Design and Measurement of Parameter-Specific Ring Oscillators

Lynn Tao-Ning Wang

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2010-159
http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-159.html

December 16, 2010

Design and Measurement of Parameter-Specific Ring Oscillators

by

Lynn Tao-Ning Wang

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

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Andrew R. Neureuther
Professor Tsu-Jae King Liu
Professor Elad Alon
Professor Fiona M. Doyle

Fall 2010

Design and Measurement of Parameter-Specific Ring Oscillators

Abstract

Design and Measurement of Parameter-Specific Ring Oscillators

By

Lynn Tao-Ning Wang

Doctor of Philosophy in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Andrew R. Neureuther, Chair

Electronic monitoring utilizing process-specific Ring Oscillators (RO) is explored as a means of identifying, quantifying, and modeling sources of variation in circuit performance due to manufacturing and layout design parameters. This approach is motivated by the need to mitigate the increased impact of process variability on circuit performance in the scaling of CMOS. To reduce such impact, there is a need to monitor, analyze, and understand process variation in order to improve current design methodologies. This work contains the first measured silicon results for the utilization of parameter-specific modification of ring oscillator layouts to electronically monitor particular process variation. Design and testing for this work were made possible through the Berkeley Wireless Research Center. The working circuits were fabricated by ST Micro in a 45 nm fabrication process that was under development.

The design was based on a process design kit (PDK) provided by ST Micro. The lithography simulation was carried out using generic models in Mentor Graphics Calibre. Five systematic process effects were considered: etch, focus, misalignment, and capping layer and Shallow Trench Isolation (STI) stress. In all cases, inverter layouts were modified in order to increase sensitivity to a particular parameter within design rule constraints. Due to the dependence of layout geometries on parasitic capacitance, it was also necessary to pre-correct measurements for this effect. A total of 32 different inverter layouts were designed.

For monitoring etch, the presence or absence of adjacent dummy gates and pre-correction for residual lithographic effects were used. For monitoring gate focus, spillover functions and Pattern Matching were used as an initial guide to place added surroundings to the gate; this was followed by systematic optimization via Mentor Graphics Calibre simulations. A bias was added to the gate length, but it was assumed that other Optical Proximity Correction (OPC) methods typically used in production, such as scatter-bar insertion, would not be applied. A focus sensitivity of 1.5 times that of a dense gate was achieved. For monitoring gate-to-active misalignment, a set of 5 pre-programmed offsets on an ‘H-shaped diffusion’ was designed. The RO frequency versus offset relationship in design showed a parabolic shape with a speed up of 3.2% for 15nm misalignment. The capping layer and lateral STI stress monitors were designed

based on changes in the lateral size of the source and drain, including those of asymmetrical source/drain areas.

Each RO monitor consists of 13 inverter stages of one of the aforementioned inverter types. The thirty-two inverter layouts were formed into an instantiation that was then repeated 12 times in a local block. This RO block was instantiated 3 times within the 2x2 mm^2 chip such that there are 36 instantiations for each RO monitor type per chip. This chip was merged with other chips to form a large overall field size, ensuring that the printed test chips were likely more than 1 cm apart.

Seventeen chips were received, packaged, and automatically tested. Every ring oscillator worked for every chip. The measured range of the across-wafer variation was 11.1% for control-case RO with minimum sized gate area. Fifteen chips had a small block-to-block variation of ~1.3%. For a given monitor on a typical chip, the variation among the 36 RO instantiations normalized to its mean is 0.2-0.3 %, with the larger value occurring for the smaller gate areas. When these values were multiplied by the square root of the product of 36 instances time 26 transistors per RO, the average threshold slope (AVT) of 2.3mV/um was obtained in an equivalent Pelgrom model.

The measured RO frequency sensitivity to gate focus monitors shows that they are about 4% slower than the control ROs. This decrease is attributed to parasitic effects as well as the non-uniform 'hour-glass' shape produced at the top and bottom of the gate from the horizontal extensions used to increase focus sensitivity. The pre-programmed gate-to-active misalignment monitors show a 2-4 nm overlay error for 17 chips. The fact that the experimental measurements are less sensitive than predicted during the design stage is in part attributed to the fact that the wafer was run under unusually good control without any programmed treatments such as defocus. This observation is supported by the fact that the measured range of RO frequency was typically centered and 1/6-1/4^{th} of the SS-FF guard band. The unanticipated requirement to apply strong OPC techniques with scatter-bars to the monitor designs in order to guarantee that they would not impact product yield also resulted in considerable sensitivity loss. This loss not only occurred for the layout monitors with isolated gates, but also for the active 'H-shape' misalignment monitors where process variation of the active layout including its height and curvature at the off-set gate position was reduced.

The Nitride Contact Etch Stop Liner (CESL) strain-induced monitors show a ring oscillator frequency increase of 5.3% and 13.9% for 1.8X and long length source/drain diffusion (LOD) respectively as compared to minimum LOD, after the normalization of raw data to simulation data so as to correct for parasitic effects. This increase is due to increased CESL-induced strain for large LOD. For the same LOD, asymmetrical designs show a 3% ring oscillator frequency increase for larger source LOD than that of larger drain LOD, indicating transistor injection velocity as well as mobility is important. These layout geometries are simulated by Nuo Xu in order to model CESL's impact on mobility and injection velocity. The measured RO frequencies show that the 45nm devices operate in a regime between the mobility and velocity injection models.

The random variation of RO circuit performance for a given layout monitor within a chip is examined for 3 sources of variations: changes in gate length (ΔL), gate oxide thickness (ΔT_{ox}), and channel doping (ΔN_{ch}). The strategy here is to make a linear approximation of the measured RO frequency sensitivity to these 3 parameters under 5 distinct combinations of operating voltages and temperatures using the 45nm PDK BSIM4/PSP models. The strategy is implemented using least mean square (LMS) analysis. Measured block means were used on one

outlier chip that showed 5% slower RO frequency in a third block. For all of the blocks, the LMS results indicate that the source of random within-chip variation is dominated by random dopant fluctuations in comparison with changes in ΔL and ΔT_{ox}. Since the decrease in RO frequency for the third block was similar for the two focus and control monitor-pairs, the decrease in measured RO frequency is unlikely to be due to a change in focus.

This thesis demonstrates that "parameter-specific ring oscillators" are suitable for multiple critical applications in quantifying systematic and random effects in the co-optimization of process development and circuit design. While parameter-specific RO monitors provide a permanent record of process effect, they are best used during process development and calibration, when less stringent design rules, no-OPC drop-ins, and programmed treatments can be accommodated, yielding inverter layouts with higher as well as directly verifiable sensitivity to process variation.

To my family

Acknowledgements

I am heartily thankful to my research advisor, Professor Andrew R. Neureuther, for his encouragement, supervision, and support from the preliminary work to the concluding chapters of this dissertation. He was always available whenever I needed help, from solving a difficult research problem to rehearsing a conference presentation. I learned a lot from him.

This dissertation is a collaborative effort among the lithography, device, and circuits groups at UC Berkeley. I worked with many professors, and I am grateful for their guidance. Professor Borivoje Nikolic guided me through my first tapeout, provided me with technical feedback on digital circuit designs, and encouraged me to publish at various conferences. Professor Tsu-Jae King Liu gave me technical advice on correlating ring oscillator frequency results to device characteristics and helped me improve upon my technical writing skills. I received a lot of great feedback on my research from Professor Elad Alon. He also taught me how to make better technical presentations. I also appreciated Professor Fiona M. Doyle's comments, which greatly enhanced this dissertation.

I am also very thankful for the help from several colleagues. Liang-Teck Pang taught me about electronic monitoring. Seng-Oon Toh helped and guided me in testing. Nuo Xu and Darsen Lu ingrained in me an immense amount of device knowledge that was necessary for the completion of this dissertation.

This dissertation would not have been possible without the contributions of the students, faculty, and sponsors of the Berkeley Wireless Research Center (BWRC) and the wafer fabrication donation of ST Micro. I would like to make special references to Yorick Trouiller from ST Micro. Without his cooperation, I would not have obtained such relevant data. I really appreciate the sponsors of my research: the UC Discovery Grant ele07-10283 under the IMPACT project and SRC Research Tasks 1443.

I am grateful to my research group members for their friendship and support during the past years: Daniel Ceperley, Eric Chin, Chris Clifford, Marshal Miller, Wojtek Poppe, and Juliet Rubinstein. I thank them for enriching my graduate school life by leading me to get involved with WICSE, PhotoBears, intramural soccer, and rock climbing. In addition, I would like to thank Jae-Seok Yang from Samsung, who changed my life by introducing me to CAD, lithography, and most importantly, Professor Neureuther. I would also like to acknowledge Kenji Yamazoe, a stellar visiting scholar from Canon, who provided me with immense help and support during my qualifying exam as well as the latter part of my graduate school career. His technical advice and encouraging words were invaluable to me.

Lastly, I owe my deepest gratitude to my family for their unconditional love and moral support.

Table of Contents

Acknowledgements ii
Table of Contents iii
Table of Figures vii
Table of Tables xii
Chapter 1: Introduction 1
1.1 Motivation 1
1.2 Dissertation Content and Contributions 2
Chapter 2: Characterizing Variability 4
2.1 Modeling Within-Chip/Across-Chip, Random/Systematic and Dynamic/Static Variability 4
2.2 Sources of Process and Layout-Induced Variation 7
2.2.1 Variations in the Lithography Exposure System 8
2.2.2 Variation in Patterning and Other Manufacturing Steps 11
2.3 Process Characterization Methodologies 12
2.3.1 Frequency 12
2.3.2 Current: $I_{d, SAT}$ and $I_{d, LEAK}$ 13
2.4 CAD Tools for Monitor Design Enablement 13
2.4.1 Parametric Yield Simulator (PYS) 14
2.4.2 Pattern Matching Physics 15
2.5 Design of Gate Lithography Focus Monitors using the Parametric Yield Simulator and Pattern Matching 18
2.6 Summary 22
Chapter 3: 45nm Ring Oscillator Monitors Experimental Design 23
3.1 RO Frequency Sensitivity Analyses for 90nm Test Chip Measurements 23
3.1.1 RO Frequency Sensitivity to Gate Length and Electrical Parasitics 24
3.1.2 RO Frequency Sensitivity to Gate Lithography Dose and Focus 25
3.1.3 Prediction of RO Sensitivity from PYS Lithography Modeling 27
3.2 Ring Oscillator Schematics and Floor Plans for 45nm 28

3.3 Summary of the RO Inverter Layouts and the Nomenclature 30
3.4 Pattern Matching Guidance for Design of the RO Monitors for Lithography 32
3.4.1 A Strehl Ratio Test for Process Variation (PV) Band Calibration 32
3.4.2 Using the Pattern Matcher to Optimize RO Monitor Design 33
3.5 Specific Designs for Etch and Lithography Monitors 34
3.5.1 RO Monitors for Etch Dependence of Poly 34
3.5.2 RO Monitors for Focus Dependence of Poly 35
3.5.3 RO Monitors for Gate-to-Active Overlay Using Pre-Programmed Offsets 36
3.6 Designing Nitride CESL-Induced Strain and STI-Induced Stress RO Monitors 39
3.7 Summary 40
Chapter 4: 45nm Ring Oscillator Testing methodologies and Noise Analysis 41
4.1 Automated RO Frequency Measurement Setup and Testing Methodologies 41
4.1.1 An overview of the RO Frequency Measurement Test Setup 41
4.1.2 Adjustment of Operating Voltages and Temperatures 43
4.1.3 RO Measurement Sequence 44
4.2 Measurement System Noise and Drift 45
4.2.1 RO Measurement Jitter 45
4.2.2 RO Measurement Drift 46
4.3 Measured Block-to-Block Variation 48
4.4 Measured Across-Wafer Variation 50
4.5 Summary 55
Chapter 5: Experimental RO Frequency Sensitivity to Gate Etch, Gate Lithography Focus, and Gate-to-Active Overlay 56
5.1 Simulated RO Frequency Variability using Process Corners and Process Variation Bands 56
5.1.1 Simulated Process Corners versus Measured RO Frequencies 57
5.1.2 PV Bands and SPICE Simulations versus Measured RO Frequency 57
5.2 RO Frequency Sensitivity to Gate Etch 58
5.2.1 Measured RO Frequency Sensitivity to Gate Etch 58
5.2.2 Simulated RO Frequency Sensitivity to Gate Etch using Process Corners 60
5.2.3 Simulated RO Frequency Sensitivity to Gate Etch Monitors using PV Bands and SPICE 62

5.3 RO Frequency Sensitivity to Gate Lithography Focus 62
5.3.1 Measured RO Frequency Sensitivity to Gate Lithography Focus 63
5.3.2 Simulated RO Frequency Sensitivity to Gate Lithography Focus using Process Corners 65
5.3.3 Simulated RO Frequency Sensitivity to Gate Lithography Focus using PV Bands and SPICE 66
5.3.4 Across-Wafer Correlation Plots 67
5.4 RO Frequency Sensitivity to Gate-to-Active Misalignment 68
5.5 Summary 69
Chapter 6: RO Frequency Sensitivity to Nitride CESL and STI-Induced Stress 71
6.1 RO Frequency Sensitivity to Nitride CESL-Induced Strain 71
6.1.1 RO Frequency Sensitivity versus LOD' 72
6.1.2 Measured RO Frequency Distributions for Two Different 45nm Tapeouts 74
6.1.3 RO Frequency Sensitivity Averaged Across 15 Typical Chips versus LOD' 75
6.1.4 RO Frequency Sensitivity to Asymmetrical Source/Drain 75
6.1.5 RO Frequency versus Mobility and Injection Velocity Models 77
6.2 RO Frequency Sensitivity to STI-Induced Stress 78
6.3 Summary 79
Chapter 7: Random Noise Analysis: Gate Length, Gate Oxide Thickness, and Doping Variation 81
7.1 Quantile-Quantile (QQ) and Autocorrelation Plots of RO Frequency Measurements 81
7.1.1 QQ Plots 82
7.1.2 The Autocorrelation Plots 83
7.2 Block-to-Block Analysis for Outlier Chips 84
7.3 RO Frequency Sensitivity to Operating Voltages (V_{dd}) and Temperature (T) 85
7.4 Inferring Process Parameter from Random Sources of Variability using Least Mean Square (LMS) Analysis 87
7.4.1 The Index Definitions 87
7.4.2 The Variable Definitions and RO Frequency Sensitivity to L, T_{ox}, and N_{ch} Extraction 88
7.4.3 The Matrix Definitions 88
7.4.4 The Least Mean Square Solution 89

7.5 Error Analysis for the LMS Solution 92
7.6 Summary 94
Chapter 8 95
8.1 Design 95
8.2 Measurement Methodology 96
8.3 Initial Measurement Results 96
8.4 Monitoring of Gate Etch, Gate Lithography, and Gate-to-Active Misalignment 97
8.5 Monitoring of Nitride CESL and STI-Induced Stress 98
8.6 Random Noise Analysis using Least Mean Squares 98
8.7 Overall Perspective 99
Bibliography 102

Table of Figures

Chapter 2

Figure 2.1: The transistor cross-section and sources of random variability 5
Figure 2.2: A RO frequency measurement fitted across a wafer shows systematic parabolic behavior [13]. 5
Figure 2.3: The typical process-corner design methodology is not well-correlated with layout-induced systematic variation [1]. 6
Figure 2.4: Random telegraph signals (RTS), a dynamic source of variability, cause the threshold voltage, V_{th}, to vary with time [14][15]. 7
Figure 2.5: A typical lithography imaging system [18] 8
Figure 2.6: The DOF concept is a range of lens-wafer distances such that the respective CDs are maintained within specifications. 9
Figure 2.7: Bossung plot shows that DOF and exposure dose control are quite important in maintaining a uniform CD [20]. 10
Figure 2.8: The fabrication steps of poly-silicon gates show that CDs suffer from etch loading effects [1] 11
Figure 2.9: The distribution of mechanical stress beneath the gate that is induced by the contact etch stop liner, which is also commonly referred to as the capping layer [23]. 12
Figure 2.10: Parametric Yield Simulator (PYS) flow [5] 14
Figure 2.11: The Pattern Matcher Concept and Flow [34] 15
Figure 2.12: The above kernels in (a) are computed by taking the inverse Fourier Transform of the corresponding Zernike terms [38], while (b) is an example of a focus monitor. 16
Figure 2.13: A typical lithography system. 16
Figure 2.14: In the Pattern Matcher, electric field spillover is combined onto an observation pixel (X_o) from surrounding pixels (X_{Si}), as shown in (a) via a convolution computation. In (b), examples of different kernels matched onto a point on the poly (red) layout are shown. 18
Figure 2.15: The post-OPC simulated CD shift of 80nm gates for four different levels of defocus reveals that some pitches exhibit high sensitivity to focus while others are insensitive to focus. 19
Figure 2.16: Bossung plot of $L_{effective}$ vs. DOF across different dose conditions for an 80nm isolated line [5] 19
Figure 2.17: A simulated aerial image using PYS for a reticulated monitor is shown in (a), and the monitor's response through focus and dose is illustrated in (b) [5]. 20
Figure 2.18: A simulated aerial image of a Zernike aberration monitor using PYS is shown in (a), and the monitor's response through focus and dose is given in (b). Since the center 90^O probe region is hyper-sensitive to defocus, the gate CD is expected to change dramatically in response to the defocus in the scanner [5]. 21

Chapter 3

Figure 3.1: The RO Frequency versus the Channel Length simulations for three monitors show that the presence or the lack of dummy gates has little effect on extracted results for a drawn gate length. The sensitivity at (1,1) is -1.8. 24
Figure 3.2: Sensitivity to the individual NMOS or PMOS device is half as big as the sensitivity to both devices. 25
Figure 3.3: A simulated plot of the $L_{effective}$ versus defocus for different dose conditions for an *isolated* 90nm gate shows a $\Delta L_{effective, focus, NMOS}$ =5.5%. 26
Figure 3.4: A simulated plot of the $L_{effective}$ versus defocus for different dose conditions for a *dense* 90nm gate shows a $\Delta L_{effective, focus, NMOS}$ =3.0%. 27
Figure 3.5: Tile 3a's simulated plot of ring oscillator frequency versus defocus for different dose conditions is used to examine the measured across-wafer RO frequency spread..... 27
Figure 3.6: A simulated plot of the RO frequency versus defocus for different dose conditions shows that the pattern dependent shift among monitors Tiles 3a, 4a, and 5a is likely due to a non-lithographic source, such as etch. 28
Figure 3.7: The schematic of a ring oscillator tile arrayed in a row is shown in (a). A 2-D array with a set of vertical and horizontal scan-chains along the perimeter in order to enable row and column select can be seen in (b) [1]..... 29
Figure 3.8: The layout view of the 45nm ring oscillator array is shown in (a). Here, (b) shows the layout in (a) replicated at 3 different locations across the 2x2 mm^2 chip in order to capture the variation due to the lithography scanner's slit and scan directions. 30
Figure 3.9: Summary of inverter layouts 31
Figure 3.10: A Strehl ratio test for a light square measuring (0.4)λ\NA per side at tophat and annular illumination using Calibre simulation..... 33
Figure 3.11: (a) shows a coherent defocus MTLP matched at the poly layer of the layout. (b) shows a defocus MTLP with annular illumination that matches at the diffusion layer of the layout..... 34
Figure 3.12: Etch RO monitors are designed to show nearly identical changes in gate length due to defocus. 35
Figure 3.13: (a) and (c) contain the control RO monitors, and (b) and (d) contain the 1.5x focus sensitive RO monitors at 40nm and 60nm, respectively..... 36
Figure 3.14: (a) shows the simulation contours for a minimum-sized active area. (b) presents the H-shaped active area gate-to-active misalignment RO monitor. 37
Figure 3.15: (a) shows the simulation contours for a minimum-sized active area. (b) presents the wedge-shaped active area gate-to-active misalignment RO monitor. 37
Figure 3.16: Normalized width versus pre-programmed offsets for the H-shaped monitors at 0nm, +10nm, and +15nm, shown in (a) and for the wedge-shaped monitors at 0nm, +5nm, and +10nm, shown in (b) 38

Figure 3.17: This plot of normalized ring oscillator frequency as a function of gate width, with and without capacitive load, shows that added capacitors increase the slope from 0.2 to 0.8. 39

Chapter 4

Figure 4.1: Measurement setup [1] 42
Figure 4.2: Chip and PCB Photo..... 42
Figure 4.3: Test setup for operating voltage (V_{dd}) experiments: the on-board regulators in (b) are adjusted for different voltage settings using a screw driver. 43
Figure 4.4: Test setup for temperature experiments 44
Figure 4.5: RO monitors are measured sequentially from 1-36..... 44
Figure 4.6: The number of samples per RO in one RO block for a typical chip ranges from 120 to 140..... 45
Figure 4.7: The maximum standard deviation to mean percentage is 0.009%. 46
Figure 4.8: Maximum percentage difference between the measured and re-measured data without intervention for a chip is 0.14%. 47
Figure 4.9: The percentage difference of measured and re-measured data for a chip is greater if the test setup is reassembled. 48
Figure 4.10: The measured RO frequency for 36 RO instantiations for one typical chip, that of Chip 8. 49
Figure 4.11: The measured RO frequency normalized to the mean for that chip shows random within-block variation. 49
Figure 4.12: The mean measured RO frequency normalized to the mean for the RO monitor designs CS40, CW40, IW50, and IW60 show an across-wafer variation range of 11.11%, 8.84%, 7.35%, and 5.99%, respectively. 51
Figure 4.13: Within-chip $(\sigma/\mu)_{36}$ distribution shows that outlier chips 4 and 9 are noisier when compared to typical chips. 51
Figure 4.14: RO monitors employing smaller-area inverters show greater variation in regard to measured RO frequency..... 52
Figure 4.15: Simulated RO frequency versus threshold voltage for RO monitors CW40, IW50, IW60, and CS40 show a slope of -2.51/mV, -1.98/mV, -1.94/mV, -1.75/mV, respectively..... 53
Figure 4.16: The Pelgrom plots for three 45nm technologies 54

Chapter 5

Figure 5.1: The SPICE corner simulations correlate with the mean measured RO frequency..... 57
Figure 5.2: PV band and SPICE simulation results for CW40 (a) and IW60 (b). 58
Figure 5.3: The mean measured RO frequencies reveal that the dummy polys reduce RO frequency by 3%. The across-wafer peak-to-peak variation for IW60, EW60 and CW60 are ~6-7%. 59

Figure 5.4: The measured $(\sigma/\mu)_{36}$ which excludes outlier chips 4 and 9 shows that the $(\sigma/\mu)_{36}$ is ~0.2%. 59
Figure 5.5: Systematic block-to-block effects are evident in the outlier chips: 6 and 9. 60
Figure 5.6: A comparison of measured and process-corner simulated RO frequency for CW60, EW60, and IW60, respectively. 61
Figure 5.7: The mean measured RO frequency for the 40nm drawn gate length shows that the hammerheads result in about a 4% lower RO frequency. The across-wafer peak-to-peak variation for the control (CS40) and focus monitor (FS40) vary by <1%, when outlier chips 4 and 9 are removed. 63
Figure 5.8: The mean measured RO frequency for 60nm drawn gate length shows that the hammerheads result in a 4% lower RO frequency. The across-wafer peak-to-peak variation for the control (IW60) and focus monitor (FW60) varies by <1%. 64
Figure 5.9: The measured $(\sigma/\mu)_{36}$ is larger for RO monitors with a smaller gate area. Additional hammerheads have negligible impact on $(\sigma/\mu)_{36}$. 64
Figure 5.10: The measured RO frequency for CS40 and FS40 across 17 chips compared with the simulated RO frequency reveals that the change in RO frequency between layouts is due to mainly parasitic effects from the gate. 65
Figure 5.11: The measured RO frequency for IW60 and FW60 across the 17 chips compared with the simulated RO frequency shows that the change in RO frequency between layouts is due to parasitic effects. 66
Figure 5.12: Correlation plots for FS40 versus FW60 (a), FW60 versus IW60 (b), CW60 versus IW60 (c), and $NW40_2$ versus $NW40_4$ (d) are presented. 68
Figure 5.13: These H-Shaped misalignment RO monitor measurement results for chips 5, 8 and 13, show a 1.5% change in RO frequency. 69
Figure 5.14: The deduced overlay error is 2-4nm across the wafer. 69

Chapter 6

Figure 6.1: The mean measured RO frequency versus LOD' for 17 chips (excluding asymmetrical monitors) 73
Figure 6.2: The measured $(\sigma/\mu)_{36}$ for NW40 $_{(1,2,3,4,6,7)}$ is 0.2%, excluding the outlier chips. 74
Figure 6.3: The mean measured RO frequency of 15 typical chips versus LODs (excluding outlier Chips 4 and 9). 75
Figure 6.4: For the same LOD', asymmetrical monitors with greater sources operate 3% faster than monitors with bigger drains. 76
Figure 6.5: For the same LODs, asymmetrical and symmetrical monitors show minor differences in measured RO frequency, indicating that the impact of the drain on device performance is small. 77

Figure 6.6: The percentage change in RO frequency versus LOD' shows that the 45nm ST Micro devices operate in a regime between that predicted by traditional (mobility) and injection velocity models. 78
Figure 6.7: Across-wafer RO frequency measurements for STI monitors show that shifting NMOS down by 0.3um has a negligible effect, while shifting PMOS up by 0.3um speeds up the RO frequency by 3%. 79

Chapter 7

Figure 7.1: These QQ plots for RO Monitor CW40 show that Chip 9 is an outlier chip because the tails of its measured RO frequency distribution deviate from the reference line. 83
Figure 7.2: The lack of correlation among measured RO frequencies for RO monitor CW40 on the 4 chips shows that there is little spatial correlation among the RO monitors. 84
Figure 7.3: The normalized mean measured RO frequency at V_{dd} =0.8V to 1.5 V (a) and T=20^{O}C to 90^{O}C (b) for the 4 chips show that RO frequency is more sensitive to changes in V_{dd} than in T. 86
Figure 7.4: The measured $(\sigma/\mu)_{36}$ from RO frequency at V_{dd}= 1.1V to 1.5V and T =20^{O}C to 90^{O}C for the 4 chips shows that $(\sigma/\mu)_{36}$ saturates with increasing V_{dd} and remains relatively constant as the temperature increases. 87
Figure 7.5: The percent change in L, T_{ox}, and N_{ch} versus RO instantiation for Chip 9 RO monitor CW40. Maximum N_{ch} variation is ~+ 6%. Block-to-block variation (every 12 RO) is present... 90
Figure 7.6: The percent change in L, T_{ox}, and N_{ch} versus RO instantiation for Chip 9 RO monitor CW60. Maximum N_{ch} variation here is + 4%, and block-to-block variation (every 12 RO) is present. 91
Figure 7.7: The differences between measured RO frequency and the estimated solutions show a magnitude on the order of 10^{-3}. 92
Figure 7.8: The simulated RO frequency deviates from that of the measured as V_{dd} deviates from the standard operating voltage, 1.1 V. 93
Figure 7.9: The measured and simulated RO frequency show good correlation with increased operating temperature. 94

Table of Tables

Chapter 2

Table 2.1: Across-chip and within-chip variation for 90nm and 45nm ST Technology [1] 4
Table 2.2: The layout-induced variation reported for 90nm and 45nm technology nodes [1] 6
Table 2.3: Design-parameter variation is caused by process-induced or layout-dependent effect. 7
Table 2.4: Relevant layout-induced variations 8
Table 2.5: A summary of the CAD tools used in this work..... 14
Table 2.6: A summary of $\Delta L_{effective}$ for gate lithography monitors..... 22

Chapter 3

Table 3.1: Nomenclature used for RO monitors..... 30
Table 3.2: The block map for the RO monitors—RO monitors in (*) have additional capacitors inserted between each inverter stage..... 32

Chapter 5

Table 5.1: A comparison of etch monitors for two technology nodes..... 62
Table 5.2: The comparison of simulated and measured across-wafer variation shows good agreement. 62
Table 5.3: A comparison of simulated versus measured across-wafer variation..... 67

Chapter 6

Table 6.1: The normalized layout dimensions for nitride CESL-induced strain RO monitors 72
Table 6.2: A comparison of RO monitors repeated at the 45nm processes for two tapeouts..... 74
Table 6.3: The normalized layout dimensions for STI-induced stress monitors 78

Chapter 7

Table 7.1: A comparison of block-based mean for focus, etch, and stress monitors and their respective control ROs for chips 4 and 9. 85
Table 7.2: Chip 9's measurement conditions and its corresponding sensitivity to process parameters for RO monitor CW40..... 89
Table 7.3: Chip 9's measurement conditions and its corresponding sensitivity to process parameters for RO monitor CW60..... 90

Chapter 8

Table 8.1: A summary of RO frequency sensitivity for parameter-specific monitors..... 100

Chapter 1
Introduction

1.1 Motivation

For the past 30 years in the semiconductor industry, advances in CMOS technology have been occurring at an exponential rate, with density doubling as cost per function has been decreasing by approximately 50% every two years. However, variation in electrical performance relative to mean frequency has been increasing with the introduction of smaller feature sizes. For example, the measured 3-standard-deviation-to-mean-ratio has doubled from 90nm to 45nm technology [1]. The introduction of advanced techniques like stress-enhancements, and advanced process treatments, such as laser-assisted annealing, has resulted in the increasing importance of layout-induced variability [2] [3]. Thus, variability has become a challenge to further CMOS scaling and to maintaining high yield.

Monitoring and characterizing variability can have an important practical impact. Circuit designers are concerned about their most vulnerable circuits and the sources that are contributing to their loss of operating margins. A good example is the Static Random Access Memory (SRAM) and the asymmetries observed in the write margins. Understanding the causes of performance variation would allow circuit designers to implement circuit solutions to reduce systematic variation in their designs. Currently, process-induced circuit performance variation is controlled by ensuring that designs obey design rule checks in layout and satisfy process corner models in simulations. In typical design methodologies, obeying these two check points results in a working chip. However, the process corner approach incorporates all manner of variation, including that of across chip, which may be overly pessimistic.

The impact of variability on circuit performance needs to be quantitatively monitored, measured, and characterized in order to reduce its magnitude and negative effect on yield. Since they reflect circuit performance variation caused by systematic and random physical effects, ring oscillators are good monitors of variability [4]. Ring oscillators enable real-time on-chip monitoring with only a few contact pads and are small in terms of area, simple to design, and easy to implement. When used in conjunction with on-chip scan-chains and multiplexers, ring oscillator frequency can be measured easily, rapidly and accurately.

A natural generalization of the ring oscillator automated measurement approach is that of introducing layout modifications of the inverter gates that enhance the sensitivity of the electrical performance to particular process parameters. A suitable starting point for this enhancement is to use the physical understanding of fabrication processes to induce systematic levels of a given process effect through layout design. The understanding of physical effects must be moved up from the wafer to the layout level, where various layout choices can be made to increase sensitivity to a given process parameter while attempting to reduce the response to other parameters. The choice of layout changes can be guided by computer-aided-design (CAD) simulation tools. This strategy has been used to design individual NMOS devices as process defocus monitors [5]. An electronic circuit in the form of ring oscillators designed with this methodology that can be automatically measured is introduced in this thesis and termed “Parameter-Specific Ring Oscillators”.

This work is a synergy of device/process, circuits, and computer-aided-design (CAD). Device/process physics and CAD tool simulation are used to design and tune the layouts to be sensitive to a specific process parameter. These layout monitors are then implemented into a 13-stage ring oscillator circuit platform designed by Pang to enable automatic electronic characterization to take place [1]. These designs are taped out through the Berkeley Wireless Research Center (BWRC) and are fabricated at ST Micro. Measured RO frequencies show promising sensitivity to process parameters, correlation with SRAM read-write performance asymmetries and new device models (including stress effects) developed by Xu [6][7].

More specifically, this dissertation explores the feasibility of electronically diagnosing process-induced circuit performance variation that is caused by five physical causes: lithography focus, lithography gate-to-active overlay, gate etch, nitride CESL-induced stress, and STI-induced stress. A set of 32 different layout monitors were designed using ST Micro's 45nm Process Design Kit (PDK) and process simulations using Mentor Graphics Calibre. The layout dimensional changes induced by variation in process conditions are extracted using the process simulator. These extracted changes then serve as inputs into the SPICE simulators in order to predict ring oscillator frequency sensitivity for specific process parameters.

A 45nm test chip was designed and fabricated in collaboration with ST Micro to characterize and quantify the effectiveness of parameter-specific ring oscillator layout monitors. The 36 ring oscillator instantiations were made on a small 2x2 mm^2 chip in 3 blocks of 12 replications for each of the 32 layout monitor designs. Ring oscillator frequencies for 17 different chips on one wafer were electronically measured and interpreted for parameter specificity. These measurements are used to identify the sources of circuit performance variation as well as to quantify systematic, residual random, within-chip, and across-chip variation.

1.2 Dissertation Content and Contributions

Chapter 2 details the literature and work most relevant to this dissertational research. It begins by describing components of variability and their impact on circuit performance. It explains the sources of systematic variation in manufacturing processes and layout dependencies. It also discusses why Ring Oscillators (RO) are good process-variation monitors and why fast-CAD tools like Pattern Matching are needed to facilitate the design of parameter-specific inverter layout for ring oscillators. Lastly, this chapter reports on lithography focus layout monitors designed for CDSEM and electric-probe monitoring.

Chapter 3 investigates the challenge of how off-axis illumination likely influences pattern-dependent effects. An extension to Pattern Matching for inclusion of off-axis illumination is developed and applied to the design of defocus inverter layout monitors for ring oscillators. Here defocus sensitivities are confirmed with the aid of Mentor Graphics Calibre process variation (PV) band simulation. This chapter also explores the design challenge of obtaining a 2-4 times focus sensitivity as compared to control-case ROs within design rule checks (DRC). Through an iterative design cycle of DRC, Pattern Matching, and Calibre simulation, RO monitors with 1.5x gate lithography focus sensitivity improvement compared to the control are achieved.

Chapter 4 is devoted to the study of measured noise in ring oscillator frequency caused by the automated testing methodologies and variation due to systematic across-wafer and random within-chip sources for control-case RO monitors. This chapter justifies expedient data analysis assumptions that will be utilized in later chapters. One main challenge delved into in this chapter is that of determining the source of random within-chip variation. Analysis shows that, since

measured noised correlates with the inverse of square root of the channel area, within-chip variation is likely due to electrical fluctuations in the devices themselves. Measured RO frequency standard deviation to mean ratio is converted to a Pelgrom plot that shows an average threshold slope (AVT) of 2.3, with respect to the inverse square root of the channel area.

Chapter 5 reports on and analyzes the measured RO frequency sensitivity to gate etch, gate lithography focus, and gate-to-active misalignment. The main challenge here is that of determining the sources of discrepancies between simulations and measured RO frequency sensitivity. For such detective work to occur, various methodologies are used, such as comparing measured RO frequency sensitivity to process corner simulations, comparing the process variation (PV) bands generated by the generic lithography models to those provided by ST Micro, and comparing the measured results with SPICE-simulated RO frequency sensitivities mapped from PV bands. Measured RO frequency sensitivity shows a decrease in gate and active focus effects when compared to that of the generic simulations. This is in part attributed to the fact that the wafer appears to have been run under unusually good control. The measured range of RO frequency is typically centered and 1/6-1/4th of that of the SPICE-simulated Slow Slow-Fast Fast (SS-FF) guard band.

Chapter 6 describes the measured ring oscillator sensitivity to Nitride CESL-strain and STI-stress induced variation, including the effect of asymmetrical source/drain active areas. One of the main challenges in determining the impact of stress/strain on RO performance lies in isolating the stress-induced effects from that of the parasitic effects. Measured RO frequencies are thus normalized with that simulated from SPICE using the 45nm PDK provided by ST Micro to extract the parasitic effects. Normalized measured RO frequencies show that a 13.9% increase in RO frequency can be achieved with long source/drain active area as compared to that of the minimum area. RO monitors with larger sources than drains operate 3% faster than the RO with larger drains than sources. This observation indicates that the impact of injection velocity is present in 45nm devices. Measured and simulated RO frequency sensitivities are compared, in which the simulated sensitivities are predicted using a new RO delay model developed by Xu with the inclusion of stress, mobility, and injection velocity effects [7].

While the previous two chapters report on and analyze measured RO frequency sensitivity, Chapter 7 shows the challenge of isolating process parameters that cause random residual effects. Thus, the main contribution of this chapter is the application of the least mean square (LMS) approach to isolate these sources. This methodology assumes that the main sources of variation are likely attributed to changes in 3 parameters: gate length, gate oxide thickness, and doping. Since ring oscillators are expected to show different sensitivity under varying operating conditions, in order to improve the accuracy of this approach, RO monitors are measured under 5 combinations of varying temperatures and operating voltages. Using SPICE simulations, the methodology first linearizes RO frequency with respect to a process parameter. Then, using LMS, the measured RO frequency is related to these parameters, and the relative contribution of each parameter is determined.

Chapter 8 summarizes the contributions of this dissertation and offers perspectives in regard to the future work. A consolidated overall summary of simulated and measured parameter-specificities is presented.

Chapter 2
Characterizing Variability

Characterizing the impact of variability is quite important in guaranteeing working chip designs. This literature review identifies the current circuit performance issues, describes the techniques for quantifying measured variability, and gives an overview of process-aware computer-aided-design (CAD) simulation tools. This chapter is organized into 6 sections. Section 2.1 describes the components of variability and their impact on circuit performance. Section 2.2 reports on the known sources of variability from the perspective of semiconductor manufacturing. Section 2.3 discusses the various techniques available to monitor variability as well as the circuits developed for monitoring variability. Section 2.4 reports on the CAD tools that enable the design of layout monitors to occur, while Section 2.5 gives examples of single transistor NMOS gate lithography focus monitors that are designed using a combination of the CAD tools that are described in Section 2.4. Finally, Section 2.6 summarizes the chapter.

2.1 Modeling Within-Chip/Across-Chip, Random/Systematic and Dynamic/Static Variability

Process variability is categorized based on intra-chip (within-chip) and inter-chip (across-chip) distinctions. These variations are caused by random or systematic physical sources. Circuit performance variability caused by random variation is non-deterministic, and the performance could change rapidly over small distances (i.e. no correlation length). Circuit performance variability caused by systematic variation induces deterministic shifts, which vary gradually over a wafer and even within a chip. Table 2.1 presents an example of quantitative within-chip and across-chip variation (3 standard deviation, σ, to mean, μ, ratio) for 90nm and 45nm ST Micro Technology [1].

	90nm	**45nm**
Across-Chip (3σ/ μ)	15%	15%
Within-Chip (3σ/ μ)	3.5%	6.6%

Table 2.1: Across-chip and within-chip variation for 90nm and 45nm ST Technology [1]

Both within-chip and across-chip variation are spatially correlated, i.e. circuit performance variability is highly dependent on the location of the circuit with respect to the position of the chip or on the wafer. The length of scale for systematic effects is process dependent here. Lithography variation (other than layout dependencies), is about 1mm, while the length of scale for Chemical Mechanical Polishing (CMP) is about 40-100um [8].

Random variation is unpredictable and is attributed to device parameter fluctuation. Atomic-scale variation causes random non-uniformity in gate length (L), gate oxide thickness (T_{ox}), and dopant density in the channel region (N_{ch}) [9][10][11]. With scaling and, thus, a decrease in channel area, the number of random dopant atoms decreases, exacerbating the random dopant fluctuation (RDF) impact and causing larger variation in the threshold voltage, V_{th} [10]. The Pelgrom model, which correlates standard deviation in V_{th} with the inverse square root of the gate area, is commonly used to characterize random variation [12].

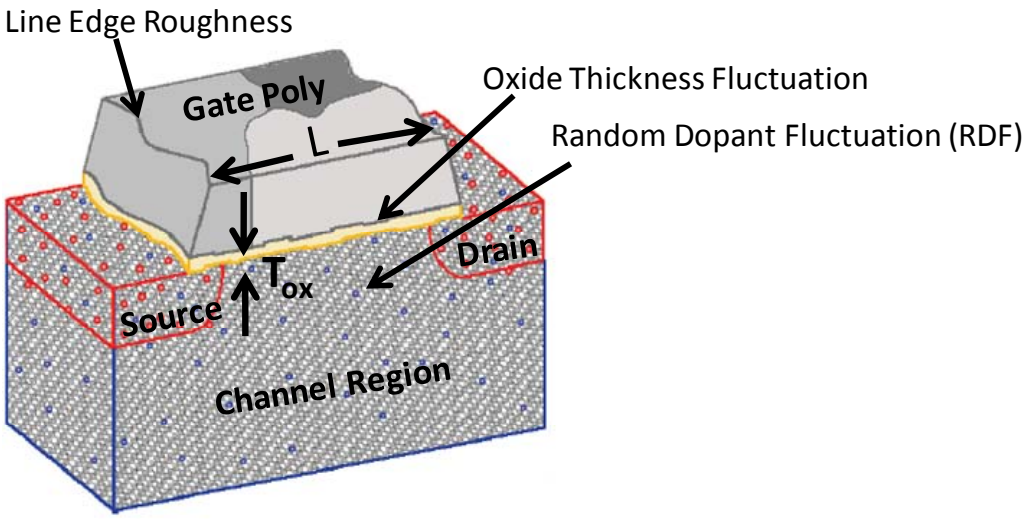


Figure 2.1: The transistor cross-section and sources of random variability

Much of systematic variability is predictable and is attributed to non-uniformity in the following manufacturing steps: lithography/etch, deposition/oxide growth, chemical-mechanical polishing (CMP), and rapid thermal annealing [3]. The evidence of systematic variability is shown in Figure 2.2 by plotting the measured RO frequency versus the position on the wafer for a 45nm technology node [13]. Recently, layout dependent variation from stress-enhancement process technology, such as tensile or compressive Contact Etch Stop Liner (CESL) and Shallow Trench Isolation (STI) have become dominant sources of systematic variations [1].

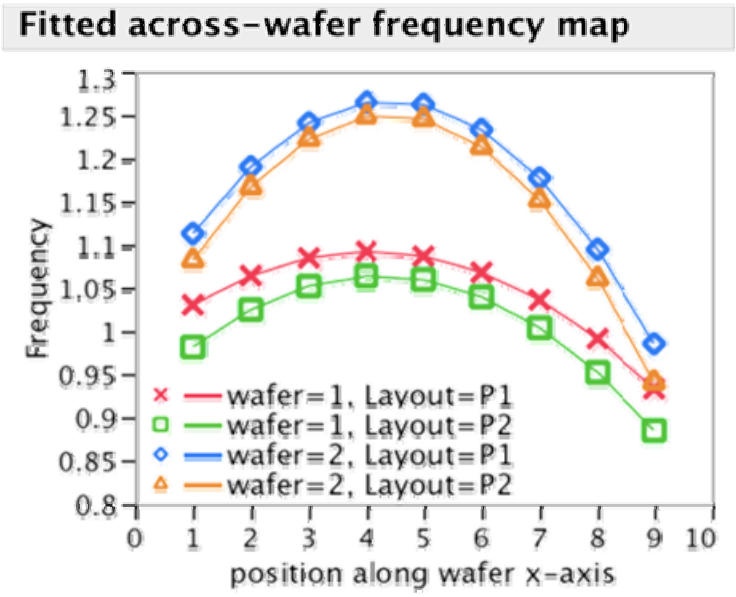


Figure 2.2: A RO frequency measurement fitted across a wafer shows systematic parabolic behavior [13].

While systematic variation is generally predictable, it is usually unknown at design time or is too complex to be included in circuit simulations in current design methodologies. Thus, current designs are typically created to satisfy worst-case corners, consisting of the total within

and across-chip variation. Process-corner design methodology is often too pessimistic and does not correlate well with layout-induced systematic effects. Figure 2.3 illustrates the problematic points associated with the current design methodology.

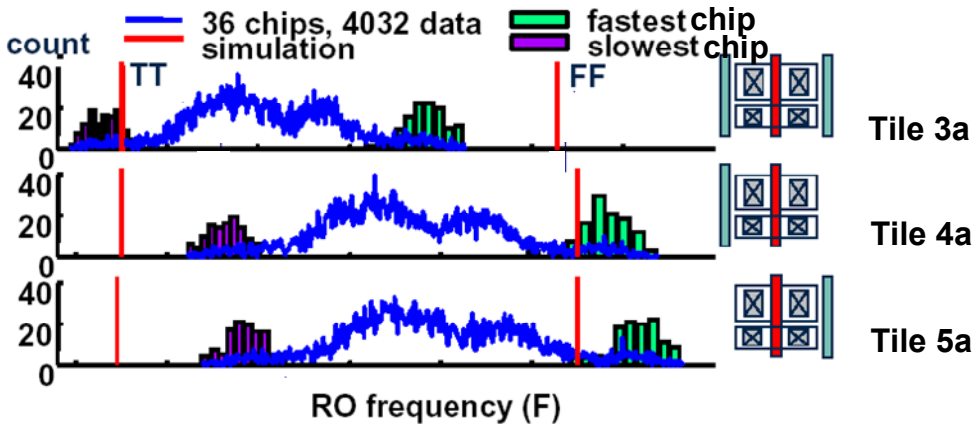


Figure 2.3: The typical process-corner design methodology is not well-correlated with layout-induced systematic variation [1].

Figure 2.3 reports the results from the 90nm testchip for RO monitors: Tiles 3a, 4a, and 5a [1]. Tile 3a has two dummy polys, while Tiles 4a and 5a have one dummy poly. The SS-TT-FF corner model is rather conservative in suggesting a spread that is twice that of the across wafer measurement. In addition, this model fails to capture the layout-dependent effects. Compared to the measured RO frequency distribution for Tile 3a, as the 10% RO frequency increase is the same for the asymmetric monitor Tiles 4a and 5a, it is unlikely that this result is due to an imaging effect, such as coma. Process variation band simulations of the gate that was imaged through focus shows a change of 4%. This observation is much smaller than the measured 10% shift, indicating that this increase is unlikely to be due to a lithography error (Chapter 3).The measured across-chip RO frequency $3(\sigma/\mu)$ spread is ~20% and is 5x wider than the fastest/slowest chip. Such a result shows that across-wafer systematic effects dominate over within-chip random distribution.

In addition to the measured RO frequency sensitivities to the three monitors, which is described in Figure 2.3, the systematic results from the screening of the Ring Oscillator (RO) layout monitors developed by Pang for 90nm and 45nm ST Micro Technology are summarized in Table 2.2 [1]. At 90nm, the main source of variation is attributed to etch/lithography, while at 45 nm the dominant source of variation is attributed to a nitride CESL-induced strain.

Layout	**90nm**	**45nm**
Source /Drain Stress	N/A	Δf_{RO} ~5%
STI	N/A	Δf_{RO} ~3%
Etch/Lithography	Δf_{RO} >10%	Δf_{RO} ~2%

Table 2.2: The layout-induced variation reported for 90nm and 45nm technology nodes [1]

Variability can also be categorized as being dynamic or static. Dynamic variation is time-varying and changes with respect to the utilized operating conditions. Sources that contribute to dynamic variation include: Random Telegraph Signals (Figure 2.4), Negative Bias Temperature

Instability (NBTI) [16], temperature, a voltage drop across the wire resistance, or capacitive coupling. Static variation is independent of the time and operating conditions. Static process variation is studied in this research.

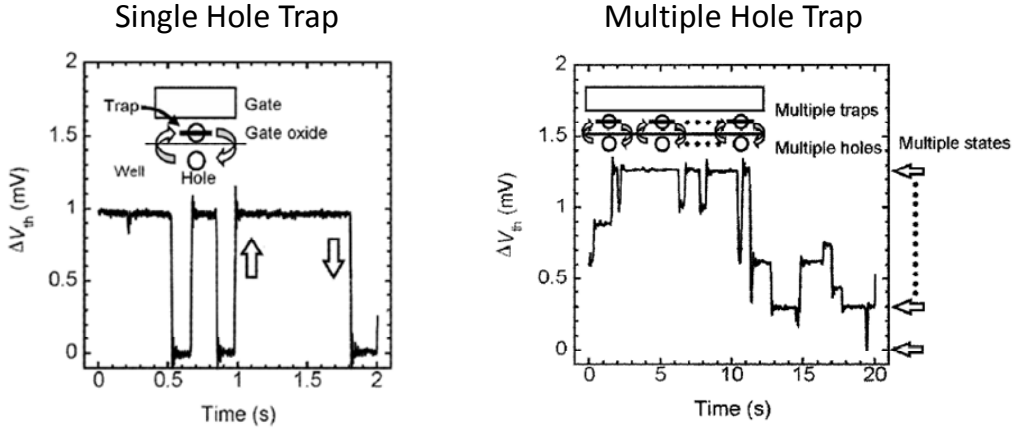


Figure 2.4: Random telegraph signals (RTS), a dynamic source of variability, cause the threshold voltage, V_{th}, to vary with time [14][15].

2.2 Sources of Process and Layout-Induced Variation

Circuit design parameters are affected by sources of variation from the manufacturing steps and layout geometries. The design parameters reported in this work are: gate length (L), gate oxide thickness (T_{ox}), threshold voltage (V_{th}), and mobility/injection velocity (μ/ ν_T). Table 2.3 summarizes the systematic sources that induce design parameter variability.

Circuit Design Parameter	**Sources of Variability from Manufacturing**
Gate length (L) and Gate width (W)	Lithography, etch, layout-dependent proximity effects, overlay error
Gate Oxide Thickness (T_{ox})	Oxide growth non-uniformity
Threshold Voltage (V_{th})	Non-uniformity in anneal temperature, channel area variations
Mobility/injection velocity (μ/ ν_T)	Layout dependency in the source/drain area, non-uniformity in strain/stress distribution from source/drain asymmetries

Table 2.3: Design-parameter variation is caused by process-induced or layout-dependent effect

The main sources of process variation that are relevant to this work are described in the following sections. Section 2.2.1 discusses the possible variability sources in lithography, in particular exposure dose, focus, and misalignment. Section 2.2.2 discusses the sources of variability that are induced by other manufacturing steps, such as chemical mechanical polishing, stress engineering, and etch. A summary of the relevant process variation is described in Table 2.4.

Process Variations	Description
Etch	Etch effects refer to the dependence of gate length on etch loading.
Focus	Focus effects refer to the dependence of gate length on lens-wafer distances.
Misalignment	Misalignment effects refer to the dependence of gate length and gate width due to a gate-to-active overlay error.
Stress/Strain	In 45nm, due to non-uniform stress in the silicon substrate, the use of strained-silicon technology introduces mobility and injection velocity variation. The stress that is induced is layout-dependent here. Sources of stress include: shallow trench isolation and a strained capping layer.

Table 2.4: Relevant layout-induced variations

2.2.1 Variations in the Lithography Exposure System

The two key requirements for lithography in manufacturing integrated circuits are critical dimension (CD) and overlay control [17]. Many different feature sizes in layout designs must remain precise within each field, across a wafer, and from wafer to wafer. Maintaining CD and overlay control are quite important to the overall transistor performance. Variability in CD is caused by two main physical sources: (1) the depth of focus (DOF) and (2) the exposure dose. Variability in misalignment is due to the error of aligning a reticle to the features on the wafer during lithography exposure steps.

A typical lithographic imaging system (Figure 2.5) is governed by two fundamental quantities: the resolution and the depth of focus (DOF).

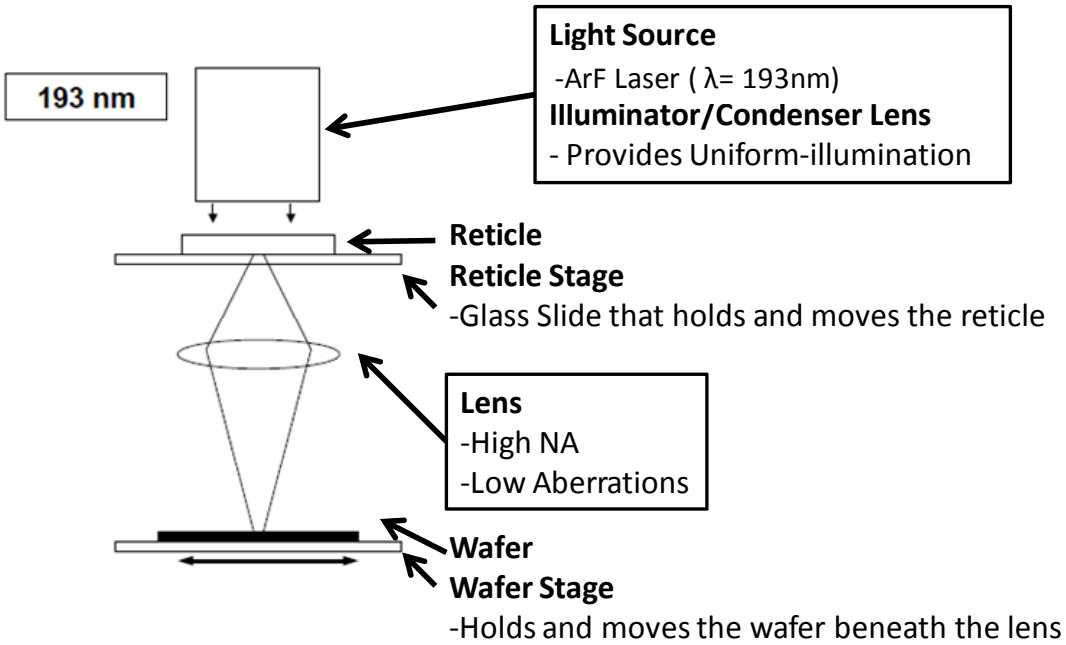


Figure 2.5: A typical lithography imaging system [18]

The smallest resolution achieved, or the smallest printable critical dimension, is defined by the Raleigh criterion and the resist process:

$$CD = k_1 \frac{\lambda}{NA} \tag{2.1}$$

where λ is the wavelength, NA is the numerical aperture of the lens, and k_1 is the "k-factor" for a given process. The k-factor changes if a different photoresist process is being used.

The DOF is a range of lens-wafer distances over which the CDs and resist profiles are maintained within specifications.

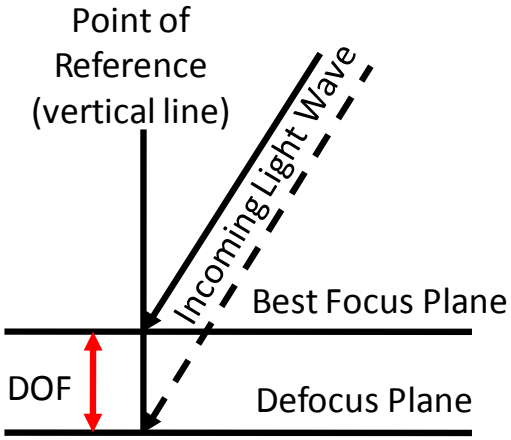


Figure 2.6: The DOF concept is a range of lens-wafer distances such that the respective CDs are maintained within specifications.

A rigorous derivation of the DOF is defined as:

$$DOF = \frac{1}{4}\frac{\lambda}{n}\frac{1}{1-\sqrt{1-\left(\frac{NA}{n}\right)^2}} \tag{2.2}$$

where λ is the wavelength, n is the refractive index, and NA is the numerical aperture of the lens.

If NA << 1, then DOF is approximated as:

$$DOF = \pm 0.5 \frac{\lambda}{NA^2} \tag{2.3}$$

Currently, lithography stepper systems use a step-and-scan technique: the mask (reticle) and the wafer are simultaneously moved in opposite directions such that a slit of the light scans the entire mask and projects the image onto the wafer [8]. Maintaining the wafer stage, in order that the lens-wafer distance is uniform throughout the step-and-scan procedure, is extremely difficult. Furthermore, in printing smaller feature sizes thus improving the resolution (refer to equation (2.1)), NA is increased in order to capture higher orders of light waves, subsequently decreasing the DOF (refer to equation (2.3)). Thus, the problem of defocus non-uniformity is

particularly difficult in optical lithography, where the DOF has become so small that there is a concern about whether optical steppers are capable of maintaining the image in focus.

In addition to concerns about the DOF control, the lithography system must provide uniform exposure dose illumination to maintain CD control across the exposure field (Figure 2.7) [19][20].

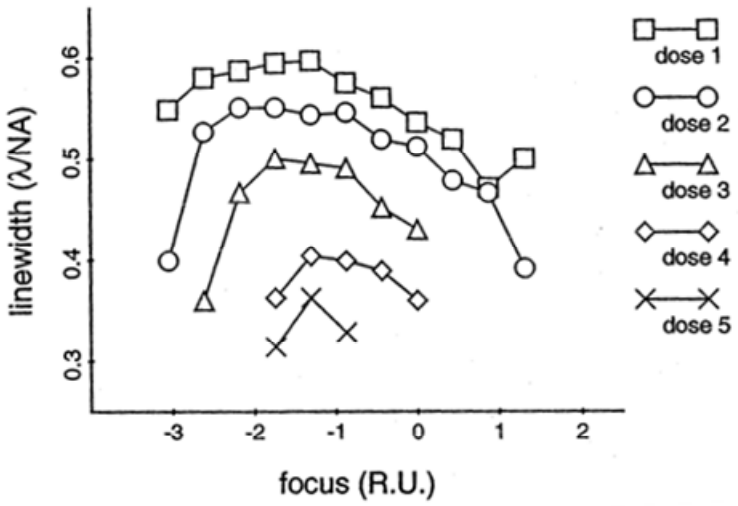


Figure 2.7: Bossung plot shows that DOF and exposure dose control are quite important in maintaining a uniform CD [20].

A uniform dose on the wafer is maintained by controlling the wafer stage scanning speed and the intensity of the laser source. The exposure dose for a given slit width, W_s, is expressed as:

$$W_s = V_m \frac{n}{f} \tag{2.4}$$

where V_m is the maximum wafer stage scan speed, n is the number of pulses required to achieve the specified dose, and f is the laser source repetition rate. The exposure dose variation range is typically 1-2% [8].

Alignment of the reticle to the wafer is also an important source of variability. Integrated circuits are manufactured through a series of patterning steps. Each new layer must be placed on top of the preceding layers, and proper alignment of the new layer to the previous circuit pattern, already imaged on the wafer, is essential for correct electrical contact to occur. This alignment error, commonly referred to as an overlay error, must be kept within certain tolerances. Current state-of-the art ArF lithography scanners can maintain overlay errors of 3-4 nm [21].

Other sources of variation in the lithography patterning steps include post-exposure bake (PEB), which causes systematic across-chip variation [22]. The resist here is first spun onto the wafer, exposed, and then baked. After it is exposed through the reticle, a PEB is used to drive additional chemical reactions or the diffusion of components within the resist film. If the temperature is uneven, this non-uniformity will result in critical dimension (CD) variation. This effect is manifested in a parabolic profile across wafers.

2.2.2Variation in Patterning and Other Manufacturing Steps

Other patterning and manufacturing steps are important sources of process-induced variation. This section lists and discusses such sources and their potential impact on transistor performance, organized by their respective magnitude of radius of influence.

- Chemical Mechanical Polishing (CMP) causes a difficult challenge for maintaining overlay-error control. The purpose of CMP is to produce planarized surfaces. CMP is helpful to lithography because CMP reduces thin-film effects; however, CMP makes overlay control difficult because it reduces the contrast of alignment targets [8].

- Dry etching is an important step in the patterning of the poly-silicon gates. This process suffers from microscopic loading effects in which densely-spaced gates experience a lower etch rate than isolated ones. This difference causes dense gates to have larger CDs than those of isolated ones.

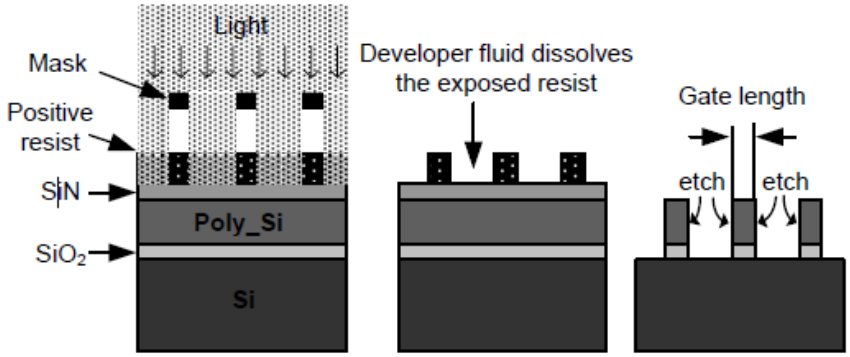


Figure 2.8: The fabrication steps of poly-silicon gates show that CDs suffer from etch loading effects [1]

- Applying mechanical stress to transistors improves their carrier mobilities. NMOS and PMOS devices respond differently to tensile and compressive stress, depending on the transistor type [23]. To apply mechanical stress to NMOS devices, a nitride contact etch stop liner (CESL) is deposited using Chemical Vapor Deposition (CVD). This capping layer has tensile strain. The discontinuity of the film at the edge of the MOSFET gate/spacer impacts the stress/strain in the channel. Figure 2.9 shows the mechanical stress distribution beneath the channel area. While the additional capping layer enhances NMOS performance, it also increases layout-induced variability due to different diffusion areas.

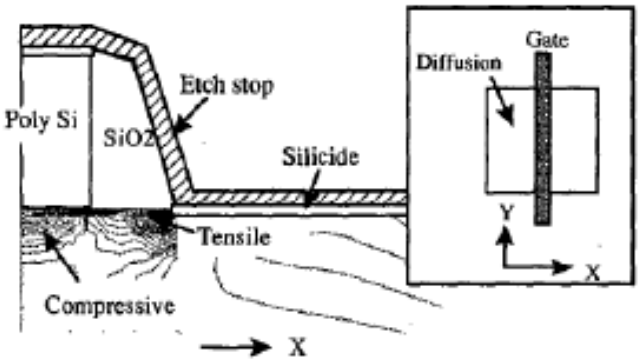


Figure 2.9: The distribution of mechanical stress beneath the gate that is induced by the contact etch stop liner, which is also commonly referred to as the capping layer [23].

2.3 Process Characterization Methodologies

In order to monitor process-induced variation, a variety of process characterization techniques can be employed, such as: optical scatterometry, scanning electron microscopy (SEM), electrical linewidth measurements (ELM), and automated electronic testing. These process characterization methods are used for process control as well as to evaluate process corners.

Optical methods like scatterometry are used to inspect features that are large enough for optical linewidth measurement systems. Their high throughput advantage over SEMs is a major reason why they are used for minimizing tool-induced shifts in quantifying overlay errors and process drifts. Most optical methods have been replaced by the use of SEMs as the primary tool for measurements. SEMs have high resolution levels. They are often used for process monitoring and to calibrate OPC models. Though the process of using SEMs is slow, with the utilization of automation they can be used for extensive studies.

Electrical Linewidth Metrology (ELM) is an electrical measurement method that serves to supplement SEM measurements. It is used to characterize systematic CD variation from a variety of sources, for example, illumination non-uniformity across the slit, wafer level effects during the Post Exposure Bake (PEB), and etch steps [24][25][26]. Since this method has a speed advantage over SEMs, it is used when a large number of linewidth measurements are needed in a small amount of time.

Electronic testing in the form of ring oscillators, Static Random Access Memories (SRAMs), and leakage test structures has commonly been used to examine circuit performance variation induced by physical effects in greater detail. Electronic testing is both easily automated and provides a permanent silicon record on the wafer. It also offers the important benefit of examining the overall performance, as the final electrical performance of a circuit is what ultimately matters.

2.3.1 Frequency

Ring Oscillators (RO) are commonly used to characterize circuit performance variability by examining physical effects. They are small in area and simple to implement. RO frequency can be measured easily, rapidly, and accurately through automation when they are used. In

conjunction with scan-chains and multiplexers, only a few input and output pads are needed for automatic measurements.

Since ROs with a large number of inverter-stages can average out the random variations of the individual stages, systematic variation is easily captured by long RO-inverter chains. Random variation is easily captured by short RO-inverter chains because ROs with a small number of stages experience less averaging at each inverter stage.

RO frequency is sensitive to gate length, mobility/injection velocity, threshold voltage, and parasitic capacitances. RO frequency is approximated as:

$$f_{RO} \propto \frac{I}{CVN_{stages}} \tag{2.5}$$

where f_{RO} is frequency, I is the transistor current, C is the load capacitance for a single stage, V is the supply voltage, and N_{stage} is the number of RO-inverter stages.

Ring oscillators have been used to characterize the effect of interconnect [4], parasitic capacitances [27], and even the exposure dose [28]. Modified ROs can also to be used to measure individual transistor gate delay with a high degree of accuracy [29][30].

2.3.2 Current: $I_{d, SAT}$ and $I_{d, LEAK}$

Techniques that measure subthreshold leakage current ($I_{d, LEAK}$) are very sensitive to threshold voltage, V_{th}. For example, gate leakage and V_{th} were measured in an addressable array of devices in IBM's 65nm SOI technology [31]. Like frequency, saturation current ($I_{d,SAT}$) is sensitive to similar design metrics, excluding capacitance that also affects circuit performance. For instance, transistor saturation current has been measured from an addressable array of transistors, and the transistor gate lengths are inferred from $I_{d,SAT}$ [32].

2.4 CAD Tools for Monitor Design Enablement

Computer-aided-design (CAD) tools are used to simulate process-induced variation at the design stage prior to fabrication. They are often used to decipher measured silicon results and will be used to explore ways to enhance parameter specificity later in this work. Fortunately, Mentor Graphics Calibre using generic models for exposure tools and resist effects is available. ST Micro's 40nm Process Design Kit (PDK) with electrical extraction and BSIM4/PSP models is available under NDA with BWRC. The CAD tools that are relevant to this work are summarized in Table 2.5. Of particular help are the Parametric Yield Simulator (PYS) developed by Poppe, and Pattern Matching, which will be delved into in Sections 2.4.1 and 2.4.2 [5][33].

CAD Tools	Purpose
SPLAT (NA < 1)	Rigorous Lithography Simulations for NA < 1
Mentor Graphics Calibre	Rigorous Lithography Simulations
ELDO, SPICE	Circuit Simulations
Sentaurus	Stress/Strain TCAD Device Simulations
Parametric Yield Simulator (Simulation flow that contains: Calibre, Non-Rectangular Transistor Modeling/BSIM, and HSPICE)	Simulation flow that converts 2-D gate lithography contours into a 1-D effective gate lengths for HSPICE simulations
Pattern Matching	Fast-CAD Lithography Simulations

Table 2.5: A summary of the CAD tools used in this work

2.4.1 Parametric Yield Simulator (PYS)

The Parametric Yield Simulator (PYS) is a comprehensive simulation flow that translates a 2-D gate lithography contour into a 1-D effective gate length to be injected into circuit simulators like HSPICE [5]. Automated simulation flow from GDS to HSPICE outputs is implemented using Mentor Graphics Calibre Work Bench, BSIM, and HSPICE, in which automation is enabled by Perl and TCL scripts. A non-rectangular transistor model using BSIM is the basis of the PYS, as it translates a physical shape (the output from Calibre simulations) into an equivalent gate length that serves as an input to HSPICE.

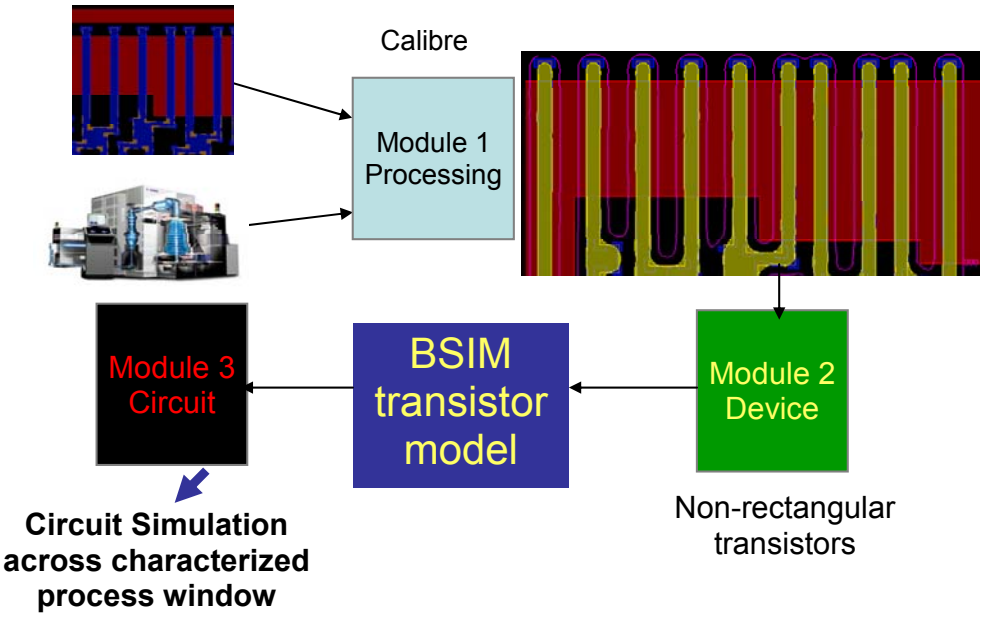


Figure 2.10: Parametric Yield Simulator (PYS) flow [5]

As illustrated in Figure 2.10, the PYS flow contains 3 simulation modules. Module 1 (Processing) uses Calibre to simulate the lithography contours under specific process conditions. Module 2 (Device) uses the simulated geometrical gate shapes from Calibre and builds a BSIM transistor model to calculate equivalent transistor gate lengths. The third module (Circuit) injects

all of the calculated transistor lengths into an HSPICE netlist for circuit simulations. The flow here is automated using scripts for simulating circuit performances across a process window.

2.4.2 Pattern Matching Physics

The concept of modeling process effects at the mask plane level as convolutions with maximal lateral influence kernel functions was introduced in a Pattern Matcher system by Gennari [34]. The influence functions for aberrations, including focus and polarization, were introduced by Robins and McIntyre [35] [36]. Extensions for illumination effects and for focus as a large aberration are included by Rubinstein, Wang, and Miller [37][38][39]. The lateral influence function for focus as adjusted for partial coherence for illuminations in Chapter 3 will be used to guide the selection of focus sensitive monitors prior to their systematic analysis using Calibre.

As shown in Figure 2.11, the Pattern Matcher contains a pattern generator that uses lens aberrations (including focus) and mask illumination to create the lateral influence pattern of the kernel. The Pattern Matcher then makes an analog weighted convolution of the chip layout and reports the degree of similarity, or the Pattern Match Factors (PMF). Circuit designers can use such outputs from the Pattern Matcher to determine the sensitivities that their layouts have to residual process effects.

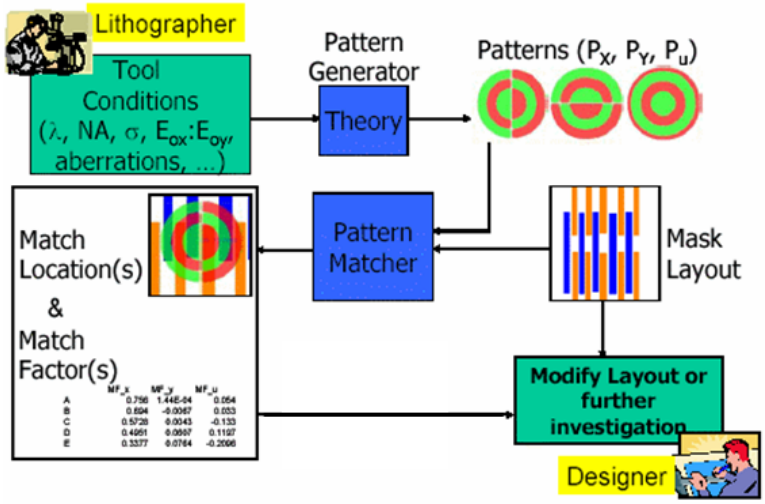
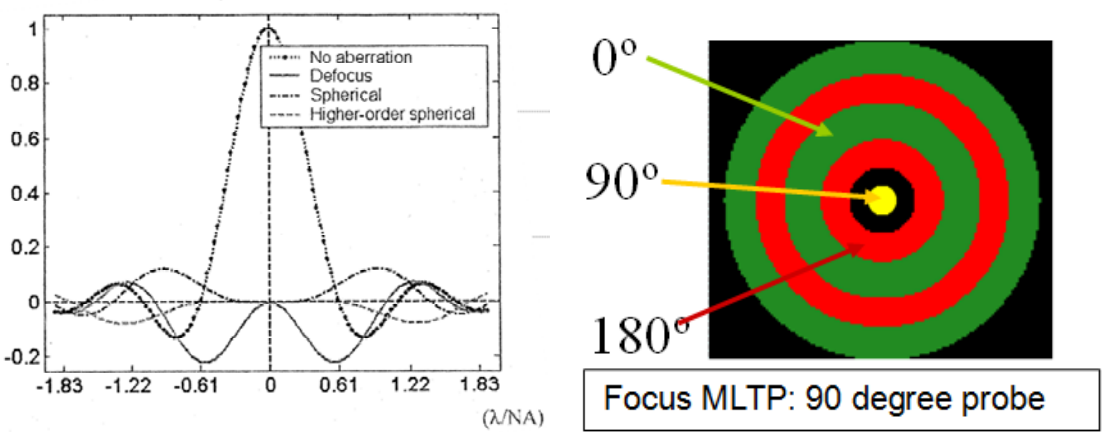


Figure 2.11: The Pattern Matcher Concept and Flow [34]

Aberrations like focus contribute lateral spillovers among features in addition to, and independent of, those due to proximity effects. The spillover function for proximity and several of the even aberrations in Figure 2.12 from Robins et al [40] are shown to be the Fourier transform of the Zernike terms in the pupil. These kernels characterize the lateral electromagnetic spillover due to no aberration (Z0), defocus (Z3), spherical (Z8), and higher-order spherical Zernike. Note that the defocus contribution is orthogonal to the unabated electric field point-spread-function. These spillover functions are actually 2-D patterns that are rotationally symmetrical with radial cut-line intensity, as can be seen in Figure 2.12(a). Pattern Matcher's pattern generation functions create these spillover functions and discretize them in pixels on a grid for convolution.



(a) (b)

Figure 2.12: The above kernels in (a) are computed by taking the inverse Fourier Transform of the corresponding Zernike terms [38], while (b) is an example of a focus monitor.

The physics behind Pattern Matching is explained below. A standard lithography system is shown in Figure 2.13.

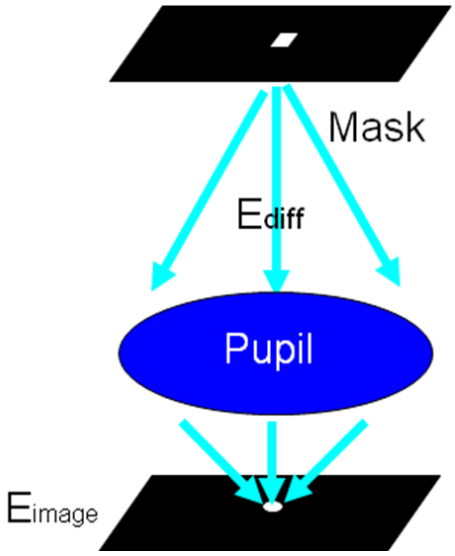


Figure 2.13: A typical lithography system

Based on the system in Figure 2.13, the electric field at the image plane, E_{image} is represented as a 2-D inverse Fourier Transform (IFT):

$$E_{image}(x,y) \propto \iint_{Pupil} E_{diff}(f,g)e^{-j2\pi(fx+gy)}e^{-jOPD(f,g)}circ\left(\sqrt{f^2+g^2}\right)dfdg \qquad (2.6)$$

where E_{diff} is the diffracted light from the mask, e^{-jOPD} is the Optical Path Difference (OPD) caused by an aberration in the lithography system that results in a small phase change, and the circ function represents the pupil lens, which acts as a low frequency filter.

Assuming that the aberration is small, Taylor Expansion can be used to represent the OPD exponential term,

$$e^{-jOPD(f,g)} \approx 1 - jOPD(f,g) \tag{2.7}$$

Since the mask in Figure 2.13 contains an infinitesimally small pinhole, a delta function centered at the origin, δ(0,0), is used to represent the mask. Its IFT results in $E_{diff}(f,g) = 1$.
The above assumptions simplify Equation (2.6) to:

$$E_{image}(x,y) \approx IFT\left[circ\sqrt{f^2+g^2}\right] + IFT[jOPD(f,g)] \tag{2.8}$$

where the first term is the electric field produced from proximity (Airy Function), and the second term is the electric field spillover produced from the aberration.

Using proximity electric fields as a reference, the change in E_{image} can be represented as:

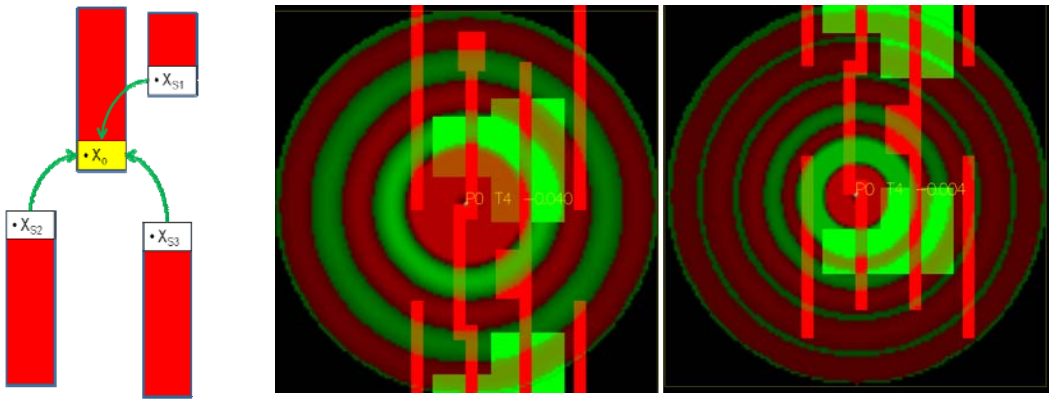
$$\Delta E_{image}(x,y) \approx IFT[jOPD(f,g)] \tag{2.9}$$

The patterns or kernels used in Pattern Matching are pixel-based representations of Equation (2.9), allowing one to use the IFT of the OPD (i.e. IFT[Zernike Polynomial]) and to work at the image plane instead of at the mask plane [38].

To assess the degree of spillover, the Pattern Matcher convolves the layout with the kernel represented in Equation (2.9). The Pattern Matcher computes a Pattern Match Factor (PMF), which is defined as a normalized 2D discrete correlation computation. This computation is similar to an image convolution. The PMF at layout position (i, j) for an X by Y kernel is calculated as:

$$PMF(i+x, j+y) \approx \sum_x \sum_y Layout(x+i, y+j) \cdot Kernel(x,y) \tag{2.10}$$

The resulting PMF is an indication of the severity of electromagnetic spillover due to lithography aberrations at a particular location (i,j) in a layout. Figure 2.14 offers some examples of the results from pattern matching simulations.



(a) (b)

Figure 2.14: In the Pattern Matcher, electric field spillover is combined onto an observation pixel (X_o) from surrounding pixels (X_{Si}), as shown in (a) via a convolution computation. In (b), examples of different kernels matched onto a point on the poly (red) layout are shown.

Since image quality correlates with intensity rather than the electric field, Miller et al developed a new Pattern Matching metric called a composite match factor (CMF) that more accurately correlates Pattern Matching outputs with changes in intensity [39].

2.5 Design of Gate Lithography Focus Monitors using the Parametric Yield Simulator and Pattern Matching

Defocus is a source of variability that is increasingly more important with increasingly high NA lithography. The effect of defocus dependencies on layout geometries has been studied extensively in the literature [41]. A basis for the automated electronic testing in this thesis is the study of how the individual transistor layouts can be used in designing a response to a focus made with Parameter Yield Simulations [5]. Using Calibre lithography simulations, Poppe's PYS platform systematically screens for layout geometries that are sensitive to focus. The platform converts the 2-D lithography images into a 1-D effective gate length using a non-rectangular sliced-transistor model which is then plotted across a process window for analysis.

One example of the layout-dependent effect is that of defocus. Figure 2.15 shows the simulated response of five-line gate arrays to defocus as a function of pitch, where the critical dimension (CD) shift is determined by taking the difference between the CD of the center (dense) line and that of the right most (semi-dense) line [5].

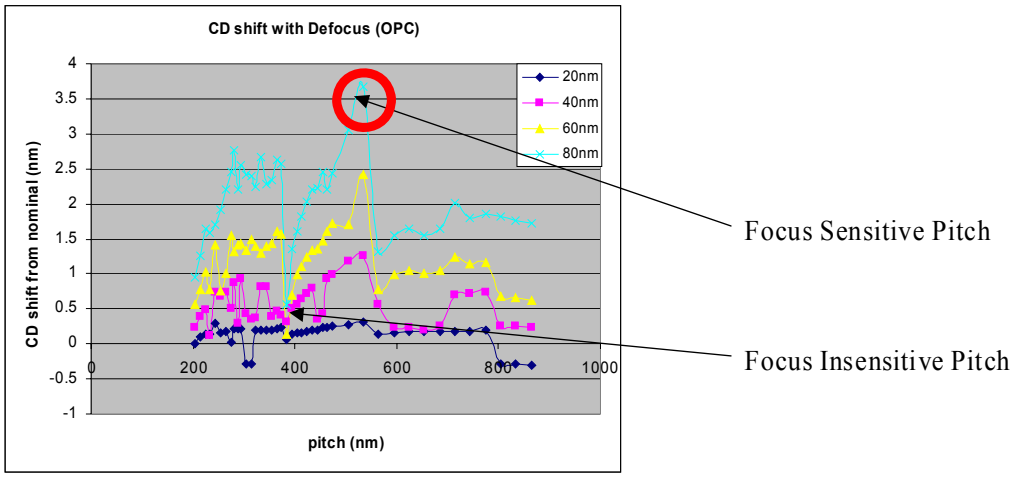


Figure 2.15: The post-OPC simulated CD shift of 80nm gates for four different levels of defocus reveals that some pitches exhibit high sensitivity to focus while others are insensitive to focus.

In Figure 2.15, a high sensitivity to focus for a particular gate pitch is observed. Hence, since transistors at this gate pitch have a higher measured CD error than the rest, they can serve as defocus monitors. Wafers can be intentionally exposed to measure the response of these monitors to various exposure dose/defocus conditions and to identify the most sensitive pitches.

Since isolated gates are sensitive to focus, the response of an isolated 80nm gate through focus and exposure dose is observed and will serve as a reference for comparison with other hyper-sensitive monitors (Figure 2.16).

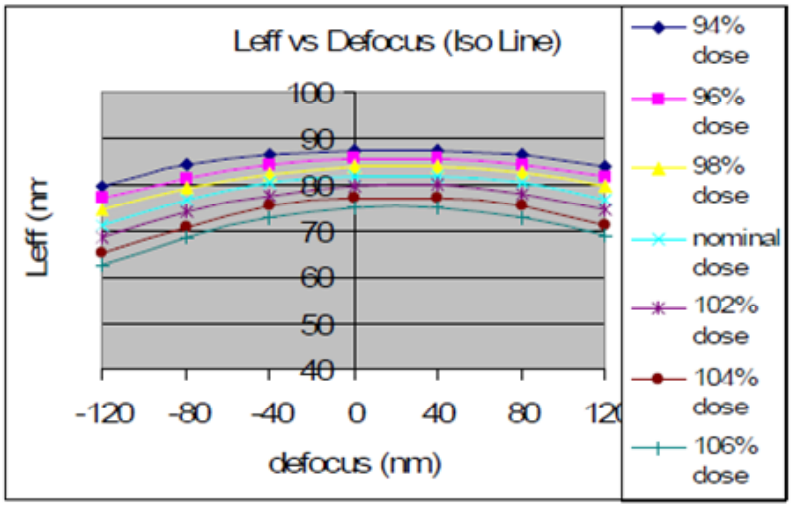
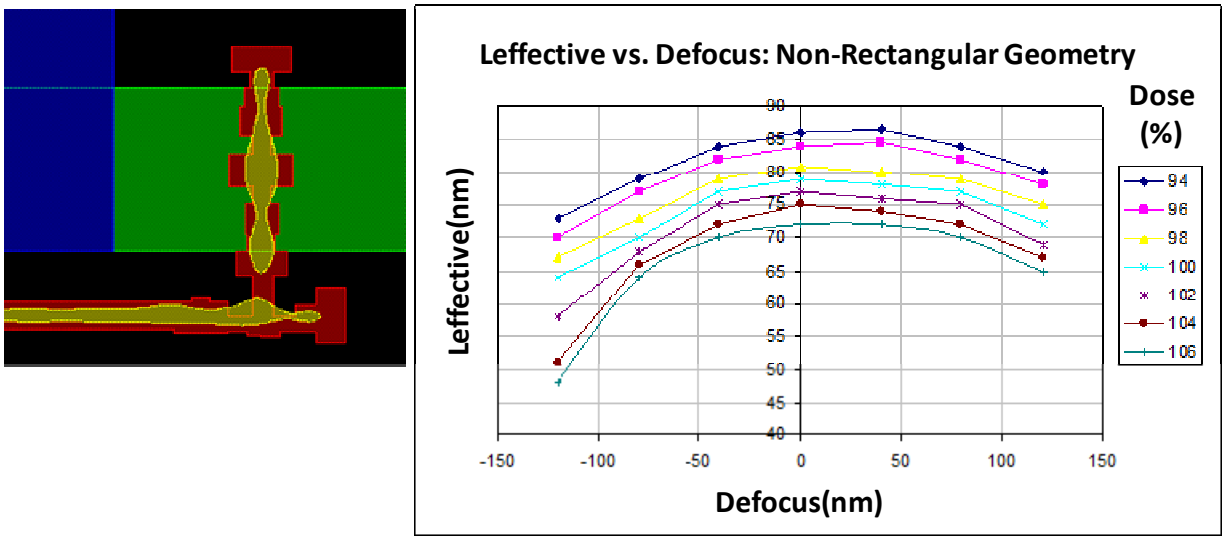


Figure 2.16: Bossung plot of $L_{effective}$ vs. DOF across different dose conditions for an 80nm isolated line [5]

For the purpose of comparison, the process variation (PV) band is defined from -80nm to 40nm DOF at a 100% (typical) dose. The $\Delta L_{effective}$ = 3% for an isolated 80nm gate. Poppe and Rubinstein developed layout monitors for gate lithography focus by adding features to an isolated gate in order to enhance focus sensitivity.

A reticulated monitor is a non-rectangular monitor that consists of hammerheads, a bump, and a poly elbow, where the synergy of these carefully added features causes two significant pinch points through focus [5]. Figure 2.17(a) shows that when a monitor is simulated with a 6% overdose and a -120nm defocus, it exhibits drastic changes in the CD.

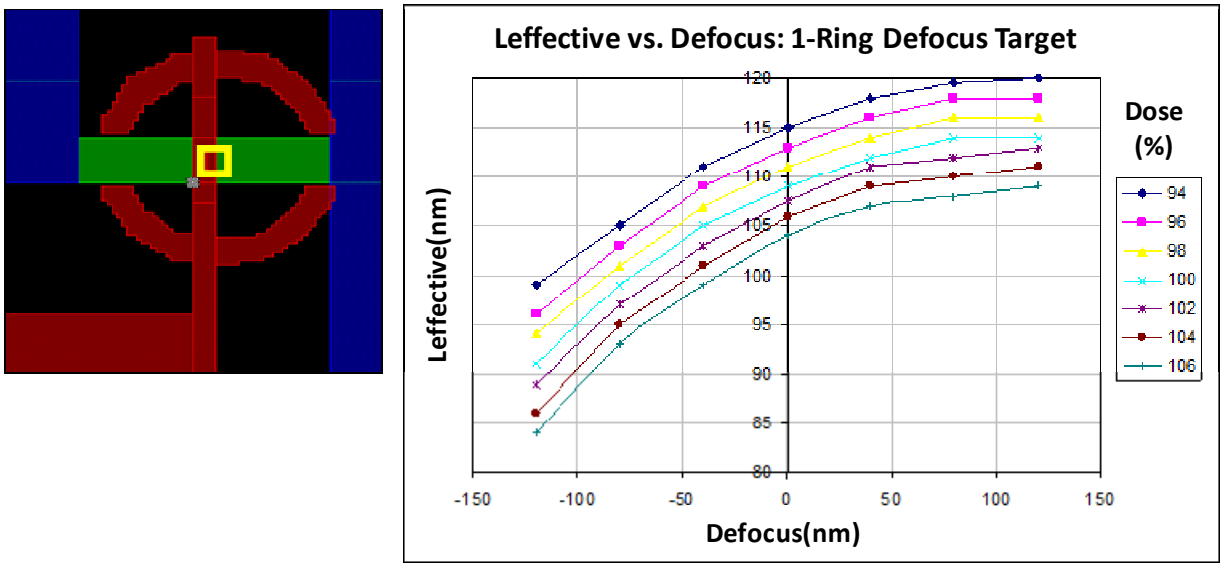


(a) (b)

Figure 2.17: A simulated aerial image using PYS for a reticulated monitor is shown in (a), and the monitor's response through focus and dose is illustrated in (b) [5].

Figure 2.17(b) shows a Bossung Plot of $L_{effective}$ versus DOF across different dose conditions. A $\Delta L_{effective}$ = 12.8% change is observed within the PV bands defined above. The curvature change is twice as sensitive as the isolated 80nm gate reference.

Much higher sensitivity can be achieved with the use of a phase shift region of 90^{O} on the mask that tends to respond linearly rather than quadratically to focus. For example, probe-based Zernike aberration monitors have been shown to be hyper-sensitive to defocus but they require SEM for metrology [43]. The defocus monitor developed by Robins has been modified by Rubinstein and Poppe into a transistor to enable electrical measurement [5][44]. This monitor uses an attenuated phase shift mask and a center probe with a 90^{O} phase etch. The result is a layout monitor that can detect the sign as well as magnitude of defocus. A hyper-sensitive layout monitor developed by Poppe using a 90^{O} phase shift probe is shown in Figure 2.18.



(a) (b)

Figure 2.18: A simulated aerial image of a Zernike aberration monitor using PYS is shown in (a), and the monitor's response through focus and dose is given in (b). Since the center 90^O probe region is hyper-sensitive to defocus, the gate CD is expected to change dramatically in response to the defocus in the scanner [5].

Figure 2.18(a) shows a hyper-sensitive layout monitor. This monitor consists of a gate with the source on the left and the drain on the right. The box represents a 90^O phase shift area on top of the gate. The light coming through the boxed region interacts with the spillovers from defocus and, depending on the direction of defocus, the phase shift area either digs in or out of the gate. Figure 2.18(b) presents a Bossung plot of $L_{effective}$ versus DOF across different dose conditions for the layout monitor presented in Figure 2.18(a). Note the linear relationship across focus in Figure 2.18(b). $\Delta L_{effective} = 27.8\%$ is observed within the PV bands defined above. This change is 5 times more sensitive than that of the point of reference, the isolated 80nm gate.

To be representative of a production process in which Optical Proximity Correction (OPC) is used to improve gate length control, the point of reference is shifted to that of the focus variations of the dense gates. It is well known that the variation across the field for isolated gates is typically twice that for dense gates. Table 2.6 now allows the percentage change to be related to dense gates. Here reticulated monitors have about a 2x greater sensitivity to focus as compared to that of the isolated gates, and they are 4x more sensitive as compared to dense gates. With the use of a 90^O phase shift mask, probe-based Zernike aberration monitors are 5x more sensitive than that of an isolated gate and 10x more sensitive than that of the dense gates. Thus, considerable leverage exists for the construction of transistor layouts for focus sensitivities that can be measured through electrical monitoring.

Gate Lithography Focus Layout Monitors	$\Delta L_{effective}$ (PV Band: -80nm to 40nm DOF)	Sensitivity Improvement using Dense Gates as the Control
Isolated (Reference)	5.5%	2
Reticulated (non-rectangular)	12.8%	4
Phase Shift	27.8%	10

Table 2.6: A summary of $\Delta L_{effective}$ for gate lithography monitors

2.6 Summary

In current design methodologies, the impact of circuit performance variation is usually not carefully characterized in circuit simulations. Instead, current designs are typically made to satisfy design rule checks and worst-case corners. The increasing number of design rule checks with scaling exacerbates design time. Process corner design methodology is often too pessimistic and is not well-correlated with layout-induced systematic effects. Thus, process variation needs to be carefully monitored and analyzed in order to both improve the current design methodology and to optimize process conditions in manufacturing.

Simulations can be used to design process-sensitive monitors as well as to help decipher measured results leveraging industrial CAD tools and promising fast-CAD approaches like Pattern Matching. Preliminary simulated results and electrical measurements demonstrate that probe-based Zernike aberration gate lithography focus monitors can be designed to show 12x more focus sensitivity than that of a dense gate.

Such work leads to the design of parameter-specific ring oscillator monitors for the electronic monitoring of sources of process-induced variability. A set of layout monitors developed by Pang is used for screening process conditions. These layouts can be modified to enhance the effects of gate etch, gate focus, gate-to-active misalignment, and nitride CESL/STI-induced stress. Pattern Matching, Calibre, and Sentaurus are particularly useful tools for quantitatively modeling diverse lithographic and stress impacts. The variety of variability sources increases with new device stressors as well as the number of process steps.

Chapter 3
45nm Ring Oscillator Monitors Experimental Design

This chapter extends the RO layout monitoring approach to enhance the sensitivity and selectivity to particular process parameters by adding physical modeling and simulation of the sources of variability in order to tune the layouts. These parameter-specific and simulation-calibrated ring oscillator inverter layouts are described for the identification and quantitative modeling of the sources of circuit performance variation for gate etch, gate lithography, gate-to-active overlay, nitride contact stop layer (CESL)-induced strain, and Shallow Trench Isolation (STI)-induced stress. Generic conceptual results are given based on experiences garnered from the preparation of proprietary layouts that pass design rule checks for a 45 nm ST Micro process under development. Poly and diffusion layout sensitivity enhancement choices have been guided by fast-CAD Pattern Matching. The accuracy of the fast-CAD estimate by the Pattern Matcher for lithography is then tuned in a detailed manner using simulations in Mentor Graphics Calibre. RO monitors are also described to characterize the stress effects associated with diffusion area size.

This chapter is organized into 7 sections. Section 3.1 analyzes the RO frequency sensitivity to 90nm RO monitors designed by Pang so as to screen for various sources of process variation [1]. Section 3.2 introduces the 45nm RO array circuit design schematics and layout floor planning. Section 3.3 summarizes the 45nm parameter-specific RO monitors. Sections 3.4 and 3.5 discuss the methodology to assess and guide the design of RO layouts using lithography-based simulations, including Pattern Matching. Section 3.6 describes the design of nitride CESL and STI stress RO monitors. Finally, Section 3.7 summarizes the chapter.

3.1 RO Frequency Sensitivity Analyses for 90nm Test Chip Measurements

The purpose of this section is to determine whether lithographic effects can explain the across-wafer and layout-to-layout RO frequency differences observed by Pang from his set of screening layouts with various dummy gates, as can be seen in Figure 2.2 in Chapter 2 [1]. One may recall that the measured within-chip $3(\sigma/\mu) = 3\%$ and an across-wafer $3(\sigma/\mu) = 20\%$ are observed. The analysis begins by investigating the RO frequency sensitivity to gate length changes and the differences in electrical parasitic effects among the three layouts. Next, the physical changes in gate length with respect to focus, exposure dose, and layout are examined using a generic lithography model. Then, the lithographic gate length changes are mapped to RO frequency changes. This mapping then allows one to infer a quantitative estimate of the expected across- wafer RO frequency variation that is due to lithography. It also enables the frequency shift from layout-to-layout differences to be quantitatively compared to the shift seen in Figure 2.2.

3.1.1 RO Frequency Sensitivity to Gate Length and Electrical Parasitics

The parasitic impact on the RO frequency due to the presence or absence of dummy gates is examined for three layout monitors using ST Micro's 90nm Process Design Kit. In Figure 3.1, the RO frequency versus channel length relationship for Tiles 3a, 4a, and 5a is simulated in the following manner. First, the drawn channel length for all of the ring oscillator inverter layouts is modified by the same fraction. Next, the SPICE netlist for the modified ring oscillator layout is extracted. Lastly, the ring oscillator frequency is simulated using the extracted netlist.

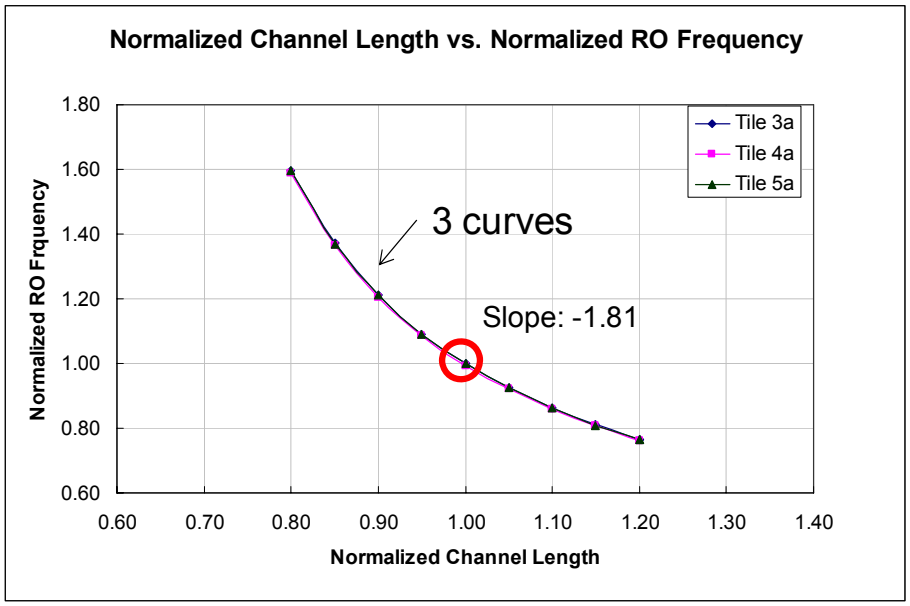


Figure 3.1: The RO Frequency versus the Channel Length simulations for three monitors show that the presence or the lack of dummy gates has little effect on extracted results for a drawn gate length. The sensitivity at (1,1) is -1.8.

Simulation results in Figure 3.1 show that a 10.1% change in drawn channel length causes an 18.1% change in RO frequency as given by the slope at (1,1). Since the three RO frequency versus channel length sensitivity curves show minor differences here, this observation indicates that the RO frequency increase observed in the measured data for Tiles 4a and 5a in Figure 2.2 is not due to changes in parasitic effects due to the presence or absence of dummy gates.

For further references in modeling non-rectangular transistor effects using the Parametric Yield Simulator (PYS), the RO frequency sensitivity to changes in the channel length for the NMOS and PMOS devices are considered separately. Following the simulation procedure stated above, the sensitivity to only the NMOS or PMOS is examined, where the drawn channel length of one device remains fixed while the other is varied.

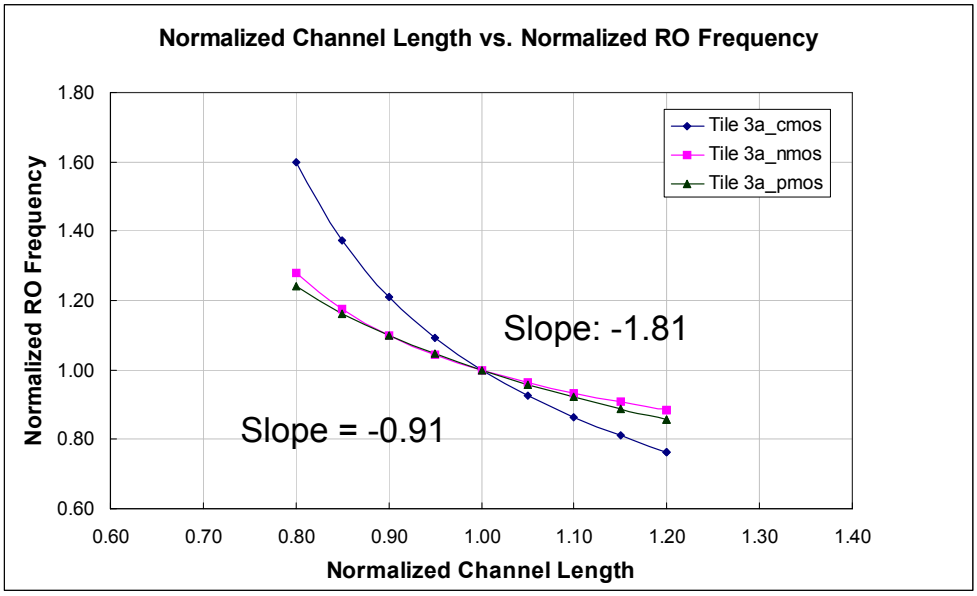


Figure 3.2: Sensitivity to the individual NMOS or PMOS device is half as big as the sensitivity to both devices.

The sensitivity to gate length change is 0.91 if only either the NMOS or the PMOS channel length is changed. When both of these gates are changed together by the same fraction, the sensitivity is 1.81. The combined effect of varying both the drawn channel length of the NMOS and PMOS devices yields twice the sensitivity as that of varying the NMOS's or the PMOS's channel lengths in an individual manner.

3.1.2 RO Frequency Sensitivity to Gate Lithography Dose and Focus

The PYS is used to model gate length with the inclusion of lithography effects. The layouts designed by Pang [1] for the 90nm ST Micro PDK are inputs to the PYS. A generic 90nm BSIM3 model card developed for predictive modeling by Arizona State University for a bulk CMOS technology is used for this conversion [45].

Lithography modeling in the PYS is generic and without calibration against actual silicon results for the ST Micro 90nm process. More specifically, for the 193 nm dry lithography, a numerical aperture (NA) of 0.78 and annular illumination with sigma_out= 0.88 and sigma_in= 0.4 is assumed. This produces a Raleigh depth of focus (RU) of 159 nm. The simulated focus range is $\pm$100 nm, and the simulated dose range is $\pm$6% from nominal (Dose=100). The impact of coma is not included in the lithography modeling because the measured RO frequency for the asymmetrically biased layout monitors (i.e. Tiles 4a and 5a) show minimal differences. A variable threshold resist model is utilized to obtain contours using Calibre.

Using PYS simulation, Bossung plots of the variation in effective non-rectangular gate length ($L_{effective}$) versus defocus and exposure dose are shown in Figure 3.3 for an isolated 90nm gate and in Figure 3.4 for a dense 90nm gate. These simulation results exhibit the expected parabolic Bossung shape. Due to the properties of the resist model, an asymmetric shift of +20nm is observed in the Bossung plots. For purposes of sensitivity analysis, observations are referenced from $\pm$ 2% of the exposure dose from nominal (i.e. Dose=100) and at a range of -80

nm to +120 nm defocus, which is a 100 nm range from the edge to the center of focus, or approximately $\pm$ 2/3 of a RU.

The percentage of the variation in the $L_{effective}$ caused by defocus ($L_{effective,\ focus}$) are calculated from the curvatures under nominal dose conditions (Dose=100%), while variations caused by the exposure dose ($L_{effective,\ dose}$) are calculated at +20nm defocus. For the sensitivity range of -80nm to a +120nm depth of focus, an isolated monitor (Figure 3.3) shows a $\Delta L_{effective,\ focus,\ NMOS}$ = 5.5%, while the dense layout (Figure 3.4) shows only a $\Delta L_{effective,\ focus,\ NMOS}$= 3.0% for the same range. For a sensitivity range of a $\pm$2% exposure dose, the isolated monitor (Figure 3.3) shows a $\Delta L_{effective,\ dose,\ NMOS}$= 8.1%, while the dense monitor shows only a $\Delta L_{effective,\ dose,\ NMOS}$ = 6.8%. In general, the additional bias from the dummy gates decreases the $\Delta L_{effective}$ sensitivity to focus and the exposure dose by 1-2%.

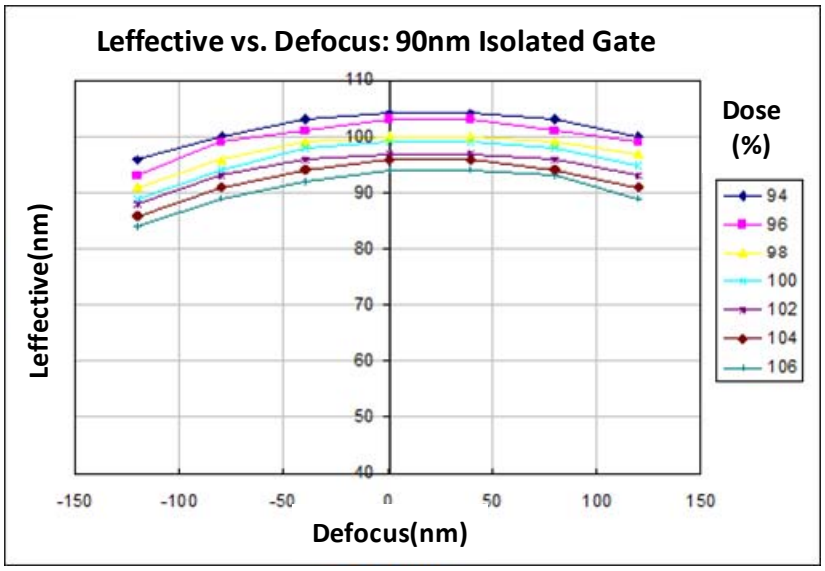


Figure 3.3: A simulated plot of the $L_{effective}$ versus defocus for different dose conditions for an *isolated* 90nm gate shows a $\Delta L_{effective,\ focus,\ NMOS}$ =5.5%.

Tile 3a

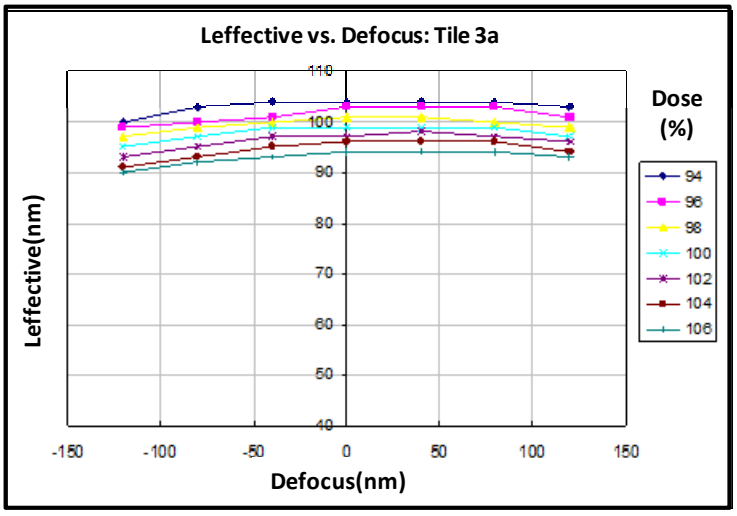
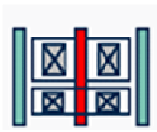


Figure 3.4: A simulated plot of the $L_{effective}$ versus defocus for different dose conditions for a *dense* 90nm gate shows a $\Delta L_{effective,\ focus,\ NMOS}$ =3.0%.

3.1.3 Prediction of RO Sensitivity from PYS Lithography Modeling

Figure 3.5 shows a Bossung plot for Tile 3a of the RO frequency versus the DOF and exposure dose for NMOS devices only. This figure, as well as that of Figure 3.6, is obtained by taking the $L_{effective}$ of the NMOS transistors and mapping the $L_{effective}$ to the RO frequency using the RO frequency versus channel length relationship from Figure 3.2. It does not include any changes in the PMOS devices and should predict only about half of the variation experimentally observed. The Bossung plots of the RO frequency versus that of the DOF are not perfectly parabolic due to discretization errors from the PYS when converting lithography contours to $L_{effective}$.



Tile 3a

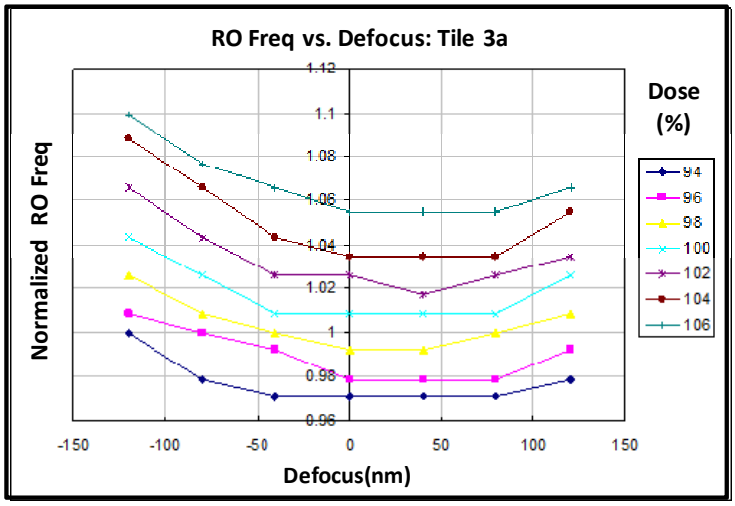


Figure 3.5: Tile 3a's simulated plot of ring oscillator frequency versus defocus for different dose conditions is used to examine the measured across-wafer RO frequency spread.

Figure 3.5 is used to quantitatively determine whether lithography is a possible source of across-wafer RO frequency variation for RO monitor Tile 3a. From Figure 2.2, the measured across-wafer 3(σ/μ) RO frequency variation is approximately 18-20%. Assuming a process window of a ±2% dose and an -80nm to 120nm defocus, Figure 3.5 shows a variation in the RO frequency, $\Delta f_{RO, NMOS} = 6\%$ ($\Delta f_{RO, DOSE, NMOS} = 4\% + \Delta f_{RO, FOCUS, NMOS} = 2\%$). This variation must be doubled when the PMOS effects are included; thus, the simulated across-wafer spread due to lithography is $\Delta f_{RO, CMOS} = 12\%$. This estimate shows that lithography variation from exposure dose and defocus contribute significantly to the 18-20% measured across-wafer spread, but that other independent sources likely contribute more. The measured across-wafer variation, σ_{total}, is approximated as the square root of the sum of the variation due to lithography, $\sigma^2_{lithography}$, and the variation due to other sources, σ^2_{other}. Since $\sigma_{lithography}/\sigma_{total} = 12\%/18\% \leq 0.7$, this implies that $\sigma_{other} > \sigma_{lithography}$.

Figure 3.6 is used to make a quantitative estimate on the layout-to-layout differences and to examine the pattern-induced RO frequency shift among the three different RO monitors. Here, little focus and dose-dependent RO frequency shift is observed. At most, a $\Delta f_{RO, NMOS} = 4\%$ is observed for a few very high defocus conditions; this variation is doubled when the PMOS effects are included ($\Delta f_{RO, CMOS} = 8\%$). Thus, lithography variation does not account for the observed pattern dependent systematic shift in RO frequency, and the measured Δf_{RO} that equals 10% is likely associated with etch.

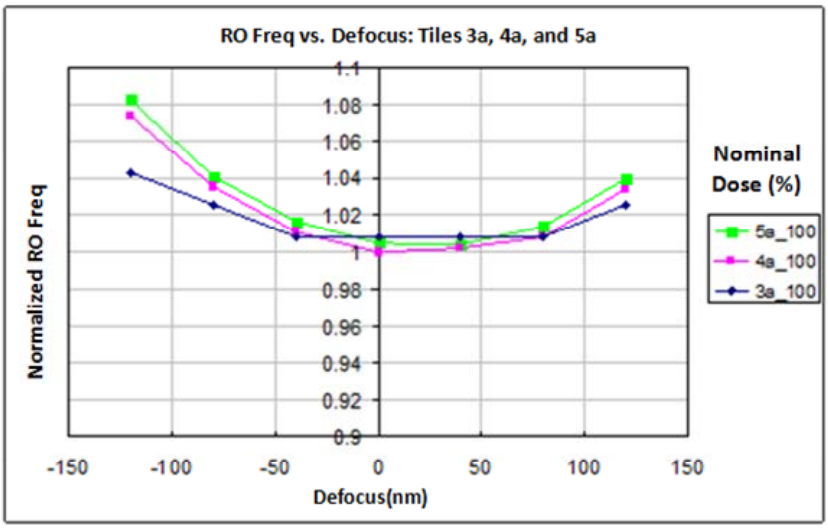
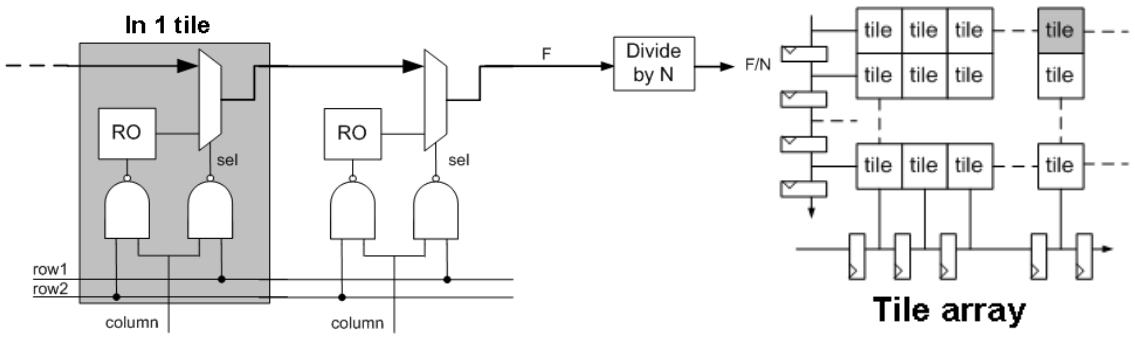


Figure 3.6: A simulated plot of the RO frequency versus defocus for different dose conditions shows that the pattern dependent shift among monitors Tiles 3a, 4a, and 5a is likely due to a non-lithographic source, such as etch.

3.2 Ring Oscillator Schematics and Floor Plans for 45nm

To accommodate for the 32 parameter-specific ring oscillators in a 45nm test chip design, a scan-chain architecture utilized by [4] was adopted. Individual 13-stage ring oscillators are used, and each ring oscillator is considered as a tile. A ring oscillator tile also contains additional control logic both to enable the ring oscillator and to multiplex the ring oscillator signal from the ring oscillator output to the final output pad, as shown in Figure 3.7(a). These ring oscillator tiles are arrayed into a matrix, as can be seen in Figure 3.7(b). The on-chip

frequency divider reduces the ring oscillator frequency before the ring oscillator signal reaches the final output pad. This divider allows off-chip automatic measurements to be made in the megahertz range. Note that just a few pads are required to measure the thousands of ring oscillator circuits.

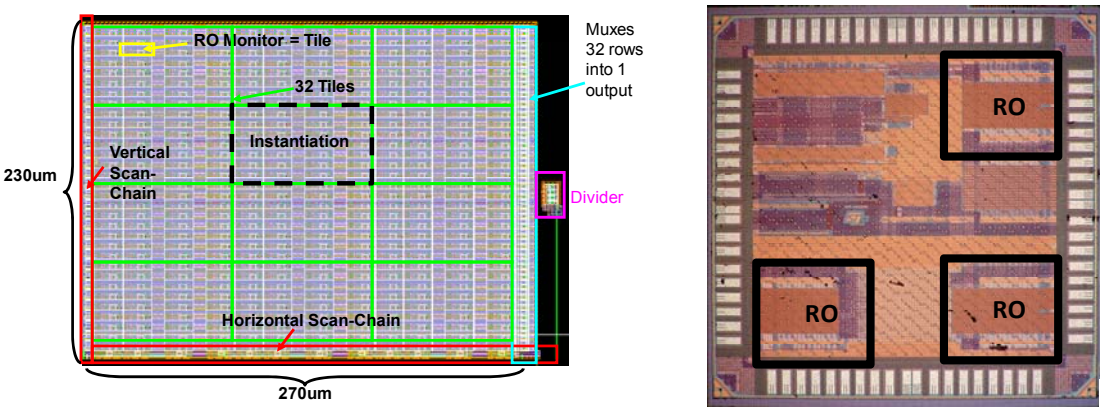


(a) (b)

Figure 3.7: The schematic of a ring oscillator tile arrayed in a row is shown in (a). A 2-D array with a set of vertical and horizontal scan-chains along the perimeter in order to enable row and column select can be seen in (b) [1].

The layout corresponding to the schematics in Figure 3.7 is shown in Figure 3.8(a). The vertical and horizontal scan-chains, multiplexers, and a frequency divider are located along the perimeter of the final array. On the left of the layout is the set of vertical scan-chains for the selection of the row tiles, while at the bottom is the set of horizontal scan-chains used to select the column tiles. On the right of the layout is a set of multiplexers which route the chosen RO signal down to a single signal line that feeds into the on-chip frequency divider.

In the middle of the layout, the smaller rectangular box represents a tile, and the larger rectangular box represents a set of 32 tiles, given the term "instantiation." One instantiation is replicated 3 times horizontally and four times vertically for a total of 12 identical instantiations of each RO type in a block. In order to capture the lithography scan and across slit direction variation, the block layout (which measures 230 x 270 um^2) is replicated in 3 locations on a $2x2mm^2$ chip in the manner shown in Figure 3.8(b).



(a) (b)

Figure 3.8: The layout view of the 45nm ring oscillator array is shown in (a). Here, (b) shows the layout in (a) replicated at 3 different locations across the 2x2 mm^2 chip in order to capture the variation due to the lithography scanner's slit and scan directions.

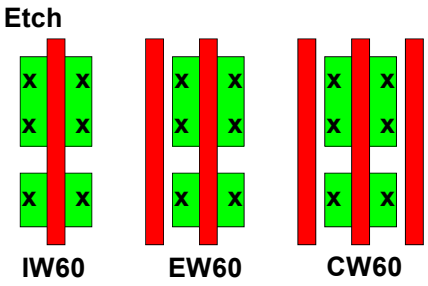
3.3 Summary of the RO Inverter Layouts and the Nomenclature

The nomenclature used for the 45nm RO monitors is detailed in Table 3.2. Cartoon examples of the inverter layouts for the different RO monitors are given in Figure 3.9.The ROs are grouped into families according to the process parameter being monitored: C=Control (with two dummy gates on each side of the gate), I=Isolated gate, E=Etch, F=Focus, M=Misalignment, N=Nitride CESL strain, and S = STI stress.

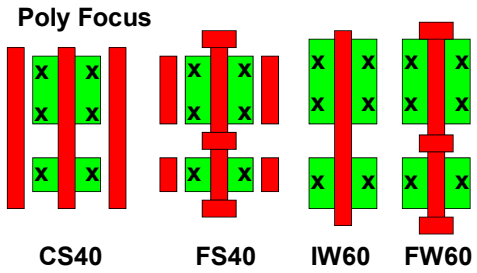
The naming convention is as follows: (1) family name, (2) gate width (W for a wide gate width, where W_{NMOS} = 300nm, and S for a short gate width, where W_{NMOS} =130nm; W_{PMOS}/W_{NMOS} = 1.4), (3) drawn gate length (40nm, 50nm, or 60nm), and (4) subscripts enumerating the different variations of the same monitor. Monitor names without subscripts indicate that there is only one variation for that monitor. For example, CW40 represents a Control monitor with a Wide gate and a drawn gate length of 40nm.

Family Name	**Gate Width (nm)**	**Drawn Gate Length (nm)**
C = Control **I= Isolated** **E =Etch** **F= Focus** **M= H-Shaped Misalignment** **W = Wedge-shaped Misalignment** **N= Nitride CESL** **S = Shallow Trench Isolation**	W: W_{NMOS}= 300 S: W_{NMOS}= 130	40 50 60

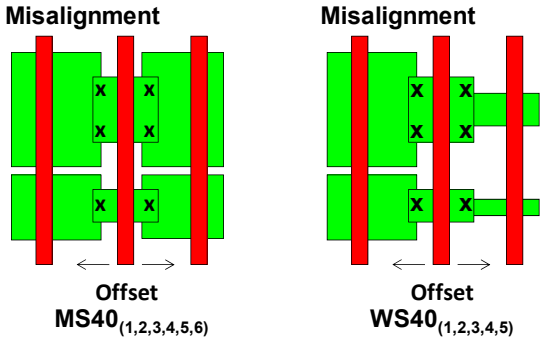
Table 3.1: Nomenclature used for RO monitors



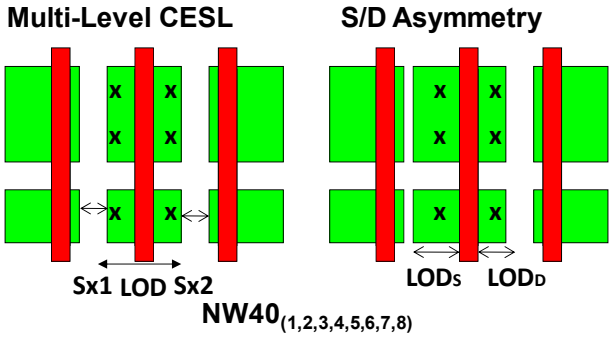
(a): Inverter layouts for etch monitors



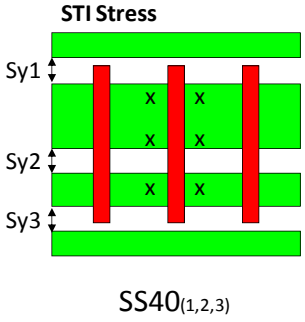
(b): Inverter layouts for gate focus monitors



(c): Inverter layout for misalignment monitors



(d): Inverter layouts for CESL-induced strain monitors



(e): Inverter layout for STI-induced stress monitors.

Figure 3.9: Summary of inverter layouts

Table 3.2 presents the block map for all of the 32 different types of RO monitors.

$NW40_1$	EW40	**$SS40_1$***	**$WS40_5$***
$NW40_2$	CW60	**$SS40_2$***	**CS40'***
$NW40_3$	EW60	**$SS40_3$***	**$MS40_1$***
$NW40_4$	IW50	**$SS40_4$***	**$MS40_2$***
$NW40_5$	CS40	**$WS40_1$***	**$MS40_3$***
$NW40_6$	FS40	**$WS40_2$***	**$MS40_4$***
CW40 ($NW40_7$)	IW60	**$WS40_3$***	**$MS40_5$***
$NW40_8$	FW60	**$WS40_4$***	**$MS40_6$***

Table 3.2: The block map for the RO monitors—RO monitors in (*) have additional capacitors inserted between each inverter stage.

3.4 Pattern Matching Guidance for Design of the RO Monitors for Lithography

A generic annular illumination optical model at the 45nm technology node is used in both the generation of the pattern matching kernel and in assessment of the sensitivity improvement by incremental layout modification using Mentor Graphics Calibre. The designs shown in this section are created with a generic model without regard to the OPC but have a gate length bias of 10nm per edge to approximately compensate for the OPC. The versions of these RO monitor designs for inclusion in the test chip are simulated by ST Micro using proprietary 45nm optical, resist, and OPC models.

3.4.1 A Strehl Ratio Test for Process Variation (PV) Band Calibration

A Strehl ratio test is used to calibrate focus limits with generic parameters for Calibre simulation. The calibration test consists of plotting the peak intensity from a transparent square measured 0.4(λ\NA) on a side. An NA=1.15 and annular illumination with sigma_out=0.9 and sigma_in= 0.75 are used to set the focus limits (although top-hat illumination is also shown). The intensity peaks at and decreases on either side of the best focus. Due to the properties of the generic resist model, for the generic simulation model used, the peak is shifted to -30nm. Since 2/3 of a Rayleigh Unit (RU) corresponds to a 0.9 intensity drop, the depth of focus (DOF) range from the annular curve is found to be -67 nm to +7 nm. These focus settings for $\pm$2/3 RU are used throughout this work to assess layout sensitivity to focus for both poly and active layer imaging.

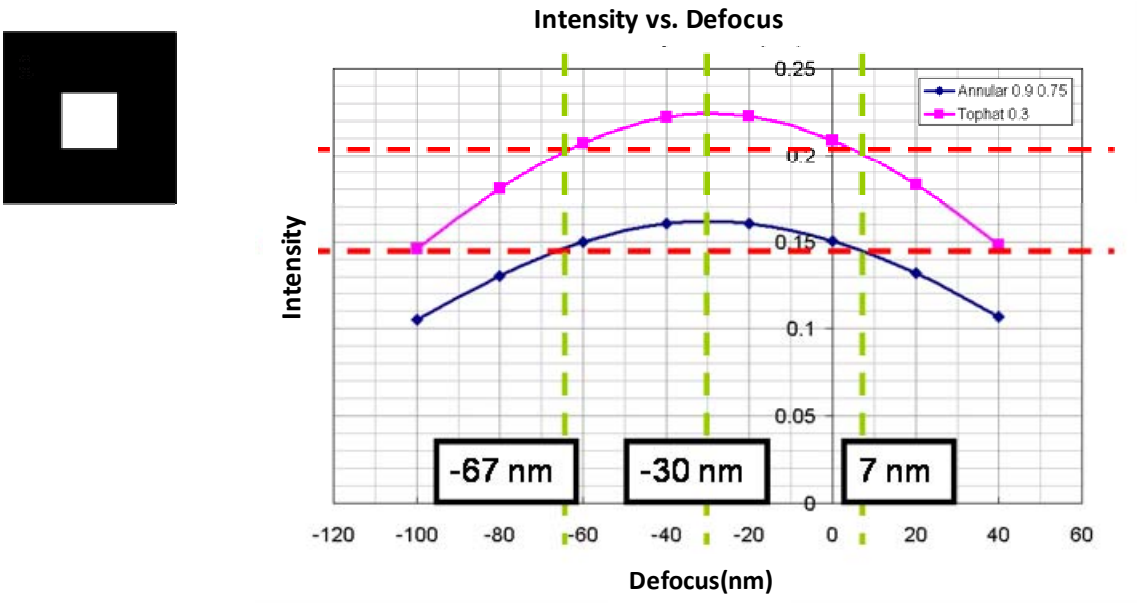


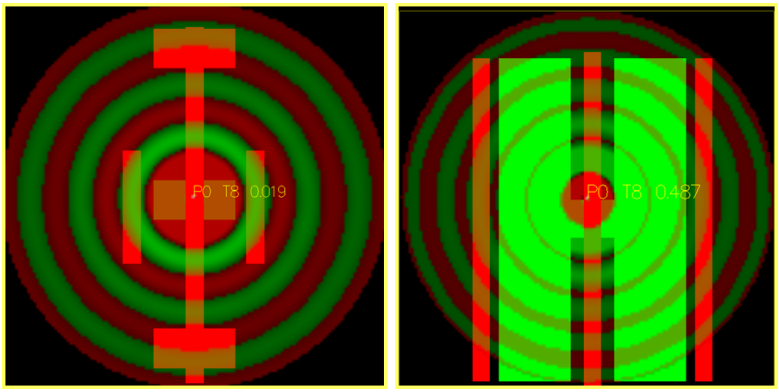
Figure 3.10: A Strehl ratio test for a light square measuring (0.4)λ\NA per side at tophat and annular illumination using Calibre simulation.

3.4.2 Using the Pattern Matcher to Optimize RO Monitor Design

The starting point in creating focus RO monitors is to overlay the maximum lateral test pattern (MLTP) at the middle of the gate in the ring oscillator layout. The MLTP are described in Section 2.4.2 for a given lithography effect, such as gate lithography focus. Then, by modifying the poly or diffusion layout, the similarity to the lateral influence function and, thus, the sensitivity to focus, can be enhanced. The increase in PMF, i.e. the convolution of the kernel with the layout, is used as a guide for the sensitivity improvement for the process effect, such as the lithography focus. Only changes that are consistent with design rules are allowed.

Figure 3.11 illustrates the method for enhancing the focus sensitivity of the layouts. The goal in Figure 3.11(a) is to enhance the sensitivity of the gate to the lithography focus. In this case, the MLTP for focus is overlaid and centered on the gate. The focus pattern suggests adding additional poly features to block the light in the green rings that reduce focus sensitivity, allowing the light through the red rings to increase focus sensitivity. Using this guidance, the initial choices for additional rectangles can then be adjusted in simulation with Calibre to obtain the layout design.

However, it is quite important to modify the MLTP to include the effect of the partial coherence from the annular illumination, as discussed in [36]. The effect of this modification can be seen in the change in the PMF in going from Figure 3.11(a) to Figure 3.11 (b). The goal in Figure 3.11(b) is to increase the sensitivity of the transistor width with focus during printing of the diffusion layer. The MLTP is located where the poly crosses the active (diffusion) mask in this H-shaped diffusion area that is prone to corner rounding with defocus. In this case, annular illumination is used, and the defocus MTLP has been multiplied by the mutual coherence for annular illumination [4]. This function is also oscillatory and produces more nulls and thus a greater green area in the kernel function.



(a) (b)

Figure 3.11: (a) shows a coherent defocus MTLP matched at the poly layer of the layout. (b) shows a defocus MTLP with annular illumination that matches at the diffusion layer of the layout.

3.5 Specific Designs for Etch and Lithography Monitors

3.5.1 RO Monitors for Etch Dependence of Poly

At 90nm, lithography/etch dependence on proximity to the neighboring dummy gates was determined to be an issue [1]. This same style of RO layout with a single and two-sided dummy gate is replicated for the 45nm process. In order to decouple the etch effects from the lithography effects, the optical effects on these layout monitors are calibrated with Calibre. The contours are obtained from aerial image simulations for 60 nm gates in the process variation band conditions determined above. Simulations in Figure 3.12 show nearly identical changes in gate length due to focus. Thus, it can be concluded that if there are any measured ring oscillator frequency shifts, they are more likely due to contributions from etch rather than the lithography focus.

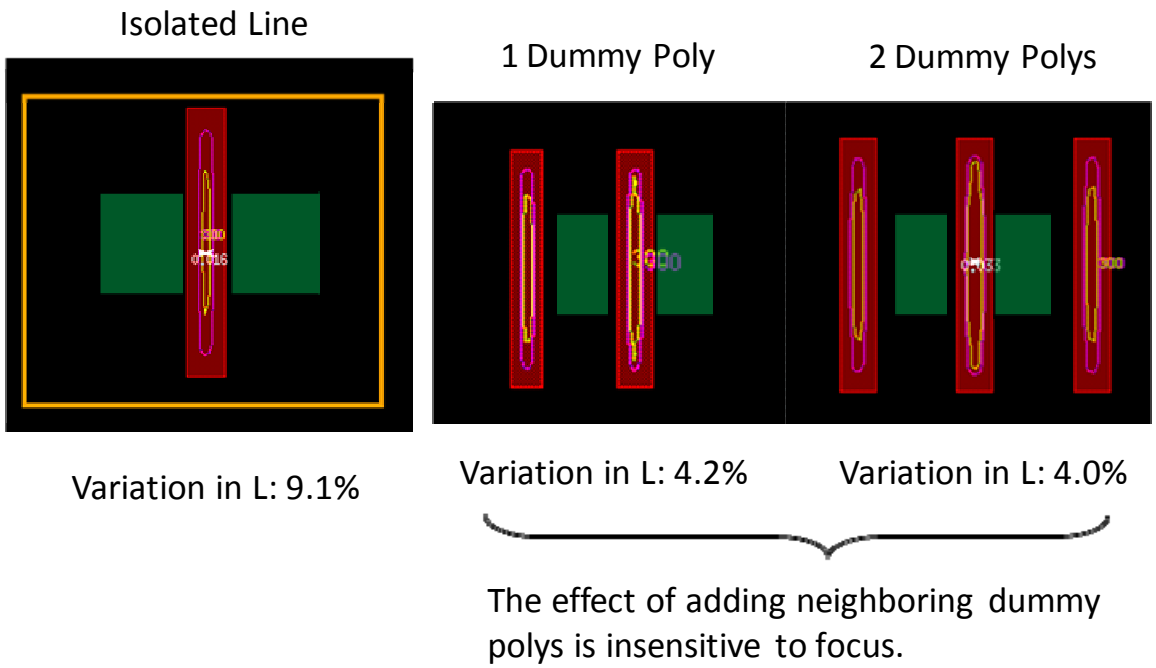


Figure 3.12: Etch RO monitors are designed to show nearly identical changes in gate length due to defocus.

3.5.2 RO Monitors for Focus Dependence of Poly

This section looks at the design and process variation (PV) band behavior of gate focus. The concept of using pattern matching, as noted in Section 3.4.2, is applied to the initial layout design. After these layouts are made design-rule-check (DRC) compatible, they are tuned using Calibre for an optimal response. The focus monitors are designed at 40nm and 60nm drawn gate lengths. The MLTP for focus is overlaid, as shown in Figure 3.11, and the layout is adjusted to increase the similarity of the poly overlap to the color of the central red area. Hammerheads are added to the gate. The dummy polys are truncated where they are on green portions of the lateral influence function in order to make these RO monitors specifically more sensitive to focus. As required by DRC, the 40nm monitors have dummy polys, while the 60nm monitors are not required to have them. As compared to the control, the new RO monitors show an increase in sensitivity to a focus of 1.5x in the simulation and still meet DRC (Figure 3.13).

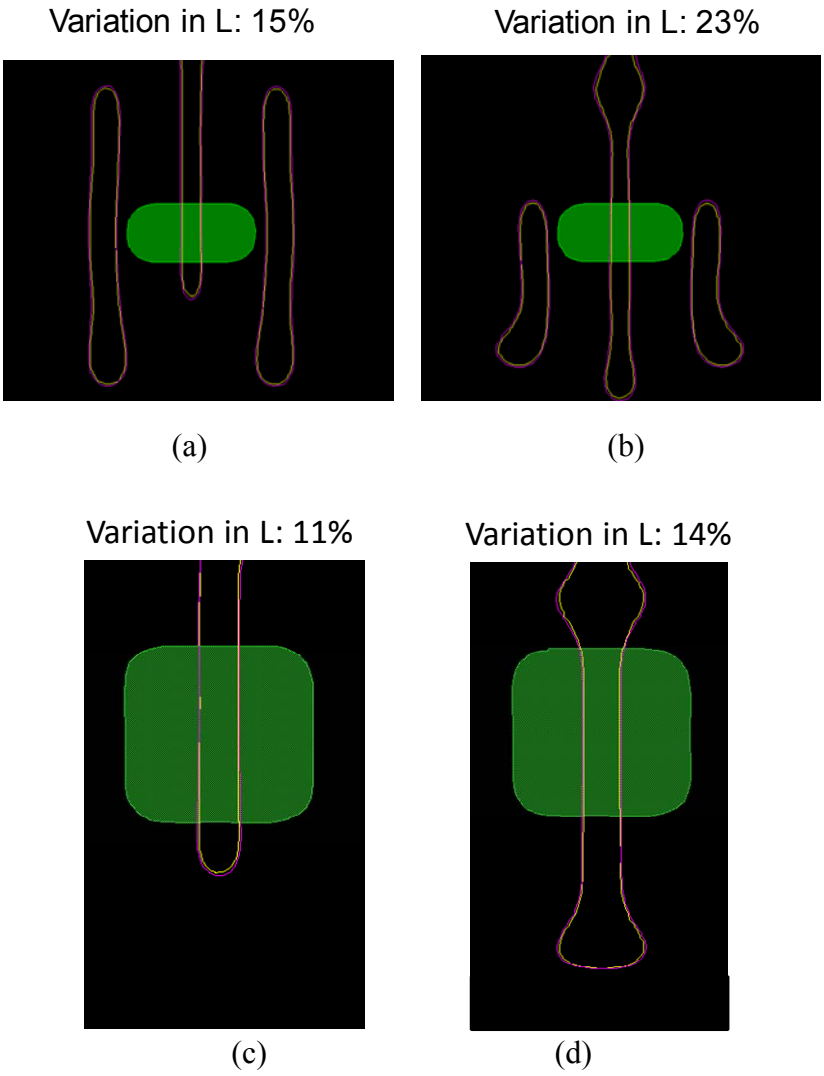
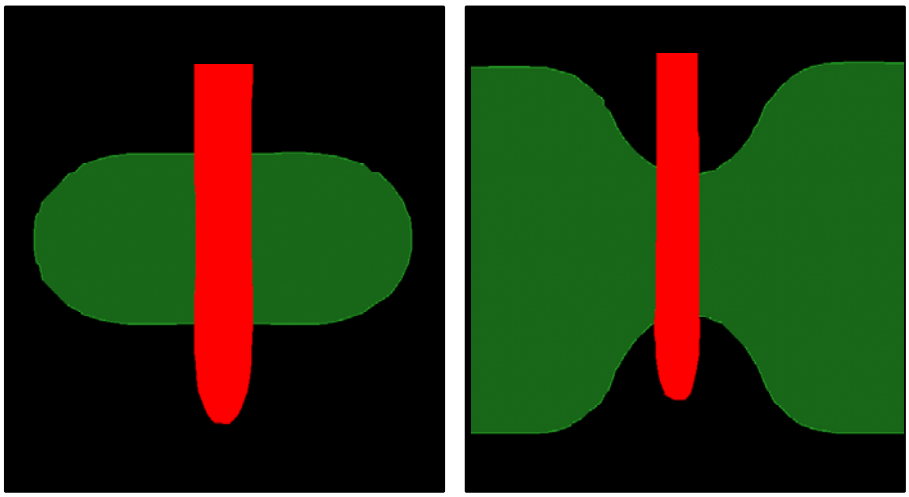


Figure 3.13: (a) and (c) contain the control RO monitors, and (b) and (d) contain the 1.5x focus sensitive RO monitors at 40nm and 60nm, respectively.

3.5.3 RO Monitors for Gate-to-Active Overlay Using Pre-Programmed Offsets

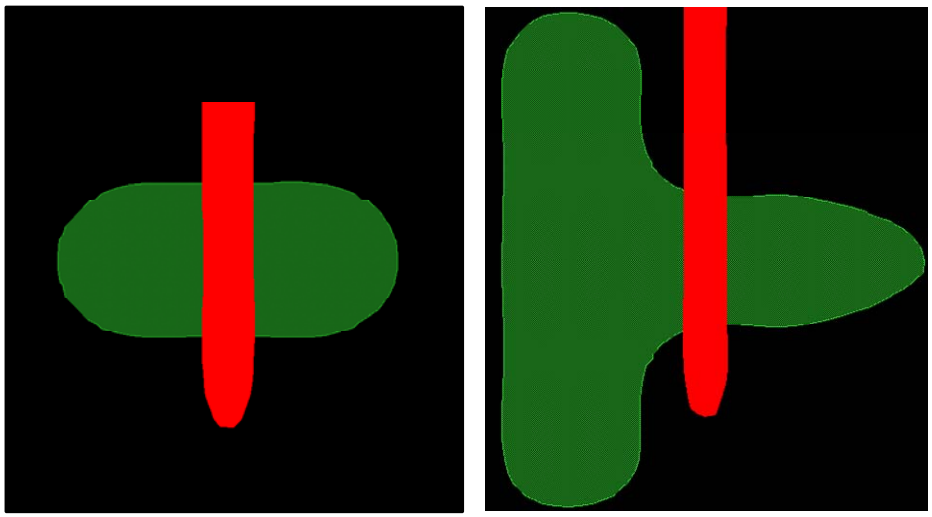
Gate-to-Active Overlay monitors are designed based on the lithographic corner rounding of an H-shaped and a wedge-shaped active area. Figure 3.14 shows that a typical minimum sized gate is fairly immune to pre-programmed gate-to-active offset. Contours from image simulation using Calibre for the rectangular active area show only a 1% change in gate width, with a deliberate offset of 15nm from the center of the gate. For an H-shaped active area, the same offset causes a stronger impact. As shown in Figure 3.14(b), the addition of the large diffusion area increases the curvature, and the same amount gate-to-active offset shows a 4% change in width, increasing the sensitivity 4-fold as compared to that of the minimum width gate.



(a) (b)

Figure 3.14: (a) shows the simulation contours for a minimum-sized active area. (b) presents the H-shaped active area gate-to-active misalignment RO monitor.

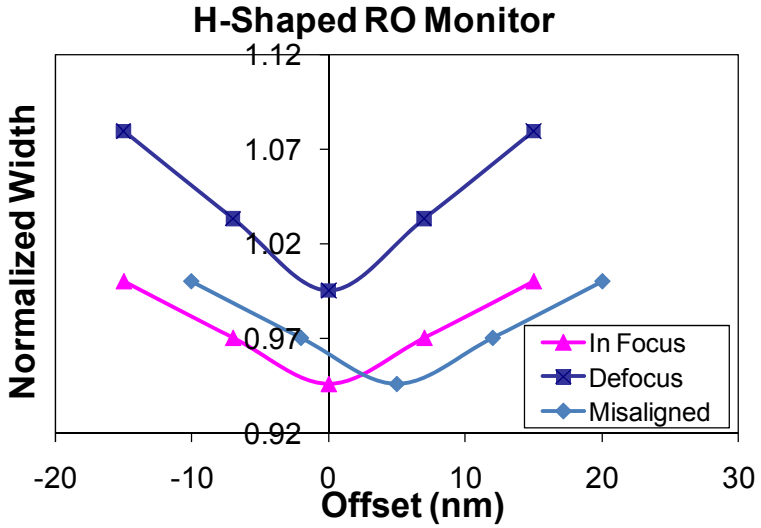
The relative sizing of transistors in SRAMs often introduces wedge-shaped diffusion areas, creating sensitivity to misalignment. This inspired the creation of the wedge-shaped RO monitor, which can be seen in Figure 3.15. Here the active area decreases as the gate position shifts from left to right. A 10nm shift of the gate leads to a gate width change of 3%, thus increasing the sensitivity by 3 times when compared to that of the rectangular diffusion layout for the minimum width gate seen in Figure 3.15(b).



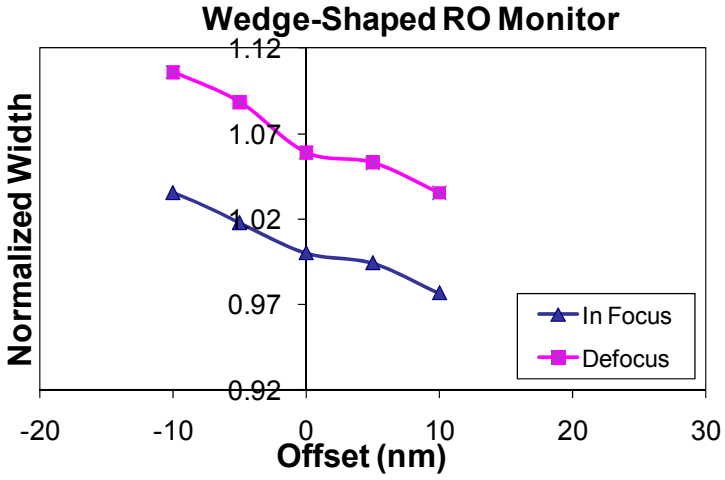
(a) (b)

Figure 3.15: (a) shows the simulation contours for a minimum-sized active area. (b) presents the wedge-shaped active area gate-to-active misalignment RO monitor.

H-shaped and wedge-shaped programmable misalignment monitors are designed with five pre-programmed gate-to-active offsets. Pre-programmed misalignment mapped to changes in gate width is shown in Figure 3.16. For each pre-programmed offset, the gate width at the offset is extracted from the Calibre simulation contours at the two levels of defocus, -30nm and -67nm. The H-shaped layout monitor is characterized by a parabolic curve, and the minimum identifies the best alignment. The wedge-shaped monitors return a characteristic curve with a slope of 0.8% change in RO frequency per 1nm defocus at a 0nm gate-to-active offset, as shown in Figure 3.16. A misalignment of several nm will occur during production. The shift in the symmetric point for the H-shaped monitor will quickly and unmistakably identify the magnitude of the overlay error. Defocus for the H-shape monitor theoretically increases the parabolic curvature.



(a)



(b)

Figure 3.16: Normalized width versus pre-programmed offsets for the H-shaped monitors at 0nm, ±10nm, and ±15nm, shown in (a) and for the wedge-shaped monitors at 0nm, ±5nm, and ±10nm, shown in (b)

In order to mitigate the self-loading compensation of a gate width effect, a loading technique is introduced. The RO frequency sensitivity to changes in gate width is reduced here due to the capacitive loading and drive current scale similarly and thus tending to cancel each other out. This effect can be mitigated by using additional sizeable capacitance loading that is independent of the gate widths such that the fractional change in capacitance is small. This mitigation can be achieved by inserting large, fixed capacitors in between the inverter stages. One option is to implement these capacitors using transistors with source and drain tied to the power supply or ground, as shown in the schematic in Figure 3.17. Figure 3.17 also shows a plot of RO frequency versus normalized gate width via simulations in SPICE with and without the added transistor capacitors. Here the drawn inverter layout width is varied from 80% to 120% in order to observe the sensitivity of the frequency to gate width change. Without the addition of capacitors, the sensitivity at (1, 1) is 0.2. With the addition of capacitors, the sensitivity increases to 0.8. Thus, 80% of the transistor width variation shown in Figure 3.16 should occur in RO frequency variation. The ring oscillator frequency is about an order of magnitude slower with the additional capacitors.

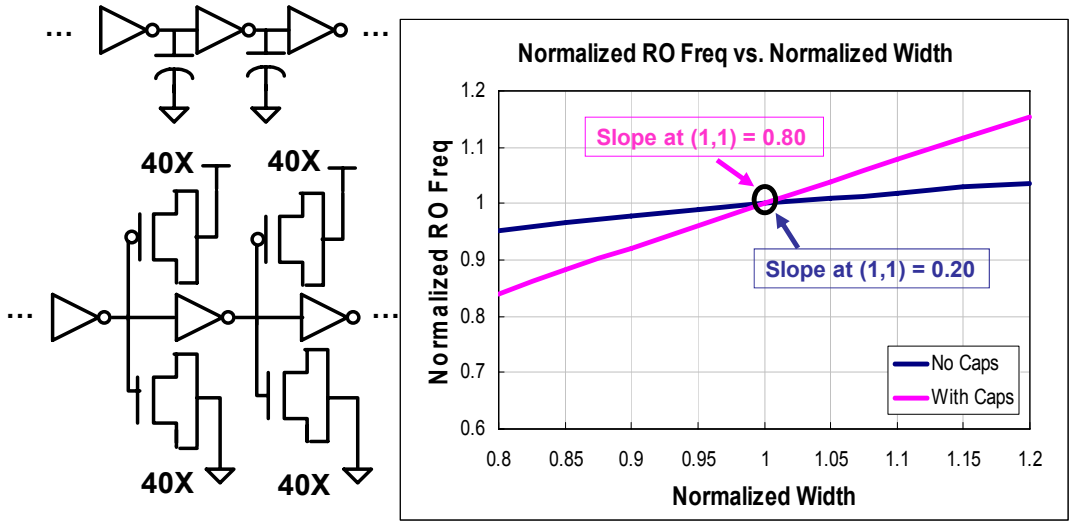


Figure 3.17: This plot of normalized ring oscillator frequency as a function of gate width, with and without capacitive load, shows that added capacitors increase the slope from 0.2 to 0.8.

3.6 Designing Nitride CESL-Induced Strain and STI-Induced Stress RO Monitors

In order to characterize the impact of strained-silicon technology, RO monitors are designed at 45nm technology so as to examine the variation due to source/drain capping layer stress, STI stress, source/drain asymmetry, and gate stress. Previously measured 45nm test chip results demonstrated that an increase of 160nm in the source/drain length of diffusion (LOD) induced a 5% change in the RO frequency caused by the increase in capping-layer-induced strain for larger LOD [46]. However, this experimental screening was done at only two levels. In addition, prior work by Khakirfirooz demonstrated that with the scaling of devices, source-side drain of a device has a larger impact on transistor injection velocity than drain-side strain [47].

Leveraging these prior works, 8 new RO monitors with multiple levels of LODs, including source/drain asymmetries, are designed to systematically characterize the impact of

strained-silicon technology on transistor performance. The intent here is to capture the detailed length behavior for comparison with concurrent modeling studies being carried out by Xu [7].

Three STI RO monitors are used to study STI influence in lateral and vertical directions of the channel. In the lateral direction, the monitors are designed with various levels of STI widths that lie in between the transistor and its adjacent dummy gates. In the vertical direction, monitors are designed by varying the STI widths that lie in between NMOS and PMOS devices. Since PMOS devices are expected to be more sensitive to changes in STI width when oriented in the <110> channel direction and NMOS in the <100> direction, measured RO frequency sensitivity from these monitors can be used to determine channel orientation [23].

3.7 Summary

Circuit performance characterization using ring oscillators accessed via scan-chains has been generalized by adjusting layout features to enhance their response to particular physical effects. In so doing, parameter-specific quantitative monitors are created. A set of 32 RO monitor layouts was designed with generic image models that met design rules under the assumption that OPC would not be used. Pattern Matching with known lateral spillover functions gives helpful guidance. For example, by adding poly features outside of the device area, the sensitivity of poly to focus is doubled. Combinations involving layout changes on multiple layers and multiple levels of programmed offsets are shown to be very effective through simulations. Wedge-shaped misalignment monitors are used to produce a sensitivity of a 0.8% change in RO frequency per 1 nanometer alignment and, in the case of the H-shaped monitor, the misalignment can be determined without calibration. In practice, the misalignment monitors based on the transistor width changes are implemented with a 0.8 sensitivity to gate width change by adding large fixed capacitors so as to reduce the impact of self-loading. Five levels of pre-programmed gate-to-active misalignment are used to identify the direction of the overlay-induced shift in a parabolic response. Stress is a new frontier where intuitive layout changes are now being utilized and will later be coupled with systematic process simulations. There were eight RO layout monitors designed to characterize the RO frequency sensitivity to active area lengths and asymmetries. Here, three RO monitors were designed to monitor STI-induced stress.

Chapter 4
45nm Ring Oscillator Testing Methodologies and Noise Analysis

This chapter reports on initial RO frequency measurements and describes the testing methodologies used for the automating measurement collection. It also includes an analysis of noise attributed to the testing environment. Systematic across-wafer and random within-chip variation for control-case RO monitors are discussed and used to guide the analysis of the systematic and random components of parameter-specific RO in the subsequent chapters.

This chapter is organized into 5 sections. Section 4.1 describes the procedure and test setup for measurement collection. Section 4.2 discusses the measurement jitter and drift from the testing system and environment. Section 4.3 reports on the initial RO frequency measurements and shows that within-chip block-to-block variation is negligible, meaning that measured RO frequency for 36 RO instantiations can be used to interpret RO frequency sensitivity to parameter specificity. Section 4.4 discusses the across-wafer variation, within-chip standard deviation to mean ratio, and the estimated residual random variation in the threshold voltage of individual devices for control-case RO monitors. Section 4.5 summarizes the chapter.

4.1 Automated RO Frequency Measurement Setup and Testing Methodologies

4.1.1 An overview of the RO Frequency Measurement Test Setup

Automation of the RO frequency measurements here is conducted using a test setup in the Berkeley Wireless Research Center (BWRC) developed by L.-T. Pang with assistance from Susan Mellers. Figure 4.1 shows the test setup for RO frequency measurement collection [1]. The following instruments are used: a pattern generator (Tektronix DG2020A), an oscilloscope (Agilent Infineum, with a 20 $x10^9$ sample per second resolution), and a computer (logic analyzer, with a 150 picosecond resolution). In addition, a printed circuit board is used to hold a packaged chip, and on-board voltage regulators are utilized in order to adjust the power supply voltages.

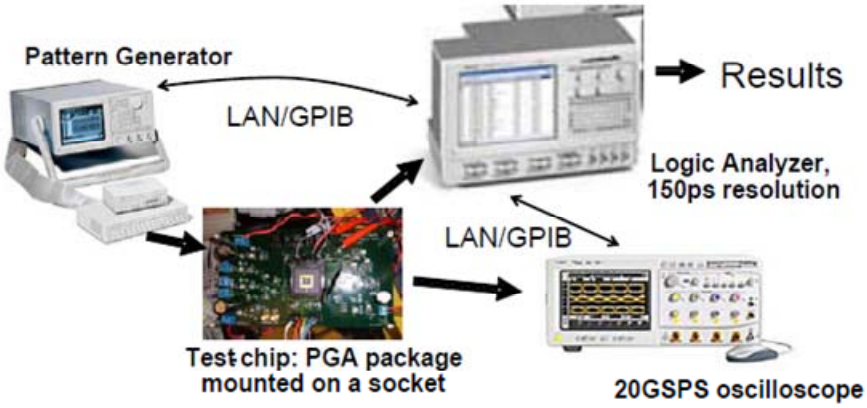


Figure 4.1: Measurement setup [1]

The pattern generator generates the address bits and the control signals to enable a ring oscillator (RO). The oscilloscope measures the RO frequency, and the computer records the measurement. Perl script commands are used to coordinate the instruments for automatic measurements. These commands are sent via the GPIB (National Instruments GPIB-ENET/100) interface to the instruments.

The test chips are packaged in ceramic PGA packages. Each packaged chip is manually placed into a socket that is soldered onto the custom PCB for measurements. The customized PCB for measuring the 45nm test chip is shown in Figure 4.2. The PCB contains components that generate regulated supply voltages from external 5V and 10V power supplies. Additional components on the PCB include: voltage regulators, capacitors, resistors, inductors, pins, connectors, switches, and pins.

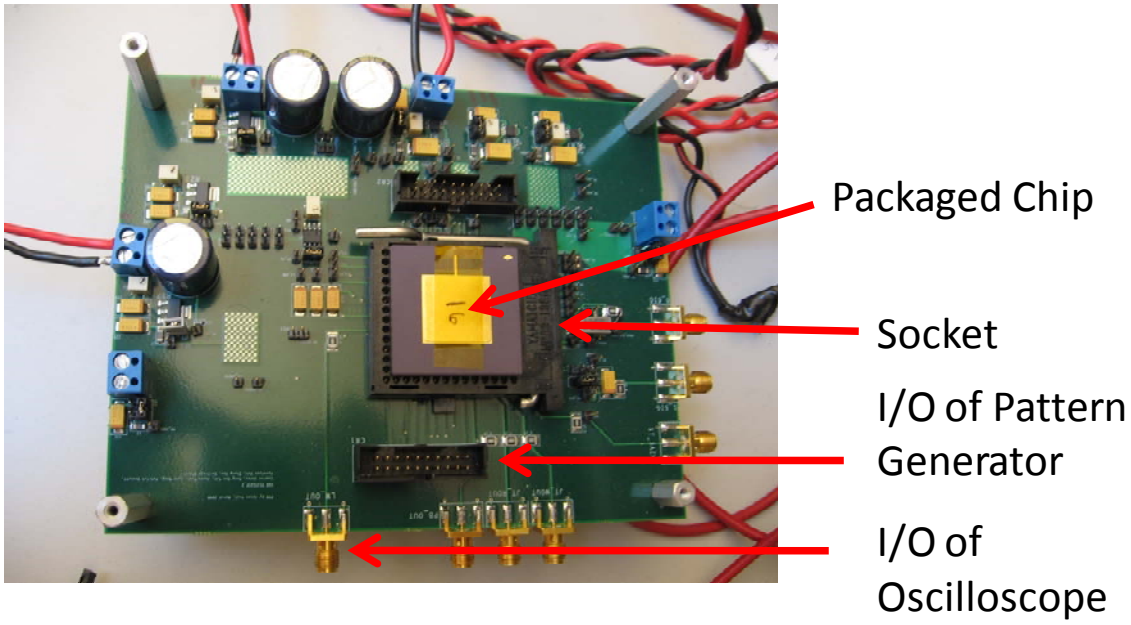
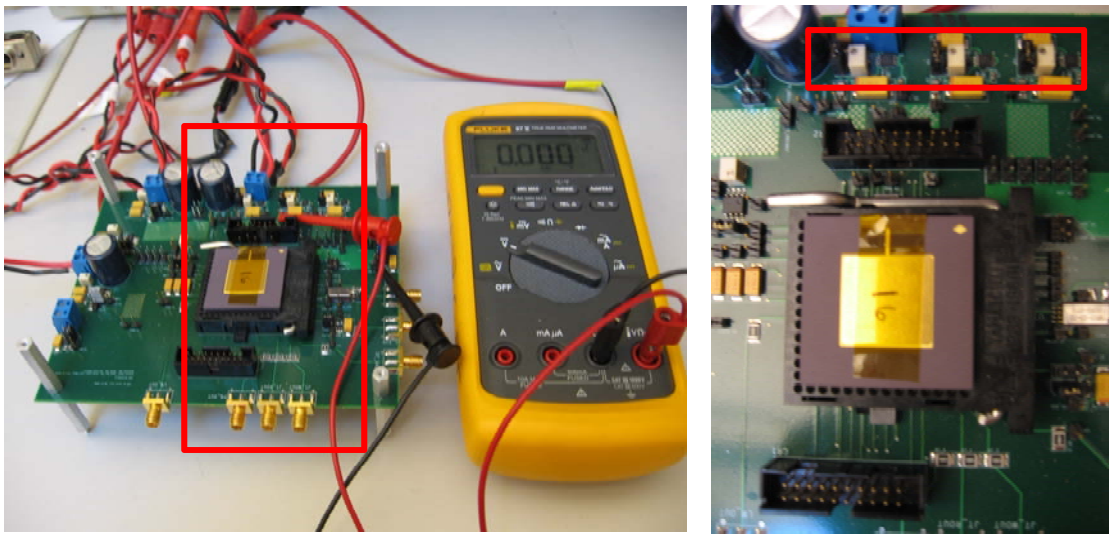


Figure 4.2: Chip and PCB Photo

4.1.2 Adjustment of Operating Voltages and Temperatures

To change the operating voltage V_{dd}, the on-board voltage regulators are manually adjusted using a screwdriver and can be set to an accuracy of $\pm$1mV. Changes in operating voltage are monitored with a Digital Multi-Meter (DMM). Figure 4.3 shows the test setup with the DMM in place.



(a) (b)

Figure 4.3: Test setup for operating voltage (V_{dd}) experiments: the on-board regulators in (b) are adjusted for different voltage settings using a screw driver.

To increase the chip's operating temperature, an electric hotplate is placed underneath the PCB to heat up the chip. A temperature sensor is placed on the chip to monitor its temperature. Figure 4.4 shows the test setup for such temperature experiments. Since the testing environment is not isolated, the temperature of the chip fluctuates with that of the environment. A $\pm$ 5 ^{O}C fluctuation is typically observed during this measurement. The measured RO frequency sensitivity to voltages and temperatures are presented in Chapter 7, revealing that the error contributed by the testing environment is negligible.

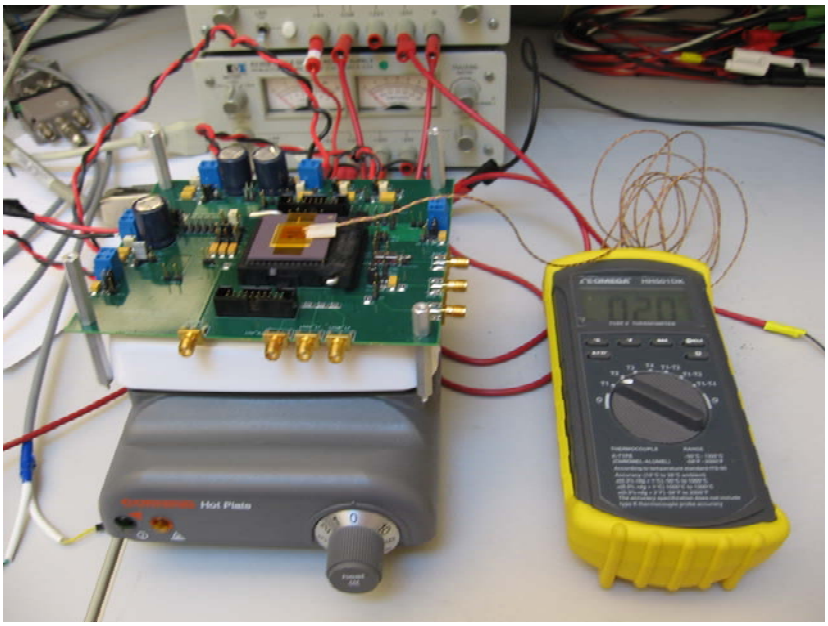


Figure 4.4: Test setup for temperature experiments

4.1.3 RO Measurement Sequence

Each chip contains 3 RO blocks, and each block contains 12 (4x3 array) complete sets of RO instantiations such that each of the RO set contains 32 types of parameter-specific RO monitors. Thus, for each chip, each type of RO monitor has 36 instantiations. The RO monitors are measured sequentially, traversing columns before rows, as enumerated in Figure 4.5.

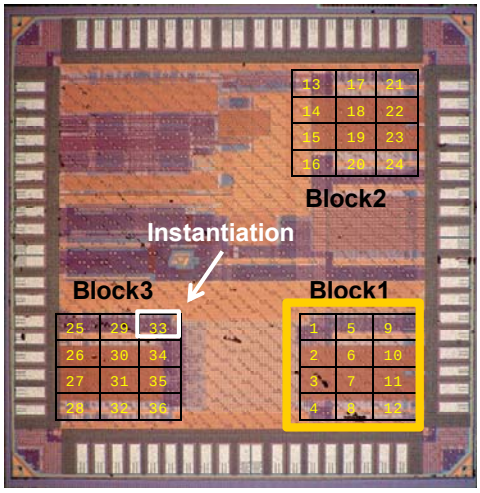


Figure 4.5: RO monitors are measured sequentially from 1-36.

4.2 Measurement System Noise and Drift

A valid testing methodology is one in which the measured results are reproducible. While due to many sources of variability induced by the testing environment, the reproduced results may not perfectly match those that were previously measured, they still must agree within a certain margin of error. This section discusses two sources of error attributed to the testing system and environment: measurement jitter and drift.

4.2.1 RO Measurement Jitter

To reduce the anticipated jitter noise from the sampling by the oscilloscope, the number of repeated oscilloscope measurements (i.e. samples) for each RO at a given scan chain address is immediately repeated. Figure 4.6 shows the number of samples taken by the oscilloscope for each 32 RO type for one RO block on a typical chip. In order to reduce the anticipated noise and jitter, the oscilloscope samples each RO 120 to 140 times. Through Perl scripts, the number of samples is set to be greater than 100 prior to saving the mean measured RO frequency to the computer. Since the oscilloscope operates asynchronously with the Perl scripts, the number of samples varies from RO to RO.

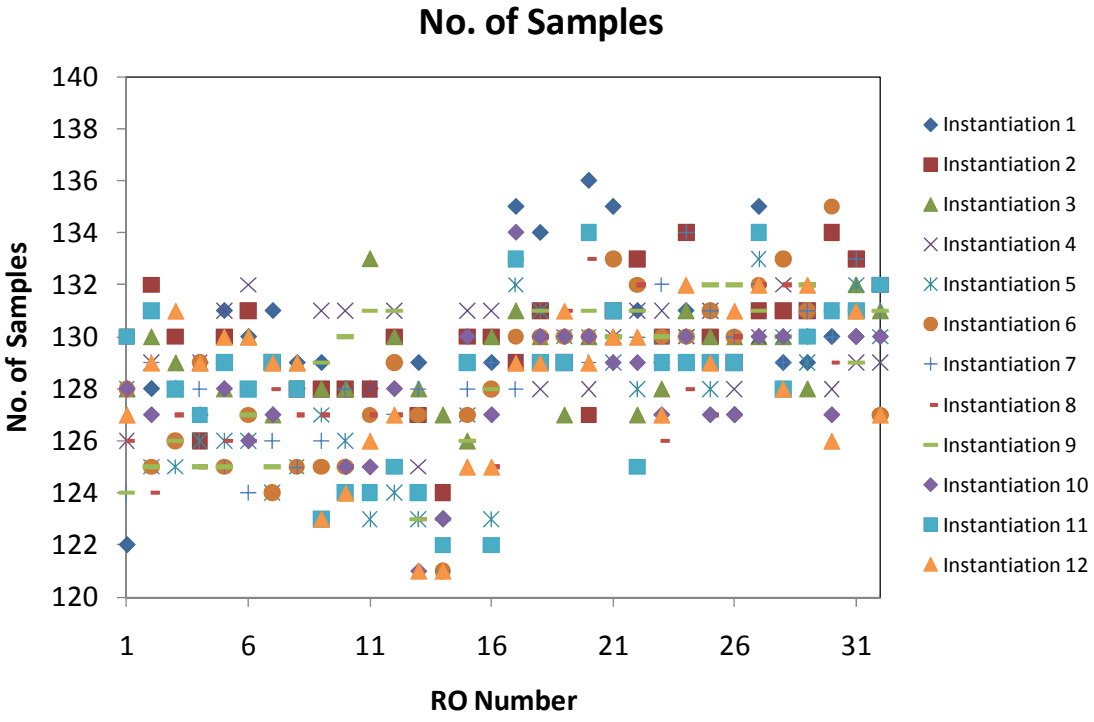


Figure 4.6: The number of samples per RO in one RO block for a typical chip ranges from 120 to 140.

In order to analyze the impact of the noise in the oscilloscope measurements, the amount of noise for each RO sample set in Figure 4.6 is quantified in Figure 4.7. Figure 4.7 presents a plot of the standard deviation to mean percentage for each RO. The amount of noise for each sample set is small, with the largest standard deviation to mean percentage at only 0.009%. Hence, this result reveals that the amount of jitter from the oscilloscope is negligible.

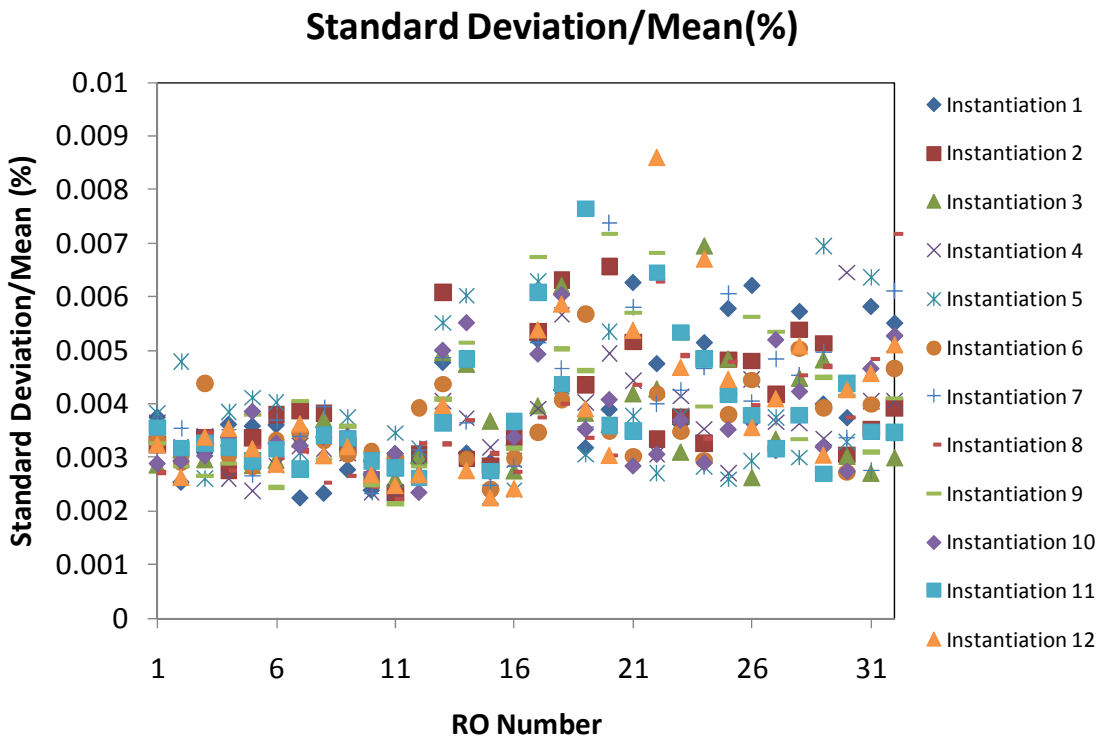


Figure 4.7: The maximum standard deviation to mean percentage is 0.009%.

Since the measured jitter from the oscilloscope is sufficiently low, the measurement time can be saved. Each RO block takes 1 hour to measure; since there are 3 blocks per chip, each chip takes approximately 3 hours to measure. Figure 4.7 shows that the number of samples can be reduced by at least 10x, thus speeding up the measurement process by nearly 10x. In the future, using tens instead of hundreds of samples is recommended so as to reduce the measurement time. For the subsequent sections of this chapter, the mean of the over one hundred RO frequency samples is represented as *one* RO frequency measurement.

4.2.2 RO Measurement Drift

Another source of variability attributed to the testing environment is that of measurement drift, where re-measured results and those that are previously measured differ systematically. This section explores the impact of measurement drift under two conditions: (1) the repeated measurement of a RO on one chip with no intervention, and (2) repeated measurement of a RO on one chip with intervention, where the test setup is taken apart and reassembled back to its previous state for the second set of measurements.

The testing methodology used to generate Figure 4.8 is as follows. For one chip, all of the 32 RO types (each with 36 instantiations) are measured. Three hours later, after all of the RO monitors have been measured once, the RO monitors are measured again. The percentage difference of the measured and re-measured results is shown in Figure 4.8, where $\mu_{36,1}$ and $\mu_{36,2}$ are the mean measured RO frequency for 36 RO instantiations of that particular RO type taken at two different time stamps. The percent error increases for RO monitors with a smaller channel area (RO No. 17 to RO No. 32). Note that the measured RO frequency drift is systematic such

that $\mu_{36,1}$ is always greater than $\mu_{36,2}$. The maximum percentage difference is 0.14%, and temperature change and power supply drift may be a factor.

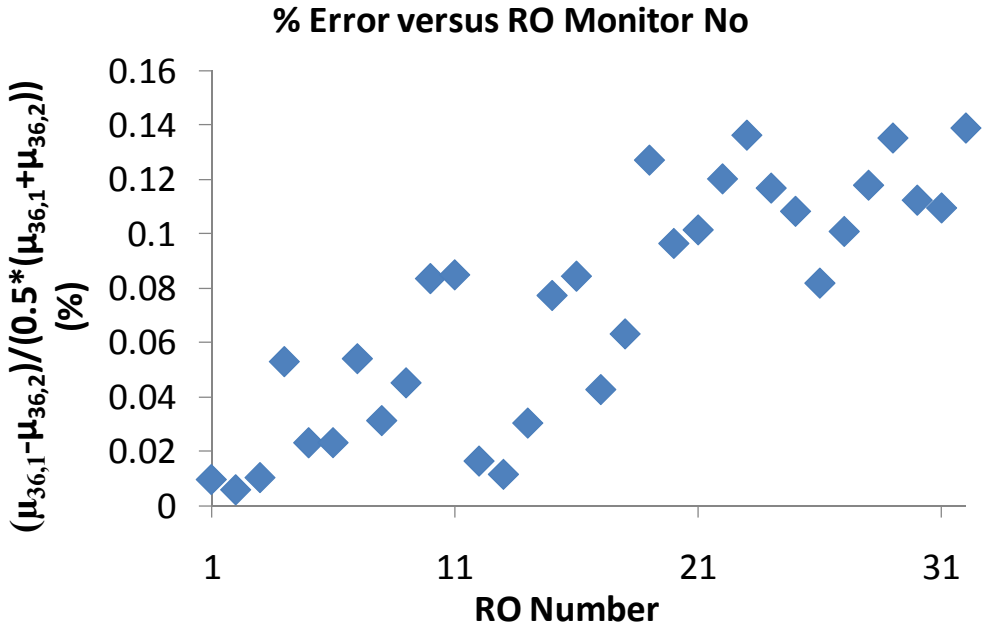


Figure 4.8: Maximum percentage difference between the measured and re-measured data without intervention for a chip is 0.14%.

The testing methodology used to generate Figure 4.9 is as follows. For one chip, all of the RO monitors are measured. The testing equipment is taken apart and reassembled to its previously state, and the same RO monitors are re-measured. The percentage difference of the measured and re-measured data is shown in Figure 4.9. As expected, the observed drift in Figure 4.9 is greater than that seen in Figure 4.8. The latter drift increased to 0.16%, and this may have been due to hand re-adjustment of power supply voltages. Again the noise is larger for the RO monitors with a smaller channel area.

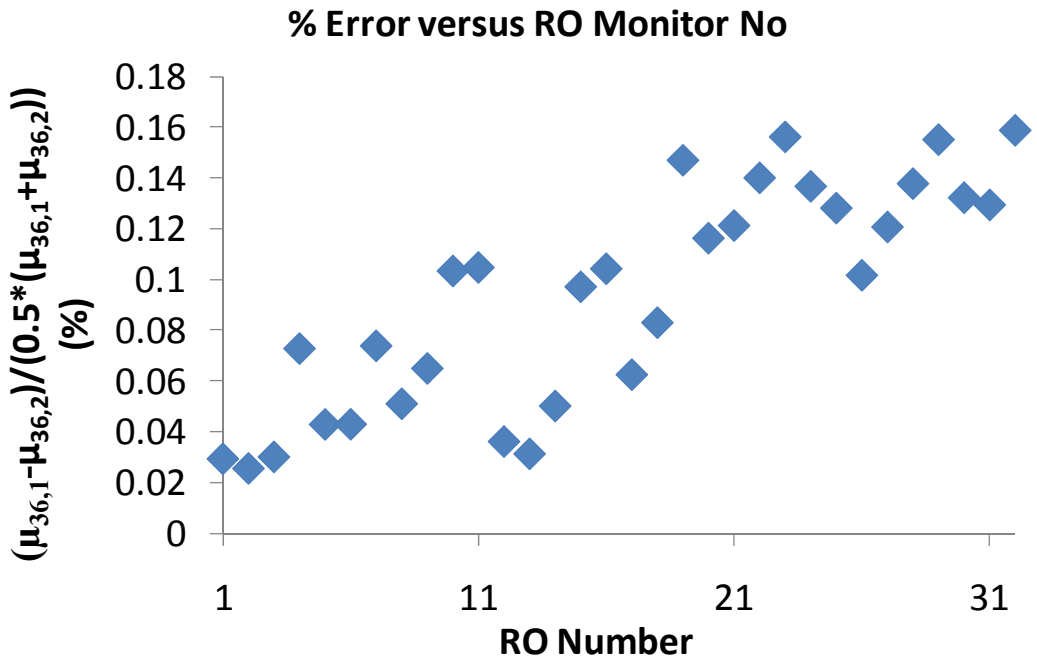


Figure 4.9: The percentage difference of measured and re-measured data for a chip is greater if the test setup is reassembled.

4.3 Measured Block-to-Block Variation

To capture the lithography scanner's scan-and-slit variation, each RO block is designed to be instantiated at three distinct locations in a chip. Since the radius of influence for lithography is on the order of millimeters, and the chip area measures only 2x2 mm^2, a small block-to-block variation may be expected.

Measured block-to-block variation is examined for 3 control-case RO monitors: CW40, IW50, and IW60, for one typical chip. Figure 4.10 shows the measured RO frequency normalized to its own mean for 36 RO instantiations on a typical chip, that of Chip 8. Each of the 12 RO monitors is instantiated at a different RO block location (refer to block locations in Figure 4.5).

Note that for the same RO layout the pattern within each of the 3 blocks is uncorrelated, suggesting that position within the block has little effect on the measured RO frequency sensitivity. Please also note that for a given RO block, the pattern for each of the 3 control-case layouts is uncorrelated, suggesting that the location of the RO instantiation within the set of 32 RO monitors has little effect on measured sensitivity. Block 3 shows slightly faster measured RO frequency than that of the other blocks, in which the mean measured RO frequency between blocks has a maximum difference of ~1.10%. The magnitude of across-block mean variation is then compared to that of the within-block random variation. For each block (1-3), the maximum-to-minimum peak-to-peak range is 4.56%, 4.24%, and 4.32%, respectively. The block-to-block average difference is 4x smaller than within-block random variation and approaches the expected mean random distribution for 12 samples, i.e. 4.4%/sqrt(12).

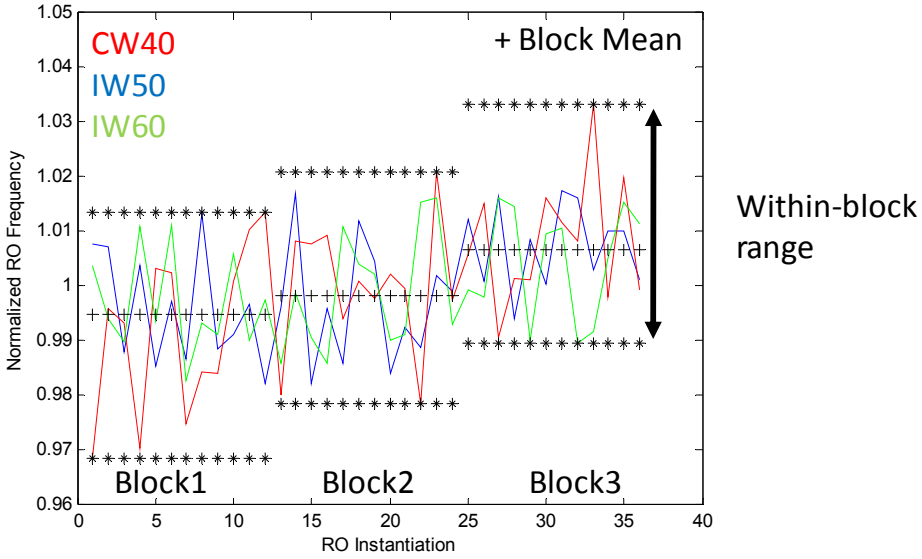


Figure 4.10: The measured RO frequency for 36 RO instantiations for one typical chip, that of Chip 8.

Block-to-block variation is examined for multiple chips. Figure 4.11 shows measured RO frequency normalized to the chip mean for a RO monitor: CW40 on 4 chips spaced at least one exposure field apart on one wafer. Each row contains the across-chip measured RO frequency for 36 RO instantiations. Each column contains the measured RO frequency for 12 RO instantiations in a 4x3 array for a RO block. The gray scale color spectrum shows a $\pm$5% difference in RO frequency, where black represents the slowest RO frequency and white represents the fastest.

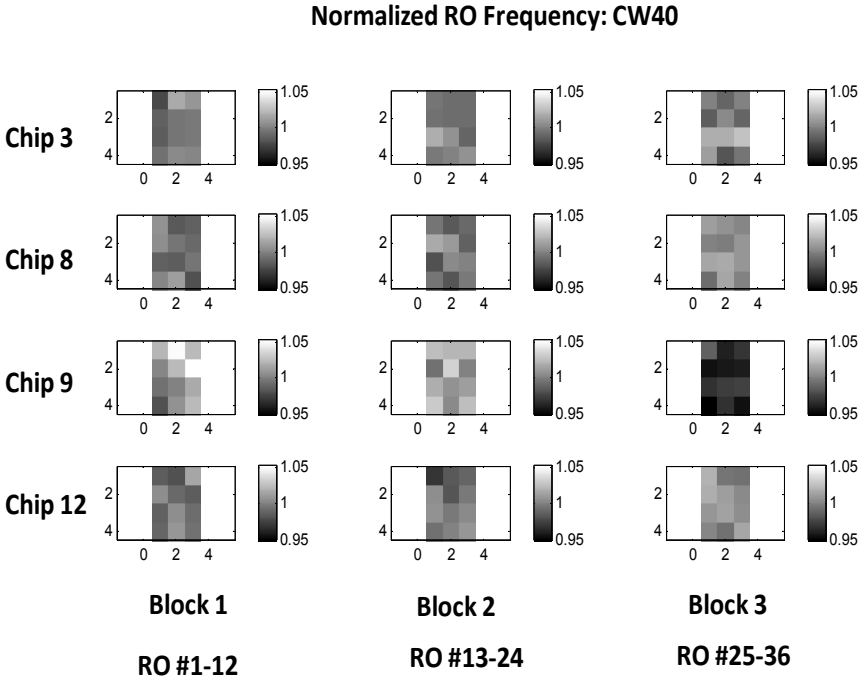


Figure 4.11: The measured RO frequency normalized to the mean for that chip shows random within-block variation.

On a given chip, the measured within-block RO frequency range is typically 4% and there is no obvious appearance of a systematic pattern. Blocks 1 and 2 show a similar mean measured RO frequency, where the typical percentage difference in their means is 0.1%. RO instantiations in block 3 usually operate at a 1.0-1.5% faster RO frequency than in the other two blocks, with the exception of Chip 9, which is an outlier chip. This block-to-block variation is small compared to the within-chip variation, which is typically 3-4%. This result again shows that random within-block variation can be 4x greater in magnitude than block-to-block systematic variation. This noise data for repeated measurements and for comparisons between blocks allows one to set the following guidelines for the analysis of the parameter-specific RO frequency measurements. Noise contributions from measurement jitter, drift and block-to-block variation are negligible. Chip-level averages (based on 36 RO instantiations) can be used to assess the process-specific sensitivity for each RO design where sensitivities greater than 1% are expected. Yet, the smaller random residuals should be analyzed on a block-to-block basis in order to ensure that random variation occurs.

4.4 Measured Across-Wafer Variation

To examine the across-wafer effects, the measured chip's mean and within-chip variation across the 17 chips are examined. Across-wafer and within-chip variation for control-case RO monitors are two important base-lines. Here RO monitors CS40, CW40, IW50, and IW60 are used and their standard deviation to mean ratio is compared. Unfortunately, since a wafer map is not available, the measured RO frequency is plotted versus chip number. The within-chip variation gives a first order estimate of the randomness that is likely associated with that from the individual devices.

Figure 4.12 shows the mean RO frequency, normalized to the mean across 17 chips for control-case RO monitors. All 4 control-case RO monitors have similar systematic oscillatory behavior with the largest measured RO frequency range of 11.11%, occurring for the smallest gate length and smallest gate width, i.e. the smallest channel area. The double humped pattern may be a systematic across wafer factor due to a bull's eye type of effect.

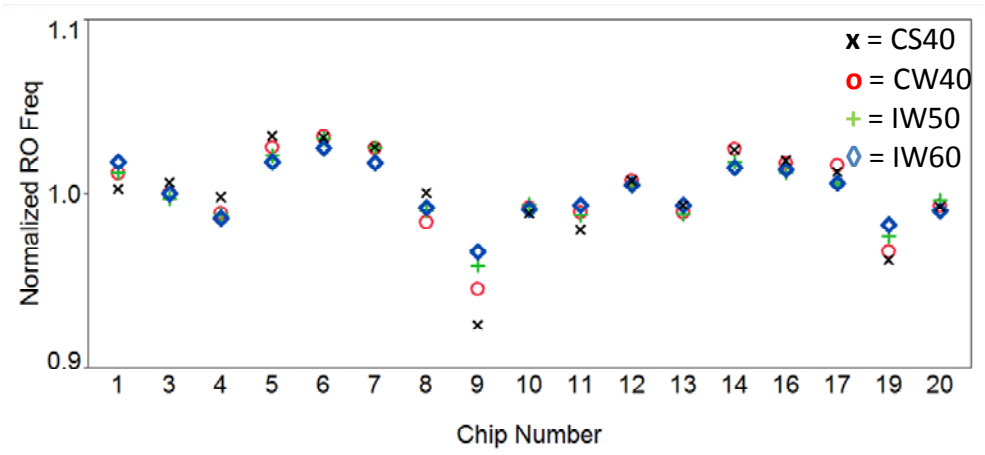


Figure 4.12: The mean measured RO frequency normalized to the mean for the RO monitor designs CS40, CW40, IW50, and IW60 show an across-wafer variation range of 11.11%, 8.84%, 7.35%, and 5.99%, respectively.

Within-chip variation is examined by quantifying the standard deviation to mean ratio of measured RO frequency for 36 RO instantiations, denoted as $(\sigma/\mu)_{36}$. In Figure 4.13, $(\sigma/\mu)_{36}$ for each chip is shown for the three gate lengths and two gate widths control-case layouts in Figure 4.12. Here $(\sigma/\mu)_{36}$ is about 0.2% for most chips, but is unusually high for chips 4 and 9. Excluding chips 4 and 9, the average $(\sigma/\mu)_{36}$ across 17 chips for CS40, CW40, IW50, and IW60 are: 0.30%, 0.20%, 0.19%, and 0.18%, respectively. For the same gate length, the measured $(\sigma/\mu)_{36}$ for RO monitors with a smaller gate width is higher, at 0.3%.

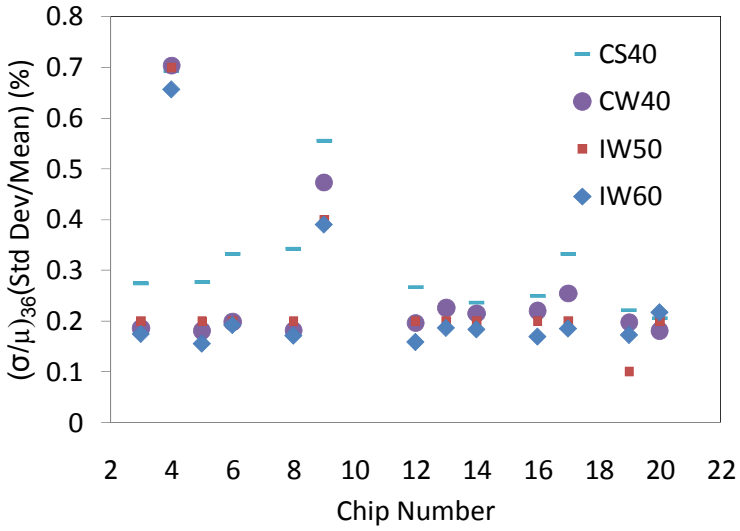


Figure 4.13: Within-chip $(\sigma/\mu)_{36}$ distribution shows that outlier chips 4 and 9 are noisier when compared to typical chips.

A correlation of noise and channel area is shown in Figure 4.13. The measured $(\sigma/\mu)_{36}$ versus the inverse square root of the transistor channel area for a typical chip is plotted (Figure 4.14). A linear regression fit shows a slope of 0.2% increase in $(\sigma/\mu)_{36}$ per a 1 um^{-1} decrease in the square root of the area.

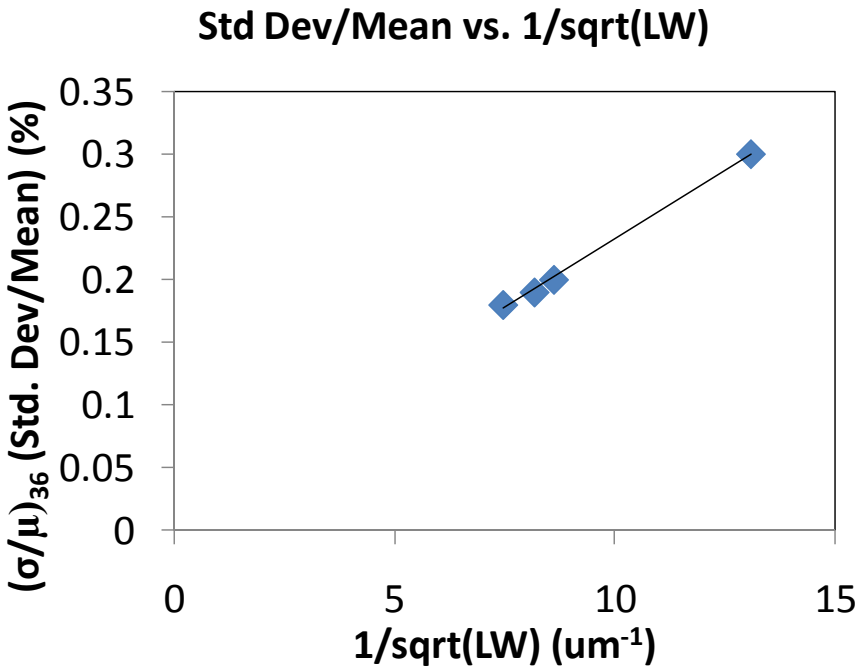


Figure 4.14: RO monitors employing smaller-area inverters show greater variation in regard to measured RO frequency.

The inverse square root of channel area behavior in Figure 4.14 suggests that the cause is likely due to the electrical variation in individual devices themselves. To verify the feasibility of this assumption, the RO frequency behavior in Figure 4.14 is converted to the quantitative variation of individual devices, in particular, to the variation in threshold voltage—commonly referred in the literature as the Pelgrom model [12]. Two steps are necessary in order to make this conversion. The first step is to determine the RO frequency sensitivity to threshold voltage. The second is to correct for the number of devices independently contributing to the RO frequency variation.

To complete the first step, RO frequency versus changes in threshold voltage (V_{th}) for control-case monitors is simulated in SPICE using the BSIM4/PSP model card from ST Micro's 45nm PDK. The change seen in V_{th} is assumed to correlate between NMOS and PMOS devices as well as among all inverter stages. The simulation results are shown in Figure 4.15.

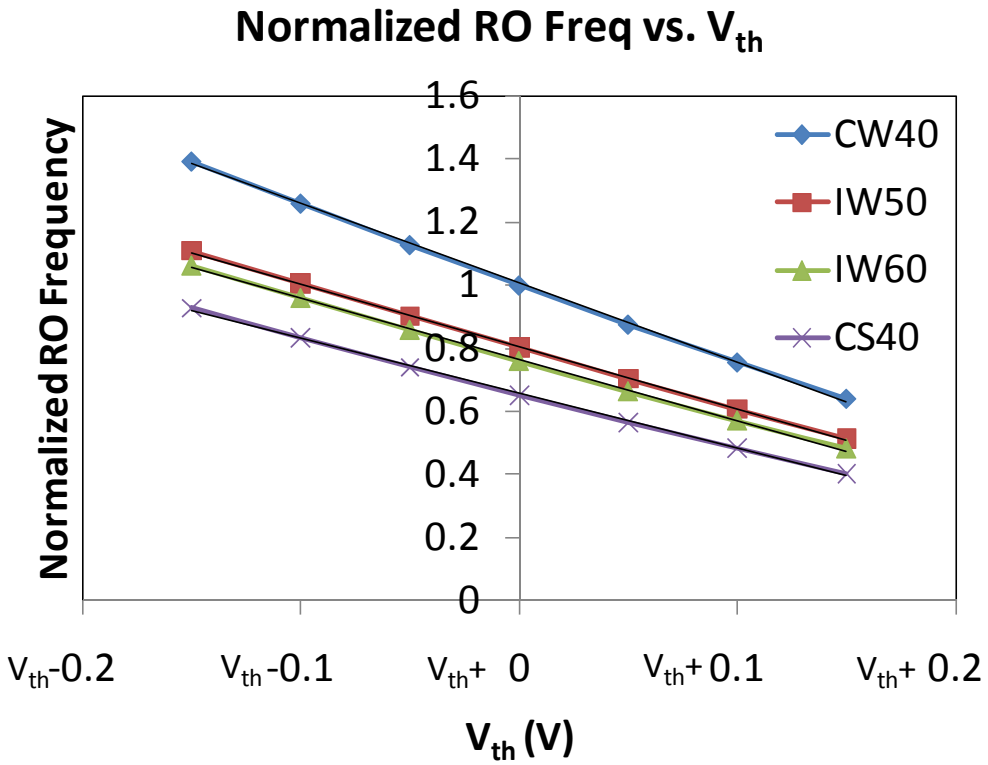


Figure 4.15: Simulated RO frequency versus threshold voltage for RO monitors CW40, IW50, IW60, and CS40 show a slope of -2.51/mV, -1.98/mV, -1.94/mV, -1.75/mV, respectively.

Using the information in Figure 4.15, the sensitivity factor, s, which describes the normalized change in RO frequency to threshold voltage, is defined and computed.

$$s = \frac{\frac{\Delta f_{RO}}{f_{RO}}}{\frac{\Delta V_{th}}{V_{th,o}}} \tag{4.1}$$

Assuming a threshold voltage change ΔV_{th} = 10mV, the ΔV_{th} to the threshold voltage $V_{th,o}$ ratio is 2.5%. The corresponding simulated change in RO frequency, Δf_{RO}, can be extracted from Figure 4.15 and is normalized by the mean of the simulated RO frequency, f_{RO}. For the RO monitors CS40, IW60, IW50, and CW40, the sensitivity factor, s, is: 1.07, 1.02, 1.01, and 0.99, respectively. The sensitivity factor in Equation (4.1) for coherent devices is then applied to scale and thus convert the measured $(\sigma/\mu)_{36}$ to the standard deviation of the threshold voltage, $\sigma_{Vth,36}$, for transistors that are potentially not correlated.

$$\sigma_{V_{th,36}} \approx \frac{\left(\frac{\sigma}{\mu}\right)_{36}}{s} \tag{4.2}$$

The second important step here is to estimate the V_{th} standard deviation for one device, $\sigma_{Vth,}$ from many devices. Two assumptions are made. The first, that the RO instantiations are independent, signifies that a single RO has a sqrt(n_{RO}) more noise. Since there are 36 RO instantiations, n_{RO} = 36. The second assumption is that, since each transistor in a RO is independent, the noise in one device is sqrt($n_{Devices}$) times larger. In this case, $n_{Devices}$ = 2 devices/RO stage x 13 RO stages = 26 devices.

$$\sigma_{V_{th}} \approx \sigma_{36 \cdot V_{th}} \cdot \sqrt{n_{RO}} \cdot \sqrt{n_{Devices}} \tag{4.3}$$

Following the above procedure, Figure 4.14 is converted into a Pelgrom plot of σ_{Vth} versus the inverse square root of the transistor channel area shown in Figure 4.16.

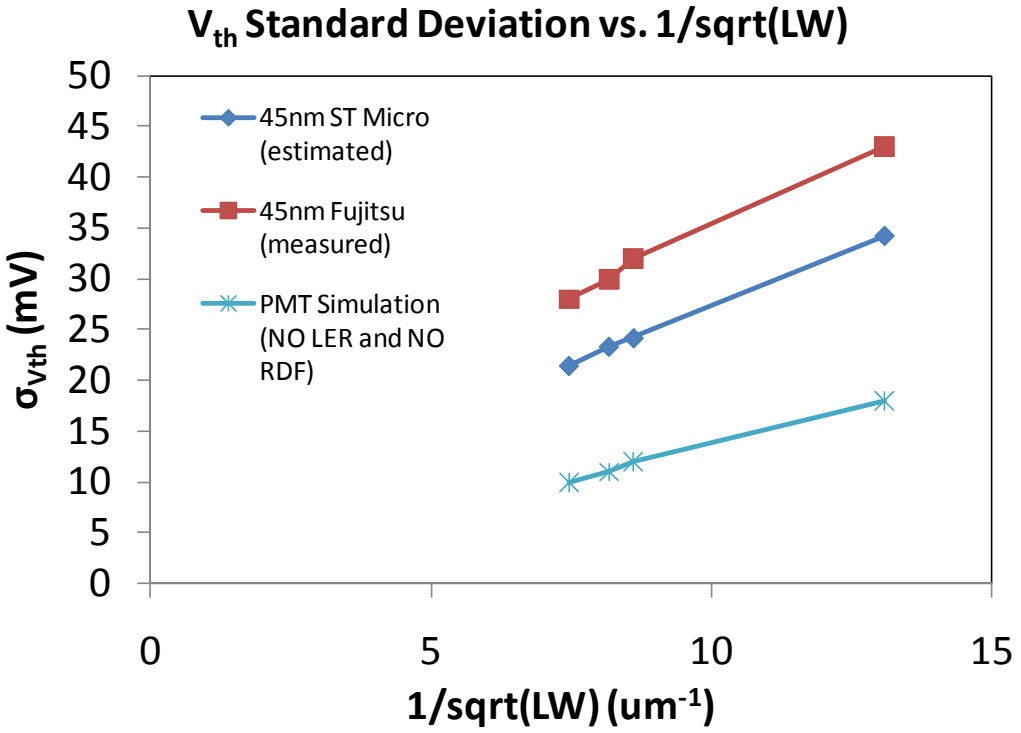


Figure 4.16: The Pelgrom plots for three 45nm technologies

For the purpose of comparison, the Pelgrom plots from Fujitsu's 45nm Technology [48] and Arizona State University's 45nm Predictive Modeling Technology (PMT) are referenced [49] in Figure 4.16. Since the Pelgrom plot for Fujitsu's Technology is based on silicon measurements, it thus reflects realistic manufacturing conditions. The Pelgrom plot for PMT is based on physical modeling and simulations; it reflects ideal manufacturing conditions and can be used to predict the amount of device noise attributed to sources of variation, such as line-edge-roughness (LER) and random dopant fluctuations (RDF). As shown in Figure 4.16, the PMT simulations contain no LER and no RDF effects. Figure 4.16 reveals that the approximations and assumptions made to observe electrical variation in a device stemming from measured noise in the RO frequency are valid and within the range set in the established literature [46]. The estimated average threshold slope, AVT, is 2.3, while the AVT extracted from the Fujitsu and the PMT references is 2.6 and 1.6, respectively.

4.5 Summary

Initial results for RO measurements are used to quantify levels of variation and to establish guidelines from which to proceed in the analysis of parameter-specific monitors and residual noise in subsequent chapters. Two main sources of measurement error are discussed: jitter and drift. Jitter from the oscilloscope is shown to be negligible with a maximum error of 0.009% error, and drift shows a maximum error of 0.16% difference in measured RO frequency. Block-to-block variation is examined and is also shown to be negligible, where the systematic shift in RO frequency among blocks is 4x smaller than the random noise that exists within-block. Thus, mean measured RO frequency across 36 RO instantiations is used to examine across-wafer and within-chip variation for control-case monitors, where measured $(\sigma/\mu)_{36}$ shows a 0.2-0.3% variation. These measured $(\sigma/\mu)_{36}$ values also show a slight channel area dependence of 0.2% versus sqrt(area) $^{-1}$. This observation then used to study the electrical device noise, shows an estimated AVT of 2.3.

Chapter 5

Experimental RO Frequency Sensitivity to Gate Etch, Gate Lithography Focus, and Gate-to-Active Overlay

Experimentally measured RO frequencies are reported herein for RO inverter layouts designed in accordance with 45nm generation logic technology design rules using ST Micro's 45nm PDK to monitor variations in gate etch, gate lithography and gate-to-active overlay. Measured RO frequency's sensitivity to process parameters is determined by comparing RO sensitivity with systematic pre-programmed variation in one or more inverter layout parameters with those of a RO control.

This diagnostics work is challenging, as the layouts are required to undergo optical proximity correction (OPC), and access to the proprietary OPC model was unavailable. Only a generic model is used to simulate the lithography process. Fortunately, in addition to the 45nm process corner simulators used here, the simulated Process Variation (PV) bands via Mentor Graphics Calibre are provided as a courtesy by ST Micro. These PV bands are used in multiple data mining methodologies to determine the physical sources that cause the differences among measured RO frequency sensitivity. They are used in conjunction with SPICE to compare simulations with measured RO frequency across-wafer spread. These PV bands generated by proprietary models are also compared with those that were generated by a generic model in order to examine the impacts of OPC, gate focus, gate etch, and gate exposure dose on RO frequency performance.

This chapter is organized into 5 sections. Section 5.1 presents how analysis from process corners and PV bands in conjunction with SPICE compares with the experimental RO frequency for 40nm, 50nm, and 60nm drawn gate lengths. Section 5.2 reports on the empirically measured RO frequency sensitivity for gate etch, while Section 5.3 does so for gate lithography focus, and Section 5.4 looks at gate-to-active misalignment. Finally, Section 5.5 summarizes this chapter.

5.1 Simulated RO Frequency Variability using Process Corners and Process Variation Bands

The measured RO frequency across-wafer variability is compared to that simulated for control case monitors at 3 different drawn gate lengths: 40nm, 50nm, and 60nm (CW40, IW50, IW60). Gate lengths drawn to be 40nm are actually fabricated to be 45nm. In keeping with the requirements of the design rules, the 40nm inverter layouts are designed with two dummy polys. The 50nm and 60nm inverter layouts are designed with isolated gates.

Two simulation methodologies are explored. The first methodology uses SPICE to simulate the RO frequency for control-case RO monitors at three different process corners: fast (FF), typical (TT), and slow (SS). The SS-TT and TT-FF guard bands each predict an across-wafer variation of 15% or a total range of 30%.

The second methodology extracts changes in gate length (ΔL) from Process Variation (PV) Band simulations using Mentor Graphics Calibre. The RO frequency versus L relationship is then obtained via SPICE simulations. Using this relationship, ΔL is mapped to change in RO frequency (Δf_{RO}). This methodology predicts an across-wafer variation range of 6-9%, which is approximately one fourth that of what is predicted by the SS-FF process corner guard bands.

5.1.1 Simulated Process Corners versus Measured RO Frequencies

Simulated RO frequencies at FF, TT, and SS corners from RO monitors CW40, IW50 and IW60 are compared to the mean measured RO frequency of 36 RO instantiations across the 17 chips in Figure 5.1. Figure 5.1 plots the normalized RO frequency versus the gate length. Here, the measured RO frequencies are normalized to the simulated RO frequency for the RO monitor with a 50 nm gate length.

The SPICE simulations using PSP/BSIM4 compact models at the TT and SS corners show that the measured variation is very small, ~1/5-1/3 of the SS-FF simulated guard bands. As may be recalled from Chapter 4, the measured RO frequency range for CW40 and IW60 RO monitors is 8.84% and 5.99%, respectively. Still, the mean measured RO frequency shows good correlation with that of the TT simulation.

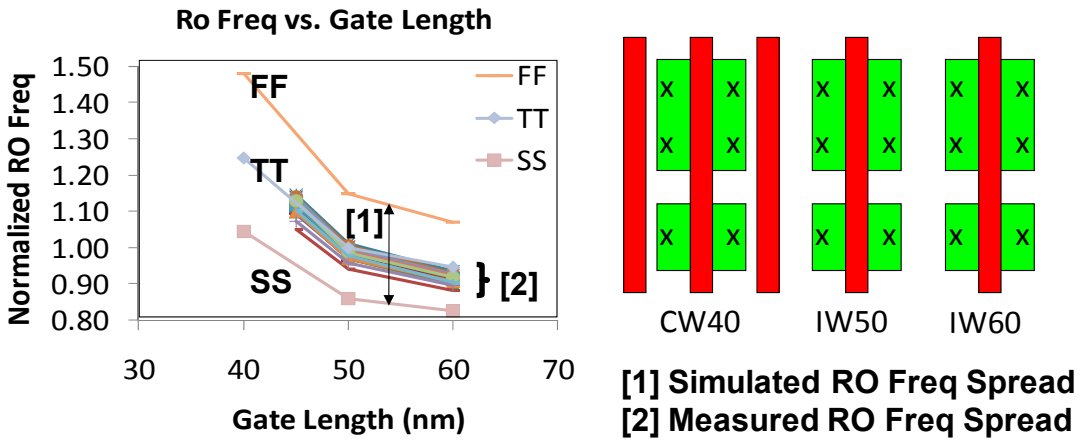
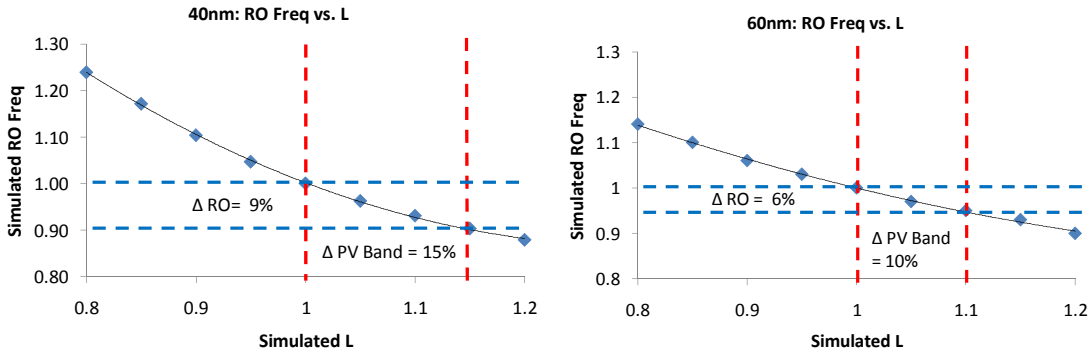


Figure 5.1: The SPICE corner simulations correlate with the mean measured RO frequency.

5.1.2 PV Bands and SPICE Simulations versus Measured RO Frequency

The simulated RO frequency variation using PV bands and SPICE for RO monitors CW40 and IW60 are shown in Figures 5.2(a) and 5.2(b), respectively. Figure 5.2 plots the simulated RO frequency versus the gate length (L). The ΔL's extracted from PV bands combined with the L sensitivity simulated in Figure 5.2 show that the Δf_{RO} ranges are 9% and 6%, respectively. These Δf_{RO} values show good correlation with the mean measured RO frequency range of 8.84% and 5.99% for CW40 and IW60 RO monitors, respectively.



(a) (b)

Figure 5.2: PV band and SPICE simulation results for CW40 (a) and IW60 (b).

Although the PV-bands-SPICE methodology shows a more accurate correlation to measured RO frequency variation, this methodology involves more steps than the process corner approach. The process corner methodology provides a simpler yet rather pessimistic approach for the simulation of measured across-wafer RO frequency variation. Both methodologies can serve to be useful under different circuit applications and design phases.

5.2 RO Frequency Sensitivity to Gate Etch

5.2.1 Measured RO Frequency Sensitivity to Gate Etch

The presence and the absence of adjacent dummy polys was used to investigate pattern dependent etch effects. A gate length of 60nm was utilized. Since OPC likely also corrects for etch, the measured RO frequency sensitivity will only reflect the post-OPC residual effects.

The measured mean RO frequency for monitors: CW60, EW60, and IW60, each with 36 RO instantiations per chip, is examined for 17 chips. It is then normalized to the measured mean RO frequency of CW60 across 17 chips. Figure 5.3 shows the normalized RO frequency versus the chip number for the three RO monitors.

The control-case monitor, CW60, serves as a reference for the following general observations. On average, removing a single dummy poly speeds up the RO frequency by about 1%. Removing both dummy polys speeds up RO frequency by 3%. For each chip, the measured range varies from as little as 3% to 5%, with the largest variation occurring for the slowest, or chip 9. For each monitor, the measured across-wafer variation range is ~6-7%.

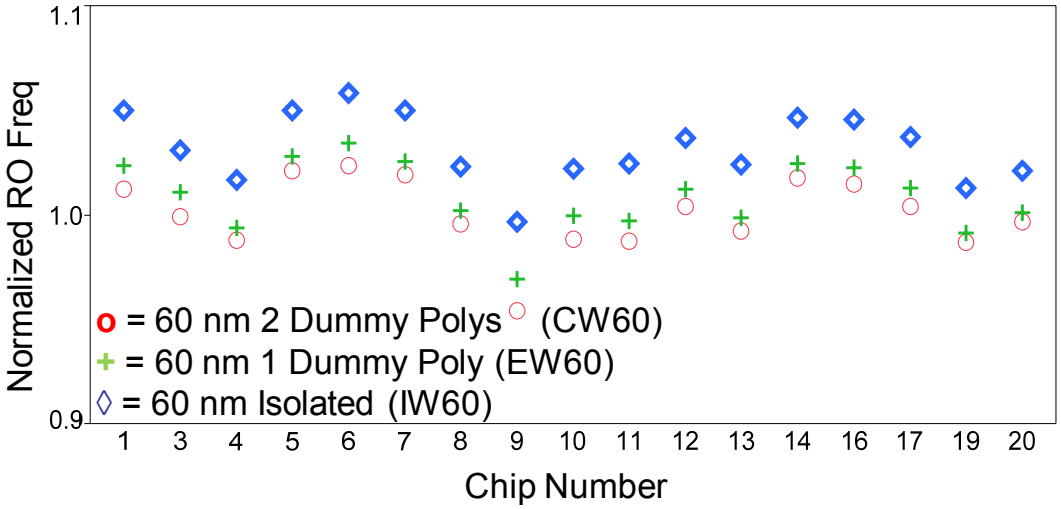


Figure 5.3: The mean measured RO frequencies reveal that the dummy polys reduce RO frequency by 3%. The across-wafer peak-to-peak variation for IW60, EW60 and CW60 are ~6-7%.

Figure 5.4 analyzes the measured noise of the RO frequency for the monitors that are presented in Figure 5.3. The measured standard deviation to mean ratio for the 36 RO instantiations, $(\sigma/\mu)_{36,}$ versus the chip number is shown. Excluding the outliers, the $(\sigma/\mu)_{36}$ for all 3 RO monitors is ~0.2%, showing that the dummy gates have minimal impact on measured noise. Outlier chips 4 and 9 have a higher $(\sigma/\mu)_{36}$ when compared to that of typical chips, and further analysis of these two chips is presented Figure 5.5

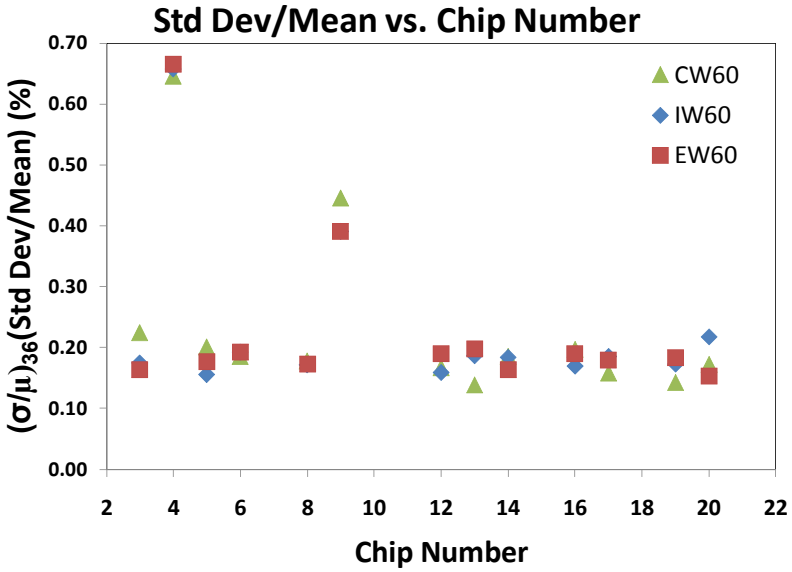


Figure 5.4: The measured $(\sigma/\mu)_{36}$ which excludes outlier chips 4 and 9 shows that the $(\sigma/\mu)_{36}$ is ~0.2%.

Figure 5.5 gives the normalized measured RO frequency versus the 36 RO instantiation for monitor CW60 plotted for Chips 4, 6, and 9. The measured RO frequency is normalized to

the mean measured RO frequency of CW60 across 3 chips. While both chips 4 and 6 operate at a faster RO frequency, chip 4 has a high noise level due to a systematic ~ 8% RO frequency decrease in RO block 3. In general, Chip 9 operates at a slower RO frequency and is also high in noise due to the 5% RO frequency slow down occurring in RO block 3. This behavior in RO block 3 is further examined in regard to etch, focus, and stress effects in Chapter 7.

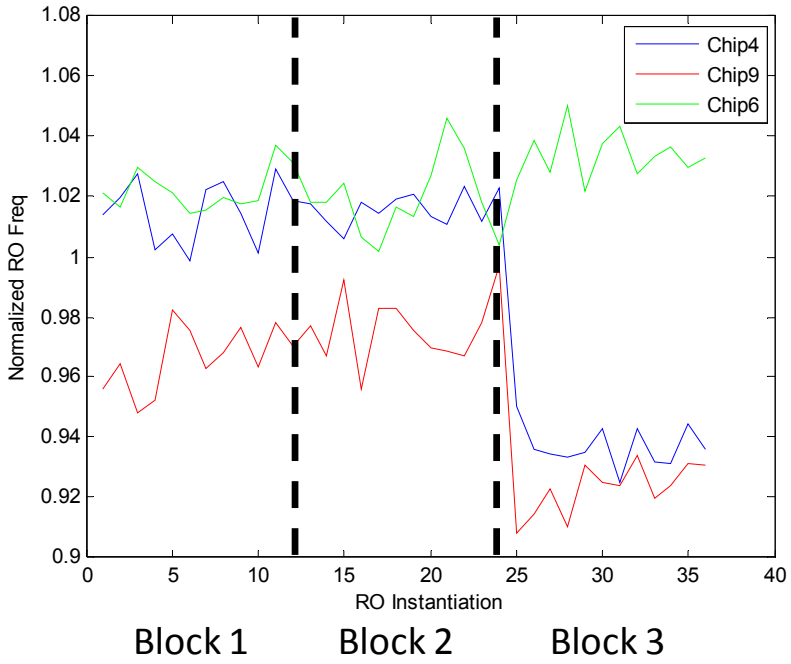


Figure 5.5: Systematic block-to-block effects are evident in the outlier chips: 6 and 9.

5.2.2 Simulated RO Frequency Sensitivity to Gate Etch using Process Corners

The behavior of the etch RO monitors is now examined using SPICE simulations and by plotting the measured RO frequency distributions. This analysis was done by extracting the netlists from the layout and simulating them in SPICE. The measured RO frequencies are normalized to the mean measured RO frequency of CW60 across 15 chips (outlier chips are removed here).

Measured RO frequency sensitivity for etch monitors is compared to that of simulated process corners at TT and SS. Figure 5.6 shows a histogram of normalized RO frequency for RO monitors CW60, EW60, and IW60, respectively. The color light blue represents the measured across-wafer distribution of normalized measured RO frequency for all 15 chips, with a total of 540 (15 chips x 36 RO instantiations) measured RO frequencies. Here the color green represents the measured RO frequency for a slow chip (Chip 19) while the color dark blue represents that of a fast chip (Chip 6). The red spikes represent SPICE simulated RO frequency at different process corners.

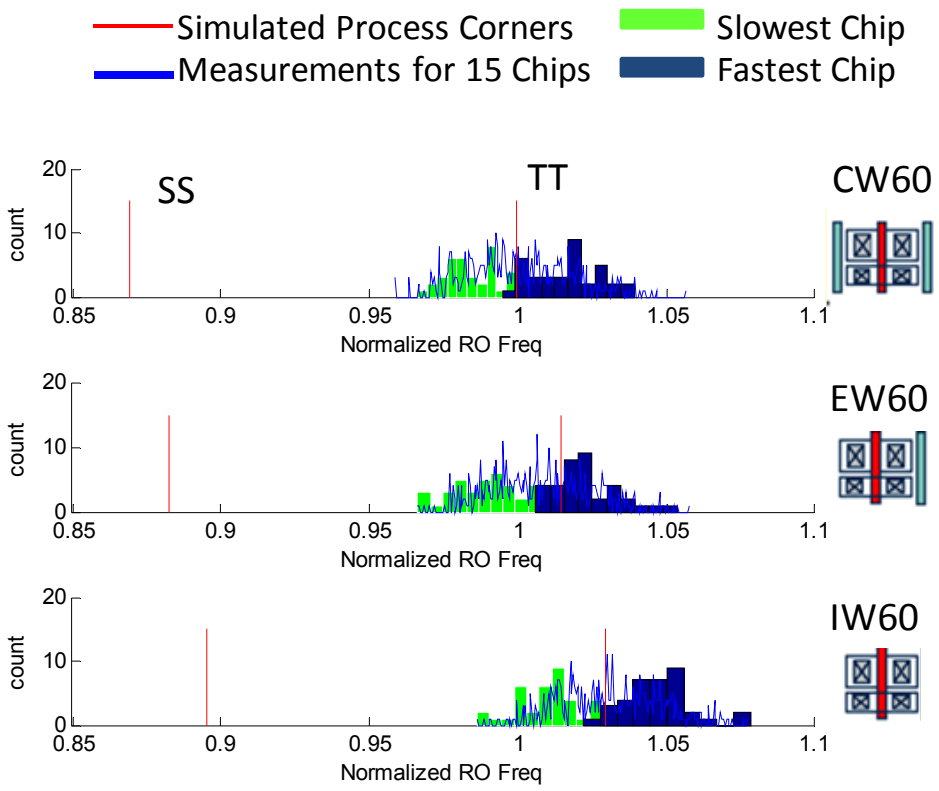


Figure 5.6: A comparison of measured and process-corner simulated RO frequency for CW60, EW60, and IW60, respectively.

Figure 5.6 shows a simulated SS-TT spread of 15%. For each RO monitor, the across-wafer measured spread, $3(\sigma/\mu)_{540}$, for all 15 chips is only ~6%. The measured across-wafer variation is only ~1/3 of the simulated SS-TT and ~1/6 of the SS-FF guard band. The within-chip variation, $3(\sigma/\mu)$, for both the slow chip (Chip 19) and the fast chip (Chip 6) is 3.3%-3.5%. Thus, the distribution here is about 2x bigger across-wafer than within-chip.

Figure 5.6 illustrates that the removal of one dummy poly results in a RO frequency shift in mean measured RO frequency, $\Delta\mu_{540}$, of 0.82%, while simulations at the TT corner predict a 1% shift. Removing two dummy polys results in a measured $\Delta\mu_{576}$=3.13%, while simulations at TT show a $\Delta\mu$= 2%. It was necessary to correct for the increase in the stray capacitance between the drain and the dummy. With this capacitive correction, the decrease in RO frequency attributable to etch was 1.2%.

In recalling the measured RO frequency observations from the 90nm node in Chapter 2 one can see that, as compared to Pang's 90nm results, the 45 nm process has less measured systematic across-wafer variation. The across-wafer variation improved from 15% to about 7%, while the typical within-chip variation remained at ~3%. Among the 3 etch monitors, the measured RO frequency variation improved from 10% at 90nm to 3% at 45nm. Table 5.1 summarizes the measured RO frequency variation for the two technology nodes.

	90nm [1]	45nm (L_{drawn} = 60nm)
Layout-Dependent Shift (Δf_{RO})	10%	3%
Across-Wafer (3σ/μ)	15%	7%
Within-Chip (3σ/μ)	3.1%	3.5%

Table 5.1: A comparison of etch monitors for two technology nodes

5.2.3 Simulated RO Frequency Sensitivity to Gate Etch Monitors using PV Bands and SPICE

Using the methodology described in Section 5.1.2, the PV bands generated by ST Micro and SPICE simulations are used in order to simulate the across-wafer RO frequency variation of the etch RO monitors. Table 5.2 compares the simulated RO frequency with that of the measured and shows good agreement.

Monitor	Simulated ΔL from PV (*Generic models)	Simulated Δf_{RO} from PV	Measured Across-Wafer Peak-to-Peak Variation (Δf_{RO}) (*No outlier chips)
IW60	11%	6%	6%
EW60	11%	6%	6%
CW60	11%	6%	6%

Table 5.2: The comparison of simulated and measured across-wafer variation shows good agreement.

As may be recalled from Figure 3.14 in chapter 3, as compared to that of dense-gate monitors, ΔL is 2x more sensitive to focus when no dummy gates are present. However, this change is not seen in the PV bands provided by ST Micro (Table 5.2). This lack of change indicates that ST Micro's PV bands include OPC, i.e. assist features, such as SRAFs. OPC has been used to control ΔL induced by lithography focus for RO monitors IW60 and EW60. Moreover, since minimal changes in ΔL are observed in the PV bands for the 3 etch monitors, it can be concluded that the measured systematic RO frequency shift of 1.2% among layout is likely due to the remaining etch loading and parasitic effects.

5.3 RO Frequency Sensitivity to Gate Lithography Focus

The intent of this design is to enhance the measured RO frequency sensitivity at the center of the gate by adding additional poly features to the gate and its surroundings in order to modify the focus spillover to the center of the gate. The gate lithography focus RO monitors are designed with 40 nm and 60nm drawn gate lengths. As may be recalled from Chapter 3, the gate patterns have been decorated with hammerheads and dummy polys placed at specific locations in order to exacerbate the focus effects by a factor of 1.5. These monitor designs were created in line with the suggested action made by Pattern Matching and refined by aerial image simulations using Mentor Graphics Calibre. Unfortunately, the experimental RO frequency sensitivity versus preprogrammed defocus treatments is not available. Consequently, only the measured RO frequency variation in across-wafer performance is available to explore the possible focus

effects. For example, the RO frequency sensitivity from the gate lithography focus monitors is compared to that of control ROs. However, this indirect analysis of focus-enhanced sensitivity is faced with many challenges. Section 5.1 reveals that the wafer is fabricated under typical conditions with well-controlled variability, indicating little process-induced variation. Section 5.2 shows that the simulated changes in gate length for etch monitors are similar, indicating strong OPC with scatter-bars was applied and thus decreasing the sensitivity of the design.

5.3.1 Measured RO Frequency Sensitivity to Gate Lithography Focus

Figure 5.7 shows the mean measured RO frequency normalized to the mean RO frequency of FS40 across all the 17 chips versus the chip number for RO monitors CS40 and FS40. The hammerheads slow down the RO frequency by 4 %. This decrease is caused by the horizontal polys at the top and bottom edges of the gate contributing parasitic effects. The measured across-wafer differences for CS40 and FS40 varies by < 1%, when outlier chips 4 and 9 are removed.

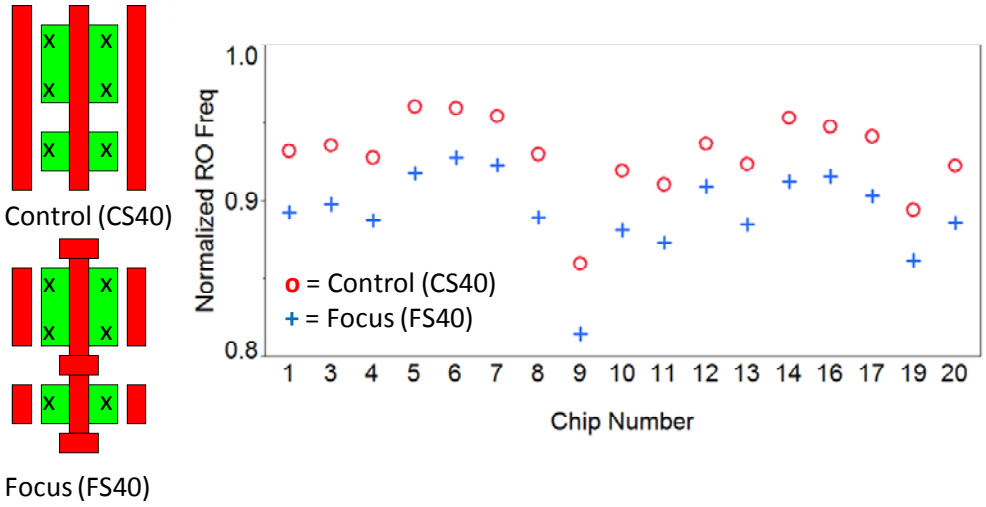


Figure 5.7: The mean measured RO frequency for the 40nm drawn gate length shows that the hammerheads result in about a 4% lower RO frequency. The across-wafer peak-to-peak variation for the control (CS40) and focus monitor (FS40) vary by <1%, when outlier chips 4 and 9 are removed.

Figure 5.8 shows the mean measured RO frequency normalized to the mean RO frequency of FW60 across the 17 chips versus the chip number for RO monitors IW60 and FW60. On average, the hammerheads slow down the RO frequency by 4%. The measured across-wafer peak-to-peak range for IW60 and FW60 varies about <1%, when outlier chips 4 and 9 are excluded.

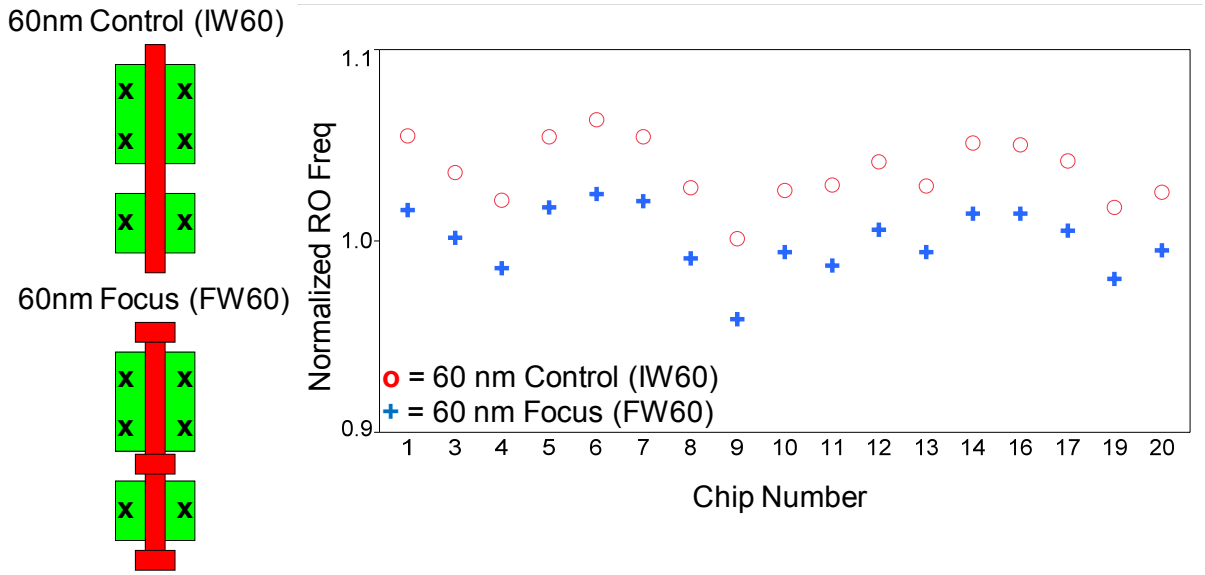


Figure 5.8: The mean measured RO frequency for 60nm drawn gate length shows that the hammerheads result in a 4% lower RO frequency. The across-wafer peak-to-peak variation for the control (IW60) and focus monitor (FW60) varies by <1%.

The noise in the measured RO frequency in Figures 5.7 and 5.8 is examined in Figure 5.9. Figure 5.9 reveals that for typical chips, the measured $(\sigma/\mu)_{36} = 0.3\%$ for the small width 40nm monitors, and $(\sigma/\mu)_{36} = 0.2\%$ for the large width 60nm monitors. The former is larger and can be attributed to smaller gate area. The focus monitors, FS40 and FW60, show higher measured noise than in the corresponding control monitors. However, the limited wafer tilt in a stepper that induces a focus variation should not have a big effect on ROs that are spaced 1-2 mm apart. Thus, this measured noise increase in the focus monitors is more likely due to the nature of the monitor itself rather than a focus effect.

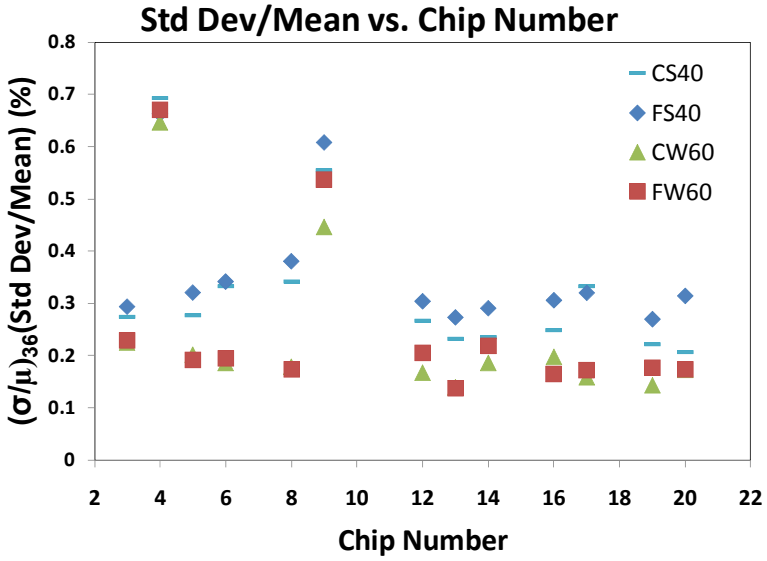


Figure 5.9: The measured $(\sigma/\mu)_{36}$ is larger for RO monitors with a smaller gate area. Additional hammerheads have negligible impact on $(\sigma/\mu)_{36}$.

5.3.2 Simulated RO Frequency Sensitivity to Gate Lithography Focus using Process Corners

The behavior of the focus monitors is now examined by plotting the measured distribution for 17 Chips, including the outlier chips. The measured RO frequency sensitivity of CS40 versus FS40 and IW60 versus FW60 is compared to that of the simulated at the SS and TT process corners in Figures 5.10 and 5.11. For each monitor, 612 RO frequency measurements (i.e. 17 Chips x 36 RO instantiations) are plotted. Please refer to Section 5.2.2 for an in depth explanation of the plots' legend.

The data in Figure 5.10 shows the measured RO frequency distributions for CS40 and FS40 normalized to the mean measured RO frequency of FS40 across all 17 chips. The simulated range between SS and TT guard-bands is 20%, while the measured across-wafer variation, $3(\sigma/\mu)_{612}$, is ~10% for CS40 and ~12% for FS40. If the outlier chips 4 and 9 are removed, the $3(\sigma/\mu)_{540}$ decreased to ~10%. Thus, the measured across-wafer variation is only ~1/2 of the simulated SS-TT and just ~1/4 of the total SS-FF guard band. For typical chips, the measured with-in chip variation, $3(\sigma/\mu)$, is ~6% (Chip 6). In regard to the outlier chips, the measured within-chip variation increases to 10% (Chip 4 and Chip 9). The plot shows a maximum layout-dependent shift in the mean measured RO frequency, $\Delta\mu_{612,} = 3.81\%$, whereas simulations extracted at the TT corner predict a $\Delta\mu_{612} = 4.60\%$. The systematic reduction in measured RO frequency in FS40 compared to that of CS40 is mainly due to additional gate capacitance and somewhat dependent on the focus-induced hourglass gate shape.

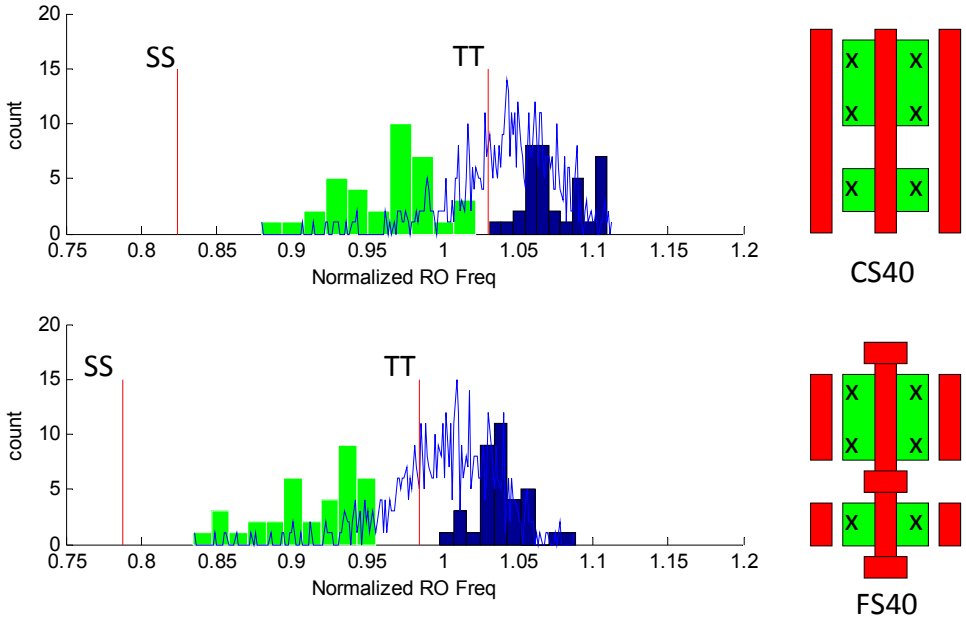


Figure 5.10: The measured RO frequency for CS40 and FS40 across 17 chips compared with the simulated RO frequency reveals that the change in RO frequency between layouts is due to mainly parasitic effects from the gate.

Figure 5.11 shows the SPICE simulations and measured RO frequency distributions for IW60 and FW60 normalized to the mean measured RO frequency of FW60 across the 17 chips.

The simulated range between the SS-TT guard-bands is 15%, while the measured across-wafer variation, $3(\sigma/\mu)_{612}$, is just ~6% for IW60 and ~7% for FW60. If the outlier chips 4 and 9 are removed, the $3(\sigma/\mu)_{540}$ decreased to ~6%. Thus, the measured across-wafer variation is only ~1/4 of the simulated SS-FF guard band. For typical chips, the measured within-chip variation, $3(\sigma/\mu)$, is ~3%. For outlier chips, the measured within-chip variation goes up to 6%. The plots show a maximum shift in mean measured RO frequency, $\Delta\mu_{612}$, of 3.55% between layouts, whereas the simulations extracted at the TT corner predict a Δf_{RO} =5.77%. Again, the systematic reduction in measured RO frequency in FW60 compared to that of IW60 is likely due to additional parasitic effects and somewhat affected by the change in the gate length due to a focus-induced hourglass gate shape effect.

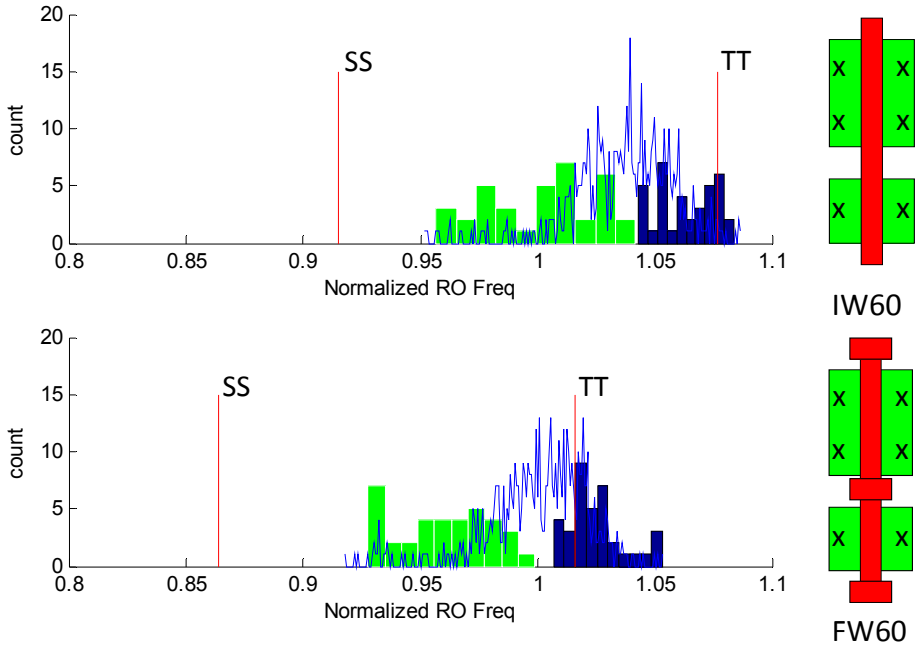


Figure 5.11: The measured RO frequency for IW60 and FW60 across the 17 chips compared with the simulated RO frequency shows that the change in RO frequency between layouts is due to parasitic effects.

5.3.3 Simulated RO Frequency Sensitivity to Gate Lithography Focus using PV Bands and SPICE

For purposes of comparison, the PV bands simulated using the generic lithography model and ST Micro's 45nm PDK SPICE simulations are used to simulate across-wafer RO frequency variation. Table 5.3 compares the simulated with the experimental RO frequency.

Monitor	Simulated ΔL from PV (*Generic model)	Simulated Δf_{RO} from PV	Measured Across-Wafer Peak-to-Peak Variation (Δf_{RO}) (*No outlier chips)	Simulated ΔL from PV (*Proprietary model)
CS40	15%	10%	10%	10%
FS40	23%	15%	10%	11%
IW60	11%	6%	6%	8%
FW60	14%	8%	6%	9%

Table 5.3: A comparison of simulated versus measured across-wafer variation.

Measured across-wafer variation shows negligible differences among the focus monitors and their respective controls. The production wafer may have had a tighter control on focus than the $\pm$2/3 RU assumed in our generic models. If the fabrication conditions were $\pm$0.35 RU, the sensitivity would be 4x lower. In addition, the decision to OPC the monitors likely decreased their sensitivity to focus. This was verified after the fact by examining the PV bands from ST Micro which were generated with proprietary OPC and process models. Table 5.3 shows that the PV band change of FW60 relative to that of IW60 is negligible. The PV bands from ST Micro also indicate that there exist little defocus in the stepper system, and the extracted differences are due to variation in exposure dose.

5.3.4 Across-Wafer Correlation Plots

As was seen earlier, changes in the focus monitors tended to track with changes in the control. To see how well these effects tracked, correlation plots of normalized variations were examined. The correlation of measured RO frequency for FS40, FW60, IW60, CW60, $NW40_2$, and $NW40_4$ is examined across 15 chips, without outlier chips. The mean measured RO frequency for FS40 versus FW60, FW60 versus IW60, CW60 versus IW60, and $NW40_2$ versus $NW40_4$ is plotted in Figures 5.12 (a), (b), (c), and (d), respectively. For each monitor, the measured mean is normalized to the monitor's mean across 15 chips.

The data for FS40 versus FW60 in Figure 5.12 (a) has a slope of 1.5, indicating that the monitor with a 40nm gate length is 1.5x more sensitive than that of a 60nm to variability. An r-squared value of 0.93 is observed. This value indicates a good correlation between the two focus monitors. This shows that they are measuring similar process effects.

The data for FW60 versus IW60 has a slope of 1.0, indicating that FW60 and IW60 monitors are equally sensitive to process variability. An r-squared value of 0.96 is observed. Again, the good correlation between the focus monitor and the isolated 60nm gate indicates that they are measuring similar process effects.

The data for CW60 versus IW60 shows a slope of 0.96. Since the r-squared value reduced to 0.87, the cause of this decrease in correlation is examined. One hypothesis is that the dense monitor CW60 could possibly experience a different capacitance load from the adjacent dummy poly due to due to gate-to-active misalignment. To test it, a correlation plot 5.12(d) of $NW40_2$ versus $NW40_4$ is made. The drain of $NW40_4$ is located 200nm away from its dummy poly, while for $NW40_2$, this distance is decreased to 30nm. The correlation between the two monitors with different dummy poly-to-drain distances shows a high correlation of r-square

value of 0.95 which likely eliminates the possibility of confounding focus effects with misalignment effects.

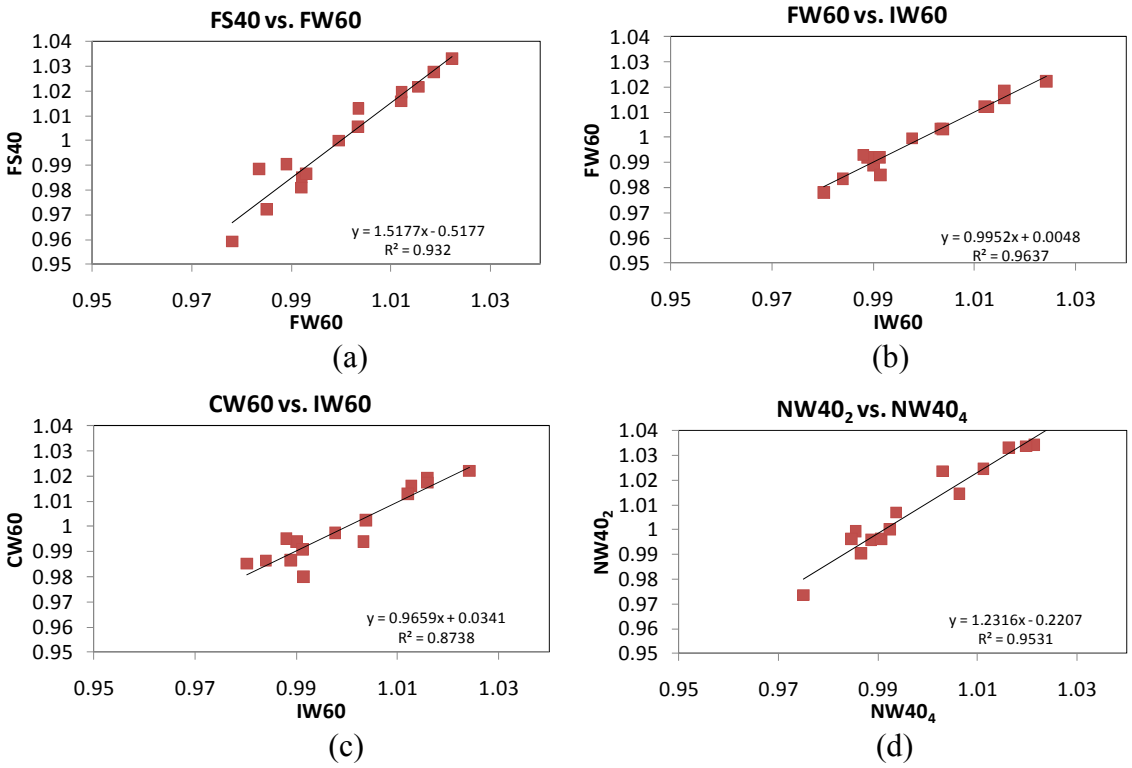


Figure 5.12: Correlation plots for FS40 versus FW60 (a), FW60 versus IW60 (b), CW60 versus IW60 (c), and $NW40_2$ versus $NW40_4$ (d) are presented.

5.4 RO Frequency Sensitivity to Gate-to-Active Misalignment

As discussed in Chapter 3, the gate-to-active misalignment monitors here were designed based on the lithographic rounding of an H-shaped active area. The measured RO frequency for five RO monitor designs with 40nm drawn gate length and drawn gate-to-diffusion offsets of 0 nm, $\pm$10nm and $\pm$15nm are reported here for the H-shaped RO monitors. The measured RO frequency for the H-shaped monitor is presented in Figure 5.13, where each data point is computed by taking the measured mean from the 36 RO frequency instantiations, where the measured $(\sigma/\mu)_{36} = 0.3\%$.

A parabolic dependence on the drawn offset is observed here. The location of the minimum quantifies the overlay error, and the highest curvature indicates more defocus. The RO frequency dip for best overlay is ~1.5%, or a roughly a 2x decrease in the RO frequency sensitivity that was expected from that of the simulated discussed in Chapter 3. The inclusion of OPC in the fabrication reduced the sensitivity of the RO monitors. The PV bands show that OPC controls the width of the device around the aligned position. Nevertheless, the measured sensitivity is still significant when compared to measured $(\sigma/\mu)_{36} = 0.3\%$ for each data point on the plot.

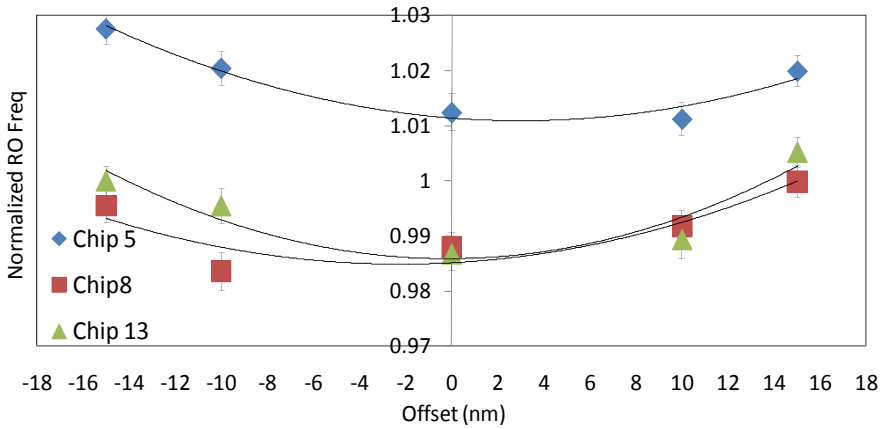


Figure 5.13: These H-Shaped misalignment RO monitor measurement results for chips 5, 8 and 13, show a 1.5% change in RO frequency.

The deduced overlay error versus chip number is shown in Figure 5.14. Here, a double hump trend that is similar to that seen in the plot of the mean measured RO frequency versus the previously shown chip number (Figure 4.12) for control-case RO monitors.

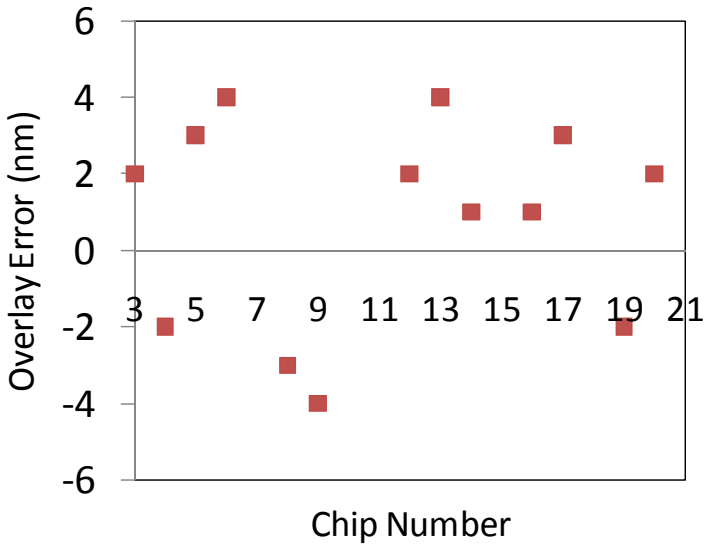


Figure 5.14: The deduced overlay error is 2-4nm across the wafer.

5.5 Summary

Adjustments in inverter layout design parameters can enhance the sensitivity of RO frequency to variation in gate etch, gate focus, and gate-to-active misalignment to be well above the level of variation due to random sources. The measured across-wafer standard deviation to mean ratio for the 36 RO instantiations, $(\sigma/\mu)_{36,}$ is ~0.2% to 0.3%. The etch monitors show a 3% RO frequency reduction with the removal of dummy polys and are well predicted by PV band and SPICE simulations. As both PDK simulations and PV band contours show negligible changes with or without the presence of dummy polys, this decrease is due to residual parasitic and etch loading effects. Although the gate lithography focus monitors here are designed with

1.5x focus sensitivity compared that of the control RO in simulations, the measured RO frequency shows a decrease in sensitivity. This discrepancy is due to the fact that the measured across-wafer variation is only 1/6-1/4th of the SPICE-simulated SS-FF guard bands, indicating a well-controlled process. In addition, ST Micro's PV bands show very little focus variation. Thus, without pre-programmed focus treatment, full calibration of the focus monitors is not possible, since the level of defocus may be too low to be measured. Given the quality of the wafer, the lack of focus treatments, the OPC of the focus monitor, and little focus sensitivity in the resulting post-OPC PV band, there was no direct way to assess the sensitivity of the focus monitor. The chip-to-chip variation among focus and control monitors were examined. A typical correlation with r-squared value of > 0.9 was obtained, indicating that both monitors are measuring similar process effects. Furthermore, the slope of 1 reveals that the focus and control monitors are equally sensitive to process effects. The H-shaped gate-to-active misalignment monitors show the expected parabolic shape. A 2-4 nm overlay error is measured and has a double humped behavior similar to that of the measured mean RO frequency versus the chip number for the control-case ROs shown in Figure 4.12.

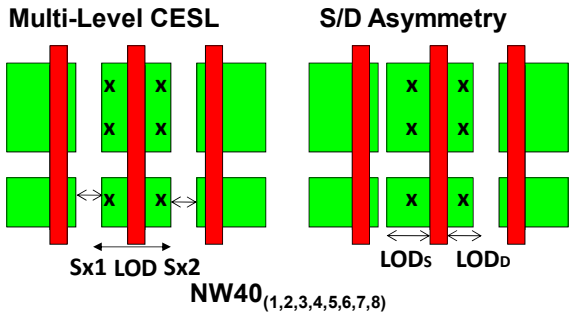
Chapter 6
RO Frequency Sensitivity to Nitride CESL- and STI-Induced Stress

Experimentally measured RO frequencies are reported for RO inverter layouts that are designed to monitor circuit performance variations caused by variability in the Nitride Contact Etch Stop Liner (CESL) and Shallow Trench Isolation (STI). For this particular process, strong uniaxial tensile strain is created by using the nitride CESL layer to increase NMOS mobility [46]. In addition, weak tensile stress is created using subatmospheric chemical vapor deposition oxide (SACVD) for STI [46]. Layouts were designed in accordance with 45nm generation logic technology design rules (using a 45nm PDK) and were measured on pre-production silicon from ST Micro. Measured RO frequency sensitivity to process parameters is determined by comparing the sensitivities for RO with systematic pre-programmed variation in one or more inverter layout parameters with the control RO. In order to determine the complex interactions of strain/stress, device mobility, and injection velocity, measured RO frequency is compared in a collaborative manner to that of the simulated RO frequency generated from stress/strain contours using Sentaurus [7].

This chapter is organized into 3 sections. Section 6.1 analyzes the RO frequency sensitivity to Nitride CESL strain for monitors with symmetrical and asymmetrical source/drain areas. Section 6.2 reports on the RO frequency sensitivity for STI monitors. Finally, Section 6.3 summarizes this chapter.

6.1 RO Frequency Sensitivity to Nitride CESL-Induced Strain

Nitride CESL monitors, NW40 $_{(1,2,3,4,5,6,7,8)}$, are used to examine the impact of the lengths of diffusions (LOD), asymmetrical source/drains areas, and lateral Shallow Trench Isolation (STI) on measured RO frequency sensitivity. The monitors are designed with 40nm drawn gate lengths and 8 various LOD and lateral STI separations. The LOD dimensions are given in Table 6.1, in which the layout dimensions are normalized by approximately 3 drawn gate lengths. Here, LOD_S and LOD_D are defined as the LOD of the source and the drain of the device, respectively. Here, LOD' is defined as the sum of normalized LOD_S, LOD_D, and the drawn gate length. *Sx1* is defined as the STI width between LOD_S and its adjacent dummy diffusion. Similarly, *Sx2* is defined as the STI width between the LOD_D and its adjacent dummy diffusion.



	LOD_S	LOD_D	Sx1	Sx2
$NW40_1$	6.64	6.64	0.00	0.00
$NW40_2$	2.64	2.64	1.32	1.32
$NW40_3$	1.82	1.82	2.14	2.14
$NW40_4$	1.00	1.00	2.95	2.95
$NW40_5$	1.73	1.00	1.09	1.82
$NW40_6$	1.73	1.73	1.09	1.09
$NW40_7$	1.00	1.00	1.82	1.82
$NW40_8$	1.00	1.73	1.82	1.09

Table 6.1: The normalized layout dimensions for nitride CESL-induced strain RO monitors

Of these 8 CESL-induced strain monitors, two monitors, $NW40_6$ and $NW40_7$, are repeated from a set of screening layouts that were previously designed by L.T. Pang for a 45nm ST Micro tapeout. These RO monitors provide a direct comparison between process maturities at two different stages during the 45nm development cycle. Two RO monitors, $NW40_5$ and $NW40_8$, are designed with asymmetrical source/drains to examine the impact of non-uniformity in strain distribution on device performance. Four monitors, $NW40_1$, $NW40_2$, $NW40_3$, and $NW40_4$, are designed so as to characterize RO frequency performance versus that of LOD.

Since modifying the LOD causes a change in the drain-substrate capacitance, for an accurate analysis of the impact of strain-induced effects on measured RO frequency to occur, it is necessary to compensate for parasitic effects by using the 45nm PDK simulations to extract and approximate their impact on RO frequency performance. The simulations show that such parasitic effects may change RO frequency by 2-15%, depending on the LOD size. For this work, in order to correct for layout-dependent parasitic components, the mean measured RO frequency is normalized to a SPICE-simulated RO frequency by dividing the measured with that of the simulated.

6.1.1RO Frequency Sensitivity versus LOD'

The dependence of mean measured RO frequency on LOD' after normalization to SPICE-simulated RO frequency for symmetrical CESL strain-induced monitors is examined here. As may be recalled, LOD' is defined as the sum of the normalized LOD_S, LOD_D, and the drawn gate length. Figure 6.1 plots the normalized mean measured RO frequency for 6 levels of LOD' for 17 chips.

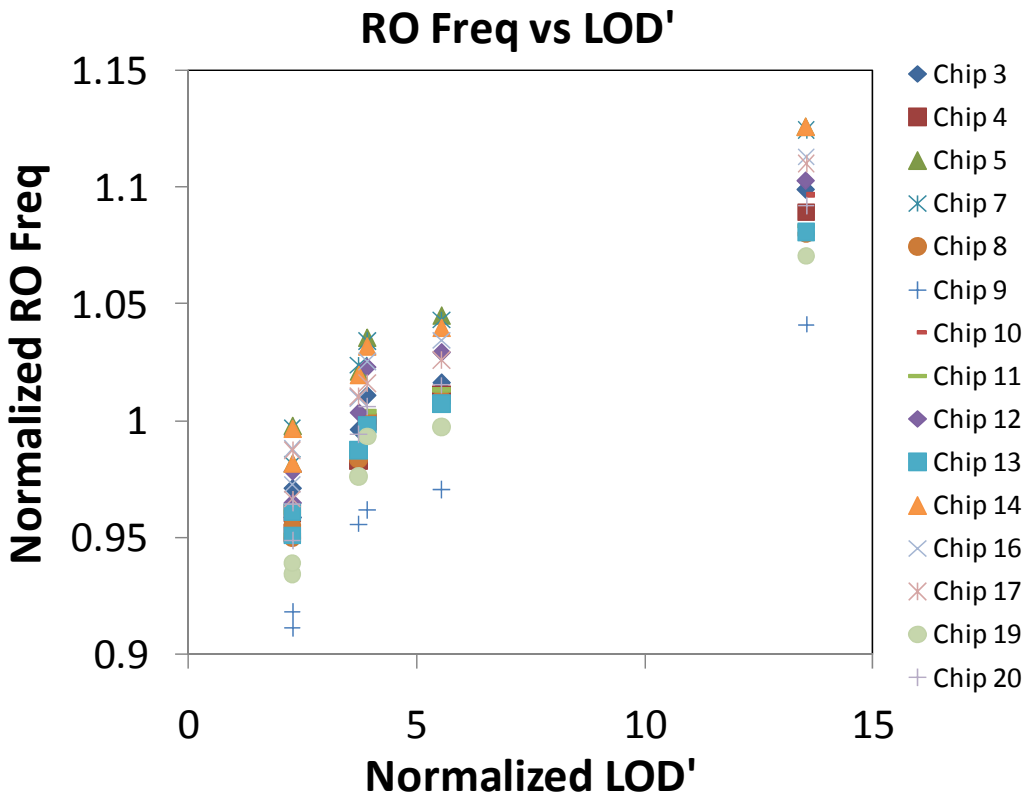


Figure 6.1: The mean measured RO frequency versus LOD' for 17 chips (excluding asymmetrical monitors)

Figure 6.1 shows a vertical spread of ~10% RO frequency variation. However, the layout-to-layout differences systematically show a measured RO frequency increase of 5% when the LOD' is approximately doubled from the minimum value of 2.27 to 3.90, and the enhancement can be as high as 15% increase for a very long LOD.

The measured $(\sigma/\mu)_{36}$ values for the RO frequencies shown in Figure 6.1 are plotted in Figure 6.2. The measured within-chip variation for all 6 types of nitride CESL RO monitor is again $(\sigma/\mu)_{36} = 0.2\%$, showing that the value of the LOD' has a negligible impact on the level of random noise. The noise level for these wide gate designs is similar to that reported in Chapter 4.

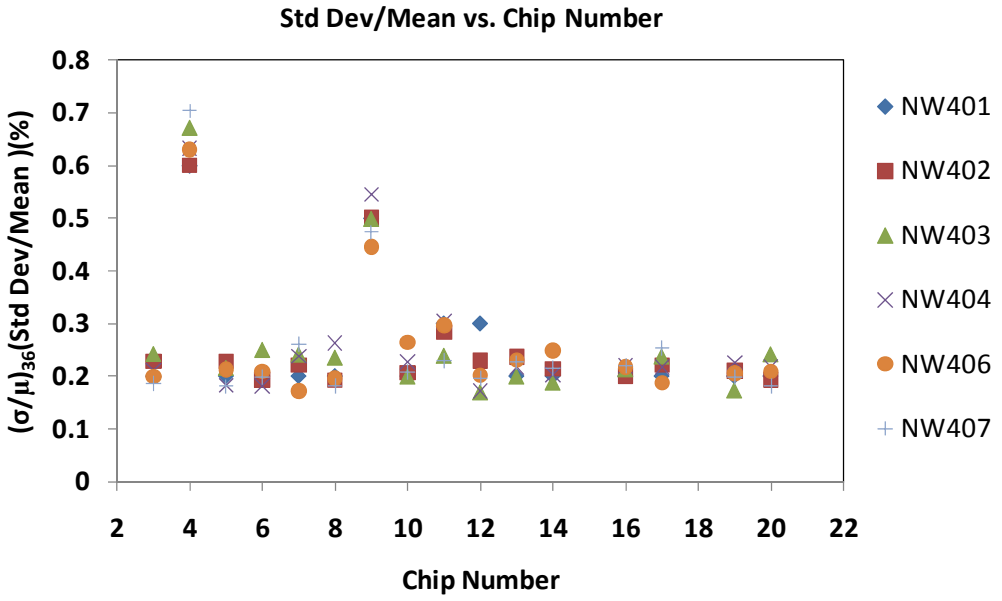


Figure 6.2: The measured $(\sigma/\mu)_{36}$ for NW40 $_{(1,2,3,4,6,7)}$ is 0.2%, excluding the outlier chips.

6.1.2 Measured RO Frequency Distributions for Two Different 45nm Tapeouts

The measured RO frequencies from CESL-induced strain RO monitors, $NW40_6$ and $NW40_7$, are compared to that of a previous tapeout and production run in order to examine the 45nm production characteristics from a development cycle. The CESL-induced stress monitors designed by Pang were included in a tapeout during the early development cycle of the 45nm ST Micro process [1]. In this work, Pang's monitors were replicated and included in a tapeout in which the process has been perfected to almost production-grade. Table 6.2 summarizes the measured RO frequency sensitivity and distributions.

	45nm [1]	**45nm**
Layout-Dependent Shift (Δf_{RO})	5%	5%
Across-Wafer for $NW40_6$ ($3\sigma/\mu$)	15%	8%
Within-Chip for $NW40_6$ ($3\sigma/\mu$)	6.6%	3.1%

Table 6.2: A comparison of RO monitors repeated at the 45nm processes for two tapeouts

Table 6.2 shows that the 5% CESL-induced RO frequency shift with the increase in LOD is still present. However, the measured across-wafer and typical within-chip variation for monitor $NW40_6$ decreased by about 50%. These reductions show significant improvements in controlling the impact of process variation, such as lithography and etch, in the mature 45nm process.

6.1.3 RO Frequency Sensitivity Averaged Across 15 Typical Chips versus LOD'

To further improve the quantitative estimate of RO frequency versus LOD behavior, the mean of the measured RO frequency, which is shown in Figure 6.1, is computed across 15 chips, excluding outlier chips 4 and 9. Figure 6.3 shows that the layout-to-layout differences systematically cause a RO frequency increase of 5.3% when the LOD' is approximately doubled from the minimum value of 2.27 to 3.90. The enhancement can be as high as a 13.9% RO frequency increase for a quite long LOD. RO monitor $NW40_4$, which has the same LOD'=2.72 as that of $NW40_7$ but with 200nm more lateral STI width, operates 1.3% slower than $NW40_7$.

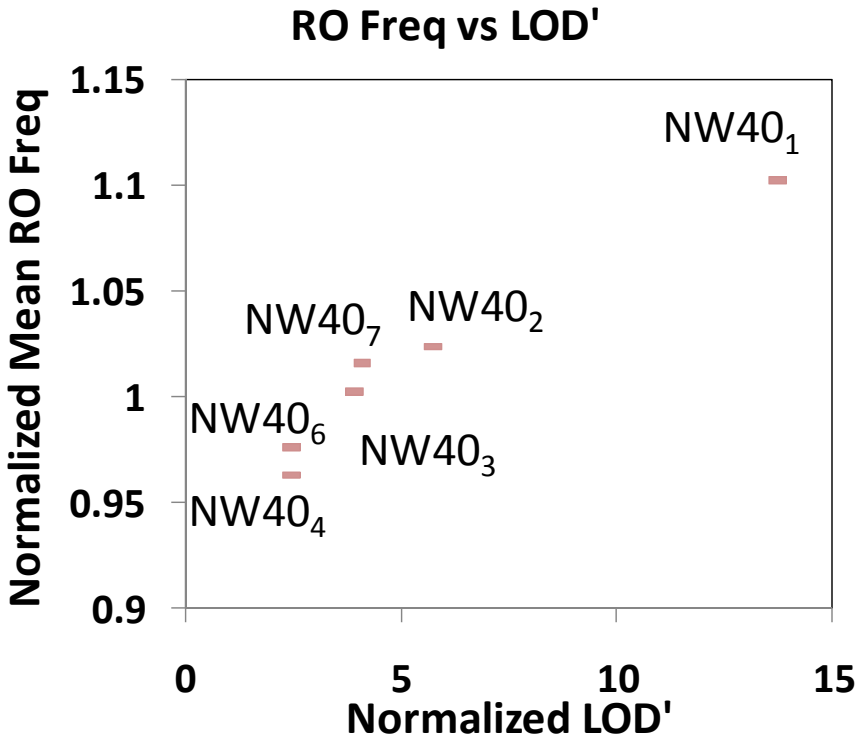


Figure 6.3: The mean measured RO frequency of 15 typical chips versus LODs (excluding outlier Chips 4 and 9)

6.1.4 RO Frequency Sensitivity to Asymmetrical Source/Drain

The RO monitors with asymmetrical source/drain $NW40_5$ and $NW40_8$ are now examined. Figure 6.4 shows the mean measured RO frequency normalized to SPICE simulations for the 17 different chips. The LOD_S or LOD_D are biased by an additional ~100nm of the LOD, presumably producing more CESL-induced stress but less STI on that side.

For the same LOD', asymmetrical source/drain RO monitors systematically show a 3% greater RO frequency speed-up for a longer LOD on the source side versus on the drain side. Considering that the noise level of the measurement is $(\sigma/\mu)_{36} = 0.2\%$, these observed effects are significant.

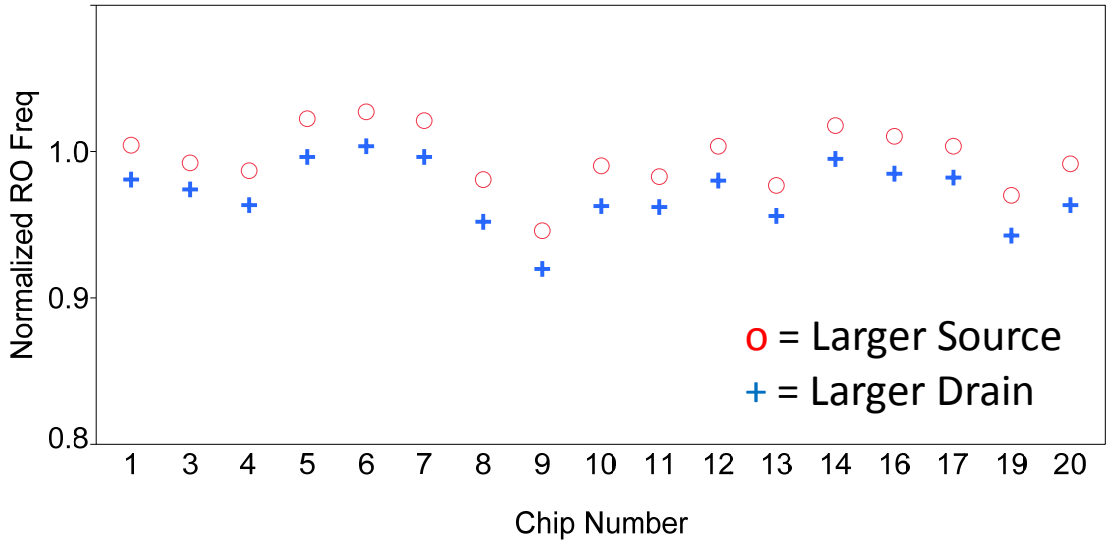


Figure 6.4: For the same LOD', asymmetrical monitors with greater sources operate 3% faster than monitors with bigger drains.

The asymmetrical monitors $NW40_5$ and $NW40_8$ are now compared to that of the control-case monitors, $NW40_6$ and $NW40_7$, respectively. Figure 6.5 plots the normalized mean measured RO frequency for RO monitors: $NW40_4$, $NW40_5$, $NW40_6$, and $NW40_7$ versus LOD_S across 15 chips (excluding outlier chips 4 and 9). The monitors $NW40_4$ and $NW40_5$ have the same LODs =1.73 with different LOD_D, where the measured difference in RO frequency is 0.3%. RO monitors $NW40_7$ and $NW40_8$ have the same LODs = 1.00 with different LOD_D, where the measured difference in RO frequency is 0.02%. Since increasing LODs produces 90% of the effect of increasing LOD_D and LODs combined, and as increasing LOD_D has essentially no effect, it can be concluded that this observation is in agreement with [45], which attributes the injection velocity to be an important factor. Thus, RO frequency sensitivity can be used to monitor the impact of injection velocity on device performance.

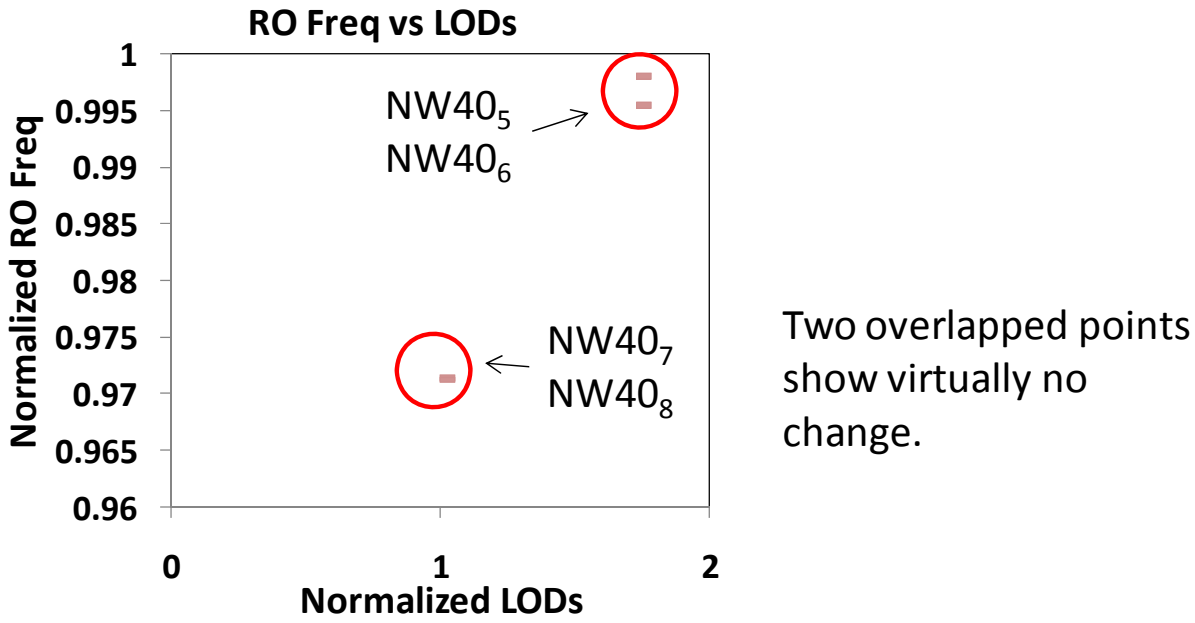


Figure 6.5: For the same LODs, asymmetrical and symmetrical monitors show minor differences in measured RO frequency, indicating that the impact of the drain on device performance is small.

6.1.5 RO Frequency versus Mobility and Injection Velocity Models

This section describes the collaborative work resulting from the comparison of measured and simulated RO frequency sensitivity from nitride CESL-strain induced monitors using the stress contours and delay models that were carried out by Xu [7]. Xu's work focuses on the use of the TCAD tool, Sentaurus, to study the stress profiles for different layout geometries. The parameters used in such stress modeling are generic and without calibration against silicon data. Nevertheless, by studying the stress distributions for different layout dimensions, a physical intuition can be developed and used to better understand the measured RO frequency sensitivity.

Xu characterizes the impact of CESL-induced strain on injection velocity and mobility through simulation. He then develops an intrinsic delay model that incorporates the effects of the mobility or injection velocity theory. The new injection velocity model is compared to that of the traditional mobility model.

Xu applied this modeling approach to the measured RO frequency results for symmetrical layouts. Figure 6.6 plots the measured and simulated change in RO frequency versus the normalized LOD'. Using a logarithmic fit, the simulated models are extrapolated for long LOD'. As expected, the roll-off in measured RO frequency shows that ST Micro's 45nm devices are operating in a regime in between traditional (mobility) and injection velocity models. Hence, this study suggests that a new extension of the PDK device model is needed for CESL stress effects.

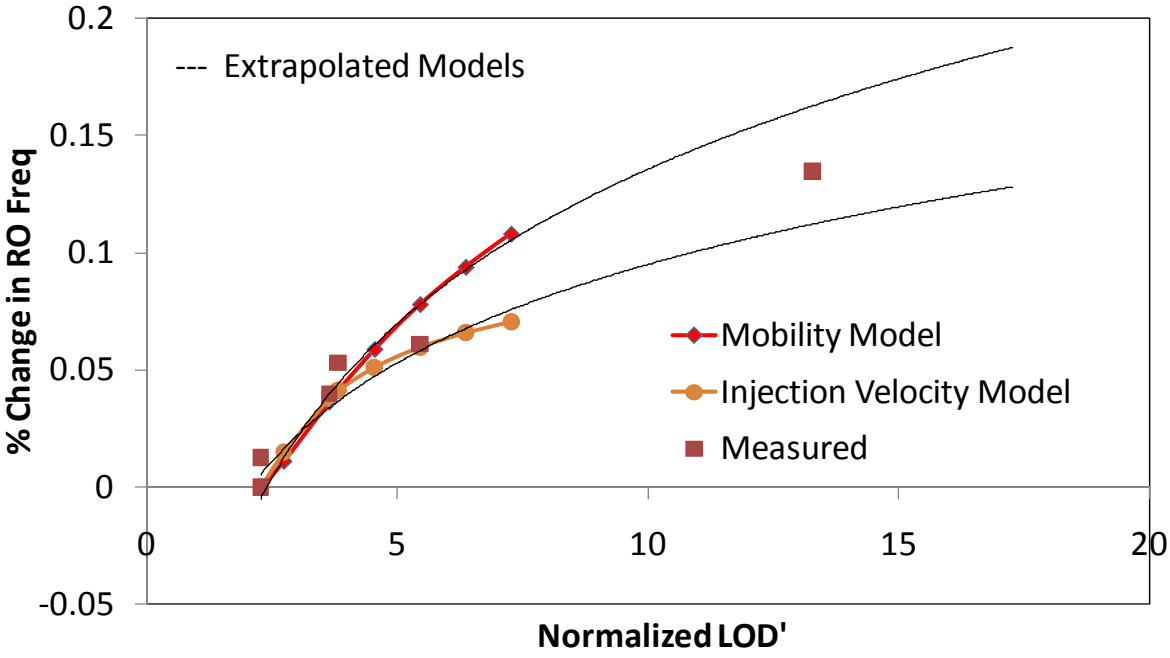
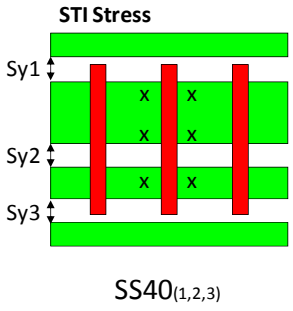


Figure 6.6: The percentage change in RO frequency versus LOD' shows that the 45nm ST Micro devices operate in a regime between that predicted by traditional (mobility) and injection velocity models.

6.2 RO Frequency Sensitivity to STI-Induced Stress

The measured RO frequency sensitivity to STI width in a vertical direction is now considered. In this work, three STI monitors ($SS40_{1,2,3}$) are designed with various STI widths between the fixed power rail and PMOS active (*Sy1*), between PMOS active and NMOS active (*Sy2*), and between the NMOS active and the fixed ground rail (*Sy3*). The normalized dimensions for *Sy1*, *Sy2*, and *Sy3* are shown in Table 6.3, where layout dimensions are normalized to approximately three drawn gate lengths. Here, for purposes of comparison, $SS40_1$ is the control monitor.



	$SS40_1$	$SS40_2$	$SS40_3$
Sy_1	5.6	3.2	5.6
Sy_2	4.8	7.2	7.7
Sy_3	6.1	6.1	3.2
Total STI Width	16.5	16.5	16.5

Table 6.3: The normalized layout dimensions for STI-induced stress monitors

The total STI width (i.e. sum of Sy_1, Sy_2, and Sy_3) shown in Table 6.3 is fixed by power and ground rails. The $SS40_2$ monitor is designed such that the PMOS devices shift up towards the power rail so that NMOS devices experience a net increase in surrounding the STI width.

The $SS40_3$ monitor is designed in such a manner that, when NMOS devices are shifted down towards the ground rail, the PMOS devices experience a net increase in STI width.

The measured RO frequency sensitivity from the RO monitors, $SS40_2$ and $SS40_3$, are examined relative to that of the control $SS40_1$. Figure 6.5 shows the mean measured RO frequency for these 3 RO monitors across the 17 chips. In order to correct for the parasitic effects, the measured RO frequency is normalized using SPICE simulation.

Shifting the PMOS devices up towards the power rail by 0.3um results in a 3% RO frequency increase, while shifting the NMOS devices down towards the ground rail has little impact. This observation indicates that the NMOS devices show more sensitivity to STI stress. The bulk silicon piezoresistance coefficients suggest that these devices are oriented with <100> channel direction [8], and this corresponds to the channel direction published in the literature by ST Micro for the 45nm technology node [46]. For each RO frequency measurement, the measured $(\sigma/\mu)_{36} = 0.3\%$, which is consistent for the minimum width devices shown in Chapter 4.

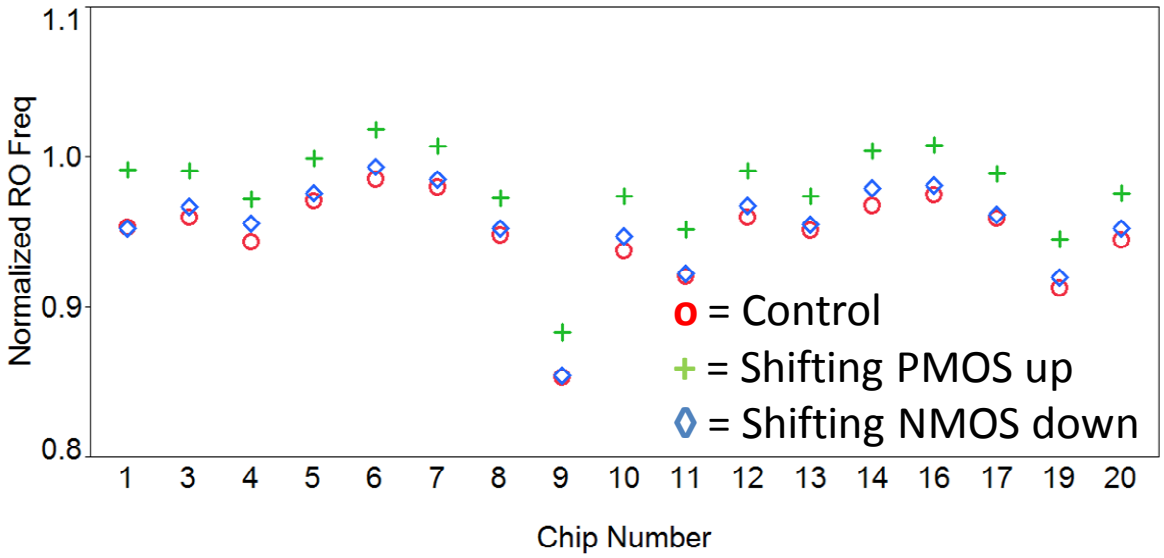


Figure 6.7: Across-wafer RO frequency measurements for STI monitors show that shifting NMOS down by 0.3um has a negligible effect, while shifting PMOS up by 0.3um speeds up the RO frequency by 3%.

6.3 Summary

Adjustments in inverter layout design parameters can enhance the sensitivity of the RO frequency variation in CESL-induced strain effects and STI-induced stress effects to be well above the level of random variation, which is 0.2-0.3%. Nitride CESL strain monitors show up to a 13.9% increase in RO frequency for a long LOD compared to that of the minimum LOD. About a 1% effect was observed with 200nm changes in the lateral STI spacing. For the same LOD, asymmetrical monitors with bigger sources operate 3% faster than monitors with bigger drains, indicating that the LOD on the source side is dominant. A 3% speed-up is observed when the PMOS devices are shifted up towards the power rail, indicating that NMOS devices are more

sensitive than PMOS devices in regard to STI-induced stress. By referencing piezoresistance coefficients, a channel direction of <100> is deduced, and this result supports the results seen in the published literature by ST Micro [44]. Measured RO frequency sensitivity to CESL and STI strain/stress can be used to determine the impact of mobility/injection velocity on device performance as well as to enhance transistor modeling for nano-scale devices.

Chapter 7
Random Noise Analysis: Gate Length, Gate Oxide Thickness, and Doping Variation

This chapter investigates the nature of random noise components. The noise level for a single RO monitor layout design measurement across a given small chip typically has a standard deviation to mean ratio of $(\sigma/\mu)_{36}$= 0.2% and 0.3% for the large and small width devices, respectively. In this chapter, measurements at different voltages and temperatures will be used to more specifically identify the contribution of variation in gate length (ΔL), gate-oxide thickness (ΔT_{ox}), and channel doping (ΔN_{ch}). Voltage and temperature are the only available adjustable parameters, with the latter requiring the implementation of a temperature chuck. Such RO frequency measurements under various operating conditions are conducted using the testing methodologies described in Chapter 4.1.

The known dependence of the current-voltage (I-V) characteristics on gate lengths, gate oxide thickness, and channel doping versus the voltage and temperature has somewhat distinct and observable behavior. From the device equations, increased voltage is expected to increase the sensitivity to ΔL much more than to ΔT_{ox} or ΔN_{ch}; increased temperature is expected to produce a minor decrease in all sensitivities, with the decrease for ΔT_{ox} and ΔN_{ch} being slightly higher.

A methodology that assumes a linear approximation of the measured RO frequency variation to process parameter sensitivity is utilized in order to determine the sources of within-chip variability. The main sources of random variation are assumed to be attributable to the 3 random process parameters described above. Since the RO block sizes are physically small, RO frequency variation due to exposure dose and focus are assumed to be negligible. A similar methodology has been successfully applied to measured RO frequency for ST Micro's 90nm node, having showed that ΔL was the dominant source of random variation [1].

This chapter is organized into 6 sections. Section 7.1 investigates the quantile-quantile (QQ) and autocorrelation plots of measured RO frequency for one control-case RO monitor. Section 7.2 examines the sources attributable to the block-to-block variation for outlier chips 4 and 9. Section 7.3 analyzes the measured RO frequency sensitivity to different operating voltages and temperatures. Section 7.4 describes how a Least Mean Square (LMS) approach is used to distinguish the sources of within-chip variation attributed to ΔL, ΔT_{ox}, and ΔN_{ch}. Section 7.5 discusses the errors among measured and simulated RO frequencies with respect to changes in operating voltages and temperatures. Section 7.6 summarizes the chapter.

7.1 Quantile-Quantile (QQ) and Autocorrelation Plots of RO Frequency Measurements

The Quantile-Quantile (QQ) and autocorrelation plottings of measured RO frequency for RO monitor CW40 are used to distinguish the random within-chip effects for typical and outlier chips. The purpose of analyzing quantile-quantile plots is to determine whether the normalized measured RO frequency is drawn from a normal (i.e. Gaussian) distribution. The autocorrelation

plottings are used to find systematically repeating patterns in measured RO frequency. The analysis of the QQ and autocorrelation plots is presented in Sections 7.1.1 and 7.1.2, respectively.

7.1.1 QQ Plots

The QQ plots in Figure 7.1 present the normalized measured RO frequency versus the quantiles of the standard normal distribution for RO monitor CW40 on four chips: 8, 3, 9 and 12. Each QQ plot contains 36 measured RO frequencies. For each chip, the measured RO frequency is normalized to that of the chip's mean.

For each QQ plot in Figure 7.1, the measured RO frequency distribution is compared to that of a reference line. If the normalized measured RO frequency comes from a normal or Gaussian distribution, the plot should be linear. For typical chips (i.e. Chips 8, 3, and 12) , the measured within-chip RO frequency distribution is more Gaussian-like than that of Chip 9, whose QQ plots show a linear trend. For Chip 9, the tails of the measured RO frequency distribution slightly deviate from the reference line. In addition, three distinct slopes can be observed, i.e. the slope is steeper at the center of the distribution compared to that of the tails. These observations lead to the conclusion that the measured RO frequencies from Chip 9 contain the systematic block-to-block effects that were previously discussed in Chapters 4 and 5.

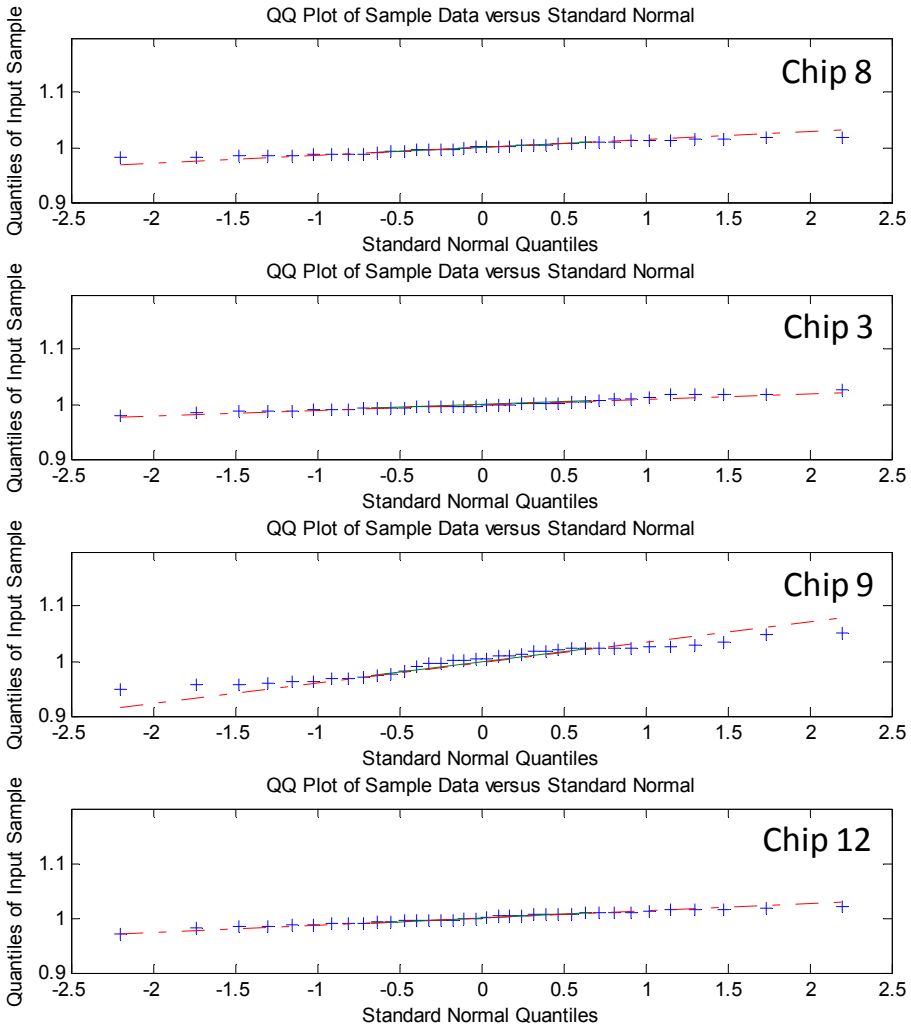


Figure 7.1: These QQ plots for RO Monitor CW40 show that Chip 9 is an outlier chip because the tails of its measured RO frequency distribution deviate from the reference line.

7.1.2 The Autocorrelation Plots

The autocorrelation plots in Figure 7.2 show the cross-correlation of the normalized measured RO frequency with its shifted (lagged) self. Such is used to quantify the similarity between normalized measured RO frequencies as a function of the spatial separation between them. The measured RO frequency is normalized to that of its chip's mean.

The autocorrelation plots for one chip and among the four chips show distinct behaviors. Thus, minimal spatial correlation exists among the individual RO monitors. Chip 9 shows the strongest periodic behavior when compared to that of other chips. The quarter period of 12 ROs again indicates that the measured RO frequencies contain systematic block-to-block effects.

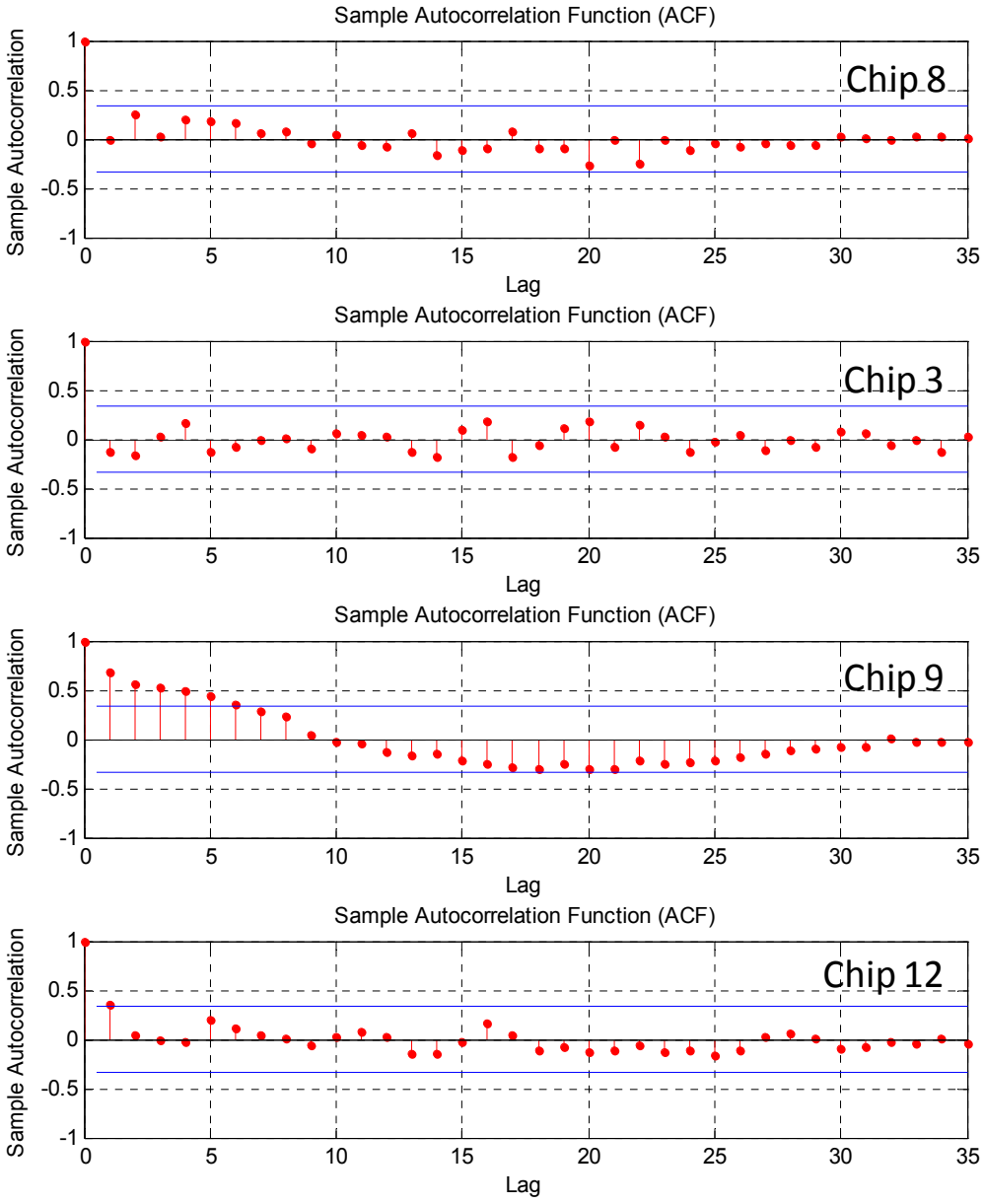


Figure 7.2: The lack of correlation among measured RO frequencies for RO monitor CW40 on the 4 chips shows that there is little spatial correlation among the RO monitors.

7.2 Block-to-Block Analysis for Outlier Chips

The systematic decrease in measured RO frequency observed in RO block 3 for Chips 4 and 9 is examined in regard to focus, etch, and stress effects. As was previously noted in Figure

5.5, outlier chips 4 and 9 show a measured RO frequency reduction of 5-8% in block 3, compared to that in other blocks.

Block-based mean measured RO frequency from RO monitors CS40, FS40, IW60, FW60, CW60, CW40, and $NW40_1$ for chips 4 and 9 is shown in Tables 7.1 (a) and (b), respectively. For each parameter-specific monitor, the measured block means are normalized to the chip's mean for the control-case RO, which are CS40, IW60, and CW40. As one may recall, each block contains 12 RO instantiations. The measured range for the block-based mean is computed for each monitor.

Chip 4	**CS40**	**FS40**	**IW60**	**FW60**	**CW60**	**CW40**	**$NW40_1$**
Block 1	1.07	1.05	1.03	1.03	1.00	1.03	0.92
Block 2	1.06	1.01	1.03	1.02	1.00	1.01	0.91
Block 3	0.98	0.94	0.95	0.95	0.92	0.95	0.84
Measured Range	7.90%	8.63%	7.86%	8.00%	7.67%	6.91%	7.87%

(a)

Chip 9	**CS40**	**FS40**	**IW60**	**FW60**	**CW60**	**CW40**	**$NW40_1$**
Block 1	1.06	1.02	1.01	0.98	0.97	1.01	0.89
Block 2	1.09	1.03	1.02	0.98	0.98	1.02	0.89
Block 3	1.02	0.96	0.96	0.93	0.93	0.97	0.84
Measured Range	4.97%	6.09%	5.28%	5.34%	5.04%	4.66%	5.18%

(b)

Table 7.1: A comparison of block-based mean for focus, etch, and stress monitors and their respective control ROs for chips 4 and 9.

For all monitors, both chips show a mean measured RO frequency reduction of ~8% for chip 4 and ~5% for chip 9 in block 3, compared to that in other blocks. Since this frequency reduction was not observed on 15 of the 17 chips, it is unlikely to be caused by a systematic within-chip effect. These results show that etch, lithography, and stress monitors are all affected in the same manner as that of their respective control ROs. The measured ranges between monitors differ at most 1%. This result indicates that none of these factors likely contribute to this frequency decrease. Since block 3 is alone on the left side of the chip, it is surmised that it might be near the edge of the wafer where thin-films of device and lithography materials or even CMP of interconnect may be non-uniform.

7.3 RO Frequency Sensitivity to Operating Voltages (V_{dd}) and Temperature (T)

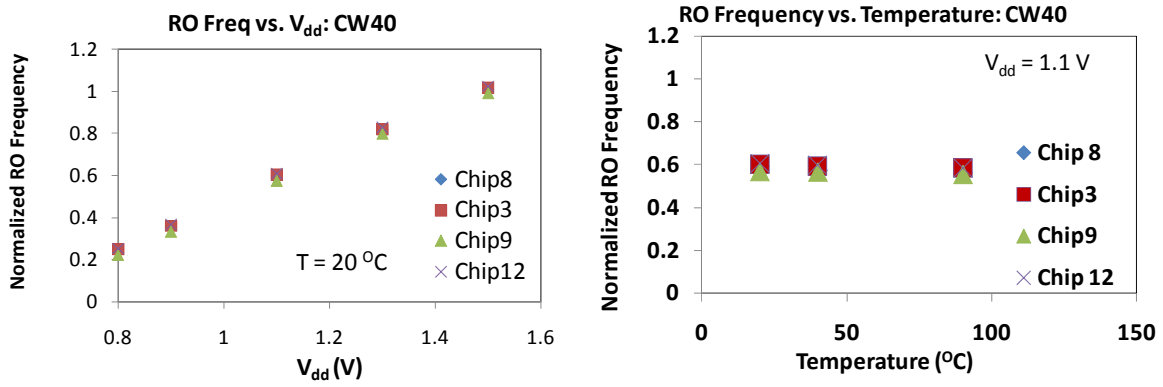
The RO frequency sensitivity to physical parameter changes depends on measurement conditions. Thus, the measured RO frequency and its noise sensitivity to operating voltages (V_{dd}) and temperatures (T) are examined as a potential method of determining the levels of noise in the various physical parameters contributing to the random component in ring oscillator performance. The dependence of RO frequency on measurement conditions is shown in Figure 7.3.

Measured RO frequency sensitivity to different operating voltages is observed at various V_{dd} and T biases. The V_{dd} bias is adjusted to the following voltages: 0.8 V, 0.9V, 1.1V, 1.3V, and 1.5 V. The T bias is adjusted to the following temperatures: 20 ^{O}C, 40 ^{O}C, and 90 ^{O}C. The mean of 36 RO frequency measurements for RO monitor CW40 operating at different V_{dd} and T

plotted for the 4 chips is shown in Figure 7.3 (a) and (b), respectively. The RO frequency measurements are normalized to Chip 8's mean at V_{dd} = 1.5V, T = 20 ^{O}C.

As expected, RO frequency speeds up as operating voltages increase and decreases slightly with increasing temperatures. At constant T, the measured RO frequency increases by 80% as V_{dd} increases from 0.8 V to 1.1 V. At constant V_{dd}, RO frequency decreases slightly by 10% as T increases from 20 ^{O}C to 90 ^{O}C.

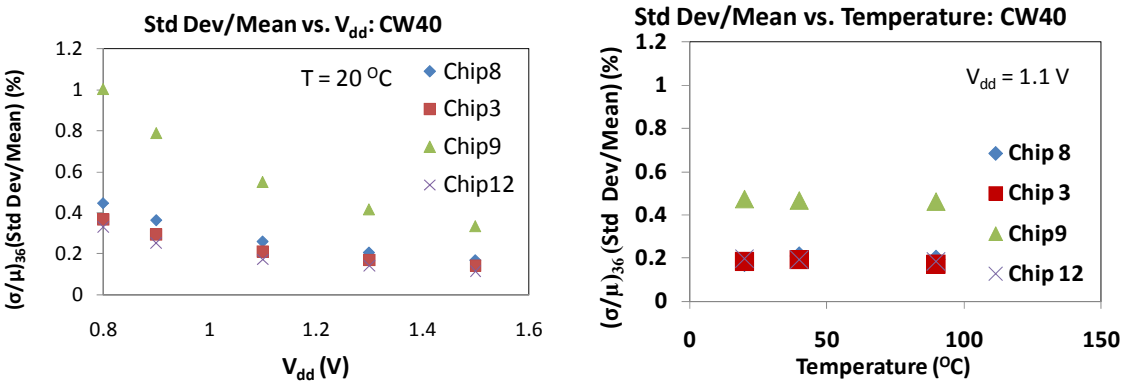
The slopes of the curves in Figure 7.3 give an indication of how the tolerances on the measurement apparatus affect the accuracy of the measurements. The slopes of the normalized measured RO frequency to V_{dd} and T are 1.1/V and 0.0002/^{O}C, respectively. As one may recall from Chapter 4.1, the resolution limit from the operating voltage and temperature measurement setups are $\pm$1 mV and $\pm$ 5 ^{O}C, respectively. A change in $\pm$1 mV or $\pm$ 5 ^{O}C induces a negligible $\pm$0.001 change in normalized RO frequency. Thus, the noise contribution due to the finite limit in resolution from the digital multi-meter and temperature sensor has minimal impact on the measured RO frequency.



(a) (b)

Figure 7.3: The normalized mean measured RO frequency at V_{dd} =0.8V to 1.5 V (a) and T=20^{O}C to 90^{O}C (b) for the 4 chips show that RO frequency is more sensitive to changes in V_{dd} than in T.

The measured $(\sigma/\mu)_{36}$ for the RO frequency results presented in Figure 7.3 (a) and (b) is presented in Figure 7.4 (a) and (b), respectively. Figure 7.4 (a) reveals that the measured $(\sigma/\mu)_{36}$ saturates as V_{dd} increases from 0.8 V to 1.5V. As expected, Chip 9 has the highest measured $(\sigma/\mu)_{36}$. A significant rate of change in the measured $(\sigma/\mu)_{36}$ mainly occurs at low V_{dd}, where the slope from V_{dd} = 0.8 V to 1.1 V is ~1.3x bigger than the slope from V_{dd} = 1.1V to 1.5V for typical chips. For Chip 9, the ratio of the two slopes can be as high as ~2.8x. Figure 7.4(b) shows that the measured $(\sigma/\mu)_{36}$ remains relatively constant for different operating T.



(a) (b)

Figure 7.4: The measured $(\sigma/\mu)_{36}$ from RO frequency at V_{dd}= 1.1V to 1.5V and T =20^{O}C to 90^{O}C for the 4 chips shows that $(\sigma/\mu)_{36}$ saturates with increasing V_{dd} and remains relatively constant as the temperature increases.

7.4 Inferring Process Parameter from Random Sources of Variability using Least Mean Square (LMS) Analysis

A least mean square (LMS) approach is used to relate the measured RO frequency variation to three process parameters: L, T_{ox}, and N_{ch}, in a linear formulation: MX=Y. Here, M is a matrix of the computed sensitivity of RO frequency to each of the three parameters at each of the 5 measurement conditions. The Y matrix contains the normalized RO frequency measurements. The matrix X is the normalized solution for ΔL, ΔT_{ox}, and ΔN_{ch} for each of the 36 RO monitors. The solution is computed via SVD in Matlab for each RO monitor type on one chip.

This approach is dependent on the RO frequency sensitivity to process parameters, which varies with the measurement conditions. For this work, these coefficients in the M matrix are derived from the linear approximation of RO frequency to the three process parameters L, T_{ox}, and N_{ch} under five measurement conditions. Utilizing SPICE and PSP/BSIM4 model cards from ST Micro's 45nm PDK, RO frequency versus changes in L, T_{ox}, and N_{ch} relationships are found. Using these simulated relationships, the slopes at the mean measured RO frequencies are extracted and used as coefficients in the M matrix.

7.4.1 The Index Definitions

This section defines the indices that are used to describe the LMS formulation. Let *i* represent measurement condition number, *j* represent RO monitor instantiation, and *k* represent the Random Process Parameter number. In this analysis, $i \in \{1, 2, 3, 4, 5\}$ since there are 5 measurement conditions. Here, $j \in \{1,2, 3, \ldots, 34, 35, 36\}$ because there are 36 identical RO monitors instantiated across chip and, thus, 36 RO frequency measurements; every 12 RO instantiation represents one RO block. Here, $k \in \{1, 2, 3\}$ since 3 process parameters are being studied (i.e. L, T_{ox}, and N_{ch}).

7.4.2 The Variable Definitions and RO Frequency Sensitivity to L, T_{ox}, and N_{ch} Extraction

This section defines the variables used in the LMS formulation and describes the extraction procedure used to obtain the normalized RO frequency versus process parameter sensitivity.

- Let F_{ij} represent the j^{th} measured RO frequency under the i^{th} measurement condition.
- Let Fo_i represent the mean measured RO frequency over j under the i^{th} measurement condition for one RO block.
- Let L_{typ}, Tox_{typ}, and Nch_{typ} represent the typical values of L, T_{ox}, and N_{ch} obtained from PSP/BSIM4 model cards. These parameters are independent of measurement conditions. For example, L_{typ} is 40 nm.
- Let $\frac{\partial F}{\partial L}$, $\frac{\partial F}{\partial Tox}$, and $\frac{\partial F}{\partial Nch}$ represent the simulated RO frequency sensitivity to L, T_{ox}, and N_{ch}, respectively. They are derived by extracting the slopes at F_{oi} and are used as coefficients in the M matrix after normalization to unit-less quantities.

7.4.3 The Matrix Definitions

The variables defined above can now be formulated into a least mean square problem: MX=Y. Matrices, Y_{ij} (with dimensions i x j), X_{kj} (with dimensions k x j), and M_{ik} (with dimensions i x k) are defined such that:

Y_{ij} represents the measured RO frequency normalized by Fo_i for the j^{th} measured RO frequency under the i^{th} measurement conditions.

X_{kj} represents the amount of variation attributed to the k^{th} process parameter in the j^{th} measured RO frequency.

M_{ik} represents the sensitivity of RO frequency attributed to the k^{th} process parameter under the i^{th} measurement condition obtained from SPICE and PSP/BSIM4 simulations.

$$Y_{ij} = \left[\frac{F_{ij} - Fo_i}{Fo_i}\right] \tag{7.1}$$

$$X_{kj} = \begin{bmatrix} \frac{\Delta L_j}{L_{typ}} \\ \frac{\Delta Tox_j}{Tox_{typ}} \\ \frac{\Delta Nch_j}{Nch_{typ}} \end{bmatrix} \tag{7.2}$$

$$M_{ik} = \begin{bmatrix} \frac{\partial\left(\frac{F}{Fo_i}\right)}{\partial\left(\frac{L}{L_{typ}}\right)} & \frac{\partial\left(\frac{F}{Fo_i}\right)}{\partial\left(\frac{Tox}{Tox_{typ}}\right)} & \frac{\partial\left(\frac{F}{Fo_i}\right)}{\partial\left(\frac{Nch}{Nch_{typ}}\right)} \end{bmatrix}$$

(7.3)

If $Y = MX$, to determine an estimated solution, $\dot{X}$:

$$\dot{X} = M^{-1} \cdot Y \tag{7.4}$$

where $\dot{X}$ is the least square estimate of X computed through SVD.

7.4.4 The Least Mean Square Solution

Using the methodology described in the previous sections, the normalized sensitivity of RO frequency to L, T_{ox}, and N_{ch} can be found at each measurement condition, *i*. Tables 7.2 and 7.3 enumerate the measurement conditions and their corresponding sensitivities from Chip 9 for RO monitors, CW40 and CW60, respectively. Both tables show that as the operating voltage increases, RO frequency is more sensitive to changes in L, T_{ox}, and $N_{ch.}$ This increase is due to the fact that drain-induced barrier lowering (DIBL) effects exhibit a more prominent increase at differing rates at higher operating voltages.

As gate length increases from 40nm to 60nm, the sensitivity of RO frequency to L shows a drastic 3x decrease, the sensitivity to N_{ch} is reduced by about 20%, and the sensitivity to T_{ox} is reduced by about 10%.

M Gradient Matrix	$\frac{\partial(F)}{\partial(L)}$	$\frac{\partial(F)}{\partial(Tox)}$	$\frac{\partial(F)}{\partial(Nch)}$
i = 1, V_{dd}=0.8, T= 20 ^{O}C	-0.47	-0.75	-0.53
i = 2, V_{dd}=0.9, T= 20 ^{O}C	-0.71	-1.03	-0.60
i = 3, V_{dd}=1.1, T= 20 ^{O}C	-1.17	-1.36	-0.73
i = 4, V_{dd}=1.1, T = 40 ^{O}C	-1.15	-1.22	-0.72
i = 5, V_{dd}=1.1, T = 90 ^{O}C	-1.14	-1.15	-0.67

Table 7.2: Chip 9's measurement conditions and its corresponding sensitivity to process parameters for RO monitor CW40

M Gradient Matrix	$\frac{\partial(F)}{\partial(L)}$	$\frac{\partial(F)}{\partial(Tox)}$	$\frac{\partial(F)}{\partial(Nch)}$
i = 1, V_{dd}=0.8, T= 20 ^{O}C	-0.14	-0.66	-0.42
i = 2, V_{dd}=0.9, T= 20 ^{O}C	-0.21	-0.93	-0.49
i = 3, V_{dd}=1.1, T= 20 ^{O}C	-0.41	-1.24	-0.59
i = 4, V_{dd}=1.1, T = 40 ^{O}C	-0.38	-1.11	-0.58
i = 5, V_{dd}=1.1, T = 90 ^{O}C	-0.36	-1.09	-0.54

Table 7.3: Chip 9's measurement conditions and its corresponding sensitivity to process parameters for RO monitor CW60

Least-mean-square analysis is applied to the measured RO frequency. Chip 9 was chosen as an example here because it has interesting block-to-block variation. In this case, to reduce block-to-block effects, the measured RO frequency is normalized to the corresponding RO block mean. The 3-parameters solution of RO monitors: CW40 and CW60 on Chip 9 are each shown for the 36 RO instantiations in Figures 7.5 and 7.6, respectively. Both figures plot the normalized percentage change attributed to each parameter: L, T_{ox}, and N_{ch} versus the RO instantiations. In general, the source of within-chip variation is mainly attributed to random dopant fluctuation (ΔN_{ch}) effects.

Specifically, Figure 7.5 shows that within-block variation trends differ from block-to-block. Variation in Block 1 shows that the ΔN_{ch} impact is 6x greater than that of ΔL. ΔL, ΔT_{ox}, and ΔN_{ch} contribute $\pm$ 1 %, $\pm$ 1%, and $\pm$ 6%, respectively. Within-block variation for block 2 shows that ΔN_{ch} impact is 2x bigger than that of ΔL, in which ΔL, ΔT_{ox}, and ΔN_{ch} contribute $\pm$ 1.5 %, $\pm$ 1%, and $\pm$ 3%, respectively. Within-block variation for Block 3 shows that the impact of ΔN_{ch} and changes in ΔL are relatively similar with ΔL, ΔT_{ox}, and ΔN_{ch} contributing $\pm$ 1 %, $\pm$ 1.5%, and $\pm$ 2%, respectively. Block 3 shows the least amount of total variability, and this correlates with the measured RO frequency spread in Figure 4.11.

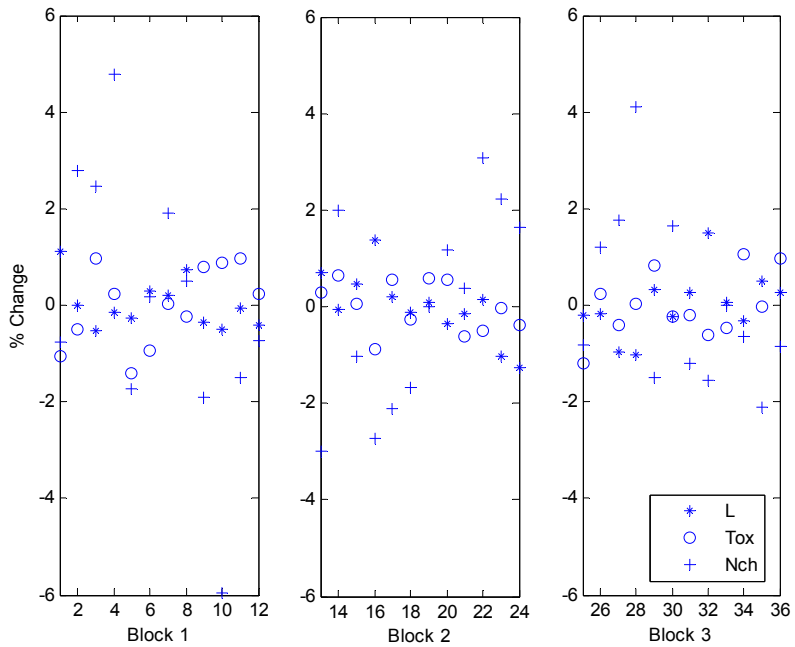


Figure 7.5: The percent change in L, T_{ox}, and N_{ch} versus RO instantiation for Chip 9 RO monitor CW40. Maximum N_{ch} variation is ~$\pm$ 6%. Block-to-block variation (every 12 RO) is present.

Now the 60nm data in Figure 7.6 is considered. In Chapter 4, the measured noise for CW60 is observed to be similar to that of CW40, though perhaps a little smaller. Figure 7.6 shows that with a 1.5x increase in gate length, the impact of ΔN_{ch} is reduced in Blocks 1 and 3,

where ΔN_{ch} is about 20% smaller than that of 40nm LMS solutions. More importantly, the impact of ΔL did not decrease by 3x as expected; this effect may be due to the cross talk between noise sources induced by slight errors in the M matrix, such as the effects of N_{ch} mixing into L. Thus, it can be concluded that the LMS methodology may not have a high enough precision to detect effects below a 0.1% change in measured RO frequency.

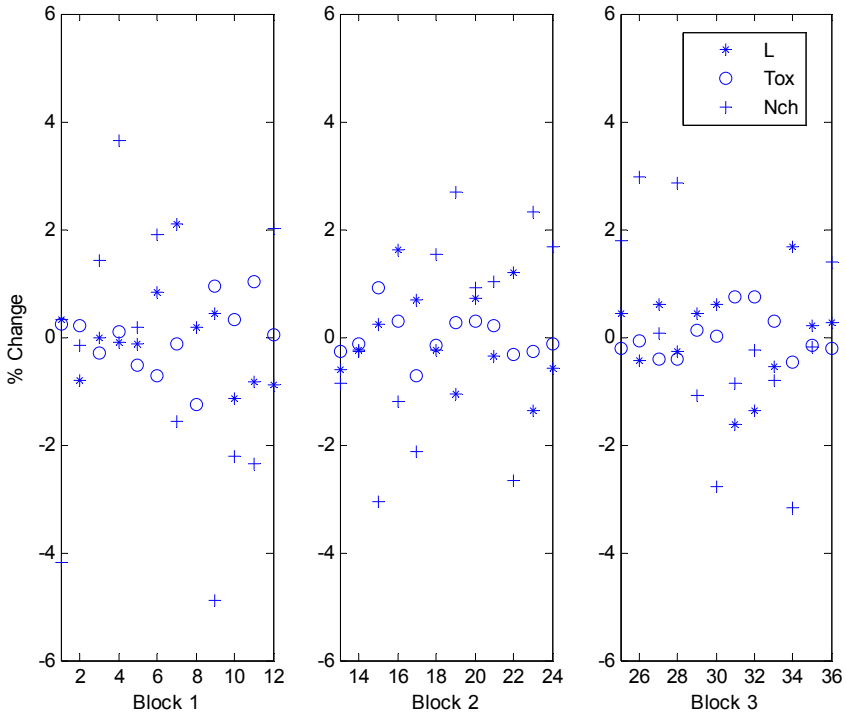


Figure 7.6: The percent change in L, T_{ox}, and N_{ch} versus RO instantiation for Chip 9 RO monitor CW60. Maximum N_{ch} variation here is $\pm$ 4%, and block-to-block variation (every 12 RO) is present.

The residual error from the estimated 3-parameter solutions is examined. Here, estimated RO frequency solutions from MX-matrices are subtracted from that of the measured Y-matrices in order to quantify the error between the measured and estimated RO frequency solutions for RO monitor CW40. The differences for each 3 RO block are shown in Figure 7.7. The magnitude of the differences between the measured and the estimated solutions is on the order of $\sim 10^{-3}$. The same magnitude of residual error is observed for RO monitor CW60.

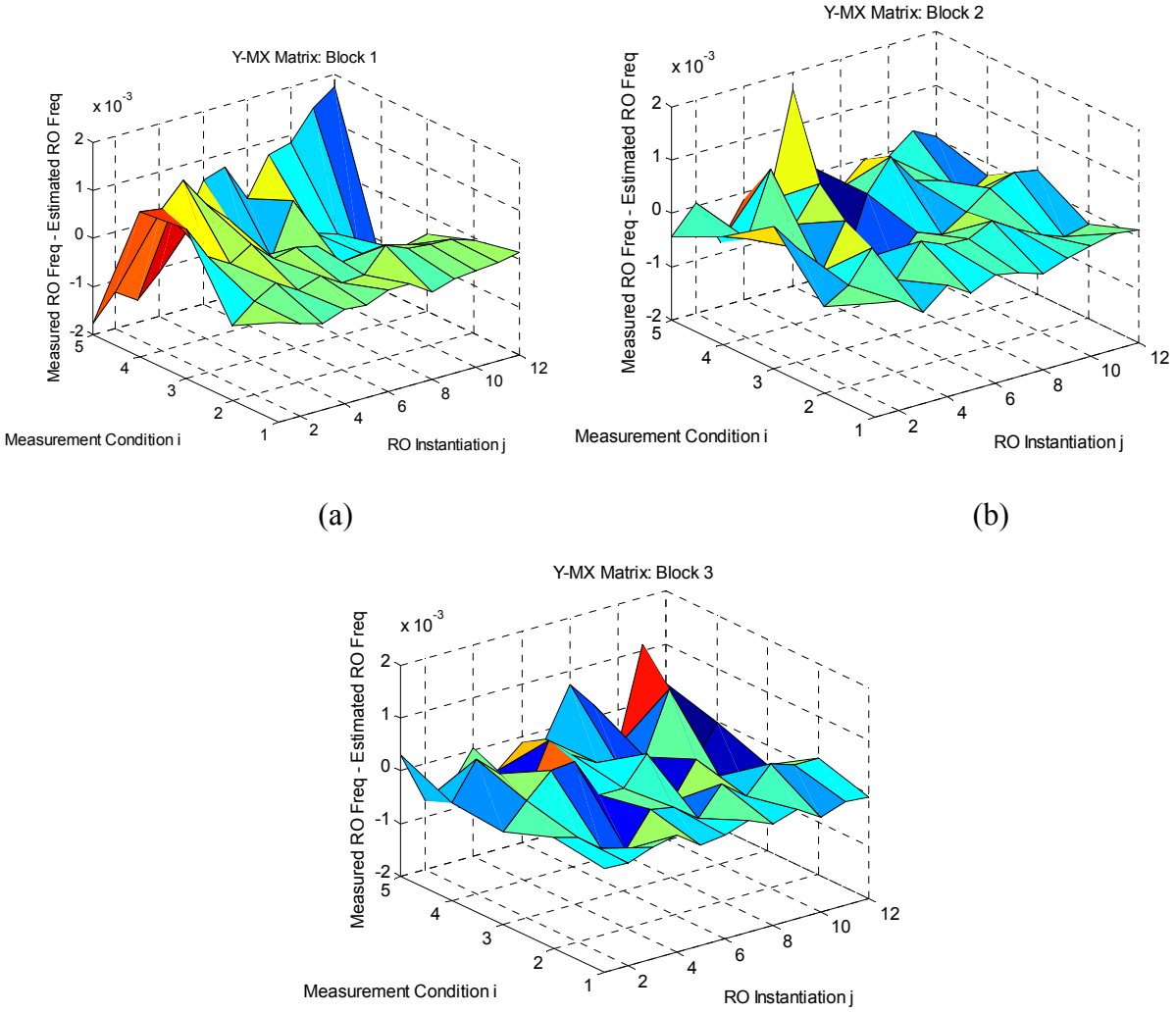


Figure 7.7: The differences between measured RO frequency and the estimated solutions show a magnitude on the order of 10^{-3}.

7.5 Error Analysis for the LMS Solution

Two sources of error are used to quantify the validity of the LMS approach. The first source of error is associated with the properties of the matrix that may affect the algorithm or the floating point errors contributed by the computation platform. The second source of error is associated with the computed sensitivity from simulations: $\frac{\partial F}{\partial L}$, $\frac{\partial F}{\partial Tox}$, and $\frac{\partial F}{\partial Nch}$. These are in addition to the accuracy of the measurements, potential nonlinearities, and confounding effects between parameters.

The condition number of a matrix is a property of a matrix that quantifies the singularity of the matrix [48]. The larger the condition number, the more singular is the matrix. Using the condition number as a metric, matrix properties are examined for the *M* matrix presented in Tables 7.1 and 7.2. The condition numbers for the M matrices are ~50. This small condition number indicates that the matrix is non-singular; singular matrices usually have condition

numbers in the 10^5 magnitude [48]. Thus, since the error stemming from the properties of the matrix is determined to be negligible here, computer accuracy is not an issue.

The main source of error appears to be in computing the sensitivity of RO frequency to L, T_{ox}, and N_{ch} due to the discrepancy between the simulated RO frequency and that which is measured. Figures 7.8 and 7.9 examine how well the measured mean RO frequency over 36 RO instantiations for one RO monitor (CW40) on one chip (Chip 9) correlates with that of the simulated RO frequency at the SS corner as V_{dd} and T vary. The measured and simulated RO frequencies are normalized to the simulated RO frequency at standard operating conditions, i.e. V_{dd}=1.1 V and T=20 ^{O}C.

Figure 7.8 shows the normalized mean measured and simulated RO frequency versus V_{dd} for RO monitor CW40 from Chip 9. As V_{dd} deviates from 1.1V, the simulated RO frequency also deviates from that of the measured RO frequency. The normalized measured and simulated RO frequency to V_{dd} slopes are 1.9/V and 2.2/V, respectively. The error between the two slopes is 15%, and this difference can cause an error to be as high as ~20% in absolute value. Thus, in order to reduce the error between the measured and the simulated RO frequency, sensitivity for the LMS application is extracted at V_{dd} = 0.8V, 0.9V, and 1.1 V.

Despite limiting the range of operating voltage for LMS applications, an 8% difference is observed at V_{dd}= 0.8 V. The sensitivity in the M matrix requires the computation of the slopes of RO frequency to a parameter using the PDK, and the slopes of RO frequency to voltage is affected by 15%. Thus, it is expected that using the PDK to estimate the process parameter would yield a similar 15% error.

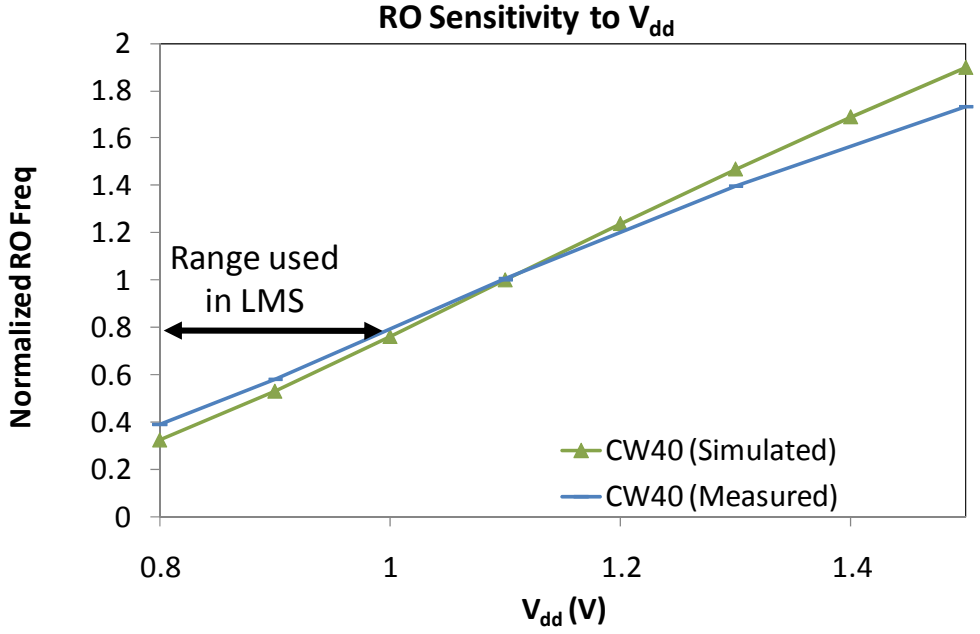


Figure 7.8: The simulated RO frequency deviates from that of the measured as V_{dd} deviates from the standard operating voltage, 1.1 V.

Figure 7.9 plots the normalized mean measured and simulated RO frequency versus T for the same RO monitor shown in Figure 7.8. The measured and simulated RO frequency has a consistent slope. As the operating temperature increases, the errors between the measured and the simulated RO frequency decrease slightly, by 1.54%.

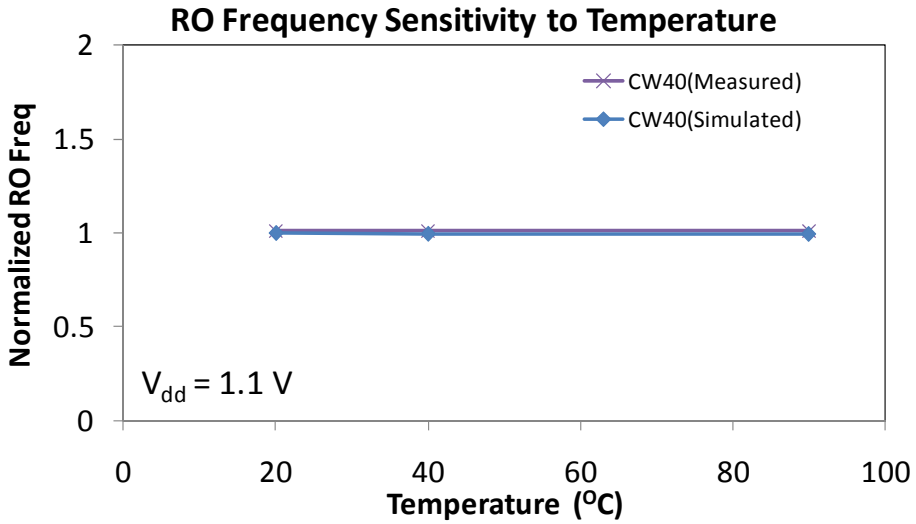


Figure 7.9: The measured and simulated RO frequency show good correlation with increased operating temperature.

7.6 Summary

Within-chip variation attributed to random sources, such as random changes in gate length, gate oxide thickness, and channel doping, have been examined. The quantile-quantile and autocorrelation plots are used to ensure that the measured RO frequencies stem from Gaussian distributions. The systematic decrease in measured RO frequency observed in RO block 3 for outlier chips 4 and 9 is examined in regard to etch, lithography, and stress effects. The results show that these monitors are all affected in the same manner as that of their respective control ROs, indicating that these factors unlikely contributed to this frequency decrease. For an outlier chip, the RO frequency is linearized to 3 process parameters using simulations from a 45nm ST Micro PDK. Such sensitivity is then related to measured RO frequency variation for each RO instantiation using a least mean square (LMS) approach. The least-mean square-estimated solutions show that random dopant fluctuations have 4-6x more impact on RO frequency variability than that of changes in gate length and oxide thickness. Error analysis shows a ~10^{-3} residual error between measured and LMS-estimated RO frequency changes. The main source of error appears to be a 15% difference in the slope of RO frequency versus voltage in the PDK model from that of the measurement, producing a similar error in using the PDK to calculate parameter sensitivity. An analytical approach using only voltage that neglects the variation in T_{ox} and computes only ΔL and ΔN_{ch} effects for CS40, CW40, and CW60 is suggested in order to separate any variation that is due to gate length and doping effects.

Chapter 8
Conclusions

This dissertation explores the feasibility of using electronic measurements of parameter-specific ring oscillators to identify and quantify circuit performance variation that is attributable to specific manufacturing and layout design parameters. Experimentally measured results from a 45nm process were made possible through the design and testing in the Berkeley Wireless Research Center of silicon chips fabricated in a collaborative effort by ST Micro. The measured results were compared to their expected sensitivity, as determined by simulation. Residual random variation was examined for within block, chip, and wafer variation, as well as for possible variation contributions in gate length, gate oxide thickness, and channel doping. The sections below summarize such contributions and offer a perspective in regard to future work in relation to each stage of this dissertation. The final section gives a consolidated overall perspective.

8.1 Design

Ring Oscillator (RO) layout monitors were created in order to characterize 5 physical effects: gate etch, gate focus, gate-to-active misalignment, nitride CESL-induced strain, and STI-induced stress. A total of 32 different layouts have been used to monitor these 5 effects. Each monitor consists of 13 stages of inverters, in which the parameter-specific sensitivity is enhanced while still meeting the design rules from the PDK kit. The enhancement of the layout-dependent lithography effects was designed using simulation with Mentor Graphics Calibre. The parameters used in the exposure and resist models are based on generic assumptions. A bias of 10 nm was added to each side of the gate length, but it was assumed that other OPC methods typically used in production, such as scatter-bar insertion, would not be applied. For the focus monitors, Pattern Matching was utilized in order to guide the design, and this was followed by tuning with Calibre simulation. A 1.5x enhancement of gate focus sensitivity compared to that of a dense gate control RO was achieved by strategically placing additional poly features on and surrounding the gate. With respect to the gate-to-diffusion misalignment monitors, a set of 5 monitors with pre-programmed offsets at 5 different misalignment levels were designed so as to detect the direction and magnitude of overlay error. In order to study the RO frequency sensitivity to CESL-induced strain, to non-uniformity of strain distribution, and to lateral STI-induced stress, RO monitors were designed at 8 various lengths of source/drain diffusion (LOD), including those with asymmetric source/drain areas. In addition, three STI monitors were designed to examine RO frequency sensitivity to changes in the vertical STI width.

In hindsight, increasing the variety of layout monitor designs is suggested. This increase is helpful because it will offer more reference points from which to triangulate and quantify any causes of variation. For example, while collaborating with Seng-Oon Toh in triangulating the effects in SRAM read-write performance asymmetries, it was discovered that SRAMs were very sensitive to both directions of misalignment and that no vertical misalignment monitors were developed. In enhancing the parameter-specific lithography effects, it would be very beneficial if the monitor designs would be allowed to selectively violate design rules without violating mask making rules, just as is allowed with OPC. In characterizing the CESL and STI effects, a more systematic set of curves could have been developed if the adjacent dummy poly did not

implicitly force constraints on the LOD and STI. Extra care should also be taken in designing asymmetrical layouts to be sure to flip the symmetry of the layout with the signal propagation direction, as this was overlooked in regard to the wedge-shaped diffusion for monitoring misalignment.

8.2 Measurement Methodology

To measure the RO, a packaged chip was manually mounted onto a PCB where measurement equipment was connected via GPIB cables and other wires. This was carried out at BWRC with the assistance of scripts used by L.T. Pang and staff member Susan Mellers. In general, the test equipment has a high degree of accuracy, and errors from test settings cause a negligible RO frequency change. The voltage monitors and the temperature sensors/environments have precisions of $\pm$1 mV and $\pm$5 ^{O}C, respectively. The remounting and re-measuring of chips cause a 0.1%-0.16% measured RO frequency drift, which is higher for the smaller gate area devices. Currently, each chip takes ~3 hours to measure all of the 32 monitor designs times the 36 instantiations or 1152 RO monitors.

In hindsight, these measurements could be reduced to 20 minutes with little loss of accuracy. In this work, in order to reduce noise in signal capture, the frequency measurement of each RO was repeated 120-140 times. The typical noise is, however, so low at 0.009% that a sample size of 10 would still have much less variation than the 0.2-0.3% associated with the individual RO. Extending the temperature range from the current limits 20 ^{O}C to 90 ^{O}C would be helpful as the sensitivity of the RO frequency to L, T_{ox} and N_{ch} is quite low.

8.3 Initial Measurement Results

Seventeen chips from one wafer were received and every one of the 19,5840 RO that were tested worked. Fifteen chips were found to be typical and showed both no block-to-block (within-chip) effects and no correlation in measured RO frequency among the 36 instantiations on a chip for the same monitor design. Hence, the chip mean for monitoring systematic effects is used, where the noise floor from the standard deviation of 36 RO instantiations to the mean ratio was 0.2-0.3%. These ratios, when plotted versus the inverse square root of the gate area, show a linear behavior similar to that of the Pelgrom model. A direct comparison was made by scaling the ratios with the sensitivity to the threshold voltage and then multiplied by the square root of the number of RO devices. This comparison showed an average effective threshold voltage slope of 2.3 mV um^{-1}, similar to that which was reported in the literature for the 45nm technology node. The measured across-wafer variation for small gate area devices showed an 11.1% range that had a double humped behavior that might be produced by a bulls-eye effect. Measured across-wafer ($3\sigma/\mu$) variation for a control-case RO monitor was 8%, a significant improvement over the 15% variation measured by L.T. Pang from a previous silicon run during the development this technology one year earlier.

These raw RO frequency measurements include layout-dependent effects due to changes in parasitics as well as changes due to process parameter specificity. To separate out the parasitic effects, the measured RO frequency was normalized to SPICE-simulated RO frequency by dividing the measured with that of the simulated. For RO monitors with dummy polys, this parasitic capacitance correction changes measured RO frequency by 2%.

In hindsight, the residual device noise of individual transistors was quite high. Fortunately the absence of within-chip systematic effects allowed the average of all 36 ring

oscillators to be used with a noise level well below the parameter-specific effects. Had this not been the case, more ring oscillators would have been beneficial. In comparing the circuit behavior versus the inverse square root of the gate area to that of the devices using the Pelgrom model, it is recommended that a broader variety of layouts be used with the range of the inverse square root of the gate area to be at least $5um^{-1}$ to $20um^{-1}$.

8.4 Monitoring of Gate Etch, Gate Lithography, and Gate-to-Active Misalignment

The measured RO frequency sensitivity to gate etch, gate focus, and gate-to-active misalignment monitors were examined. The directly measured etch monitors showed that the absence of two adjacent dummy polys caused a 3.13% RO frequency shift. The PV bands provided by ST Micro did not show a change in gate length among these layouts indicating that lithography is not likely the source of this variation. This similarity between isolated and dense unfortunately indicates that strong OPC with scatter-bars was used. It was necessary to correct for the increase in the stray capacitance between the drain and the dummy poly. SPICE simulation at TT produced a 1% and 2% RO frequency decrease for single and double dummy poly removal, respectively. With this capacitive correction, the decrease in RO frequency attributable to process effects such as etch was 1.2%.

The measured RO frequency sensitivity to gate focus monitors showed that they were about 4% slower than the control. This is attributed to parasitic effects as well as the non-uniform 'hour-glass' shape produced at the top and bottom of the gate from the horizontal extensions used to increase focus sensitivity. Data was available only from one production wafer, which due to the presence of other products, did not have programmed treatments for defocus and dose. The fact that the experimental measurements were less sensitive than predicted during the design stage is in part attributed to the fact that the wafer appears to have been run under unusually good control. This observation is supported by the fact that the measured range of RO frequency was typically centered and 1/6-1/4^{th} of that of the SPICE-simulated SS-FF guard band. Also, the PV bands of the designs provided by ST Micro showed that the dose exposure effects dominate over that of focus and predicted little defocus in the lithography stepper system. In addition, the unanticipated requirement to apply strong OPC techniques with scatter-bars to protect other products on the wafer instead of the gate length biases used in design also greatly reduced the sensitivity.

Given the quality of the wafer, the lack of focus treatments, the OPC of the focus monitor, and little focus sensitivity in the resulting post-OPC PV band, there was no direct way to assess the sensitivity of the focus monitor. The chip-to-chip variation among focus and control monitors were examined. A typical correlation with r-square value of > 0.9 was obtained. While this indicates that the chip-to-chip variation of 10% among the seventeen $2x2mm^2$ chips is highly correlated between different monitor layouts, an enhanced slope of the focus monitors compared to that of the control was not observed.

The measured RO frequency sensitivity from the 5 pre-programmed at 0nm, $\pm$10nm, and $\pm$15nm gate-to-active offsets for the H-shaped RO monitor showed only a 1.5% change, instead of the 3.2% change that was observed for the simulated design. The OPC of the gate structure did not adversely affect this monitor. However, in examining PV bands for the curved diffusion shape at the neck in the crossbar of the H, it appeared that OPC had flattened the contour of the

drawn diffusion width which thus reduced the parabolic shape from the off-sets and decreased the sensitivity to actual misalignment.

In hindsight, calibrating the monitors was quite challenging because no programmed treatments, such as defocus or dose, were available. More importantly, OPC greatly impacted and nearly cancelled the design principles. It would be quite interesting to use these same designs without OPC to observe their sensitivities. From this non-OPC designs for production wafers could be produced which are compatible with product. It would be very beneficial if the monitor designs would be allowed to selectively violate design rules without violating mask making rules, just as is allowed with OPC itself. Inherent trade-offs were found in looking for single dummy etch and lithography effects along the upper (forward) layout and lower (reverse) layouts for the signal paths within the ring oscillators. To test for coma effects, the dummy poly should always be on the same, e.g. left, side. But when the device drain is in the direction of signal flow, extra parasitic fringe capacitance to the dummy poly occurs on the section of the RO in which the signal is propagating left. Fortunately extraction captures this effect and provides a means of first order compensation but it is challenging to design monitors that can be accurate without parasitic capacitance compensation.

8.5 Monitoring of Nitride CESL and STI-Induced Stress

The measured RO frequency sensitivity to changes in CESL-induced stress was analyzed by characterizing the RO frequency versus Length of Diffusion (LOD) behavior for 8 RO monitors. By increasing the LOD from the minimum size by 1.8x on both the source and drain sides, the measured RO frequency sped up by 5.3%. Measurements showed that RO frequency saturated for large LOD values. At 6x the minimum LOD, a 13.9% increase in measured RO frequency was observed. In collaboration with Nuo Xu, a comparison with simulation shows that experiment rolls-off with LOD faster than predicted by mobility alone and a model for injection velocity effects is needed. This conclusion was further verified by the fact that the addition of ~70 nm of LOD to the minimum drain produced little effect, while adding ~70 nm to the source increased the RO frequency by 3%. A ~200nm change in lateral STI width induced a smaller 1.3% RO frequency decrease.

The RO frequency sensitivity to 3 STI-induced stress layout monitors was also examined. It was found that moving PMOS up caused a 3% increase in measured RO frequency, while moving NMOS downwards caused no change. This observation was consistent with published <100> channel orientation by ST Micro, indicating that the speed increase was likely due to the increase in the STI width surrounding the NMOS (induced by moving the PMOS) improving the NMOS performance.

In hindsight, in order to fully validate a new device model that include the combined effects of injection velocity and mobility, more intermediate levels in the range of 1.8-6x the minimum LOD are recommended. To facilitate further studies of RO frequency sensitivity to STI-induced stress, enlarging the distance between power-to-ground rails in layout is suggested, such that the monitors could be exposed to a larger change in STI width.

8.6 Random Noise Analysis using Least Mean Squares

To further explore random effects, the RO monitors were operated at different operating voltages and temperatures in order to help separate sources of variation that can be attributed to gate length (ΔL), gate oxide thickness (ΔT_{ox}), and channel doping (ΔN_{ch}). The sensitivities to

these 3 parameters were evaluated under 5 measurement conditions in the 0.8 to 1.1V and 20^{O}C to 90^{O}C range using 45nm ST Micro PDK models. All sensitivities were found to increase along with increasing V_{dd}, with the sensitivity to L increasing the most and sensitivity to N_{ch} increasing the least in absolute value. These sensitivities were used to solve for ΔL, ΔT_{ox}, and ΔN_{ch} of each RO instantiation using a least mean square approach. For 40nm, these values were $\pm$1%, $\pm$1%, and $\pm$6%, respectively. For 60nm, variation in ΔN_{ch} decreased by 20% which is consistent with the sqrt(1.5) increase in gate length. The effect of ΔL like that of T_{ox} is small. These small values may be representative of the accuracy of the technique rather than actual physical effects.

The systematic decrease in measured RO frequency observed in RO block 3 for outlier chips 4 and 9 was examined in regard to etch, lithography, and stress effects. These chips show a mean measured RO frequency reduction of 5-8% in block 3, compared to that in other blocks. Since this frequency reduction was not observed on 15 of the 17 die it is not a systematic within-die effect. The results show that etch, lithography, and stress monitors are all affected in the same manner as that of their respective control ROs, indicating that it is likely that these factors contributed to this frequency decrease. Since block 3 is alone on the left side of the chip, it is surmised that it might be near the edge of the wafer where thin-films of device and lithography materials or even CMP of interconnect may be non-uniform.

In hindsight, to further improve the noise analysis methodology, 7-9 measurement conditions over a larger range of parameters are recommended. If it may be assumed that the oxide is uniform, voltage only measurements in the range 0.8 to 1.5V might be used in conjunction with devices with a 2x range of device widths such as 40 nm to 80 nm. This is because the sensitivity of the ring oscillator frequency to L decreases with $1/L^2$ while that for N_{ch} decreases as 1/sqrt(L). Moreover, it may be interesting to correlate the standard deviation of the channel doping variation found from this sensitivity analysis with the inverse square root of the gate area.

8.7 Overall Perspective

The simulated and measured RO frequency sensitivities from parameter-specific RO monitors are summarized in Table 8.1. Gate-etch monitors showed a 1.2% RO frequency speed up when one or both dummy polys were removed after correction for changes in parasitic capacitances. Gate-focus monitors were designed to show a 1.5x sensitivity improvement compared to that of a control RO. The lack of pre-programmed focus treatments prevented effective calibration of the focus monitors. Moreover, the PV bands from ST Micro showed that there was little defocus in the stepper system, and measured RO frequency for 17 chips revealed that the across-wafer variability was only 1/6-1/4^{th} of the SS-FF SPICE-simulated guard bands, indicating that the across-wafer variability was well-controlled. Thus, the level of defocus may be too small to detect. For gate-to-active misalignment monitors, overlay errors were deduced to be 2-4 nm. The unanticipated requirement to apply strong OPC with scatter-bars decreased the sensitivity of the change in measured RO frequency by half compared to the design intent. The, CESL-induced strain monitors showed a speed up of 5.3% and 13.9% for a 1.8x increase and 6x increase in LOD, respectively. The lateral change in STI width produced a 1.3% frequency slow down. The STI-induced stress monitors showed a 3% RO frequency increase when the NMOS devices were exposed to more change in vertical STI surround.

Process Parameter	Simulation	Measured	Reasons for Discrepancy	Improvement Ideas
Gate Etch	N/A	Δf_{RO}=1.2%		Dummy on source side only
Gate Focus	Δf_{RO}=1.5x	No defocus data available. Monitor designs are OPC-ed, and the level of defocus may be too small to calibrate.	Diluted by OPC	F/E matrix calibration, no OPC, allow mask making design rules, use 90^O phase shift mask
Gate-to-Active misalignment	Δf_{RO}=3.2%	Δf_{RO}=1.5% $\Delta x_{overlay}$ = 2-4 nm	Diluted by OPC	No OPC, wedge-shaped monitors revised, add vertical misalignment monitors
CESL-Induced Strain	N/A	$\Delta f_{RO,\ 1.8x}$ = 5.3% $\Delta f_{RO,\ 6x}$ = 13.9%		~10 intermediate levels of LOD = 1.8x to 6x than that of minimum sized LOD, more monitors to examine lateral STI width and source/drain asymmetries
STI-Induced Stress	N/A	$\Delta f_{RO,\ horizontal}$=1.3% $\Delta f_{RO,\ vertical}$=3%		Enlarge rail-to-rail distance in layout

Table 8.1: A summary of RO frequency sensitivity for parameter-specific monitors

Much has been learned that could improve the utilization of parameter-specific ring oscillators in electronic monitoring. It would be quite interesting to use these same designs without OPC and observe the results where a different focus treatment is applied. A wafer with a programmed dose would also be helpful in making comparisons with the simulated PV bands for dose and focus. Enhancing gates with 90^O phase shift features is a sure way to further improve focus sensitivity by another 2-fold, while also electronically detecting the direction of focus.

Additional interesting approaches could improve measurement specificity for devices and the self-compensation of performance of electrical circuits. Parameter-specific ring oscillator circuits could be designed to selectively measure NMOS or PMOS devices in an individual manner. This could be done by slowly resetting the devices in a few gate delays after the transition signal passes to produce an effect similar to a stadium wave. These ring oscillators could also be implemented as on-chip process monitors that provide real-time feedback to corrective algorithms and hardware solutions.

This work demonstrates that parameter-specific ring oscillators are suitable for multiple critical applications in quantifying systematic and random effects in the co-optimization of process development and circuit design. Potential applications include developing device models for the role of stress in velocity injection and mobility for capping layer and STI stressors,

determining process variations in early process development, calibrating process models such as OPC and etch, obtaining distributions of residual process parameters, providing quantitative guidelines for designers, improving process-aware design tools, and monitoring mature processes. Hopefully, the both methodologies as well as the insights from the experimental results developed in this dissertation will be useful in putting this technology in practice.

Bibliography

[1] L. T. Pang, "Measurement and Analysis of CMOS Circuit Variability," Ph.D. Dissertation, University of California, Berkley, 2008.

[2] M. Kanno, A. Shibuya, M. Matsumura, K. Tamura, H. Tsuno, S. Mori, Y. Fukuzaki, T. Gocho, H. Ansai, and N. Nagashima, "Empirical Characteristics and Extraction of Overall Variation for 65-nm MOSFETs and Beyond," Symposium of VLSI Circuits Tech. Dig., pp. 88-89, 2007.

[3] V. Josh, K. Agarwal, D. Sylvester, D. Blaauw, "Analyzing Electrical Effects of RTA-driven Local Anneal Temperature Variation," Proc. Design Automation Conference (ASP-DAC), 2010 15th Asia and South Pacific, pp. 739-744, 2010.

[4] D. Boning, J. Panganiban, K. Gonzalez-Valentin, C. MdCowell, A. Gattiker, and F. Liu, "Test structures for delay variability," ACM/IEEE TAU Workshop, 2002.

[5] W. J. Poppe, "Collaborative Platform for DFM," Ph.D. Dissertation, University of California, Berkeley, 2007.

[6] S. O. Toh, B. Nikolic, and Z. Guo, "Dynamic SRAM Stability Characterization in 45nm CMOS," Seng Oon Toh, Borivoje Nikolic, Zheng Guo, Symposium VLSI Circuits Tech. Dig., 2010.

[7] N. Xu, "Physically Based Modeling of Stress-Induced Performance Variations for Nanoscale MOSFETs," M.S. Thesis, University of California, Berkeley, 2010.

[8] H. J. Levinson, "Principles of Lithography," SPIE Press, Bellingham, Washington, USA, 2001.

[9] P. Oldiges, Q. Lin, K. Petrillo, M. Sanchez, and M. H. M. Ieong, "Modeling Line Edge Roughness Effects in sub 100 Nanometer Gate Length Devices," Proc. SISPAD, pp. 131–134, 2002.

[10] A. Asenov, S. Kaya, and J. H. Davies, "Intrinsic Threshold Voltage Fluctuations in Nano-MOSFETs due to Local Oxide Thickness Variations," IEEE Trans. Electron Devices, Vol. 49, No. 1, pp. 112-119, 2002.

[11] D. J. Frank, Y. Taur, M. Ieong, and H.-S. P.Wong, "Monte Carlo Modeling of Threshold Variation due to Dopant Fluctuations," Symposium VLSI Circuits Tech. Dig., pp. 171–172, 1999.

[12] M. J. M. Pelgrom, A. C. J. Duinnaijer, and A. P. G.Welbers, "Matching Properties of MOS Transistors," IEEE J. Solid-State Circuits, Vol. 24, No.5, pp. 1433–1440, Oct. 1989.

[13] K. Qian, C. Spanos, and B. Nikolic, "Hierarchical Modeling of Spatial Variability with a 45nm Example," Proc. SPIE 7275, pp. 727505-1- 727505-12, 2009.

[14] S. O. Toh, Y. Tsukamoto, Z. Guo, L. Jones, T.-J. K. Liu, and B. Nikolic, "Impact of Random Telegraph Signals on Vmin in 45nm SRAM," International Electron Device Meeting Tech. Dig., 2009.

[15] N. Tega, H. Miki, F. Pagette, D. J. Frank, A. Ray, M. J. Rooks, W. Haensch, and K. Torii, "Increasing Threshold Voltage Variation due to Random Telegraphic Noise in FETs as Gate Lengths Scale to 20nm, Symposium on VLSI Technology Digest of Technical Papers, pp. 50-51, 2009.

[16] Y. Tsukamoto, S. O. Toh, C. Shin, A. Mairena, T.-J. K. Liu, and B. Nikolić, "Analysis of the Relationship between Random Telegraph Signal and Negative Bias Temperature Instability," IRPS, 2010.

[17] L. Wilson, “International Technology Roadmap for Semiconductors” <http://www.itrs.net/>, Web., 2009
[18] S. Renwick, “Progress in Lithography: Krf lithography through EUV,” ICCAD Short Course, 2008.
[19] J. W. Bossung, “Projection Printing Characterization,” Proc. SPIE, Vol. 100, pp. 80-84, 1977.
[20] A. K.-K. Wong, “Resolution Enhancement Techniques in Optical Lithography,” Bellingham, WA, USA, SPIE Press, 2001.
[21] Nikon, <http://www.nikon.com/index.htm>, Web, 2010.
[22] D. Steele, A. Coniglio, C. Tang, and B. Singh, “Characterizing Post Exposure Bake Processing for Transient and Steady State Conditions, in the Context of Critical Dimension Control,” Proc. SPIE, Vol. 4689, pp. 517–530, 2002.
[23] K. Rim, “Mobility Enhancements,” International Electron Device Meeting Tech. Dig Short Course, 2007.
[24] Q. Zhang, C. Tang, T. Hsieh, N. Maccrae, B. Singh, K. Poolla and C. J. Spanos, “Comprehensive CD Uniformity Control across Lithography and Etch”, Proc. SPIE, Vol. 5752, pp. 692, 2005.
[25] J. Allgair, G. Chen, S. Marples, D. Goodstein, J. Miller, F. Santos, “Feature Integrity Monitoring for Process Control using a CD SEM.” Proc. SPIE, Vol. 3998, pp. 227-31, 2002.
[26] C. Tabery, M. Craig, G. Burbach, B. Wagner, S. McGowan, P. Etter, S. Roling, C. Haidinyak, and E. Ehrichs, “Process Window and Device Variations Evaluation Using Array-Based Characterization Circuits.” International Symposium on Quality Electronic Design, pp. 6, 2006.
[27] M. Bhushan, A. Gattiker, M. B. Ketchen, K. and K. Das, “Ring Oscillators for CMOS Process Tuning and Variability Control,” IEEE Transaction on Semiconductor Manufacturing, Vol. 19, No.1, 2006.
[28] A.H. Gabor, et al, “Improving the Power-Performance of Multicore Processors Through Optimization of Lithography and Thermal Processing”, Proc. SPIE, Vol. 6521, pp. 6521-18, 2007.
[29] A. Bassi, A. Veggetti, L. Croce, and A. Bogliolo, “Measuring the Effects of Process Variations on Circuit performance by Means of Digitally-Controllable Ring Oscillators,” International Conference on Microelectronic Test Structures, pp. 214–217, 2003.
[30] B. Zhou and A. Khouas, “Measurement of Delay Mismatch due to Process Variations by Means of Modified Rng Oscillators,” IEEE International Symposium on Circuits and Systems, Vol. 5, pp. 5246–5249, 2005.
[31] K. Agarwal and S. Nassif, “Characterizing Process Variation in Nanometer CMOS,” Proc. Design Automation Conference, pp. 396–399, 2007.
[32] S. Ohkawa, M. Aoki, and H. Masuda, “Analysis and Characterization of Device Variations in an LSI chip using an Integrated Device Matrix Array,” IEEE Transaction on Semiconductor Manufacturing, Vol. 17, No. 2, pp. 155–165, 2004.
[33] F. Gennari, “Linking TCAD and EDA through Pattern Matching,” Ph.D. Thesis, University of California, Berkeley, 2004.
[34] A. R. Neureuther and F. Gennari, “No-Fault Assurance: Linking Fast Process CAD and EDA,” Proc. SPIE, Vol. 4889, pp. 871-878, 2002.

[35] G. Robins, A. R. Neureuther, K. Adam, "Measuring Optical Image Aberrations with Pattern and probe Based Targets," J. Vac. Sci. Technol. B., Vol. 20, No.1, pp. 338-343, 2002.
[36] G. McIntyre, J. Holwill, A. R. Neureuther, L. Capodieci, Y. Zou, H. Levinson, and J. Kye, "Screening Layouts for High-NA Polarization Effects using Pattern Matching," J. Vac. Sci. Technol. B., Vol. 23, No.6, pp. 2646 – 2652, 2005.
[37] J. Holwill, G. McIntyre, W. Poppe, and A. Neureuther, "Layout 'Hot Spots' for Advancing Optical Technologies, " Proc. SPIE 6154, pp. 61543M, 2006.
[38] L. T.-N. Wang, A. Yeh, L. Kem, and A. R. Neureuther, "Illustration of Illumination Effects on Proximity, Focus Spillover, and Design Rules," Proc. SPIE 7275, pp. 72750B, 2009.
[39] M. Miller, K. Yamazoe, A. R. Neureuther, "Extensions of Boundary Layer Modeling of Photomask Topography Effects to fast-CAD using Pattern Matching," Proc. SPIE 7488, pp. 74883H , 2009.
[40] G. Robins, A. R. Neureuther, K. Adam, "Measuring Optical Image Aberrations with Pattern and probe Based Targets," J. Vac. Sci. Technol. B., Vol. 20, No.1, pp. 338-343, 2002.
[41] I. Ahsan, O. Glushchenkov, R. Logan, E. J. Nowak, H. Kimura, G. Berg, J. Herman, E. Maciejewski, A. Chan, A. Azuma, S. Deshpande, B. Dirahoui, G. Freeman, A. Gabor, M. Gribelyuk, S. Huang, M. Kumar, K. Miyamoto, D. Mocuta, A. Mahorowala, E. Leobandung, and H. Utomo, "RTA-driven intra-die variations in stage delay, and parametric sensitivities for 65nm technology," Symposium VLSI Technology Tech. Dig., pp. 170–171, 2006.
[42] H.-Y. Liu, L. Karklin, Y.-T. Wang, Y. C. Pati, "Application of Alternating Phase-Shifting Masks to 140-nm Gate Patterning: Linewidth Control Improvements and Design Optimization," Proc. SPIE, Vol. 3236, pp. 328, 1997.
[43] G. Robins, "Interferometric Pattern and Probe-Based Aberration Monitors," Ph.D. Dissertation, University of California, Berkeley, 2005.
[44] J. Rubinstein (Holwill), and A. Neureuther, "Self-Interferometric Electrical Image Monitors," Proc. SPIE 6152, pp. 615215, 2006.
[45] Y. Cao, "Predictive Technology Model," <http://ptm.asu.edu/>, Web. , 2008.
[46] L.-T. Pang and B. Nikolic, "Measurement and Analysis of Variability in 45nm Strained-Si CMOS Technology," Custom Integrated Circuits Conference Dig. Of Tech. Papers, pp. 129-132, 2008.
[47] A. Khakifirooz and D. A. Antoniadis, "Transistor Performance Scaling: The Role of Virtual Source Velocity and Its Mobility Dependence," International Electron Device Meeting Tech. Dig., pp. 667–670, 2006.
[48] T., Miyashita, T. Kubo, Y.S. Kim, M. Nishikawa, Y. Tamura, J. Mitani, M. Okuno, T. Tanaka, H. Suzuki, T. Sakata, T. Kodama, T. Itakura, N. Idani, T. Mori, Y. Sambonsugi, A. Shimizu, H. Kurata, and T. Futatsugi, "A study on Millisecond Annealing (MSA) Induced Layout Dependence for Flash lamp Annealing (FLA) and Laser Spike Annealing (LSA) in Multiple MSA Scheme with 45nm High-Performance Technology," International Electron Devices Meeting Tech. Dig., pp.1-4, 2009.
[49] Y. Ye, F. Liu, S. Nassif, and Y. Cao, "Statistical Modeling and Simulation of Threshold Variation under Dopant Fluctuations and Line-Edge Roughness," Proc. Design Automation Conference, pp. 900-905, 2008.

[50] S. Boyd, “EE263: Introduction to Linear Dynamical Systems,” Lecture 1, pp. 207, Stanford University, 2008.
[51] M. Born and E. Wolf, “Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light,” Cambridge University Press, 7 edition, pp. 472-476, October 1999.