Statistical methodologies for modelling the impact of process variability in ultra-deep-submicron SRAMs

Kaya Can Akyel

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Nano-Electronique et Nano-Technologie**

Arrêté ministériel : 7 août 2006

Présentée par

Kaya Can AKYEL

Thèse dirigée par **Gérard GHIBAUDO** et
Co-encadrée par **Lorenzo CIAMPOLINI**

préparée au sein du **Laboratoire IMEP-LAHC**
dans **l'École Doctorale EEATS Electronique, Electrotechnique, Automatique et Traitement du Signal**

Statistical Methodologies for Modeling the Impact of Process Variability in Ultra-Deep-Submicron SRAMs

Thèse soutenue publiquement le **17 Décembre 2014,**
devant le jury composé de :

M. Ian O'CONNOR
Professeur à Ecole Centrale de Lyon (Président)
M. Jean-Michel PORTAL
Professeur des universités, Aix-Marseille (Rapporteur)
M. Amara AMARA
Professeur, ISEP-Paris (Rapporteur)
M. Gérard GHIBAUDO
Directeur de Recherche à l'IMEP-LAHC, Grenoble (Directeur de thèse)
M. Lorenzo CIAMPOLINI
Ingénieur STMicroelectronics, Crolles (Co-encadrant)

"Statistics is the grammar of science."

KARL PEARSON

Acknowledgments

It is hard to acknowledge three years of work in a few paragraphs, especially taking into account the numerous people with whom I have had the pleasure of working.

First of all, I would like to thank the people behind my motivation for becoming a Ph.D. candidate : Jean-Marc Daveau, Sylvain Clerc and Philippe Roche. When I was an intern on their team during the last year of my master's degree, I was conveniently brainwashed by these research experts, who somehow managed to succeed in convincing me to continue my studies towards a Ph.D. degree. Even though I chose to do my Ph.D. with another team, they were always on my mind. So, after more than three years, I can finally say: Thank you for brainwashing me, it is totally worth it.

Three years is a sufficient amount of time to really get to know your teammates. I was really lucky to be the part of ST Microelectronics Crolles SRAM Team. I thank Francois Veillet, Faress Tisaffi Drissi, Magali Lamy, Caroline Biasi, Ludovic Goualou, Jean-Christophe Lafront, Olivier Menut, Jean-Philippe Noel, Matthieu Le Boulaire, Mohamed Reda Boumchedda, Joseph Nguyen, Guy Debar and Arnaud Wenzel for first accepting me as one of them at the beginning and then supporting my work both professionally and personally during these last 3 years.

I was very lucky to work under two different team leaders: Christophe Lecocq and David Turgis. I have learned a lot from both and always enjoyed any kind of discussions that I have held with them. I thank Christophe in particular for the trust that he has had in me and for allowing to me be part of industry related discussions that I will remember as a very important experience in my education. I specially thank David for managing me and my work in the warmest and friendliest way possible.

Everybody needs some time to relax and this was possible for me at ST thanks to the "CCDS Wednesday football mail list". I have started my Ph.D. as a rookie football player and I have ended it as an experienced right back. It would take too long to acknowledge each member of the mailing list individually, so please accept a "common thank you" for all.

Working towards a Ph.D. degree is not always easy, but it gets easier, and even fun, when you are working with good work friends. I thank Hani Sherry,

Milovan Blagojevic, Dajana Danilovic, Camilo Salazar, Cecilia Mezzomo, Nicolas Beaudouin, Antoine Delmas, Tekfoy Lim, Bertrand Pelloux-Prayer, Julien Kieffer, Cyril Bottoni, Fady Abouzeid and Mathieu Vallet for their friendship.

I specially thank Olivier Thomas from LETI for being a "non-official" advisor of my thesis, and providing me all I need to decorate my modeling works with silicon measurements. I have really appreciated our collaboration and I am looking forward to the possibility of it happening again in the future.

I acknowledge my academic advisor Gerard Ghibaudo with all respect and love that I have for him. Three years were not enough for me to understand how he can deal with dozens of Ph.D. candidates as effectively as he does. I have not had any other academic advisor in my short life, but I can bet that Gerard is one of the best.

From my first day to the last, someone was always there for me regardless of the subject matter on my mind, either professional or personal. I thank my industrial advisor, my teammate, my cubicle-mate and my friend Lorenzo Ciampolini. This work would never have reached success without him and his efforts, and furthermore, I would never have become the Ph.D that I am today without his guidance.

Finally, I want to thank my family for their support during my academic career and beyond.

Abstract

The downscaling of device geometry towards its physical limits exacerbates the impact of the inevitable atomistic phenomena tied to matter granularity. In this context, many different variability sources raise and affect the electrical characteristics of the manufactured devices. The variability-aware design methodology has therefore become a popular research topic in the field of digital circuit design, since the increased number of transistors in the modern integrated circuits had led to a large statistical variability affecting dramatically circuit functionality.

Static Random Access Memory (SRAM) circuits which are manufactured with the most aggressive design rules in a given technology node and contain billions of transistor, are severely impacted by the process variability which stands as the main obstacle for the further reduction of the bitcell area and of its minimum operating voltage. The reduction of the latter is a very important parameter for Low-Power design, which is one of the most popular research fields of our era. The optimization of SRAM bitcell design therefore has become a crucial task to guarantee the good functionality of the design at an industrial manufacturing level, in the same time answering to the high density and low power demands. However, the long time required by each new technology node process development means a long waiting time before obtaining silicon results, which is in cruel contrast with the fact that the design optimization has to be started as early as possible. An efficient SPICE characterization methodology for the minimum operating voltage of SRAM circuits is therefore a mandatory requirement for design optimization. This research work concentrates on the development of the new simulation methodologies for the modeling of the process variability in ultra-deep-submicron SRAMs, with the ultimate goal of a significantly accurate modeling of the minimum operating voltage Vmin. A particular interest is also carried on the time-dependent sub-class of the process variability, which appears as a change in the electrical characteristics of a given transistor during its operation and during its life-time.

This research work has led to many publications and one patent application. The majority of findings are retained by STMicroelectronics SRAM development team for a further use in their design optimization flow.

Contents

List of Figures **3**

List of Tables **12**

Glossary **13**

1 General Introduction **16**

2 Variability in Ultra-Deep-Submicron CMOS **19**
2.1 Introduction to Variability 19
2.2 Variability Sources in CMOS 20
2.2.0.1 Systematic Variability 20
2.2.0.2 Random Variability 22
2.2.0.3 Time Dependency in Variability 24
2.3 The Improvement Techniques for Variability and New Device Architectures 26
2.4 Conclusion 29

3 The Impact of the Variability in Static Random Access Memory **31**
3.1 Introduction to SRAMs limitations in modern System-On-Chips 31
3.2 SRAM Bitcell Architecture and Common Operations 33
3.3 Alternate Bitcell Architecture 36
3.4 A Simple Model for the Bitcell Variability Space 37
3.5 SRAM Bitcell Failure Analysis 40
3.5.1 SRAM Bitcell Static (DC) Analysis 42
3.5.1.1 Static Noise Margin 43
3.5.1.2 Write Margin 44
3.5.2 SRAM Bitcell Dynamic Analysis 44
3.5.2.1 Read-Ability 47
3.5.2.2 Write-Ability 47
3.5.2.3 Multiple-Pulse Analysis and Figure of Merits 48
3.6 6T-SRAM Bitcell Failure Mechanisms 51
3.7 The Minimum Operating Voltage V_{min} 55

4 SRAM Bitcell Variability Space Modeling for V_{min} Estimation 57
4.1 Bitcell Variability Space Modeling using Monte Carlo SPICE simulations . . . 57
4.2 SRAM Static V_{min} Analysis . . . 64
4.2.1 6T Bitcells Static V_{min} Measurements and SPICE Modeling Results . . . 64
4.2.2 Ultra-Low-Voltage SRAM Static V_{min} Measurements and Modeling Results . . . 67
4.3 Bitcell Variability Space Modeling using Smart Algorithm : Hypersphere Most Probable Failure Point Search Methodology 69
4.3.1 Bitcell Failure Probability Calculation . . . 73
4.4 SRAM Dynamic Vmin Analysis . . . 78
4.4.1 Bitcell Dynamic Fail/Pass SPICE Analysis using Hypersphere Algorithm . . . 78
4.4.1.1 Read-Ability Test . . . 79
4.4.1.2 Write-Ability Test . . . 83
4.4.1.3 Read-After-Write Test . . . 86
4.4.2 SRAM Dynamic Failures on Silicon . . . 90
4.5 Application Example: Hypersphere MPFP Search for Investigations on SNM Yield Loss at High-Voltage in 28nm UTBB FD-SOI SRAM bitcells . . . 93
4.6 Smart Dynamic Back-Biasing Bitcell V_{min} Boost in UTBB FD-SOI . . . 95
4.7 Conclusion . . . 101

5 Random Telegraph Signal Noise in 28nm UTBB FD-SOI and the impact on 6T SRAM 103
5.1 Time-Dependent Random Telegraph Signal Noise Variability 103
5.2 SPICE-level RTS Noise Modeling in UTBB FD-SOI . . . 106
5.2.1 RTS Trap Characteristics and Particularity of UTBB FD-SOI . . . 106
5.2.2 Front- and Back-gate Coupling Aware 2-Dirac Charge Inversion model . . . 111
5.2.3 RTS-aware 6T SRAM SPICE netlist generation in Matlab . . . 115
5.3 Measurements and Simulation Results . . . 118
5.3.1 Hardware Setup . . . 118
5.3.2 Results . . . 121
5.4 Conclusion . . . 125

6 General Conclusion 127
6.1 Key Contributions . . . 127
6.2 Future Work . . . 129

List of Figures

2.1 CMOS cross section with major sources of process variability [4] 21
2.2 Lithography induced variability and OPC based improvement. The layout images (left) and the Scanning Electroin Microscope (SEM) images (right) of the manufactured patterns are shown [12]. 22
2.3 (a) Average number of dopants evolution in MOS Transistor channel w.r.t technology node [18] (b) 3D View of a numerical MOS transistor model simulating the number of dopants in the channel in 65nm and 45nm technology nodes [12]. 23
2.4 Different types of traps associated with Si-SiO_2 interface [25]. 25
2.5 I_{ON}/I_{OFF} evolution with respect to the technology node in the planar CMOS technologies for high performance and low power devices [50]. 27
2.6 The threshold voltage variability (σ_{VTH}) with the different components w.r.t the gate length for planar CMOS technologies [51]. 28
2.7 Schematics of Bulk, FD-SOI and FinFet transistors. 29

3.1 SRAM bitcell scaling trend from 65nm to 32nm techonology node for the performance bitcell (squares) and for dense bitcell (diamonds) showing the 50% area reduction [78]. 33
3.2 SRAM minimum operating voltage reported in ISSCC and VLSI Conferences between 2004 - 2010 (crosses) and ITRS predictions at 2001 and at 2009 (straight lines) [79]. 33
3.3 6T-SRAM Bitcell schematics: Two cross-coupled inverters (top) for the storage of the data are connected to the external world via two access transistors (pass-gates) which are activated with a word-line signal. The transistor-level schematic (bottom) shows also the pull-up (PMOS), pull-down (NMOS) and pass-gate (NMOS) transistors. 35

3.4 A block level of SRAM system. Column and row decoder circuitries are driven by the addressing latch which selects the required bitcells. Data Register and I/O bufffers are used to write to cells and are disabled during read cycles when the sense amplifier outputs the stored data. The clock circuitry and the word line driver which determines the word line pulse width, are not shown. 36
3.5 The Ultra-Low Voltage 10T bitcell architecture [80, 81]. . . . 38
3.6 The 2-Dimension variabilty space of an inverter formed by T1 and T2. Variability impact on each transistor is quantified by a shift in the threshold voltage, δV_{thT1} and δV_{thT2}. A positive shift result in a slower device while a negative shift results in a faster device, giving 4 different combinations. I_0 represents the nominal inverter without the variability representing the origin O of the 2-Dimension space. I_1, I_2, I_3, I_4, I_5 represent different variants of I_0 appearing as a consequence of variability. 39
3.7 The different steps in the process developpement of a technology node in semi-conductor manufacturing industry. The SRAM SPICE model cards reaches the final maturity when the silicon extracted bitcell figures match with CAD results, both respecting the specifications for that given technology node. 41
3.8 The 6T-bitcell schematic for SNM simulation. Two separate noise voltage sources Vn_L and Vn_R are connected to the bitcell internal nodes. 43
3.9 The butterfly curve illustrating VTCs of two cross-coupled inverters. a)VTC of nominal and balanced bitcell without statistical variability b)VTCs of a bitcell under statistical variability c)The spread of VTCs under a large statistical variability. 45
3.10 The WM SPICE simulation waveforms for a bitcell initially storing '0' in the node L. The BLR is swept from Vdd to 0V while BLL and WL are held at "1". The WM is the BLR voltage at which the internal node toggle occurs. 46
3.11 The Read-Ability (RA) extraction example for a bitcell initially storing '0' in the node L. The BLL voltage, V_{BLL}, is discharged during the read. For a successful read, the voltage difference between BLL and BLR has to be equal or larger than the SA_{offset} at the end of the WL pulse. 47

3.12 The Write-Ability (WA) test for a bitcell initially storing '1' in the node L. The voltage difference between internal nodes has to be equal or larger than the threshold value p*V_{dd} at the end of the write operation. The time at which the test is performed, $t_{measure}$ depends on the timing requirements, for exemple, in a aggressive high-frequency test, $t_{measure}$ is equal to the twice WL pulse width. 48
3.13 The Write-Ability (WA) failure for a bitcell initially storing '1' in the node L. The voltage difference between internal nodes at $t_{measure}$ is not large enough at the end of the WL pulse and the content is lost. 49
3.14 The bitcell internal nodes voltage waveforms leading to read-stability failure during multiple read access. A first read (RD1) is successfully performed, but the non stable states of the internal nodes at the begining of the second read (RD2) lead the bitcell to a stability failure (internal nodes toggle). . 49
3.15 The bitcell internal nodes voltage waveform during read after write operations. a) The read helps to the completion of the node R thus no fail occurs. b) The non stable-states of the intenal nodes lead to read stability failure during the read. . . 50
3.16 The 2-Dimension variabilty space of an inverter formed by T1 and T2, with two failure zones F_0 and F_1 caused by the mismatch beween T0 and T1. I_0 represents the nominal inverter without the variability. I_1 and I_4 represent the variants of I_0 occuring in the failure zones of the variability space, whereas I_2 and I_3 occur close to the failur zones but they are still in the safe zone of the variability space. 51
3.17 The current flow in a bitcell during a read operation causing a stability (blue arrows) and read-ability (orange arrows) failures. Arrows sizes are proportional to the amount of current flows across the device. reached. 53
3.18 The current flow during a write operation causing a discharge failure (blue arrows) and completion failure (orange arrows). Arrows sizes are proportional to the amount of current flowing across the device. 53

4.1 Probability Density Function of a normally distributed values with mean 0 and standard deviation 1: The dark blue area is less than one standard deviation away from the mean and accounts for 68.2% of the values, while two standard deviations from the mean (medium and dark blue) account for 95.4%, and three standard deviations (light, medium, and dark blue) account for 99.7%. 60

4.2 Probability Density Function (blue bars) of the Single Port High Density Bitcell Read Current in 40nm node obtained with 4096 Monte Carlo runs performed using typical SPICE process corners at 27 °C. The straight red line represents the normal distribution curve drawn using the mean and standard deviation of the simulation results. 60
4.3 Probability Density Function (bars) of SNML, SNMR and SNM margins, as obtained from approximately 10^5 measurements at nominal Vdd and ambient temperature on Single-Port, High-Density bitcell manufactured in C45. Minimum and maximum SNM results for 10^5 Monte Carlo simulations are also presented, for a standard simulator (crosses) and fast MC extension (triangles). 62
4.4 Same SNM measurements probability density function as in figure 4.3, presented in log-scale (bars). CAD results with the Gaussian approximation (crosses) and the large-sigma model for tail estimation (red line) are also shown. 63
4.5 Cumulative distribution functions of the measured V_{min} (measured yield) versus the operating voltage for a) Single-Port Register b) Single-Port High-Density bitcells manufactured in C40 technology. V_{min}@95 is the Vdd at which one has 95% of yield (solid line). The two graphs have a common voltage scale. 65
4.6 a) CAD threshold voltage alignment (empty symbols) with respect to the silicon (filled symbols) measurements at each process corners for pull-down and pull-up transistors. b) Aligned-CAD (empty symbols) vs. Silicon (filled symbols) bitcell average Icell. 66
4.7 Cumulative distribution functions of the simulated Vmin (simulated yield) as obtained with C40 models versus the operating voltage for a) Single-Port Register b) Single-Port High-Density bitcells. V_{min}@95% is the Vdd at which 95% of yield (solid line) is reached. The two graphs have the same voltage scale as those of figure 4.5. 67
4.8 CAD vs. Silicon V_{min}@95% for the two C40 bitcells at each process corner. The higher V_{min} of all corners is the bitcell V_{min}@95%. The grid lines have the same spacing as in figure 4.5 and figure 4.7. 68
4.9 Cumulative distribution function of the lowest operating voltage of a population of 32kb ULV memory cut. Experimental data (circles and crosses) are compared to Monte Carlo results (diamonds and squares) obtained at two different process corners. The dot-lines indicate 50% and 95% yield levels. 69

4.10 Most Probable Failure Point illustration in a 2-dimensional variability space. MPFP is the failure point with the smallest norm lying on the fail/pass boundary. The smallest norm indicates highest occurence probability. 71
4.11 The Hypersphere (HS) Most-Probable Failure Point (MPFP) search algorithgm. The origin of the hypersphere is the nominal bitcell. 72
4.12 An example of MPFP search in a 2D variability space for two skew variables n_1 and n_2. The hypersphere surface in 2D is a circumference. The algorithm detects successively F1, F2 and F3 and finally F4 as MPFP. 73
4.13 Failure probability estimation propsed in [98]. The method takes into consideration only the shaded area of the failure region and exludes the grayed parts, thus results in an underestimation of the failure probability. 74
4.14 Considering an event A that is characterized by a random variable X and its the probabability density function $f(x)$, the interested low probability tail region $A_{interest}$ can be oversampled using a different random variable $\hat{X}$ that has its probability density function $\hat{f}(x)$ centered at $A_{interest}$. This is the key idea behind Importance Sampling. 76
4.15 Different well architectures in 28nm UTBB FD-SOI [99]. In Dual-Well bitcell the NMOS is lying on a P-Well and PMOS on a N-Well as in the bulk CMOS technology. In the Single-P-Well architecure, NMOS and PMOS shares the same P-Well, which allows lowering the Vth of PMOS. The P-Well back-plane voltage V_{Bp} is set to 0V, which puts the PMOS in forward-body bias mode 79
4.16 The 6T bitcell simulation netlist schematic for transisient simulation. The bitcell under test is initialized to store "1" in the node L, whereas other cells of the same column store the oppposite content. The number of cells in a given column is denoted as nb_rows. 80
4.17 SPHD SPW bitcell high-frequency RA MPFP MSVs with 100 mV SA_{offset}, at V_b=0V. Simulations are perfomed for 64,128 and 256 rows. RA fails are tied to slow PGR and PGR skew factor decreases with the increase in the column lenght. The MPFP MSV are same for DW bitcell (not shown), since the NMOS PGR is not affected by the well change. 81
4.18 SPREGLV bitcell high-frequency RA MPFP MSVs with 100 mV SA_{offset}, at V_b=0V. Simulations are perfomed for 64,128 and 256 rows. RA fails are tied to slow PGR. PGR skew factor decreases with the increase in the column lenght. change. . . 82

4.19 RA failures low-frequency MPFP MSVs at Vdd=0.5 and 0.6V and Vb=0V, for 64, 128 and 256 rows. Tthe failures are mainly caused by stability failures as in static conditions. . . 82
4.20 SPHD bitcell high-frequency WA MPFP MSVs for A)dual-well and B)single-p-well bitcells at various Vdd, for Vb=0V. . 84
4.21 SPHD DW bitcell WA MPFP MSVs at high-frequency (purple bars) and low-frequency (gray bars), at Vb=0V. Larger WL pulse width results in a change of the failure mechanism. 84
4.22 SPHD bitcell high-frequency WA and RA yield for both DW and SPW architectures. The sigma-yield are calculated by perfoming Importance Sampling around the MPFP. 85
4.23 SPREGLV SPW bitcell high-frequency WA MPFP MSVs for at V_{dd}=0.5V and 0.6V, at Vb=0V. Results for p=0.5 (relaxed) and p=0.8 (aggressive) are shown. 86
4.24 SPHD bitcell RAW MPFP MSVs at high-frequency operations at V_b=0V, for DW on the top and SPW on the bottom. 87
4.25 SPHD bitcell high-frequency WA, RA and RAW sigma-yields for DW and SPW architectures, at V_b=0V. RAW V_{min} for 64 and 128 rows is 70mV lower than single pulse test. DW bitcell V_{min} becomes read limited with 256 rows. The SPW V_{min} bitcell is limited by write discharge failures thus the multiple pulse analysis gives the same yield as the single-pulse analysis. 88
4.26 SPREGLV bitcell RAW MPFP MSVs at high-frequency operations at V_b=0V. 89
4.27 SPREGLV bitcell high-frequency WA, RA and RAW sigma-yields, at V_b=0V. RAW test leads to a lower V_{min} then WA test with p=0.8. With 256 rows, bitcell V_{min} becomes read limited. 90
4.28 SPHD SPW SRAM macro RA measurement BER vs. V_{dd} for various WL pulse widths, at V_b=0V [114]. The RA yield degrades with shorter WL pulse. 91
4.29 SPHD SPW SRAM macro RA measurement BER vs. V_b at V_{dd}=0.5V and 5ns WL pulse width [114]. The BER is improved with the increase in V_b (strengthen NMOS), confirming that the failure mechanism is tied to the slow Pass-Gate device as it is observed in simulations. 92
4.30 SPHD SPW SRAM macro WA measurement BER vs. V_b, at 5μs WL pulse width under various V_{dd} [114]. WA measurement BER evolves with respect to the V_b and exposes two different write failure mechanisms at $V_b < 0$ and at $V_b > 0$. . 93

4.31 SPHD SPW SRAM macro RA, WA and Read-Stability (RS) V_{min} vs. WL pulse width, at V_b=0V [114]. It is shown that the limiting operation switches to write from read for WL pulse larger that 10 ns. RS yield do not evolve with the presented WL pulse width range, confirming the suitability of this criterion for static-like tests. 94
4.32 SPREGLV SPW bitcell SNM μ/σ with respect to V_{dd}, The yield drop occurs around V_{dd}=0.9V, and becomes dramatic at higher V_{dd}. 95
4.33 On the top, the current flow in the half-cell leading to a SNM failure and causing the loss of the high-logic level. On the bottom, SPREGLV SPW bitcell SNM MPFP MSVs at diferent V_{dd}. SNM MPFP MSV at V_{dd}=1V has a much smaller PUL skew component comapred to MPFP MSVs at lower V_{dd} values. 96
4.34 SPREGLV SNM and WM simulations μ/σ ratio at worse case corners and temperatures (FS 125 °C for read, SF -40 °C for write), with V_b tied to 0V (crosses) and with adaptive V_b (squares) in which V_b is decreased to -1V above V_{dd}=1V. . . 97
4.35 SPREGLV SNM and WM 6-σ yields at worse-case corners and temperature with blind usage of body-biasing. The minimum V_{min} that can be reached is 0.565V. 98
4.36 SPREGLV SNM and WM 6-σ yields at worse-case corners and temperature with smart usage of body-biasing. 0.489V V_{min} can be reached. 98
4.37 SPREG SNM and WM 6-σ yields at worse-case corners and temperature with blind usage of body-biasing. The minimum V_{min} that can be reached is 0.591V. 99
4.38 SPREG SNM and WM 6-σ yields at worse-case corners and temperature with smart usage of body-biasing. 0.474V V_{min} can be reached. 99
4.39 SPHD SNM and WM 6-σ yields at worse-case corners and temperature with blind usage of body-biasing. The minimum V_{min} that can be reached is 0.691V. 100
4.40 SPHD SNM and WM 6-σ yields at worse-case corners and temperature with smart usage of body-biasing. 0.550V V_{min} can be reached. 100

5.1 Two-state drain current fluctuation caused by RTS noise generated by a single trap. The low current level corresponds to a filled-trap state, whereas the high current level is associated with an empty-trap state [116]. 104
5.2 Contribution of RTS Noise on SRAM design margins vs. technology nodes [134]. 106

5.3 Schematic diagram of RTS noise by a single trap in bulk N-Channel CMOS device. RTS is attributed to trapping/detrapping events caused by defects near the silicon - oxide interface. High and low states correspond to carrier capture and emission. 107
5.4 Schematic diagram of RTS in N-Channel UTBB FD-SOI device. Trapping/detrapping events occur at the silicon/gate dielectric interface and at the silicon/BOX interface. 108
5.5 Cumulative probability distribution of number of traps in 28nm UTBB FD-SOI SPHD 6T SRAM Bitcell NMOS Pull-Down (PD) and PMOS Pull-Up(PU) transistors, at both FG and BG dielectrics. It is assumed that the number of traps in a given dielectric follows Poisson distribution. 109
5.6 Cumulative probability distribution of δV_{th} of FG dielectric traps in 28nm UTBB FD-SOI SPHD 6T SRAM Bitcell NMOS Pull-Down (PD) and PMOS Pull-Up(PU) transistors. It is assumed that traps are uniformly distributed in the gate dielectric. 110
5.7 Schematic diagram of the capacitive network between the front-gate and the back-gate of an UTBB FD-SOI transisotr. 112
5.8 TCAD simulations (crosses) vs. 2-Dirac model (straight lines) for Q_i, τ_c and F_{OX}. 113
5.9 $< \tau_c >$ and $< \tau_e >$ evolution with respect to V_g and V_b for $E_{trap} = 0.25$eV (low energy) and 0.6eV (high energy) for FG dielectric. at ambient temperature. 114
5.10 $< \tau_c >$ and $< \tau_e >$ evolution with respect to V_g and V_b for $E_{trap} = 0.25$eV (low energy) and 0.6eV (high energy) for BG dielectric. at ambient temperature. 115
5.11 The equivalent RTS-aware transistor model with multiple traps: N traps at FG, M traps at BG. A voltage source modeling V_{th} fluctuation caused by $i-$th trap is denoted as R_i. The overall V_{th} modulation is the sum of the independent δV_{th}s. 116
5.12 RTS-aware simulation flow chart. 116
5.13 MATLAB RTS PWL Generator scheme. 117
5.14 The proposed optimization algorithm to model bias-dependent behavior of RTS time constants. Very slow τ_c that occurs at V_g=0V is canceled out and time constats are calculated only if $V_g > (V_{dd} - \epsilon)$. 118
5.15 An RTS-aware 6T SPW bitcell netlist where each device has one FG trap and one BG trap. The bloc "R" represents the PWL voltage source modeling V_{th} fluctuation cuased by a single trap. 119
5.16 On the top: PWL Voltage source waveform of a single trap. On the bottom: The overall V_{th} modulation waveform with multiple traps. 119

5.17 RTS measurement setup that is used in the dynamic characterization module. 120
5.18 Single Write WA measurements BER vs. V_b for no overdrive, only-FG-overdrive, only-BG-overdrive and both FG-and BG overdrive, at 80ns WL pulse width and $Vddm_{test}$ = 0.34V. The BER improvement with overdrive is interpreted as the impact of the RTS in the PMOS PU1, which slows down the device, thus cancels some of discharge failures. Results are normalized with respect to those obtained at V_b =-0.1V . . . 122
5.19 Single Write WA simulation BER vs. V_b for no RTS, only-FG-RTS, only-BG-RTS and both FG-and BG RTS, at 80ns WL pulse width and $Vddm_{test}$ = 0.34V. Results are normalized with respect to those obtained at V_b =-0.1V. 122
5.20 Single Write WA measurements BER vs. simulation BER with respect to V_b for FG overdrive and the FG-RTS simulation, at 320 ns WL pulse width and $Vddm_{test}$ = 0.34V. Results are normalized with respect to those obtained at V_b =-0.1V. 123
5.21 Multiple Write (WAW) WA measurements BER with respect to V_b for no overdrive and only-FG-overdrive, at 320 ns WL pulse width and $Vddm_{test}$ = 0.34V. The impact of RTS is observed at extreme V_b values, as a BER decrease at negative V_b (discharge failure zone) and as a BER increase at positive V_b (completion failure zone). 123
5.22 Multiple Write (WAW) WA simulation BER at $Vddm_{test}$ = 0.34V and 320 ns WL pulse width for no-overdrive and only-FG-overdrive. Two exreme V_b conditions are simulated. At negative V_b BER decreases with overdrive (discharge failures) and at positive V_b, BER increase with overdrive ((completion failures). 124
5.23 Multiple Write (WAW) WA measurements BER with respect to V_{dd} for no overdrive and only-FG-overdrive, at 108 ns WL pulse width and V_b = 0.5V. The overdrive results in a 30mv increase in the minimum Vddm that can be reached without having write failures. 125
5.24 The illustration of the RTS-sensitive zone in SRAM bitcell in a 2D variability space. RTS is not a limiting variability souce in 28nm UTBB FD-SOI technology and becomes visible when the memory is already in the failure zone. 126

List of Tables

4.1 The number of sigma vs. covered area under the bell-shaped normal distribution curve. 61
4.2 Smart Body-Biasing V_{min} gain vs. Bitcell Area 101

Glossary

- ABB : Adaptive Body Biasing
- AC : Alternative Current
- BG : Back-Gate
- BL : Bit-Line
- BLL : Bit-Line Left
- BLR : Bit-Line Right
- BOX : Buried Oxide
- CDF : Cumulative Density Function
- CMOS : Complementary Metal Oxide Semiconductor
- DC : Direct Current
- DFM : Design For Manufacturing
- EM : Electromigration
- FF : Fast-Fast Process Corner
- FS : Fast-Slow Process Corner
- FG : Front-Gate
- FD-SOI: Fully Depleted Silicon on Insulator
- ITRS : International Technology Roadmap for Semiconductors
- LER : Line Edge Roughness
- MFPF : Most Probable Failure Point
- MSV : Mean Shift Vector
- NBTI : Negative Bias Temperature Instability

- OPC : Optical-Proximity correction
- PG : Pass-Gate
- PBTI : Positive Bias Temperature Instability
- PVT : Process-Voltage-Temperature
- PDF : Probability Density Function
- PU : Pull-Up
- PD : Pull-Down
- RDF : Random Dopant Fluctuations
- RTS : Random Telegraph Signal
- RD : Read
- RA : Read-Ability
- RAW : Read-After-Write
- SA : Sense Amplifier
- SCE : Short Channel Effects
- SPHD : Single Port High Density
- SPREG : Single Port Register File
- SPREGLV: Single Port Register File Low-Voltage
- SS : Slow-Slow Process Corner
- SF : Slow-Fast Process Corner
- SNM : Static Noise Margin
- SNML : Static Noise Margin Left
- SNMR : Static Noise Margin Right
- SRAM : Static Random Access Memory
- TOX : Thickness of oxide layer
- TT : Typical Process Corner
- ULV : Ultra Low Voltage

- UTBB FD-SOI: Ultra Thin Body and Buried Oxide Fully Depleted Silicon on Insulator
- UWVR : Ultra Wide Voltage Range
- WR : Write
- WAW : Write After Write
- WA : Write Ability
- WM : Write Margin
- WL : Word Line
- VTC : Voltage Transfer Function

Chapter 1

General Introduction

In the ultra-deep-submicron technology era, the downscaling of device geometry towards its physical limits exacerbates the impact of the inevitable atomistic phenomena tied to matter granularity. In this context, many different variability sources raise and affect the electrical characteristics of the manufactured devices, even though when they are identically-designed. The process variability is thereby seen as the biggest obstacle for further downscaling, since it introduces additional manufacturing and design challenges at each new technology node. Starting from the early 80's, many works in the literature have investigated the impact of the process variability at a device level. During the last decade, the variability impact in the circuit level, which was traditionally seen as a concern for analog circuits, has become a popular research topic also in the field of digital circuit design, since the increased number of transistors in the modern integrated circuits had led to a large statistical variability affecting dramatically circuit functionality. A particular interest is carried out on to the Static Random Access Memory (SRAM) circuits which are manufactured with the most aggressive design rules in a given technology node and contain billions of transistor. SRAM circuits are therefore seen as the principal victim of the process variability, since their stability and performance are severely impacted by the process variability which stands as the main obstacle for the further reduction of the bitcell area and of its minimum operating voltage. The reduction of the latter is a very important parameter for Low-Power design, which is one of the most popular research fields of our era. The optimization of SRAM bitcell design therefore has become a crucial task to guarantee the good functionality of the design at an industrial manufacturing level, in the same time answering to the high density and low power demands. However, the long time required by each new technology node process development means a long waiting time before obtaining silicon results, which is in cruel contrast with the fact that the design optimization has to be started as early as possible. An efficient SPICE characterization methodology for the minimum

operating voltage of SRAM circuits is therefore a mandatory requirement for design optimization.

This research work concentrates on the development of the new simulation methodologies for the modeling of the process variability in ultra-deep-submicron SRAMs, with the ultimate goal of a significantly accurate modeling of the minimum operating voltage V_{min}. An existing Monte Carlo simulation based methodology is improved to increase the modeling accuracy of the distribution tails under large variations. The proposed modeling approach is validated though a large set of silicon measurements that have been performed for the process monitoring database in C40 technology node, showing the good accuracy in the estimation of the static silicon V_{min} taking into account the process variability impact at an industrial production level. However, the Monte Carlo based analysis is not always suitable for yield investigations during the design optimization.

In advanced technology nodes, in contrast to the conventional static test criteria, SRAMs have also to be analyzed under dynamic conditions, which increases evidently the analysis complexity, since it does not only introduce new test metrics, but also introduces a new complexity level that appears as the dependency of the failures to the time during which the given bitcell-under-test is connected to the external world. A smart algorithm based on the use of hyperspherical surfaces has been developed to investigate a simplified version of the full variability space, offering a simple-but-efficient modeling of the process variability impact on a given bitcell design. The proposed algorithm is not just used for an accurate modeling under large variations, but also as an investigation tool allowing for the extraction of different mismatch mechanisms underlying behind the V_{min} limiting failures in a given bitcell. The efficiency of the proposed methodology is demonstrated through simulations and has been validated elsewhere through silicon measurements. The knowledge acquired about the different bitcell V_{min} limiting failure mechanisms have served the purpose of providing solutions during the 28nm Ultra-Thin Body and Buried Oxide Fully Depleted Silicon On Insulator (UTBB FD-SOI) technology development to overcome SRAM bitcells V_{min} limitations.

Finally, a particular interest is carried out on the time-dependent subclass of the process variability, which appears as a change in the electrical characteristics of a given transistor during its operation and during its lifetime. The Random Telegraph Signal (RTS) noise, originating from the capture and emission of charge carriers by defects in the surrounding dielectric layers, is known as the one of the main time-dependent variability sources in ultra-deep-submicron technologies. Considering the high vulnerability of SRAMs to the transistor mismatch and the growing impact of the RTS noise with the downscaling, the integration of the RTS noise into the SPICE-level bitcell design optimization has become a requirement. A transistor-level bias-dependent RTS noise model peculiar to UTBB FD-SOI

technology, considering the front- and back-gate coupling of the device, has been developed and integrated in the form of SPICE netlists, allowing to perform RTS-aware simulations. The accuracy of the model to describe the change trends in the bitcell stability metrics due to RTS noise has been verified through silicon measurements.

This research work has led to many publications and one patent application. The majority of findings are retained by STMicroelectronics SRAM development team for a further use in their design optimization flow.

Chapter 2

Variability in Ultra-Deep-Submicron CMOS

This chapter presents the variability phenomena that occur in the semiconductor manufacturing process. First, different class of process variability and the sources of the different contributors are discussed. Later, improvement techniques that have been used to reduce the variability in conventional bulk CMOS technologies and new device architectures that have been proposed to overcome bulk CMOS limitations, are presented.

2.1 Introduction to Variability

The semiconductor industry has delivered increasing performance and reduced cost for Complementary Metal Oxide Semiconductor (CMOS) technology over the last 40 years following the Moores law [1]. This was possible by achieving the device scaling requirements and thus cramming more transistors in the same area size at each new technology node. However at sub-100nm technologies, the device scaling became more and more challenging at each new technology node due to the increasing complexity in the manufacturing process and to the resulting variability impact in the electrical characteristics of the manufactured transistors. In this chapter, this phenomenon called variability, which affects the semiconductor manufacturing process as variations in the electrical characteristics of identically-designed transistors, is discussed.

Considering an integrated circuit (IC) fabricated in different fabrication facilities, the identically-designed transistors would have different electrical characteristics, since the equipments across different sites do not have exactly identical outputs. Moreover, considering wafers manufactured in the same fabrication site, due to small deviations in the manufacturing condi-

tions over a given time interval or a spatial distance, the electrical characteristics of transistors would vary from one lot to another lot (lot-to-lot), from one wafer to another wafer (wafer-to-wafer) in a same lot or from one die to the another die (die-to-die) in a same wafer. The output of a manufacturing equipment can even vary rapidly over distances smaller than the dimension of a die resulting in within-die variations. In other words, at an industrial manufacturing level, it is impossible to keep the uniformity in the electrical characteristics of the manufactured identically-designed transistors. The first work presenting the variability in the semiconductor manufacturing process is published by the co-inventor of the transistor, William Shockley, in 1961 [2]. A study covering the impact of the process variations on the transistor threshold voltage sensitivity is published in 1974 [3].

A terminology describing the different class of the variability in CMOS transistors and circuits has first to be defined for a better understanding of the different sources and their consequences. The first class is the systematic process variability, which is associated with the manufacturing equipment and appears as a parameter drift lot-to-lot, wafer-to-wafer, across wafer and across chip. The systematic variability is associated to the design layout and introduced by lithography, strain and well proximity effects. The second class is the random variability which is due to the discreteness of charge and granularity of matter. The random variability will be still present, even if the manufacturing process is ideal, i.e. in the absence of the systematic process variability, since it depends on the atomistic phenomena.

In this chapter, first the different variability sources in CMOS technology and their consequent limitations are presented. Later, the techniques used to reduce the variability impact at manufacturing-level as well as on design-level are discussed.

2.2 Variability Sources in CMOS

2.2.0.1 Systematic Variability

The systematic variability sources are attributed to the different steps in the manufacturing process [5]. Historically, the systematic variability, which results in parameter drifts in identically-designed MOSFETs separated by long distance, or fabricated in different time, is seen as the dominant variability source. These parameter drifts are deterministic shifts in space and time of process parameters in same direction and are in general spatially correlated. In the detail, the systematic variability causes shifts in the mean value of the sensitive design parameters, including channel length (L), channel width (W), layer thickness, resistivity, doping density and body effect. The term "global" is also commonly used to describe this class of variability. Furthermore, the continuous downscaling has introduced significant changes in the device structure, in the processing and in the material leading

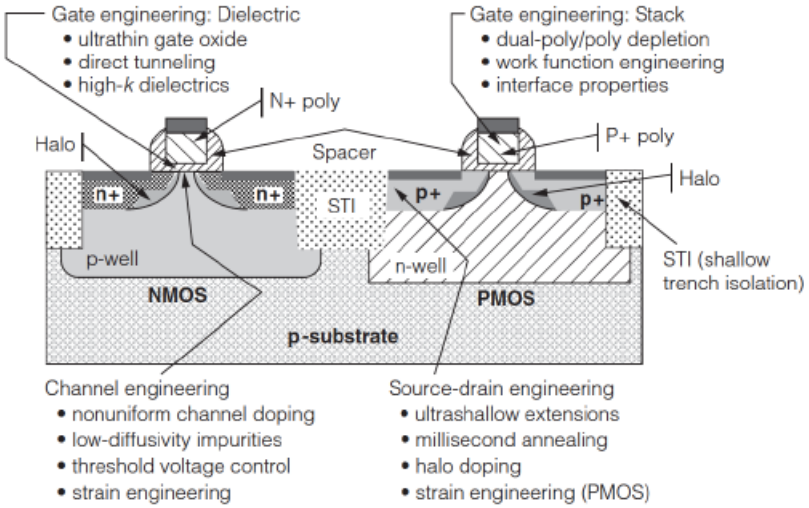

Figure 2.1: CMOS cross section with major sources of process variability [4]

to very complex manufacturing in advanced bulk CMOS device. Different process and engineering steps that are used to manufacture a bulk CMOS device is illustrated in figure 2.1. As a consequence the systematic variability evolves with each technology node. Although the systematic variability is well characterized and in many case predictable during the design and the manufacturing level, it cannot be eliminated. The main contributors of the systematic variability are classified as the following [6]:

- Variations due to the physical implant and anneal process introduced through halo implantation, the accuracy and the purity of dose and the variations in the peak anneal temperature [7].
- Chemical metal polish variability in the polishing of Shallow Trench Isolation (STI) [8] leading to gate height variation in both polysilicon and metal gates [9].
- Variation in film thickness impacting oxide thickness, gate stacks, wire and dielectric layer height, due to the deposition and growth process as well as the chemical-mechanical planarization step [5].
- Temperature non-uniformities in the critical post-exposure bake and etch steps [5].

The systematic variability has also a design-dependent sub-class which is mainly related to the manufacturability of the design layout. The overlay error, mask error, shift in wafer scan speed, rapid thermal anneal and the

dependence of stress on layout have become notable sources of the systematic variations. The photolithography and the etching processes contribute significantly to variations in nominal lengths and widths of the device due to the complexity required to fabricate lines that are much narrower than the wavelength of light used to print them [10] (Figure 2.2). Optical-proximity correction (OPC) [11], phase-shift masking (PSM), layout induced strain and well-proximity effects are other contributors for the systematic variability. It has to be noted that the systematic variability do not have a time-dependency meaning that the electrical characteristic change of a given device remains same over the life-time of the transistor.

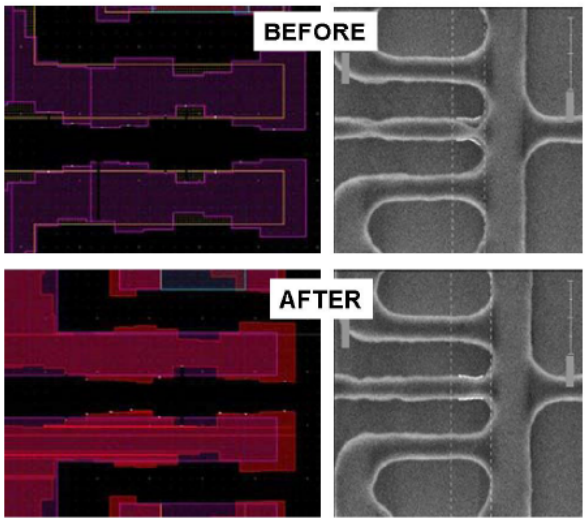

Figure 2.2: Lithography induced variability and OPC based improvement. The layout images (left) and the Scanning Electroin Microscope (SEM) images (right) of the manufactured patterns are shown [12].

2.2.0.2 Random Variability

The random variability is caused by the granularity of charge and matter and has strongly exacerbated by the aggressive downscaling of devices. Unlike the systematic variability, the randomness causes parameter changes in an individual instance of a given device in an unpredictable manner. The random variability produces differences in the electrical characteristics of microscopically identical transistors. In other words, the random variability creates parameter changes in identically-designed MOSFETs across a very short distance, even for two neighbor transistors separated by the smallest possible distance in a given technology node. In the conventional bulk CMOS technology, the main contributor for the random variability is the

Random Dopant Fluctuations (RDF) [7, 13–15], which results from the discreteness of dopant atoms in a transistor channel. The number of dopants is inversely proportional to the channel geometry, thus it decreases exponentially in each new technology node (Figure 2.3). As the device dimensions are scaled down, the reduced number of dopants in the channel and their random position had a significant impact on device electrical characteristics. The number of dopants is a discrete statistical quantity and two transistors sitting side by side have different electrical characteristics because of the randomness, resulting in device-to-device variability. The first order effect of RDF being a random shift in the threshold voltage, the electron transport variability is also observed as a second order.

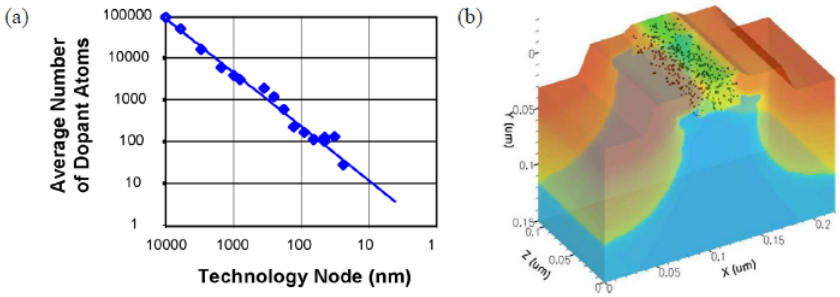

Figure 2.3: (a) Average number of dopants evolution in MOS Transistor channel w.r.t technology node [18] (b) 3D View of a numerical MOS transistor model simulating the number of dopants in the channel in 65nm and 45nm technology nodes [12].

The second major source for the random variability in CMOS technology is the Line-Edge Roughness (LER) [15–17] which stems from subwavelength lithography causing variation in the critical dimension of the manufactured device size. The impact of LER increases as transistor dimensions shrink. LER introduces significant variability in subthreshold current as well as threshold voltage and causes the degradation of I_{ON}/I_{OFF} ratio [18].

The Interface roughness and oxide thickness (TOX) variation [19] is another variability source introducing significant process variability in sub-65nm technologies through Si/SiO_2 and polysilicon-gate/SiO_2 interfaces. The Polysilicon Granularity [15,20–22] increases the uncertainty in gate doping enhancing the gate dopant diffusion along the grain boundaries, leading to a non-uniform polysilicon gate doping and a potential localized penetration of the dopants through the gate oxide into channel region. The use of high-K gate dielectric is introduced below 32nm technology node to provide a thicker physical TOX reducing the direct-tunneling gate leakage current while ensuring an ultra-thin electrical TOX required for the continual scaling of MOSFET devices [23]. This change in the process introduces a significant

variability due to the interface roughness between Si and the high-K dielectric and between the high-K dielectric and the metal gate causing mobility degradation and the equivalent TOX variation [19]. In addition, the random polycrystalline-like texture of high-K dielectrics causes fluctuations in the channel potential under the gate which also increases the variability. [4].

The impact of the main sources of random variability on the threshold voltage of MOSFETs has been shown to be relatively statistically-independent [24], so that the threshold voltage variability can be analytically modeled by a statistical addition of the individual variability sources [6]. The random variability can not be canceled by nature, since it is tied to the atomistic phenomena and may even cause a time-dependent impact meaning that the electrical characteristics of a given device can change over the product-life time.

2.2.0.3 Time Dependency in Variability

The variability is named as static (non-time dependent) if the impact remains same along the transistor life, or dynamic (time-dependent) if the impact evolves along the transistor life. Previously described variability sources are all static so they do not have time-dependency. The term "dynamic" indicates that the device electrical characteristics evolve with respect to the environmental conditions (supply voltage, temperature etc.). The dynamic variability in semiconductor devices stems from defects in the dielectric layer, aging and wear-out mechanisms.

The splendid progress in CMOS technology has been made possible by the unique properties of silicon dioxide (SiO_2), which is an excellent electrical insulator and capable of forming a nearly perfect electrical interface with its substrate [25]. The demands from the gate-oxide have grown with downscaling in each new technology node and the quality of the gate-oxide needs to be of the highest quality, which has resulted in the use of a stack of high-K materials instead of thermally grown SiO_2 as mentioned in the previous section. However, there are imperfections in the high-k dielectric to silicon interface due to the inevitable atomistic phenomenon leading to defects in the dielectric layer named as "traps". [26] (Figure 2.4). These traps are originated from structural oxidation, metal impurities and different kinds of bond breaking processes, hot carrier stress or other phenomena [27]. The traps in the dielectric layer are in electrical communication with the underlying silicon; a trap can capture a majority charge carrier from the channel or can emit a previously captured carrier to the channel. This phenomenon also named as "trapping-detrapping" in the dielectric layer is known as Random Telegraph Signal (RTS) Noise [28]. The influence of the traps in the dielectric layer on the performance of a semiconductor device is determined by the density of states and the probability that these states are occupied by a charge carrier. The discrete levels of a two-level fluctuation

of a given trap, filled or empty, correspond to a high and low conductance in the channel causing changes in drain current over time. For a given trap, the physics behind the times in the filled and empty states are defined by the Shockley-Read-Hall statistics [29]. Therefore, the trapping-detrapping in the dielectric layer is a statistical variability source in CMOS devices, but its impact is time-dependent in distinct from the previously presented variability sources, since the times in the filled and empty states for a given trap are bias-dependent. The RTS noise is investigated in this manuscript with a particular interest on its modeling and the findings are presented in the last chapter.

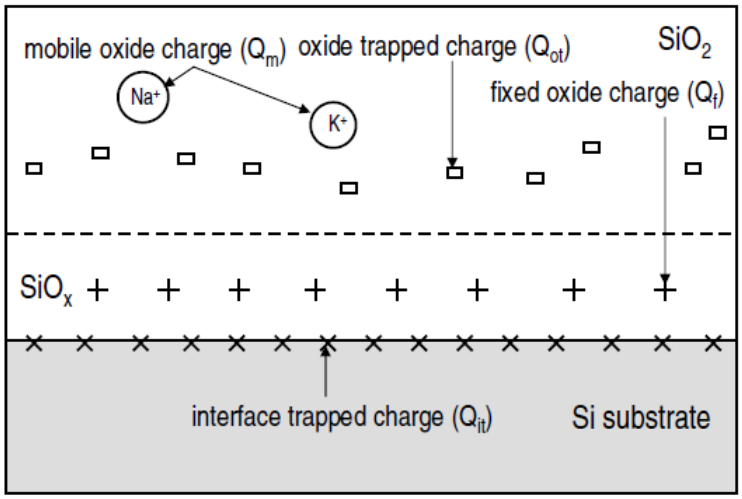

Figure 2.4: Different types of traps associated with Si-SiO_2 interface [25].

The CMOS technology asserts different aging and wear-out mechanisms that have to be anticipated during the design process. The Negative Bias Temperature Instability (NBTI) which arises from the generation of interface states and positive trapped charge at the silicon/dielectric interface or in the oxide layer while the device is in operation. NBTI reduces the performance of p-channel MOSFETs by increasing the threshold voltage of the device [30–32]. With the introduction of high-k metal gates, the Positive Bias Temperature Instability (PBTI) [33] has appeared as a new degradation mechanism. The PBTI affects n-channel MOSFETs when positively biased. Hot-electron effect (HotE) degrades n-MOSFET on-current by injecting additional charge into the gate oxide which must be overcome in order to turn the device on [34]. The electromigration (EM) [35] is gradual displacement of metal atoms, depleting the interconnect of conductor atoms over an extended period. EM arises from the high current densities in excess of the reliable limit of the wire causing the drift of metal ions in the direc-

tion of the electron flow. As the structure size decreases, the significance of EM increases. The aging and the wear-out mechanisms are strongly related to how often and how long the device is on and to the chip environmental operating conditions.

2.3 The Improvement Techniques for Variability and New Device Architectures

With the continued down-scaling of MOSFET devices towards their ultimate physical limits, the variability had become a critical design parameter for VLSI circuits and emerged as a major technological barrier for further downscaling. Traditionally, the systematic variability have been the main concern in digital circuits and handled in the design process through the worst-case modeling [36] [37,38], whereas concern for within-chip statistical variations has been in the domain of analog circuit design [39]. However, advanced CMOS integrated circuits are large enough (large number of transistors) that device and interconnect parameter variations within a chip are as important as chip-to-chip variations. Furthermore, each new technology node introducing new engineering challenges to overcome scaling limitations results in very complex manufacturing (figure 2.1). As a result, the within-die fluctuations have exceeded the die-to-die variations [40, 41] and have become the main threat to the performance and functionality for digital circuits [40].

Since 90nm technology node, the variability has taken seriously into consideration in the design process in order to ensure the design functionality over a large number of manufactured circuits. Several improvement techniques have been developed in order to overcome variability-related limitations. The systematic variability is strongly mitigated with the improvements in the manufacturing equipment, the maturation of the technology and the co-operation between process and design techniques, as well as purely design based improvements [42]; the systematic variability is managed using design for manufacture (DFM) [43] tools as well as through the regularized design [44] which optimizes layouts [5, 45]. The improvement in OPC largely mitigated the lithography induced systematic variability, but the introduction of strain for mobility enhancement exacerbates this last one due to its non-uniform application introducing additional variations which depend on channel length.

However, as the technology is downscaled, the random variability within dies overwhelms the systematic variability. The random variability is by nature much harder to cope, since it is not related to the quality of manufacturing equipment, or to the manufacturability of the design, but related to atomistic limitations [46, 47]. The miniaturization along different technology nodes made very difficult to keep a good electrostatic control of the

channel, which as a consequence increases the leakage current (I_{OFF}) and thus decreases the I_{ON}/I_{OFF} ratio (Figure 2.5). A more important doping is needed to improve channel electrostatic which made increase RDF, which therefore stands as the principal contributor for local threshold voltage variability [15, 22]. RDF together with other stochastic variability sources like LER become a huge obstacle not only for the downscaling of device geometry, but also for the downscaling of the operating voltage; which is a very important parameter for Low-Power design [48]. The miniaturization dramatically increases the short-channel effect (SCE) [49] thus degrades the control of the channel electrostatic, while increasing the local threshold voltage variability (σ_{VTH}). The introduction of high-K metal gate stack at 45-32nm provides serious improvements for limiting both SCE and RDF, but it remains inefficient below 28nm technology. The impact of the different variability components on the threshold voltage variability is illustrated in Figure 2.6.

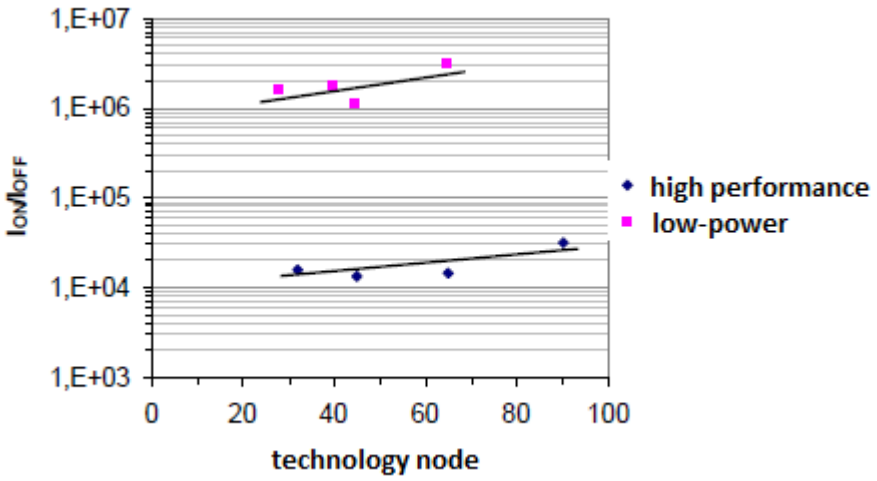

Figure 2.5: I_{ON}/I_{OFF} evolution with respect to the technology node in the planar CMOS technologies for high performance and low power devices [50].

The ageing impact in the CMOS circuits may lead to a significant performance lost [52–55] over time and therefore has to be anticipated during the design process. Furthermore, the resulting impact may totally destroy the circuit functionality and in that case, some design-level correction methods are required to compensate the ageing impact. Supply voltage scaling [56], which consists in gradually increasing the operating voltage, has been proposed to compensate the NBTI induced degradation. However, since the voltage scaling increase the dissipation of the dynamic power quadratically, which in turn increases the temperature, it ultimately accelerates the NBTI

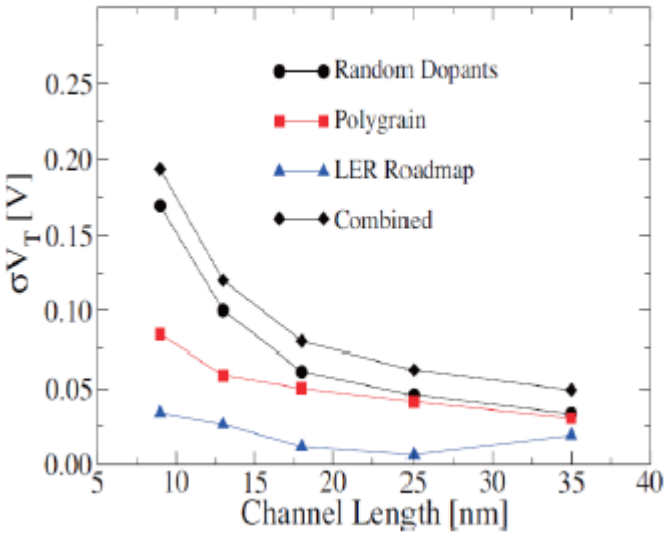

Figure 2.6: The threshold voltage variability (σ_{VTH}) with the different components w.r.t the gate length for planar CMOS technologies [51].

effect. The increase in the power consumption is also not acceptable for designs that have power consumption constraint. Adaptive body-biasing techniques (ABB) [57,58] have been proposed to adjust transistor strengths through modifying the gate-source biasing in order to maintain same attributes as in the beginning of the product life and to produce a stable drain current regardless of the threshold voltage shift caused by the ageing. ABB requires the use of the variability sensors and the detect-and-correct circuits which as a consequence introduce an area penalty.

Therefore new architectural choices had to be made to overcome bulk CMOS limitations in order to continue to follow Moore's Law. Two promising new process technologies have been developed in this purpose.

FinFET [59,60] is a multiple-gate transistor manufactured as a 3-D structure and consists of a vertical silicon fin, which is wrapped by the gate. The thickness of the fin, measured in the direction from source to drain, determines the effective channel length of the device. The wrap-around gate structure provides a better electrostatic control over the channel thus reduces the leakage current and overcomes other SCEs. The 3-D structure gives more volume than a planar gate for the same planar area, enabling high density integration, but requires fundamental changes in the manufacturing process and re-designing cell libraries and IP block. The width of FinFET transistors (W_{Si}) is quantized by nature of manufacturing meaning that the transistors cannot be sized freely, which is not designer-friendly.

The Fully Depleted Silicon-on-Insulator (FD-SOI) [61, 62] is a planar

technology that relies on two primary innovations: An Ultra-thin layer of insulator, called buried oxide (BOX), is positioned on top of the base silicon and a very thin silicon film implements the transistor channel. Thanks to its thinness, there is no need to dope the channel thus making the transistor fully-depleted and suppressing the main variability contributor of bulk CMOS, RDFs. The short-channel electrostatics is determined mostly by the silicon film thickness T_{Si} allowing a better short-channel control compared to the bulk technology, which allows a good scalability as well as the reduction of the LER impact on threshold voltage variability. The BOX layer lowers the parasitic capacitance between the source and the drain, efficiently confining the electron flow from the source to the drain and reducing performance-degrading leakage currents. The BOX also make the body biasing much more efficient compared to the bulk technology thanks to the reduced parasitic leakage, enabling control of transistor flavor through polarizing the substrate underneath the device [62]. One additional source of variability in FD-SOI compared to bulk devices is the T_{Si} variation across the wafer [63]. However, SOI wafers fabricated using SmartCut technology [64] have demonstrated good within-wafer-uniformity ($< \pm 5$A), allowing low threshold voltage variability. For a clear comparison and a better understanding between different device architectures, the schematic of bulk CMOS, FinFET and FD-SOI devices are illustrated on Figure 2.7.

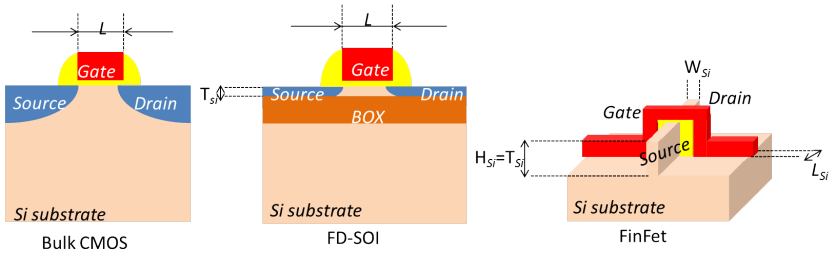

Figure 2.7: Schematics of Bulk, FD-SOI and FinFet transistors.

2.4 Conclusion

With the aggressive downscaling in CMOS technologies, the variability has become a critical design parameter for designers to ensure the performance and the reliability over large number of manufactured circuits. In older technologies, the systematic variability which is mainly related to the manufacturing equipment quality, was dominant and handled during the design process through worst-case modeling. However, below 65nm technology node, the random variability, which is mostly related to the atomistic mechanisms, has become the dominant variability source introducing local

variations in identically-designed transistors leading to unpredictable impacts on circuit operations.

The random variability is first taken into consideration as a design concern in analog circuits [65]. With the aggressive downscaling, the digital circuits are also concerned by the random variability, since the high integration and the low-power requirements increase the vulnerability to the transistor mismatch within the same die. Due to the extremely high statistical variability in conventional bulk MOSFETs, alternative technologies such as FinFET and Ultra-Thin Body and Buried Oxide (UTBB) Fully Depleted Silicon-On-Insulator (FD-SOI) are developed to allow the continuation of CMOS scaling, the improvement of performance, the reduction of the leakage current and of the statistical variability impact. Both architectures allow reducing the variability impact with respect to the conventional bulk transistor, but the impact can not be eliminated. The variability remains as the biggest obstacle for performance and density improvements in advanced technology nodes. A good understanding of the variability impact on circuit-level and a variability-aware design methodology have therefore become a requirement for design optimization before manufacturing.

Chapter 3

The Impact of the Variability in Static Random Access Memory

This chapter first presents a brief summary on the role of SRAMs in modern System-On-Chips and outlines the basic structures and operations of an SRAM circuit. This is followed by the presentation of multi-dimensional variability space model that is used to reproduce in simulations real-world SRAMs that are strongly impacted by variability during the their manufacturing. Later, SRAM figure of merits that are used to evaluate stability and performance are discussed. SRAM bitcell characterization techniques and the concept of the minimum operating voltage V_{min} are presented after a particular focus on the different bitcell failure mechanisms.

3.1 Introduction to SRAMs limitations in modern System-On-Chips

The variability impact on digital circuits has been a very popular research topic for last 15 years [66–68]. This impact manifests in digital logic circuits in the form of delay and leakage power variability [69]. Considering a signal path passing though multiple gates and each gate having its particular delay, the resulting distribution of the collection of delays is assumed as Gaussian-like [69]. The overall delay reduces as the number of gate in the path increases, since variations along a long path can be averaged out [70]. However, analogue and analogue-like systems like Static Random Access Memory (SRAM) and latch registers are more complex, since they rely for their operation on balanced pairs of transistors. The SRAM circuits are particularly concerned from the variability phenomenon, since they are fabricated with most aggressive design rules for density improvements and

the manufacturing yield degrades dramatically due to the increased in-die variability [71]. Therefore unlike digital logic circuits, the SRAM requires additional correction and redundancy circuits to overcome the statistical variability impact [72, 73], which costs more area in the same chip. The impact of variability on SRAM circuits has therefore become an important research topic among the academics and the semiconductor industries. The published works refer to the statistical modeling of the variability impact and its characterization both on simulation and silicon measurements, as well as to the improvement techniques to increase the SRAM immunity to variability.

The highly demanding low-power market requests to deal with very sophisticated applications and thus expects more performance and higher storage capacity from on-chip memories. In modern System on Chip (SoC) applications, SRAMs can be used as cache memories, temporary buffers and large capacity storage RAMs, therefore occupying a significant portion of the chip area [74]. This stems from the fact that 20-40% of all program instructions require memory [75] and SRAM is the only efficient and fast storage system for processor caches for the amount of data required by a processor [76]. This heavy-use of SRAMs makes them the main contributor for the overall power consumption of a given chip [77]. The equation (3.1) presents the sum of static and dynamic power of a given digital circuit, which depends on the leakage current I_{leak}, the operating voltage V_{dd}, the activity factor α indicating the fraction of the circuit that is switching, the overall equivalent capacitance C and the cycle frequency F.

$$P_{tot} = P_{stat} + P_{dyn} = (1 - \alpha).I_{leak}.V_{dd} + 1/2.\alpha.C.V_{dd}^2.F \qquad (3.1)$$

The equation (3.1) evidences that reducing V_{dd} will reduces both static and dynamic power. Considering the fact that SRAMs are the main contributor for the overall power consumption in a given chip, reducing the memory operating voltage will lead to a significant overall power reduction. [73].

At each new technology node, SRAM bitcell footprint is shrunk following the Moore's Law by a factor of two as shown in figure 3.1, allowing for a higher memory density (potentially double). As a consequence, today's SRAM arrays contain more than a billion transistors, thus their manufacturing leads to a very large variability in the worst-case values of the transistor electrical parameters. In addition, the local random variability impact is also amplified by the aggressive downscaling of the device sizes. This unavoidable variability phenomenon stands as the biggest obstacle for further density improvements, as well as the reduction of the memory operating voltage, since the variability impact is magnified in low-voltage memory operations. Figure 3.2 illustrates the stagnation in the SRAM operating voltage scaling as it is recognized by the International Technology Roadmap for Semiconductors (ITRS).

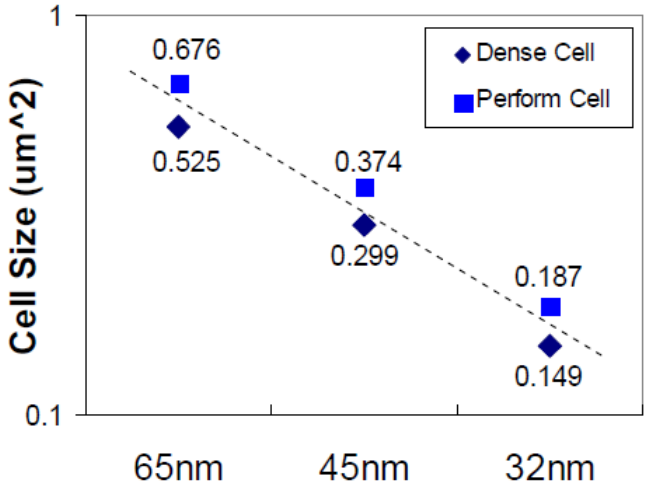

Figure 3.1: SRAM bitcell scaling trend from 65nm to 32nm techonology node for the performance bitcell (squares) and for dense bitcell (diamonds) showing the 50% area reduction [78].

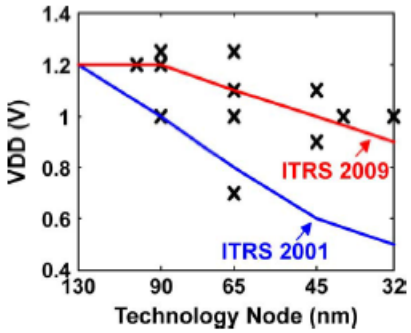

Figure 3.2: SRAM minimum operating voltage reported in ISSCC and VLSI Conferences between 2004 - 2010 (crosses) and ITRS predictions at 2001 and at 2009 (straight lines) [79].

3.2 SRAM Bitcell Architecture and Common Operations

The conventional SRAM bitcell architecture is the standard 6-Transistors (6T) SRAM bitcell, which consists of two cross-coupled inverters for the content storage and two access transistors for the communication with the external world, as shown in figure 3.3. The cross-coupling is performed as that the output of the first inverter is connected to the input of the second

inverter and the output of the second inverter is connected to the input of first inverter, forming a bistable latching circuitry. The latch input and output nodes, L and R in figure 3.3, represent the bitcell internal nodes. The state of an internal node is forced by the complementary state of the other allowing for the data storage. For an optimal performance, the inverters have to be symmetrical and well-balanced in their behavior. The PMOS-NMOS transistors forming an inverter are named as Pull-Up (PU) and Pull-Down (PD) respectively. The access transistors, also called as Pass-Gate (PG) transistor, are controlled via the World-Line (WL) signal and connect the bitcell internal nodes to the Bit-Lines (BL). BLs are external access points to the bitcell, where the information can be read from or written to the cell. To ensure that the access time from both sides is equal in a given bitcell, the access transistors have to be well matched, representing one of the vulnerable points of SRAM to the statistical variability.

An SRAM array is formed by a matrix of cells and each cell is associated to the storage of one bit, hence the name bitcell comes from. Aside from the bitcells, a SRAM contains a significant amount of peripheral circuitry including world line pulse generation, addressing logic, sense amplifier, pre-charge/line buffer and multiplexer circuitry. A block level of SRAM schematic is shown in figure 3.4.

A write operation (WR) in SRAM consist in forcing the bitcell content to the values set on the BLs. Considering the bitcell illustrated in figure 3.3 which has a low-logic level in the node L and high-logic level in the node R, a successful WR results in the toggle of the internal nodes in order to have high-logic level in L and low-logic level in R. The high-logic level and the low-logic level represents the equivalent voltage level that a internal node needs to be able store "1" or "0" logic values respectively. The write is performed by first setting left bit-line (BLL) to high-logic level and discharging the right bit-line (BLR) to 0V. Then the access transistors are turned on via the WL signal providing access to the bitcell internal nodes. The high voltage level in the node R is discharged through the access transistor PG2 and the BLR. When the voltage level of the node R goes below the trip point of inverter formed by PU1 and PD1, the bitcell content toggles; the bitcell stores now a "1" in the node L and a "0" in the node R.

A read operation (RD) from 6T SRAM is performed as follows: Both bit-lines are pre-charged to high-logic level and then access transistors are turned on enabling the connexion between the bit-lines and the bitcell internal nodes. Considering the same bitcell of figure 3.3, during a RD 0 from the node 'L', the BLL is discharged through the PG1 and PD1 transistors generating a voltage difference between BLL and BLR. In case of SRAM arrays which contain sense-amplifiers in the peripheral circuity, the BL voltage difference is captured by a sense-amplifier and the required voltage difference to enable sensing depends on the sense-amplifier design. It is possible to perform successful read without sensing, but this requires a full

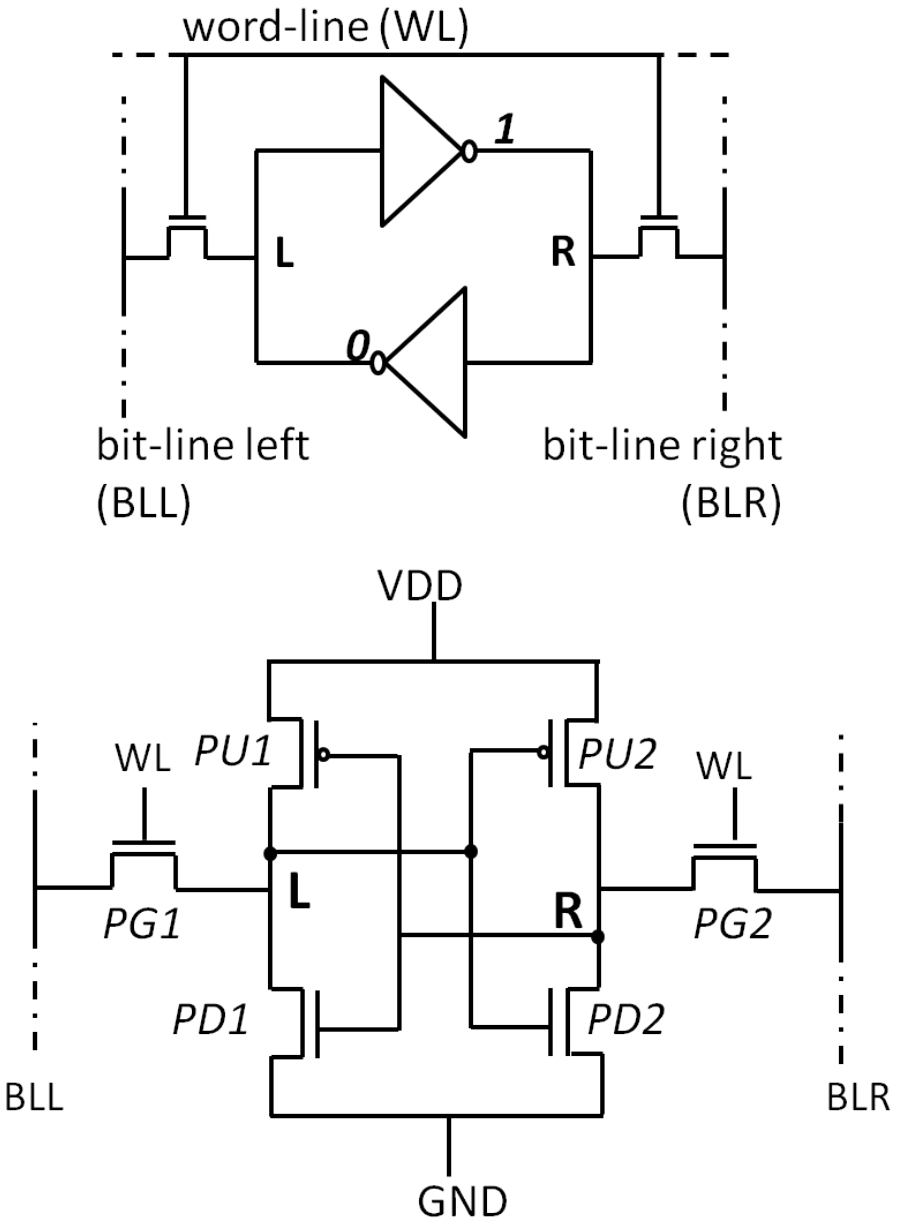

Figure 3.3: 6T-SRAM Bitcell schematics: Two cross-coupled inverters (top) for the storage of the data are connected to the external world via two access transistors (pass-gates) which are activated with a word-line signal. The transistor-level schematic (bottom) shows also the pull-up (PMOS), pull-down (NMOS) and pass-gate (NMOS) transistors.

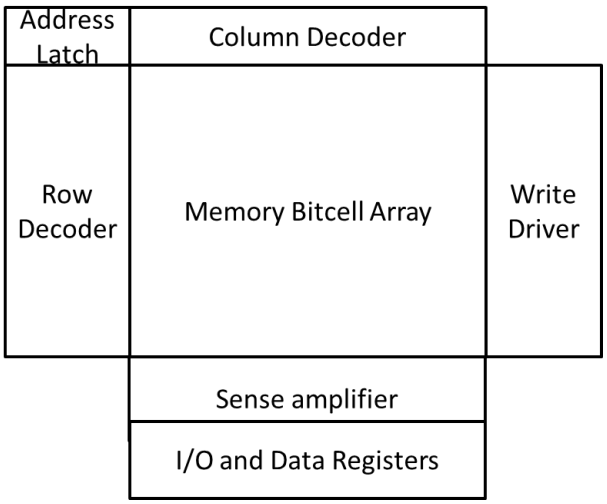

Figure 3.4: A block level of SRAM system. Column and row decoder circuitries are driven by the addressing latch which selects the required bitcells. Data Register and I/O bufffers are used to write to cells and are disabled during read cycles when the sense amplifier outputs the stored data. The clock circuitry and the word line driver which determines the word line pulse width, are not shown.

swing in the read circuitry, which causes a significant slowing-down in the read speed. Full-swing read is a popular method for ultra-low-voltage (ULV) SRAMs [80,81], since conventional sense-amplifiers can not operate in ULV. A novel ULV sense-amplifier design has been proposed in the literature to overcome this limitation [82].

3.3 Alternate Bitcell Architecture

Aside 6T bitcell architectures, to answer to the low-voltage requirements, new derived bitcell architectures as 8T [83], 10T [80] [84] or 11T [85] have been proposed in the literature. The new architectures mostly offer the separation of the write and the read operations improving the bitcell stability and thus increasing immunity to the statistical variability, which evidently allows the downscaling of the operating voltage. Per contra, the increased number of transistors in the bitcell enlarges the bitcell footprint and causes an area penalty. The high density requirements to increase the memory capacity raised the design of smaller (denser) bitcell architectures as 5T [86] and 4T [87]. The smaller bitcell architectures allow evidently a higher integration, but on the other hand, the reduced number of transistor degrades the bitcell stability. Variations in the electrical characteristics of the identically-designed transistors are also amplified by aggressive scaling

of the device sizes. Therefore small bitcell limits the downscaling of the operating voltage.

This research work is concentrated on the conventional 6T bitcell SRAMs, while a 10T bitcell is used in particular for ulta-low-voltage studies. The Ultra-Low Voltage (ULV) 10T bitcell architecture designed for ultra-wide voltage range (UWVR) applications [80, 81] is presented in figure 3.5. The write and reap operations are separated by the use of dual pass-gates to access bitcell internal nodes. The inner pass-gates, PG1 and PG2, are controlled by a second world-line signal WLC, whereas the outer read pass-gates, PG1R and PG2R are controlled by WL signal. During write operation, WCL and WL are asserted to open both the outer and inner pass-gates connecting the bitcell to BLs. The dual pass-gate implementation forms outer internal nodes, Lo and Ro, whose voltage levels are equal to the inner internal nodes, L and R. Footer gate transistors, FG1 and FG2, are connected to the outer internal nodes and their gates are controlled by the opposite inner internal node voltage. Considering a bitcell that stores a '0' in the node 'L', thus both Li and L voltage levels are equal to low-logic level voltage, the read operation is performed as follows: First the BLs are pre-charged to high-logic value as for a conventional 6-T SRAM read, then the outer pass-gates PG1R and PG2R are turned-on via WL signal establishing the connexion between outer internal nodes and BLs. PG1R and the footer gate FG1 then discharges the BLL through Lo, without disrupting L and thus increasing the read stability, since the inner internal nodes are protected from charge sharing as the inner pass-gates PG1 and PG2 controlled by WCL are closed.

3.4 A Simple Model for the Bitcell Variability Space

While the downscaling in the transistor sizes allows the designers to achieve considerable density improvements, it dramatically increases the in-die variability of the smallest MOS device used in SRAM bitcells. In the same time, since todays electronics devices feature millions of SRAM bitcells and chips might be manufactured in millions, the resulting large variability of the worst-case electrical parameters values increase the risk to obtain a significant amount of cells that are far from their nominal attribute. The overall impact of the in-die and within-die variability leads to memory failures and lowers the overall chip yield. The stability and performance analysis of worst-case SRAM cells has thus a very relevant technological impact, because it is so tied to the chip yield.

The real-life bitcells, the ones that are manufactured, can not be seen as identical to the nominal design due to the process variability. Each manufactured bitcell has its proper transistors with their proper electrical characteristics; some are close to, some are very far from the nominal design. In

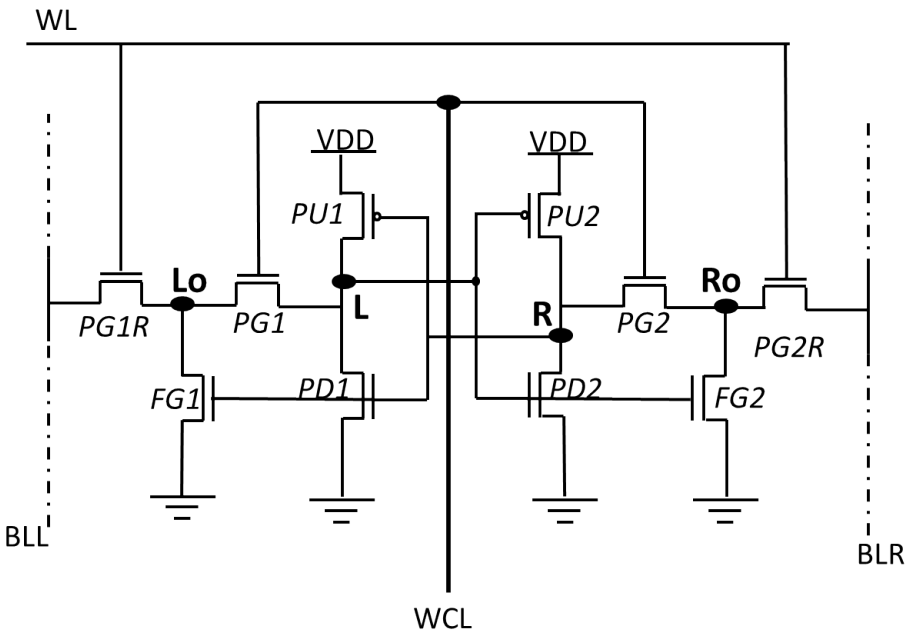

Figure 3.5: The Ultra-Low Voltage 10T bitcell architecture [80, 81].

real-life bitcells, the variability impact is emerged in many physical design parameter as the oxide thickness, the width and length of channel, the number of dopants, all affecting the electrical characteristics of manufactured devices. An ideal modeling of the variability in one transistor requires an accurate modeling of all different variability sources. At circuit-level, each transistor in the circuit has to be modeled, which increases significantly the complexity and the related computing cost, not just because of the increased number of variation parameter, but also due to the fact that MOS characteristics are heavily non-linear. The threshold voltage (V_{th}) plays a fundamental role in device electrical characterization and stands as the most significant parameter for transistor classifications. The overall variability impact thus can be modeled, as a first-order approximation, by variations in the device V_{th}. This approximation is a common approach which is also used in reliability analysis for modeling transistor aging.

Assuming that the overall variability impact in one transistor is modeled as variations in the device V_{th}, the collections of many manufactured bitcells, each having their proper transistors with shifted V_{th}, can be seen as multi-dimensional space, in which each dimension represents the amount of V_{Vth} variations in one transistor. In other words, manufacturing of a given SRAM bitcell can be modeled as a generation of many other bitcells, that are named henceforth as the "variants" of the nominal bitcell, where the nominal bitcell represents the bitcell having exactly the expected attributes as it is designed.

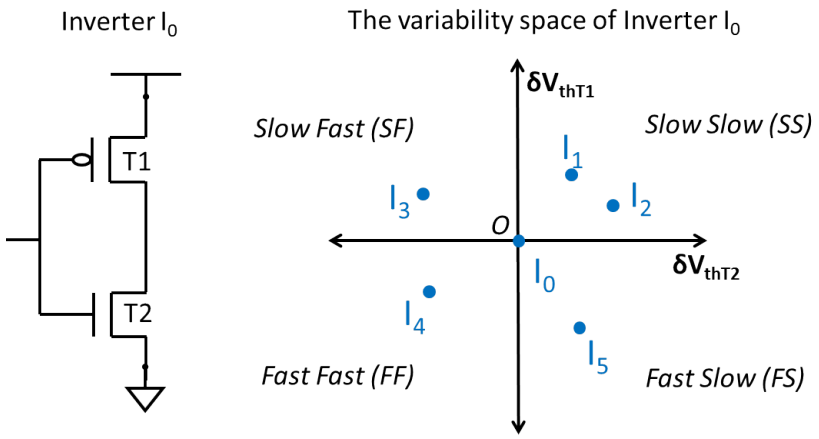

Figure 3.6: The 2-Dimension variabilty space of an inverter formed by T1 and T2. Variability impact on each transistor is quantified by a shift in the threshold voltage, δV_{thT1} and δV_{thT2}. A positive shift result in a slower device while a negative shift results in a faster device, giving 4 different combinations. I_0 represents the nominal inverter without the variability representing the origin O of the 2-Dimension space. I_1, I_2, I_3, I_4, I_5 represent different variants of I_0 appearing as a consequence of variability.

The space covering the nominal bitcell and its variants is called as the bitcell variability space and the number of dimension is equal to the number of device impacted by the variability. Figure 3.6 illustrates the variability space of an inverter formed by 2 transistors giving a 2-Dimension (2D) space. The horizontal and vertical axes of the 2D space represent the amount of the V_{th} shift for the transistors forming the inverter, PMOS T1 and NMOS T2 respectively. The shift amount is denoted by δV_{th}. A positive shift results in a slower (S) device while a negative shift results in a faster (F) device. Each quadrants of the 2D space represent therefore a different combination for NMOS and PMOS process, that are named as the process corners SS, SF, FF, FS. A particular inverter I_x, which is a variant of the nominal inverter I_0 placed at the origin O, is described by its V_{th} shift components, δV_{thT1} and δV_{thT2}. In Figure 3.6, I_1, I_2, I_3, I_4, I_5 illustrates particular inverters appeared as a consequence of the variability and they are all variants of I_0.

A 6T SRAM bitcell has therefore a 6-dimension (6D) variability space, which is impossible to illustrate. The origin of this 6D space represents the nominal bitcell as it is designed. It is evident that the increase in the number of dimensions of the variability space increases the analysis complexity and thus its related cost in terms of computing power and time.

3.5 SRAM Bitcell Failure Analysis

A SRAM bitcell have to be designed in a such way that it provides a non destructive read and a reliable write operation. The high vulnerability of SRAM cells to the variability is originated from this required balance between write and read operations. To perform an optimal write and read, the cross-coupled inverters have to be symmetrical and well-balanced, since the bitcell is accessed in the same way for both read and write operations (section 3.2). Considering that an SRAM design is manufactured on many dies, the global (inter-die) variability pushes the whole die towards the different regions of the variability space, where the SRAM bitcells attributes differ substantially from the nominal bitcell. These regions are illustrated in figure 3.6 with the 4 process corners SS, SF, FF, FS. The real-life bitcells that are also affected by the local (intra-die) random variability, might suffer more stronger mismatching between the cross-coupled inverters and the access transistors, which as a result cause one of the many possible SRAM failure. SRAM failures are classified as:

- Stability Failures: This occurs when the bitcell is supposed to keep the stored content, but instead the content is lost, which can happen due to the thermal noise or during the read operation if the inverters are not well-balanced.
- Write Failures: It occurs during a write operation when the bitcell is not able to toggle its content, i.e. is unable to perform a write into the internal node.
- Read-Access Failures: A failure occurs if the required voltage difference between bit-lines is not captured by the sense-amplifier during a read operation within the timing requirements.
- Leakage or Retention Failures: The leakage can cause the lost of the content while the cell is in retention mode, i.e. when the access transistor are "off". The cell leakage impacts also the static power consumption, which is a very important for low-power design.

Ideally, a SRAM design can be characterized on silicon giving a very precise image of the variability impact under its real-life operating conditions. However, the silicon test represents high time and money cost. Considering an SRAM array, in a best-case scenario, the overall needed time can be split into 3 months for design process, 3 months for manufacturing and then another month for the test. Therefore, in order to perform the very first silicon test of the design, the one has to wait for 7 month in best-case and then has to enter again into the same cycle for any optimization that has to be done in the design. This long process time is evidently unacceptable at the industry

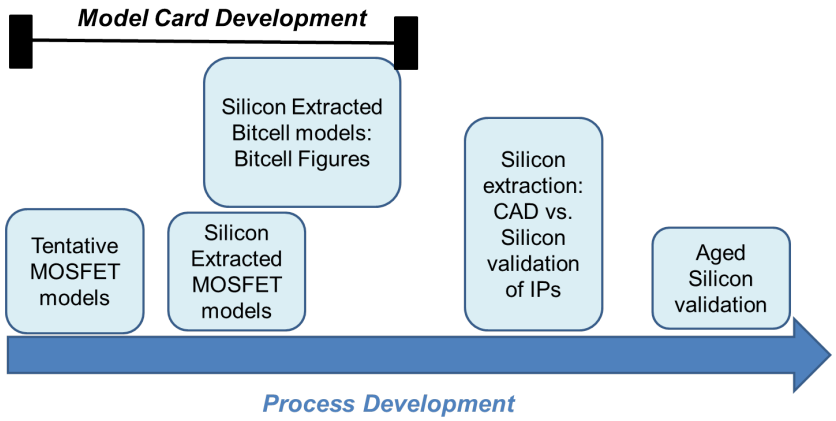

Figure 3.7: The different steps in the process developpement of a technology node in semi-conductor manufacturing industry. The SRAM SPICE model cards reaches the final maturity when the silicon extracted bitcell figures match with CAD results, both respecting the specifications for that given technology node.

level. A design has to be efficiently characterized in advance, before performing any silicon test. This is done through the SPICE-level characterization, which requires the use of accurate SPICE transistor model cards that has to model on-silicon behavior of the transistor with highest accuracy, for a wide range of supply voltage and temperature. Figure 3.7 illustrates the different steps of the process development of a new technology node in the semiconductor industry. The SPICE model card development starts as well with the process development with tentative MOSFET model cards. When the process reaches some low-but-sufficient maturity level, first silicon extracted MOSFET models are delivered. SRAM transistors have different electrical characteristics than the transistors used for other logic circuits, since they are manufactured with more aggressive design rules. Therefore, particular model cards are build for SRAM. In the process cycle, the delivery of SRAM model cards can be assumed to be around the same time as the MOSFET models. The one should note that the SPICE model cards reaches its maturity well before the process reaches its final maturity, giving to the designers the opportunity to perform SPICE characterizations independently of process maturity level. An accurate SPICE-level characterization methodology should take into consideration the large process variability and parasitic effects that exist on silicon. The accuracy of the characterization is directly related to the accuracy and to the efficiency of the used methodology which has to be validated on silicon. The SPICE-level design characterization represents therefore a very important step for the design optimization phase.

The SRAM memory design starts at a bitcell level and the efficiency of the bitcell design impacts directly the memory array design. In other words, the memory array characteristic can be estimated within a given accuracy by performing SPICE characterization of the bitcell that is used in the memory array. The propagation of the delay across the memory circuit, the impact of the peripheral circuitry and finally the large statistical variability limits the estimation accuracy but the estimation remains sufficiently acceptable if a solid methodology is used. A SRAM bitcell is commonly characterized by its stability, its read and leakage currents and its response to the timing requirements. The stability indicates the ability of the bitcell to be read from and to be written into safely. The read current indicates the read-ability of the bitcell representing the amount of current that flows from the bit-line into the internal node during a read operation generating the required bit-line voltage difference as described in section 3.5.2.1. The leakage current is a critical parameter impacting the bitcell stability in the retention mode, i.e. when the bitcell is not accessed, and its static power consumption [88]. The timing requirements are mostly evaluated using a critical path or the full memory array [89, 90], since they are meaningless at the bitcell level. The write time indicating the minimum required time to toggle internal nodes of bitcell during a write operation within a limited accuracy, since it does not take into consideration the delay in the array.

A SRAM bitcell can be characterized in two ways. The Direct Current (DC) analysis, or also called static, consists in using a quasi-static approach, in which solutions are found assuming a quasi-infinite access to the bitcell internal nodes. The Alternative Current (AC), or dynamic, analysis methodology involves the evolution of the WL pulse width, thus implies solutions that are found on the time domain. The transient response of the circuit evolves also with the parasitic capacitance and resistance of the bitcell layout. The next subsections presents static and dynamic analysis figure of merits. Although, silicon characterization equivalents for each figure of merit can be determined, this report mainly concentrates on SPICE characterization and the following figure of merits are presented for SPICE use.

3.5.1 SRAM Bitcell Static (DC) Analysis

Traditionally, the static analysis is the commonly used method for bitcell characterization, either electrical or simulation. It consists in measuring the design margins which quantify the ability of a particular bitcell to keep the stored data or to toggle its content under DC conditions. The bitcell nodes are supposed to be in one of these two stable states, 1 or 0, and then the margins are calculated with respect to them. The WL signal is always "on" and has a infinite duration so that the analysis is called "static"; the bitcell is always accessed and there is no time-dependency. The failures occurring

during a static analysis are defined as the static failures.

Two well-known figure of merits exist in literature for DC analysis: The Static Noise Margin (SNM) [91] for read stability evaluation, measuring the bitcell ability to keep its content while a read operation occurs and the Write-Margin [92, 93] for measuring the ability of the bitcell to toggle its content while a write operation occurs.

3.5.1.1 Static Noise Margin

The Static Noise Margin (SNM) is a standard measure of the bitcell stability, which is disturbed during a read operation. SNM measures the ability of the bitcell to keep the stored data against the read operation and it consists in extracting the voltage transfer characteristics of the cross-coupled inverters composing the bitcell. SNM is the minimum of two separately quantified design margins on the two inverters: SNML and SNMR for the left and right inverters respectively. SNML and SNMR are roughly correlated, a large SNML corresponding in general to a small SNMR and vice-versa. SNM simulation is performed with the bit-lines and the word-line held high. The SNM voltage represents the maximum voltage which can be present, during the read operation, on either one of the internal nodes without causing a content toggle.

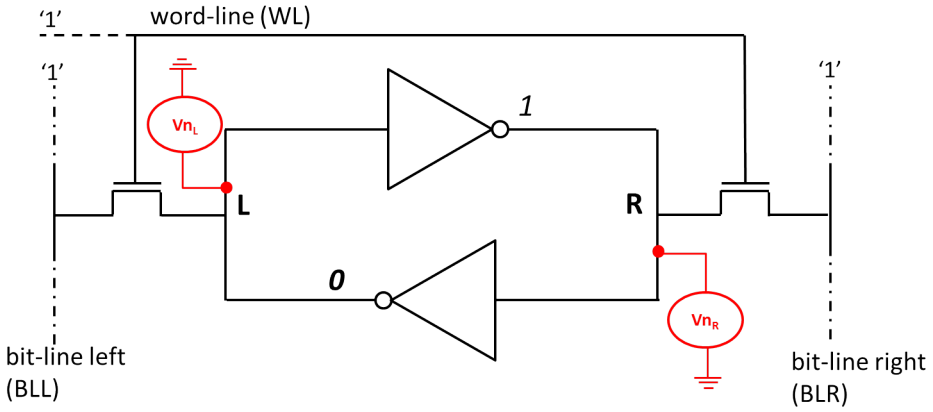

Figure 3.8: The 6T-bitcell schematic for SNM simulation. Two separate noise voltage sources Vn_L and Vn_R are connected to the bitcell internal nodes.

The SNM of a given bitcell is measured by connecting two voltage sources to each internal nodes, as shown in figure 3.8. The applied bias voltage V_n is then swept from 0V to the supply voltage Vdd and the voltage at the opposite internal node is measured. The same procedure is repeated for the opposite internal node. The Voltage Transfer Function (VTC) for each inverter is drawn in the same plot forming the "butterfly curve", as

illustrated in 3.9. Then the SNML and the SNMR are defined as the largest squares that can be fitted within each of the two loops of the "butterfly curve" and the final SNM of a given bitcell is the minimum of these two squares. There are three possible bitcell states: 1, 0 or the cell metastable point which is illustrated by the three crossover points in figure 3.9.a. Figure 3.9.a shows the butterfly curve for a nominal bitcell without the presence of the variability, in which the SNM squares within each loop are equal and the crossover point between the two curves has same x and y axis value, since the inverters are symmetric. Figure 3.9.b shows the butterfly curve for a bitcell under variability, in which the loops are unbalanced and the SNML is much larger than the SNMR. Figure 3.9.c shows the evaluation of the inverter VTCs under a large process variability.

3.5.1.2 Write Margin

The Write Margin (WM) test tells how hard is to write into the bitcell. This work uses Bit-Line Margin (BLM) [94] method for SPICE simulations, which is illustrated in figure 3.10 for a given bitcell storing '0' in node L. The BLR voltage is swept from VDD and 0 while the WL and the BLL are held "on", the WM is defined as the BLR voltage at which the internal node toggle occurs. As for SNM, the same procedure is repeated at the opposite internal node, holding BLR at '1' and sweeping BLL from VDD to 0. The final WM is the minimum of these two separately quantified WMs.

Other WM measurement methods exist in the literature: Write Static Noise Margin (WSNM) [95] method is based on drawing dc characteristics of the cross-coupled inverters in form of a butterfly curve as for the SNM test. World-line Margin (WLM) [96] consist in detecting the write step by applying a voltage sweep on Word-Line instead of bit-line and monitoring the internal node voltage. Combined World-Line Margin(CWLM) [97] consists in combining the bit-line and word-line voltage sweeps. In each method, the test is repeated for both internal nodes as in the SNM test and the final WM value of the bitcell is the minimum of the two separately measured WM. The BLM method is sufficiently accurate in SPICE, however the CWLM method is more reliable for silicon measurements. In CWLM, the final margin value is given by the WL signal, which is the gate signal of the access transistors. Therefore, the CWLM method gives a linear dependence between the WM and V_{th} of the access transistor, which is almost non-varying during measurements, allowing an efficient quantification of the effects of process variation and operating environment (temperature, power supply voltage).

3.5.2 SRAM Bitcell Dynamic Analysis

With the increase in the processor's speed, which goes upto few GHz in today's SoCs, SRAM operating frequency had to increase relative to the

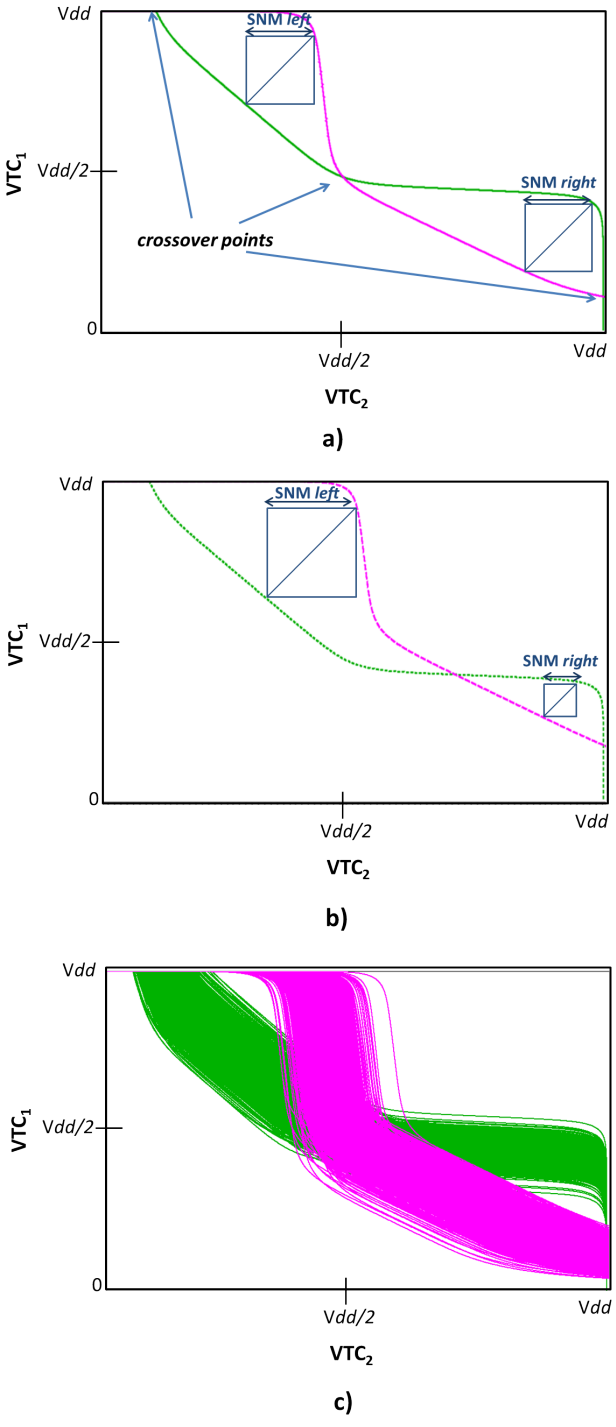

Figure 3.9: The butterfly curve illustrating VTCs of two cross-coupled inverters. a)VTC of nominal and balanced bitcell without statistical variability b)VTCs of a bitcell under statistical variability c)The spread of VTCs under a large statistical variability.

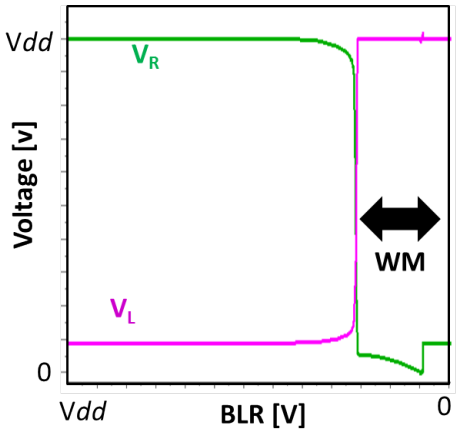

Figure 3.10: The WM SPICE simulation waveforms for a bitcell initially storing '0' in the node L. The BLR is swept from Vdd to 0V while BLL and WL are held at "1". The WM is the BLR voltage at which the internal node toggle occurs.

processor speed. High-Speed (HS) SRAM arrays have become requirements to offer good functionality at very short WL pulse. At high-frequency operations, a static test might not be enough accurate, since the mechanisms behind the static and dynamic failures and the required design optimizations are different. The dynamic SRAM Bitcell analysis is performed with a finite duration WL pulse, or with successive finite WL pulses. The mechanisms behind the bitcell failures may therefore evolve with respect to the WL duration and the retention time between successive pulses. The failures occurring under dynamic conditions are defined as the dynamic failures.

Static analysis of an SRAM bitcell features only two stable states, the first with an internal node being close to Vdd and the other node close to ground and the second vice versa. Any bitcell at rest is supposed to be in one of these two states. At high-frequency, successive operations might take place [98] on the same bitcell and the internal nodes may not have the required time between successive operations to reach their stable states. Therefore the bitcell cannot any more be supposed to start its operations at rest. The finite pulse duration of the Word-Line (WL) limits the time available to access a bitcell and thus makes it harder to write, but also harder to toggle it during reading, with respect to the static case. The bitcell parasitic capacitance and resistance, which are a function of the bitcell layout and of the column and row architectures, also slow down any changes in the bitcell content. The dynamic stability depends thus on the device variability, on the operating frequency, on the bitcell layout, on the memory row and column architectures, and, finally, on the access history of each bitcell. Therefore, new figure of merits have to be introduced to include

not only the WL pulse length dependency, but also parasitic capacitances impacting bitcell performance and speed.

3.5.2.1 Read-Ability

In contrast to the static analysis, in dynamic analysis, the stability of a given bitcell is not the main concern during a read operation. The content toggle during the read operation is less critical if the WL pulse is relatively short, so that the time during which the internal nodes are accessed is not sufficiently long enough to cause a non-desired toggle. However, this short time might cause a non-successful read of the bitcell content, if the required voltage difference between bit-lines is not reached. In this work, it is assumed that the read operation is performed through sensing and the required voltage difference is henceforth called as the SA_{offset}, indicating the minimum required voltage difference to activate SA. Figure 3.11 illustrates the RA test in which the figure of merit is defined as follows: The BLs voltage difference ΔV_{BL}=V_{BLR}-V_{BLL} is extracted at the end of WL pulse. A RA failure is detected if ΔV_{BL} is less than SA_{offset}. It has to be noted that if for some reason a read-stability error occurs during the RA test, ΔV_{BL} at the end of the WL pulse would be negative, thus less than SA_{offset}, and the stability error can be captured. In a read operation, the discharge of a BL is slowed down due to the parasitic capacitance of other bitcells in the column which share thus the same BLs.

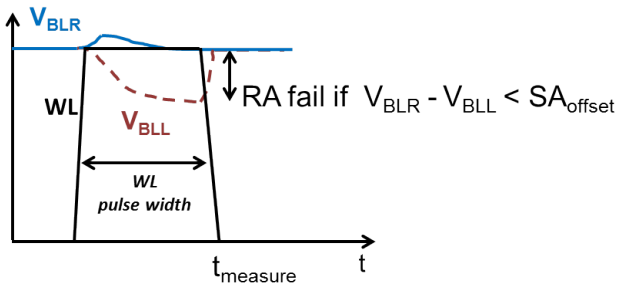

Figure 3.11: The Read-Ability (RA) extraction example for a bitcell initially storing '0' in the node L. The BLL voltage, V_{BLL}, is discharged during the read. For a successful read, the voltage difference between BLL and BLR has to be equal or larger than the SA_{offset} at the end of the WL pulse.

3.5.2.2 Write-Ability

The failures during the write operation are considered as minor concern in the static analysis, since the infinitely long WL pulse helps writing into the bitcell. This is not always true in dynamic analysis, since the WL pulse

width is inversely proportional to the operating frequency. Furthermore, for a successful write operation, the internal nodes toggle during the WL pulse may not sufficient. After the toggle, a sufficiently enough voltage difference between the internal nodes has also to be reached to ensure that the internal nodes state remain stable state and a non desired toggle would not occur before or during the upcoming operation. Figure 3.12 illustrates a successful write operation in a bitcell initially storing '1' in the node L. Figure 3.13 illustrates the internal node voltage waveforms of the another bitcell occurred as a consequence of variability. The internal nodes L and R voltages, V_L and V_R respectively, do not reach low and high logic level voltages before the $t_{measure}$ leading later to a non-desired toggle. The figure of merit for the WA test is defined as follows: The internal node voltage difference ΔV_N=V_R-V_L is extracted at $t_{measure}$. The value of $t_{measure}$ depends on the timing requirements. A WA failure occurs if ΔV_N is less than a given percentage p of V_{dd}. The value of p is assumed to be between 50-80% yielding in a parameter-dependent figure of merit. The value of p and its impact on the failure estimation is investigated later in this report in section 4.4. A write failure that might be caused by the non-toggle of the bitcell internal nodes within the WL pulse would be also captured, since in that case ΔV_N will be negative, thus less than p*V_{dd}.

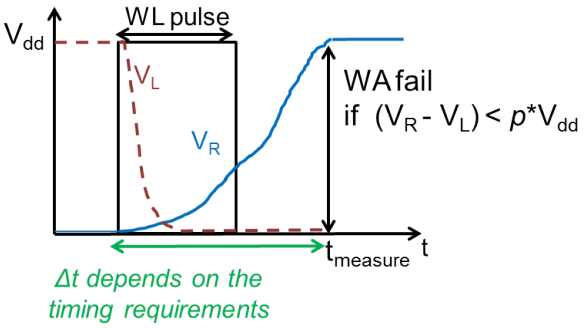

Figure 3.12: The Write-Ability (WA) test for a bitcell initially storing '1' in the node L. The voltage difference between internal nodes has to be equal or larger than the threshold value p*V_{dd} at the end of the write operation. The time at which the test is performed, $t_{measure}$ depends on the timing requirements, for exemple, in a aggressive high-frequency test, $t_{measure}$ is equal to the twice WL pulse width.

3.5.2.3 Multiple-Pulse Analysis and Figure of Merits

In real-world SRAM operations, a bitcell usually might go through a series of operations. Some particular combinations may actually introduce failures that would not be captured if only single WL pulse was used. Con-

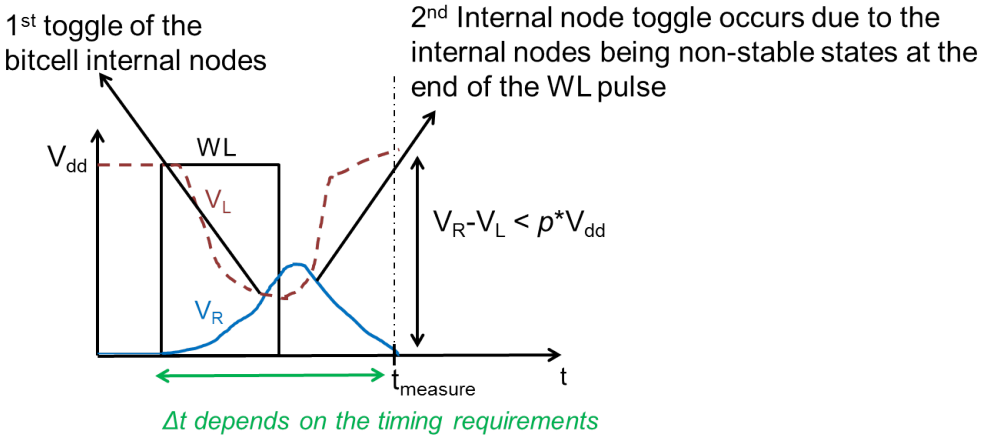

Figure 3.13: The Write-Ability (WA) failure for a bitcell initially storing '1' in the node L. The voltage difference between internal nodes at $t_{measure}$ is not large enough at the end of the WL pulse and the content is lost.

sidering high-frequency memory operating conditions, a given bitcell may be accessed for successive reads. This multiple access in a short time space can affect the bitcell internal nodes stability. A bitcell that is stable enough to withstand one read operation, might actually fail after two or more successive read operations [98] [79]. Figure 3.14 illustrates the internal node voltages waveform of a given bitcell during multiple read access, in which a read stability failure occurs after the second consecutive read. The Read-After-Read (RaR) figure of merit consists in performing an RA test after each new WL pulse. The RaR failures in a given bitcell might evolve with the number of successive WL pulses and with their period.

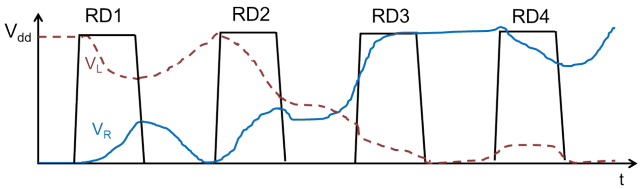

Figure 3.14: The bitcell internal nodes voltage waveforms leading to read-stability failure during multiple read access. A first read (RD1) is successfully performed, but the non stable states of the internal nodes at the beginning of the second read (RD2) lead the bitcell to a stability failure (internal nodes toggle).

During consecutive access to SRAM bitcell, a read operation may follow a write operation. Considering a bitcell that struggles in having stable internal

node states after a write operation, an upcoming read operation may either help or hinder the completion of write depending on the mismatch present in the bitcell. These two possible scenarios are illustrated in figure 3.15 as obtained from two different Monte Carlo simulations runs. In both scenarios, the first WL pulse is a write 0 into the node 'L'. In figure 3.15.a, the node R voltage did not reach Vdd within the WL period, thus the state of R is not stable at the arrival of RD WL Pulse. The pre-charged bit-lines lead to a current flow from BLR to R helping the node R to be pulled-up to Vdd. In figure 3.15.b, the internal nodes state is far from being stable and the upcoming read, thus the current flow into the bitcell internal nodes, causes a read stability failure. The read-after-write (RaW) figure of merit consist in performing a read-ability test after the read operation. If a read-ability failure is detected, the failure may be caused by the write as well as from the read operation.

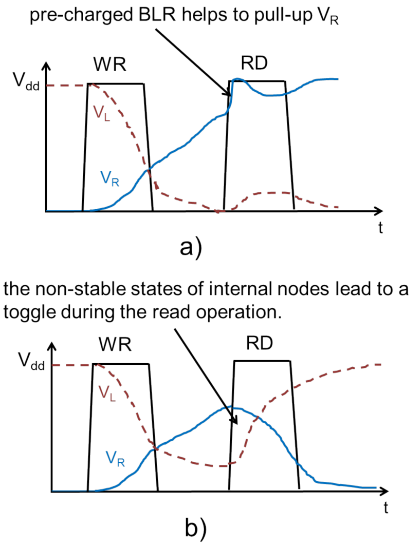

Figure 3.15: The bitcell internal nodes voltage waveform during read after write operations. a) The read helps to the completion of the node R thus no fail occurs. b) The non stable-states of the intenal nodes lead to read stability failure during the read.

If a write operation takes place after a read operation, the initial condition of the bitcell cannot be assumed to be one of its stable states and in general this helps the write operation [98] and is therefore generally discarded from the dynamic analysis.

3.6 6T-SRAM Bitcell Failure Mechanisms

A nominal bitcell must be designed in such a way that it is immune to static and dynamic failures to withstand the effect of variability. The simplified variability space model can be used to study efficiently different bitcell failures, since bitcells failing under a given criterion can be represented by a particular zone in the variability space. Figure 3.16 illustrates the concept of failure zone in the variability space, for the inverter of figure 3.6, in that one adds a figure of merit deciding about the failure criteria. The failures zones represents the zones of the variability space in which the amount of V_{th} variations in T1 and T2 cause the inverter to fail. Considering a switching error, i.e. the inverter is not able to invert its input, a failure occur if the balance in the NMOS-PMOS pair is broken down. In other words, failures might occur when the V_{th} variations result in fast PMOS-slow NMOS or slow NMOS-fast PMOS couples. These zones are illustrated by F_0 and F_1, respectively. The real-life inverters are modeled using the simple variability space model through the generation of the inverters I_1, I_2, I_3, I_4, who are all variants of the nominal inverter I_0. In this example, the inverters I_1 and I_4 occur in the failure zones F_0 and F_1, respectively, whereas I_2 and I_3 occur in the safe zone of the variability space, and do not fail.

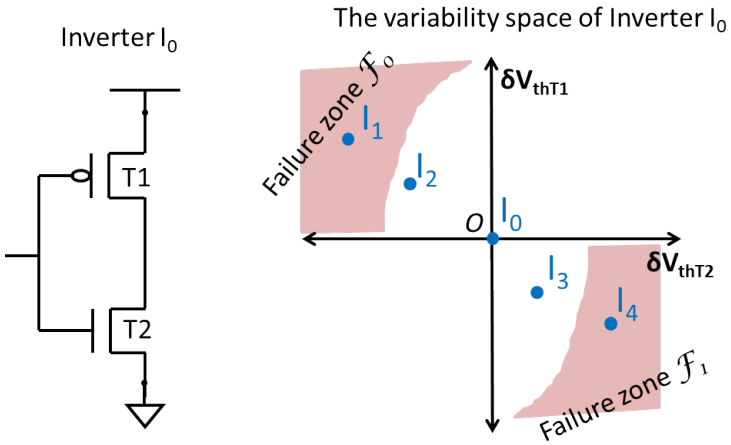

Figure 3.16: The 2-Dimension variabilty space of an inverter formed by T1 and T2, with two failure zones F_0 and F_1 caused by the mismatch beween T0 and T1. I_0 represents the nominal inverter without the variability. I_1 and I_4 represent the variants of I_0 occuring in the failure zones of the variability space, whereas I_2 and I_3 occur close to the failur zones but they are still in the safe zone of the variability space.

For a 6T SRAM bitcell and its 6D variability space, the number of failure zones is related to the design choices and to the operating conditions.

The mismatch combination yielding to a particular bitcell failing under a given criterion is tied in this work to some *failure mechanism*. SRAM bitcell failure mechanisms are largely dependent on the WL pulse width, since the latter determines the amount of the time during which the bitcell internal nodes remain connected to the bit-lines; this connexion generating a current flow that impacts internal nodes. Mainly, the failure mechanism under the long World-Line pulse are associated to static failures and the failure mechanisms under a short world-line pulse memory operations are associated with dynamic failures.

The read-related failure mechanisms are strongly dependent on the WL pulse width. Considering a long WL pulse, the bitcell internal nodes remain connected to the bit-lines during significantly long time and this represents a worst-case scenario for read stability failures, since the bit-lines are precharged to high-logic level before the read operations and the large access time to the internal nodes may cause an unwanted amount of BL discharge. If this amount of voltage that is discharged from BL is sufficiently high to toggle internal nodes, a read-stability failure occurs. The read-stability failures are triggered by the mismatch combination associated to fast (strong) NMOS and slow (weak) PMOS pair. Considering the particular bitcell illustrated in figure 3.17, the presence of variability resulting in a fast NMOS and slow PMOS devices, the read stability failure mechanism is as follows: The fast PG1 discharges the bit-line leading to a voltage increase in the node L. This voltage increases strengthens the initially-fast NMOS PD2 drivability and weakens the initially-slow PMOS PU2, leading the bitcell to struggle to keep high-logic level in the node R, and, if the node L voltage is sufficiently high, the cross-coupled inverter system toggles causing a loss of the stored bit. The current flow in the bitcell causing a read stability failure is illustrated in figure 3.17 via blue arrows. The arrows sizes are proportional to the amount of the current flows across the device. Overall, a bitcell with a fast-NMOS and slow-PMOS pair due to the variability is highly vulnerable to the read stability failures, in particular under long world-line pulse. Under a short WL pulse, the read-stability of a given bitcell is not anymore the first concern, since there is much less time available to disturb internal node states [99] [100]. On the other hand, this short WL duration limits the time for the bit-lines to reach the required voltage difference to complete the read operation, causing a read-ability failure. Considering again the bitcell of figure 3.17, a read-ability failure occurs if the PG1 (the device on duty of discharging the BLL) is too slow, therefore the voltage difference between BLL and BLR do not reach the required level within the WL pulse. The current flow of this failure mechanism is also illustrated in figure 3.17 via the orange arrow.

A write failure in 6T SRAM bitcell occurs if the bitcell content can not be successfully flipped within the timing requirements. This can be caused by a non-toggle of the internal nodes within the WL pulse and the under-

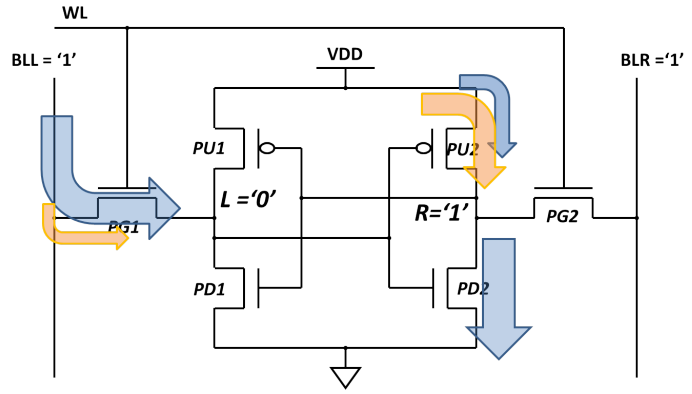

Figure 3.17: The current flow in a bitcell during a read operation causing a stability (blue arrows) and read-ability (orange arrows) failures. Arrows sizes are proportional to the amount of current flows across the device. reached.

lying mechanism is called as *write discharge failure*. Considering the bitcell illustrated in figure 3.18, a discharge failure may occur if the bitcell has a fast PU2 and slow PG2 pair. The underlying mechanism is as follows: The slow PG2 struggles to discharge the high-logic level of node R to the BLR, whereas the fast PU2 keeps "pulling-up" the node R to VDD, preventing the internal node toggle. The current flow in a bitcell leading to a write discharge failure is illustrated in figure 3.18 via blue arrows. Discharge failures are directly related to the strength ratio between pull-up and pass-gate devices.

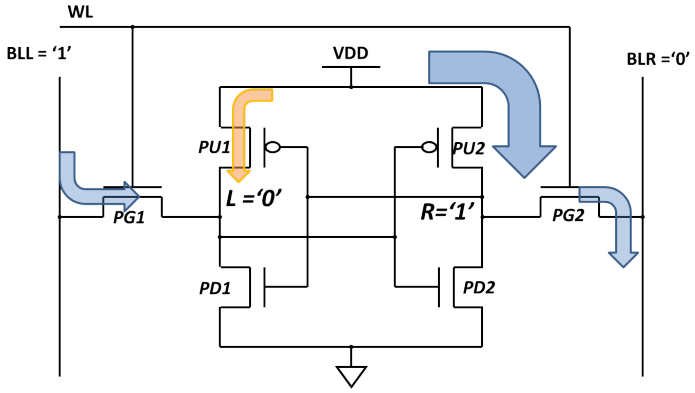

Figure 3.18: The current flow during a write operation causing a discharge failure (blue arrows) and completion failure (orange arrows). Arrows sizes are proportional to the amount of current flowing across the device.

A short world-line pulse raises the write complexity, since the internal

nodes state evolves with respect to the WL pulse width and to the time left between two successive memory operations. Therefore the write operation is highly concerned by the WL pulse width and the period. As already mentioned in section 3.5.2.2, the concern is not only for the non-toggling of the bitcell content, but also for not having the internal nodes at their stable states within the required time space, even if the toggle occurs within the WL pulse. This failure mechanism is generated by a lack of sufficient voltage difference between the internal nodes at the end of the write operation. This lack of voltage leaves the bitcell with non-stable internal node which presents a high risk for content lost during the upcoming WL pulses. The failure is caused by a slow PMOS device on the side of the internal node initially storing '0' and called as *completion* failure. Considering the bitcell illustrated in figure 3.18 initially storing '0' in the node L, a *write completion failure* occurs after the internal nodes toggle if the PU1 struggles to "pull-up" the node L to the VDD, leaving the bitcell internal nodes in a non-stable state at the end of the write operation which may lead the internal nodes towards a second toggle before, or during, the upcoming WL pulse causing the lost of the previously written content. The current flow that might lead to a completion failure is illustrated in figure 3.18 via orange arrows.

The mechanisms behind the write discharge and read stability failures are opposite in terms of mismatch between NMOS and PMOS. A particular bitcell being vulnerable to read-stability failures due to the presence of variability, has a high immunity to write discharge failures and vice-versa. For long WL pulse memory operations, considering a bitcell design that is well-balanced for DC conditions, the write failures are less critical then the read stability failures, since the long "access time" to the internal nodes helps the write operation, therefore, most of the bitcell designs are read-optimized. Write-Assist techniques are also commonly used to improve the write-ability of the bitcell [101], allowing for larger read margins in the bitcell design.

The leakage failures occur when the WL signal is off, i.e. the bitcell is in a retention mode and there is no access to the internal nodes. The immunity of a bitcell to leakage failures is degraded with longer "retention" mode. The leakage failures are not seen as a first-order impact of the statistical variability, since they are not triggered by mismatch, but by the systematic variability. The leakage current in a given transistor increases with the decrease of V_{th}, therefore an SRAM having fast-NMOS and fast-PMOS devices represent higher concern against leakage failures. The increase in the temperature is another factor that increases the leakage current.

As mentioned in 3.5.2.3, successive WL pulses can occur during memory operations and this can lead the bitcell to failures, even if the first operation had ended successfully. The failure mechanism analysis in successive operations is more complex compared to the single-WL pulse analysis, since failures depend strongly on the internal nodes state before and after each

WL pulse. The evolution of the internal nodes state depends on the type of the operation that succeeds the previous operation, the time left between these successive operations, and, finally, the column and row architecture, thus the capacitive coupling that affects the bitcell. The failure mechanisms, which would be one of the mechanisms that are discussed for single-WL pulse analysis, might also change with respect to the number of successive operations.

3.7 The Minimum Operating Voltage V_{min}

On the top of all characteristics, there lies the minimum operating voltage V_{min} indicating the minimum possible supply voltage under which the memory array remains functional, which means that there is no content lost with a sufficiently high read current during a read operation and the bitcell is able to toggle its internal nodes during the write operation. The particularity in the V_{min} evaluation is that it can be calculated with respect to different dynamic or static figure of merits. Moreover, the large amount of bitcells in an SRAM array yields in a large V_{min} distribution. The spread of the V_{min} variation across a memory array is directly related to the array size and the resulting distribution wander away from the known distribution laws for larger memory sizes. Therefore, the V_{min} is strongly dependent on the design specifications as the memory cut size, the maximum and minimum operating temperatures, the operating frequency etc.

Since the bitcell V_{min} is direct contributor of the memory array power consumption and the SRAM is biggest contributor for the overall chip power consumption, a huge interest has been shown for lowering V_{min}. A way for lowering V_{min} is to design new bitcell architectures using an increased number of transistors as it is already discussed in section 3.2, which enlarge on the other hand the bitcell footprint. Adding read and write assist circuitry [100] is another efficient way for V_{min} improvement, but again in trade-off with area penalty. In advanced technology nodes, the body biasing techniques are also used for adjusting the balance between the read and write operations via adjusting NMOS and PMOS transistors strength ratio. Although the body-biasing benefits in conventional bulk technologies are limited by device physics, it is very efficient for design-level V_{min} optimization in FDSOI technology thanks to the use of buried oxide under the thin silicon film allowing the device body to act as a second gate and to change transitor flavor.

The variability in an SRAM array including more than billions of transistors leads to a very large statistical distributions of SRAM failures. [102]. Therefore the V_{min} of that given array, which represents the V_{min} of a particular bitcell in that array having the worse immunity against bitcell failures, is impacted dramatically by the large variability. The V_{min} mea-

surement is by nature possible only with a mature manufacturing process, since it requires measurements on a very large amount of dies to have a statistical validity. This is possible only if the manufacturing yield is high enough to allow functionality of a memory cut containing millions of bitcells on a statistical set of dies. However, the design starts well in advance with respect to the time in which manufacturing reaches such a maturity. An accurate estimate of the V_{min} using SPICE models plays a very critical role, since it allows designers to discover the weakness and the limits of their designs at CAD level. Such an anticipated design optimization process helps to cut overall production cost, allowing the designers to optimize and enhance their designs before being able to perform silicon tests, i.e. saving time and money.

The V_{min} modeling and the accuracy of modeling methodology have become therefore key factors during the design optimization. For large memory arrays, the V_{min} modeling consists in modeling of very rare events which occur with very small probabilities ($< 10^{-9}$) and thus requires the use of advanced statistical analysis. A sufficiently high modeling accuracy is crucial for efficient design optimization, since the V_{min} represents the minimum safe supply voltage value below which the memory array lose its functionality. The next chapter presents investigations on the variability-related V_{min} limitations of SRAM bitcells. An improvement of the existing modeling methodology for static V_{min} as well as a novel modeling methodology for more complex dynamic V_{min} are both presented.

Chapter 4

SRAM Bitcell Variability Space Modeling for V_{min} Estimation

This chapter presents two different SPICE simulation methodologies for the modeling of the SRAM bitcell minimum operating voltage V_{min}. First, an existing SRAM static design margin modeling methodology based on Monte Carlo simulations is presented. The methodology is improved in order to increase the modeling accuracy at distribution tail. Secondly, a smart algorithm that is based on hypersherical surfaces and simplified bitcell variability space analysis is presented. The proposed methodology is first tested for SRAM bitcells operating under transient conditions. The key advantage is stated as the efficiency in extracting the underlying mismatch mechanisms behind failures that limit SRAM bitcell V_{min} and their dependencies on the operating conditions. Silicon measurements are used to validate both modeling methodologies at different technology nodes. Finally, the knowledge acquired about V_{min}-limiting mismatch mechanisms is used for optimization of SRAM bitcell usage during the 28nm UTBB FD-SOI technology development.

4.1 Bitcell Variability Space Modeling using Monte Carlo SPICE simulations

For an efficient SRAM design optimization at SPICE-level, as a ultimate condition, the bitcell variability space has to be sufficiently accurate such that it gives a realistic image of the real-life manufactured SRAMs. In the same time, this variability space has to be reproducible with a feasible computing and time cost. The term "design optimization" can be therefore used, only if these two conditions are satisfied. This is done by adding the

systematic variability into the SPICE transistor model cards through the use of "process corners", which describes the models of 5 particular points of the process variability space; 4 process corners Slow-Slow (SS), Fast-Fast (FF), Slow-Fast (SF) and Fast-Slow (FS) as already mentioned in section 3.4, plus the typical (TT) process. At each process corner, the random variability is modeled by normally distributing the device threshold voltage (V_{th}), since the overall impact of the different random variability sources results in a Gaussian-like V_{th} distribution [103], within a certain accuracy. The standard deviation of the V_{th} distribution, denoted as σ_{Vth}, read as:

$$\sigma_{Vth} = \frac{A_{Vt}}{\sqrt{LW}}$$

where L and W are the channel length and width respectively, A_{Vt} is a technology dependent mismatch constant that varies roughly linearly with the technology node length. The variability in Vth increases therefore with the device scaling at each new technology node.

The Monte-Carlo simulation method [104] is a widely used technique for the modeling and the analysis of correlated and uncorrelated variations in analog and digital integrated circuits. The Monte-Carlo method is known as a broad class of computational algorithms that rely on repeated random sampling to obtain numerical estimation of the occurrence probability of a given phenomenon; typically running so-called randomized simulations many times over allowing one to obtain the unknown distribution of an observable. The algorithms are often used in physical and mathematical problems and are most useful when it is difficult or infeasible to solve a problem analytically, or to solve it with a deterministic algorithm. In an ideal world, an accurate SRAM bitcell analysis would require the same number of Monte Carlo runs as the number of bitcells in the SRAM array, and then iterating the same several times to average out the results reducing the statistical error which decreases with the total number of total runs.

A Monte Carlo estimation $\hat{X}$ of the random variable X, is represented by a confidence interval Z and an error percentage ϵ. It read as

$$P(X = \hat{X}) = X \in \left[\hat{X} - \epsilon\hat{X}, \hat{X} + \epsilon\hat{X}\right] = Z$$

The relation between the number of needed iterations, N_{MC}, and the targeted occurrence probability, p, with the desired confidence interval Z and the error percentage ϵ is determined as:

$$N_{MC} = \left(\frac{Z_\lambda}{\epsilon}\right)^2 \frac{(1-p)}{p} \tag{4.1}$$

where Z_λ represents the equivalent standard deviation allowing to cover $\lambda\%$ of estimates around the estimated value assuming that the estimation itself follows a normal distribution; the normal distribution law is described

in detail later in this chapter. The confidence interval is usually set as Z=95%, which gives a Z_λ equal to ±1.96 normalized standard deviation. In other words, 95% of confidence interval with 10% of error percentage for the estimate of a random variable X means the following :

$$P(X = \hat{X}) = X \in \left[\hat{X} - 0.1\hat{X}, \hat{X} + 0.1\hat{X}\right] = 0.95$$

Therefore using equation (4.1) and considering a targeted p=1.10^{-5}, with 95% confidence interval and 10% error percentage, approximately 38.10^6 iterations have to be run. This very large number of iterations is not feasible on an industrial basis due to the time cost. On the other hand, a limited number of Monte Carlo runs is enough to extract fist two moment: the mean μ and the standard deviation σ of the unknown distribution. Theses informations are sufficient to estimate tail values, i.e. low probability occurrence, assuming that the unknown distribution fits well a normal distribution. Figure 4.1 illustrates the probability density function of the normal distribution with μ=0 and σ=1, which is known as the "bell-shaped curve". The curve is centered on the mean value μ, which is also the value with the highest occurrence probability. The left and right tails are symmetric and the area under the bell curve limited by a given number of sigmas on both sides of μ is the probability that the random variables assume a value lying in that interval. For example, the 68.2% of values around the mean in a given normally distributed set are lying between $\pm 1\sigma$. The ratio between the area under bell curve limited by $\pm n^*\sigma$ and the overall area under the bell curve is given by

$$erf\left(\frac{n}{\sqrt{2}}\right) \tag{4.2}$$

where *erf* is the error function [105].

A designer may need to estimate the worst-case value of a given bitcell metric under a given process corner, voltage and temperature (PVT) conditions in order to check the design compliance with its specifications. If the metric under test is the bitcell read current, denoted as I_{cell}, figure 4.2 illustrates the I_{cell} probability density function (PDF) of the 40nm Single-Port High-Density bitcell obtained with 4096 Monte Carlo simulations at typical corners and 27 °C, using ST Microelectronics SPICE model cards. The red straight line represents the normal curve drawn using the mean and the standard deviation of the simulation results, showing a reasonable fit for the read current variability under this conditions. The mean of the I_{cell} distribution is simulated as 15.6μA and the variability results in worst-case tail values (leftmost tail) less than 10μA, or a value that is $\approx 3\sigma$ far from the mean. This confirms that the distribution is nearly normal. Considering a 4Kb of SRAM cut and a normal-like distribution for a given figure of merit,

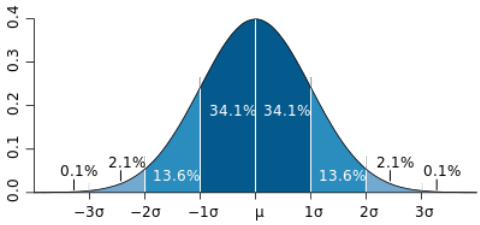

Figure 4.1: Probability Density Function of a normally distributed values with mean 0 and standard deviation 1: The dark blue area is less than one standard deviation away from the mean and accounts for 68.2% of the values, while two standard deviations from the mean (medium and dark blue) account for 95.4%, and three standard deviations (light, medium, and dark blue) account for 99.7%.

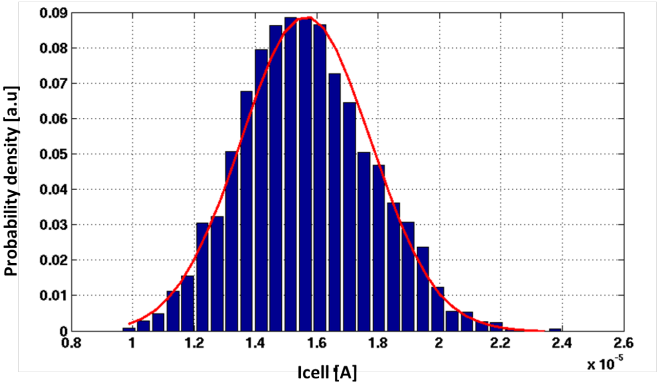

Figure 4.2: Probability Density Function (blue bars) of the Single Port High Density Bitcell Read Current in 40nm node obtained with 4096 Monte Carlo runs performed using typical SPICE process corners at 27 °C. The straight red line represents the normal distribution curve drawn using the mean and standard deviation of the simulation results.

the worse case bitcells that occur at left and right symmetric tails of the distribution have both an occurrence probability of

$$\frac{2}{4000} = \frac{1}{2000}.$$

One of these bitcells occurs in fact at the best-case tail so that it does not represent any concern. The distance between the mean and the other bitcell that is at the worse-case tail can be therefore calculated from equation (4.2) as:

$$n = \sqrt{2} erf^{-1}(1 - \frac{1}{2000}) = 3.49\sigma.$$

The table 4.1 presents the equivalent number of sigma (n) distance from the mean that is needed to reach worse-case tail values for a given SRAM cut size assuming a normal-like figure of merit distribution.

cut size	$erf\left(\frac{n}{\sqrt{2}}\right)$	n
4Kb	0.99951172	3.49
8Kb	0.99975586	3.67
16Kb	0.99987793	3.84
32Kb	0.99993896	4.01
64Kb	0.99996948	4.17
128Kb	0.99998474	4.32
256Kb	0.99999237	4.48
512Kb	0.99999619	4.62
1Mb	0.99999809	4.76
2Mb	0.99999905	4.90
4Mb	0.99999952	5.04
8Mb	0.99999976	5.17
16Mb	0.99999988	5.29

Table 4.1: The number of sigma vs. covered area under the bell-shaped normal distribution curve.

However, in advanced technology nodes, attempting to fit the distribution metric with a basic normal fit may lead to a significant estimation error, especially for large memory arrays containing billions of transistors. Although it is a common approach to assume that SRAM bitcell performance and stability metrics follow normal distribution law, the large variability impact may results in particular distributions that follow the normal law up to few σ, but the occurrence probability of the tail values may deviate from the normal law. This fact represents a very high risk in SRAM analysis, since the bitcells occurring at the distribution tail are the ones who limit the good functionality of an SRAM array and they therefore have to be known with an admissible accuracy during the design optimization.

As previously mentioned in section 3.5.1, the Static Noise Margin (SNM) is one of the decisive metric indicating the bitcell stability during a read operation, and described as the minimum of two separately quantified margins performed for both inverters forming the SRAM bitcell. Figure 4.3 presents PDF of SNM measurements that has been performed for the process monitoring database of Single-Port High-Density (SPHD) bitcell in 45nm technology node. 10^5 dies has been measured across multiple wafers and lots giving a very realistic image of the variability impact at industrial level manufacturing. The first two distributions show the measurements of the separately measured SNML and SNMR, and the third one illustrates SNM,

which is the the minimum of SNML and SNRM in measurements. SNML and SNMR distribution follow normal law and they are roughly correlated, but the SNM distribution has a longer leftmost tail, and a smaller mean and standard deviation. The cross symbols represent the minimum and maximum SNM results obtained with 10^5 Monte Carlo simulations showing the good accuracy of ST Microelectronics SPICE model cards. The results obtained with a Fast-Monte Carlo extension [106] plugged on the same simulator are also shown (triangles) confirming the good accuracy of model cards.

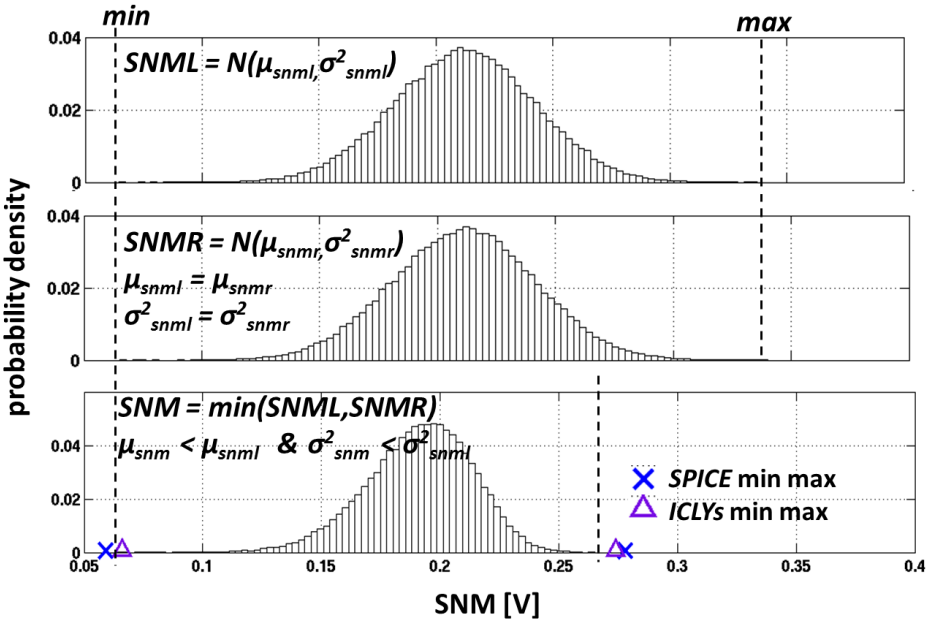

Figure 4.3: Probability Density Function (bars) of SNML, SNMR and SNM margins, as obtained from approximately 10^5 measurements at nominal Vdd and ambient temperature on Single-Port, High-Density bitcell manufactured in C45. Minimum and maximum SNM results for 10^5 Monte Carlo simulations are also presented, for a standard simulator (crosses) and fast MC extension (triangles).

Figure 4.4 presents the same PDF of SNM measurements, but in a log-scale. The die distribution follows the normal law only until approximately -4σ, while farther samples have a much higher probability than the extrapolated values. A larger standard deviation would be needed to fit the distribution tail. Figure 4.4 presents an additional normal curve (large-sigma line), that has mean μ_{SNM} and standard deviation

$$\sigma_{LS} = \frac{\left(\sigma_{SNM} + \sigma_{SNML(R)}\right)}{2} \tag{4.3}$$

Using the large-sigma method, the estimation error at left-most tail, which is the interested region, is significantly reduced. Therefore, the proposed large-sigma method has to be applied to increase modeling accuracy of SRAM static design margins for large memory cuts, i.e. for cut sizes in which the distribution tail modeling requires investigations of further samples than 4σ variations. It is worth to say that the large-sigma method is also valid for Write-Margin modeling, since the latter is also determined as the minimum of separately quantified margins (section 3.5.1) and represents the same statistical variability dynamics.

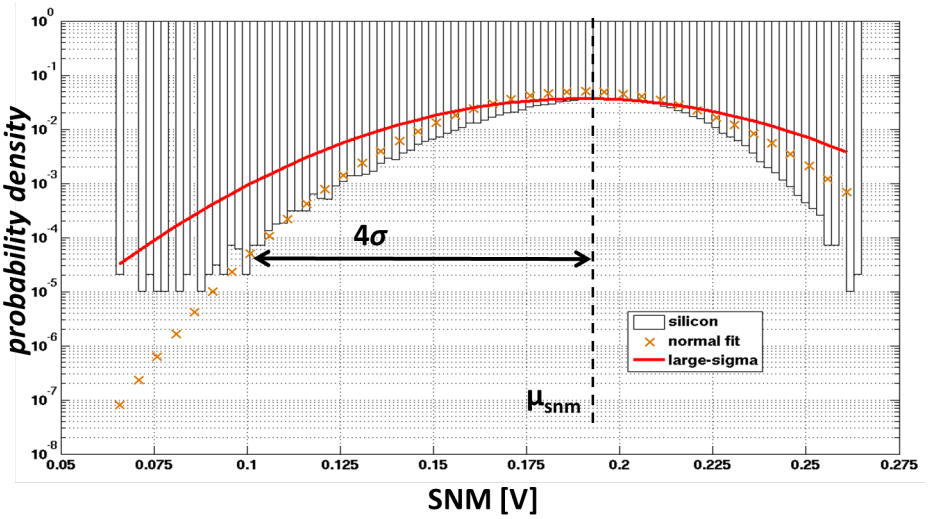

Figure 4.4: Same SNM measurements probability density function as in figure 4.3, presented in log-scale (bars). CAD results with the Gaussian approximation (crosses) and the large-sigma model for tail estimation (red line) are also shown.

The term "yield loss" is a commonly used term in semiconductor manufacturing indicating the fraction of the number of manufactured chips that do not respect the design specifications, in particular at worst case PVTs of the test metric. The supply voltage reduction trend to reduce power consumption of SRAM circuits has resulted in new yield loss category that is named as voltage induced yield loss or V_{min} induced yield loss [73]. In the following section, the static V_{min} modeling methodology based on the proposed large-sigma method for silicon V_{min} prediction at industrial level manufacturing.

4.2 SRAM Static V_{min} Analysis

4.2.1 6T Bitcells Static V_{min} Measurements and SPICE Modeling Results

V_{min} measurements have been performed on a particular corner lot that has been manufactured in C40 technology by intentionally skewing the process centering from the typical process, to yield FS, SF, SS and FF corners. Two memory cuts of 5Mb Single-Port, REGister (SPREG) and 5Mb SPHD memories have been manufactured on all process corners. The Marinescu test algorithm [107] is used to measure V_{min} with respect to low-frequency write and read operations on 60 dies at each corner, at ambient temperature. In this work, the target V_{min} value is set as the supply voltage V_{dd} at which a given SRAM has 95% of V_{min} yield, and denoted as V_{min}@95%. The 5% of yield less is seen as acceptable considering that the V_{min} is estimated on the worst-case process corner, which represents a 3-σ deviation from the typical process.

Figure 4.5 presents the Cumulative Distribution Functions (CDF) of the V_{min} measurements for each process corner. Due to the confidentiality, the measurement results are published in this report with arbitrary units. The ultimate value for each CDF is the Vdd at which 95% of yield (solid line) is reached. The bitcell V_{min}@95% is the maximum over all process corners V_{min}@95%. The variation in V_{min} across different process corners gives us information about the limiting bitcell operation, which can be either write (SF as worse-case corner) or read (FS as worse-case process corner) depending on how the bitcell was designed and which failure mechanism is dominant (section 3.6). Figure 4.5 shows that the highest V_{min} is obtained at SF process corner, indicating that both bitcells are write-limited. The spread of V_{min} across the different process corners differs between SPREG and SPHD, but for both bitcells, the best V_{min} is obtained at FF process corner. For SPHD bitcell, which is the bitcell with the smallest devices and thus the largest variability, the results indicate that the local variability overwhelms the process variations, since the spread of V_{min} across different process corners is narrower than the spread at a single corner. This finding confirms that the random variations may become much critical than the systematic variation for aggressively scaled transistors.

The bitcell V_{min}@95% has been modeled in simulations with respect to read and write static fails. The existing SPICE model cards have been aligned preliminarily to the measured threshold voltages of the processed corners. Figure 4.6 (a) present the saturation threshold voltage of the so-manufactured NMOS and PMOS devices (averages over 50 devices). The spice model corners have been manually skewed in order to be aligned to these values. The good agreement between CAD models and silicon data across the corners was validated by comparing the average I_{cell} value for

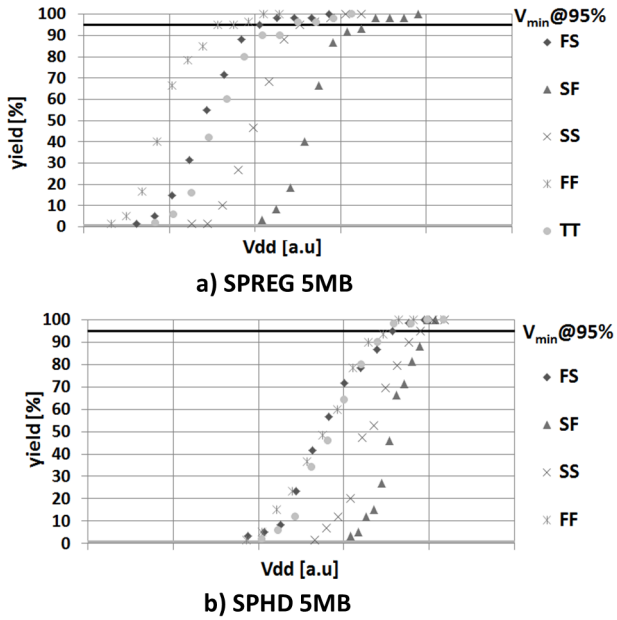

Figure 4.5: Cumulative distribution functions of the measured V_{min} (measured yield) versus the operating voltage for a) Single-Port Register b) Single-Port High-Density bitcells manufactured in C40 technology. V_{min}@95 is the Vdd at which one has 95% of yield (solid line). The two graphs have a common voltage scale.

each corner, as shown in Figure 4.6 (b).

The read and write operation V_{min} are separately calculated using SNM and WM tests (section 3.5.1), respectively, and the final bitcell V_{min} is the maximum of these two. Both SNM and WM distributions tails are modeled using the proposed large-sigma method. The modeling methodology is used as follows: MC simulations are performed at 25 °C with STMicroelectronics C040 SPICE models. The first two moments (μ and σ) of the SNM and WM distributions are extracted together with the second moment of the half-cell distribution (left or right inverter). σ_{LS} is calculated as in (4.3), and use to estimate the margin distribution. It is easy to use table 4.1 to obtain the probability of a fail, i.e. if a margin distribution is normal, and the μ/σ ratio is 4.9, the average fail probability is one over 2Mb. However, this does not mean that, if one manufactures cuts of 2Mb, all cuts will contain one fail. The number of fails per cuts will instead follow a Poisson distribution, which is taken into account in the existing model to estimate the final yield of a given a memory cut size.

Figure 4.7 presents the Cumulative Distribution Functions (CDF) of

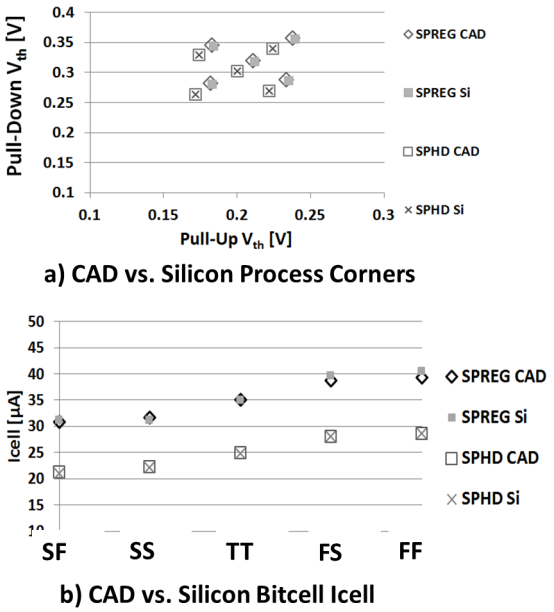

Figure 4.6: a) CAD threshold voltage alignment (empty symbols) with respect to the silicon (filled symbols) measurements at each process corners for pull-down and pull-up transistors. b) Aligned-CAD (empty symbols) vs. Silicon (filled symbols) bitcell average Icell.

the modeled V_{min} at each process corner for two bitcells. The same Vdd grid spacing as in figure 4.5 is used. For all bitcells under evaluation, the simulation results show that the highest V_{min} is obtained at SF process corner. This is in good agreement with the previous considerations based on the experimental data, showing that all bitcells are write-limited. The best V_{min} is obtained at FS process corner for SPREG and SPHD bitcells, which is evidently a consequence of the design architectural choices, in other words those bitcells are read-optimized. It has be noted that the spread of V_{min} across the different process corners differs in some cases from the CAD results.

Figure 4.8 summarizes the comparison between CAD-based modeling and measurements at each process corner. The largest V_{min}@95% estimation error is 46 mV and 48 mV at FS corner respectively for SPREG and SPHD bitcells. On the other hand, the estimation error for the bitcell V_{min}@95% (the highest V_{min}@95% among all process corner) is 15mV for SPREG bitcell and 31 mV for SPHD bitcell, both at SF process corner.

The main source of estimation error looks to be due to the large sensitivity of the results with respect to the device V_{th}. In particular, the CAD

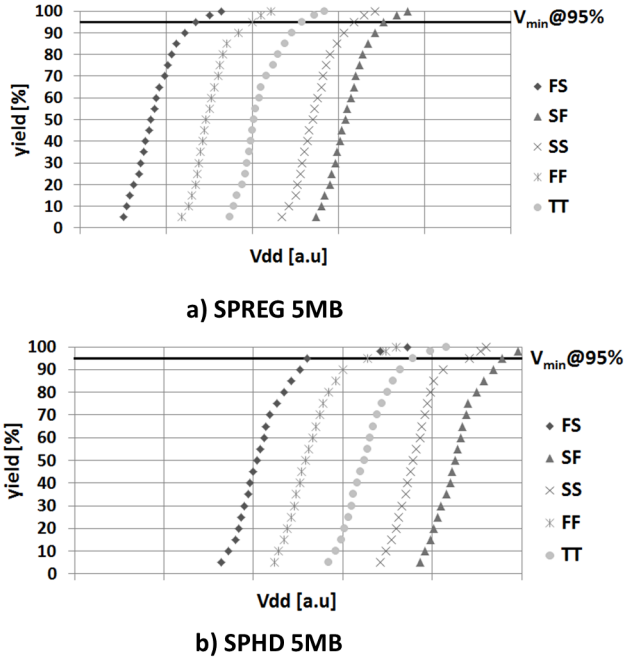

Figure 4.7: Cumulative distribution functions of the simulated Vmin (simulated yield) as obtained with C40 models versus the operating voltage for a) Single-Port Register b) Single-Port High-Density bitcells. V_{min}@95% is the Vdd at which 95% of yield (solid line) is reached. The two graphs have the same voltage scale as those of figure 4.5.

vs. silicon alignment for each process corner is done by averaging the V_{th} over only 50 devices, which limits the accuracy of the alignment process, since it includes a significant statistical error. The results show that, even if the overall spread of silicon V_{min} across different process corners cannot be modeled with a good accuracy, the worst-case silicon V_{min}, i.e. the limiting value, which is the V_{min}@95% yield, can be estimated within 30 mV accuracy.

4.2.2 Ultra-Low-Voltage SRAM Static V_{min} Measurements and Modeling Results

Considering SRAM manufacturing at an industrial level, in which many different SRAM designs using different architectures are manufactured, a SPICE methodology for design optimization has also to be adaptable for all of the manufactured architectures. The static design margin modeling methodology presented in the previous section has been applied to the ST Microelectronics ULV 10T bitcell [80,81]. This alternate bitcell architecture

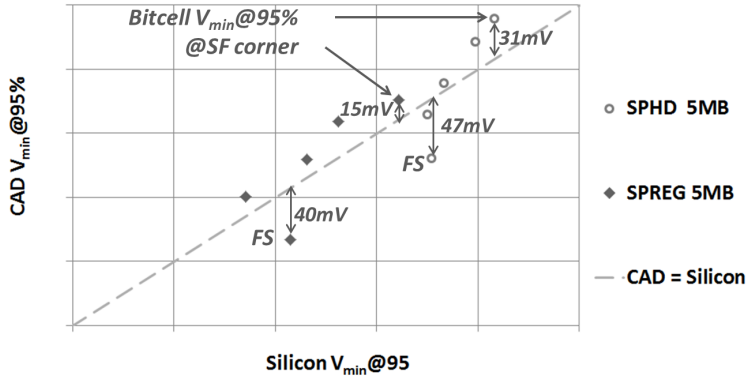

Figure 4.8: CAD vs. Silicon V_{min}@95% for the two C40 bitcells at each process corner. The higher V_{min} of all corners is the bitcell V_{min}@95%. The grid lines have the same spacing as in figure 4.5 and figure 4.7.

is described in section 3.3. The V_{min} estimation is a very critical concern in ULV field, since the main motivation of the ULV design is to save power by reducing the operating voltage Vdd. As a consequence, a sufficiently accurate estimation of the V_{min} in SPICE becomes crucial to optimize for ULV a bitcell design.

Figure 4.9 presents the 10T ULV bitcell V_{min} measurements (circles and crosses) obtained from a 32Kb cut in C65. V_{min} measurements has been performed using two different tests: SCAN test presenting optimistic results, and the Checkerboard (CHKB) test which gives more pessimistic results. The SCAN test algorithm covers only stuck-at type faults while the CHKB test triggers also leakage faults (due to the used check board-like pattern) since it presents a worst-case scenario for leakage paths between the bitcell and the bit-lines. Therefore, the V_{min} measured with CHKB test is higher than the V_{min} measured with SCAN test, since it covers a larger amount of memory faults. If one numerically differentiates data presented in 4.9 to obtain its probability density function, one will see that its maximum and minimum values lie at $\approx \pm 4\sigma$ from the mean. This is in good agreement with the normality property expressed by table 4.1 for a 32Kb cut size. Considering our finding presented in figure 4.4, which indicates that the SNM distribution follows the normal law within $\pm 4\sigma$, it has be noted that the use of the large-sigma modeling is not appropriate since the yield loss would be too pessimistic compared to the reality. Therefore, the read and write V_{min} are modeled in the same way as it is done for C40 6T bitcells, but without using the large-sigma approach. Instead, σ of the WM and SNM distributions are directly used. The diamonds and squares in figure 4.9 present the V_{min} cumulative distribution functions. The modeling methodology is applied on two different process corners, first on typical process in the presence

of global variations (TTG), and then on Worst-Case process corner (WCC) in the presence of only random local variations. Figure 4.9 shows us that the pessimist modeling (crosses) is in line with the aggressive test results, while the optimistic test yield (circles) is coherent with our optimistic modeling application, allowing us to estimate the V_{min}@95% aggressive test yield with an error of 20 mV which is a very good value because of the uncertainties arising in the use of the model cards at ULV operating region that gives an accuracy within 50 mV. This definitely validates for small cuts the existing V_{min} modeling methodology.

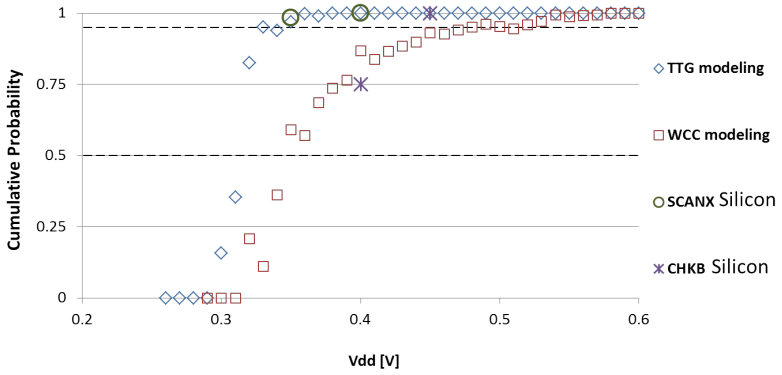

Figure 4.9: Cumulative distribution function of the lowest operating voltage of a population of 32kb ULV memory cut. Experimental data (circles and crosses) are compared to Monte Carlo results (diamonds and squares) obtained at two different process corners. The dot-lines indicate 50% and 95% yield levels.

4.3 Bitcell Variability Space Modeling using Smart Algorithm : Hypersphere Most Probable Failure Point Search Methodology

In the previous section, it is shown that the proposed static design margins modeling methodology gives a sufficiently enough accuracy considering SRAM bitcells operating under DC conditions. However, these margins are calculated with an infinite WL pulse duration and one might question their validity under transient operating conditions. New figures of merit grouped under the name "ability metrics" have been proposed, as already described in detail in section 3.5.2. The ability metrics use the "fail or pass" test for a given criterion instead of calculating margins with respect to the bitcell stable states. The "fail or pass" test introduce statistical modeling complexity, since the bitcell variability space can not be simply modeled with

analytical distribution laws. Moreover, considering an SRAM bitcell operating at high frequency, the number of possible failure mechanisms and the chance of transition between the different mechanisms increase. Therefore, the basic Monte Carlo method leads to unfeasible time and computing costs considering that the target of the analysis is not only building the bitcell variability space, but also investigating it in a more efficient way across all possible operating conditions. The term "efficient" is chosen intentionally to point out that the designer has to travel the variability space, i.e. searching fails in the variability space, with as less as possible complexity. The basic Monte Carlo approach is expected to explore each direction in the variability space with equal probability and this implies that a large number of Monte Carlo runs will be done for configurations that are very far away from fail. This bottleneck gives rise the need for an in-house smart variability space search algorithm.

It has to be noted that the SRAM bitcell failure is monotonic with respect to the variations, which means that when the bitcell is in failure zone, a stronger variation will keep it in the failure zone, while a lighter variation will push it back to the pass zone. Using this property, we can assume that the pass region has simple connectivity, from the topological view. In other words, the boundary between fail and pass is a simple hypersurface in the bitcell variability space. Amongst all points of this surface, one lies closest to the origin with respect to all other points. By definition of the bitcell variability space, this is the Most Probable Failure Point [98]. An interesting way to describe rare events behavior is the Large Deviation Theory [108] that is put into words as "*When a rare event happens, it happens in the most likely manner*". By definition, the MPFP is the point at which failure just occurs with the least amount of variations (highest probability), and, as larger variations will keep the bitcell in failure, thus their occurrence probability will be lower than the MPFP. We can therefore neglect by the large deviation theory these less-possible points and concentrate an effort to search the MPFP, which is illustrated for a 2-dimensional variability space in figure 4.10, in which the pass region and the failure region are highlighted. The MPFP is the closest point to the origin of the space lying on the boundary between the failure and pass regions.

Therefore, if the MPFP of a given failure criterion can be found in the bitcell variability space, the bitcell failure probability for that given criterion can be also estimated. The MPFP-based SRAM failure probability estimation methods have been popular in recent years [98, 109] [110]. The analysis presents two main steps: Identification of MPFP and the exact failure probability extraction. [98] and [109] propose MPFP search through norm minimization algorithms based on the Monte Carlo approach. Both optimize their algorithms by supposing that the failure region in the variability space is judiciously selected. In other words, the mismatch combination of the different transistors in the bitcell for a given failure criterion is assumed to be

known. Even though this assumption is practical and the mismatch combinations are generally known, different combinations can be observed under some conditions; for example in high-frequency multiple WL pulse analysis, the failures are mainly dependent on the dynamic states of the internal nodes and thus any assumption based on stable states may be wrong. The proposed search algorithm of this work do not make any assumption about the failure region position in the variability space and finds it by analyzing all possible directions. However, if the direction that has to be analyzed is known, the proposed algorithm can be also run particularly in that region.

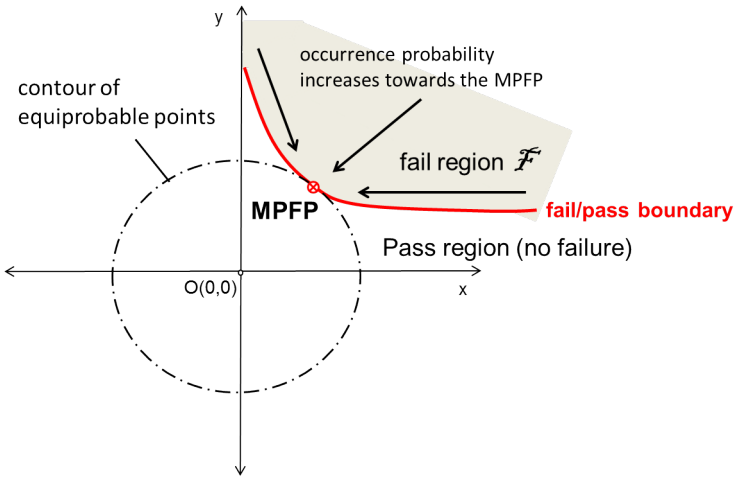

Figure 4.10: Most Probable Failure Point illustration in a 2-dimensional variability space. MPFP is the failure point with the smallest norm lying on the fail/pass boundary. The smallest norm indicates highest occurence probability.

In section 3.4, the bitcell variability space concept is presented as a modeling approximation for the random variability impact in SRAM transistors, assuming that the change in the electrical characteristics of a given transistor can be represented by a change in its threshold voltage (V_{th}). The 6-dimensional variability space of a 6T SRAM bitcell is therefore represented by six V_{th} skews, and each skew follows Gaussian probability density function with null mean and a given standard variation. Considering the 6T bitcell with two cross-coupled inverters formed by PD1, PU1, PD2, PU2, and two access transistors PG1 and PG2, a particular variant bitcell in the variability space is then described by the vector

$$(n_1\sigma_{Vth_{PD1}}, n_2\sigma_{Vth_{PU1}}, n_3\sigma_{Vth_{PG1}}, n_4\sigma_{Vth_{PD2}}, n_5\sigma_{Vth_{PU2}}, n_6\sigma_{Vth_{PG2}})$$

, and normalizing each skew by its standard deviation, we obtain 6-dimensional Mean-Shift Vector (MSV) [109] $\vec{n} = (n_1, n_2, n_3, n_4, n_5, n_6)$. Assuming that

the skews are independent random variables, the norm of the MSV (the Euclidean distance from the origin O) is inversely proportional to the occurrence probability of the bitcell. If the bitcell fails with respect to a given criterion, the point representing it in the variability space belongs to that criterion fail region. The MPFP is then simply the point in that fail region that has the smallest norm. It is therefore possible to find the MPFP by finding the MSV with the smallest norm that belongs to the failing region. This is possible using the *hypersphere MPFP search algorithm* presented in figure 4.11.

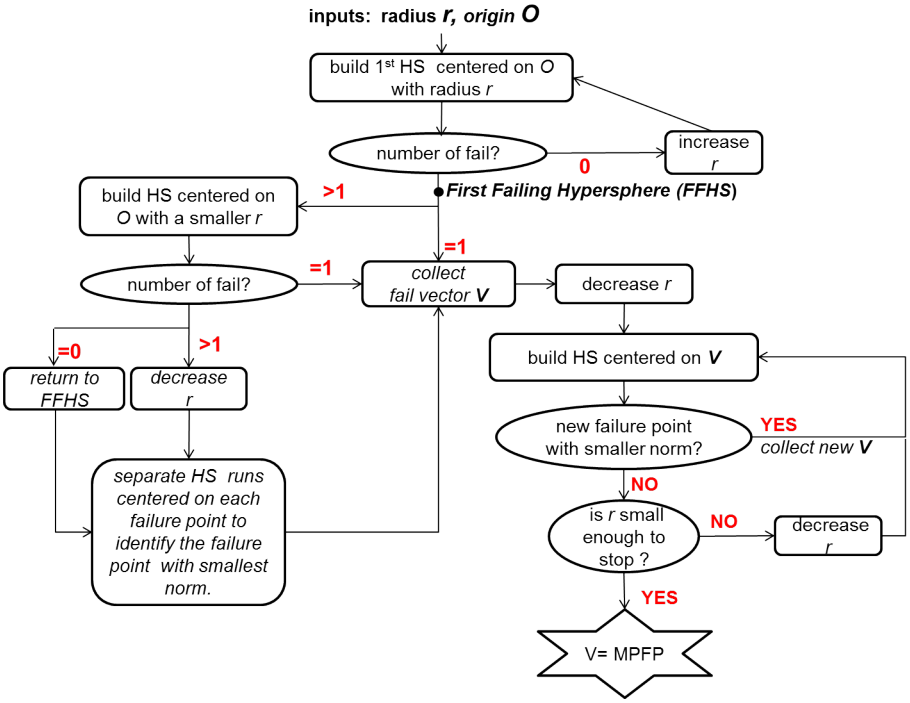

Figure 4.11: The Hypersphere (HS) Most-Probable Failure Point (MPFP) search algorithgm. The origin of the hypersphere is the nominal bitcell.

The analysis starts by building an hypersphere with radius r_0 centered on the origin O, which represents the nominal bitcell without any variability. The r_0 is an input parameter and its value can be optimized to speed-up the analysis and to start as close as possible to the failure region. The equiprobable bitcell variants lying on the surface of the hypersphere are tested for a given failure criterion. The analysis is carried out at a set of points meshed at fixed spacing on the hypersphere surface. The value of r_0 is increased until a failing point is detected. When a failing point is detected, it is defined as the center of a new hypersphere with a radius $r_1 < r_0$, which is meshed at a finer spacing. A new analysis is carried out on the new mesh of points, and if more than one failing point is detected with equal

probabilities, i.e on the surface of the same hypersphere, each failing point is separately analyzed, and the one leading to the failure point with the smallest norm is selected. As evidently from figure 4.12, the procedure is iterated by decreasing the radius r_i down to a minimum value which sets the accuracy of the search algorithm. Figure 4.12 illustrates the MPFP search algorithm in a 2-dimensional variability space. The analysis starts at the origin O, and finds the MPFP at F_4 via analysis performed respectively at failure points F_1, F_2, F_3. Although the method has been developed independently, it is similar to those presented in [110], but it is based on a quicker brute-force approach on the hypersphere surface, instead of a Monte Carlo approach on the full hypersphere volume.

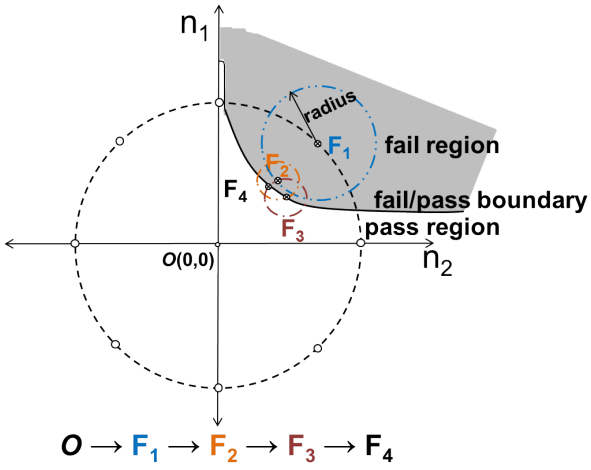

Figure 4.12: An example of MPFP search in a 2D variability space for two skew variables n_1 and n_2. The hypersphere surface in 2D is a circumference. The algorithm detects successively F1, F2 and F3 and finally F4 as MPFP.

4.3.1 Bitcell Failure Probability Calculation

The norm of the MSV is tied to the occurrence probability, since a smaller norm means a higher probability. Considering that each component of MSV is a normally distributed random variable, we can build analytically the probability density function of a given norm $\|\vec{n}\|$. In fact, the square root of the sum of squares of m standard independent normal random variables follows the chi-distribution law with degree m [111]. A worse-case estimation of the failure probability is then $P(\|\vec{n}\| > \|\vec{n}_{mpfp}\|)$. However, we would obtain the occurrence probability of all bitcells that are lying outside the hypersurface with radius $\|\vec{n}_{mpfp}\|$, which thus results in a very large overestimation of the failure probability as visible in figure 4.13 for a 2-dimensional

space.

In [98], the authors propose the following: Considering the random variable $X=(x_1, ..., x_m)$ with $x_i = \eta(\mu_i, \sigma_i^2)$ and the MPFP $(n_1\sigma_1, ..., n_m\sigma_m)$ of a given failure region F, the failure probability is approximated to:

$$P(X \in F) = \prod_{i=1}^{M} P(x_i \geq n_i\sigma_i) \tag{4.4}$$

The assumption in the equation (4.4) is illustrated in figure 4.13 for a 2-dimensional space : The failure region is considered as the shaded area and the grayed area is neglected, which thus results in an underestimation of the failure probability.

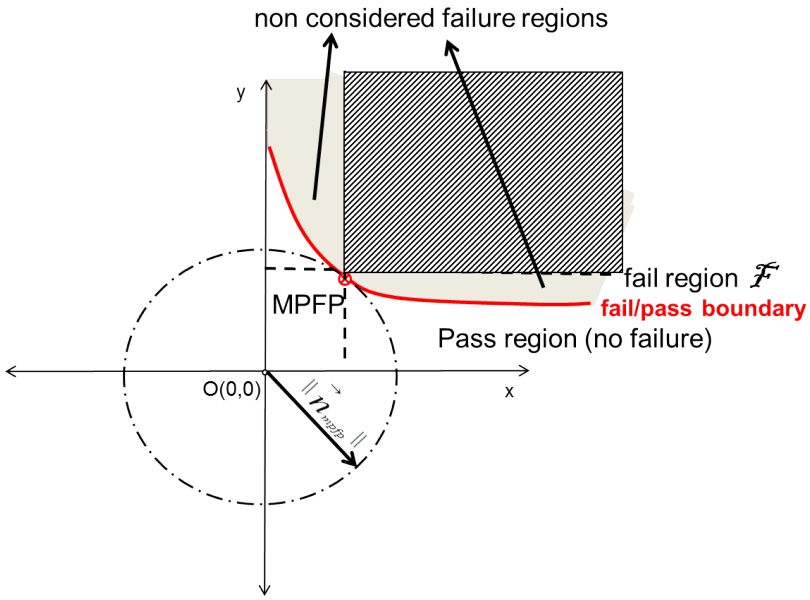

Figure 4.13: Failure probability estimation propsed in [98]. The method takes into consideration only the shaded area of the failure region and exludes the grayed parts, thus results in an underestimation of the failure probability.

Considering the MPFP as a variant bitcell occurring at the tail of the given failure metric distribution, the failure probability can be numerically estimated using importance sampling methods [109]. Importance sampling is a common method in statistical analysis for increasing the estimation accuracy in the distribution tails, where the occurrence probabilities are so

low that the standard Monte Carlo simulation is not practical. Considering a particular distribution, the importance sampling consists in a modified MC where one samples from a distribution in such a manner to overweight the important region. Having oversampled the important region, any estimate has to be corrected numerically to produce non-oversampled results. A fundamental advantage is that a very precise distribution tail form can be extracted with no analytical assumptions made for distribution law.

Let us consider a random variable X taking value in a certain domain. The probability that X is in a particular region of interest $A_{interst}$ $Pr(X \in A_{interest})$ is indicated as p_{fail}. The Monte Carlo method generates N independent samples $X^1, ..., X^N$ of X. The estimate of p_{fail} using Monte Carlo method is simply

$$\hat{p}_{fail} = \frac{1}{N} \sum_{k=1}^{N} \left(X^k \in A_{interest} \right)$$

where the expression in parenthesis is a logical expression that takes the value 1 when $X^k \in A_{interest}$ and 0 otherwise. As said before, if p_{fail} is very small, the classic Monte Carlo method needs too many samples to have a good accuracy, because it requires approximately $1/p_{fail}$ samples to have a reasonable probability of observing at least one sample belonging to $A_{interest}$. Let us consider a different random variable $\hat{X}$ with a distribution such that $\hat{p} = Pr(\hat{X} \in A_{interest})$ is not rare (say $\hat{p}$ is close to 0.5). If we denote the density function of X as f(x) and the density function of $\hat{X}$ as $\hat{f}(x)$, one can write the weight function as:

$$w(x) = \frac{f(x)}{\hat{f}(x)}, \text{ for all x.}$$

In such a scenario, $\hat{p}_{fail}$ can be found quickly by performing a Monte Carlo analysis of $\hat{X}$ and then adjusting the estimate using w. This is the key idea behind the importance sampling, which is illustrated in figure 4.14. Evidently, if $f(x)$ is a known distribution law, the use of importance sampling for p_{fail} estimation does not present any gain. However, if we consider a SRAM bitcell failure criterion as a random variable, which is itself a function of a set of 6 different random variables, p_{fail} estimation becomes a complex problem.

The analytical expression of the weight function w(x) is determined as in [109]: Let us first consider some function of a set of independent random variables X $(Y_1, ..., Y_M)$ where each Y_i has a Gaussian distribution with mean μ_i and standard deviation σ_i. In a SRAM, X represents a given bitcell failure criterion and Y_i is the threshold voltage of the $i-th$ transistor. Now consider $\hat{X}$=$(\hat{Y}_1, ..., \hat{Y}_M)$ where $\hat{Y}_i$ has a Gaussian distribution with same standard deviation σ_i as Y_i, but with a mean $\mu_i + s_i$, where s_i is the skew value chosen such that $\hat{X}_i$ has its mean towards $A_{interest}$ region, as shown

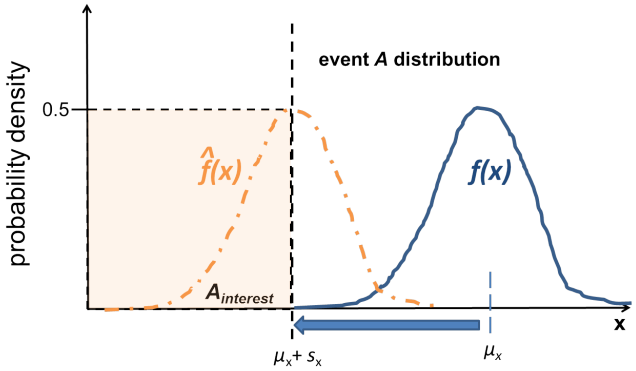

Figure 4.14: Considering an event A that is characterized by a random variable X and its the probabability density function $f(x)$, the interested low probability tail region $A_{interest}$ can be oversampled using a different random variable $\hat{X}$ that has its probability density function $\hat{f}(x)$ centered at $A_{interest}$. This is the key idea behind Importance Sampling.

in figure 4.14. Since both Y_i and $\hat{Y}_i$ have a Gaussian distribution, their probability density functions denoted as $f(y_i)$ and $\hat{f}(y_i)$, respectively, can be written as:

$$\begin{aligned} f(y_i) &= \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma_i^2}\right) \\ \hat{f}(y_i) &= \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(y_i-(\mu_i+s_i))^2}{2\sigma_i^2}\right) \end{aligned}$$

Therefore the weight function can be calculated as:

$$\begin{aligned} w(y_1,...,y_M) &= \frac{f(y_1,...,y_M)}{\hat{f}(y_1,...,y_M)} \\ &= \frac{\exp\left(-\sum_{i=1}^{M}\frac{(y_i-\mu_i)^2}{2\sigma_i^2}\right)}{\exp\left(-\sum_{i=1}^{M}\frac{(y_i-\mu_i-s_i)^2}{2\sigma_i^2}\right)} \\ &= \exp\left(-\sum_{i=1}^{M}\frac{s_i(2(y_i-\mu_i)-s_i}{2\sigma_i^2}\right) \end{aligned} \tag{4.5}$$

Since w is known, considering that the failure region of a given SRAM bitcell test criterion is the region where $X \in A_{interest}$, the failure probability is estimated in the following way: First, the failure probability of the skewed variable $\hat{X}$ is numerically estimated by running Monte Carlo simulations centered at the MPFP. The resulting failure probability estimate, denoted

as $\hat{p}_{mpfp}$, is expected to be ≈0.5. For N Monte Carlo iterations centered at the MPFP, one has

$$\hat{p}_{mpfp} = \frac{1}{N}\sum_{k=1}^{N}(\hat{X}^k \in A_{interest}) \tag{4.6}$$

where the index k represents k-th Monte Carlo run centered at the MPFP. Considering a failing sample $\hat{X}^k$, its contribution to $\hat{p}_{mpfp}$ is 1/N and its contribution to $\hat{p}_{fail}$ is therefore w(X^k)/N. In other words, $P(X \in A_{interest})$ can be estimated using Importance Sampling and is denoted as $\hat{p}_{is}$ which is equal to:

$$\hat{p}_{is} = \frac{1}{N}\sum_{k=1}^{N} w(X^k)(\hat{X}^k \in A_{interest}) \tag{4.7}$$

Another concern for the accuracy of the failure probability estimation is the required number of Monte Carlo iterations centered at MPFP. As it is proposed in [109], the estimator accuracy can be quantified by the estimator variance. The variance of the estimator $\hat{p}$, denoted as VAR($\hat{p}$) is

$$VAR(\hat{p}) = \frac{1}{N}\left(E[p^2] - E[p]^2\right) \tag{4.8}$$

where $E[*]$ denotes the expectation value operator. From equation (4.7) and (4.8), the variance of the importance sampling estimator can be calculated as:

$$\begin{aligned} VAR(\hat{p}_{is}) &= \frac{1}{N}\left(E[(\hat{X}^k \in A_{interest})^2 w(X^k)^2] - \hat{p}_{is}^2\right) \\ &= \frac{1}{N}\left(E[(\hat{X}^k \in A_{interest}) w(X^k)^2] - \hat{p}_{is}^2\right) \\ &= \frac{1}{N}\left(\frac{1}{N}\sum_{k=1}^{N}(\hat{X}^k \in A_{interest}) w(X^k)^2 - \hat{p}_{is}^2\right) \end{aligned}$$

in which $(\hat{X}^k \in A_{interest})^2 = (\hat{X}^k \in A_{interest})$, since $\hat{X}$ assumes only the $\{0, 1\}$ values. Given the variance of the estimator VAR($\hat{p}_{is}$), the figure of merit for estimator accuracy, $\rho(\hat{p}_{is})$, is defined as

$$\rho(\hat{p}_{is}) = \frac{\sqrt{VAR(\hat{p}_{is})}}{\hat{p}_{is}} \tag{4.9}$$

In this work, $\rho(\hat{p}_{is})$ is used as follow: The expression $(\hat{X}^k \in A_{interest})$ represents a Bernoulli trial [112], since it is a "fail or pass" test giving only 1 (fail), or 0 (pass) as output. Therefore, one can write equation (4.8) for $\hat{p}_{is}$ as:

$$VAR(\hat{p}_{is}) = \frac{1}{N}\left[(\hat{p}_{is}(1-\hat{p}_{is})\right] \tag{4.10}$$

Assuming a normal distribution law with the required confidence λ and error percentage ϵ, from equation (4.10), one can write equation (4.1) as:

$$\hat{p}_{is} = \frac{Z_\lambda}{\epsilon}\sqrt{\frac{\hat{p}_{is}(1-\hat{p}_{is})}{N}} = \frac{Z_\lambda}{\epsilon}\sqrt{VAR(\hat{p}_{is})}$$

Thus from equation (4.9), we obtain:

$$\rho(\hat{p}_{is}) = \frac{\epsilon}{Z_\lambda}$$

For example, if 90% confidence ($\lambda = 0.9, Z_\lambda = 1.644$) and $\pm$10% error ($\epsilon = 0.1$) are required, the Monte Carlo simulations centered at MPFP have to be run until $\rho(\hat{p}_{is}) \approx 0.06$ is reached.

Finally, it is worth to say that the proposed MPFP search through Fail/Pass test that is initially developed for transient analysis, can be also applied for static design margin estimation. For example, if SNM test is considered, the failure region can be set as the region where $SNM \leq 0$, and then the hypersphere MPFP search algorithm can be run for the given bitcell design.

4.4 SRAM Dynamic Vmin Analysis

4.4.1 Bitcell Dynamic Fail/Pass SPICE Analysis using Hypersphere Algorithm

The Hypersphere Most Probable Failure Point (MPFP) search algorithm is applied on the 28nm UTBB FD-SOI technology SRAM bitcells for dynamic V_{min} investigations, in particular for the investigation of the WL pulse dependency of read and write operations. The analyses are carried out on two different bitcells : Single-Port $0.120\mu m^2$ High-Density (SPHD) and Single-Port 0.197 μm^2 REGister Low-Voltage (SPREGLV) bitcells. The SRAM portfolio of ST Microelecronics includes two SPHD bitcell with two different well architectures: First one is the bulk-like Dual-Well (DW) bitcell in which NMOS is lying on a P-Well and PMOS is lying on N-Well. The second one is the Single-P-Well (SPW) architecture [99] in which NMOS and PMOS share a common P-Well, which is peculiar to the UTBB FD-SOI technology. This feature is feasible thanks to the Buried-Oxide (BOX) under the thin silicon film. BOX enables the use of the transistor body as a second gate [50, 62] allowing in principal to form a second channel at the interface of the silicon film and the BOX, and a stronger channel inversion occurs if the well under the BOX is doped with same type of atoms as the drain and source. Therefore, the use of P-Well (p-type back-plane)

for PMOS allows lowering its V_{th}. Both well architectures are depicted in figure 4.15. It is worth to say that a design with Single-N-Well architecture is also possible, in which the common N-Well would allow lowering NMOS V_{th}. The SPREGLV bitcell exists only with SPW architecture.

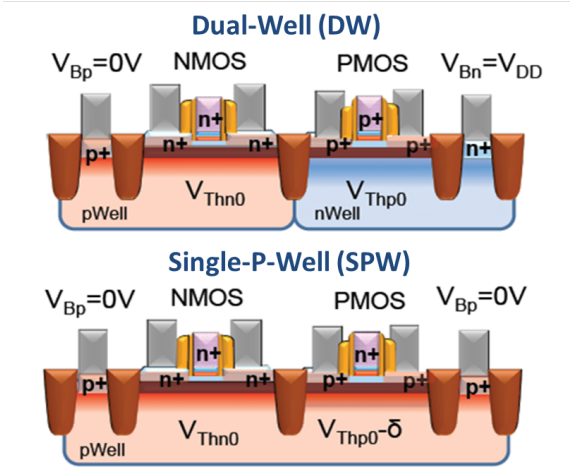

Figure 4.15: Different well architectures in 28nm UTBB FD-SOI [99]. In Dual-Well bitcell the NMOS is lying on a P-Well and PMOS on a N-Well as in the bulk CMOS technology. In the Single-P-Well architecure, NMOS and PMOS shares the same P-Well, which allows lowering the Vth of PMOS. The P-Well back-plane voltage V_{Bp} is set to 0V, which puts the PMOS in forward-body bias mode

Figure 4.16 illustrates the simulation netlist, in which a column of the memory array is instantiated. The bitcell under test content is initialized with a "1" in the node L and the content of the other bitcells are complementary to the bitcell under test. This setup represents a worst-case condition for a read operation, because while only the bitcell under test discharges BLR, the leakage currents of the load bitcells tend to discharge BLL. Simulations are performed under both high-frequency and low-frequency operating conditions, at ambient temperature and typical process corner. In high-frequency operations, the WL duration is set to one half of the clock period, which is scaled with V_{dd} as extracted from a critical path of ARM A9 Cortex, assuming 2 GHz at 1V. Low-frequency operations are simulated by dividing ARM A9 Cortex frequencies by a factor of 10.

4.4.1.1 Read-Ability Test

The RA test that is described in section 3.5.2.1 is applied by assuming 100mV for SA_{offset}, which can be seen as an aggressive value for low-voltage

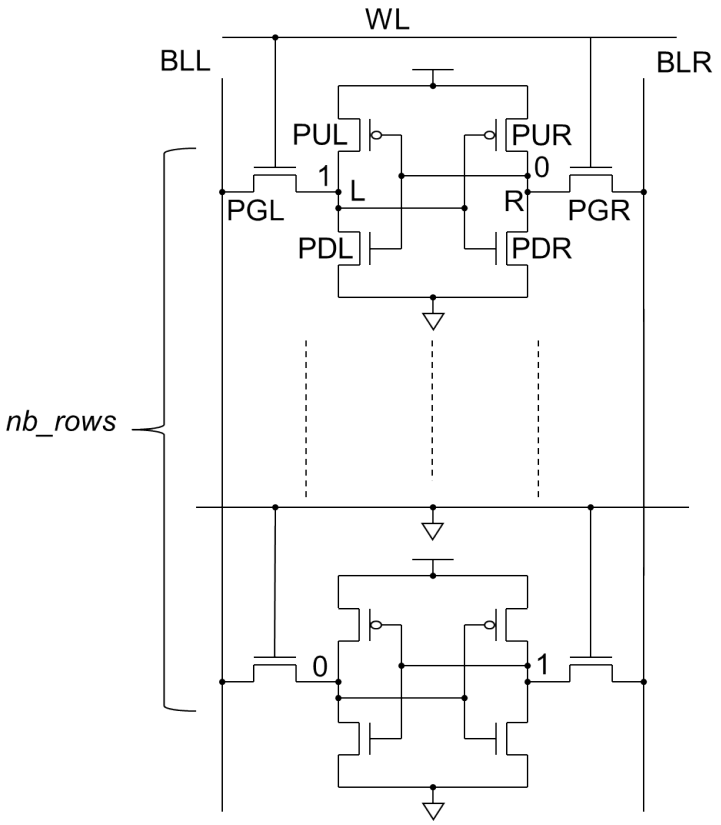

Figure 4.16: The 6T bitcell simulation netlist schematic for transisient simulation. The bitcell under test is initialized to store "1" in the node L, whereas other cells of the same column store the oppposite content. The number of cells in a given column is denoted as nb_rows.

operations in the UTBB FD-SOI technology [113].

Figure 4.17 shows high-frequency RA MPFP MSVs of SPHD SPW bitcell, for 64,128 and 256 rows, at V_{dd}=0.5V and 0.6V, with V_b=0V. The skew factor of each device forming the MPFP MSV is represented by bars. A positive skew indicates a weakened device, whereas a negative skew indicates a strengthened device. MSVs show that the failure mechanism is tied to slow PGR, indicating that the failures are *RA failures* as described in section 3.5.2.1: an insufficient bit-line voltage difference to activate SA at the end of the WL pulse. The PGR skew factor of the MPFP is getting smaller for longer columns, so does the norm of the MSV, thus the MPFP becomes closer to the origin of the hypersphere, which means an increase of the RA failure probability. This increase is due to not only the increased leakage currents that makes harder to reach the desired BL voltage difference at the end of WL pulse, but also to the change in the parasitic load (capacitance

and resistance) with larger columns. The RA failures remain same in DW bitcell, since the PGR NMOS device remains intact after the well change under the PMOS.

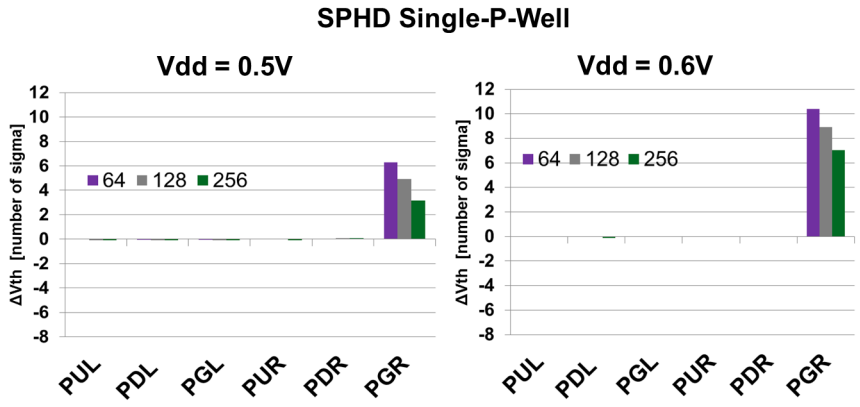

Figure 4.17: SPHD SPW bitcell high-frequency RA MPFP MSVs with 100 mV SA_{offset}, at V_b=0V. Simulations are perfomed for 64,128 and 256 rows. RA fails are tied to slow PGR and PGR skew factor decreases with the increase in the column lenght. The MPFP MSV are same for DW bitcell (not shown), since the NMOS PGR is not affected by the well change.

Figure 4.18 presents high-frequency RA MPFP MSVs of SPREGLV bitcell, for 64,128 and 256 rows, at V_{dd}=0.5V and 0.6V, with V_b=0V. As for the SPHD bitcell, the RA failures are tied to the slow PGR device, and the the PGR skew factor of the MPFP is getting smaller for longer columns. Comparing figure 4.17 with 4.18, one should note that the MSV norms of MPFPs of the SPREGLV bitcell are larger than those of SPHD bitcell, thus SPREGLV offers safer operations at Low-Voltage.

At high-frequency operations, RA failures in the SPHD bitcell are caused by a non sufficient bit-line voltage difference due to the slow PGR. Figure 4.19 presents RA test results for the same bitcell at low-frequency operations. At a clock period 10 times slower than high-frequency operations, the read failures are mainly caused by fast PG and slow PD on the side of node storing low-logic value and fast PD on the side of the node storing high-logic value. The failure is related to stability issues similar to the stability failures in static conditions (section 3.6), but without any contribution of PMOS devices. A particular case occurs at V_{dd}=0.5V for 256 rows, in which RA failures are caused by the slow PGR as in high-frequency simulations, as a consequence of the high leakage currents and high parasitic load, as already discussed. RA MPFP MSVs of DW bitcell remain same, since failures are tied only to NMOS devices which are not impacted by the well change.

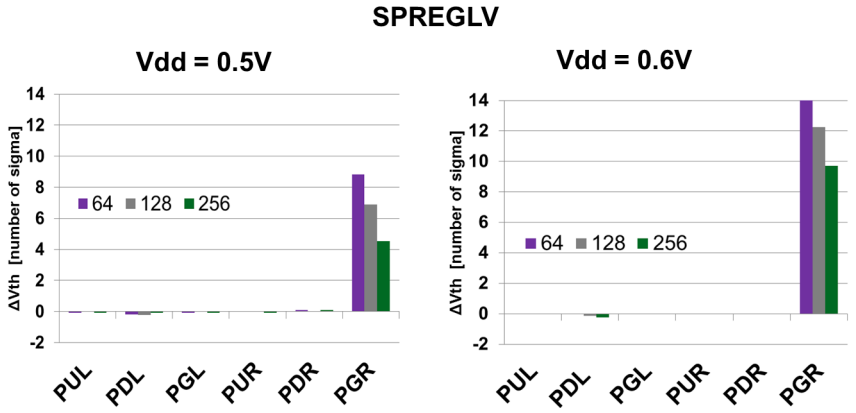

Figure 4.18: SPREGLV bitcell high-frequency RA MPFP MSVs with 100 mV SA_{offset}, at V_b=0V. Simulations are perfomed for 64,128 and 256 rows. RA fails are tied to slow PGR. PGR skew factor decreases with the increase in the column lenght. change.

Finally, a comparison between figure 4.17 and 4.19 allows to show the effects of the WL pulse duration on read failures.

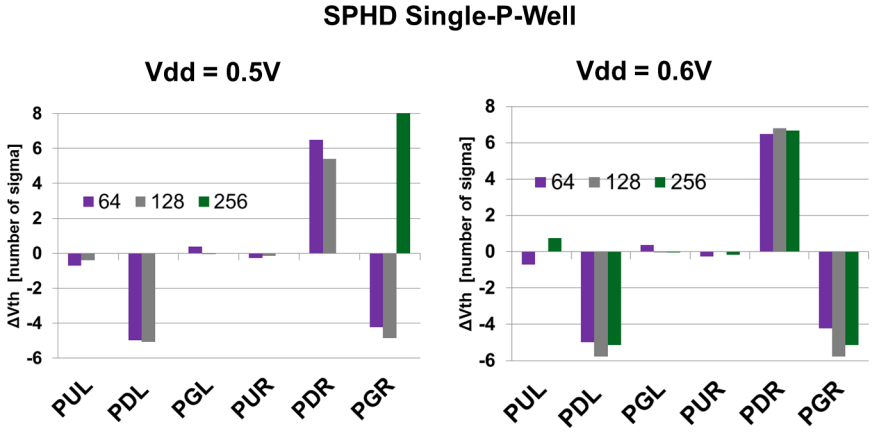

Figure 4.19: RA failures low-frequency MPFP MSVs at Vdd=0.5 and 0.6V and Vb=0V, for 64, 128 and 256 rows. Tthe failures are mainly caused by stability failures as in static conditions.

4.4.1.2 Write-Ability Test

Before starting to discuss Write-Ability (WA) simulation results, a reminder for the different WA failure mechanisms that are presented in section 3.6 may be needed. The WA discharge failures represents the non toggle of the bitcell internal nodes within the WL pulse and are associated to slow PG-fast PU couple on the side of the internal node initially storing '1'. The WA completion failures are caused by the non-stable state of the internal nodes after the toggle occurs and are associated to a slow PU device on the side initially storing '0', which prevents the node reaching high-logic level.

The number of load bitcells sharing the same column as the bitcell under test does not affect write operation in first order, since BLs values are forced to 0 and 1 during the WL pulse. However, the completion of the write operation after internal nodes toggle may be impacted by parasitic loads and leakage currents, but we assume that this impact remain as a second-order and the following analysis are performed with 128 rows which represents neither an aggressive nor relaxed architecture. It is worth to say that the write operation would be strongly impacted by the number of columns in a given memory array, which will lead to a delay in the WL signal depending on the position of the bitcell in that row. This is not taken into account in the bitcell-level simulation.

The WA failure criterion (section 3.5.2.2) is applied on SPHD bitcell with failure threshold p=0.5 and 0.8. Figure 4.20 presents high-frequency WA MPFP MSVs at V_b=0V, for V_{dd}=0.5V, 0.6V and 0.7V, for both DW (A) and SPW (B) architectures. The results shows that in this voltage range the DW bitcell WA failures are dominated by completion failures, whereas the SPW bitcells WA failures are caused by discharge failures. The failure mechanism is different due to the initially strong PMOS device of the SPW bitcell, which cancels out completion failures. The DW bitcell results show that less PUR skew is needed with p=0.8, since the latter represents more aggressive test condition compared to the p=0.5.

Figure 4.21 presents in the same plots SPHD high frequency and low-frequency WA MPFP MSVs at V_{dd}=0.5, 0.6 and 0.7V, with V_b=0V and p=0.8. Only DW bitcell results is shown, since SPW bitcell WA failures are dominated by discharge failures, which does not evolve with the WL pulse duration. Gray bars present the high frequency results (same results as in figure 4.20 for p=0.8) while purple bars show low frequency results. Independently from V_{dd}, the high-frequency completion failure mechanism turns into discharge failures at low-frequency, showing the high dependency of the write failure mechanism on the WL pulse duration. The low frequency results show also that the discharge failure mechanism is more critical in static-like conditions.

Figure 4.22 presents the summary of the high-frequency WA and RA yield of the SPHD bitcell. The σ-yield is determined following table 4.1

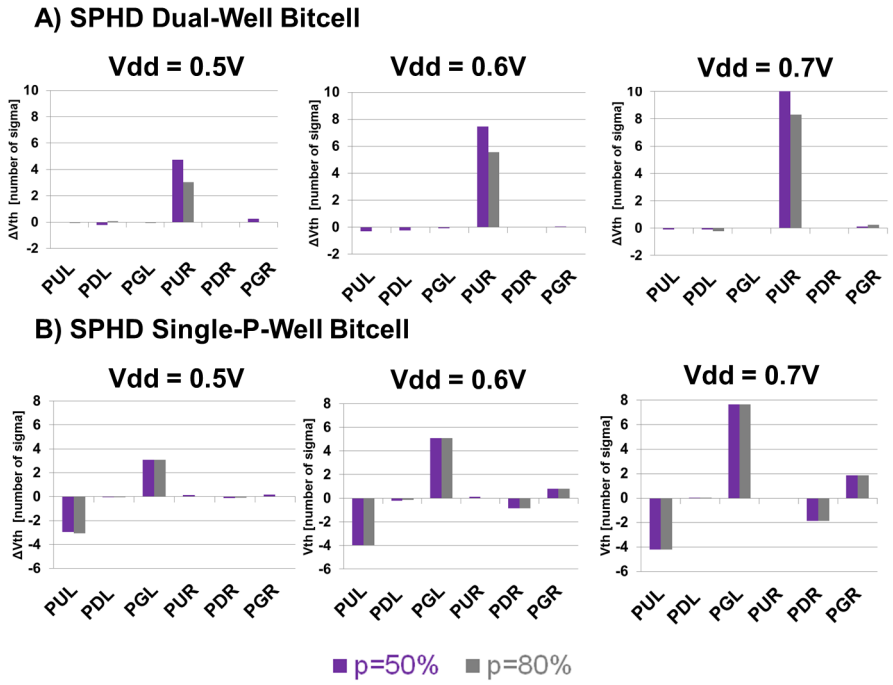

Figure 4.20: SPHD bitcell high-frequency WA MPFP MSVs for A)dual-well and B)single-p-well bitcells at various Vdd, for Vb=0V.

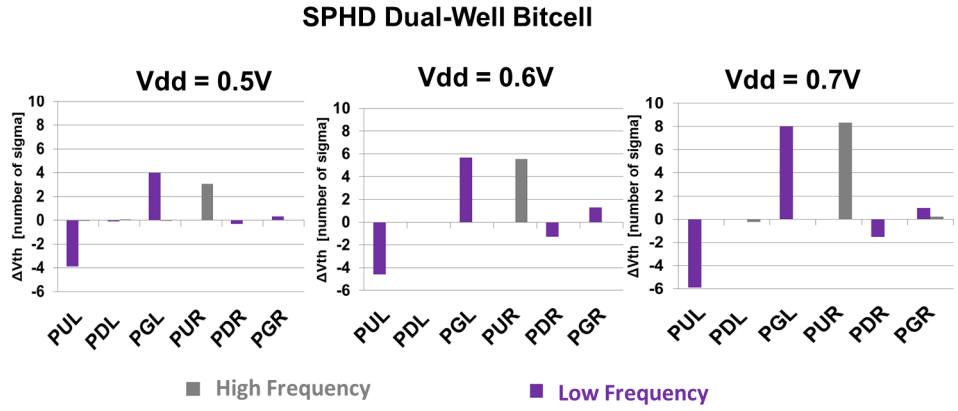

Figure 4.21: SPHD DW bitcell WA MPFP MSVs at high-frequency (purple bars) and low-frequency (gray bars), at Vb=0V. Larger WL pulse width results in a change of the failure mechanism.

as the equivalent standard normal deviation that gives the probability extracted with Importance Sampling centered at the MPFP. The WA yield is only presented with p=0.8, since it represents more aggressive results. It is shown that by reducing column size from 256 to 64 rows, the read V_{min} can be improved by 80mV. On the other hand, using SPW architecture instead DW, the write V_{min} is improved by 40mV thanks to the strengthened PMOS which improves the completion of the write operation. It is also observed that both architectures are write-limited under these operating conditions, since WA yield is systematically lower than RA yield.

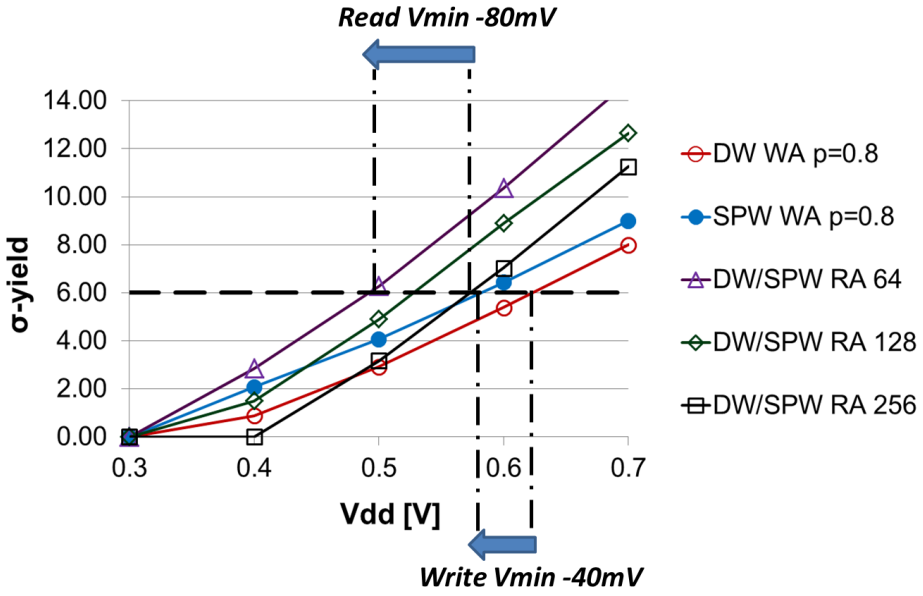

Figure 4.22: SPHD bitcell high-frequency WA and RA yield for both DW and SPW architectures. The sigma-yield are calculated by perfoming Importance Sampling around the MPFP.

Figure 4.23 presents high-frequency WA MPFP MSVs for SPREGLV bitcell at V_{dd}=0.5V and 0.6V, with V_b =0V. WA criterion is applied with p=0.5 and 0.8. At V_{dd}=0.5V, WA test results with the relaxed failure threshold p=0.5 shows that WA failures are generated by slow PGL-fast PUL pair indicating that the failures are caused by the discharge failure mechanism. However, the mechanism is extracted as completion when WA test is applied with aggressive failure threshold p=0.8, in which failures are generated by positively skewed PUR device. If we analyze the results of p=0.8, we see that there is an increase in the PUR skew factor between V_{dd}=0.5V and 0.6V, which shows the improvement of the WA yield with respect to V_{dd}. Considering the results at V_{dd}=0.6V, the decrease in the

PUR skew factor between p=0.5 and p=0.8 demonstrates the degradation of WA yield with respect to the aggressiveness of the test. It is worth to conclude by noting that the parameter dependency of the single-pulse WA test introduces an uncertainty in the results.

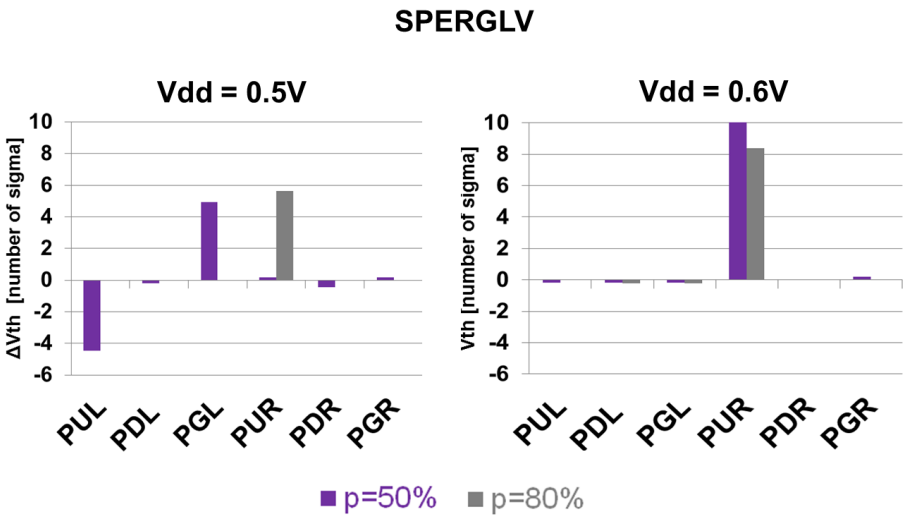

Figure 4.23: SPREGLV SPW bitcell high-frequency WA MPFP MSVs for at V_{dd}=0.5V and 0.6V, at Vb=0V. Results for p=0.5 (relaxed) and p=0.8 (aggressive) are shown.

4.4.1.3 Read-After-Write Test

High-frequency operations results of the SPHD and SPREGLV bitcell show that WA failure probability estimation is highly dependent on the value of p, which can lead to overestimation or underestimation of the failure probability, and even in some particular cases, it can expose different failure mechanisms as the dominant one. At this point, considering that studied bitcells V_{min} are all limited by write operations at high-frequency, the analysis accuracy remains limited by the use of parameter-dependent failure criterion. Furthermore, the analysis based on a single-WL pulse is not a sufficiently realistic representation of the high-frequency memory operations, since in the latter, bitcell internal nodes may undergo a new operation before reaching their stable state. Considering that a "readable" bitcell means in the same time that the bitcell was previously written successfully, the multiple-WL pulse Read-After-Write (RAW) test gives the opportunity for more realistic evaluation. In this section, the RAW test that is previously presented in section 3.5.2.3 is performed on both SPHD DW and SPW

bitcells. 128 rows are instantiated for both analysis. This intermediate value is chosen, since RAW is used in this analysis as a figure of merit for write operation and a subsequent read operation should not affect (help or hinder) write yield dramatically. One should expect that leakage currents and parasitic loads would help the completion of a write operation with this initial conditions.

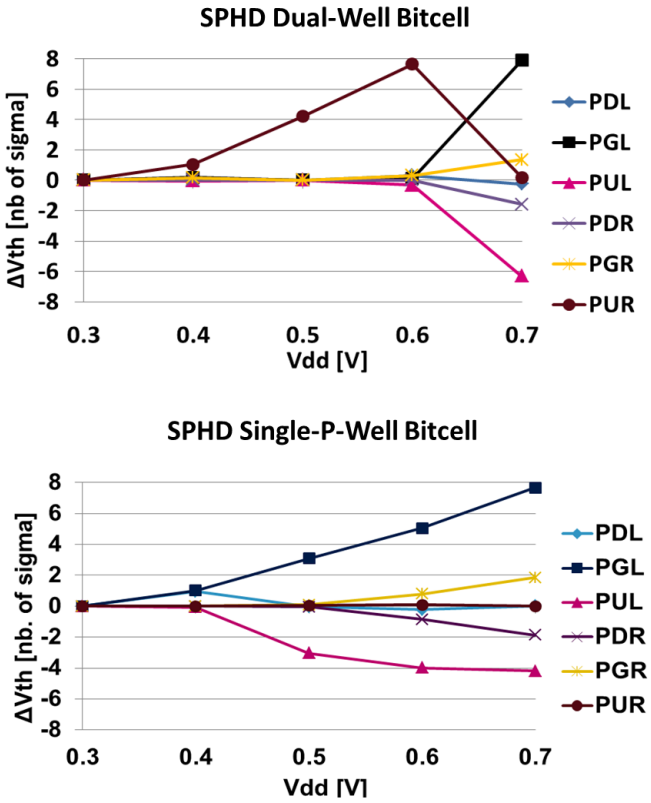

Figure 4.24: SPHD bitcell RAW MPFP MSVs at high-frequency operations at V_b=0V, for DW on the top and SPW on the bottom.

Figure 4.24 represents SPHD bitcell RAW test MPFP MSVs at high-frequency operations for various V_{dd} values, at V_b=0V, for DW architecture on the top and for SPW architecture on the bottom. In the DW bitcell, the main failure mechanism is extracted as the write completion failures below 0.7V V_{dd}, since the MPFP MSV has only one dominant component which is PUR. At 0.7V V_{dd}, it is observed that the main failure mechanism switches to discharge failures, which was not observed in the single pulse WA test that is presented in figure 4.20 top. In the SPW bitcell, the failure mechanism is extracted as the discharge write failures over the whole range

of V_{dd}, which is in coherent with single pulse WA test as presented in figure 4.20 bottom, since discharge failures prevent the bitcell internal nodes toggle during the WL pulse duration and thus they can not evolve with any subsequent operation unlike completion failures.

Figure 4.25 presents SPHD DW and SPW bitcells WA, RA and RAW yields in the same plot, in which WA and RA results are the same one as in figure 4.22. The DW bitcell V_{min}, which is limited by write completion failures in the single-pulse analysis, is simulated as 70mV lower for 64 and 128 rows using RAW test criteria. This improvement is due to the fact that multiple-pulse RAW test allows canceling fake completion failures thus leads to a higher yield. However, it is shown that the DW bitcell is read limited with 256 rows. The use of RAW test does not influence SPW bitcell analysis, since the write discharge failures related V_{min} limitation of the SPW bitcell do not evolve with the number of successive operations.

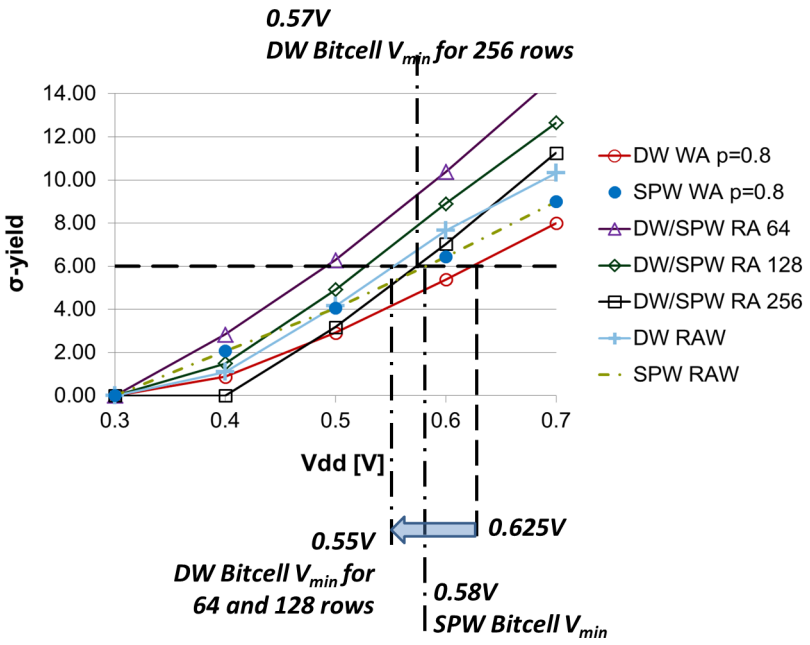

Figure 4.25: SPHD bitcell high-frequency WA, RA and RAW sigma-yields for DW and SPW architectures, at V_b=0V. RAW V_{min} for 64 and 128 rows is 70mV lower than single pulse test. DW bitcell V_{min} becomes read limited with 256 rows. The SPW V_{min} bitcell is limited by write discharge failures thus the multiple pulse analysis gives the same yield as the single-pulse analysis.

Figure 4.26 presents SPREGLV RAW test MPFP MSVs at high-frequency

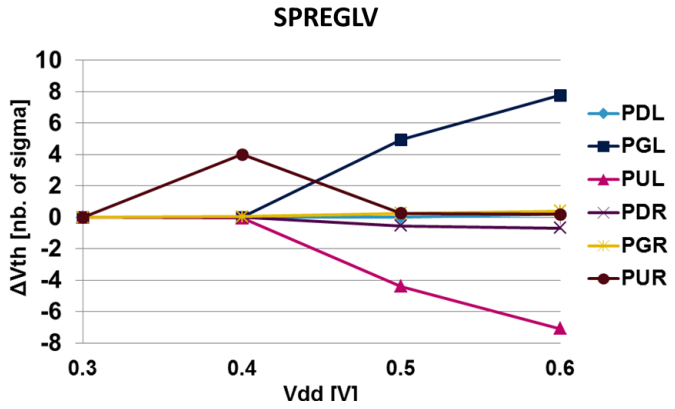

Figure 4.26: SPREGLV bitcell RAW MPFP MSVs at high-frequency operations at V_b=0V.

operations for various V_{dd} values, at V_b=0V. At 0.4V V_{dd}, the failures are tied to slow PUR indicating that the failure mechanism is write completion. At 0.5V and 0.6V of V_{dd}, the failures are caused by slow PGL and fast PUL which indicates that the failure mechanism is the write discharge. The failure mechanisms with respect to the V_{dd} extracted with RAW test, are completely opposite to the ones extracted with single-pulse WA test, which exhibits the uncertainty arising from the use of parameter-dependent single-pulse WA test for high-frequency operation evaluation.

Figure 4.27 presents SPREGLV WA, RA and RAW yields in the same plot. For 64 and 128 rows, the bitcell V_{min} is write limited, and the RAW test leads to a lower V_{min} than WA test with p=0.8. This improvement is due to the fact that multiple-pulse RAW test allows canceling fake completion failures thus leads to a higher yield. The bitcell becomes read-limited with 256 rows due to high leakage.

For the examples that are shown in the dynamic V_{min} analysis, the amount of V_{min} shifts are relatively small, and can be even seen as insignificant considering the inevitable modeling error arising from the use of SPICE model cards. The V_{min} limitation of a given bitcell is highly dependent on the operating conditions, the chosen architecture and the used test criteria, thereby the underlying mismatch mechanisms behind failures, which limit the bitcell V_{min}, evolve severely as well with the change in these parameters. The presented analysis have shown that the Hypersphere MPFP search algorithm coupled with Importance Sampling allows extracting these dependencies. However, a silicon verification is needed for validation.

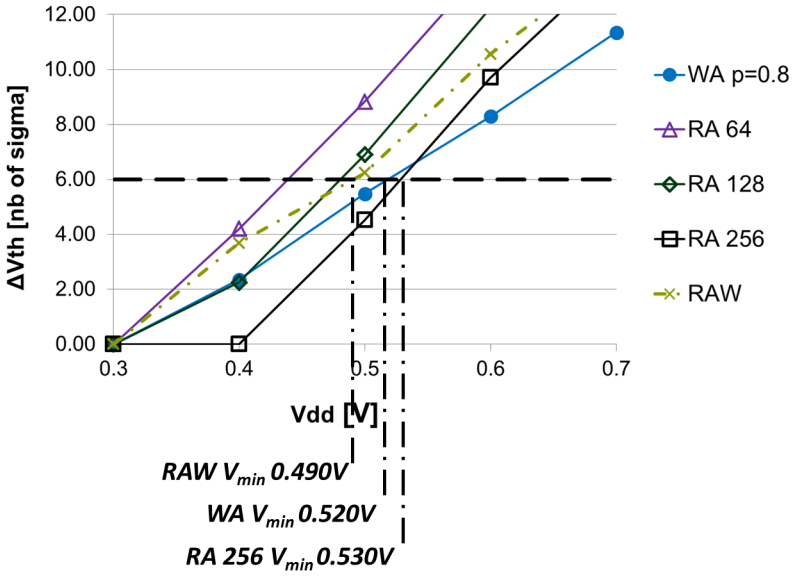

Figure 4.27: SPREGLV bitcell high-frequency WA, RA and RAW sigma-yields, at V_b=0V. RAW test leads to a lower V_{min} then WA test with p=0.8. With 256 rows, bitcell V_{min} becomes read limited.

4.4.2 SRAM Dynamic Failures on Silicon

A 140kB SPHD SPW macro has been tested by O.Thomas et al. using a dynamic characterization module that has been developed in the Berkley Wireless Research Center [114]. The proposed measurement methodology allows extracting the dynamic behavior of SRAM failures and therefore presents a good opportunity to validate the findings of the previous section. The presented measurements in this section are courtesy of O.Thomas.

Figure 4.28 presents RA measurement BER with respect to V_{dd} for various WL pulse widths, at V_b=0V. The shortest WL pulse leads to highest BER as expected, since it is associated to a minimum time available to build the bitline voltage difference required by the SA. For this particular chip, with a 5 μs WL pulse, the memory remains totally immune against RA failures down to 0.4 V V_{dd}. However, if a 2.5ns WL pulse is used, the memory becomes non functional below 0.75V V_{dd}.

Figure 4.29 presents RA measurement BER evolution with respect to V_b, at V_{dd}=0.5V and 5ns WL pulse. The BER decreases with the increase in the V_b, confirming that the failures are tied to the slow NMOS Pass-Gate device as observed in simulations, since the stronger V_b allows strengthening the NMOS device improving the RA yield.

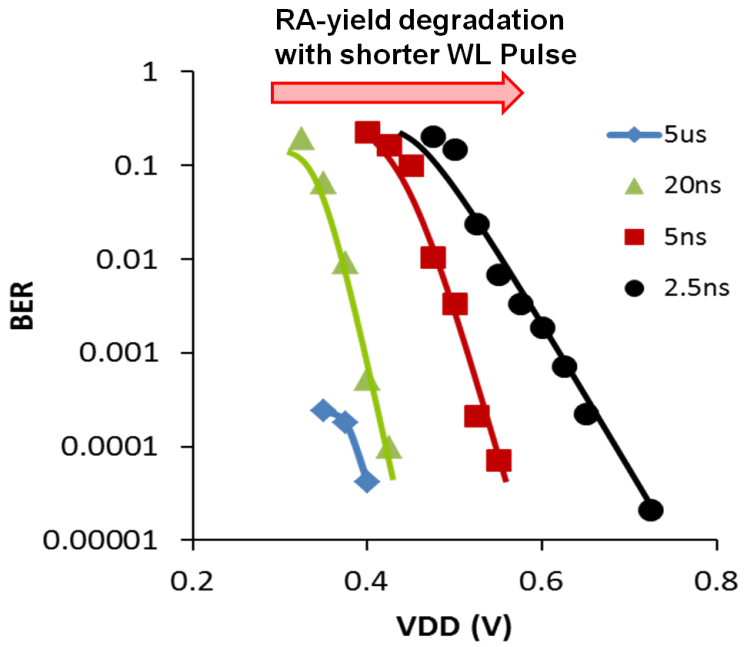

Figure 4.28: SPHD SPW SRAM macro RA measurement BER vs. V_{dd} for various WL pulse widths, at V_b=0V [114]. The RA yield degrades with shorter WL pulse.

Figure 4.30 presents WA measurement BER with respect to the V_b for various V_{dd}, with a 5μs WL pulse. The two different write failures mechanism are both observed. At $V_b < 0$, the strong NMOS and the weak PMOS place memory to the discharge failure zones, thus the BER increases with the decrease in V_b. At $V_b > 0$, the initially weak PMOS places the memory to the completion failures zones, thus the BER increase with the increase in V_b. Thanks to the SPW memory macro, which allows changing NMOS and PMOS flavors in opposite senses and control NMOS and PMOS strength ratio through V_b, both write failure mechanisms are observed confirming our observations on simulation. The results also show that the optimum balance between NMOS and PMOS is reached at V_b=0V.

Figure 4.31 presents RA,WA and Read-Stability (RS) V_{min} with respect to the WL pulse width, at V_b=0V. For this particular chip, V_{min} is limited by the RA failures for a WL pulse shorter than 10ns. With longer WL pulses, WA failures becomes dominant for V_{min} limitation, demonstrating the WL pulse width dependency of V_{min}. It is also shown that the RS V_{min} do not evolve with the presented WL pulse width range, which means that RS related failures occurs under longer WL pulses, confirming the suitability of this criteria for static-like tests. Another important point is that the RA and RS results are superposed for WL pulse width values 100ns and longer, which

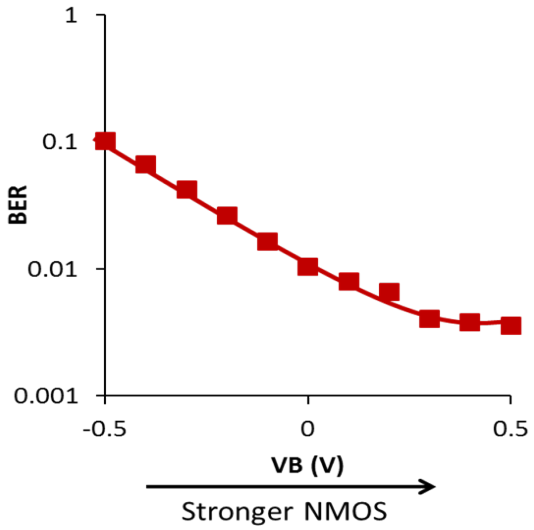

Figure 4.29: SPHD SPW SRAM macro RA measurement BER vs. V_b at V_{dd}=0.5V and 5ns WL pulse width [114]. The BER is improved with the increase in V_b (strengthen NMOS), confirming that the failure mechanism is tied to the slow Pass-Gate device as it is observed in simulations.

means that the RA failure mechanism switches to stability failures; this WL pulse width dependency of the RA failure mechanism has been already observed in simulations. The WA V_{min} keeps evolving in the given WL pulse width range, with a slope more and more insignificant (approaching nearly 0). This means that the long WL pulse width value which leads to a static-like write operation is not yet reached, but considering that the WA V_{min} is approaching a constant value, the static-like WL pulse width value is expected to be slightly longer than 10000ns for this particular chip.

The measurement results published in [114] have allowed to validate our findings on the bitcell failure mechanisms obtained through the simulation results that are performed using the Hypersphere MPFP search algorithm. However, silicon validation of an accurate V_{min} estimation can not be performed unlike the previously presented C40 static V_{min} modeling, since the required large set of statistical measurements are not yet available. Furthermore, presented measurements are performed on only few dies of the same wafer that is manufactured with a non industrialized process, so that the results are not affected by the spatial process variations, but only by random variations present in the die. As a consequence, these silicon results can not be fitted by current STMicroelectronics SPICE models. Nevertheless, the silicon results are sufficient to confirm the efficiency of the proposed modeling methodology to capture different failures mechanisms underlying behind the failures which limit the memory V_{min}, thereby the methodology can be

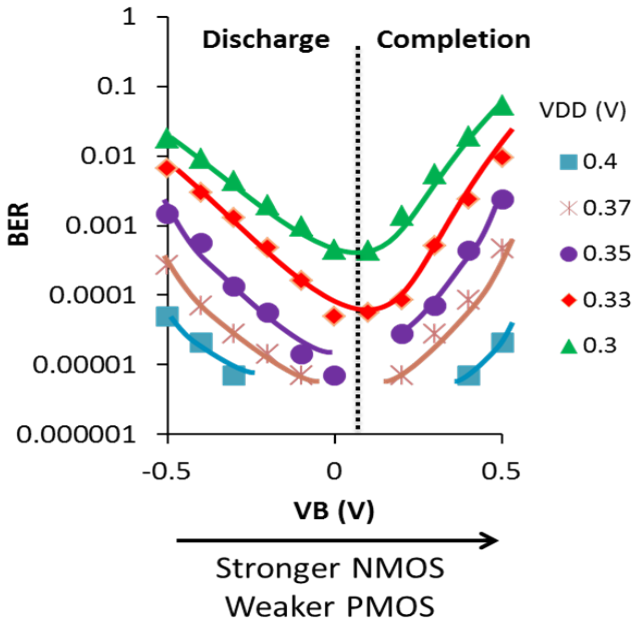

Figure 4.30: SPHD SPW SRAM macro WA measurement BER vs. V_b, at 5μs WL pulse width under various V_{dd} [114]. WA measurement BER evolves with respect to the V_b and exposes two different write failure mechanisms at $V_b < 0$ and at $V_b > 0$.

used for bitcell design analysis and optimization.

4.5 Application Example: Hypersphere MPFP Search for Investigations on SNM Yield Loss at High-Voltage in 28nm UTBB FD-SOI SRAM bitcells

It has been previously shown in figure 3.7 that the maturity of the SRAM SPICE model cards evolves with the maturity of the technology, which is possible only if a joint contribution from the process integration engineers and from memory designers is established. During the development of 28nm UTBB FD-SOI technology, it was found that all bitcells of STMicroelectronics SRAM portfolio show a drop in SNM-related yield at high-voltage memory operations. Figure 4.32 illustrates SNM mean/sigma ratio with respect to V_{dd} for SPREGLV bitcell at V_b=0V. Above V_{dd}=0.9V, SNM yield starts to drop significantly. If 6-σ threshold is chosen as the validation criteria for manufacturability, 6-σ yield cannot be held anymore above V_{dd}= 1.4V. It should be noted that SPREGLV bitcell is optimized for low-power applica-

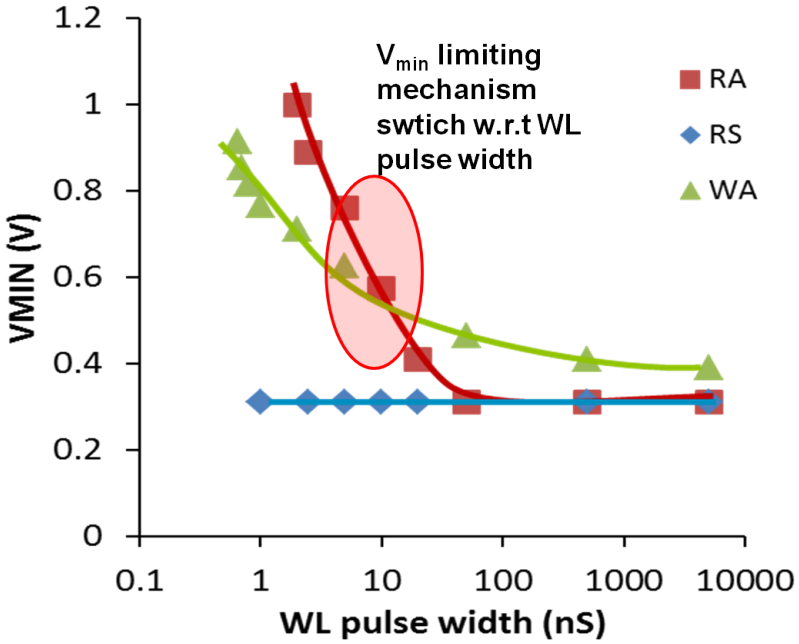

Figure 4.31: SPHD SPW SRAM macro RA, WA and Read-Stability (RS) V_{min} vs. WL pulse width, at V_b=0V [114]. It is shown that the limiting operation switches to write from read for WL pulse larger that 10 ns. RS yield do not evolve with the presented WL pulse width range, confirming the suitability of this criterion for static-like tests.

tions, thus it has the highest SNM compared to other bitcells of the SRAM portfolio, and other bitcells having less margin are susceptible to suffer more dramatic consequences.

The study on SNM-yield drop at high-voltage is carried out on using the Hypersphere MPFP search tool, which allows investigating mismatch mechanisms underlying behind bitcell failures. Figure 4.33 illustrates on the top, the half-cell storing high-logic level. The current flow mechanism leading to a content lost during a read operation (WL='1' and BLs='1') is illustrated by arrows, which is the same static read failure mechanism as described in section 3.6, in which the content lost is related mainly to the battle between a fast Pull-Down and a slow Pull-Up transistors. The bottom figure presents SNM MPFP MSVs at V_{dd} between 0.8V to 1.3V, with V_b held at 0V. At V_{dd}=0.8V and 0.9V, MSVs are formed by negatively skewed PDL and PGR and positively skewed PUL and PDR, confirming that the failure mechanism is tied to a battle between PMOS and NMOS. Between 0.8V and 0.9V V_{dd}, the absolute value of each skew forming the MPFP MSV is increasing showing the yield improvement. However, at V_{dd}=1V and above, a significant decrease in the PUL skew factor is observed. In

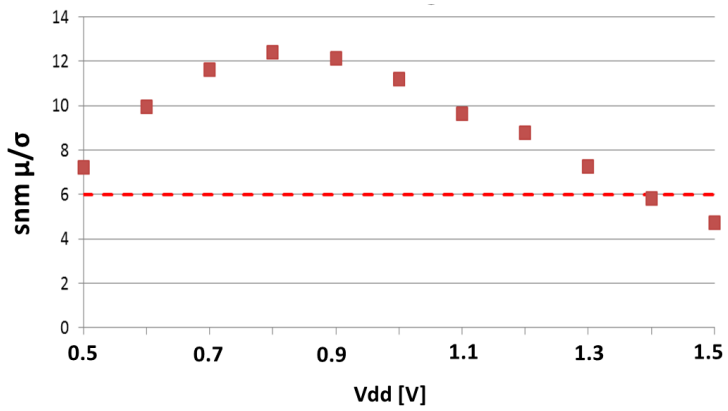

Figure 4.32: SPREGLV SPW bitcell SNM μ/σ with respect to V_{dd}, The yield drop occurs around V_{dd}=0.9V, and becomes dramatic at higher V_{dd}.

other words, at V_{dd}=1V and above, the norm of MSV decreases, meaning that the failure probability becomes higher, thus the SNM yield lowers. This behavior change is due to the fact that NMOS become too strong at this high V_{dd} value compared to PMOS, and the balance between PDL and PUL that has to be assured to keep the stored content is broken more easily compared to lower V_{dd} operations.

The body-biasing feature of UTBB FD-SOI can be a solution to work around the high-voltage SNM-yield drop issue. The body-biasing which allows efficiently adjusting the strength ratio between NMOS and PMOS, is used to keep the required strength balance against stability failures. Figure 4.34 illustrates the adaptive body-biasing (BB) use. The V_b which was initially 0V, is reduced to -1V above 1V V_{dd}, allowing the weakening of the too strong NMOS and the strengthening of the PMOS. Both WM and SNM yields are plotted for their worst-case corner and temperatures. The body voltage change at 1V V_{dd} and above allows shifting significantly SNM yield towards higher values, preventing the loss of the 6-σ margin within the required V_{dd} range. On the other hand, the weakened NMOS and strengthened PMOS lead to lower WM yield, but the decrease is not critical, since write operation is largely safe (very high yield) at high-voltage operations.

4.6 Smart Dynamic Back-Biasing Bitcell V_{min} Boost in UTBB FD-SOI

As demonstrated in previous chapters, a given bitcell V_{min} is limited by its read or write yield depending on the operating conditions, and also shown at many occasion that when the bitcell operates close to static conditions, the failures become more and more related to the broken balance between

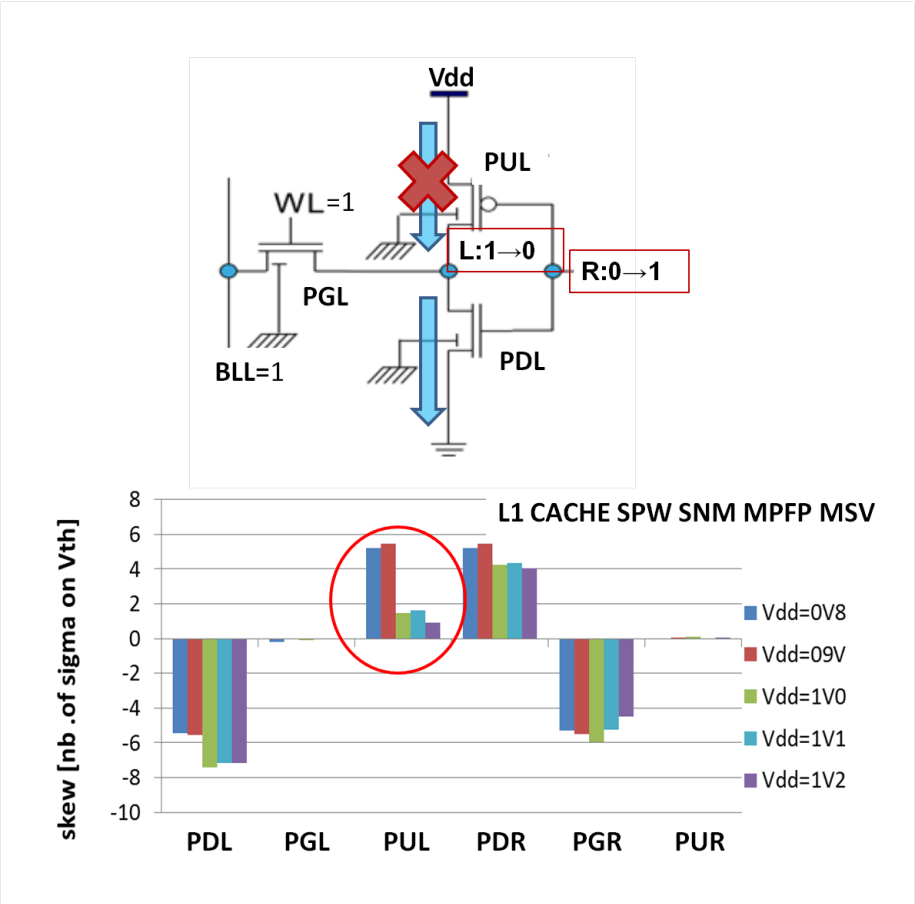

Figure 4.33: On the top, the current flow in the half-cell leading to a SNM failure and causing the loss of the high-logic level. On the bottom, SPREGLV SPW bitcell SNM MPFP MSVs at diferent V_{dd}. SNM MPFP MSV at V_{dd}=1V has a much smaller PUL skew component comapred to MPFP MSVs at lower V_{dd} values.

NMOS and PMOS devices. Last section shows that the use of body-biasing feature of UTBB FD-SOI technology can allow preventing SNM-yield loss at high-voltage operations. The use of body-biasing can be extended to improve write yield for write-limited bitcells. More general, an adaptive body-biasing can compensate the variability impact at different process-corners and temperatures, in order to improve memory V_{min}. This idea of smart body biasing for V_{min} improvement have brought along a patent application, as described on three bitcells: Single Port REGister File Low-Voltage (SPREGLV), Single Port REGister File (SPREG) and Single Port High Density (SPHD). Considering the conventional CMOS bulk nomenclature for transistor body voltages, i.e. *vdds* for PMOS body and *gnds* for

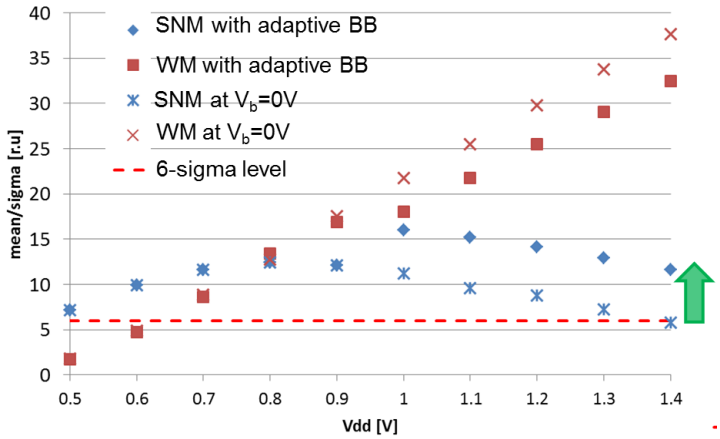

Figure 4.34: SPREGLV SNM and WM simulations μ/σ ratio at worse case corners and temperatures (FS 125 °C for read, SF -40 °C for write), with V_b tied to 0V (crosses) and with adaptive V_b (squares) in which V_b is decreased to -1V above V_{dd}=1V.

NMOS body, a Single-P-Well architecture requires

$$vdds{=}gnds{=}\ V_b$$

The analysis is carried out on as the following: Monte Carlo simulations are performed for each bitcell with different body-biasing values. SNM and WM $\frac{\mu}{\sigma}$ ratios are extracted. Only worst-case process corners and temperature conditions of SNM and WM are simulated (SF,FS,-40 °C and 125 °C). For this analysis, the so-called V_{min} represents simply the V_{dd} voltage at which $\frac{\mu}{\sigma}$ ratio is equal to 6. First, the lowest V_{min} that can be reached among different biasing conditions is determined as the "Blind Usage" V_{min}. Secondly, the V_{min} analysis is performed separately at each PVT, and the lowest V_{min} of each PVT is determined as the "Smart Usage" V_{min}.

Figure 4.35 presents SPREGLV blind usage V_{min} at various V_b at different PVT's, which yields a the minimum V_{min} 0.565V when V_b is tied to V_{dd}. Figure 4.36 presents for the same bitcell, SNM and WM V_{min} that can be obtained using smart biasing. The V_{min} can be reduced to 0.489V, which is obtained at SF corner and -40 °C, with V_b=1.2V. It is worth to say that 0.489V represents maximum of separately obtained V_{min} at each PVT using smart biasing. For example, the V_{min} at FS 125 °C is 0.448V with V_b=0V, but this value can not be stated as the overall V_{min}, since the minimum V_{min} that can be reached at SF -40 °C with all possible biasing values, is higher.

Figure 4.37 and figure 4.39 present SPREG and SPHD V_{min} with blind usage of the body-biasing, for which 0.591V and 0.691V V_{min} values are

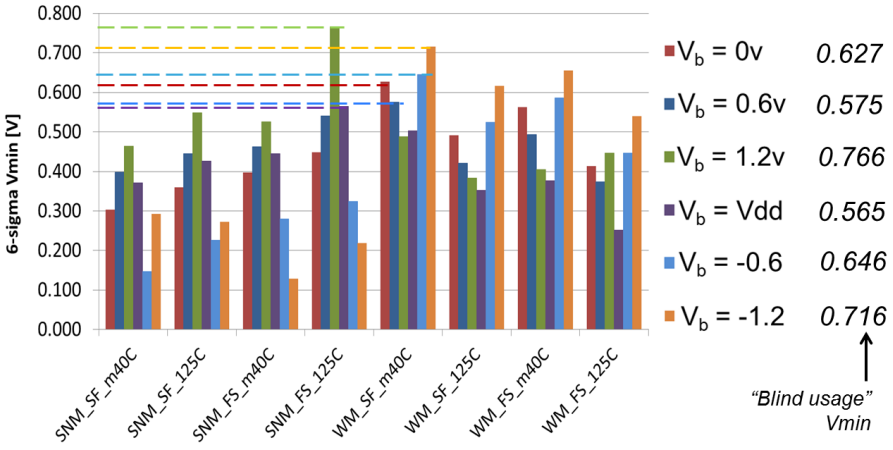

Figure 4.35: SPREGLV SNM and WM 6-σ yields at worse-case corners and temperature with blind usage of body-biasing. The minimum V_{min} that can be reached is 0.565V.

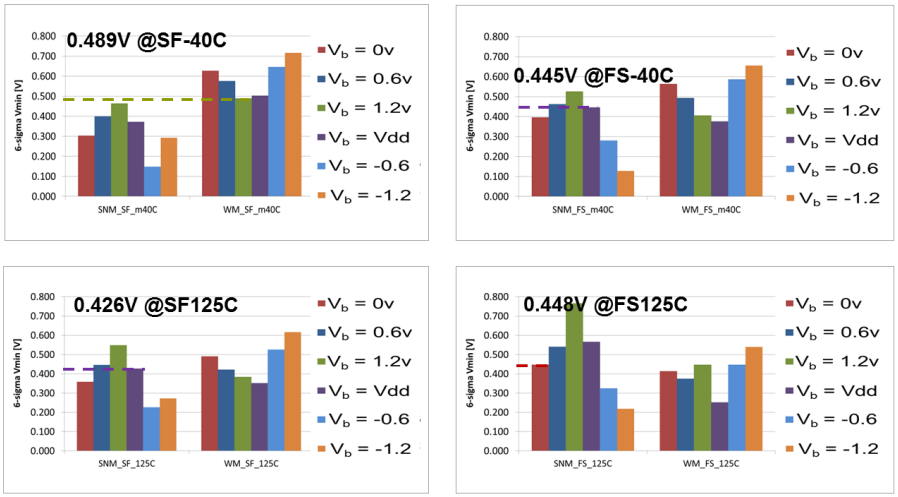

Figure 4.36: SPREGLV SNM and WM 6-σ yields at worse-case corners and temperature with smart usage of body-biasing. 0.489V V_{min} can be reached.

determined, respectively. Figure 4.38 and figure 4.40 present V_{min} simulation results with smart biasing for same bitcells. It is shown that the smart

biasing allows reaching 0.474V V_{min} for SPREG bitcell and 0.550V V_{min} for SPHD bitcell.

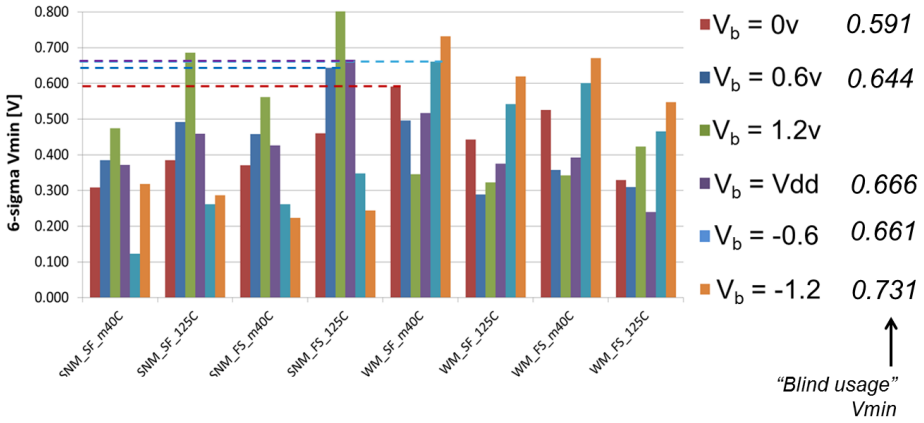

Figure 4.37: SPREG SNM and WM 6-σ yields at worse-case corners and temperature with blind usage of body-biasing. The minimum V_{min} that can be reached is 0.591V.

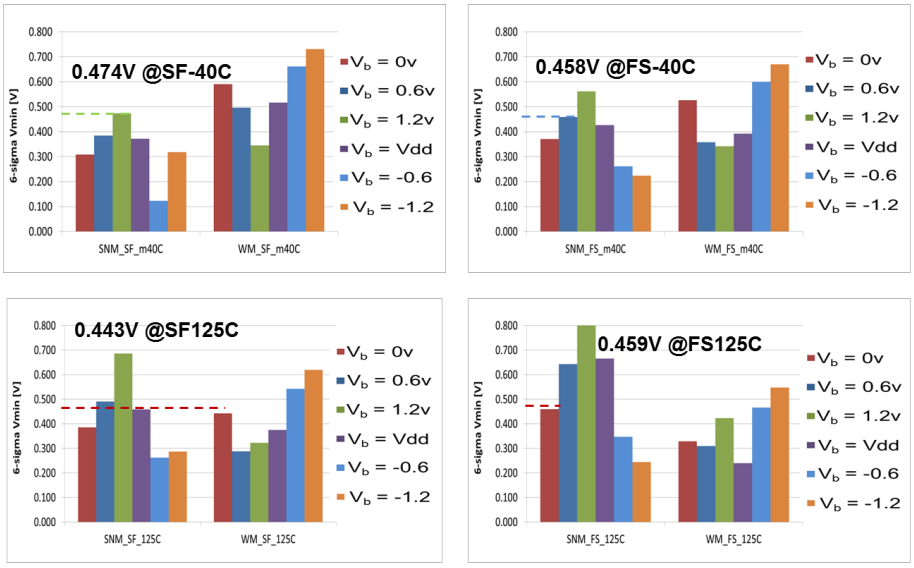

Figure 4.38: SPREG SNM and WM 6-σ yields at worse-case corners and temperature with smart usage of body-biasing. 0.474V V_{min} can be reached.

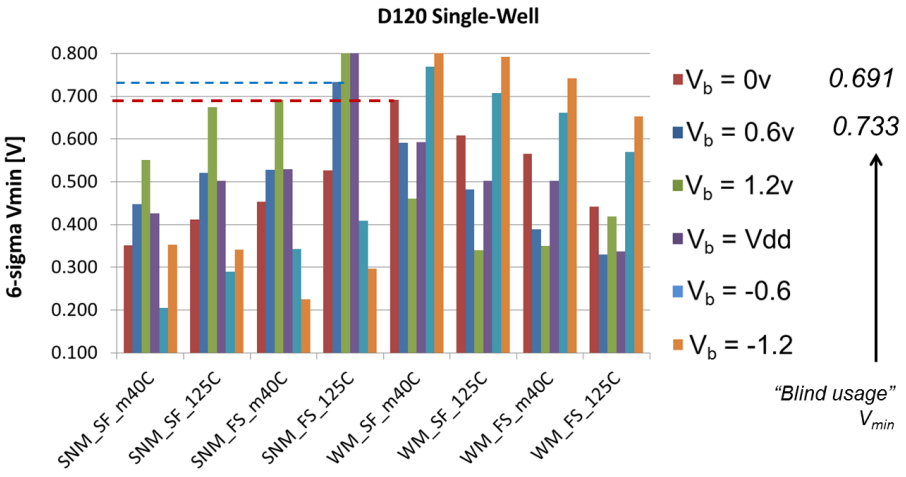

Figure 4.39: SPHD SNM and WM 6-σ yields at worse-case corners and temperature with blind usage of body-biasing. The minimum V_{min} that can be reached is 0.691V.

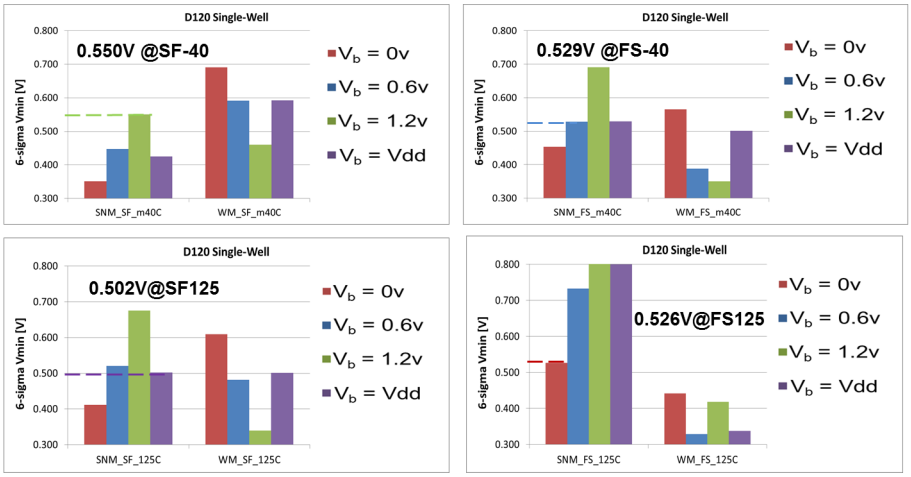

Figure 4.40: SPHD SNM and WM 6-σ yields at worse-case corners and temperature with smart usage of body-biasing. 0.550V V_{min} can be reached.

Table 4.6 presents for each bitcell, the V_{min} gain that can be obtained using smart body-biasing with respect to the blind use of body biasing. The gain is more important when the bitcell area is smaller, or in other words, the smart biasing is more efficient when the bitcell has a higher blind use V_{min}. The highest V_{min} gain is simulated as 140mV for SPHD bitcell, which

is a very large improvement for the most critical bitcell.

Bitcell (Area μm^2)	V_{min} with Blind Biasing [V]	V_{min} with Smart Biasing [V]	Gain [V]
SPREGLV (0.197)	565	489	86
SPREG (0.152)	591	474	117
SPHD (0.120)	691	550	140

Table 4.2: Smart Body-Biasing V_{min} gain vs. Bitcell Area

4.7 Conclusion

In this chapter, the SRAM bitcell V_{min} limitation due to the process variability is studied from different aspects, with the ultimate goal of proposing accurately enough SPICE-level V_{min} modeling methodologies for efficient design optimization. First, a Monte Carlo method based static V_{min} modeling that uses the full variability models that are integrated in SPICE model cards is presented. The modeling methodology is validated through C40 silicon measurements showing the good accuracy in modeling the worst-case silicon V_{min}. Static V_{min} modeling methodology is also applied on a Ultra-Low-Voltage bitcell which has an alternate 10-transistor architecture. A refinement of the method is to use "Large-Sigma" estimation, which, as demonstrated on C45 silicon, models distribution tails with a much higher accuracy and fits better data for large memory cuts and for a industrial-level process variability. The good agreement between silicon measurements and modeling results shows that the proposed methodology is independent of the bitcell architecture and the operating conditions, and the only requirement is to have sufficiently accurate SPICE model cards.

A smart algorithm that is based on a Most Probable Failure Point (MPFP) search on hyperspherical surfaces and the Importance Sampling, is developed to overcome analysis complexity that appears in the advanced technology nodes, and are mainly related to use of the basic Monte Carlo method. The key assumption is that the random variability leading to mismatch in identically designed neighbor transistors is modeled through a normally distributed threshold voltage variation. Using only this variability source, the variability of the real-world SRAM bitcells is represented by a multi-dimensional variability space centered on the nominal bitcell design. The Hyperpshere Most Probable Failure Point search algorithm is first used for the UTBB FD-SOI SRAM bitcells dynamic V_{min} investigations. The findings about the different failure mechanism and their dependency on the operating conditions are also validated by silicon measurements that has been performed on a SRAM macro as part of an another work.

The knowledge acquired during the SRAM failure mechanisms investigations has served the purpose of bitcell usage optimization during the 28nm UTBB FD-SOI technology development in which the Hypersphere MPFP search algorithm is used as an investigation tool. In UTBB FD-SOI technology, the use of the body-biasing in the single-well architecture bitcells is presented as an efficient solution to overcome V_{min} limitations, since it allows to adjust NMOS-PMOS strength ratio using their common body voltage. It is demonstrated that the smart use of body-biasing, i.e. dynamically adapting the biasing value with respect to operating conditions, can lead to a significant V_{min} gain. A patent for this invention has been filled during the thesis work.

Chapter 5

Random Telegraph Signal Noise in 28nm UTBB FD-SOI and the impact on 6T SRAM

In the previous chapter, it is shown that SRAM circuits, which are designed with the most aggressive design rules in a given technology node, are very vulnerable to static variability. Besides the static variability which is tied to non-time dependent spatial dispersions, the variability in semiconductor devices may also originate from dynamic shift of the device electrical characteristics [29] over its life-time depending on its operating conditions. This dynamic variability phenomenon introduces new challenges for circuit designers, since it might have dramatic consequences on circuit performance. In this chapter, the impact of the Random Telegraph Signal Noise, which is seen as an important dynamic variability concern in advanced technology nodes, is studied. The investigation are carried out on UTBB FD-SOI SRAM 6T Bitcells. First, the RTS noise analytical model peculiar to UTBB FD-SOI technology taking into account the front- and back-gate coupling, is presented. The analytical model is then used to generate RTS-aware SRAM bitcell SPICE netlist. Finally, Monte Carlo simulation results are compared with silicon measurements for validation of the proposed RTS noise model.

5.1 Time-Dependent Random Telegraph Signal Noise Variability

Random Telegraph Signal (RTS) [28] noise arises from trapping and de-trapping of charge carriers (electrons or holes) by defects (also names "traps") located in the silicon/dielectric interface and in the dielectric it-

self. The traps in the gate dielectric may originate from oxidation-induced defects, metal impurities and different kinds of bond breaking processes, caused by radiation, hot carrier stress or other phenomena tied to NBTI [25]. Capture and emission of a charge carrier by a trap results in the discrete modulation of the channel conductance, which causes the channel current to be similar to a random telegraph signal [115]. The discrete levels of a two-level RTS correspond to high- and low-conductivity in the channel. In particular, an empty trap corresponds to a high-current level, whereas a filled trap corresponds to a low-current level in the channel. A two-state drain current fluctuation caused by a single trap is illustrated in figure 5.1.

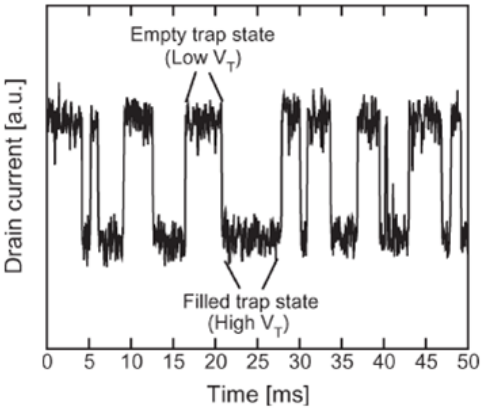

Figure 5.1: Two-state drain current fluctuation caused by RTS noise generated by a single trap. The low current level corresponds to a filled-trap state, whereas the high current level is associated with an empty-trap state [116].

The impact of trapping/detrapping has been first observed as a concern in analog circuits in form of low frequency noise [28,117,118]. In small MOS devices with a low number of free carriers, LF noise performance is dominated by RTS noise on top of the ever present bulk noise [119, 120]. With the development of ultra-deep submicron CMOS technologies, the RTS noise leads to large current fluctuation which might have dramatic consequences on circuit operations. Theoretically, the RTS noise amplitude scales with the inverse of the channel area ($W.L$) and the oxide thickness t_{ox}. This means that RTS noise amplitude increase trend should follow device scaling, since L,W and t_{ox} are in principal scaled by same factor [121]. However, the stagnation of t_{ox} scaling in advanced technology nodes due to the increased gate leakages currents, has aggravated the increase in RTS amplitude. The relation between RTS amplitude and the channel geometry is due to the fact that RTS noise is tied to the fluctuations in the number of free carriers in the channel, and the latter is proportional to the channel area [122]. In highly-scaled technologies, RTS noise amplitude is exacerbated by the discreteness

of dopants in the channel, since the resulting percolation paths through the valleys in the potential landscape dominating the current flow are strongly influenced by the occupation of traps in the oxide located within proximity of these paths [123].

As it will be shown in this chapter, besides the amplitude of RTS noise, the influence of traps on the performance of a semiconductor device is determined by the density of traps and the probability that these traps are occupied by a charge carrier. A trap can change its occupancy by either capturing or emitting a charge carrier. The Shockley-Read-Hall (SRH) theory [124] originally describing generation-recombination of bulk states has been adopted to describe the trapping-detrapping behavior of the traps located at the silicon/dielectric interface. For the traps located in the oxide, the mechanism is a two step process which involves capture and then tunneling through the oxide. The physical characteristics of trap will be detailed later in the next section.

RTS noise is seen as an important dynamic variability concern in advanced technology nodes [125–129], and is projected to be a significant source of transistor variability affecting the yield of highly scaled SRAM cell [130], since SRAM bitcells use small channel area transistors and they are highly vulnerable against mismatch. The contribution of RTS noise in SRAM design margins across different technology nodes is illustrated in figure 5.2 showing the growing impact of RTS noise. Therefore, study of RTS noise in highly-scaled SRAM has been very popular research topic in the last decade. The authors of [130] show that in 45nm technology node, the V_{th} shift caused by RTS noise is at same order as the one caused by random dopant fluctuations (RDF) and can even exceed RDF at some rare cases. In their work, the measurement data obtained from individual transistors are used to predict the impact of V_{th} shift caused by RTS noise on highly-scaled SRAM static margins using analytical models. The authors in [131] demonstrated the impact of RTS on 64kB SRAM array V_{min} through static measurements and quantified the impact as around 50mV in the 45nm technology node. In [132], the transistor-level modeling of RTS noise impact with multiple-traps for SPICE simulations is presented and the proposed approach is applied on SRAM static margins simulations. However, physical characteristics and dynamics of traps are extracted from measurements performed on individual transistors, which limits the accuracy of bias-dependency of their model. In [133], authors investigate large-signal bias and temperature dependencies of the RTS impact on SRAM through measurements. A considerable conclusion of their work is that the impact of RTS noise on SRAM stability differs with respect to the number of successive operations.

All findings in the literature point that the RTS noise represents high risk for stability and performance metrics of highly-scaled SRAM bitcells. Therefore, the dynamic RTS noise variability has to be considered in the

SRAM design optimization. This raises the need for an efficient SPICE-level RTS characterization tool that can be used in time domain analysis with industrial EDA softwares.

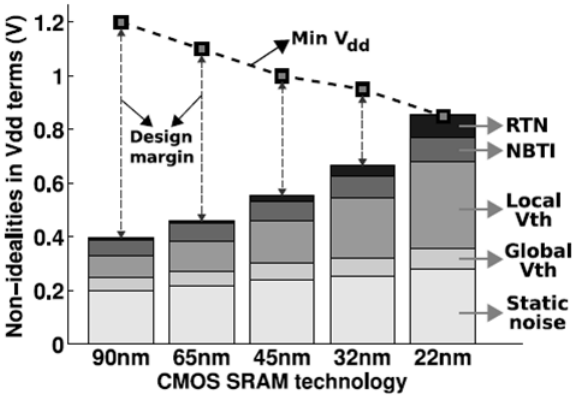

Figure 5.2: Contribution of RTS Noise on SRAM design margins vs. technology nodes [134].

5.2 SPICE-level RTS Noise Modeling in UTBB FD-SOI

5.2.1 RTS Trap Characteristics and Particularity of UTBB FD-SOI

The trapping and de-trapping of charge carriers in CMOS devices is observed as an increase or decrease of the channel conductivity leading to I_{ds} fluctuations. Considering an N-channel MOSFET in which the majority carriers are electrons, the capture of one electron by a trap located in the gate dielectric causes the loss of a majority carrier in the channel modulating the intrinsic transistors parameters which leads as a consequence to a decrease of the I_{ds} current. On the other hand, the subsequent emission of the electron will lead to an increase in the I_{ds} current. The same mechanism is valid for a PMOS device, in which majority charge carriers are holes. The trapping/de-trapping event caused by a single trap in a N-channel bulk CMOS is illustrated in figure 5.3. Using the frame shown in figure 5.3, a particular trap in the gate dielectric is characterized by its distance from the silicon/dielectric interface x_t, its energy level in the silicon band gap, its capture cross section σ, its activation energy E_a.

The relative amount of fluctuation on the current flowing through the transistor channel due to one single trap at the interface has been observed

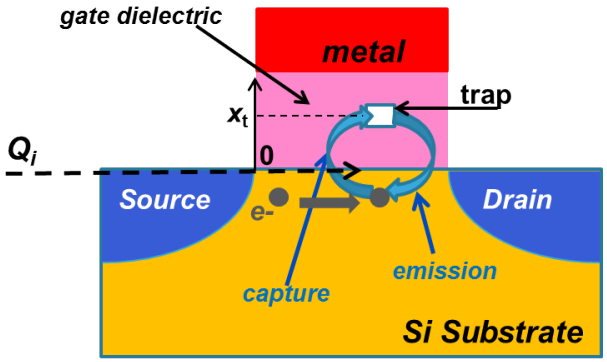

Figure 5.3: Schematic diagram of RTS noise by a single trap in bulk N-Channel CMOS device. RTS is attributed to trapping/detrapping events caused by defects near the silicon - oxide interface. High and low states correspond to carrier capture and emission.

in [135] to be described as:

$$\frac{\delta I_{ds}}{I_{ds}} = \frac{g_m}{I_{ds}} \frac{-q}{W_{eff} L_{eff} C_{ox}} \tag{5.1}$$

where g_m is the transconductance in AV^{-1}, I_{ds} is the source-to-drain current in A, W_{eff} and L_{eff} are the effective transistor dimensions in m, C_{ox} is the gate dielectric capacitance per area in $\mathrm{F/m}^2$, and q is the elementary charge given by $1.602\mathrm{x}10^{-19}\mathrm{C}$.

An equivalent way to consider the effect of a single trap is to think in terms of threshold voltage fluctuation. Considering that the voltage fluctuation produced by a charge trapped at the silicon/dielectric interface as a fluctuation in the device threshold voltage V_{th}, from [27] the one can write

$$\delta V_{th} = \delta V_{fb} = \frac{-q}{W_{eff} L_{eff} C_{ox}}$$

where V_{fb} is the flatband voltage. Recalling the definition

$$g_m = \frac{\delta V_g}{\delta I_{ds}}$$

and using the simple approximation $\delta V_{fb} = \delta V_g$ the equation (5.1) becomes

$$\delta V_{th} = \delta V_g = \frac{\delta I_{ds}}{g_m} = \frac{-q}{W_{eff} L_{eff} C_{ox}} \tag{5.2}$$

representing the equivalent input gate voltage fluctuation of the channel that is caused by one single trap. In transistor-level modeling, it is more suitable to work with δV_g rather than δI_{ds}, since gate voltage is one of the

inputs of 4-terminals SPICE transistor model cards. From equation (5.2), according to [28], we can obtain the analytical expression of the threshold voltage fluctuation as a function of the location of the trap in the oxide:

$$\delta V_{th} = \frac{q}{W_{eff} L_{eff} C_{ox}} \left(1 - \frac{x_t}{t_{ox}}\right) \tag{5.3}$$

where $0 \leq x_t \leq tox$. x_t is the location of trap indicating that how deep it is in the oxide thickness as illustrated in figure 5.3, and t_{ox} is the oxide thickness. Considering the device of figure 5.3, a filled trap (captured electron) will increase the device V_{th} by δV_{th}, whereas the subsequent emission of the electron will decrease V_{th} by δV_{th}. From equation (5.3), a trap close to the silicon interface will cause V_{th} fluctuations with larger amplitude than those caused by a trap in the depth of the dielectric.

The cross section of UTBB FD-SOI transistor is illustrated in figure 5.4. The body of an UTBB FD-SOI transistor acts as a second gate, placed below the buried oxide (BOX) under the thin silicon film, and participates substantially to the control of silicon film charge. In the detail, the excellent body factor in UTBB-FDSOI technology allows dynamically changing the device flavor through the use of body-biasing techniques [62] [50]. In other words, a UTBB FD-SOI transistor is formed of two gates: front and back gates (FG and BG) and has as a consequence two sources of RTS noise in distinct from bulk transistors : the front gate dielectric and the BOX at the back-gate. As a result, the trapping/detrapping events occur at both interfaces.

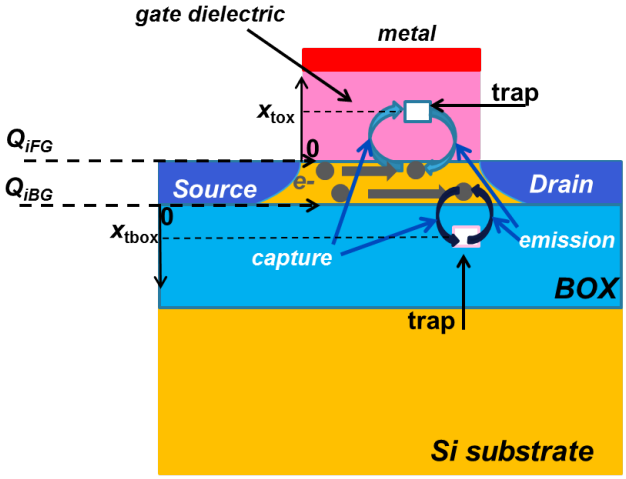

Figure 5.4: Schematic diagram of RTS in N-Channel UTBB FD-SOI device. Trapping/detrapping events occur at the silicon/gate dielectric interface and at the silicon/BOX interface.

The average number of traps in a given dielectric layer, N_{avg}, is related

not only to the dielectric geometry, i.e. to its width W, to its length L and to its thickness t_{ox}, but also to the oxide quality, which is quantified by its average volume trap density N_T [136]:

$$N_{avg} = N_T W_{eff} L_{eff} t_{ox} \left(E_{max} - E_{min}\right) \tag{5.4}$$

where $(E_{max} - E_{min})$ represents the active energy of the silicon band gap (1.11 eV at 300K). Figure 5.5 presents the cumulative distributions of the number of traps in 28nm UTBB FD-SOI 6T SPHD SRAM Bitcell NMOS Pull-Down (PU) and Pull-Up (PU) transistors. It is assumed that number of traps in a given dielectric follows Poisson distribution with the mean N_{avg} calculated using equation (5.4). The average trap densities at FG and BG dielectrics, N_{Tox} and N_{Tbox}, are chosen as $10^{18} cm^{-3} eV^{-1}$ and $10^{17} cm^{-3} eV^{-1}$ respectively [137], indicating that BOX has a better oxide quality than the FG gate dielectric, which can be explained by the use of thermally grown SiO_2 in BOX, instead of the complex stack of high-K dielectrics in the FG. More traps are expected in NMOS PD transistors, since PMOS PU are smaller (W=51 nm L=40n) then NMOS PD (W=106 nm L=40nm). It has to be noted that the larger dimensions of BOX (t_{BOX}=25nm vs. t_{ox}=1.7nm) is compensated by the improved oxide quality. However, even with its better oxide quality, BOX is prone to have more traps than the FG dielectric.

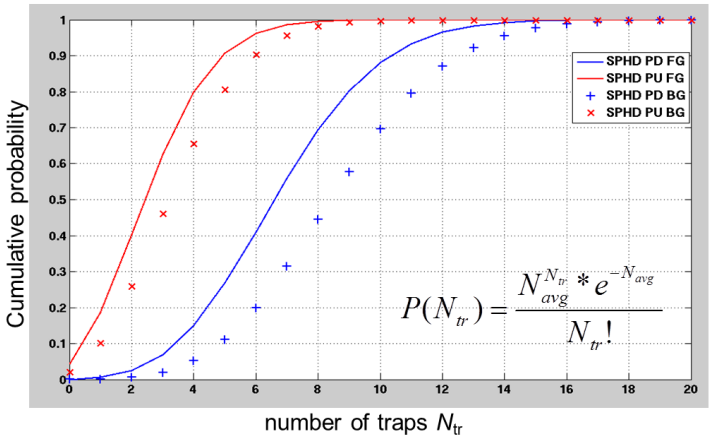

Figure 5.5: Cumulative probability distribution of number of traps in 28nm UTBB FD-SOI SPHD 6T SRAM Bitcell NMOS Pull-Down (PD) and PMOS Pull-Up(PU) transistors, at both FG and BG dielectrics. It is assumed that the number of traps in a given dielectric follows Poisson distribution.

Assuming that traps are uniformly distributed in a given dielectric [29], figure 5.6 presents cumulative distribution of δV_{th} for FG dielectric traps in PD and PU devices. Fluctuation amplitudes are larger in the PMOS due

to smaller geometry of the dielectric (equation (5.3)). It is worth to say that the overall impact of many traps in a given device is cumulative and indicated in this work as ΔV_{th}. Equation (5.3) can be applied to BG traps, if t_{BOX} and C_{BOX} are used instead of t_{ox} and C_{ox}.

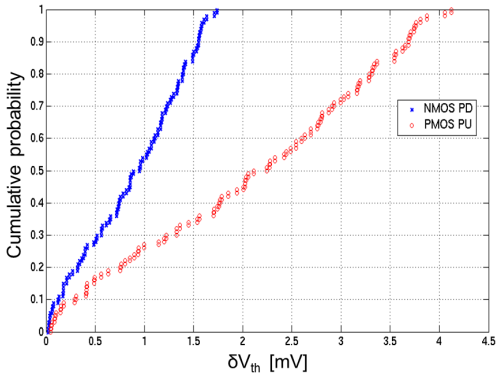

Figure 5.6: Cumulative probability distribution of δV_{th} of FG dielectric traps in 28nm UTBB FD-SOI SPHD 6T SRAM Bitcell NMOS Pull-Down (PD) and PMOS Pull-Up(PU) transistors. It is assumed that traps are uniformly distributed in the gate dielectric.

In order to get the dynamical picture of RTS, one should be able to calculate how a trap state evolves between filled and empty during device operations. The trap occupancy evolves as a stochastic process over time with capture τ_c and emission τ_e times following exponentially falling distribution [29]. Using the SRH [124] recombination model, the average capture and emission times satisfy the following relations :

$$< \tau_c > \quad = \quad \frac{q}{f_e \sigma Q_i} \tag{5.5}$$

$$< \tau_e > \quad = \quad < \tau_c > \frac{Q_i}{Q_{itrap}} \exp\left(\frac{x_t F_{ox}}{k_B T}\right) \tag{5.6}$$

where k_B is the Boltzmann constant expressed in eV/K, T is the absolute temperature in Kelvin, q is the elementary charge, f_e is the (constant) escape frequency for tunneling into traps of the dielectric, F_{ox} is the electric field in the dielectric. The cross section σ of a given trap at ambient temperature can be estimated as

$$\sigma = \sigma_0 . \exp\left(\frac{-x_t}{\lambda}\right)$$

where σ_0 is the cross-section pre-factor [138] and λ is the tunneling attenuation distance in the dielectric. Q_{itrap} is the inversion charge at 50% trap

occupancy ($< \tau_c >=< \tau_e >$), which is related to the position of the trap in the energy band with respect the valance band, and finally Q_i is the charge inversion at the silicon/dielectric interface as illustrated in figure 5.3 a bulk device. In the UTBB FD-SOI transistor, charge inversion occurs at both silicon/dielectric interfaces as illustrated in figure 5.4. Hence, equations (5.5) and (5.6) should be duplicated and written for each interface. In particular, Q_i will become Q_{iFG} for silicon/front-gate dielectric interface and Q_{iBG} for silicon/BOX interface. The same duplication is then also applied for F_{ox} giving us F_{oxFG} and F_{oxBG}. Therefore, an accurate modeling of RTS noise average time constants implies accurate modeling of the charge inversion at both gates of a UTBB FD-SOI transistor.

5.2.2 Front- and Back-gate Coupling Aware 2-Dirac Charge Inversion model

As shown in [139], the inversion charge Q_i at each interface placed at x_i can be modeled as a Dirac's delta function, whose amplitude depends on the surface potential V_s at that interface. The dynamic behavior of RTS comes from the fact that V_s evolves with the applied gate and body bias, which evolve in time. Considering an SRAM bitcell transistor which can experience large and very fast bias swings within a trapping cycle of a given trap, the modeling of V_s becomes crucial for an accurate RTS Noise time constants modeling. In detail, Q_i at a given interface placed at x_i reads

$$Qi = qDOS \exp\left(\frac{V_s}{k_B T}\right)\delta\left(x - x_i\right) \tag{5.7}$$

where DOS is the effective density of states emulating the silicon film at each silicon/dielectric interface, and determined by

$$DOS = n_i \frac{t_{si}}{2}$$

where n_i is the intrinsic carrier concentration in silicon (1.45^{10} cm^{-3}) and t_{si} is the thickness of the silicon film (7nm in 28nm UTBB FD-SOI [61]). The model is named as *2-Dirac* with respect to the shape of the charge distribution between both silicon/dielectric interfaces.

Figure 5.7 illustrates the capacitive network between FG and BG of an UTBB FD-SOI device [62] [50]. The standard equation of charge coupling in a SOI MOSFET [140] is adapted to UTBB FD-SOI as:

$$\begin{aligned} Q_i(V_{sFG}) &= C_{Si}(V_{sBG} - V_{sFG}) + C_{OX}(V_g - V_{sFG}) \\ Q_i(V_{sBG}) &= C_{Si}(V_{sFG} - V_{sBG}) + C_{BOX}(V_b - V_{sBG}) \end{aligned}$$

in which flat band voltages are omitted for simplicity. Recalling equation (5.7), the system of equation can be written as:

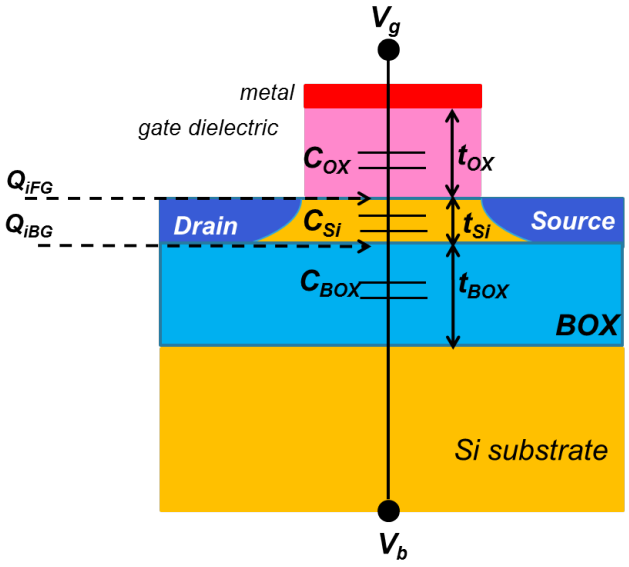

Figure 5.7: Schematic diagram of the capacitive network between the front-gate and the back-gate of an UTBB FD-SOI transisotr.

$$\begin{aligned} Q_i(V_{sFG}) &= qDOS \exp\left(\frac{V_{sFG}}{k_B T}\right)\delta\left(x - x_i\right) \\ &= C_{Si}(V_{sBG} - V_{sFG}) + C_{OX}(V_g - V_{sFG}) \qquad (5.8) \\ Q_i(V_{sBG}) &= qDOS \exp\left(\frac{V_{sBG}}{k_B T}\right)\delta\left(x - x_i\right) \\ &= C_{Si}(V_{sFG} - V_{sBG}) + C_{BOX}(V_b - V_{sBG}) \qquad (5.9) \end{aligned}$$

Using the non-linear system formed by equations (5.8) and (5.9), we can solve V_s from 1D models of FG and BG voltages (V_g and V_b). For a given solution of V_{sBG} and V_{sFG}, F_{OX} at FG and BG then can be modeled as:

$$\begin{aligned} F_{OXFG} &= \frac{V_g - V_{sFG}}{t_{OX}} \\ F_{OXBG} &= \frac{V_b - V_{sBG}}{t_{BOX}} \end{aligned}$$

The analytical *2-Dirac* model is implemented in both Mathcad [141] and MATLAB [142] softwares using their integrated system solver functions, assuming 0V as initial values of V_{sFG} and V_{sBG}. In order to validate the accuracy of the simplified surface potential modeling, TCAD simulations using more complex electrostatic equations are performed with FlexPDE

[143]. Figure 5.8 presents results for Q_i, τ_c and F_{OX} at both FG and BG for a NMOS FET , showing the good agreement of the proposed model with TCAD simulations. We can therefore conclude that the proposed model has a sufficient accuracy at all tested operating conditions. The results shows for both gates that the increase in Q_i will decrease τ_c, which means that the traps are more probably filled (smaller τ_c and larger τ_e) at strong channel inversion. Trapping at each gate is also accelerated by the strong inversion in the other gate. The larger thickness (25 nm vs. 1.7 nm [61]) and the better quality of BG oxide result in few decades slower τ_c compared to FG.

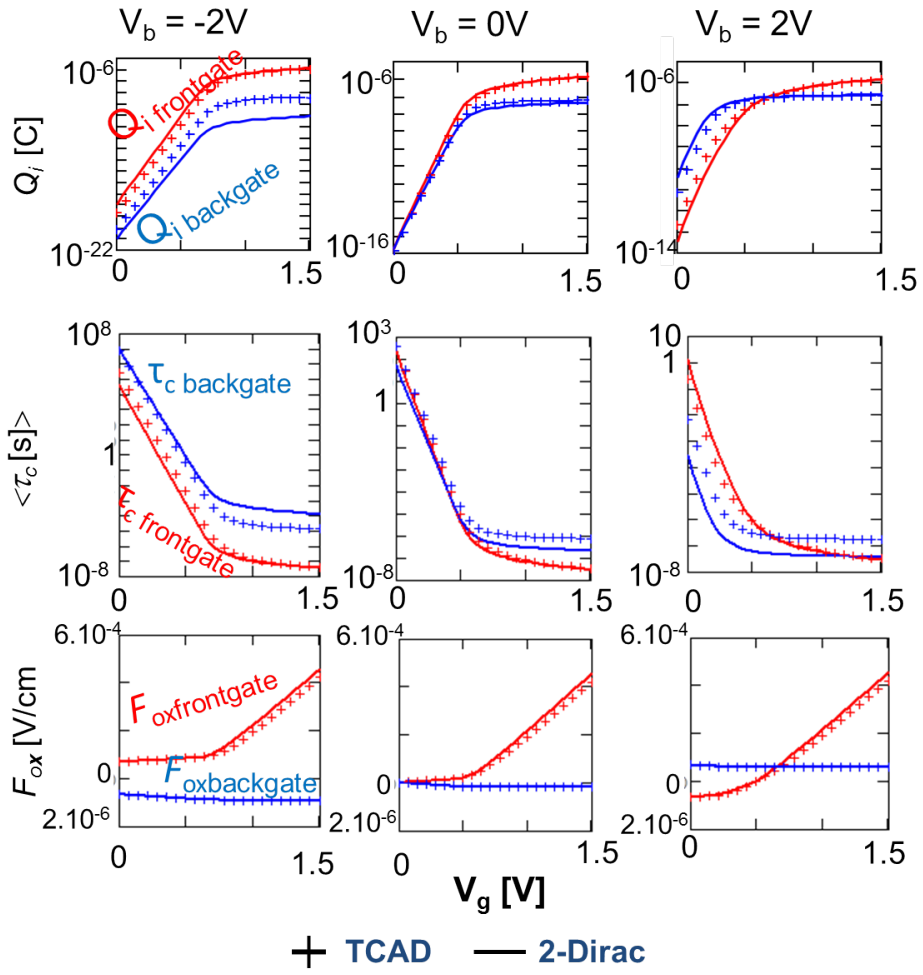

Figure 5.8: TCAD simulations (crosses) vs. 2-Dirac model (straight lines) for Q_i, τ_c and F_{OX}.

The *2-Dirac* model is used to calculate the fundamental trap characteristics and their average time constants, as they evolve under different bias conditions. Figure 5.9 illustrates $< \tau_c >$ and $< \tau_e >$ evolution of a given

trap in a NMOS FG dielectric with respect to V_g, at different V_b values, at ambient temperature, for two different trap energy levels. The results tell that, as already observed in figure 5.8, traps have more tendencies to be filled (captured) at strong inversion in both gate sides. Considering that the average period between two successive capture and emission events is simply

$$< T >=< \tau_c > + < \tau_e >$$

the one can define define the trap frequency F_T as the inverse of the period:

$$F_T = \frac{1}{< T >}.$$

A trap with higher energy has shorter $< \tau_e >$, for the same $< \tau_c >$, which as a result increases F_T. If F_T is higher than, or same order as, the memory operating frequency, the transistor mismatch may evolve during a memory operation or between two successive operations which may lead to dramatic consequences in SRAMs. The impact of V_b is more significant on $< \tau_c >$, until some value above which $< \tau_c >$ remains quasi-stable independent of the value of the V_b.

Figure 5.10 illustrates $< \tau_c >$ and $< \tau_e >$ evolution of a given trap in a NMOS BG dielectric with respect to V_b, at different V_g values, at ambient temperature, for two different trap energy levels. It is shown that BG traps average time constants are mainly set by the applied V_g, leading to a 7 decades shorter $< \tau_c >$ between V_g=0V and V_g=1V. As for FG traps, a trap with higher energy leads to a shorter $< \tau_e >$, however the impact is insignificant compared to FG traps.

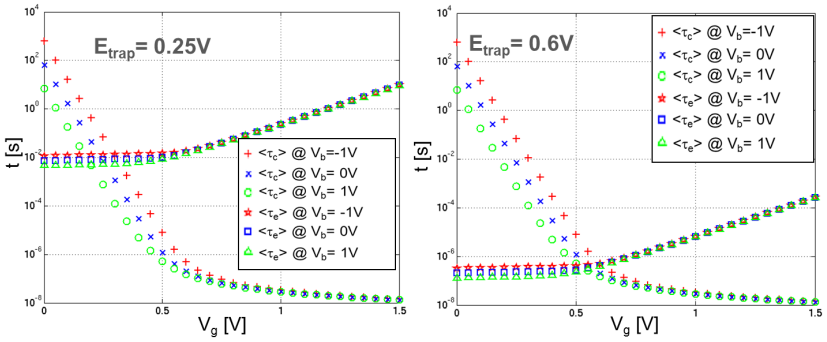

Figure 5.9: $< \tau_c >$ and $< \tau_e >$ evolution with respect to V_g and V_b for E_{trap} = 0.25eV (low energy) and 0.6eV (high energy) for FG dielectric. at ambient temperature.

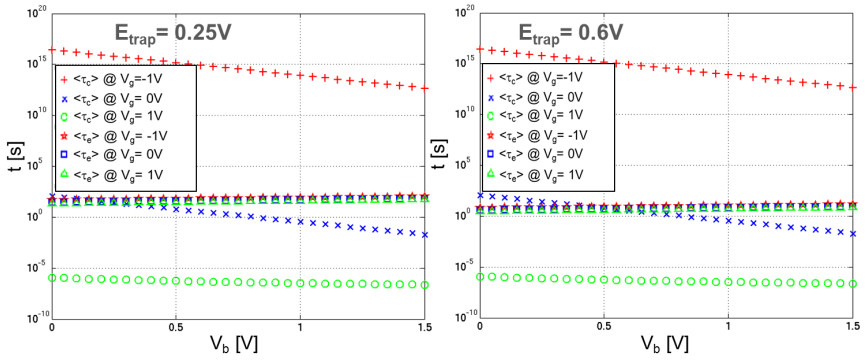

Figure 5.10: $< \tau_c >$ and $< \tau_e >$ evolution with respect to V_g and V_b for E_{trap} = 0.25eV (low energy) and 0.6eV (high energy) for BG dielectric. at ambient temperature.

5.2.3 RTS-aware 6T SRAM SPICE netlist generation in Matlab

Equation (5.3) shows that the effect of a trapped charges can be described in terms of V_{th} shifts, and a significant simplification can be done by describing the V_{th} shift as a shift of the gate voltage. In this way, a trapped charge in a given silicon/dielectric interface of a transistor can be modeled by a voltage sources connected to the gate node like in [132]. For example, V_{th} fluctuations caused by a single trap in the FG dielectric is added into SPICE transistor model by connecting a Piece-Wise Linear (PWL) voltage source to the front-gate of the transistor. How these fluctuation waveforms are generated will be described later in this section. The equivalent RTS-aware transistor model with multiple traps at both FG and BG, is shown in figure 5.11. N different voltage sources are connected in series modeling N FG traps, and M different voltage sources are connected in series modeling M BG traps. A voltage source modeling V_{th} fluctuation caused by $i-$th trap is denoted as R_i. The overall V_{th} modulations at each gate, ΔV_{FG} and ΔV_{BG}, are equal to the sum of independent δV_{th}s.

The stochastic behavior of RTS is accounted for by randomly generating the capture and emission times using exponentially falling distributions in time domain. As a consequence, traps are not stationary and their occupancy rate is changing with bias conditions. The bias dependency, which distinguishes this work from [132], is extracted from a nominal simulation, i.e. a run without RTS, performed at first. The extracted FG and BG gate voltage waveforms, $V_g(t)$ and $V_b(t)$, are used as inputs, together with dielectrics geometries, for RTS generator program in MATLAB. Figure 5.12 presents the proposed RTS-aware simulation flow.

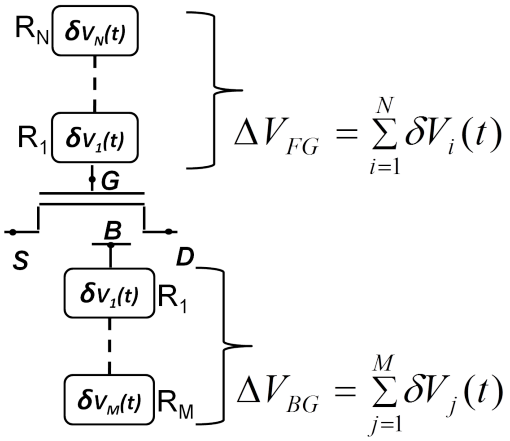

Figure 5.11: The equivalent RTS-aware transistor model with multiple traps: N traps at FG, M traps at BG. A voltage source modeling V_{th} fluctuation caused by $i-$th trap is denoted as R_i. The overall V_{th} modulation is the sum of the independent δV_{th}s.

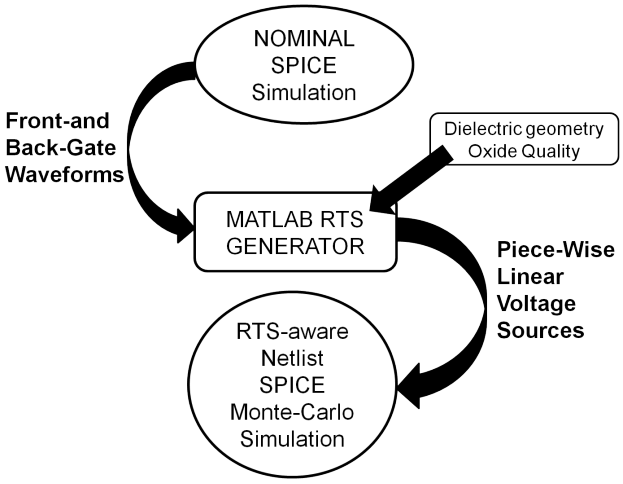

Figure 5.12: RTS-aware simulation flow chart.

Figure 5.13 presents in detail the MATLAB RTS PWL Generator, which is composed of two sub-generators:

- Trap Generator generates the trap profile of each device: position of traps in the dielectric layer, their energies and their cross sections. The generated profile can be saved for further use through the user input trap profile. The number of traps in a given dielectric is a Poisson random variable with the mean N_{avg} calculated using equation (5.4).
- PWL Voltage Source Generator generates the PWL Voltage Controlled

Voltage Source netlist for each trap, that will be connected to the corresponding device's gate.

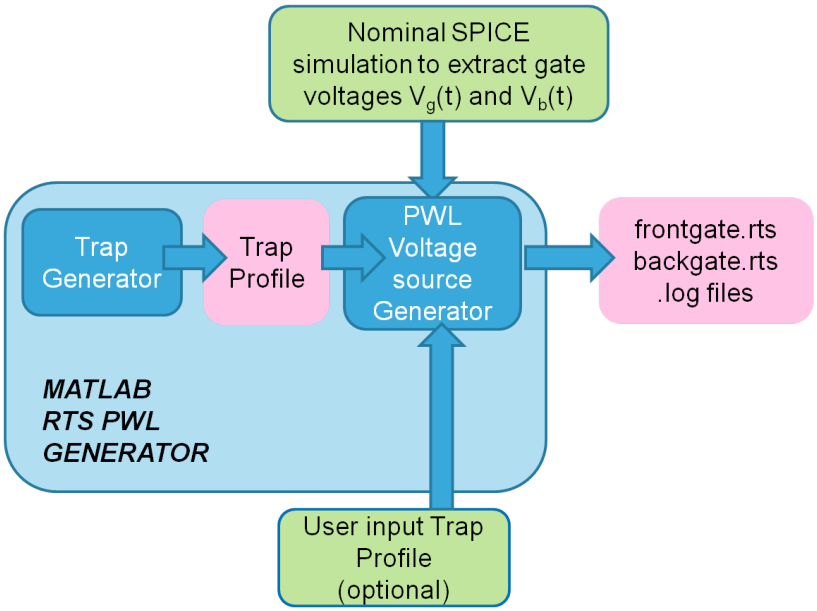

Figure 5.13: MATLAB RTS PWL Generator scheme.

For each trap in each device, the PWL Voltage source Generator generates bias-dependent $< \tau_c >$ and $< \tau_e >$ using $V_g(t)$ and $V_b(t)$ input waveforms as look-up tables. $\tau_c(t)$ and $\tau_e(t)$ are then randomly generated respecting their exponentially falling distribution. Each computed τ value is added to the sum of previously computed τ's, allowing to move forward step by step in time domain, until the overall simulation time, t_{max}, is reached. It is worth to say that the overall simulation time t_{max} is extracted by the priorly performed nominal simulation. The dynamic-like bias-dependent behavior of time constants is obtained by canceling very slow time constants that occurs when bias voltage is below a given threshold $V_{dd} - \epsilon$. This proposed optimization algorithm that allows to simulate dynamic behavior of $< \tau_c >$ and $< \tau_e >$, is illustrated in figure 5.14. Considering the simulation time t_i, where the gate voltage $V_g(i)$ is at 0V, the τ_{ci} computed at t_i can be very slow (few seconds), since equation (5.7) will result in a nearly zero Q_i. If this τ_{ci} is very slow, the next computation step that will be at $t_i + \tau_{ci}$ exceeds t_{max}, meaning that there will be no trapping/de-trapping during the simulation, even if the $V_g(t)$ switches to high-voltage level. However, considering short channel transistors, the capture and emission time constants response to bias change is the same order as the electron diffusion time ($\approx 1ps$), meaning that the time constants have to be updated when a bias change occurs. A simple way to simulate this bias-dependency with-

out affecting the simulation speed is to cancel very slow time constant that occurs at very low gate bias. This approach is not affecting the model accuracy, since these slow time constants would not be anymore valid when a bias change occurs considering real-world SRAM operating frequencies.

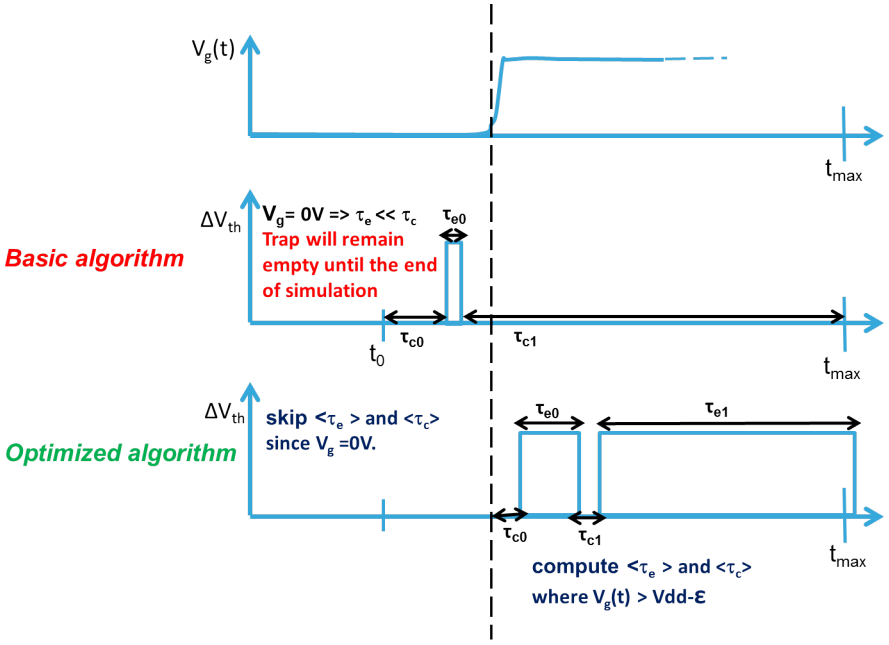

Figure 5.14: The proposed optimization algorithm to model bias-dependent behavior of RTS time constants. Very slow τ_c that occurs at V_g=0V is canceled out and time constats are calculated only if $V_g > (V_{dd} - \epsilon)$.

Figure 5.15 presents RTS-aware 6T SPW bitcell netlist. A PWL Voltage source associated to a single trap is illustrated with a block denoted as R. Each transistor of the illustrated bitcell has one trap at FG and one trap at BG. The output waveform of a single R (single trap) is illustrated in figure 5.16 on the top. The trap initialization time, t_0, is first chosen randomly, then τ_{c0}, τ_{e0}, τ_{c1}, τ_{e1}, τ_{c2}, τ_{e2} are computed successively (using the previously described optimization algorithm) until the sum of the τ's reaches t_{max}. On the bottom of figure 5.16, the overall V_{th} modulation waveform that is caused by multiple traps in a given device, is shown.

5.3 Measurements and Simulation Results

5.3.1 Hardware Setup

The RTS investigations are carried out on a 143Kb SPW SPHD SRAM array through Write Ability measurements using the dynamic characterization module that is mentioned in section 4.4.2. The measurement is com-

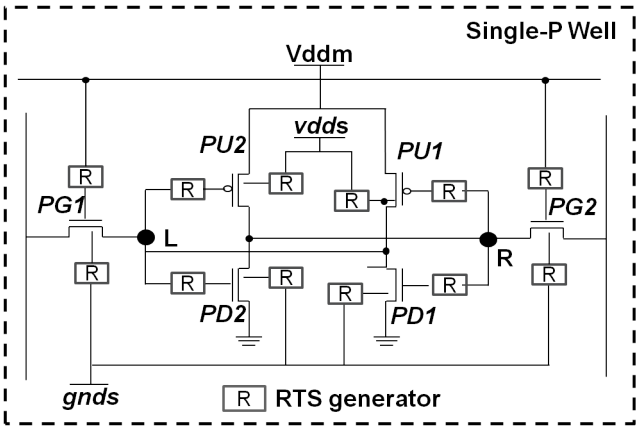

Figure 5.15: An RTS-aware 6T SPW bitcell netlist where each device has one FG trap and one BG trap. The bloc "R" represents the PWL voltage source modeling V_{th} fluctuation cuased by a single trap.

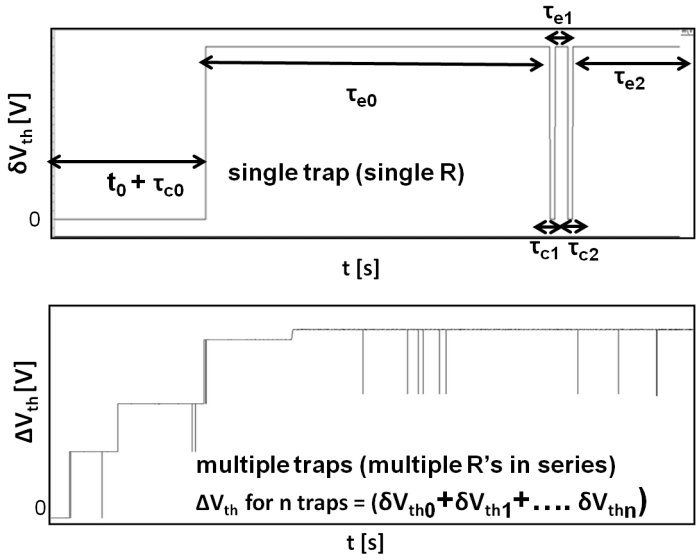

Figure 5.16: On the top: PWL Voltage source waveform of a single trap. On the bottom: The overall V_{th} modulation waveform with multiple traps.

posed of 4 phases: Initialization, overdrive, write (WR) and check. The measurement setup is presented in figure 5.17. In the initialization phase, the memory (same bitcell as in figure 5.15) is written at nominal operating voltage ($Vdd_{nominal}$) and low-frequency with the internal node "L" to "1". An overdrive phase of 1.5 seconds at 1.9V is performed to accelerate trapping and to increase the number of filled traps. Considering the memory initial conditions, the overdrive will impact more PU1 and PD2, since

these two devices are turned on during the overdrive, thus their channels are in strong inversion. The array operating voltage Vddm is reduced to the Vddm_{test} for the WR phase, which is performed column wise to avoid read disturb errors. Vddm_{test} is held at very low value to push the bitcells towards the failure zone, where the sensitivity to RTS will be critical. In the check phase, the memory is read at Vdd_{nom} and low frequency to avoid read access errors. The fraction of cells having a WR failure is denoted the Bit-Error-Rate (BER). It is worth to note that the used measurement allows separate the effects of the dynamic variability and the static variability that is present in the measured die.

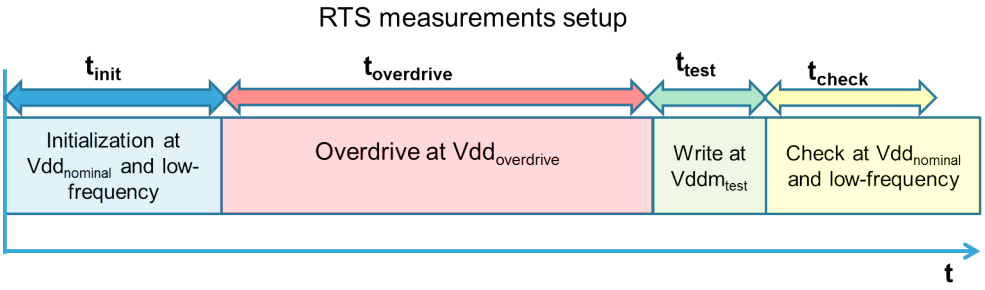

Figure 5.17: RTS measurement setup that is used in the dynamic characterization module.

The analysis is performed in the following way: BER is measured with and without overdrive, and the difference of the 2 results is interpreted as being due to RTS. 4 different conditions are tested: no overdrive, only-FG-overdrive, only-BG-overdrive and both FG-and BG overdrive, allowing to study separate and joint impacts of both gates. In simulation, both RTS-aware netlist and the nominal netlist (without RTS) are simulated with Monte Carlo method using the STMicroelectronics SPICE model cards, at ambient temperature and typical corners. The overdrive phase is also simulated with RTS-aware netlist to increase the probability of trapping.

In simulation, for each device, it is assumed that the number of traps follows a Poisson distribution, traps are uniformly distributed in the dielectric, and, for each trap, the cross-section pre-factor σ_0 is randomized between 1.10^{-14} and 1.10^{-17}.

Before starting to discuss the results, a reminder about Write Ability failure mechanisms (section 3.6) is useful, since the RTS can be only observed in failing bitcells. A discharge write failure is caused by a strong PMOS which will prevent discharging the node initially storing "1", and a completion write failure is caused by a weak PMOS which will prevent the node initially storing 0 reaching 1. In other words, the WA of a bitcell depends on the strength ratio between the NMOS and PMOS. In this work, the body biasing is used to adjust NMOS-PMOS strength ratio, and

this causes a transition between the WA failure mechanisms. Besides, this is done in a simple way controlling only one body-supply voltage, thanks to the single-well architecture. A negative V_b will place the bitcell in discharge failure zone, since it strengthens (forward body-bias) PMOS, whereas a positive V_b will place the memory in the completion failure zone, since it weakens (reverse body-bias) PMOS. According to this, the presence of RTS noise can be visible as the difference of BER between the nominal test and the overdriven test: If a bitcell is failing due to strong PMOS, the overdriven PMOS that is weakened due to RTS can prevent the failure. On the other hand, if a bitcell is failing due to weak PMOS, the overdriven PMOS may lead to additional failures.

5.3.2 Results

Figure 5.18 shows WA measurement BER for a single write operation, in which a '0' is written into the node 'L' at 80ns WL pulse width and 0.34V Vddm_{test}. Results for 4 different conditions are shown. It is shown that at strongly negative V_b, the only-FG-overdrive leads to a significant BER improvement, since the memory is in its discharge failure zone and the weakened PMOS due to RTS cancels out some of the discharge failures. On the other hand, only-BG-overdrive do not show a significant impact compared to the nominal test (no overdrive), indicating that the impact of the BG traps remains as a second order and the main source of RTS noise is the FG dielectric. This is also confirmed by the FG- and BG-overdrive test, which gives similar results as only-FG-overdrive and the slight difference can be explained by statistical error in the measurements. Therefore, although BG traps have larger amplitudes (up to 60mV), the body factor (≈ 0.06 [50, 62]) attenuates their final impact on the device V_{th}.

Figure 5.19 shows WA simulation BER for a single write operation, at 108ns WL pulse width and 0.34V Vddm_{test}. The RTS-aware netlist is used to perform only-FG-RTS, only-BG-RTS and FG-and BG-RTS tests. As in figure 5.18, RTS-aware simulations are compared with the nominal netlist (no RTS) results. The same WA BER trend as in the measurements is reproduced, showing the good accuracy of the proposed RTS-aware simulation method.

The good accuracy of the proposed RTS-aware simulation method is also illustrated in figure 5.20, which shows the normalized measurements and simulation WA BER, performed at 320ns pulse width and Vddm_{test}=0.34V. The one should conclude that the WA BER improvement trend that is observed on silicon, is accurately reproduced using RTS-aware netlist.

With single write WA tests, the RTS noise presence is observed as improvement of BER, when the memory is placed in its discharge failure zone. Figure 5.21 illustrates Write-After-Write (WAW) WA measurements BER with respect to the V_b, at Vddm_{test}=0.34V and 320ns WL pulse-width.

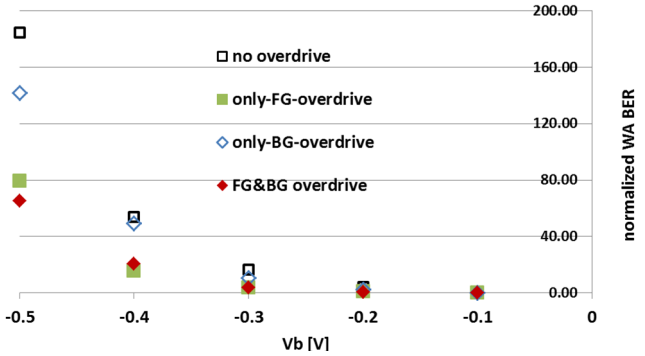

Figure 5.18: Single Write WA measurements BER vs. V_b for no overdrive, only-FG-overdrive, only-BG-overdrive and both FG-and BG overdrive, at 80ns WL pulse width and $Vddm_{test}$ = 0.34V. The BER improvement with overdrive is interpreted as the impact of the RTS in the PMOS PU1, which slows down the device, thus cancels some of discharge failures. Results are normalized with respect to those obtained at V_b =-0.1V

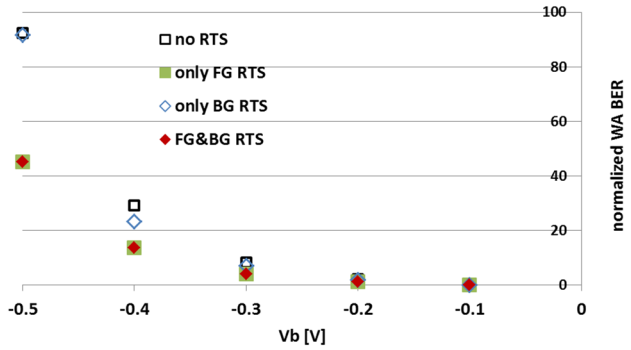

Figure 5.19: Single Write WA simulation BER vs. V_b for no RTS, only-FG-RTS, only-BG-RTS and both FG-and BG RTS, at 80ns WL pulse width and $Vddm_{test}$ = 0.34V. Results are normalized with respect to those obtained at V_b =-0.1V.

WAW consists in performing a WR0 which is followed by a WR1 in the next WL pulse, and the memory content is checked to read '1' from the node 'L'. At $V_b < 0$, the same BER change trend as in the single write WA mesurements is observed, since the initially strong PU1 places the memory in the discharge failures zone and these discharges failures occur in the WR0 giving the same picture as the single write test. In distinct to single write WA measurements, results for $V_b > 0$ are also shown, where only-FG-overdrive results present degradation in BER. At strongly positive V_b, the PMOS is initially reverse body-biased (weakened) switching the main failure mechanism to completion failures and WR0 is therefore not anymore con-

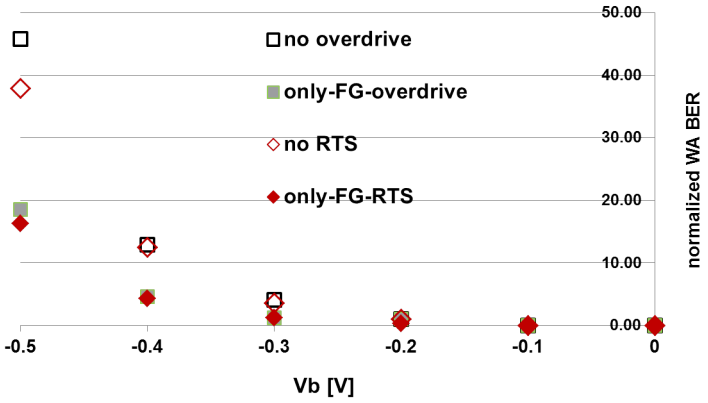

Figure 5.20: Single Write WA measurements BER vs. simulation BER with respect to V_b for FG overdrive and the FG-RTS simulation, at 320 ns WL pulse width and $Vddm_{test}$ = 0.34V. Results are normalized with respect to those obtained at V_b =-0.1V.

cerned by discharge failures. Since the memory initial conditions are same as in the single write test ('L' at '1'), the devices that are mostly impacted by RTS noise are also same (PU1 and PD2), and the overdriven PU1 plays also an active role in the completion of the WR1. The BER degrades with overdrive, since the PU1 is weakened due to RTS noise increasing completion failures that occurs in WR1.

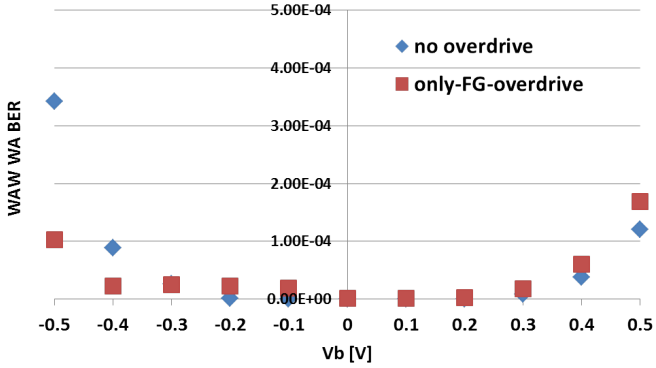

Figure 5.21: Multiple Write (WAW) WA measurements BER with respect to V_b for no overdrive and only-FG-overdrive, at 320 ns WL pulse width and $Vddm_{test}$ = 0.34V. The impact of RTS is observed at extreme V_b values, as a BER decrease at negative V_b (discharge failure zone) and as a BER increase at positive V_b (completion failure zone).

Figure 5.22 shows WAW simulations with and without RTS-aware netlist,

at two extreme V_b conditions, Vb=-0.4V and Vb=0.4. The WA test is performed after each WL pulse. At Vb -0.4V, the discharge BER in WR0 is reduced by 62%. The total BER is decreased by 70% telling that discharge in WR1 is also improved. At positive V_b, WR1 completion BER is increased by 38%, a result that justifies the increase in total BER by 35%. In conclusion, the RTS noise in the overdriven PU1 impacts the WA BER as either an improvement or degradation depending on the main failure mechanism.

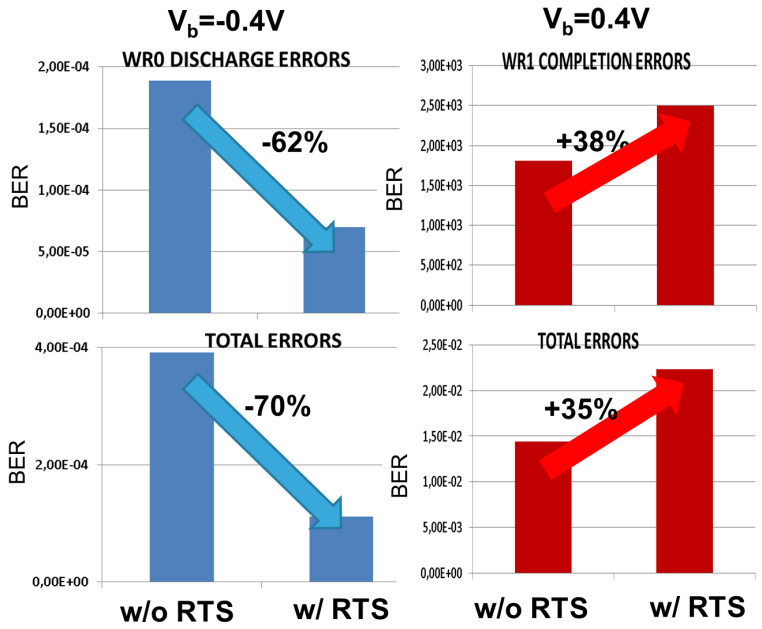

Figure 5.22: Multiple Write (WAW) WA simulation BER at $Vddm_{test}$ = 0.34V and 320 ns WL pulse width for no-overdrive and only-FG-overdrive. Two exreme V_b conditions are simulated. At negative V_b BER decreases with overdrive (discharge failures) and at positive V_b, BER increase with overdrive ((completion failures).

Figure 5.23 presents WAW BER measurement vs. V_{dd} at 108ns pulse width and 0.5V V_b, with (squares) and without (diamonds) overdrive. It is shown that the minimum Vddm at zero BER with respect to WAW WA can increase due to RTS, in aggressive conditions. A 30mV increase is observed for this particular chip, and, even if a larger increase can be supposed to arise in a statistical amount of chips, the shift looks relevant but not critical for this technology node. The strongly positive V_b justifies the increase of the BER due to RTS, since the memory is in its completion failure zone and a weakened PMOS due to overdrive leads to additional failures.

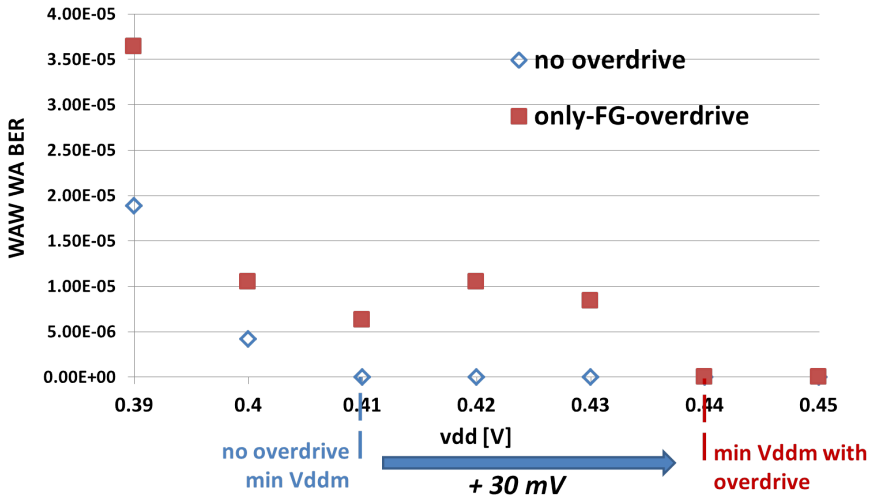

Figure 5.23: Multiple Write (WAW) WA measurements BER with respect to V_{dd} for no overdrive and only-FG-overdrive, at 108 ns WL pulse width and V_b = 0.5V. The overdrive results in a 30mv increase in the minimum Vddm that can be reached without having write failures.

5.4 Conclusion

In this chapter, we investigated the impact at circuit level of RTS on a 6T SRAM manufactured in 28nm UTBB FD-SOI technology node. A new modeling approach considering the UTBB FD-SOI back-gate dielectric as a second source for RTS and taking into account front-and back-gate coupling has been presented. The proposed analytical model is first validated by TCAD simulations. The RTS is added in SPICE netlist as a V_{th} fluctuations through voltage sources connected to the device gate, for which the stochastic time-domain waveforms are generated in MATLAB. The bias-and time-dependency of carrier trapping is extracted from a nominal simulation. Single and multiple-writes measurement tests are performed, in which an overdrive is first applied to accelerate trapping. The presence of RTS noise impact is observed as either an improvement or degradation of the Write-Ability Bit-Error-Rate, depending on the write failure mechanism at that given measurements conditions. The body voltage V_b is used to adjust NMOS-PMOS strength ratio, allowing to switch between two different write failure mechanism.

It is demonstrated that the proposed RTS-aware SPICE-level model reproduces well the measured BER trends. Results show that the RTS is not a critical failure source in 28nm UTBB FD-SOI 6T SRAM bitcells and that the bitcell has to be pushed in his failure zone to expose RTS-related issues,

which is illustrated in figure 5.24 for a 2D variability space. However, in very aggressive conditions, it is shown that RTS could limit further reduction of V_{dd} for SRAM circuits, thus it should be considered as a serious variability concern for future, downscaled UTBB FD-SOI devices.

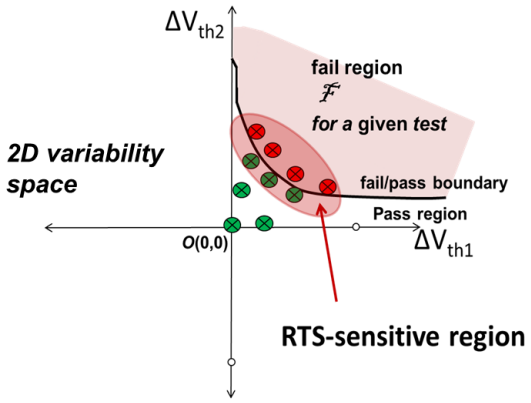

Figure 5.24: The illustration of the RTS-sensitive zone in SRAM bitcell in a 2D variability space. RTS is not a limiting variability souce in 28nm UTBB FD-SOI technology and becomes visible when the memory is already in the failure zone.

Chapter 6

General Conclusion

In the ultra-deep-submicron era, SRAM design is by nature very challenging due to the fact that the design must answer both high-density and low power requirements. The aggressive downscaling in transistor geometry and the operating voltage exacerbate together the impact of process variability on the stability and performance metrics of SRAM circuits. Considering that the manufacturing yield of memory circuits represents an important milestone for the maturity of a given technology node, a joint optimization is needed between process technology and memory design, and this has even become a mandatory requirement for the process development in the advanced technology nodes. This work shows novel variability-aware statistical methodologies for SRAM bitcell simulation that is to be used during SRAM design optimization phase of process development, with the ultimate goal of providing a sufficiently accurate modeling of the SRAM bitcell minimum operating voltage. The compatibility with industrial EDA softwares and the easiness of integration into the already existing simulation flows are stated as key qualities of proposed statistical methodology, evidently together with modeling accuracy and time- and computation-cost efficiency. Besides the modeling of the conventional static variability subclass of process variability, which is by definition not time-dependent, a particular interest is carried out on the modeling of the time-dependent sub-class of the process variability. Random Telegraph Signal noise which is seen as a critical dynamic variability concern for sub-28nm SRAM circuits, is first modeled at UTBB FD-SOI transistor-level and is later successfully integrated into SRAM bitcell SPICE netlist.

6.1 Key Contributions

The key contributions of this work are as follows:

- An existing Monte Carlo method based static design margin modeling is improved in order to increase the modeling accuracy under large

variations. The methodology improvement is derived from investigations that are carried out on a very large amount of Static Noise Margin silicon measurements performed on C45 SPHD SRAM bitcell.

- The proposed methodology is first applied on C40 SPHD and SPREG bitcells for V_{min} modeling. The worst-case V_{min} across all process corners is simulated with 30 mV error with respect to silicon measurements that are performed on intentionally skewed process corners during the process monitoring.

- The static design margin modeling methodology is also applied on C65 Ultra Low Voltage 10 Transistor bitcell and V_{min} is estimated with 20 mV error with respect to the silicon measurements showing the adaptability of the proposed methodology to different SRAM bitcell architectures.

- In order to perform more complex variability analysis addressing very large cut size and both static and transient operating conditions, a smart algorithm based on hypespherical surface analysis of the multi-dimensional bitcell variability space is developed. The proposed *Hypersphere Most Probable Failure (MPFP) search* methodology is coupled with Importance Sampling method allowing accurate simulation bit-error-rate extraction under large variations with a feasible computation time.

- *The Hypersphere MPFP search* tool allows extracting different mismatch mechanisms that underlie behind bitcell failures, offering a better understanding of bitcell failures and their dependency on operating conditions. The simulation-based findings are also validated through silicon measurements that are performed by CEA/LETI. In particular, the tool has been used during 28nm UTBB FD-SOI technology development to investigate SNM drop issue that had appeared at high-voltage SRAM operations.

- In order to anticipate novel variability-related limitation in further downscaled SRAMs, the Random Telegraph Signal (RTS) Noise, which results in time-dependent mismatch in a given SRAM bitcell, is studied. A transistor-level modeling that takes into account the particular two-gates device architecture of UTBB FD-SOI technology is presented. The transistor-level modeling is integrated into SPICE SRAM bitcell netlist in order to perform bias-dependent RTS-aware simulations. Silicon measurements are performed to validate the accuracy in modeling the impact of RTS noise on SRAM stability metrics using the proposed RTS-aware netlist.

6.2 Future Work

This work has shown that a variability-aware SRAM bitcell design using industrial EDA softwares can be performed with a sufficient accuracy with respect to the real-life manufactured SRAM bitcells. One can say that it is always possible to provide higher accuracy or lower computation cost and this can be true considering the nature of modeling works. We have concluded that the provided accuracy and the computation time for V_{min} modeling under static process variability is admissible for a semiconductor industry and therefore no perspective work for modeling improvement will be mentioned. However, it is worth to say that the normally distributed threshold voltage approximation for modeling process variability impact on a single transistor tends to be not valid anymore in further downscaled technologies ($\approx 10nm$ gate length) and as a consequence, modeling methodologies based on this approximation may not be valid neither.

On the other hand, considering our findings about the impact of the RTS noise and the growing weight of the dynamic variability with respect to the one of the static variability, the development of bias-dependent RTS-aware SPICE model cards modeling accurately RTS noise has to be seen as a priority by semiconductor industry and EDA software companies for further downscaled technologies.

More generally, considering today's SRAMs which may operate up to few GHz and the exacerbated parasitic effects that are present in the high-density memory arrays, bitcell stability under transient conditions becomes main concern for V_{min} limitation and may lower significantly V_{min} yield that is measured using static stability tests. Semiconductor industries therefore need to update their conventional bitcell characterization flow in order to integrate dynamic stability metrics.

It has also explicitly shown that the variability-aware bitcell design optimization is a mandatory requirement in advanced technology nodes. The memory designers have to be aware of potential risks related to process variability and also understand the underlying reasons behind variability-related failures. Based on this fact, the knowledge acquired on the different bitcell failure mechanism and their dependency on operating conditions can be used to design "highly-optimized" bitcells for specific applications. In particular, the very promising body-biasing feature together with the low threshold voltage variability provided by UTBB FD-SOI technology can offer great opportunities for optimum memory designs that address all of the high-density, high-performance and low-power requirements. This optimization should not be only considered in bitcell design, but also for novel assist circuitry techniques.

Publications and Patents

- O.Thomas, B. Zimmer, B. Pelloux-Prayer, N. Planes, K.C. Akyel, L. Ciampolini, P. Flatresse and B. Nikolic, "6T SRAM design for wide voltage range in 28nm FDSOI", IEEE SOI Conference 2012, pp 1-2.
- K.C. Akyel, L. Ciampolini, O. Thomas, B. Pelloux-Prayer, S. Kumar, P. Flatresse, C. Lecocq and G. Ghibaudo, "Multiple-Pulse Dynamic Stability and Failure Analysis of 6T-SRAM Low-Voltage Bitcells in 28nm FDSOI", Circuits and Systems (ISCAS), 2013 IEEE International Symposium on, May 2013, pp 1452-1455.
- F. Abouzeid, A. Bienfait, K.C. Akyel, A. Feki, S. Clerc, L. Ciampolini, F. Giner, R. Wilson and P. Roche, "Scalable 0.35V to 1.2V SRAM bitcell design from 65nm CMOS to 28nm FDSOI", IEEE Journal of Solid-State Circuits, July 2014, pp 1499-1505.
- K.C Akyel, L. Ciampolini, O. Thomas, D.Turgis and G. Ghibaudo, "Impact of Random Telegraph Signals on 6T High-Density SRAM in 28nm UTBB FD-SOI", IEEE European Solid State Device Conference (ESSDERC) 2014.
- K.C Akyel, L. Ciampolini, O. Thomas, D. Turgis and G. Ghibaudo, "28nm UTBB FD-SOI Front-and Back-Gate Coupling Aware Random Telegraph Signal Impact Analysis on a 6T SRAM", IEEE S3S Conference 2014.
- "Smart dynamic back-bias for the bitcell Vmin boost, 13-GR1-0652FR01, Inventors: Christophe Lecocq, Kaya Can Akyel, Amit Chhabra, Dipya Dibti.

Bibliography

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 28, no. 8, 1965.

[2] W. Shockley, "Problems related to p-n junctions in silicon," *Solid-State Electronics*, vol. 2, pp. 35–67, January 1961.

[3] W. Schemmert and G. Zimmer, "Threshold voltage sensitivity of ion-implanted m.o.s transistors due to process variations," *Electronics Letters*, vol. 10, pp. 151–152, may 1974.

[4] S. K. Saha, "Modelling process variability in scaled CMOS technology," *IEEE Design and Test of Computers*, vol. 27, pp. 8–16, march 2010.

[5] L.-T. Pang, K. Qiang, C. Spanos, and B. Nikolic, "Measurements and analysis of variability in 45nm strained-si CMOS technology," *IEEE Journal of Solid-State Circuits*, vol. 44, August 2009.

[6] P. Asenov, *Accurate Statistical Circuit Simulation in the presence of statistical variability.* PhD thesis, University of Glasgow, march 2013.

[7] T. Tanaka, T. Usuki, Y. Momiyama, and T. Sugii, "Direct meaurement of vth fluctuation caused by impurity positioning," in *Very Large Scale Integration Sypmosium on Technology Digest of Papers*, pp. 136–137, 2000.

[8] R. Tian and X. Tang, "Dummy-feature placement for chemical-mechanical polishing uniformity in a shallow trench insulation process," *IEEE Transactions on Computer Aided Design*, vol. 21, no. 1, pp. 63–71, 2002.

[9] M. Kahre, "High-k/metal gate technology: A new horizon.," in *Proceedings of Custom Integrated Circuits Conference*, pp. 417–420, 2007.

[10] A. B. Khang and Y. Pati, "Subwavelength lithography and its potential impact on design and eda," in *Proceedings of 36th Design Automation Conference*, pp. 799–804, June 1999.

[11] A. K. Wong, "Microlithography: trends, challenges, solutions, and their impact on design," *IEEE Micro*, vol. 23, pp. 12–21, March/April 2003.

[12] K. J. Kuhn, "Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS," in *IEEE Electron Device Meeting (IEDM)*, pp. 471–474, 2007.

[13] M. Steyaert, J. Bastos, R. Roovers, P. Kinget, W. Sansen, B. Graindourze, A. Pergoot, and E. Janssens, "Threshold voltage mistmatch in short-channel MOS transistors," *Electronic Letters*, vol. 30, no. 18, pp. 1546–1548, 1994.

[14] P. Stolk, F. P. Widdershoven, and D. B. M. Klaassen, "Modeling statistical dopant fluctuations in MOS transistors," *IEEE Transactions on Electron Devices*, vol. 45, no. 9, pp. 1960–1971, 1998.

[15] K. Bernstein, "High-performance CMOS variability in the 65-nm regime and beyong," *IBM J. Research and Development*, vol. 50, no. 4-5, pp. 443–449, 2006.

[16] 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 *Proceedings of International Conference SISPAD*, pp. 131–134, 2000.

[17] A. Asenov, S. Kaya, and A. Brown, "Intrinsic paremeter fluctuations in decananometer MOSFETs introduced by gate line edge roughness," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1254–1260, may 2003.

[18] S. Xiong and J. Bokor, "A simulation study of gate line edge roughness effects on doping profiles of short-channel mosfet devices," in *IEEE Transctions on Electron Device*, vol. 51(2), pp. 228–232, February 2004.

[19] A. Asenov, S. Kaya, and J. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations," *IEEE Transctions on Electron Devices*, vol. 49, no. 1, pp. 112–119, 2002.

[20] R. Difrenza, J. Vildeuil, P. Llinares, and G. Ghibaudo, "Impact of grain number fluctuations in the MOS transistor gate on matching performance," in *IEEE ICMTS*, pp. 244–249, 2003.

[21] A. Cathignol, K. Rochereaub, and G. Ghibaudo, "Impact of a single-grain boundary in the polycrstalline silicon gate on sub-100-nm bulk

MOSFET characeristics. implication on matching properties," in *Proceedings of the 7th International Conference ULIS*, pp. 145–148, 2006.

[22] C. Mezzomo, A. Bajolet, A. Cathignol, R. D. Frenza, and G. Ghibaudo, "Characterization and modeling of transistor variability in advanced CMOS technologies," *IEEE Transactions on Electron Devices*, vol. 58, no. 8, pp. 2235–2248, 2011.

[23] D. Park, Y. King, Q. Lu, T. king, C. Hu, A. Kalnitsky, S.-P. Tay, and C. Cheng, "Transistor characteristics with Ta_2O_5 gate dielectric," *IEEE Electroin Devices Letters*, vol. 19, p. 441, 1998.

[24] A. Cathignol, B. Cheng, D. Chanemougarme, A. R. Brown, K. Rochereau, G. Ghibaudo, and A. Asenov, "Quantative evaluation of statistical variability sources in a 45-nm technological node lp n-mosfet," *IEEE Electroin Devices Letters*, vol. 29(6), pp. 609–611, june 2008.

[25] J. Kolhatkar, *Steady-State and Cyclo-stationary RTS Noise in MOSFETs*. PhD thesis, University of Twente, 2005.

[26] G. Ribes, J. Mitard, M. Denais, S. Bruyere, F. Monsieur, C. Parathasaray, E. Vincent, and G. Ghibaudo, "Review on high-K dielectrics reliability issues," *IEEE Transactions on Device and Materials Reliability*, vol. 5(1), pp. 5–19, March 2005.

[27] E. H. Nicollian and J. Brews in *MOS (Metal Oxide Semiconductor) Physics and Technology*, Wiley, 1982.

[28] G. Ghibaudo and T. Boutchacha, "Electrical noise and RTS fluctuations in advanced CMOS devices," *Microelectronics Reliability*, vol. 42, pp. 573–582, may 2002.

[29] L. Brusamarello, I. Gilson, and R. da Silva and, "Statistical RTS fluctuations in advanced CMOS devices," *Microelectronics Reliability*, vol. 49, no. 9-11, pp. 1064–1069, 2009.

[30] C. E. Blat, e.H. Nicollian, and E. H. Poindexter, "Mechanism of negative bias temperature instability," *Journal of Applied Physics*, vol. 69, no. 3, pp. 1712–1720, 1998.

[31] B. Kaczer, T. Grasser, P. Roussel, J. Franco, R. Degraeve, L.-A. Ragnarsson, E. Simoen, G. Groeseneken, and H. Reisinger, "Origin of NBTI variability in deeply scaled pFETs," in *IEEE International Reliability Physics Symposium*, pp. 26–32, 2010.

[32] A. Brown, V. Huard, and A. Asenov, "statistical simulation of progressive NBTI degradation in a 45-nm technology pMOSFET," in *IEEE Transactions on Electron Devices*, vol. 57, pp. 2320–2323, 2010.

[33] J. F. Zhang and W. Eccleston, "Positive bias temperature instability in MOSFETs," *IEEE Transactions on Electron Devices*, vol. 45, no. 1, pp. 116–124, 1998.

[34] E. Takeda, "Hot-Carrier effects in submicrometer MOS VLSIs," *IEEE Proceedigns*, vol. 131, pp. 153–162, 1984.

[35] D. Young and A. Christou, "Failure mechanism models for electromigration," *IEEE Transactions on Reliability*, vol. 43, no. 2, pp. 186–192, 1994.

[36] S. W. Director and w. Maly, "Statistical approach to VLSI," *Advances in CAD for VLSI*, vol. 8, 1994.

[37] J. Power, B. Donnellan, A. Mathewson, and W. Lan, "Relating statistical mosfet model paremeter variabilities to ic manufacturing process fluctuations enabling realistic worse case design," in *IEEE Trans. Semiconductor Manufacturing*, august 1994.

[38] J. Zhang and M. Styblinski, "Yield variability optimization of integrated circuits," in *Kluwer*, 1994.

[39] P. Drennan, "Understanding mosfet mismatch for analog design," *IEEE Journal of Solid-State Circuits*, vol. 38, March 2003.

[40] S. G. Duvall, "Statistical circuit modeling and optimization," in *5th International Workshop on Statistical Metrology*, pp. 56–63, June 2000.

[41] K. Nagase, S. Ohkawa, M. Aoki, and H. Masuda, "Variation status in 100nm CMOS process and below," in *Proceedings of Conference of Microelectronic Test Structure*, pp. 257–261, march 2004.

[42] K.Kuhn, C. Kenyon, A. Kornfeld, M. Liu, A. Maheshwari, S. Weikai, S. Sivakumar, G. Taylor, P. VanDerVoorn, and K. Zawadzki., "Managing process variation in Intel's 45nm CMOS technology," *Intel Technology Journal*, vol. 12, pp. 93–1089, june 2008.

[43] C. Chiang and J. Kawa in *Design for Manufacturability and yield for nanoscale CMOS*, Springer, 2007.

[44] A. J. Strojwas, "Challenges in modelling layout schematic effects in compact device models," in *MOS-AK.GSA Workshop*, 2010.

[45] L.-T. Pang and B. Nikolic, "Impact of layout on 90nm CMOS process parameter fluctuations," in *Symposium on VLSI Circuits Digital Tech. Papers*, pp. 69–70, june 2006.

[46] S. Nassif, "Within chip variability analysis," in *IEEE International Electron Device Meeting*, pp. 283–286, December 1998.

[47] A. Asenov, A. Cathignol, B. Cheng, K. P. McKenna, A. Brown, A. L. Shluger, D. Chanemougame, K. Rochereau, and G. Ghibaudo, "Origin of the asymmetry in the magnitude of the statistical variability of n- and p-channel Poly-Si Gate bulk MOSFETs," *IEEE International Electron Device Meeting*, vol. 29, no. 8, pp. 913–915, 2008.

[48] F. Arnaud, A. Thean, M. Eller, M. Lipinski, Y. Teh, M. Ostermayr, K. Kang, N. Kim, K. Ohuchi, J.-P. Han, D. Nair, J. Lian, S. Uchimura, S. Kohler, S. Miyaki, P. Ferreira, J.-H. Park, M. Hamaguchi, K. Miyashita, R. Augur, Q. Zhang, K. Strahrenberg, S. El-Ghouli, J. Bonnouvrier, F. Matsouoka, R. Lindsay, J. Sudijono, F. Johnson, J. Ku, M. Sekine, A. Steegen, and R. Sampson, "Competitive and cost effective high-K based 28nm CMOS technology for low power applications," *IEEE International Electron Device Meeting*, pp. 28.2.1–28.2.4, 2009.

[49] C. F. Nieh, K. C. Hu, H. Chang, L. Wang, L. Huang, Y. Sheu, C. W. T. Lee, S. Chen, and J. Gong, "Millisecond anneal and short-channel effect control in Si CMOS transistor performance," *IEEE Electron Devices Letters*, vol. 27, no. 12, pp. 86–87, 2006.

[50] J.-P. Noel, *Optimisation de dispositifs FDSOI pour la gestion de la consommation et de la vitesse: application aux memoires et fonctions logiques*. PhD thesis, University of Grenoble, 2011.

[51] A. Asenov, "Simulation of statistical variability in nano MOSFETs," in *Symposium on VLSI Technology Digets of Technical Papers*, pp. 86–87, 2007.

[52] Y. Leblebici, "Design considerations for CMOS digital circuits with improved hot-carrier reliability," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 7, pp. 1014–1024, 1996.

[53] H. Yonezawa, J. Fang, Y. Kawakami, N. Iwanishi, L. Wu, A.-H. Chen, N. K. an P. Chen, Y. Chune-Sin, and L. Zhihong, "Ratio based hot-carrier degradation modeling for aged timing simulation of millions of transistors digital circuits," in *Electron Device Meeting Technical Digest*, pp. 93–96, 1998.

[54] B. Paul, K. Kang, H. Kufluoglu, M. Alam, and K. Roy, "Impact of NBTI on temporal performance degradation of digital circuits," *IEEE Electron Device Letters*, vol. 26, no. 8, pp. 560–562, 2005.

[55] W. Wang, S. Yang, S. Bhardwaj, S. Liu, and Y. Cao, "The impact of NBTI effect on combinational circuit: Modeling, simulation, and analysis," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 18, no. 2, pp. 173–183, 2010.

[56] T. K. Y. Lee, "A fine-grained technique of nbti-aware voltage scaling and body biasing for standard cell based designs," in *16th Asia and South Pacific Design Automation Conference*, pp. 603–608, 2011.

[57] J. Yuan and H. Tang, "CMOS RF design for reliability using adaptive gate-source biasing," *IEEE Transactions on Electron Devices*, vol. 55, pp. 2348–2353, september 2008.

[58] K. Kang, S. Park, K. Kim, and K. Roy, "On-chip variability sensor using phase-locked loop for detecting and correcting parametric timing failures," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 18, pp. 270–280, February 2010.

[59] H. Xuejue, L. Wee-Chin, K. Charles, D. Hisamoto, C. Leland, J. Kedzierski, E. Anderson, H. Takeuchi, c. Yang-Kyu, K. Asano, V. Subramanian, K. Tsu-Jae, J. Bokor, and H. Chenming, "sub 50-nm FinFet: PMOS," in *International Electron Device Meeting Technical Digest*, pp. 67–70, 1999.

[60] E. Karl, W. Yih, N. Yong-Gee, G. Zheng, F. Hamzaoglu, U. Bhattacharya, K. Zhang, K. Mistry, and M. Bohr, "A 4.6GHz 162mb SRAM design in 22-nm trigate CMOS technology with integrated active VMIN enhancing assist circuitry," in *International Solid-State Circuits Conference (ISSCC)*, pp. 230–232, february 2012.

[61] N. Planes, O. Weber, V. Barral, S. Haendler, D. Noblet, D. Croain, M. Bocat, P. Sassoulas, X. Federspiel, A. Cros, A. Bajolet, E. Richard, B. Dumont, P. Perreau, D. Petit, D. Golanski, C. Fenouillet-Beranger, N. Guillot, M. Rafik, V. Huard, S. Puget, X. Montagner, M.-A. Jaud, O. Rozeau, O. Saxod, F. Wacquant, F. Monsieur, D. Barge, L. Pinzelli, M. Mellier, F. Boeuf, F. Arnaud, and M. Haond, "28nm FDSOI technology platform for high-speed low-voltage digital applications," in *Technology Symposium on Very-Large Scale Integration (VLSIT)*, pp. 133–134, 2012.

[62] J. Noel, O. Thomas, M.-A. Jaud, O. Weber, T. Poiroux, C. Fenouillet-Beranger, P. Rivallin, P. Scheiblin, F. Andrieu, M. Vinet, O. Rozeau, F. Boeuf, O. Faynot, and A. Amara, "Multi-vt UTBB FDSOI device architecture for low-power CMOS circuit," *IEEE Transactions on Electron Devices*, vol. 58, no. 8, pp. 2473–2482, 2011.

[63] J. Mazurier, O. Weber, F. Andrieu, A. Toffoli, O. Rozeau, T. Poiroux, F. Allain, P. Perreau, C. Fenouillet-Beranger, O. Thomas, M. Belleville, and O. Faynot, "On the variability in planar FDSOI technology: From mosfets to SRAM cells," *IEEE Transactions on Electron Devices*, vol. 58, no. 8, pp. 2326–2336, 2011.

[64] C. Maleville and C. Mazure, "Smart Cut technology: from 300mm ultrathin soi production to advanced engineered substrates," *Solid-State Electronics*, vol. 48, no. 6, p. 855, 2004.

[65] A.-J. Annema, B. Nauta, R. V. Langevelde, and H. Tuinhout, "Analog circuits in ultra-deep-submicron CMOS," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 132–143, 2005.

[66] Y. Li, C.-H. Hwang, and T.-Y. Li, "Random-dopant-induced variability in nano-CMOS devices and digital circuits," *IEEE Transactions on Electron Devices*, vol. 56, no. 8, pp. 1588–1597, 2009.

[67] J. S. Gyvez and H. P. Tuinhout, "Threshold voltage mismatch and intra-die leakage current in digital CMOS circuits," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 1, pp. 157–168, 2004.

[68] S. Mukhopadhyay, K. Kim, K. A. Jenkiins, C.-T. Chuang, and K. Roy, "An on-chip test structure and digital measurement method for statistical characterization of local random variability in a process," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 9, pp. 1951–1962, 2008.

[69] S. Nassif, "Delay variability: sources, impacts and trends," in *IEEE International Solid-State Circuits Conference*, pp. 368–369, 2000.

[70] D. Blaauw, K. Chopra, A. Srivastava, and L. Scheffer, "Statistical timing analysis: From basic principles to state of art," *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, vol. 27, no. 4, pp. 589–607, 2008.

[71] Z. Guo, A. Carlson, L.-T. Pang, K. duong, T.-J. K. Liu, and B. Nikolic, "Large-scale SRAM variability characterization in 45nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 44, november 2009.

[72] D. Burnett, K. Erington, C. Subramanian, and K. Baker, "Implications of fundamental threshold voltage variations for high-density SRAM and logic circuits," in *Proceedings of VLSI Technology Symposium*, pp. 15–16, 1994.

[73] G. Shamanna, R. Gaurav, Y. Raghavendra, P. Marfatia, and B. Kshatri, "Using ECC and redudancy to minimize vmin induced yield loss in 6T SRAM," in *IEEE International Conference on IC Design & Technology*, pp. 1–4, 2012.

[74] B. Giraud, O. Thomas, and A. Amara in *Planar double-Gate Transistor*, Springer, 2009.

[75] J. L. Hennessy and D. A. Paterson in *Computer Architecture : a Quantative Approach: Physics and Technology*, Morgan-Kaufman, 1982.

[76] W. A. Wulf and S. A. McKee, "Hitting the memory wall, implications of the obvious," *ACM SIGARCH Computer news*, vol. 23, no. 1, pp. 20–14, 1995.

[77] S. Cosemans, W. Dehaene, and F. Cathoor, "A low-power embedded SRAM for wireless applications," *IEEE Journal of Solid-State Circuits*, vol. 42, july 2007.

[78] H. Yang, R. Wong, R. Hasumi, Y. Gao, N. Kim, D. Lee, S. Badrudduza, D. Nair, M. Ostemayr, H. Kang, H. Zhuang, J. Li, L. Kang, X. Chen, A. Thean, F. Arnaud, L. Zhuand, C. Schiller, D. Sun, Y. Teh, J. Wallner, Y. Takasu, K. Stein, S. Samavedam, D. Jaeger, C. Baiocco, M. Sherony, M. Khare, C. Lage, J. Pape, J. Sudijono, A. Steegen, and S. Stiffler, "Scaling of 32nm low power SRAM with high-K metal gate," in *IEEE International Electron Device Meeting*, pp. 1–4, december 2008.

[79] Z. G. Seng Oon Toh, T.-J. K. Lui, and B. Nikolic, "Characterization of dynamic SRAM stability in 45 CMOS," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 11, pp. 2702–2712, 2011.

[80] S. Clerc, F. Abouzeid, G. Gasiot, D. Gauthier, and P. Roche, "A 65nm SRAM achieving 250mV retention and 350mV, 1MHz, 55fJ/bit access energy, with bit-interleaved radiation soft error tolerance," in *Proceedings of the 38th European Solid-State Circuits Conference*, pp. 313–316, 2012.

[81] F. Abouzeid, S. Clerc, B. Pelloux-Prayer, and P. Roche, "0.42-to-1.20V read assist circuit for SRAMs in CMOS 65nm," in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (3S3)*, pp. 1–2, 2013.

[82] A. Feki, D. Turgis, J. Lafont, and B. Allard, "280 mV sense amplifier designed in 28nm UTBB FD-SOI technology using back-bias control," in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (3S3)*, pp. 1–2, 2013.

[83] N. Verna and A. Chandrakasan, "A 65nm 8t sub-Vt SRAM employing Sense-Amplifier redundancy," in *Proceedings of International Solid-State Circuits Conference Digest of Technical Papers*, pp. 328–606, february 2007.

[84] M. Ya-Qi, J. Zheng, Z.-Y. Zhang, Q.-S. Yao, Y. Wang, and Y.-P. Zhang, "40nm 10T SRAM cell with independent SNM WM and suppress active and leakage power," in *"Proceedings of Solid-State and Integrated Circuit Technology*, pp. 1136–1138, Shangai 2010.

[85] Y.-W. Chiu, M.-H. Tu, J.-K. Zhao, S.-J. Jou, and C.-T. Chuang, "A 40nm 0.32 V 3.5MHz 11T singe-ended bit-interleaving subthreshold SRAM with data-aware write-assist," in *IEEE International symposium on Low Power Electronics and Design*, pp. 51–56, september 2013.

[86] I. Carlson, S. Anderson, and S. N. an A. Alvandpour, "A high density, low leakage, 5T SRAM for embedded caches," in *Proceedings of the 30th European Solid-State Circuits Conference*, pp. 215–218, September 2004.

[87] J.-P. Noel, O. Thomas, C. Fenouillet-Beranger, M.-A. Jaud, and A. Amara, "Robust multi-Vt 4T SRAM cell in 45nm thin BOx fully-depleted SOI technology with ground plane," in *IEEE International Conference on IC Design and Technology*, pp. 191–194, may 2009.

[88] N. Kim, D. Blaauw, and T. Mudge, "Quantative analysis and optimization techniques for on-chip cache leakage power," in *IEEE Transactions on Very Large Scale Integration Systems*, vol. 13, pp. 1147–1156, october 2005.

[89] M. Qazi, M. Tikekar, L. Dolecek, D. Shah, and A. Chandrakasan, "Technique for efficient evaluation of SRAM timing failure," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 21, no. 8, pp. 1558–1562, 2013.

[90] Y.-Y. Chen, S.-Y. Huang, and Y.-C. Chang, "Rapid and accurate timing modelinf for SRAM compiler," in *IEEE International Workshop on Memory Technology, Design, and Testing*, pp. 73–76, 2009.

[91] 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, 1987.

[92] K. Kim, J.-J. Kim, and C. Chuang, "Asymetrical SRAM cells with enhanced read and write margins," in *International Symposium on VLSI Technology and Systems and Applications*, pp. 1–2, 2007.

[93] H. Makino, S. Nakata, H. Suzuki, S. Mutoh, M. Miyama, T. Yoshimura, S. Iwade, and Y. Matsuda, "Reexamination fo SRAM cell write margin definitions in view of predicting the distribution," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 58, no. 4, pp. 230–234, 2011.

[94] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "A 3GHz 70-Mb SRAM in 65-nm CMOS technology with integrated column-based dynamic power supply," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 146–151, 2006.

[95] A.Bhavnagarwala, S. Kosonocky, C. Radens, K. Stawiasz, R. Mann, Q. Ye, and K. Chin, "Fluctuation limits & scaling oppurtunities for CMOS SRAM cells," in *International Electron Device Meeting Technical Digest*, pp. 659–662, november 2005.

[96] K. Takeda, H. Ikeda, Y. Hagihara, M. Nomura, and H. Kobtake, "Redefinition of write margin for next-generation SRAM and write-margin monitoring circuit," in *Proceedings of IEEE International Solid-State Circuits Conference Digest of Technical Papers*, pp. 2602–2611, february 2006.

[97] N. Gierczynski, B. Borot, and N. Planes, "A new combined methodology for write-margin extraction of advanced SRAM," in *Proceedings of IEEE International Conference on Microelectronics Test Structures*, pp. 91–100, 2007.

[98] D. Khalil, M. Khellah, N. M. Kim, Y. Ismail, T. Karnik, and V. K. De, "Accurate estimation of SRAM dynamic stability," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 16, no. 12, pp. 1639–1647, 2008.

[99] O. Thomas, B. Zimmer, B. Pelloux-Prayer, N. Planes, K. C. Akyel, L. Ciampolini, P. Flatresse, and B. Nikolic, "6T SRAM design for wide voltage range in 28nm FDSOI," in *International IEEE SOI Conference*, pp. 1–2, october 2012.

[100] B. Zimmer, S. O. Toh, Y. Lee, O. Thomas, K. Asanovic, and B. Nikolic, "SRAM assist techniques for operation in a wide voltage range in 28-nm CMOS," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 59, no. 12, pp. 1549–7747, 2012.

[101] K. Nii, M. Yabuuchi, Y. Tsukamoto, Y. 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 SRA family with robust read/write stabilizing circuitry under dvfs environment," in *IEEE Symposium on Very Large Scale Integration Circuits*, pp. 212–213, 2008.

[102] K. Agarwal and S. Nassif, "The impact of random device variation on SRAM cell stability in sub-90-nm CMOS technologies," *IEEE Trans-*

actions on Very Large Scale Integration Systems, vol. 16, pp. 86–87, 2008.

[103] C. Millar, D. Reid, G. Roy, S. Roy, and A. Asenov, "Accurate statistical description of random dopant-induced threshold voltage variability," *IEEE Electron Device Letters*, vol. 29, pp. 946–948, 2008.

[104] N. Metropolis and S. Ulam, "The Monte Carlo Method," *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, 2014.

[105] L. C. Andrews in *Special Functions of Mathematics for Engineers*, SPIE–The International Society for Optical Engineering, 1997.

[106] Infiniscale, *ICLYs*. http://www.infiniscale.com/.

[107] O. Kebichi and M. Nicolaidis, "A tool for automatic generation of BISTed and transparent BISTed RAMs," in *Proceedings of the IEEE International Conference on Computer Design*, pp. 570–575, 1992.

[108] S. R. S. Varadhan, "Asymptotic probability and differential equations," *Communications on Pure and Applied Mathematics*, vol. 19, pp. 261–286, 1966.

[109] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, "Breaking the simulation barrier: SRAM evaluation through norm minimization," in *IEEE/ACM International Conference on Computer-Aided-Design*, pp. 322–329, november 2008.

[110] T. Kida, Y. Tsukamoto, and Y. Kihara, "Optimization of importance sampling monte carlo using consecutive mean-shift method and its application to SRAM dynamic stability analysis," in *13th International Symposium on Quality Electronic Design (ISQED)*, pp. 572–579, March 2012.

[111] M. Evans, N. Hastings, and B. Peacock in *Statistical Distribution, 3rd edition.*, ch. Chi distribution, Wiley, 2000.

[112] A. Papoulis in *"Bernouilli Trials". Probability, Random Variables, and Stochastic Processes*, pp. 57–63, McGraw-Hill, 1984.

[113] G. Moritz, B. Giraud, J.-P. Noel, D. Turgis, and A. Grover, "Optimization of a voltage sense amplifier operating in ultra wide voltage range with back bias design techniques in 28nm UTBB FD-SOI technology," in *2013 International Conference on IC Design & Technology (ICICDT)*, pp. 53–56, 2013.

[114] O. Thomas, B. Zimmer, S. O. Toh, L. Ciampolini, N. Planes, R. Ranica, P. Flatresse, and B. Nikolic, "Dynamic single-P-well SRAM bitcell characterization with back-bias adjustment for optimized wide-voltage-range SRAM operation in 28nm UTBB FD-SOI," in *International Electron Device Meeting*, 2014 forthcoming.

[115] K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng, "Random telegraph noise of deep-submicrometer MOSFETs," *IEEE Electron Device Letters*, vol. 11, no. 2, pp. 90–92, 1990.

[116] C. M. Compagnoni, R. Gusmeroli, A. S. Spinelli, A. L. Lacaita, M. Bonanomi, and A. Visconti, "Statistical model for random telegraph noise in flash memories," *IEEE Transactions on Electron Devices*, vol. 55, no. 1, pp. 388–395, 2008.

[117] B. Razavi, "A study of phase noise in CMOS oscillators," *IEEE Journal of Solid State Circuits*, vol. 31, no. 3, pp. 331–343, 1996.

[118] C. Leyris, S. Pilorget, M. Marin, and M. Minondo, "Random telegraph signal noise spice modeling for circuit simulators," in *37th Solid State Device Research Conference, ESSDERC*, pp. 187–190, 2007.

[119] T. G. M. Kleinpenning, "On 1/f noise and random telegraph noise in very small electronic devices," *Physica B: Condensed Matter*, vol. 164, pp. 331–334, 1990.

[120] L. K. J. Vandamme, D. Sodini, , and Z. Gingl, "On the anomalous behavior of the relative amplitude of RTS noise," *Solid State Electronics*, vol. 42, no. 6, pp. 901–905, 1998.

[121] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. Leblanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *Journal of Solid-State Circuits*, vol. 9, pp. 256–268, 1974.

[122] S. O. Toh, *Nanoscale SRAM Variability and Optimization*. PhD thesis, University of California at Berkley, December 2011.

[123] A. Asenov, R. Balasubramaniam, A. R. Brown, and J. H. Davies, "RTS amplitudes in decananometer MOSFETs : 3D simulation study," *IEEE Transactions on Electron Devices*, vol. 50, no. 3, pp. 839–845, 2003.

[124] W. Shockley and W. T. Read in *Statistics of the Recombination of Holes and Electrons*, vol. 87, pp. 835–842, American Physical Society, 1952.

[125] H. Kurata, K. Otsuga, A. Kotabe, S. Kajiyama, T. Osabe, Y. Sasago, S. Narumi, K. Tokami, S. Kamohara, and O. Tsuchiya, "The impact of random telegraph signals on the scaling of multilevel flash memories," in *Symposium on Very Large Scale Integration Circuits*, pp. 125–126, 2006.

[126] N. Tega, H. Miki, T. Osabe, A. Kotabe, K. Otsuga, H. Kurata, S. Kamohara, K. Tokami, Y. Ikeda, and R. Yamada, "Anomalously large threshold voltage fluctuation by complex random telegraph signal in floating gate flash memory," in *Technical Digest of International Electron Device Meeting*, pp. 491–494, 2006.

[127] R. Gusmeroli, C. M. Compagnonim, A. Riva, A. S. Spinelli, A. L. Lacaita, M. Bonanomi, and A.Visconti, "Defects spectroscopy in SiO2 by statistical random telegraph noise analysis," in *Technical Digest of International Electron Device Meeting*, pp. 483–486, 2006.

[128] H. Miki, T.Osabe, N. Tega, A. Kotabe, H. Kurata, K.Tokami, Y. Ikeda, S. Kamohara, and R. Yamada, "Quantitative analysis of random telegraph signals as fluctuations of threshold voltages in scaled flash memory cells," in *Proceedings of International Reliability Physics Symposium*, pp. 29–35, 2007.

[129] K. Abe, S. Sugawa, S.Watabe, N.Miyamoto, A.Teramoto, Y. Kamata, K. Shibusawa, M. Toita, and T. Ohmi, "Random telegraph signal statistical analysis using a very largescaled array TEG with 1M MOSFETs," in *Symposium on Very Large Scale Integration Technology*, pp. 210–211, 2007.

[130] N. Tega, H. Miki, M. Yamaoka, H. Kume, T. Mine, T. Ishida, Y. Mori, R. Yamada, and K. Torii, "Impact of threshold voltage fluctuations due to random telegraph noise on scaled-down SRAM," in *Proceedings of International Reliability Symposium*, pp. 541–546, 2008.

[131] S. O. Toh, Y. Tsukamoto, G. Zheng, L. Jones, K. L. Tsu-Jae, and B. Nikolic, "Impact of random telegraph signals on Vmin in 45nm SRAM," in *International Electron Device Meeting*, pp. 1–4, 2009.

[132] M. Tanizawa, S. Ohbayashi, T. Okagaki, K. Sonoda, K. Eikyu, Y. Hirano, K. Ishikawa, O. Tsuchiya, and Y. Inoue, "Application of a statistical compact model for random telegraph noise to scaled-SRAM Vmin analysis," in *Symposium on VLSI Technology Digest of Technical Papers*, pp. 95–96, 2010.

[133] S. O. Toh, T.-J. K. Liu, and B. Nikolic, "Impact of random telegraph signaling noise on SRAM stability," in *Symposium on VLSI Technology Digest of Technical Papers*, pp. 204–205, 2011.

[134] K. Aadithya, S. Venogopalan, and A. D. J. Roychowdhury, "MUSTARD: A coupled, stochastic/deterministic, discrete/continuous technique for predicting the impact of random telegraph noise on SRAMs and drams," in *Design Automation Conference*, pp. 292–297, 2011.

[135] E. Simoen, B. Dierickx, C. Claeys, and G. Declerck, "Explaining the amplitude of RTS noise in submicrometer mosfets," *IEEE Transactions on Electron Devices*, vol. 39, no. 2, pp. 422–429, 1992.

[136] E. G. Ioannidis, A. Bajolet, T. Pahron, N. Planes, F. Arnaud, R. A. Bianchi, M. Haond, D. Golanski, J. Rosal, C. Fenouillet-Beranger, P. Perreau, C. A. Dimitriadis, and G. Ghibaudo, "Low frequency noise variability in high-k/metal gate stack 28nm bulk and FD-SOI CMOS transistors," in *International Electron Device Meeting*, pp. 18.6.1–18.6.4, 2011.

[137] C. G. Theodorou, E. G. Ioannidis, F. Andrieu, T. Poiroux, O. Faynot, C. A. Dimitriadis, and G. Ghibaudo, "Low-frequency noise source in advanced UTBB FD-SOI MOSFETs," *IEEE Transactions on Electron Devices*, vol. 61, no. 4, pp. 1161–1167, 2014.

[138] M. J. Kirton, M. J. Uren, S. Collins, M. Schulz, A. Karmann, and K. Scheffer, "Individual effects at the si:sio2 interface," *Semiconductor Science and Technology*, vol. 4, no. 3, pp. 1116–1126, 1989.

[139] K. C. Akyel, L. Ciampolini, O. Thomas, D. Turgis, and G. Ghibaudo, "Impact of random telegraph signals on 6t high-density SRAM in 28nm UTBB FD-SOI," in *IEEE European Solid-State Device Research Conference*, 2014.

[140] H.-K. Lim and J. G. Fossum, "Threshold voltage of thin-film silicon-on-insulator (SOI) MOSFET's," *IEEE Transacions on Electron Device*, vol. 30, no. 10, pp. 1244–1251, 1983.

[141] PTC, *Mathcad*. http://www.ptc.com/product/mathcad/.

[142] Mathworks, *MATLAB*. http://www.mathworks.com/.

[143] PDE Solutions Inc., *FlexPDE*. http://www.pdesolutions.com/.