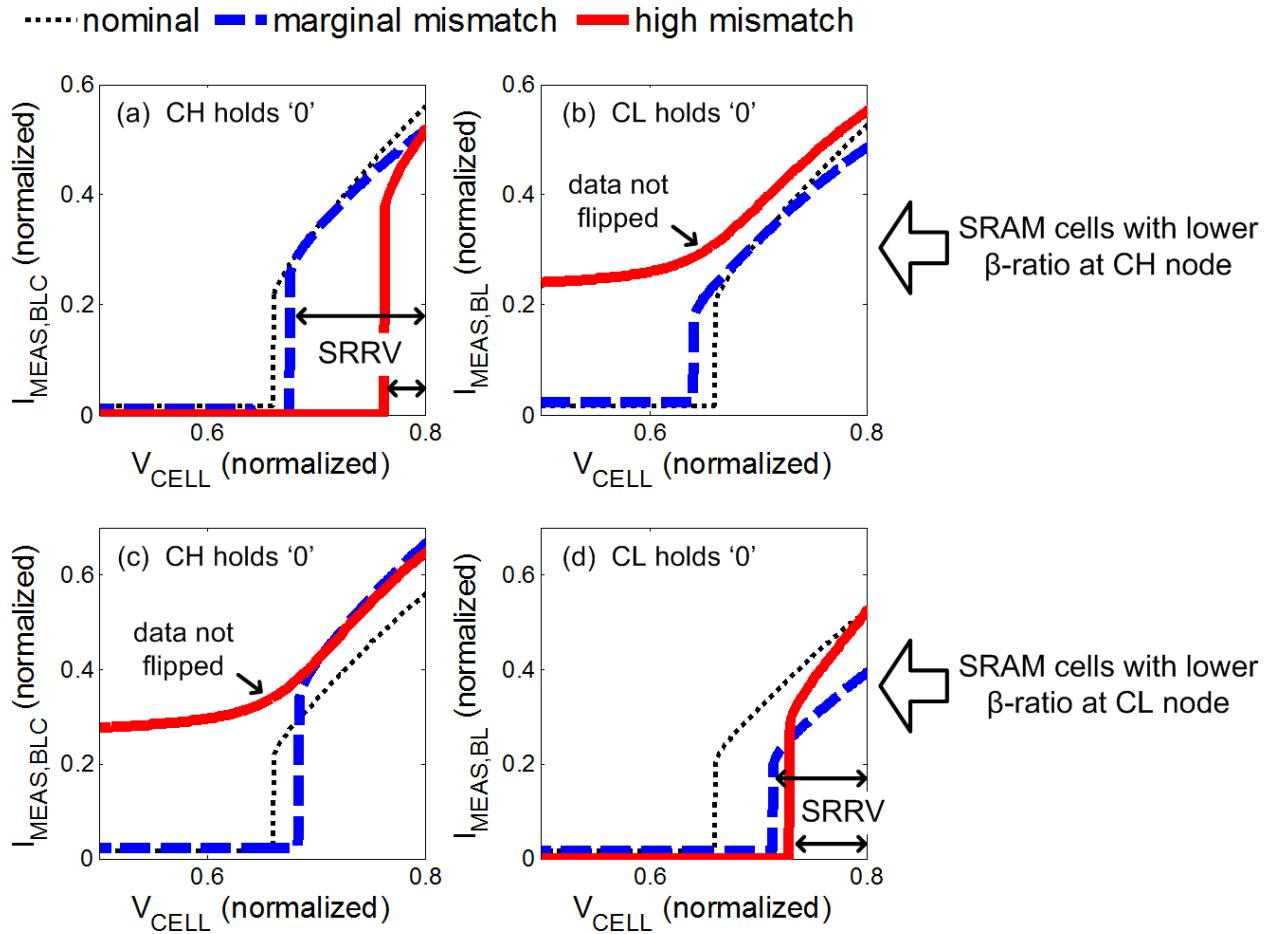


Figure 2.28: (a) SRRV transfer curves for storing a '0' at the less read-stable *CH* node; all transfer curves exhibit sharp fall off in $I_{MEAS,BLC}$. (b) SRRV transfer curves for storing a '0' at the more read-stable *CL* node; only the transfer curve corresponding to a marginal intrinsic mismatch exhibit a sharp fall-off in $I_{MEAS,BL}$, while the transfer curve corresponding to a high intrinsic mismatch shows a smooth bend in $I_{MEAS,BL}$. (c) SRRV transfer curves for storing a '0' at the more read-stable *CH* node; only the transfer curve corresponding to a marginal intrinsic mismatch exhibit a sharp fall-off in $I_{MEAS,BLC}$, while the transfer curve corresponding to a high intrinsic mismatch shows a smooth bend in $I_{MEAS,BLC}$. (c) SRRV transfer curves for storing a '0' at the less read-stable *CL* node; all transfer curves exhibit sharp fall off in $I_{MEAS,BL}$.

read upset between holding a '0' at CH ('1' at CL) and holding a '0' at CL ('1' at CH). Depending on the degree of this asymmetry, a data disturbance, in the form of a bit flip, can either occur for both data polarities or only for the less read-stable data polarity when the cell supply is dropped. As a result, the SRRV can be characterized for both data polarities in some SRAM cells while only for the less read-stable data polarity in other SRAM cells. Figure 2.28a-b highlights the SRRV transfer curves for two SRAM bitcells with worse read stability when CH holds a '0' - i.e. bitcells with lower cell β -ratio at the CH node. The two highlighted transfer curves (in bold) correspond to a bitcell with marginal intrinsic mismatch and a bitcell with high intrinsic mismatch. The SRRV transfer curve for a nominal SRAM cell, with no intrinsic mismatch, is added for comparison. Figure 2.28a shows that when a '0' is stored at the less read-stable CH node, both transfer curves exhibit a sharp fall-off in the BLC current ($I_{MEAS,BLC}$), indicating a clear SRAM cell data disturbance in the form of a bit flip. However, when a '0' is stored at the more read-stable CL node, only the transfer curve corresponding to a marginal intrinsic mismatch exhibit a sharp fall-off in the BL current ($I_{MEAS,BL}$); while the transfer curve corresponding to a high intrinsic mismatch shows a smooth bend in $I_{MEAS,BL}$ (Figure 2.28b). In the latter case, due to a heavily skewed read stability favoring the storage of a '0' at CL , a clear data disturbance, in the form of a bit flip, does not occur when the cell supply is dropped beyond data retention and the SRAM cell enters a metastable state. Consequently, the SRRV can only be characterized for storing a '0' at the less read-stable CH node. Figure 2.28c-d presents similar SRRV transfer curves for two SRAM bitcells with worse read stability when CL holds a '0' - i.e. lower cell β -ratio at the CL node - showing that SRRV can always be characterized for storing a '0' at the less read-stable CL node, but not for storing a '0' at the more read-stable CH node. To gauge the SRAM read stability, the SRRV value extracted for the less read-stable data polarity is used - this is equivalent to taking the side of the smaller maximum-square when extracting the SNM (Section 2.2.1).

Word-line Read Retention Voltage

When the word-line is driven high during a read/write cycle, both the SRAM cell under direct read access and all un-accessed SRAM cells driven by the asserted word-line undergo a read stress. This read stress can be exacerbated by boosting the word-line voltage beyond V_{DD} . Therefore, the read stability of an SRAM cell can also be measured by the largest word-line boost without upsetting cell data retention. This is denoted as the word-line read retention voltage (WRRV) [65]. Figure 2.29a graphically illustrates the measurement setup for characterizing WRRV. After the SRAM cell is first initialized to a known state, both BL and BLC are precharged at V_{DD} and WL is activated to emulate a read operation - identical to the SRRV characterization. The WL voltage is then ramped above V_{DD} ¹⁰, and kept below the gate-oxide breakdown voltage set by the technology, while the BL current at the '0' storage side ($I_{MEAS,BL}$) is monitored. When the WL voltage is boosted sufficiently high above V_{DD} , the SRAM cell state is disturbed due to an exacerbated read stress as N_{AXL} dominates N_L and pulls V_{CL} above the trip point of inverter $P_R - N_R$. The cell disturbance is captured as a sudden drop in the measured current $I_{MEAS,BL}$. The simulated transfer curve,

¹⁰ V_{WL} should be ramped from below V_{SS} when characterizing SRAM cells with negative read margin.

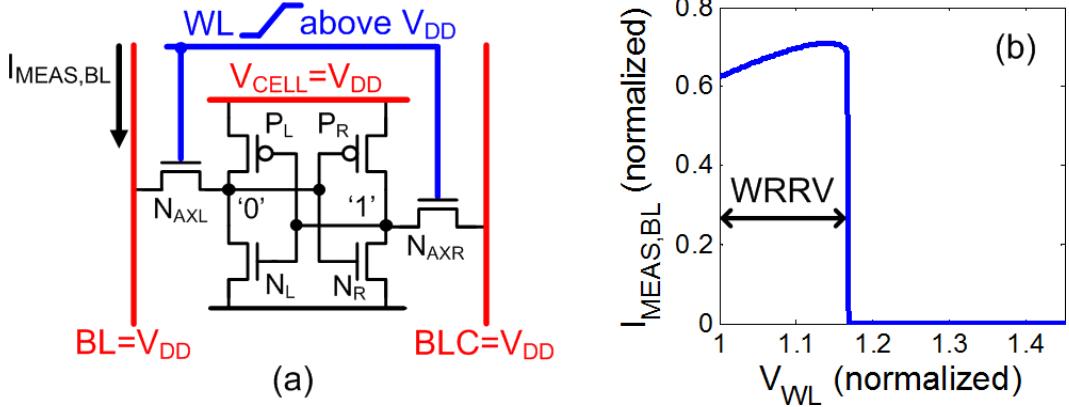


Figure 2.29: (a) Measurement setup for characterizing WRRV. (b) Definition of WRRV from simulated transfer curve.

of $I_{MEAS,BL}$ as a function of V_{WL} , is plotted in Figure 2.29b. The WRRV of an SRAM cell is quantified as the difference between the WL voltage causing $I_{MEAS,BL}$ to suddenly drop and V_{DD} . Similar to SRRV, when $WRRV = 0$, the SRAM cell is biased for a nominal read operation with WL , BL , BLC and V_{CELL} all biased at V_{DD} . $WRRV > 0$ indicates that V_{WL} can be boosted above V_{DD} , which increases the V_{GS} of the pass-gate transistor (at the '0' storage side), without disturbing the stored data. Therefore, the WRRV effectively measures the maximum tolerable reduction in the cell β -ratio - through an increase of the pass-gate transistor V_{GS} , while keeping the operating condition of the pull-down transistor relatively unchanged - before causing data corruption during the read cycle.

When the read stability of the SRAM bitcell becomes heavily skewed to favor the storage of either data polarity - holding a '0' at CH ('1' at CL) or holding a '0' at CL ('1' at CH) - the favored polarity will be preserved even under very high word-line boost. As a result, the monitored current $I_{MEAS,BL}$, depending on the degree of within-cell mismatch, may never drop significantly. Consequently, similar to the case for SRRV characterization, the WRRV can be characterized for both data polarities in some SRAM cells while only for the less read-stable data polarity in other SRAM cells. Figure 2.30a-b highlights the WRRV transfer curves for two SRAM bitcells with worse read stability when CH holds a '0' - i.e. bitcells with lower cell β -ratio at the CH node. The two highlighted transfer curves (in bold) correspond to a bitcell with marginal intrinsic mismatch and a bitcell with high intrinsic mismatch. The WRRV transfer curve for a nominal SRAM cell, with no intrinsic mismatch, is added for comparison. Figure 2.30a shows that when a '0' is stored at the less read-stable CH node, both transfer curves exhibit a sharp fall-off in the BLC current ($I_{MEAS,BLC}$), indicating a clear SRAM cell data disturbance. However, when a '0' is stored at the more read-stable CL node, only the transfer curve corresponding to a marginal intrinsic mismatch exhibit a sharp fall-off in the BL current ($I_{MEAS,BL}$); while the transfer curve corresponding to a high intrinsic mismatch shows a smooth $I_{MEAS,BL}$ (Figure 2.30b). In the latter case, due to a heavily skewed read stability favoring the storage of a '0' at CL , the cell state is not disturbed by the overdriven WL . As a result, the WRRV can only be characterized for storing a '0' at the less read-stable CH node. Figure 2.30c-d presents similar WRRV transfer curves for two SRAM bitcells with worse read stability when CL

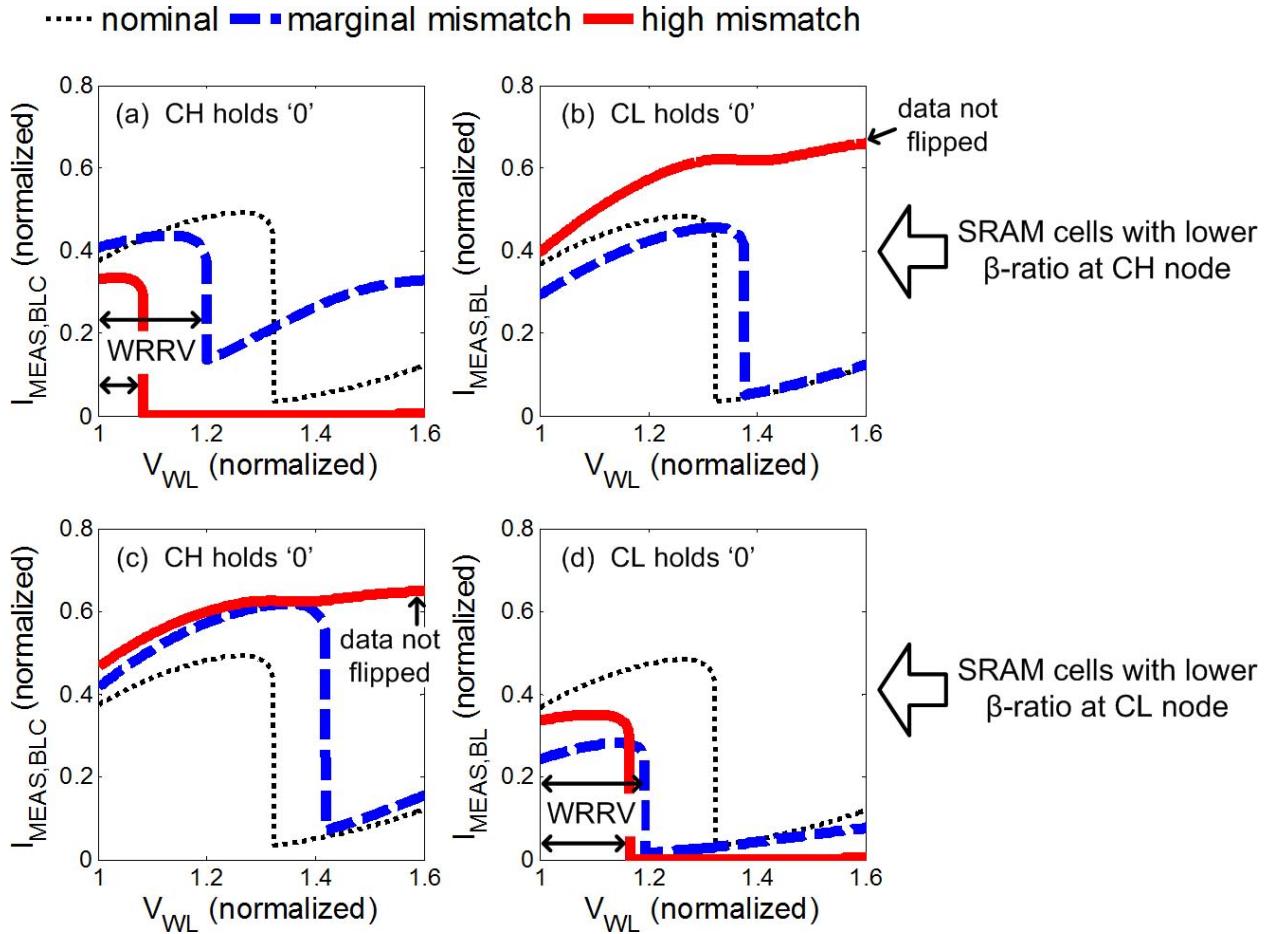


Figure 2.30: (a) WRRV transfer curves for storing a '0' at the less read-stable *CH* node; all transfer curves exhibit sharp fall off in $I_{MEAS,BLC}$. (b) WRRV transfer curves for storing a '0' at the more read-stable *CL* node; only the transfer curve corresponding to a marginal intrinsic mismatch exhibit a sharp fall-off in $I_{MEAS,BL}$, while the transfer curve corresponding to a high intrinsic mismatch shows a smooth $I_{MEAS,BL}$. (c) WRRV transfer curves for storing a '0' at the more read-stable *CH* node; only the transfer curve corresponding to a marginal intrinsic mismatch exhibit a sharp fall-off in $I_{MEAS,BLC}$, while the transfer curve corresponding to a high intrinsic mismatch shows a smooth $I_{MEAS,BLC}$. (c) WRRV transfer curves for storing a '0' at the less read-stable *CL* node; all transfer curves exhibit sharp fall off in $I_{MEAS,BL}$.

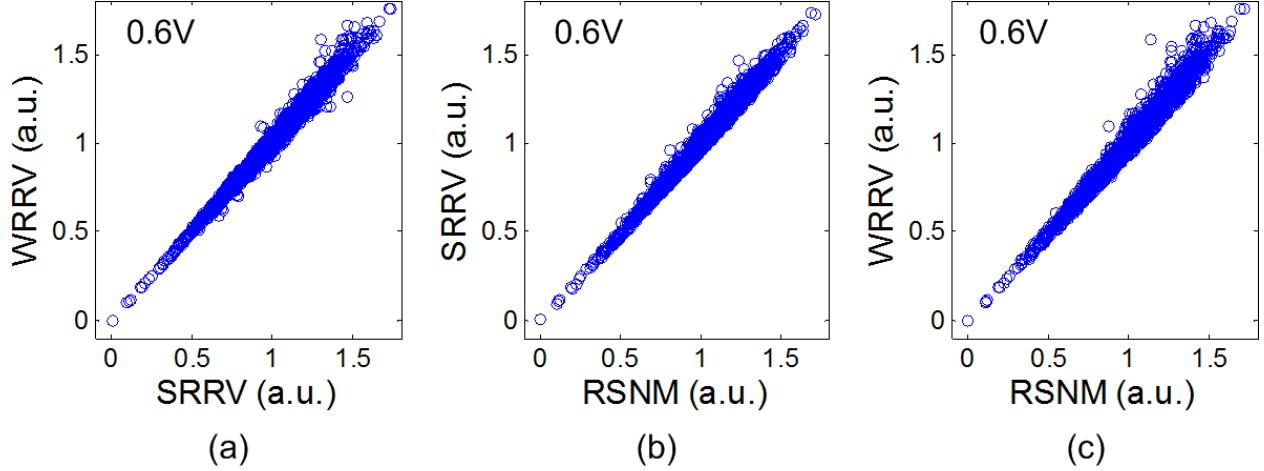


Figure 2.31: Scatter plots for (a) SRRV versus WRRV, (b) SRRV versus RSNM, and (c) WRRV versus RSNM at $V_{DD} = 0.6V$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process.

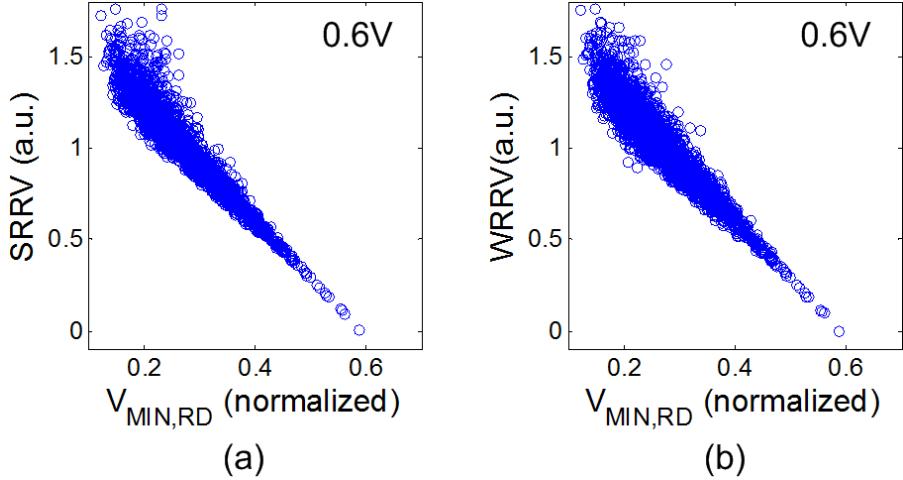


Figure 2.32: Scatter plots for (a) SRRV versus $V_{MIN,RD}$ and (b) WRRV versus $V_{MIN,RD}$ at $V_{DD} = 0.6V$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process.

holds a '0' - i.e. lower cell β -ratio at the CL node - showing that WRRV can always be characterized for storing a '0' at the less read-stable CL node, but not for storing a '0' at the more read-stable CH node. To gauge the SRAM read stability, the WRRV value extracted for the less read-stable data polarity is used.

Large-Scale Read Stability Metrics versus Conventional RSNM

With the large-scale read stability metrics defined, it is of interest to see how these metrics correlate with the conventional RSNM and the per-cell $V_{MIN,RD}$. Figure 2.31 shows the scatter plots for SRRV versus WRRV (in a), SRRV versus RSNM (in b), and WRRV

versus RSNM (in c) at $V_{DD} = 0.6V$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process. Results indicate excellent correlations between all three metrics. Figure 2.32 shows the scatter plots for SRRV and WRRV versus the per-cell $V_{MIN,RD}$. Excellent correlations are established between SRRV/WRRV and $V_{MIN,RD}$. This indicates that both SRRV and WRRV can effectively track the SRAM $V_{MIN,RD}$ and are therefore well suited for $V_{MIN,RD}$ estimation.

2.3.3 Large-Scale Writeability Metrics

Bit-line Write Trip Voltage

During the write cycle, the bit-lines are driven differentially according to the data input and the word-line is driven high. The writeability of an SRAM cell in a functional SRAM array can be gauged by the maximum bit-line voltage, at the '1' storage node, able to flip the cell state during a write cycle [52, 58, 64, 65, 68]. This is denoted as the bit-line write trip voltage (BWTV). Figure 2.33a graphically illustrates the measurement setup for characterizing BWTV. To capture the BWTV of an SRAM cell, the SRAM cell is first initialized to a known state, after which the cell supply (V_{CELL}), WL , BL , and BLC are all biased at V_{DD} . The BLC voltage at the '1' storage side is then ramped down while the BL current at the '0' storage side ($I_{MEAS,BL}$) is monitored. As the BLC voltage is ramped low, the pass-gate N_{AXR} overcomes P_R and the '1' storage is dropped below the inverter $P_L - N_L$ trip point, resulting in a successful write operation. This flip in the data polarity is signified by a sudden drop in $I_{MEAS,BL}$. Figure 2.33b plots the simulated transfer curve of $I_{MEAS,BL}$ as a function of the BLC voltage. The BWTV is quantified as the BLC voltage that induces a sudden change in $I_{MEAS,BL}$. When $BWTV = 0$, the SRAM cell is biased for a nominal write operation with WL , BL (or BLC), and V_{CELL} biased at V_{DD} and BLC (or BL) biased at V_{SS} . $BWTV > 0$ indicates that a successful write operation can take place even with a BLC (or BL) voltage higher than V_{SS} - i.e. with a decreased V_{GS} and a decreased drain-source voltage (V_{DS}) for the pass-gate transistor (at the '1' storage side), compared to a nominal write operation. Therefore, the BWTV effectively measures the maximum tolerable reduction in the cell α -ratio - through a reduction of the pass-gate transistor V_{GS} and V_{DS} , while maintaining the operation condition of the pull-up transistor - to still successfully write the SRAM cell.

Since the measurement setup for BWTV requires first exerting a read stress on the SRAM cell under test (CUT), a read disturb may occur in the SRAM CUT before the BWTV can be captured when testing at a reduced V_{DD} . However, due to intrinsic mismatch of transistors within an SRAM bitcell, a read disturbance at a high enough supply voltage typically only happens for the less read-stable data polarity while a read disturbance for the other data polarity may happen at a lower V_{DD} or may never happen at all (Figure 2.28). In this case, the BWTV can continue to be characterized for the more read-stable data polarity, which typically corresponds to the data polarity that is more difficult to overcome during a write operation.

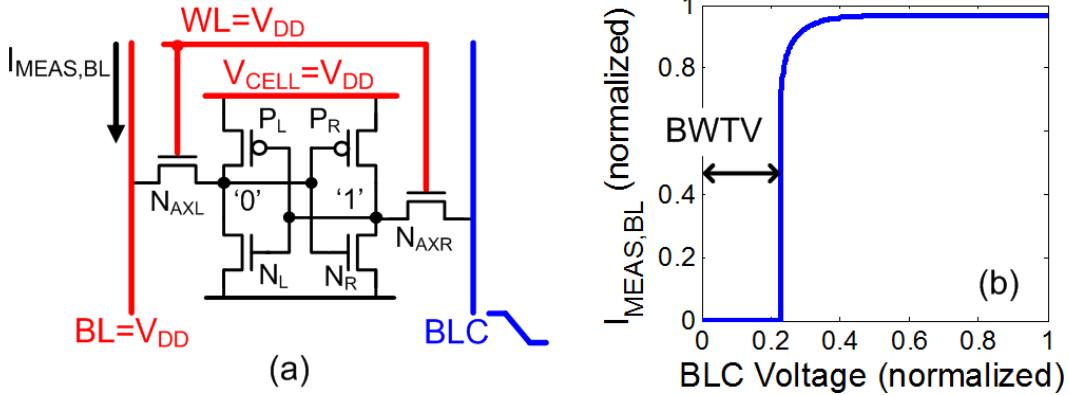


Figure 2.33: (a) Measurement setup for characterizing BWTV. (b) Definition of BWTV from simulated transfer curve.

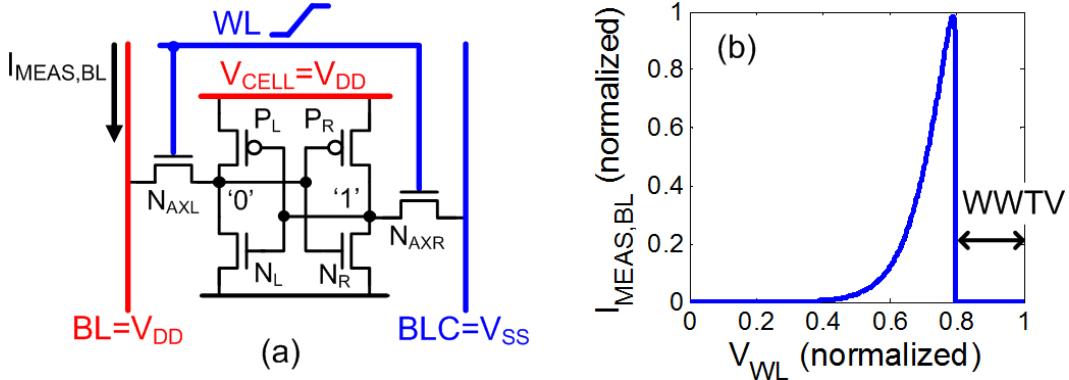


Figure 2.34: (a) Measurement setup for characterizing WWTV. (b) Definition of WWTV from simulated transfer curve.

Word-line Write Trip Voltage

The writeability of an SRAM cell can also be characterized by first configuring the bit-lines according to the data input and then ramping up the word-line [58, 64, 65]. The minimum word-line voltage able to flip the SRAM cell state during a write cycle, denoted as the word-line write trip voltage (WWTV), can be used to gauge the SRAM writeability. Figure 2.34a graphically illustrates the measurement setup for characterizing WWTV. Again, the SRAM cell is first initialized to a known state. The cell supply (V_{CELL}) and BL , at the '0' storage side, are then biased at V_{DD} while BLC , at the '1' storage side, is biased at V_{SS} - in accordance with the data input. Finally, V_{WL} is ramped high while the BL current at the '0' storage side ($I_{MEAS,BL}$) is monitored. As V_{WL} is increased, the BL current initially resembles the $I_D - V_G$ curve of the pass-gate N_{AXR} . When the WL voltage is sufficiently high, the cell state flips, leading to a successful write operation. This flip in the data polarity is signified by a sudden drop in the magnitude of $I_{MEAS,BL}$. Figure 2.34b plots the simulated transfer curve of $I_{MEAS,BL}$ as a function of V_{WL} . The WWTV is quantified as the value $V_{DD} - V_{WL}$, where V_{WL} is the minimum WL voltage causing the sudden drop in $I_{MEAS,BL}$. Similar to BWTV, when $WWTV = 0$, the SRAM cell is biased for a nominal write operation with

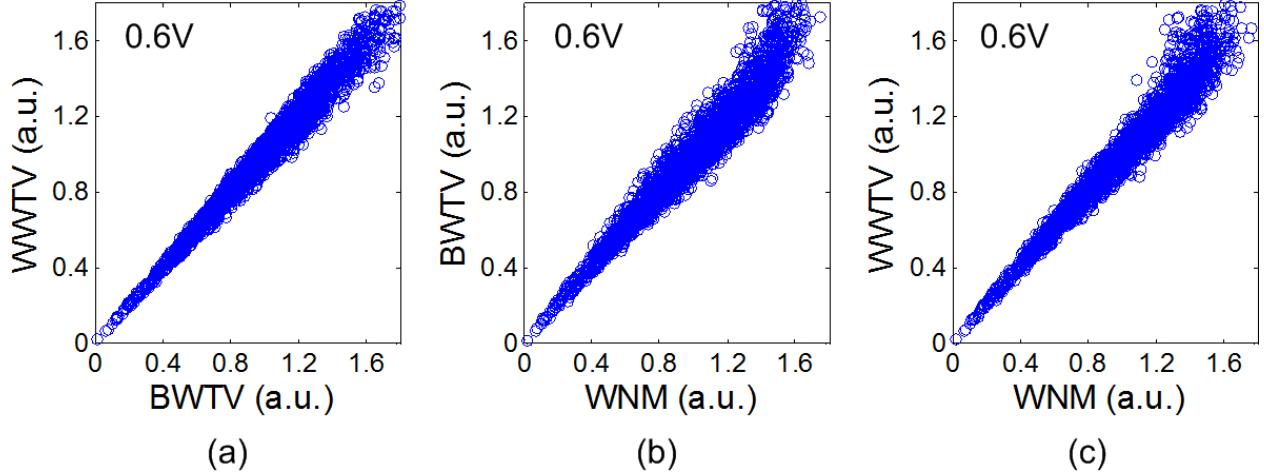


Figure 2.35: Scatter plots for (a) BWTV versus WWTV, (b) BWTV versus WNM, and (c) WWTV versus WNM at $V_{DD} = 0.6V$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process.

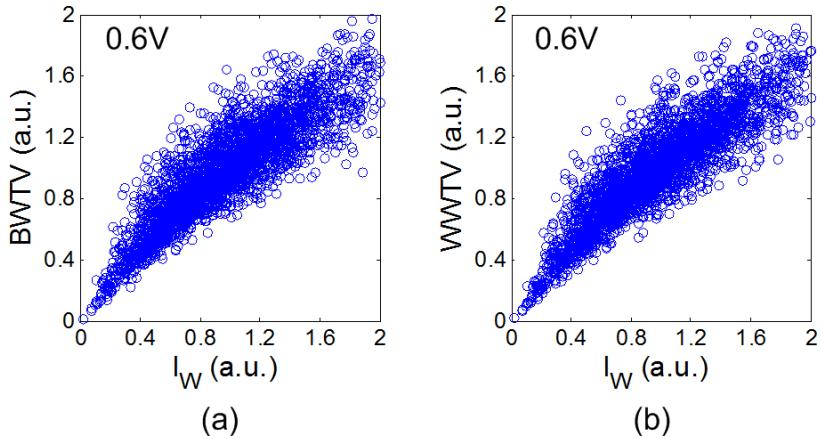


Figure 2.36: Scatter plots for (a) BWTV versus I_W and (b) WWTV versus I_W at $V_{DD} = 0.6V$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process.

WL , BL (or BLC), and V_{CELL} biased at V_{DD} and BLC (or BL) biased at V_{SS} . $WWTV > 0$ indicates that a successful write operation can take place even when $V_{WL} < V_{DD}$ - i.e. when the V_{GS} for the pass-gate transistor (at the '1' storage side) is decreased, compared to a nominal write operation. Therefore, the WWTV effectively measures the maximum tolerable reduction in the cell α -ratio - through a reduction of the pass-gate transistor V_{GS} , while maintaining the operation condition of the pull-up transistor - to still successfully write the SRAM cell. The most notable advantage of the WWTV measurement is that, unlike during the BWTV characterization, the SRAM CUT is not put under a read stress at the onset of the measurement (with $V_{WL} = 0V$). Therefore, the WWTV can continue to be characterized for SRAM cells under aggressively scaled supply voltages (V_{DD}).

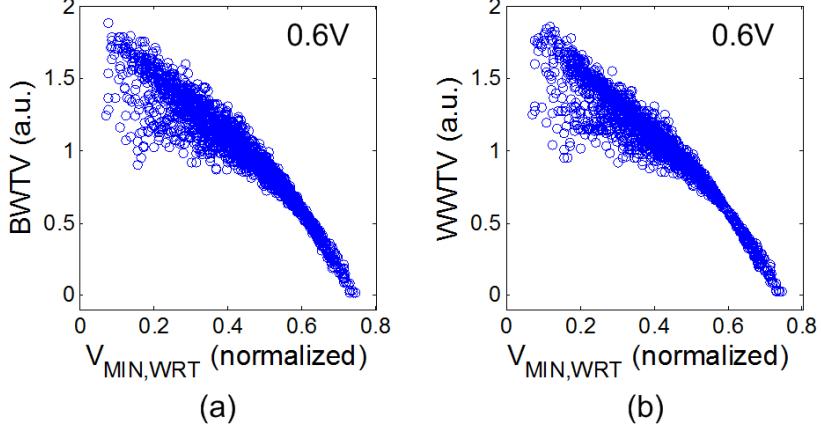


Figure 2.37: Scatter plots for (a) BWTW versus $V_{MIN,WRT}$ and (b) WWTW versus $V_{MIN,WRT}$ at $V_{DD} = 0.6V$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process.

Large-Scale Writeability Metrics versus Conventional WNM and I_W

With the large-scale writeability metrics defined, it is of interest to see how these metrics correlate with the conventional WNM and I_W as well as the per-cell $V_{MIN,WRT}$. Figure 2.35 shows the scatter plots for BWTW versus WWTW (in a), BWTW versus WNM (in b), and WWTW versus WNM (in c) at $V_{DD} = 0.6V$ obtained from 3k-sample MC simulations using a commercial low-power 45nm CMOS process. Results indicate excellent correlations between the BWTW, the WWTW, and the conventional WNM. Figure 2.36 shows the scatter plots for BWTW versus I_W and WWTW versus I_W obtained from 3k-sample MC simulations. Results show some dispersion between the BWTW/WWTW and the I_W values at high writeability due to similar reasons that limit the correlation between I_W and WNM. However, excellent correlations between BWTW/WWTW and I_W are established near writeability failure. In addition, Figure 2.35 and Figure 2.36 show that all four metrics share a common point of failure (i.e. the origin). Figure 2.37 shows the scatter plots for BWTW and WWTW versus the per-cell $V_{MIN,WRT}$. Excellent correlations are established between BWTW/WWTW and $V_{MIN,WRT}$. This indicates that both BWTW and WWTW can effectively track the SRAM $V_{MIN,WRT}$ and are therefore well suited for $V_{MIN,WRT}$ estimation.

2.3.4 Per-cell V_{MIN} Extraction using Direct Bit-line Access

In addition to read stability and writeability characterization, the direct bit-line access scheme can be adopted to characterize the minimum DC operating voltage of each SRAM bitcell during standby, read, and write cycles. Figure 2.38a shows the flow-chart diagram for characterizing the SRAM V_{MIN} during a static read operation. Each iteration of this measurement starts with a data initialization under a high supply voltage (V_{DD}) to guarantee a successful write operation - a boosted WL voltage can be used in conjunction for extra assurance. The SRAM cell is then configured for a low voltage read operation with V_{CELL} , V_{WL} , and both bit-line voltages all biased at a lower V_{DD} , which is gradually reduced for each iteration of the measurement process. Finally, V_{DD} is raised for a high voltage read

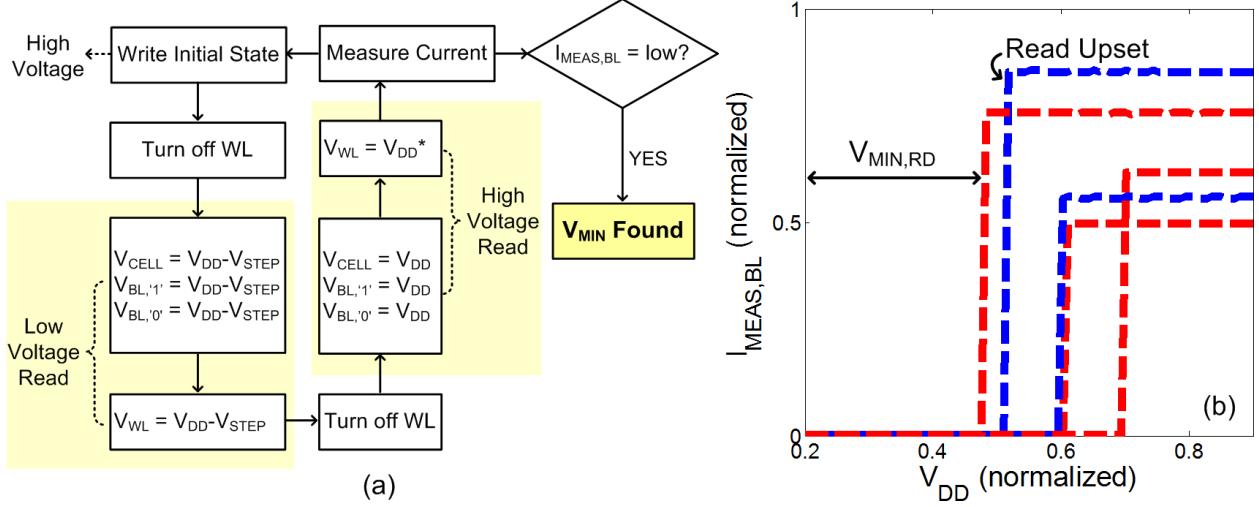


Figure 2.38: (a) Flow chart for $V_{MIN,RD}$ characterization and (b) bit-line currents at different V_{DD} points for $V_{MIN,RD}$ extraction.

operation and the BL current at the '0'-initialized storage node ($I_{MEAS,BL}$) is measured. The measured current should be high - equal to I_{READ} - if no data disturbance took place during the low voltage read operation. Therefore, V_{MIN} can be characterized as the maximum supply voltage (V_{DD}) before $I_{MEAS,BL}$ drops (Figure 2.38b). The per-cell SRAM V_{MIN} during the standby mode can be characterized using the same procedure as in Figure 2.38a by keeping V_{WL} low, at V_{SS} , during each low voltage read operation - i.e. to emulate a low voltage hold operation. To eliminate an accidental data disturbance, the WL is turned off between low voltage and high voltage operations. Additionally, V_{WL} can be reduced, below V_{DD} (denoted as V_{DD}^* in Figure 2.38a), during the high voltage read operation to further eliminate the chance for accidental data disturbance during the high voltage read¹¹.

Figure 2.39a shows the flow chart diagram for measuring SRAM V_{MIN} during a static write operation. The procedure is very similar to that for the read V_{MIN} characterization - but instead of a low voltage read operation, each iteration performs a low voltage write operation where V_{CELL} , V_{WL} , and the bit-line voltage at the '0'-initialized storage node are biased at a lower V_{DD} and the bit-line voltage at the '1'-initialized storage node is biased at V_{SS} . Each low voltage write is immediately followed by a high voltage read where the bit-line current at the '0'-initialized storage node ($I_{MEAS,BL}$) is measured. The measured current should be low if data is successfully written during the low voltage write. V_{MIN} can be characterized as the maximum operation voltage while $I_{MEAS,BL}$ remains low (Figure 2.39b). Note that V_{MIN} characterization using direct bit-line measurements is slower than

¹¹If a lower V_{WL} is used, the measured current will be lower than I_{READ} , but still high enough to differentiate between a binary '1' or '0'.

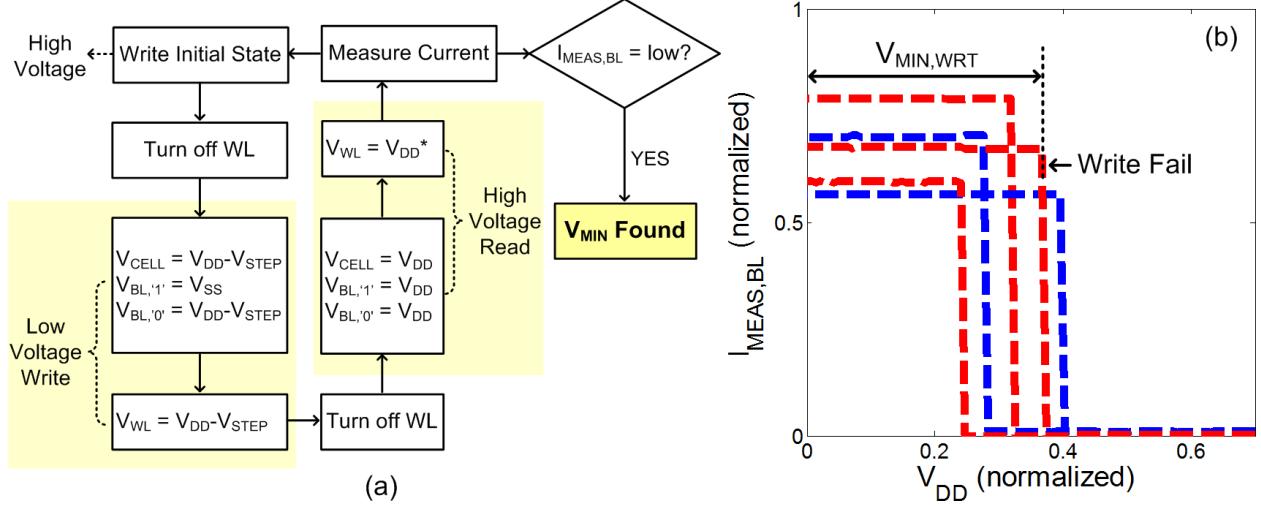


Figure 2.39: (a) Flow chart for $V_{MIN,WRT}$ characterization and (b) bit-line currents at different V_{DD} points for $V_{MIN,WRT}$ extraction.

the typical on-chip digital SRAM tester¹², using similar read-after-read and read-after-write sequences as described above, because of the need to monitor the bit-line current. However, since the direct bit-line V_{MIN} characterization can be performed alongside the large-scale read stability and writeability measurements with no additional hardware overhead, it is used, in this work, to establish correlations between bitcell failure and the bitcell read and write characteristics.

¹²Note that the V_{MIN} characterization presented in this section are performed statically. Typical on-chip digital SRAM testers (such as the SRAM built-in self test or BIST) often perform dynamic V_{MIN} characterizations as well - using similar read-after-read and read-after-write sequences as described above, under a frequency constraint. To perform such a characterization, precise timing controls of the bit-lines and the word-line are needed; in this case, a read failure is characterized as a read upset during the word-line pulse with the bit-lines discharging accordingly (i.e. not clamped at V_{DD}) and a write failure corresponds to the inability to successfully update the cell state during the word-line pulse. In addition, a read access failure can be characterized as an inability to discharge the bit-lines to specified levels during the word-line pulse of a read test cycle. Since static stability metrics are studied here, only static V_{MIN} characterizations are performed.

Chapter 3

Variability Characterization Test Chip

3.1 Introduction

After the detailed examination of the conventional SRAM design metrics and the introduction of the large-scale SRAM design metrics, this chapter focuses on the implementation of both the conventional and the large-scale SRAM variability characterization on two commercial low-power 45nm CMOS test chips. Section 3.2 details the implementation of the all-internal-node access scheme for characterizing the conventional SRAM design metrics and Section 3.3 details the implementation of the direct bit-line access scheme for characterizing the large-scale SRAM design metrics. Section 3.4 presents the overview for both 45nm CMOS test chips.

3.2 Implementation for Characterizing the Conventional SRAM Design Metrics

3.2.1 Top Level SRAM Macro Construction

The all-internal-node access scheme is implemented on small SRAM macros to characterize the conventional SRAM design metrics - HSNM/RSNM, WNM, and I_W [33, 64, 65]; this is similar to [20, 21]. In this design, all 10 internal nodes - V_{CELL} ¹, $V_{SS,CELL}$, V_{WL} , V_{BL} , V_{BLC} , V_{CL} , and V_{CH} - of each SRAM cell under test (CUT) are wired out through a hierarchy of switches to allow VTC and N-curve measurements as well as I-V characterization of all 6 transistors in the SRAM bitcell. The area overhead of this design is higher than [20, 21] as a result of accessing more internal nodes. N-well and P-well biasing - V_{NW} and V_{PW} - in each SRAM macro is shared with the functional SRAM arrays (Section 3.3) to investigate the effect of body biasing on the SRAM read stability and writeability as well as on the transistor V_{TH} . Figure 3.1 presents the high level circuit diagram of the all-internal-node access characterization scheme for the SRAM macros. Each SRAM macro consists of a 20-row

¹ V_{CELL} , $V_{SS,CELL}$, and V_{WL} are separately accessed for the two halves of an SRAM bitcell - accounting for 6 accessed internal nodes.

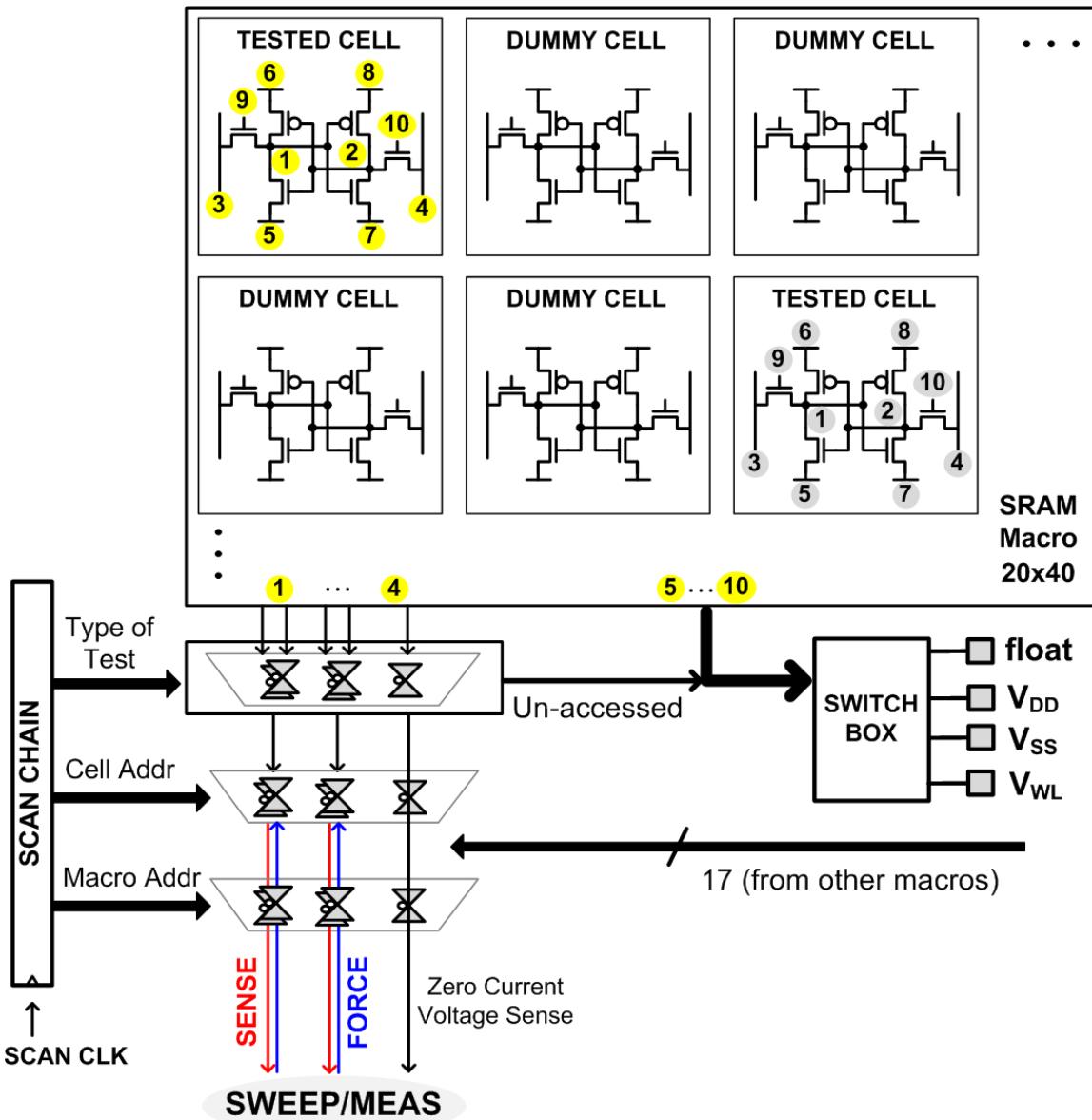


Figure 3.1: Circuit diagram of the all-internal-node access characterization scheme for the SRAM macros.

by 40-column array², with one bitcell accessed per column and per row. To provide enough metal spacing for routing out the 10 internal nodes of each SRAM CUT, every other column in the array is skipped³, yielding 20 SRAM CUTs per macro.

The switch network for the all-internal-node access scheme is implemented using wide, long-channel, thick-oxide CMOS transmission gates driven by a higher separate supply voltage to suppress the leakage from the un-accessed cell nodes. Figure 3.2a shows the schematic of the thick-oxide CMOS transmission gate used in the switch network. The transmission gate is sized to minimize its V_{DS} drop for a given level of current drive - corresponding to the current sourced into or out from each cell node, while maintaining compactness. The NMOS transistor is sized more than $5\times$ larger than the PMOS transistor to exploit the higher current drive strength of the NMOS transistor. Since a higher (1.8V) supply voltage drives the gate of the NMOS transistor, a full output voltage range - from 0V to 1.1V - can be driven by the NMOS alone. The smaller PMOS transistor is inserted to assist the larger NMOS transistor when driving the higher voltages within that range. In addition, the gate of the NMOS transistor can be overdriven at above 1.8V to further reduce the V_{DS} . Figure 3.2b-c plots the simulated V_{OUT} versus I_D transfer curves for the transmission gate switch, with a 200mV gate overdrive (i.e. the gate voltage is at 2.0V), when driving input voltages of 1.1V and 0V. Results show that the transmission gate can drive up to $100\mu A$ of drain current, which is higher than the expected current at each cell node, while keeping its V_{DS} below 12mV (when driving an input voltage of 1.1V) and below 5mV (when driving an input voltage of 0V). A separate simulation indicates that less than a 4mV reduction in the transistor V_{DS} is achieved - for driving an input voltage of 1.1V - when the PMOS transistor width is doubled.

Although transistor sizing and gate overdrive are used to minimize the V_{DS} drop of each transmission gate within the switching hierarchy, they cannot completely mitigate its effects. To mitigate the effects of V_{DS} drop within the switching hierarchy, the 4-terminal Kelvin sensing method is adopted to make use of independent force (current) and sense (voltage) paths to access all critical cell nodes. The critical cell nodes can be identified as all cell nodes requiring a precise voltage bias while sourcing or sinking a non-zero current. For VTC and N-curve measurements, the 4-terminal Kelvin sensing method is applied at V_{CELL} , $V_{SS,CELL}$, both bit-lines and one of the storage nodes (CL or CH) - the other storage node is either left floating (for N-curve measurements) or wired out for zero-current voltage sensing (for VTC measurements) and does not require the 4-terminal Kelvin connection. For individual transistor I-V characterization, the 4-terminal Kelvin sensing method is applied at the source and drain terminals. In addition to characterizing the conventional SRAM design metrics and the individual transistor I-V behavior, each SRAM macro is also capable of measuring the large-scale SRAM design metrics presented in Section 2.3. This is important as it enables the direct correlation between the measured conventional SRAM design metrics and the measured large-scale SRAM design metrics. For measuring large-scale SRAM design metrics, the 4-terminal Kelvin sensing method is applied at V_{CELL} , $V_{SS,CELL}$, and both bit-lines. All un-accessed cell nodes from each SRAM cell are selectively left floating or biased

²SRAM macros for $0.299 \mu m^2$ and $0.252 \mu m^2$ bitcells consist of 20-rows by 60-columns in each array (Section 3.2.2).

³SRAM macros for $0.299 \mu m^2$ and $0.252 \mu m^2$ bitcells require skipping 2 columns (Section 3.2.2).

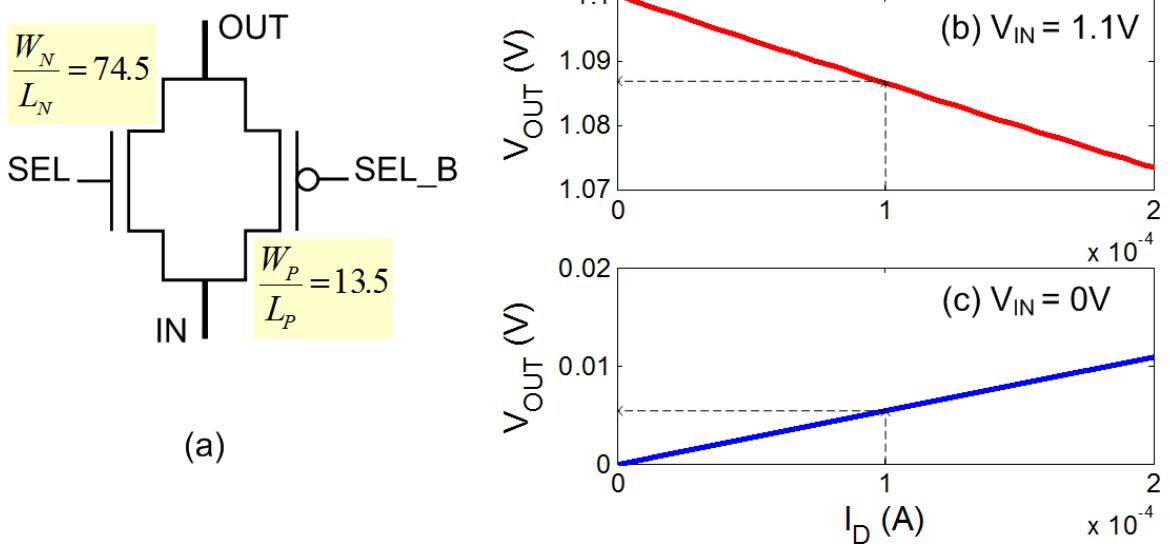


Figure 3.2: (a) Schematic of the thick-oxide CMOS transmission gate used in the switch network for both the all-internal-node access scheme and the direct bit-line access scheme. Simulated V_{OUT} versus I_D of the switch for (b) $V_{IN} = 1.1V$ and (c) $V_{IN} = 0V$.

at V_{DD} or V_{SS} .

Each internal node is accessed through 3 levels of switching hierarchy. The gate leakage in the switch network is limited by the usage of thick-oxide transmission gates. The subthreshold leakage is minimized by the stack effect in the hierarchy - i.e. 3 levels of switching hierarchy translates to 3 transmission gates connected in series. The only significant source of leakage in the switch network comes from the drain to body leakage, which sets the lower limit of measurable current at a few to a few tens of nA and does not affect the read/write margin and the transistor I-V measurements⁴. Finally, all static control signals are supplied by the scan chain to minimize the I/O pin count.

Figure 3.3 shows the layout view of an SRAM macro constructed for a 20-row by 40-column array. Only a small fraction of the total layout area accounts for the actual SRAM array, while a much larger fraction of the total layout area is attributed to the switch network and some digital logic - this again highlights the major drawback of characterizing SRAM variability on silicon using the conventional SRAM design metrics.

3.2.2 SRAM Cell Layouts for All-Internal-Node Access

Accessing the internal nodes of an SRAM cell requires the removal of the upper metal layers for node access and the insertion of metal wires for routing. Three different SRAM bitcells are implemented in this low-power 45nm CMOS process - yielding SRAM cell sizes of $0.374 \mu\text{m}^2$, $0.299 \mu\text{m}^2$, and $0.252 \mu\text{m}^2$. Figure 3.4 shows the layout cartoon for a $0.374 \mu\text{m}^2$ bitcell with all 10 internal nodes wired out - the 10 internal nodes are labeled in accordance with the schematic in Figure 3.1. Due to the extra spacing required beyond the left- and

⁴The lower limit of measurable current does set the lower limit for the transistor leakage current measurements.

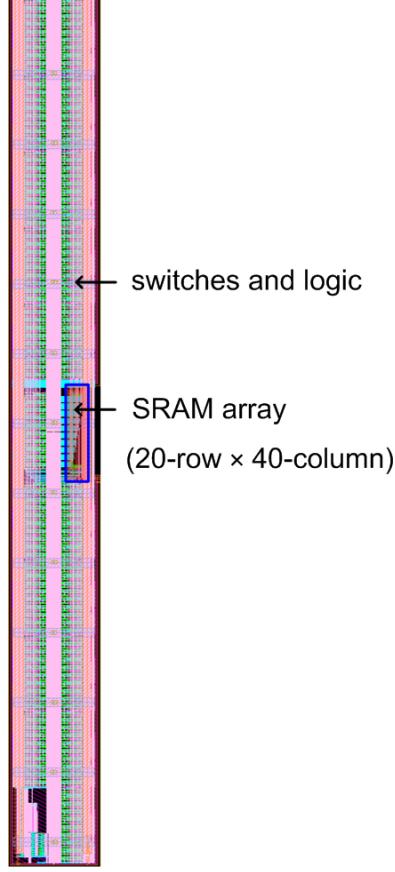


Figure 3.3: Layout view of an SRAM macro constructed for a 20-row by 40-column array.

right-boundary of the SRAM cell, only every other column in each array can be accessed for characterization. Figure 3.5 shows the layout cartoon for a $0.299 \mu\text{m}^2$ bitcell with all 10 internal nodes wired out. Due to a reduction in the SRAM cell size, the spacing required beyond the left- and right-boundary of the SRAM cell is higher than for the $0.374 \mu\text{m}^2$ bitcells. As a result, only 1 out of every 3 columns can be accessed for characterization and a 20-row by 60-column SRAM array is required for a total of 20 SRAM CUTs. For the $0.252 \mu\text{m}^2$ bitcells, the spacing becomes so stringent such that the direct access to all 10 internal nodes is not possible. A careful study, however, reveals that if the poly-gate of the NMOS pull-down and the PMOS pull-up transistors in the SRAM CUT is extended to contact the corresponding poly-gate in a neighboring bitcell, the neighboring bitcell storage nodes will be shorted to the storage nodes of the SRAM CUT and can be accessed instead⁵. Figure 3.6 shows the layout cartoon for a $0.252 \mu\text{m}^2$ bitcell with all 10 internal nodes wired out; note the poly-gate extension at both the left- and right- halves of the SRAM cell.

⁵Care should be taken to short out the terminals of the NMOS pull-down transistor and the PMOS pull-up transistor of the neighboring bitcell so that they do not contribute any current to the N-curve characterization.

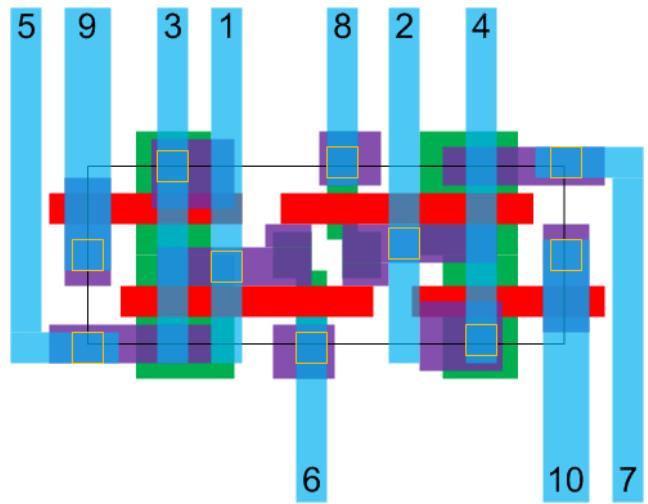


Figure 3.4: Layout cartoon for a $0.374 \mu\text{m}^2$ bitcell with all 10 internal nodes wired out.

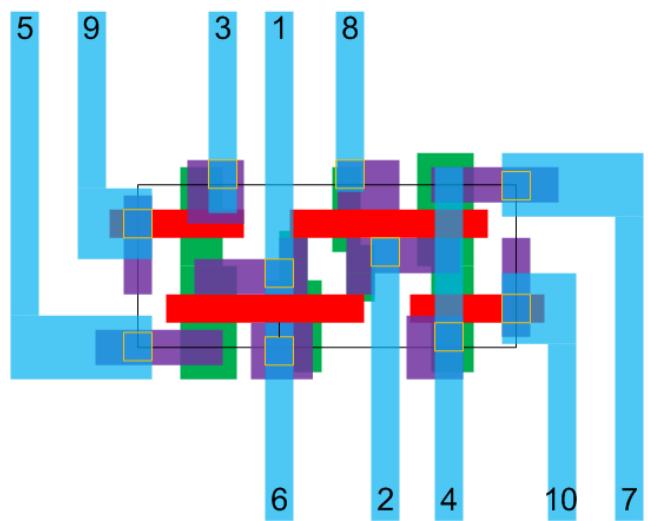


Figure 3.5: Layout cartoon for a $0.299 \mu\text{m}^2$ bitcell with all 10 internal nodes wired out.

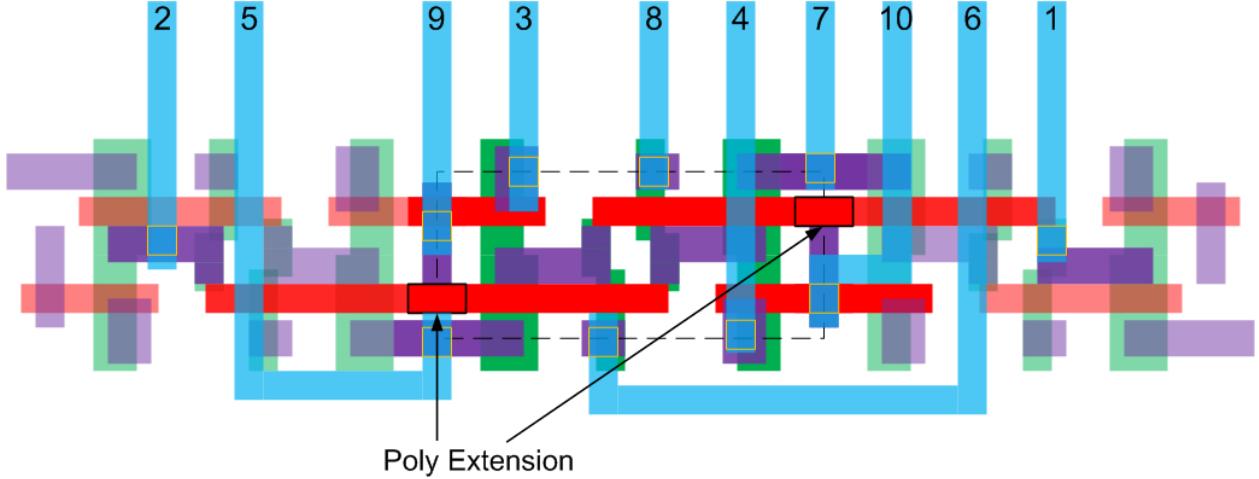


Figure 3.6: Layout cartoon for a $0.252 \mu\text{m}^2$ bitcell with all 10 internal nodes wired out. The SRAM CUT is outlined by the dotted line.

3.3 Implementation for Characterizing the Large-Scale SRAM Design Metrics

3.3.1 Top Level Overview

The direct bit-line access scheme is implemented for functional SRAM arrays to characterize the large-scale read stability and writeability metrics [64, 65]. Figure 3.7 presents the high level circuit diagram of the direct bit-line characterization scheme. The lower right portion of the circuit diagram shows a typical functional SRAM array with the row decoder and the column read/write circuitry. A level shifter with a V_{SS} slightly below 0V, denoted as $V_{SS,NEG}$ in Figure 3.7, is inserted within the later stages of the row decoder to allow a sufficient range of word-line voltages from 0V to +400mV above $V_{DD,NOMINAL}$ (equal to 1.1V in this process). The value of the analog word-line voltage is set by $V_{DD,WL}$ (See Figure 3.13 in Section 3.3.4). The SRAM array is implemented with independent cell supply (V_{CELL}), cell V_{SS} ($V_{SS,CELL}$), N-well bias (V_{NW}), and P-well bias (V_{PW}). All 4 terminals can be used either for voltage sweeping or for setting bias conditions. During the direct bit-line measurements, the column read/write circuitry can be shut off by a low R/W enable signal. The bit-lines are accessed through a hierarchy of bit-line switches implemented using wide, long-channel, thick-oxide CMOS transmission gates - identical to the switches used in the all-internal-node access scheme described in Figure 3.2. As mentioned in Section 3.2.1, the gate leakage of the switches are limited by the usage of thick-oxide transistors and the sub-threshold leakage in the switch network is minimized by the stack effect. The only significant source of leakage in the switch network comes from the drain to body leakage, which sets the lower limit of measurable current at a few to a few tens of nA and does not affect the large-scale read/write margin measurements.

Each bit-line is accessed through 4 levels of hierarchy with a maximum of 16 switches sharing the same node. In order to accurately set voltages at the bit-lines, the V_{DS} drop, whose effect is similar to the IR-drop associated with a series resistance, in the switch

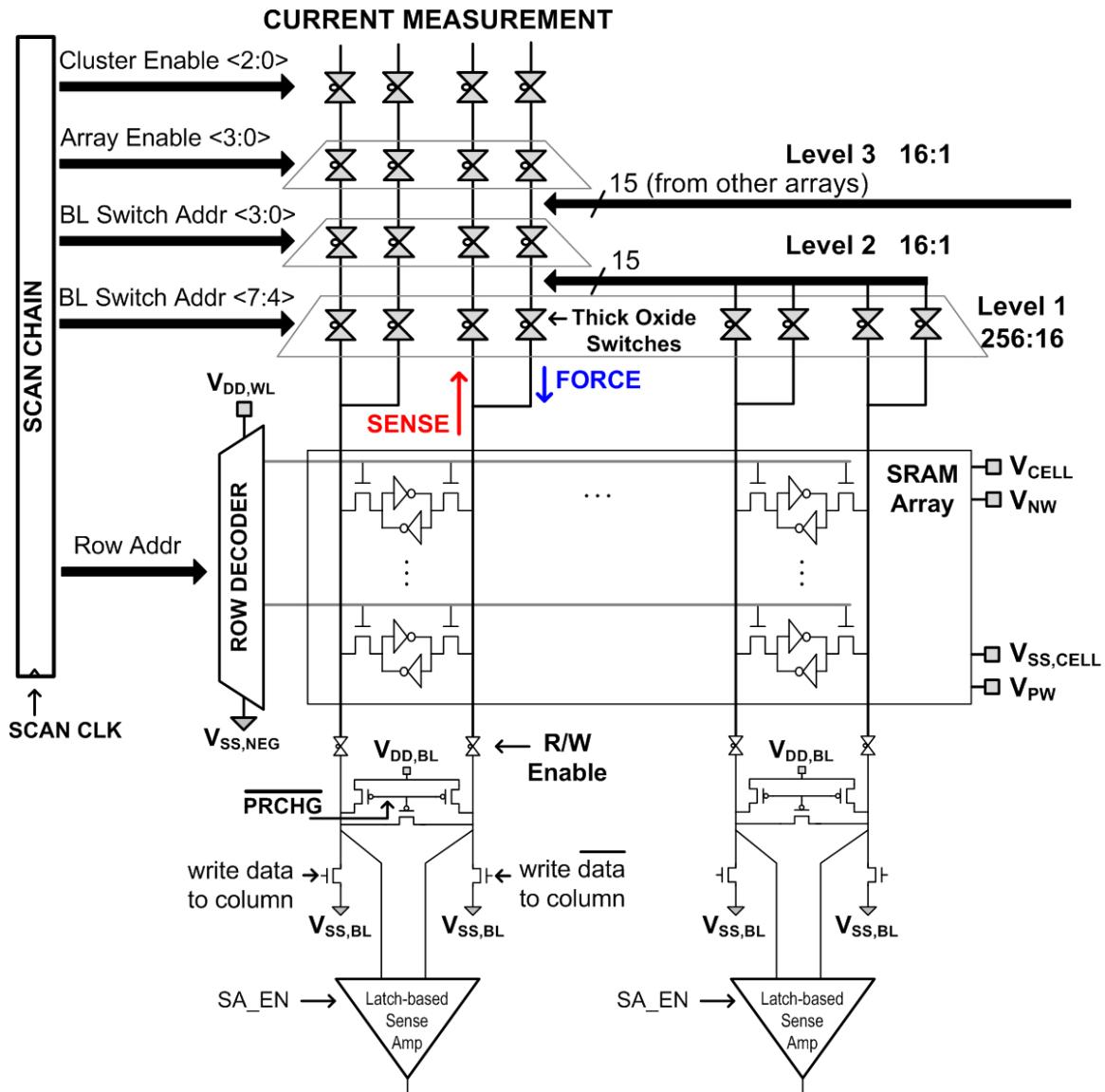


Figure 3.7: Circuit diagram of the direct bit-line characterization scheme for the functional SRAM arrays.

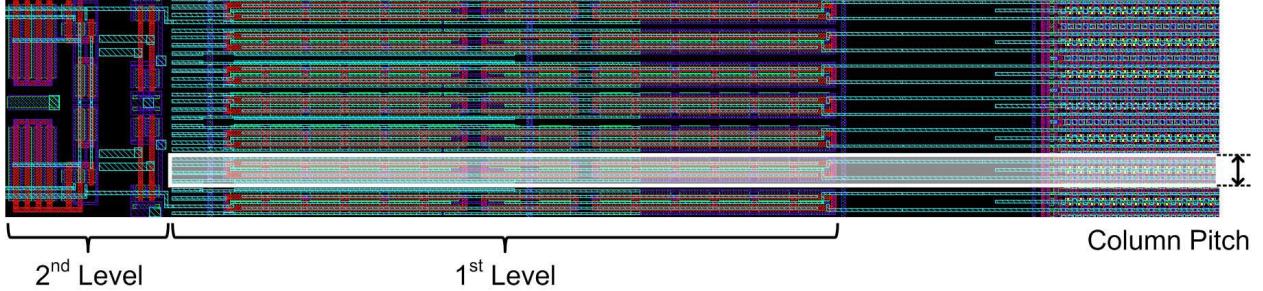


Figure 3.8: Layout view (up to the M2 layer) showing the construction of the first level of the bit-line switch hierarchy within the column pitch of an SRAM sub-array using the $0.374 \mu\text{m}^2$ bitcell design.

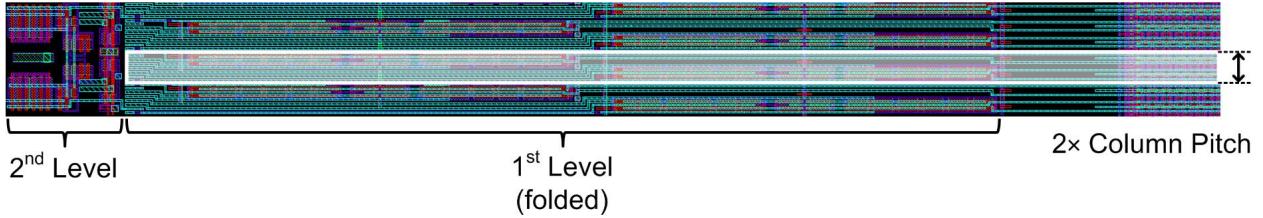


Figure 3.9: Layout view (up to the M2 layer) showing the construction of the first level of the bit-line switch hierarchy folded to fit within $2\times$ the column pitch of an SRAM sub-array using the $0.299 \mu\text{m}^2$ bitcell design.

hierarchy must be eliminated. Overdriving the gates of the thick oxide transistors can help to decrease the V_{DS} drop in each switch, as illustrated in Figure 3.2, but cannot completely mitigate its effects. Thus, the 4-terminal Kelvin sensing method is adopted to make use of independent force (current) and sense (voltage) paths to access each bit-line. This effectively eliminates the V_{DS} drop (i.e. the IR-drop) in the switch hierarchy. All static control signals are supplied by the scan chain to minimize the I/O pin count.

3.3.2 SRAM Arrays

Each 256kb functional SRAM array is partitioned into four 256-row by 256-column sub-arrays⁶. Each 256-row by 256-column sub-array is further split into two 128-row by 256-column mini-arrays to make space for a row of well contacts, as required by the design rules in this process. 2 levels of bit-line switches are required to access all 256 columns of each sub-array, as indicated in Figure 3.7. The 2-level bit-line switch array is inserted adjacent to each SRAM sub-array and perpendicular to the columns on the side that is not occupied by the column circuitry. Since all 256 columns are accessed through the direct bit-line connection, the first level bit-line switch for each bit-line pair - consisting of both force and sense paths for BL and BLC - must fit within the column pitch of the SRAM sub-array, which is set by the SRAM cell width. The layout view in Figure 3.8 shows the

⁶The first test chip (see Section 3.4) also included 128kb functional SRAM arrays constructed from four 128-row by 256-column sub-arrays.

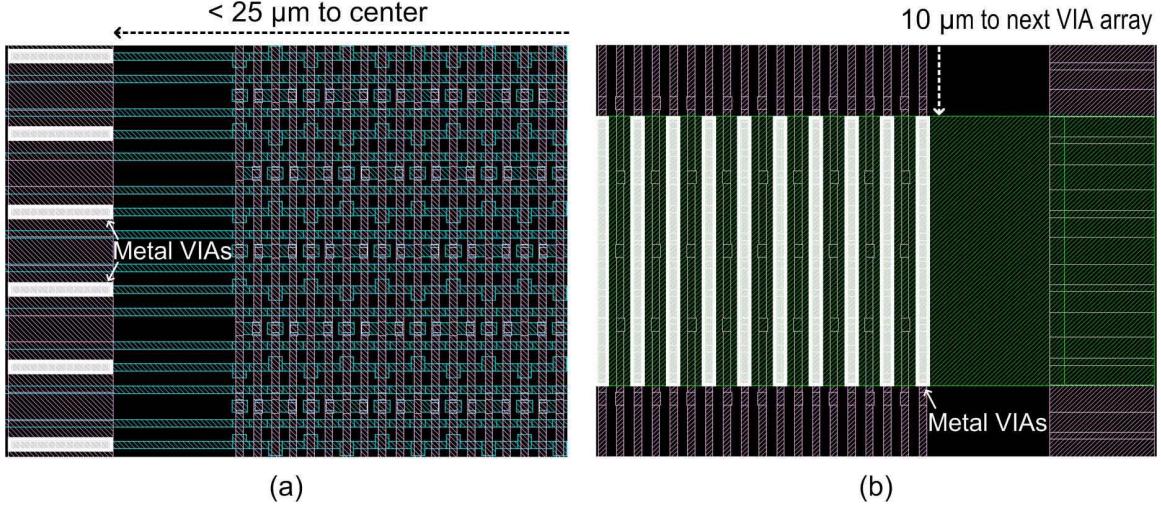


Figure 3.10: (a) Layout view showing the M2-M3 connection of V_{CELL} outside the 128×256 mini-array. (b) Layout view showing the M3-M4 connection of $V_{SS,CELL}$ inside the 128×256 mini-array.

construction of the first level of the bit-line switch hierarchy within the column pitch of an SRAM sub-array using the $0.374 \mu\text{m}^2$ bitcell design. Due to a significant width reduction in the $0.299 \mu\text{m}^2$ and the $0.252 \mu\text{m}^2$ bitcell designs, the first level bit-line switch for each bit-line pair cannot fit within the single bitcell column pitch using the general logic design rules in this process. Therefore, the first level bit-line switches from neighboring bit-line pairs are folded to relax the column pitch requirement by a factor of 2. The layout view in Figure 3.9 shows the construction of the first level of the bit-line switch hierarchy folded to fit within $2 \times$ the column pitch of an SRAM sub-array using the $0.299 \mu\text{m}^2$ bitcell design - this layout view is exemplary of the SRAM sub-arrays consisting of the $0.252 \mu\text{m}^2$ bitcells.

In order to minimize the IR-drop in the power supplies, both V_{CELL} and $V_{SS,CELL}$ must be densely gridded. In this design, V_{CELL} from each SRAM bitcell is routed in M2 along with the bit-lines, while $V_{SS,CELL}$ is routed in M3 along with the word-line. As a result, V_{CELL} can only be connected up from M2 outside each 128×256 mini-array. The layout view in Figure 3.10a shows the connection of V_{CELL} from M2 to M3 outside the 128×256 mini-array. Each 128×256 mini-array is less than 50 μm in the horizontal direction (parallel to the columns), resulting in a worst scenario IR-drop for the two center rows - at a distance of less than 25 μm away from the M2-M3 VIA array. The layout view in Figure 3.10b shows the connection of $V_{SS,CELL}$ from M3 to M4 inside the 128×256 mini-array. The M4-to-M7 grid for both V_{CELL} and $V_{SS,CELL}$ consist of metal layers 3 μm in width with a 10 μm spacing⁷. In addition to the supply IR-drop, the IR-drop along each bit-line must also be considered. While each 256-cell column, for the $0.374 \mu\text{m}^2$ bitcells, is about 100 μm tall (routed in M2), approximately 30Ω bit-line resistance is needed for every 1mV IR-drop (at $V_{DD}=1.1\text{V}$), given the expected levels of the bit-line current during characterization. This IR-drop is expected to significantly drop as V_{DD} is decreased (where most of the measurements are conducted

⁷This spacing is required to setup the grid structure for routing four supply voltages - V_{CELL} , $V_{SS,CELL}$, V_{NW} , and V_{PW} .

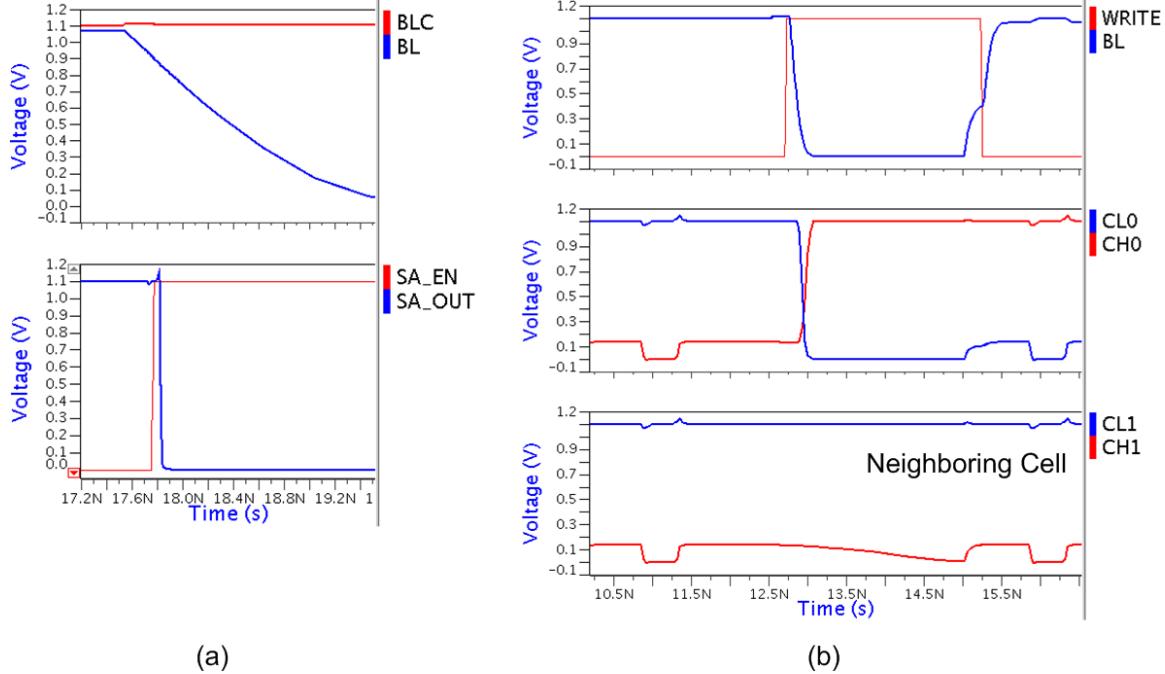


Figure 3.11: Simulated waveforms during (a) the read cycle and (b) the write cycle showing correct functionality.

- see Chapter 4), due to a decrease in the bit-line current (e.g. at $V_{DD} = 0.7V$, $\sim 150\Omega$ of bit-line resistance is needed for every $1mV$ IR-drop).

3.3.3 SRAM Array Functionality

In order to verify the functionality of the SRAM array, simulations for the read and write cycles are performed. Figure 3.11a plots the simulated waveforms during a read cycle for both bit-lines (*BL* and *BLC*), the sense amplifier enable signal (*SA_EN*), and the sense amplifier output signal (*SA_OUT*). The discharging *BL* is correctly captured as *SA_OUT* is driven low after asserting the *SA_EN* signal. Figure 3.11a plots the simulated waveform during a write cycle. The waveforms show that the *WRITE* signal correctly pulls *BL* to V_{SS} corresponding to the data input and the accessed storage nodes (*CL0* and *CH0*), corresponding to the bitcell with an asserted *WL*, correctly toggles while the neighboring bitcell storage nodes (*CL1* and *CH1* from a different row) remain unaffected. Since the goal of this design is to characterize the DC large-scale read stability and writeability margins, only correct functionality is verified while the access time performance of the SRAM array is not evaluated and generous cycle times are used during measurements.

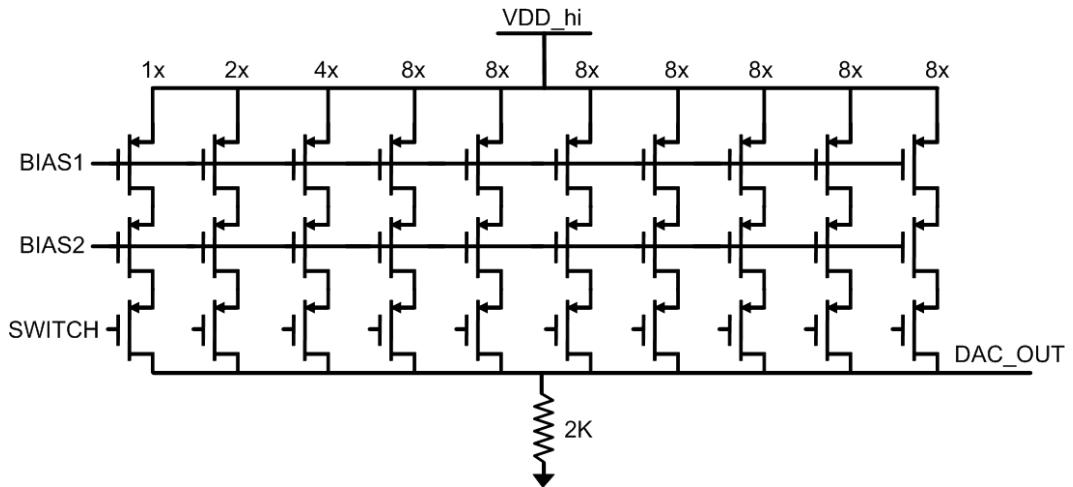


Figure 3.12: Circuit diagram of a 3LSB-3MSB segmented DAC implemented to perform on-chip word-line sweep.

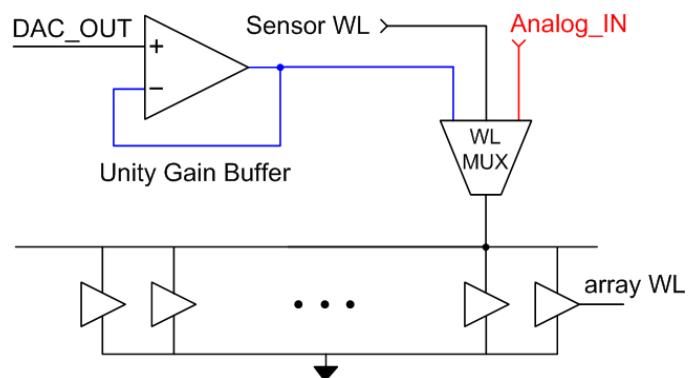


Figure 3.13: Circuit diagram showing how the output of the 6-bit DAC is multiplexed to drive the word-line.

3.3.4 On-Chip 6-bit Digital-to-Analog Conversion

An optional 6-bit digital-to-analog converter (DAC) is implemented to perform on-chip word-line sweep⁸. The circuit topology for a 3LSB-3MSB segmented DAC is adapted from [50,51] and uses cascode PMOS current sources to allow an output voltage range down to 0V. Figure 3.12 shows the circuit diagram for the 3LSB-3MSB segmented DAC. The top two PMOS transistors in each current source are self-biased by an on-chip reference circuit and the bottom PMOS transistor is used to switch ON and OFF the current source. The 6-bit DAC consists of 63 unit current sources implemented using a common centroid layout structure to compensate for the process gradients. The switching gate of each unit current source is connected in accordance with the binary multiplier shown in Figure 3.12.

To drive the word-line voltage, the output of the DAC needs to drive the supply voltage of the final word-line decode stage. Since the output of this DAC structure cannot drive a low impedance node, a unity gain buffer, implemented using a folded cascode operation amplifier, is used. The circuit diagram in Figure 3.13 shows the setup for driving the supply voltage of the final word-line decode stage⁹. An analog multiplexer is implemented, using large thick-oxide CMOS transmission gates, to drive the supply voltage of the final word-line decode stage to either an off-chip analog input voltage, the DAC output voltage (driven by the unity gain buffer), or the word-line voltage dynamically tuned by an on-chip variation sensor [32]. The V_{DS} drop in the switches of the analog multiplexer is not an issue as the supply voltage it is driving should not source any static current; therefore, the 4-terminal Kelvin connection is not required for the off-chip analog input voltage.

In order to accurately drive the DAC output voltage at the word-line, the folded cascode operation amplifier must have high gain. Due to the unity gain feedback configuration, which presents a worst case feedback for stability [60], the amplifier must also have enough phase margin. In addition, the phase margin should be engineered to allow for process variations and to avoid significant overshoot while settling, since this can translate to an overshoot in the word-line voltage and cause an accidental data flip during read/write margin characterization. The folded cascode operation amplifier design is adopted from [111,112]. Figure 3.14a plots the simulated frequency response waveforms for the folded cascode operation amplifier used in the unity gain buffer - a gain of 60dB is achieved at low frequencies and the phase margin is approximately 88°, which does not offer the fastest settling but guards against significant overshoot at the output of the unity gain buffer. Figure 3.14b illustrates the characterization of the WWTV by driving the word-line voltage with the 6-bit DAC output. Since a 6-bit DAC can only deliver 64 discrete voltages, the range of word-line sweep must be reduced to enhance the measurement resolution - in Figure 3.14b, the range of the word-line sweep is reduced to 400mV – 800mV, which achieves a maximum resolution of 6.25mV.

⁸However, due to equipment limitations (for using the GPIB [71] - general purpose interface bus - interface), the 6-bit DAC does not enhance measurement speed and therefore is not used during testing.

⁹A level shifter is inserted before the final word-line decode stage to allow word-line voltages above $V_{DD,NOMINAL}$, as mentioned in Section 3.3.1.

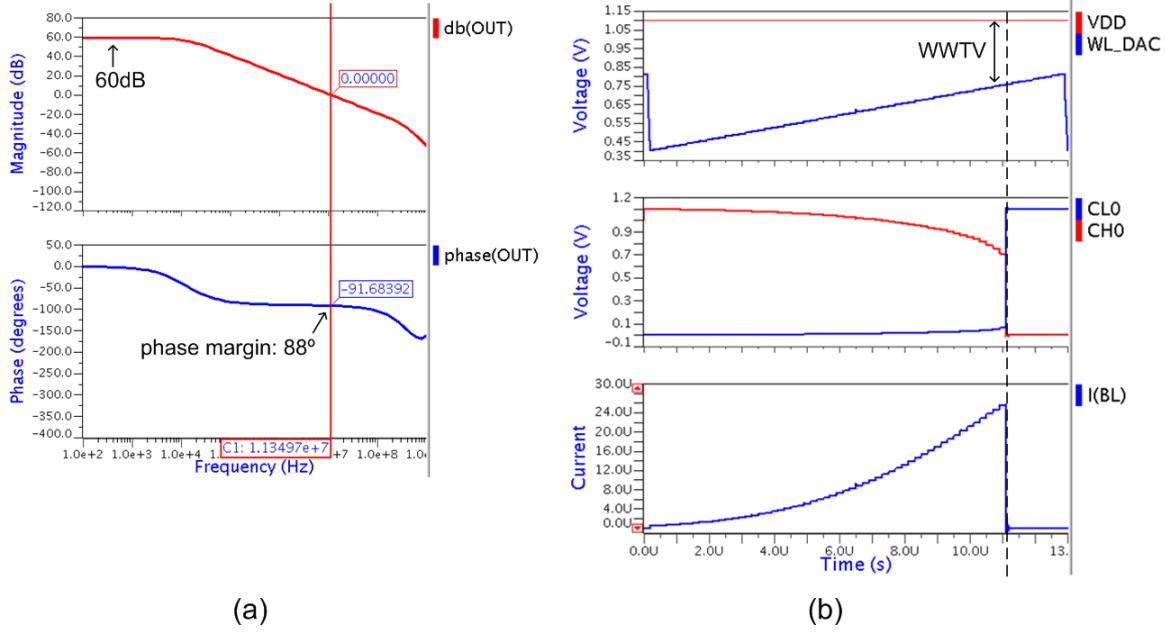


Figure 3.14: (a) Simulated gain and phase frequency response waveforms for the folded cascode operation amplifier used in the unity gain buffer. (b) Simulated transient waveforms showing the characterization of WWTV using the 6-bit DAC. Note: the *BL* current is scaled and does not reflect actual current values from the simulation.

3.3.5 Area Penalty of the Direct Bit-Line Characterization

The overall area overhead of the bit-line switch network in this prototype is approximately 20%¹⁰. This area overhead can be further reduced with an optimized layout of the bit-line switch network and/or by reducing the depth of the switch hierarchy. In addition, the array efficiency can be enhanced by using SRAM arrays with larger column heights. The proposed direct bit-line characterization scheme requires that the worst-case on-current of a single pass-gate transistor connected to a bit-line be higher than the sum of leakage currents of all pass-gate transistors connected to the complementary bit-line. This requirement is typically less stringent than the constraint set by the SRAM read access performance and therefore should not limit the column segmentation of the SRAM array. However, in the case where the SRAM read access constraint is relaxed and the bit-line leakage is high, direct bit-line measurements at lower operating voltages may be challenging as the bit-line on- and off-currents become harder to distinguish when detecting a data flip. This can be solved by returning to a higher operating voltage that ensures read stability for bit-line current measurements (i.e. a high voltage read operation) after stressing the SRAM cell with the appropriate sweeping voltage at a lower supply - similar to the V_{MIN} characterization loop described in Section 2.3.4. Due to a reasonably low overhead, the proposed direct bit-line characterization can either be implemented in an early SRAM development vehicle or, occasionally, on a working chip to monitor the process variability.

¹⁰The 6-bit DAC is not included in this estimation as it is not required for direct bit-line characterization.

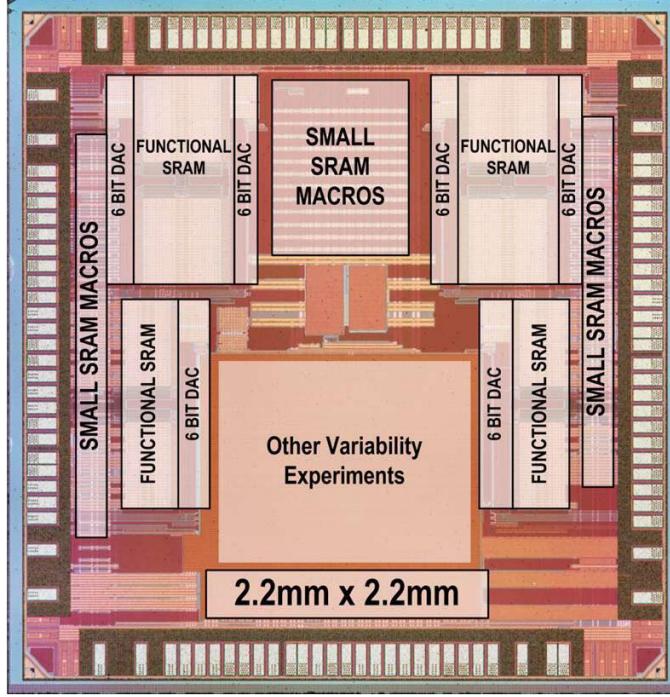


Figure 3.15: Die photo of the first low-power 45nm CMOS test chip.

3.4 45nm Low-Power CMOS Test-Chips

The die photo of a $2.2\text{ mm} \times 2.2\text{ mm}$ test chip [32, 64, 65] implemented in a low-power strained-Si 45nm CMOS process [29, 59, 76] with 7 metal layers is presented in Figure 3.15. The bitcell provided in this process for this particular test chip is a high-speed (i.e. high read current) SRAM cell, with a cell area of $0.374\text{ }\mu\text{m}^2$. The test chip consists of two 128kb functional SRAM arrays (each partitioned into four 128-row by 256-column sub-arrays) and two 256kb functional SRAM arrays (each partitioned into four 256-row by 256-column sub-arrays) all configured for large-scale read stability, writeability, and per-cell V_{MIN} characterization¹¹ - yielding 768kb of measurable SRAM cells per chip. It also includes eighteen 20×40 small SRAM macros configured with all-internal-node access for conventional SRAM VTC, N-curve, and individual transistor I-V measurements - 20 bitcells in each SRAM macro have all 10 internal nodes externally accessible through a switch network. In addition, a 6-bit 3LSB-3MSB segmented DAC is shared by every two 256×256 SRAM sub-arrays¹² for optional on-chip word-line sweep.

A second test chip (Figure 3.16) is implemented in an updated 45nm CMOS process offering 2 extra SRAM bitcell designs - achieving cell areas of $0.299\text{ }\mu\text{m}^2$ and $0.252\text{ }\mu\text{m}^2$. The chip area is $2.8\text{ mm} \times 2.5\text{ mm}$. This test chip consists of three 256kb functional SRAM arrays, corresponding to the three different SRAM bitcell designs - with cell areas of $0.374\text{ }\mu\text{m}^2$, $0.299\text{ }\mu\text{m}^2$, and $0.252\text{ }\mu\text{m}^2$. Each 256kb SRAM array is partitioned into four 256-row by 256-

¹¹The cell read current (I_{READ}) is also characterized with the direct bit-line access scheme.

¹²Each 128×256 sub-array is implemented with its own 6-bit DAC.

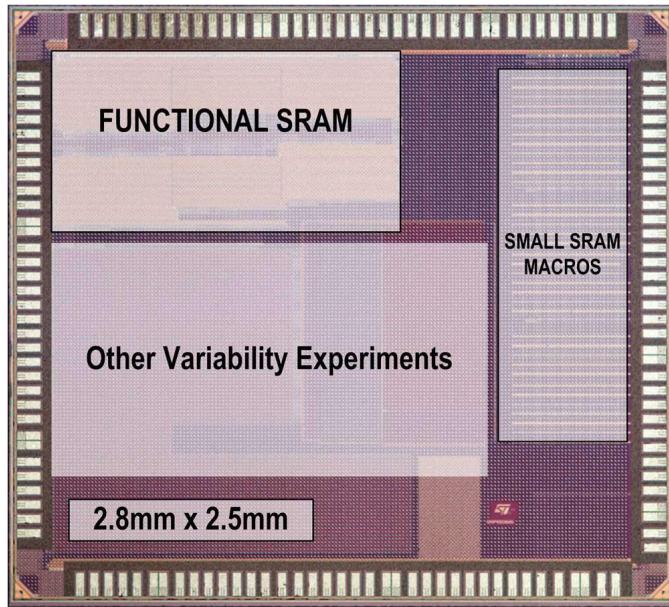


Figure 3.16: Die photo of the second low-power 45nm CMOS test chip. This test chip allows the characterization of 3 different SRAM bitcell designs, yielding cell areas of $0.374 \mu\text{m}^2$, $0.299 \mu\text{m}^2$, and $0.252 \mu\text{m}^2$.

column sub-arrays, all configured for large-scale characterization¹³. It also includes small SRAM macros configured for conventional SRAM VTC, N-curve, and individual transistor I-V measurements. Eighteen clusters of 20×40 SRAM arrays are implemented for each bitcell design (20×60 SRAM arrays are implemented for the $0.299 \mu\text{m}^2$ and the $0.252 \mu\text{m}^2$ bitcells, as indicated in Section 3.2.2), of which 20 bitcells per cluster are wired for characterization.

¹³On-chip DAC was not implemented for this test chip.

Chapter 4

Analysis of Measured Variability

4.1 Introduction

This chapter presents the measurement results from the low-power 45nm CMOS test chips introduced in Chapter 3. Section 4.2 presents and compares the measured results for both the conventional and the large-scale SRAM read stability and writeability metrics. The per-cell V_{MIN} measurements are presented in Section 4.3, where a direct correlation between the measured large-scale read/write margins and the per-cell V_{MIN} is established. In addition, a method to estimate the V_{MIN} of a functional SRAM array using the large-scale read/write margin measurements is described. Section 4.4 presents the read current (I_{READ}) measurements. Sources of systematic variations and their impacts on the SRAM cell stability are studied in Section 4.5. Finally, Section 4.6 analyzes the impact of several read and write assist circuits on the SRAM cell stability. Most of the measurement results presented in this chapter are extracted from the first 45nm CMOS test chip; therefore, unless otherwise noted, all measured results correspond to the 0.374 μm^2 bitcell design.

4.2 Read Stability and Writeability Measurements

4.2.1 Measurements and Correlations for SRAM Read Stability and Writeability

Conventional SRAM Design Metrics

The measured transfer curves extracted from SRAM macros with all-internal-node access for characterizing the conventional SRAM design metrics is presented in Figure 4.1. Scatter plots are generated to evaluate the correlations between the measured RSNM and the read stability metrics extracted from the measured N-curves (i.e. SVNM, SINM, and SPNM) and is presented in Figure 4.2. The measured results are in agreement with the simulated results presented in Section 2.2.1 - excellent correlation is established between the SVNM and the RSNM measurements, whereas the correlations between SINM and RSNM (and between SPNM and RSNM) suffer due to a bias difference between the SINM and the RSNM extraction (as discussed in Section 2.2.1). The distribution densities of the different

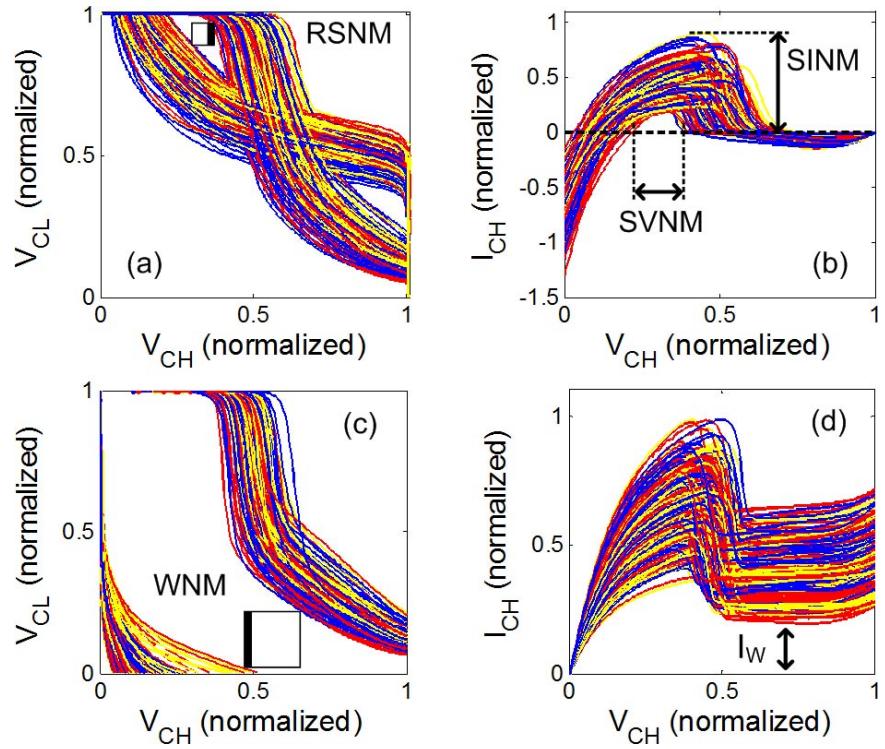


Figure 4.1: Measured (a) butterfly-curves for RSNM extraction, (b) N-curves for SVNM and SINM (as well as SPNM) extraction, (c) VTC pairs for WNM extraction, and (d) N-curves for I_W extraction from SRAM macros using all-internal-node access.

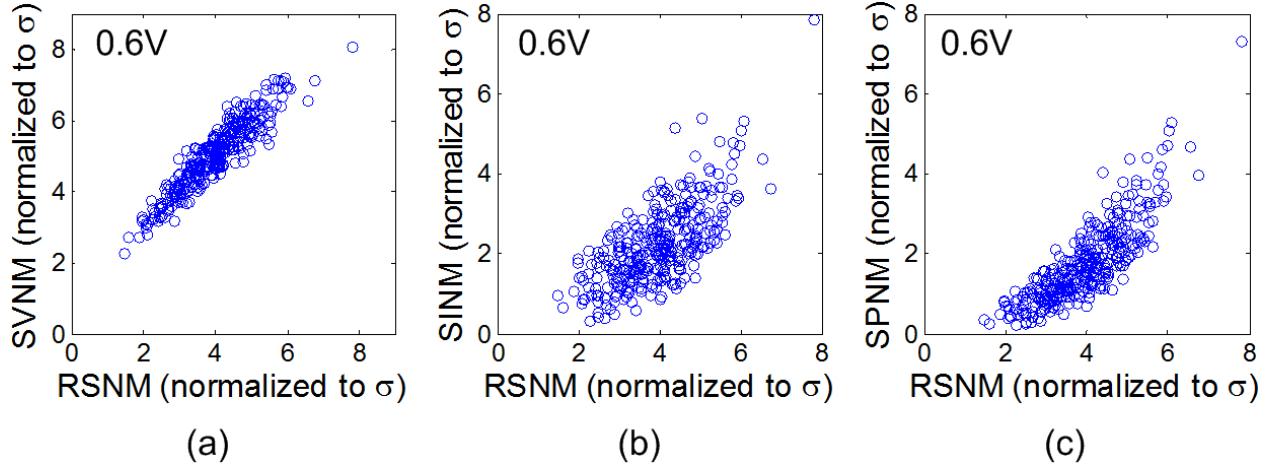


Figure 4.2: Scatter plots for (a) SVNM versus RSNM, (b) SINM versus RSNM, and (c) SPNM versus RSNM measured at $V_{DD} = 0.6V$ from the same SRAM macros using all-internal-node access.

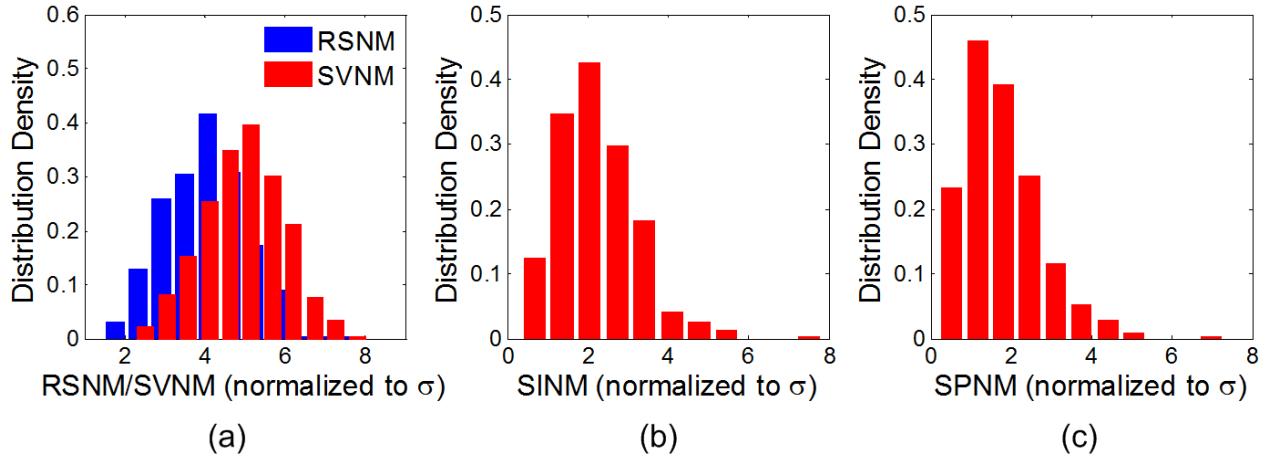


Figure 4.3: Distribution densities of (a) RSNM/SVNM, (b) SINM, and (c) SPNM measured at $V_{DD} = 0.6V$ from the same SRAM macros using all-internal-node access. Note: the y-axis scale differs from Figure 2.12 because each metric is normalized to its σ value.

metrics, measured at $V_{DD} = 0.6V$, is presented in Figure 4.3. Results show that while the RSNM and SVNM distributions remain Gaussian at $V_{DD} = 0.6V$, the distributions of both SINM and SPNM start to resemble log-normal distributions as transistors approach the subthreshold region of operation. Figure 4.4 presents the scatter plot for I_W versus WNM measured at $V_{DD} = 0.6V$ from the same SRAM macro. The measured results agree with the simulated results presented in Section 2.2.2, indicating excellent correlation between the I_W and the WNM measurements. Each plot contains the measured data from 360 SRAM cells, which equals the total number of SRAM CUTs, wired for all-internal-node access, per test chip (i.e. 18×20).

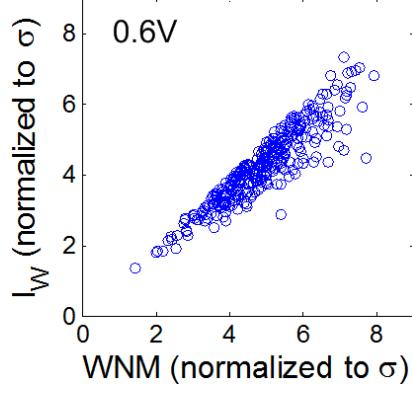


Figure 4.4: Scatter plot for I_W versus WNM measured at $V_{DD} = 0.6V$ from the same SRAM macros using all-internal-node access.

Large-Scale SRAM Design Metrics

The measured transfer curves extracted from functional SRAM arrays with direct bit-line access for characterizing the large-scale SRAM design metrics is presented in Figure 4.5. Figure 4.6 illustrates the impact of within-cell mismatch on measured SRRV and WRRV transfer curves. All transfer curves in Figure 4.6 are extracted from SRAM bitcells with a lower cell β -ratio at the *CH* storage node - corresponding to the simulated transfer curves in Figure 2.28a-b and Figure 2.30a-b. As expected from the simulation results, when the less read-stable *CH* node holds a '0', both measured SRRV and WRRV transfer curves exhibit a sharp fall-off in the *BLC* current ($I_{MEAS,BLC}$), indicating a clear SRAM cell data disturbance in the form of a bit flip. However, when the more read-stable *CL* node holds a '0', only some measured transfer curves exhibit a sharp fall-off in $I_{MEAS,BL}$ while other transfer curves do not.

Scatter plots are generated to evaluate the correlations between the different large-scale read stability and writeability metrics as well as between the large-scale design metrics and the conventional design metrics. Figures 4.7-4.8 present the scatter plots for 4 pairs of design metrics. Each pair of design metrics is measured for the same set of SRAM cells first at $V_{DD} = 0.8V$ and then at $V_{DD} = 0.5V$ to expose low read stability and writeability¹ - this is done in order to identify the point of failure for each design metric and to investigate the correlations between the different design metrics near the point of failure. The WRRV-RSNM pair (Figure 4.7a) and the WWTW- I_W pair (Figure 4.8a) are measured from the same SRAM macros - for a total of 360 SRAM cells - using all-internal-node access². I_W is presented here rather than WNM because large sweep margins are needed at $V_{DD} = 0.8V$ to expose convexity in the write-VTC for SRAM cells with higher writeability, otherwise the WNM values saturate and the correlation becomes exacerbated due to an error in the extraction process. A 200mV N-well bias (V_{NW}) is applied when measuring the SRAM writeability (i.e. WWTW- I_W pair) in the SRAM macros for the case of $V_{DD} = 0.5V$; this is

¹This is similar to the simulated scatter plots presented in Chapter 2.

²As mentioned in Section 3.2.1, large-scale SRAM design metrics can be measured in SRAM macros with all-internal-node access to establish correlations against the conventional SRAM design metrics in measurement.

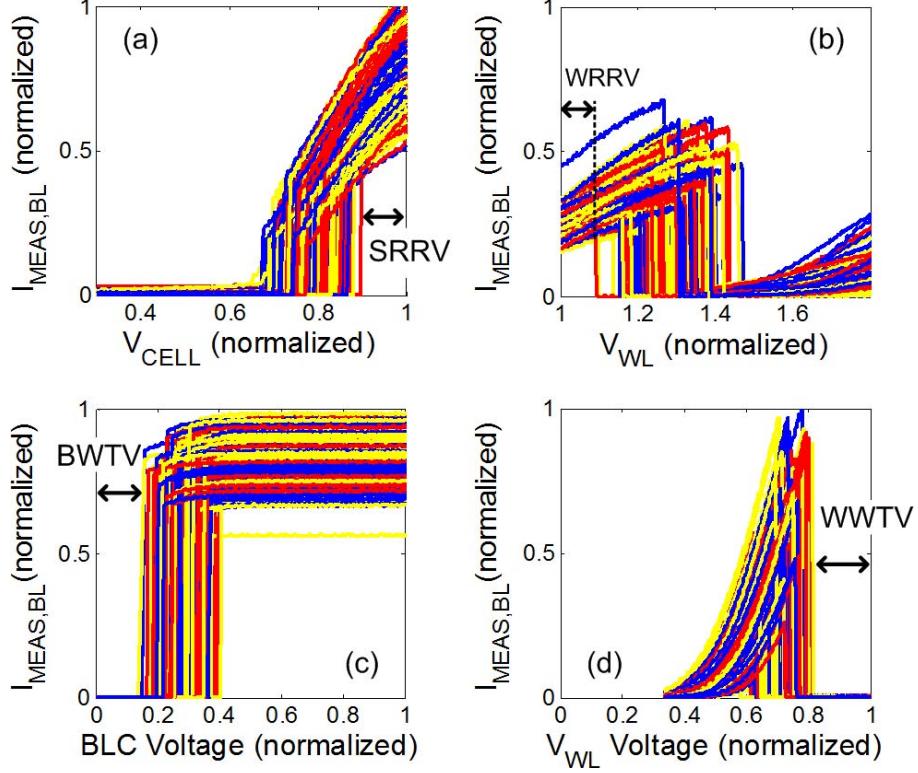


Figure 4.5: Measured transfer curves for (a) SRRV extraction, (b) WRRV extraction, (c) BWTV extraction, and (d) WWTW extraction from functional SRAM arrays using direct bit-line access.

done to further reduce the SRAM writeability and expose write failures by decreasing the V_{TH} of the PMOS pull-up transistors through a forward body bias (FBB). The SRRV-WRRV pair (Figure 4.7b) and the BWTV-WWTW pair (Figure 4.8b) are measured from the same functional SRAM array using direct bit-line access³. At $V_{DD} = 0.8V$, the μ of each measured metric sits comfortably above 6σ and a slight dispersion is observed in the measured data of each metric pair - this is in agreement with the simulation results presented in Chapter 2. This dispersion is generally smaller at lower measured values and larger at higher measured values - giving each scatter plot a drum-stick shape. However, when the supply is dropped to $0.5V$ ⁴ and the SRAM bitcell is pushed to the edge of stability, excellent agreement is established within each metric pair, especially near the zero crossing (i.e. the origin). This indicates that the three read stability metrics (RSNM, SRRV, and WRRV) and the three writeability metrics (I_W , BWTV, and WWTW) share a common point of failure and have excellent agreement near failure. Therefore, both the conventional and the large-scale SRAM design metrics can be used for SRAM failure estimation.

³Figures 4.7b and 4.8b are plotted for 4096 SRAM cells.

⁴A $200mV$ V_{NW} is applied, in addition to the reduced V_{DD} , for the measurements of WWTW and I_W in Figure 4.8a.

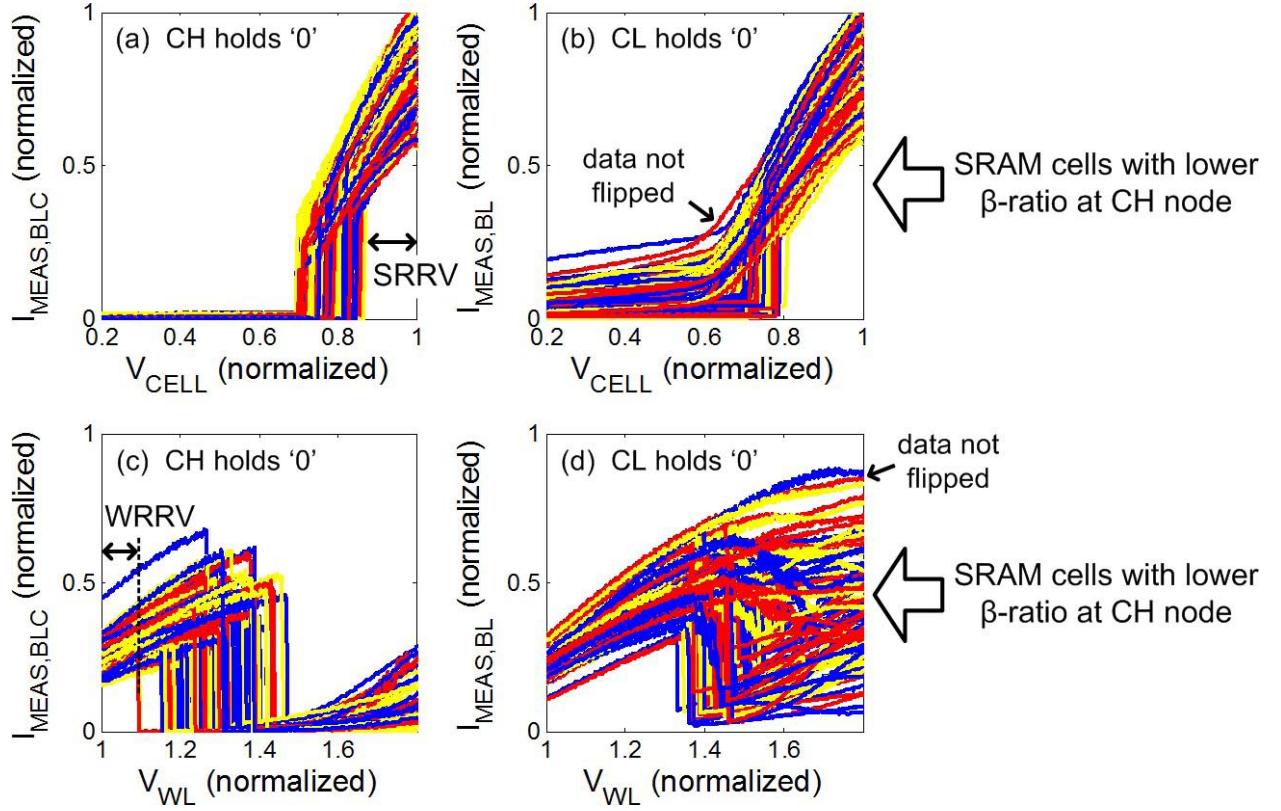


Figure 4.6: (a) Measured SRRV transfer curves for storing a '0' at the less read-stable *CH* node; all transfer curves exhibit sharp fall off in $I_{MEAS,BLC}$. (b) Measured SRRV transfer curves for storing a '0' at the more read-stable *CL* node; only some transfer curves exhibit a sharp fall-off in $I_{MEAS,BL}$ while other transfer curves do not. (c) Measured WRRV transfer curves for storing a '0' at the less read-stable *CH* node; all transfer curves exhibit sharp fall off in $I_{MEAS,BLC}$. (d) Measured WRRV transfer curves for storing a '0' at the more read-stable *CL* node; only some transfer curves exhibit a sharp fall-off in $I_{MEAS,BL}$ while other transfer curves do not.

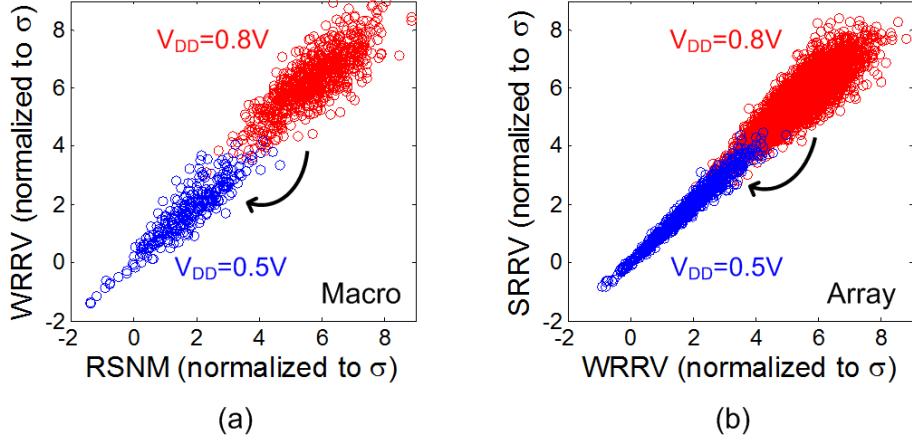


Figure 4.7: (a) Scatter plot for WRRV versus RSNM measured from the same SRAM macro using all-internal-node access at $V_{DD} = 0.8V$ and $0.5V$. (b) Scatter plot for SRRV versus WRRV measured from the same functional SRAM array using direct bit-line access at $V_{DD} = 0.8V$ and $0.5V$.

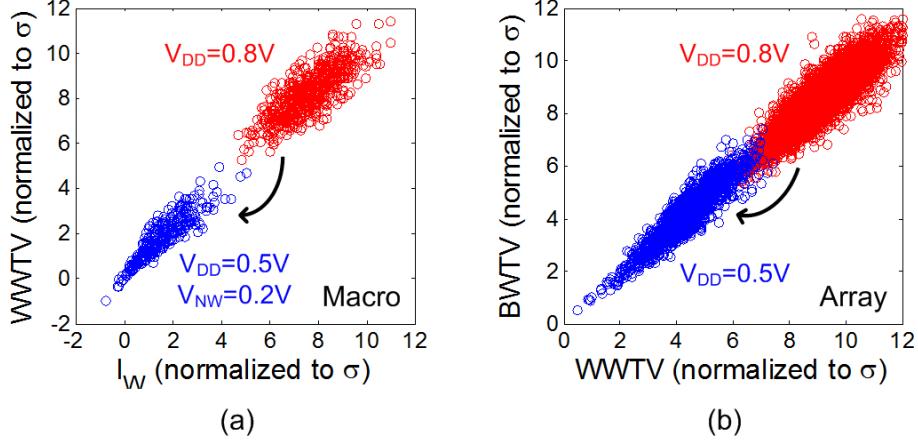


Figure 4.8: (a) Scatter plot for WWTV versus I_W measured from the same SRAM macro using all-internal-node access at $V_{DD} = 0.8V$ and at $V_{DD} = 0.5V$ with $V_{NW} = 0.2V$. (b) Scatter plot for BWTV versus WWTV measured from the same functional SRAM array using direct bit-line access at $V_{DD} = 0.8V$ and $0.5V$.

4.2.2 Measured Distributions for SRAM Read Stability and Writeability

Figure 4.9 presents the semi-log plots for the distribution densities of both the read stability and the writeability metrics measured at $V_{DD} = 0.7V$ fitted to Gaussian distributions. The distribution densities of RSNM, SRRV, and WRRV is presented in Figure 4.9a, and the distribution densities of I_W , BWTV, and WWTV is presented in Figure 4.9b; each metric is normalized to its σ value. RSNM and I_W are measured from SRAM macros via all-internal-node access for 360 SRAM cells; each metric is measured for both data polarities,

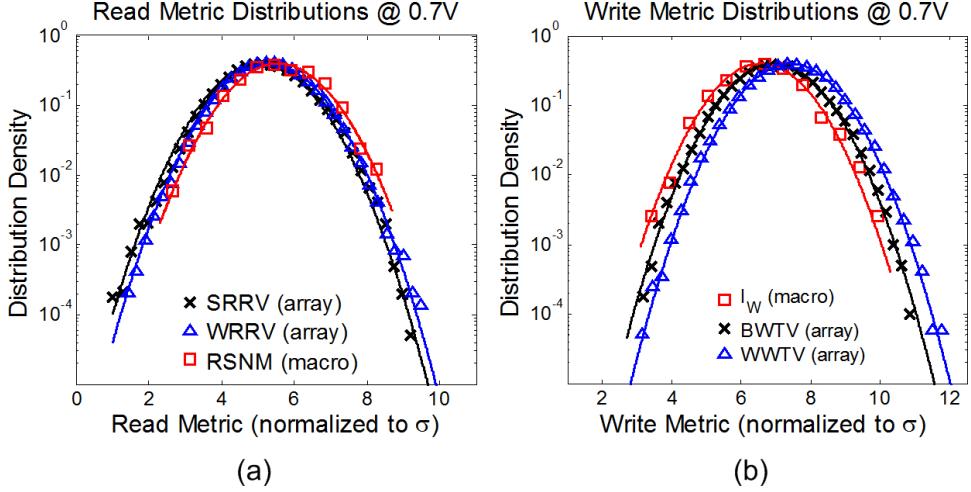


Figure 4.9: Semi-log plots for (a) the measured read metric distributions using RSNM, SRRV, and WRRV at $V_{DD} = 0.7V$; and (b) the measured write metric distributions using I_W , BWTW, and WWTW at $V_{DD} = 0.7V$.

yielding 720 data points. SRRV, WRRV, BWTW, and WWTW are measured from a 64kb functional SRAM sub-array using direct bit-line access; BWTW and WWTW are captured for a single data polarity while SRRV and WRRV are extracted for the less read-stable data polarity. Figure 4.9 shows that the large-scale characterization captures several orders of magnitude more statistical data than the conventional silicon characterization while requiring much less hardware overhead⁵.

Figure 4.10 shows the normal probability plots [34] for SRRV, WRRV, and RSNM measured at $V_{DD} = 0.7V$. The large-scale read stability metrics, SRRV and WRRV, are measured from a 64kb functional SRAM sub-array. Figure 4.10a-b exhibit good normality at the center of the distributions for both measured metrics. Slight deviations, from a normally distributed function, are observed near both the upper and the lower tails of the distributions - at approximately above and below $\pm 3\sigma$. In both cases, the normal probability plots show a slight bending upwards and to the left, indicating a slightly right skewed distribution. This is in agreement with ordered statistics as both SRRV and WRRV are extracted for the less read-stable data polarity, which is equivalent to taking the minimum of two Gaussian distributions and yields a right skewed distribution. The conventional read stability metric, RSNM, is measured in SRAM macros via all-internal-node characterization - 720 data points are extracted for both data polarities from 360 SRAM cells. Figure 4.10c exhibits good normality up to $\pm 3\sigma$ in the measured RSNM⁶.

Figure 4.11 presents the normal probability plots for BWTW, WWTW, and I_W measured at $V_{DD} = 0.7V$. The large-scale writeability metrics, BWTW and WWTW, are measured from a 64kb functional SRAM sub-array. Figure 4.11a-b exhibit good normality down to more than -4σ for both writeability metrics. The conventional writeability metric, I_W , is measured in SRAM macros via all-internal-node characterization and its measurements,

⁵The hardware overhead for both schemes is discussed in Chapter 3.

⁶The RSNM data in Figure 4.10c is extracted for both data polarities in each SRAM cell and therefore does not show a right skewed distribution.

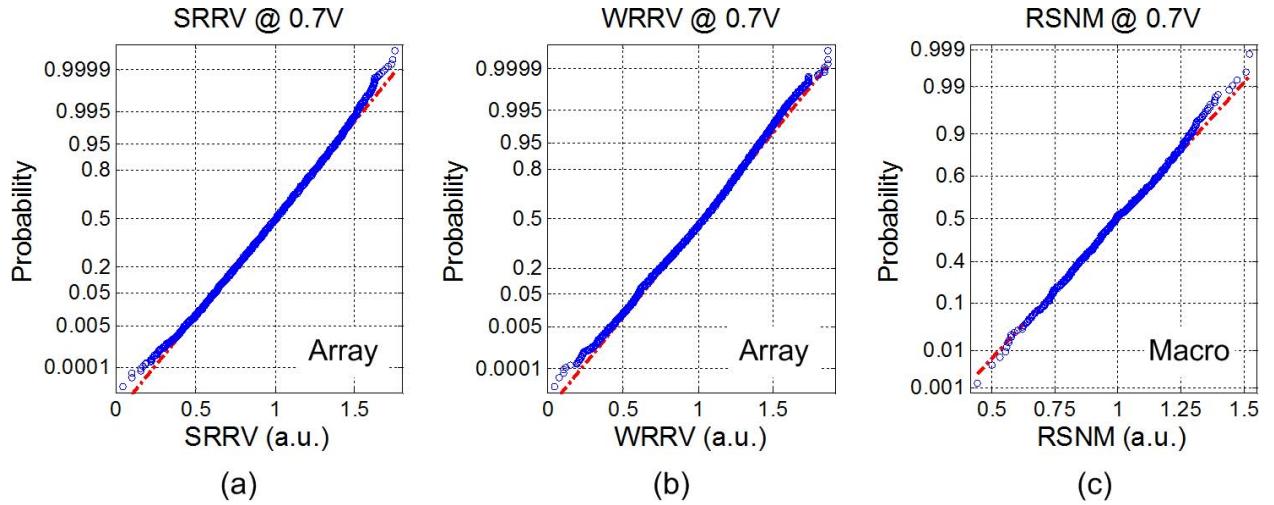


Figure 4.10: Normal probability plots for (a) SRRV, (b) WRRV, and (c) RSNM measured at $V_{DD} = 0.7V$.

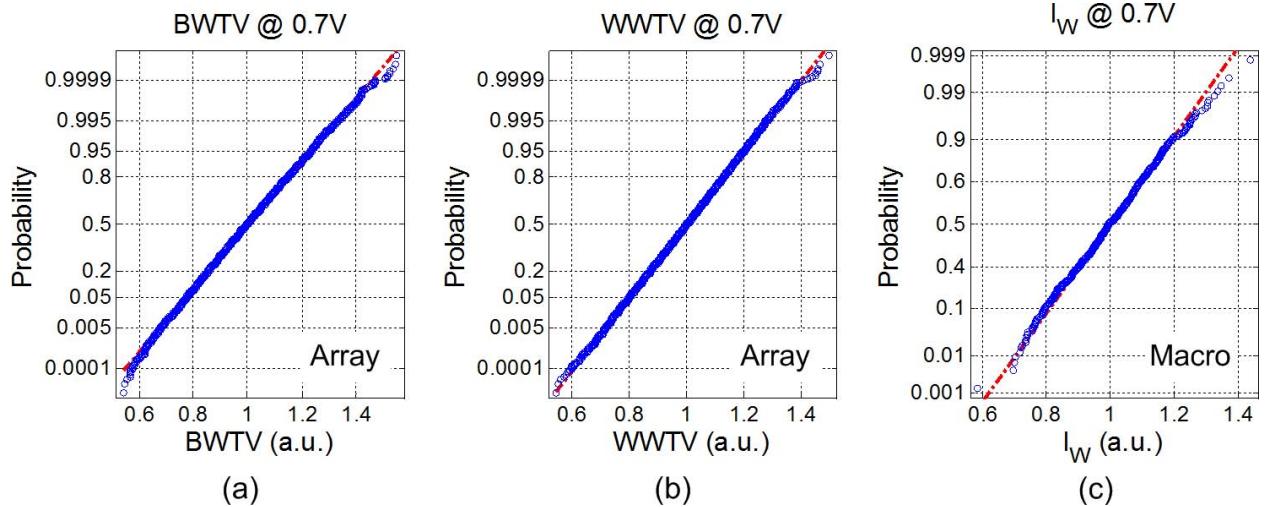


Figure 4.11: Normal probability plots for (a) BWTV, (b) WWTV, and (c) I_W measured at $V_{DD} = 0.7V$.

as illustrated in Figure 4.11c, show no significant deviation from a normal distribution down to -3σ .

4.2.3 Understanding the μ/σ Value

The distribution densities plotted in Figure 4.9 indicate that each read stability metric, as well as writeability metric, measured at $V_{DD} = 0.7V$ exhibit a slightly different μ/σ value. The difference in the μ/σ value between the large-scale SRAM design metrics and the conventional SRAM design metrics can be attributed to the fact that the conventional SRAM design metrics are measured in standalone SRAM macros and may not reflect the functionality of SRAM cells in dense arrays - in particular, one source of systematic variabil-

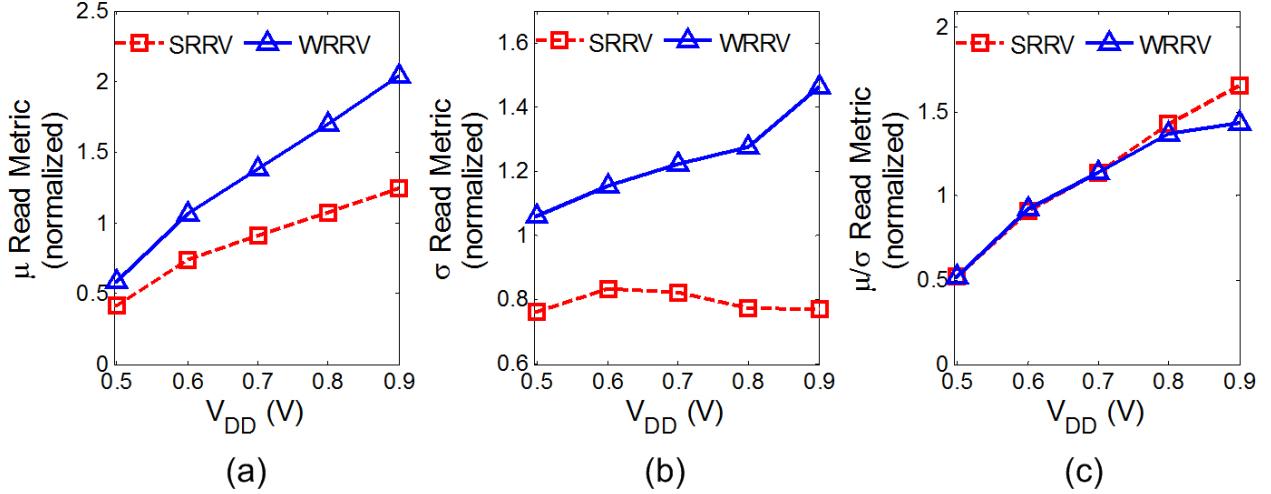


Figure 4.12: Measured (a) μ , (b) σ , and (c) μ/σ of SRRV and WRRV as a function of V_{DD} .

ity is speculated in Section 4.5.1 to modulate the SRAM read stability and writeability in the SRAM macros but not in the functional SRAM arrays. However, Figure 4.9 also indicates a slight mismatch in the μ/σ value amongst the large-scale SRAM design metrics measured for the same 64kb SRAM sub-array. A careful examination of the different measurements reveals a difference in the operation of the pass-gate transistor at the point of read/write margin extraction - i.e. where the data polarity flips - for the different large-scale metrics.

Figure 4.12 plots the μ , the σ , and the μ/σ value of the measured SRRV and WRRV as a function of V_{DD} . Significant differences in both the μ and the σ values of SRRV and WRRV is revealed in Figure 4.12a-b. This can be explained by examining the WRRV measurement process. During the WRRV measurement, as the WL voltage is driven above the SRAM cell supply voltage (V_{CELL}), the gate overdrive of the pass-gate transistor at the '0' storage node, which is the root cause of a cell disturbance during the read cycle, gradually saturates as the '0' storage node rises above V_{SS} and follows the increasing WL voltage. This is manifested in a reduced sensitivity in the measured bit-line current to the WL voltage as WL is driven above V_{CELL} . Figure 4.13 graphically illustrates this phenomenon, where a reduced bit-line current sensitivity is indicated by a decreasing slope in the measured current as the WL voltage is increased beyond V_{CELL} . This sensitivity is further reduced as the operating voltage is increased because the inverter trip point at the '1' storage side is increased and the SRAM cell can withstand a greater rise at the '0' storage node before data disturbance. Figure 4.13 confirms this as the slope, near data disturbance, in the bit-line current transfer curve measured at $V_{DD} = 0.9V$ is lower than the case for $V_{DD} = 0.7V$. Consequently, a higher difference between V_{WL} and V_{CELL} is needed for the WRRV characterization than for the SRRV characterization, despite having both bit-lines biased at a higher voltage than V_{CELL} when measuring SRRV⁷. As a result, the WRRV measurements show a higher μ than the SRRV measurements - as indicated in Figure 4.12a. In addition, a reduced sensitivity in the pass-gate transistor strength to the WL voltage also implies that a larger spread in V_{WL} is needed to produce the same spread in the on-current conducted by the pass-gate transistor.

⁷This is for the case when SRRV > 0.

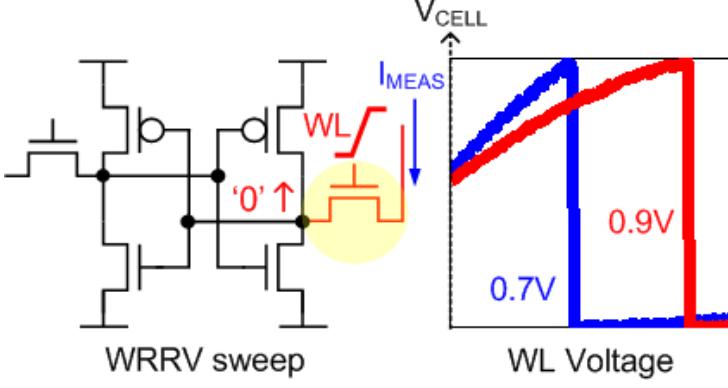


Figure 4.13: The bit-line current sensitivity to the WL overdrive is reduced due to a rise in the '0' storage node voltage and is more pronounced as V_{DD} is increased.

Therefore, the WRRV measurements display a larger σ than the SRRV measurements - as indicated in Figure 4.12b. Since the bit-line current sensitivity is V_{DD} -dependent, the σ value of the SRRV measurements is also expected to vary with V_{DD} . Figure 4.12b shows that while the σ value of the WRRV measurements (σ_{WRRV}) exhibit a heightened dependence on V_{DD} , the σ value of the SRRV measurements (σ_{SRRV}) is nearly supply-independent.

The difference in the μ/σ value in the measurements of SRRV and WRRV at a given V_{DD} depends on the relative augmentations in the values of their μ and σ . Figure 4.12c shows that, at $V_{DD} = 0.7$ and below, both SRRV and WRRV measurements result in similar μ/σ values as the shifts in their μ and σ values compensate each other. At $V_{DD} = 0.8V$ and beyond, the μ/σ value of the WRRV measurements drops below that of the SRRV measurements due to a greater rise in σ_{WRRV} . Measurements of WWRV at above $V_{DD} = 0.9V$ are not conducted to keep the WL voltage below 1.5V to avoid transistor gate-oxide breakdown. Consequently, the WRRV is not suitable for read stability characterization at higher operating voltages but can be useful for read stability characterization near failure. Overall, the SRRV can be measured over a more complete range of operating voltages and has a nice property of a supply-independent σ value, which makes it easy for $V_{MIN,RD}$ estimation - as discussed in Section 4.3.2.

Figure 4.14 plots the μ , the σ , and the μ/σ value of the measured BWTV and WWTV as a function of V_{DD} . Marginal differences in both the μ and the σ values of BWTV and WWTV is revealed in Figure 4.14a-b. Although both BWTV and WWTV measure the write trip voltage of the SRAM cell, a closer examination of the characterization processes reveals that their measurements stress the pass-gate transistor differently. During the bit-line sweep of the BWTV characterization, the strength of the pass-gate transistor at the '1' storage side is modulated through adjusting both its gate-source voltage (V_{GS}) and drain-source voltage (V_{DS}) while the pass-gate transistor at the '0' storage side remains in saturation until the data polarity is flipped. Since the source node of the pass-gate transistor at the '1' storage node is ramped from V_{DD} to V_{SS} , the pass-gate transistor is first put under a full- V_{DD} reverse body bias (RBB) and the magnitude of the applied RBB drops with the bit-line voltage during the BWTV characterization. Additionally, the V_{TH} of the pass-gate transistor, at the '1' storage side, is also modulated by a drain-induced barrier lowering

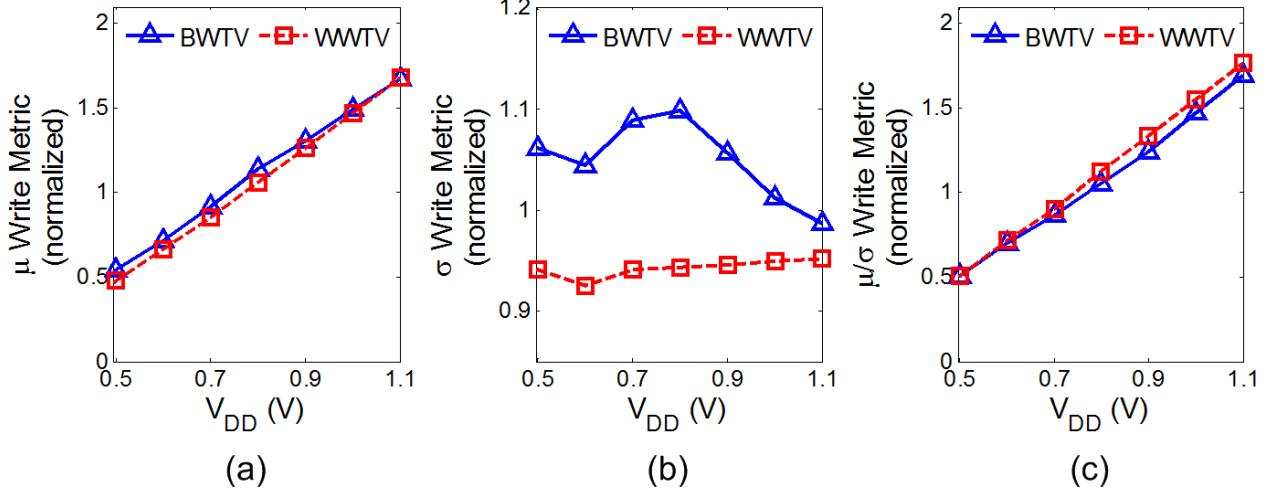


Figure 4.14: Measured (a) μ , (b) σ , and (c) μ/σ of BWTW and WWTW as a function of V_{DD} .

(DIBL) [152] effect due to a non-zero V_{DS} . Converse to the RBB effect, the effect of DIBL increases during the bit-line sweep as the pass-gate V_{DS} increases. Since RBB increases the transistor V_{TH} and DIBL reduces it, the pass-gate transistor V_{TH} is reduced and its current drive, relative to the PMOS pull-up transistor, is increased with the decreasing bit-line voltage during the BWTW characterization. In addition, due to a decrease in the value of BWTW with a decreasing V_{DD} , the pass-gate transistor current drive, relative to the PMOS pull-up transistor, is increased as V_{DD} is decreased at the point where BWTW is extracted - i.e. at the bit-line voltage causing the data polarity to flip. Consequently, the μ value of the measured BWTW (μ_{BWTW}) deviates slightly from a linear dependence on V_{DD} in the positive direction as V_{DD} is decreased - as shown in Figure 4.14a. On the other hand, the word-line sweep during the WWTW characterization modulates the strengths of both pass-gate transistors through adjusting only the gate-source voltage (V_{GS}), leading to a fixed pass-gate transistor V_{TH} until the data polarity is flipped. As a result, Figure 4.14a shows a linear dependence on V_{DD} [58, 64, 134] for the μ value of the measured WWTW (μ_{WWTW}).

The increased within-die variation of transistor V_{TH} due to RBB [97] and the varying degree of RBB and DIBL applied at each bit-line voltage for BWTW extraction increases the σ value of the BWTW measurements (σ_{BWTW}). Consequently, Figure 4.14b indicates a higher σ value for the BWTW measurements than for the WWTW measurements, where the pass-gate transistor V_{TH} remains constant during each characterization sweep. In addition, the σ_{BWTW} is shown, in Figure 4.14b, to have elevated V_{DD} -dependence because the degree of RBB and DIBL varies as the value of BWTW changes with V_{DD} ⁸. On the other hand, σ value for the WWTW measurements (σ_{WWTW}) is shown to be relatively supply-independent. Finally, the difference in the μ/σ value in the measurements of BWTW and WWTW at a given V_{DD} depends on the relative augmentations in the values of their μ and σ . Figure 4.14c shows that, at below $V_{DD} = 0.7$, both BWTW and WWTW measurements result in similar

⁸The exact dependence of the σ_{BWTW} on V_{DD} depends on the cumulative impact of RBB and DIBL on the BWTW values at each V_{DD} .

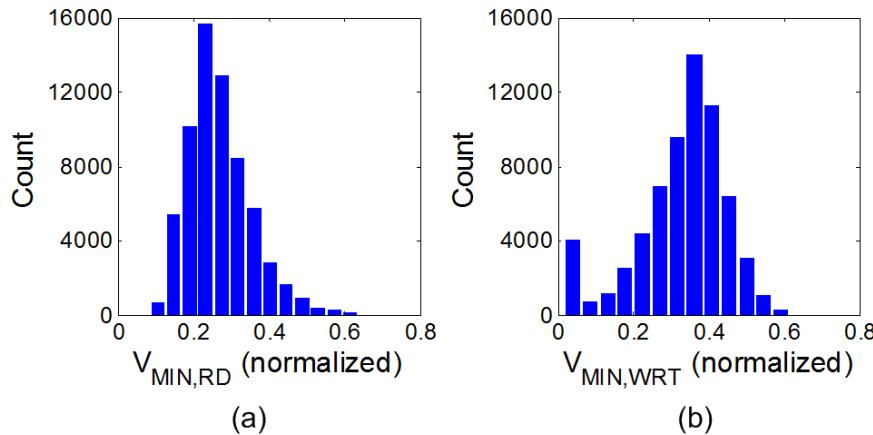


Figure 4.15: Distributions of (a) $V_{MIN,RD}$ and (b) $V_{MIN,WRT}$ measured in a 64kb functional SRAM sub-array.

μ/σ values as the shifts in their μ and σ values compensate each other. At $V_{DD} = 0.7V$ and above, the μ/σ value of the BWTW measurements drops slightly below that of the WWTW measurements due to a decrease in the slope of μ_{BWTW} as a functional of V_{DD} . Due to a more linear dependence on V_{DD} and a supply-independent σ value, the WWTW can more effectively quantify the impact of V_{DD} on SRAM writeability and can be more easily used for $V_{MIN,WRT}$ estimation - as discussed in Section 4.3.2.

4.3 Measurements and Estimation of the SRAM Minimum Operating Voltage

4.3.1 V_{MIN} Measurements

The minimum operating voltage (V_{MIN}) distributions for static read and write operations measured in a 64kb functional SRAM sub-array are presented in Figure 4.15. In both cases, V_{MIN} measurements are extracted from each SRAM bitcell for the less read stable or the less writeable data polarity - i.e. the maximum value of V_{MIN} is taken for each bitcell. Figure 4.15a shows that process variability causes the $V_{MIN,RD}$ distribution to have a long tail in the direction of higher $V_{MIN,RD}$ values - similar to the case for the data retention voltage (DRV) distributions [118,147]. This figure also indicates that a majority of the measured bitcells can achieve read data retention at very low operating voltages - below the transistor V_{TH} , and thus yield a log-normal shaped distribution, at the lower tail, due to an exponential dependence of transistor currents on transistor V_{TH} . However, the array $V_{MIN,RD}$ value is determined by the maximum point in the long tail of the $V_{MIN,RD}$ distribution and is notably above the transistor V_{TH} . Figure 4.15b shows the measured $V_{MIN,WRT}$ distribution. Due to the high writeability of the $0.374 \mu\text{m}^2$ bitcells in this process, a weak write [90], in the form of a 100mV reduction in the word-line bias⁹, is applied to generate

⁹A reduced word-line voltage is applied rather than a raised bit-line voltage (above V_{SS} at the '1' storage node) to avoid altering the pass-gate V_{TH} .

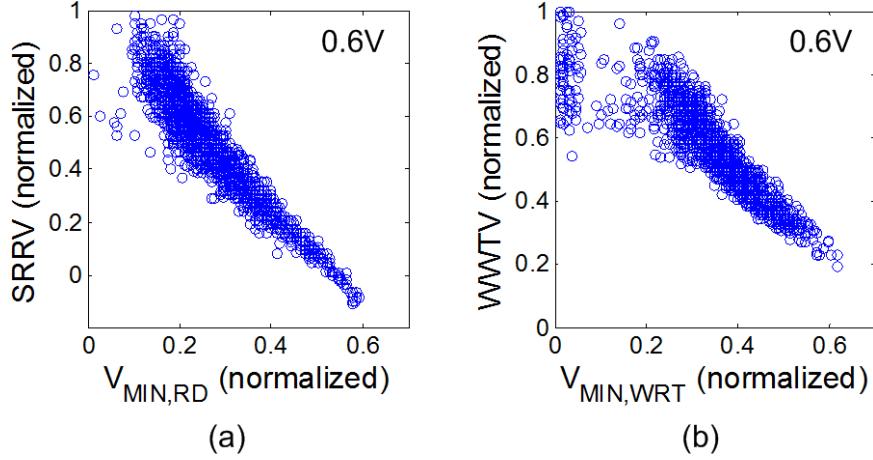


Figure 4.16: Scatter plots for (a) SRRV versus $V_{MIN,RD}$ and (b) WWTV versus $V_{MIN,WRT}$ measured in a 64kb functional SRAM sub-array demonstrating excellent correlation near failure. SRRV and WWTV are measured at $V_{DD} = 0.6V$; a 100mV word-line weak write is applied during $V_{MIN,WRT}$ characterization.

a reasonable amount of write failures above the transistor V_{TH} . The measured distribution shows two peaks, with a high write failure count near $V_{DD} = 0V$ corresponding to SRAM cells with very high writeability, where the bit flip is caused by a standby retention failure. It is important to note that, although a retention failure cannot be easily distinguished from a write failure under very low operating voltages, the array $V_{MIN,WRT}$ value is determined by the maximum point in rightmost tail of the $V_{MIN,WRT}$ distribution; therefore, the peaking of the $V_{MIN,WRT}$ distribution near $V_{DD} = 0V$ can be disregarded.

A direct correlation between the per-cell V_{MIN} measurements and the large-scale SRAM read/write margin measurements is presented in Figure 4.16 for a 64kb functional SRAM sub-array. Figure 4.16a shows the scatter plot of SRRV versus $V_{MIN,RD}$, where the SRRV is measured at $V_{DD} = 0.6V$ to expose near-failure read stability. The scatter plot of WWTV (measured at $V_{DD} = 0.6V$ to expose near-failure writeability) versus $V_{MIN,WRT}$ (measured with a 100mV word-line weak write) is shown in Figure 4.16b. Measured results indicate excellent agreement between the per-cell $V_{MIN,RD}/V_{MIN,WRT}$ and the large-scale SRRV/WWTV, particularly at high $V_{MIN,RD}/V_{MIN,WRT}$ and low SRRV/WWTV values where the bitcell approaches read stability/writeability failure. Since the SRRV is measured at $V_{DD} = 0.6V$, the zero crossing of the SRRV measurements, in Figure 4.16a, maps exactly to $V_{MIN,RD} = 0.6V$ (before normalization)¹⁰. Figure 4.16b shows a large cloud near the y-axis at high WWTV and low $V_{MIN,WRT}$ values, corresponding to the $V_{MIN,WRT}$ values captured due to standby retention failures. The excellent agreement between the extracted V_{MIN} and the large-scale read/write margin measurements (particularly the SRRV and the WWTV) suggest that the SRAM design metrics measured using direct bit-line characterization can be used to estimate V_{MIN} .

¹⁰The zero crossing of the WWTV measurements does not correspond to $V_{MIN,WRT} = 0.6V$ due to the application of a 100mV word-line weak write.

4.3.2 V_{MIN} Estimation

Section 4.3.1 points out that the minimum operating voltage (V_{MIN}) of an SRAM array is determined by the most stable bitcell, which involves extracting the extreme maximum value of the $V_{MIN,RD}$ and the $V_{MIN,WRT}$ distributions. Consequently, to characterize the V_{MIN} of large SRAM arrays, every bitcell in each array must be examined to capture the entire $V_{MIN,RD}$ and $V_{MIN,WRT}$ distributions. This is particularly true for $V_{MIN,RD}$ given its very long tail in the positive direction (Figure 4.15a). Therefore, it is desirable to formulate a method to quickly estimate the V_{MIN} of an SRAM array of arbitrary size without having to examine each individual bitcell. A mixture importance sampling (IS) based approach is developed in [78] to quickly estimate rare failure events in SRAM by re-centering the sample space within the failure region, which can be realized through shifting the means of the transistor V_{TH} in the 6-T bitcell. This approach is extended in [46] by adopting a norm minimization algorithm to shift the means of the V_{TH} of all six transistors in a bitcell along the most likely path to failure. Alternately, the Extreme Value Theory (EVT) [89] is adopted in [127] to generate the tail of the distribution of a rare event, by blocking samples that are likely to occur, and then fitting the tail to a Generalized Pareto Distribution (GPD) for failure prediction. This method is applied in [127] to estimate the worst case write time for a 64-bit SRAM column and is extended in [128] to estimate the SRAM DRV. While these methods can provide very fast failure predictions, their accuracies depend on the transistor models. As processes become increasingly complex and harder to control, designers can no longer solely rely on model accuracy to fully capture the random effects in large cache memories. Since these methods require a well controlled simulation framework to explore statistics in the failure region, they cannot be easily applied using measured results.

In Section 4.3.1, a direct relationship between the large-scale SRAM design metrics and the per-cell V_{MIN} is established through measurements. The excellent correlation of the large-scale metrics with V_{MIN} indicate that they can be used for V_{MIN} estimation. Since each large-scale metric can fit reasonably well to a Gaussian distribution¹¹, the μ and σ values can be extracted and applied to estimate the probability of failure at each V_{DD} value using simple statistics. This is similar to the approach adopted in [20], to predict the read/write fail count for a sample 1k SRAM cells, and in [147], which uses a linear fit of μ (as a function of V_{DD}) and a static σ for the hold static noise margin (HSNM) to estimate the probability of data retention failure as a function of V_{DD} .

Since the actual read margin (and write margin) of each SRAM bitcell is defined as the minimum of two margins extracted for each data polarity, the resulting random variable is the minimum of two Gaussian variables, which no longer fits to a normal distribution. Order statistics can be used to define the probability density function (*PDF*) of the minimum of n independent and identically distributed (i.i.d.) random variables. Let X_1 and X_2 denote two i.i.d. normal random variables with a common mean, μ_C , and a common variance, σ_C^2 - i.e. $X_1 \sim N(\mu_C, \sigma_C^2)$ and $X_2 \sim N(\mu_C, \sigma_C^2)$. The *PDF* of $X = \min(X_1, X_2)$ is given by order

¹¹SRRV (and WRRV), in this case, should be evaluated for both data polarities and stored independently. Although the SRRV (and WRRV) cannot be successfully measured for both data polarities, which may cause the measured data to deviate slightly from a Gaussian distribution beyond more than $\pm 3\sigma$, the μ and σ values should still be accurately extracted.

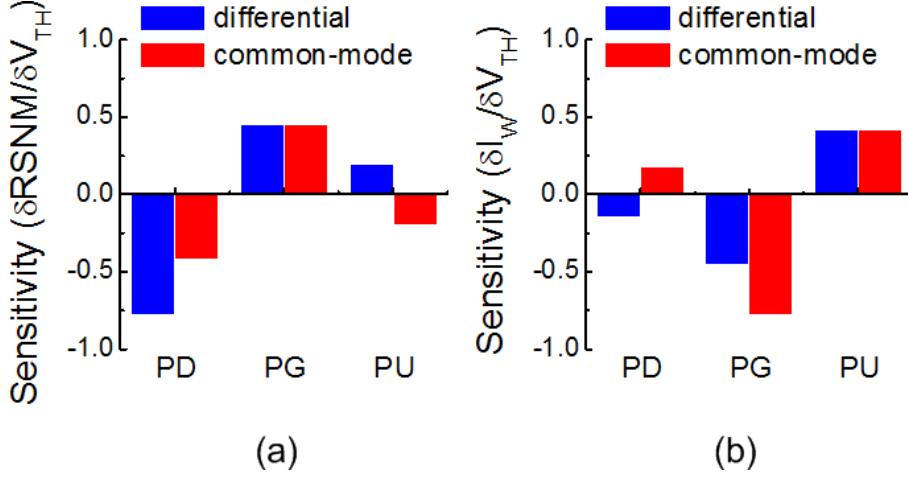


Figure 4.17: Sensitivity of (a) RSNM and (b) WNM to differential and common-mode variations in the pull-down, pass-gate, and pull-up transistor pairs within an SRAM bitcell [32].

statistics as

$$f(x) = 2 \times f(x_1) \times [1 - F(x_1)] \quad (4.1)$$

where $f(x_1)$ is the *PDF* and $F(x_1)$ is the cumulative distribution function (*CDF*) of X_1 . The *CDF* of $X = \min(X_1, X_2)$ [27] can be expressed as

$$F(x) = 2 \times F(x_1) - F^2(x_1) \quad (4.2)$$

where $F(x_1)$ is the *CDF* of X_1 .

Equation 4.1 is adopted recently in [27, 61, 147] to estimate the probability of static stability failure in SRAM. However, this equation makes the assumption that the noise margins measured for the two data polarities of a bitcell are independent of each other and are identically distributed - i.e. they have the same mean and the same variance. Sections 2.3.1 and 4.2 illustrate that due to random mismatch within each bitcell, the bitcell typically favors the storage of one data polarity over the other. It has been previously shown in [32, 136] that while the read stability of an SRAM bitcell is equally sensitive to both differential (mismatch) and common-mode (systematic) variations in the pass-gate transistors, the overall sensitivity of the read stability is dominated by a strong sensitivity to differential variations in the pull-down transistors; thus resulting in a negative correlation between the read stability margins of the two data polarities¹². Conversely, [32] shows that while the writeability of an SRAM bitcell is equally sensitive to both differential and common-mode variations in the pull-up transistors, the overall sensitivity of the writeability is dominated by a strong sensitivity to common-mode variations in the pass-gate transistors; thus resulting in a positive correlation between the writeability margins for the two data polarities. Figure 4.17 graphically summarizes the findings in [32]. The resulting positive correlation between the two RSNMs - RSNM1 and RSNM2 - and negative correlation between the two WNMs - WNM1 and WNM2 - are established, using 3k-sample MC simulations, in Figure 4.18. The MC simulation environment is identical to that used in Chapter 2 - with common-mode

¹² [27] also pointed out this negative correlation.

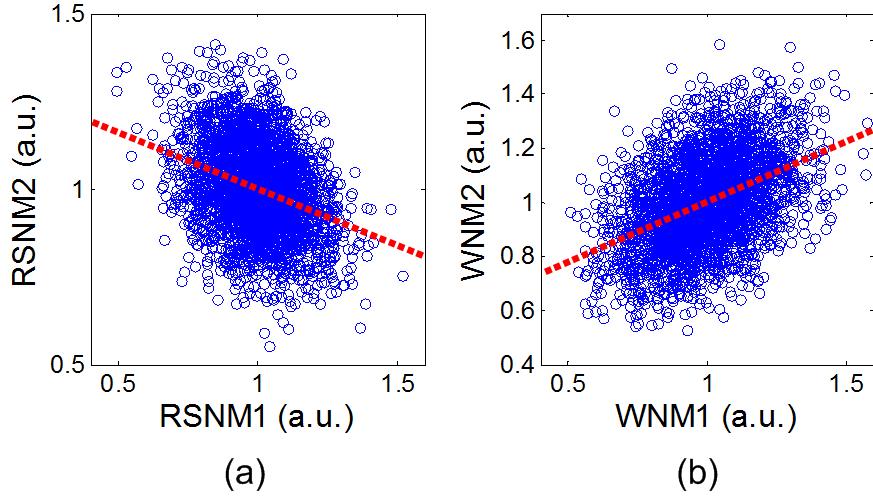


Figure 4.18: (a) Scatter plot of RSNM1 versus RSNM2, along with a linear fit, showing a negative correlation between the read stability of the two data polarities. (b) Scatter plot of WNM1 versus WNM2, along with a linear fit, showing a positive correlation between the writeability for the two data polarities.

global variations in L_G , W , T_{OX} , and V_{TH} as well as random mismatch in V_{TH} for all transistors. Furthermore, while the stability margins acquired through simulations (for the two data polarities) may yield approximately the same means and variances, as shown in [27], physical differences between the layouts of each half cell may produce a lithography-induced systematic shift between the means of the two read/write margins of a bitcell [64, 65] (Section 4.5.2). Therefore, a more general expression for the *PDF* of $X = \min(X_1, X_2)$, where $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are not necessarily i.i.d., is needed to more accurately estimate the SRAM read/write failure probability.

The mathematical expression for the exact distribution of the minimum of two Gaussian random variables is provided by the statistics literature [17, 42, 96, 140]. Let (X_1, X_2) represent a bivariate normal random vector where $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ denote two normal random variables with means (μ_1, μ_2) , variances (σ_1^2, σ_2^2) , and a correlation coefficient ρ . The *PDF* of $X = \min(X_1, X_2)$ can be expressed as

$$f(x) = f_1(x) + f_2(x) \quad (4.3)$$

with

$$f_1(x) = \frac{1}{\sigma_1} \times \phi\left(\frac{x - \mu_1}{\sigma_1}\right) \times \Phi\left(\frac{\rho(x - \mu_1)}{\sigma_1 \sqrt{1 - \rho^2}} - \frac{x - \mu_2}{\sigma_2 \sqrt{1 - \rho^2}}\right) \quad (4.4)$$

$$f_2(x) = \frac{1}{\sigma_2} \times \phi\left(\frac{x - \mu_2}{\sigma_2}\right) \times \Phi\left(\frac{\rho(x - \mu_2)}{\sigma_2 \sqrt{1 - \rho^2}} - \frac{x - \mu_1}{\sigma_1 \sqrt{1 - \rho^2}}\right) \quad (4.5)$$

where $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$ is the *PDF* and $\Phi(x) = (1/2)[1 + \text{erf}(x/\sqrt{2})]$ ¹³ is the *CDF* of the standard normal distribution, i.e. $N(0, 1)$.

¹³ $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ represents a special function known as the error function.

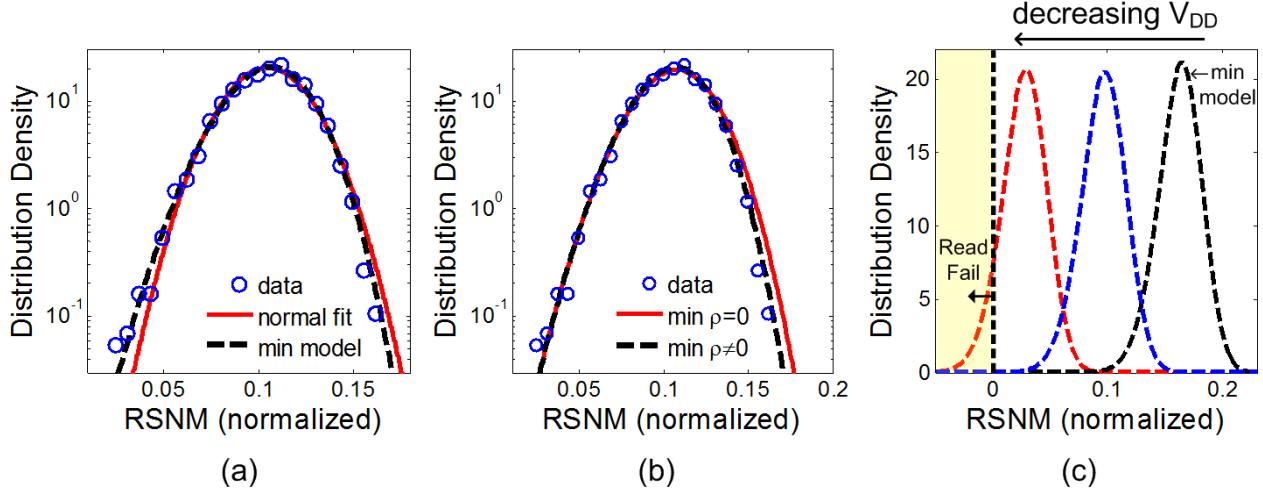


Figure 4.19: Semi-log plot of the distribution density of the actual RSNM - taken as the minimum of two RSNMs - extracted from a 3k-sample MC simulation fitted in (a) using the normal *PDF* and the *PDF* defined by equations 4.3-4.5; and in (b) using the *PDF* defined by equations 4.3-4.5 with $\rho = 0$ and with the extracted ρ value. The worst-case tail matches nicely to the *PDF* defined by equations 4.3-4.5 with and without modeling the ρ . (c) Fitted *PDFs* using equations 4.3-4.5 for three different values of V_{DD} . The probability of read stability failure at each value of V_{DD} is equal to the area under the *PDF* and to the left of the line $y = 0$.

It is worthwhile to note that in the special case where X_1 and X_2 are i.i.d., equations 4.3-4.5 can be simplified as

$$f(x) = 2 \times \frac{1}{\sigma_1} \times \phi\left(\frac{x - \mu_1}{\sigma_1}\right) \times \left[1 - \Phi\left(\frac{x - \mu_1}{\sigma_1}\right)\right] \quad (4.6)$$

by applying the conditions $\mu_1 = \mu_2$, $\sigma_1 = \sigma_2$, and $\rho = 0$. Note that equation 4.6 is equivalent to equation 4.1, which is provided by order statistics for the minimum of two i.i.d. random variables.

To validate the *PDF* defined by equations 4.3-4.5, a semi-log plot is generated, in Figure 4.19a-b, for the distribution density of the actual RSNM - taken as the minimum of two RSNMs for each bitcell - extracted from a 3k-sample MC simulation. The distribution is fitted to the normal *PDF*, the *PDF* defined by equations 4.3-4.5 with the extracted ρ value, and the *PDF* defined by equations 4.3-4.5 with $\rho = 0$. Results show that the *PDF* defined by equations 4.3-4.5 with an extracted ρ value can best match the entire RSNM distribution, whereas the normal *PDF* does not match well to either tail of the RSNM distribution. However, the *PDF* defined by equations 4.3-4.5 with $\rho = 0$ can almost exactly match the *PDF* generated using the extracted ρ value at the worst-case tail. Since the V_{MIN} of large SRAM arrays is determined by the tail of its distribution, a good matching at the worst-case tail of the read/write margin distributions, at each V_{DD} , is sufficient [27]. Therefore, it appears that *PDF* defined by equations 4.3-4.5 with $\rho = 0$ can also be used for SRAM V_{MIN} estimation. Applying the condition $\rho = 0$ to equations 4.3-4.5, the corresponding *PDF* can

be redefined as

$$f(x) = \frac{1}{\sigma_1} \times \phi\left(\frac{x - \mu_1}{\sigma_1}\right) \times \left[1 - \Phi\left(\frac{x - \mu_2}{\sigma_2}\right)\right] + \frac{1}{\sigma_2} \times \phi\left(\frac{x - \mu_2}{\sigma_2}\right) \times \left[1 - \Phi\left(\frac{x - \mu_1}{\sigma_1}\right)\right] \quad (4.7)$$

or equivalently

$$f(x) = f(x_1) \times [1 - F(x_2)] + f(x_2) \times [1 - F(x_1)]. \quad (4.8)$$

It is important to note that, while Figure 4.19b suggests that a good matching to the worst-case tail of the read/write margin distributions can be achieved without accounting for the correlation between the two noise margins of an SRAM bitcell, the utility of equation 4.1 is still limited when a difference is present between the means and/or the variances of the two margins.

With the above definitions, the read/write noise margin (NM) of an SRAM bitcell can be represented by $X_{NM} = \min(X_{NM1}, X_{NM2})$ with its *PDF* defined by either equations 4.3-4.5 (for the general case) or by equation 4.8 (with $\rho = 0$). Although the *CDF* of X_{NM} , in both cases, cannot be easily expressed mathematically, contrary to the case for the minimum of two i.i.d. random variables (equation 4.2)¹⁴, the probability of read stability or writeability failure can be numerically calculated by integrating the *PDF* from $-\infty$ to 0 (which is equivalent to numerically solving for the *CDF* - $F(x)$ - at $x = 0$) - in other words

$$P(FAIL) = P(NM \leq 0) = \int_{-\infty}^0 f(x) dx. \quad (4.9)$$

Using equation 4.9, the probability of either a read stability or a writeability failure can be estimated at each V_{DD} . This is graphically illustrated in Figure 4.19c, where the *PDF* (defined using equations 4.3-4.5) of the RSNM is plotted for three different values of V_{DD} and the probability of read stability failure at each V_{DD} is estimated by the area under the *PDF* and to the left of the line $y = 0$. Because the *CDF* of $X_{NM} = \min(X_{NM1}, X_{NM2})$ for the general case is not easily expressed mathematically (as mentioned above), an expression for the $V_{MIN,RD}/V_{MINWRT}$ is not formulated. However, the values of $V_{MIN,RD}/V_{MINWRT}$ can be easily extracted by looking up the V_{DD} value for the corresponding yield, which can be expressed as $1 - P(FAIL)$. In order to estimate the probability of failure as a function of V_{DD} , equation 4.9 must be solved at each V_{DD} using a different *PDF*, which can be fully defined by a mean and a variance. It is, therefore, sufficient to model the mean (μ) and the standard deviation (σ) of each read/write metric as a function of V_{DD} . This is similar to how HSNM is treated in [147], which adopts the *PDF* defined by equation 4.1. Likewise, in [20], the mean and sigma values are applied to an error function to predict the read/write fail count from a sample of 1k SRAM cells. However, due to the non-Gaussian nature of the actual read/write margin distributions, fail counts predicted directly from an error function may not reflect the actual fail counts of an actual SRAM array, where a failure indicates the inability to retain or write either data polarity.

The above method is first examined against MC simulations, with previously stated conditions, using a commercial low-power 45nm CMOS process for the case of $V_{MIN,RD}$ estimation using both RSNM and SRRV. SRRV is selected, rather than WRRV, because σ_{SRRV}

¹⁴Although a simple mathematical expression is derived for the *CDF* of the minimum of two i.i.d. normal random variables, the solution of this expression still requires a numerical calculation of the error function. However, the solution can also be estimated using the error function's Taylor series.

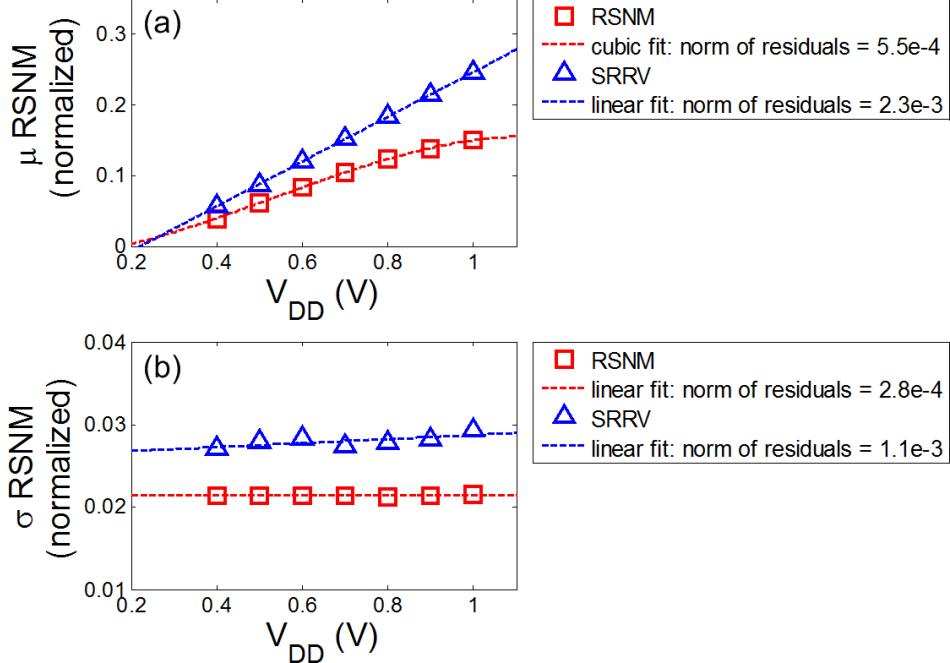


Figure 4.20: The simulated (a) μ and (b) σ for RSNM and SRRV, using 3k-sample MC simulations, as a function of V_{DD} along with the corresponding polynomial fit and the norm of the residuals.

is shown to be relatively supply-independent and can be easily approximated using a linear fit (Section 4.2.3). In addition, μ_{SRRV} is approximately a linear function of V_{DD} . Figure 4.20 plots the simulated μ and σ for RSNM and SRRV, using 3k-sample MC simulations, as a function of V_{DD} . μ_{RSNM} is fitted using a 3rd order polynomial, while σ_{RSNM} , μ_{SRRV} , and σ_{SRRV} are fitted linearly. The negative correlation between RSNM1 and RSNM2, measured for each SRAM bitcell, is graphically illustrated in Figure 4.21a (similar to Figure 4.18a). Figure 4.21b plots the extracted ρ value between RSNM1 and RSNM2 as a function of V_{DD} along with its quadratic fit. ρ_{RSNM} is fitted to further compare the accuracy of $V_{MIN,RD}$ estimation when generating the *PDF* using equations 4.3-4.5 versus using equation 4.8.

Recall from Sections 2.3.1 and 4.2 that, when the within-cell mismatch is high, the SRRV can only be characterized for a single data polarity. Therefore, although a negative correlation is expected between the two read margins of each bitcell due to a strong sensitivity to differential variations in the pull-down transistors (Figure 4.17a), the SRRV can only be extracted for the less read-stable data polarity - i.e. only the smaller read margin can be extracted. As a result, ρ_{SRRV} can only be extracted for bitcells with a marginal within-cell mismatch, for which both SRRV1 and SRRV2 can be characterized. Since the marginal within-cell mismatch is modulated by a systematic variation, a slightly positive ρ_{SRRV} is extracted - this is graphically illustrated in Figure 4.22a. Figure 4.22b presents the semi-log plot of the distribution density of the SRRV extracted from a 5k-sample MC simulation¹⁵. The worst-case tail of the distribution matches very well against the *PDFs* generated using

¹⁵A 5k-sample MC simulation is used here to extract adequate samples of both SRRV1 and SRRV2 for accurate ρ extraction.

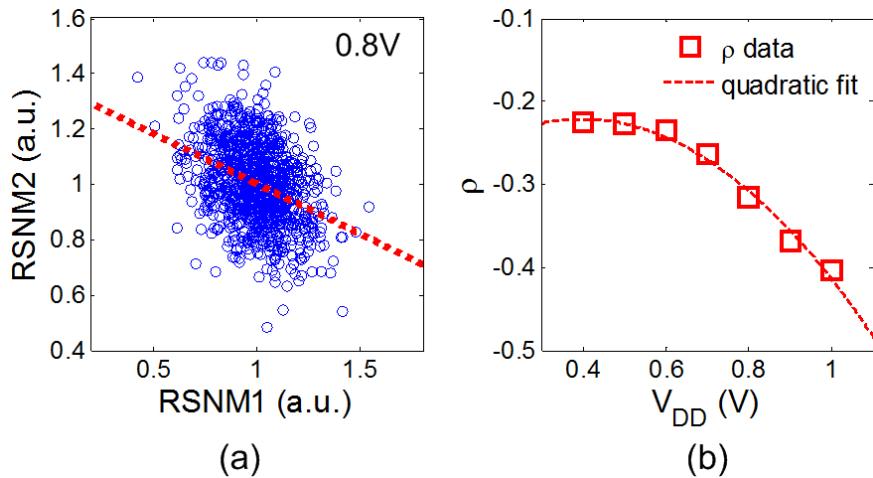


Figure 4.21: (a) The scatter plot for RSNM2 versus RSNM1, along with a linear fit, showing a negative correlation. Here, $V_{DD} = 0.8V$ is selected without any particular reason. (b) The coefficient of correlation, ρ , between RSNM1 and RSNM2 as a function of V_{DD} along with the quadratic fit.

both equations 4.3-4.5 and equation 4.8; although the *PDF* generated using equations 4.3-4.5 does present a better overall match to the distribution¹⁶.

Figure 4.23a-b plots the read fail probability as a function of V_{DD} , estimated using RSNM, SRRV, and SVNM. The SRAM $V_{MIN,RD}$ is plotted, in Figure 4.23c, against the number of functional (or error-free) SRAM cells¹⁷, expressed in units of σ - using the transformation

$$\sigma = \sqrt{2} \times \text{erf}^{-1}(1 - P(\text{FAIL})) \quad (4.10)$$

where $erf^{-1}(\cdot)$ denotes the inverse error function. The estimates are compared against a 100k-sample MC simulation. Results indicate that the estimation using RSNM and SRRV can achieve excellent agreement to the MC simulation - with less than 0.7% error at 3σ , 0.5% error at 4σ , and 1.5% error at 4.42σ - while offering a $5\times$ speedup compared to a 100k-sample MC simulation. This speedup can be increased by fitting the μ and σ values using less V_{DD} points - in this example, 3k-sample MC simulations are run at 7 V_{DD} points (from 0.4V to 1.0V). Since μ_{RSNM} is fit to a 3rd order polynomial, a minimum of 4 data points are needed to generate a unique fitting; whereas only 2 data points are needed to fit μ_{SRRV} . In addition, the speedup is more significant when comparing to a larger-sample MC simulation. It is important to note that the accuracy of this method does have a dependence on the normality of the individual noise margin distributions, since the *PDFs* are defined

¹⁶ Although a negative correlation is expected between the two read margins of each bitcell, a better match is established against the *PDF* generated using equations 4.3-4.5 with $\rho_{SRRV} > 0$ because the missing SRRV data that leads to $\rho_{SRRV} > 0$ is also missing from the distribution itself - therefore, a positive correlation is present between the distributions for SRRV1 and SRRV2. However, the value of ρ_{SRRV} may not measure this correlation exactly - also due to the missing SRRV data.

¹⁷This is similar to what is done for DRV in [147]. Here, the expression for the general case *PDF* is used - with equations 4.3-4.5 - to estimate $V_{MIN,RD}$ using both conventional and large-scale SRAM metrics.

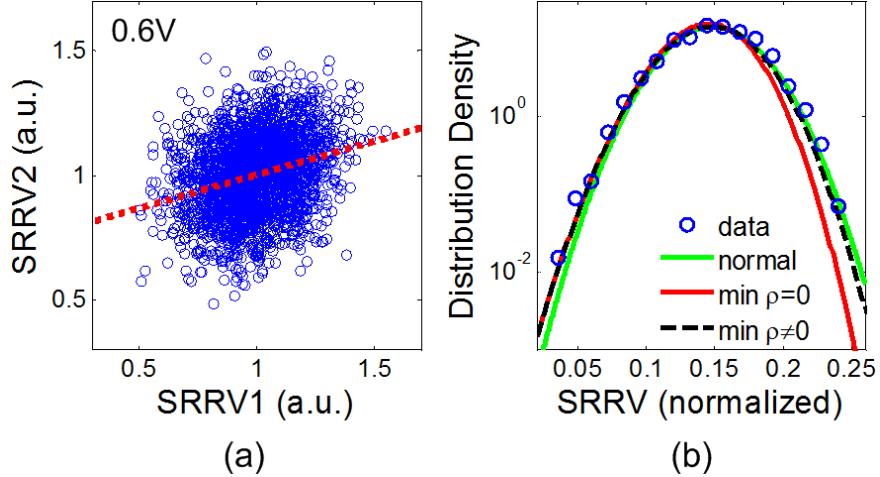


Figure 4.22: (a) The scatter plot for SRRV2 versus SRRV1 with a linear fit, for SRAM cells allowing the characterization of SRRV for both data polarities, showing a positive correlation. (b) Semi-log plot of the distribution density of SRRV extracted from a 5k-sample MC simulation fitted using the normal *PDF* and the *PDFs* defined by equations 4.3-4.5 and by equation 4.8. The worst-case tail matches nicely to the *PDFs* defined by either equations 4.3-4.5 or equation 4.8. A 5k-sample MC simulation is used to extract adequate samples of both SRRV1 and SRRV2 for accurate ρ extraction.

for the minimum of two normal random variables¹⁸.

Figure 4.23b-c indicates that the estimation using SVNM, on the other hand, matches poorly against the MC simulation. Recall from Section 2.2.1 that the SVNM is unable to characterize negative read margin values. To isolate this effect, the estimation using SVNM is performed twice - with μ and σ fitted within a full range of V_{DD} values (e.g. 0.4V to 1.0V) in (1) and within only the higher V_{DD} values (e.g. 0.7V to 1.0V) in (2), to avoid having to characterize a negative read margin. However, only a marginal improvement is achieved using (2) - as shown in Figure 4.23b-c. To investigate this further, the distributions of RSNM, SRRV, SVNM, and SINM for a single data polarity is plotted in Figure 4.24 (using 3k-sample MC simulations). Each metric is characterized at three different supply voltages to examine how its distribution changes with V_{DD} . Figure 4.24 shows that the distributions of all four metrics shift to the left, as expected, when V_{DD} is decreased - i.e. the read margins decrease with a decreasing V_{DD} . However, while the shapes of the RSNM and the SRRV distributions remain unchanged - indicating a shift in μ with a constant σ , the distributions of both SVNM and SINM are altered as V_{DD} is reduced. This is not surprising for SINM, whose distribution, as mentioned in Section 2.2.1, is expected to become log-normal at low V_{DD} s - making it unsuitable for $V_{MIN,RD}$ estimation using the above method. In the case of SVNM, both μ and σ changes as V_{DD} is reduced. In addition, the SVNM distribution, as illustrated in Figure 4.24b, deviates from Gaussian and becomes slightly right skewed at low V_{DD} values (e.g 0.4V) - this is likely caused by the inability to characterize a negative SVNM. Therefore, although Section 2.2.1 indicates that the SVNM is able to effectively track the

¹⁸ [147] shows a matching accuracy, using the *PDF* defined by equation 4.1, out to $7 - 8\sigma$ against the statistical method provided in [127, 128] for the estimation of SRAM DRV.

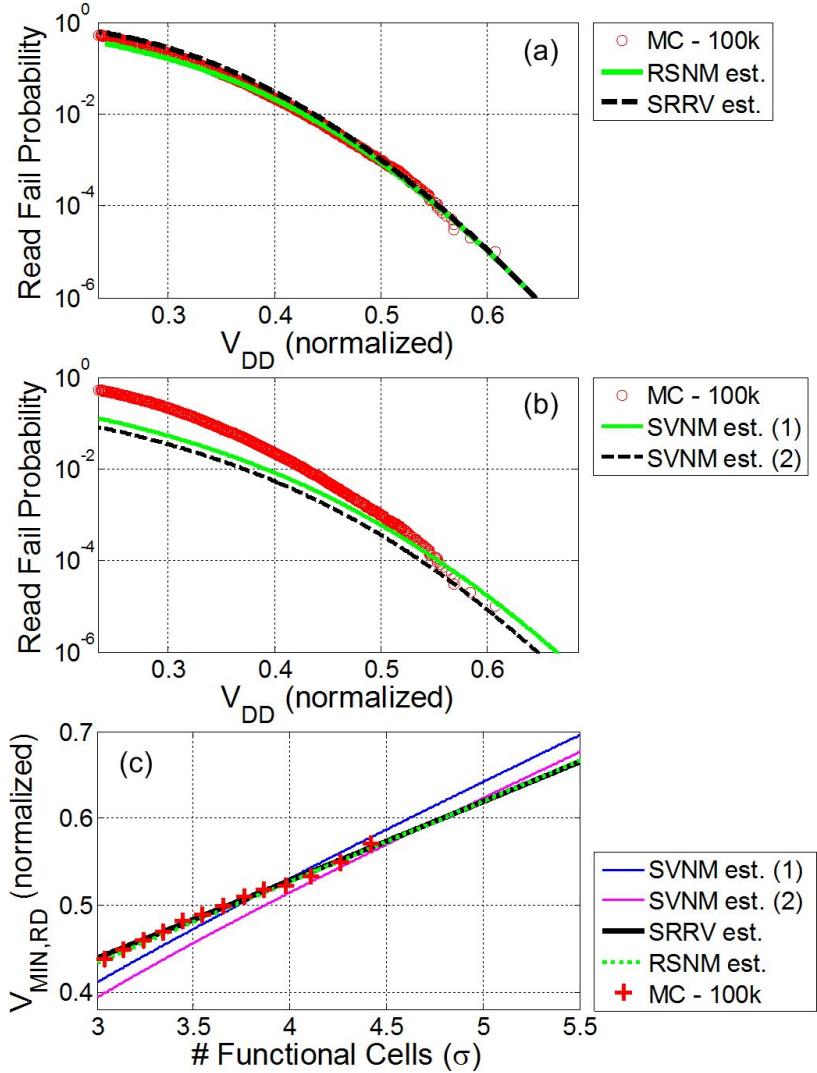


Figure 4.23: Semi-log plot for the read fail probability as a function of V_{DD} , both extracted from a 100k-sample MC simulation and estimated (a) using RSNM and SRRV, and (b) using SVNM. (c) $V_{MIN,RD}$ as a function of the number of functional SRAM cells, in units of σ , extracted from a 100k-sample MC simulation and estimated using RSNM, SRRV, and SVNM. The estimations using RSNM and SRRV matches very well against the results from MC, whereas the estimation using SVNM does not. The estimation using SVNM is done twice - with μ and σ fitted using (1) a full range of V_{DD} values and (2) only higher V_{DD} values.

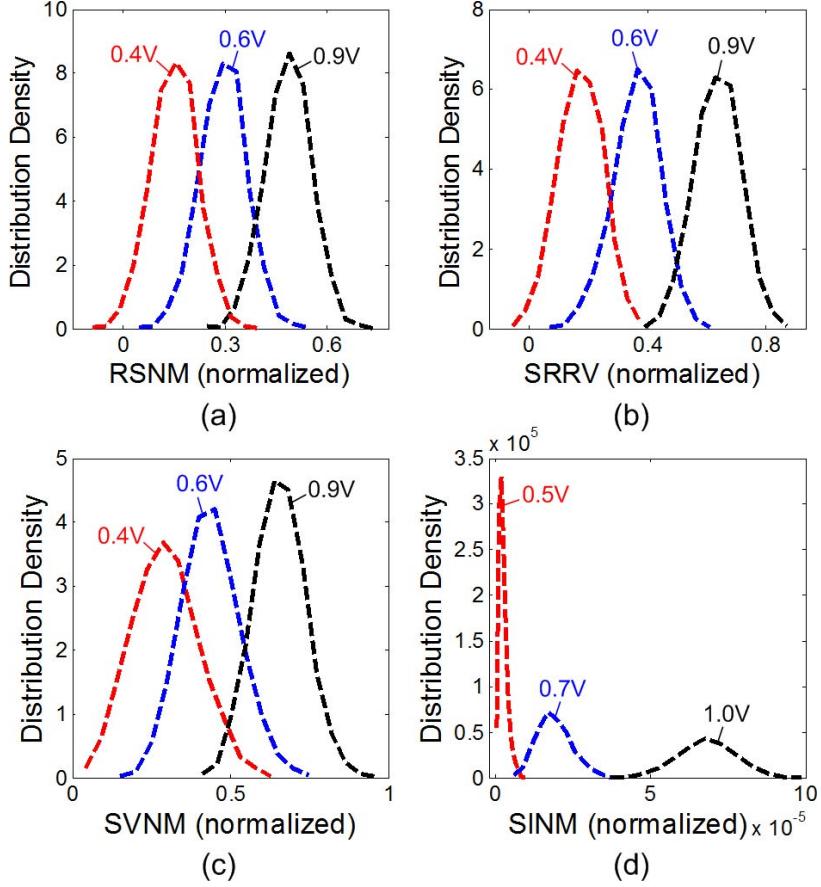


Figure 4.24: Distribution densities, at three different supply voltages, of (a) RSNM, (b) SRRV, (c) SVNM, and (d) SINM extracted from 3k-sample MC simulations for a single data polarity. The distributions of both SVNM and SINM become non-Gaussian as V_{DD} is reduced.

SRAM $V_{MIN,RD}$, its utility in $V_{MIN,RD}$ estimation is limited.

Before digging further, Figure 4.25 is presented to complement Figure 4.24 for the write metrics. This figure shows that, similar to RSNM and SRRV, the distributions of WNM and WWTW remain Gaussian down to very low supply voltages (i.e. 0.4V and beyond). Furthermore, their σ values remain relatively unchanged - i.e. neither the heights nor the widths of the distributions change with a decreasing V_{DD} . Consequently, WNM and WWTW, which were previously shown to effectively track the SRAM $V_{MIN,WRT}$ (Chapter 2), are good candidates for $V_{MIN,WRT}$ estimation using the above method¹⁹. Conversely, as discussed in Section 2.2.2, the distribution of I_W deviates from Gaussian and becomes log-normal as V_{DD} is decreased - making it unsuitable for $V_{MIN,WRT}$ estimation using the above method.

Figure 4.26 plots the read fail probability as a function of V_{DD} as well as $V_{MIN,RD}$

¹⁹It is important to recall that WNM characterization may require adequate sweep margins to expose convexity in the write-VTC. As a result, WNM characterization may not be possible at higher values of V_{DD} .

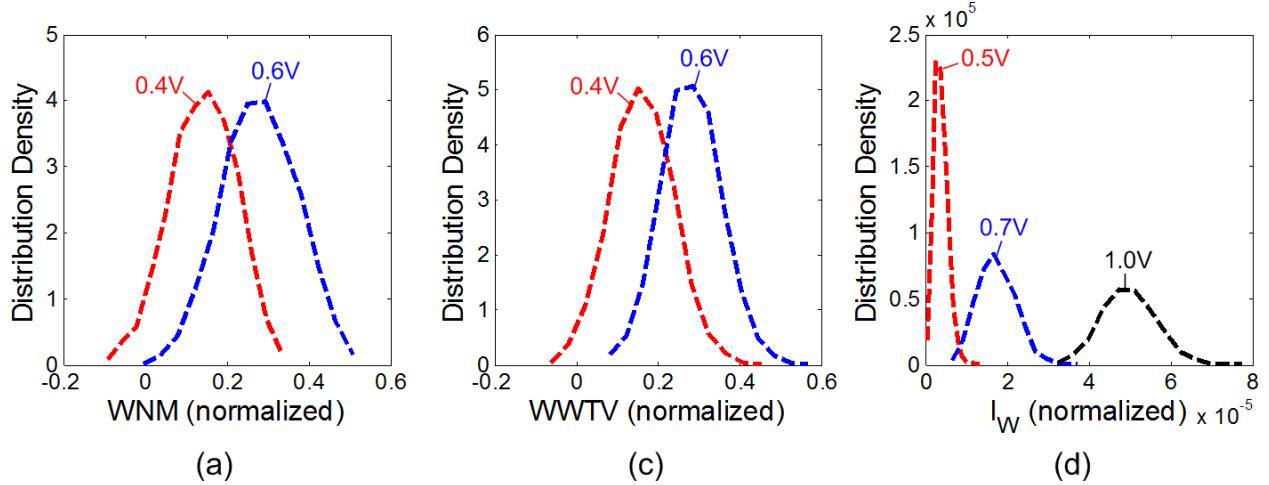


Figure 4.25: Distribution densities of (a) WNM, (b) WWTV, and (c) I_W extracted from 3k-sample MC simulations for a single data polarity. The distribution of I_W become non-Gaussian as V_{DD} is reduced.

as a function of the number of functional SRAM cells (in units of σ). The presented results are estimated using RSNM with μ fitted either linearly or to a 3rd order polynomial. The comparison justifies the usage of a 3rd order polynomial fit for μ_{RSNM} by showing that it can indeed achieve more accuracy than a linear fit.

As previously mentioned, the worst-case tail of the read/write margin distributions can be sufficiently matched without accounting for the correlation between the two noise margins of an SRAM bitcell. To further validate this, the read fail probability as a function of V_{DD} is estimated using RSNM, with its *PDF* modeled by either equations 4.3-4.5 or equation 4.8. Figure 4.27a indicates that both sets of equations produce the same read fail probabilities down to a very low supply voltage ($\sim 0.3V$). At $V_{DD} \sim 0.3V$, the estimated fail probabilities start to deviate slightly as the bodies of the *PDFs* approach the $y = 0$ line and start to impact the read fail probability (Figure 4.19c). Figure 4.27b, which plots $V_{MIN,RD}$ against the number of functional bitcells, indicates that both sets of equations produce the same $V_{MIN,RD}$ estimates from 2σ onwards. Therefore, for large SRAM arrays, equation 4.8 is sufficient for failure analysis.

Although V_{MIN} estimation can be performed without accounting for ρ , the utility of equation 4.1, which assumes i.i.d., may still be limited when a difference exists between the means and/or the variances of the two noise margins. While the RSNM1 (SRRV1) and RSNM2 (SRRV2) data gathered from MC simulations at the different supply voltages appear to be identically distributed, a closer examination of the statistics reveals an average difference of $\sim 2.3\%$ between σ_{RSNM1} and σ_{RSNM2} over the range $V_{DD} = 0.4V - V_{DD} = 1.0V$ due to random mismatch. This difference is slightly more pronounced between σ_{SRRV1} and σ_{SRRV2} - reaching an average difference, in the same direction, of $\sim 3.5\%$ - likely due to the fact that the SRRV characterization is discrete - determined by the stepping

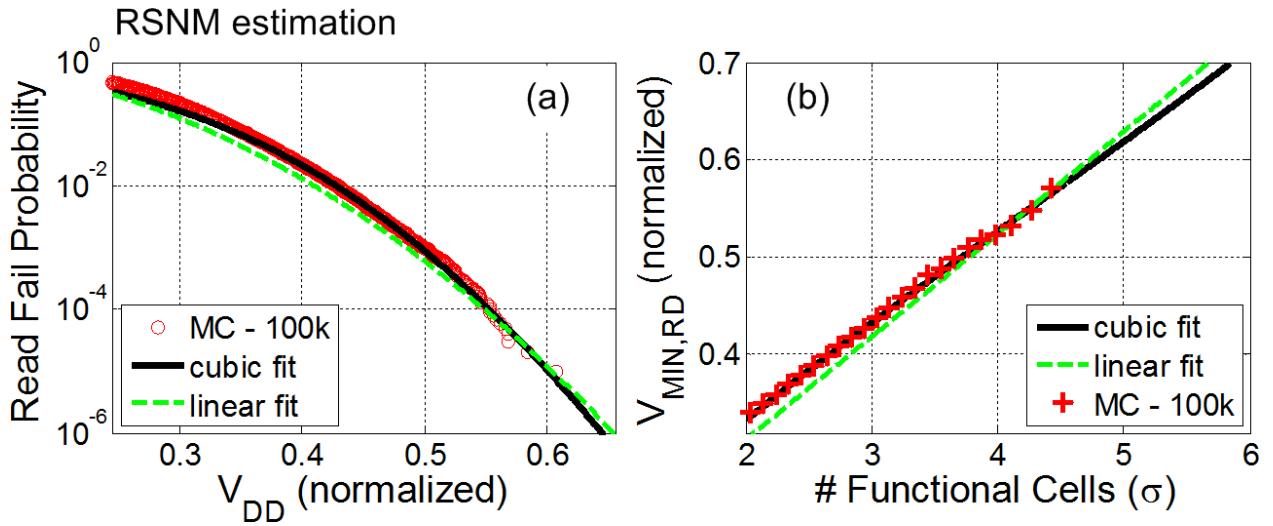


Figure 4.26: (a) Semi-log plot for the read fail probability as a function of V_{DD} ; and (b) $V_{MIN,RD}$ as a function of the number of functional SRAM cells, in units of σ - both extracted from a 100k-sample MC simulation and estimated using RSNM, with μ fitted either linearly or to a 3rd order polynomial.

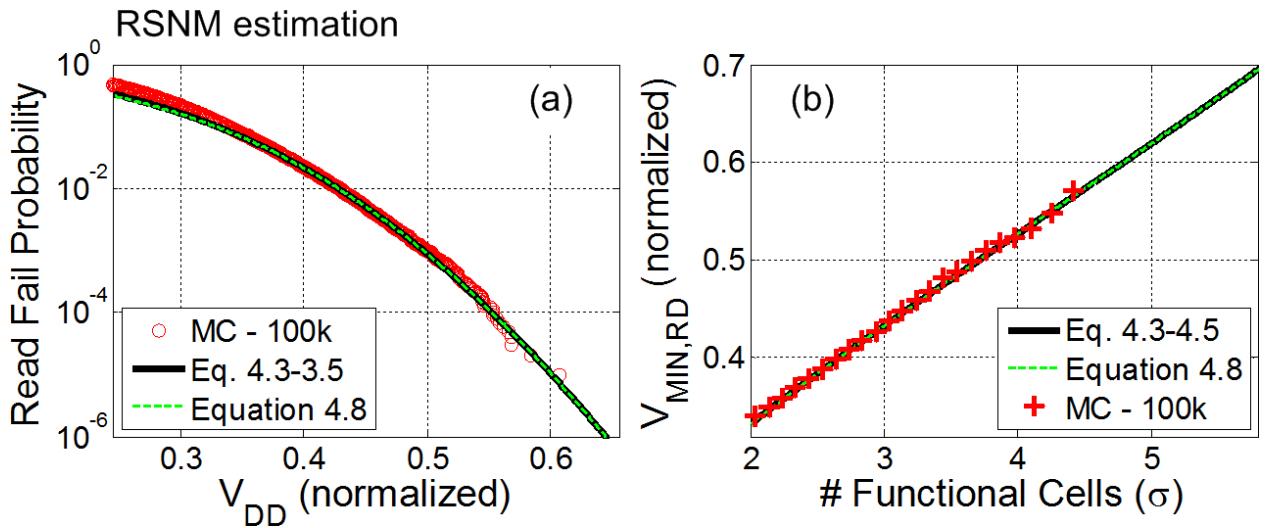


Figure 4.27: (a) Semi-log plot for the read fail probability as a function of V_{DD} ; and (b) $V_{MIN,RD}$ as a function of the number of functional SRAM cells, in units of σ - both extracted from a 100k-sample MC simulation and estimated using RSNM, with its PDF modeled by either equations 4.3-4.5 or equation 4.8.

size²⁰, whereas the RSNM extraction is continuous. In the meantime, the μ values for both RSNM and SRRV, although matched to within 1% difference, shift together in the opposite direction (i.e. if $\sigma_1 > \sigma_2$, then $\mu_1 < \mu_2$) - enhancing the effect of the difference in the σ values. Figure 4.28 summarizes the impact of this small σ difference on the accuracy of

²⁰A stepping size of 1mV is used in the MC simulations.

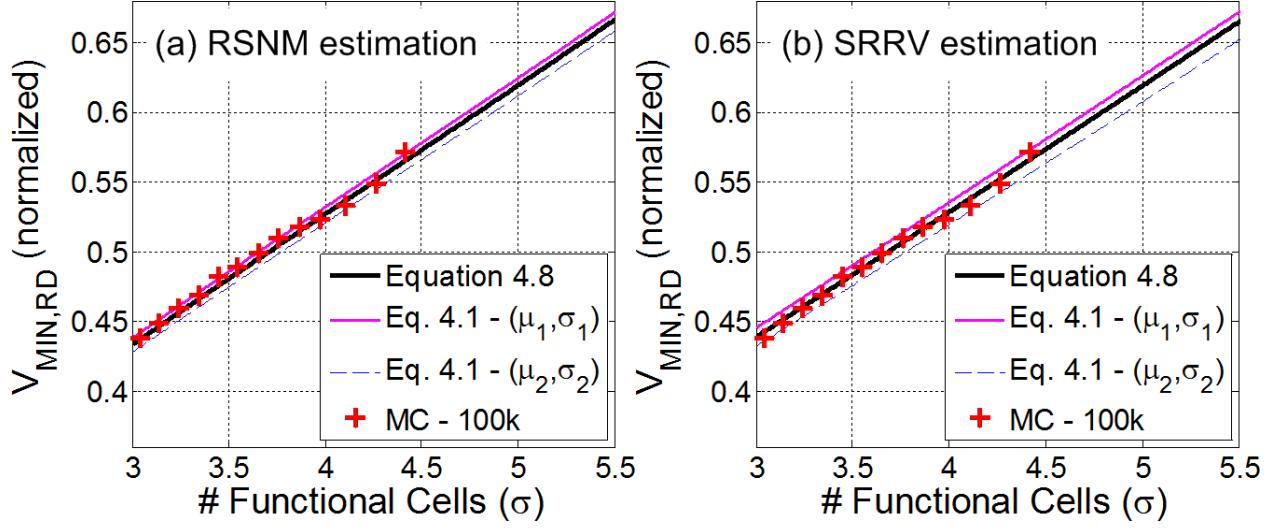


Figure 4.28: $V_{MIN,RD}$ as a function of the number of functional SRAM cells, in units of σ , both extracted from a 100k-sample MC simulation and estimated using - (a) RSNM with its *PDF* modeled by either equation 4.8 or equation 4.1; and (b) SRRV with its *PDF* modeled by either equation 4.8 or equation 4.1.

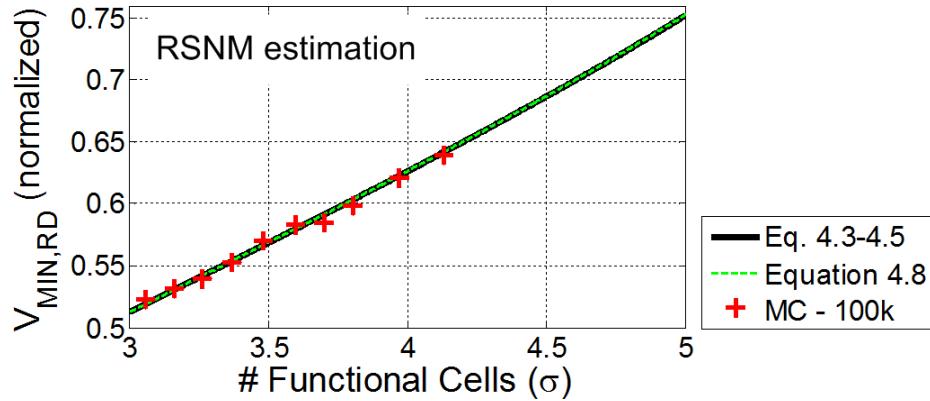


Figure 4.29: $V_{MIN,RD}$ as a function of the number of functional SRAM cells, in units of σ , both extracted from a 100k-sample MC simulation and estimated using RSNM, with its *PDF* modeled by either equations 4.3-4.5 or equation 4.8. A systematic mismatch is introduced to the bitcell through a differential adjustment in the L_G of the pull-down transistor pair - producing a $\sim 8\%$ shift in μ and a $\sim 6\%$ shift in σ between SRRV1 and SRRV2.

$V_{MIN,RD}$ estimation using equation 4.1 versus using equation 4.8. The results from both RSNM- and SRRV-based estimations are presented. While both the RSNM- and SRRV-based $V_{MIN,RD}$ estimations using equation 4.8 achieve excellent matching against the MC simulation (same as shown in Figure 4.23), the estimations using equation 4.1 - either taking (μ_1, σ_1) or (μ_2, σ_2) - deviates slightly from the results of the MC simulation as well as the estimations using a more complete definition of the *PDF* - i.e. equation 4.8. Therefore, even with small differences between (μ_1, σ_1) and (μ_2, σ_2) , equation 4.8 can provide a more accurate estimation than equation 4.1.

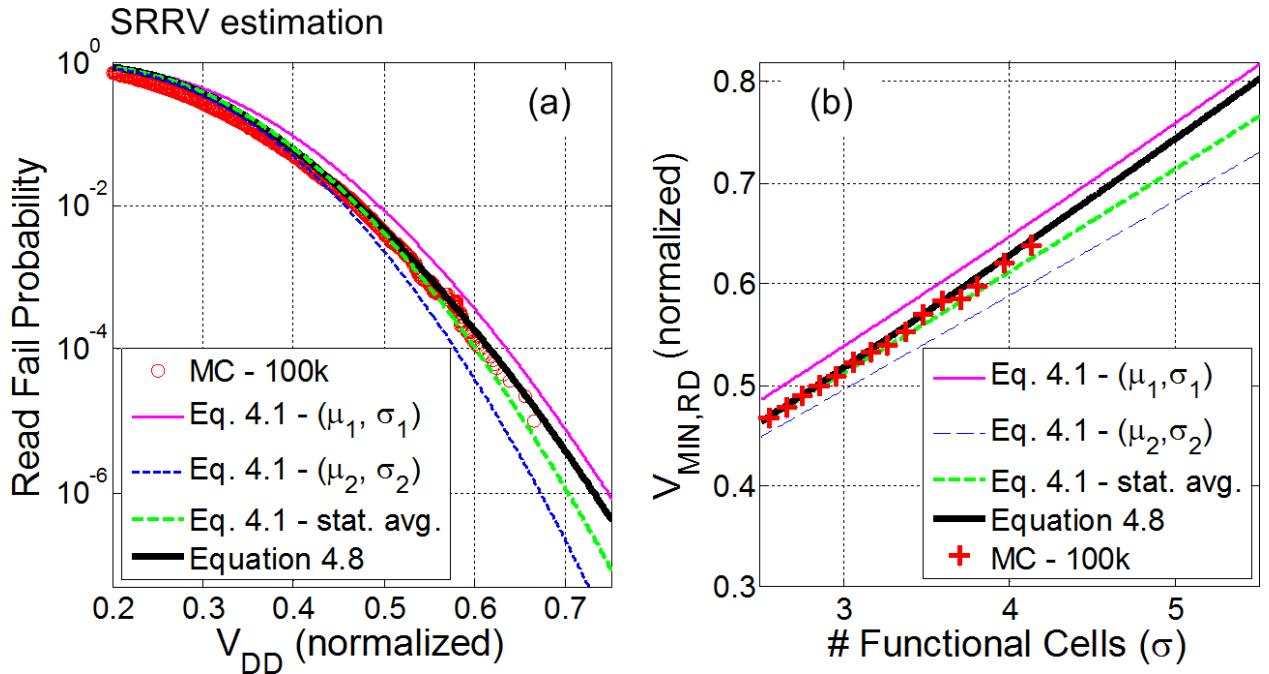


Figure 4.30: (a) Semi-log plot for the read fail probability as a function of V_{DD} ; and (b) $V_{MIN,RD}$ as a function of the number of functional SRAM cells, in units of σ - both extracted from a 100k-sample MC simulation and estimated using SRRV, with its PDF modeled by either equation 4.8 or equation 4.1. A systematic mismatch is introduced to the bitcell through a differential adjustment in the L_G of the pull-down transistor pair - producing a $\sim 8\%$ shift in μ and a $\sim 6\%$ shift in σ between SRRV1 and SRRV2.

While the error in failure estimation using equation 4.1 is only marginal in Figure 4.28, this error is expected to increase in the presence of a systematic within-cell mismatch that produces a difference in the means and/or the variances of the two noise margins. To examine this, an intentional systematic mismatch is introduced to the SRAM bitcell through a differential adjustment in the L_G of the pull-down transistor pair, producing approximately an 8% shift in μ and a 6% shift in σ (in opposite directions) of the two noise margins within the bitcell - characterized using SRRV²¹. Before examining the impact of this systematic mismatch on the accuracy of failure prediction using equation 4.1, the nearly exact matching between $V_{MIN,RD}$ estimates using equations 4.3-4.5 versus using equation 4.8, in the presence of a systematic within-cell mismatch, is reproduced in Figure 4.29. Therefore, the impact of the correlation between the two noise margins is independent of whether the two noise margins are identically distributed. Figure 4.30 plots the read fail probability as a function of V_{DD} as well as $V_{MIN,RD}$ against the number of functional SRAM cells, in units of σ , extracted from a 100k-sample MC simulation (for the bitcell mentioned above). The SRRV-based estimation is performed using equation 4.8 and equation 4.1 for both (μ_1, σ_1) and (μ_2, σ_2) . In addition, the statistical averages of μ and σ - i.e. $\mu_{AVG} = (\mu_1 + \mu_2)/2$ and $\sigma_{AVG} = \sqrt{(\sigma_1^2 + \sigma_2^2)/2}$, assuming independence - is used with equation 4.1 for estimation.

²¹The percentages in μ and σ shifts are averaged over the range $V_{DD} = 0.4V - V_{DD} = 1.0V$.

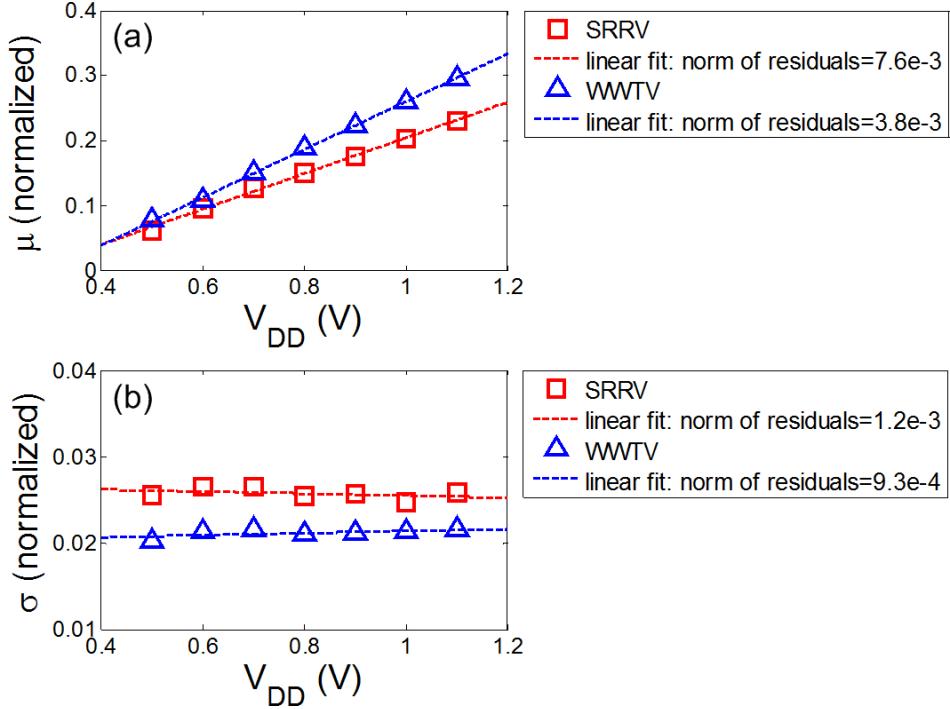


Figure 4.31: The (a) μ and (b) σ of SRRV and WWTW, measured for 8k bitcells from a 64kb SRAM sub-array, as a function of V_{DD} along with the corresponding linear fit and the norm of the residuals.

As expected, the estimation using equation 4.8 achieves excellent matching against the 100k-sample MC simulation. Conversely, the estimation using equation 4.1 either overestimates failure - with (μ_1, σ_1) - or underestimates failure - with (μ_2, σ_2) . While the estimation using equation 4.1 with the statistical averages of μ and σ does provide a better estimation than either (μ_1, σ_1) or (μ_2, σ_2) , Figure 4.30b indicates that its results start to deviate significantly from the estimation using equation 4.8 at $>\sim 4\text{-}5\sigma$. Therefore, while the accuracy of failure predication using equation 4.1 can be acceptable (Figure 4.28) for SRAM arrays subject to a purely random within-cell mismatch, its utility becomes noticeably limited when a systematic within-cell mismatch is also present.

Measured Results

As processes become increasingly complex and harder to control, simulation-based failure analysis can no longer be relied upon for large cache memories and the direct analysis of silicon data becomes crucial. Figure 4.31 plots the μ and σ of SRRV and WWTW, measured for 8k bitcells from a 64kb SRAM sub-array, as a function of V_{DD} . All four parameters match nicely to a linear fit, where both σ_{SRRV} and σ_{WWTW} are approximately supply-independent. The SRRV distributions, as the minimum of SRRV1 and SRRV2, is

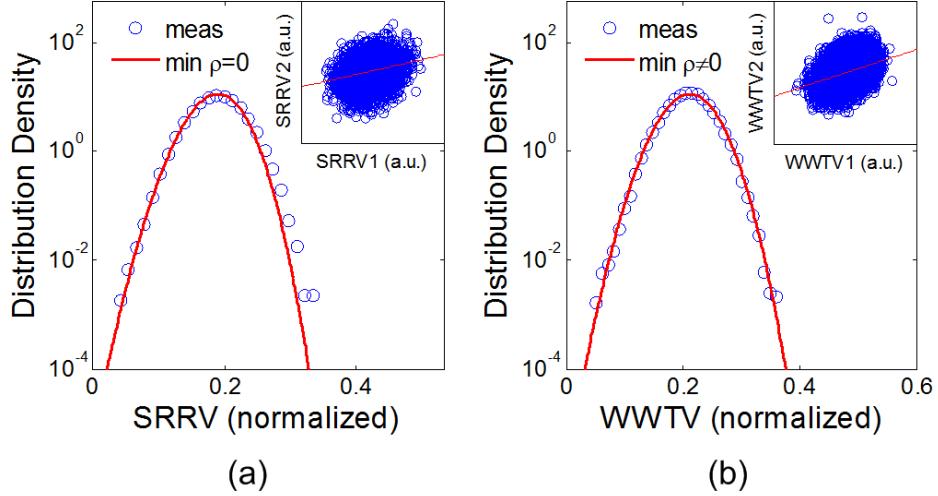


Figure 4.32: Semi-log plot of the distribution densities of (a) SRRV, modeled using equation 4.8; and (b) WWTW, modeled using equations 4.3-4.5. Both SRRV and WWTW, in this example, are measured for a 64kb SRAM sub-array. The worst-case tail matches well in both cases. The scatter plots of SRRV2 versus SRRV1 and WWTW2 versus WWTW1 are included to show the positive correlation from measurement.

modeled using equation 4.8 with $\rho = 0^{22}$; the WWTW distributions, as the minimum of WWTW1 and WWTW2, is modeled, in contrast, using equations 4.5-4.1. Figure 4.32 plots the distribution densities of SRRV and WWTW, along with the models for their *PDF*, measured for a 64kb SRAM sub-array. The measured results indicate excellent matching of the worst-case tails and, in the case of WWTW (which uses the exact *PDF* from equations 4.5-4.1), over the entire distribution for 65,536 data points. The probabilities of read stability failure and write stability failure are estimated using SRRV and WWTW measurements, modeled by equations 4.5-4.1 and equation 4.8, respectively. The estimated results are compared against measurements from a 64kb SRAM sub-array and is summarized in Figure 4.33. Good matching is achieved between the estimated $V_{MIN,RD}/V_{MIN,WRT}$ values and the measurement data - with a maximum error of less than 6% from 2σ to $> 4\sigma$.

4.4 Read Current Measurements

To gauge the read performance of bitcells in functional SRAM arrays, the per-cell read current (I_{READ}) is measured through the direct bit-line access scheme. Figure 4.34 plots the measured $\sigma_{I_{READ}}/\mu_{I_{READ}}$, from four separate 64kb SRAM sub-arrays at $V_{DD} = 1.1V$, as a function of $1/\sqrt{W \times L}$ for three different SRAM bitcell designs with cell areas of $0.374 \mu\text{m}^2$, $0.299 \mu\text{m}^2$, and $0.252 \mu\text{m}^2$ - where W and L represent the nominal drawn width (W_{DRAWN}) and drawn L_G (L_{DRAWN}) of the pass-gate transistor for each bitcell design. According to Pelgrom's model [113], σ_{VT_H} is inversely proportional to the square root of the transistor

²²Recall that the SRRV can only be characterized for one data polarity in bitcells with elevated within-cell mismatch; therefore, a positive correlation is measured (Figure 4.32a) even though the correlation between the two read margins in a bitcell is expected to be negative.

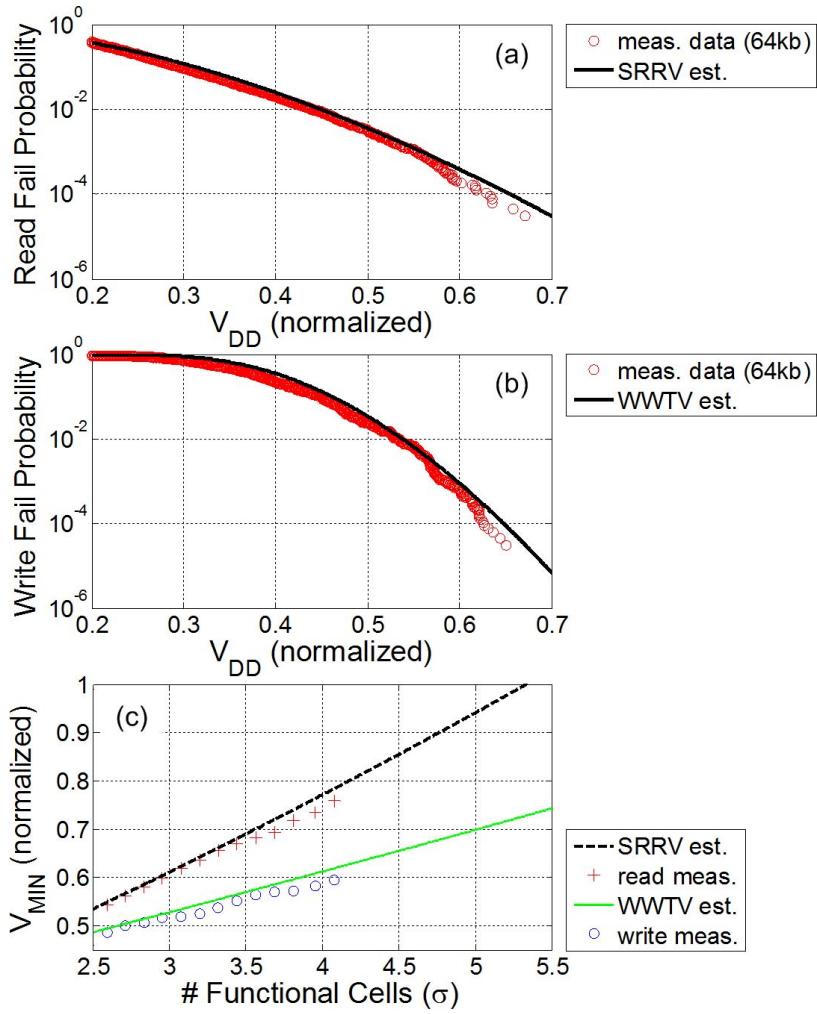


Figure 4.33: Semi-log plot for (a) the read fail probability as a function of V_{DD} , measured for a 64kb SRAM sub-array and estimated using SRRV; and (b) the write fail probability as a function of V_{DD} , measured for a 64kb SRAM sub-array and estimated using WWTV. (c) $V_{MIN,RD}/V_{MIN,WRT}$ as a function of the number of functional SRAM cells, in units of σ - both measured and estimated. To reduce writeability and expose higher $V_{MIN,WRT}$ values, NMOS RBB and PMOS FBB are applied during $V_{MIN,WRT}$ and WWTV measurements - word-line weak write is not applied because WWTV characterization requires direct word-line control. The data in (c) are so normalized to display both $V_{MIN,RD}$ and $V_{MIN,WRT}$ results in the same graph. Results indicate good matching between the estimated values and the measurement data.

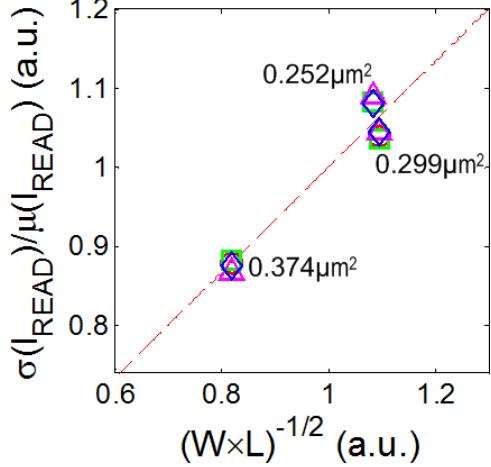


Figure 4.34: $\sigma_{I_{READ}}/\mu_{I_{READ}}$, measured from four separate 64kb SRAM sub-arrays, as a function of $1/\sqrt{W \times L}$ for three different SRAM bitcell designs with cell areas of $0.374 \mu\text{m}^2$, $0.299 \mu\text{m}^2$, and $0.252 \mu\text{m}^2$.

channel area $W \times L$ - i.e. $\sigma_{VTH} \propto 1/\sqrt{W \times L}$. At $V_{DD} = 1.1V$, I_{READ} is dominated by the pass-gate transistor current [52] and has a nearly linear dependence on the V_{TH} of the pass-gate transistor, operating in the velocity saturation mode, and the pull-down transistor, operating in the linear mode. Since each SRAM bitcell design offers a unique pass-gate transistor size, the measured $\sigma_{I_{READ}}/\mu_{I_{READ}}$ should demonstrate an inverse proportionality to the square root the pass-gate transistor channel area $W \times L$ in accordance to Pelgrom's model. It should be noted that while the $0.252 \mu\text{m}^2$ bitcell design has a smaller cell area, due to a more aggressive scaling of the design rules, its pass-gate transistors are sized slightly larger than the pass-gate transistors of the $0.299 \mu\text{m}^2$ bitcell design. Figure 4.34 indicates that while the measured data for both the $0.374 \mu\text{m}^2$ - $0.299 \mu\text{m}^2$ pair and the $0.374 \mu\text{m}^2$ - $0.252 \mu\text{m}^2$ pair exhibit an inverse proportionality, the measured data for the $0.299 \mu\text{m}^2$ - $0.252 \mu\text{m}^2$ pair do not. This suggests that either the effective width (W_{EFF}) and the effective L_G (L_{EFF}) of the pass-gate transistors for the $0.252 \mu\text{m}^2$ bitcells are smaller than for the $0.299 \mu\text{m}^2$ bitcells²³; or local random variations are dominated by other effects, such as line edge roughness (LER) and gate oxide interface roughness [112] rather than by random dopant fluctuation (RDF); or a combination of both.

Figure 4.35 presents the normal probability plot for I_{READ} measured at $V_{DD} = 1.1V$, $V_{DD} = 0.7V$, and $V_{DD} = 0.5V$ from a 64kb SRAM sub-array. As previously mentioned, I_{READ} has a nearly linear dependence on the V_{TH} of the pass-gate transistor and the pull-down transistor at $V_{DD} = 1.1V$. Consequently, the measured I_{READ} at $V_{DD} = 1.1V$ exhibits good normality up to more than $\pm 4\sigma$ [52]. As the supply voltage is dropped to $0.7V$, the pass-gate transistors of certain bitcells are no longer velocity saturated; simultaneously, the pull-down transistors of a fraction of the bitcells in the sub-array enter the saturation mode, due to a voltage rise at the '0' storage node coupled with the high transistor V_{TH} in this process. As a result, I_{READ} no longer is linearly dependent on the V_{TH} of either the pass-gate

²³i.e. the difference between W_{DRAWN}/L_{DRAWN} and W_{EFF}/L_{EFF} is higher for the $0.252 \mu\text{m}^2$ bitcell, possibly a result of its more aggressively scaled design rules.

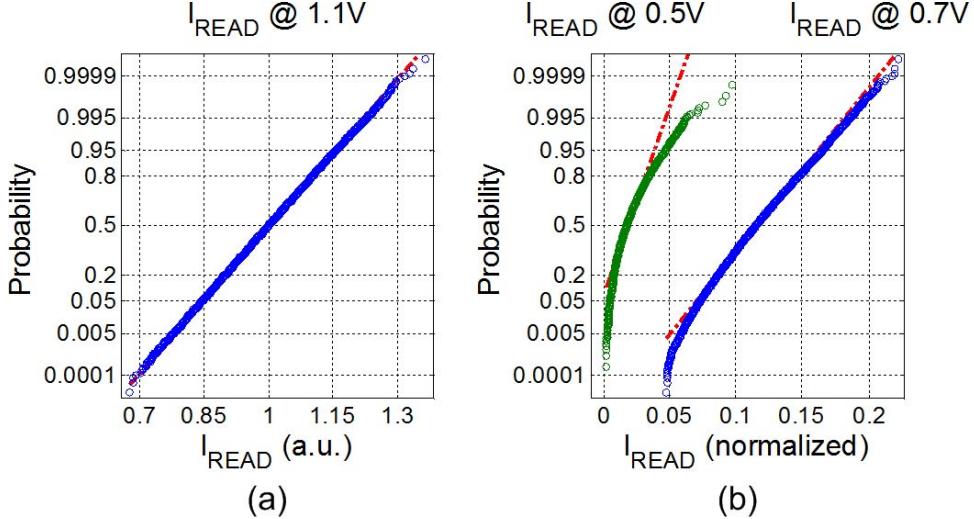


Figure 4.35: Normal probability plots for (a) I_{READ} measured at $V_{DD} = 1.1V$ and (b) I_{READ} measured at $V_{DD} = 0.7V$ and $V_{DD} = 0.5V$. The data in (b) is normalized to the mean of I_{READ} measured at $V_{DD} = 1.1V$.

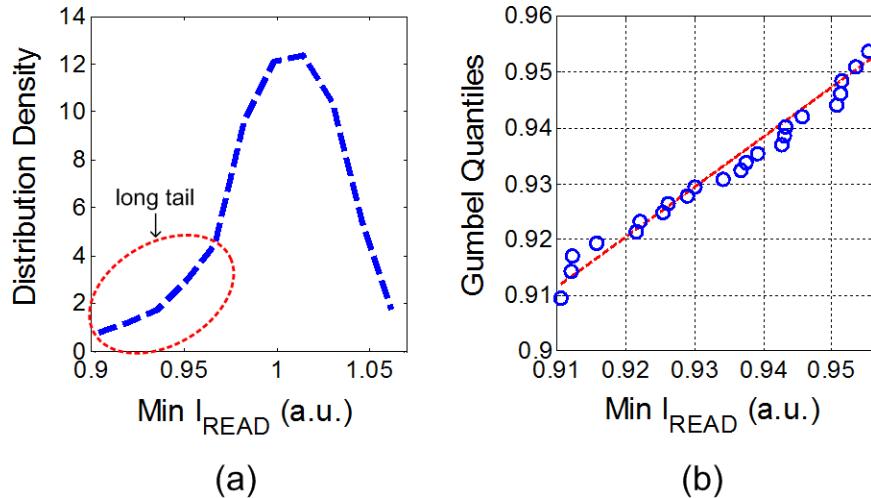


Figure 4.36: (a) The distribution density of the measured minimum I_{READ} over a 1kb SRAM block showing a long tail to the left. (b) Gumbel probability plot for the lower tail of the distribution for the measured minimum I_{READ} over a 1kb SRAM block.

transistors or the pull-down transistors. Thus, the lower tail of the I_{READ} distribution at 0.7V exhibits a bending downwards and to the left (Figure 4.35b), indicating a left skewed distribution. As the supply voltage is further reduced to 0.5V, the transistors within the bitcells approach the weak inversion and/or the subthreshold regions of operation and the I_{READ} distributions begin to show a log-normal shape.

Since the read performance of an SRAM array is limited by the slowest SRAM cell, assuming an error-intolerant design with no ECC and/or redundancy to correct for errors from a slow read, it is, therefore, valuable to model the worst-case bitcell I_{READ} over a

large number of arrays or even dies [5]. In addition, the modeling of the worst-case bitcell I_{READ} over a large number of SRAM blocks sharing a common sense amplifier is important for providing adequate timing margin before the activation the sense amplifier [7, 8]. To model the worst-case value of a sample taken from a continuous Gaussian distribution - such as I_{READ} , measured at $V_{DD} = 1.1V$ - Extreme Value Theory (EVT) [89] can be applied. In particular, it has been shown [5] that the worst-case bitcell I_{READ} over a large number of dies can be modeled using the Gumbel Distribution [62]. To verify this through silicon measurements of the per-cell I_{READ} , 1kb SRAM blocks from a 256kb SRAM array (consisting of four 64kb sub-arrays) are used as an example to represent a cell-group. The distribution density of the measured minimum I_{READ} over each 1kb cell-group (which involves taking the minimum of 2,048 data points - i.e. 2 measurements per bitcell) is presented in Figure 4.36a, showing a long tail to the left - illustrating a need for extreme order statistics. Figure 4.36b shows that the long tail in the minimum I_{READ} distribution fits nicely to a Gumbel distribution.

4.5 Impact of Systematic Variability on SRAM Cell Stability

Through the large-scale characterization of SRAM stability in the functional SRAM arrays, as well as the conventional characterization in the SRAM macros, several sources of process-induced systematic variability are identified. First, the effects of a shallow trench isolation (STI) induced stress on the bitcell stability, identified through measurements in the SRAM macros with all-internal-node access, is described. Then, a systematic within-cell mismatch, identified through measurements in the functional SRAM arrays, is described and the impact of the SRAM cell orientation on its stability and performance is presented. Finally, the impact of die-to-die (D2D) and wafer-to-wafer (W2W) variability on the SRAM V_{MIN} is summarized.

4.5.1 Effects of Shallow Trench Isolation (STI) Induced Stress on SRAM Cell Stability

Figure 4.37a presents the layout view for a 20×40 SRAM array, with gate-poly in the vertical direction, wired for all-internal-node access for conventional SRAM metrics and transistor I-V characterization. Each array inside the SRAM test macro is separated from the thick-oxide switch network by wide regions of shallow trench isolation (STI) on all four sides. Figure 4.37b-c summarizes the impact of a STI-induced stress on the transistor performance in this low-power $45nm$ process. Sub-atmospheric chemical vapor deposition (SACVD) oxide is used in this process as STI gap-fill. Transistor channels in this process are oriented in the $<100>$ direction, making the PMOS transistors insensitive to stress while, at the same time, enhancing the hole mobility. Due to the usage of SACVD oxide as the gap-fill, the trenches exert a weak tensile strain on the NMOS transistors orthogonal (i.e. transverse) [139] to the direction of current flow, rather than a strong compressive strain [29, 112]. Figure 4.37b reveals a systematic decrease in the I_{DSAT} of the NMOS pass-gate transistors and the NMOS

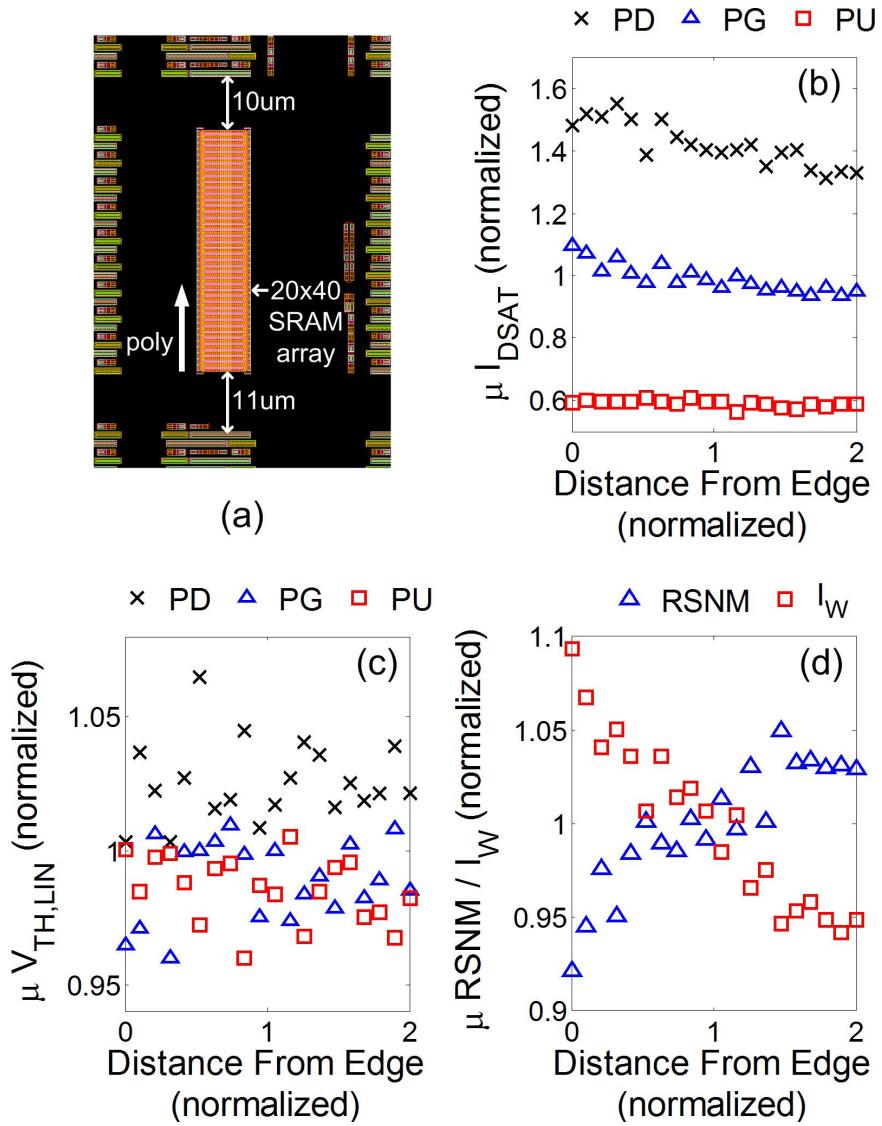


Figure 4.37: (a) Layout view for a 20×40 SRAM array, with gate-poly in the vertical direction, wired for all-internal-node access. Each array inside the test macro is surrounded by wide regions of STI in all directions. (b) μ of the measured I_{DSAT} for pull-down, pass-gate, and pull-up transistors as a function of the distance from the edge of the array (normalized to the average distance). (c) μ of the measured $V_{TH,LIN}$ for pull-down, pass-gate, and pull-up transistors as a function of the distance from the edge of the array. (d) μ of RSNM and I_W as a function of the distance from the edge of the array. All measurements are taken from SRAM macros via all-internal-node access.

pull-down transistors away from the periphery of the SRAM array within each test macro, while the I_{DSAT} of the PMOS pull-up transistors remain unaffected - both agreeing with the speculations. The average recorded drop in the NMOS I_{DSAT} from the edge to the center of the SRAM arrays is roughly 10%. In addition, Figure 4.37c indicates that the $V_{TH,LIN}$ of both NMOS and PMOS transistors display no systematic dependence on their

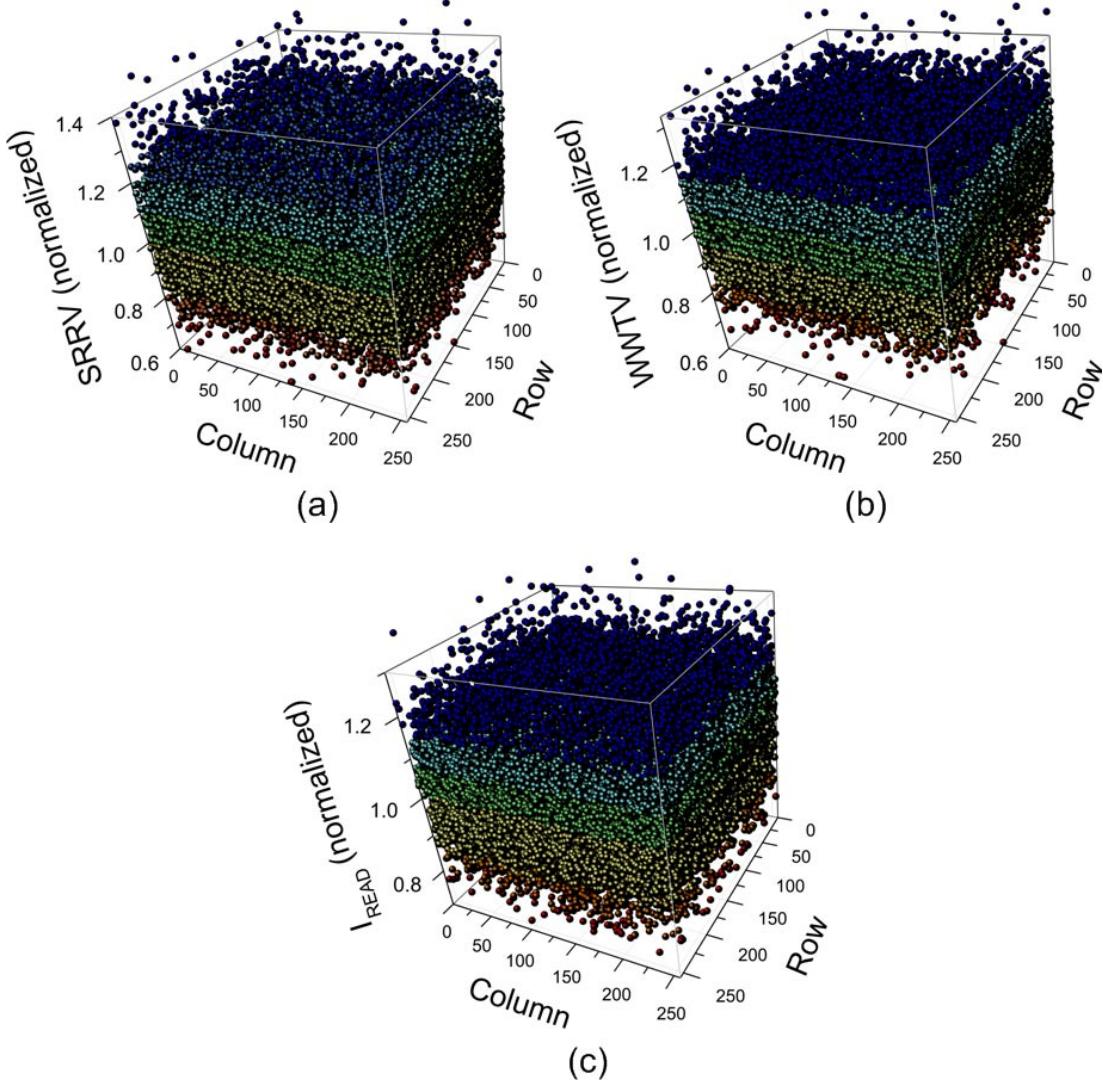


Figure 4.38: Measured (a) SRRV, (b) WWTW, and (c) I_{READ} as a function of row and column position within a 256×256 (64kb) functional SRAM sub-array.

positions within each array. This confirms an enhancement in the electron mobility of the NMOS transistors due to the tensile strain induced by the large SACVD-filled STI regions peripheral to each SRAM array within the test macro. The impact of this STI-induced stress on the SRAM read stability and writeability is summarized in Figure 4.37d. Due to a decrease in the NMOS transistor strength, while the PMOS transistor strength stays unaffected, the bitcell RSNM increases away from the STI interface of the SRAM array, whereas the I_W drops. Due to a more direct impact of the NMOS-to-PMOS transistor ratio on the writeability of the SRAM cell, the drop in I_W from the edge to the center of the array is roughly 15% whereas the rise in RSNM is just over 10%.

Figure 4.38 plots the measured SRRV, WWTW, and I_{READ} as a function of the row and the column position within a 256×256 (64kb) functional SRAM sub-array. Although significant systematic shifts in the SRAM RSNM and I_W are observed in the SRAM macros

with all-internal-node access, no clear systematic drifts are recorded in SRRV, WWTW, and I_{READ} measurements from the functional SRAM arrays as they are densely surrounded by peripheral circuitry as well as dummy fills. Figure 4.38 indicate that all measured metrics - SRRV, WWTW, and I_{READ} - vary, in a random fashion, with cell position within each 64kb SRAM array. This emphasizes, again, the importance of characterizing the stability of SRAM cells in their natural environment.

4.5.2 Within-cell Mismatch and Cell Orientation

SRAM cells are typically mirrored both horizontally and vertically to maximize the array density, yielding 4 different cell orientations, as illustrated in Figure 4.39a. In this 4-cell cluster, orientations A and D share the same layout, with reversed storage nodes; likewise, orientations B and C share the same layout, with reversed storage nodes²⁴. Figure 4.39c-e summarizes the effect of the within-cell mismatch and the cell orientation on SRAM stability and performance. The measurement results from two test chips, scattered across the same wafer, are highlighted. The locations of the test chips within the wafer are identified in Figure 4.39b. Since a high within-cell mismatch causes the more read-stable data polarity to always be preserved, the large-scale read metric for that particular data polarity cannot be extracted (Sections 2.3.1 and 4.2). Therefore, to more clearly show a difference in the read stability, the frequency of read disturbance - measured as the fractional occurrence of a data disturbance when either storage node (CL or CH) holds a '0', as a function of the cell storage node and the cell orientation, is plotted in Figure 4.39c²⁵. Figure 4.39d-e plots the μ of the measured WWTW and I_{READ} as a function of the cell storage node and the cell orientation - where the WWTW measurement for each storage node indicates the margin measured when writing a '1' into that node, which initially holds a '0'.

Measurement data reveal up to $4\times$ difference in the read disturb frequency, $\sim 4\%$ shift in the measured μ_{WWTW} , and $\sim 8\%$ shift in the measured μ_{IREAD} between the two data polarities of the bitcell. This shift is consistent throughout the test chip, suggesting a systematic mismatch between the two halves of the SRAM cell, which may be attributed to a difference in the direction and the location of the notches in the NMOS and PMOS diffusions of the two half-cell layouts. Additionally, the effects of poly-gate to active-source/drain misalignment may also play a role. To see this more clearly, consider the layout cartoon in Figure 4.40, overlaid with a drawing of the lithography contour for the NMOS and PMOS diffusions, simulated using Calibre [28]. First notice that both the directions and the locations of the notches in both the NMOS and PMOS diffusions are opposite for the two half-cells. In addition, due to diffusion rounding [66], any misalignment between the poly-gate and the active-source/drain can be manifested as a change in the transistor channel width [104, 156]. For minimum geometry transistors used in SRAM cells, this may also result in a change in the transistor V_{TH} due to the reverse narrow width effect [152] - i.e. V_{TH} decreases as the channel width is reduced. The extent of the reverse narrow width effect depends on the transistor channel width and is, therefore, different for each of the three transistors in a

²⁴Storage nodes are labeled such that the CL and CH side of orientation A (B) have identical layouts as the CL and CH side of orientation D (C).

²⁵Here, a read disturbance is recorded for the storage node, initially holding a '0', to first toggle to a '1'.

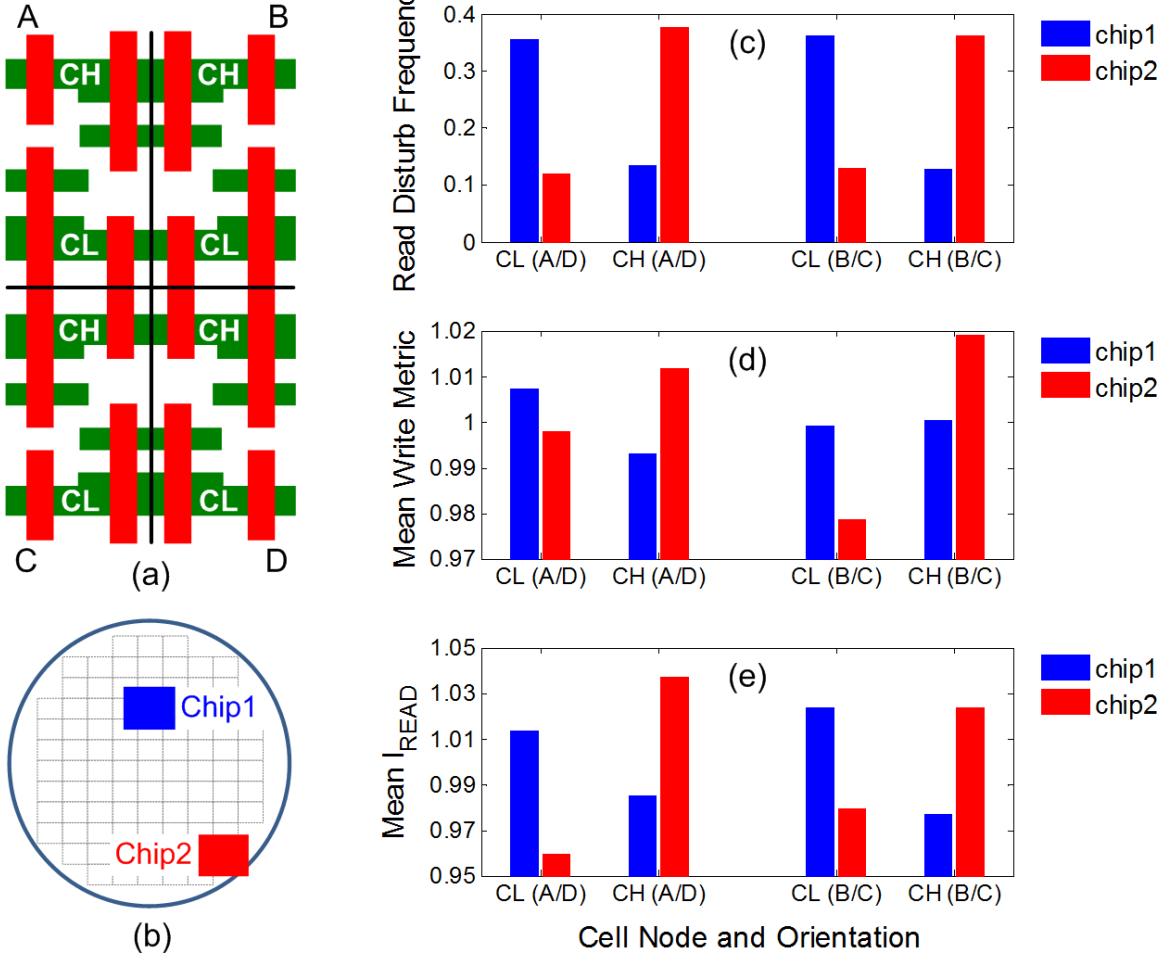


Figure 4.39: (a) A 4-cell cluster in an SRAM array showing the 4 cell orientations; the storage nodes of orientations C and D are reversed in the drawing for clarification. (b) Wafer map identifying the measured chips. Measured (c) read disturb frequency, (d) μ of WWTW, and (e) μ of I_{READ} for two test chips on the same wafer as a function of the cell storage node and the cell orientation.

half-cell. The net effect of this misalignment is different for each half-cell due to the differential placement of the transistors between the two half-cells - i.e. the pull-down transistor for the bottom half-cell is placed on the left side of the pass-gate transistor, whereas, for the top half-cell, the pull-down transistor is placed on the right side of the pass-gate transistor. Since the misalignment shifts the poly-gate, relative to the active-source/drain, in the same direction across each die, a systematic within-cell mismatch may result across the test chip. As a result, diffusion-notch-free (DNF) SRAM cells [81, 104, 156], which limits the cell β -ratio to 1 (or closed to 1, if a different L_G is used for the pull-down transistor versus the pass-gate transistor), have been proposed as an alternative. In addition, Figure 4.39c-e indicate that the direction of the systematic within-cell mismatch can be the same or opposite between two test chips taken from the same wafer - this is not surprising as the gate-poly to active-source/drain misalignment may be different for different dice on the same wafer.

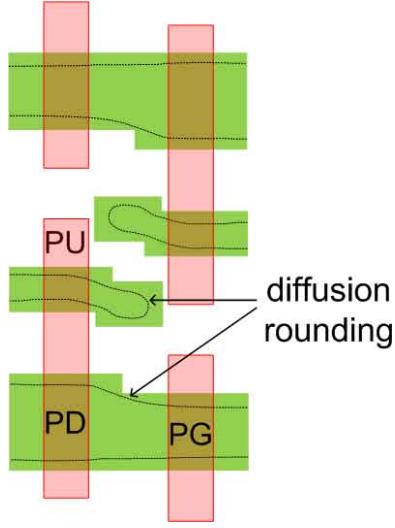


Figure 4.40: Layout cartoon of an SRAM cell showing the corner rounding of the PMOS diffusions and the NMOS diffusions.

The measurement data also indicates that the directions of the shifts in the read disturb frequency, WWT_V, and I_{READ} are correlated - i.e. a higher read disturb frequency (for a storage node initially holding a '0') typically corresponds to a higher writeability (for writing a '1' into that storage node) and a higher I_{READ} (when that storage node holds a '0').

As the SRAM cells are mirrored across the direction of the gate-poly (from orientation A/D to orientation B/C), a slight alteration in the degree of the within-cell mismatch is observed. This can also be attributed to a slight misalignment between the poly-gate and the active-source/drain, which adjusts the channel widths differently for the same storage side of neighboring bitcells. As the poly lines are shifted to the right or the left, both the NMOS pull-down transistors and the NMOS pass-gate transistors on the same storage side of the bitcell experience either a common increase or a common decrease in the channel width, due to the diffusion rounding, depending on the cell orientation. As long as the degree of poly-gate misalignment stays fairly uniform throughout the SRAM array, the fluctuations in the cell β -ratio (i.e. the strength ratio of pull-down to pass-gate transistors) should be small. Therefore, the observed alteration in the degree of within-cell mismatch (between orientations A/D and orientations B/C) is the smallest for the read disturb frequency. This alteration is slightly greater for I_{READ} , which has a more direct dependence on the pass-gate transistor drive strength than the pull-down transistor drive strength. Since the PMOS diffusion is narrower (due to a narrower channel width) than the NMOS diffusion (for either the pass-gate or the pull-down transistor), the V_{TH} of the pull-up transistor is expected to vary more with the misalignment due to a stronger reverse narrow width effect. Therefore, this alteration is most pronounced in the cell writeability, which directly depends on the strength of the PMOS pull-up transistor.

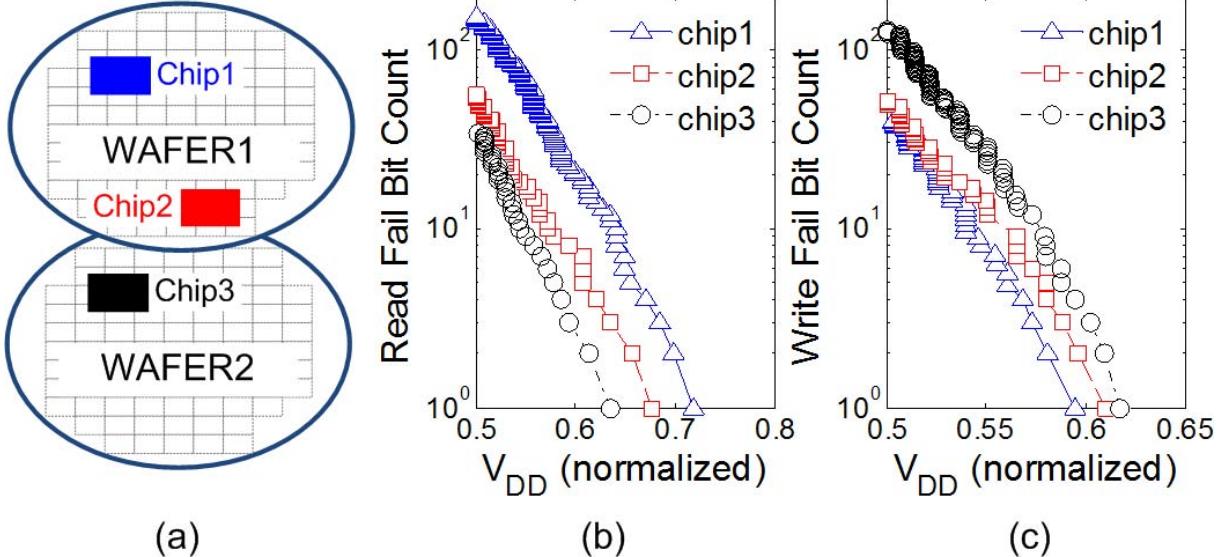


Figure 4.41: (a) Locations of the measured test chips on two different wafers. (b) Fail bit count as a function of V_{DD} during a static read operation, and (c) fail bit count as a function of V_{DD} during a static write operation, measured for 64kb SRAM sub-arrays on three test chips from two different wafers.

4.5.3 Die-to-Die (D2D) and Wafer-to-Wafer (W2W) Variability

To assess the impact of die-to-die (D2D) and wafer-to-wafer (W2W) variability on the yield of functional SRAM arrays in this low-power 45nm process, the per-cell V_{MIN} is characterized for both the read and the write operations. Figure 4.41 plots the fail bit count as a function of V_{DD} for (static) read and write operations, measured for a 64kb SRAM sub-array from three test chips, separately located on two different wafers. The locations of the three test chips are identified in Figure 4.41a - chip1 and chip2 are scattered across wafer1 and chip3 shares the same location as chip1 on wafer2. The two wafers have a nominal 4nm difference in the effective transistor channel length (L_{EFF}), corresponding to two different process corners [112], where wafer1 represents the faster wafer. A 100mV word-line weak write is applied during the $V_{MIN,WRTIE}$ measurements for all three test chips. Results indicate that, due to D2D variability, chip1 shows a 3% reduction in the write-fail-free V_{DD} and a 6% rise in the read-upset-free V_{DD} compared to chip2. In addition, a notable W2W systematic shift in the measured V_{MIN} is also observed as chip3, on wafer2, displays a 5% rise in the write-fail-free V_{DD} and a 9% reduction in the read-upset-free V_{DD} compared to chip1, which is identically placed on wafer1. Notice that, in both cases, $V_{MIN,RD}$ and $V_{MIN,WRT}$ shift in opposite directions. This is expected, as illustrated in Figure 4.17, because the read stability and the writeability margins have opposite sensitivities to common-mode systematic variations in the different transistor pairs.

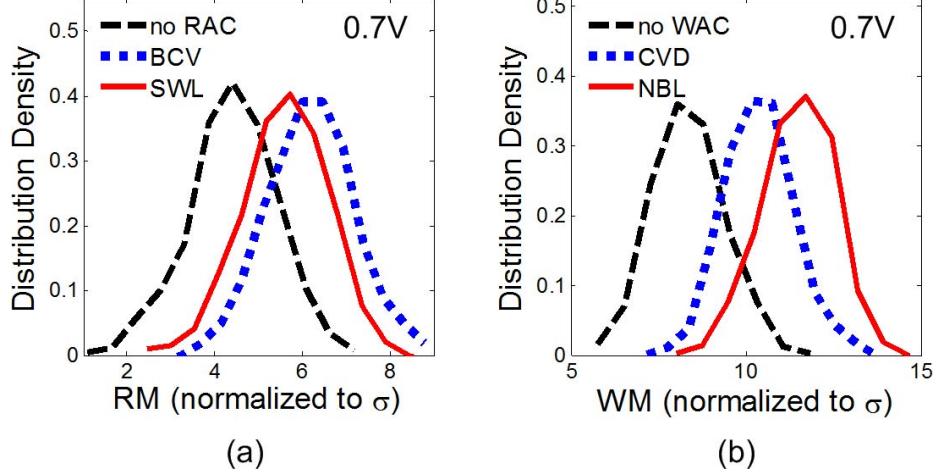


Figure 4.42: Distribution densities of (a) measured read margins (using either SRRV or WRRV) without RAC, and with BCV and SWL; and (b) measured write margins (using WWTV) without WAC, and with CVD and NBL. Measurements are taken for 2k-samples from a 64kb SRAM sub-array at $V_{DD} = 0.7V$.

4.6 Enhancement of SRAM Cell Stability using Assist Circuits

While process-induced variability, as the name suggests, is rooted in the process technology, circuit techniques can offer a quick and relatively cheap solution compared to process optimization, especially in a fast-paced semiconductor industry where first-to-market, in many situations, makes a huge difference in the success of any given product. Although these circuit techniques, often referred to as assist circuits [116], do not treat the problem at its source and cannot reduce random variability caused by LER, gate oxide interface roughness, and RDF - which increases with scaling as per $1/\sqrt{W \times L}$ (i.e. Pelgrom's model); they can shift the circuit operating point such that the effects of variability are less constraining. Therefore, these assist circuits, oftentimes, do not alter the shape of the measured noise margin distributions²⁶, but, rather, shift the distributions away from failure.

To assess the impact of assist circuits on the SRAM cell stability in this low-power $45nm$ process, two conventionally used read assist circuit (RAC) and two conventionally used write assist circuit (WAC) schemes are applied to a 64kb SRAM sub-array. To enhance the SRAM read stability, V_{CELL} can be raised to increase the cell β -ratio by increasing the gate overdrive of the pull-down transistor - this technique is referred to as the boosted cell V_{DD} (BCV) scheme [41, 156, 162] (Figure 4.43a). Alternatively, the V_{WL} can be reduced to decrease the gate overdrive of the pass-gate transistor (also increasing the cell β -ratio) - this technique is referred to as the suppressed word-line (SWL) scheme [100, 104] (Figure 4.44a). To enhance the SRAM writeability, on the other hand, V_{CELL} can be reduced [162] to decrease gate overdrive of the pull-up transistor and, thus, increasing the cell α -ratio - this

²⁶Assist schemes that change the substrate bias may increase or decrease the variance of the distributions through the body effect, which has been shown to affect the degree of variability [97].

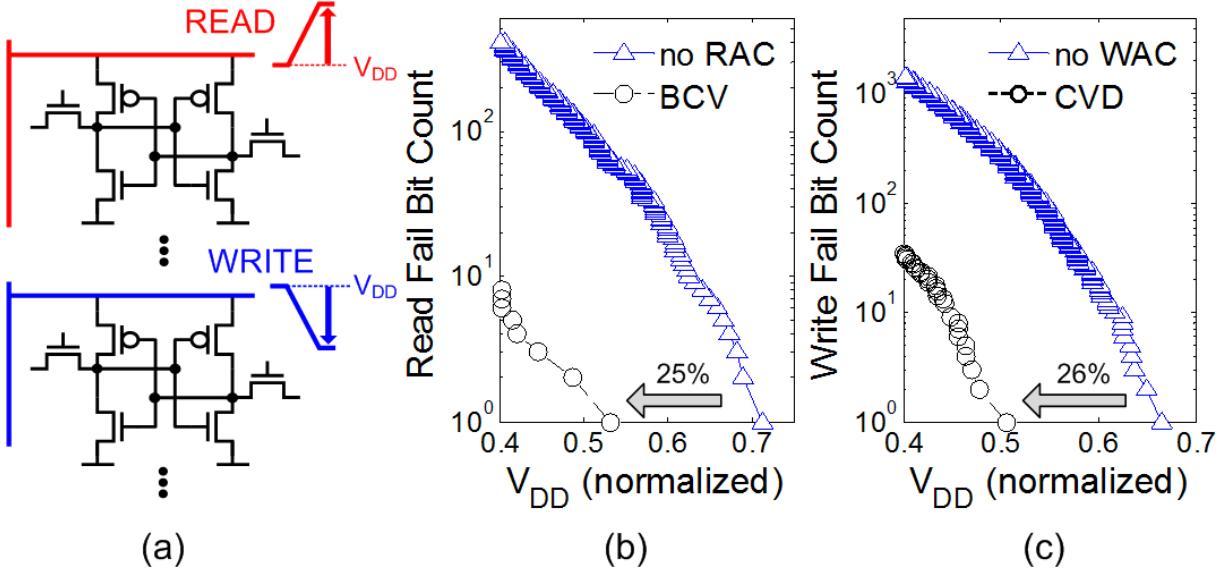


Figure 4.43: (a) Simplified schematic of the boosted cell V_{DD} (BCV) and the cell V_{DD} down (CVD) scheme for read and write assist. (b) Fail bit count as a function of V_{DD} during a read operation, measured for the same 64kb SRAM sub-array with no read assist circuits (RAC) and with a 100mV BCV. (b) Fail bit count as a function of V_{DD} during a write operation measured for the same 64kb SRAM sub-array with no write assist circuits (WAC) and with a 100mV CVD.

technique is referred to as the cell V_{DD} down (CVD) scheme (Figure 4.43a). Alternatively, the bit-line voltage at the '1' storage side can be pulled below V_{SS} to increase the gate overdrive of the pass-gate transistor to also enhance the cell α -ratio - this technique is referred to as the negative bit-line (NBL) scheme [100, 125, 144] (Figure 4.44a). A negative bit-line voltage is preferred, rather than raising the word-line voltage, because it does not increase the word-line stress for half-selected SRAM cells²⁷. Figure 4.42 presents the distribution densities of the measured read margins and write margins for 2k-samples of SRAM bitcells within a 64kb sub-array at $V_{DD} = 0.7V$. The measurements are taken for three cases - (i) without RAC/WAC, (ii) with a 100mV BCV/CVD, and (iii) with a 100mV SWL/NBL. For case (ii), since BCV requires access to the V_{CELL} terminal, SRRV cannot be characterized and WRRV is used to gauge the read stability. Figure 4.42 shows that the distributions of both the read and write margins are shifted to the right by activating the RAC/WAC, while the shapes of the distributions remain essentially unchanged - indicating a mean shift in the distributions with a constant variance²⁸. Results also indicate that the NBL scheme more effectively shifts the write margin distribution to the right than the CVD scheme. This is because the NBL not only increases the gate-source overdrive (V_{GS}) of the pass-gate transistor, but also reduces its source-body bias (V_{SB}), by pulling its source node below V_{SS} . Therefore, a forward body bias (FBB) is applied to reduce the V_{TH} of the pass-gate transistor, and

²⁷An SRAM cell is said to be half-selected if it is selected through the word-line row decoder but not through the bit-line column decoder - e.g. when writing a neighboring cell in the same row.

²⁸It should be noted that the N-well bias, V_{NW} , is shorted to V_{CELL} when activating both BCV and CVD as to not increase the variability due to a reverse body bias (RBB).

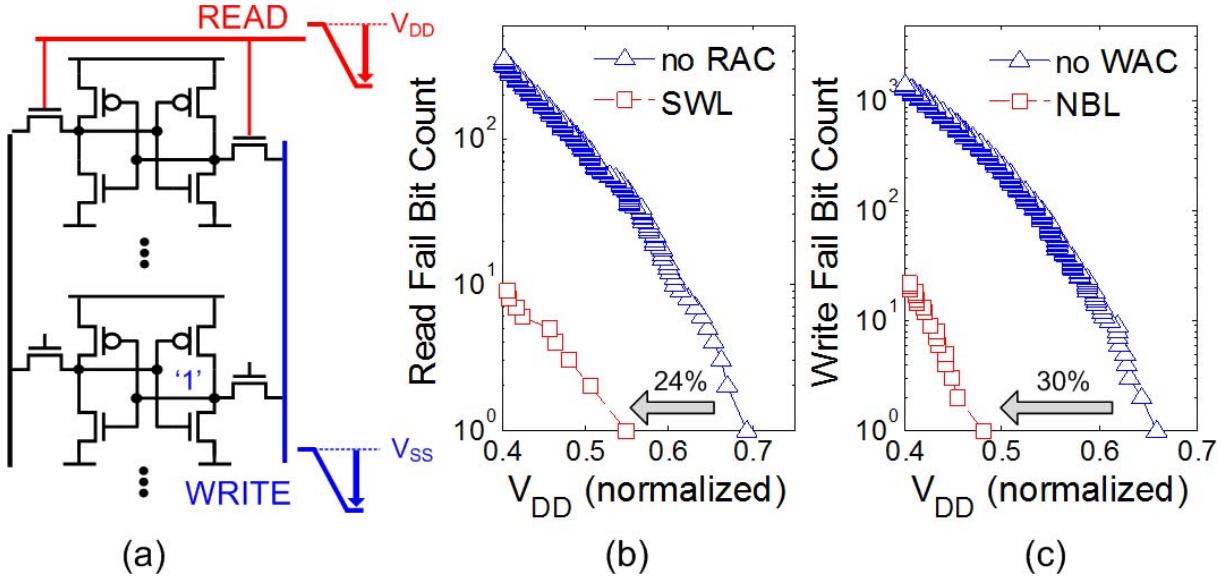


Figure 4.44: (a) Simplified schematic of the suppressed word-line (SWL) and the negative bit-line (NBL) scheme for read and write assist. (b) Fail bit count as a function of V_{DD} during a read operation measured for the same 64kb SRAM sub-array with no read assist circuits (RAC) and with a 100mV SWL. (b) Fail bit count as a function of V_{DD} during a write operation measured for the same 64kb SRAM sub-array with no write assist circuits (WAC) and with a 100mV NBL.

thus further enhances the cell α -ratio. The efficacy of BCV versus SWL, however, cannot be easily deciphered from Figure 4.42a because WRRV is used to gauge the read stability when BCV is applied and SRRV is used to gauge the read stability when SWL is applied - recall, from Section 4.2.3, that the μ/σ value shifts differently for SRRV and WRRV as the SRAM read stability is enhanced.

To further assess efficacy of the different assist circuits, the per-cell V_{MIN} is characterized²⁹. Figures 4.43b-c and 4.44b-c plot the fail bit count as a function of V_{DD} for (static) read and write operations, before and after applying a 100mV BCV/SWL and a 100mV CVD/NBL. Measurements indicate that a 100mV BCV and a 100mV SWL achieve similar $V_{MIN,RD}$ enhancements - 25% and 24%, respectively. On the other hand, a 100mV NBL achieves slightly better $V_{MIN,WRT}$ enhancements than a 100mV CVD - 30% versus 26% - and is in agreement with Figure 4.42b.

When selecting between the different RAC and WAC options, however, it is not sufficient to only compare their efficacy in terms of V_{MIN} enhancement. For instance, while both BCV and SWL are shown to achieve similar $V_{MIN,RD}$ enhancements, application of the SWL scheme may negatively impact the read access performance due to a degraded V_{GS} of the pass-gate transistor during the read cycle. Similarly, although NBL is shown to achieve better $V_{MIN,WRT}$ enhancements than CVD, a small positive V_{GS} is applied for all un-accessed SRAM cells in the same column when NBL is activated, which may lead to increased bit-line leakage current during write cycles. In addition, CVD can be easily

²⁹Note that all $V_{MIN,WRT}$ measurements are performed with a 100mV word-line weak write.

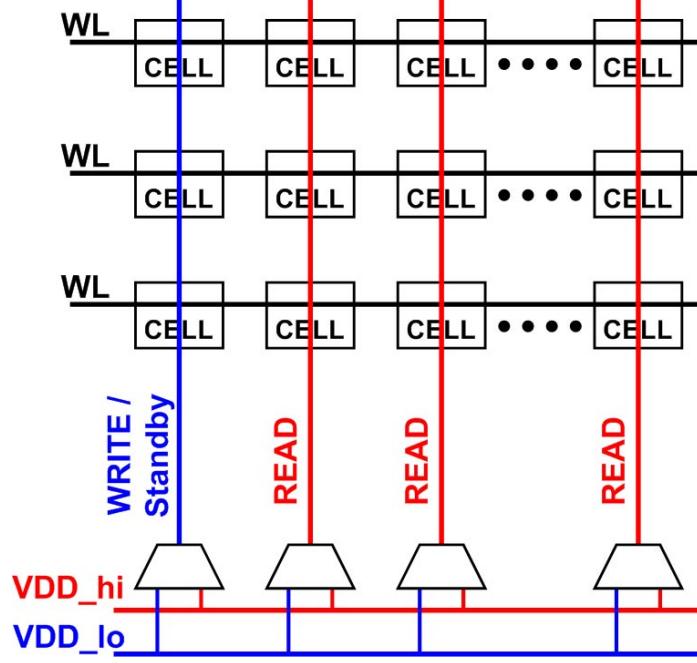


Figure 4.45: Circuit diagram for a column based biasing scheme, implemented in [162], to independently achieve high read stability and writeability.

integrated with BCV to simultaneously enhance the SRAM read stability and writeability - by boosting the cell supply during the read cycle and suppressing the cell supply during the write cycle - for SRAM cells utilizing the thin-cell topology with column-based supply routing (Figure 4.45) [162]. On the other hand, the NBL technique can remain effective for writeability enhancement of dual-port SRAM, whereas CVD has compatibility issues with dual-port SRAM because it degrades the read stability of a simultaneously accessed cell from the same column [100, 144]. Furthermore, scalability issues may also need to be considered - since both NBL and BCV increase the voltage across the transistor gate oxide³⁰, their scalability may be limited due to a reduction in the maximum tolerable voltage across the gate oxide as its thickness scales down. Lastly, other issues may also include the area penalty and the design complexity of each scheme - perhaps the bitcell can be designed to favor either read stability or writeability, with the weaker of the two enhanced by a single assist circuit; etc. Therefore, it is important to carefully consider each aspect of the assist circuits, in addition to its efficacy in V_{MIN} enhancement, when selecting the best-fit scheme for a specific design or application.

4.7 Summary

This chapter presents the measurement results for both the conventional and the large-scale SRAM design metrics from a strained-Si low-power 45nm CMOS test chip. The large-

³⁰While NBL increases the voltage across the gate oxide of the pass-gate transistor, BCV increases the voltage across the gate oxide of the pull-up transistor.

scale characterization of SRAM variability is attractive for early stages of SRAM development due to its ability to capture massive statistical data at a very low design and area overhead, compared to the conventional method. In addition, with the large-scale characterization method, a direct correlation between the measured read/write margins and the per-cell V_{MIN} is established. Due to the excellent agreement between the measured read/write margins and the per-cell V_{MIN} near the regions of read stability/writeability failure, quick and accurate V_{MIN} estimation using the measured large-scale read/write metrics - in particular, SRRV and WWTW - is possible.

While methods using mixture importance sampling (IS) and Extreme Value Theory (EVT) can provide very fast estimation of the SRAM V_{MIN} , they cannot be easily applied using measured results and their accuracy is heavily dependent on the transistor models. Therefore, a method to estimate the SRAM V_{MIN} using a statistical transformation of the *PDFs* of the read/write margins is presented, where a general and exact model for the *PDF* of the minimum noise margin is described. This method is first examined against a 100k-sample MC simulation and then verified through measurements in a 64kb SRAM sub-array. While Section 4.2.3 suggests that the μ/σ value is metric dependent and raises the question of the suitability of using μ and σ for V_{MIN} estimation, the silicon validation of the estimated $V_{MIN,RD}$ and $V_{MIN,WRT}$ in Section 4.3.2 indicate that accurate V_{MIN} estimation is possible when using the right read/write metrics. In particular, metrics, whose μ and σ values can be accurately fitted, using a relatively low-order polynomial, over a wide range of V_{DD} values, are shown to be good candidates for estimating V_{MIN} .

Several sources of systematic variability and their impacts on the SRAM cell stability are studied. In particular, a systematic shift in the SRAM read/write margins, due to a STI-induced stress, is found in the measurements taken from the SRAM test macros with all-internal-node access. This effect is not observed in the large-scale measurements taken from the functional SRAM arrays and, therefore, further emphasizes the importance of characterizing the stability of SRAM cells in their natural environment. A systematic within-cell mismatch is identified using large-scale characterization of functional SRAM arrays. This mismatch can be attributed to a rounding effect near the notches in the NMOS and PMOS diffusions, which, couple with a slight gate-poly to active-source/drain misalignment, results in a systematic variability in the V_{TH} of all six transistors due to the reverse narrow width effect. To combat this, diffusion-notch-free (DNF) SRAM cells have been recently proposed in literature. Finally, the impact of several conventionally used read and write assist circuits on the SRAM cell stability is examined through measurements - achieving $\sim 24\%-30\%$ improvement in the SRAM $V_{MIN,RD}/V_{MIN,WRT}$.

Chapter 5

Robust SRAM Design using FinFETs

5.1 Introduction

Scaling of the classical bulk-Si MOSFET structure down into the sub-20nm regime presents several key challenges. The control of short channel effects (SCE) requires heavy channel doping ($> 10^{18} cm^{-3}$) and heavy super-halo implants to suppress sub-surface leakage currents. Consequently, carrier mobilities are severely degraded due to impurity scattering and a high transverse electric field in the on-state. Furthermore, the increased depletion charge density leads to a larger depletion capacitance and, hence, a larger subthreshold slope. Thus, for a given off-state leakage specification, the on-state drive current is degraded. In addition, the off-state leakage current is enhanced due to band-to-band tunneling (BTBT) between the body and the drain.

While these challenges impose a direct restriction on the performance versus power optimization of memory (and logic) circuits, the ultimate limitation on the scaling of bulk-Si based SRAM will be determined by the yield. In particular, the SRAM read stability has been demonstrated [22, 32, 136] to have a high sensitivity to the random V_{TH} mismatch in its transistors - especially in the pull-down transistors (Figure 4.17a). While the SRAM writeability is shown [32] to have the highest sensitivity to common-mode systematic variations in the pass-gate transistor V_{TH} , it also exhibits a significant sensitivity to the random V_{TH} mismatch in the pass-gate and the pull-up transistors (Figure 4.17b). Among the various sources of device parameter variability, random dopant fluctuation (RDF) dominates as the primary supplier of σ_{VTH} [9, 135] in planar bulk-Si and partially-depleted silicon-on-insulator (PD-SOI) MOSFETs and will continue to do so, at least, until $L_G < 20nm$ [10]. Table 5.1 [31] summarizes the predicted increase in σ_{VTH} due to RDF (whose magnitude is inversely proportional to the channel area [113]), down to the 22nm technology node, following the scaling specifications from the International Technology Roadmap for Semiconductors (ITRS) [73].

To maintain yield in the presence of increased σ_{VTH} , the traditional 6-T SRAM cell has been scaled at a slower pace, since larger channel areas make the bitcell more immune to variations; this has been a dominant approach in the 65nm and the 45nm technology nodes. Recently, bitcell designs implemented with extra (i.e. more than six) transistors have been proposed to enhance the cell margins. The most prevalent alternative to the 6-T SRAM bitcell is the 8-T dual-port SRAM bitcell (Figure 5.1a) [37], which employs two extra

Table 5.1: Expected RDF-induced V_{TH} variation, expressed as a percent of the 90nm node value (for $W/L = 2$), following the ITRS scaling specifications.

Node	65nm	45nm	32nm	22nm
σ_{VTH} (% of 90nm node)	149%	211%	294%	426%

transistors ($N1$ and $N2$), incurring a $\sim 30\%$ area penalty (compared to a 6-T bitcell with a β -ratio of 2), to form a separate read path. Since the storage node (CL) is accessed through a high impedance node (i.e. gate of $N1$) during the read operation, no read stress is exerted on the SRAM cell. However, half-selected cells continue to suffer from an unintentional read stress during the write cycle; therefore, this bitcell design necessitates the update of an entire row during each write operation - either limiting the row width to be a single word or requiring a read-modify-write configuration [38]. To address the half-select issue, a 10-T dual-port cross-point SRAM bitcell is implemented (Figure 5.1b) [35]. This bitcell has the capability of accessing a single cell, with a separate row word-line and a column word-line, and, therefore, do not suffer from half-select and has reduced leakage during the read- and write-cycles. It also has the advantage (compared to the 8-T dual-port bitcell) of a differential read. However, each write access is performed through a series of two pass-gates - $N2$ and $N3$. Additionally, a 100% area penalty is incurred, compared to the 6-T bitcell - equivalent to backtracking one full technology node (in terms of density). To be fair, both the 8-T dual-port bitcell and the 10-T dual-port cross-point bitcell are shown to function at supply voltages unachievable using the 6-T design and may also benefit from a better scalability. In addition, they allow simultaneous read-and write-access, and, therefore, may have more applications than the traditional 6-T bitcell. More recently, an 8-T single-port cross-point bitcell is also implemented (Figure 5.1c) [154] to address the half-select issue without incurring the extra area penalty of the 10-T design. While all these cell designs (including the 6-T bitcell with larger transistors) are able to offer improved yield, they all suffer from increased cell areas, which undermine the fundamental objective of technology scaling - to increase density. Alternatively, assist techniques [41, 100, 104, 116, 125, 144, 156, 162] have been implemented to widen the SRAM design margins by shifting the SRAM operating point away from failure - as discussed in Section 4.6. These assist circuits aim to increase the array robustness of smaller bitcell designs, but inevitably degrade the array efficiency and, thus, the array density. Furthermore, all these techniques are designed, not to address the source of the problem (i.e. increased σ_{VTH}), but rather, as an attempt to hide its effects.

To combat the increasing σ_{VTH} in planar bulk-Si (and PD-SOI) MOSFETs, techniques for adaptive body biasing (ABB) have been implemented [142] to reduce the frequency and leakage variations in logic and, more recently [94], to reduce parametric failures in SRAM. In [94], two bias levels are generated to allow three different body bias voltages - $-500mV$, $0V$, and $+500mV$. A single body bias is then selectively applied, based on leakage monitoring, to all SRAM arrays in each test chip, via large PMOS bypass transistors, to combat die-to-die (D2D) V_{TH} shifts. The area penalty of the PMOS bypass transistors is reported to be $\sim 5\%$, in addition to the area required for leakage monitoring and bias generation, which is not reported. This technique can be adopted to combat systematic within-die (WID) V_{TH} shifts by segmenting the SRAM array into smaller blocks sharing common well

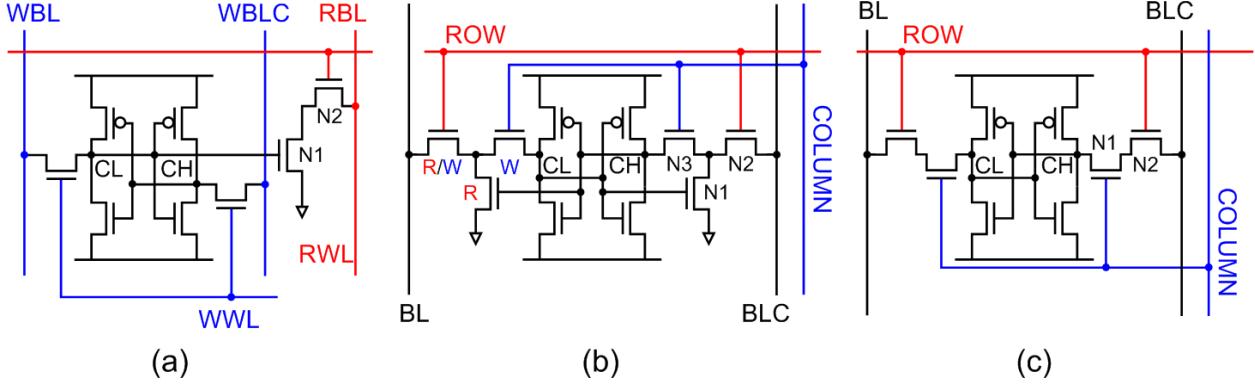


Figure 5.1: Schematic of a (a) 8-T dual-port SRAM bitcell, (b) 10-T dual-port cross-point SRAM bitcell, and (c) 8-T single-port cross-point SRAM bitcell.

potentials. In addition, the resolution of V_{TH} adjustments can be enhanced with more body bias levels. However, the area penalty of this technique increases with both the number of bias levels and the number of shared wells. In addition, although these schemes have been shown to effectively reduce the impact of D2D (and WID [142]) systematic variability, random V_{TH} mismatch continues to limit the scaling of memory arrays. Furthermore, the range of V_{TH} tuning, using body bias, is shown [79] to decrease with bulk-Si CMOS scaling, and, therefore, limits the scalability of ABB. Ultimately, maintaining tight control of V_{TH} will require a device architecture in which V_{TH} is determined by parameters with lower variability - i.e. physical channel dimensions and the gate work function. Fully-depleted SOI (FD-SOI), FinFET, triple-gate, and gate-all-around devices using light channel doping have all been proposed to reduce σ_{VTH} and thereby enable continued scaling of CMOS, particularly memory, circuits.

In this chapter, FinFET based SRAM bitcells [30, 31, 63, 75, 103] are investigated as an alternative for nanoscale memory design. The advantages of the FinFET technology for SRAM design is summarized in Section 5.2. In addition, the methodology used to assess the different designs is described. Section 5.3 explores several FinFET based 6-T bitcell designs, including a dynamic pass-gate feedback (PGFB) architecture, first introduced in [63]. The design of a robust 4-T SRAM cell using FinFETs is described in Section 5.4. Finally, the results are summarized in Section 5.5.

5.2 FinFET Technology for SRAM Design

5.2.1 Advantages of the FinFET Technology

To overcome the scaling limitations imposed by bulk-Si (and PD-SOI) MOSFETs, recent studies have focused on advanced MOSFET structures such as the fully-depleted SOI (FD-SOI), where the depletion region extends throughout the entire thickness of the channel layer. These structures enable more aggressive transistor scaling due to a more effective control of SCE by utilizing a very thin body. Scaled FD-SOI MOSFETs can eliminate the need for channel dopants, resulting in lower transverse electric field in the on-state and negligible

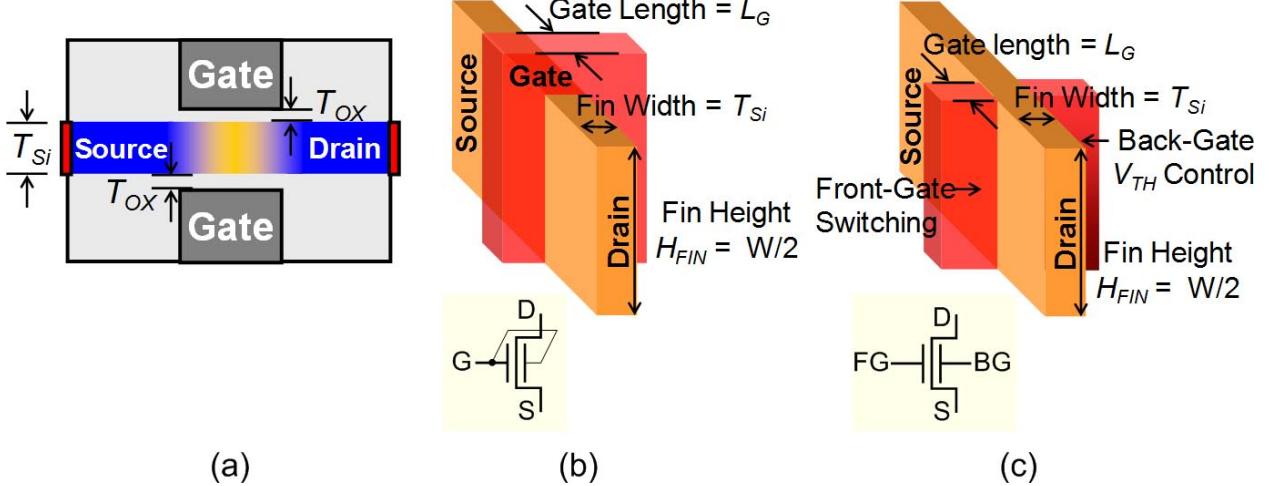


Figure 5.2: (a) Cross-sectional schematic of the FinFET structure. The gates of the FinFET can either (b) swing together in double-gated (DG) operation or (c) swing independently in independently-gated (IG) operation.

impurity scattering; hence higher carrier mobilities. In addition, devices with undoped channels have negligible depletion charge and capacitance, yielding a steep subthreshold slope. A reduced drain-to-body capacitance also enables higher circuit performance while consuming less dynamic power; in memory design, this translates to a reduction in the bit-line capacitive loading. Most importantly, the absence of channel dopants minimizes the V_{TH} variations due to RDF and thereby reduces σ_{VTH} . It has been reported [157] that SRAM cells constructed using thin-body FD-SOI devices can achieve improved stability. The planar structure of FD-SOI MOSFETs can also achieve a wide and continuous range of transistor widths, enabling optimal β - and α -ratios in SRAM design. Additionally, existing bulk-Si designs can be adopted in the FD-SOI technology with less design effort compared to non-planar MOSFET structures. However, the scalability of FD-SOI is limited as silicon films thickness of approximately $T_{Si} < L_G/4$ have been shown to be necessary for good short channel behavior down to $L_G = 18nm$ [39]. In addition, channel thicknesses of less than $5nm$ are expected to suffer from quantum confinement effects [57], resulting in degraded on-currents and increased V_{TH} sensitivities to T_{Si} variations. As a result, the scaling of FD-SOI beyond the 22nm node will be difficult.

More recently, the FinFET structure, graphically illustrated in Figure 5.2, has been developed [69] as an alternative to further improve scalability. The FinFET structure utilizes a vertical Si fin (rather than a planar Si surface) as the channel/body, which can be manufactured with conventional lithography and etching processes. The gate electrode of the FinFET straddles the fin (Figure 5.2b). The fin width is the effective body thickness, and the fin height is the effective channel width. In the on-state, current flows between the source and the drain along the gated sidewall surfaces of the Si fin. The FinFET structure can achieve similar improvements (as FD-SOI) in carrier mobility and subthreshold slope when an undoped channel is used. Similar to the FD-SOI MOSFET structure, the V_{TH} of FinFETs is determined by the silicon thickness and the gate work function. Therefore, it shares the same benefit of eliminating RDF-induced V_{TH} variations. In addition, both the

Table 5.2: Transistor parameters used for Taurus simulations.

Parameter	FinFET	Bulk-Si
L_G (nm)	22	22
L_{SD} (nm)	24	24
T_{OX} (Å)	11	11
T_{Si} (nm)	15	—
V_{DD} (V)	1.0	1.0
Channel Doping, N_{BODY} (cm^{-3})	10^{16}	4×10^{18}
H_{FIN} (nm)	30	—
S/D Doping Gradient (nm/dec)	4	4

depletion and the junction capacitances are effectively eliminated, thus reducing the bit-line capacitive loading. Due to the double-gate structure, FinFETs have been shown to achieve good short channel behavior with a relaxed body thickness requirement of $T_{Si} < 2L_G/3$ [13]. Although low V_{TH} values can be difficult to achieve simultaneously for both NMOS and PMOS FinFETs in logic circuits, a single gate material with a mid-gap work function can achieve symmetric high V_{TH} values for both NMOS and PMOS FinFETs in low-leakage applications, such as SRAM. A unique advantage of the FinFET structure is that the gates on either side of the fin can be electrically isolated to allow for independent operation, by selectively removing the gate material in the region directly on top of the fin (Figure 5.2c) [88]. In double-gated (DG) operating mode, the two gates are electrically shorted to switch the FinFET ON/OFF; whereas in independently-gated (IG) operating mode, the front-gate (FG) can be used to switch the FinFET ON/OFF while the back-gate (BG) can be used to adjust its V_{TH} . The IG operation offers dynamic performance tunability which can be leveraged to improve stability trade-offs in SRAM design [26, 30, 31, 63, 74, 82]. The FinFET structure, therefore, presents a promising device architecture for continued SRAM scaling due to its robust V_{TH} control and the opportunity for better stability trade-offs through the IG operation.

5.2.2 Methodology

Mixed-mode device simulation [138] using the drift-diffusion model for carrier transport and the density gradient model to account for quantum-mechanical effects in nanoscale MOSFETs is employed to simulate the DC transfer characteristics of SRAM cells under different biasing conditions. Because the high-field transient velocity overshoot effects are ignored, the drain current values may be underestimated. However, the trends and differences between device technologies and their impact on SRAM read/write margins should still be valid because they depend on the relative strengths of two transistors and not their absolute I_{ON} . While the simulated access times may deviate from actual values due to errors in estimating the I_{ON} together with unknown interconnect properties, they are expected to accurately illustrate relative performance. It is expected that the effect of parasitic resistances and capacitances will limit circuit performance in deeply scaled CMOS technologies. Series resistance and extrinsic contact resistance are included in this study, which lessens the improvements associated with the intrinsic device structure.

Table 5.3: 45nm node general logic design rules used for SRAM bitcell layouts.

Design Rules	Line / Space (nm)
Active	50 / 70
Poly	50 / 70
Contact	60 / 70
Metal1	60 / 60
Via1	60 / 60
M-x	70 / 70
Via-x	65 / 75
Design Rules	Extension / Space (nm)
Poly - Related Active	80 / 50
Poly - Unrelated Active	- / 25

The FinFET structure used in this study is graphically illustrated in Figure 5.2. The key design parameters for both planar bulk-Si MOSFETs and FinFETs are summarized in Table 5.2. L_{SD} and T_{Si} are optimized for the FinFET (consistent with the thickness requirement for good short channel behavior); T_{OX} and source/drain doping gradient are estimated from scaling trends. Because completely undoped silicon substrates are expensive and challenging to obtain, a low yet realistic channel doping of $10^{16} cm^{-3}$ is assumed for the FinFET. The FinFETs in this study are chosen to be symmetric, with identical oxide thicknesses and gate work functions for the front- and back-gates - this is motivated by the relatively high process complexity associated with asymmetric FinFETs, requiring either precise lithographic alignment (less than $T_{Si}/2$) or tilted implantations, which may become even more challenging due to the high aspect ratios of tall and densely packed fins. FinFETs fabricated on a standard (100) wafer have channels on the fin-sidewalls that are oriented along (110) planes, for standard layouts. To capture the effect of fin-sidewall surface orientation on FinFET performance, the carrier mobilities in Taurus [138] are calibrated using experimental data for the (110) surface [159].

To study the layout implications of various bitcell designs, 45nm node logic design rules, generated as a linearly scaled version of the 90nm node design rules for general logic, are summarized in Table 5.3. Because the design rules for general logic are more conservative than the typical SRAM design rules, the presented cell areas are larger than predicted by the roadmap; but they should indicate the relative compactness of different designs. Therefore, the cell areas are expressed as a multiple of F^2 , where F denotes the M1 half-pitch and the multiple represents the SRAM cell area factor - as used in [73].

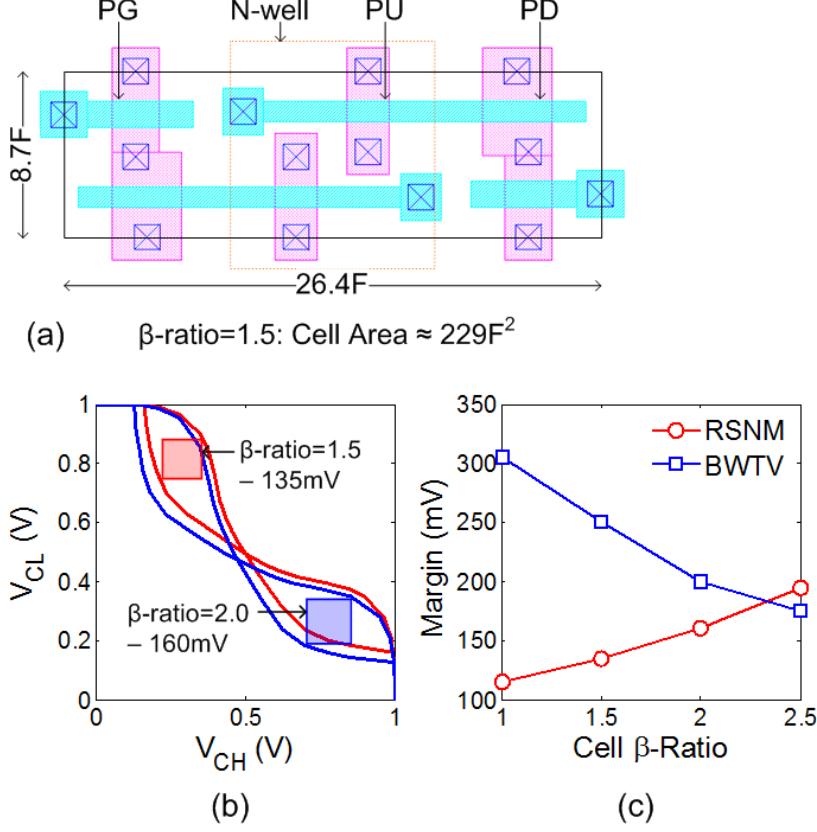


Figure 5.3: (a) Thin-cell layout for a conventional 6-T bulk-Si based SRAM cell with β -ratio = 1.5. The dark outline indicates the area of one memory cell. (b) Read butterfly-curves for a conventional 6-T bulk-Si based SRAM cell with β -ratio = 1.5 and β -ratio = 2.0. (c) Impact of cell β -ratio on the cell read- and write-margins (RSNM is used as the read metric and BWTV is used as the write metric). β -ratio is adjusted in (b) and (c) by changing the channel widths of the pull-down transistors.

5.3 6-T FinFET based SRAM design

5.3.1 Conventional Bulk-Si Based 6-T SRAM Cell

The thin-cell layout (up to the M1 via) for a conventional 6-T bulk-Si based SRAM cell with β -ratio = 1.5¹ is presented in Figure 5.3a. The dark outline indicates the memory cell boundary. It is important to note that the compactness of the bulk-Si based SRAM cell is significantly limited by the spacing requirement for the N-well and the two extra contacts required on each side of the bitcell to contact the P- and N-type diffusions.

Cell β -ratio is leveraged, by adjusting either the channel widths of the pull-down transistors or the L_G of the pass-gate transistors, to set the cell read stability margin. High- V_{TH} transistors are used to suppress cell leakage and enhance the cell read stability. Further increase in V_{TH} of the bulk-Si transistors may not translate to lower leakage due to band-to-

¹ β -ratio in this chapter denotes the size-ratio between the transistors, rather than the absolute strength ratio.

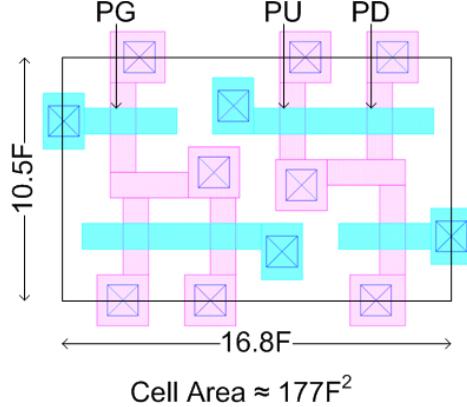


Figure 5.4: Thin-cell layout for a conventional double-gated (DG) FinFET based 6-T SRAM cell with β -ratio = 1. The dark outline indicates the area of one memory cell.

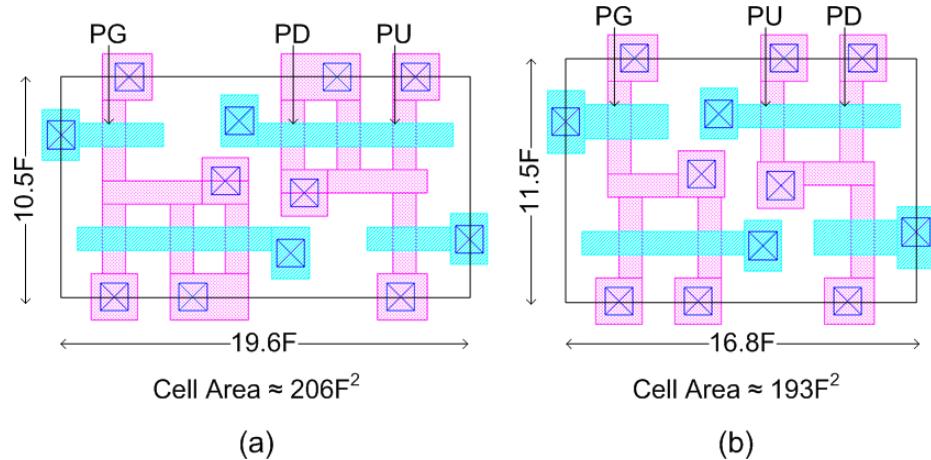


Figure 5.5: Thin-cell layout for a conventional double-gated (DG) FinFET based 6-T SRAM cell with (a) 2 fins in the pull-down transistors and (b) $L_{G,\text{pass-gate}} = 2 \times L_{G,\text{pull-down}}$. The dark outline indicates the area of one memory cell.

band tunneling (BTBT). Figure 5.3b presents the butterfly-curves for bulk-Si based bitcells with β -ratios of 1.5 and 2, through the adjustment of the pull-down transistor channel width - β -ratio = 1.5 yields a 6-T bitcell with RSNM= 135mV and a cell area of approximately $229F^2$.

While increasing the channel widths of the pull-down transistors enhances the read stability, the writeability is compromised due to a reduction in the inverter trip point. Increasing the L_G of the pass-gate transistors can achieve similar enhancements in the read stability; however, both writeability and I_{READ} are compromised. Figure 5.3c summarizes the impact of channel width adjustments in the pull-down transistors on the cell read margin - RSNM - and the cell write margin - BWTV (Section 2.3.3).

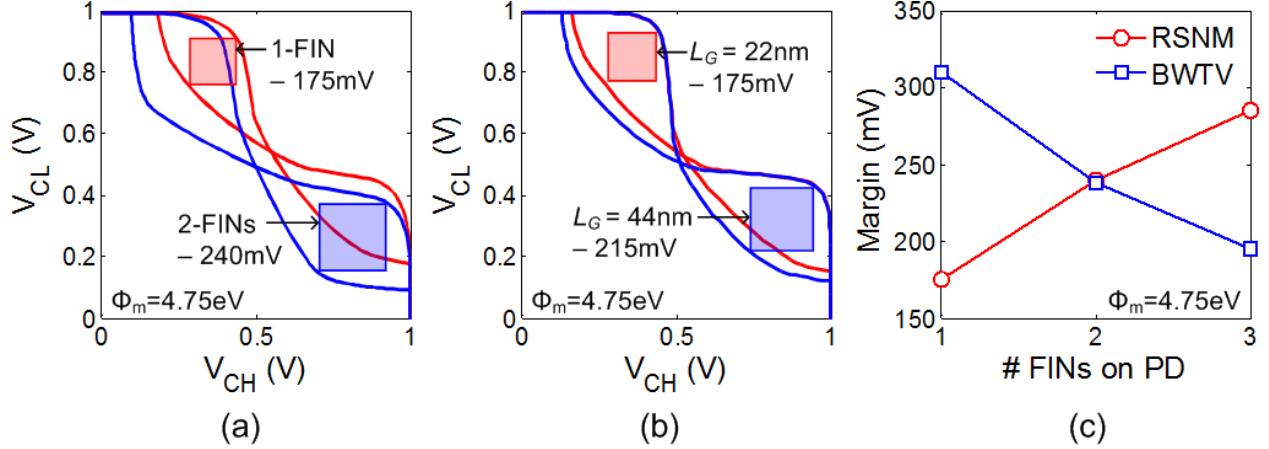


Figure 5.6: Read butterfly-curves for a conventional DG FinFET based 6-T SRAM cell with (a) 1 fin and 2 fins in the pull-down transistors, and (b) $L_G = 22\text{nm}$ and $L_G = 44\text{nm}$ for the pass-gate transistors. (c) Impact of cell β -ratio, determined by the number of pull-down transistor fins, on the cell read- and write-margins (RSNM is used as the read metric and BWTV is used as the write metric).

5.3.2 Conventional Double-Gated (DG) FinFET 6-T SRAM Cell

Several FinFET based SRAM cell architectures are explored to demonstrate the flexibility of designing SRAM using FinFETs. The conventional double-gated (DG) design is first presented. Figure 5.4 illustrates the thin-cell layout for a conventional DG FinFET based 6-T SRAM cell with β -ratio = 1. It should be noted that FinFET based SRAM bitcells will generally achieve denser layouts than similarly sized bulk-Si SRAM cells, because they can avoid the P- to N-well spacing rules and two contacts within the cell can be eliminated by directly connecting the NMOS and PMOS drains. A conservative source/drain contact scheme, using large landing pads, is assumed in this study. Elimination of the source/drain landing pads (e.g. by using local interconnects) can improve the FinFET layout efficiency, but also increases the parasitic capacitance [124].

Similar to the bulk-Si based design, the read stability of the DG 6-T bitcell can be enhanced by up-sizing the pull-down transistors or increasing the L_G of the pass-gate transistors. Since the channel widths of FinFETs are determined by the number of fins, only discrete sizing is available [103] - however, a pitch-halving technique, using spacer lithography [40], can significantly reduce the layout penalty of multi-fin transistors. Increasing the pass-gate transistor L_G has less impact on cell area but increases the word-line capacitance and also negatively impacts I_{READ} , resulting in slower access time. The thin-cell layouts for DG 6-T bitcells with 2 fins on each pull-down transistor and with longer L_G for each access transistor are shown in Figure 5.5. Fabricated FinFET SRAM cells based on these layouts have been previously reported [18].

Figure 5.6a-b presents the butterfly-curves for the DG 6-T bitcell with β -ratios of 1 and 2. With a RSNM of 175mV at a cell area of approximately $177F^2$, the FinFET based bitcell with single-fin pull-down transistors achieves a 30% improvement in RSNM, at $V_{DD} = 1\text{V}$, and a more compact cell layout, as compared to its bulk-Si based counterpart

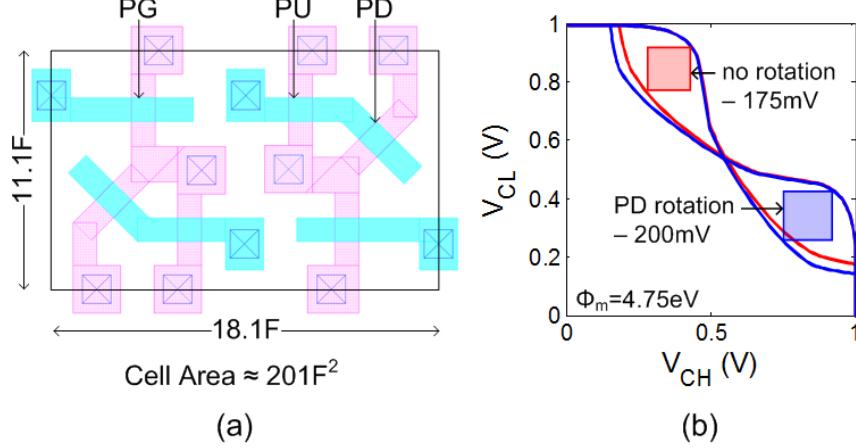


Figure 5.7: (a) Thin-cell layout for a DG FinFET based 6-T SRAM cell with fin-rotation to increase the effective cell β -ratio. The outline indicates the area of one memory cell. (b) Read butterfly-curves for a DG FinFET based 6-T SRAM cell with fin-rotation, showing improved RSNM.

with β -ratio = 1.5. A 37% further improvement in the RSNM, with a 17% area penalty, can be achieved by adding 1 extra fin to each pull-down transistor (with no pitch-halving technique). Alternatively, a 23% improvement in the RSNM can be achieved, with an 8% hit in the cell area, by doubling the L_G of the pass-gate transistors. High- V_{TH} transistors are implemented in the FinFET designs, to suppress cell leakage and enhance the cell read stability, by utilizing a gate material with a 4.75eV work function for both the NMOS and PMOS. Using a single gate material also improves manufacturability since it is challenging to implement different gate work functions (Φ_m) for closely spaced p-channel and n-channel fins - the high aspect ratio of the FinFETs makes it difficult to selectively tune Φ_m along the sidewalls of the fins, e.g. by masked ion implantation.

When the cell β -ratio is increased, either by adding fins to the pull-down transistors or increasing the L_G of the pass-gate transistors, the cell writeability degrades - either due to a reduction in the inverter trip point or a decrease in the pass-gate strength. The trade-off between the cell read- and write-margins as a function of the number of fins on the pull-down transistors is presented in Figure 5.6c.

5.3.3 Double-Gated (DG) FinFET 6-T SRAM Cell with Fin-Rotation

It is known that the electron mobility along (100) planes is higher than along (110). Therefore, the effective cell β -ratio, and thus the cell read stability, can be enhanced by rotating the NMOS pull-down transistors to have channel surface along the (100) plane. Unlike the planar bulk-Si based designs, FinFET based designs with channel surfaces both along (110) and (100) planes can be easily fabricated by rotating the (110) fins by 45° for the (100) orientation [36, 63]. As a tradeoff, printing rotated fins may be lithographically more challenging and may result in enhanced process variability. The layout of a DG 6-T SRAM cell with fin-rotation and the corresponding RSNM enhancement are shown in Figure 5.7. By rotating the fins of the pull-down transistors by 45°, a 14% improvement in the RSNM,

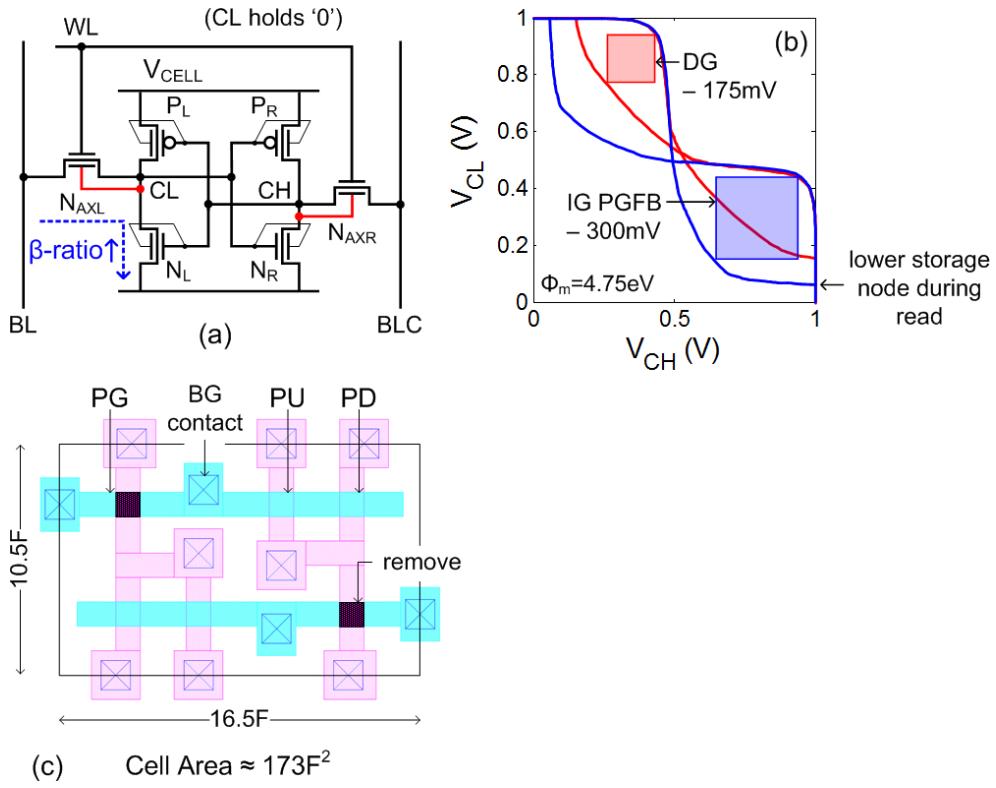


Figure 5.8: (a) Schematic for an IG FinFET based 6-T SRAM cell with dynamic PGFB. (b) Read butterfly-curves for an IG FinFET based 6-T SRAM cell with dynamic PGFB, showing significantly improved RSNM. (c) Thin-cell layout for an IG FinFET based 6-T SRAM cell with dynamic PGFB, indicating zero area penalty compared to the conventional DG 6-T design. The dark outline indicates the area of one memory cell. Note the use of BG-FinFET NMOS pass-gate transistors involves gate separation, as indicated in the layout by the dark region over their fins.

at $V_{DD} = 1V$, can be achieved with a 14% area penalty. Fabricated FinFET SRAM cells based on this design have been previously reported [18].

5.3.4 Independently-Gated (IG) FinFET 6-T SRAM Cell with Dynamic Pass-Gate Feedback

Whereas adaptive body biasing becomes less effective with bulk-Si MOSFET scaling [79], back-gate (BG) biasing of a thin-body MOSFET remains effective for dynamic control of V_{TH} with transistor scaling, and can provide improved control of short-channel effects as well [72]. The strong BG effect in FinFETs can thus be leveraged to optimize SRAM stability through a dynamic adjustment of the effective cell β -ratio.

Figure 5.8a presents the schematic of an independently-gated (IG) FinFET based 6-T SRAM cell with dynamic pass-gate feedback (PGFB) [30, 31, 63]. The storage nodes are connected to the BG of the pass-gate transistors to selectively decrease their current drive, and thus increasing the effective cell β -ratio - i.e. the logic '0', stored in node CL ,

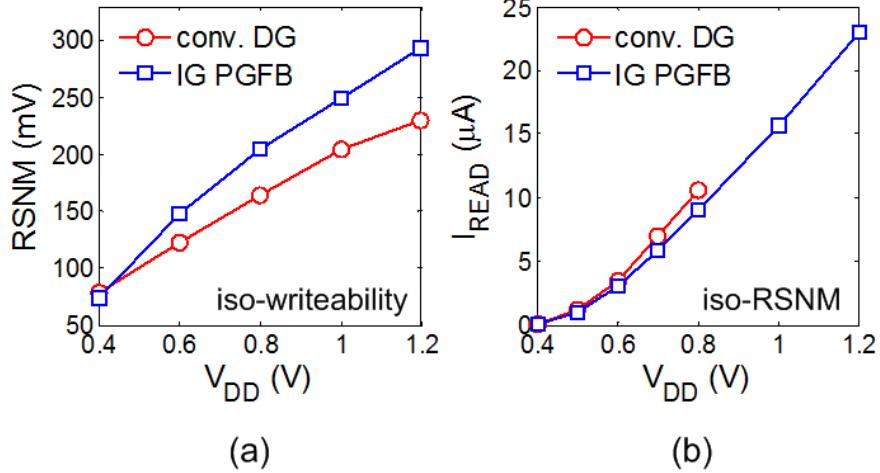


Figure 5.9: (a) Iso-writeability comparison of RSNM and (b) iso-RSNM comparison of I_{READ} , over a wide range of V_{DD} , between the conventional DG 6-T bitcell and the IG PGFB 6-T bitcell. Writeability and read stability are equalized at each supply voltage, between the two designs, using Φ_m tuning.

biases the BG of pass-gate transistor N_{AXL} ; thereby decreasing its strength relative to the pull-down transistor N_L . Because the cell retains its state during a read operation or a half-select condition, the higher β -ratio is maintained throughout the access, and the read static noise margin is enhanced. Figure 5.8b indicates a 71% improvement in RSNM over the conventional DG 6-T design with identical Φ_m at $V_{DD} = 1V$. During a write operation, the logical '1', stored in node CH , biases the BG of the pass-gate transistor and helps it to discharge the storage node until the cell state flips. Therefore, although the current drive of a BG-connected pass-gate transistor is expected to be lower than that of a DG-connected pass-gate transistor, the writeability of the IG PGFB 6-T bitcell does not degrade much over the DG 6-T design. In addition, the IG PGFB 6-T SRAM bitcell also benefits from improved soft-error reliability, due to the increased storage node capacitance. Furthermore, this simple BG connection can be made, by extending the gate-poly of the pass-gate transistor to the gate contact of the inverter-pair (Figure 5.8c), without incurring any area penalty over the conventional DG 6-T design - in fact, a 2% reduction is achieved due to the elimination of the 80nm gate-poly extension beyond the active region (fin), as required by the DG-connected pass-gate transistor, yielding a cell area of only $173F^2$.

Iso-writeability comparison of RSNM [31] between the conventional DG 6-T bitcell and the IG PGFB 6-T bitcell is presented in Figure 5.9a. Φ_m tuning is used on the conventional DG 6-T design to match the writeability of the IG PGFB 6-T design at each point over a wide V_{DD} range ($0.4V - 1.0V$). Results indicate significantly higher read stability, by $\sim 20\%$ (in RSNM), for the IG PGFB 6-T design at $V_{DD} > 0.4$. At $V_{DD} \leq 0.4V$, the higher Φ_m used for the DG 6-T design offsets the RSNM improvement of the dynamic PGFB. Nevertheless, Figure 5.9a indicates an improved read/write margin tradeoff for using the dynamic PGFB over Φ_m tuning. It should be noted here that the successful fabrication and measurement of an IG PGFB 6-T bitcell, based on this design, has been reported in [48].

Since the dynamic PGFB aims to reduce the pass-gate transistor strength at the

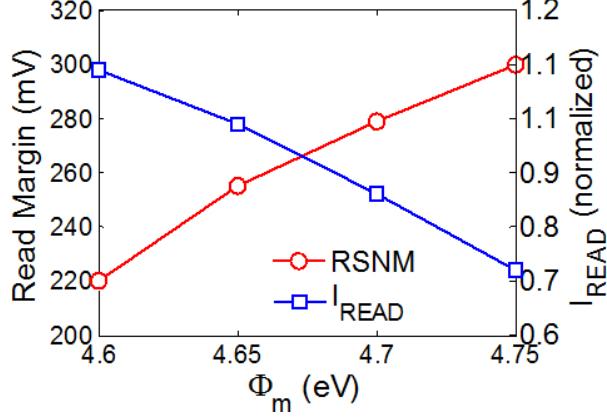


Figure 5.10: RSNM and I_{READ} , of an IG 6-T SRAM bitcell with dynamic PGFB, as a function of Φ_m . The values of I_{READ} are normalized to that of a conventional DG 6-T design with $\Phi_m = 4.75\text{eV}$.

'0' storage side, the cell I_{READ} is inevitably degraded. This presents a fundamental trade-off, to a certain extent, since a higher I_{READ} increases the charge (Q) to reverse the cell state. However, Figure 5.9b indicates that the degradation in I_{READ} is small, less than $\sim 15\%$, when compared to a conventional DG 6-T design with matched RSNM up to $V_{DD} = 0.8\text{V}$ ² [31]. This is because the lower Φ_m , in the IG PGFB 6-T design, reduces the V_{TH} of NMOS transistors and increases their current drive. In addition, the '0' storage node, in the IG PGFB 6-T design, stays closer to V_{SS} than the conventional DG 6-T design (Figure 5.8b); thus giving the BG-connected pass-gate transistor a higher gate-source overdrive. Furthermore, the read access time consists of two delay components - the word-line drive and the bit-line discharge. While the bit-line discharge is a function of the cell I_{READ} , the word-line drive depends on the word-line resistance and capacitance. Since only the front-gate (FG) of each pass-gate transistor is driven by the word-line, the word-line capacitance is lower for the IG PGFB 6-T design; thus reducing the word-line drive component of the read access delay.

If a better read performance is desired, Φ_m tuning can be used for the IG PGFB 6-T design to increase I_{READ} while still maintaining a high level of read stability. Figure 5.10 summarizes the impact of Φ_m tuning on the read stability and I_{READ} of an IG PGFB 6-T SRAM bitcell. As illustrated, with $\Phi_m = 4.65\text{eV}$, I_{READ} of the IG PGFB 6-T design can be made approximately equal to the conventional DG 6-T design, while still maintaining a $> 250\text{mV}$ RSNM. In addition, the writeability of the IG PGFB 6-T design also improves with a lower Φ_m - due to an increased PMOS V_{TH} and a decreased NMOS V_{TH} . Figure 5.11a illustrates the impact of Φ_m tuning on the read stability and the writeability of an IG PGFB 6-T SRAM bitcell. With $\Phi_m = 4.65\text{eV}$, a $\sim 250\text{mV}$ BWTV can be achieved simultaneously with a $> 250\text{mV}$ RSNM and a similar I_{READ} as the DG 6-T design (with $\Phi_m = 4.75\text{eV}$).

Alternatively, a column-based biasing technique (Section 4.6) [162] can be employed to independently enhance SRAM read stability and writeability using optimized cell supply

²Iso-RSNM comparisons of I_{READ} are not established beyond 0.8V as Φ_m tuning of the conventional DG 6-T design cannot achieve the same RSNM as the IG PGFB 6-T design for $V_{DD} > 0.8\text{V}$.

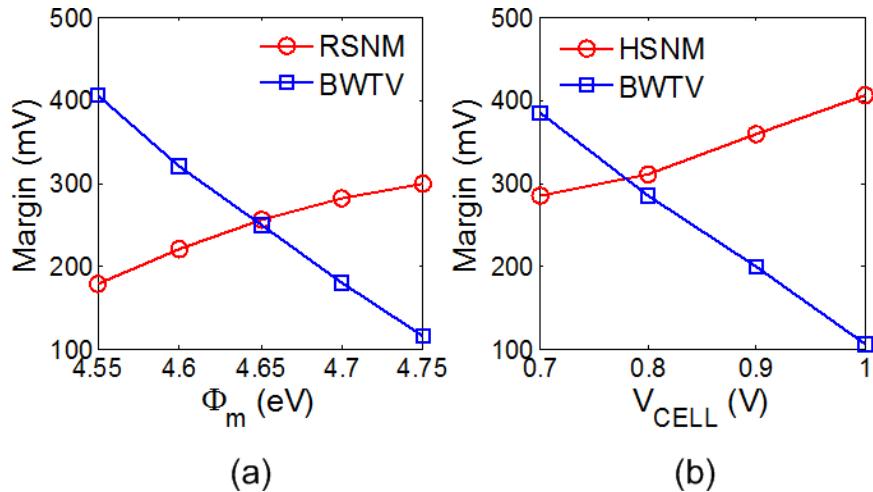


Figure 5.11: (a) RSNM and BWTM, of an IG 6-T SRAM bitcell with dynamic PGFB, as a function of Φ_m . (b) HSNM and BWTM, of an IG 6-T SRAM bitcell with dynamic PGFB, as a function of V_{CELL} .

voltages (V_{CELL}). Using this technique, the contention between read stability and writeability can be replaced by a trade-off between standby stability and writeability, which offers a much bigger window for optimization. Figure 5.11b summarizes the writeability enhancements, for an IG PGFB 6-T bitcell, by reducing V_{CELL} and the corresponding impact on the cell HSNM. The drawback of this method, however, is the need to generate and distribute two different voltages.

Array Design Considerations

Due to the BG-connected pass-gate transistors, the worst-case bit-line discharge condition may be exacerbated for the IG PGFB 6-T bitcell design. In a typical SRAM array, the worst-case bit-line discharge happens when the accessed bitcell stores the opposite polarity from all neighboring bitcells in the same column. The effective bit-line discharge current ($I_{BL,DISCHARGE}$) is simply $I_{READ} - \sum I_{LEAK,1} - \sum I_{LEAK,0}$ - where I_{READ} is the read current, at the '0' storage side, of the accessed bitcell; $\sum I_{LEAK,1}$ is the total bit-line leakage current at the '0' storage side of the accessed bitcell (for which all un-accessed bitcells store a '1'); and $\sum I_{LEAK,0}$ is the total bit-line leakage current at the '1' storage side of the accessed bitcell (for which all un-accessed bitcells store a '0'). For an SRAM array constructed using IG PGFB 6-T bitcells, $\sum I_{LEAK,1}$ may become significant at lower bit-line voltages as a logic '1' biases the BG of the pass-gate transistors. Specifically, the un-accessed pass-gate transistors experience a drive voltage of up to $V_{DD} - V_{BL}$ at the BG, which sources significant current onto the bit-line when $V_{BL} < V_{DD} - V_{TH,BG}$. This prevents the pass-gate transistors from completely shutting off; hence preventing the bit-line to discharge fully (to V_{SS}). Therefore, a sense amplifier is required to generate a zero voltage at the output. A good sense amplifier design typically requires only a small bit-line differential voltage (usually $\sim 10\%$ of V_{DD}) to determine the cell state, for which $\sum I_{LEAK,1}$ is small. Therefore, the degradation in the bit-line discharge speed should not be significant (for typical column heights). Figure 5.12a

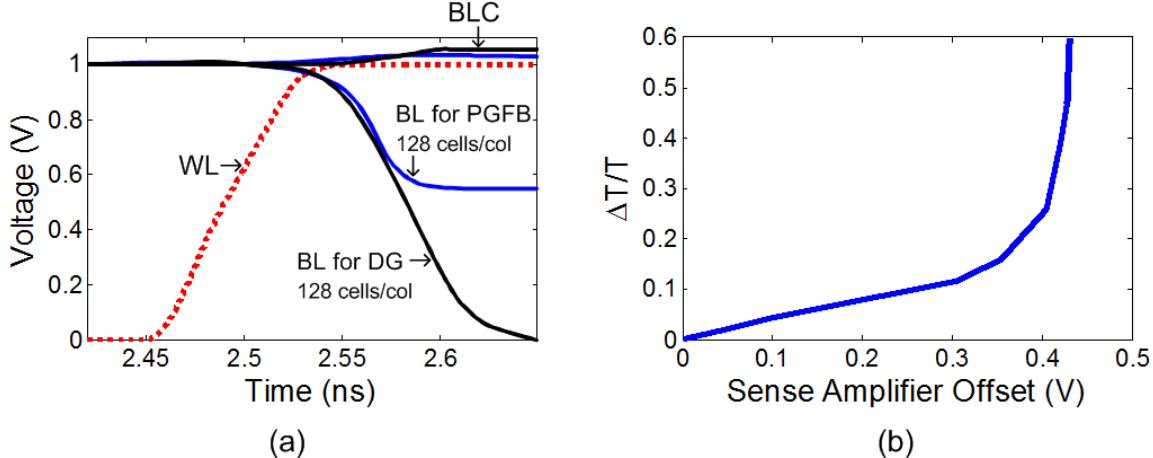


Figure 5.12: (a) Bit-line voltage simulations for the conventional DG 6-T SRAM design and the IG PGFB 6-T SRAM design, with 128 bitcells per column. (b) Impact of dynamic PGFB on sensing speed - where $\Delta T/T$ is the normalized difference in the bit-line discharging time between a conventional 6-T design and a IG PGFB 6-T designs; and the sense amplifier offset is the tolerable offset voltage at the inputs of the sense amplifier. Less than 5% impact on sensing speed is incurred when using sense amplifiers with less than 100mV offset voltage.

compares the worst-case bit-line discharging patterns for a conventional single-fin DG 6-T SRAM column and an IG PGFB 6-T SRAM column (with column heights of 128). Φ_m is set to 4.75eV for the DG bitcell and 4.65eV for the IG PGFB bitcell, to achieve similar I_{READ} . As expected, the bit-line is able to discharge fully for the DG design, whereas the bit-line can only discharge to just under 600mV for the IG PGFB design. However, the bit-line discharging speed is weakly affected by the dynamic PGFB connection. Figure 5.12b summarizes the impact of dynamic PGFB on the sensing speed. For an ideal sense amplifier with zero input offset, the slightest bit-line differential can be distinguished; hence ΔT approaches zero. Conventional latch-based sense amplifiers can achieve less than 100mV (10% of V_{DD}) offset voltage - the yield an optimization of latch type SRAM sense amplifier is discussed in [151]; and [95] extends this design to a double-gated FinFET technology using independent gating. Figure 5.12b indicates that less than 5% impact on the sensing speed is incurred, for the IG PGFB 6-T design, when using sense amplifiers with less than 100mV offset.

In both IG PGFB and conventional DG 6-T designs, the effective worst-case $I_{BL,DISCHARGE}$ is reduced with an increasing column height. On the other hand, decreasing the column height incurs more area overhead from the sense amplifiers. Column multiplexing can be used to optimize the array area efficiency by allowing the read/write circuitry and the sense amplifiers to be shared among multiple columns. However, the non-zero resistance of the bit-line multiplexers degrades the column performance.

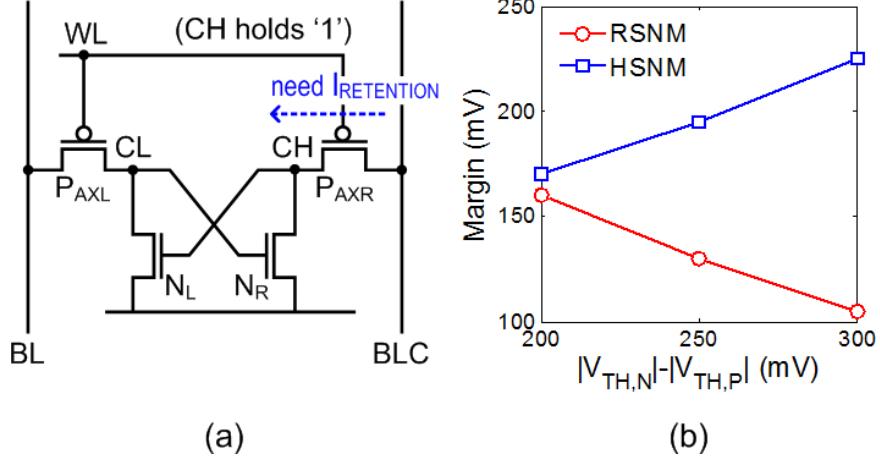


Figure 5.13: (a) Schematic for a conventional loadless 4-T SRAM bitcell. (b) RSNM and HSNM of a conventional loadless bulk-Si based 4-T SRAM bitcell, with β -ratio= 2, as a function of the difference in NMOS to PMOS V_{TH} .

5.4 4-T FinFET based SRAM design

To seek further enhancements in the array density, 4-T SRAM bitcells are considered. Figure 5.13a shows the schematic for a conventional loadless 4-T SRAM bitcell [87, 101]. It consists of a cross-coupled NMOS pair (N_L , N_R) and two PMOS pass-gate transistors (P_{AXL} , P_{AXR}) for read/write access. During standby, both bit-lines (BL and BLC) and the word-line are biased at V_{DD} , shutting off both PMOS pass-gate transistors. To retain the stored data, the PMOS pass-gate transistors must provide enough leakage current, commonly referred to as the data retention current ($I_{RETENTION}$), to compensate for all the leakage paths from the '1' storage node (CH). This is typically guaranteed by giving the PMOS pass-gate transistors a much lower V_{TH} than the NMOS pull-down transistors [101]. During the read operation, word-line is driven low (to V_{SS}), and the PMOS pass-gate transistors act as PMOS loads for the cross-coupled pseudo-NMOS inverters. Similar to the 6-T bitcell, stronger pass-gate transistors tend to destabilize the 4-T bitcell by causing the '0' storage node (CL) to rise above the inverter trip point (which is approximately equal to the V_{TH} of the NMOS pull-down transistors). As a result, a contention is formed between the cell read stability and the cell hold stability. Figure 5.13b summarizes the trade-off between the cell RSNM and HSNM, for a bulk-Si based 4-T bitcell, when adjusting the value of $|V_{TH,N}| - |V_{TH,P}|$, where $V_{TH,N}$ and $V_{TH,P}$ denote the V_{TH} of NMOS and PMOS transistors, respectively³. Results indicate that, even with β -ratio= 2, it is challenging to simultaneously achieve good read stability and hold stability for a bulk-Si based 4-T bitcell (in this process). In addition, although $I_{RETENTION}$ is only needed at the '1' storage node, a leaky PMOS pass-gate transistor at the '0' storage node draws significantly more current ($\gg I_{RETENTION}$) from the other bit-line due to its higher V_{DS} (approximately equal to V_{DD}) - recall that

³A β -ratio (denoting the size-ratio of the NMOS to PMOS transistors, in this case) of 2 is applied during the simulation, where the L_G of the PMOS pass-gate transistors are doubled.

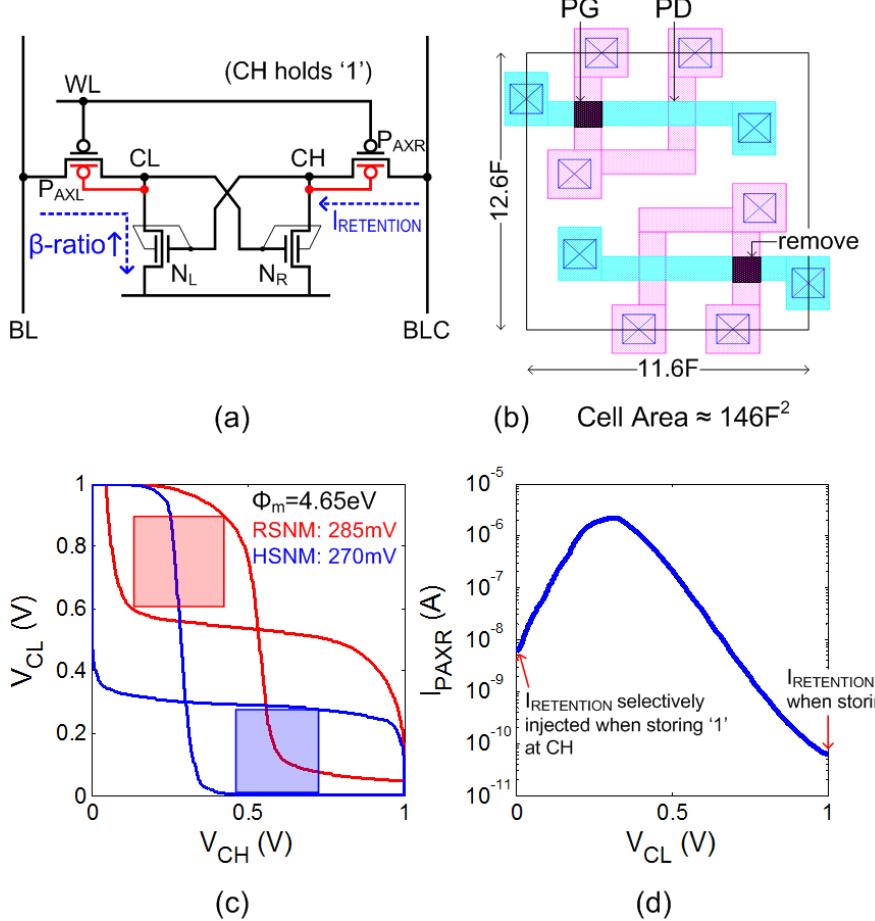


Figure 5.14: (a) Schematic and (b) layout for an IG FinFET based loadless 4-T SRAM bitcell with dynamic PGFB. Here, β -ratio = 1 is assumed. (c) Read and standby butterfly-curves for an IG FinFET based 4-T SRAM cell with dynamic PGFB (β -ratio = 1 and $\Phi_m = 4.65\text{eV}$), showing significantly improved RSNM and HSNM. (d) PMOS pass-gate current as a function of the opposing storage node voltage, illustrating the selective injection of $I_{\text{RETENTION}}$ when the storage node holds a '1'.

$I_{\text{LEAK}} \propto \left(1 - e^{\frac{-V_{DS}}{V_{TH}}}\right)$ [43]. This dramatically increases the static power consumption of the bulk-Si based 4-T bitcell, making it unsuitable for high-density and low-power applications.

The 4-T SRAM design presents another opportunity for the IG operation of the FinFET technology. It is shown [63, 157] that dynamic control of the PMOS V_{TH} can offer a means for selectively adjusting $I_{\text{RETENTION}}$, and also provides higher effective β -ratio for the 4-T SRAM design. Figure 5.14a presents the schematic for an IG FinFET based loadless 4-T SRAM bitcell with dynamic PGFB [63]. The storage nodes are connected to the BG of the pass-gate transistor on the opposite side to selectively reduce the V_{TH} of the pass-gate transistor at the '1' storage side; thus injecting $I_{\text{RETENTION}}$ only at the '1' storage node (Figure 5.14d). Consequently, the IG PGFB 4-T design can achieve a much lower static power consumption compared to the bulk-Si based 4-T design. In addition, leaky PMOS

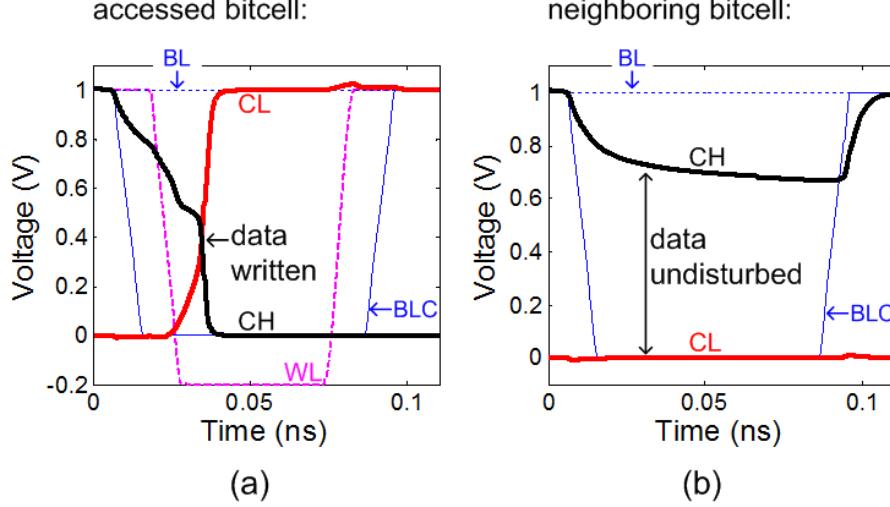


Figure 5.15: Write cycle simulation for the IG PGFB 4-T design (with $\Phi_m = 4.65eV$) illustrating (a) the successful write operation for an accessed bitcell and (b) the successful data retention for a neighboring bitcell (in the same column). $V_{WL} = -200mV$ is applied during the write cycle.

pass-gate transistors are no longer needed in the IG PGFB 4-T design; thus eliminating the contention between the cell read stability and the cell hold stability. Furthermore, the effective β -ratio is increased at the '0' storage side, as a logic '1' biases the BG of the PMOS pass-gate transistor on that side. Figure 5.14b presents the layout for the IG PGFB 4-T bitcell - a cell area of $146F^2$ is achieved, marking a $> 15\%$ reduction compared to the IG PGFB 6-T bitcell with comparable read stability (Figure 5.14a). It should be noted that the writeability of a 4-T bitcell is guaranteed due to its inherent instability during a write cycle.

Since retention of a logic '1', in 4-T SRAM bitcells, relies on a retention current that is sourced from the bit-line, that particular bit-line (at the '1' storage side) is necessarily biased at V_{DD} . However, during a write operation, the bit-lines are driven differentially, corresponding to the data input; thereby reversing the direction of $I_{RETENTION}$ for all un-accessed bitcells in the same column and storing the inverse data polarity (as the data input). In addition, the magnitude of the reversed $I_{RETENTION}$ is much higher due to the higher V_{DS} applied across the pass-gate transistor. As a result, the stability of all un-accessed bitcells in the same column (storing the inverse data polarity) is compromised. For a conventional bulk-Si based 4-T SRAM design, this neighbor write instability can be addressed through careful bit-line timing - i.e. as long as the bit-lines are restored to V_{DD} before the storage bit is flipped, the cell state can be preserved. The amount of time that either bit-line can stay low depends on the strength of the pass-gate transistor relative to the pull-down - the weaker the pass-gate transistor, the longer the bit-lines can stay discharged. However, weaker pass-gate transistors inevitably lead to increased write delays⁴, thus requiring the bit-lines to stay discharged for longer periods, and may also compromise the cell standby stability.

Because the fundamental condition of $|V_{TH,P}| < |V_{TH,N}|$ is no longer required for

⁴This can be avoided by pulling V_{WL} below V_{SS} during the write cycle; however, this compromises the half-select cell read stability.

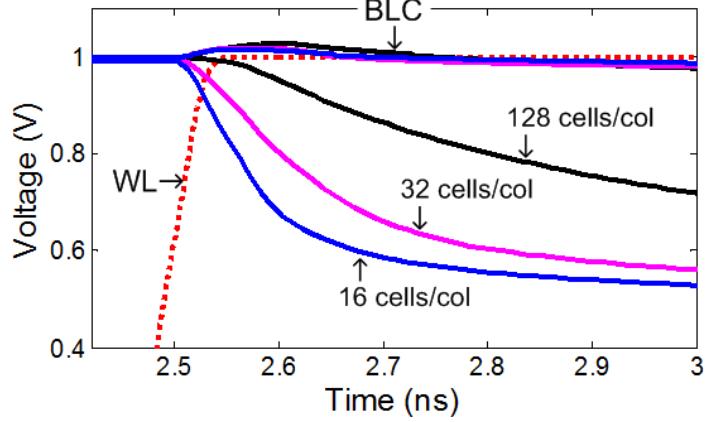


Figure 5.16: Bit-line voltage simulations for the IG PGFB 4-T SRAM design with varying column heights.

the IG PGFB 4-T design, a positive neighbor write stability margin can be achieved by giving the PMOS transistors a higher V_{TH} (compared to the NMOS transistors) - i.e. if $|V_{TH,P}| > |V_{TH,N}|$, the PMOS pass-gate transistor (discharging the '1' storage node) will shut off before the NMOS pull-down transistor (driven by the '1' storage node) stops conducting⁵. In addition, this margin can be optimized by adjusting $V_{TH,P}$ relative to $V_{TH,N}$. However, increasing $|V_{TH,P}|$ may lead to increased write delay (as mentioned previously). To address this, a negative V_{WL} can be used during the write cycle - although this may compromise the read stability of the half-select bitcells, the much better read stability of the IG PGFB 4-T design can still maintain adequate margin; additionally, increasing $|V_{TH,P}|$, relative to $|V_{TH,N}|$, also enhances RSNM, which reduces the impact of a negative V_{WL} on the read stability of the half-select bitcells. The write cycle simulation in Figure 5.15 illustrates the successful write operation (with $V_{WL} = -200\text{mV}$) for an accessed bitcell and the successful data retention for a neighboring bitcell in the same column.

The IG PGFB 4-T bitcell presents a similar trade-off as the IG PGFB 6-T bitcell in optimizing the column segmentation - i.e. the BG-connected PMOS pass-gate transistors also prevents the bit-lines from discharging fully. Figure 5.16 presents the bit-line voltage simulations for the IG PGFB 4-T SRAM design with varying column heights. Since PMOS pass-gate transistors are less efficient (compared to NMOS pass-gate transistors) at pulling down, due to a decreasing $|V_{GS}|$ as V_{BL} drops, the bit-lines discharge at a slower rate compared to the IG PGFB 6-T design⁶. In addition, the amount of bit-line discharge is smaller compared to the IG PGFB 6-T design. It should be noted that while the bit-lines can discharge to a lower voltage, approximately equal to $|V_{TH,P}|$, for a conventional bulk-Si based 4-T design, the discharge rate is expected to be significantly slower than a conventional 6-T design for similar reasons. Consequently, 4-T SRAM arrays are typically intended for lower performance, higher density applications [157]. Since these applications typically have very stringent power budgets, very low static power consumption is required. While the IG PGFB

⁵In other words, before reaching the inverter trip point of the other half-cell.

⁶For a given channel area, a PMOS transistor is also expected to deliver less on-current than an NMOS transistor.

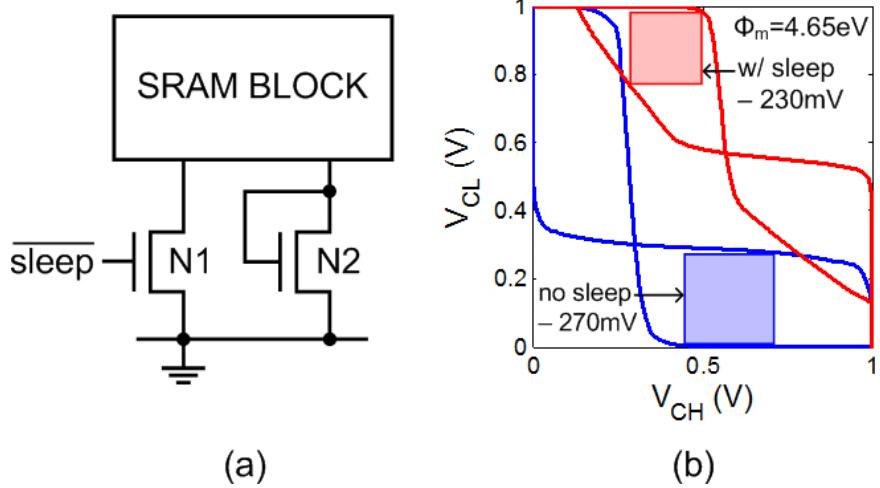


Figure 5.17: (a) Schematic for a gated- V_{SS} leakage reduction scheme. (b) The impact of leakage reduction on the HSNM of a IG PGFB 4-T SRAM bitcell.

Table 5.4: Simulated HSNM and per-cell standby leakage currents for the IG PGFB 4-T design.

Cell Design	HSNM(mV)	I _{CELL,STANDBY} (nA)
IG PGFB 4-T (no gated- V_{SS})	270	5.9
IG PGFB 4-T (w/ gated- V_{SS})	230	0.076

FinFET 4-T design achieves much lower static power consumption than the bulk-Si based 4-T design, an adequate $I_{RETENTION}$ places a lower bound on its static leakage current.

To achieve a much lower static power consumption suitable for low-power applications, a gated- V_{SS} leakage reduction scheme, using sleep transistors [143], can be adopted. Figure 5.17a shows the schematic of a simple gate- V_{SS} leakage reduction scheme using NMOS sleep transistors, integrated into each SRAM block, to control the SRAM V_{SS} . During standby, N1 is turned off and the SRAM V_{SS} is boosted by the V_{TH} of the diode-connected N2; whereas during a read/write operation, N1 is activated and the SRAM V_{SS} is pulled low (to the global V_{SS}). Figure 5.17b indicates that the implementation of the gated- V_{SS} leakage reduction scheme incurs less than a 15% degradation in the cell HSNM - achieving a 230mV HSNM when the gated- V_{SS} scheme is activated. Table 5.4 indicates that the gated- V_{SS} scheme can limit the per-cell standby leakage current, of the IG PGFB 4-T bitcell design, to under 80pA - justifying the usage of IG PGFB 4-T bitcells for low-power, high-density applications with lower-performance requirements.

5.5 Summary

This chapter evaluates FinFET based SRAM as an alternative in nanoscale memory design. Both 6-T and 4-T SRAM bitcells are analyzed using mixed-mode Taurus simulations. Results indicate that the FinFET technology can offer improved stability and flexibility over bulk-Si MOSFET in SRAM design and presents a promising device architecture for continued

Table 5.5: Cell area, RSNM (and HSNM), and per-cell standby leakage currents for various bitcell designs.

Cell Design	Area(F^2)	RSNM/HSNM(mV)	$I_{CELL,STANDBY}(nA)$
Bulk-Si 6-T (β -ratio = 1.5)	229	135	6.6
DG 6-T (1-fin PD)	177	175	0.19
DG 6-T (2-fin PD)	206	240	0.26
DG 6-T ($L_{G,pass-gate} = 2 \times L_{G,pull-down}$)	193	215	0.19
DG 6-T (PD fin-rotation)	201	200	0.19
IG PGFB 6-T	173	300	0.19
Bulk-Si 4-T ($\Delta V_{TH} = 250mV$)	210 ^a	130	132
IG PGFB 4-T (no gated- V_{SS})	146	285 / 270	5.9
IG PGFB 4-T (w/ gated- V_{SS})	146 ^b	285 / 130	0.076

SRAM scaling beyond the 22nm node. It is shown that a conventional DG 6-T FinFET based bitcell (with β -ratio = 1) can provide immediate improvement (30%) in RSNM over the bulk-Si counterpart (with β -ratio = 1.5), in addition to a more compact cell layout. The cell RSNM can be further improved by 71%, at little performance and zero area penalty, via a dynamic pass-gate feedback (PGFB) to adaptively adjust the pass-gate transistor strengths - achieving a 300mV RSNM, while keeping the per-cell standby leakage current at below 0.2nA (by using high V_{TH} - $\Phi_m = 4.75eV$). It is shown that 4-T SRAM design can also take advantage of the independently-gated (IG) operation of the FinFET technology. An IG PGFB 4-T FinFET SRAM cell can simultaneously achieve adequate read stability and hold stability margins, while dissipating only a fraction of the static power consumed by a bulk-Si based design. Compared to the IG PGFB 6-T design, the 4-T bitcell can achieve > 15% area reduction with less than 80pA/cell of leakage current during standby - making it extremely attractive for high-density, low-power embedded memory applications. Table 5.5 summarizes the cell area, the RSNM⁷, and the per-cell standby leakage current for the various bitcell designs. Furthermore, due to the elimination of RDF-induced σ_{VTH} , the FinFET based designs can also achieve better immunity against process variability than bulk-Si based designs (as illustrated in Figure 5.18). A simplified variability analysis is provided here - a more detailed analysis, using the concept of cell sigma, is presented in [31].

In addition to the bitcell designs presented in this chapter, several other variations of IG 6-T [26,30,31,48,74,82,108,109] and 8-T [74,83] FinFET based SRAM bitcells have since

⁷HSNM is also shown for the 4-T designs.

^aThe layout for this bitcell is not shown.

^bThere is a per-column area overhead for this implementation.

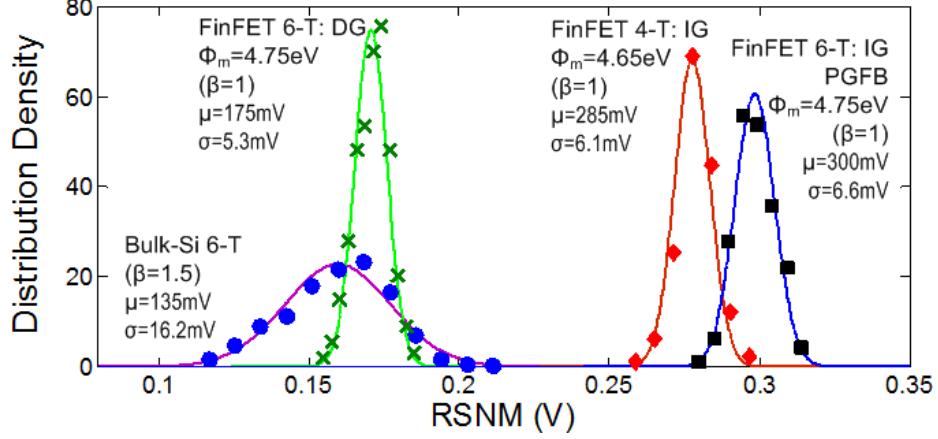


Figure 5.18: Impact of process variations on the cell RSNM for various bulk-Si and FinFET SRAM bitcells. The Monte Carlo (MC) simulations are run in mixed-mode using Taurus. Geometric variations in L_G and T_{Si} (with $3\sigma(L_G) = 3\sigma(T_{Si}) = 10\%L_G$) are considered for FinFETs, whereas only RDF is considered for bulk-Si MOSFETs [63]. The RSNM extracted for FinFET based designs show much tighter distributions (i.e. smaller σ) than that of the bulk-Si based design.

been studied in literature. In particular, an IG 6-T SRAM cell with pull-up write gating (PUWG), where a BG bias is applied to the PMOS pull-up transistors during the write cycle to enhance the cell writeability, is presented in [30, 31]. This technique can be implemented with zero cell area penalty and is shown to complement the PGFB technique to offer both enhanced read stability and writeability. Alternatively, IG 6-T bitcells implemented with a separate control signal to bias the BG of the pass-gate transistor [109], the pull-down transistor⁸ [26], or both [74, 108], have been investigated for stability enhancements and/or leakage reduction. Furthermore, the IG operation is adopted for the read path transistors ($N1$ and $N2$ in Figure 5.1a) of an 8-T dual-port SRAM bitcell, in [74], to enhance the read access performance without increasing leakage power.

Endo *et al.*, in [48], presented successfully fabricated IG FinFET SRAM bitcells based on four different designs - IG 6-T with PGFB (referred to as PG-SN), IG 6-T with BG of the pass-gate connected to the opposite storage node (PG-OSN), IG 6-T with a common bias for the BG of both pull-down and pass-gate transistors (Flex- V_{TH} - similar to [74, 108]), and IG 6-T with BG bias only for the pass-gate transistors (Flex-PG - similar to [109]). It is shown that the Flex- V_{TH} design can provide dramatic leakage current reduction while (almost) maintaining the read and write margins (compared to a conventional DG 6-T design). The Flex-PG design is shown to achieve excellent trade-off between read and write margins; however, similar to the Flex- V_{TH} design, it requires the generation and routing of an additional bias voltage. While the PG-OSN design can achieve the best read margin, it has a negative write margin, at $V_{DD} = 1.0V$, due to a very weak pass-gate transistor (at the '1' storage side). Finally, the PG-SN (or the IG PGFB 6-T) design demonstrates excellent read margin (although not as good as the PG-OSN design), with a moderate write margin,

⁸In this design, the BG of the pull-down transistor is biased at V_{SS} for leakage reduction.

which is shown to significantly improve using the CVD assist scheme (Section 4.6 and Figure 5.11b).

While successfully fabricated FinFET based bitcells have been demonstrated [18, 48], the fin LER still presents a significant challenge - although it can be mitigated by using spacer lithography [45]. More recently, Delprat *et al.* and SOITEC have demonstrated, in [44], the capability to uniformly manufacture 15nm thick SOI layers with $\pm 0.5\text{nm}$ thickness variation, showing readiness for the 22nm technology node. According to [86, 126], FD-SOI transistors can demonstrate steeper subthreshold slope and reduced RDF than the FinFET, making FD-SOI a promising candidate for enabling embedded SRAM scaling beyond the 22nm node. Although FD-SOI may offer less design flexibility, compared to the DG FinFET, substrate biasing techniques can still be used to implement the Flex-PG or the Flex- V_{TH} bitcells. While a general consensus has yet to be reached between FinFET and PD-SOI, one thing is for sure - the continuation of SRAM scaling until the end of the roadmap will require technology and circuit co-design.

Chapter 6

Conclusion

Continued increase in the process variability is perceived to be a major roadblock for future technology scaling. Its impact is particularly pronounced in large memory arrays due to both the utilization of minimum sized transistors and their extremely large data capacity. Therefore, memory design presents an extreme example of variability-aware design. To satisfy the functionality of hundreds of millions of SRAM cells in current on-die cache memories, the design has to provide more than 6 standard deviations of margin to parameter variations. This is becoming increasingly challenging to satisfy, and presents a major problem for continued scaling of memory density. Concurrently, high-end μ Ps have been increasing the amount of on-die cache to improve the performance - e.g. the current-generation AMD Opteron [130] and Intel Xeon [131] server μ Ps feature over 10^7 and 10^8 bitcells, respectively, on the L3 cache. Increased size of cache arrays requires accounting for even wider process extremes in the design. Consequently, the ability to monitor and characterize (on-chip) the variations in SRAM functionality and performance becomes critical for both gaining a deeper understanding of the sources of variability and for developing more robust circuits and topologies for the next-generation embedded SRAM memory.

6.1 Key Contributions

This work encompasses three key contributions to facilitate the variability-aware design of embedded SRAM:

- ◊ **A methodology to characterize, directly, the impact of process variability on the functionality of large SRAM-based cache memories is developed.**

The large-scale characterization of SRAM variability is attractive for early stages of SRAM development due to its ability to capture massive statistical data at a very low design and area overhead, compared to the conventional method. Due to its low overhead, this methodology can also be implemented, occasionally, on a working chip to monitor the process variability. In addition, the large-scale characterization method can complement standard SRAM built-in self test (BIST) methods by correlating BIST failures to the measured bit cell read/write margins. Furthermore, irregular bit cell characteristics measured through direct bit-line access can be mapped to the cell location and verified using nano-probing to determine its source. This methodology is

further extended for the characterization of SRAM V_{MIN} during read and write cycles. As a result, a direct correlation between measured SRAM read/write margins and the per-cell V_{MIN} in a functional SRAM array is established. Due to the excellent agreement between the measured read/write margins and the per-cell V_{MIN} near the regions of read stability/writeability failure, quick and accurate V_{MIN} estimation using the measured large-scale read/write metrics - in particular, SRRV and WWTW - is possible and presented. The large-scale characterization of functional SRAMs can also efficiently evaluate different assist schemes and identify sources of systematic mismatch within the die, from die to die, and from wafer to wafer. Moreover, this method can be easily extended to capture more than 6 standard deviations of parameter variations by increasing the SRAM array size, and therefore can serve as a valuable addition to the next-generation SRAM development vehicle.

- ◊ The correlations between the various conventional and large-scale metrics, as well as per-cell V_{MIN} , are studied, in detail, through Monte Carlo simulations and chip measurements; and speculations for the utility of the different metrics for V_{MIN} estimation are made.

A close examination of the different conventional SRAM read stability and writeability metrics reveals that, while the various metrics share the same zero crossing, they may have very poor correlations at higher supply voltages - this was demonstrated, particularly, between the classical RSNM and the N-curve read metrics, SINM and SPNM. In addition, the various metrics were compared against the per-cell V_{MIN} , characterized for both read and write cycles. RSNM and WNM are shown to have excellent correlations against $V_{MIN,RD}$ and $V_{MIN,WRT}$ at moderately low supply voltages. The correlation between WNM and $V_{MIN,WRT}$ suffer at higher supply voltages due to an extraction error for WNM, caused by the inability to exhibit convexity in the write VTC curves. SINM and SPNM, on the other hand, are shown to have poor correlations against $V_{MIN,RD}$ even at low supply voltages. While I_W demonstrates better correlation against both WNM and $V_{MIN,WRT}$ at moderately low supply voltages, its distribution deviates dramatically from Gaussian and becomes log-normal at low supply voltages - making it unsuitable for $V_{MIN,WRT}$ estimation using the method presented in Chapter 4¹. Although SVNM exhibit good correlations against both RSNM and $V_{MIN,RD}$ and its distribution does not deviate significantly from Gaussian, its inability to quantize a negative read margin leads to inaccurate $V_{MIN,RD}$ estimation.

On the other hand, excellent correlations are established between the large-scale read/write metrics and the conventional RSNM/WNM, and between the large-scale read/write metrics and the per-cell V_{MIN} (as mentioned above). The large-scale metrics, particularly SRRV and WWTW, are shown to be excellent candidates for V_{MIN} estimation. This investigation provides valuable understanding of the different read stability and writeability metrics, as well as how, and whether, each metric should be used for yield predication.

¹ Aside from the poor correlation against $V_{MIN,RD}$, SINM and SPNM are also unsuitable for $V_{MIN,RD}$ estimation for this same reason.

- ◊ New SRAM bitcell designs, using thin-body double-gated (DG) FinFETs, are proposed.

As the $1/\sqrt{W \times L}$ dependent RDF-induced σ_{VTH} has been, and is expected to continually, get worse with the scaling of bulk-Si MOSFETs, FinFET based SRAM is investigated as an alternative in nanoscale memory design. It is shown that the FinFET technology can offer improved stability and flexibility over bulk-Si MOSFET in SRAM design and presents a promising device architecture for continued SRAM scaling beyond the 22nm node. Due to a more efficient control of the SCE via a thin-body, a conventional DG 6-T FinFET based bitcell can provide immediate improvement in RSNM over the bulk-Si counterpart, in addition to a more compact cell layout. The cell RSNM can be further improved, at little performance and zero area penalty, via a dynamic pass-gate feedback (PGFB) to adaptively adjust the pass-gate transistor strengths. The successful fabrication of this bitcell design has been reported in [48]. While 4-T SRAM design in bulk-Si suffers from significantly enhanced leakage currents and a small design window (to achieve both adequate read and standby stability), the independently-gated (IG) operation of the FinFET technology can enable the practical design of a 4-T SRAM cell. It is shown that an IG PGFB 4-T FinFET SRAM cell can simultaneously achieve adequate read stability and hold stability margins, in addition to a more compact layout, while dissipating only a fraction of the static power consumed by a bulk-Si based design - making it extremely attractive for high-density, low-power embedded memory applications. Furthermore, a Monte Carlo analysis shows that FinFET based bitcell designs can indeed achieve tighter noise margin distributions over the bulk-Si designs, due to an elimination of the RDF-induced σ_{VTH} . The continuation of SRAM scaling until the end of the roadmap will require technology and circuit co-design. The FinFET technology is particularly attractive for nanoscale SRAM design not only for its reduced σ_{VTH} and better SCE control, but also for the architectural flexibility enabled by its unique IG operation.

6.2 Future Work

While this work on the large-scale characterization of functional SRAMs may lay an important foundation for enabling the next-generation SRAM development, some improvements are still needed. To speed up the characterization process and be included on a functional chip, the characterization process should be fully automated. While an on-chip DAC is implemented in this prototype, its utility is limited without an on-chip current monitor. A method to perform on-chip leakage current measurements using a single-slope analog-to-digital converter (ADC) is developed in [110, 111]²; this design can be improved to achieve a larger range in the measurable current to complement the on-chip DAC. In addition, a register bank will be needed as a temporary storage for the measured data.

Although static DC stability metrics can offer good estimates for SRAM functionality, they cannot replace the more realistic dynamic stability metrics. Consequently, there has

²This leakage current monitor is actually implemented on the same (first) 45nm test chip [112] presented in this work.

been quite a few recent studies on the definition and modeling of dynamic SRAM stability [15, 47, 80, 146, 161]; however, a clear consensus has yet to emerge and more studies against silicon measurements are needed. Currently, work is being conducted to implement dynamic stability characterization methods on silicon, including a novel approach using tunable ring oscillators [141], within our group at the University of California, Berkeley.

While successfully fabricated FinFET SRAM bitcells have been reported recently [18, 48], and spacer lithography techniques [33, 45] have been proposed to further reduce CD variability, high-volume manufacturing of FinFET based designs is still nowhere in sight. Meanwhile, 6-T SRAM design using a thin-BOX FD-SOI process has been recently investigated [86, 126]. While both FinFET and thin-BOX FD-SOI technologies have their respective advantages over the other, it is not yet clear which technology will emerge first. In addition, while both technologies indicate clear advantages over the bulk-Si MOSFETs with respect to SCE and RDF, new processes will inevitably bring about new sources of variability, which may offset the benefits of the new processes until they become mature.

6.3 Final Words

Process variations are here to stay and will forever impact the ways that circuits are designed. However, as more problems arise, even more solutions are being proposed. Designers have long projected the limits for scaling [25, 54, 137], yet breakthroughs are, and will continually be, made to extend it [23, 129, 132]; while fluctuation limits have been predicted [22], techniques for fluctuation tolerance are being implemented [20]. Although the road for technology scaling has narrowed by its challenges, it seems that research will continue to dig a path forward.

Bibliography

- [1] P. A and M. Sachdev, *CMOS SRAM Circuit Design and Parametric Test in Nano-Scaled Technologies: Process-Aware SRAM Design and Test*, ser. Frontiers in Electronic Testing. Dordrecht, Netherlands: Springer Netherlands, 2008, vol. 40.
- [2] J. Abraham, “Overcoming timing, power bottlenecks,” *EE Times*, Apr. 2003.
- [3] K. Agawa, H. Hara, T. Takayanagi, , and T. Kuroda, “A bitline leakage compensation scheme for low-voltage SRAMs,” *IEEE Journal of Solid-State Circuits*, vol. 36, no. 5, pp. 726–734, May 2001.
- [4] M. Agostinelli, J. Hicks, J. Xu, B. Woolery, K. Mistry, K. Zhang, S. Jacobs, J. Jopling, W. Yang, B. Lee, T. Raz, M. Mehalel, P. Kolar, Y. Wang, J. Sandford, D. Pivin, C. Peterson, M. DiBattista, S. Pae, M. Jones, S. Johnson, and G. Subramanian, “Erratic fluctuations of SRAM cache VMIN at the 90nm process technology node,” in *IEEE International Electron Devices Meeting Tech. Dig.*, Washington, DC, Dec. 2005, pp. 655–658.
- [5] R. Aitken and S. Idgunji, “Worst-case design and margin for embedded SRAM,” in *Design, Automation, and Test in Europe Conference and Exhibition*, Nice, France, Apr. 2007, pp. 1289–1294.
- [6] Y. Aoki, T. Toyabe, S. Asai, and T. Hagiwara, “CASTAM: a process variation analysis simulator for MOS LSI’s,” *IEEE Transactions on Electron Devices*, vol. 31, pp. 1462–1467, Oct. 1984.
- [7] U. Arslan, M. P. McCartney, M. Bhargava, X. Li, K. Mai, and L. T. Pileggi, “Variation-tolerant SRAM sense-amplifier timing using configurable replica bitlines,” in *IEEE Custom Integrated Circuits Conference*, San Jose, CA, Sep. 2008, pp. 415–418.
- [8] I. Arsovski and R. Wistort, “Self-referenced sense amplifier for across-chip-variation immune sensing in high-performance content-addressable memories,” in *IEEE Custom Integrated Circuits Conference*, San Jose, CA, Sep. 2006, pp. 453–456.
- [9] A. Asenov, “Random dopant induced threshold voltage lowering and fluctuations in sub-0.1um MOSFETs: A 3-D atomistic simulation study,” *IEEE Transactions on Electron Devices*, vol. 45, pp. 2505–2513, Dec. 1998.
- [10] ———, “Simulation of statistical variability in nano MOSFETs,” in *Symposium on VLSI Technology Dig. of Tech. Papers*, Kyoto, Japan, Jun. 2007, pp. 86–87.

- [11] A. Asenov, S. Kaya, and J. H. Davies, “Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations,” *IEEE Transactions on Electron Devices*, vol. 49, pp. 112–119, Jan. 2002.
- [12] P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, M. Hussein, J. Hwang, D. Ingerly, R. James, J. Jeong, C. Kenyon, E. Lee, S.-H. Lee, N. Lindert, M. Liu, Z. Ma, T. Marieb, A. Murthy, R. Nagisetty, S. Natarajan, J. Neirynck, A. Ott, C. Parker, J. Sebastian, R. Shaheed, S. Sivakumar, J. Steigerwald, S. Tyagi, C. Weber, B. Woolery, A. Yeoh, K. Zhang, and M. Bohr, “A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and $0.57 \mu\text{m}^2$ SRAM cell,” in *IEEE International Electron Devices Meeting Tech. Dig.*, San Francisco, CA, Dec. 2004, pp. 657–660.
- [13] S. Balasubramanian, “Nanoscale thin-body MOSFET design and applications,” Ph.D. dissertation, University of California, Berkeley, Berkeley, CA, 2006.
- [14] M. Ball, J. Rosal, R. McKee, W. Loh, T. Houston, R. Garcia, J. Raval, D. Li, R. Hollingsworth, R. Gury, R. Eklund, J. Vaccani, B. Castellano, F. Piacibello, S. Ashburn, A. Tsao, A. Krishnan, J. Ondrussek, and T. Anderson, “A screening methodology for VMIN drift in SRAM arrays with applications to sub-65nm nodes,” in *Proc. IEEE International Symposium on Quality of Electronic Design*, San Francisco, CA, Dec. 2006, pp. 1–4.
- [15] A. Bansal, R. Rao, J.-J. Kim, S. Zafar, J. H. Stathis, and C.-T. Chuang, “Impacts of NBTI and PBTI on SRAM static/dynamic noise margins and cell failure probability,” *Microelectronics Reliability*, vol. 16, no. 12, pp. 642–649, Jun. 2009.
- [16] J. Barth, W. Reohr, P. Parries, G. Fredeman, J. Golz, S. Schuster, R. Matick, H. Hunter, C. Tanner, J. Harig, H. Kim, B. Khan, J. Griesemer, R. Havreluk, K. Yanagisawa, T. Kirihata, and S. S. Iyer, “A 500MHz random cycle 1.5ns-latency, SOI embedded DRAM macro featuring a 3T micro sense amplifier,” in *IEEE International Solid-State Circuits Conference Dig. of Tech. Papers*, San Francisco, CA, Feb. 2007, pp. 486–617.
- [17] A. P. Basu and J. K. Ghosh, “Identifiability of the multinormal and other distributions under competing risks model,” *Journal of Multivariate Analysis*, vol. 8, pp. 413–429, Sep. 1978.
- [18] F. Bauer, K. von Arnim, C. Pacha, T. Schulz, M. Fulde, A. Nackaerts, M. Jurczak, W. Xiong, K. T. San, C.-R. Cleavelin, K. Schrufer, G. Georgakos, and D. Schmitt-Landsiedel, “Layout options for stability tuning of SRAM cells in multi-gate-FET technologies,” in *European Solid-State Circuits Conference*, Munich, Germany, Sep. 2007, pp. 392–395.
- [19] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer, “High-performance CMOS variability in the 65-nm regime and beyond,” *IBM Journal of Research and Development*, vol. 50, no. 4/5, pp. 433–449, Jul. 2006.

- [20] A. J. Bhavnagarwala, S. Kosonocky, C. Radens, Y. Chan, K. Stawiasz, U. Srinivasan, S. Kowalczyk, and M. Ziegler, “A sub-600mV, fluctuation tolerant 65nm CMOS SRAM array with dynamic cell biasing,” *IEEE Journal of Solid-State Circuits*, vol. 43, no. 4, pp. 946–955, Apr. 2008.
- [21] A. J. Bhavnagarwala, S. Kosonocky, C. Radens, K. Stawiasz, R. Mann, Q. Ye, and K. Chin, “Fluctuation limits and scaling opportunities for CMOS SRAM cells,” in *IEEE International Electron Devices Meeting Tech. Dig.*, Washington, DC, Dec. 2005, pp. 659–662.
- [22] A. J. Bhavnagarwala, X. Tang, and J. D. Meindl, “The impact of intrinsic device fluctuations on CMOS SRAM cell stability,” *IEEE Journal of Solid-State Circuits*, vol. 36, no. 4, pp. 946–955, Apr. 2001.
- [23] M. Bohr, “The new era of scaling in an SoC world,” in *IEEE International Solid-State Circuits Conference Dig. of Tech. Papers*, San Francisco, CA, Feb. 2009, pp. 23–28.
- [24] ——, “Presentation: Silicon technology for 32nm and beyond System-on-Chip (SoC) products,” in *Intel Developer Forum*, San Francisco, CA, Sep. 2009.
- [25] S. Borkar, “Design challenges of technology scaling,” *IEEE Micro*, vol. 19, no. 4, pp. 23–29, Jul. 1999.
- [26] T. Cakici, K. Kim, and K. Roy, “FinFET based SRAM design for low standby power applications,” in *Proc. IEEE International Symposium on Quality of Electronic Design*, San Jose, CA, Mar. 2007, pp. 127–132.
- [27] B. H. Calhoun and A. P. Chandrakasan, “Static noise margin variation for sub-threshold SRAM in 65-nm CMOS,” *IEEE Journal of Solid-State Circuits*, vol. 41, no. 7, pp. 1673–1679, Jul. 2006.
- [28] Calibre®, Mentor Graphics®.
- [29] C. L. Cam, F. Guyader, C. de Buttet, P. Guyader, G. Ribes, M. Sardo, S. Vanbergue, F. Boeuf, F. Arnaud, E. Josse, and M. Haond, “A low cost drive current enhancement technique using shallow trench isolation induced stress for 45-nm node,” in *Symposium on VLSI Technology Dig. of Tech. Papers*, Honolulu, HI, Jun. 2006, pp. 82–83.
- [30] A. Carlson, Z. Guo, S. Balasubramanian, L.-T. Pang, T.-J. King, and B. Nikolić, “FinFET SRAM with enhanced read/write margins,” in *Proc. IEEE International SOI Conference*, Niagara Falls, NY, Oct. 2006, pp. 105–106.
- [31] A. Carlson, Z. Guo, S. Balasubramanian, R. Zlatanovici, T.-J. K. Liu, and B. Nikolić, “SRAM read/write margin enhancements using FinFETs,” *IEEE Transactions on VLSI Systems*, 2009.
- [32] A. Carlson, Z. Guo, L.-T. Pang, T.-J. K. Liu, and B. Nikolić, “Compensation of systematic variations through optimal biasing of SRAM word-lines,” in *IEEE Custom Integrated Circuits Conference*, San Jose, CA, Sep. 2008, pp. 411–414.

- [33] A. E. Carlson, “Device and circuit techniques for reducing variation in nanoscale SRAM,” Ph.D. dissertation, University of California, Berkeley, Berkeley, CA, 2008.
- [34] J. Chambers, W. Cleveland, B. Kleiner, and P. Tukey, *Graphical Methods for Data Analysis*. Wadsworth, 1983.
- [35] I. Chang, J.-J. Kim, S. P. Park, and K. Roy, “A 32kb 10T subthreshold SRAM array with bit-interleaving and differential read scheme in 90nm CMOS,” in *IEEE International Solid-State Circuits Conference Dig. of Tech. Papers*, San Francisco, CA, Feb. 2008, pp. 388–622.
- [36] L. Chang, M. Ieong, and M. Yang, “CMOS circuit performance enhancement by surface orientation optimization,” *IEEE Transactions on Electron Devices*, vol. 51, pp. 1621–1627, Oct. 2004.
- [37] 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, K. W. Guarini, and W. Haensch, “Stable SRAM cell design for the 32nm node and beyond,” in *Symposium on VLSI Technology Dig. of Tech. Papers*, Kyoto, Japan, Jun. 2005, pp. 14–16.
- [38] L. Chang, Y. Nakamura, R. K. Montoye, J. Sawada, A. K. Martin, K. Kinoshita, F. H. Gebara, K. B. Agarwal, D. J. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek, “A 5.3GHz 8T-SRAM with operation down to 0.41V in 65nm CMOS,” in *Symposium on VLSI Circuits Dig. of Tech. Papers*, Kyoto, Japan, Jun. 2007, pp. 252–253.
- [39] Y.-K. Choi, K. Asano, N. Lindert, V. Subramanian, T.-J. K. Liu, J. Bokor, and C. Hu, “Ultrathin-body SOI MOSFET for deep-sub-tenth micron era,” *IEEE Electron Device Letters*, vol. 21, no. 5, pp. 254–255, May 2000.
- [40] Y.-K. Choi, T.-J. K. Liu, and C. Hu, “A spacer patterning technology for nanoscale CMOS,” *IEEE Transactions on Electron Devices*, vol. 249, number =.
- [41] Y. Chung and S.-W. Shim, “An experimental 0.8V 256-kbit SRAM macro with boosted cell array scheme,” *RTRI Journal*, vol. 29, no. 4, pp. 457–462, Aug. 2007.
- [42] H. A. David and H. N. Nagaraja, *Order Statistics 3rd Ed.* New York, NY: John Wiley & Sons, 2003.
- [43] V. De, Y. Ye, A. Keshavarzi, S. Narendra, J. Kao, D. Somasekhar, R. Nair, and S. Borkar, “Techniques for leakage power reduction,” in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, W. Bowhill, and F. Fox, Eds. Piscataway, NJ: IEEE Press, 2001.
- [44] D. Delprat, F. Boedt, C. David, P. Reynaud, A. Alami-Idrissi, D. Landru, C. Girard, and C. Maleville, “SOI substrate readiness for 22/20 nm and for fully depleted planar device architectures,” in *Proc. IEEE International SOI Conference*, Foster City, CA, Oct. 2009, pp. 1–4.

- [45] A. Dixit, K. G. Anil, E. Baravelli, P. Roussel, A. Mercha, C. Gustin, M. Bamal, E. Grossar, R. Rooyackers, E. Augendre, M. Jurczak, S. Biesemans, , and K. D. Meyer, “Impact of stochastic mismatch on measured SRAM performance of FinFETs with resist/spacer-defined fins: Role of line-edge roughness,” in *IEEE International Electron Devices Meeting Tech. Dig.*, San Francisco, CA, Dec. 2006, pp. 709–712.
- [46] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, “Breaking the simulation barrier: SRAM evaluation through norm minimization,” in *IEEE/ACM International Conference Computer-Aided Design*, San Jose, CA, Nov. 2008, pp. 322–329.
- [47] W. Dong, P. Li, and G. M. Huang, “SRAM dynamic stability: Theory, variability and analysis,” in *IEEE/ACM International Conference Computer-Aided Design*, San Jose, CA, Nov. 2008, pp. 378–385.
- [48] K. Endo, S.-I. Ouchi, Y. Ishikawa, Y. Liu, T. Matsukawa, K. Sakamoto, J. Tsukada, K. Ishii, H. Yamauchi, E. Suzuki, and M. Masahara, “Enhancing SRAM cell performance by using independent double-gate FinFET,” in *IEEE International Electron Devices Meeting Tech. Dig.*, San Francisco, CA, Dec. 2008, pp. 1–4.
- [49] M. Ercken, L. Leunissen, I. Pollentier, G. P. Patsis, V. Constantoudis, and E. Gogolides, “Effects of different processing conditions on line edge roughness for 193nm and 157nm resists,” in *Metrology, Inspection, and Process Control for Microlithography XVIII*, ser. Proc. SPIE, R. M. Silver, Ed., Feb. 2004, vol. 5375, pp. 266–275.
- [50] D. Fang, “A mixed signal interface for maskless lithography,” Master’s thesis, University of California, Berkeley, Berkeley, CA, 2004.
- [51] D. Fang, R. Roberts, and B. Nikolić, “A 6-b DAC and analog DRAM for a maskless lithography interface in 90 nm CMOS,” in *Proc. IEEE Asian Solid-State Circuits Conference*, Hangzhou, China, Nov. 2006, pp. 423–426.
- [52] T. Fischer, D. Amirante, P. Huber, T. Nirschl, A. Olbrich, M. Ostermayr, and D. Schmitt-Landsiedel, “Analysis of read current and write trip voltage variability from a 1-MB SRAM test structure,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 21, pp. 534–541, Nov. 2008.
- [53] T. Fischer, C. Otte, D. Schmitt-Landsiedel, E. Amirante, A. Olbrich, P. Huber, M. Ostermayr, T. Nirschl, and J. Einfeld, “A 1 Mbit SRAM test structure to analyze local mismatch beyond 5 sigma variation,” in *IEEE International Conference on Microelectronic Test Structures*, Tokyo, Japan, Mar. 2007, pp. 63–66.
- [54] D. J. Frank, R. H. Dennard, E. J. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong, “Device scaling limits of Si MOSFETs and their application dependencies,” *Proc. of the IEEE*, vol. 89, no. 3, pp. 259–288, Mar. 2001.
- [55] D. J. Frank, Y. Taur, M. Ieong, , and H.-S. P. Wong, “Monte Carlo modeling of threshold variation due to dopant fluctuations,” in *Symposium on VLSI Circuits Dig. of Tech. Papers*, Kyoto, Japan, Jun. 1999, pp. 171–172.

- [56] G. Gasiot, D. Giot, and P. Roche, “Alpha-induced multiple cell upsets in standard and radiation hardened SRAMs manufactured in a 65 nm CMOS technology,” *IEEE Transactions on Nuclear Science*, vol. 53, pp. 3479–3486, Dec. 2006.
- [57] L. Ge and J. G. Fossum, “Analytical modeling of quantization and volume inversion in thin Si-film DG MOSFETs,” *IEEE Transactions on Electron Devices*, vol. 49, pp. 287–294, Feb. 2002.
- [58] N. Gierczynski, B. Borot, N. Planes, and H. Brut, “A new combined methodology for write-margin extraction of advanced SRAM,” in *IEEE International Conference on Microelectronic Test Structures*, Tokyo, Japan, Mar. 2007, pp. 97–100.
- [59] B. L. Gratiet, P. Gouraud, E. Aparicio, L. Babaud, K. Dabertrand, M. Touchet, S. Kremer, C. Chaton, F. Foussadier, F. Sundermann, J. Massin, J.-D. Chapon, M. Gatefait, B. Minghetti, J. de Caunes, and D. Boutin, “Process control for 45nm CMOS logic gate patterning,” in *Metrology, Inspection, and Process Control for Microlithography XXII*, ser. Proc. SPIE, J. A. Allgair and C. J. Raymond, Eds., Mar. 2008, vol. 6922, p. 6922OZ.
- [60] P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and design of analog integrated circuits*. New York, NY: John Wiley & Sons, 2001.
- [61] E. Grossar, M. Stucchi, K. Maex, and W. Dehaene, “Read stability and write-ability analysis of SRAM cells for nanometer technologies,” *IEEE Journal of Solid-State Circuits*, vol. 41, no. 11, pp. 2577–2588, Nov. 2006.
- [62] E. J. Gumbel, “Les valeurs extrêmes des distributions statistiques,” *Ann. Inst. H. Poincaré*, vol. 5, no. 2, pp. 115–158, 1935.
- [63] Z. Guo, S. Balasubramanian, R. Zlatanovici, T.-J. K. Liu, and B. Nikolić, “FinFET-based SRAM design,” in *Proc. IEEE/ACM International Symposium on Low Power Electronics and Design*, San Diego, CA, Aug. 2005, pp. 2–7.
- [64] Z. Guo, A. Carlson, L.-T. Pang, K. Duong, T.-J. K. Liu, and B. Nikolić, “Large-scale read/write margin measurement in 45nm CMOS SRAM arrays,” in *Symposium on VLSI Circuits Dig. of Tech. Papers*, Honolulu, HI, Jun. 2008, pp. 42–43.
- [65] ——, “Large-scale SRAM variability characterization in 45nm CMOS,” *IEEE Journal of Solid-State Circuits*, vol. 44, no. 11, pp. 3174–3192, Nov. 2009.
- [66] P. Gupta, A. B. Kahng, Y. Kim, S. Shah, and D. Sylvester, “Investigation of diffusion rounding for post-lithography analysis,” in *Asia and South Pacific Design Automation Conference*, Seoul, Korea, Mar. 2008, pp. 480–485.
- [67] F. Hamzaoglu, K. Zhang, Y. Wang, H. J. Ann, U. Bhattacharya, Z. Chen, Y.-G. Ng, A. Pavlov, K. Smits, and M. Bohr, “A 153Mb-SRAM design with dynamic stability enhancement and leakage reduction in 45nm high-k metal-gate CMOS technology,” in *IEEE International Solid-State Circuits Conference Dig. of Tech. Papers*, San Francisco, CA, Feb. 2008, pp. 376–621.

- [68] R. Heald and P. Wong, “Variability in sub-100nm SRAM designs,” in *IEEE/ACM International Conference Computer-Aided Design*, San Jose, CA, Nov. 2004, pp. 347–352.
- [69] X. Huang, W.-C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Y.-K. Choi, K. Asano, V. Subramanian, T.-J. K. Liu, J. Bokor, and C. Hu, “Sub 50-nm FinFET: PMOS,” in *IEEE International Electron Devices Meeting Tech. Dig.*, Washington, DC, Dec. 1999, pp. 67–70.
- [70] T. Ichikawa and M. Sasaki, “A new analytical model of SRAM cell stability in low-voltage operation,” *IEEE Transactions on Electron Devices*, vol. 43, pp. 54–61, Jan. 1996.
- [71] IEEE-488, also known as GPIB (General Purpose Interface Bus), IEEE.
- [72] M. Ieong, E. Jones, T. Kasnarsky, Z. Ren, O. Dokumaci, R. Roy, L. Shi, T. Furukawa, Y. Taur, R. Miller, and H.-S. P. Wong, “Experimental evaluation of carrier transport and device design for planar symmetric/asymmetric double-gate/ground-plane CMOS-FETs,” in *IEEE International Electron Devices Meeting Tech. Dig.*, Washington, DC, Dec. 2001, pp. 19.6.1–19.6.4.
- [73] ITRS 2001-2007 edition, 2007.
- [74] R. V. Joshi, K. Kim, R. Q. Williams, E. J. Nowak, and C.-T. Chuang, “A high-performance, low leakage, and stable SRAM row-based back-gate biasing scheme in FinFET technology,” in *International Conference on VLSI Design*, Bangalore, India, Jan. 2007, pp. 665–672.
- [75] R. V. Joshi, R. Q. Williams, E. J. Nowak, K. Kim, J. Beintner, T. Ludwig, I. Aller, and C.-T. Chuang, “FinFET SRAM for high-performance low-power applications,” in *Proc. European Solid-State Device Research Conference*, Leuven, Belgium, Sep. 2004, pp. 69–72.
- [76] 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. L. 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,” in *IEEE International Electron Devices Meeting Tech. Dig.*, San Francisco, CA, Dec. 2006, pp. 1–4.
- [77] A. B. Kahng and Y. C. Pati, “Subwavelength lithography and its potential impact on design and EDA,” in *ACM/IEEE Design Automation Conference*, New Orleans, LA, Jun. 1999, pp. 799–804.
- [78] 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,” in *ACM/IEEE Design Automation Conference*, San Francisco, CA, Jul. 2006, pp. 69–72.

- [79] A. Keshavarzi, J. W. Tschanz, S. Narendra, V. De, W. R. Daasch, K. Roy, M. Sachdev, and C. F. Hawkins, “Leakage and process variation effects in current testing on future CMOS circuits,” *IEEE Design and Test of Computers*, vol. 19, pp. 36–43, Oct. 2002.
- [80] D. Khalil, M. Khellah, N.-S. Kim, Y. Ismail, T. Karnik, and V. De, “Accurate estimation of SRAM dynamic stability,” *IEEE Transactions on VLSI Systems*, vol. 16, no. 12, pp. 1639–1647, Dec. 2008.
- [81] M. Khellah, N. S. Kim, Y. Ye, D. Somasekhar, T. Karnik, N. Borkar, F. Hamzaoglu, T. Coan, Y. Wang, K. Zhang, C. Webb, and V. De, “A 65nm SoC embedded 6T-SRAM design for manufacturing with read and write cell stabilizing circuits,” in *Symposium on VLSI Circuits Dig. of Tech. Papers*, Honolulu, HI, Jun. 2006, pp. 48–49.
- [82] J.-J. Kim, K. Kim, and C.-T. Chuang, “Independent-gate controlled asymmetrical SRAM cells in double-gate MOSFET technology for improved read stability,” in *Proc. European Solid-State Device Research Conference*, Montreux, Switzerland, Sep. 2006, pp. 73–76.
- [83] Y. B. Kim, Y.-B. Kim, and F. Lombardi, “Low power 8T SRAM using 32nm independent gate FinFET technology,” in *IEEE International SOC Conference*, Newport Beach, CA, Sep. 2008, pp. 247–250.
- [84] W. Kong, P. C. Parries, G. Wang, and S. S. Iyer, “Analysis of retention time distribution of embedded DRAM - a new method to characterize across-chip threshold voltage variation,” in *IEEE International Test Conference*, Santa Clara, CA, Oct. 2008, pp. 1–7.
- [85] K. J. Kuhn, “Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS,” in *IEEE International Electron Devices Meeting Tech. Dig.*, Washington, DC, Dec. 2007, pp. 471–474.
- [86] T.-J. K. Liu, C. Shin, M. H. Cho, X. Sun, B. Nikolić, and B.-Y. Nguyen, “SRAM cell design considerations for SOI technology,” in *Proc. IEEE International SOI Conference*, Foster City, CA, Oct. 2009, pp. 1–2.
- [87] R. F. Lyon and R. R. Schediwy, “CMOS static memory with a new four-transistor memory cell,” ser. Proc. Stanford Conference on Advanced Research in VLSI, P. Losleben, Ed. Cambridge, MA: MIT Press, 2001, pp. 111–131.
- [88] L. Mathew, Y. Du, A. V.-Y. Thean, M. Sadd, A. Vandooren, C. Parker, T. Stephens, R. Mora, R. Rai, M. Zavala, D. Sing, S. Kalpat, J. Hughes, R. Shimer, S. Jallepalli, G. Workman, W. Zhang, J. G. Fossum, B. E. White, B.-Y. Nguyen, and J. Mogab, “CMOS vertical multiple independent gate field effect transistor (MIGFET),” in *Proc. IEEE International SOI Conference*, Charleston, SC, Oct. 2004, pp. 187–189.
- [89] A. McNeil, “Estimating the tails of loss severity distributions using extreme value theory,” *ASTIN Bulletin*, vol. 27, pp. 117–137, 1997.

- [90] A. Meixner and J. Banik, "Weak write test mode: An SRAM cell stability design for test technique," in *IEEE International Test Conference*, Washington, DC, Nov. 1997, pp. 1043–1052.
- [91] K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, M. Buehler, A. Cappellani, R. Chau*, C.-H. Choi, G. Ding, K. Fischer, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He, J. Hicks, R. Heussner, D. Ingerly, P. Jain, R. James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Liu, J. Maiz, B. McIntyre, P. Moon, J. Neirynck, S. Pae, C. Parker, D. Parsons, C. Prasad, L. Pipes, M. Prince, P. Ranade, T. Reynolds, J. Sandford, L. Shifren, J. Sebastian, J. Seiple, D. Simon, S. Sivakumar, P. Smith, C. Thomas, T. Troeger, P. Vandervoorn, S. Williams, and K. Zawadzki, "A 45nm logic technology with high-k + metal gate transistors, strained silicon, 9 Cu interconnect layers, 193nm dry patterning, and 100% Pb-free packaging," in *IEEE International Electron Devices Meeting Tech. Dig.*, Washington, DC, Dec. 2007, pp. 247–250.
- [92] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, Apr. 1965.
- [93] ——, "No exponential is forever: but "forever" can be delayed!" in *IEEE International Solid-State Circuits Conference Dig. of Tech. Papers*, San Francisco, CA, Feb. 2003, pp. 20–23.
- [94] S. Mukhopadhyay, K. Kang, H. Mahmoodi, A. Datta, D. Park, and K. Roy, "Self-repairing SRAM for reducing parametric failures in nanoscaled memory," in *Symposium on VLSI Circuits Dig. of Tech. Papers*, Honolulu, HI, Jun. 2006, pp. 132–133.
- [95] S. Mukhopadhyay, H. Mahmoodi, and K. Roy., "Design of high performance sense amplifier using independent gate control in sub-50nm double-gate MOSFET," in *Proc. IEEE International Symposium on Quality of Electronic Design*, San Jose, CA, Mar. 2005, pp. 490–495.
- [96] S. Nadarajah and S. Kotz, "Exact distribution of the max/min of two gaussian random variables," *IEEE Transactions on VLSI Systems*, vol. 16, pp. 210–212, Feb. 2008.
- [97] S. Narendra, D. Antoniadis, and V. De, "Impact of using adaptive body bias to compensate die-to-die vt variation on within-die vt variation," in *Proc. IEEE/ACM International Symposium on Low Power Electronics and Design*, San Diego, California, Aug. 1999, pp. 229–232.
- [98] S. Nassif, "Delay variability: Sources, impacts and trends," in *IEEE International Solid-State Circuits Conference Dig. of Tech. Papers*, San Francisco, CA, Feb. 2000, pp. 368–369.
- [99] S. Natarajan, M. Armstrong, M. Bost, R. Brain, M. Brazier, C.-H. Chang, V. Chikarmane, M. Childs, H. Deshpande, K. Dev, G. Ding, T. Ghani, O. Golonzka, W. Han, J. He, R. Heussner, R. James, I. Jin, C. Kenyon, S. Klopcic, S.-H. Lee, M. Liu, S. Lodha, B. McFadden, A. Murthy, L. Neiberg, J. Neirynck, P. Packan, S. Pae,

- C. Parker, C. Pelto, L. Pipes, J. Sebastian, J. Seiple, B. Sell, S. Sivakumar, B. Song, K. Tone, T. Troeger, C. Weber, M. Yang, A. Yeoh, and K. Zhang, “A 32nm logic technology featuring 2nd-generation high-k + metal-gate transistors, enhanced channel strain and $0.171 \mu\text{m}^2$ SRAM cell size in a 291Mb array,” in *IEEE International Electron Devices Meeting Tech. Dig.*, San Francisco, CA, Dec. 2008, pp. 1–3.
- [100] K. Nii, M. Yabuuchi, Y. Tsukamoto, S. Ohbayashi, Y. Oda, K. Usui, T. Kawamura, N. Tsuboi, T. Iwasaki, K. Hashimoto, H. Makino, and H. Shinohara, “A 45-nm single-port and dual-port SRAM family with robust read/write stabilizing circuitry under DVFS environment,” in *Symposium on VLSI Circuits Dig. of Tech. Papers*, Honolulu, HI, Jun. 2008, pp. 212–213.
- [101] K. Noda, K. Matsui, K. Takeda, and N. Nakamura, “A loadless CMOS four-transistor SRAM cell in a 0.18- μm logic technology,” *IEEE Transactions on Electron Devices*, vol. 48, pp. 2851–2855, Dec. 2001.
- [102] E. J. Nowak, “Maintaining the benefits of CMOS scaling when scaling bogs down,” *IBM Journal of Research and Development*, vol. 46, no. 2/3, pp. 169–180, Mar. 2002.
- [103] E. J. Nowak, B. A. Rainey, D. M. Fried, J. Kedzierski, M. Ieong, W. Leipold, J. Wright, and M. Breitwisch, “A functional FinFET-DGCMOS SRAM cell,” in *IEEE International Electron Devices Meeting Tech. Dig.*, San Francisco, CA, Dec. 2002, pp. 411–414.
- [104] S. Ohbayashi, M. Yabuuchi, k. Nii, Y. Tsukamoto, S. Imaoka, Y. Oda, T. Yoshihara, M. Igarashi, M. Takeuchi, H. Kawashima, Y. Yamaguchi, K. Tsukamoto, M. Inuishi, H. Makino, K. Ishibashi, and H. Shinohara, “A 65-nm SoC embedded 6T-SRAM designed for manufacturability with read and write operation stabilizing circuits,” *IEEE Journal of Solid-State Circuits*, vol. 42, no. 4, pp. 820–829, Apr. 2007.
- [105] P. Oldiges, Q. Lin, K. Petrillo, M. Sanchez, M. Ieong, and M. Hargrove, “Modeling line edge roughness effects in sub 100 nanometer gate length devices,” in *International Conference on Simulation of Semiconductor Processes and Devices*, Seattle, WA, Sep. 2000, pp. 131–134.
- [106] M. Orshansky, L. Milor, P. Chen, K. Keutzer, and C. Hu, “Impact of spatial intra-chip gate length variability on the performance of high-speed digital circuits,” *IEEE Transactions on Computer-Aided Design of Circuits and Systems*, vol. 21, no. 5, pp. 544–553, May 2002.
- [107] M. Orshansky, S. Nassif, and D. Boning, *Design for Manufacturability and Statistical Design: A Constructive Approach*, ser. Integrated Circuits and Systems, A. Chandrakasan, Ed. New York, NY: Springer Science + Business Media, LLC, 2008.
- [108] S. O’uchi, M. Masahara, K. Endo, Y. X. Liu, T. Matsukawa, K. Sakamoto, T. Sekigawa, H. Koike, and E. Suzuki, “FinFET-based flex-vth SRAM design for drastic standby-leakage-current reduction,” *IEICE Transactions on Electronics*, vol. E91-C, no. 4, pp. 534–542, Apr. 2008.

- [109] S. O'uchi, M. Masahara, K. Sakamoto, K. Endo, Y. X. Liu, T. Matsukawa, T. Sekigawa, H. Koike, and E. Suzuki, “Flex-pass-gate SRAM design for static noise margin enhancement using FinFET-based technology,” in *IEEE Custom Integrated Circuits Conference*, San Jose, CA, Sep. 2007, pp. 33–36.
- [110] L.-T. Pang, “Measurement and analysis of variability in CMOS circuits,” Ph.D. dissertation, University of California, Berkeley, Berkeley, CA, 2008.
- [111] L.-T. Pang and B. Nikolić, “Measurements and analysis of process variability in 90nm CMOS,” *IEEE Journal of Solid-State Circuits*, vol. 44, no. 5, pp. 1655–1663, Aug. 2009.
- [112] L.-T. Pang, K. Qian, C. Spanos, and B. Nikolić, “Measurement and analysis of variability in 45nm strained-Si CMOS technology,” *IEEE Journal of Solid-State Circuits*, vol. 44, no. 8, pp. 2233–2243, Aug. 2009.
- [113] M. Pelgrom, A. Duinmaijer, and A. Welbers, “Matching properties of MOS transistors,” *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.
- [114] H. Pilo, “Discussion session: SRAM design in 90nm era,” in *IEEE International Solid-State Circuits Conference*, San Francisco, CA, Feb. 2005.
- [115] ———, “2006 IEDM SRAM short course,” in *IEEE International Electron Devices Meeting*, San Francisco, CA, Dec. 2006.
- [116] H. Pilo, J. Barwin, G. Braceras, C. Browning, S. Burns, J. Gabric, S. Lamphier, M. Miller, A. Roberts, and F. Towle, “An SRAM design in 65nm and 45nm technology nodes featuring read and write-assist circuits to expand operating voltage,” in *Symposium on VLSI Circuits Dig. of Tech. Papers*, Honolulu, HI, Jun. 2006, pp. 15–16.
- [117] J. D. Plummer, M. D. Deal, and P. B. Griffin, *Silicon VLSI Technology*. Upper Saddle River, NJ: Prentice-Hall, 2000.
- [118] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, “SRAM leakage suppression by minimizing standby supply voltage,” in *Proc. IEEE International Symposium on Quality of Electronic Design*, San Jose, CA, Mar. 2004, pp. 55–60.
- [119] Rashmi, A. Kranti, and G. A. Armstrong, “6-T SRAM cell design with nanoscale double-gate SOI MOSFETs: Impact of source/drain engineering and circuit topology,” *IOP Semiconductor Science and Technology*, vol. 23, Jul. 2008.
- [120] S. Rusu, “Keynote: Trends and challenges in high-performance microprocessor design,” in *Electronic Design Processes*, Monterey, CA, Apr. 2004.
- [121] W. Schemmert and G. Zimmer, “Threshold-voltage sensitivity of ion-implanted M.O.S. transistors due to process variations,” *Electronics Letters*, vol. 10, no. 9, pp. 151–152, May 1974.

- [122] D. K. Schroder and J. A. Babcock, “Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing,” *Journal of Applied Physics*, vol. 94, no. 1, pp. 1–18, Jul. 2003.
- [123] E. Seevinck, F. List, and J. Lohstroh, “Static-noise margin analysis of MOS SRAM cells,” *IEEE Journal of Solid-State Circuits*, vol. 22, no. 5, pp. 748–754, Oct. 1987.
- [124] H. Shang, L. Chang, X. Wang, M. Rooks, Y. Zhang, B. To, K. Babich, G. Totir, Y. Sun, E. Kiewra, M. Ieong, and W. Haensch, “Investigation of FinFET devices for 32nm technologies and beyond,” in *Symposium on VLSI Technology Dig. of Tech. Papers*, Honolulu, HI, Jun. 2006, pp. 54–55.
- [125] N. Shibata, H. Kiya, S. Kurita, H. Okamoto, M. Tan’no, and T. Douseki, “A 0.5V 25MHz 1mW 256kb MTCMOS/SOI SRAM for solar-power-operated portable personal digital equipment - sure write operation by using step-down negatively overdriven bit-line scheme,” *IEEE Journal of Solid-State Circuits*, vol. 41, no. 3, pp. 728–742, Mar. 2006.
- [126] C. Shin, M. H. Cho, Y. Tsukamoto, B.-Y. Nguyen, B. Nikolić, and T.-J. K. Liu, “SRAM yield enhancement with thin-BOX FD-SOI,” in *Proc. IEEE International SOI Conference*, Foster City, CA, Oct. 2009, pp. 1–2.
- [127] A. Singhee and R. A. Rutenbar, “Statistical blockade: A novel method for very fast Monte Carlo simulation of rare circuit events, and its application,” in *Design, Automation, and Test in Europe Conference and Exhibition*, Nice, France, Apr. 2007, pp. 1–6.
- [128] A. Singhee, J. Wang, B. H. Calhoun, and R. A. Rutenbar, “Recursive statistical blockade: An enhanced technique for rare event simulation with application to SRAM circuit design,” in *International Conference on VLSI Design*, Hyderabad, India, Jan. 2008, pp. 131–136.
- [129] T. Skotnicki, J. A. Hutchby, T.-J. King, H.-S. P. Wong, and F. Boeuf, “The end of CMOS scaling: Toward the introduction of new materials and structural changes to improve MOSFET performance,” *IEEE Circuits and Devices Magazine*, vol. 21, no. 1, pp. 16–26, Jan. 2005.
- [130] Specifications for the OpteronTM Istanbul Microprocessor, AMD.
- [131] Specifications for the Xeon® Dunnington Microprocessor, Intel®.
- [132] X. Sun, Q. Lu, V. Moroz, H. Takeuchi, G. Gebara, J. Wetzel, S. Ikeda, C. Shin, and T.-J. K. Liu.
- [133] I. E. Sutherland, R. F. Sproull, and D. F. Harris, *Logical Effort: Designing Fast CMOS Circuits*. San Francisco, CA: Morgan Kaufmann, 1999.

- [134] K. Takeda, H. Ikeda, Y. Hagihara, M. Nomura, and H. Kobatake, “Redefinition of write margin for next-generation SRAM and write-margin monitoring circuit,” in *IEEE International Solid-State Circuits Conference Dig. of Tech. Papers*, San Francisco, CA, Feb. 2006, pp. 2602–2611.
- [135] K. . Takeuchi, T. Fukai, T. Tsunomura, A. T. Putra, A. Nishida, S. Kamohara, and T. Hiramoto, “Understanding random threshold voltage fluctuation by comparing multiple fabs and technologies,” in *IEEE International Electron Devices Meeting Tech. Dig.*, Washington, DC, Dec. 2007, pp. 467–470.
- [136] K. Takeuchi, R. Koh, and T. Mogami, “A study of the threshold voltage variation for ultra-small-bulk and SOI CMOS,” *IEEE Transactions on Electron Devices*, vol. 48, pp. 1995–2001, Sep. 2001.
- [137] Y. Taur, “CMOS design near the limit of scaling,” *IBM Journal of Research and Development*, vol. 46, no. 2/3, pp. 213–222, Mar. 2002.
- [138] Taurus v.2003.12, Synopsys, Inc.
- [139] S. E. Thompson, M. Armstrong, C. Auth, S. Cea, R. Chau, G. Glass, T. Hoffman, J. Klaus, Z. Y. Ma, B. McIntyre, A. Murthy, B. Obradovic, L. Shifren, S. Sivakumar, S. Tyagi, T. Ghani, K. Mistry, M. Bohr, and Y. El-Mansy, “A logic nanotechnology featuring strained-silicon,” *IEEE Electron Device Letters*, vol. 25, no. 4, pp. 191–193, Apr. 2004.
- [140] Y. L. Tong, *The Multivariate Normal Distribution*. New York, NY: Springer-Verlag, 1990.
- [141] J. Tsai, S. O. Toh, Z. Guo, L.-T. Pang, T.-J. K. Liu, and B. Nikolić, “SRAM stability characterization using tunable ring oscillators in 45nm CMOS,” in *IEEE International Solid-State Circuits Conference Dig. of Tech. Papers*, San Francisco, CA, Feb. 2010, p. accepted for publication.
- [142] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, “Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage,” in *IEEE International Solid-State Circuits Conference Dig. of Tech. Papers*, San Francisco, CA, Feb. 2002, pp. 422–478.
- [143] J. Tschanz, S. Narendra, Y. Ye, B. Bloechel, S. Borkar, and V. De, “Dynamic-sleep transistor and body bias for active leakage power control of microprocessors,” in *IEEE International Solid-State Circuits Conference Dig. of Tech. Papers*, San Francisco, CA, Feb. 2003, pp. 102–481 vol.1.
- [144] D. P. Wang, H. J. Liao, H. Yamauchi, Y. H. Chen, Y. L. Lin, S. H. Lin, D. C. Liu, H. C. Chang, and W. Hwang, “A 45nm dual-port SRAM with write and read capability enhancement at low voltage,” in *IEEE International SOC Conference*, Hsin Chu, Taiwan, Sep. 2007, pp. 211–214.

- [145] G. Wang, K. Cheng, H. Ho, J. Faltermeier, W. Kong, H. Kim, J. Cai, C. Tanner, K. McStay, K. Balasubramanyam, C. Pei, L. Ninomiya, X. Li, K. Winstel, D. Dobuzinsky, M. Naeem, R. Zhang, R. Deschner, M. J. Brodsky, S. Allen, J. Yates, Y. Feng, P. Marchetti, C. Norris, D. Casarotto, J. Benedict, A. Kniffm, D. Parise, B. Khan, J. Barth, P. Parries, T. Kirihata, J. Norum, and S. S. Iyer, “A $0.127 \mu\text{m}^2$ high performance 65nm SOI based embedded DRAM for on-processor applications,” in *IEEE International Electron Devices Meeting Tech. Dig.*, San Francisco, CA, Dec. 2006, pp. 1–4.
- [146] J. Wang, S. Nalam, and B. H. Calhoun, “Analyzing static and dynamic write margin for nanometer SRAMs,” in *Proc. IEEE/ACM International Symposium on Low Power Electronics and Design*, Bangalore, India, Aug. 2008, pp. 129–134.
- [147] J. Wang, A. Singhee, R. A. Rutenbar, and B. H. Calhoun, “Statistical modeling for the minimum standby supply voltage of a full SRAM array,” in *Proc. European Solid-State Circuits Conference*, Munich, Germany, Sep. 2007, pp. 400–403.
- [148] Y. Wang, H. J. Ahn, U. Bhattacharya, Z. Chen, T. Coan, F. Hamzaoglu, W. M. Hafez, C.-H. Jan, P. K. Ans S. H. Kulkarni, J.-F. Lin, Y.-G. Ng, I. Post, L. Wei, Y. Zhang, K. Zhang, and M. Bohr.
- [149] Y. Wang, U. Bhattacharya, F. Hamzaoglu, P. Kolar, Y.-G. Ng, L. Wei, Y. Zhang, K. Zhang, and M. Bohr, “A 4.0GHz 291Mb voltage-scalable SRAM in 32nm high-k metal-gate CMOS with integrated power management,” in *IEEE International Solid-State Circuits Conference Dig. of Tech. Papers*, San Francisco, CA, Feb. 2009, pp. 456–457.
- [150] C. Wann, R. Wong, D. J. Frank, R. Mann, S.-B. Ko, P. Croce, D. Lea, D. Hoyniak, Y.-M. Lee, J. Toomey, M. Weybright, and J. Sudijono, “SRAM cell design for stability methodology,” in *IEEE International Symposium on VLSI-TSA*, Hsinchu, Taiwan, Apr. 2005, pp. 21–22.
- [151] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, “Yield and speed optimization of a latch type voltage sense amplifier,” *IEEE Journal of Solid-State Circuits*, vol. 39, no. 7, pp. 1148–1158, Jul. 2004.
- [152] S. Wolf, *Silicon Processing for the VLSI Era, Vol. 3: The Submicron MOSFET*. Sunset Beach, CA: Lattice Press, 1995.
- [153] V. K. Wong, C. H. Lock, K. H. Siek, and P. J. Tan, “Electrical analysis to fault isolate defects in 6T memory cells,” in *Proc. IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits*, Raffles City, Singapore, Jul. 2002, pp. 101–104.
- [154] M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Ohbayashi, Y. Nakase, and H. Shinohara, “A 45nm 0.6V cross-point 8T SRAM with negative biased read/write assist,” in *Symposium on VLSI Circuits Dig. of Tech. Papers*, Kyoto, Japan, Jun. 2009, pp. 158–159.

- [155] Y. Yamagata, “Short course presentation: Embedded memory technology for low power systems,” in *IEEE International Electron Devices Meeting Tech. Dig.*, Washington, DC, Dec. 2005.
- [156] M. Yamaoka, K. Osada, and K. Ishibashi, “0.4V logic-library-friendly SRAM array using rectangular-diffusion cell and delta-boosted-array voltage scheme,” *IEEE Journal of Solid-State Circuits*, vol. 39, no. 6, pp. 934–940, Jun. 2004.
- [157] M. Yamaoka, K. Osada, R. Tsuchiya, M. Horiuchi, S. Kimura, and T. Kawahara, “Low power SRAM menu for SOC application using yin-yang-feedback memory cell technology,” in *Symposium on VLSI Circuits Dig. of Tech. Papers*, Honolulu, HI, Jun. 2004, pp. 288–289.
- [158] H. Yamauchi, “Tutorial: Variation-tolerant SRAM circuit designs,” in *IEEE International Solid-State Circuits Conference*, San Francisco, CA, Feb. 2009.
- [159] M. Yang, E. P. Gusev, M. Ieong, O. Gluschenkov, D. C. Boyd, K. K. Chan, P. M. Kozlowski, C. P. DEmic, R. M. Sicina, P. C. Jamison, and A. I. Chou, “Performance dependence of CMOS on silicon substrate orientation for ultrathin oxynitride and HfO_2 gate dielectrics,” *IEEE Electron Device Letters*, vol. 24, no. 5, pp. 339–341, May 2003.
- [160] M. Yoshimoto, K. Anami, H. Shinohara, T. Yoshihara, H. Takagi, S. Nagao, S. Kayano, and T. Nakano, “A divided word-line structure in the static RAM and its application to a 64K full CMOS RAM,” *IEEE Journal of Solid-State Circuits*, vol. 18, no. 5, pp. 479–485, Oct. 1983.
- [161] B. Zhang, A. Arapostathis, S. Nassif, and M. Orshansky, “Analytical modeling of SRAM dynamic stability,” in *IEEE/ACM International Conference Computer-Aided Design*, San Jose, CA, Nov. 2006, pp. 315–322.
- [162] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, “A 3GHz 70Mb SRAM in 65nm CMOS technology with integrated column-based dynamic power supply,” *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 146–151, Jan. 2006.
- [163] K. Zhang, U. Bhattacharya, L. Ma, Y.-G. Ng, B. Zheng, M. Bohr, and S. Thompson, “A fully synchronized, pipelined, and re-configurable 50Mb SRAM on 90nm CMOS technology for logic applications,” in *Symposium on VLSI Circuits Dig. of Tech. Papers*, Kyoto, Japan, Jun. 2003, pp. 253–254.