Characterization and Analysis of Process Variability in Deeply-Scaled MOSFETs

by

Karthik Balakrishnan

Submitted to the Department of Electrical Engineering and Computer Science

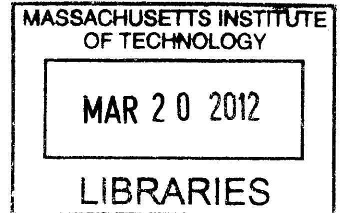
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2012


ARCHIVES
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
MAR 20 2012
LIBRARIES

© Massachusetts Institute of Technology 2012. All rights reserved.

Author ....
Department of Electrical Engineering and Computer Science
November 1, 2011

Certified by ....
Duane S. Boning
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ....
Leslie A. Kolodziejski
Chairman, Department Committee on Graduate Students

Characterization and Analysis of Process Variability in Deeply-Scaled MOSFETs

by

Karthik Balakrishnan

Submitted to the Department of Electrical Engineering and Computer Science
on November 1, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Variability characterization and analysis in advanced technologies are needed to ensure robust performance as well as improved process capability. This thesis presents a framework for device variability characterization and analysis. Test structure and test circuit design, identification of significant effects in design of experiments, and decomposition approaches to quantify variation and its sources are explored. Two examples of transistor variability characterization are discussed: contact plug resistance variation within the context of a transistor, and AC, or short time-scale, variation in transistors. Results show that, with careful test structure and circuit design and ample measurement data, interesting trends can be observed. Among these trends are (1) a distinct within-die spatial signature of contact plug resistance and (2) a picosecond-accuracy delay measurement on transistors which reveals the presence of excessive external parasitic gate resistance. Measurement results obtained from these test vehicles can aid in both the understanding of variations in the fabrication process and in efforts to model variations in transistor behavior.

Thesis Supervisor: Duane S. Boning
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I would first like to thank my research advisor, Prof. Duane Boning. Duane, you have provided me with excellent research guidance and direction throughout this doctoral program. I have learned a tremendous amount from you about not just technical research, but also how to overcome difficult challenges and be in a position to be successful.

Three other professors provided excellent support and guidance over the last few years and their invaluable input helped me tremendously to complete this research. Prof. Dimitri Antoniadis helped to pinpoint the cause of systematic variations in the contact resistance measurement data and also gave numerous suggestions for strengthening the foundation of the AC variability work. Prof. Luca Daniel was helpful in placing my work in the larger framework of a characterization-modeling-mitigation context and provided suggestions for test structure design that yielded meaningful results. Prof. Vladimir Stojanovic's expertise in integrated circuit design aided me to better analyze the AC variability measurement results and determine their significance towards ICs. To all of them, I am extremely grateful for their time and patience in helping me complete this Ph.D research and thesis.

I would like to thank many of the research staff members at IBM Research. The summer internship opportunity extended to me by Dr. Keith Jenkins, Dr. Vijay Narayanan and Dr. Mukesh Khare turned into an ongoing collaboration which eventually resulted in the AC variability characterization part of my thesis work. Keith and Vijay, you were both very helpful and I thank you for your support and guidance. I also thank Dr. Leland Chang, Dr. Paul Solomon, and Dr. Jae-Joon Kim of IBM Research for their technical support.

My family, of course, deserve my infinite gratitude for supporting me throughout. My mom, dad, and brother have all been very supportive of me and shown me lots of love and I could not have accomplished this without them. To my fiancée Lavanya — you have been extremely loving and caring from the time we met and I cannot thank you enough.

My friends have all been great to me and I have a huge amount of respect for them. To my best friend Cyrus, I can't describe how good of a friend you are so I won't try. Michael, Daihyun, and Nigel — your friendship throughout the years has been invaluable and you have all helped me in so many different ways. Thanks to all of you for everything.

I am grateful to all the past and present members of the Boning research group, with whom I've shared many conversations about endless topics, and who have also helped me in numerous ways.

I acknowledge the support of the Interconnect Focus Center (IFC), one of five research centers funded under the Focus Center Research Program, a Semiconductor Research Corporation Program.

Contents

1 Introduction **25**

1.1 Thesis Organization 31

2 Addressing Process Variation in Deeply-Scaled Technologies **33**

2.1 Characterization of Process Variation in Devices 34

2.1.1 Classification of Transistor Parameters 35

2.1.2 Challenges in Variability Characterization 36

2.1.3 Test Structures for Device Variability Characterization 37

2.1.4 Statistical Metrology to Enable Variation Analysis 40

2.2 Modeling Device Variation 42

2.2.1 Variation-Aware Modeling for Devices 42

2.2.2 Variation Modeling for Unit Processes 44

2.3 Mitigating Process Variation 45

2.3.1 Mitigation at the Process Level 45

2.3.2 Mitigation at the Device Level 46

2.4 Summary 47

3 Contact Plug Resistance Variability **49**

3.1 Contacts in a Device Context 50

3.2 Background Work 51

3.2.1 Individual Contact Measurement 52

3.2.2 Failure and Defect Analysis 54

3.2.3 Analytical Modeling 55

3.2.4 Arrayed Test Structures 55
3.3 Test Structure for Contact Plug Resistance Variability Characterization 56
3.3.1 Contact Plug Resistance Measurement Circuit 56
3.3.2 Simultaneous Contact and Device Measurement Circuit 57
3.3.3 Measurement Accuracy 60
3.3.4 Design of Experiments 60
3.4 Variation Decomposition Methodology 62
3.4.1 Spatial Correlation Computation 62
3.4.2 Decomposition of Variation Sources 63
3.4.3 Analysis of Variance (ANOVA) 65
3.5 Statistical Analysis Results 65
3.5.1 Overall Trends 66
3.5.2 Die-to-Die Trends 67
3.5.3 Within-Die Systematic Layout-Dependent Trends 69
3.5.4 Within-Die Systematic Position-Dependent Trends 75
3.5.5 Random Spatially Uncorrelated Variation 77
3.5.6 Spatial Correlation Via Sparse Regression 79
3.6 Simultaneous Bank Measurement Results 80
3.7 Need for Variability Models 83
3.8 Summary 85

4 Array-Based Test Structure for AC Variability Characterization 87

4.1 Array-Based Test Circuit 89
4.1.1 DUT Array 90
4.1.2 Design Optimization for AC Variability Measurement 94
4.1.3 Signal Propagation 96
4.1.4 Delay Measurement Circuit 98
4.1.5 Measurement Setup and Methodology 100
4.1.6 Measurement Accuracy 101
4.1.7 Test Chip 102

4.2 Summary 102

5 Ring Oscillator-Based Test Structure **105**

5.1 Introduction 105

5.2 Transistor Propagation Delay 106

5.3 Test Circuit Description 108

5.4 Delay Measurement Circuit 111

5.5 Test Circuit for Compensation of SOI Variations 112

5.6 Simulation Results 112

5.6.1 RO-Based Test Circuit Accuracy and Sensitivity 114

5.6.2 Sensitivity of t_{meas} in the Absence of AC Variations 117

5.7 Test Chip 119

5.8 Measurement Results 120

5.8.1 Off-Chip Measurement Accuracy 122

5.8.2 Single-Die Results 122

5.8.3 Wafer-Averaged Results 124

5.8.4 Multiple Operating Voltage Results 125

5.8.5 Potential Circuit Impact and Implications 127

5.9 Summary 130

6 Conclusions **131**

6.1 Contributions 131

6.2 Conclusions 133

6.3 Future Work 134

6.3.1 Characterization 135

6.3.2 Variation-Aware Modeling 135

6.3.3 Mitigation of Variation 136

List of Figures

1-1 Polysilicon CD window versus technology node in Intel's manufacturing process [1]. Shrinking upper and lower bounds on allowable critical dimensions present a significant challenge to transistor scaling. 26

1-2 Random dopant fluctuation [2], causing the number of dopants and their locations in the within the channel of a transistor to vary from transistor to transistor. 27

1-3 Number of dopants decrease as a function of technology node, which means that random dopant fluctuations in advanced technology nodes cause increased deviations relative to the mean [2]. 27

1-4 Constant standard deviation with scaling and linear relationship between input parameter and output performance. An input parameter which has these characteristics does not pose a variation or yield-related challenge to technology scaling. 28

1-5 Constant relative standard deviation with scaling and linear relationship between input parameter and output performance. An input parameter which has these characteristics does not pose a variation or yield-related challenge to technology scaling. 29

1-6 Constant standard deviation with scaling and superlinear relationship between input parameter and output performance. An input parameter which has these characteristics poses significant variation and yield-related challenges to technology scaling because the relative variation in output performance increases as the technology scales. 29

1-7 Constant relative standard deviation with scaling and superlinear relationship between input parameter and output performance. An input parameter which has these characteristics poses variation and yield-related challenges to technology scaling because the relative variation in output performance increases as the technology scales. 30

2-1 Multi-pronged approach for addressing process variation in deeply-scaled technologies: modeling, characterization and mitigation. 34

2-2 Simulations showing the effect of polysilicon pattern density variations on spatial RTA temperature distribution [3]. Before optimization to obtain uniform polysilicon pattern density across the chip area, the simulated RTA temperature is significantly higher for regions of low polysilicon pattern density than for regions of high polysilicon pattern density. After optimization, the temperature gradient across the chip is reduced. 39

2-3 Transistor matching as it relates to device area [4]. The standard deviation of the difference in threshold voltage between two identically designed transistors is inversely proportional to the square root of the transistor area. Therefore, the scaling of transistors to smaller dimensions increases the threshold voltage mismatch between them. 43

2-4 Example of how DFM is used in Intel's SRAM cell design from 90nm to 45nm nodes [3]. The implementation of single-orientation polysilicon, relaxed pitch features, and a polysilicon endcap process which results in square polysilicon ends can be seen as the SRAM cell scales from the 90nm node to the 45nm node. 46

3-1 Contact in the context of a transistor and relevant extrinsic parasitic resistances [5]. The parasitic resistance components which most significantly involve the contact are $R_{CONTACT}$, $R_{SILICIDE}$, and $R_{INTERFACE}$. However, other choices, such as that to use elevated source-drain regions, can significantly impact the magnitude of these parasitics and others. 51

3-2 Resistor-grid model of metal-semiconductor contact [6]. R' is the resistivity of the doped silicon source/drain junction and G' is the resistivity of the metal-semiconductor contact interface. 53

3-3 Three-dimensional transistor view with path of current flow through contact to determine contact resistance of middle contact under test (yellow). The gate is switched off so no current flows from the source to the drain of the transistor itself. 57

3-4 Test circuit to measure contact plug resistances in an arrayed set of DUTs. Three transmission gate switches are used to control access to the V_{OUTL}, V_{OUTH}, and I_F. Off chip-analog-to-digital converters are used to sense the output voltages. 58

3-5 Test circuit to measure both contact plug resistances and transistor I-V characteristics for an arrayed set of DUTs. An additional transmission gate switch is used to control the gate voltage of the transistor DUT and off-chip operational amplifiers are used to force the source and drain voltages to their desired values. The device current is measured through the measurement of voltage across an off-chip resistor that is located in the current path of the DUT transistor. 59

3-6 Geometry-based variables in the DOE (half transistor shown for simplicity). Contact-to-gate distance, (d_{cg}), contact-to-diffusion edge distance (d_{cd}), and metallization layer to contact overlap for the y-dimension (d_o) are varied to determine any possible impact on contact plug resistance. 61

3-7 Design of experiments for contact resistance variability analysis. Values are chosen such that many DUT geometries exist at or near "nominal" case of $d_{cg} = 80nm$, $d_{cd} = 40$nm, and $d_o = 10$nm. 61

3-8 Spatial correlation analysis-based variation decomposition methodology. Spatially correlated contact plug resistance values can indicate the presence of a systematic trend, which can be subtracted from the measured data points to obtain a residual resistance map, for which the same analysis can be performed until there is no significant spatial correlation detected. 64

3-9 ANOVA results on wafer-level measurement data of contact plug resistance. The largest sum of squares terms are those coming from the source-drain width parameter, the die parameter, and the error term for unexplained variance. The sum of squares of the three interaction terms are much smaller in comparison. 65

3-10 Contribution of each variation source to total variation in contact plug resistance. More than half of the total variance can be attributed to die-to-die variations, while over 25% of the variance comes from the layout-dependent systematic component. Random within-die variation represents roughly 15% of the total variance. 66

3-11 Distribution of measured contact plug resistances across one die. . . . 67

3-12 Normal probability plot of contact plug resistance measurements over one die. 68

3-13 Distribution of measured contact plug resistances over the entire wafer, which has a mean of 14.36Ω and a standard deviation of 0.92Ω. . . . 68

3-14 Normal probability plot of contact plug resistance measurements over the entire wafer show that the distribution is not Gaussian. In this case, this is due to the presence of various systematic effects due to various factors. 69

3-15 Wafer map of average die contact plug resistance for 43 measured die. Given the observed data, no statistically significant wafer-level trend is observed. Some outlier die are located at the corners of the wafer. . 70

3-16 A plot of contact plug resistance resistance mean and standard deviation versus d_{cg} shows an increase in mean contact plug resistance for those contacts which are located further away from the polysilicon gate of the transistor. However, the standard deviation of the measured resistance does not change as a function of d_{cg}. 71

3-17 A plot of resistance mean and standard deviation versus d_{cd} shows an increase in mean contact plug resistance for those contacts which are located further away from the edge of the diffusion region of the transistor. However, the standard deviation of the measured resistance does not change as a function of d_{cd}. 71

3-18 A 3D structure closely replicating the test structure design is used for device simulations. Determining the current flow and electrostatic potentials at the surface of the silicide region can help to understand systematic trends in the measurement data. 72

3-19 A contour plot of electrostatic potential at silicon surface of a narrow diffusion region shows that current crowding occurs near the top of contact B, resulting in some difference in average electrostatic potential between contacts B and C. 73

3-20 A contour plot of electrostatic potential at silicon surface of a wide diffusion region shows that less current crowding occurs because of the large amount of diffusion area through which current can flow. In this case, the difference in average electrostatic potential between contacts B and C changes from its value in the case of a narrow diffusion region. 74

3-21 A plot of contact plug resistance as a function of source-drain width shows that the average measured plug resistance increases with larger source-drain widths. For very large widths, the average resistance approaches an asymptotic value. 75

3-22 Residual resistance as a function of column (x-location) shows two distinct regions of plug resistance. Contact plugs located at $x > 1210\mu m$ have an average resistance which is 1.3% lower than those located at $x < 1210\mu m$. 76

3-23 (a) Spatial correlation analysis computed with systematic die-to-die effects included, (b) Spatial correlation analysis computed with systematic die-to-die effects removed, (c) Spatial correlation analysis computed for single die located at wafer edge 78

3-24 Spatial maps of contact plug resistance for both a normal die and die located at edge are shown. The outlier die located at the edge of the wafer has an additional systematic spatial trend. 80

3-25 DCT coefficients from applying S-OMP-based algorithm on raw measurement data reveal the periodic patterns present in the DUT array due to the repeating order of DUT types in the layout. 81

3-26 A die-level map showing systematic layout-dependent trends in contact plug resistance, created from the extracted DCT coefficients from the measurement data, matches the distribution map of DUT types across the chip. 81

3-27 A scatter plot of normalized contact plug resistance versus normalized transistor current, measured at $V_{gs} = 1.0V$ and $V_{ds} = 0.2V$. A positive correlation of 0.33 exists between the two variables. 82

3-28 Correlation coefficients between measured device currents at various operating points and measured contact plug resistance. Correlations are strongest in the linear region of operation (low values of V_{gs} and high values of V_{ds}). 83

3-29 A plot of device current as a function of w_{sd} demonstrates that, while I_d is shown to be correlated with the contact plug resistance, the cause is due to unintentional stress which is a function of the distance from the gate to the STI edge. 84

4-1 Some AC-relevant parasitics in a conventional MOSFET. The characteristics of such parameters are difficult to capture by performing DC measurements on the transistor, and therefore other characterization techniques which involve transients or high frequency operation are necessary. 88

4-2 A proposed test circuit design approach which measures delay variation among multiple transistors, but for which the delay is primarily due to targeted AC variation sources rather than all sources including DC sources such as threshold voltage and channel length. 89

4-3 Array-based test circuit schematic consisting of a clock source, an array of DUTs, and a delay detector. The relative delay mismatches through all DUTs in the array are measured by comparing the arrival time of node B, the DUT output, with the arrival time of node C, a common reference. 90

4-4 Schematic of a transmission gate DUT array. In this case, both the DUT select enable device and the DUT are the same transmission gate. 91

4-5 Schematic of an NMOS DUT array. The input clock has access to the gate of one of the NMOS DUTs, controlled by the DUT select input and the transmission gates, and the output node is connected to a weak PMOS pull-up transistor to enable the output to swing high. 91

4-6 Schematic of a PMOS DUT array. The input clock has access to the gate of one of the PMOS DUTs, controlled by the DUT select input and the transmission gates, and the output node is connected to a weak NMOS pull-down transistor to enable the output to swing low. 92

4-7 Tradeoff involving number of DUTs in array versus AC variability captured. A large number of DUTs results in a large load capacitance, which makes the overall transition at the drain dominated by DC variation sources. On the other hand, a small load capacitance results in a DUT delay which is too small and whose variability can be overwhelmed by external variation sources. 93

4-8 DUT array optimization for AC variability characterization shows that 128 DUTs ensures that 98% of the variance in delay is attributable to AC variation sources. 95

4-9 Simulated DUT delay distributions for different array sizes and variability sources. In the case of 8 DUTs, the distribution of delays when AC variation sources are imposed differs significantly from that when only DC variation sources are imposed. However, in the case of 1024 DUTs, the distributions are more similar to each other. 96

4-10 Variation in relative delay as a function of number of DUTs - DC variations imposed versus all variations imposed. The point at which the distributions deviate from one another is qualitatively marked in red. 97

4-11 A buffered H-tree for input signal propagation into transmission gate array. 98

4-12 Delay measurement technique using a logic gate followed by a first-order low-pass RC filter (NAND can be replaced with NOR depending on whether the falling or rising edge needs to be characterized). . . . 99

4-13 Waveforms describing the operation of the delay measurement technique. Nodes B and C are the inputs to a NAND gate, whose output is shown in D. The low-pass filter then produces an average DC voltage, V_{DC}. 100

4-14 Limitations on accuracy of delay measurement technique. For up to 30ps of relative delay mismatch, the error in measurement is bounded by 2ps. 101

4-15 The array-based test circuit layout is divided into three blocks: PMOS DUT array, NMOS DUT array, and transmission gate DUT array. A scan chain is implemented in the vertical direction which controls DUT access for all blocks. 103

5-1 Propagation delay metrics which characterize the short time-scale behavior of a transistor. 107

5-2 Ring oscillator-based test circuit for AC variability characterization, which operates in two modes and requires two clock period measurements and a delay measurement in order to characterize the DUT. . . 108

5-3 Waveforms at PMOS DUT terminals during *pass* mode, in which both transitions at the drain of the DUT are triggered by transitions at the source of the DUT. 109

5-4 Waveforms at PMOS DUT terminals during *wait* mode, in which one transition at the drain of the DUT is triggered by a transition at the source of the DUT, while the other transition at the drain of the DUT is triggered by a transition at the gate of the DUT. 110

5-5 Array of RO blocks for statistical characterization of DUT AC performance. Both the ring oscillator period and delay measurement pins are shared outputs, while each RO is accessed through an enable signal controlled by a scan chain. 110

5-6 Delay measurement using a logic gate and RC filter, which converts the delay between two signals into a pulse whose duty cycle is proportional to the delay, and the converts the pulse into a DC voltage whose value is also proportional to the delay. 111

5-7 Modification of *pass* mode operation to minimize variations due to SOI history effect difference between modes. The XOR gate before the *pass* switch only affects the *pass* mode of operation, while leaving the *wait* mode of operation unchanged. 113

5-8 Waveform at PMOS DUT terminals during *pass* mode in SOI history effect-compensated test circuit. The average duty cycle of V_{gs} is similar in both the *pass* and *wait* modes of operation by creating periods during which time the DUT gate is switched off when it does not affect the propagation of the source signal to the drain. 113

5-9 Ring oscillator waveforms for NMOS DUT type show how the transitions occur during the two modes of operation. 114

5-10 Ring oscillator waveforms for PMOS DUT type show how the transitions occur during the two modes of operation. 116

5-11 Quantile-quantile plot showing t_{meas} distributions under DC variations and all variations. The distribution of t_{meas} when the DUT is subject to all variation sources deviates from the case of only DC variation sources at a low standard deviation value. 116

5-12 Simulation results showing a plot of the distribution of t_{meas} when only certain DC variation sources are present. These results indicate that threshold voltage is the parameter to which the output parameter t_{meas} is most sensitive. 118

5-13 A normal probability plot showing the simulated decomposition of different DC variation sources and how sensitive t_{meas} is to each of them. Threshold voltage variation is the DC source predominantly captured by the output parameter, t_{meas}. 118

5-14 Ring oscillator layout showing the device under test (DUT), inverter stages, a logic block, and the resistor used for the delay measurement block. 119

5-15 Test circuit layout which includes four ring oscillator blocks which characterize NMOS DUTs and PMOS DUTs in both standard and SOI-compensated configurations. 120

5-16 Schematic of identifier RO block, which includes an external gate resistor in series with the gate of the DUT. 121

5-17 Layout of identifier RO block, which includes a $30k\Omega$ polysilicon resistor connected to the gate of the DUT. 121

5-18 Measured output parameters for PMOS DUTs on a single chip show the values of the three measurement parameters for each DUT. 123

5-19 t_{meas} for PMOS DUTs on a single chip, calculated from the direct measurement results in Figure 5-18. Identifier DUTs which exhibit larger values of t_{meas} are clearly distinguishable from other data points. 123

5-20 t_{meas} for all DUTs averaged over 40 die on wafer shows the presence of identifier DUTs more clearly, in addition to some weak systematic trends due to power supply variations. 124

5-21 Schematic for PMOS DUT-based RO, showing all transistors and gates as well as their relative sizes. 125

5-22 Schematic for NMOS DUT-based RO, showing all transistors and gates as well as their relative sizes. 126

5-23 t_{meas} for multiple V_{DDL} values for PMOS DUT shows that the identifier is distinguishable at all voltages, but some voltage-dependent effects also change the measurement values relative to one another. 128

5-24 t_{meas} for multiple V_{DDL} values for NMOS DUT shows that the identifier is distinguishable at all voltages, but some voltage-dependent effects also change the measurement values relative to one another. 128

5-25 Simulation study of a 7-stage ring oscillator frequency variation due to AC variation sources equal to that measured from the test chip, scaled to a 32nm technology node. Results indicate that the frequency has a $\frac{\sigma}{\mu} = 1.6\%$. 129

List of Tables

2.1 Classification of device-related parameters to enable the understanding of challenges involved in device variability characterization. 35

3.1 Layout design parameter values chosen for the DOE: 4 factors and 55 DUT types representing a subset of all possible combinations of these 4 factors. 62

5.1 Transistor and gate parameters for ring oscillator-based test circuit. . 127

Chapter 1

Introduction

In 1965, Gordon Moore observed that every 18 months, the density of transistors on a die increased by a factor of two [7]. This observation, which has been rebranded "Moore's Law" by the microelectronics industry consumers, has propelled the microelectronics industry forward at an astonishing pace over the past 30 years. However, the challenges of integrating billions of transistors on a single die are becoming increasingly difficult to overcome. Fabricating two nominally identical transistors so that they behave identically is not possible due to imperfections and non-uniformities in the manufacturing process, also known as process variations. With smaller transistors and increased transistor density, the effect of process variations is more significant and meeting performance and yield specifications is increasingly challenging.

One example of process variations becoming more significant with scaling involves transistor gate length. Transistor gate length is a key parameter, along with gate pitch, that ultimately determines overall transistor density. For this reason, the minimum feature size able to be fabricated for a given process technology, which is used to create a transistor gate, is also known as the gate "critical dimension" (CD). A technology node is defined as the minimum half-pitch between two features that is printable for that given technology. The technology node therefore serves as a measure of achievable transistor density for a given technology. Shown in Figure 1-1 is the polysilicon CD target window for Intel at their different technology nodes to achieve yield and performance specifications for that node [1]. As the technology node be-

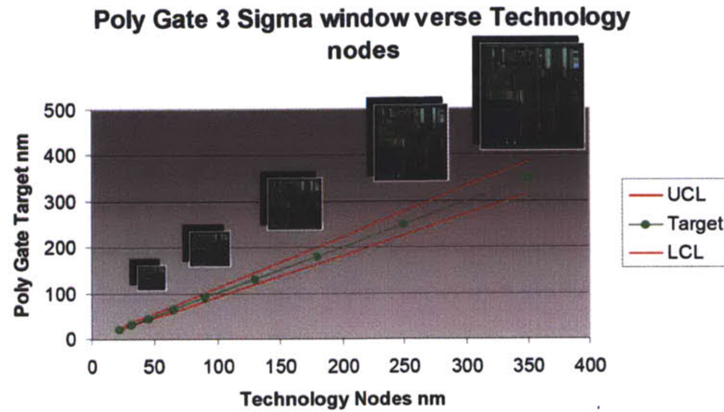


Figure 1-1: Polysilicon CD window versus technology node in Intel's manufacturing process [1]. Shrinking upper and lower bounds on allowable critical dimensions present a significant challenge to transistor scaling.

comes more (smaller number), the window of allowable polysilicon CD, represented by the red line upper and lower control limits, shrinks. For previous technology nodes such as 0.35μm, process variations which cause the CD to differ across multiple transistors is more tolerable because there is more margin for variations in CD. However, for advanced technology nodes such as 32nm and 22nm, the CD can only vary by a small amount, which is difficult to achieve in the presence of process variations. As an example, from Intel's 130nm to the 32nm technology node, both the scaling factor for the channel length and the bounds on the upper and lower control limits, have been the same (around 0.7) [8]. This indicates that the percentage tolerance limits on channel length as calculated from the nominal value are staying the same as technology scales.

Fabricating nominally identical transistors which behave identically has become more difficult due to scaling for other reasons as well. For example, the number of dopants in the transistor channel must be controlled to within certain boundaries to ensure performance and yield specifications are met. Figure 1-2 shows Intel's simulated distribution and location of dopants within the channel of a transistor [2]. Random dopant fluctuation (RDF) is a form of process variation resulting in variations in the number or location of dopant atoms implanted into the channel of each transistor. For previous technology nodes where the channel had a large area and

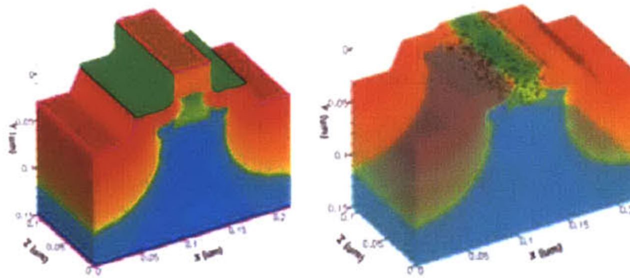


Figure 1-2: Random dopant fluctuation [2], causing the number of dopants and their locations in the within the channel of a transistor to vary from transistor to transistor.

the average number of dopants was large, the statistics of large numbers meant that the distribution of the number of dopants for multiple transistors was very tight with a small variance. In that case, RDF did not have a significant effect. However, when the transistor shrinks, the average number of dopants in the channel decreases, as shown in Figure 1-3 [2]. Consequently, there is less of an averaging effect when observing the number of dopants across multiple transistors. These increased deviations relative to the mean are larger, which in turn makes RDF more problematic. This is a difficult challenge because overcoming process variations in order to control the number or location of dopants in the transistor channel so precisely is difficult.

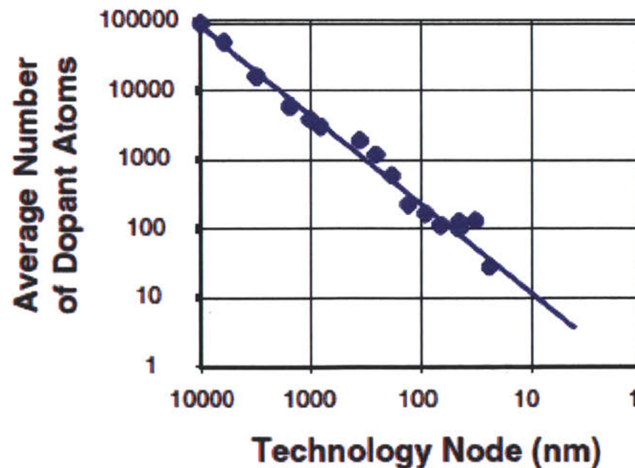


Figure 1-3: Number of dopants decrease as a function of technology node, which means that random dopant fluctuations in advanced technology nodes cause increased deviations relative to the mean [2].

With continued Moore's Law scaling, one can expect that process variations will play a more significant role in ultimately determining the performance and yield of a chip designed using a particular technology. To demonstrate this more clearly, it is useful to determine how both the absolute and relative variations of input parameters scale with technology, as well as the nature of the relationship between the input parameters and the performance and yield. First, the relative variation of a device parameter, represented by the quantity $\frac{\sigma}{\mu}$, may scale with technology node in multiple ways. It may decrease, increase, or remain constant with technology scaling. Second, the relationship between the device parameter and the performance metric of interest may be linear, sublinear, or superlinear. Four examples are shown for hypothetical input parameters and relationships to the output performances in Figures 1-4, 1-5, 1-6, and 1-7.

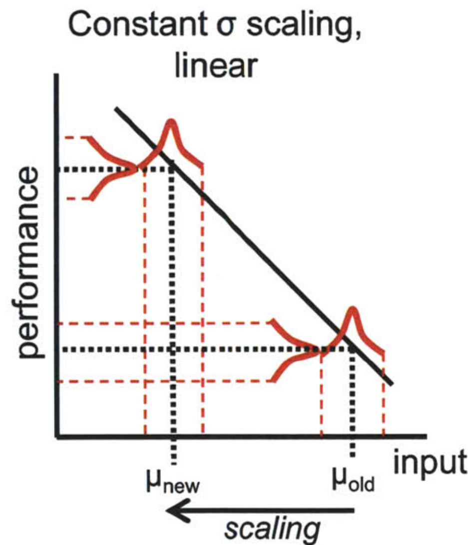


Figure 1-4: Constant standard deviation with scaling and linear relationship between input parameter and output performance. An input parameter which has these characteristics does not pose a variation or yield-related challenge to technology scaling.

Each device parameter which impacts performance, and whose variation impacts yield, can be categorized in this manner. For example, the scaling of channel length variation with technology can be characterized as having a constant relative standard deviation with respect to the mean channel length for a given technology. In addi-

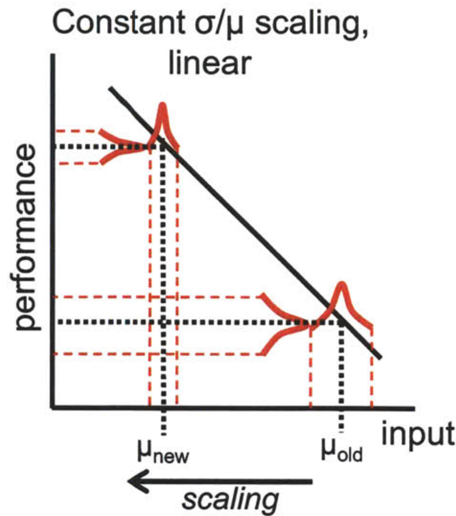


Figure 1-5: Constant relative standard deviation with scaling and linear relationship between input parameter and output performance. An input parameter which has these characteristics does not pose a variation or yield-related challenge to technology scaling.

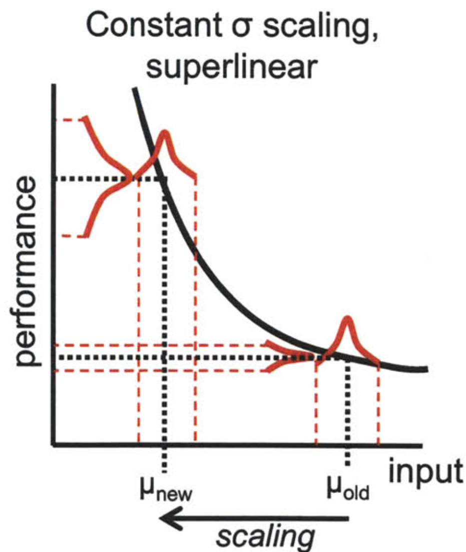


Figure 1-6: Constant standard deviation with scaling and superlinear relationship between input parameter and output performance. An input parameter which has these characteristics poses significant variation and yield-related challenges to technology scaling because the relative variation in output performance increases as the technology scales.

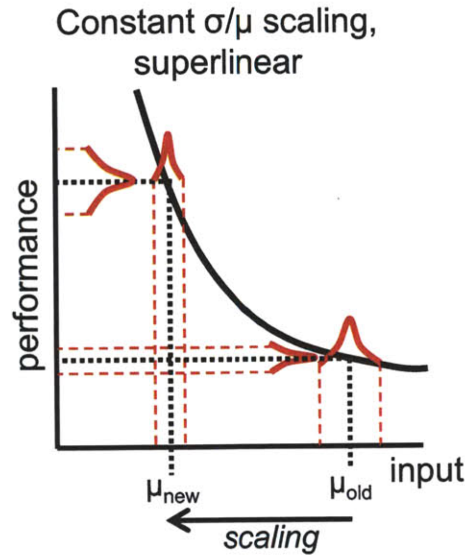


Figure 1-7: Constant relative standard deviation with scaling and superlinear relationship between input parameter and output performance. An input parameter which has these characteristics poses variation and yield-related challenges to technology scaling because the relative variation in output performance increases as the technology scales.

tion, the relationship between channel length and performance, or more specifically saturation current in this case, can be described as superlinear. Because the device saturation current depends inversely on the channel length of the device, a percent deviation from a smaller nominal channel length will result in a larger percent deviation in saturation current than that caused by the same percent deviation from a larger nominal channel length. Similarly, the number of dopants in the channel play a significant role in determining threshold voltage, and consequently, leakage current. The number of dopants scaled with technology in a constant standard deviation manner, but the relationship between threshold voltage and leakage current is exponential in nature. Therefore, it can be concluded that random dopant fluctuation will cause more variation with technology scaling. Considering these effects along with variations in other device parameters, it is apparent that process variations will play a more significant role in determining yield and performance as the technology continues to scale.

Another source of motivation is that new devices and process technologies are

being explored in order to continue Moore's Law scaling. These novel approaches are likely to be sensitive to process variations, both well-studied and new. With that serving as motivation, this thesis contributes test circuit-based methodologies to characterize such variations and statistical analysis tools to better understand them.

1.1 Thesis Organization

Challenges associated with addressing process variation are presented in Chapter 2. Previous work in the development of test structures to characterize transistor variations is described and a classification of transistor parameters for the purposes of variation-related analysis is introduced. Two ways of coping with variation, namely modeling and mitigation, are discussed. On the modeling front, methodologies to incorporate variation in existing component models as well as techniques for fast circuit simulation using such variation-aware models are described. Variation-based models for unit process steps in IC manufacturing are also discussed. On the mitigation front, existing techniques to reduce transistor variation in two different areas are presented: design for manufacturability (DFM) and process control.

Chapter 3 presents a test structure-based methodology for characterizing contact plug resistance. After motivating the need for such work, a test structure is presented along with a variation decomposition methodology. Then, silicon measurement results from a test chip are presented and various trends are described. The chapter concludes with a discussion on the need for variability-aware models for future technologies.

The need for the analysis of AC, or short time-scale, performance variations in transistors is motivated and an array-based test structure to characterize them is presented in Chapter 4. Such short time-scale variations in transient behaviors can be caused by variations in device geometries, parasitics, or other device parameters. A design-time optimization is used to make the test circuit sensitive to individual device AC variations, and simulation results are shown which illustrate the effectiveness of the measurement technique. Furthermore, the implementation and fabrication of a test chip are outlined along with details regarding the measurement setup and

methodology.

Then, Chapter 5 continues the discussion on AC variation analysis by introducing another test circuit to characterize the same. A ring oscillator-based test structure is introduced and simulations are shown which confirm the high sensitivity of the measurement technique to AC variation sources. Then, silicon measurement results from a test chip are presented and analyzed.

Finally, Chapter 6 concludes with a summary of this thesis and thoughts for future work in this area to address the challenges outlined earlier.

Chapter 2

Addressing Process Variation in Deeply-Scaled Technologies

Addressing the impact of process variation in deeply-scaled technologies requires a multi-pronged approach involving variability characterization, variation-aware modeling, and techniques for mitigation. Such an approach is necessary whether the issues involving process variation are tackled at the process, device, circuit, or system level. This discussion will focus on the challenges of addressing variation in devices, although a fair bit of overlap will inevitably exist with the process and circuit levels. A diagram of how such a multi-pronged approach might work is shown in Figure 2-1. The problem of addressing process variation in deeply-scaled transistors begins with variability characterization, which will be discussed in more detail in Section 2.1. The combination of test structure design and statistical metrology serves this function. Then, the results of the variability characterization and analysis can be fed into two different areas. One area is in variation modeling (Section 2.2). For device-level variation trends, the variations caused by different processing steps such as lithography, etch, oxide growth, ion implantation, annealing, and polishing can be modeled. In addition, the devices themselves lend themselves to variability-aware compact modeling, which works towards modeling variations in device performance due to variations in transistor parameters. Finally, the results of variability characterization of devices can also lead to techniques for mitigation. In the device and

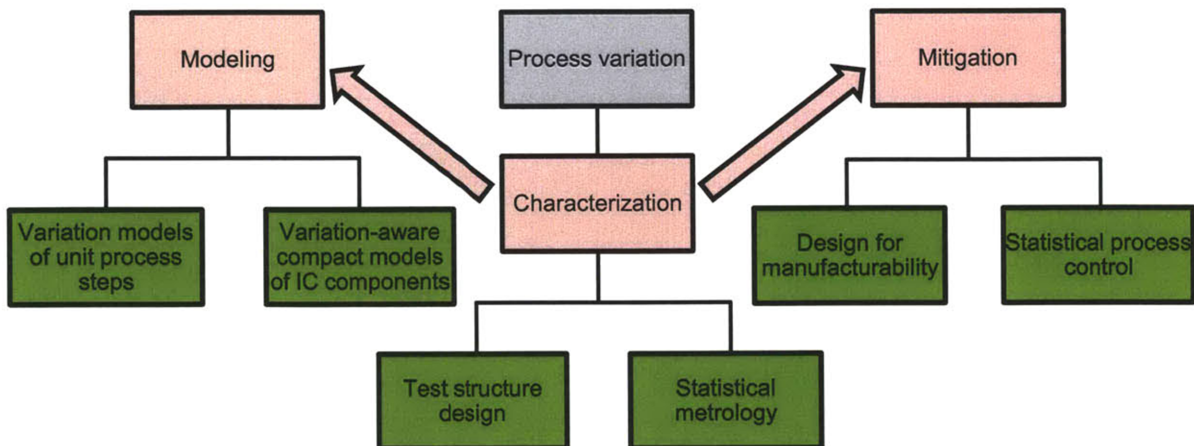


Figure 2-1: Multi-pronged approach for addressing process variation in deeply-scaled technologies: modeling, characterization and mitigation.

circuit spectrum, design for manufacturability, or DFM, is based on improving circuit performance and yield by employing design techniques informed by variability data. In the IC manufacturing process, statistical process control is used to improve yield by ensuring that steps in the manufacturing process will meet relevant specification bounds. This control is guided by results of variability characterization of the process steps through the use of test structures. A detailed discussion on mitigating variation will be presented in Section 2.3.

2.1 Characterization of Process Variation in Devices

The characterization of process variation in deeply-scaled devices involves two major components: test structure design/measurement and variability decomposition. Before delving into the realm of existing research in these areas, it is useful to classify transistor parameters into different groups to better understand the challenges of device variability characterization. The classification outlined in the following section also helps to understanding the compact modeling of variation in devices, as the characterization and modeling efforts are generally closely coupled.

2.1.1 Classification of Transistor Parameters

The characterization of a set of transistors and the determination of the distribution of one or more parameters requires an understanding of four mostly non-intersecting sets of transistor parameters: physical device parameters, device model parameters, device-measurable parameters, and geometry-based layout parameters. Some of these key parameters are shown in Table 2.1 and are discussed in more detail in Section 2.1.1.

Relevant Device-Related Parameters			
Physical Device	Device Model	Device-Measurable	Layout
channel doping	threshold voltage	saturation current	channel length
dopant locations	carrier mobility	drain leakage current	channel width
oxide thickness	intrinsic gate cap.	gate leakage current	source/drain areas
channel length	source/drain resistance	DIBL coefficient	well proximity
channel width	other parasitic RC	sub-threshold swing	distance to shapes
		cutoff frequency	pattern density
		unity gain frequency	

Table 2.1: Classification of device-related parameters to enable the understanding of challenges involved in device variability characterization.

Physical device parameters are physical (non-electrical or structural) parameters whose values are the direct result of the process steps involved during transistor fabrication. Channel doping concentration, N_A, locations of the dopant atoms, gate oxide thickness, t_{ox}, effective channel length, L_{eff}, and channel width, W, are some fundamental transistor parameters which may differ from transistor to transistor due to variations in the manufacturing process.

Device model parameters are those which are derived from physical device parameters and geometry-based layout parameters and used to model transistor behavior. Some of these parameters include threshold voltage, V_T, electron or hole mobility, μ, intrinsic gate capacitance, C_G, source-drain resistance, R_{sd}, and various parasitic resistances and capacitances associated with the extrinsic portion of the transistor.

Device-measurable parameters are those which can be relatively easily characterized by low or high-frequency electrical voltage or current-based measurements of a device. Some of these parameters include drain saturation current, $I_{D,sat}$, off-

state leakage current, I_{off}, drain induced barrier lowering (DIBL) coefficient, η, subthreshold swing, S_{s-th}, cutoff frequency, f_T, and unity gain power frequency, f_{max}. These measurements can usually be made using dedicated probe pads for each transistor during in-line testing.

Geometry-based layout parameters are those which can be manipulated at the integrated circuit-design level. While, for a particular transistor type, a circuit designer cannot change the gate oxide thickness or channel doping for each individual transistor, geometry-based layout parameters may be changed. Some of these layout parameters may include the drawn channel length, L_{drawn}, the channel width, W, the area of the source and drain regions, the number of contacts and their locations within the source and drain regions, the proximity of the transistor to a well, the separation distances to nearby polysilicon, active, and shallow trench isolation (STI) shapes, and effective pattern density.

2.1.2 Challenges in Variability Characterization

Two challenges in variability characterization and modeling stem from the previous discussion involving the classification of transistor parameters.

First, the relationship between device-measurable parameters and device model parameters is interdependent and correlated. As a result, building a test structure which determines the variances of a set of device-measurable parameters may not necessarily lead to the determination of the variances of a set of device model parameters. Careful test structure design and circuit simulation are often required to determine variances in model parameters with confidence.

Second, the relationship between the physical device structure parameters and the device model parameters is also interdependent and correlated by nature. Therefore, attributing a variation in a device-measurable parameter to variations in one or more physical device parameters requires a carefully planned design of experiments, adequate replication, and a sound variation decomposition methodology.

2.1.3 Test Structures for Device Variability Characterization

With the challenges in variability characterization now described, it is important to discuss test structures to characterize such device variability which have been developed over the years. While the characterization of individual transistors by using dedicated pads has its advantages in ease of design and measurement, statistical data necessary for variation analysis cannot be obtained without adequate replication. The variability characterization of deeply-scaled transistors can be performed in two ways. The first can be described as isolation-based characterization, in which an isolated parameter which has an impact on the overall transistor variation is characterized. Examples of such parameters are threshold voltage and channel length. The second is to obtain a more holistic or broad set of measurements on each of multiple transistors. This would include test structures that, for example, measure the I-V characteristics of multiple transistors in an array.

Isolation-based Test Structures

One important parameter in devices as they have scaled has been the threshold voltage (V_T). Because variability in device performance and leakage has increased substantially due to V_T variation as technology has scaled, a number of test structures have been developed to characterize it. For example, [9] uses a test structure comprised of an array of devices whose individual off-state leakage currents are measured by an on-chip integrating analog-to-digital converter. Then, device equations are used to obtain relative intrinsic threshold voltage values for each device. Another approach, presented in [10], focuses on monitoring V_{GS} for each transistor in an individually addressable array for a fixed current and then correlating that value back to the threshold voltage of the device.

Another important transistor parameter is its gate length, which can change across transistors due to variations in the lithography and etch processes. In order to determine the critical dimension (CD) variability for a given technology, a test structure was designed in [11] which measured the CD of multiple polysilicon lines through

electrical resistance measurements.

Another variation-related issue, particularly for analog circuit and memory applications, is matching between two or more identically designed devices. Therefore, a significant amount of work has been done to characterize the mismatch between transistors. For example, [12] discusses a test circuit that uses current mirrors to characterize transistor mismatch in the sub-threshold region of operation. More recently, a large addressable array of devices was characterized to assess the impact of different doses of implantation on the mismatch in both individual device threshold voltage and leakage current [13]. In this work, the leakage current of each device was measured while using techniques to cancel the off-state leakage of the other devices in the array which were not being measured.

More recently, studies have also been done to analyze the impact of other sources of variation. One such source is the rapid thermal annealing (RTA) process. Different pattern densities of polysilicon or shallow trench isolation (STI) may change the annealing temperature and therefore transistor properties. This is illustrated in simulation results performed by Intel, which shows that the annealing temperature during the RTA process varies across a die when dummy polysilicon is not used to create a uniform polysilicon density (Figure 2-2). However, a uniform pattern density makes the temperature profile across the die more uniform. To investigate the consequences of this effect, a test structure was designed to determine the impact of different pattern densities on doped poly-silicon sheet resistance, gate length, transistor currents, and ring oscillator frequencies [14]. Each structure was carefully designed to maximize the impact of potential RTA-induced variations by modulating the pattern densities accordingly. In addition, the impact of shallow trench isolation (STI) edge effects on transistor variability was also characterized in [15]. In this work, a mismatch sweep analysis technique is used, in which intentionally dissimilar pairs of transistors are laid out and measurement results are used to quantify the impact of STI-induced stress variations.

While most of the previously discussed structures focus on front-end-of-line (FEOL) variations, a significant amount of research has also been done on investigating back-

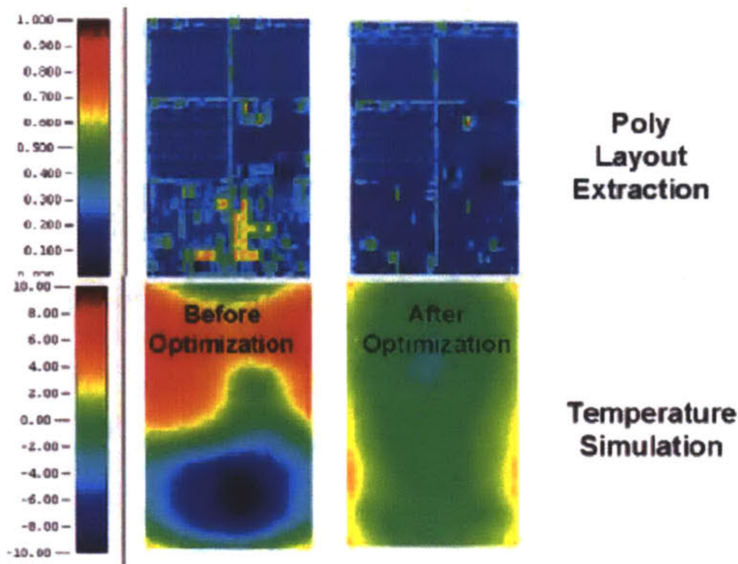


Figure 2-2: Simulations showing the effect of polysilicon pattern density variations on spatial RTA temperature distribution [3]. Before optimization to obtain uniform polysilicon pattern density across the chip area, the simulated RTA temperature is significantly higher for regions of low polysilicon pattern density than for regions of high polysilicon pattern density. After optimization, the temperature gradient across the chip is reduced.

end-of-line (BEOL) variations. One such example involves the building of a test structure to investigate variations in the chemical-mechanical polishing (CMP) process due to layout-based pattern-dependent effects such as feature area, density, and pitch [16].

Holistic-based Test Structures

A variety of test structures have been designed which obtain I-V characteristics of multiple transistors located in an array or a bank. Often, one primary objective of such a test structure is to quickly gather data for a large number of devices so as to enable in-line characterization. Such is the case in [17] and [18], where a scribe line test structure is built for product wafer monitoring which obtains I-V characteristics for transistors quickly by using parallel testing methods and pad multiplexing. Similarly, an integrating ADC-based approach for obtaining I-V characteristics of multiple transistors was employed in [19]. In addition, an array of devices which

included transistors, capacitors, resistors, and ring oscillators, was designed in [20] as a comprehensive test vehicle for technology characterization. Leakage-minimization and noise immunity techniques were employed in order to obtain high-accuracy current and voltage measurement for the various test blocks.

2.1.4 Statistical Metrology to Enable Variation Analysis

Once measurement data is obtained from any test vehicle, one primary objective is to use the measurement results to obtain information about the potential variation sources at work, their relative magnitudes, and relationships among them. To perform such a task generally requires the use of statistical tools to determine the relationships between input design parameters, such as transistor size, surrounding layout geometries, or layout pattern density, and measured output parameters, such as threshold voltage, channel length, saturation current, gate capacitance, or transistor delay.

In [21], the concept of statistical metrology as it applies to the semiconductor manufacturing process is introduced. Furthermore, the term "statistical metrology" is defined as "the body of methods for understanding variation in micro-fabricated structures, devices, and circuits." For the purposes of this discussion, we use the term as a way to express the methodologies by which statistical measurements are interpreted to obtain useful information. Reference [21] also motivates the need for statistical data analysis techniques to aid in increasing process yield by analyzing the case of interlayer dielectric thickness (ILD) variations due to variations in the chemical mechanical polishing (CMP) process. More recently, motivation for improved statistical analysis tools has come from the need to determine the root cause of large device variations which have a significant impact on yield and performance when the cause is difficult to predict due to the interactions of multiple process steps and parameters. A case study to determine the source of a large bipolar junction transistor leakage current variation in Motorola's manufacturing process is presented in [22]. In this study, a "blind" approach where the measurement data is analyzed in a pure statistical fashion without interpreting the meaning of any of the measurements was able to determine the root cause more quickly than the conventional design of experiments

(DOE) approach. Ideally, the use of a good DOE combined with good pure statistical techniques should serve to optimize process control and process optimization.

One such methodology which combines a DOE with statistical analysis techniques is used to analyze wafer-level, die-level, and wafer-die interaction components of variation in the CMP process by employing filtering, spline, and regression-based approaches as well as spatial Fourier transform methods [23]. Another use of statistical metrology is in analyzing the line edge roughness (LER) of critical dimension features. In [24], the line edge roughness of multiple features are characterized using scanning electron microscopy (SEM). The measurement results were then fitted to an analytical model for LER, considering the impact of correlations of edge roughness between both sides of the feature. Such an analysis can also be considered a spatial variation analysis, but at a much shorter length scale.

Lately, several efforts to analyze measurement results have been focused on the within-die spatial variation component. For example, I-V characteristics of multiple transistors in an array were used to fit channel length, threshold voltage, and mobility parameters in a BSIM4 model, as described in [25]. Then, spatial correlation analysis was performed to determine that, in the 65nm node, unlike that of a previously characterized 130nm technology, the spatial correlation of channel length was negligible. In addition, a mathematical construct to ensure the extraction of valid spatial correlation functions was presented in [26]. In this work, techniques to extract both a valid spatial-correlation function and a valid spatial-correlation matrix in the presence of measurement noise are presented. Finally, a technique for the extraction and modeling of non-stationary spatial variations is presented in [27]. Edge-detection algorithms for detecting sharp transitions in measurement data, methods for chip-partitioning into non-overlapping regions, and the development of a quantitative measure of stationarity are presented.

2.2 Modeling Device Variation

With the development of various test structures to measure process-induced variations in devices and methodologies to extract key statistical trends, it becomes important to model these device or structure variations for two main purposes. The first is to enable process control and process optimization, especially through the use of variation-based models of unit processes in the manufacturing line. The second is to enable the mitigation of process variation effects through techniques such as design for manufacturability (DFM). In addition, developing compact models for variation at the device level will enable the development of similar models at the circuit level, which then lends itself to circuit-level variation mitigation techniques. The discussion of variation modeling is divided into two parts: variation modeling for devices and variation modeling for unit processes.

2.2.1 Variation-Aware Modeling for Devices

Because the literature on variation-aware device modeling is vast and wide-ranging, a few examples which illustrate some of the main modeling techniques commonly employed will be discussed. The introduction of the Pelgrom model for transistor matching, motivated by the need for transistor matching for analog applications, has been a key enabler for more advanced variation models for devices. The Pelgrom model for transistor matching states that the standard deviation of the difference in threshold voltage of two nominally identical devices is inversely proportional to the square root of the transistor area, with a proportionality constant, A_{VT}. Shown in Figure 2-3 is a plot of $\sigma_{\Delta V_t}$ versus $\frac{1}{\sqrt{WL}}$ for a set of n-channel MOSFETs fabricated in a 0.18μm process. Analytical models, particularly those which help to understand threshold variation, have been instrumental in understanding how to design transistors with less variation and how to design circuits for manufacturability and yield. For example, the dependence of transistor threshold voltage on different channel-depth doping profiles was studied in [28] by analyzing and simulating continuous doping profiles. Studies on threshold voltage variation have also involved the "atomistic"

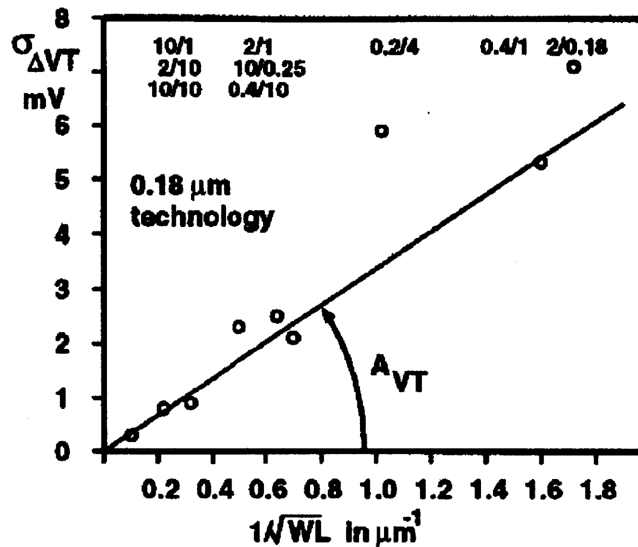


Figure 2-3: Transistor matching as it relates to device area [4]. The standard deviation of the difference in threshold voltage between two identically designed transistors is inversely proportional to the square root of the transistor area. Therefore, the scaling of transistors to smaller dimensions increases the threshold voltage mismatch between them.

simulation-based approach, in which a 3D atomistic simulation of a transistor is performed, and by solving Poisson's equation at each point in the 3D mesh [29]. Using this approach, it was determined that, not only the number of dopants, but also the position of the dopants within the channel, determine and affect the threshold voltage of a transistor. Bottom-up variability modeling approaches such as these, which generally do not require silicon measurements, are useful in calibrating and fitting compact device models such as BSIM and PSP to measurement data due to the additional insight which they can provide.

Variation models based on silicon measurement data, both for threshold voltage and for transistors as a whole, have been developed. For example, Intel published results of their measurement and modeling of threshold voltage variations in a 150nm high performance logic technology [30]. Intrinsic and extrinsic sources which contributed to threshold voltage variation were modeled, leading to the conclusion that some component of the variation depends on transistor width or length, although the majority of the variation is still due to random dopant fluctuation, which de-

pends on the device area. In [31] and [32], techniques such as principle component analysis (PCA) and the backward propagation of variance (BPV) are used in order to develop variability-aware compact models from silicon measurement data for the BSIM4 and PSP models, respectively. While these techniques are effective to some degree, the correlation among parameters in the compact models as well as the large number of parameters make it difficult to model variation in a stable and accurate fashion. Improving such variability extraction and fitting techniques therefore remains a challenge and continues to draw a great deal of attention. The challenges and possible future strategies for variability-aware compact modeling in both BSIM and PSP are outlined in [33]. One such challenge, among others, is the need to accommodate non-Gaussian distributions for model parameters in order to model the tails of distributions accurately.

Other variation-aware modeling approaches, such as those for the back-end-of-line (BEOL) process, have also been widely studied. For example, a capacitance solver which enables variation-aware extraction by using an incremental approach is described in [34]. The ability to quickly extract parasitic variations in the interconnect is greatly improved by the proposed floating random walk-based algorithm.

In addition, new sources of variation which may have an impact in advanced technologies have also been investigated. In high-K metal-gate technologies, the variation in the metal-gate work function caused by varying grain orientations contributes to threshold voltage variation. This effect is modeled at a device level in [35], and the implications of such variations on sub-threshold leakage current and SRAM performance and yield are demonstrated. As new concepts in device integration lead to new transistors, other variations will likely require similar efforts in device variation modeling.

2.2.2 Variation Modeling for Unit Processes

Variation modeling for processes in the semiconductor manufacturing process are important for maintaining process control and maximizing yield. While such models can also extend to the manufacturing tools themselves, analyzing effects such as drift over

time and tool-to-tool variations, the focus of this discussion is on the manufacturing process step itself. One example of a wide range of efforts to model variation in the manufacturing process is in that of chemical mechanical polishing (CMP). In CMP, either an oxide or metal is planarized across the wafer surface by removing the extra material by polishing the wafer with a pad. However, due to differences in pattern density across the wafer, the residual thickness of the patterned features may not all be the same. The case of interlayer dielectric (ILD) thickness variation is analyzed in [36] and a closed-form model for the variation is presented.

Lithography and the patterning of sub-wavelength features is also made more challenging by variation sources such as the exposure dose variations and focus variations. Consequently, it has become important to accurately model variations in the lithography process, particularly to better perform optical proximity correction (OPC) and employ other resolution enhancement techniques (RET). In [37], an analytical model for the variations in feature shapes due to defocus and dose variations is described. Furthermore, the model is used to develop a variation-aware OPC algorithm which can result in printed features that more closely represent the intentions of the design.

2.3 Mitigating Process Variation

Process variations can be mitigated at multiple levels in the design hierarchy. While numerous techniques for mitigating variation at the circuit level and the system level have been developed, the section will focus on the techniques for mitigating variation at the process level and the device level.

2.3.1 Mitigation at the Process Level

Statistical process control and feedback control is one way to mitigate variation at the process level. For example, a technique for using feedback control for a plasma etch process is described in [38]. In [39], the temperature post-exposure bake (PEB) process was optimized according to the density of features in the region. While taking into account these density variations, added variations caused later in the etching

process were also handled in the optimization scheme. This results in better CD uniformity across a wafer according to simulation and silicon measurement results.

2.3.2 Mitigation at the Device Level

The mitigation of process variation at the device level is closely coupled with circuit design techniques through the concept of design for manufacturability (DFM). One set of DFM techniques involve transistor layout methodologies which focus on mitigating process variation. For example, placing dummy polysilicon gates at the left and right sides of transistors is one way ensure that the local layout non-uniformities which can create different stress profiles for different transistors do not occur. More recently, other layout DFM approaches have included fixed-grid and single-orientation polysilicon and metal lines, multiple contacts and vias connecting different interconnect metallization layers, and the use of stricter overall design rules for spacing between features. An example of the use of some of these DFM rules used by Intel in their SRAM cell design is shown in Figure 2-4 [3]. In fact, this example illustrates

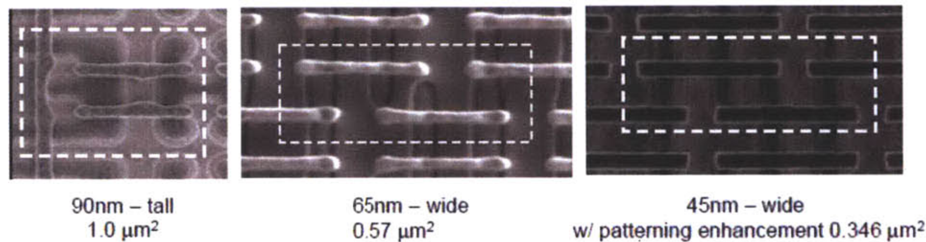


Figure 2-4: Example of how DFM is used in Intel's SRAM cell design from 90nm to 45nm nodes [3]. The implementation of single-orientation polysilicon, relaxed pitch features, and a polysilicon endcap process which results in square polysilicon ends can be seen as the SRAM cell scales from the 90nm node to the 45nm node.

both DFM-related process variation techniques, such as the use of single-orientation polysilicon and relaxed pitch features, and process-level mitigation techniques such as the use of a special polysilicon end-cap process which results in square ends rather than rounded ends. The implementation of such design rules is designed to mitigate the systematic variation caused by irregular geometries which change from transistor to transistor. DFM strategies can also be used to improve the number of yielding

die on a wafer for various product applications. For example, the authors in [40] use a design-time methodology in which yield-friendly layouts of IP cores are optimally mixed into the design, trading off area, performance, and power.

2.4 Summary

This chapter has outlined the challenges in addressing process variation in deeply-scaled semiconductor manufacturing technologies. The characterization, modeling, and mitigation of these variation sources have been well-documented in the literature thus far. However, there is an increasing need for new techniques and ideas in these areas as the variation problem becomes larger. The complexity of the manufacturing process and the engineering required to design a robust modern transistor further motivate the need for new ways to overcome variation for future technology generations.

Chapter 3

Contact Plug Resistance Variability

Variability characterization and modeling in advanced technologies are needed to ensure robust performance as well as improved process capability – and methods for the measurement, analysis, and mitigation of variation in devices, interconnects, and circuits are starting to emerge [41] [42]. Key elements of these "statistical metrology" methods include test structure and test circuit designs to gather the large amounts of data required; statistical analysis techniques, including identification of significant effects in spatial design of experiments, variation decomposition, and spatial correlation approaches to quantify variation and its sources; and finally, modeling to understand the impact and implications of variation on devices, circuits, and systems. In this chapter, we review and extend these approaches for one important component: contact plug resistance in advanced MOSFETs.

A great deal of work focuses on the characterization and modeling of variability in MOSFETs. In particular, threshold voltage and channel length are two device parameters which have been studied extensively. With increased scaling, however, the parasitic components of the MOSFET are playing a more significant role in determining the performance of a device. Parasitic resistance components resulting from the gate-to-source/drain overlap regions, source/drain extensions, and contact regions are growing in magnitude relative to the intrinsic resistance associated with the channel of

the MOSFET. High frequency device characteristics are also increasingly determined by parasitic capacitances associated with the charge storage not only in the channel, but elsewhere in the device as well. As new materials and processing techniques are being investigated to reduce these parasitic effects, it will become important to accurately assess and model the variability in these contacts.

Furthermore, a variability characterization technique and a decomposition methodology associated with it is important for contacts and vias in emerging technologies. Whereas in conventional silicon CMOS processes, the contact resistance can be as low as 10Ω, in an emerging technology such as carbon nanotube vias, the resistance can be significantly higher and more variable [43, 44, 45]. The circuit and variation analysis tools developed in this chapter for the analysis of traditional tungsten plugs in silicon CMOS processes may be adapted to emerging technologies where variations may be more significant.

The rest of this chapter is organized as follows. Section 3.1 discusses the importance of a contact in the context of device operation. Section 3.2 reviews existing methods for contact resistance characterization and modeling. Section 3.3 describes a new arrayed test structure for contact plug resistance variability characterization. Sections 3.4, 3.5 and 3.6 present statistical analysis techniques and measurement results. Section 3.7 underscores the need for variability-based models, and Section 3.8 concludes.

3.1 Contacts in a Device Context

In the context of a transistor, the contact plays a critical role in that it serves as the connection between the transistor itself and the surrounding interconnect or other devices. In other words, the contact is necessary in order to send and receive electrical signals to and from the source, drain, and gate terminals of a transistor. For the source and drain terminals, the contact is generally comprised of a silicide material to form the interface between the semiconductor active region and the metal which is the contact plug. A different type of contact or silicide may be used in order to

contact a polysilicon gate, and these may also differ from those used to contact metal gates which exist in a high-K metal gate technology. An example of how source-drain contacts are used in a transistor is shown in Figure 3-1. The contact plug itself

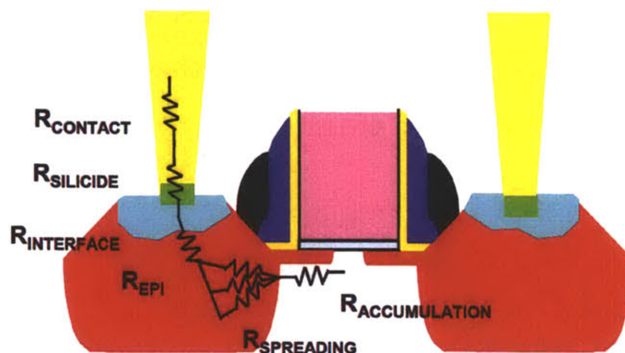


Figure 3-1: Contact in the context of a transistor and relevant extrinsic parasitic resistances [5]. The parasitic resistance components which most significantly involve the contact are $R_{CONTACT}$, $R_{SILICIDE}$, and $R_{INTERFACE}$. However, other choices, such as that to use elevated source-drain regions, can significantly impact the magnitude of these parasitics and others.

is shown in yellow, which is then connected to the silicide (green) which is formed just above the active region. The silicide-diffusion interface, shown in light blue, is modeled with a resistance, $R_{INTERFACE}$. An electrical signal must travel from the edge of the transistor channel to the top of the contact or set of contacts before it reaches the first metallization layer, upon which it can traverse through various vias and upper-level metallization layers before it must go through another contact or set of contacts to access another transistor terminal. The work in this chapter focuses on the characterization of the contact plug resistance of multiple contacts in a transistor context and will be described in more detail in the upcoming sections.

3.2 Background Work

The interface between the device and local metallization layers is critical in understanding the performance and robustness of a MOSFET. Consequently, increasing attention has been paid to the analysis of contact resistance, which represents a growing proportion of the total on-resistance associated with a transistor. ITRS projections

predict that the contact resistance will double every technology generation [46]. This is partially due to the higher aspect ratios required for contact plugs in advanced technologies. Methods for contact characterization and analysis can be categorized into four areas: fabrication and measurement of individual contacts, failure and defect analysis, analytical modeling, and arrayed test structures for characterization of parametric variability. Details of each of these areas are discussed in the subsequent sections.

3.2.1 Individual Contact Measurement

A key step in the development of any process involves the characterization of individual components, such as transistors, diffusion resistors, or metallization layers. In this context, the fabrication of contacts and the measurement of its resistance has been performed and reported in the literature several times. Measurement techniques for the measurement of a single contact can generally be classified as using either a) the transfer-length method, or b) the four-terminal probe method, both of which will be described in the following subsections. In addition to these techniques, SEM images can be used to inspect the physical quality of an individual contact. Contact chains can also be used to measure the series resistance of multiple contacts and therefore obtain an average contact resistance [47]. These techniques involving individual contact measurements tends to consume a great deal of both on-chip and off-chip resources due to the use of large, dedicated pad structures; as a result, opportunities for variability characterization, which requires measurements of many contacts, are limited.

Transfer Length Method

In terms of characterizing contact resistivity, one of the first major breakthroughs was the transfer-length method, or TLM, which enabled the determination of contact resistivity by obtaining the resistances of many different sized contacts and using analytical equations to determine the resistivity [48]. The derivation of the analyt-

ical equation stems from a resistor-grid model of the metal-semiconductor contact interface as shown in Figure 3-2.

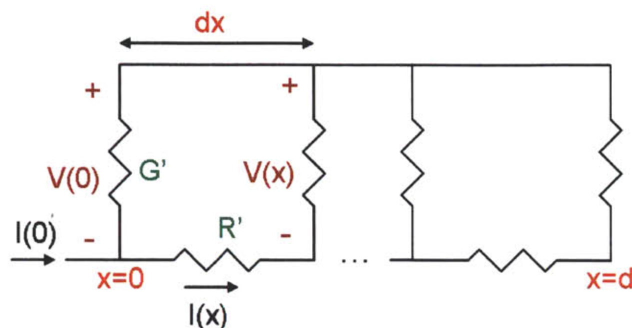


Figure 3-2: Resistor-grid model of metal-semiconductor contact [6]. R' is the resistivity of the doped silicon source/drain junction and G' is the resistivity of the metal-semiconductor contact interface.

Here R' is the resistivity of the doped silicon source/drain junction, and G' is the resistivity of metal-semiconductor contact interface. The metal resistivity is negligible compared to the other two, so it is neglected in this analysis. Based on simple Kirchoff's law current equations, the contact resistance, R_c, is

$$R_c = \frac{R_{sc} L_T}{W_c} coth \frac{d}{L_T} \tag{3.1}$$

where $L_T = \sqrt{\rho_c / R_{sc}}$ is the characteristic contact transfer length, R_{sc} is the sheet resistance of the diffusion layer directly beneath the contact area, W_c is the contact width, and d is the contact length [6]. The transfer length method was extended for the analysis of silicided diffusion regions in [49]. More recently, technology scaling has demanded smaller dimensions for contacts, as well as design rules which allow for only one specific contact size. Nevertheless, recent efforts to accurately characterize the contact resistivity have been successful despite these trends [50]. For example, in [51], the TLM was combined with analysis of various geometries to determine the specific contact resistance of NiSi and PtSi silicides.

Four-Terminal Probe Method

Another common method used to determine contact resistance is the four-terminal Kelvin resistor method. This test structure uses four terminals which are all connected to the device under test — in this case, a contact [52]. With this structure, it is relatively simple to obtain the contact resistance, but more difficult to obtain the specific contact resistivity. This is because the current flow through the contact is non-uniformly distributed throughout its area, which presents problems when coupling this method with the TLM in order to obtain ρ_c. A test structure was developed in a $0.8\mu m$ technology to determine the distribution of contact resistances for an array of 4k contacts in [53]. The results indicated a Gaussian distribution of contact resistances for any given contact size. Some work has also focused on refining each of these methods by eliminating parasitic components which could lessen the accuracy of the resistivity measurement, or on developing a unified approach to accurately extract specific contact resistance [54] [55].

3.2.2 Failure and Defect Analysis

A more comprehensive approach for individual contact characterization is failure and defect analysis, which is useful for functional yield estimation. Thus, a variety of test structures are available to characterize resistances, in order to quantify the failure or defect rate of contacts or vias. These structures seek to capture the functional yield of the contacts, e.g., by detecting open contacts with extremely high resistance, rather than trying to determine parametric yield by obtaining a distribution of individual contact resistances. The advent of these types of test structures has been due to the increasing complexity and density of integrated circuits and the possibility that a small number of contact or via failures could jeopardize proper operation. In [56], a passive multiplexing approach was developed in order to quickly and accurately characterize contact and via fail-rates. Using bit-lines and word-lines and testing a number of combinations of interconnect paths, highly resistive contacts were detected. The cause for the open contact and via failures was found to be partly due to a lack

of tungsten within the plug which was to be filled. Another test structure, described in [57], implemented an efficient methodology to characterize open contact and via failures by using a pyramidal architecture scheme. SEM results indicated that the highly resistive contacts and vias were caused by three major process failures: voiding failures, etching failures, and resistive failures. These techniques for failure and defect analysis are useful in capturing functional yield, but the increasing parametric variability in advanced technologies requires other techniques for parametric variability characterization.

3.2.3 Analytical Modeling

For contact resistance modeling, the TLM remains useful as an accurate method for modeling contact resistance. However, due to the differences in device geometries as a result of scaling, such models have been adjusted to include sidewall interface resistance, dopant redistribution calculations, and various nuances in process steps. For example, in [58], the contact resistance is modeled as part of an effort to accurately model the entire series resistance of a device. An effort to include the sidewall contact resistance contribution is made in [59], since the contact width is shrinking. Analytical modeling of contact resistance is oftentimes folded into the analytical modeling of extrinsic resistance in a MOSFET. Such modeling efforts are beneficial in understanding the physical nature of the contact resistance and its interactions with other variables, but it is also important for these efforts to be extended to account for the increasing variability which is present in advanced technologies.

3.2.4 Arrayed Test Structures

Arrayed test structures are helpful in gathering variability data for contact resistances. In [60], a test structure was designed to assess the resistance of individual contacts located within transistors. Results showed that the distribution of resistances is Gaussian, but that the means can also change due to different device layout configurations. When the data gathering process is sufficiently fast, this rapid char-

acterization of many contacts can be helpful in assessing variability. The remainder of this chapter will focus on a test chip which has also been designed and measured for contact plug resistance variability characterization.

3.3 Test Structure for Contact Plug Resistance Variability Characterization

A test chip has been implemented in a 90nm bulk CMOS technology to determine the characteristics of contact plug resistance variability and the layout-based design parameters which can affect it. The resistance is obtained by a current-force, voltage-sense approach, similar to the four-terminal probe method used for individual contact characterization, which is multiplexed across over 40,000 devices under test. A contact measurement bank comprised of 36,864 DUTs is implemented where only contact plug resistances are measured. In addition, a simultaneous measurement bank comprised of 3,760 DUTs is implemented where measurements are performed for both contact plug resistance and current through the transistor in which the contact is located. Section 3.3.1 describes the test circuit used for the contact measurement bank, Section 3.3.2 describes the test circuit used for the simultaneous measurement bank, Section 3.3.3 discusses the measurement accuracy of the test circuits, and Section 3.3.4 describes the design of experiments used for the test structure.

3.3.1 Contact Plug Resistance Measurement Circuit

Figure 3-3 shows a three-dimensional view of the device under test and the current flow through it. The current is forced into the silicide region through one contact and out of the region through the other contact (yellow), which is the contact to be characterized. The voltage is tapped across the two terminals of the contact, emanating in the two voltage outputs, V_{OUTL} and V_{OUTH}. The current is guided to a sink device transistor where it then flows to ground. The resistance is directly proportional to $V_{OUTH} - V_{OUTL}$ with a proportionality constant of $1/I_F$, the inverse of the forced

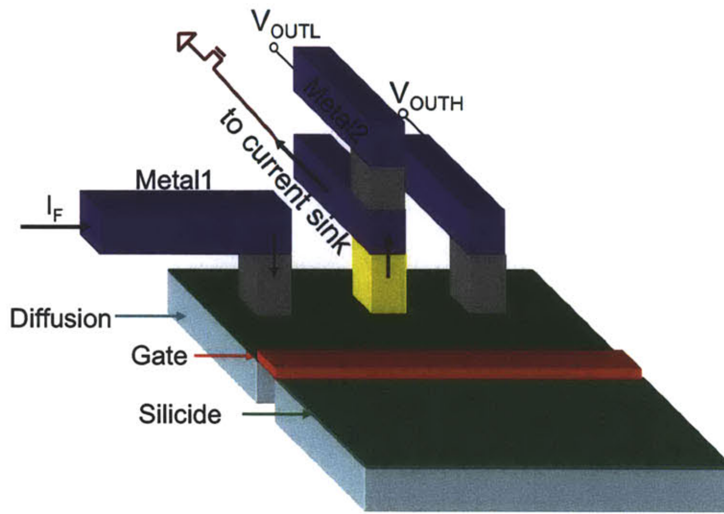


Figure 3-3: Three-dimensional transistor view with path of current flow through contact to determine contact resistance of middle contact under test (yellow). The gate is switched off so no current flows from the source to the drain of the transistor itself.

current. The multiplexing scheme, shown in Figure 3-4, features three DUT access transmission gates which are enabled by the outputs of row and column decoders. In addition to this scheme, each row contains its own high-V_T transmission gate to adequately minimize leakage current paths. The analog-to-digital conversion of the voltage outputs is performed off-chip, and digital output enables fast characterization of contact plug resistances.

3.3.2 Simultaneous Contact and Device Measurement Circuit

The test circuit for the simultaneous measurement of contacts and devices is shown in Figure 3-5. The gate of the transistor in which the contact is located is connected to a multiplexing transmission gate switch, M4, which is then connected off-chip to a configurable gate voltage. The sources of all DUTs are connected to the output of an operational amplifier through a snaking wire which then loops back and connects to the negative input of the same op-amp. The positive input is connected to an external source-meter which provides the desired source voltage on the DUT for measurement.

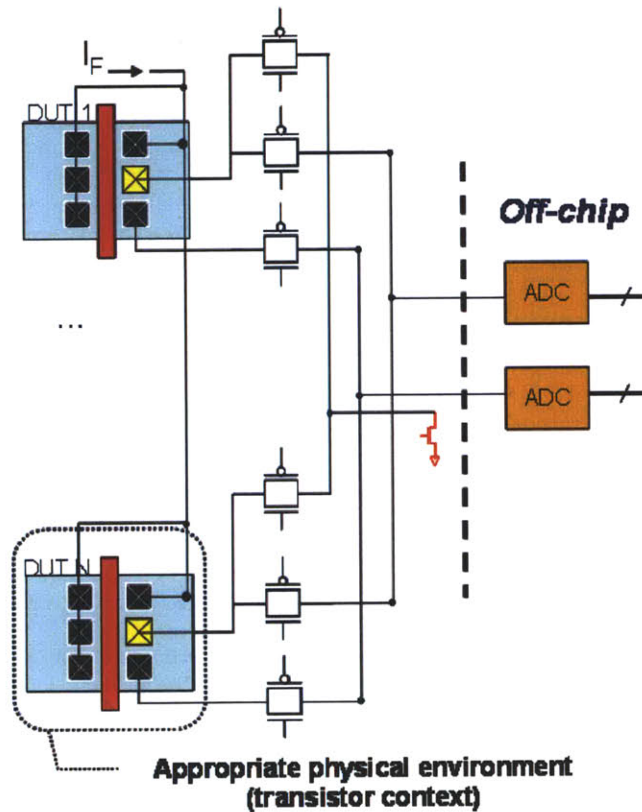


Figure 3-4: Test circuit to measure contact plug resistances in an arrayed set of DUTs. Three transmission gate switches are used to control access to the V_{OUTL}, V_{OUTH}, and I_F. Off chip-analog-to-digital converters are used to sense the output voltages.

This configuration, used previously in [61] is used to eliminate the IR drop within the interconnect between the pad of the forced source voltage and the DUT source. The drain node of the transistor is also forced in a similar way, except that the voltage is only applied to contact A (top-most contact). This is because the contacts are not shorted together through M1 because the contact plug resistance measurement still needs to be made.

Three off-chip resistors are used to measure the I-V current through the DUT. The resistance value chosen for a particular current depends on the magnitude of the current and is controlled through D_{meas}. For smaller currents, the smallest resistor is used, while for larger currents, the largest resistor is used. The voltages across these resistors is measured through an off-chip analog-to-digital converter (ADC). Then this voltage is divided by the value of the resistor to determine the current through the DUT.

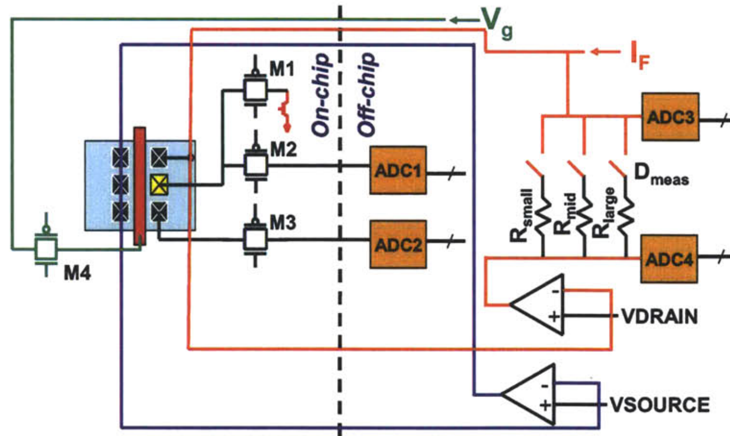


Figure 3-5: Test circuit to measure both contact plug resistances and transistor I-V characteristics for an arrayed set of DUTs. An additional transmission gate switch is used to control the gate voltage of the transistor DUT and off-chip operational amplifiers are used to force the source and drain voltages to their desired values. The device current is measured through the measurement of voltage across an off-chip resistor that is located in the current path of the DUT transistor.

3.3.3 Measurement Accuracy

The measurement accuracy of contact plug resistances in the contact-only bank is determined by two factors. The first factor is the difference in voltage which appears between the two terminals of the DUT contact of interest and the voltage that appears at the inputs of the off-chip analog-to-digital converter. The second factor in determining measurement accuracy is the difference between the current which flows through the DUT contact of interest and the current which is sensed at the drain of the sink transistor using an off-chip source-meter. Simulations indicate that the combined measurement error resulting from these two sources is bounded by 0.01Ω.

3.3.4 Design of Experiments

The design of experiments for this test chip includes five key layout design parameters: contact-to-gate distance (d_{cg}), contact-to-diffusion edge distance (d_{cd}), metallization layer to contact overlap for the y-dimension (d_o), the number of contacts in the source diffusion region (N_s), and the number of contacts in the drain diffusion region (N_d). These parameters are shown pictorially with reference to the DUT in Figure 3-6. In addition, DUTs with contacts located in both NMOS and PMOS transistors are used. The contact which is measured is always on the drain side of the device. A DUT type is one which consists of some particular combination of these five layout design parameters. The chip contains a total of 55 types of devices under test. The chip also contains 256 rows and 144 columns of DUTs, for a total of 36,864 resistance measurements.

Figure 3-7 shows pictorially the first three layout parameters to be examined and the corresponding sets of values chosen. A comprehensive list of the input variables used in the design of experiments and their corresponding possible values is shown in Table 3.1.

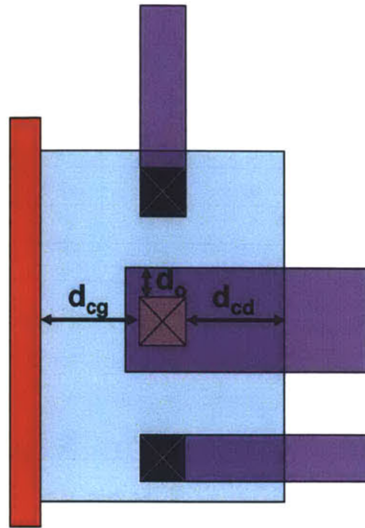


Figure 3-6: Geometry-based variables in the DOE (half transistor shown for simplicity). Contact-to-gate distance, (d_{cg}), contact-to-diffusion edge distance (d_{cd}), and metallization layer to contact overlap for the y-dimension (d_o) are varied to determine any possible impact on contact plug resistance.

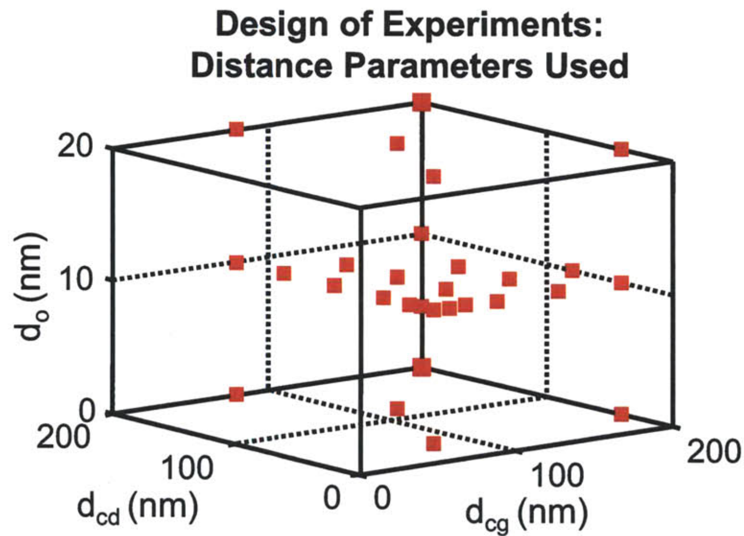


Figure 3-7: Design of experiments for contact resistance variability analysis. Values are chosen such that many DUT geometries exist at or near "nominal" case of $d_{cg} = 80nm$, d_{cd} = 40nm, and d_o = 10nm.

Layout Design Parameter Values			
d_{cg} (nm)	d_{cd} (nm)	d_o (nm)	$N_{C,left}, N_{C,right}$
80	40	0	3,3
90	50	10	4,4
100	60	20	5,5
120	80	-	3,4
160	120	-	4,3
200	160	-	-
-	200	-	-

Table 3.1: Layout design parameter values chosen for the DOE: 4 factors and 55 DUT types representing a subset of all possible combinations of these 4 factors.

3.4 Variation Decomposition Methodology

In order to analyze the measurement data, a variation decomposition methodology has been developed which uses the concept of spatial correlation to uncover within-die systematic spatial trends. Because the arrangement of DUTs within the die is periodic, spatial correlation analysis can also reveal trends in contact plug resistance based on DUT type. This decomposition methodology is used to determine the magnitude and nature of the different effects which contribute to variation in the measurement data.

3.4.1 Spatial Correlation Computation

Before the overall methodology is described, it is important to define the concept of spatial correlation as it applies to this measurement data. A spatial correlation coefficient is computed for a set of distances, starting from $0\mu m$ to the longest distance between any pair of DUTs on a single chip, at a reasonably chosen interval to see any possible trends. The spatial correlation coefficient for each distance is computed as follows. Let $R(x, y, c)$ represent the contact plug resistance of the DUT located at (x, y) on chip c. Then, for a given distance, d, a set of all non-intersecting pairs $\{R_i, R_j\}$ for which Equation 3.2 is satisfied is computed.

$$d - \epsilon \leq \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \leq d + \epsilon \tag{3.2}$$

Then, the pairwise spatial correlation coefficient, $\rho(d)$, is computed for the set of pairs using Equation 3.3.

$$\rho = \frac{\sum(R_i - \mu_{R_i})(R_j - \mu_{R_j})}{\sigma_{R_i}\sigma_{R_j}} \tag{3.3}$$

When computed for many distances, the spatial correlation coefficients, $\rho(d)$, can be plotted versus the DUT separation distance, d. The magnitude of these coefficients and the associated confidence intervals can help to determine the nature of systematic trends if they exist.

Although there is a distinction between spatially correlated versus uncorrelated variation and systematic versus random variation, the results of spatial correlation analysis can still help to determine systematic trends within the measurement data. Therefore, a large spatial correlation coefficient may not indicate a pure spatial trend but rather a systematic effect possibly due to layout patterns, within-die edge effects, or die-to-die variations.

3.4.2 Decomposition of Variation Sources

Once spatial correlation analysis shows that a possible systematic trend exists, the $\rho(d)$ is inspected to determine the possible systematic effect at work. Figure 3-8 shows how spatial correlation analysis is used to enable the identification of systematic variations.

Starting from the original set of data, the spatial correlation coefficients are computed. If there are statistically significant coefficients, this indicates a possible systematic source of variation. Upon identification of the variation source, a residual data set is computed by subtracting the mean of the set of data to which each $R(x, y, c)$ belongs from its original value. This process continues until no statistically significant spatial correlations can be found in the residual measurement data. The sources of variation identified for this data set, in order from most significant systematic contribution to least significant systematic contribution to total variance, are the following: die-to-die variation, within-die layout-dependent systematic varia-

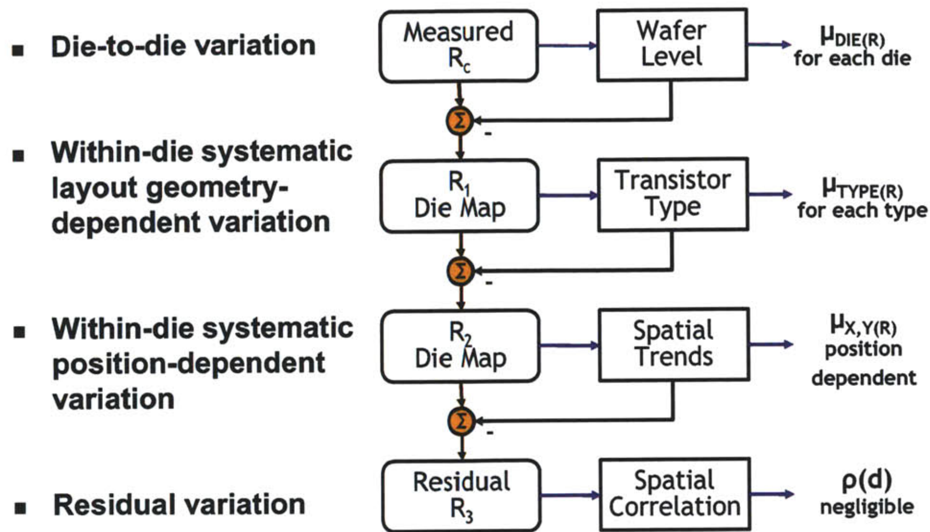


Figure 3-8: Spatial correlation analysis-based variation decomposition methodology. Spatially correlated contact plug resistance values can indicate the presence of a systematic trend, which can be subtracted from the measured data points to obtain a residual resistance map, for which the same analysis can be performed until there is no significant spatial correlation detected.

tion, within-die position-dependent variation. The remaining variation is categorized as random, spatially uncorrelated variation.

Equation 3.4 describes the decomposition of the resistance value of a single DUT into its respective components.

$$R = \mu_{DIE(R)} + \mu_{TYPE(R)} + \mu_{X,Y(R)} + \epsilon \tag{3.4}$$

Each measured DUT resistance, R, has the following properties within the context of the decomposition. First, the die in which the DUT is located is denoted as $DIE(R)$. Next, the geometric type as determined by the values for each variable in the DOE is $TYPE(R)$. Finally, the within-die x and y coordinates of the DUT with reference to the entire chip, with the bottom left corner of the chip area being $(0, 0)$ are $X(R)$ and $Y(R)$, respectively.

3.4.3 Analysis of Variance (ANOVA)

In order to determine the existence and/or nature of any cross-term effects on the value of a DUT resistance, it is necessary to perform an analysis of variance on the measurement data. A three-way ANOVA with single factor and two-factor interactions were considered, where the three factors were the DUT source-drain width, w_{sd}, the M1-CA overlap distance in the y-direction, d_o, and the die on which the DUT is located. Results of this analysis, shown in Figure 3-9 indicate that there are no statistically significant coefficients for the interaction terms which contribute to the variance of the measurement data. The die on which the DUT is located, the

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
sd_width	6.5508	3	2.18361	2353.02	0
do	0	0	0	0	NaN
die	5.6987	25	0.22795	245.64	0
sd_width*do	0.2367	6	0.03945	42.51	0
sd_width*die	0.1782	175	0.00102	1.1	0.1838
do*die	0.0286	50	0.00057	0.62	0.9848
Error	5.8696	6325	0.00093		
Total	32.6003	6590			

Figure 3-9: ANOVA results on wafer-level measurement data of contact plug resistance. The largest sum of squares terms are those coming from the source-drain width parameter, the die parameter, and the error term for unexplained variance. The sum of squares of the three interaction terms are much smaller in comparison.

source-drain width of the DUT in which the contact is located, and other variations unexplainable by the factors (likely random variation), contribute the most to the total variation of contact plug resistance.

3.5 Statistical Analysis Results

Measurement results are reported for a total of 23 die. The results will be described in five sections: overall trends, die-to-die trends, within-die systematic layout-dependent trends, within-die systematic position-dependent trends, and random spatially uncorrelated variations which are highlighted by the comparison of measurements from

outlier and non-outlier die. The percentage of total variation contributed by each of these variation sources, in addition to that contributed by other small systematic effects, is shown in Figure 3-10.

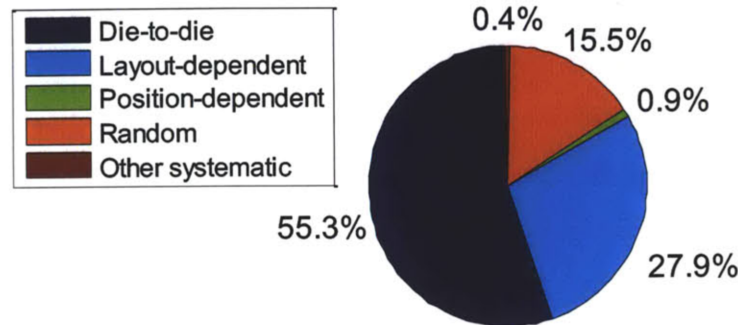


Figure 3-10: Contribution of each variation source to total variation in contact plug resistance. More than half of the total variance can be attributed to die-to-die variations, while over 25% of the variance comes from the layout-dependent systematic component. Random within-die variation represents roughly 15% of the total variance.

3.5.1 Overall Trends

The summary of the statistics on contact plug resistance are as follows. The mean contact plug resistance over all measurements made on all die on the wafer is 14.36Ω. The standard deviation of contact plug resistance over all measurements made is 0.92Ω, which results in a $\frac{\sigma}{\mu} = 6.15\%$. This indicates that, for this 90nm technology, while the percentage variation on the contact plug resistance is substantial, its overall impact on the transistor source-drain parasitic resistance variation is still small due to the comparatively small contact plug resistance mean. However, the methodologies used in order to uncover systematic and spatial trends in the measurement data are applicable to a wide variety of potential data sets. Furthermore, transistor dimension scaling and the use of new types of contact plugs, materials, and silicides in future technologies may increase the amount of observed variation in contact plug resistance. For this measured data, the distribution of contact plug resistance over a single die is nearly Gaussian, while the distribution of contact plug resistance over the entire

wafer is not quite Gaussian due to the presence of multiple variation sources which affect the contact plug resistance values.

Figure 3-11 shows the distribution of contact plug resistances over a single die. In addition, Figure 3-12 shows a normal probability plot of the same data points, normalized to standard deviation values of contact plug resistance. This plot indicates that the data points are nearly normally distributed, but the presence of some systematic effects introduces non-normality into the data, including a slightly larger than Gaussian upper tail.

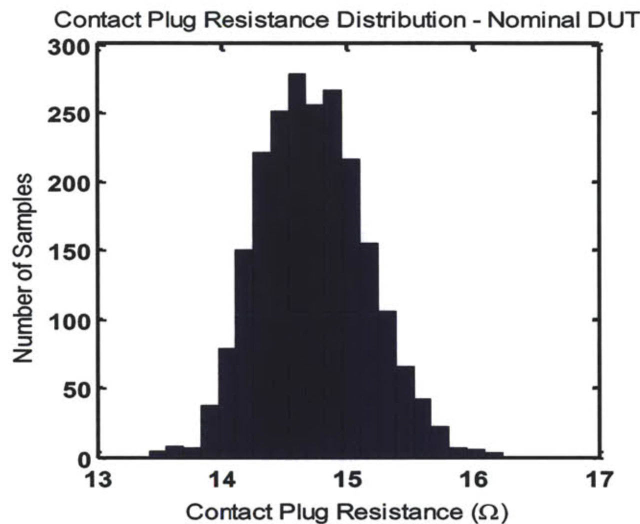


Figure 3-11: Distribution of measured contact plug resistances across one die.

The wafer-level distribution of measured contact plug resistance is shown in Figure 3-13. The normal probability plot for the same distribution can be seen in Figure 3-14, where indications of non-normality can be seen at several points in the distribution, likely as a result of one or more systematic variation sources.

3.5.2 Die-to-Die Trends

To observe die-to-die variation, the wafer level-normalized average die contact plug resistance is plotted on a spatial wafer map in Figure 3-15 for 43 measured die. For this plot, 43 die are shown because the DUTs have been measured from the simultaneous measurement bank. This is to more clearly show any possible wafer-

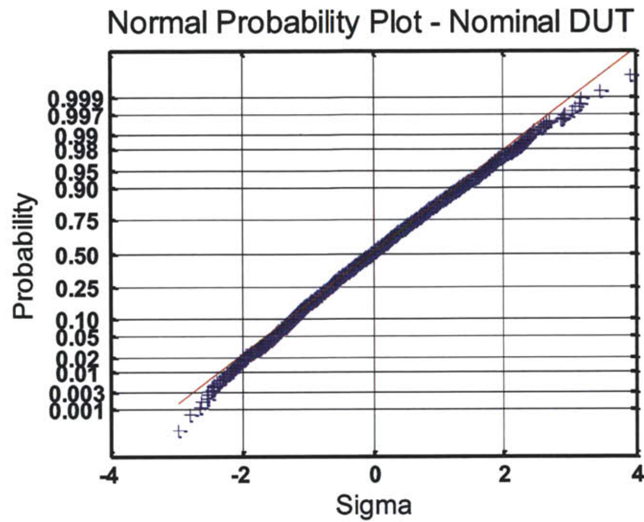


Figure 3-12: Normal probability plot of contact plug resistance measurements over one die.

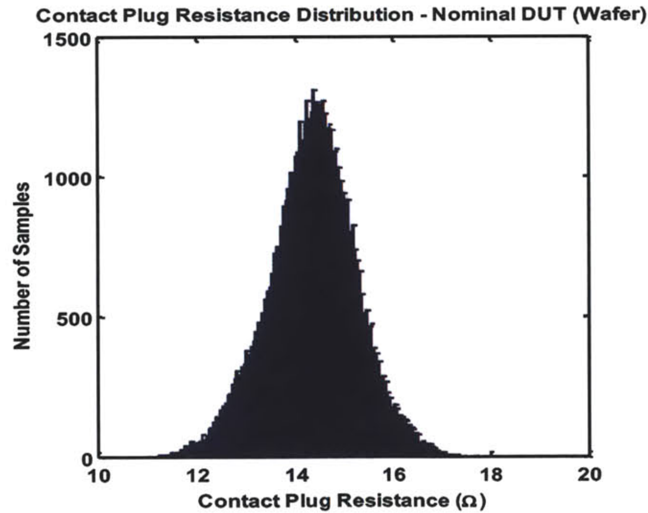


Figure 3-13: Distribution of measured contact plug resistances over the entire wafer, which has a mean of 14.36Ω and a standard deviation of 0.92Ω.

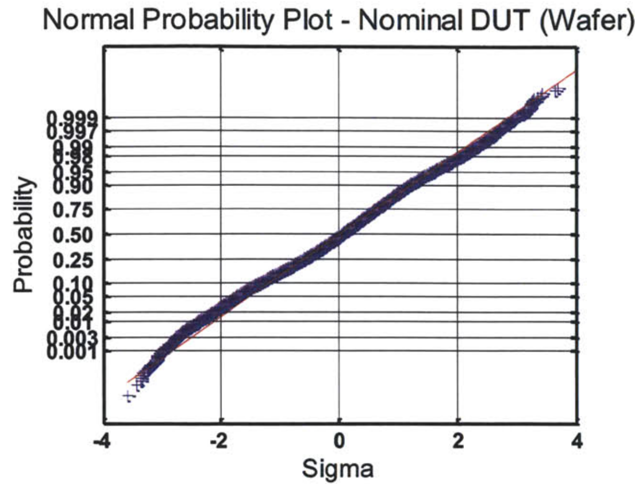


Figure 3-14: Normal probability plot of contact plug resistance measurements over the entire wafer show that the distribution is not Gaussian. In this case, this is due to the presence of various systematic effects due to various factors.

level trends in the data. However, the data used in the decomposition come from the contact-only bank, for which 23 die are measured. Because no systematic wafer-level trend is seen, we can assume that the die-to-die variation is largely random and therefore the die mean can be subtracted to remove the effect from the data. With more measured die on the wafer and/or more wafers, it may be possible to determine that a systematic wafer-level variation exists with a higher confidence. In this case, a wafer-level variation model which had a systematic component could be constructed. Then, the corresponding variations could be subtracted out using such a model in order to arrive at the residual variation data.

3.5.3 Within-Die Systematic Layout-Dependent Trends

This analysis focuses on variations arising from the design of experiments in layout parameters. Because a large number of replicated DUT types are available (both within each die, and for 23 die), it is possible to obtain quite tight confidence intervals on the estimation of mean and variance of resistance for each DUT type, enabling us to identify any systematic effect of different layout parameters on the mean and variance of the contact plug resistance. Two notable effects are those of d_{cg} and

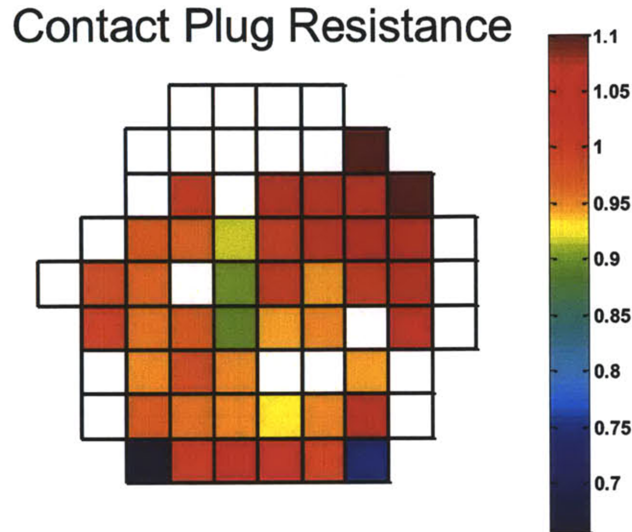


Figure 3-15: Wafer map of average die contact plug resistance for 43 measured die. Given the observed data, no statistically significant wafer-level trend is observed. Some outlier die are located at the corners of the wafer.

d_{cd}. Figure 3-16 shows the resistance mean and standard deviation as a function of the distance between the contact and the polysilicon gate with all other layout parameters held constant. The values of the other layout parameters are $d_{cd} = 40nm$, $d_o = 10nm$, and $N_s = N_d = 3$, and only contacts within NMOS devices are included. Figure 3-17 shows the normalized resistance mean and normalized standard deviation as a function of the distance between the contact and the edge of the diffusion region. In this case, the other layout parameters which are held constant are The values of the other layout parameters are $d_{cg} = 80nm$, $d_o = 10nm$, and $N_s = N_d = 3$, and only contacts within NMOS devices are included here as well. Both the mean and standard deviation are plotted with 95% confidence intervals represented by the error bars attached to each point. Resistance values are normalized to the global wafer mean.

Results show an increase in resistance with both increasing d_{cg} and d_{cd}. However, there appears to be no clear significant change in the variance when plotted against these distances.

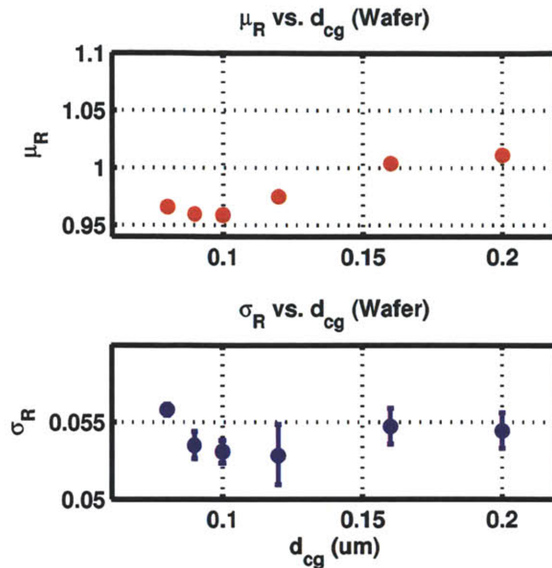


Figure 3-16: A plot of contact plug resistance resistance mean and standard deviation versus d_{cg} shows an increase in mean contact plug resistance for those contacts which are located further away from the polysilicon gate of the transistor. However, the standard deviation of the measured resistance does not change as a function of d_{cg}.

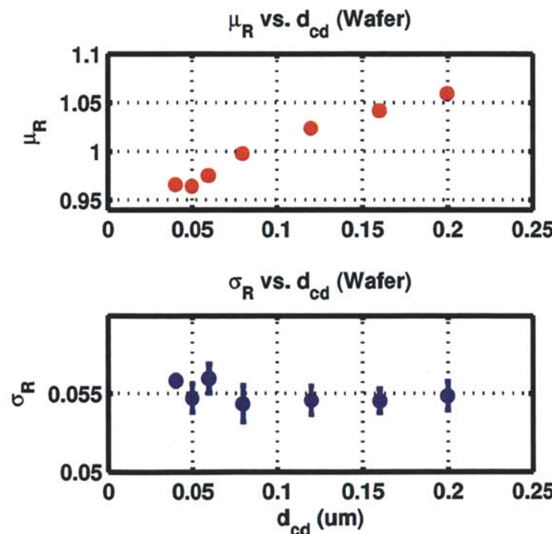


Figure 3-17: A plot of resistance mean and standard deviation versus d_{cd} shows an increase in mean contact plug resistance for those contacts which are located further away from the edge of the diffusion region of the transistor. However, the standard deviation of the measured resistance does not change as a function of d_{cd}.

3D Device Simulations

Device simulations are performed to better understand the nature of this mean shift in the plug resistance. Using the 3D Sentaurus device simulator, the resistance is observed as a function of both d_{cd} and d_{cg}. The structure shown in Figure 3-18 is simulated to determine the distribution of electrostatic potential at the silicon interface in the source-drain region. Current enters the transistors through an orthogonal

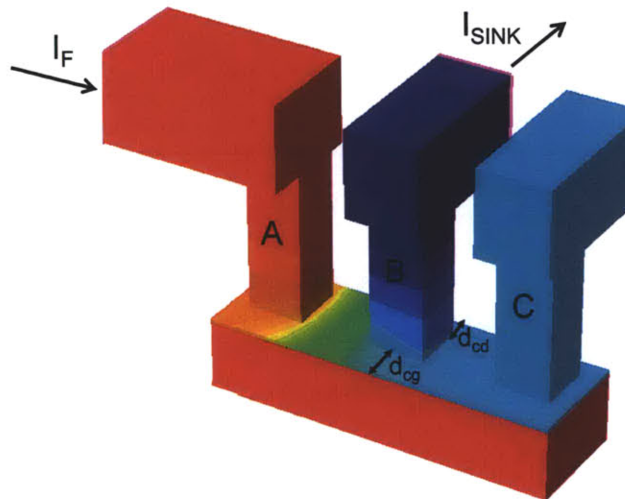


Figure 3-18: A 3D structure closely replicating the test structure design is used for device simulations. Determining the current flow and electrostatic potentials at the surface of the silicide region can help to understand systematic trends in the measurement data.

M1 wire just as in the implemented test circuit and flows only through the contacts A and B. No current flows through contact C because it is simply a tap whose voltage should be identical to that of the bottom contact B.

When this structure is simulated in two different configurations, a different distribution of electrostatic potentials is seen at the bottom silicon surface of contact B. For a "narrow" structure, which corresponds to the DUT type with design parameters of d_{cg} = 80nm and d_{cd} = 40nm, the current flow is restricted because of the relatively narrow source-drain region, as shown in Figure 3-19. Consequently, the electrostatic potential surrounding contact B is asymmetric and the electrostatic potential throughout the area below contact C is low. This results in a smaller dif-

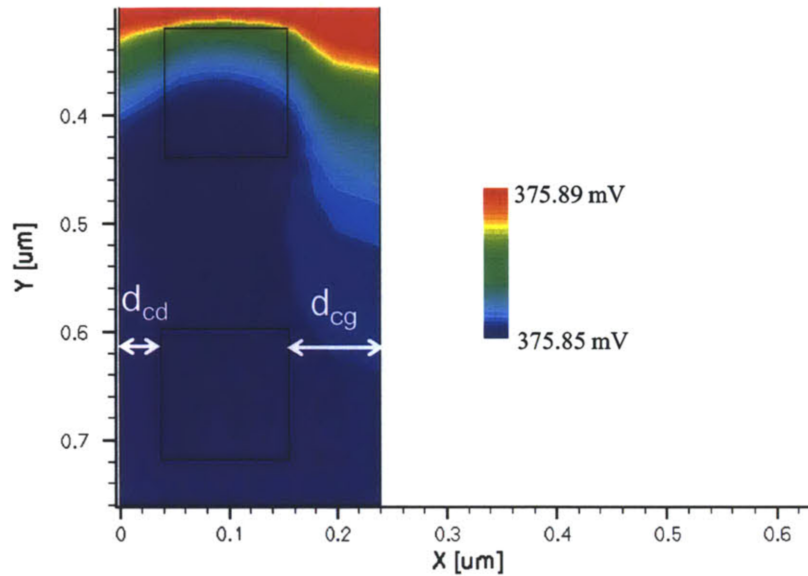


Figure 3-19: A contour plot of electrostatic potential at silicon surface of a narrow diffusion region shows that current crowding occurs near the top of contact B, resulting in some difference in average electrostatic potential between contacts B and C.

ference between the average electrostatic potentials at the surfaces of contacts B and C, which in turn results in a measured contact resistance that is smaller than what one would ideally measure by observing a single contact in isolation.

In contrast, when a "wide" structure is used, the opposite effect is seen. For the simulation shown in Figure 3-20, the design parameters are d_{cg} = 280nm and d_{cd} = 160nm.

Because of the large area in the x-direction for current to flow, the electrostatic potential at the bottom surface of contact B is more symmetric than in the case of the "narrow" structure. The potential at contact C matches closely with that of the boundary of contact B. The difference between the average electrostatic potentials at the bottom of contacts B and C as a result is larger than in the case of the "narrow" structure. This larger difference in voltage then maps to a measured resistance which is larger than what one would ideally measure by observing a single contact in isolation, since the resistance is measured as $\frac{\Delta V}{I}$.

When the actual measured contact plug resistances from the test chip are plotted,

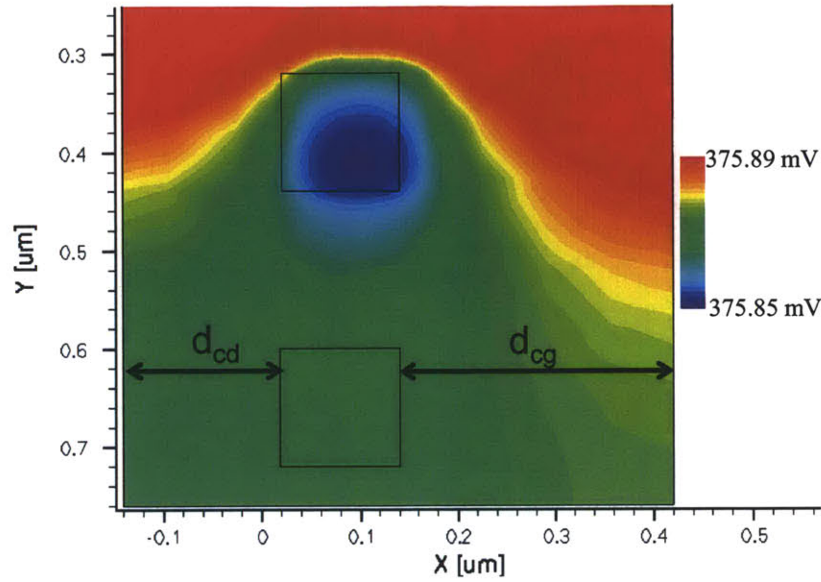


Figure 3-20: A contour plot of electrostatic potential at silicon surface of a wide diffusion region shows that less current crowding occurs because of the large amount of diffusion area through which current can flow. In this case, the difference in average electrostatic potential between contacts B and C changes from its value in the case of a narrow diffusion region.

a clear trend is seen as shown in Figure 3-21. This trend is consistent with the 3D device simulation results, which suggest that the amount of current crowding due to the source-drain widths of the DUT causes different contact plug resistance measurements.

Because the underlying variable governing the current flow through the contacts is actually the total width of the source-drain region, the measurement results from the test chip have been replotted using this parameter as the control variable, where the source-drain width, w_{sd} is described by Equation 3.5, where w_c represents the width of the contact, which in the case of this technology is 120nm.

$$w_{sd} = d_{cg} + d_{cd} + w_c \tag{3.5}$$

The asymptotic nature of the measured resistance as the w_{sd} increases confirms the conceptual notion that, for increasingly larger source-drain widths, the overall effect on the measured resistance diminishes.

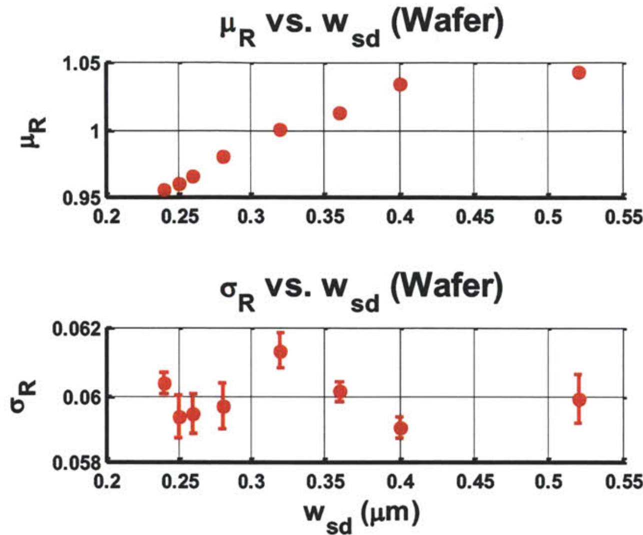


Figure 3-21: A plot of contact plug resistance as a function of source-drain width shows that the average measured plug resistance increases with larger source-drain widths. For very large widths, the average resistance approaches an asymptotic value.

3.5.4 Within-Die Systematic Position-Dependent Trends

When the systematic layout-dependent induced variability components (summarized in the Section 3.5.4) are subtracted off the total resistance distribution die pattern, the remaining portion of the data exposes any systematic (repeatable) within-die trends. To do this, the type mean-subtracted resistance is computed for each DUT. The type mean-subtracted resistance is the measured resistance minus the average resistance of all DUTs of the same type as itself. Some insight is obtained when the resulting mean-subtracted resistance is plotted for each column, with all DUTs in that column averaged. This is done in Figure 3-22 for each of the 144 rows with 95% confidence intervals on the column averages. The resistances are plotted as a fraction shift from the wafer global mean. Clearly, two regions of resistance exist with a sharp change at around $x = 1210\mu m$. The average over all contact plugs on the left side of the chip, where $x < 1210\mu m$, is noticeably greater than the average over all devices on the right side of the chip, where $x > 1210\mu m$. Furthermore, the change in average plug resistance is quite abrupt. While all of the previously seen trends with regards to layout design parameters can still be seen for DUTs in each of the two sides of the

chip, there appears to be an offset. The average resistance measured for DUTs on the right side of the chip is 1.3% lower that for those located on the left side of the chip.

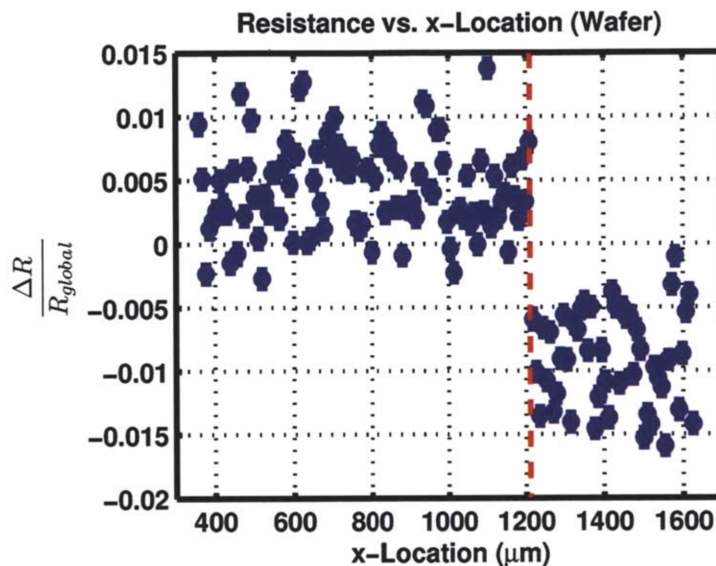


Figure 3-22: Residual resistance as a function of column (x-location) shows two distinct regions of plug resistance. Contact plugs located at $x > 1210\mu m$ have an average resistance which is 1.3% lower than those located at $x < 1210\mu m$.

Though the source of this variation is unclear, it is possible to identify some steps in the manufacturing process which may result in a change of this magnitude and abruptness. One possibility is that an unintentional error or offset occurred during the software-based mask generation process. For example, a previously unnoticed software bug may have caused the optical proximity correction algorithm to shape contacts slightly differently for one side of the reticle versus another. The magnitude of such a difference which would corresponds to the measured difference in plug resistance would be approximately 1.4nm in a single direction.

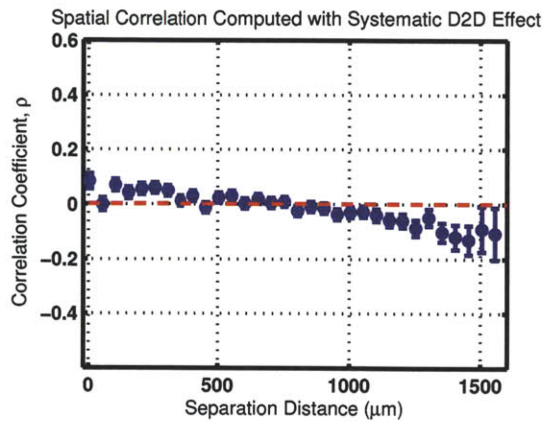
Another possibility may be related to variability resulting from a step-and-scan exposure system used during lithography processes [62]. Because the variability only occurs in the x-direction and not in the y-direction, the optical scan direction would have been the vertical y direction, while the slits would have been organized in the horizontal x direction. The two distinctly different regions in terms of contact plug

resistance suggest that a total of two slits were part of this particular die (part of a larger multi-project die) on the reticle. The resulting effect is a critical dimension change from one side of the die to the other side of the die. This critical dimension change may then manifest itself in terms of a change in the cross-sectional area of the contact plug, thus changing the total resistance. If this is indeed the case, the contact width would have changed by approximately 0.7% in order to explain the observed resistance change.

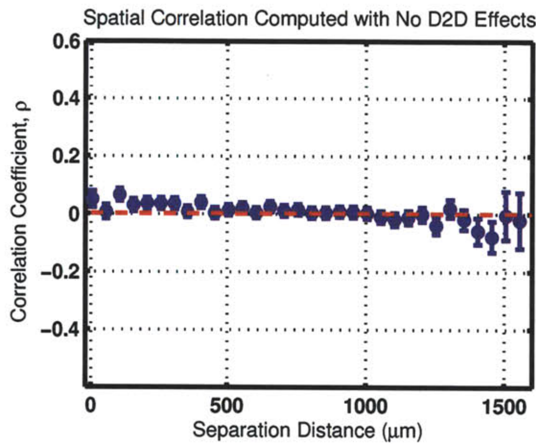
3.5.5 Random Spatially Uncorrelated Variation

While the examination of systematic components of contact plug resistance variability is important, the analysis of random components is also critical for both modeling and for understanding the variation sources. When the systematic components are mean-subtracted from the resistance data as previously described in the variation decomposition methodology (Figure 3-8), what remains is "unexplained" or random variation in contact plug resistance. To analyze the nature of this variability, and in particular to understand if some remaining correlated spatial variation remains, contact-to-contact spatial separation distance correlation analyses can be performed. An important observation below is that systematic spatial trends (such as those discussed earlier) can appear to be spatially correlated variations, if they are not previously identified and removed. Thus, spatial correlation analysis can serve as a tool by which potential spatial trends can be "flagged" for investigation, removal, or mitigation.

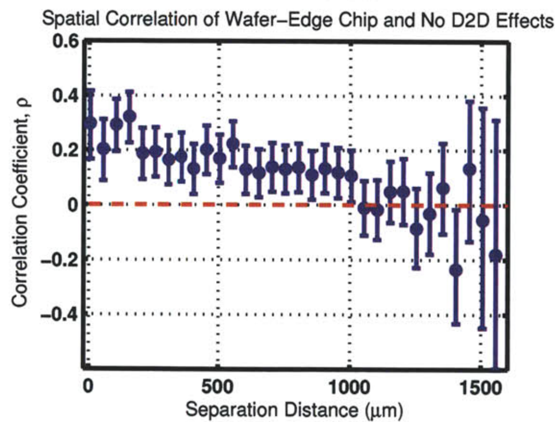
First, to understand the effects of non-removed systematic components on spatial correlation analysis, the distance-dependent spatial correlation coefficient has been plotted for the nominal device type on all die with 95% confidence intervals on each coefficient in Figure 3-23(a). Because the systematic die-to-die component is not removed (mean-subtracted), we see some small positive correlation for small distances as well as some small negative correlation for large distances. In this case, what appears to be spatially correlated random variation is actually systematic variation which has not been removed prior to the correlation analysis.



(a) Spatial correlation analysis computed with systematic die-to-die effects included



(b) Spatial correlation analysis computed with systematic die-to-die effects removed



(c) Spatial correlation analysis computed for single die located at wafer edge

Figure 3-23: (a) Spatial correlation analysis computed with systematic die-to-die effects included, (b) Spatial correlation analysis computed with systematic die-to-die effects removed, (c) Spatial correlation analysis computed for single die located at wafer edge

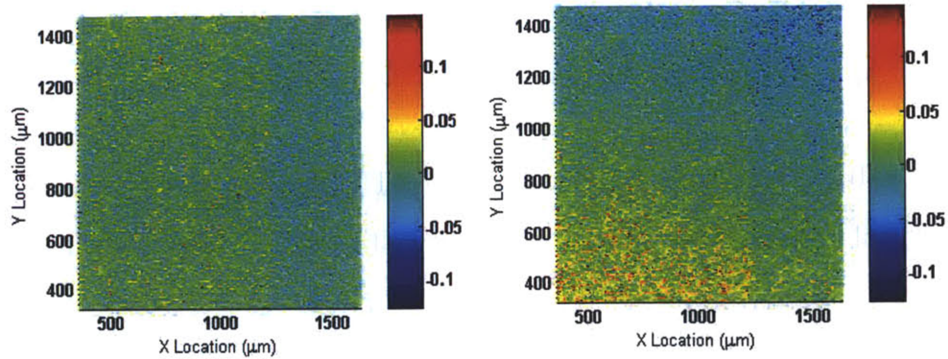
In Figure 3-23(b), the same analysis is done, except the (x, y) location mean is subtracted from each point before the correlation is computed. Here, the small positive correlation at small distances could be the result of either some subtle systematic trends which have not been accounted for, or truly spatial correlation. An illustration of the effect of a substantial spatial trend is seen in Figure 3-23(c), which shows the correlation coefficient as a function of separation distance for contacts located only on the die which is at the edge of the wafer. The 95% confidence intervals in this plot are substantially wider, because the data set includes resistance values from only one die as opposed to all the die. Here, we see both positive correlation at small separation distances and negative correlation at large separation distances, despite the fact that the location means have been subtracted off. This large spatial correlation is due to the unique systematic gradient in contact plug resistance on this die, shown earlier in Figure 3-24(b).

Wafer-Level Edge Effects

To examine the nature of the wafer-level edge effect determined by the spatial correlation analysis, it can be useful to look at the raw measurement data. While nearly all of the other die show a systematic pattern similar to that seen in Figure 3-24(a), the particular die located at the edge of the wafer shows a systematic pattern as depicted in Figure 3-24(b). This indicates that some wafer-level or wafer-edge variation may be affecting the contact plug resistance in this die. Once again, measurements of die from multiple wafers would be useful in determining the cause of this variation.

3.5.6 Spatial Correlation Via Sparse Regression

An alternative approach to uncovering systematic trends in measurement data is based on analysis in the frequency domain. A frequency-domain analysis is performed using a discrete cosine transform (DCT) based method, as described in [63]. When the coefficients are determined using an error-minimizing algorithm such as Simultaneous Orthogonal Matching Pursuit (S-OMP) [64], periodic systematic trends



(a) Normalized resistance map of sample die

(b) Normalized resistance map of die located at edge of wafer

Figure 3-24: Spatial maps of contact plug resistance for both a normal die and die located at edge are shown. The outlier die located at the edge of the wafer has an additional systematic spatial trend.

can be determined. Using this algorithm on the raw contact plug resistance measurements reveals the systematic within-die spatial trend in the data. Figure 3-25 shows the DCT coefficients when the S-OMP based algorithm is used on the data. The results from the algorithm show that the dominant source of within-die systematic variation is that caused by the different layout types of the DUTs. This is clear when comparing the systematic variation determined by the DCT coefficients in Figure 3-26(a) with Figure 3-26(b), which plots the type of each DUT versus its location on the chip.

3.6 Simultaneous Bank Measurement Results

To determine if there is a significant interaction between the contact plug resistance and the DUT current, the simultaneous bank data is analyzed. Results indicate that the device current is correlated with the contact plug resistance, as shown in Figure 3-27. However, the direction of the trend is opposite to what one might expect. For lower values of contact plug resistance, the device current tends to be smaller, while for higher values of contact plug resistance, the device current tends to be larger. The correlation coefficient between the two parameters is 0.33. In addition, a 95% confidence interval bound is drawn in red around the data points.

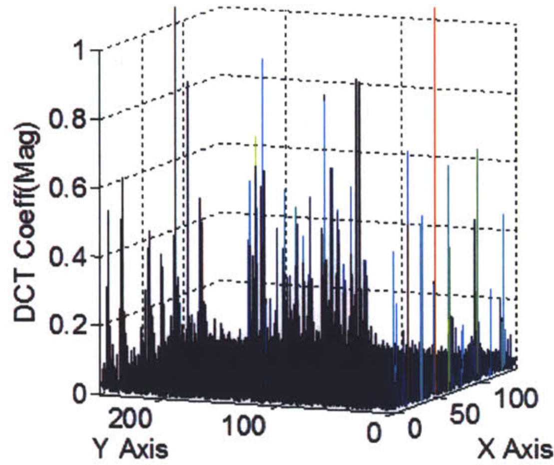
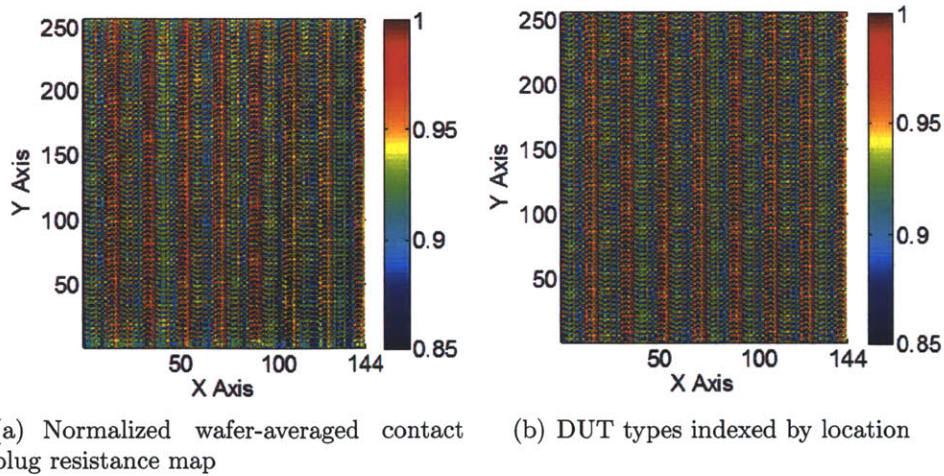


Figure 3-25: DCT coefficients from applying S-OMP-based algorithm on raw measurement data reveal the periodic patterns present in the DUT array due to the repeating order of DUT types in the layout.



(a) Normalized wafer-averaged contact plug resistance map

(b) DUT types indexed by location

Figure 3-26: A die-level map showing systematic layout-dependent trends in contact plug resistance, created from the extracted DCT coefficients from the measurement data, matches the distribution map of DUT types across the chip.

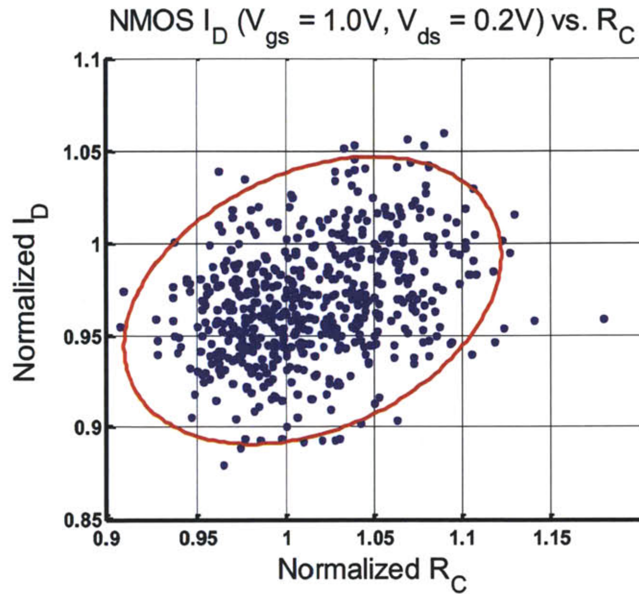


Figure 3-27: A scatter plot of normalized contact plug resistance versus normalized transistor current, measured at $V_{gs} = 1.0V$ and $V_{ds} = 0.2V$. A positive correlation of 0.33 exists between the two variables.

When a set of such a correlation coefficient is computed for multiple values of V_{gs} and V_{ds}, it can be plotted as shown in Figure 3-28. Figure 3-28 shows that the highest correlations between contact plug resistance and device currents occur in the linear region of operation, where V_{gs} is large and V_{ds} is small. Based on this trend, we can conclude that the cause of this correlation is the different amounts of unintentional stress induced by the STI regions surrounding the transistor due to the difference in source-drain widths among different DUT types. The correlation between the contact plug resistance value and the drain current is a by-product of the relationship between contact plug resistance and source-drain width, discussed earlier in Section 3.5.4. This hypothesis is further supported by Figure 3-29, which shows a clear dependence of device current on DUT source-drain width, where the magnitude of current variation is similar to that observed in the plug resistance versus device current correlation scatter plot. In addition, the fact that the NMOS currents increase with larger source-drain widths (longer distance from gate to STI edge) while the PMOS currents increase with smaller source-drain widths (shorter distance from gate to STI edge) is consistent with an STI-related stress effect [65].

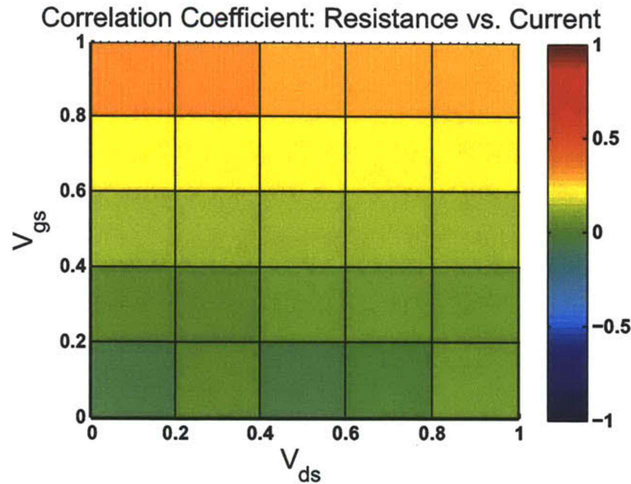
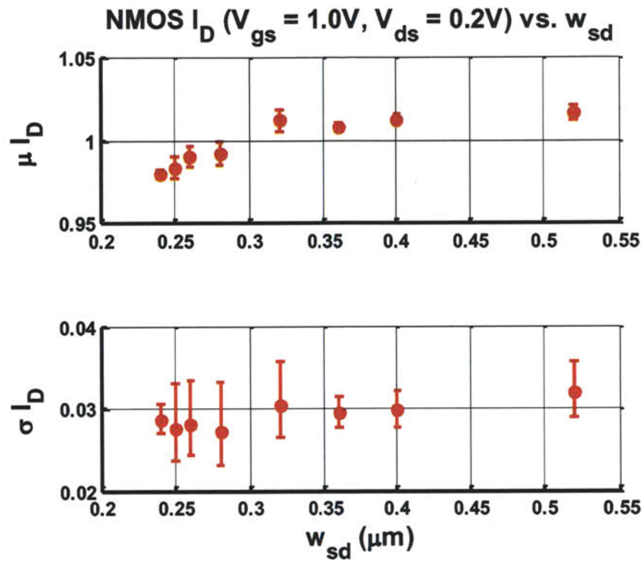


Figure 3-28: Correlation coefficients between measured device currents at various operating points and measured contact plug resistance. Correlations are strongest in the linear region of operation (low values of V_{gs} and high values of V_{ds}).

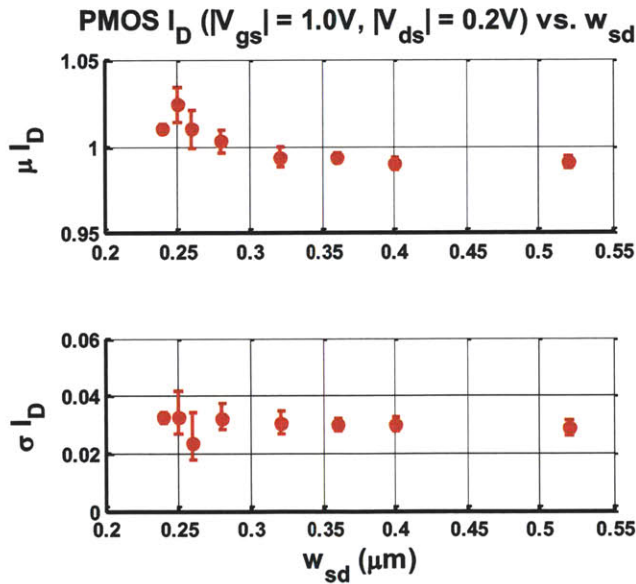
Future DOEs of this design may include more replicates of transistors with identical source-drain widths but different contact-to-gate and contact-to-diffusion edge distances in order to obtain more measurement samples which focus on the possible impact of the contact plug on device current.

3.7 Need for Variability Models

Results from the previous section motivate the need for variability models which can capture systematic effects in contacts accurately, including small but important choices made in the local layout. As technology scales, parameters such as contact resistance are becoming more critical to the operation of the transistor. The issues of increasing contact plug resistance and higher variability have presented concerns that scaling past the $65nm$ node while continuing with the same contact plug process steps may result in intolerably high resistances [66]. While efforts have been made to improve the materials and process steps which are used to form these contacts, parallel efforts are needed to model the variability present in these contacts. Advances in both numerical modeling and compact modeling are necessary to capture these variations.



(a) NMOS DUTs.



(b) PMOS DUTs.

Figure 3-29: A plot of device current as a function of w_{sd} demonstrates that, while I_d is shown to be correlated with the contact plug resistance, the cause is due to unintentional stress which is a function of the distance from the gate to the STI edge.

Current state-of-the-art device models for transistors such as BSIM are lacking in their ability to adequately model contact resistance variability as part of the transistor. While some functionality exists in terms of choosing some geometric parameters such as the type of contact (wide or point contact) and type of connection to the source/drain regions (isolated or shared), a more accurate variation-aware model will be needed to capture the impact of all these layout configurations on variability. In addition, many layout parasitic extraction tools do not consider the impact of the source/drain contacts during their analysis.

In terms of analytical or numerical modeling of contacts, work has been done to model the total extrinsic resistance of a MOSFET [67] [68]. In addition, analytical models have been derived for accurately determining the parasitic capacitances associated with the source/drain contacts [69]. The analytical models were verified by 3-D Monte Carlo simulation results. Work has also been done to perform sensitivity analysis of contacts with respect to different geometries using device simulations [70]. In addition, methods for fast variation-aware extraction of capacitances have been proposed which effectively take into account geometric perturbations [71]. Fast variation-aware extraction tools such as these will be necessary in the future due to increased variability and aggressive scaling. More work is necessary, however, to generate accurate variability-based models for contacts and to integrate them into the design framework.

3.8 Summary

Because of the growing impact of contact variability in advanced technologies, it is necessary to understand the nature of this variability. While several methods have been developed to study individual contact resistance, as well as to understand failures and defects, new methods are necessary to accurately analyze the parametric variability in these structures. In this context, a test structure has been designed, fabricated, and measured which enables the characterization of contact plug resistance variability. The application of statistical analysis techniques reveal both die-to-die

and within-die variability, as well as wafer-level edge effects, affecting contact plug resistance. In addition, spatial correlation analysis was performed to uncover further possible trends in the data. For future technologies, it will become necessary to have adequate numerical and compact models for robust design.

Chapter 4

Array-Based Test Structure for AC Variability Characterization

Variability in FET devices and interconnect have become an increasing concern with technology scaling [72]. Traditional sources of variability in devices, e.g. saturation current, threshold voltage, and channel length, have been well-studied and characterized [9][73][19]. However, with further scaling and technology development, the presence of other sources of variability is possible. Some of these other sources may only be seen at high frequencies or at short time domains [74]. Thus, they may not be captured by the measurement of the aforementioned device characteristics. Examples of some device parasitics whose variability may cause such an effect are shown in Figure 4-1.

The work presented in this chapter addresses this problem by designing a simple test circuit which specifically focuses on the measurement of device delays which are greater than those attributable to known DC device parameters. The results of these measurements will be comprised of a histogram of propagation delays through three different types of devices under test (DUTs). These results can then be compared with those generated from the simulations using models developed for known DC effects.

Existing techniques for on-chip AC device characterization include ring oscillator frequency measurements and S-parameter measurements using a network analyzer

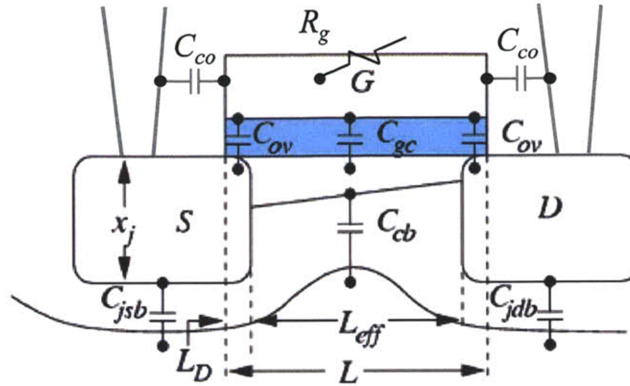


Figure 4-1: Some AC-relevant parasitics in a conventional MOSFET. The characteristics of such parameters are difficult to capture by performing DC measurements on the transistor, and therefore other characterization techniques which involve transients or high frequency operation are necessary.

[75][76][77]. For the case of the ring oscillator frequency measurements, the measured output parameter, frequency, is usually a function of the average delay among all the ring oscillator stages, rather than a single transistor delay. Therefore, the goal of gathering statistical data, which can reveal information about the performance of many nominally identical individual transistors, is difficult to achieve. When using a network analyzer to perform S-parameter measurements on a single transistor, pad limitations make it difficult for hundreds of transistors to be measured. In addition, measurement setup is often time-consuming and difficult. Obtaining statistics which can reveal information about variability is also difficult using this technique. Charge-based capacitance measurement techniques have also been used to measure MOSFET C-V characteristics [78], but a pure gate capacitance measurement may not provide a complete representation of the device AC characteristics. The approach for designing a test circuit which is capable of characterizing AC device variations is shown in Figure 4-2.

This chapter presents a simple test circuit which successfully measures the individual delays of each transistor in a large array. Section 4.1 presents the array-based test circuit, which has been designed in such a way that the overall delay reflects the AC characteristics of the device under test (DUT) rather than the DC characteristics of the DUT or the characteristics of other transistors or interconnect along the delay

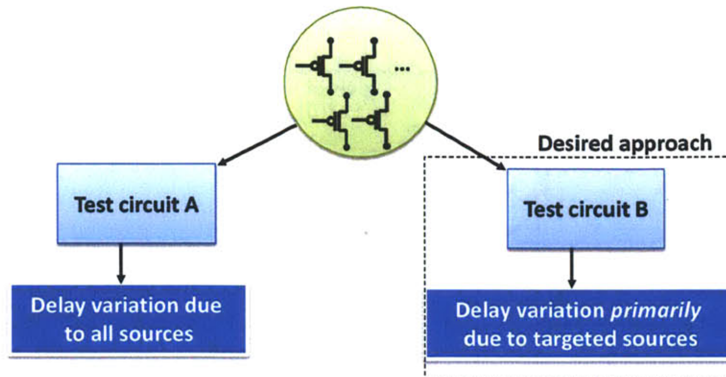


Figure 4-2: A proposed test circuit design approach which measures delay variation among multiple transistors, but for which the delay is primarily due to targeted AC variation sources rather than all sources including DC sources such as threshold voltage and channel length.

path. By optimizing the number of devices in the array and properly sizing the auxiliary devices used for signal propagation and switching, simulation results show that over 90% of the variability measured will be attributable to DUT variation rather than other device or interconnect variation. In addition, the AC characteristics of the DUT will be primarily reflected in the measured output data rather than traditional DC characteristics.

4.1 Array-Based Test Circuit

In this array-based test circuit, shown in Figure 4-3, an array of devices under test (DUTs) is implemented with the inputs and outputs of each DUT connected together. The input is fed by a clock source oscillator with the select signal of one DUT enabled and all others disabled. A scan chain is used to select the DUT which is to be measured. A delay detector, described in Section 5.4, performs a delay measurement for each DUT. Using these building blocks, the test circuit measures the relative delay of each DUT as compared to the other DUTs in such a way that the measured delay quantities primarily reflect the transistor's AC characteristics.

Three versions of DUT arrays are implemented. The first is a transmission gate array, shown in Figure 4-4. The second and third are NMOS and PMOS arrays,

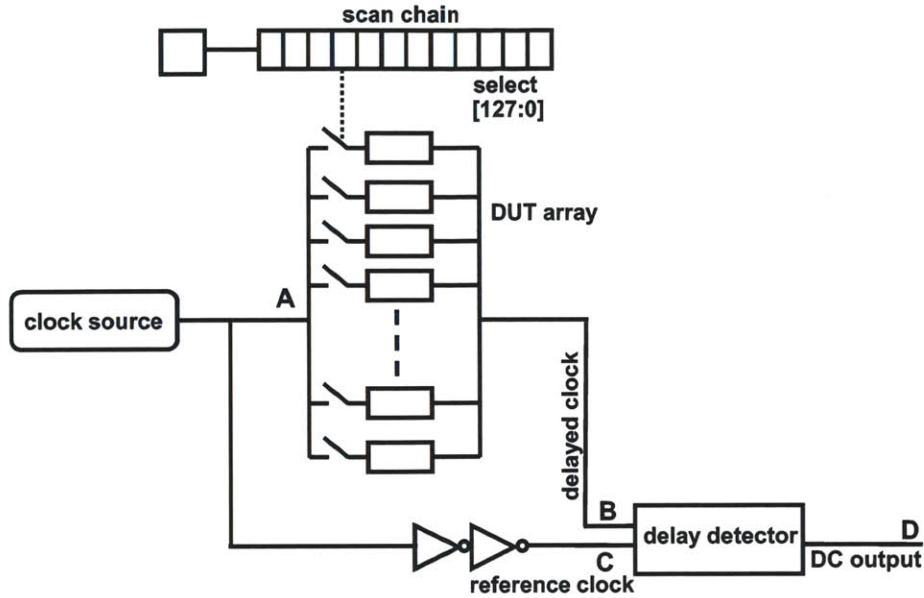


Figure 4-3: Array-based test circuit schematic consisting of a clock source, an array of DUTs, and a delay detector. The relative delay mismatches through all DUTs in the array are measured by comparing the arrival time of node B, the DUT output, with the arrival time of node C, a common reference.

respectively, with each DUT preceded by a transmission gate switch, shown in Figures 4-5 and 4-6.

4.1.1 DUT Array

The total delay difference between the clock source tapping point, labeled node A, and the delayed clock, labeled node B, can be represented by Equation 4.1.

$$D_{total} = D_{bufferedIC} + D_{DUT} \tag{4.1}$$

Because from node A, there is an optimized clock tree distribution with large buffers, any variability in these paths will not have a large impact on the total delay of the path. However, because of the small size of the DUT transistor as well as the large parasitic load which it must drive (the parasitic drain capacitances of the other DUTs and the interconnect), the DUT variability will have dominate the total delay of the path, as desired. Simulation results show that, when applying Monte Carlo-

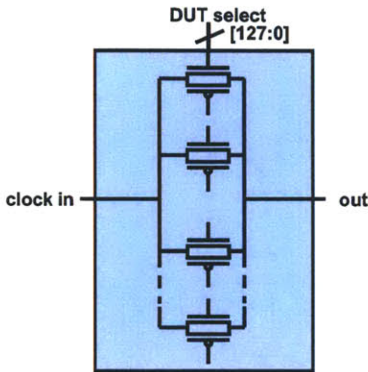


Figure 4-4: Schematic of a transmission gate DUT array. In this case, both the DUT select enable device and the DUT are the same transmission gate.

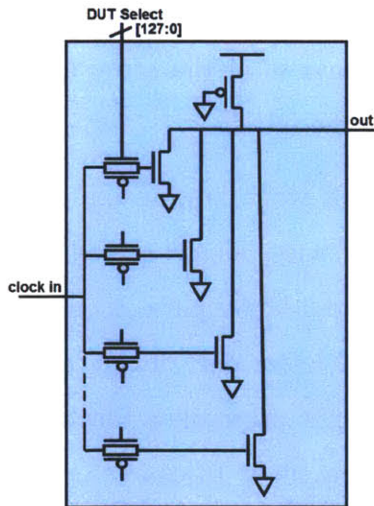


Figure 4-5: Schematic of an NMOS DUT array. The input clock has access to the gate of one of the NMOS DUTs, controlled by the DUT select input and the transmission gates, and the output node is connected to a weak PMOS pull-up transistor to enable the output to swing high.

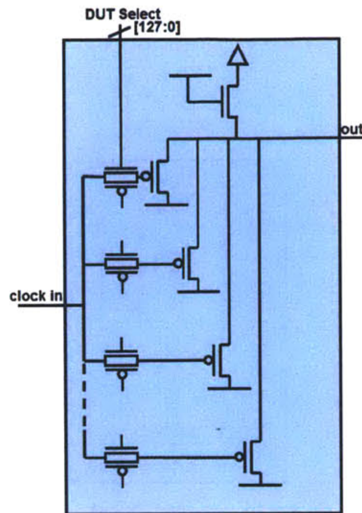


Figure 4-6: Schematic of a PMOS DUT array. The input clock has access to the gate of one of the PMOS DUTs, controlled by the DUT select input and the transmission gates, and the output node is connected to a weak NMOS pull-down transistor to enable the output to swing low.

based variation to the DUT, the overall delay changes by a factor of 60 more than it does when applying variation to all the other devices and interconnect RC values. Therefore, as the number of devices in the array increases, the more sensitive the overall delay of the path is to the DUT.

However, there is a tradeoff when adding too many devices, as illustrated by Figure 4-7. As the number of devices increases, the transition at the output node of the DUT becomes slower due to a larger parasitic capacitance on that node. This is problematic because during the latter part of the transition, the delay is dominated by the saturation current, which is primarily a function of DC transistor parameters. Because the goal of this test circuit is to characterize AC variability, which occurs on a short time scale, the interesting range of the transient response is just after the input switches at either the gate or the source. Therefore, it is advantageous to have the appropriate number of devices which both maximizes the sensitivity of the overall delay to DUT variation and maximizes the impact of DUT AC variability as compared to DC variability. The number of DUTs in the array has been optimized to achieve this, as discussed next for each type of DUT. Quantitative optimization

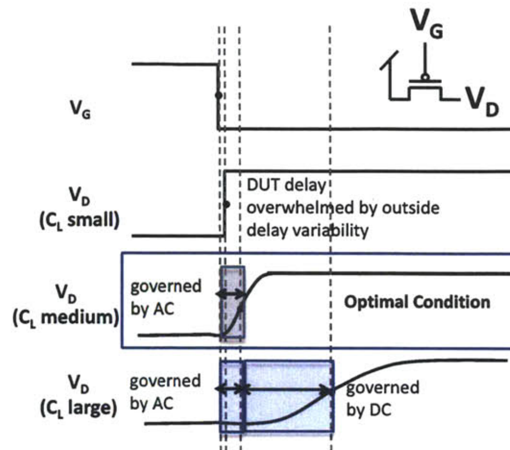


Figure 4-7: Tradeoff involving number of DUTs in array versus AC variability captured. A large number of DUTs results in a large load capacitance, which makes the overall transition at the drain dominated by DC variation sources. On the other hand, a small load capacitance results in a DUT delay which is too small and whose variability can be overwhelmed by external variation sources.

results are shown in Section 4.1.2.

Transmission Gate DUT Array

For the transmission gate DUT array, an array of transmission gate devices is used. The W/L ratios of both the NMOS and PMOS transistors in the transmission gates are 32. These dimensions maximize the amount of possible AC variability between these devices while minimizing the amount of threshold voltage, channel length, and saturation current variability between them. The transmission gate is used as a DUT cell in itself because simulation results show that it will yield information regarding the AC variability of the combined pair of NMOS and PMOS devices which comprise it. In addition, this array is required because they become switches for the NMOS and PMOS DUT arrays, but they may have small delay variations of their own.

NMOS and PMOS DUT Arrays

For the NMOS DUT array, an array of NMOS devices is used, with each device preceded by a transmission gate switch. When a DUT is not selected, the transmission

gate is switched off so that no current passes through the DUT. When the DUT is selected, the transmission gate is turned on and it acts as a DC switch. A shared PMOS pull-up device, which is always drawing current, is employed in order to ensure that the output swing starts from the supply voltage for the high-to-low output transition. Because the pull-up device is shared among all the DUTs, the same amount current is drawn for each of the devices as it is being measured. And thus, variability is not introduced due to the pull-up device. The NMOS DUTs are sized with a W/L ratio of 32, while the PMOS pull-up device is sized with a W/L ratio of 6. The DUT sizes are chosen such that the relative magnitude of possible AC variation as compared to DC variation would be large. The PMOS DUT array is simply the dual of the NMOS DUT array. While the statistics obtained from this array will represent a convolution of the transmission gate variance and the NMOS or PMOS DUT variance, because the transmission gate is driving such a small load (only the DUT), the overwhelming proportion of the variance in delay (over 99%) will be due to the DUT variance.

4.1.2 Design Optimization for AC Variability Measurement

In order to quantitatively arrive at the optimal condition shown in Figure 4-7, a design optimization is performed on the number of DUTs in an array. If the number of DUTs in the array is too small, the amount of replication will be insufficient for adequate statistical confidence. For larger array sizes, two 1,000 point Monte Carlo simulations are performed for each DUT array size. In the first, a DC variability model is used for the DUTs. In the second, both a DC variability model and an AC variability model are used for the DUTs. The AC variability model is based on the possible range of AC effects within the transistor for which characterization is desirable. The variance in DUT delay is determined for the case of each model applied. Figure 4-8 plots the percentage of variability reflected by AC variation in the DUT delay as a function of the number of DUTs in the array. Because there is a sharp drop-off in the fraction of total variation that is due to possible AC effects as a function of the number of DUTs in the array, the optimal point to choose for

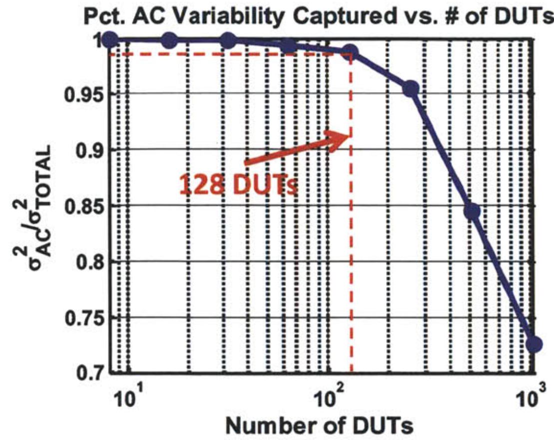


Figure 4-8: DUT array optimization for AC variability characterization shows that 128 DUTs ensures that 98% of the variance in delay is attributable to AC variation sources.

the number of DUTs occurs just before this drop-off. This way, we have sufficient replication for statistical significance, sufficient delay through the DUT such that it is not overwhelmed by outside variations, and the capturing of variability which is primarily due to possible AC effects (over 98%). For this reason, 128 DUTs are used in each array.

To illustrate this example further, Figure 4-9 shows simulation results of delay distributions due to DC variation sources and DC and AC variation sources for two different array sizes: 8 DUTs and 1024 DUTs. In both histograms, many outliers of large DUT delays which exist in the case of all variation have not been plotted so as to show the main trend more clearly. In the case of 8 DUTs, the AC characteristics distinguish themselves from the DC characteristics when observing the delay distribution of DUTs. This means that the measured results from the test structure would reveal the presence of any AC variability within the transistors down to the extent to which it has been modeled for these simulations. In contrast, the presence or absence of AC variability does not change the distribution of delays very much in the case of 1024 DUTs, which means that the test structure would not perform nearly as well in terms of decoupling possible AC variability from known sources of DC variability.

This can be seen more clearly in Figure 4-10. Drawn in each plot is a red line which

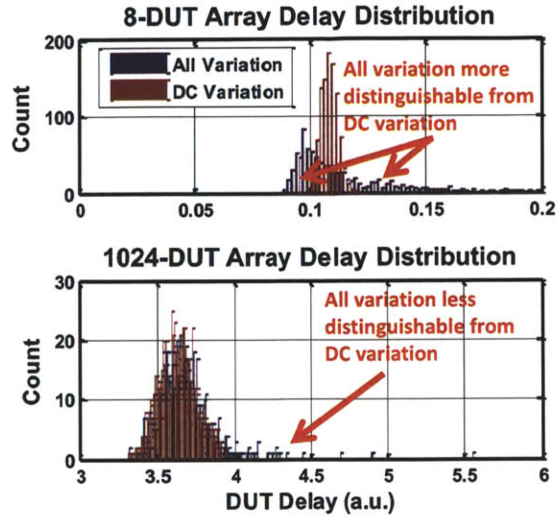
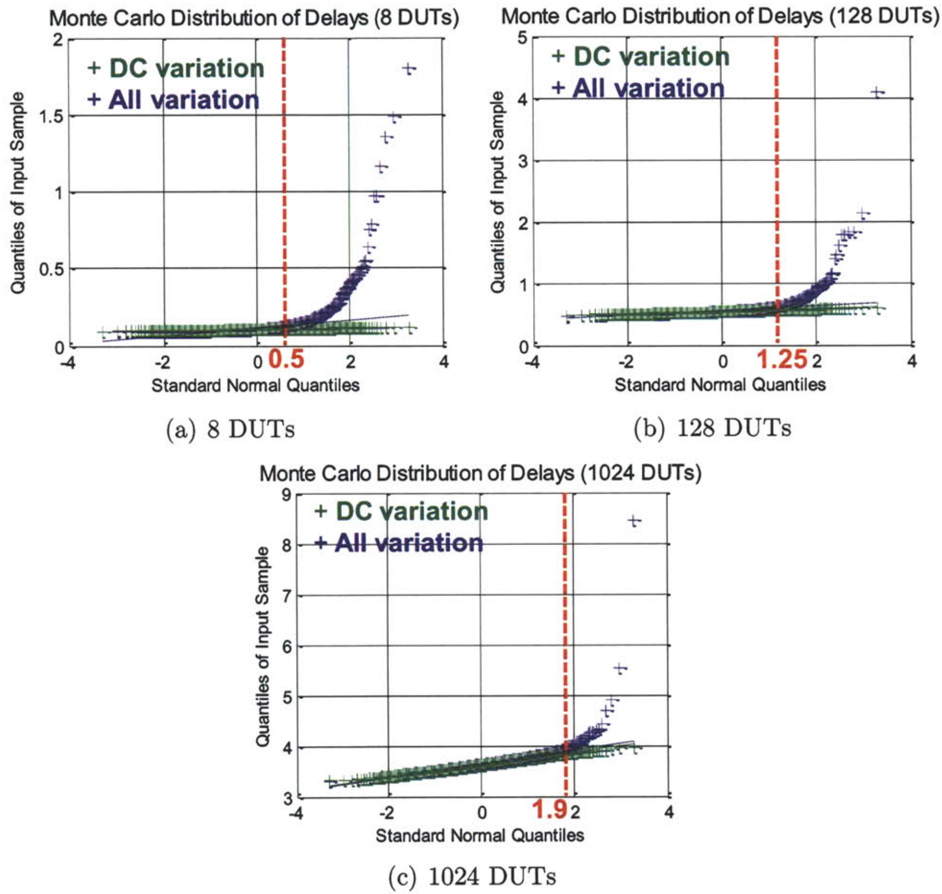


Figure 4-9: Simulated DUT delay distributions for different array sizes and variability sources. In the case of 8 DUTs, the distribution of delays when AC variation sources are imposed differs significantly from that when only DC variation sources are imposed. However, in the case of 1024 DUTs, the distributions are more similar to each other.

qualitatively describes the standard deviation at which the delay distribution of the DUT when subject to all variation sources separates itself from the delay distribution when only subject to DC variation sources.

4.1.3 Signal Propagation

Because of the need to minimize any variability due to the differences in devices and interconnect along the signal path, a balanced H-tree with optimized buffers is implemented for the input signal path. As shown in Figure 4-11, two stages of inverters are used in addition to an H-tree distribution in order to equalize as much as possible the delay between the input node, A, and any of the output branches. Simulation results using variability models which arise from known DC sources for both the buffers and the interconnect lines show that the variance in delay due to the mismatch in this H-tree signal distribution is less than 5% of the total variance, which is dominated by the variability in the DUT. This also includes the mismatch between the slew rates at the various output branches due to the variability in drive



(a) 8 DUTs

(b) 128 DUTs

(c) 1024 DUTs

Figure 4-10: Variation in relative delay as a function of number of DUTs - DC variations imposed versus all variations imposed. The point at which the distributions deviate from one another is qualitatively marked in red.

strengths of the buffers.

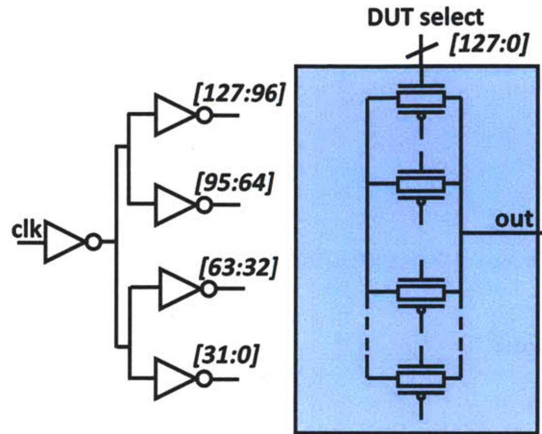


Figure 4-11: A buffered H-tree for input signal propagation into transmission gate array.

On the output side, no buffering is required since the large parasitic capacitance is beneficial towards maximizing the sensitivity of the measured delay to the DUT variability. Therefore, a simple interconnect H-tree is used to connect the drains of all the DUTs together. By using such a tree, the signal propagation delay from the output of any DUT to the shared output terminal will be matched well enough so that the resulting mismatch does not overwhelm the DUT variability. Simulations results show that the variance in delay resulting from the mismatch in such an H-tree is less than 5% of the total delay variance. Once again, this confirms that the predominant source of delay variability in this test circuit is the DUT.

4.1.4 Delay Measurement Circuit

For each array of DUTs, the delay measurement is performed as shown in Figure 4-12. Delay measurements from the input to the output of each DUT are made in the following manner. Two signals are observed at the output of the array. The first comes from the connection to the DUT outputs, while the second is a delayed version of the common clock source input. Because these two signals have the same frequency and only differ in that one is a time-delayed version of the other, one can determine the delay differential between them by connecting them as the two inputs to a logic

gate. The output of the logic gate is a pulse whose duty cycle is proportional to the delay difference between the two signals, and therefore the logic gate output can be filtered by a simple RC network, and the resulting DC voltage can be measured off-chip. When this measurement is performed for all DUTs, the relative delay variation between the DUTs can be quantified. For the transmission-gate and PMOS arrays, a

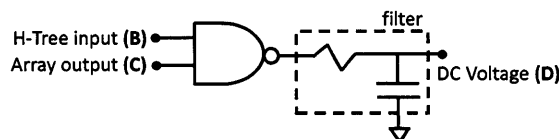


Figure 4-12: Delay measurement technique using a logic gate followed by a first-order low-pass RC filter (NAND can be replaced with NOR depending on whether the falling or rising edge needs to be characterized).

NAND gate is used in order to capture the delay through the DUT which involves the rising edge at the output node of the DUT. For the NMOS array, a NOR gate is used in order to capture the delay through the DUT which involves the falling edge at the output node of the DUT. One logic gate input is tapped directly from the beginning of the input signal H-tree and then delayed by an inverter chain which is shared by the entire array. The other logic gate input is tapped from the shared node of the output signal H-tree. The inverter chain is long enough so that the transition at node A does not occur until the transition at node B, which comes from the output of the DUT array, is complete. Because the logic gate is shared among all the DUTs within an array, the offset caused by the propagation delay through it is the same for all DUT selections. The output of this gate is a periodic waveform whose duty cycle is directly proportional to the delay between nodes B and C. Because the variability in this delay for different DUTs is dominated by the DUT variability itself, measuring the duty cycle of this waveform for every DUT will result in a statistical distribution of delays which will reveal the DUT variability.

The duty cycle of the signal at the logic gate output is measured by using an RC filter. A 190kΩ on-chip resistor is used along with an on-chip MOM capacitor with sufficiently large value to provide enough filtering to limit the output swing to 20mV peak-to-peak. The drive strength of the logic gate is large enough that the voltage

swing directly at its output node is within 0.5mV of the rail values. This ensures that the accuracy of the filtered DC voltage value is not significantly affected by any resulting nonlinearities due to insufficient drive strength of the logic gate.

The relevant waveforms in the delay measurement scheme are shown in Figure 4-13. The labels A, B, and C refer to the nodes labeled in Figure 4-3.

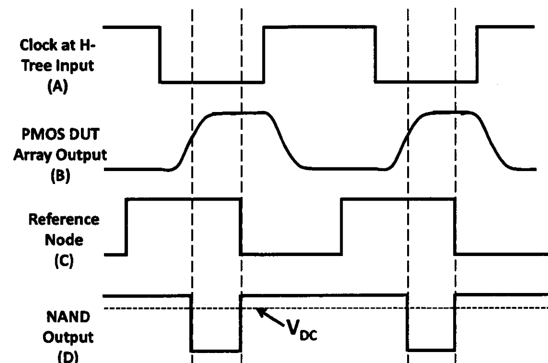


Figure 4-13: Waveforms describing the operation of the delay measurement technique. Nodes B and C are the inputs to a NAND gate, whose output is shown in D. The low-pass filter then produces an average DC voltage, V_{DC}.

4.1.5 Measurement Setup and Methodology

The measurement setup and methodology is as follows. An off-chip clock source will be sent onto the chip and into the H-tree signal propagation tree. Since the frequency of this signal is known a priori, it does not need to be measured on the chip. The limitations on the clock frequency depend on the rise and fall time at the heavily loaded DUT array output, which in the case of the NMOS and PMOS DUT arrays, is made larger by the pull-up and pull-down devices, respectively. The frequency to be used for the transmission gate array is 1GHz, while the frequency to be used for the NMOS and PMOS DUT arrays is 100MHz. Then, one of the the DUTs in an array will be selected using the scan-chain setup and the average voltage of the output of the duty cycle measurement logic gate will be measured. Because the measurement of delay variability among many devices is the goal of this test chip, absolute delay through a single DUT is not characterized. The inverter chain is shared and has the

same delay for all DUTs and the H-tree signal propagation tree has minimal delay variation along its branches. Therefore, the delay can be measured using Equation 5.5, where V_{DD} is the supply voltage and V_{SS} is the ground voltage, V_{DC} is the measured average DC voltage, and T is the period of the input signal.

$$DELAY = T\left(1 - \frac{V_{DC}}{V_{DD} - V_{SS}}\right) \tag{4.2}$$

As an example, when the input clock source is running at a frequency of 1GHz and the supply voltage is 1V, the sensitivity of this measurement technique is 1 ps/mV.

4.1.6 Measurement Accuracy

The accuracy of the delay measurement technique is limited by the NAND or NOR delay detector which is shared at the output of all the DUTs. Because the logic gate does not have infinite current drive, the linearity of the average DC voltage value with respect to the duty cycle at the input is dependent upon two parameters. First is each of the pull-up and pull-down drive strengths of the delay detector logic gate (either NAND or NOR). Second is the value of the resistor immediately following the NAND gate. Because the resistor is designed to have a significantly large value, the main limitation in accuracy comes from the finite drive strength of the gate. Figure 4-14

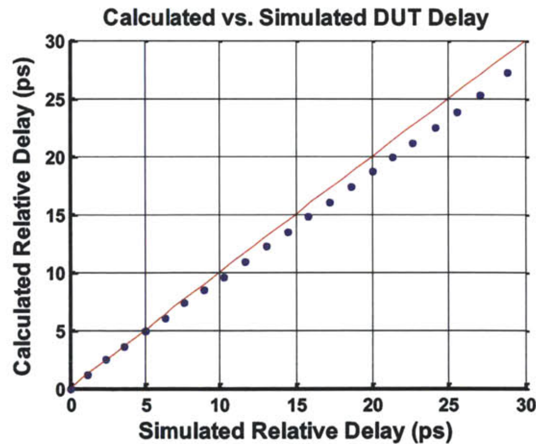


Figure 4-14: Limitations on accuracy of delay measurement technique. For up to 30ps of relative delay mismatch, the error in measurement is bounded by 2ps.

shows a plot of the delay calculated based on observing the average DC voltage at the output and using Equation 5.5 and comparing its value to the actual simulated delay between the DUT output and the reference node. The blue points show the results of the simulations, while the red line shows the ideal case of perfect measurement accuracy. The delay measurement is accurate to within 2ps for up to 30ps of relative delay variations.

4.1.7 Test Chip

A test chip has been implemented in an advanced IBM CMOS SOI process technology. Each of the three array banks contain 128 DUT cells and 4 dummy DUT cells to ensure similar surrounding areas for each DUT. The size of each DUT cell is 1.44μm x 1.00μm. The test circuit layout, shown in Figure 4-15, occupies an area of 400μm x 20μm, with each of the three DUT arrays occupying one-third of the total area. The three on-chip diffusion resistors occupy an area of 100μm x 0.5μm each. The on-chip MOM capacitors, which use three metal layers, occupy an area 200μm x 4μm each. Each DUT is accessed using a scan chain-based approach. The test circuit has been designed using just four metal layers, enabling in-line measurements relatively early in the manufacturing process.

4.2 Summary

A test structure that measures the AC variability characteristics of MOSFETs in an advanced CMOS SOI technology has been designed. Simulation results show that over 90% of the measured variance will be attributable to the variability within the device under test rather than other devices or interconnect. Furthermore, the number of DUTs in the array has been optimized to ensure that the AC characteristics of the DUT transistors will be primarily reflected in the delay measurement output rather than DC characteristics. In addition, only a single DC voltage measurement is required for each device under test, which enables a simple characterization flow.

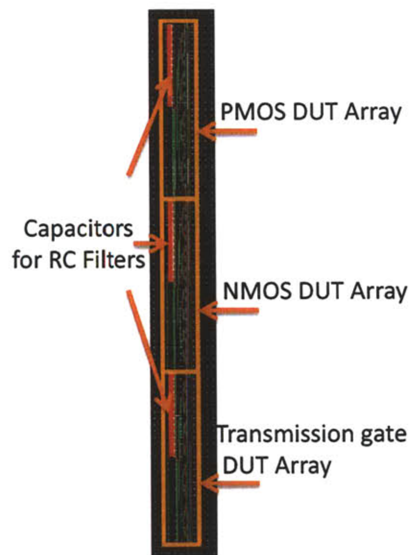


Figure 4-15: The array-based test circuit layout is divided into three blocks: PMOS DUT array, NMOS DUT array, and transmission gate DUT array. A scan chain is implemented in the vertical direction which controls DUT access for all blocks.

Chapter 5

Ring Oscillator-Based Test Structure

5.1 Introduction

Variability in FET devices have become an increasing concern with technology scaling [72]. Traditional sources of variability in devices, e.g. saturation current, threshold voltage, and channel length, have been well-studied and characterized [9][73][19]. However, with further scaling and technology development, the presence of other sources of variability is possible. Some of these other sources may only be seen at high frequencies or at short time domains [74]. Thus, they may not be captured by the measurement of the aforementioned device characteristics. This work addresses this problem by designing a ring oscillator-based test circuit which specifically focuses on the measurement of the difference between gate-to-drain and source-to-drain propagation delay in a transistor. The results of these measurements will be comprised of a histogram of propagation delay differences through different types of devices under test (DUTs). These results can then be compared with those generated from the simulations using models developed for known DC effects.

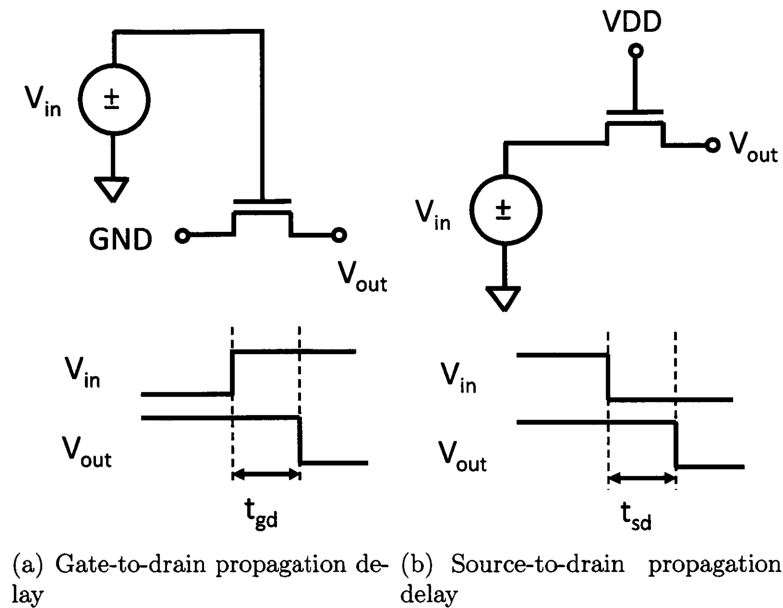
Existing ring oscillator-based techniques for on-chip device characterization are presented in [79][76][80][81]. In [79], the C-V characteristics of transistors are extracted by observing ring oscillator frequency, but the output represents the average

of many devices under test. In [76] and [80], the output also represents the average performance of multiple devices. In [81], a technique is used to determine the delay of a single inverting or non-inverting gate, and variability measurements are made on the delays of many such gates. For the ring oscillator frequency measurements for which the frequency is a function of the average performance of multiple devices, rather than that of a single device, gathering statistical data is difficult to achieve. For the case in which individual gate delays are measured, the delays depend on DC characteristics such as threshold voltage, channel length, and saturation current of the transistors, as well as AC characteristics. In this work, a circuit is designed whereby only the AC characteristics of a single transistor within a ring oscillator is measured. Furthermore, many replicates of the structure allow statistics to be obtained to determine AC variability.

When using a network analyzer to perform S-parameter measurements on a single transistor, as in [77], pad limitations make it difficult for hundreds of transistors to be measured. In addition, measurement setup is often time-consuming and difficult. Obtaining statistics which can reveal information about variability is also difficult using this technique. Charge-based capacitance measurement techniques have also been used to measure MOSFET C-V characteristics [78], but a pure gate capacitance measurement may not provide a complete representation of the device AC characteristics.

5.2 Transistor Propagation Delay

Observing the propagation delay of an ideal step input voltage between two terminals of a transistor is one way to characterize the high-frequency behavior of such a transistor. Two such metrics of propagation delay are illustrated in Figure 5-1 for an example NFET. A similar illustration can be shown for a PFET. When the source terminal of the transistor is tied to the ground voltage and the gate terminal is connected to a pulsed voltage, the delay between the rising edge at the gate terminal and the subsequent falling edge at the drain terminal can be described as the gate-



(a) Gate-to-drain propagation delay (b) Source-to-drain propagation delay

Figure 5-1: Propagation delay metrics which characterize the short time-scale behavior of a transistor.

to-drain propagation delay, or t_{gd}, of the transistor, as shown in Figure 5-1(a). In this case, the initial operating condition of the transistor of interest is $V_{GS} = 0V$ and $V_{DS} = VDD$. Similarly, when the gate terminal is connected to the supply voltage and the source terminal is connected to a pulsed voltage, the delay between the falling edge at the gate terminal and the falling edge at the source terminal can be described as the source-to-drain propagation delay, or t_{sd}, as shown in Figure 5-1(b). In this case, the initial operating condition of the transistor of interest is $V_{GS} = 0V$ and $V_{DS} = V_t$ since there is a threshold voltage drop between the source and drain nodes when there is no current through the transistor. These two metrics provide slightly different information regarding the high-frequency characteristics of the transistor. This is due to the slightly different parasitic capacitances and resistances which are involved with each metric. However, some transistor parameters such as threshold voltage and channel length are likely to influence both metrics in a similar manner and to a similar degree.

For this reason, a new metric, t_{meas}, is defined in Equation 5.1 for the case of a PMOS device, where the propagation delays involving low-to-high transitions at the

drain node are considered.

$$t_{meas} = t_{gd,H} - t_{sd,H} \tag{5.1}$$

By measuring t_{meas}, the relative sensitivity to variations in AC parameters is magnified, while the relative sensitivity to variations in DC parameters such as threshold voltage and channel length is reduced, as will be discussed in detail in Section 5.6.

5.3 Test Circuit Description

A ring oscillator is implemented with a device under test (DUT) acting as a pass gate between two of the stages in the oscillator, as shown in Figure 5-2. A PMOS DUT is used in this example and all forthcoming discussions will assume a PMOS DUT. However, an NMOS DUT can be substituted by simply changing the weak pull-down transistor to a weak pull-up transistor and by connecting the *pass* switch to V_{DD} instead of ground. The ring oscillator operates in two modes: *pass* and *wait*. In the

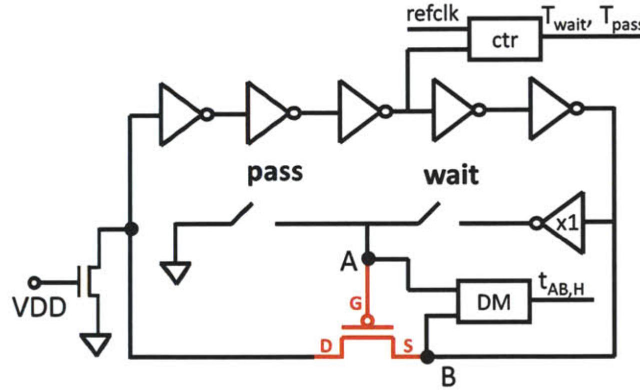


Figure 5-2: Ring oscillator-based test circuit for AC variability characterization, which operates in two modes and requires two clock period measurements and a delay measurement in order to characterize the DUT.

pass mode, the *pass* switch is enabled and the *wait* switch is disabled, so the gate of the DUT is always on and it acts as pass gate. A weak NMOS pull-down device is connected to the drain terminal of the DUT in order to assist the PMOS DUT in passing high-to-low transitions to its drain node. The relevant signal waveforms during the *pass* mode are shown in Figure 5-3. Because the gate of the DUT is always

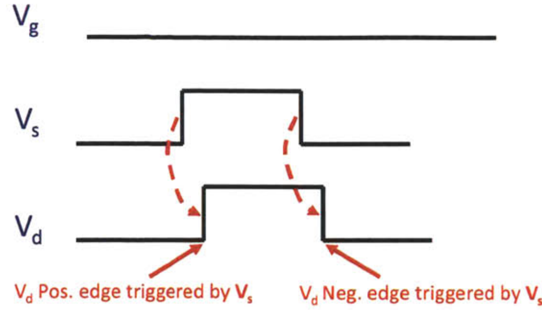


Figure 5-3: Waveforms at PMOS DUT terminals during *pass* mode, in which both transitions at the drain of the DUT are triggered by transitions at the source of the DUT.

on, both low-to-high and high-to-low transitions at the drain of the DUT are triggered by transitions at the source of the DUT. The RO period of oscillation during the *pass* mode can be characterized by Equation 5.2.

$$T_{pass} = \sum (t_{inv,L} + t_{inv,H}) + t_{sd,H} + t_{sd,L} \tag{5.2}$$

In the *wait* mode, the *pass* switch is disabled and the *wait* switch is enabled. In this mode, the gate of the DUT is controlled by a delayed and inverted version of the input to the source. When there is a high-to-low transition at the source, the gate is on and as a result there is a high-to-low transition at the drain as well. However, when there is a low-to-high transition at the source, the gate is off and so the drain does not switch from low-to-high. Instead, the signal at the drain remains low and *waits* until the gate turns on as a result of the path through the inverter or chain of odd-numbered inverters, *x1*. When the gate turns on, the drain finally then switches from low-to-high. Figure 5-4 illustrates the waveforms during the *wait* mode operation. The RO period of oscillation during the *wait* mode can be characterized by Equation 5.3, where $t_{AB,H}$ refers to the delay between the signal at the DUT source going high and the signal at the DUT gate going low.

$$T_{wait} = \sum (t_{inv,L} + t_{inv,H}) + t_{gd,H} + t_{sd,L} + t_{AB,H} \tag{5.3}$$

Substituting Equations 5.2 and 5.3 into Equation 5.1 gives the relationship between

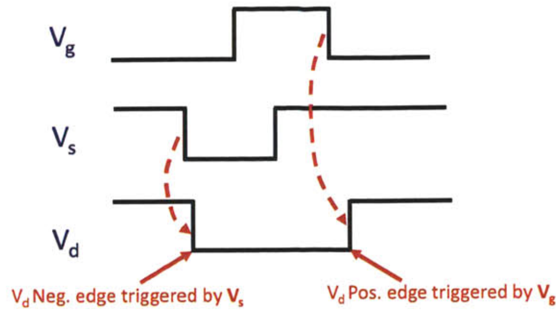


Figure 5-4: Waveforms at PMOS DUT terminals during *wait* mode, in which one transition at the drain of the DUT is triggered by a transition at the source of the DUT, while the other transition at the drain of the DUT is triggered by a transition at the gate of the DUT.

the desired quantity t_{meas} and the measurable quantities, shown in Equation 5.4.

$$t_{meas} = T_{wait} - T_{pass} - t_{AB,H} \tag{5.4}$$

The DUT parameter t_{meas} is obtained by performing RO frequency measurements, either on- or off-chip, to measure T_{wait} and T_{pass}, and by performing a delay measurement, described in Section 5.4, to measure $t_{AB,H}$.

Multiplexing is performed among multiple RO blocks for statistical characterization of individual DUTs, as shown in Figure 5-5. Each RO is accessed using a scan

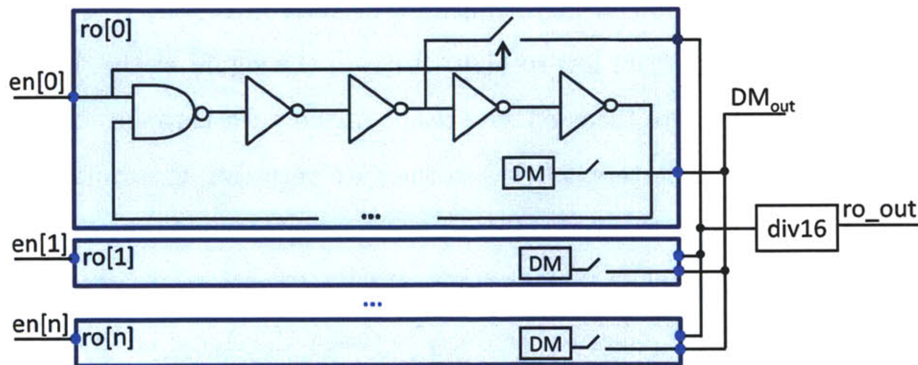


Figure 5-5: Array of RO blocks for statistical characterization of DUT AC performance. Both the ring oscillator period and delay measurement pins are shared outputs, while each RO is accessed through an enable signal controlled by a scan chain.

chain-based approach, with the RO enabled by using a NAND gate as a substitute

for one of the inverter stages. All ROs share a common frequency divide-by-16 block, whose output is propagated off-chip for frequency measurement.

5.4 Delay Measurement Circuit

The delay measurement circuit consists of a logic gate followed by an RC filter in order to obtain an average DC value, as shown in Figure 5-6. For the case of a

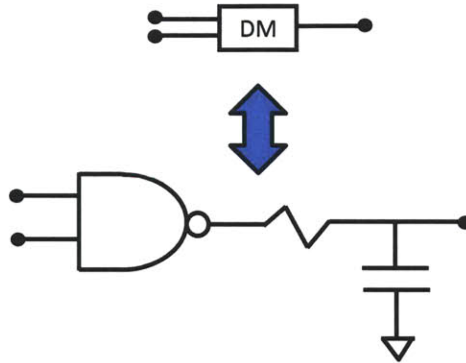


Figure 5-6: Delay measurement using a logic gate and RC filter, which converts the delay between two signals into a pulse whose duty cycle is proportional to the delay, and the converts the pulse into a DC voltage whose value is also proportional to the delay.

PMOS DUT, using a 2-input NAND gate with the two inputs being the gate and source terminals of the DUT, the output of the NAND gate will be a pulse whose duty cycle is directly proportional to the delay between the two signals. When this pulse is subsequently filtered by using the RC circuit, the average DC voltage will be proportional to the duty cycle of the NAND gate output. Therefore, the delay can be measured using Equation 5.5, where V_{DD} is the supply voltage and V_{SS} is the ground voltage, V_{DC} is the measured average DC voltage, and T is the period of either of the two input signals (both will have the same period).

$$t_{AB} = T\left(1 - \frac{V_{DC}}{V_{DD} - V_{SS}}\right) \tag{5.5}$$

Each ring oscillator, which contains one DUT, has its own delay measurement circuit which includes a 30kΩ diffusion resistor. However, the capacitance is shared

by all the DUTs and is comprised of parasitic wire and pad capacitances. The logic gate is sized such that the 3σ error in the delay measurement as a result of DUT-to-DUT logic gate mismatch is bounded by 0.5ps. The combination of the NAND drive strength and resistor value are sufficient to drive the NAND output to within 0.5mV of the rail values.

5.5 Test Circuit for Compensation of SOI Variations

In partially-depleted SOI technologies, the quantity t_{meas} will have a history effect component to it since the average floating gate-to-body voltage on the DUT during the *pass* and *wait* modes is different. Whereas in the *pass* mode for a PMOS device the gate voltage is held at 0V, in the *wait* mode the gate voltage alternates between 0V and V_{DD} with a duty cycle slightly greater than 50%. Thus, in order to produce substantially the same duty cycle during the pass mode while not altering the circuit operation, an XOR gate is configured between two of the ring oscillator stages with the output coupled to the *pass* switch, as shown in Figure 5-7. This artificially creates a duty cycle of about 50% for the DUT gate voltage but does not functionally alter the *pass* mode of operation. During those times where the output of XOR gate is high, an inverter and NMOS switch disables the pull down device and prevents the drain voltage of the DUT from being reset to ground in the pass mode. The relevant waveforms during the operation of this circuit is shown in Figure 5-8.

5.6 Simulation Results

In order to better understand and quantify the ring oscillator-best test circuit, simulations are performed for multiple reasons. All simulations of the ring oscillator-based test circuit are performed using an advanced PD-SOI CMOS technology. The discussion of the simulation results are divided into two parts. The first part describes simulations performed for two purposes: (a) to measure the accuracy of the output

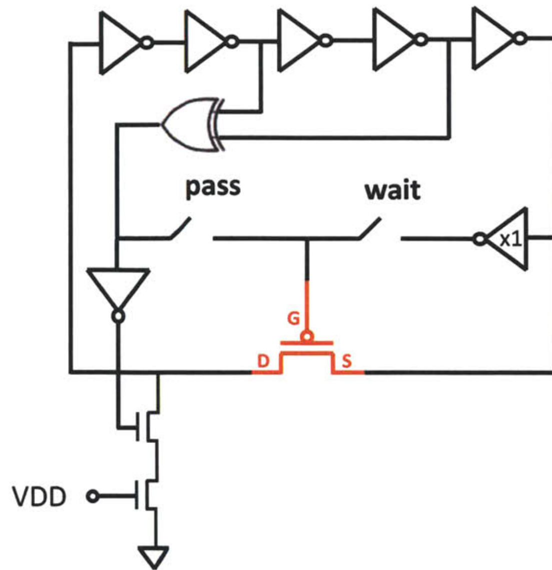


Figure 5-7: Modification of *pass* mode operation to minimize variations due to SOI history effect difference between modes. The XOR gate before the *pass* switch only affects the *pass* mode of operation, while leaving the *wait* mode of operation unchanged.

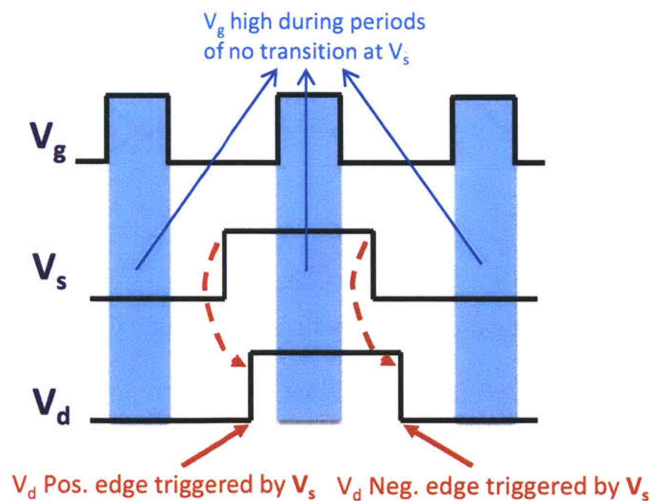
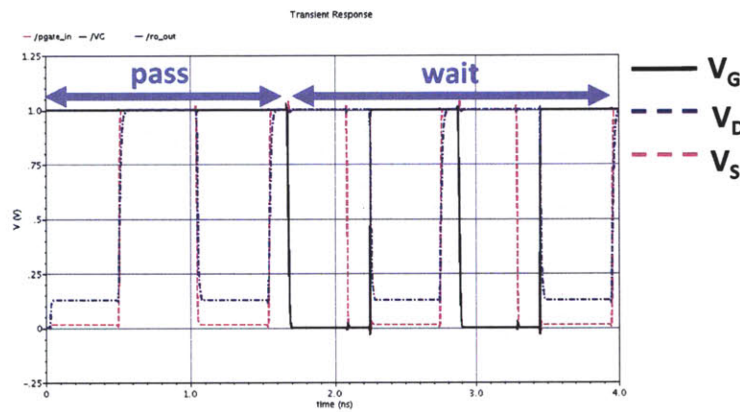


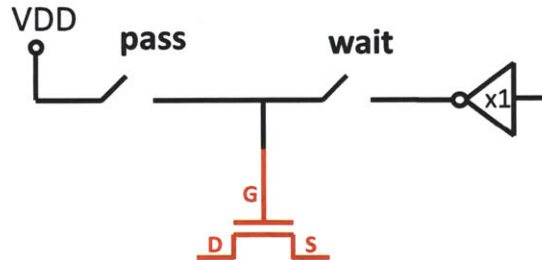
Figure 5-8: Waveform at PMOS DUT terminals during *pass* mode in SOI history effect-compensated test circuit. The average duty cycle of V_{gs} is similar in both the *pass* and *wait* modes of operation by creating periods during which time the DUT gate is switched off when it does not affect the propagation of the source signal to the drain.

t_{meas} derived from RO frequencies and a delay measurement, and (b) to verify the sensitivity of the measured output parameter, t_{meas}, towards AC variations as compared to DC variations from parameters such as threshold voltage and channel length. The next part is focused on determining the sensitivity of the output parameter, t_{meas}, in the absence of AC variations.

5.6.1 RO-Based Test Circuit Accuracy and Sensitivity



(a) NMOS DUT waveforms during *pass* and *wait* modes.



(b) Ring oscillator sub-block for NMOS DUT.

Figure 5-9: Ring oscillator waveforms for NMOS DUT type show how the transitions occur during the two modes of operation.

Figure 5-9(a) shows the simulation of an NMOS DUT in the configuration depicted by Figure 5-10(b). Here, gate of the DUT is always enabled during the *pass* mode, allowing the drain node to follow the source node without any external delays. In the *wait* mode, after the low-to-high transitions at the source and drain nodes of the DUT, the gate is disabled. Then, some time after the source node changes, the gate

once again goes high, allowing the drain node to go low. However, because of the pull-up device that is connected to the drain of the DUT, the drain does not transition all the way down to the ground voltage. Instead, the value at the drain node is approximately 130mV when it swings low. However, this does not have an effect on the measurement accuracy since the voltage to which the drain swings is nearly identical during the *pass* and *wait* modes. Shown in Figure 5-10 are the analogous waveforms for a ring oscillator containing a PMOS DUT. In this case, one difference from the ring oscillator containing the NMOS DUT is that there is a pull-down device connected to the DUT drain rather than a pull-up device. This causes the drain node to only reach approximately 0.9V rather than the supply rail of 1.0V. In addition, the high-to-low transition at the gate node triggers the drain node transition because the DUT is a PMOS device. However, the similarities between the NMOS and PMOS ring oscillators are enough that, based on the waveforms shown, the measurement technique and calculations for t_{meas} can be identical except for the need to change the edge directions.

To determine accuracy, the RO circuit is simulated with static power supply variation due to current differences in the *pass* and *wait* modes, mismatch between NAND gates as well as mismatch between RO stages. Results show that the difference between the directly simulated t_{meas} of the DUT and the derived t_{meas} using Equation 5.4 is bounded by 1ps.

For the sensitivity analysis, two sets of 1000-point Monte Carlo simulations are performed on the circuit. In the first set of simulations, only DC mismatch variations in parameters such as threshold voltage and channel length are imposed on the DUT. In the second set of simulations, both DC and AC variations are imposed on the DUT. Results show that 99.96% of the total variance in the output parameter t_{meas} is attributable to the input AC variation sources imposed on the DUT, while the remainder is attributable to DC variation sources. The AC variability model for the DUT is based on the possible range of AC effects within the transistor for which characterization is desirable. Figure 5-11 shows that the cumulative distribution of t_{meas} when subject to all variation sources is distinguishable at a value around 0.5σ

(a) PMOS DUT waveforms during *pass* and *wait* modes.

(b) Ring oscillator sub-block for PMOS DUT.

Figure 5-10: Ring oscillator waveforms for PMOS DUT type show how the transitions occur during the two modes of operation.

Figure 5-11: Quantile-quantile plot showing t_{meas} distributions under DC variations and all variations. The distribution of t_{meas} when the DUT is subject to all variation sources deviates from the case of only DC variation sources at a low standard deviation value.

when compared to that when subject to only DC variation sources.

In addition, a comparison with the array-based test circuit described in the previous chapter reveals that the ring oscillator-based test structure is much more sensitive to, and therefore more capable of measuring, AC variations. While 90% of the total variation in the output parameter is due to AC variations in the array-based test structure, almost all (99.97%) of output parameter variation in this work is due to AC variations in the DUT rather than DC variations. The tradeoff is with the simplicity of design. While the design of the array-based circuit is more straightforward and requires less time and simulation effort, the RO-based circuit is more complex and requires more careful simulation to ensure proper operation which allows for high sensitivity to AC variation sources.

5.6.2 Sensitivity of t_{meas} in the Absence of AC Variations

In the absence of significant on-chip AC variations from transistor to transistor, it becomes useful to examine the sensitivity of the RO-based test circuit to the remaining DC sources of variation. In order to do this, simulations have been performed which show how the variance of the output parameter, t_{meas}, relates to the variance of input device parameters, namely threshold voltage (V_T), channel length (L), and channel width (W).

Figure 5-12 shows the distribution of normalized t_{meas} values for the device under test when subject to different variation sources. For each histogram, a set of 1,000 Monte Carlo simulations has been done using mismatch variations only. When the DUT is subject to variations in only transistor width, the value of t_{meas} does not deviate substantially from its mean. Furthermore, when the DUT is subject to variations in only transistor length, the resulting variation in t_{meas} is slightly larger, but still relatively small. The largest spread in the value of t_{meas} occurs when the threshold voltage, V_T, is varied.

The same conclusions can be drawn from analyzing Figure 5-13. The normal probability plot of t_{meas} when subject to different DC variation sources indicates that the output parameter is most sensitive to threshold variations rather than channel

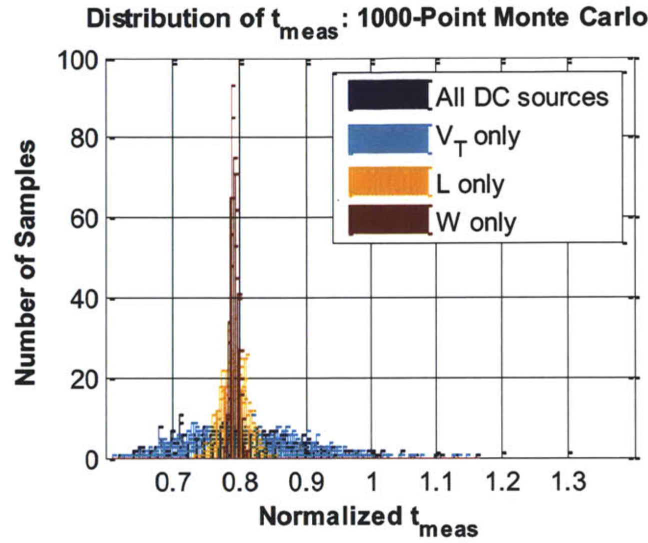


Figure 5-12: Simulation results showing a plot of the distribution of t_{meas} when only certain DC variation sources are present. These results indicate that threshold voltage is the parameter to which the output parameter t_{meas} is most sensitive.

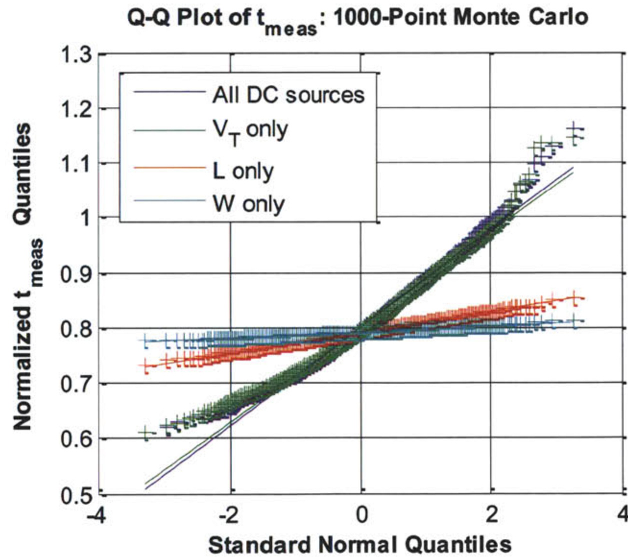


Figure 5-13: A normal probability plot showing the simulated decomposition of different DC variation sources and how sensitive t_{meas} is to each of them. Threshold voltage variation is the DC source predominantly captured by the output parameter, t_{meas}.

length or channel width variations. Therefore, in the absence of actual AC variations, the ring oscillator-based test circuit measurement output is likely to characterize variations in device threshold voltage. If AC variations are present, however, the test circuit is highly sensitive to these and enables characterization of these AC variations.

5.7 Test Chip

Two test chips have been implemented which contain this ring oscillator-based test circuit. One is in an advanced CMOS PD-SOI technology and occupies an area of $1600\mu m$ x $20\mu m$. The other is in a TSMC bulk 65nm CMOS technology and occupies an area of $1800\mu m$ x $50\mu m$. The forthcoming details regarding the test chip layout refer to the design in the advanced CMOS PD-SOI technology. The layout of a single RO is shown in Figure 5-14. The ring oscillator stages occupy significant area due to

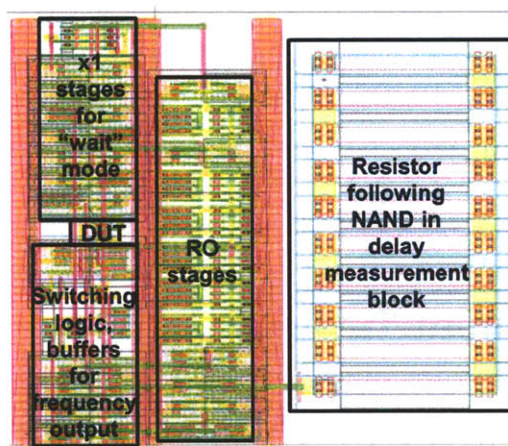


Figure 5-14: Ring oscillator layout showing the device under test (DUT), inverter stages, a logic block, and the resistor used for the delay measurement block.

the extra capacitive loads after each stage to slow down the period of oscillation. In addition, the resistor after the NAND delay measurement circuit also occupies a large area relative to that of the DUT. Four blocks, each containing 128 ring oscillator cells, are used. Both NMOS and PMOS DUTs used for each of the blocks are sized with W/L = 8, which is a common ratio used in standard cell logic for this technology. The four blocks are comprised of ROs containing NMOS and PMOS DUTs as well as

SOI-compensated ROs containing NMOS and PMOS DUTs, as shown in Figure 5-15. The test circuit has been designed using just four metal layers, enabling in-line measurements relatively early in the manufacturing process.

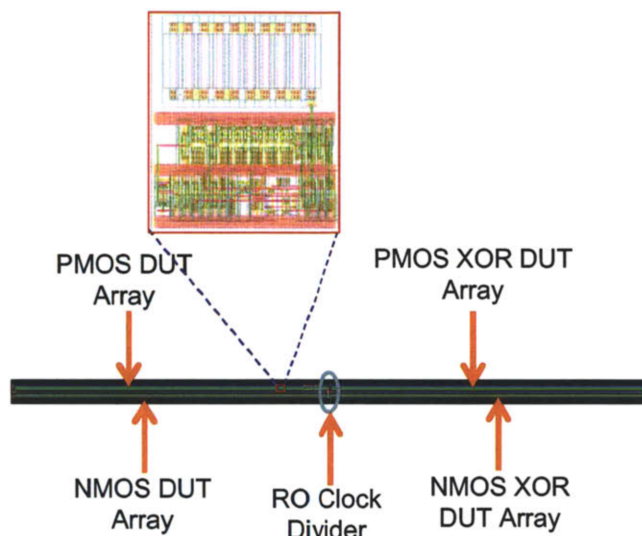


Figure 5-15: Test circuit layout which includes four ring oscillator blocks which characterize NMOS DUTs and PMOS DUTs in both standard and SOI-compensated configurations.

5.8 Measurement Results

Measurement results have been obtained for 40 die on a single wafer fabricated in a TMSC 65nm bulk CMOS technology. Because the technology is not SOI-based, the RO DUTs with the modification for SOI have not been included. Therefore, a block 96 ROs containing PMOS DUTs and a block of 96 ROs containing NMOS DUTs have been implemented. In each of the two blocks, two identifier ROs have been implemented. These identifier ROs have DUTs which contain a $30k\Omega$ p^{+}-polysicilicon resistor in series with the polysilicon gate. This is to replicate a possible source of AC variation which would be difficult to detect with either DC current-voltage measurements or a pure capacitance-based measurement. A schematic and layout of an example identifier RO are shown in Figures 5-16 and 5-17, respectively.

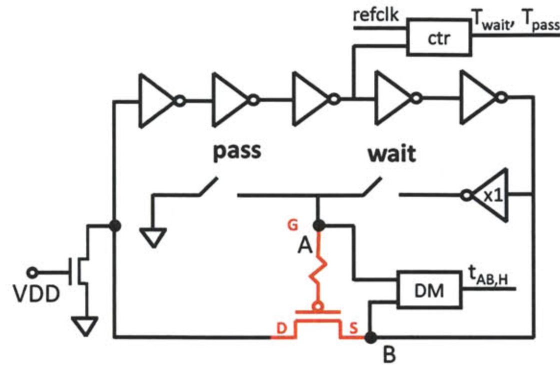


Figure 5-16: Schematic of identifier RO block, which includes an external gate resistor in series with the gate of the DUT.

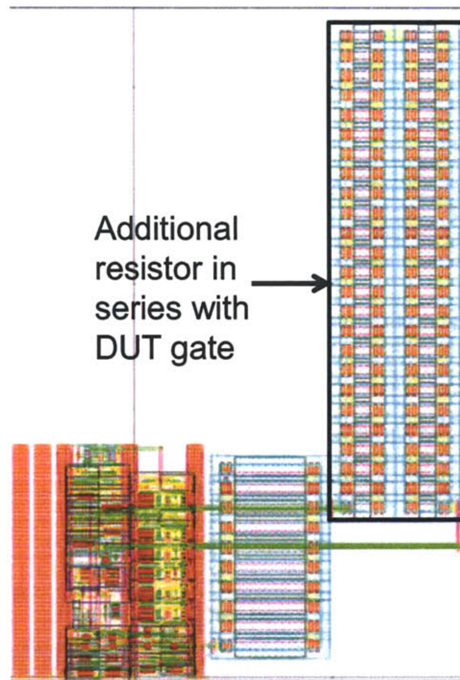


Figure 5-17: Layout of identifier RO block, which includes a $30k\Omega$ polysilicon resistor connected to the gate of the DUT.

5.8.1 Off-Chip Measurement Accuracy

The measurement accuracy for the RO clock periods is determined by a combination of the amount of time used to count, clock jitter, and possible clock drift due to external sources such as small physical movements in probe connections, drifts in the ground voltage due to sharing ground with an FPGA, and low frequency noise. The clock periods measured for a single DUT multiple times has a variance which is 2% of the variance in clock period across all DUTs. Therefore, the clock period measurement is replicable enough that the predominant source of variation seen in the measurements is actually due to the difference in DUTs rather than measurement noise. A similar analysis of the average DC voltage measurement shows that the variance due to measurement noise is 3% of the variance in average DC voltage across all DUTs. When the computation is performed to calculate t_{meas} for each DUT, the variance due to measurement noise increases to 5% of the total variance across all DUTs. These percentages can be reduced by noise mitigation techniques such as power and ground shielding both on and off-chip, as well as stabilizing any sources which result in clock period drift across millisecond time-scales.

5.8.2 Single-Die Results

A plot of the three output variables, T_{pass}, T_{wait}, and t_{AB}, are shown in Figure 5-18. Clearly, the two RO clock period measurements, T_{pass} and T_{wait}, are highly correlated across different DUTs because they are two periods measured from the same DUT. In addition, the clock period measured in the *wait* mode is always higher than that measured in the *pass* mode. This is due to the additional time that the signal has to *wait* at the source of the DUT for the gate of the DUT to turn on. Finally, t_{AB} is measured for each DUT by measuring the average DC voltage of the filtered NAND or NOR gate output and using Equation 5.5 to compute t_{AB}.

Because the variable which is highly sensitive to potential DUT AC variations is t_{meas}, its value is calculated for each DUT in the chip for PMOS devices and plotted in Figure 5-19. In this plot, the location two identifier devices for the PMOS block

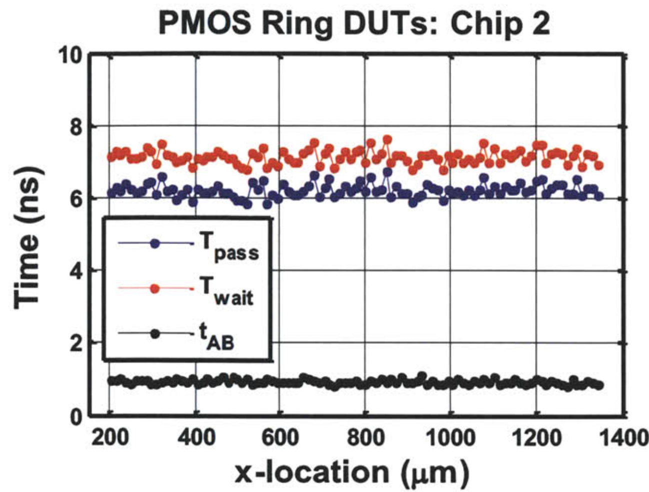


Figure 5-18: Measured output parameters for PMOS DUTs on a single chip show the values of the three measurement parameters for each DUT.

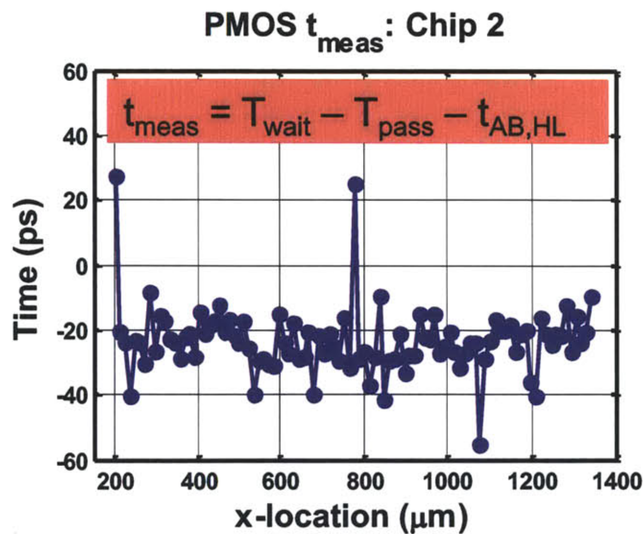


Figure 5-19: t_{meas} for PMOS DUTs on a single chip, calculated from the direct measurement results in Figure 5-18. Identifier DUTs which exhibit larger values of t_{meas} are clearly distinguishable from other data points.

is clear based on the large value of t_{meas} when compared to the other DUTs. This is expected behavior and matches simulation results because when a large resistor is placed in between the gate access point and the actual polysilicon gate, the gate-to-drain delay will increase much more than the source-to-drain delay. The increase in gate-to-drain delay is a direct result of the parasitic RC component generated from the external gate resistance. Any increase in source-to-drain delay, however, is due to a second-order feedback effect which slightly changes the gate voltage for a small period of time when the source transitions. This difference is captured by t_{meas} for the DUT measurements in the identifier ROs.

5.8.3 Wafer-Averaged Results

Figure 5-20 shows the average value of t_{meas} over all 40 die the DUT located at each x-coordinate. In this plot, the location of the identifier devices is more clear due to the averaging of measurement noise by the analysis of multiple measurements. In addition, the average value of t_{meas} is different for NMOS and PMOS devices, which is related to the device design points chosen in the technology for each of the two device types. Furthermore, two apparent systematic trends are evident in the data.

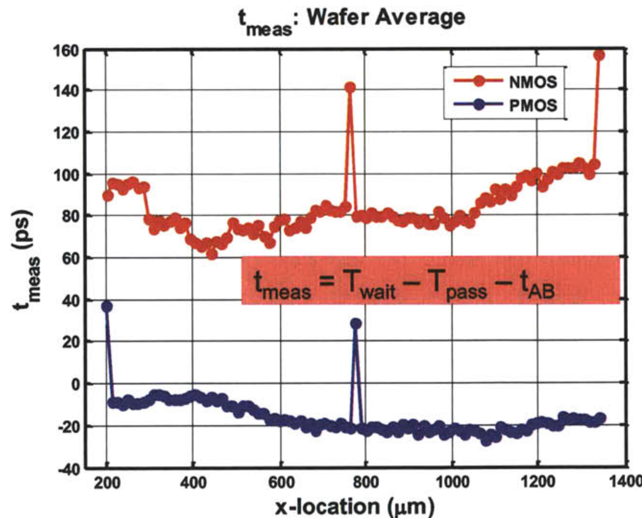


Figure 5-20: t_{meas} for all DUTs averaged over 40 die on wafer shows the presence of identifier DUTs more clearly, in addition to some weak systematic trends due to power supply variations.

First, a small trend exists where each consecutive group of 8 DUTs has a similar value for t_{meas}. This is likely due to the fact that, in the layout, the ROs are grouped in sets of 8, and each set of eight ROs shares a buffer to direct the frequency output to the clock divider. The number of buffers necessary to direct the output to the divider depends on the sub-block in which the RO of interest is located. Hence, the IR drop in the power supply, which affects the RO frequency in both the *pass* and *wait* modes of operation, is dependent on the sub-block in which the RO is located.

5.8.4 Multiple Operating Voltage Results

A robust test circuit should ideally be able to operate over a wide range of supply voltages while still obtaining meaningful and accurate measurement results. This section discusses the particular relevance of the supply voltage issue with regards to the RO-based test circuit. For the 65nm implementation of the test circuit, the schematic for the PMOS DUT-based RO is shown in Figure 5-21. In addition, the NMOS DUT-based RO is shown in Figure 5-22. Various circuit techniques are used

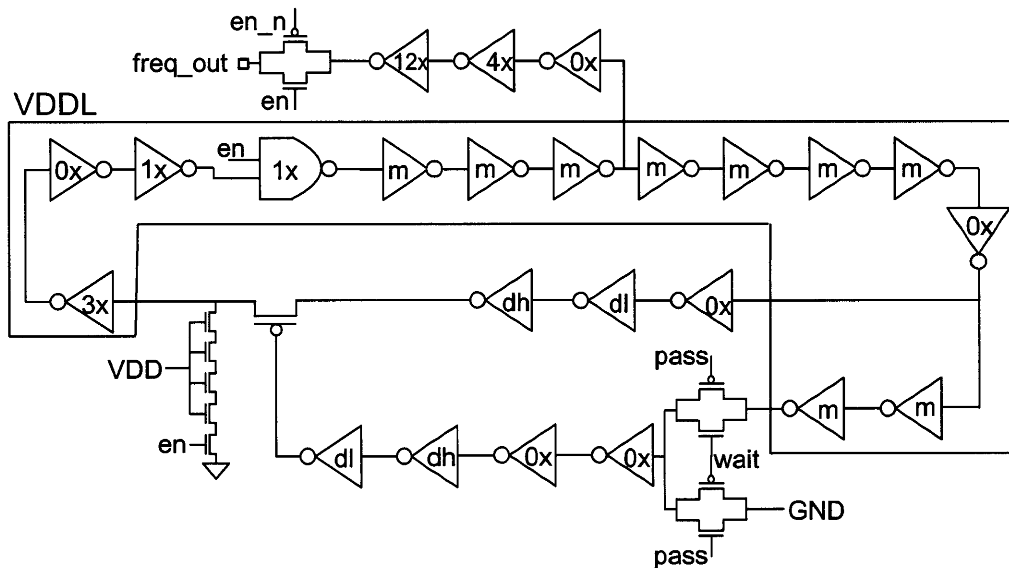


Figure 5-21: Schematic for PMOS DUT-based RO, showing all transistors and gates as well as their relative sizes.

to ensure that the ring oscillator frequency is low enough that, during the *pass* and

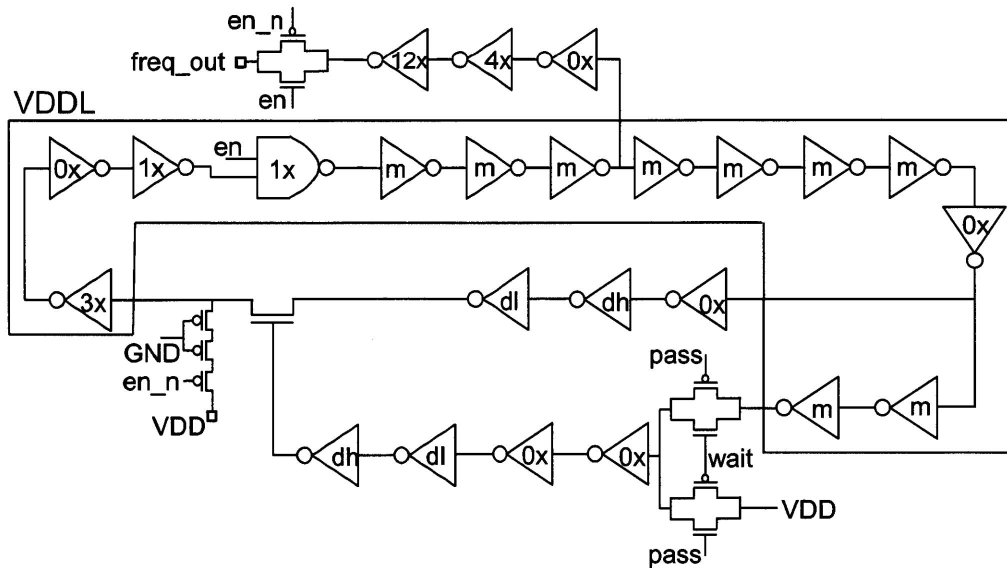


Figure 5-22: Schematic for NMOS DUT-based RO, showing all transistors and gates as well as their relative sizes.

wait modes of operation, the rising and falling voltage waveforms for each stage are identical. The first such technique is the use of a high-V_t stacked inverter and a capacitive load for the device *m*. Some critical layout-based transistor parameters for each transistor/gate in both the NMOS DUT and PMOS DUT-based ring oscillator are shown in Table 5.1. The second technique used to obtain a slow RO frequency is the use of dual supply voltages. While the gates near the DUT are operated at VDD = 1.2V, the other gates in the RO path are operated at VDDL, which can range from 0.9V to 1.0V. Without such a reduction in supply voltage, the ring oscillator would be too fast, causing differences in stage delays between the two modes of operation. If the supply voltage is reduced further, near-threshold effects manifest themselves, causing larger variations in frequency and increased sensitivity to supply voltage variations. In addition, other inconsistencies are observed during the measurement of the test circuit at low supply voltages for which the causes have not been determined.

Measurement results for t_{meas} are shown for all PMOS DUTs measured on a single chip at different values of V_{DDL} in Figure 5-23. While the average absolute value of this quantity tends to change in a non-systematic manner depending on the supply voltage used, the identification of outlier devices is still clear. These shifts in the

RO-based Test Structure Transistor Parameters				
Device	W_p/L_p (nm)	W_n/L_n (nm)	V_T	Load cap.
m	120/180	120/180	high-V_T	900/60 NMOSCAP
dl	120/60	400/60	standard-V_T	none
dh	400/60	120/60	standard-V_T	none
t-gate	200/60	200/60	standard-V_T	none
1x NAND	520/60	390/60	standard-V_T	none
pull-up	200/60	-	standard-V_T	none
pull-down	-	160/60	standard-V_T	none
DUT	200/60	200/60	standard-V_T	none
0x	260/60	195/60	standard-V_T	none
1x	520/60	390/60	standard-V_T	none
3x	1560/60	585/60	standard-V_T	none
4x	2080/60	780/60	standard-V_T	none
12x	6240/60	2340/60	standard-V_T	none

Table 5.1: Transistor and gate parameters for ring oscillator-based test circuit.

average value of t_{meas} may be attributable to a combination of effects, including but not limited to the amount of supply voltage noise and IR drop, the different behavior of the driving devices to the gate and drain nodes of the DUT caused by the difference between V_{DDL} and VDD, and different rise and fall time behavior of the output of the DUT due to the supply voltage of the 3x inverter following it. The test circuit can be improved by performing simulations to better understand these trade-offs and choose an optimal pair of supply voltages which results in high sensitivity to possible AC effects as well as consistency over a range of voltages similar to those selected. Results for t_{meas} in the case of an NMOS DUT are shown in Figure 5-24. The results are similar to those of the PMOS DUT except that the DUT-to-DUT random variance is larger, making it slightly more difficult to distinguish the presence of identifier devices.

5.8.5 Potential Circuit Impact and Implications

In order to demonstrate how it is possible to quantify the impact of AC variations on circuit-level variation, it is useful to simulate a 7-stage ring oscillator for which each device is subject to AC variations. The amount of AC variations to which each device

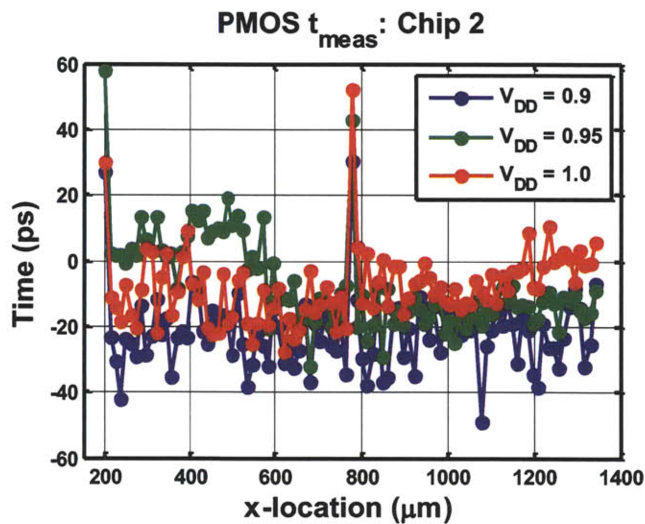


Figure 5-23: t_{meas} for multiple V_{DDL} values for PMOS DUT shows that the identifier is distinguishable at all voltages, but some voltage-dependent effects also change the measurement values relative to one another.

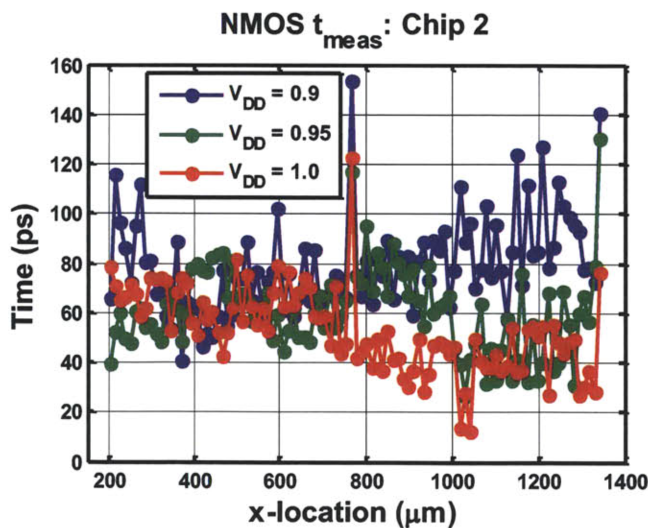


Figure 5-24: t_{meas} for multiple V_{DDL} values for NMOS DUT shows that the identifier is distinguishable at all voltages, but some voltage-dependent effects also change the measurement values relative to one another.

is subject is derived from the measurement results from the test chip on t_{meas}. The steps involved in constructing the simulation platform are the following: a) take the variance on the t_{meas} data and assume it is completely due to AC variation sources, b) calculate the amount of variation in the AC sources which would produce the magnitude of observed t_{meas} variation in the output measurement data, and c) scale the variances of the AC parameters to enable a simulation study in a 32nm CMOS technology.

A 7-stage ring oscillator implemented by scaled versions of the DUTs from the measured test chip to a 32nm technology node with a supply voltage of 1.0V is simulated and subject to AC variations. Because the percentage of t_{meas} variation due to AC sources as opposed to DC sources, systematic effects, or measurement noise is difficult to determine without additional measurement data, this study assumes that all the variation in t_{meas} is due to AC sources. This provides an upper bound on the impact of AC variations on the ring oscillator.

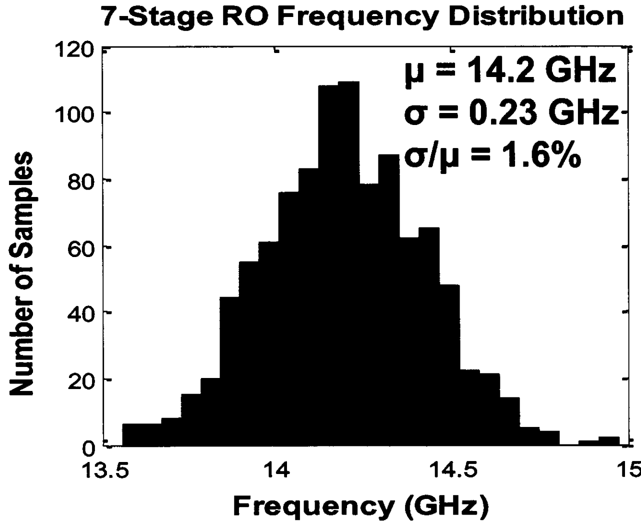


Figure 5-25: Simulation study of a 7-stage ring oscillator frequency variation due to AC variation sources equal to that measured from the test chip, scaled to a 32nm technology node. Results indicate that the frequency has a $\frac{\sigma}{\mu} = 1.6\%$.

The simulation results are shown in Figure 5-25. The results indicate that the frequency has a $\frac{\sigma}{\mu} = 1.6\%$. The same study can be done for other circuit blocks, and can be expanded to different-sized transistors in the case that additional measurement

data were available for other DUT sizes. In addition, these simulations can be refined when provided with additional information regarding the amount of AC variations actually captured in the measurement data.

5.9 Summary

This chapter presents a ring oscillator-based test circuit for characterizing AC device variations across multiple individual devices. By operating the RO in two different modes, the difference between the gate-to-drain delay (t_{gd}) and source-to-drain delay (t_{sd}) is obtained for a single DUT transistor which is gating the RO. The required measurements are two frequency measurements and a DC voltage measurement. Simulations indicate that the measurement output is highly sensitive to possible AC variations in the device under test, and silicon measurement results show that the presence of an external parasitic 30kΩ gate resistor is detected by the measurement. Such a high gate resistance would be difficult to detect with only DC current-voltage measurements or pure capacitance measurements.

Chapter 6

Conclusions

The contributions of this thesis, conclusions from the work, and possibilities for future work in this area will be discussed in this chapter.

6.1 Contributions

This thesis has developed a framework for device variability characterization and analysis. The impact of process variation in devices has been characterized with test circuits, and the measurement results have been analyzed through the use of a decomposition methodology to uncover systematic trends and determine their causes. This framework can aid both in the understanding of variations in the fabrication process and in efforts to model variations in transistor behavior. The detailed contributions are summarized as follows.

First, a test chip was designed, fabricated, and measured in a 90nm CMOS technology in order to understand parametric variations in contact plug resistance. The application of statistical analysis techniques, including a variation decomposition methodology, revealed both geometry-dependent and position-dependent within-die systematic trends. In addition, spatial correlation analysis was performed to uncover further possible trends in the data, such as the presence of any outlier die. The methodologies used to design such a test chip and the ways in which design challenges were overcome are applicable to a large subset of test structures for device or

component characterization. For example, the use of high-V_t multiplexing switches to limit leakage currents coming from a large number of off-state devices under test can be used in the design of any variability characterization test chip that is sensitive to total current values. In addition, practically speaking, the boundaries established between on-chip and off-chip components for the contact plug resistance test architecture stem from a general set of guidelines which state that non-essential components should be placed off-chip when there is no trade-off in speed, accuracy, or any other testing metric. This allows for the design of simple circuits on-chip and leverages the use of a wide range of state-of-the-art special purpose ICs which can modularize the design architecture and enable easier debugging of the testing infrastructure. The statistical techniques for variation decomposition, which involve the use of spatial variation analysis to uncover systematic trends, can be employed to analyze the results of other test structures which have a regular, repeatable spatial DUT layout patterns.

Second, a test structure was designed to measure the AC, or short time-scale, characteristics of transistors. Such AC effects are difficult to detect through standard DC measurements such as I-V characterization, or even pure capacitance-based measurements such as charge-based capacitance measurement (CBCM). Therefore, an array-based test structure was designed for which the variance in the output parameter, relative delay, was comprised of over 95% AC device variations, while the remainder was due to other variation sources. Furthermore, only a single DC voltage measurement is required for each DUT, which enables simple in-line characterization. Test structures are beneficial when attempting to assess the variance of individual transistor performance, especially early in the development of a new and advanced technology in which device and parameter variations are larger than they are later in the development of the technology.

Finally, a ring oscillator-based test structure was also developed for AC variability characterization in transistors. In order to obtain a higher sensitivity towards possible AC variations within the device under test, a metric was created that reflected AC, or short time-scale, transistor performance variations. The difference between

the gate-to-drain and source-to-drain delay of an individual device was measured by implementing a ring oscillator that operates in two different modes. This difference, defined as t_{meas}, is highly sensitive to AC variations — 99.97% of the variance in this parameter is due to variance in the AC characteristics of the DUT. Silicon measurement results from a 65nm test chip show that the addition of an external parasitic gate resistance is detectable by the test structure. Because this test structure is sensitive to possible AC variations in devices, such a circuit can be used to characterize the AC variations of devices in any advanced or emerging technology which may be sensitive to them. In addition, a framework for incorporating the measurement results of such test circuits into existing variability-aware device models was explored. Using the results from these test chips along with those from other test chips focused on characterizing DC variations such as those due to threshold voltage and channel length, one can develop a unified variation model which describes the relative magnitudes of DC and AC variation for a given technology.

6.2 Conclusions

The conclusions of this thesis are multi-fold. This work shows that test structures can be developed for the analysis of variation sources which may not be typically investigated because of its relatively small contribution to total device variation when compared to traditional sources such as threshold voltage, channel length, and oxide thickness variations. Furthermore, in order to understand the nature of these variations, techniques such as variation decomposition can be utilized. The use of a design of experiments helps to determine the nature of impact of certain variation sources.

Coupling these new test structure designs with existing structures for variability characterization can prove to be useful in analyzing variation for advanced technologies. Combining isolation-based test structures with holistic-based ones allows for both the diagnosis of specific variation sources and their impacts and the benchmarking of a technology with regards to variation.

The analysis of the two examples of variation sources discussed in this work —

contact plug resistance and AC variations — can be extended for other sources of variation. In FinFETs and tri-gate transistor technologies which use a high-K metal-gate process, for example, the metal-gate work function variation can be a significant contributor to threshold voltage variation. Additionally, the fin thickness and fin height variation will have an impact on the effective channel length and width of the transistor. Contacts to the source-drain region may have variable parasitics due to multiple sources such as variations in the doping concentration and the distance to the fin, which may impact the parasitic capacitance. Such variations can be characterized, analyzed, and modeled in a similar manner to that done in this work: (1) develop a test structure whose output parameter is highly sensitive to the variation source of interest, (2) design a multiplexing scheme without introducing new variation sources which affect the measurement result, (3) create a design of experiments which will help to understand how the variation changes under different conditions, and (4) employ statistical techniques to decompose variation obtained from silicon measurement results.

6.3 Future Work

Future work in this area would involve the refinement of the test structures developed in this work as well as the design of new test structures to analyze variability in devices. With the emergence of new device structures such as FinFETs, tri-gate transistors, and gate-all-around (GAA) silicon nanowire FETs, parametric variation can present itself in a large number of different areas. Understanding these variations will be a key in continuing Moore's Law scaling. In addition, variation-aware modeling of such transistors will be critical in enabling the design of circuits and systems that meet yield and performance specifications. Within this context of emerging devices which will likely come to the forefront of the semiconductor industry in the years to come, the possibilities for future work can be categorized into those which fall under the characterization, modeling, and mitigation realms.

6.3.1 Characterization

In the area of statistical characterization of transistors for variation analysis, the need for new isolation-based test circuits will likely continue due to the drastic changes which will have to be made to the transistor architecture in order to continue Moore's Law scaling. After an initial functional yield ramp-up, the parametric variations in these devices must be well-characterized for both process optimization and variation modeling purposes. Test structures that can isolate key variation sources as well as test structures that can benchmark the variability characteristics for a given technology will be necessary. In addition, algorithms to optimally select a design of experiments based on expected variations will be important as the number of variation sources and their interactions grows larger. Finally, statistical techniques for systematic variation decomposition of test circuit measurement results in the face of large random variations, undesired systematic trends, limited replication and DOE, and measurement noise can prove to be useful in understanding variations.

6.3.2 Variation-Aware Modeling

The opportunities for improving the variation-aware modeling of transistors are numerous. First, the development of a variation-aware compact model of a transistor whose parameters can be modulated at multiple levels of the parameter hierarchy is desirable. For example, information regarding the statistical distribution of the number of dopants in the channel obtained from atomistic simulation results could be fed into such a model as easily as the measured distribution of intrinsic threshold voltages obtained from a test chip. This kind of flexibility is key in making use of variation-related measurement data as well as physical or analytical models and simulation results. Tools can also be improved which focus on the back propagation of variance from measurement data to the variance in device parameters which are most relevant. Because of the significant amounts of interaction between a given set of measurement outputs and the device parameters involved in determining their values, such methods would be useful in taking measurement data from test circuits and

developing useful device models from those data sets.

Furthermore, the coupling between variation-aware models and circuit-level simulation can be improved. Because the nature of circuit simulation will increasingly revolve around concepts such as process variation, reliability, and yield, a robust framework for circuit simulation which incorporates statistics information is necessary. The idea of alternatives to the traditional but time-consuming Monte Carlo simulation approach, which has already been an area of interest for many years, should continue to present opportunities for future work.

6.3.3 Mitigation of Variation

In IC manufacturing, the first major steps taken towards mitigating process variation were targeted at the manufacturing process itself. Statistical process control, feedback and feed-forward mechanisms, and innovations such as the use of resolution enhancement techniques for lithography and chemical mechanical polishing for metallization were developed. More recently, steps have been taken at the device and circuit design levels to reduce variations. Design for manufacturability has become necessary to achieve yield and performance specifications. In these areas, opportunities for further research exist because of the emergence of new devices — new tradeoffs will manifest themselves and will require careful analysis to determine the DFM rules associated with them. However, the most significant need and opportunity for future work in variability mitigation is at the circuit and system levels. While the variability reduction through the use of process- and device-domain techniques is becoming increasingly incremental, the reductions as a result of compensation and adaptation at the circuit and system levels can be significantly larger. The implementation of such techniques, which in the past have largely been focused on adaptive body biasing, multiple supply voltage islands, etc., can potentially be extended to optimizing circuit topologies and designing variation-aware system architectures.

Bibliography

[1] N. M. Gearailt, "The measurement and control of process variation in high volume manufacturing semiconductor fabs," *Presentation - EMEA Academic Forum, Budapest, Hungary*, Jun. 2007.

[2] K. Kuhn, C. Kenyon, A. Kornfeld, M. Liu, A. Maheshwari, W. K. Shih, S. Sivakumar, G. Taylor, P. VanDerVoorn, and K. Zawadzki, "Managing process variations in Intel's 45nm CMOS technology," *Intel Technology Journal*, vol. 12, Jun. 2008.

[3] K. Kuhn, "Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS," *IEEE International Electron Devices Meeting*, pp. 471–474, Dec. 2007.

[4] M. Pelgrom, H. Tuinhout, and M. Vertregt, "Transistor matching in analog CMOS applications," *International Electron Devices Meeting*, pp. 915–918, Dec. 1998.

[5] K. Kuhn, "22 nm device architecture and performance elements," *IEEE International Electron Devices Meeting*, pp. 1–4, Dec. 2008.

[6] S. S. Cohen and G. S. Gildenblat, *Metal-Semiconductor Contacts and Devices*, vol. 13 of *VLSI Electronics - Microstructure Science*, ch. 4, pp. 87–110. Orlando, FL: Academic Press, Inc., first ed., 1986.

[7] G. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, vol. 86, pp. 82–85, Jan. 1998.

[8] K. Kuhn, "Variation in 45nm and implications for 32nm and beyond," *International CMOS Variability Conference*, May 2009.

[9] N. Drego, A. Chandrakasan, and D. Boning, "A test-structure to efficiently study threshold-voltage variation in large MOSFET arrays," *International Symposium on Quality Electronic Design*, pp. 281–286, Mar. 2007.

[10] K. Agarwal, S. Nassif, F. Liu, J. Hayes, and K. Nowka, "Rapid characterization of threshold voltage fluctuation in MOS devices," *IEEE International Conference on Microelectronic Test Structures*, pp. 74–77, Mar. 2007.

[11] M. Orshansky, L. Milor, and C. Hu, "Characterization of spatial intrafield gate CD variability, its impact on circuit performance, and spatial mask-level correction," *IEEE Transactions on Semiconductor Manufacturing*, vol. 17, pp. 2–11, Feb. 2004.

[12] M. Conti, G. Dalla Betta, S. Orcioni, G. Soncini, C. Turchetti, and N. Zorzi, "Test structure for mismatch characterization of MOS transistors in subthreshold regime," *IEEE International Conference on Microelectronic Test Structures*, pp. 173–178, Mar. 1997.

[13] K. Agarwal, F. Liu, C. McDowell, S. Nassif, K. Nowka, M. Palmer, D. Acharyya, and J. Plusquellic, "A test structure for characterizing local device mismatches," *Symposium on VLSI Circuits*, pp. 67–68, Feb. 2006.

[14] I. Ahsan, N. Zamdmer, O. Glushchenkov, R. Logan, E. Nowak, H. Kimura, J. Zimmerman, 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, H. Utomo, and B. Walsh, "RTA-driven intra-die variations in stage delay, and parametric sensitivities for 65nm technology," *Symposium on VLSI Technology*, pp. 170–171, 2006.

[15] N. Wils, H. Tuinhout, and M. Meijer, "Characterization of STI edge effects on CMOS variability," *IEEE Transactions on Semiconductor Manufacturing*, vol. 22, pp. 59–65, Feb. 2009.

[16] B. Stine, D. Ouma, R. Divecha, D. Boning, J. Chung, D. Hetherington, C. Harwoo, O. Nakagawa, and S.-Y. Oh, "Rapid characterization and modeling of pattern-dependent variation in chemical-mechanical polishing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 11, pp. 129–140, Feb. 1998.

[17] C. Hess, S. Saxena, H. Karbasi, S. Subramanian, M. Quarantelli, A. Rossoni, S. Tonello, S. Zhao, and D. Slisher, "Device array scribe characterization vehicle test chip for ultra fast product wafer variability monitoring," *IEEE International Conference on Microelectronic Test Structures*, pp. 145–149, Mar. 2007.

[18] S. Saxena, C. Hess, H. Karbasi, A. Rossoni, S. Tonello, P. McNamara, S. Lucherini, S. Minehane, C. Dolainsky, and M. Quarantelli, "Variation in transistor performance and leakage in nanometer-scale technologies," *IEEE Transactions on Electron Devices*, vol. 55, pp. 131–144, Jan. 2008.

[19] S. Realov, W. McLaughlin, and K. Shepard, "On-chip transistor characterization arrays with digital interfaces for variability characterization," *International Symposium on Quality Electronic Design*, pp. 167–171, Mar. 2009.

[20] S. Ohkawa, M. Aoki, and H. Masuda, "Analysis and characterization of device variations in an LSI chip using an integrated device matrix array," *IEEE Transactions on Semiconductor Manufacturing*, vol. 17, pp. 155–165, May 2004.

[21] D. S. Boning and J. E. Chung, "Statistical metrology: Understanding spatial variation in semiconductor manufacturing," 1996.

[22] M. Gardner and J. Bieker, "Data mining solves tough semiconductor manufacturing problems," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 376–383, 2000.

[23] B. Stine, D. Boning, and J. Chung, "Analysis and decomposition of spatial variation in integrated circuit processes and devices," *IEEE Transactions on Semiconductor Manufacturing*, vol. 10, pp. 24–41, Feb. 1997.

[24] J. Croon, G. Storms, S. Winkelmeier, I. Pollentier, M. Ercken, S. Decoutere, W. Sansen, and H. Maes, "Line edge roughness: characterization, modeling and impact on device behavior," *International Electron Devices Meeting*, pp. 307–310, 2002.

[25] W. Zhao, F. Liu, K. Agarwal, D. Acharyya, S. Nassif, K. Nowka, and Y. Cao, "Rigorous extraction of process variations for 65-nm cmos design," *IEEE Transactions on Semiconductor Manufacturing*, vol. 22, pp. 196–203, Feb. 2009.

[26] J. Xiong, V. Zolotov, and L. He, "Robust extraction of spatial correlation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, pp. 619–631, Apr. 2007.

[27] E. Dyer, M. Majzoobi, and F. Koushanfar, "Hybrid modeling of non-stationary process variations," *ACM/EDAC/IEEE Design Automation Conference*, pp. 194–199, Jun. 2011.

[28] P. Stolk, F. Widdershoven, and D. Klaassen, "Modeling statistical dopant fluctuations in MOS transistors," *IEEE Transactions on Electron Devices*, vol. 45, pp. 1960–1971, Sep. 1998.

[29] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFET's: A 3-D 'atomistic' simulation study," *IEEE Transactions on Electron Devices*, vol. 45, pp. 2505–2513, Dec. 1998.

[30] A. Keshavarzi, G. Schrom, S. Tang, S. Ma, K. Bowman, S. Tyagi, K. Zhang, T. Linton, N. Hakim, S. Duvall, J. Brews, and V. De, "Measurements and modeling of intrinsic fluctuations in MOSFET threshold voltage," *ACM International Symposium on Low Power Electronics and Design*, pp. 26–29, 2005.

[31] K. Takeuchi and M. Hane, "Statistical compact model parameter extraction by direct fitting to variations," *IEEE Transactions on Electron Devices*, vol. 55, pp. 1487–1493, Jun. 2008.

[32] X. Li, C. McAndrew, W. Wu, S. Chaudhry, J. Victory, and G. Gildenblat, "Statistical modeling with the PSP MOSFET model," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, pp. 599–606, Apr. 2010.

[33] B. Cheng, D. Dideban, N. Moezi, C. Millar, G. Roy, X. Wang, S. Roy, and A. Asenov, "Statistical-variability compact-modeling strategies for BSIM4 and PSP," *IEEE Design Test of Computers*, vol. 27, pp. 26–35, Mar. 2010.

[34] T. A. El-Moselhy, I. M. Elfadel, and L. Daniel, "A capacitance solver for incremental variation-aware extraction," *IEEE/ACM International Conference on Computer-Aided Design*, pp. 662–669, 2008.

[35] H. Dadgour, K. Endo, V. De, and K. Banerjee, "Modeling and analysis of grain-orientation effects in emerging metal-gate devices and implications for SRAM reliability," *IEEE International Electron Devices Meeting*, pp. 1–4, Dec. 2008.

[36] B. Stine, D. Ouma, R. Divecha, D. Boning, and J. Chung, "A closed-form analytical model for ILD thickness variation," *CMP Processes, Proc. CMP-MIC*, pp. 266–273, 1997.

[37] P. Yu, S. X. Shi, and D. Z. Pan, "Process variation aware OPC with variational lithography modeling," *ACM Design Automation Conference*, pp. 785–790, 2006.

[38] P. Mozumder and G. Barna, "Statistical feedback control of a plasma etch process," *IEEE Transactions on Semiconductor Manufacturing*, vol. 7, pp. 1–11, Feb. 1994.

[39] Q. Zhang, C. Tang, T. Hsieh, N. Maccrae, B. Singh, K. Poolla, and C. J. Spanos, "Comprehensive CD uniformity control across lithography and etch," *Society of*

Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 5752, pp. 692–701, May 2005.

[40] J. Kibarian, C. Guardiani, and A. Strojwas, "Design for manufacturability in nanometer era: system implementation and silicon results," *IEEE International Solid-State Circuits Conference*, pp. 268–269, Feb. 2005.

[41] D. Boning, K. Balakrishan, H. Cai, N. Drego, A. Farahanchi, K. Gettings, D. Lim, A. Somani, H. Taylor, D. Truque, and X. Xie, "Variaton," *IEEE Transactions on Semiconductor Manufacturing*, vol. 21, pp. 63–71, Feb. 2008.

[42] M. Orshansky, S. Nassif, and D. Boning, *Statistical Design and Design for Manufacturability: A Constructive Approach.* New York: Springer, 2008.

[43] X. Sun, K. Li, W. Wu, P. Wilhite, T. Saito, and C. Yang, "Contact resistances of carbon nanotube via interconnects," *IEEE Electron Devices and Solid-State Circuits*, pp. 131–135, Dec. 2009.

[44] A. Kawabata, S. Sato, T. Nozue, T. Hyakushima, M. Norimatsu, M. Mishima, T. Murakami, D. Kondo, K. Asano, M. Ohfuti, H. Kawarada, T. Sakai, M. Nihei, and Y. Awano, "Robustness of CNT via interconnect fabricated by low temperature process over a high-density current," *International Interconnect Technology Conference*, pp. 237–239, Jun. 2008.

[45] J. Coiffic, M. Fayolle, H. le Poche, S. Maitrejean, and S. Olivier, "Realization of via interconnects based on carbon nanotubes," *International Interconnect Technology Conference*, pp. 153–155, Jun. 2008.

[46] S. Demuynck, A. Nackaerts, G. Van den bosch, T. Chiarella, J. Ramos, Z. Tokei, J. Vaes, N. Heylen, G. Beyer, M. Van Hove, T. Mandrekar, and R. Schreutelkamp, "Impact of Cu contacts on front-end performance: a projection towards 22nm node," *IEEE International Interconnect Technology Conference*, pp. 178–180, 2006.

[47] Y.-C. Chen, T.-Y. Hung, Y.-L. Chang, K. Shieh, C.-L. Hsu, C. Huang, W. Yan, K. Ashtiani, D. Pisharoty, W. Lei, S. Chang, F. Huang, J. Collins, and S. Tzou, "Optimizing ALD WN process for 65nm node CMOS contact application," *IEEE International Interconnect Technology Conference*, pp. 105–107, Jun. 2007.

[48] W. Shockley, "Research and investigation of inverse epitaxial UHF power transistors," Tech. Rep. 64, Wright Patterson Air Force Base, Ohio, Sept. 1964.

[49] D. Scott, W. Hunter, and H. Shichijo, "A transmission line model for silicided diffusions: Impact on the performance of VLSI circuits," *Journal of Solid-State Circuits*, vol. 17, pp. 281–291, Apr. 1982.

[50] T. Isogai, H. Tanaka, A. Teramoto, T. Goto, S. Sugawa, and T. Ohmi, "Advanced method for measuring ultra-low contact resistivity between silicide and silicon based on cross bridge kelvin resistor," *IEEE International Conference on Microelectronic Test Structures*, pp. 109–113, Apr. 2009.

[51] N. Stavitski, M. van Dal, R. Wolters, A. Kovalgin, and J. Schmitz, "Specific contact resistance measurements of metal-semiconductor junctions," *IEEE International Conference on Microelectronic Test Structures*, pp. 13–17, Mar. 2006.

[52] S. Proctor, L. Linholm, and J. Mazer, "Direct measurements of interfacial contact resistance, end contact resistance, and interfacial contact layer uniformity," *IEEE Transactions on Electron Devices*, vol. 30, pp. 1535–1542, Nov. 1983.

[53] T. Hamamoto, T. Ozaki, M. Aoki, and Y. Ishibashi, "Measurement of contact resistance distribution using a 4k-contacts array," *IEEE Transactions on Semiconductor Manufacturing*, vol. 9, pp. 9–14, Feb. 1996.

[54] M. Finetti, A. Scorzoni, and G. Soncini, "Lateral current crowding effects on contact resistance measurements in four terminal resistor test patterns," *IEEE Electron Device Letters*, vol. 5, pp. 524–526, Dec. 1984.

[55] W. Loh, S. Swirhun, T. Schreyer, R. Swanson, and K. Saraswat, "Modeling and measurement of contact resistances," *IEEE Transactions on Electron Devices*, vol. 34, pp. 512–524, Mar. 1987.

[56] C. Hess, B. Stine, L. Weiland, T. Mitchell, M. Karnett, and K. Gardner, "Passive multiplexer test structure for fast and accurate contact and via fail-rate evaluation," *IEEE Transactions on Semiconductor Manufacturing*, vol. 16, pp. 259–265, May 2003.

[57] A. Cabrini, D. Cantarelli, P. Cappelletti, R. Casiraghi, A. Maurelli, M. Pasotti, P. Rolandi, and G. Torelli, "A test structure for contact and via failure analysis in deep-submicrometer CMOS technologies," *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, pp. 57–66, Feb. 2006.

[58] S.-D. Kim, C.-M. Park, and J. Woo, "Advanced model and analysis of series resistance for CMOS scaling into nanometer regime. I. Theoretical derivation," *IEEE Transactions on Electron Devices*, vol. 49, pp. 457–466, March 2002.

[59] S. D. Kim, S. Narasimha, and K. Rim, "A new method to determine effective lateral doping abruptness and spreading-resistance components in nanoscale MOSFETs," *IEEE Transactions on Electron Devices*, vol. 55, pp. 1035–1041, Apr. 2008.

[60] F. Liu and K. Agarwal, "A test structure for assessing individual contact resistance," *IEEE International Conference on Microelectronic Test Structures*, pp. 201–204, Apr. 2009.

[61] K. Gettings and D. Boning, "Study of CMOS process variation by multiplexing analog characteristics," *IEEE Transactions on Semiconductor Manufacturing*, vol. 21, pp. 513–525, Nov. 2008.

[62] A. K. K. Wong, A. F. Molless, T. A. Brunner, E. Coker, R. H. Fair, G. L. Mack, and S. M. Mansfield, "Linewidth variation characterization by spatial decompo-

sition," *Journal of Microlithography, Microfabrication, and Microsystems*, vol. 1, no. 2, pp. 106–116, 2002.

[63] W. Zhang, K. Balakrishnan, X. Li, D. Boning, and R. Rutenbar, "Toward efficient spatial variation decomposition via sparse regression," *To appear, IEEE/ACM International Conference on Computer-Aided Design*, Nov. 2011.

[64] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, pp. 4655–4666, Dec. 2007.

[65] K.-W. Su, Y.-M. Sheu, C.-K. Lin, S.-J. Yang, W.-J. Liang, X. Xi, C.-S. Chiang, J.-K. Her, Y.-T. Chia, C. Diaz, and C. Hu, "A scaleable model for STI mechanical stress effect on layout dependence of MOS electrical characteristics," *IEEE Custom Integrated Circuits Conference*, pp. 245–248, Sep. 2003.

[66] A. Topol, C. Sheraw, K. Wong, X. Shao, R. Knarr, S. Rossnagel, C.-C. Yang, B. Baker-O'Neal, A. Simon, B. Haran, Y. Li, C. Ouyang, S. Allen, C. Brodsky, S. Cohen, L. Deligianni, X. Chen, S. Deshpande, C. Sung, and M. Ieong, "Lower resistance scaled metal contacts to silicide for advanced CMOS," *Symposium on VLSI Technology*, October 2006.

[67] S.-D. Kim, C.-M. Park, and J. Woo, "Advanced model and analysis of series resistance for cmos scaling into nanometer regime. II. Quantitative analysis," *IEEE Transactions on Electron Devices*, vol. 49, pp. 467–472, Mar. 2002.

[68] N. Lu and B. Dewey, "Characterization, simulation, and modeling of FET source/drain diffusion resistance," *IEEE Custom Integrated Circuits Conference*, pp. 281–284, Sept. 2008.

[69] N. Mohapatra, M. Desai, S. Narendra, and V. Ramgopal Rao, "Modeling of parasitic capacitances in deep submicrometer conventional and high-k dielectric MOS transistors," *IEEE Transactions on Electron Devices*, vol. 50, pp. 959–966, Apr. 2003.

[70] J. Mueller, R. Thoma, E. Demircan, C. Bernicot, and A. Juge, "Modeling of MOSFET parasitic capacitances, and their impact on circuit performance," *Solid State Electronics*, vol. 51, pp. 1485–1493, Nov. 2007.

[71] T. El-Moselhy, I. Elfadel, and L. Daniel, "A capacitance solver for incremental variation-aware extraction," *IEEE/ACM International Conference on Computer-Aided Design*, pp. 662–669, Nov. 2008.

[72] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer, "High-performance CMOS variability in the 65-nm regime and beyond," *IBM Journal of Research and Development*, vol. 50, pp. 433–449, Jul. 2006.

[73] R. Rao, K. Jenkins, and J.-J. Kim, "A completely digital on-chip circuit for local-random-variability measurement," *IEEE International Solid-State Circuits Conference*, pp. 412–623, Feb. 2008.

[74] A. Balankutty, T. Chih, C. Chen, and P. Kinget, "Mismatch characterization of ring oscillators," *IEEE Custom Integrated Circuits Conference*, pp. 515–518, Sep. 2007.

[75] M. Bhushan, A. Gattiker, M. Ketchen, and K. Das, "Ring oscillators for CMOS process tuning and variability control," *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, pp. 10–18, Feb. 2006.

[76] L.-T. Pang, K. Qian, C. Spanos, and B. Nikolic, "Measurement and analysis of variability in 45 nm strained-Si CMOS technology," *IEEE Journal of Solid-State Circuits*, vol. 44, pp. 2233–2243, Aug. 2009.

[77] M. Koolen, J. Geelen, and M. Versleijen, "An improved de-embedding technique for on-wafer high-frequency characterization," *Bipolar Circuits and Technology Meeting*, pp. 188–191, Sep. 1991.

[78] Y.-W. Chang, H.-W. Chang, T.-C. Lu, Y.-C. King, W. Ting, Y.-H. J. Ku, and C.-Y. Lu, "Charge-based capacitance measurement for bias-dependent capacitance," *IEEE Electron Device Letters*, vol. 27, pp. 390–392, May 2006.

[79] M. Bhushan, M. Ketchen, M. Cai, and C. Kim, "Ring oscillator technique for MOSFET CV characterization," *IEEE Transactions on Semiconductor Manufacturing*, vol. 21, pp. 180–185, May 2008.

[80] T. Iizuka, J. Jeong, T. Nakura, M. Ikeda, and K. Asada, "All-digital on-chip monitor for PMOS and NMOS process variability measurement utilizing buffer ring with pulse counter," *European Solid-State Circuits Conference*, pp. 182–185, Sep. 2010.

[81] B. Das, B. Amrutur, H. Jamadagni, N. Arvind, and V. Visvanathan, "Within-die gate delay variability measurement using reconfigurable ring oscillator," *IEEE Transactions on Semiconductor Manufacturing*, vol. 22, pp. 256–267, May 2009.