Combined C-V/I-V and RTN CMOS Variability Characterization Using An On-Chip Measurement System

Simeon Dimitrov Realov

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2012

Abstract

Combined C-V/I-V and RTN CMOS Variability Characterization Using An On-Chip Measurement System

Simeon Dimitrov Realov

With the number of transistors integrated into a single integrated circuit (IC) crossing the one-billion mark and complementary metal-oxide-semiconductor (CMOS) technology scaling pushing device dimensions ever-so-close to atomic scales, variability in transistor performance is becoming the dominant constraint in modern-day CMOS IC design. Developing novel approaches for device characterization, which allow a detailed study of electrical transistor characteristics across large statistical sample sets, is crucial for the proper identification, characterization, and modeling of different physical sources of device variability. On-chip characterization methodologies have the potential to address all of these issues by enabling the characterization of large statistical device sample sets, while also allowing for high measurement quality and throughput.

In this work, a fully-integrated system for on-chip combined capacitance-voltage (C-V) and current-voltage (I-V) characterization of a large integrated test transistor array implemented in a 45-nm bulk CMOS process is presented. On-chip I-V characterization is implemented using a four-point Kelvin measurement technique with 12-bit sub-10 nA current measurement resolution, 10-bit sub-1 mV voltage measurement resolution, and sampling speeds on the order of 100 kHz. C-V characterization is performed using a novel leakage- and parasitics-insensitive charge-based capacitance measurement (CBCM) technique with atto-Farad resolution.

The on-chip system is employed in developing a comprehensive CMOS transistor variability characterization methodology, studying both random and systematic sources of quasi-static device variability. For the first time, combined C-V/I-V characterization of circuit-representative devices is demonstrated and used to extract variations in the underlying physical parameters of the device. Additionally, the fast current sampling capabilities of the system are used for the characterization of random telegraph noise (RTN) in small area devices. An automated methodology for the extraction of RTN parameters is developed, and the statistics of RTN are studied across device type, bias, and geometry.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to acknowledge the mentorship and support of my doctoral research adviser, Professor Kenneth Shepard, without whom none of this would have been possible. Ken is the reason why I chose to come to Columbia, and working with him over the years has been a pleasure and a privilege. He was always a reliable source of technical insight, and in spite of his overwhelmingly busy schedule, always made sure to set aside time to guide me through the rough patches. Ken provided me with every resource I ever required, whether it was access to cutting-edge manufacturing technology, CAD tools, test equipment, or lab space, and while it may seem trivial, this freedom has really made a difference in my work. At the end of the day, Ken is without a doubt the one person who has influenced me the most in my career up to this point, for which I will always be grateful.

I would also like to acknowledge the mentorship of Dr. Stanislav Polonsky, whom I worked with closely during my summers at the IBM T. J. Watson research center. Stas has been a great role model, a reliable source of technical expertise, and a good friend. Thanks to him, I gained exposure to the real-world aspects of the work I have been doing, which shaped my research efforts and helped me make my work more relevant.

The other members of my thesis committee also deserve a special mention. Professor Yannis Tsividis, Professor Luca Carloni, and Professor Charle Zukowski set aside time to examine my work and give me valuable feedback, for which I am thankful. I am fortunate to have had faculty with such vast expertise serve on my committee and contribute to my work.

My fellow graduate students have been another great resource during my time at

Columbia. Whether it was technical discussions, or practical help in the lab, each and every one of them has helped me in some way. Paul Pan was instrumental in helping me tape-out the test chip and deserves my gratitude and praise for that. He offered me a helping hand when I really needed it, and without him, I would have been hard-pressed to submit a working design in time to meet the deadline. Additionally, I would like to thank Jacob Rosenstein for his help in the lab and useful technical discussions. I would also like to specifically mention Matthew Jonhston and Ryan Field, who always contributed to technical discussions, sometimes in non-academic settings at odd hours of the night, and proved to be great friends outside of work as well. Additionally, Noah Sturcken, Jared Roseman, Inanc Meric, Sebastian Sorgenfrei, and Peter Levine have all contributed to my work in some way or another over the years, and have become good friends and colleges that I hope to remain in contact with in the future. Besides Ken, I have to say that the students in the group have been the best resource available to me throughout the years, and the quality of my work would not have been the same without their continual input.

Last, but not least, I would like to thank my family - my mom, Sonya, my dad, Dimitar, and my brother, Christo, for all of their support throughout the years. When things got rough, and inevitably, they always do in one's graduate school career, my family was always there for me. Without their support and constant encouragement, I would have never completed an undertaking of this magnitude.

Chapter 1

Introduction

Complementary metal-oxide-semiconductor (CMOS) technology is undoubtedly the dominant integrated circuit (IC) technology of today. In large part, the assent of CMOS can be attributed to the scalability of MOS transistors, which has persisted at an exponential rate for over five decades. However, as critical device dimensions are approaching atomic scales, issues associated with device variability are rapidly becoming a bottleneck across the entire design stack.

In order to adequately manage transistor variability in advanced CMOS technology nodes, novel comprehensive methodologies for variability characterization have to be established. Measurement techniques, which enable the fast and detailed characterization of large statistical device sample sets, are needed for this purpose.

On-chip integration of large addressable device-under-test (DUT) arrays and the associated characterization circuitry has the potential to address all of these issues. In this work, an on-chip characterization system for capacitance -voltage (C-V) and current-voltage (I-V) characterization of circuit-representative devices implemented in a 45-nm bulk CMOS process is introduced. The system is used to identify quasi-static sources of random and systematic device variability through a novel combined C-V/I-V characterization methodology. Random telegraph noise (RTN) is small-area devices is also studied with an emphasis on developing an automated methodology for the analysis of RTN waveforms, and

a comprehensive statistical model for the prediction of overall RTN amplitude is developed.

1.1 Thesis Outline

Chapter 2 begins by establishing the background for this work. A brief historical overview of CMOS circuits is presented, followed by a discussion of CMOS variability, including different approaches to modeling and characterizing variability. In this context, the on-chip combined C-V/I-V characterization approach is introduced.

Chapter 3 gives a description of the design of the on-chip combined C-V/I-V characterization system. Overall system design is discussed with an emphasis on full on-chip integration and design modularity. Implementation of the different circuit blocks in a 45-nm bulk CMOS process is described, and functionality is verified using simulation and measurement results. An experimental measurement setup is introduced.

Chapter 4 describes measurement techniques for combined on-chip C-V/I-V characterization. Accurate four-point Kelvin I-V measurements are demonstrated. A leakage- and parasitics-insensitive charge-based capacitance measurement (CBCM) technique with atto-Farad measurement resolution is presented. Combined C-V/I-V measurements are performed on large statistical sample sets of devices using the on-chip characterization system. Different electrical parameters are extracted, and the variability in these parameters is studied across device geometry. Specific emphasis is placed on using the combined information from C-V and I-V characterization to uncover the physical sources of variability in the quasi-static device characteristics. Sources of both random and systematic nature are examined.

Chapter 5 demonstrates another application of the on-chip device characterization system, where the rapid I-V measurement capability is used for time-domain characterization of RTN in small-area devices. An automated approach for the extraction of different RTN parameters from measured data is developed, and parameter statistics across bias, geometry, and device polarity are examined. Based on gathered data, an empirical statistical model for the modeling of overall RTN amplitude fluctuations is extracted and verified

across the sample range.

Chapter 6 concludes. The original contributions made in this work are summarized, and the resulting peer-review publications are presented. Future research directions are outlined.

Chapter 2

Background

2.1 Introduction

Chapter 2 serves to place the presented work in appropriate technical context. The origins and basics of CMOS integrated circuits are discussed, and the continual push driving device scaling is examined. CMOS transistor variability is presented as a major challenge in integrated circuit design. After establishing a brief historical perspective, different classifications and sources of variability are described, along with their impact on transistor characteristics and overall circuit performance. Basic approaches to modeling variability in transistor parameters are outlined. Different methods for measuring and characterizing device variability are discussed, with an emphasis on their comparative advantages and disadvantages. The proposed on-chip combined C-V/I-V variability characterization system is introduced in light of its advantages over current variability characterization methodologies.

2.2 CMOS Basics

2.2.1 CMOS Technology

Complementary metal-oxide-semiconductor (CMOS) technology is the dominant technology used today for the implementation of integrated circuits (ICs), ranging from digital memory and microprocessors to highly-integrated mixed-signal systems-on-chip (SOCs).

The complementary nature of CMOS technology is derived from the availability of both p-type (PMOS) and n-type (NMOS) field-effect transistors (FETs) as the fundamental circuit building blocks. An FET is a four-terminal semiconductor device, which uses an electric field applied through a gate terminal to induce a conduction channel in a doped semiconductor substrate, thus establishing an electrical contact between a source and a drain terminal doped with dopant atoms of the opposite polarity; a fourth terminal, referred to as the body terminal, is used to set the substrate potential. The majority carrier in an FET device refers to the type of carrier which transfers charge through the channel of the device. In NMOS transistors the majority carriers are negatively-charged electrons and in PMOS transistors the majority carriers are positively charged holes. PMOS devices are typically used as charging devices, and NMOS devices are typically used as discharging devices in the implementation of digital circuits. Basic representations of NMOS and PMOS transistor cross sections are shown in Fig.2.1.

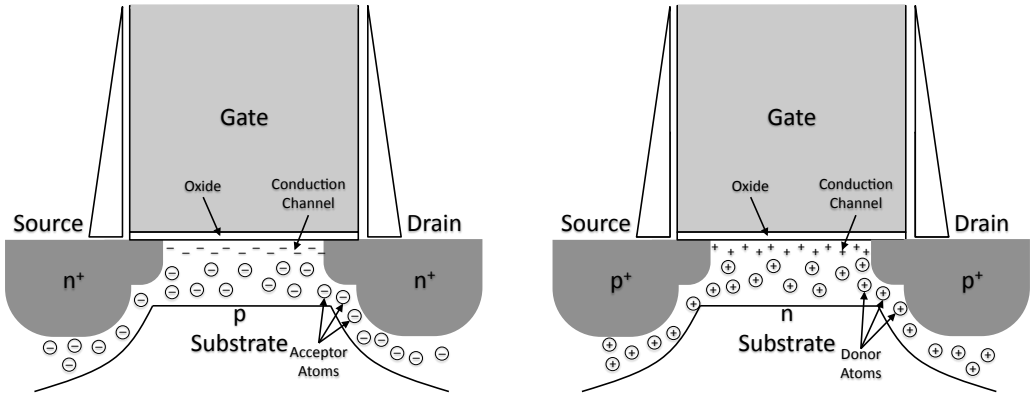


Figure 2.1: Cross-section of (a) an NMOS device and (b) a PMOS device; the devices are shown with the channel fully inverted.

2.2.2 The Appeal of CMOS

The concept of a field-effect transistor was originally introduced by Lilienfeld in 1926 [1], but it wasn't until 1960 that Kahng from Bell Labs demonstrated the first operational MOS device [2]. The appearance of the first MOS transistors coincided with the introduction of

the concept of a monolithic integrated circuit, first patented by Noyce in 1959 [3]. While both NMOS and PMOS transistors were conceived early on, the idea of combining the two types of device polarities in order to achieve minimal standby power in a digital circuit implementation was introduced in 1963 by Wanlass and Sah [4]. The ability to limit power consumption to only power required for the charging and discharging of load capacitances during switching events is one of the primary appeals of CMOS technology.

In 1965, Moore published his seminal paper,"Cramming More Components Onto Integrated Circuits" [5], which for the first time introduced the concept of technology scaling, and established what is commonly referred to as Moore's Law. Moore's Law states that the number of components integrated in a single IC will increase exponentially, doubling approximately every two years. While often labeled as a self-fulfilling prophecy, the push for continual scaling has been the driving force behind the explosive expansion of the semiconductor industry, making IC design one of the major global economic engines for the past five decades. It is interesting to note that Moore's prediction was not based on a technical argument, but rather on an economic one, stating that the reason for the exponential increase in the number of integrated components is based on the premise that higher integration leads to a correspondingly lower cost per device, keeping the cost of the IC relatively constant, while simultaneously increasing the available functionality at an exponential rate. As such, Moore's genius is perhaps not so much in noticing the technical possibilities for scaling, as much as in perceiving the ultimate economic driving force behind ever higher levels of integration. Amazingly, the IC industry has kept up with the exponential growth predicted by Moore to present day (Fig.2.2).

The technical path toward the realization of Moore's Law was in large part outlined by Dennard in 1974, when he introduced the concept of constant-field scaling [7]. The theory of constant-field scaling proposes that if MOSFETs are scaled in a manner, such that the magnitude of the electric fields that govern transistor operation remains constant, then a proportional gain in circuit performance can be expected, while at the same time a quadratic reduction of power per circuit operation is achieved, resulting in constant power density per

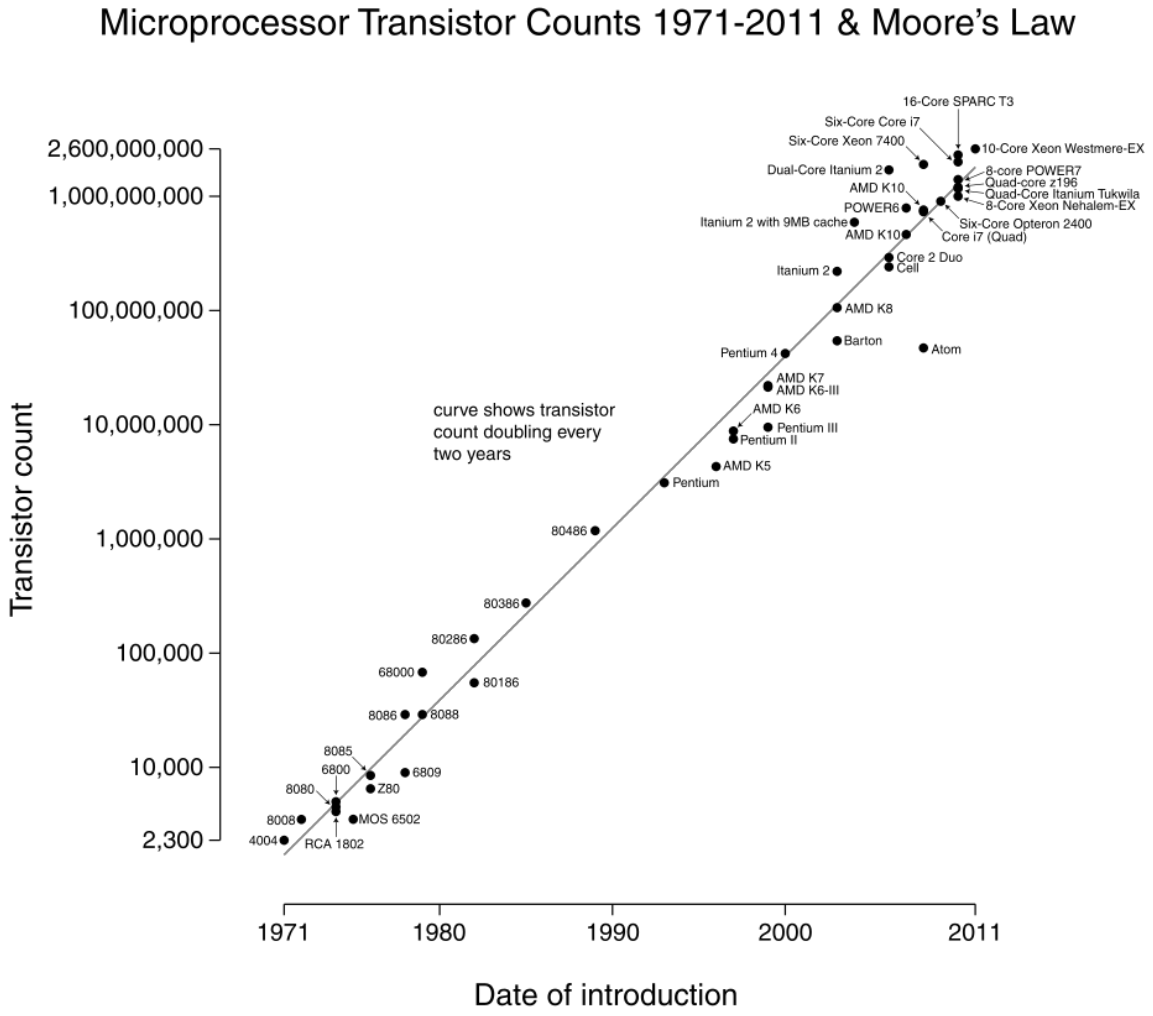


Figure 2.2: Moore's Law as demonstrated by the continual exponential growth of number of transistors integrated in a single IC up to present day [6].

Table 2.1: Dennard Constant Field Scaling [7]

Device or Circuit Parameter	**Scaling Factor**
Device dimension t_{ox}, L, W	$1/k$
Doping concentration, N_a	k
Voltage, V	$1/k$
Current, I	$1/k$
Capacitance, $\epsilon A/t_{ox}$	$1/k$
Delay time/circuit, CV/I	$1/k$
Power dissipation/circuit VI	$1/k^2$
Power density, VI/A	1

unit area. This type of scaling can be accomplished by reducing all of the physical device dimensions by a scaling factor, k, reducing all voltages (including the threshold voltage) by k, and increasing the doping concentration by k. The overall scaling trends proposed by Dennard are shown in Table 2.1. By and large, it is this notion of constant field scaling which cleared the way for the realization of Moore's law, and ultimately made CMOS technology the prevailing IC technology of today. However, strictly adhering to Dennard's scaling rules as described in [7] has proven difficult in recent years, mainly due to an inability to scale the threshold voltage of the device as the parameter approaches fundamental limits set by the laws of thermodynamics. Other scaling schemes, such as constant-voltage scaling and quasi-constant-voltage scaling [8], have been proposed in an attempt to circumvent this issue, and the general drive to scale the dimensions of the device with each new technology node, resulting in an overall exponential decrease in device dimensions over time, has remained true to this day.

2.3 CMOS Variability

While device scaling has many desirable properties, such as improved circuit performance and decreased cost per transistor, maintaining the scaling trajectory predicted by Moore's law is becoming increasingly difficult. Many factors contribute to this problem; however, variability in device performance is arguably the most difficult to overcome. In fact, Moore

[5] identifies issues related to yield as the major limiting factor in achieving ever higher levels of integration. Every manufacturing process is fundamentally limited by quality and reliability concerns, and after a certain point, trying to integrate more devices into a single IC results in decreased yield, and consequently, a higher price per working part. Problems associated with variability are particularly nefarious, since as device dimensions scale beyond the wavelength of light used for patterning and approach atomic dimensions, controlling device behavior with increasing precision becomes tremendously challenging. Different effects traced to the discrete and quantum nature of charge and the inability to accurately define device dimensions begin to dominate transistor behavior. At the same time, increasing levels of integration, with hundreds of millions and even billions of devices integrated on the same die in modern-day digital microprocessors [9,10], require ever tighter control of device parameters in order to ensure that if not all, then at least most of the devices integrated in a single die behave in a predictable manner that does not compromise the functionality of the underlying product.

2.3.1 Historical Perspective

Issues related to device variability have always been a central theme in semiconductor manufacturing. While variability may be perceived by many as an emerging concern in modern-day sub-100-nm CMOS technologies, it has in fact been a well-studied and carefully monitored quantity starting from the earliest days of CMOS design. Shockley is one of the first to examine the phenomenon of semiconductor device variation in his study of the random fluctuations in the breakdown voltage of p-n junctions [11]. His work is later extended by Keyes [12] to explain the effect of dopant fluctuations on the performance of FETs. Traditionally, fluctuations in the threshold voltage have been the primary source of device performance variability, and as such, have been most closely studied. However, as scaling trends have continued to progress over the years, many additional sources of device variability have been identified, and as the complexity of integrated circuits has grown, so has the sophistication in characterizing, modeling and measuring device variability.

2.3.2 Impact of CMOS Variability

The effect of device variability on the performance of digital logic circuits is primarily expressed in terms of overall variability in delay and power consumption. Both quantities are a function of the current-driving and capacitive characteristics of the device, with delay being proportional to capacitance and inversely proportional to current, active power consumption being directly proportional to capacitance, and static power consumption being directly proportional to leakage current. Therefore, in order to be able to examine the true impact of device variability, the effects of variability in both the current-voltage (I-V) and capacitance-voltage (C-V) characteristics of the device must be studied. As is discussed below, most work focuses on either studying overall variations in delay, or variations in the I-V characteristics of the device, whereas the impact of C-V variability, especially at circuit-representative geometries, is largely ignored. More importantly, there has been no work showing the relationship between variability in the C-V and I-V characteristics of the device, which are ultimately based on many of the same physical device properties, and are expected to show correlation.

2.3.3 Classification of CMOS Variability

The sources of device variability in a CMOS process can generally be classified as static and dynamic [13]. Dynamic sources of variability are manifested as either gradual changes in device characteristics over time, most notably bias-temperature instability (BTI) effects leading to shifts in the effective threshold voltage of a device, V_{th}, [14–16], or noise, and in particular, low-frequency random telegraph noise (RTN). RTN is the result of trapping and de-trapping of discrete charges at the channel/oxide interface of the device, and is generally modeled as sudden quantized jumps in V_{th} giving rise to corresponding fluctuations in the channel conductance [17]. This type of noise is mainly associated with small-area devices, and is closely related to the aforementioned BTI effects [18], as well as $1/f$ noise [19]. The measurement, characterization, and modeling of RTN as a source of device variability are a primary focus of this work and are discussed in detail in Chapter 5.

Static device variability refers to variability in the electrical characteristics of the transistor as a result of uncertainties in the manufacturing process, and can be either random or systematic [20]. Random variability is defined in terms of mismatch between nominally identical devices situated in immediate proximity to one another and is generally caused by random fluctuations in the underlying physical characteristics of the device. On the other hand, systematic variations manifest themselves as well-defined gradients across the die, known as within-die (WID) variations, gradients across the wafer, known as die-to-die (D2D) variations, and differences between mean parameter values across wafers, known as wafer-to-wafer (W2W) variations. By definition, systematic variability can be traced to a deterministic source. In terms of static variability, this work focuses on the measurement and characterization of random variations as well as within-die (WID) systematic variations.

2.3.4 Sources of Systematic Variability

Systematic variability is generally caused by fluctuations in different parameters of the manufacturing process. Issues related to pattern density [21, 22], channel stress [23–25], mask misalignment [26], and across-wafer gradients due to rapid thermal anneal [13], are amongst many reported. Identifying the actual sources of systematic variability can be challenging without intimate knowledge of the manufacturing process, and as such, is not a central focus of this work. However, measured gradients in electrical parameters can still be mapped to systematic variations in the physical properties of the device, as is shown in Chapter 4. Identifying the physical causes of systematic variation is critical in determining the underlying source.

2.3.5 Sources of Random Variability

There are a number of physical factors that contribute to random variability in the electrical characteristics of an FET. The general tendency dictated by Moore's law is that the variance of these physical parameters should scale proportionally to device dimensions, in order to keep yields at acceptable levels. However, as the critical dimensions of CMOS transistors

scale close to and beyond fundamental physical dimensions, such as atomic dimensions or the wavelength of light used for patterning, the task of keeping variability at bay is becoming ever more challenging. There are a number of physical sources of random device variability that manifest as variability in the electrical properties of the transistor, both in terms of its C-V and I-V characteristics. Four of the most dominant sources are introduced below.

Random dopant fluctuation

Random dopant fluctuation (RDF) refers to variations in the number of dopant atoms in a device, which arise from the inability to precisely distribute a given concentration of dopant atoms throughout the channel of the transistor. Due to the aggressive scaling in device dimensions, the total number of dopant atoms per device is rapidly diminishing, reaching values on the order of 100 in a 45-nm process [27]. As a result, even single-atom fluctuations can have an appreciable effect on V_{th}, and the impact of RDF accounts for upwards of 60% of the threshold voltage variability [20]. The effects of RDF are expected only to worsen with new technology nodes, as the relative variation in threshold voltage begins to creep up [28]. This is one of the main reasons why supply voltages in recent technology nodes have not scaled according to the principles of constant-field scaling. In order to find a long term solution to this problem, different device topologies involving undoped or lightly-doped silicon are investigated [29, 30].

RDF mainly has an impact on the threshold voltage, V_{th}, although some influence on the effective carrier mobility, μ_{eff}, can also be expected due to Coulomb scattering [31]. The Coulomb interaction between mutually-repellant carriers and dopants in the channel limits the effective mobility of the carriers. As the uncertainty in the number and distribution of dopant atoms in the channel grows, so does the corresponding uncertainty in the effective carrier mobility.

Variations in V_{th} can be expected to have an effect on both the I-V and the C-V characteristics of the device, and variations in μ_{eff} have an effect only on the current-driving characteristics of the device. In general, the effect of RDF on V_{th} is dominant,

whereas μ_{eff} is expected to be more sensitive to variations in other process parameters, namely mechanical stress in the channel [32], as discussed below.

Line-edge roughness

Line-edge roughness (LER) is another major source of variation in modern-day CMOS processes [13,27,33]. LER results from statistical variation in the number of photons incident on the sample during lithographical exposure, and the absorption rate, chemical reactivity, and molecular composition of the photoresist [13]. Roughness along the edges of the gate can be expected to result in variability in the effective dimensions of the device, as shown in Fig. 2.3.

Since the length, L, of the device is generally considerably smaller than the width, W, LER is expected to couple into the I-V characteristics of the device mainly as an effect due to a variation in L. The impact on the drain current, I_D, is established through the inverse relationship between I_D and L, as well as through second-order effects on V_{th}. In terms of its impact on the capacitance of the device, LER is expected to influence the variability in the gate capacitance through the proportional relationship between the gate area, given by the product of W and L, and the intrinsic gate-to-channel capacitance, $C_{GC,int}$.

While discussing the effects of LER on device variability, it is important to note than maintaining a scaling trend in the variability of the critical dimensions (CD) of the device is crucial to continual scaling. As a result, much effort is put into maintaining these trends, as is discussed in [27], where it is shown that the variability in CD is very closely monitored and scales at the same rate as the technology nodes, at least down to the 45-nm node. As a result, one would not expect to see a tremendous impact of LER on device performance at 45-nm, in contrast to RDF, for instance. However, being able to monitor and characterize the variations in CD is still of great interest in terms of ensuring that these scaling trends are maintained [34].

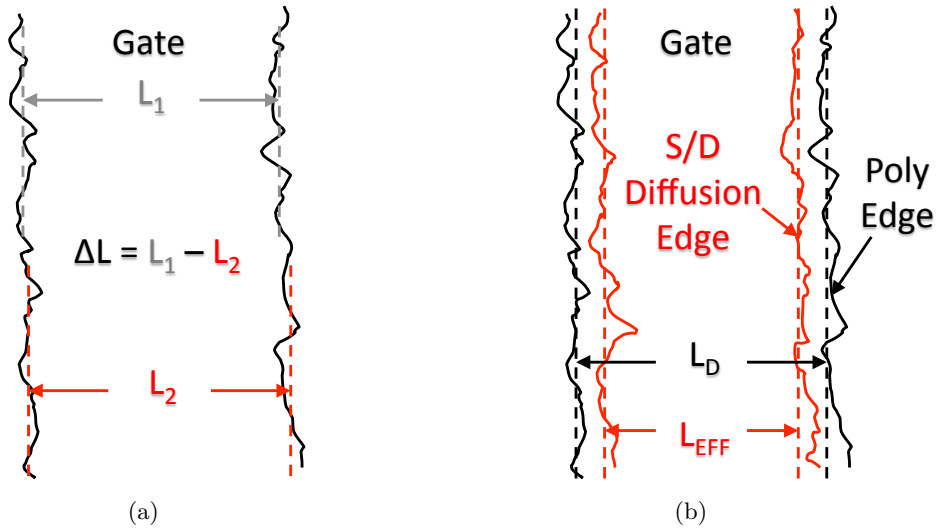
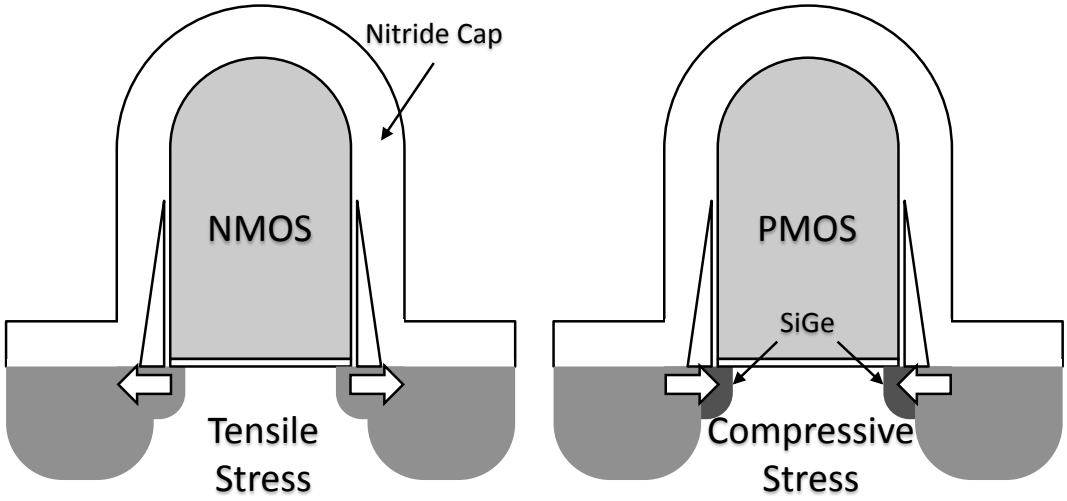


Figure 2.3: (a) An example of different effective lengths across different areas of the same poly-silicon gate due to LER, and (b) difference between LER of the gate and of the source-drain edges, which actually define the effective length of the device - the two quantities are highly correlated, but not the same.

Channel stress fluctuations

Channel stress has a fundamental impact on carrier mobility, a phenomenon widely exploited in modern-day CMOS devices in order to overcome the increasing degradation of carrier mobility caused by channel impurity scattering [35]. Due to the difference in charge carriers, NMOS devices have to be subjected to tensile stress and PMOS devices have to be subjected to compressive stress in order to enhance the effective channel mobility. While application of stress as a mobility enhancement technique is common to both device polarities, the mechanisms for producing the two distinct types of stress are different. In particular, nitride capping layers over the NMOS devices give rise to tensile stress, while embedded *SiGe* source/drain epitaxial layers give rise to compressive stress in the PMOS channels, as shown in Fig. 2.4. As a result, mobility variations due to stress can be expected to be different for each device type, as the controllability of the two types of stress would be different based on the two different stressing techniques used.

Stress-induced variability is most often considered as a source of systematic variability, where the stress caused by the device environment deterministically causes a change in



(a) (b)

Figure 2.4: Application of channel stress for mobility enhancement: (a) tensile stress in an NMOS device using of a nitride capping layer, and (b) compressive stress in a PMOS device using a $SiGe$ diffusion implant.

performance [23–25]. However, random variations in stress are also expected to contribute to random variations in μ_{eff} [32, 36, 37]. In terms of the effects on the electrical characteristics of the device, variations in stress have a significant effect on the I-V characteristics through modulation of the channel mobility. However, in terms of the C-V characteristics of the device, carrier mobility, and consequently, channel stress, are expected to have no notable impact.

Interface roughness and defects

As the oxide thickness, t_{ox}, scales down to atomic levels, reaching as low as 2.4 nm in a 45-nm CMOS process, oxide roughness of one or two atomic layers can have a significant impact on device characteristics [38]. Variations in t_{ox} have a direct impact on the oxide capacitance per unit area, C'_{ox}, and as such, affect both the C-V and I-V characteristics of the device. Additionally, gate leakage is extremely sensitive to variations in t_{ox}. Due to this sensitivity, it is expected that t_{ox} is a variable, which is extremely-well controlled, both in terms of growth of the epitaxial layers, as well as post-processing using chemical-mechanical polishing (CMP) [27]. Due to the increased sensitivity to variations in the

oxide thickness, modern-day CMOS technologies are moving towards high-k dielectric gate materials, which enable scaling of the gate capacitance while maintaining a larger physical dielectric thickness.

Regardless of the type of dielectric used, defects in the channel-oxide interface are bound to occur, giving rise to interface trap states and the associated random telegraph noise (RTN). As mentioned above, RTN is a considerable source of dynamic variability in device performance and is discussed in more detail in Chapter 5.

2.3.6 Modeling Random Variability

Pelgrom [39] makes the observation that the variance of device parameters scales inversely with the device area based on general principles regarding the spacial averaging of random variables across device area. This idea is fundamental, in that it points to the need to consistently scale variability in device parameters with each new technology node in par with the reduction in device area, in order to maintain acceptable yield levels. Of course, as mentioned earlier, maintaing high yield is fundamental to the formulation of Moore's law in its original form, where scaling is expressed in terms of cost per transistor. The two results from Pelgrom's work quoted most often are the relationships regarding the scaling of the variability in the random mismatch of threshold voltage, ΔV_{th}, and the relative current factor, $\Delta\beta/\beta$, which are given by

$$\sigma^2_{\Delta V_{th}} \propto \frac{1}{WL} \tag{2.1}$$

and

$$\sigma^2_{\Delta\beta/\beta} \propto \frac{1}{WL}, \tag{2.2}$$

where $\beta = \mu_{eff} C'_{ox} W/L$.

Drennan [40] expands on the idea by distinguishing between physical device characteristics and electrical model parameters, and stipulating that the physical parameters, such as oxide thickness, doping, device dimensions, carrier mobility, etc. are the ones that average away with increase in the physical dimensions of the device. However, more than one physical parameter can have an impact on any given electrical parameter, and for accu-

rate modeling of variability, the variance of each physical parameter has to be propagated to the variance of the electrical parameter. In particular, the variance, σ_e^2, of an electrical parameter, e, can be expressed in terms of the variance, $\sigma_{p_i}^2$, of all physical parameters, p_i, and the electrical parameter's sensitivity to them, as given by

$$\sigma_e^2 = \sum_i \left(\frac{\partial e}{\partial p_i} \right)^2 \sigma_{p_i}^2(W, L). \tag{2.3}$$

More notably, Drennan also points out that while most physical parameters indeed average out over area, with

$$\sigma_{p_i}^2 \propto \frac{1}{WL}, \tag{2.4}$$

the fluctuations in device length, L, and width, W, caused by line-edge roughness (LER), do not. Instead,

$$\sigma_{\Delta L}^2 \propto \frac{1}{W} \tag{2.5}$$

and

$$\sigma_{\Delta W}^2 \propto \frac{1}{L}. \tag{2.6}$$

This sort of analysis is also present in Pelgrom's work, where the author consideres the effect of $\sigma_{\Delta L}$ and $\sigma_{\Delta W}$ on $\sigma_{\Delta\beta/\beta}^2$, but ultimately argues that variability due to LER can be neglected if the dimensions W and L are large enough.

Pelgrom scaling and propagation of variance (POV) form the basis for most rudimentary variability modeling used in design, and are the primary variability models considered in this work.

2.4 Methods for Variability Characterization

Different techniques for variability measurement and characterization have been proposed to enable the monitoring and analysis of device variability and its impact on circuit performance. In general, when measuring device variability, a few important characterization parameters must be considered. The ability to measure large sample sets in order to extract the tails of statistical distributions with a high level of confidence is critical. Therefore,

characterization methodologies which allow high degrees of integration are preferred. Measuring large sample sets requires fast measurement acquisition times, which also makes highly integrated approaches appealing. On the other hand, if the physical sources of variability are to be identified, detailed measurements of different device characteristics need to be performed, preferably on the same device sample set. Measurement techniques which enable this are generally implemented using off-chip characterization equipment, making them time consuming and unsuitable for characterization of large device sample sets.

In this section, different popular variability characterization techniques are examined with respect to their ability to meet the specifications described above. While each has its own specific advantages, none of the popularly used techniques manage to satisfy all of the desired criteria. Consequently, a novel variability characterization methodology is proposed, which enables both rapid and detailed device characterization of large statistical sample sets.

2.4.1 Ring Oscillators

Ring oscillators (RO) and other delay-based characterization methods are widely employed in research and industry as tools for circuit and device variability measurement [41–47]. They offer a highly integrated approach to variability characterization, where variability in the delay of simple circuits (inverters or other standard gates) is measured through the means of measuring variability in the RO frequency of oscillation - a measurement, which can be done in a purely digital fashion. Such measurements can be performed rapidly and in large volumes. In an RO with N number of stages, the delay of a single stage, t_D, is related to the frequency of oscillation, f_{osc}, according to

$$t_D = \frac{1}{2Nf_{osc}}. \tag{2.7}$$

The delay metric, t_D, is arguably one of the most important metrics characterizing digital circuit performance, and it incorporates variations in both the current-driving and capacitive characteristics of the underlying devices.

However, there are significant drawbacks to RO characterization as well. As gate delay decreases with new technology nodes, generating a characteristic frequency, f_{osc}, that

is low enough to be accurately sampled necessitates RO structures with a large number of stages, especially if the frequency is not measured on-chip. This results in decreased measurement sensitivity, as the ratio of the standard deviation to the mean of the measured oscillation period, T_{osc}, decreases with increasing N, as given by

$$\frac{\sigma(T_{osc})}{\mu(T_{osc})} = \frac{\sqrt{2N}}{2N}\frac{\sigma(t_D)}{\mu(t_D)}. \tag{2.8}$$

More importantly, RO measurements integrate together all possible sources of variation into a single metric, making it impossible to determine the individual contribution of different sources of variability. In some cases, this issue can be addressed by comparing measurement results from ROs comprised of similar but different device structures, as reported in [41,48], and then assigning the relative difference in measured frequency to the differences in circuit and device topologies. However, such an approach still fails to distinguish between effects due to I-V and C-V variability.

Overall, the popularity of ROs stems from the ease of characterization they offer, the ability to integrate them alongside functional circuitry, and the direct observation of variation in the the delay metric. They can be very useful as general purpose variability monitors, but do not give sufficient information about the underlying physical sources of device variability.

2.4.2 Device Simulation and Imaging Techniques

On the opposite side of the spectrum from ROs lie a different set of variability characterization techniques, which depend on atomistic device modeling and imaging. In terms of using device simulations to model atomistic effects that lead to variability in different electrical device parameters, Asenov and his students have demonstrated a wide array of applications. They have used atomically correct device simulations based on their Glasgow simulation engine to study variability in both I-V and C-V device behavior at small geometries due to random dopant fluctuations, line-edge roughness, channel stress, and channel/oxide interface traps [33, 37, 49–54].

An atomistic device simulation methodology is well-suited for studying the impact different physical sources of variability have on electrical device characteristics. Also, given enough computational power, a Monte-Carlo approach can be used to gather a large number of statistical samples as needed to accurately extract different parameter distributions. However, simulated results have the disadvantage of predicting only phenomena incorporated in the model, and in general, the complexity of a real device can never truly be captured by any model, no matter how complex. Therefore, while atomistic device simulation is an indispensable tool in the detailed study of the relationship between the physical sources of device variability and their electrical manifestations, a methodology that is based on measurements of actual devices is still needed.

In addition to atomically correct device modeling, different device imagining techniques can also be used to directly observe physical sources of variability. As an example, scanning electron microscopy (SEM) can be used in the study of line-edge roughness by top-down imaging of transistor gates and digital extraction of gate contours from the SEM image [33, 55]. Such a technique is appealing in terms of studying in detail the different physical aspects of LER, such as the autocorrelation of the edge roughness, but is rather time consuming and taking a large number of samples is impractical. Moreover, as shown in Fig. 2.3, the LER associated with the poly gate is not the actual parameter of interest; instead, the LER associated with the source/drain diffusion edges is what affects variability in the electrical device characteristics.

In addition to SEM imaging, a new technique named laser-assisted atom probe tomography (APT) has been used to generate 3-D images of individual dopant atoms in a MOSFET structure, giving a much more informative view of the physical characteristics of the device [56, 57]. This technique is geared towards studying how atomic-scale phenomena contribute to device variability, but similarly to all other imaging techniques, it is impractical in terms of overall variability characterization due to the inability to collect a large number of samples over a wide device parameter space.

2.4.3 Electrical Device Characterization

Perhaps the most robust way of characterizing the variability in MOSFETs is through electrical measurements. Direct measurements give the best indication of the variability in the electrostatic properties of the device, which is ultimately what gives rise to the variability in circuit performance. At the same time, methodical electrical characterization employing different I-V and C-V measurements can give a good insight into the actual physical sources of device variability through established relationships between physical device characteristics and electrical device properties. Depending on the level of integration and measurement functionality, direct device characterization can provide detailed measurements of large statistical device sets at acceptable rates, and can potentially result in a variability characterization methodology which offers a balance between detail, accuracy, sample size, and measurement throughput.

Most direct characterization approaches applied to variability measurements focus on the I-V characteristics of the device. This is primarily due to the fact that dc currents are easy to measure and most aspects of variability are reflected it the I-V characteristic of the transistor. I-V variability characterization comes in three different levels of test circuit integration. The first level is the traditional direct probing approach, which involves routing individual pads to each of the terminals of the device under test (DUT) [22, 30, 58, 59]. Even if small probe pads are used, this approach still results in small statistical sample sets and is generally not applicable to large-scale variability studies. Additionally, off-chip current measurements of low current signals can be slow due to the need to charge the large parasitic capacitances associated with the measurement probes, cables, and equipment at every sampling step.

In order to enable the study of large statistical DUT sample sets, different array-based characterization approaches have been employed [60–65]; these approaches are able to achieve much higher DUT density as a result of the reduction of the number of probe pads. However, issues related to sub-optimal sampling rates due to the large parasitics associated with taking measurements off-chip still remain.

Recent work attempts to address these issues by integrating the analog measurement circuitry on-chip, alongside a large addressable DUT array [48, 66]. This approach overcomes many of the disadvantages associated with direct electrical measurements, in that it allows large device sample sets to be characterized at fast sampling speeds, while also offering a digital measurement output and removing the need for analog measurement equipment. However, the on-chip characterization system designs described in [48, 66] still require sweeping an analog input in order to perform voltage sweep measurements, and are limited to I-V characterization only.

In terms of studying the variability in the C-V characteristics of FETs, few results can be found in the literature referring to modern-day CMOS capacitor characterization. Charge-based capacitance measurements (CBCM) have been successfully employed to perform atto-Farad resolution back-end-of-line (BEOL) characterization [67,68]. More recently, CBCM approaches have been successfully applied in front-end-of-line (FEOL) characterization to characterize the voltage-dependent capacitance of MOS transistors with sub-femto-Farad resolution [69,70]. However, most work focuses not on variability, but rather on using high-accuracy C-V measurements of individual devices for parameter extraction.

When referring to capacitance variability measurements, two recent publications stand out. One is the work by Polonsky et. al. [71], where a variation of CBCM called quadrature-voltage capacitance measurement (QVCM) specifically designed to overcome C-V characterization limitations due to gate leakage through the DUT is used to characterize an array of devices in a 45-nm SOI CMOS process. However, while sub-femto-Farad resolution C-V measurements of circuit-representative devices is indeed demonstrated, the statistical sample set measured consists of only 11 DUTs and the measurements from only one set of devices at circuit-representative dimensions is presented. The other example of CMOS C-V variability measurements is by Tsuji et. al. [72], where CBCM is once again used to measure C-V curves of devices with different dimensions in statistical sample sets of 24 DUTs and even some Pelgrom variability analysis of MOSFET capacitance is reported. It should be noted, however, that this work, while published in 2011, reports results from a

seemingly older technology node, with the smallest length of devices measured set at 0.12 μm. Presumably, gate leakage is not an issue at this technology node, which greatly simplifies C-V characterization. In both instances, only the DUT arrays and the supporting CBCM switching circuitry are integrated on chip, with all measurements performed off-chip at CBCM measurement frequencies of 1 MHz or below.

2.4.4 Proposed Approach

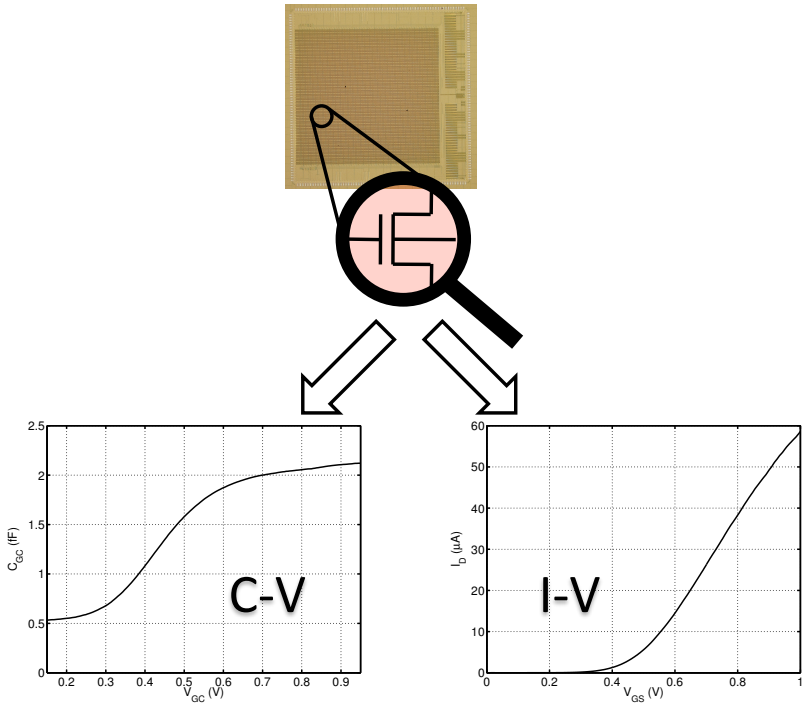


Figure 2.5: An illustration of the proposed combined C-V/I-V characterization approach; C-V and I-V characteristics for the same circuit-representative device are measured.

The CMOS variability characterization methodology proposed in this work is based on an on-chip electrical characterization approach where both the C-V and the I-V characteristics of the device are extracted simultaneously, as illustrated in Fig. 2.5. Such an approach enables complete quasi-static device characterization, allowing for the correlation between the variability in the two characteristics to be studied. This will be shown to be useful with regards to extracting information about the underlying physical phenomena responsible for the measured variability in the electrical characteristics of the device.

An on-chip system for combined C-V/I-V characterization with a digital I/O interface is designed and implemented in a 45-nm bulk CMOS process, alongside a large addressable DUT array. Complete measurement system integration aims at improving measurement throughput as need for large-volume data acquisition. Moreover, on-chip integration makes high-frequency current measurements possible at low signal levels, enabling the characterization of random telegraph noise (RTN), in addition to quasi-static C-V and I-V characterization. Large statistical sample sets of different DUT types spanning a number of design parameters are available for characterization. All studied devices have circuit-representative dimensions, allowing accurate variability measurements at relevant device sizes. A novel CBCM characterization technique with atto-Farad measurement resolution is developed for the purpose of C-V characterization of small-area DUTs. To the author's knowledge, this is the first time when such a comprehensive electrical device characterization system has been implemented at such high levels of integration, and the only published work to date presenting results on combined C-V/I-V variability characterization of large sample sets at an advanced technology node. The proposed approach addresses issues related to sample set size and measurement throughput through complete on-chip integration, as well as issues related to measurement detail and ability to identify physical sources of device variability through combined C-V/I-V characterization.

Chapter 3

On-Chip Characterization System

3.1 Introduction

Chapter 3 details the design of the on-chip variability characterization system and its individual components. The chapter begins by introducing a top-level system overview. Emphasis in placed on complete system integration and component modularity. The design of each of three main system components – addressable device-under-test (DUT) array, biasing digital-to-analog converted (DAC), and measurement unit (MU) – is described in detail. Simulation results and characterization measurements are included where applicable. Overall, the system is demonstrated to have all of the desired characteristics, including full on-chip integration of measurement and stimulus circuitry, complete digital I/O interface, and high-resolution current and voltage characterization capabilities. System implementation in a 45-nm bulk CMOS process and the associated experimental measurement setup are presented.

3.2 System Overview

Fig. 3.1 shows a simplified top-level schematic of the on-chip characterization system. The system consists of three major blocks – an on-chip switching matrix, used to individually address transistors from the device-under-test (DUT) array, a four-channel digital-to-analog

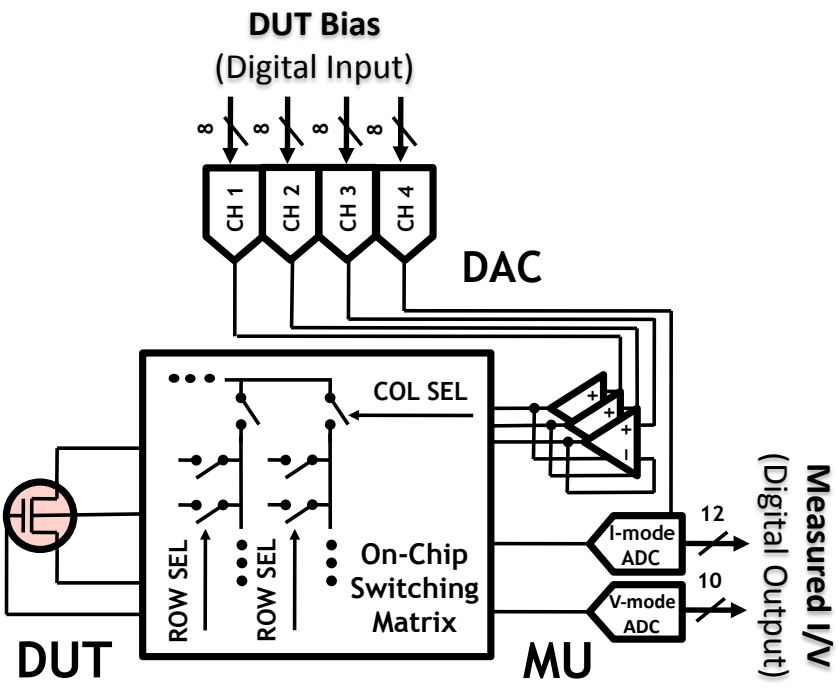


Figure 3.1: A simplified top-level schematic of the on-chip characterization system.

converter (DAC), used to supply each of the four DUT terminal bias voltages, and a measurement unit (MU), which consists of a current- and voltage-mode analog-to-digital converters (ADCs), used to perform accurate on-chip current-voltage (I-V) and charged-based capacitance-voltage (C-V) characterization. The system can be configured to characterize both NMOS as well as PMOS devices by adjusting the analog references and internal controls to account for opposite current polarities.

The design of the on-chip measurement system is based on complete on-chip integration of the entire MOSFET characterization infrastructure. The system has a digital-in/digital-out measurement interface, making it compatible with a purely digital test flow. Any need for accurate and expensive bench-top measurement equipment is completely removed. The only analog signals going to the chip are dc current and voltage reference signals as required by the DAC and ADCs, as well as dc currents for biasing on-chip analog circuits. As far as the author is aware, this is the highest level of integration achieved in any published system for on-chip device characterization. As discussed in Chapter 2,

achieving high levels of on-chip measurement integration not only removes the need for an analog signal interface, but also has the potential to greatly improve measurement throughput, which is an important parameter of any characterization setup, especially when large sample volume is considered.

One potential benefit of complete on-chip integration is the ability to design each of the individual system components according to specific characterization needs. However, this particular system is designed with generality in mind, and can in fact be used for a variety of electrical measurements on large statistical device sample sets. This flexibility is demonstrated by applying the characterization system in the context of both combined C-V/I-V dc measurements and random telegraph noise (RTN) time domain measurements, with the potential for many more experiments still there. The generality of the proposed characterization methodology is one of its main appeals, and it should be noted that the described approach can be extended to I-V and C-V variability characterization of other front-end-of-line (FEOL) components, as well as back-end-of-line (BEOL) components with very little overhead.

The on-chip characterization system is designed with maximum process compatibility in mind. Although there are a number of analog circuit blocks integrated on chip, no special analog devices or process options are used in the design of these blocks. Instead, basic thick-oxide I/O transistors are used to implement all analog functionality. These devices can be operated at 2.0 V supply, allowing for the necessarily voltage headroom to bias the digital DUTs at voltages of up to 1.1 V. Additionally, these devices have a minimum length of 0.44 μm, which results in a comparatively larger intrinsic gain, due to the lack of various short-channel effects, at the expense of reduced intrinsic speed. Since the sampling rates of interest are below 1 MHz, the reduced speed is not an issue, whereas the increased gain is essential for accurate analog-to-digital conversion.

The design is highly modular and individual circuit blocks are reused whenever possible. While this may not be the most efficient design approach, power efficiency and, to a lesser degree, area efficiency are not of significant concern in this particular design.

Instead, a more fundamental design goal is to simplify the porting of the characterization system to new technology nodes. This is accomplished in a highly-modular design, as ultimately only a small number of sensitive analog blocks have to be redesigned in the new process. Using thick-oxide devices helps in this regard as well, since these devices tend to remain largely unchanged from one technology node to the next, further reducing the necessary redesign effort.

3.3 On-Chip Switching Matrix and DUT array

3.3.1 DUT Array

A simplified representation of the integrated DUT array and the accompanying switching matrix is shown in Fig. 3.2. Column-select and row-select signals from one-hot shift registers are used to sequentially address individual DUTs from the array. The DUT array spans 40 rows of DUTs across 56 columns. Each column contains 39 identical DUTs, in addition to one empty array cell used as a null reference. Two neighboring columns contain matched DUTs of the same type, as indicated in Fig. 3.2, allowing any gradients along the height of the DUT column to be cancelled out differentially. Overall, the measurement sample set consists of 28 different DUT types with devices spanning different lengths, widths, threshold voltages, and environments, with a statistical set size of 78 DUTs per DUT type. Two test arrays containing NMOS and PMOS DUTs are weaved parallel to one another, sharing many of the global digital control signals, but with completely electrically isolated analog signal paths and dedicated on-chip characterization systems.

A detailed view of the individual DUT cell and the associated switches is shown in Fig. 3.3. The relative sizing of the switches is indicated on the schematic. Each terminal can be connected to a voltage sense path, which is ultimately routed to the integrated voltage-mode ADC. Since this is a high impedance sense path, minimum size switches are used. The source and drain terminals have dedicated switches used to contact the device during measurement. Since the current measurement path is a low-impedance path, larger switches

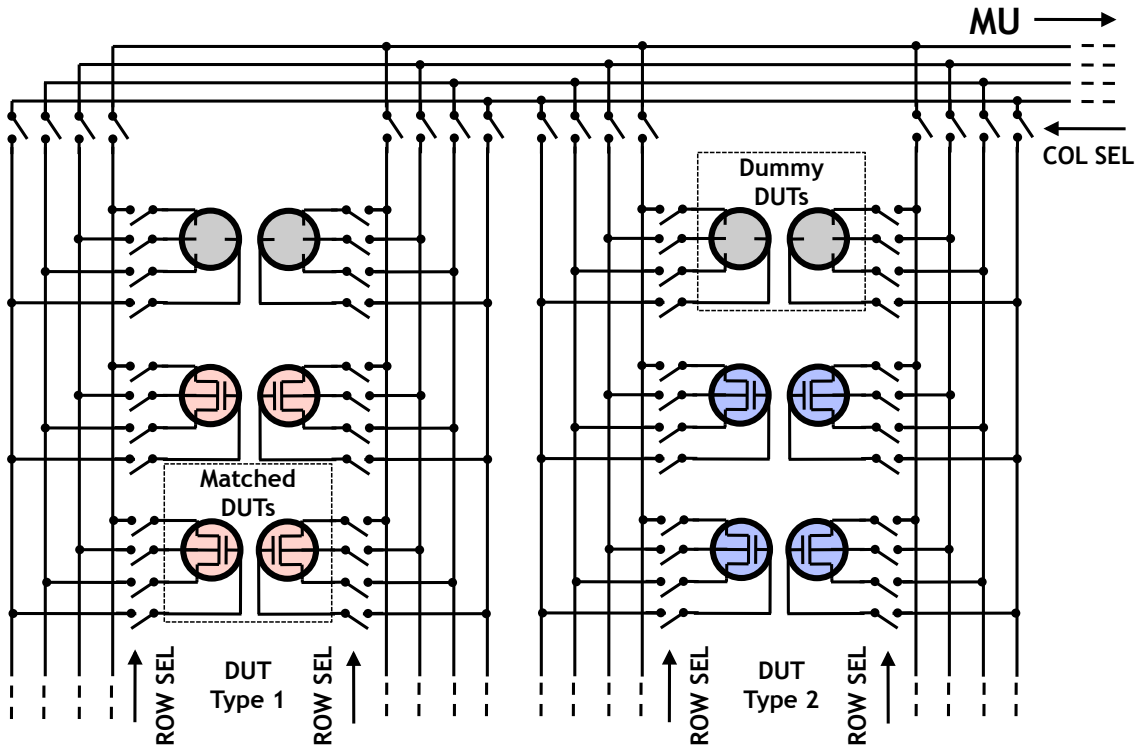


Figure 3.2: A simplified schematic of the DUT array and associated on-chip switching matrix.

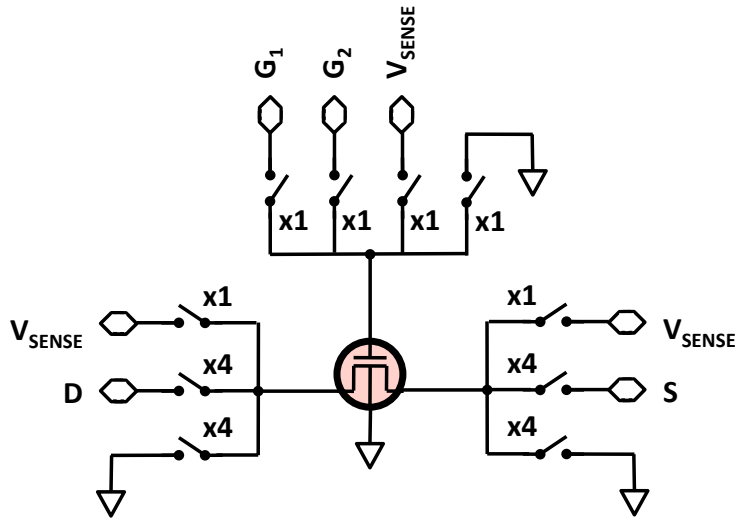


Figure 3.3: A simplified schematic of an NMOS DUT cell and the associated switches; different size switches are used depending on weather the switches connect to a high- or a low-impedance signal path.

are used. On the gate side, there are two switches, either one of which can be used to provide the gate bias. The need for two identical switches is described in detail in Chapter 4, where the charge-based capacitance measurement (CBCM) technique is discussed. Since the gate presents a high-impedance, minimum-size switches are used. Each terminal also has a path that allows it to be tied to ground. The source and drain terminals have to be discharged during the CBCM C-V characterization. Additionally, the gate terminal is tied to ground when the cell is not selected, in order to reduce leakage during an I-V measurement. In the case of an NMOS DUT, the body is tied to ground since the body potential is set by the shared substrate potential. In the case of a PMOS DUT, since each DUT has a dedicated N-well associated with it, the body can also be biased through a set of switches identical to those at the gate (not shown). Apart from this, the only other difference between the NMOS and the PMOS DUT cell is that the PMOS DUT cell enables all terminals to be shorted to the analog supply, $V_{DD,A}$, rather than ground, as needed by the opposite polarity C-V measurement.

3.3.2 Switch Design

Switches in the array are implemented using thick-oxide CMOS transmission gates, offering low leakage in the off state and and allowing high voltage swings. The schematic of a transmission gate switch is shown in Fig. 3.4(a). The NMOS and PMOS transistors are identically sized with $W/L = 3.9\ \mu m/0.44\ \mu m$ for a unit switch with a nominal on-resistance of approximately 500 Ω. In order to decrease the on-resistance of switches in low-impedance paths, the width of the transistors is proportionally increased. Equal NMOS and PMOS sizing tends to keep the resistance of the switch relatively constant over the entire bias range [73]. Since all digital control signals on-chip are generated and distributed using 1.1 V native-oxide devices, a level-shifter circuit is used to transition between 1.1 V ($V_{DD,D}$) and 2.0 V ($V_{DD,A}$) logic levels. A standard positive-feedback topology is used for the level shifter (Fig. 3.4(b)), which amplifies the digital logic signals and generates complimentary outputs as needed for the CMOS transmission gate. Each analog switch has a dedicated

level-shifter.

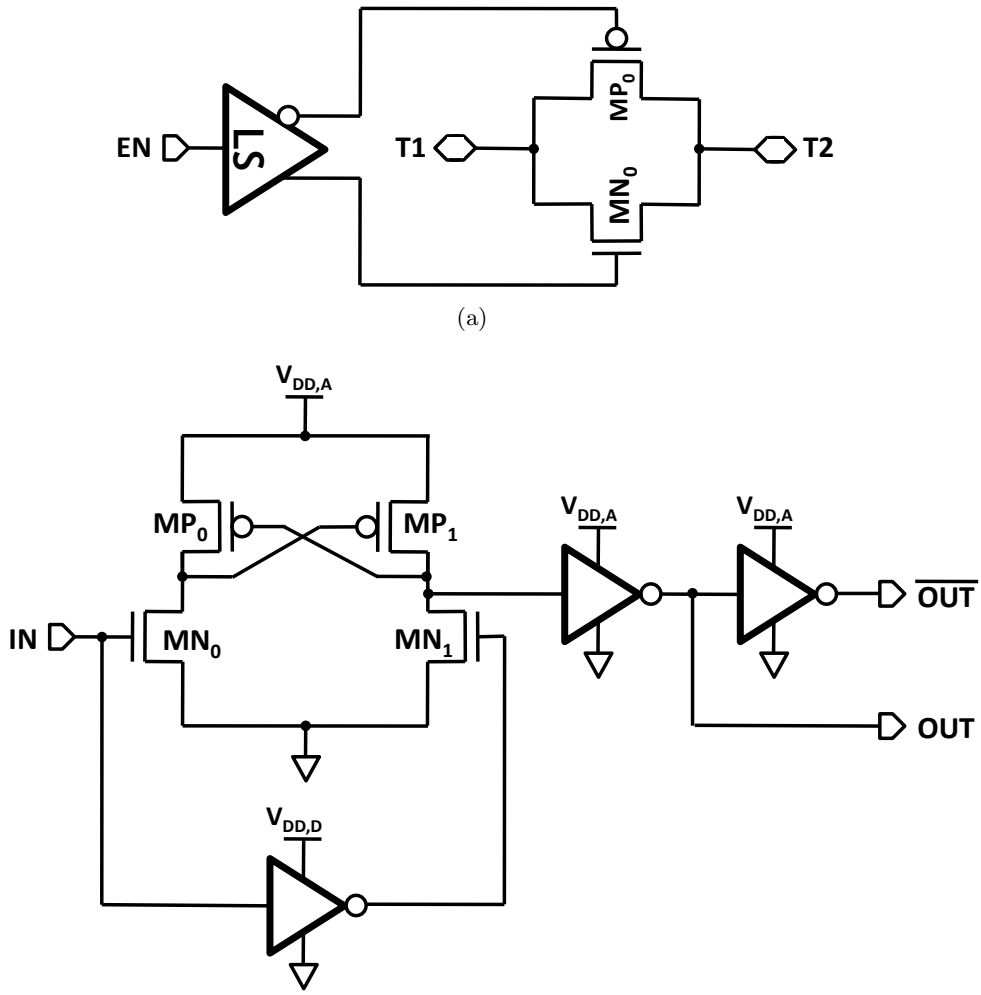


Figure 3.4: (a) Transmission gate implementation of analog switch; (b) a positive-feedback level-shifter used to step-up 1.1 V ($V_{DD,D}$) digital control signals to 2.0 V ($V_{DD,A}$) analog levels; Table 3.2 gives sizing for the transistors in the positive-feedback latch.

In order to be able to integrate a large number of DUTs in a single DUT array, it is essential that the switches in the on-chip switching matrix provide enough isolation between DUTs. To achieve the best isolation, the source and drain terminals of all DUTs but the one being measured are disconnected from the measurement circuitry, and their gates are connected to either ground or $V_{DD,D}$, depending on weather an NMOS or a PMOS DUT is measured, respectively. The series combination of the two disconnected switches and the DUT biased in its off state results in very high impedance. Fig. 3.5 shows a simulation for

Table 3.1: Level-Shifter Design Table

Component	Value
MP_0/MP_1	0.8 μm/0.44 μm
MN_0/MN_1	3.9 μm/0.44 μm

the parasitic current due to a large DUT in the off-state corrupting an I-V measurement, with the associated drain-to-source voltage swept between 0 and 1.1 V. While there is a somewhat appreciable leakage of about 27 pA, most of this leakage is not due to the swept potential and can be nulled during a calibration step using the empty DUT cell present in each column. The actual measured parasitic resistance of the DUT in the off state is shown to be $R_{OFF} = 1.37 \times 10^{15}\ \Omega$. Such a high off-resistance essentially means that the density

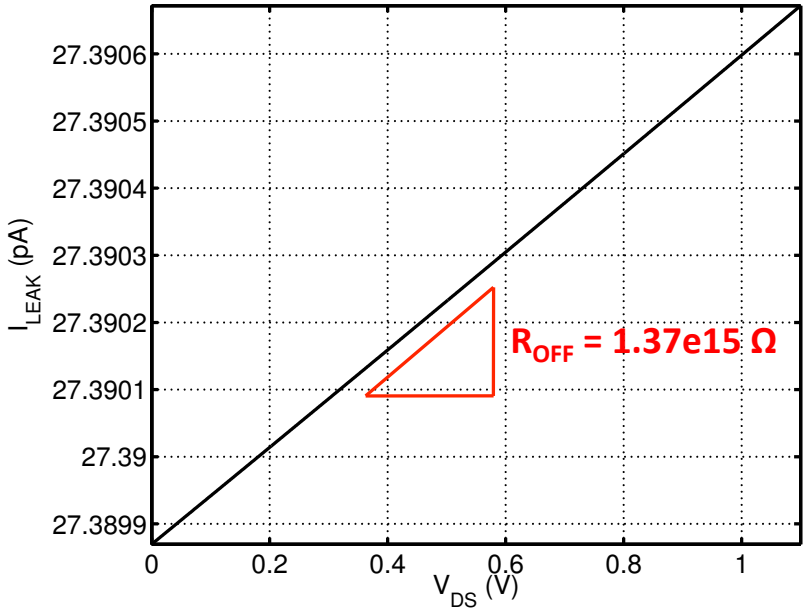


Figure 3.5: A simulation showing the excellent isolation of a DUT placed in the off state.

of the DUT array is limited only by area considerations and more complex array structures aimed at reducing switch leakage, such as the one presented in [63], are not necessary.

3.4 Biasing DAC

The four-channel resistor-string DAC shown in Fig. 3.6 is used to provide each of four DUT terminal bias voltages. The four DAC channels share the same resistor string reference,

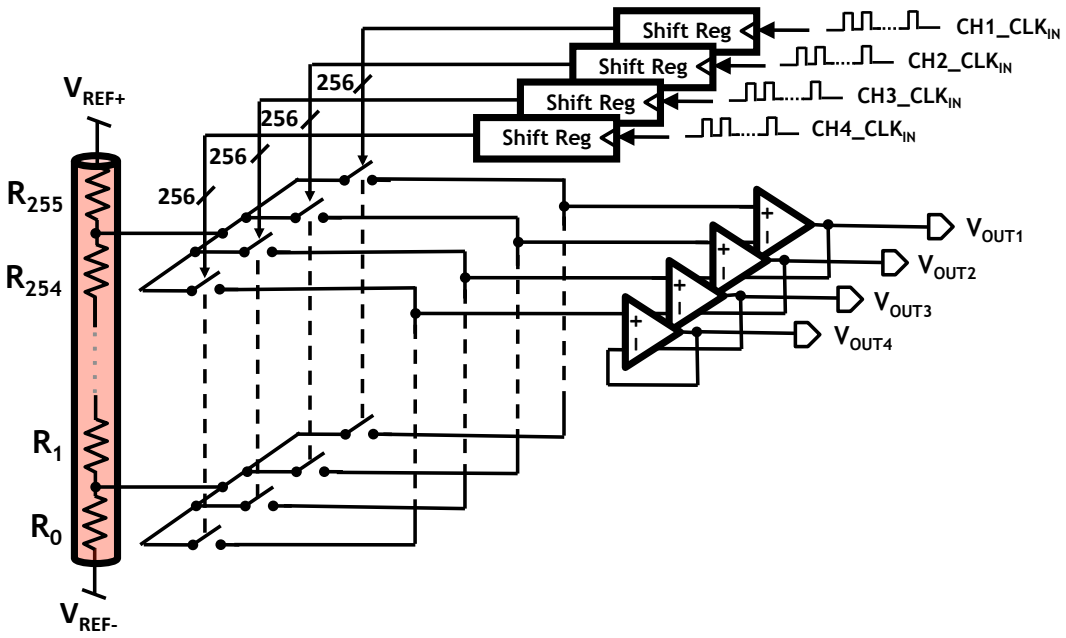


Figure 3.6: Four-channel R-string DAC with one-hot shift register control used for generating DUT bias voltages.

keeping the overall DAC area footprint small. Each channel has 256 different output levels, resulting in eight-bit precision. The control of the DAC outputs is implemented using four independent one-hot bi-directional shift registers. This unorthodox control scheme is ideally suited for generating voltage sweeps, as needed for I-V and C-V characterization, while also keeping the control circuitry simple and compact as compared to a decoder-based approach.

The unit resistor of the resistor string is implemented using a poly-silicon resistor with a nominal resistance of 67.5 Ω. The resistance value is chosen such that when all four channels are simultaneously switched to mid-code, which results in the worst RC charging time constant, the resistor string output settles to five time constants within 0.5 μs, resulting in a maximum DAC frequency of 2 MHz. The operating frequency of the DAC is designed to be one order of magnitude higher than the sampling frequency of the analog-to-digital converters, which can be as high as 200 kHz in the desired range of operation. As a result, biasing the device at each measurement step does not factor significantly into the overall characterization time.

Each of the four DAC channels is buffered using a single-stage load-stabilized analog

voltage buffer. The transistor-level schematic of the DAC buffer is shown in Fig. 3.7(a). Similarly to all other analog circuitry in this system, the voltage buffers are implemented using high-voltage thick-oxide devices. The simulated open-loop response of the buffer driving a 10.0 pF capacitive load is shown in Fig. 3.7(b). The buffer has an open-loop gain of 47.9 dB, a gain-bandwidth product (GBP) of 76.5 MHz, and a phase margin of 74.6 deg.

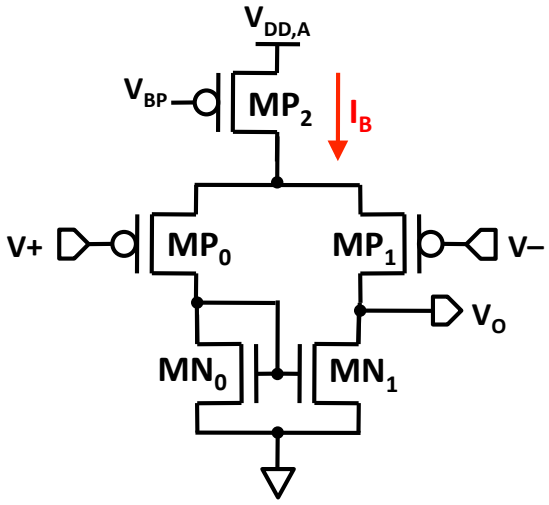
The measured differential non-linearity (DNL) and integral non-linearity (INL) of the DAC are shown in Fig. 3.8. Only two of the channels characteristics are plotted for clarity; measured performance of the other two channels is similar. The DAC reference voltages, V_{REF+} and V_{REF-}, are set to 1.4 V and 0.3 V, respectively, resulting in a least significant bit nominally set to $V_{LSB} = 4.3\ mV$. As expected, the resistor-string DAC offers excellent DNL performance, with absolute maximum DNL and INL of less than 0.1 V_{LSB}. While all terminal voltages are measured at the point of application making absolute DAC accuracy non-essential, as discussed in Chapter 4, it is still desirable to have a monotonic DAC, especially in the context of C-V measurements, where the differential of the applied voltage factors in. A DNL of less than 1.0 V_{LSB} guarantees such monotonicity. The non-linearity of the converter is mainly caused by mismatches in the shared resistor string, which is made evident by the similarity of the INL plots for the two channels.

Table 3.2: DAC Buffer Design Table

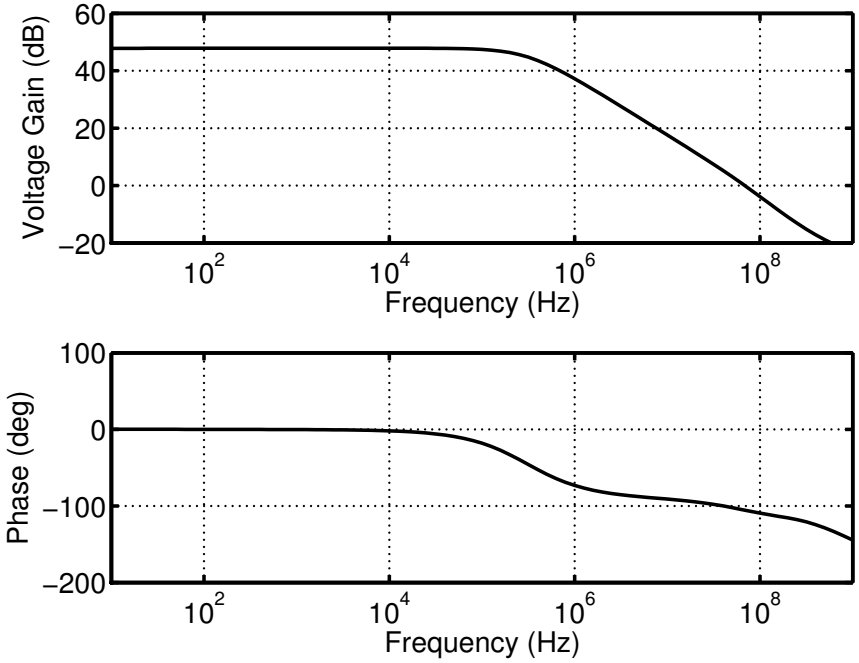
Design Component	**Value**	**Bias Current**	**Value**	**Simulation Result**	**Value**
MP_0/MP_1	600 μm/1.6 μm	I_B	0.7 mA	DC gain	47.9 dB
MP_2	400 μm/0.6 μm			Phase Margin	74.6 deg
MN_0/MN_1	300 μm/2.0 μm			GBP	76.5 MHz

3.5 Measurement Unit

The measurement unit (MU) consists of both current- and voltage-mode ADCs, as needed to perform accurate I-V and charged-based C-V device characterization, as well as voltage



(a)



(b)

Figure 3.7: (a) Schematic for the single-stage, load-stabilized DAC buffer; (b) simulated open-loop frequency response with a 10.0 pF capacitive load.

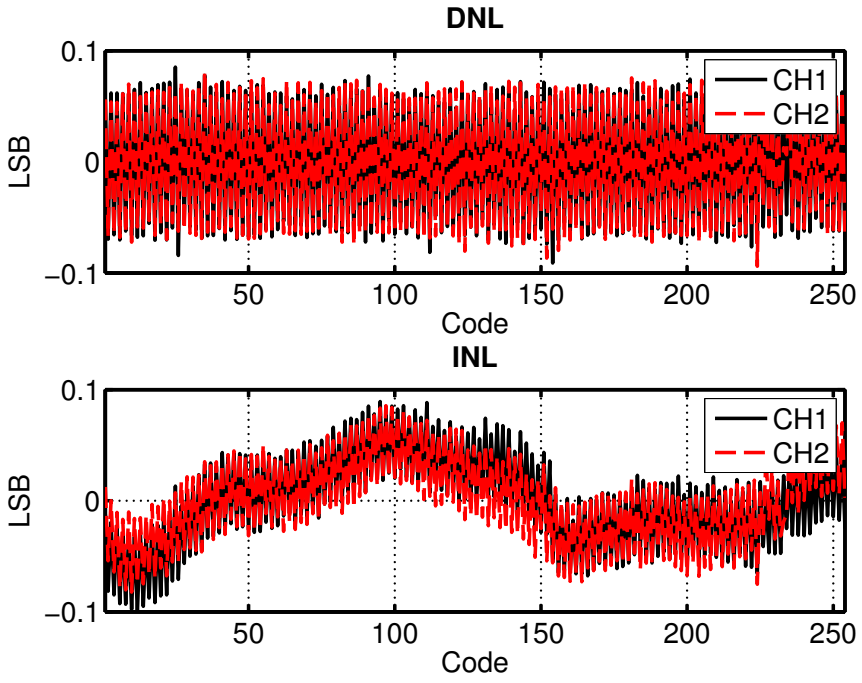


Figure 3.8: Measured linearity performance for two channels of the R-string DAC: (top) DNL and (bottom) INL; excellent linearity is observed.

buffers for biasing the DUT. The two types of ADCs are both based on a dual-slope integrator topology, which offers excellent measurement characteristics, including high tolerance for variation in the passive components [74] and high immunity to noise and other interference [75]. The MU is the most sensitive analog block in the characterization system, as its performance ultimately determines the precision of all measurements. As is demonstrated below, this analog block is designed with modularity and design reuse in mind, ensuring maximum portability to new and potentially immature technology nodes. A single general-purpose operational amplifier (op-amp) forms the basis of most analog functionality, including voltage buffering and analog-to-digital conversion. Design modularity is also extended to a higher level of abstraction in the design of the current- and voltage-mode ADCs, with both converters implemented around the same integrator core. High-precision analog-to-digital conversion is achieved without the need for any special analog process options, keeping in line with the goal of design portability.

3.5.1 Integrator Core

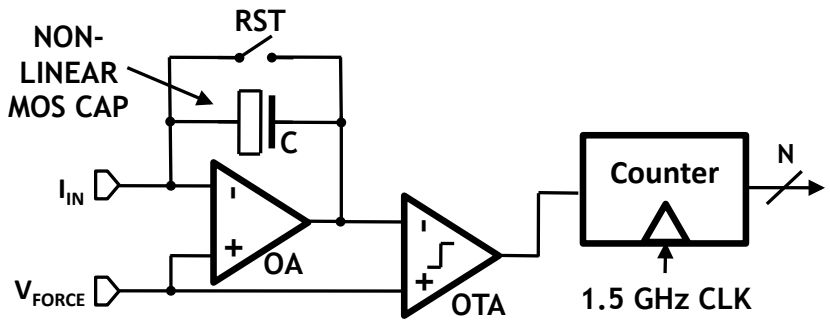


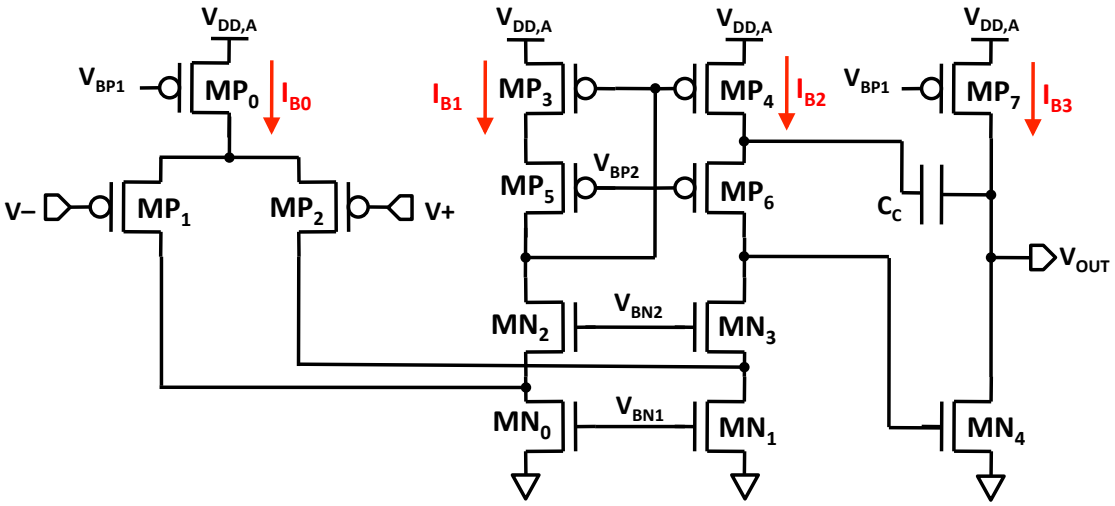
Figure 3.9: Dual-slope integrator core used in current-mode and voltage-mode ADCs.

The integrator core, shown in Fig. 3.9, consists of a two-stage op-amp, a non-linear MOS capacitor, a high-gain comparator, and a high-speed digital counter. The op-amp is designed to have a high voltage gain and a low-output impedance, as discussed below. High gain is needed to achieve high conversion accuracy while low output impedance is essential for driving low-impedance loads, such as wide DUTs biased in strong inversion. An 80 pF non-linear thick-oxide MOS capacitor is used as the main integrating element; this type of capacitor offers higher charge density as compared to other capacitor options and is readily available even in purely digital design flows. Since dual-slope integration is based solely on the concept of charge conservation, the non-linearity of the capacitor does not affect the linearity of the ADCs, as long as the capacitor does not leak or absorb charge during the conversion cycle. The comparator is implemented as a scaled-down version of the folded-cascode op-amp input stage. An 18-bit high-speed counter used to time the charge and discharge cycles of the integrator is implemented using native 45-nm devices, which allow it to operate at frequencies of 1.5 GHz and above. Leveraging the intrinsic speed of the underlying 45-nm CMOS process is essential for achieving high integrator core performance as increasing the reference clock rate results in higher measurement precision for a given sampling rate, or equivalently, a higher sampling rate for a given measurement precision.

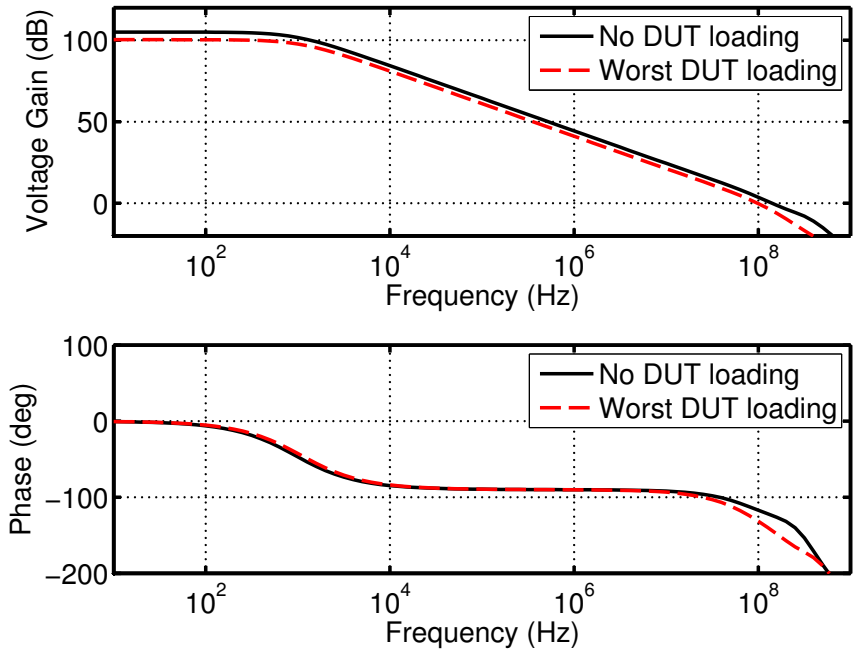
Op-Amp

A schematic for the two-stage op-amp is shown in Fig. 3.10(a). The input stage is a folded-cascode stage, which has a simulated voltage gain of 90 dB. PMOS input devices are used in order to enable an input voltage range from 0 V up to 1.7 V when operating on a 2.0 V analog supply. Additionally, PMOS transistors are expected to exhibit lower $1/f$ noise [76]. The cascode devices are biased using a low-voltage cascode biasing scheme [76], ensuring maximum swing at the outputs restricted to $\pm 2V_{ov}$ from the power rails; V_{ov} is the overdrive voltage of the cascode transistors given by $V_{ov} = V_{GS} - V_{th}$ and nominally set to 150 mV. Since most of the voltage gain of the folded-casode stage comes from the high output impedance, this stage alone cannot be used to drive a real load, as any resistance seen in the output would appear in parallel with the output resistance of the cascode devices and diminish the gain. In order to circumvent this issue, a common-source (CS) output stage is used as a low-impedance output buffer. The CS stage adds an additional 20 dB of intrinsic gain, and at the same time reduces the output impedance of the amplifier to 580 Ω. While the gain of the CS output stage is also diminished by resistive output loading, the input stage remains unaffected and provides more than sufficient gain as needed for the desired ADC resolution. Additionally, the CS output stage improves the output swing of the op-amp, bringing it to only $\pm V_{ov}$ from the power rails, where once again, $V_{ov} = 150$ mV.

The two-stage amplifier is expected to operate in unity-gain feedback mode, and as such, its unity-gain stability needs to be ensured. This is done with the help of a Miller compensation capacitor, $C_C = 32$ pF; like all other capacitors in this design, C_C is implemented using a MOS capacitor option. The Miller compensation capacitor acts to split the two dominant poles of the amplifier (one associated with each stage) and ensure adequate phase margin at unity gain. The compensation scheme used connects the Miller capacitor between the output of the second stage and the source of the cascode device in the first stage. This tends to push out the right-hand plane zero associated with Miller compensation without the need for an additional zero resistor, R_Z, connected in series with



(a)



(b)

Figure 3.10: (a) Schematic for two-stage op-amp; (b) simulated open-loop frequency response with a 100 pF load; design parameters and simulation results are summarized in Table 3.3.

C_C [76].

Fig. 3.10(b) shows simulation results for the open-loop frequency response of the amplifier. Two simulation scenarios are considered. In one case, the op-amp is loaded with a purely capacitive load and in the other case the amplifier is additionally loaded with the worst possible DUT loading conditions ($W/L = 1.0\ \mu m/0.04\ \mu m$ DUT biased in strong inversion saturation). In both cases the capacitive load is 100 pF, which includes 80 pF from the integrating capacitor, 10 pF due to additional loading of the amplifier from the input of the op-amp itself, and another 10 pF representing the input loading of the comparator, as well as any wiring parasitics along the measurement path. The simulation results are summarized in Table 3.3.

Table 3.3: Op-amp Design Table

Design Component	**Value**	**Bias Current**	**Value**	**Simulation Result**	**Value**
MP_0	6400 $\mu m/0.6\ \mu m$	I_{B0}	4.0 mA	Condition	No DUT
MP_1/MP_2	3200 $\mu m/0.6\ \mu m$	I_{B1}	2.0 mA	DC gain	105.9 dB
MP_3/MP_4	3200 $\mu m/0.6\ \mu m$	I_{B2}	2.0 mA	Phase Margin	56.3 deg
MP_5/MP_6	3200 $\mu m/0.6\ \mu m$	I_{B3}	2.0 mA	GBP	165.9 MHz
MP_7	1920 $\mu m/0.44\ \mu m$			Condition	Wide DUT
MN_0/MN_1	1600 $\mu m/0.6\ \mu m$			DC gain	100.3 dB
MN_2/MN_3	800 $\mu m/0.6\ \mu m$			Phase Margin	49.3 deg
MN_4	960 $\mu m/0.44\ \mu m$			GBP	113.5 MHz
C_C	32 pF			$v_{n,RMS}$ 100 $Hz - 1\ MHz$	41.9 μV

Since this op-amp is to be used as a general purpose op-amp throughout the design, it needs to meet a number of design goals. Unity-gain stability is crucial, as the op-amp is used in unity-gain feedback in the integrator, as well as when used as a buffer. Such stability is guaranteed by a worst-case phase margin of 49.3 deg. At the same time, a gain-bandwidth product (GBP) of more than 100 MHz is required to allow the op-amp to settle with an accuracy of up to five time constants when processing a 20 MHz input signal, such as the one used during CBCM C-V characterization. Once again, this design goal is met and exceeded. A high dc gain is desired in order to enable high-resolution A-

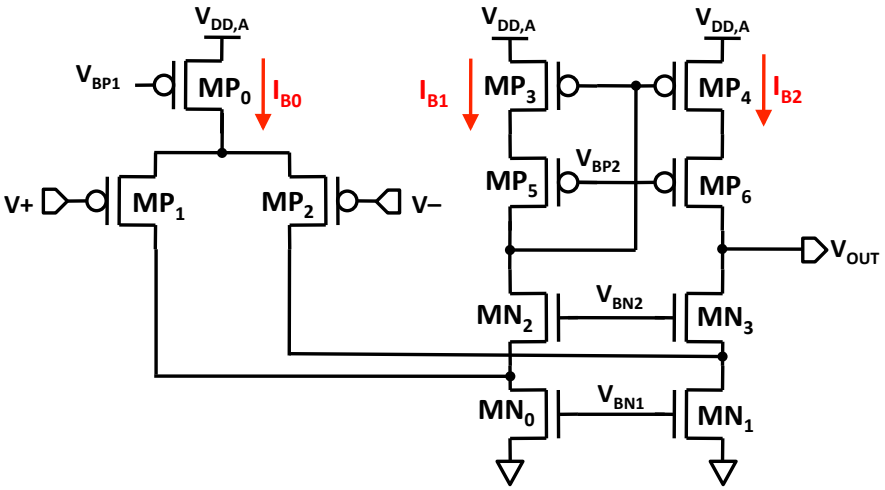
to-D conversion. The rule of thumb generally used is that 6 dB of gain is needed for each additional bit of resolution in order to guarantee that the systematic error due to the finite gain of theop-amp is less than the quantized signal-to-noise ratio (SNR). With a gain of over 100 dB, this op-amp can support up to 16 bits of resolution. Finally, the integrated input-referred voltage noise of the op-amp, including all white and $1/f$ noise sources, is simulated to be 41.9 μV RMS in the frequency band from 100 Hz to 1 MHz. This noise performance is comparable to the systematic error due to the finite gain of the op-amp and is adequate for the application at hand.

Comparator

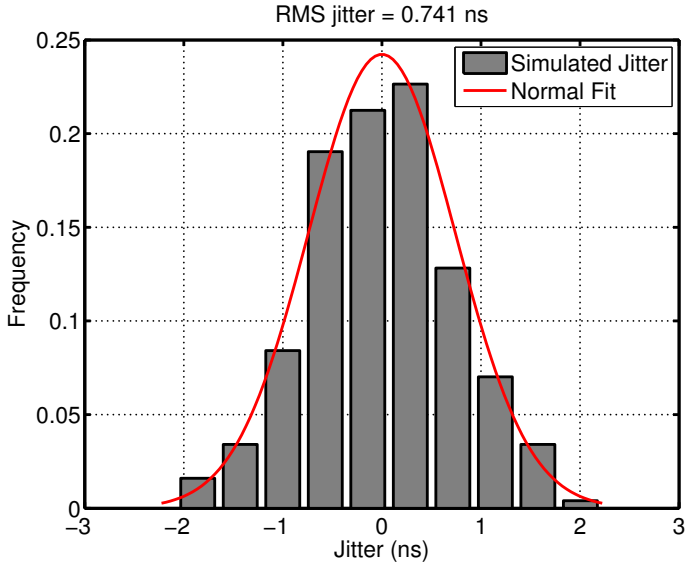
The comparator functionality is implemented using a variation of the folded-cascode input stage of the op-map (Fig. 3.11(a)). However, in order to conserve area and decrease the loading that the comparator presents to the integrator, the transistor sizes and corresponding bias currents are sized down by a factor of four.

The performance of the comparator can be characterized by simulating the comparator jitter when detecting a reference-voltage crossing by an ideal triangle wave input. The triangle wave used in simulation has a slew rate of 0.125 $V/\mu s$, resulting from a 80 pF integrator capacitor discharged by a 10 μA reference current; these conditions match the integrator nominal operating conditions. Using a time-domain noise analysis simulation, a histogram of the measured jitter can be extracted, as shown in Fig. 3.11(b). The simulated RMS jitter is 0.741 ns. This value is less than one clock cycle of a 1.5 GHz reference clock, and as such, the RMS jitter of the comparator is not expected to have a significant impact on the resolution of the ADCs. It should also be noted that the comparator exhibits a systematic delay of 116.7 ns in simulation, which is fairly significant compared to the clock cycle of the reference clock. However, this delay is constant for a given reference current and comparator reference voltage and can easily be calibrated out using a nulling measurement of an empty DUT cell.

Using a folded-cascode operational trans-impedance amplifier (OTA) as a compara-



(a)



(b)

Figure 3.11: (a) Schematic for folded-cascode comparator; (b) simulated jitter due to comparator noise when detecting the reference crossing of a triangle waveform with slew rate of 0.125 $V/\mu s$; design parameters and simulation results are summarized in Table 3.4.

Table 3.4: Comparator Design Table

Design Component	**Value**	**Bias Current**	**Value**	**Simulation Result**	**Value**
MP_0	1600 μm/0.6 μm	I_{B0}	1.0 mA	Delay	116.7 ns
MP_1/MP_2	800 μm/0.6 μm	I_{B1}	0.5 mA	RMS jitter	0.741 ns
MP_3/MP_4	800 μm/0.6 μm	I_{B2}	0.5 mA		
MP_5/MP_6	800 μm/0.6 μm				
MN_0/MN_1	400 μm/0.6 μm				
MN_2/MN_3	200 μm/0.6 μm				

tor has another practical advantage, which warrants a brief mention. With the help of a few extra switches, the OTA can be placed in a unity-gain configuration and used to buffer the internal integrator node to an output test pin. Observing the internal integrator node directly is not feasible due to the extra capacitive loading such a test path would present at the output of the integrator; the two-stage op-amp topology is conditionally stable and cannot accommodate much additional loading. However, if the comparator is used to buffer the signal, this problem is circumvented. Since the folded-cascode OTA acts as a single-stage amplifier, it can be load-stabilized by adding capacitive loading off-chip when probing the output. It is important that the comparator does not need to be stabilized on chip, as that would significantly impact its open-loop performance. Therefore, the folded-cascode OTA can serve a secondary role as a voltage buffer for debugging purposes without compromising its effectiveness as a comparator.

Up/Down Counter

An 18-bit loadable up/down asynchronous counter is implemented based on a design described in [73]. An asynchronous topology is chosen over a synchronous one as the former allows higher input clock rates. The ripple effect, which generally limits the maximum frequency of an asynchronous counter, is not an issue in the design at hand, as the counter can be allowed to ripple through before its output is sampled. Post-layout simulations of the counter show that it is operational with clock frequencies of up to 2.5 GHz. Ultimately,

the counter frequency is limited by the frequency of the clock signal that can reliably be delivered on chip, which is experimentally shown to be around 1.5 GHz. This aligns with the analysis of the comparator performance presented above, which demonstrates that the converter would not benefit from reference frequencies much higher than 1.5 GHz, as jitter from the comparator effectively limits the time resolution of the converter beyond that point.

3.5.2 Current-Mode ADC

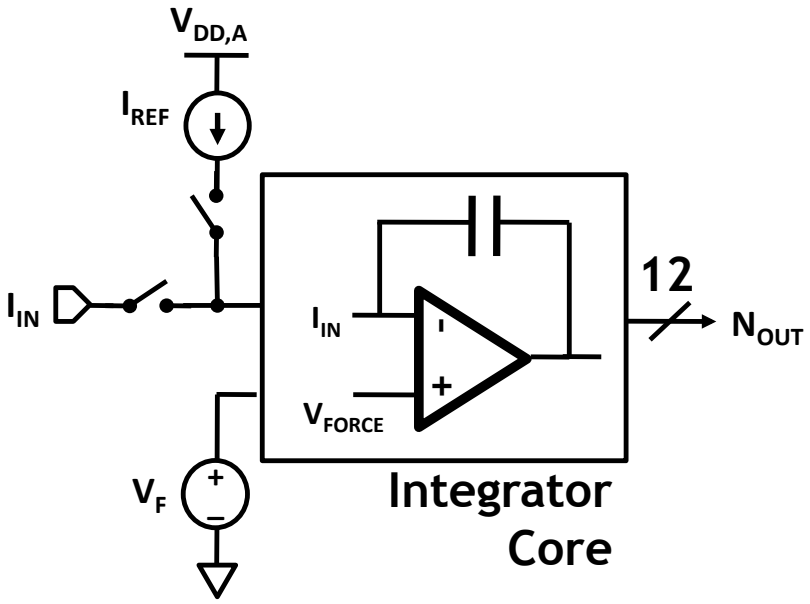


Figure 3.12: Current-mode ADC configuration using the dual-slope integrator core.

The integrator core is easily transformed into a current-mode ADC, as shown in Fig 3.12. The potential, V_F, supplied by one of the DAC channels, is forced at the input of the ADC though the negative feedback of the integrator, setting the voltage bias at the current input node. The desired pre-charge time is loaded into the counter, which counts down at the rate of the input clock. During this time, the integrator is sampling the input current, while simultaneously averaging out any high-frequency noise or interference that might be present. When the counter reaches zero, it starts counting up and a reference current with opposite polarity to that of the input is switched in, discharging the integrating capacitor

at a known rate. When the capacitor is fully discharged to its initial state, the comparator flips, signaling the end of the conversion cycle. The discharge time is recorded, and the output of the converter is given by

$$I_{IN} = N_{OUT}\frac{I_{REF}}{N_{REF}} \tag{3.1}$$

where I_{REF} is the reference current, N_{REF} is the integration time measured in reference clock cycles, and N_{OUT} is the measured digital output. From Eq. 3.1, the nominal I_{LSB} of the converter is given by $\frac{I_{REF}}{N_{REF}}$. The I_{LSB}, and consequently, the dynamic range of the converter can be adjusted by either increasing the sampling time or decreasing the reference current. While both methods act to slow down the conversion cycle, the former results in less sampled noise due to more averaging during the sampling stage of the conversion cycle, and is therefore preferred.

It is important to note that the value of the integration capacitor does not factor in Eq. 3.1. However, choosing a proper capacitor value is essential, as it affects both the sampling rate of the converter, as well as the noise sampled by the integrator. The integrator sampling function can be expressed in the frequency domain as [77]

$$H(f) = \frac{\sin \pi T f}{\pi T f}, \tag{3.2}$$

where f is the frequency variable, and T is the sampling period, given by $T = \frac{N_{REF}}{f_{CLK}}$. One way to define the sampled noise bandwidth due to $H(f)$ is to consider the first null frequency of $|H(f)|$ given by

$$f_0 = \frac{1}{T}. \tag{3.3}$$

A larger integrating capacitor enables more averaging during sampling and consequently, a proportionally narrower effective noise bandwidth and less integrated noise. At the same time, a larger capacitor also results in a longer conversion cycle and a slower sampling rate. Therefore, the integrating capacitor value presents a trade-off between sampling rate and resolution. With an integrating capacitor of 80 pF, a sampling period $T = 0.683$ μs can be achieved while still accommodating input currents as large as 40 μA. According to Eq. 3.3,

the resulting null frequency is $f_0 = 1.46\ MHz$. At the same time, $T = 0.683\ \mu s$ enables sample rates as high as 200 kHz with a reference current of 10 μA.

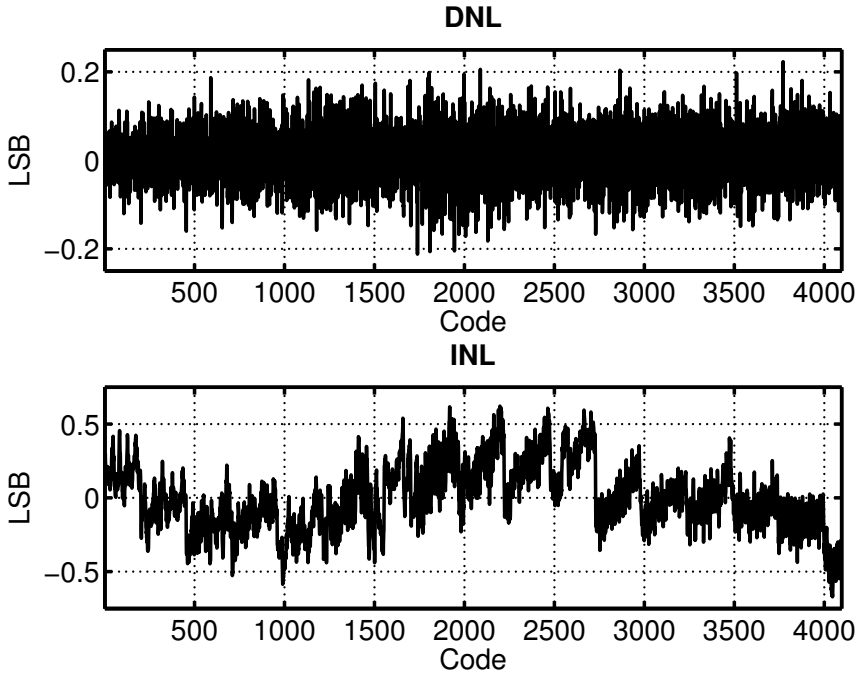


Figure 3.13: Measured DNL and INL for the current-mode ADC at 12-bit resolution with $I_{LSB} = 9.76\ nA$.

The measured DNL and INL of the current-mode ADC are shown in Fig. 3.13. The ADC operates at a 12-bit resolution with a pre-charge time of $N_{REF} = 1024$ reference clock cycles and a reference current $I_{REF} = 10\ \mu A$; this configuration yields $I_{LSB} = 9.76$ nA. The worst absolute DNL is less than 0.25 LSB and the worst absolute INL is less than 0.6 LSB, indicating true 12-bit effective resolution. This setting is appropriate for I-V characterization of small- to medium-sized DUTs operating in the linear region. Larger DUTs require the dynamic range to be adjusted by setting $N_{REF} = 512$ for $I_{LSB} = 19.52$ nA, and a maximum current range of 79.95 μA. In order to achieve the accuracy need for C-V characterization, the resolution of the ADC is boosted through a number of oversampling and noise-reduction techniques, as described in Chapter 4.

3.5.3 Voltage-Mode ADC

The implementation of the voltage-mode ADC is shown in Fig. 3.14. A simple sample-and-hold amplifier (SHA) is used to sample and buffer the input voltage. The SHA guarantees high input impedance, which is essential for proper four-point Kelvin measurements, and is based around a copy of the op-amp used in the integrator core in a unity-gain-feedback configuration. The 10 pF input capacitance of the op-amp (explicitly shown for clarity) is sufficient to reduce any errors due to charge injection from the voltage sampling switch to negligible levels. A 25 $k\Omega$ on-chip resistor is used to convert the sampled voltage to a dc current to be sampled by the integrator core. The value of the resistor is chosen such that the conversion time for a voltage measurement is comparable to that for a current measurement, and both measurements can be performed in parallel without either one presenting a sampling bottleneck.

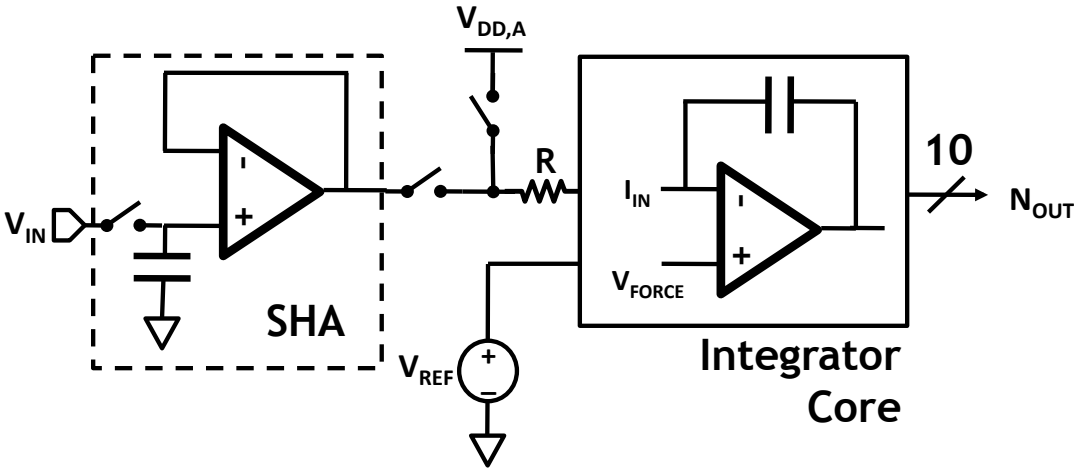


Figure 3.14: Voltage-mode ADC configuration using dual-slope integrator core; a SHA is used to provide high impute impedance.

In the configuration shown in Fig. 3.14, the ADC is used to sample voltages lower than V_{REF}, and the output of the converter is given by

$$V_{OUT} = V_{REF} - N_{OUT}\frac{V_{DD,A} - V_{REF}}{N_{REF}} \tag{3.4}$$

where V_{REF} is the reference voltage (defined with respect to a ground potential), $V_{DD,A}$ is the analog supply voltage, N_{REF} is the integration time measured in reference clock cycles,

and N_{OUT} is the measured digital output; the V_{LSB} is given by $\frac{V_{DD,A}-V_{REF}}{N_{REF}}$.

While the value of the resistor, R, does not factor into the conversion equation, the linearity of the converter is directly related to the linearity of the resistor, unlike the case with the integration capacitor. Therefore, care must be taken in choosing the proper resistor option and sizing in order to achieve the desired converter linearity. The resistor chosen in this case is a poly-silicon resistor, which offers better linearity as compared to other available options, such as an N-well resistor, for instance.

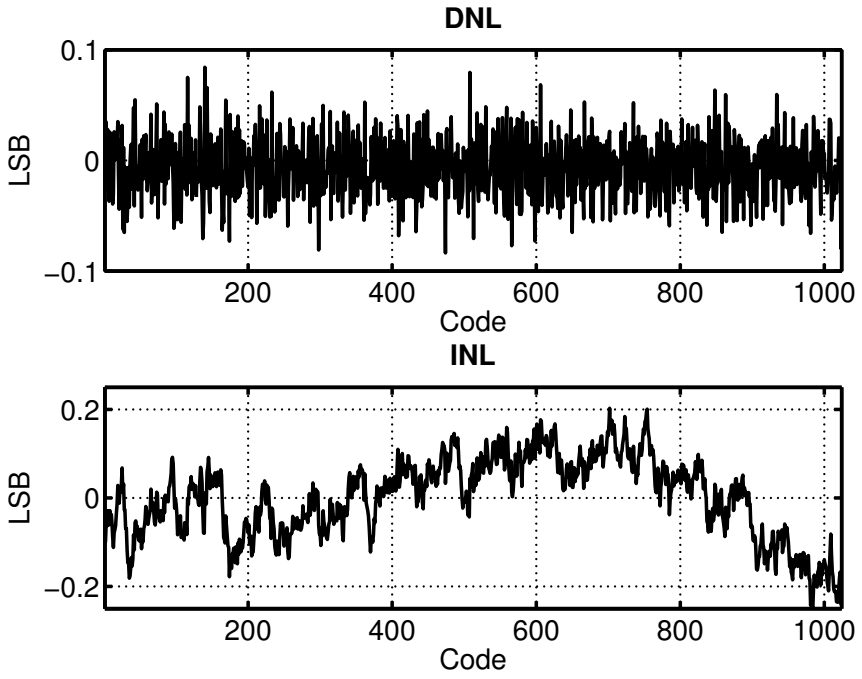


Figure 3.15: Measured DNL and INL for the voltage-mode ADC at 10-bit resolution with $V_{LSB} = 0.976\ mA$.

Fig. 3.15 shows the measured DNL and INL for the voltage-mode ADC. The characterization is preformed at a 10-bit resolution operating with a pre-charge time $N_{REF} = 512$ reference clock cycles, a reference voltage $V_{REF} = 1.5\ V$, and an analog supply voltage $V_{DD,A} = 2.0\ V$, resulting in $V_{LSB} = 0.976\ mV$. Once again, with a worst-case absolute DNL of less than 0.1 LSB and a worst-case absolute INL of less than 0.25 LSB, the voltage-mode converter is shown to operate at a true 10-bit effective resolution. At this resolution and dynamic range the ADC can cover an input range of 1.0 V, whereas the nominal V_{DD}

of the digital devices in this technology is 1.1 V. However, since the counter is much deeper than 10-bits, N_{OUT} can go above the nominal $2^{10} - 1$ reference clock cycles given by the 10-bit resolution to allow the input range to be extended to 1.1 V. Since the DNL and INL performance of the converter are consistent with what is required for 11-bit effective resolution, the partial sampling of an 11-th bit is justified.

3.5.4 Analog Buffers

Unity-gain analog buffers are used to bias the DUT terminals not connected to the current-mode ADC. The DAC buffers shown in Fig. 3.7 are not suitable for driving real loads with high accuracy, and are only used as high-input-impedance buffers between the internal resistor-string outputs and the unity-gain buffers in the MU. The MU buffers, which drive low-impedance paths going to the source/drain terminals of the DUT, are implemented using copies of the two-stage op-amp shown in Fig. 3.10. High-impedance paths biasing the gate and the body of the DUT, on the other hand, are driven by unity-gain buffers implemented using the scaled-down version of the folded-cascode op-amp input stage also used as a comparator (Fig. 3.11(a)). This single-stage OTA is stabilized using a 7.5 pF load capacitor connected between V_{OUT} and ground. The Bode plot in Fig. 3.16 shows the simulated open-loop ac response; the simulated open-loop dc gain is 78.3 dB with a phase-margin of 62.2 deg and a gain-bandwidth product of 140.8 MHz.

3.6 Test Chip

The on-chip characterization system is implemented in a 45-nm bulk CMOS process. Two identical copies of the characterization system are integrated on the same chip, as can be seen in the die micrograph shown in Fig. 3.17. One copy is used to characterize NMOS devices, while the other is used to characterize PMOS devices; both copies are operated in parallel for increased measurement throughput. While the two copies of the system are identical in their design, their analog references and internal controls are set up to accommodate opposite NMOS and PMOS current polarities, respectively. The NMOS and

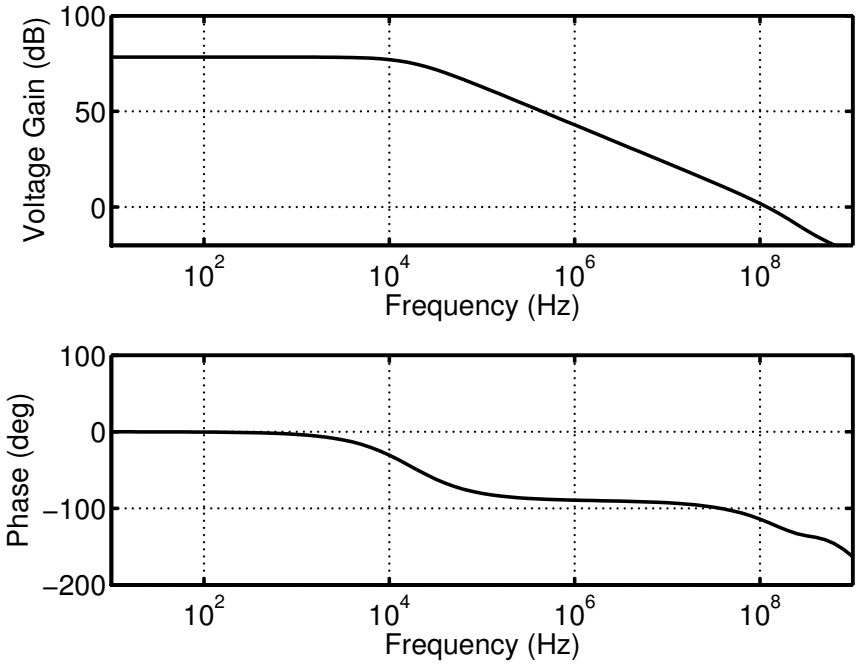


Figure 3.16: Simulated open-loop frequency response of the comparator in Fig. 3.11(a) also used as an OTA buffer; the OTA is stabilized using a 7.5 pF compensation capacitor connected between V_{OUT} and ground.

PMOS DUT arrays are interweaved parallel to one another, as described in Section 3.3.1. The majority of the chip area is occupied by the DUT arrays and associated switching matrix, with the DAC and MU circuitry integrated in the periphery of the test chip. The total chip area is 25 mm^2, and it contains the two copies of the characterization system, as well as over 4300 DUTs.

3.7 Measurement Setup

Fig. 3.18 shows the measurement setup used to run experiments using the on-chip C-V/I-V characterization system. An Opal Kelly^{TM} FPGA board [78] is used to interface between a measurement PC and the test chip. The FPGA board along with various ICs used to generate analog references and biases as needed by the test chip, are integrated on a custom-built printed circuit board (PCB) shown in Fig. 3.19; all analog signals are dc signals, as previously noted. A state machine implemented on the FPGA generates the required measurement controls and packages measurement data from the test chip to be

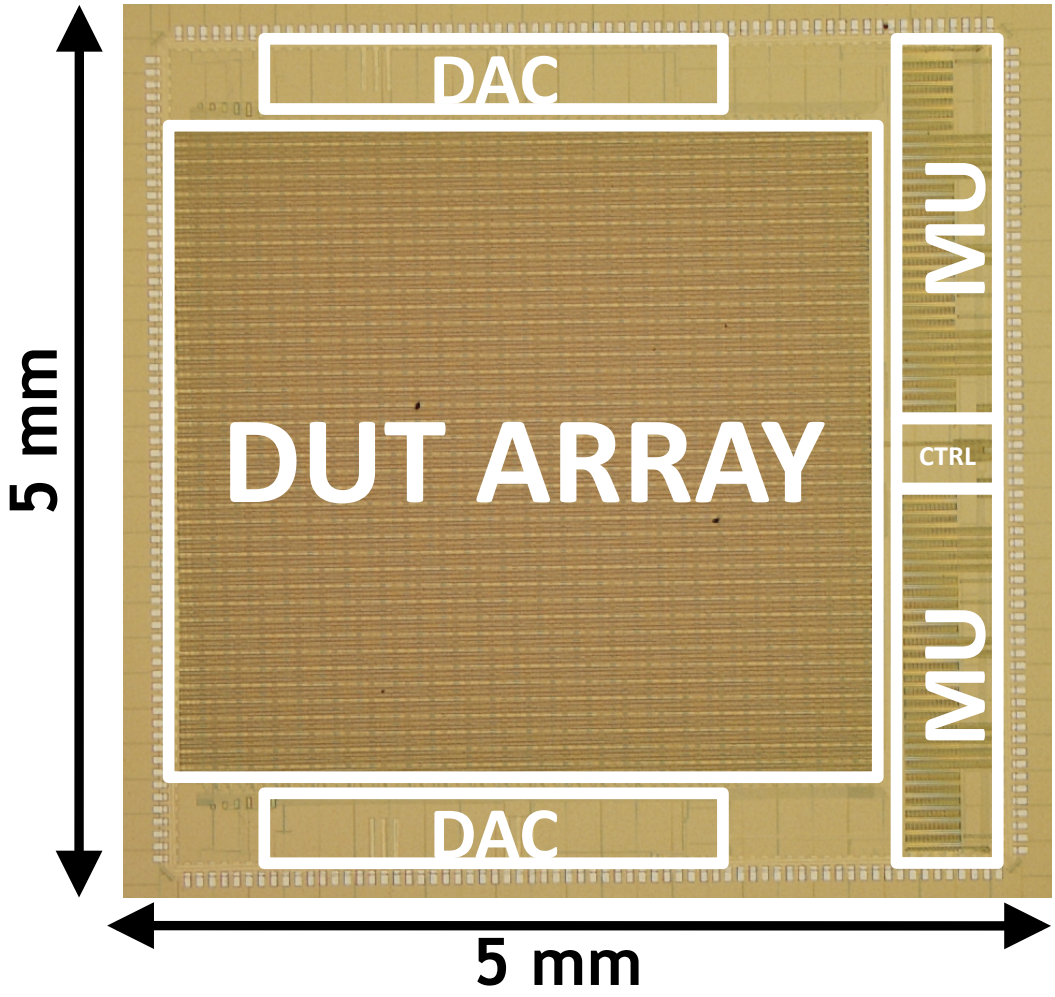


Figure 3.17: A micrograph of the system implemented in a 45-nm CMOS process; two identical copies for characterization of NMOS and PMOS devices, respectively, are integrated on the same chip and can be operated in parallel for increased measurement throughput.

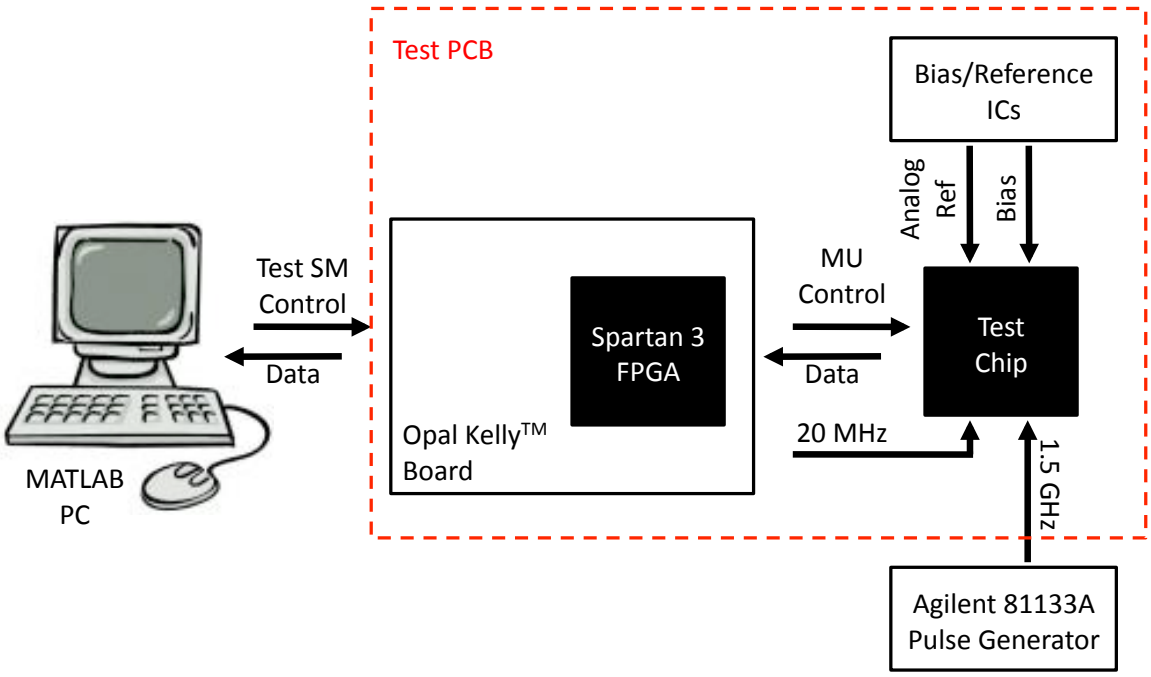


Figure 3.18: Measurement setup used to interface with the on-chip characterization system.

sent to the PC over a USB 2.0 interface. MATLAB is used to collect and analyze the measurement results. A PLL integrated on the Opal Kelly™ board is used to generate the 20 MHz reference clock needed for C-V characterization. An Agilent pulse generator [79] is used to generate the 1.5 GHz ADC reference clock.

3.8 Conclusion

The design of the on-chip variability characterization system has been presented. The entire system is integrated on-chip allowing for fast and accurate device characterization using a digital I/O interface. Individual components are designed for maximum compatibility with a purely digital CMOS process and use of specialized analog passives or transistor options is avoided. Modular design makes the system easily portable to new technology nodes. The functionality of all major system components is verified in simulation and measurement based on a 45-nm bulk CMOS implementation. The four-channel biasing DAC is shown to have 8-bit resolution with $V_{LSB} = 4.3\ mV$, the current-mode ADC is shown to have 12-bit resolution with $I_{LSB} = 9.76\ nA$, and the voltage-mode ADC is shown to have 10-bit

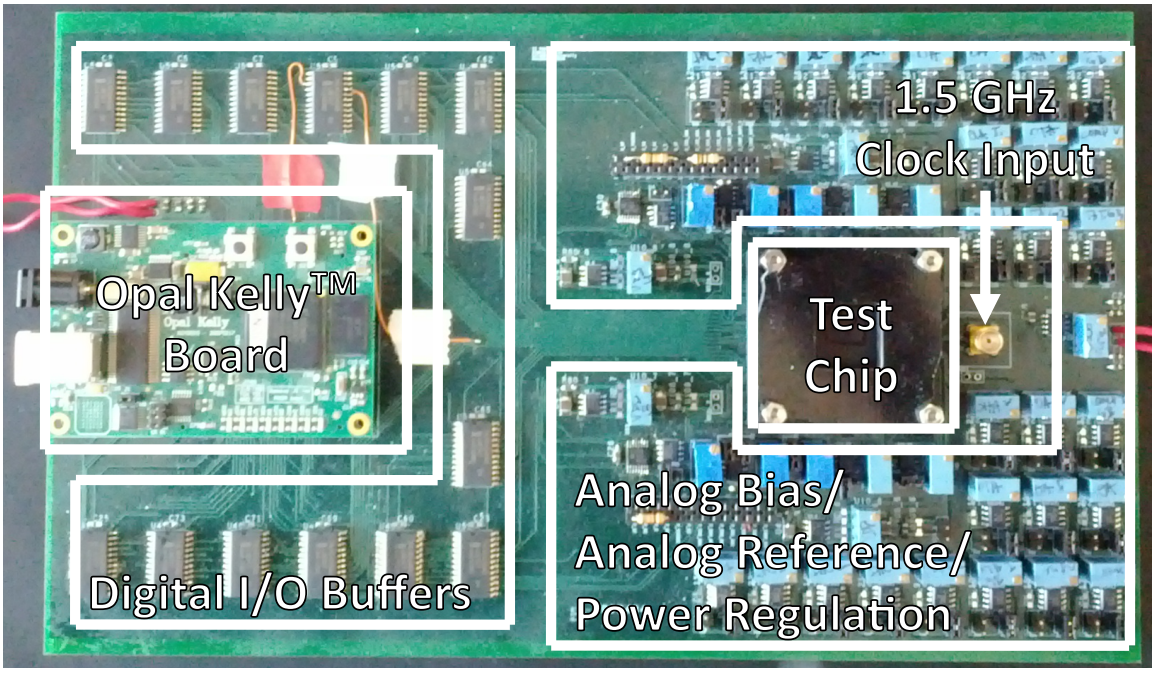


Figure 3.19: Custom PCB used for testing; different functional blocks are annotated.

resolution with $V_{LSB} = 0.976\ mV$.

Chapter 4

Combined C-V/I-V Characterization

4.1 Introduction

Chapter 4 describes the combined C-V/I-V variability characterization methodology. C-V and I-V measurement techniques for accurate on-chip quasi-static device characterization using the on-chip characterization system of Chapter 3 are introduced. Raw C-V/I-V measurement data are presented and different electrical parameter extraction techniques are discussed. C-V data is used to extract the effective MOS transistor channel length, L_{eff}, as needed for the analysis of random device variability. Random and systematic device variability is analyzed, with an emphasis on leveraging the combined C-V/I-V characterization approach to uncover the underlying physical sources of variability reflected in the observed variability of the electrical parameters.

4.2 Measurement Techniques

In order to enable accurate on-chip quasi-static device characterization, I-V and C-V measurement techniques which exploit the circuitry of Chapter 3 are developed. Issues related to non-negligible parasitic resistances through the on-chip switching matrix need to be ad-

dressed when performing I-V characterization. For the purpose of C-V characterization, a charge-based capacitance measurement (CBCM) technique, which allows capacitance measurements to be reduced down to dc current measurements, is introduced. This CBCM technique addresses issues related to gate leakage through the DUT, parasitic leakages and capacitances from the on-chip switching matrix, and measurement noise to achieve atto-Farad resolution measurements as needed for C-V variability characterization of circuit-representative devices.

4.2.1 I-V Measurements

Kelvin Sensing

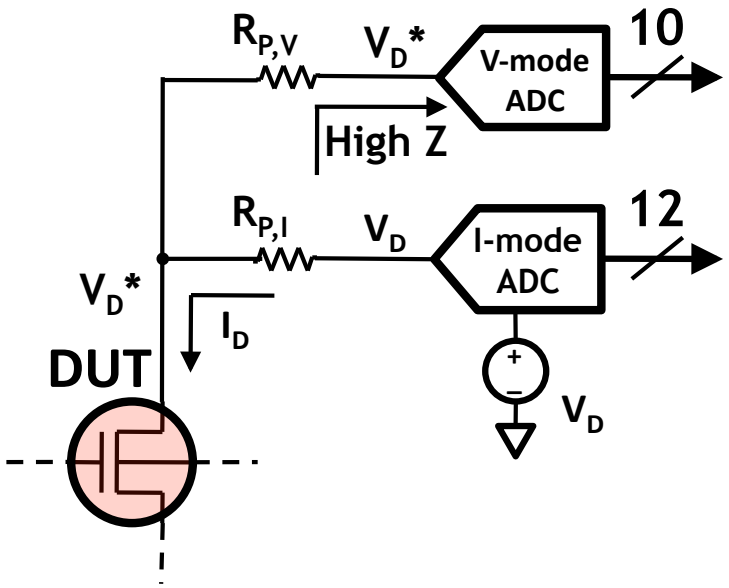


Figure 4.1: Four-point Kelvin measurements to negate parasitic IR drops through the on-chip switching matrix.

A four-point Kelvin measurement approach (Fig. 4.1) is used for accurate I-V characterization in the presence of non-negligible parasitic resistance through the on-chip switching matrix due to both the wiring parasitics and the resistance of the access switches. While the virtual ground supplied by the integrator at the input of the current-mode ADC accurately sets a bias for the current measurement to be performed, an IR drop across the parasitic resistance in the current path, $R_{P,I}$, causes the applied voltage V_D^* at the terminal of the

DUT to decrease as a function of the measured drain current, I_D, as given by

$$V_D^* = V_D - I_D R_{P,I}, \tag{4.1}$$

where V_D is the desired DUT terminal voltage bias. To eliminate the effect of this parasitic resistance, V_D^* is directly measured using a secondary sense path. The parasitic resistance in the sense path, $R_{P,V}$, does not affect the measured voltage due to the high input impedance of the voltage-mode ADC. Since the current at a DUT terminal is potentially a function of all four terminal bias voltages, the voltage at each of the DUT terminals is measured through individually designated voltage sense paths.

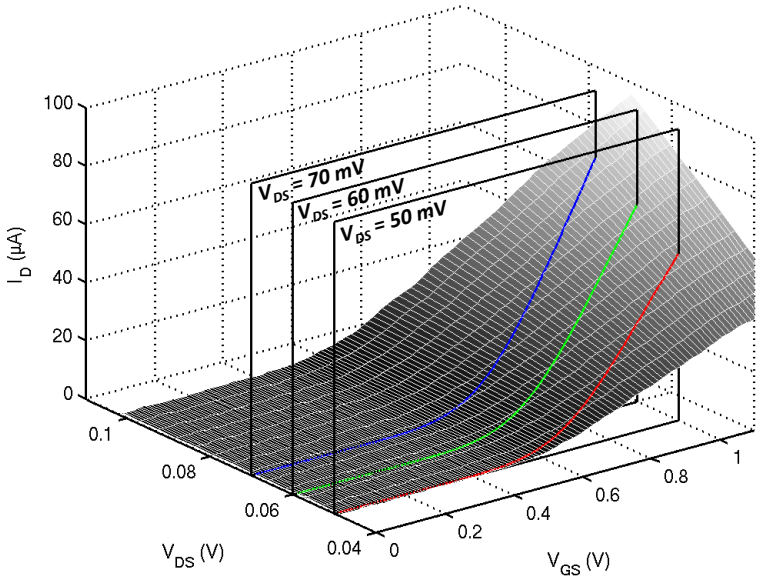


Figure 4.2: Measured I_D surface as a function of V_{GS} and V_{DS}; I_D values can be interpolated along planes of constant V_{DS} potential as shown.

While the DUT array allows voltage and current sense paths to be routed to any one of the four DUT terminals (with the exception of the body of NMOS DUTs, which is tied to the substrate potential of the chip), characterizing the drain current, I_D, as a function of the gate-to-source bias, V_{GS}, for a constant drain-to-source bias, V_{DS}, is the primary focus of this work. By sweeping the applied gate and drain voltages, a three-dimensional surface plot of I_D as a function of V_{GS} and V_{DS} can be measured, as shown in Fig. 4.2.

Since the voltage sweep steps given by the V_{LSB} of the DAC are as small as 4.3 mV, linear interpolation can be used to accurately extract the measured drain current across planes of constant V_{DS} bias. Although I_D is derived for a constant V_{DS}, V_{SB} does vary slightly over the bias range. If desired, this slight bias dependence can also be removed by introducing a sweep at the source terminal of the DUT.

Asynchronous Sampling

While traditionally analog-to-digital converters sample signals at a given characteristic sampling frequency, the dual-slope integrator ADCs described in Chapter 3 can be configured to vary the sampling frequency according to the strength of the measured signal. The period needed to complete a conversion cycle is a function of the pre-charge time, which is set by the desired dynamic range, and the discharge time, which is determined by the signal strength. When operated in a synchronous fashion, the ADC sampling frequency needs to be configured to accommodate the maximum discharge time corresponding to the maximum allowable input current. However, when the I-V measurements considered are dc measurements, maintaining a constant sampling frequency is not required. Instead, the sampling time can be adjusted according to signal strength, yielding significant characterization time savings. Such optimal signal sampling can be accomplished by asynchronously triggering the beginning of a new conversion cycle on the comparator output signaling the completion of the previous sampling cycle.

The resulting variation in sampling frequency during a linear-region I-V measurement of a $W/L = 1.0\ \mu m/0.04\ \mu m$ device across 256 V_{GS} and 16 V_{DS} points is shown in Fig. 4.3. As can be seen, due to the wide range of input signal strength, considerable characterization time savings can be realized. In this particular example, the signal-strength-optimized sampling frequency varies between 185.2 kHz and 84 kHz, resulting in an overall characterization time of 26.6 ms for a sweep of 16×256 points. If the signal is instead uniformly sampled at a frequency of 84 kHz, the overall characterization time for the same sweep would be 48.7 ms, which is 83% higher. As a result, asynchronous sampling

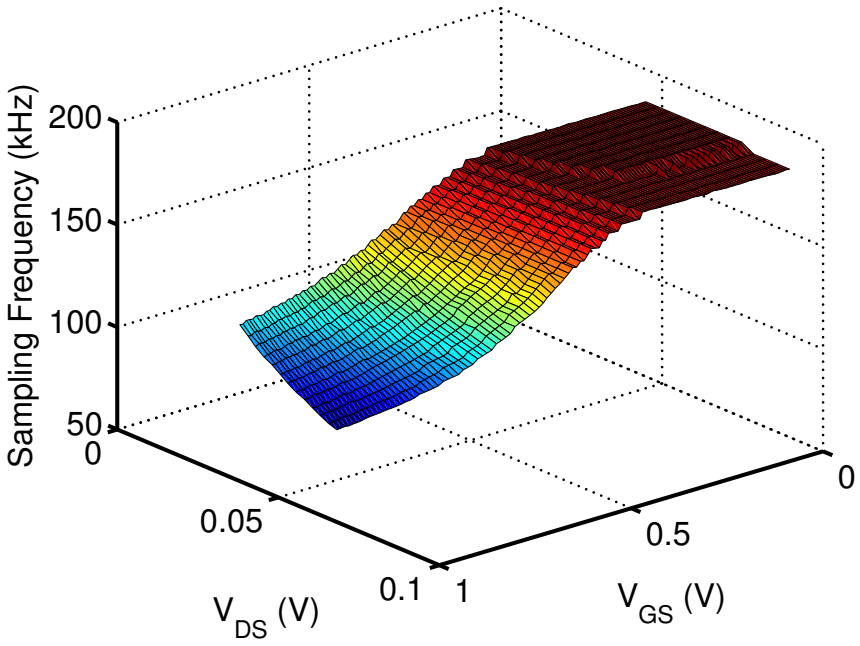


Figure 4.3: Signal-strength-optimized sampling frequency across a 16×256 point linear I-V sweep; using asynchronous sampling results in more than 80% improvement in overall characterization time.

has the potential to significantly reduce the required I-V characterization time, making the measurement of large DUT sample sets considerably more time-efficient.

4.2.2 C-V Measurements

In order to achieve accurate C-V characterization of circuit-representative devices, a novel on-chip charge-based capacitance measurement (CBCM) technique is developed. CBCM techniques reduce capacitance measurements to dc current measurements, allowing for existing I-V measurement infrastructure to be used to perform C-V characterization. The newly-developed CBCM technique addresses issues related to gate leakage through the DUT, as well as issues related to the characterization of devices integrated into a high-density on-chip DUT array, such as errors due to parasitic interconnect capacitance and array leakage. Different data post-processing techniques are used to boost the effective measurement resolution in order to achieve atto-Farad resolution C-V characterization, as needed to measure the variability of 45-nm devices with circuit-representative dimensions.

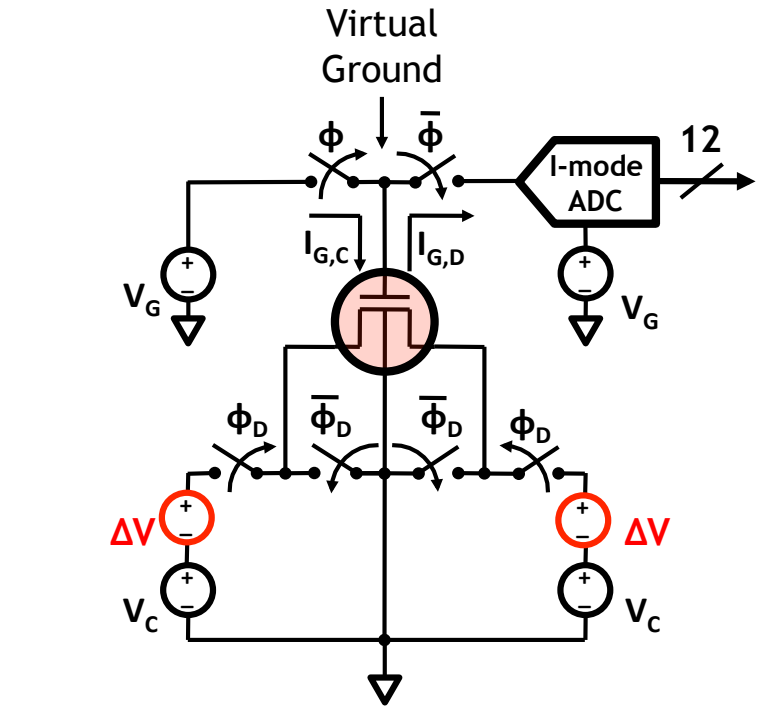
CBCM Measurement Technique

The leakage- and parasitic-insensitive on-chip CBCM technique is illustrated in Fig. 4.4 as used for characterizing the gate-to-channel MOS capacitance, C_{GC}. During the first phase of the measurement, the DUT is biased at the desired operating point, with its gate connected to a gate potential, V_G, and the drain and source connected to a channel potential, V_C. Once this bias point has been established and all of the DUT potentials are stabilized, the current-steering switch at the gate flips, making the on-chip current-mode ADC the new source for the gate bias, which remains at V_G. At this point, the current necessary to maintain the bias at the gate starts being measured. After a short on-chip-generated delay on the order of 100 *ps* (denoted as t_D in Fig. 4.4(b)), the drain and source of the DUT are shorted to ground, causing C_{GC} to be discharged. The change of charge needed to keep the potential of the gate at V_G is integrated by the current-mode ADC, along with any leakage current through the device and the associated switching matrix (not shown). Once this charge transfer has settled, the DUT gate is disconnected from the measurement circuitry, and the DUT channel is once again biased at V_C. This cycle is repeated multiple times, building up a measurable amount of charge in the current-mode ADC.

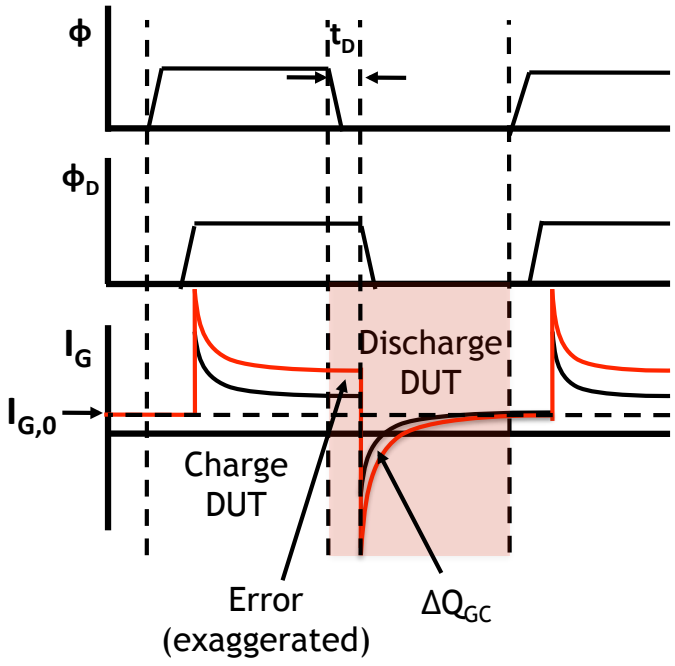
At the end of the conversion, the average measured charge, $Q_{G,M}(V_{GC})$, which is derived from the average discharge gate current, $I_{G,D}(V_{GC})$, can be expressed as

$$Q_{G,M}(V_{GC}) = \int_{T/2}^{T} I_{G,D}(t)\,\mathrm{d}t = Q_G(V_{GC}) + Q_{G,0}(V_G), \tag{4.2}$$

where T is the measurement clock period. $Q_{G,M}(V_{GC})$ consists of $Q_G(V_{GC})$, which represents the gate charge due to the discharging of C_{GC}, and the charge $Q_{G,0}(V_G)$, which represents the charge due to an integrated error current. The error current, denoted as $I_{G,0}$ is Fig. 4.4(b), is due to any leakages through the DUT and the switching matrix, as well as any charge injection from the switch at the input of the ADC, leakage currents due to mismatches between the two nominally identical gate potentials, and any charge shared from the parasitic capacitance at the gate node due to the same mismatch. The error current, $I_{G,0}$, is a function of the gate voltage, V_G, and is largely independent of the channel



(a)



(b)

Figure 4.4: (a) Illustration of the leakage-insensitive CBCM technique applied to measuring C_{GC} of an NMOS transistor, and (b) the accompanying voltage and current waveforms; the shaded portion of the I_G plot indicates the measured discharge current $I_{G,D}$.

potential, V_C, and consequently, the gate-to-channel potential, V_{GC}. This results from the fact that the error current is accumulated during the discharge phase of the cycle, where the bias conditions on the DUT are always constant – V_G at the gate, and ground at all other terminals. This is in contrast to the gate charge, $Q_G(V_{GC})$, which is a function of the bias voltage applied during the charging step of the cycle and, therefore, varies with V_{GC}.

The gate-to-channel capacitance, C_{GC}, is given by

$$C_{GC}(V_{GC}) = -\frac{\partial Q_G(V_{GC})}{\partial V_{GC}}. \tag{4.3}$$

Even though it is difficult to decouple $Q_G(V_{GC})$ from $Q_{G,0}(V_G)$, it is easy to see that since $Q_{G,0}(V_G)$ is a function of V_G only (and not V_{GC}), as long as V_G is kept constant throughout the measurement, differentiating the measured charge, $Q_{G,M}(V_{GC})$ with respect to V_{GC} yields

$$-\frac{\partial Q_{G,M}(V_{GC})}{\partial V_{GC}} = -\frac{\partial Q_G(V_{GC})}{\partial V_{GC}} = C_{GC}(V_{GC}). \tag{4.4}$$

Therefore, by keeping V_G constant while V_C is swept, the combination of the effective virtual ground at the gate and the differential nature of CBCM can be leveraged to cancel out the errors due to $I_{G,0}$. The only error signal left corrupting the measurement is the leakage current integrated for the duration t_D (see Fig. 4.4(b)), which is in fact a function of V_C. However, t_D, as mentioned above, is a delay of only about 100 ps; the error accumulated in this time interval is the same as that in a standard CBCM measurement if the measurement frequency were 5 GHz and is negligible.

Fig. 4.5 demonstrates the concept of leakage cancellation described above by examining the effect of the CBCM clock frequency, f_{CLK}, on the measured $Q_{GC,M}$ as well as on the extracted C_{GC}. Example measurements for the case of a PMOS and an NMOS transistors with $W/L = 1.0\ \mu m/0.11\ \mu m$ are considered. As expected, linearly decreasing the period of the measurement clock, T_{CLK} (or equivalently, increasing the measurement frequency, f_{CLK}), proportionally shifts the measured PMOS $Q_{GC,M}$ up and the measured NMOS $Q_{GC,M}$ down; the different directions of the shift are explained by the different polarities of the leakage current, $I_{G,0}$. More importantly, however, these shifts are constant

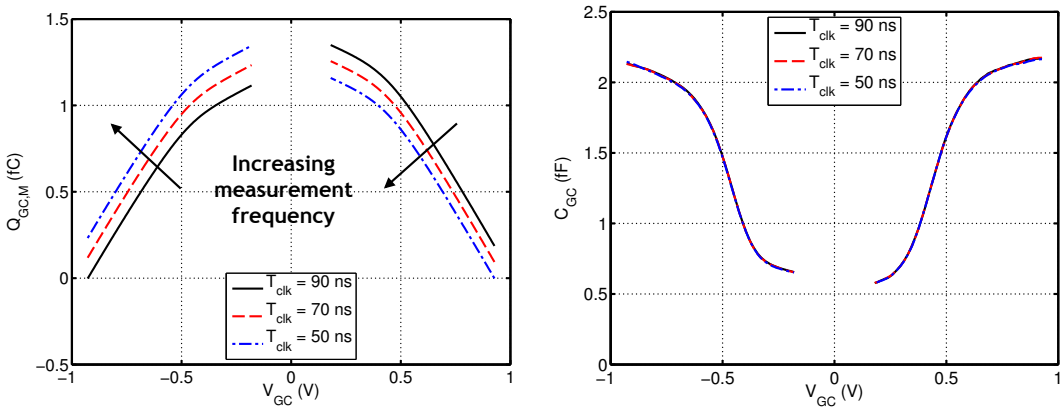


Figure 4.5: (a) Q_{GC} and (b) C_{GC} measured at different CBCM clock frequencies for a PMOS (negative V_{GC}) and an NMOS (positive V_{GC}) device; while Q_{GC} changes with frequency, the extracted C_{GC} remains constant.

across the measurement bias range, and result in no appreciable change in the extracted C_{GC}, as seen in Fig. 4.5(b).

While the measurement insensitivity to f_{CLK} is a desired effect, which demonstrates accurate cancellation of the effects of $I_{G,0}$, operating at the highest possible $f_{CLK} = 20$ MHz still has the benefit of improving overall measurement speed. Improved measurement speed decreases the time needed to characterize a device and also enables the differential $1/f$ noise cancellation technique described below. Additionally, a higher clock frequency increases the ratio of $Q_G(V_{GC})$ to $Q_{G,0}(V_G)$, thus reducing the dynamic range requirements for a given measurement precision. Performing capacitance measurements at such aggressive switching frequencies is generally not applicable to off-chip measurement techniques due to the large time constants associated with taking weak analog signals off-chip.

Measurement Noise Reduction Techniques

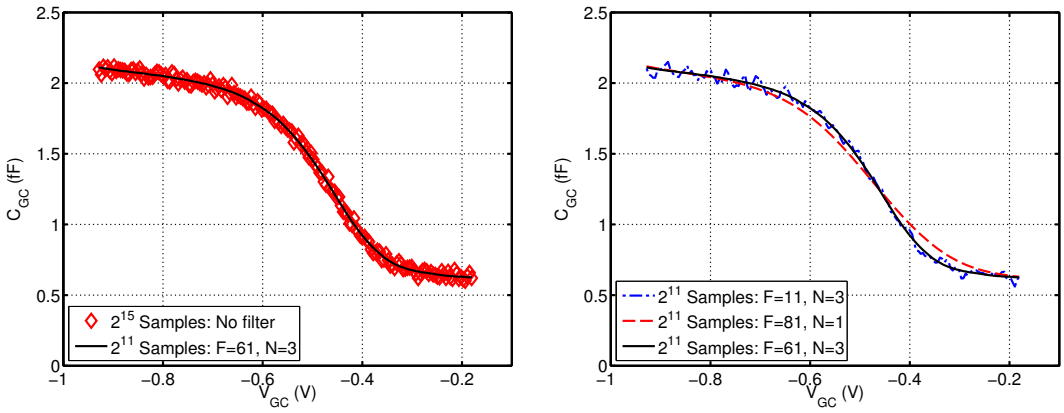
In order to achieve atto-Farad (aF) resolution, a combination of oversampling and filtering techniques is used. If a voltage step $\Delta V = 4.3$ mV is considered, the average change of charge, ΔQ, which needs to be measured in order to resolve a capacitance of 1 aF is 4.3×10^{-21} C. This is less than one elementary charge, and at a measurement frequency of 20 MHz, results in an average current of 8.6 fA.

The nominal I_{LSB} of the current-mode ADC is decreased to $I_{LSB} = 390.6\ pA$ by setting $N_{REF} = 1024$ and $I_{REF} = 400\ nA$. Since the current measured to derive $Q_G(V_{GC})$ is a dc current, oversampling can be leveraged to further reduce the effective I_{LSB} of the converter. In particular, if an oversampling ratio (OSR) of 2^{11} is used, I_{LSB} is reduced to 190.7 fA, while an OSR of 2^{15} gives an I_{LSB} of 11.92 fA. Unfortunately, even with an aggressive CBCM clock frequency of 20 MHz, a full 256-point C-V measurement at an OSR of 2^{15} has an overall run time of approximately 12 min, which is prohibitively large for high-throughput studies.

To further suppress measurement noise, a Savitzky-Golay filter [80], particularly suited for extracting smooth derivatives from measured data, can be used. The Savitzky-Golay filter works by performing a polynomial fit through a moving data window, and using the fitted polynomial parameters to estimate the derivative of the data at the mid-point of the window. The application of such a filter is tantamount to oversampling, where the knowledge that the data does not change rapidly within a given interval is used to better estimate a single-point value based on an ensemble of neighboring points. The size of the moving window, F, and the order of the fitted polynomial, N, are parameters of the filter, which must be chosen to achieve the desired noise reduction, while maintaining the high-frequency components of the C-V curve.

Fig. 4.6(a) demonstrates the effectiveness of the Savitzky-Golay filter for appropriately chosen filter parameters; measurements for a $W/L = 1.0\ \mu m/0.11\ \mu m$ PMOS device are used as an example. Even with an OSR of 2^{15}, the unfiltered data remains somewhat noisy. In comparison, using an OSR of 2^{11} and a filter with parameters set to $F = 61$ and $N = 3$ results in a low-noise measurement, which tracks the unfiltered data very well. At an OSR of 2^{11}, a 256-point C-V measurement takes only 45 sec to complete, making the acquisition of large statistical data sets feasible.

In order to explain the choice of filter parameters, Fig. 4.6(b) shows examples of processing the same measurement data using different filter settings. If $F = 81$ and $N = 1$ is used, too much of the high-frequency component of the data is lost, and the transition



(a) (b)

Figure 4.6: Application of the Savitzky-Golay digital filter – (a) effectiveness of filtering in comparison to increasing the oversampling ratio and (b) choosing optimal filter parameters.

between depletion and strong inversion is not as sharp as the transistion observed in the unfiltered data. On the other hand, if the filter settings are configured to $F = 11$ and $N = 3$, the smoothing of the filter is not high enough to sufficiently suppress the measurement noise. With $F = 61$ and $N = 3$, the filter manages to both adequately suppress the high-frequency noise, and a the same time retain the underlying behavior of the C-V curve.

Oversampling is a very effective technique for reducing the white-noise floor of the measurement, but it leaves the measurement sensitive to $1/f$ noise in the form of low-frequency drift. $1/f$ noise places a practical limit on the usefulness of oversampling as a noise reduction technique, since after a certain point, the increasing amount of accumulated drift negates the effect of increasing the oversampling ratio. Fortunately, in CBCM the measured quantity of interest is actually the difference between measurements at two adjacent bias points. If the bias conditions are swept at a high rate rate (greater than 10 kHz in this experiment), low-frequency drift does not materially affect the measured differences. As shown in Fig. 4.7, if oversampling is done at each individual bias point before moving on to the next bias point, the accumulated drift is substantial enough to distort the measured capacitance curve. However, if instead only one measurement is taken per bias point, and then the entire voltage sweep is performed at the same oversampling ratio, a consistently reproducible C-V curve is extracted.

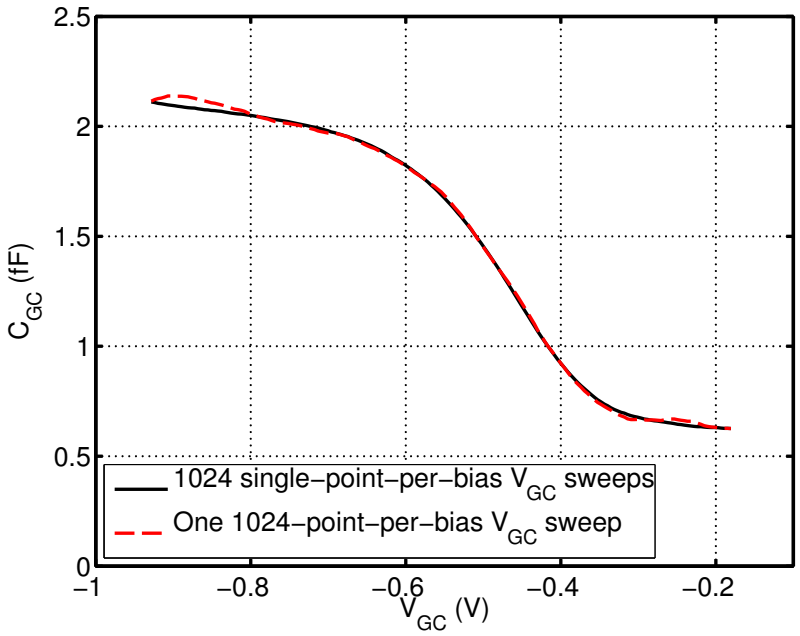


Figure 4.7: Avoiding the effects of $1/f$ noise on oversampling.

This approach to $1/f$ noise reduction also cancels out the effects of $1/f$ noise in the sensing path, including noise contributed from the CMOS switches used to access the device, as well as noise from the DUT itself. Significant random telegraph noise (RTN) is present in this 45-nm technology [81], as discussed in Chapter 5, and this low-frequency noise, caused by trapping and de-trapping of interface charge, is expected to have a direct impact on the measured gate charge, Q_G. However, if the charge trapping processes are much slower than the sweep time between bias points in the C-V measurement, their effect is canceled out along with all other sources of drift. This technique is superior to up-mixing techniques, such as the ones used by lock-in amplifiers, as those techniques generally only address $1/f$ noise contributed by the measurement circuitry itself. On-chip measurement integration significantly reduces the interconnect capacitance that needs to be overcome during each measurement step, allowing relatively fast sampling rates at very small signal levels, which in turn make the differential cancelation of $1/f$ noise possible.

C-V Measurement Precision

In order to estimate the achieved measurement precision, the C-V measurement reproducibility can be considered. When the system operates in C-V mode, I_{LSB} is reduced to 390.6 pA ($N_{REF} = 1024$ and $I_{REF} = 400\ nA$), in addition to using the oversampling and filtering techniques described above. Fig. 4.8 shows the standard deviation from the mean of a $W/L = 1.0\ \mu m/0.11/\ \mu m$ PMOS transistor C-V measurement across the entire bias range of interest. The standard deviation is taken across 16 identical measurements per bias point. The maximum observed standard deviation is 1.05 aF with mean standard deviation across all bias points of 0.87 aF. This result gives an effective noise floor for the C-V characterization performed in this work and demonstrates that aF measurement precision is achieved.

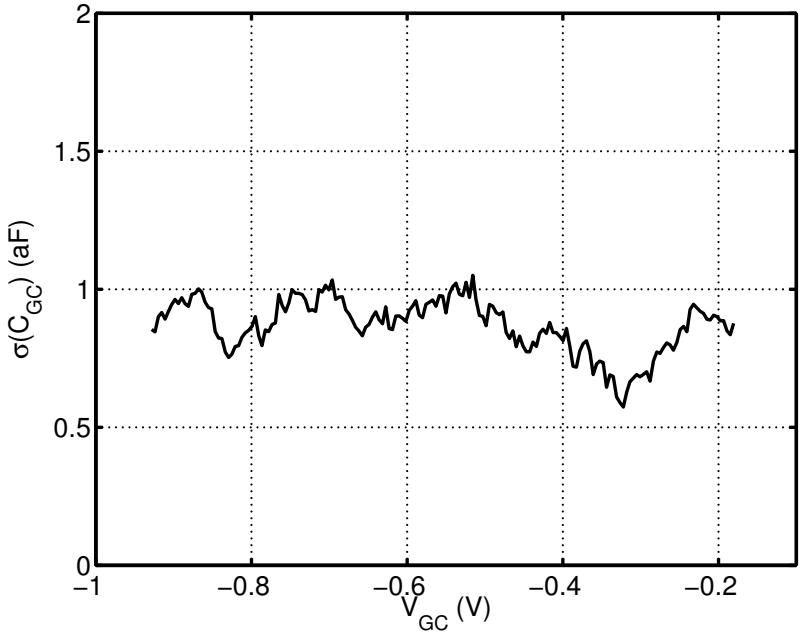
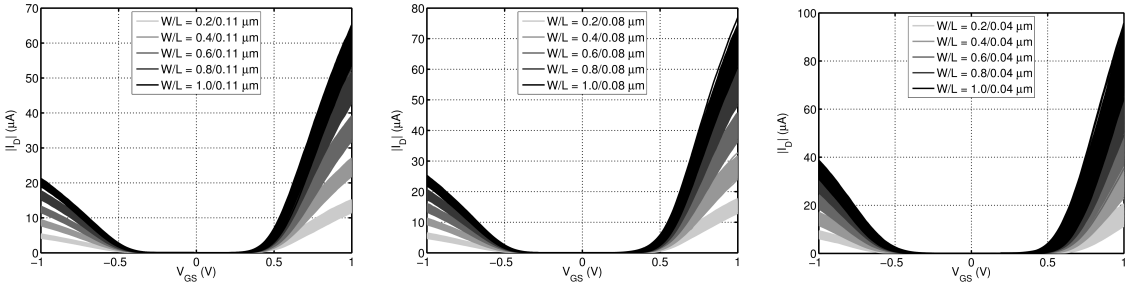


Figure 4.8: Standard deviation of 16 repeated measurements of C_{GC} across bias; the curve demonstrates the excellent reproducibility of the capacitance measurement.

4.3 Measurement Results

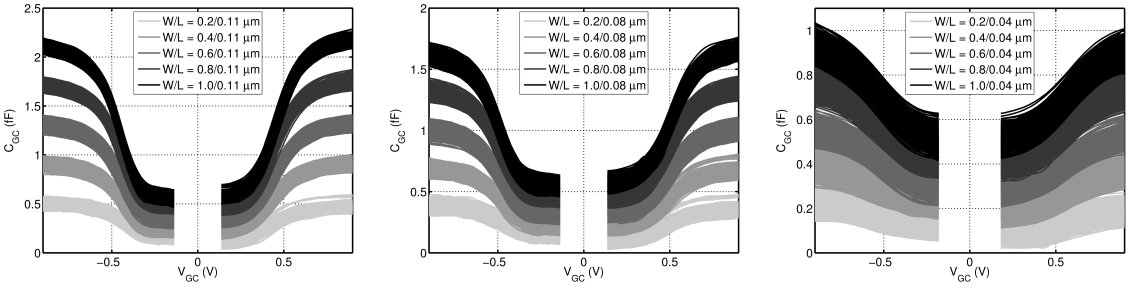
4.3.1 Raw C-V/I-V Measurements

Fig. 4.9 and Fig. 4.10 show constellation plots of I_D as a function of V_{GS} for $|V_{DS}| = 60$ mV, and C_{GC} as a function of V_{GC}, respectively. Fifteen different DUT types, spanning the parameter space of $W = [0.2 - 1.0]$ μm and $L = [0.04 - 0.11]$ μm, are measured across four chips, yielding 312 unique measurements per device type. The data contain measurements of both NMOS and PMOS devices, allowing comparisons to be drawn between the two device polarities. More importantly, I-V and C-V data are measured for the exact same device, allowing correlations between these measurements to be observed. This data form the basis for all MOS parameter extraction and variability analysis presented in this chapter.



(a) $L = 0.11$ μm (b) $L = 0.08$ μm (c) $L = 0.04$ μm

Figure 4.9: PMOS (negative V_{GS}) and NMOS (positive V_{GS}) I-V measurements across four different test chips.



(a) $L = 0.11$ μm (b) $L = 0.08$ μm (c) $L = 0.04$ μm

Figure 4.10: PMOS (negative V_{GC}) and NMOS (positive V_{GC}) C-V measurements corresponding to the I-V measurements in Fig. 4.9.

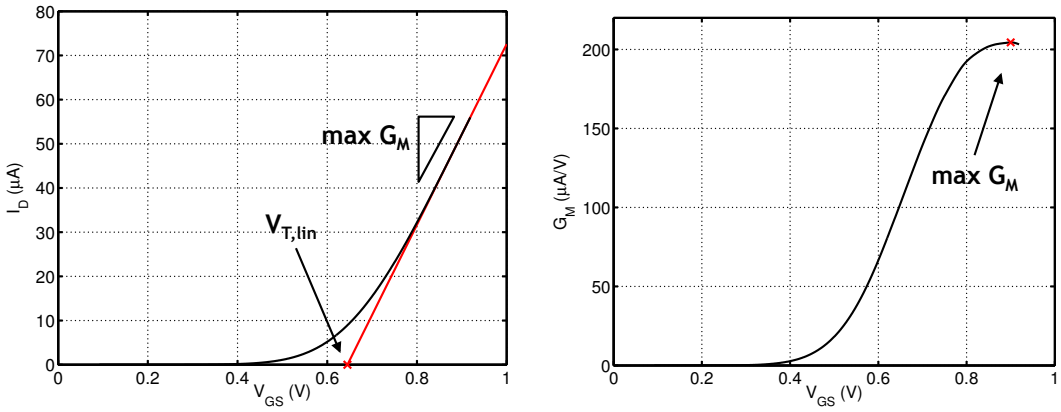
4.3.2 Parameter Extraction

I-V Parameters

I-V data are used to extract the linear threshold voltage, $V_{T,lin}$, using the extrapolation-in-the-linear-region (ELR) method [82] illustrated in Fig. 4.11. In this method, linear extrapolation is used to extrapolate I_D from the point of maximum G_M, where G_M is the large-signal device transconductance defined as

$$G_M(V_{GS}) \equiv \frac{\partial I_D}{\partial V_{GS}}. \tag{4.5}$$

The intersection point with the V_{GS} axis is interpreted as the linear threshold voltage, $V_{T,lin}$. A smooth G_M curve is extracted using Savitzky-Golay filtering. The variability in both $V_{T,lin}$ and G_M at a constant overdrive, $V_{GS} - V_{T,lin}$, can be studied.



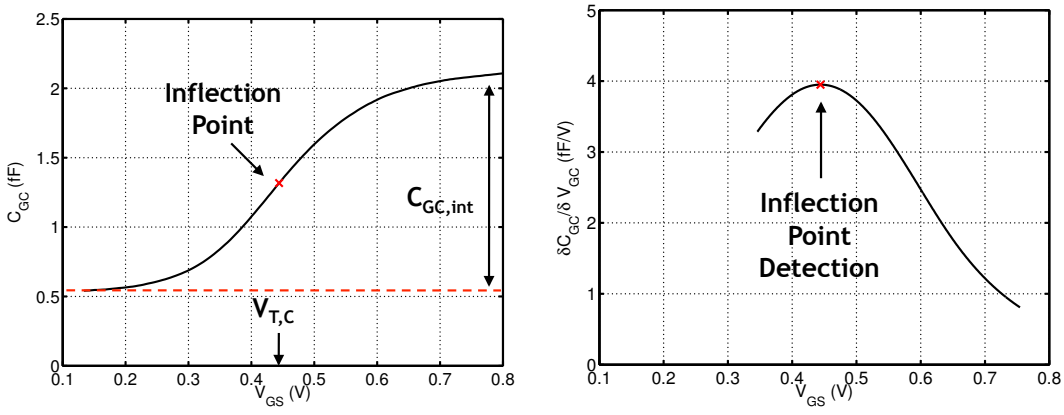
(a) I_D (b) G_M

Figure 4.11: $V_{T,lin}$ extraction from measured $I_D(V_{GS})$ data using the ELR method.

C-V Parameters

C-V data are used to extract variability in the intrinsic gate capacitance, which is expected to have a strong dependence on the effective gate area of the device. One way to extract the intrinsic gate capacitance, $C_{GC,int}$, is to look at the difference between the capacitance when the channel is inverted and the capacitance when the channel is depleted. While this definition has some problems associated with neglecting the inner fringe capacitance, as

discussed in Section 4.3.3 below, it can still be used to observe variation in the intrinsic dimensions of the device. Additionally, C-V data can be used to extract another variation of the threshold voltage, which will be denoted as $V_{T,C}$. $V_{T,C}$ is defined as the inflection point in the $C_{GC}(V_{GC})$ curve, which identifies the onset of the formation of an inversion layer. In order to detect the inflection point, a Savitzky-Golay filter is used to extract the derivative of C_{GC} with respect of V_{GC}. Fig. 4.12 illustrates the extraction of these C-V parameters from raw measurement data.



(a) C_{GC} (b) $\partial C_{GC}/\partial V_{GC}$

Figure 4.12: Extraction of $C_{GC,int}$ and $V_{T,C}$ from $C_{GC}(V_{GC})$ measurement data.

4.3.3 L_{eff} Extraction from C-V Data

When studying device variability, it is crucial to have a good understanding of the effective area of the device. In particular, the effective channel length, L_{eff}, is of great interest since it can vary significantly from the drawn length, L, due to overlap between the gate and the source/drain regions. Numerous methodologies for the extraction of L_{eff} from I-V measurement data have been proposed [83,84], but their accuracy is limited by the presence of an unknown and difficult to extract parasitic source-drain resistance, R_{SD}, and changes in effective mobility as a function of the channel length. More recently, methodologies for extracting L_{eff} from C-V data have been proposed [85, 86]. These methodologies, while also presenting some challenges, offer a more robust means of extracting the physical L_{eff}.

The extraction of L_{eff} from C-V data is based on the intrinsic channel capacitance,

$C_{GC,int}$, which forms between the gate and the inversion layer in strong inversion. $C_{GC,int}$ is given by

$$C_{GC,int} = C'_{ox} W L_{eff} = C'_{ox} W (L - \Delta L) \tag{4.6}$$

where C'_{ox} represents the oxide capacitance per unit area, W and L represent the drawn transistor dimensions, and ΔL represents the total gate overlap with both the source and the drain of the device. Therefore, if $C_{GC,int}$ can be measured across a number of different drawn lengths and widths, ΔL, as well as C'_{ox}, can be extracted.

A split C-V technique [87], such as the one used in this work, where the gate-to-channel capacitance is measured independently of other gate-referred capacitances (specifically the gate-to-body capacitance), is well-suited for the extraction of $C_{GC,int}$. In particular, if the measured capacitance at high gate bias, when the device is in strong inversion and the inversion channel is fully formed, is subtracted from the measured capacitance at low bias, when the device is in depletion and the inversion channel is fully suppressed, the resulting difference yields the intrinsic gate capacitance due to the formation of the inversion layer, while allowing any extrinsic capacitance due to fringing and gate overlap, as well as any stray capacitance from coupling between the measurement leads, to be cancelled out.

While such an interpretation is generally valid, it neglects the effect of a portion of the fringe capacitance between the gate and the source/drain regions in the device known as the intrinsic fringe capacitance, C_{if}. When the channel is depleted, fringing electric field lines between the gate and inner walls of the source/drain junction regions give rise to C_{if}, as shown in Fig. 4.13(a). However, when the channel is inverted, the inversion layer screens this capacitance, leaving only the outer fringe, C_{of}, and the overlap capacitance, C_{ov}, portions of the extrinsic gate capacitance (Fig. 4.13(b)). Since the inner fringe capacitance is proportional to W, its effect cannot easily be decoupled from the effects of ΔL, and ignoring it leads to an overestimation of the gate overlap, as shown in [85]. In particular, the capacitance measured in depletion is given by

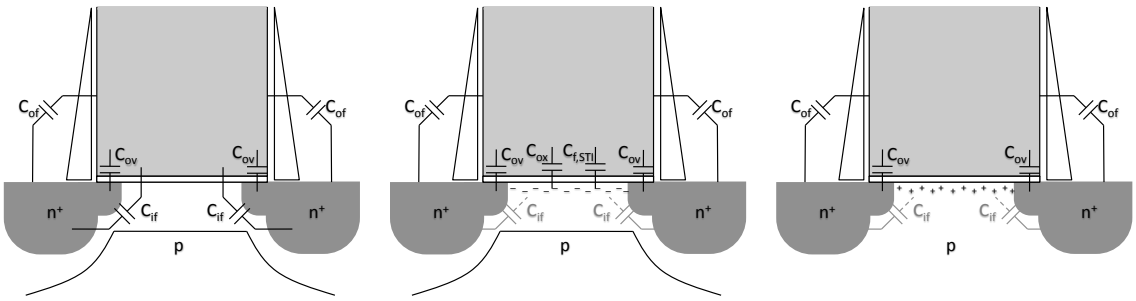
$$C_{GC,dep} = 2(C'_{ov} + C'_{of} + C'_{if})W + C_{par} \tag{4.7}$$

where C'_{ov}, C'_{of}, and C'_{if} are the overlap, outer fringe, and inner fringe capacitance per unit width, respectively, and C_{par} accounts for any parasitic coupling between the measurement leads.

In contrast, the measured gate-to-channel capacitance when the channel is fully inverted is given by

$$\begin{aligned} C_{GC,inv} =& C'_{ox}W(L-\Delta L) + 2(C'_{ov}+C'_{of})W + \\ & 2C'_{f,STI}(L-\Delta L) + C_{par} \end{aligned} \tag{4.8}$$

where $C'_{f,STI}$ is a fringing capacitance per unit length due to field lines between the gate and the inversion layer formed across the STI isolation along the length of the device. This capacitance is responsible for the threshold voltage roll-off in narrow-channel MOS transistors [88].



(a) (b) (c)

Figure 4.13: Components of the measured C_{GC} when the device is biased in (a) depletion, (b) inversion, and (c) accumulation; C_{if} is shielded in both inversion and accumulation.

Subtracting Eq. 4.7 from Eq. 4.8 results in

$$C^*_{GC} = C'_{ox}\left(W + \frac{2C'_{f,STI}}{C'_{ox}}\right)(L-\Delta L) - 2C'_{if}W \tag{4.9}$$

In order to be able to extract ΔL from Eq. 4.9 above, the effect of C_{if} must be cancelled. An observation is made that if the device is driven into accumulation (Fig. 4.13(c)), the layer of minority carriers formed close to the surface screens C_{if} [86]. This accumulation layer is electrically isolated from the source and drain and does not add any additional intrinsic gate-to-channel capacitance. As a result, the measured gate-to-channel capacitance

in accumulation is given by

$$C_{GC,acc} = 2(C'_{ov} + C'_{of})W + C_{par} \tag{4.10}$$

Therefore, subtracting Eq. 4.10 from Eq. 4.7 yields an estimate for the inner fringe capacitance as a function of device width.

Once C'_{if} has been extracted, its effects can be subtracted from Eq. 4.9 to get

$$C^*_{GC,int} = C'_{ox}\left(W + \frac{2C'_{f,STI}}{C'_{ox}}\right)(L - \Delta L) \tag{4.11}$$

which can be used to extract C'_{ox}, $C'_{f,STI}$, and ΔL.

While the NMOS DUTs share a body contact with the rest of the chip through the chip substrate and, therefore, cannot be biased in accumulation, the PMOS DUTs have isolated N-well potentials which do allow $C_{GC,acc}$ to be measured. A PMOS DUT can be biased deep in accumulation by setting the gate bias to a high potential and the body bias to a low potential, and establishing a gate-to-body potential $V_{GB} = 1.1\ V$. Fig. 4.14 shows an example measurement of C_{GC} for a $W/L = 1.0\ \mu m/0.11\ \mu m$ PMOS device transitioning from depletion to inversion, as well as biased deep in depletion and deep in accumulation. The difference between the measured capacitance in depletion and the measured capacitance in accumulation gives the inner fringe capacitance, C_{if}, as shown.

The inner fringe capacitance per unit W is assumed to be the same in both NMOS and PMOS devices, and the extracted average value of the inner fringe capacitance for each PMOS device size is added to the average measured C^*_{GC} to cancel the effect of C_{if} and bring the measurement to the form shown in Eq. 4.11. This data is then used to extract C'_{ox},$C'_{f,STI}$, and ΔL. The values derived for PMOS and NMOS devices are shown in Table 4.1.

Fig. 4.15(a) shows a plot of $C^*_{GC,int}$ as a function of the drawn dimensions, W and L, and 4.15(b) shows a plot of $C^*_{GC,int}$ as a function of $\left(W + 2C'_{f,STI}/C'_{ox}\right)(L - \Delta L)$, with solid lines representing a fit to Eq. 4.11 based on the parameters shown in Table 4.1. Drawing a comparison between the two, it is clear that the functional relationship described in Eq. 4.11 explains the measured data better. Due to the effects of poly depletion, C'_{ox} for PMOS

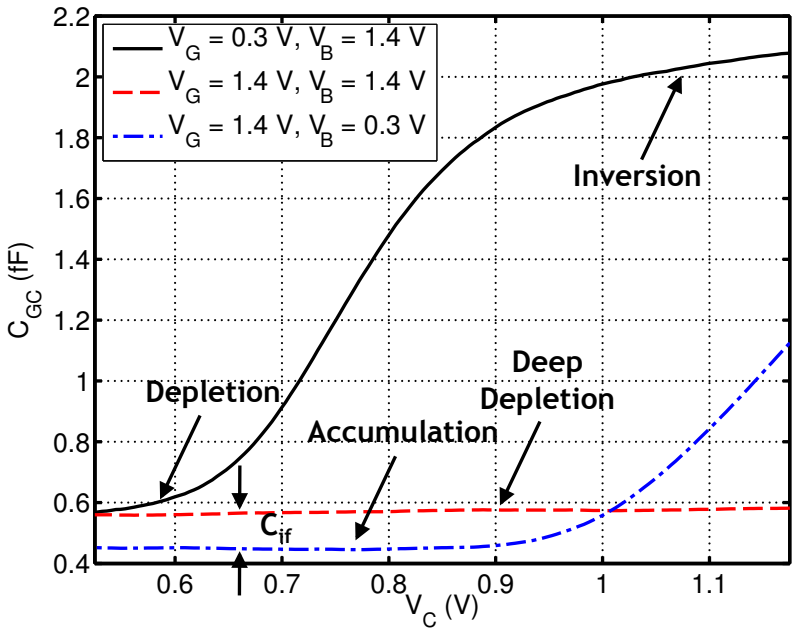


Figure 4.14: Measurement and extraction of C_{if} for a PMOS device based on C-V data measured in accumulation.

Table 4.1: Extracted Values of ΔL, $C'_{f,STI}$, and C'_{ox}

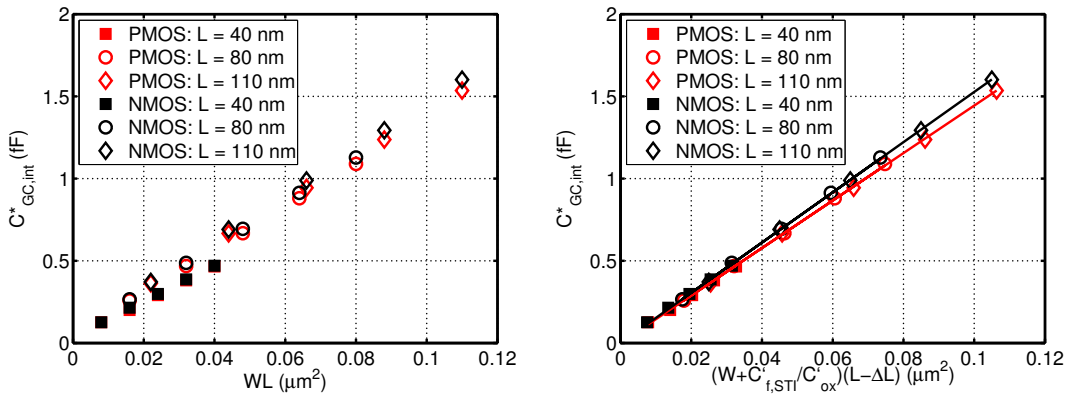
Parameter	**NMOS**	**PMOS**
ΔL (μm)	0.010	0.009
$C'_{f,STI}$ $(fF/\mu m)$	0.38	0.38
C'_{ox} $(fF/\mu m^2)$	15.3	14.4

devices is slightly smaller than C'_{ox} for NMOS devices. The PMOS devices show slightly less overlap as compared to the NMOS devices, resulting in a comparatively larger L_{eff}. The STI fringe capacitance per unit area, $C'_{f,STI}$, extracted from both sets of measurements is the same, as would be expected, validating the underlying assumption that C_{if} is the same for PMOS and NMOS transistors.

4.3.4 Analysis of Random Variability

All random parameter fluctuations are analyzed by considering the difference in the parameter of interest, ΔP, extracted from two matched DUTs in the DUT array. This difference is given by

$$\Delta P = P_1 - P_2 \tag{4.12}$$



(a) (b)

Figure 4.15: $C^*_{GC,int}$ (a) as a function of the drawn dimensions, W and L, and (b) as a function of $\left(W + 2C'_{f,STI}/C'_{ox}\right)(L - \Delta L)$, where the solid lines represent the linear fit used to extract C'_{ox},$C'_{f,STI}$, and ΔL.

where P_1 and P_2 are the extracted parameter values for the two matched DUTs. Such a pseudo-differential measurement approach cancels out the effects of any systematic parameter gradients along the columns of the DUT array. Within-die random variation is assumed to be the same across all measured chips. Gathering measurement data from four chips results in 156 differential parameter measurements, which is the sample size used in all random variability analysis presented in this chapter.

While the versatility of the combined C-V/I-V measurement approach makes it possible to analyze variability in practically any parameter governing the quasi-static transistor behavior, the analysis presented in this work focuses on variability in the threshold voltage, V_T, the large-signal transconductance, G_M, and the intrinsic gate-to-channel capacitance, $C_{GC,int}$.

V_T Variability

Fig. 4.16 shows a Pelgrom plot [39] of the variability in the linear threshold voltage, $V_{T,lin}$, extracted from I-V measurements as described in Section 4.3.2. The standard deviation of $\Delta V_{T,lin}$ is modeled as

$$\sigma_{\Delta V_{T,lin}} = \frac{A_{V_T}}{\sqrt{W L_{eff}}}. \tag{4.13}$$

In order to underscore the importance of using the effective rather than the drawn area of the device, the Pelgrom fit is performed using the effective device area corresponding to the effective channel length, L_{eff}, but it is plotted versus the drawn area, WL. As can be seen, what might initially appear as an increase in the Pelgrom slope, A_{V_T}, for the case of minimum-length devices, is in fact accounted for by the decrease in effective gate area due to ΔL, extracted from C-V measurements, as discussed in Section 4.3.3. This is just one example of using information from both C-V and I-V measurements to gain a better understanding of the underlying causes of device variability. As can be expected, parameter variability is a function of the effective and not the drawn area of the device, and all variability models should be based on L_{eff} rather than the drawn length, L.

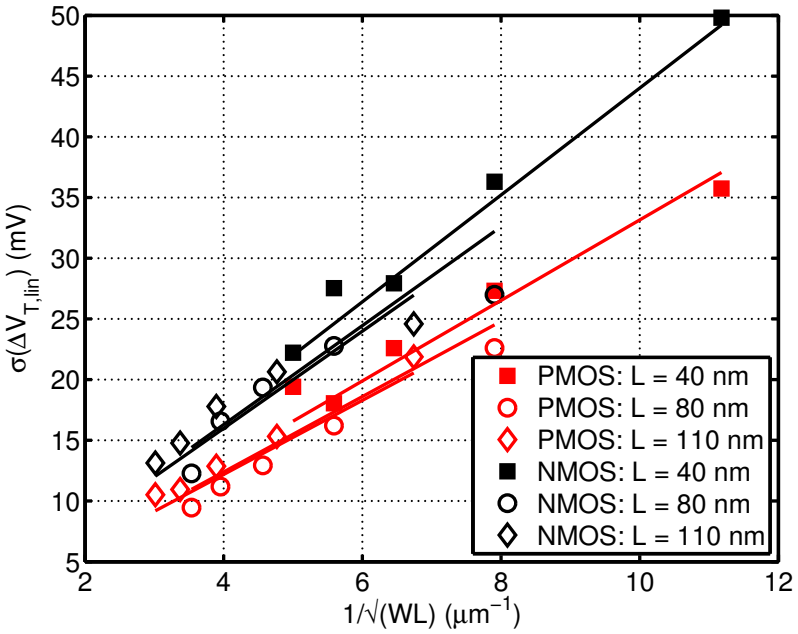


Figure 4.16: Pelgrom plot of $\sigma_{\Delta V_{T,lin}}$ for both NMOS and PMOS devices; using effective rather than drawn area results in a better fit to the measured data (solid lines).

The NMOS devices exhibit a Pelgrom slope of 4.10 $mV/\mu m$, while the PMOS devices have a smaller slope of 3.15 $mV/\mu m$. This can be traced to issues related to the fabrication of the source/drain extensions of the device, where the annealing process has been shown to result in a comparative increase in NMOS dopant atom fluctuations due to defect migration [57]. For completeness, it should be noted that Eq. 4.13 accounts only for variations in the

threshold voltage due to RDF, and does not account variations due to fluctuations in W or L. This approximation is used because simple expressions for the dependance of $V_{T,lin}$ on W and L are not readily available, and in general, these dependancies are expected to be relatively weak compared to the effects of RDF. If a more accurate model accounting for these variations is required, it can be derived using the treatment described in [40], where partial derivatives accounting for the sensitivity of $V_{T,lin}$ to changes in W and L are numerically extracted from the BSIM device models, and used to propagate the variance in W and L to partially account for some of the variability in $V_{T,lin}$.

Another version of the threshold voltage, $V_{T,C}$, can be extracted from C-V measurement data, as described in Section 4.3.2. In order to validate the combined C-V/I-V methodology, the correlation between $\Delta V_{T,C}$ and $\Delta V_{T,lin}$ can be examined. Fig. 4.17 shows correlation plots of the two parameters across the entire measurement sample set (156 device pairs per type for 15 different device types). The extracted correlation coefficients for NMOS and PMOS devices are $\rho_N = 0.67$ and $\rho_P = 0.63$, showing a reasonable correlation between the two measured parameters. Differences in $\Delta V_{T,C}$ and $\Delta V_{T,lin}$ can be accounted for by the different definitions of the threshold voltage parameter, as well as the body effect observed during C-V measurements, where the channel-to-body voltage, V_{CB} is effectively swept along with the gate-to-channel voltage, V_{GC}.

G_M Variability

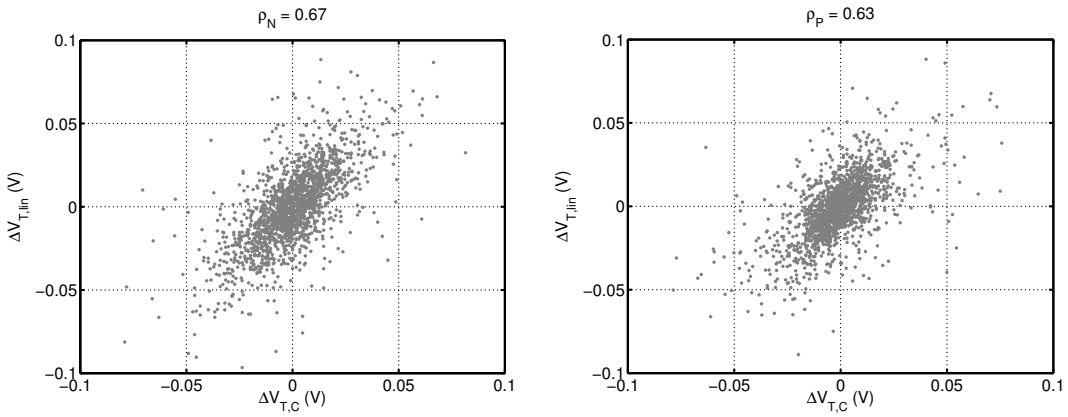
When studying the variability in G_M, the following model for the drain current, I_D is assumed

$$I_D = k\frac{W}{L_{eff}}(V_{GS} - V_{T,lin})V_{DS}, \tag{4.14}$$

where k is the product of the effective channel mobility, μ_{eff}, and oxide capacitance per unit area, C'_{ox}. Therefore, by the definition in Eq. 4.5, G_M is given by

$$G_M(V_{GS}) \equiv \frac{\partial I_D}{\partial V_{GS}} = k\frac{W}{L_{eff}}V_{DS}. \tag{4.15}$$

To study the relative variability in $\frac{\Delta G_M}{G_M}$, the effects of the variance of W, L_{eff}, and



(a) NMOS (b) PMOS

Figure 4.17: Correlation plots between $V_{T,C}$ and $V_{T,lin}$ for (a) NMOS and (b) PMOS devices; correlation coefficients of $\rho_N = 0.67$ and $\rho_P = 0.63$, respectively, demonstrate the close relationship between measured variability in the C-V and I-V characteristics of devices across the entire sample set.

k can be propagated, as described in [40]. In particular, the variance of $\frac{\Delta G_M}{G_M}$ is given by

$$\sigma^2_{\Delta G_M/G_M} = \left(\frac{\partial\frac{\Delta G_M}{G_M}}{\partial\Delta L}\right)^2 \sigma^2_{\Delta L} + \left(\frac{\partial\frac{\Delta G_M}{G_M}}{\partial\Delta W}\right)^2 \sigma^2_{\Delta W} + \left(\frac{\partial\frac{\Delta G_M}{G_M}}{\partial\frac{\Delta k}{k}}\right)^2 \sigma^2_{\Delta k/k}, \tag{4.16}$$

where

$$\sigma^2_{\Delta L} = \frac{A^2_{\Delta L}}{W}, \tag{4.17}$$

$$\sigma^2_{\Delta W} = \frac{A^2_{\Delta W}}{L_{eff}}, \tag{4.18}$$

and

$$\sigma^2_{\Delta k/k} = \frac{A^2_{\Delta k/k}}{WL_{eff}}. \tag{4.19}$$

The partial derivatives in Eq. 4.16 can be derived from Eq. 4.15 as follows:

$$\begin{aligned} \frac{\partial\frac{\Delta G_M}{G_M}}{\partial\Delta L} &\approx \frac{1}{L_{eff}} \\ \frac{\partial\frac{\Delta G_M}{G_M}}{\partial\Delta W} &= \frac{1}{W} \\ \frac{\partial\frac{\Delta G_M}{G_M}}{\partial\frac{\Delta k}{k}} &= 1. \end{aligned} \tag{4.20}$$

Plugging Eq. 4.17, Eq. 4.18, Eq. 4.19, and Eq. 4.20 into Eq. 4.16 results in

$$\sigma^2_{\Delta G_M/G_M} = \frac{A^2_{\Delta L}}{WL^2_{eff}} + \frac{A^2_{\Delta W}}{W^2L_{eff}} + \frac{A^2_{\Delta k/k}}{WL_{eff}}. \tag{4.21}$$

Table 4.2: Extracted Values of $A_{\Delta L}$, $A_{\Delta W}$, and $A_{\Delta k/k}$ from Eq. 4.21

Parameter	**NMOS**	**PMOS**
$A_{\Delta L}$ $(\mu m^{3/2})$	0.4×10^{-3}	0.5×10^{-3}
CI	$[0 - 1.1] \times 10^{-3}$	$[0.3 - 0.8] \times 10^{-3}$
$A_{\Delta W}$ $(\mu m^{3/2})$	2.0×10^{-3}	0.6×10^{-3}
CI	$[1.1 - 2.9] \times 10^{-3}$	$[0 - 2.1] \times 10^{-3}$
$A_{\Delta k/k}$ (μm)	8.2×10^{-3}	6.8×10^{-3}
CI	$[7.1 - 9.3] \times 10^{-3}$	$[6.1 - 7.5] \times 10^{-3}$

Fig. 4.18 shows a Pelgrom plot of the measured standard deviation of the relative transconductance, $\sigma^2_{\Delta G_M/G_M}$, for both NMOS and PMOS devices across the span of W and L_{eff}. Solid lines represent a fit to Eq. 4.21, and the extracted parameters along with their 95% confidence intervals are listed in Table 4.2. While the variability in k accounts for most of the G_M variability in either device type, values for $A_{\Delta L}$ and $A_{\Delta W}$ are also extracted. Extracting values for these two parameters is important, as it gives an indication of the amount of line-edge roughness (LER) present in the technology. However, due to the dominance of the $A_{\Delta k/k}$ term, the uncertainty in determining the LER parameters is too large.

It is interesting to note that the $A_{\Delta k/k}$ term for PMOS transistors, and consequently, the overall $\sigma_{\Delta G_M/G_M}$, is less than that for NMOS transistors in this technology, even as the higher μ_{eff} and slightly higher C'_{ox} of the NMOS transistors result in a higher value for k and should tend to decrease $\sigma_{\Delta k/k}$. If it is assumed that variations in μ_{eff} dominate variations in k, then the data presented in Fig. 4.18 indicates that the mobility in the PMOS channel is better controlled. This could be related to the fact that different strain techniques are used to boost the mobility in the NMOS and PMOS channels, as mentioned in Chpater 2, and would indicate that compressive strain, used in PMOS devices, is better controlled than tensile stress, used in NMOS devices. Additionally, Coulomb scattering is expected to give rise to more μ_{eff} variability in NMOS devices as compared to PMOS devices due to the comparatively larger RDF, as observed in the analysis of $\Delta V_{T,lin}$ variations.

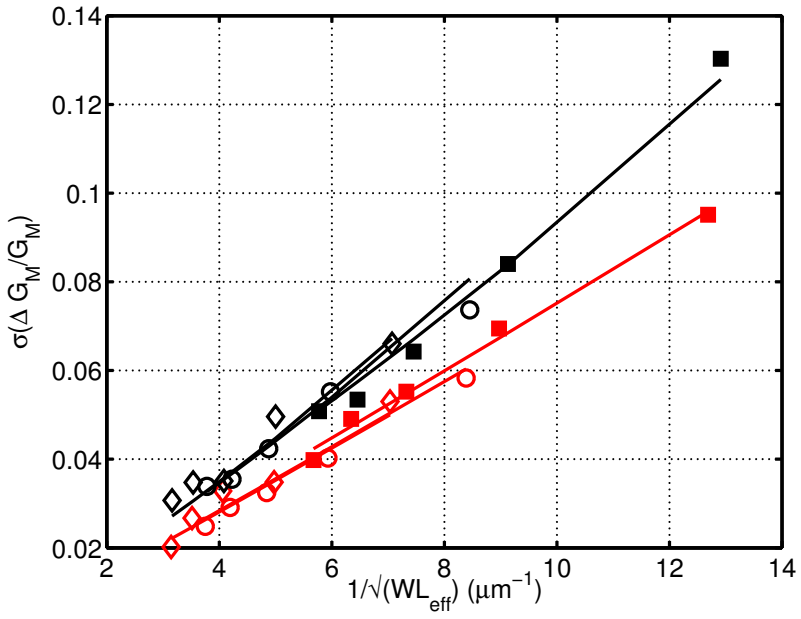


Figure 4.18: Pelgrom plot of $\sigma_{\Delta G_M/G_M}$ against the inverse square root of effective device area, $1/\sqrt{WL_{eff}}$; solid lines represent a fit to Eq. 4.21.

C_{GC} Variability

Similarly to the treatment of the variability in G_M described above, propagation of variance can be used to gain a better understanding of the causes of variability in the intrinsic gate-to-channel capacitance, $C_{GC,int}$. In order to be able to compare data from both NMOS and PMOS measurements, the measurements of the intrinsic gate-to-channel capacitance are done by subtracting the capacitance in depletion from the capacitance in strong inversion, as described by Eq. 4.9. However, for simplicity, the form

$$C_{GC,int} \approx C'_{ox} W L_{eff} \tag{4.22}$$

will be used, where W and L_{eff} represent the effective dimensions of the device, and C'_{ox} represents the oxide capacitance per unit area, as extracted in Section 4.3.3. Applying propagation of variance gives

$$\sigma^2_{\Delta C_{GC,int}} \approx \left(\frac{\partial \Delta C_{GC,int}}{\partial \Delta L}\right)^2 \sigma^2_{\Delta L} + \left(\frac{\partial \Delta C_{GC,int}}{\partial \Delta W}\right)^2 \sigma^2_{\Delta W} + \left(\frac{\partial \Delta C_{GC,int}}{\partial \Delta C'_{ox}}\right)^2 \sigma^2_{\Delta C'_{ox}}, \tag{4.23}$$

where

$$\sigma^2_{\Delta C'_{ox}} = \frac{A^2_{\Delta C'_{ox}}}{W L_{eff}}, \tag{4.24}$$

and $\sigma^2_{\Delta L}$ and $\sigma^2_{\Delta W}$ are given by Eq. 4.17 and Eq. 4.18, respectively. Plugging in Eq. 4.22, Eq. 4.24, Eq. 4.17, and Eq. 4.18 into Eq. 4.23, the expression

$$\frac{\sigma^2_{\Delta C_{GC,int}}}{C'^2_{ox}} \approx A^2_{\Delta L} W + A^2_{\Delta W} L_{eff} + A^2_{\Delta C'_{ox}/C'_{ox}} W L_{eff} \tag{4.25}$$

is derived, where

$$A_{\Delta C'_{ox}/C'_{ox}} = \frac{A_{\Delta C'_{ox}}}{C'_{ox}}. \tag{4.26}$$

It should be noted that this expression allows the extraction of the LER parameters $A_{\Delta L}$ and $A_{\Delta W}$, as well as the relative variance of the oxide capacitance per unit area given by

$$\sigma^2_{\Delta C'_{ox}/C'_{ox}} = \frac{A^2_{\Delta C'_{ox}/C'_{ox}}}{W L_{eff}}. \tag{4.27}$$

All of these parameters are also present in the treatment of $\sigma_{\Delta G_M/G_M}$ discussed above. The ability to extract and compare variability in physical device parameters from both C-V and I-V data is uniquely enabled by the combined C-V/I-V characterization methodology described in this work.

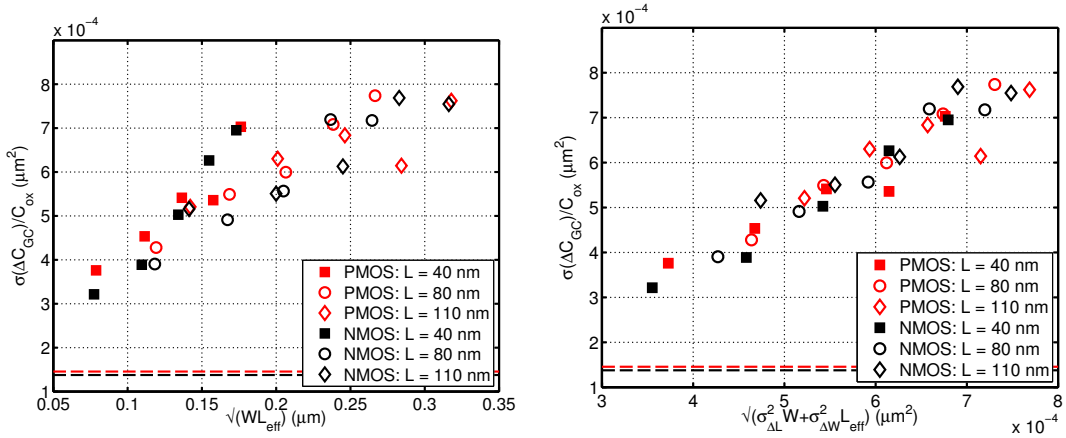
In the process of fitting the measured $C_{GC,int}$ variability data to the model shown in Eq. 4.25, it is found that the $\sigma^2_{\Delta C'_{ox}/C'_{ox}}$ term does not contribute at a statistically significant level and can be ignored in the case of both NMOS and PMOS data analysis. This result points to the fact that the major source of variation in gate-to-channel capacitance matching is not traced back to variations in C'_{ox}, but to variations in the effective dimensions of the device instead. Such a conclusion is logical, as the variations studied are very localized due to the proximity of the matched DUTs, and the gate oxide film is not expected to vary significantly over a small distance. Unlike the case of $\sigma_{\Delta G_M/G_M}$, where variations in the parameter k account for most of the measured variability, in the case of $\sigma_{\Delta C_{GC,int}}$ essentially all of the variability comes from $\sigma_{\Delta L}$ and $\sigma_{\Delta W}$. Consequently, analyzing $\sigma_{\Delta C_{GC,int}}$ proves to be a much more effective way of extracting the variation in device geometry due to LER.

Fig. 4.19 shows $\frac{\sigma_{\Delta C_{GC,int}}}{C'_{ox}}$ plotted agains the effective device area, $\sqrt{WL_{eff}}$, as well as against $\sqrt{\sigma^2_{\Delta L} W + \sigma^2_{\Delta W} L_{eff}}$. The former is equivalent to a classical Pelgrom treatment of $\sigma_{\Delta C_{GC,int}/C_{GC,int}}$ modeled as proportional to $\frac{1}{\sqrt{WL_{eff}}}$; such an interpretation of the measured variability in $C_{GC,int}$ across geometry is presented in [72], but even there the data

Table 4.3: Extracted Values of $A_{\Delta L}$ and $A_{\Delta W}$ from Eq. 4.25

Parameter	**NMOS**	**PMOS**
$A_{\Delta L}$ $(\mu m^{3/2})$	0.6×10^{-3}	0.6×10^{-3}
CI	$[0.5 - 0.7] \times 10^{-3}$	$[0.5 - 0.7] \times 10^{-3}$
$A_{\Delta W}$ $(\mu m^{3/2})$	1.1×10^{-3}	1.3×10^{-3}
CI	$[0.8 - 1.5] \times 10^{-3}$	$[0.9 - 1.8] \times 10^{-3}$

reported fails to give an accurate fit to the basic Pelgrom expression. On the other hand, a much better alignment between data from minimum-length devices and the rest of the DUT parameter space is seen in Fig. 4.19(b), where only the perimeter effects due to LER are considered.



(a) (b)

Figure 4.19: $\frac{\sigma_{\Delta C_{GC,int}}}{C'_{ox}}$ (a) as a function of $\sqrt{WL_{eff}}$, and (b) taking LER into account; modeling $\frac{\sigma_{\Delta C_{GC,int}}}{C'_{ox}}$ as dependent on the device perimeter rather then the device area aligns minimum-length device data (filled squares) with the rest of the measurements; dashed lines indicate the C-V measurement noise floor, conservatively defined as twice the maximum measured error in Fig. 4.8.

The extracted data are fairly noisy, as the standard deviations measured are close to the overall measurement precision, marked with dashed lines in Fig. 4.19. However, due to the large number of points, $A_{\Delta L}$ and $A_{\Delta W}$ can be extracted fairly precisely, as shown in Table 4.3. A comparison between the values in Table 4.3 and Table 4.2 shows that the LER parameters extracted using the $C_{GC,int}$ variability data fall within the confidence interval of those extracted using the G_M variability data, but offer much higher precision. It is interesting to note that L_{eff} is controlled about twice as well as W, pointing to the fact

that in general more care and resources are put into controlling the length of the device. This is logical, as variations in L_{eff} have a higher impact on overall device performance, since the length of the device is normally much smaller than its width. Additionally, the fact that $\frac{\sigma^2_{\Delta C_{GC,int}}}{C'^2_{ox}}$ is insensitive to $\sigma^2_{\Delta C'_{ox}/C'_{ox}}$ implies that variations in μ_{eff} are the primary source of variation in $\frac{\Delta k}{k}$, as stipulated in the discussion of variation in G_M. Once again, this type of analysis underscores the benefit of being able to confidently cross-reference results based on both C-V and I-V characterization of the device, and demonstrates the utility of the combined C-V/I-V variability characterization approach.

4.3.5 Analysis of Systematic Variability

Another aspect of device variability, which can be studied using the proposed combined C-V/I-V characterization methodology, is systematic variability across the die. Up to this point, all analysis is based on modeling the matching of parameters, as described by Eq. 4.12, canceling out the effects of systematic variability in order to focus on random variability instead. However, systematic variability can also be studied, by extracting parameter gradients across the die. Such gradients are not random in nature, and do not scale with the area or the perimeter of the device; instead, they can be traced to a deterministic source which affects all devices in a similar manner independent of their geometry, as discussed in Chapter 2.

In order to examine parameter gradients across the DUT array, each parameter of interest is measured, and then measurements along a column of the chip are normalized according to

$$\boldsymbol{P}_n = \frac{\boldsymbol{P} - \mu(\boldsymbol{P})}{\sigma(\boldsymbol{P})} \tag{4.28}$$

where $\boldsymbol{P}$ is a vector of extracted parameter values along the height of the column, $\mu(\boldsymbol{P})$ is its mean value, $\sigma(\boldsymbol{P})$ is its standard deviation, and $\boldsymbol{P}_n$ is the resulting normalized parameter vector. Using this type of normalization enables the study of gradient vectors along a DUT array consisting of different types of DUTs.

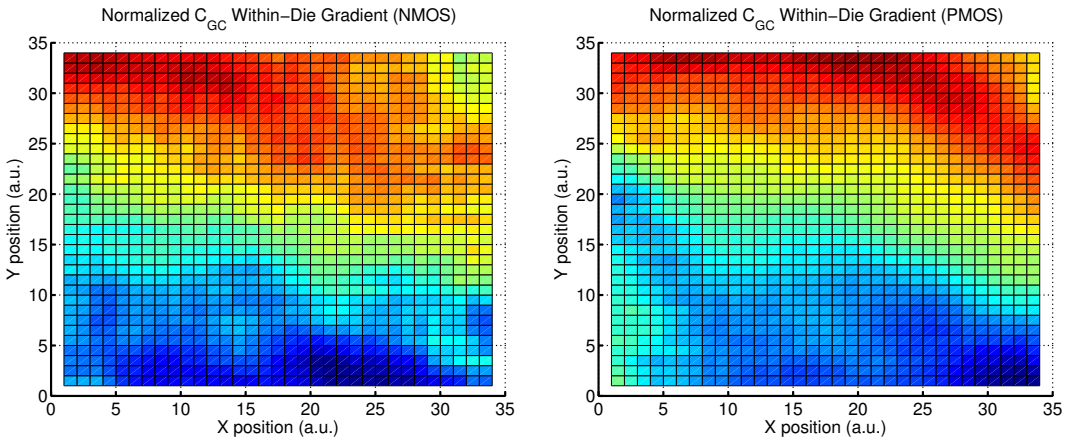
Heat maps for extracted normalized gradients of $C_{GC,int}$ across the DUT array for both NMOS and PMOS devices are shown in Fig. 4.20(a) and Fig. 4.20(b), respectively. A well-defined gradient is observed, which points to a systematic source of variation in the intrinsic gate-to-source capacitance. This source of variation appears to be common to both NMOS and PMOS devices, as indicated by a correlation coefficient between the two measurements of $\rho = 0.91$.

If gradients in the normalized large-signal transconductance, G_M, are considered, similar patterns emerge, as seen in Fig. 4.20(c) and Fig. 4.20(d). Interestingly, the gradients in Fig. 4.20(a) and Fig. 4.20(c), and those in Fig. 4.20(b) and Fig. 4.20(d), exhibit strong negative correlation, with correlation coefficients of $\rho = -0.66$ and $\rho = -0.82$ in the case of PMOS and NMOS measurements, respectively.

When considering the functional form of $C_{GC,int}$ shown in Eq. 4.6 and that of G_M shown in Eq. 4.15, it is clear that the only inverse correspondence between the two is through the device length variable, L. Therefore, by leveraging the combined C-V/I-V characterization approach, the source of the negatively-correlated systematic variation in the two parameters is identified as a systematic variation in the length of the device, L. Since similar gradients are present in all measured dice, this systematic error can be traced to a lithographical or mask-alignment issue. Systematic variations of device length across the reticle are also reported in [22]. It should be noted that the gradients presented in Fig. 4.20(a) and Fig. 4.20(b) are much smoother than those in Fig. 4.20(c) and Fig. 4.20(d), due to the fact that the latter set of measurements is obscured by a relatively high variation in μ_{eff} on top of the underlying systematic length variation. This once again demonstrates the benefit of studying variation in the gate-to-channel capacitance, as it gives a much cleaner representation of variations in the intrinsic device geometry.

4.4 Conclusion

Measurement techniques for accurate I-V and C-V device variability characterization using the combined C-V/I-V on-chip characterization system discussed in Chapter 3 have been



(a) $C_{GC,int}$, NMOS

(b) $C_{GC,int}$, PMOS

(c) G_M, NMOS

(d) G_M, PMOS

Figure 4.20: Normalized gradients across the die for (a) NMOS and (b) PMOS $C_{GC,int}$, and (c) NMOS and (d) PMOS G_M.

demonstrated. A four-point Kelvin I-V measurement technique is shown to overcome issues related to parasitic on-chip resistance, and a CBCM C-V characterization technique is shown to be insensitive to leakage and parasitic capacitance. With the help of oversampling, filtering, and noise cancellation, atto-Farad resolution C-V characterization of circuit-representative devices in large sample sets is achieved. In order to demonstrate the utility of the proposed combined C-V/I-V characterization approach, random variability in V_T, G_M, and $C_{GC,int}$ is studied, with an emphasis on using information from both sets of electrical measurements to extract underlying physical causes of device variability. C-V characterization is shown to be particularly suitable for extracting parameters related to the effective area of the device, such as L_{eff}, and variability in effective device dimensions due to LER. In the case of systematic gradients across the chip, C-V/I-V correlation is used to identify a variation in L as the major contributing factor. While quasi-static C-V/I-V data can be used to study a much larger array of electrical parameter variations by using more sophisticated device models and measurements in different regions of operation, the analysis presented in this chapter is sufficient to demonstrate the effectiveness of the proposed approach.

Chapter 5

Random Telegraph Noise Characterization

5.1 Introduction

In addition to characterizing quasi-static device variability, as discussed in Chapter 4, the on-chip characterization system described in Chapter 3 can also be used for time-domain characterization of random telegraph noise (RTN). This chapter begins with an overview of RTN. The study of this phenomenon is put in a brief historical context and the basic mechanisms behind RTN is small-area devices are discussed. A direct on-chip RTN characterization approach is introduced, and a methodology for extraction of RTN parameters from measured time domain data is developed. Based on the acquired data, a statistical model for the prediction of overall drain current fluctuations due to RTN is established.

5.2 Overview of RTN in Semiconductors

Random telegraph noise is a low-frequency noise phenomenon in semiconductor devices, which in the context of MOS transistor operation manifests itself as sudden discrete random jumps in the drain current amplitude, as shown in the example measured waveform in Fig. 5.1. RTN has been an issue of growing concern in modern-day CMOS technology nodes,

especially as minimum channel length scales down to 45-nm and below [18, 89–92]. Much effort has been directed towards the characterization and statistical modeling of RTN.

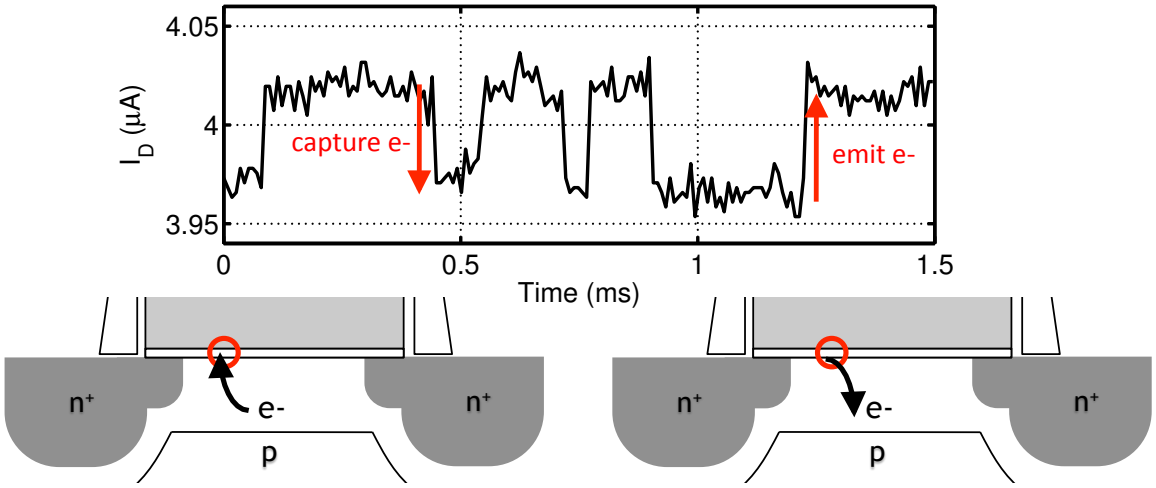


Figure 5.1: An example of a measured two-level RTN waveform along with an illustration of the underlying carrier trapping process.

5.2.1 Historical Perspective

Random telegraph noise (also referred to as burst or popcorn noise) has long been a studied low-frequency noise phenomenon in semiconductor devices. It has been observed in work dating all the way back to the early days of semiconductor development, with RTN signals identified as a source of low-frequency noise and instability in reverse-biased p-n junctions [93–95]. With regards to gated MOS device structures, some of the earliest efforts towards measurement and characterization of random telegraph noise date back to 1969 and the work of Hsu, et. al. [96]. A relationship between random telegraph noise, and a more common form of low-frequency noise known as $1/f$, or pink, noise has been suggested [19, 97], where the source of $1/f$ noise is identified as a superposition of a large number of RTN signals with varying amplitudes and capture/emission time constants.

In recent years, RTN has gradually emerged as a major source of concern in advanced CMOS technology nodes. Analog functionality is generally more sensitive to device noise, and RTN has been identified as a limiting factor in the design of small-area analog amplifiers, such as those used in CMOS imagers [98, 99]. However, as minimum device dimensions

shrink, RTN is becoming an issue in digital circuit design as well. Traditionally, floating gate flash memory devices are particularly sensitive to RTN due to their reliance on charge trapping as a mechanism for data storage [17, 91, 100, 101]; as is discussed below, RTN is associated with trapping and de-trapping of charges at the channel/oxide interface. However, more recently, RTN has also been identified as a potential source of failure in SRAM circuits [89, 102, 103], as well as a contributor to delay variation in logic circuits [104]. This trend is only expected to worsen with scaling of the device area, making the measurement, characterization, and statistical modeling of random telegraph noise a popular research subject in recent years.

5.2.2 Source of RTN: Mobility vs. Carrier Density Modulation

The prevalent view in the literature is that RTN in MOS transistors is induced by the random trapping and de-trapping of charges in potential traps near the channel/oxide interface [49, 91, 97, 105]. However, the effect of trapped charges at the Si/SiO_2 interface on charge transport through the channel is not necessarily agreed upon. In particular, the observed discrete fluctuations in the drain current, I_D, are ascribed to modulation of either the number of carriers in the channel [106], the the mobility of the carriers in the channel [107], or both [108]. In the number-of-carriers interpretation, a single trapped charge is thought to randomly modulate the effective interface charge density of the transistor, causing discrete shifts in the flatband/threshold voltage of the device, which in turn give rise to the discrete random fluctuations observed in the drain current. At the same time, any fixed charge close to the channel can also be expected to have an effect on carrier mobility through Coulomb scattering, which gives a mechanism for the modulation of mobility as a source of RTN. Clearly, both interpretations are likely to be true, especially since one does not preclude the other. However, measurements performed in this work tend to point towards a fluctuation in the number of carriers in the channel as the dominant mechanism giving rise to random telegraph noise. This is an important point that will be examined in more detail in Section 5.3 and in Section 5.4, as it refers to evidence gathered from measurements of

both NMOS and PMOS devices, and the implications that it carriers with regards to the statistical modeling of RTN.

5.2.3 Multi-level RTN

While the RTN signal shown in Fig. 5.1 is due to a single trapped charge, in the more general case, RTN can be caused by the superposition of two or more active traps in the same device [81,105,109]. Conceptually, multi-trap RTN can be interpreted as the transition point between random telegraph noise and $1/f$ noise, as discussed above. An example of an RTN signal due to a three-trap system is shown in Fig. 5.2. Since the effects of each individual trap can be considered to be additive [105], predicting the overall impact of RTN on device performance is as much dependent on an accurate model of the number of traps per device as it is on the prediction of single-trap amplitudes. Therefore, developing a methodology to extract the number of traps from measured RTN waveforms is vital to the development of a robust statistical model predicting the overall drain current amplitude fluctuations due to RTN.

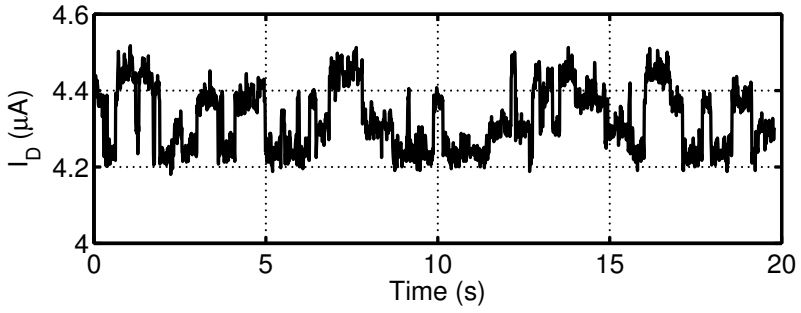


Figure 5.2: An example of a measured multi-level RTN waveform due to the superposition of the effects of three individual traps.

5.2.4 Scaling Trends

The effect of RTN is expected to increase with decreasing device size [105], similarly to other sources of device variability, such as threshold voltage variability due to random dopant fluctuations (RDF). However, in comparison to RDF, RTN amplitude is known to have a stronger dependence on device size [103], and is thus expected to increase proportionally

faster. More importantly, unlike the normally distributed RDF, overall RTN amplitude distributions have been shown to exhibit long tails, which could make it a major source of failure in high-density designs, such as SRAM blocks, at 22-nm and below [89]. All of these facts point towards the need for a better understanding of the statistical behavior of RTN, especially as it relates to device area and scaling. Uncovering the functional dependance of the overall RTN amplitude on the width, W, and length, L, of the device, is of particular interest.

5.3 Measurement and Characterization of RTN

5.3.1 Measurement Approach

The measurement approaches to RTN characterization used throughout the literature can be divided into two categories: direct measurements and stress-based measurements.

Direct measurement

One of the main obstacles in direct characterization of RTN comes from the necessity to measure small currents at high sampling rates. The need for low current measurements comes from the fact that RTN needs to be characterized in small-area devices, where its impact is expected to be most noticeable. These devices tend to be narrow, and have comparatively low current drives. Additionally, it is desirable that the device-under-test (DUT) is biased in the linear region of operation, where it tends to exhibit current-driving behavior well-approximated by the simple equation

$$I_D = \mu C'_{ox} \frac{W}{L} \left(V_{GS} - V_{th} - \frac{V_{DS}}{2} \right) V_{DS}. \tag{5.1}$$

Eq. 5.1 provides a straightforward methodology for extracting ΔV_{th} variations from measured ΔI_D variations, as needed to model RTN as the result of modulation of the number of carriers in the channel. The linear region of operation is also preferred, because of diminished short-channel effects, which can otherwise make data difficult to interpret across geometry, especially as devices span different channel lengths.

At the same time, RTN trap events can have short characteristic capture and emission times, which means that the measurement sampling speeds should be fast enough to be able to accurately observe these events. In normal testing conditions, achieving the desired sampling rates of at least 50-100 kHz at the required resolution of a few nA can be rather challenging and usually requires direct probing [109]. Direct probing substantially reduces the amount of DUTs available for characterization, which makes the extraction of accurate statistical data very difficult.

The on-chip I-V characterization system described in Chapter 3 is ideally suited for high-frequency, high-resolution time-domain measurement of RTN. Integrating the measurement circuitry on the same die as the transistor DUT array drastically reduces interconnect capacitance and enables fast sampling rates for low-amplitude signals, even in the presence of non-negligible resistance through the on-chip switching matrix. Additionally, the high-density DUT array offers access to a number of configurations of small-area MOS transistors, with different transistor polarities, threshold voltages, and relative dimensions. The DUTs are organized in large sample sets, enabling the analysis of the statistics of RTN across a wide parameter space.

Stress-based measurement

In addition to the direct measurement approach used in this work, stress-based techniques for the characterization of RTN have also been proposed [18, 97, 110]. Electrical stress RTN measurements are akin to bias-temperature instability (BTI) measurements, and in fact, a close relationship between BTI and RTN has recently been established [18]. BTI effects, associated with wearing out the device over time, are traced to defects induced at the channel/oxide interface similar to those giving rise to RTN, and also lead to a similar deterioration of the threshold voltage.

The general idea behind stress-based RTN characterization is that electrically stressing the device by applying a large gate bias, sometimes as high as two times the nominal V_{DD} supply voltage, causes interface traps to be filled with high probability. As the device

is put back in nominal bias conditions, a drain current relaxation curve due to charges gradually leaving the interface traps is observed. If the device indeed exhibits RTN, the relaxation curve is composed of discrete jumps, rather than a smooth decay [18], each jump signifying the emptying of a single potential trap. While stress-based techniques could have a number of potential benefits, such as decreasing the monitoring time needed to observe a trap or allowing a more seamless extraction of individual trap amplitudes, they have the added disadvantage of putting the device in an unnatural state. Stressing the device could potentially lead to an overestimation of the effects of RTN by detecting traps which would normally never be occupied in normal operating conditions, or even inducing new traps as a result of hot carrier effects. Therefore, the work presented here focuses on the direct measurement approach, and stress-based methods are only discussed for the sake of completeness.

5.3.2 Measurement Setup

In order to ensure optimal performance, the on-chip characterization system has to be configured to measure low current signals with high dynamic range and at the best timing resolution possible. The I_{LSB} of the current-mode ADC is configured to 2.44 nA by setting the sampling time, N_{REF}, to 4096 clock cycles, and the reference current, I_{REF}, to 10 μA. At a 12-bit resolution, the maximum current that can be measured is 10 μA, which is sufficient considering the bias range and DUT dimensions of interest. The ADC is operated in asynchronous mode, with each new conversion cycle beginning as soon as the previous cycle has been completed, thus ensuring optimal timing resolution. As a result, the sampling speed of the ADC varies as a function of signal strength, as illustrated in Fig. 5.3. For the signal levels and resolution considered here, sampling speeds vary between 125 kHz and 75 kHz. Since the time-domain behavior of RTN is also of interest, the ADC sampling frequency is recorded along with the data to enable accurate time- and frequency-domain representation of the sampled waveforms. In order to prevent sampling frequency fluctuations as a result of signal amplitude fluctuations due to RTN, the measurement frequency

is allowed to change only between bias points.

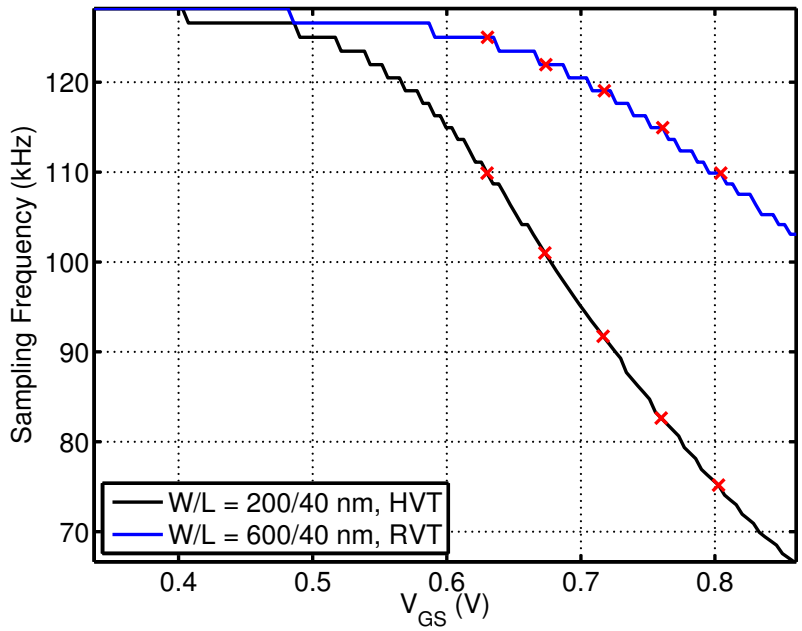


Figure 5.3: Varying sampling frequency due to signal-level-optimized ADC operation - sampling frequencies for a high-current and a low-current NMOS DUT are plotted versus V_{GS} bias; red crosses mark sampling rates at the RTN measurement bias points.

The gate-to-source bias range, V_{GS}, covered by the RTN measurements consists of five equally-spaced bias points, starting at 0.63 V and ending at 0.80 V, with a drain-to-source potential, V_{DS}, set to 50 mV. This bias range is chosen so that the device operates in the linear region and in strong inversion, where Eq. 5.1 gives a reasonable small-signal approximation for the relationship between I_D and V_{th}.

The sample size at each bias point is 2^{21} sample points, which results in a measurement duration between 17 and 28 seconds per bias point. While RTN traps can have characteristic capture and emission times of up to a few hours or more [105], characterizing a large number of devices at such long time intervals for a number of bias points and across a number of different geometries is impractical. In addition, when a subset of the data is sampled at intervals up to four times longer appreciable differences in the final results are not observed, implying that the vast majority of traps present in the population are detectable using the shorter sampling interval.

The measurement DUT sample set consists of an orthogonal set of minimum-length

devices ($L = 40\ nm$, $W = 200, 400, 600\ nm$) and minimum-width devices ($W = 200\ nm$, $L = 40, 80, 110\ nm$), enabling the study of RTN properties as a function of device geometry. Both NMOS and PMOS devices are measured, where the bias for the two device polarities is set such that $V_{GS_{NMOS}} = V_{SG_{PMOS}}$ and $V_{DS_{NMOS}} = V_{SD_{PMOS}}$. The DUTs are organized in statistical sets of 78 devices per chip, and a total of four chips are measured for an overall statistical sample set of 312 DUT per DUT type. The size of the sample set is sufficiently large to enable observing statistics at the 95-percentile level.

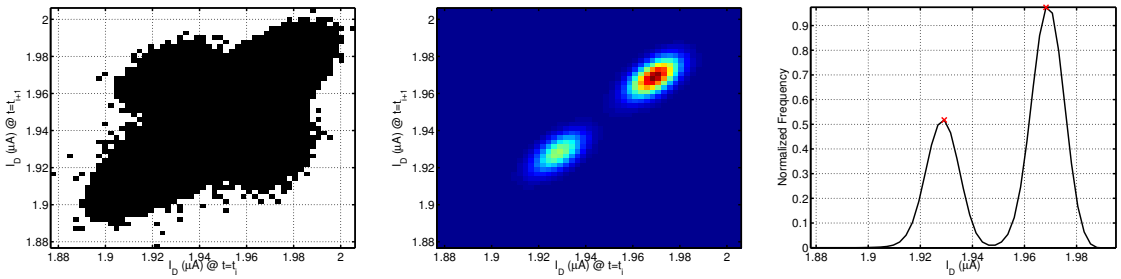
5.3.3 Parameter extraction

Due to the high volume and random nature of the measured data, a fully-automated analysis methodology has to be developed to extract RTN parameters from $I_D(t)$ measurements. In particular, the quantities of interest that need to be extracted are the number of observed traps, N_T, and the RTN amplitude associated with individual traps. Since RTN is modeled as a ΔV_{th} effect, and a technique to map ΔI_D fluctuations to ΔV_{th} fluctuations is also discussed. Finally, a hidden Markov model (HMM) is used to extract time-domain RTN behavior in order to identify different types of traps present in the DUT population.

Time Lag Plot

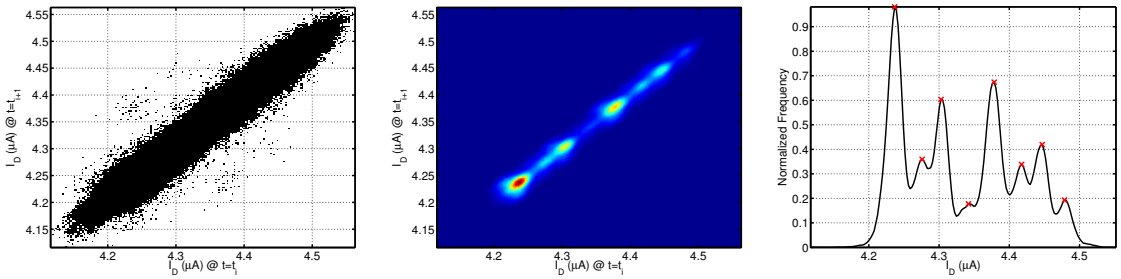
A time lag plot (TLP), also known as a lag scatter plot, is a tool for analyzing autocorrelation in time-series data, which can be used as a tool for analyzing time-domain RTN measurements [109]. TLPs are constructed by plotting data sampled at the i^{th} time interval, t_i, versus data sampled at $i^{th} + 1$ time interval, t_{i+1}. As the $I_D(t)$ waveform lingers at different RTN levels, the measured data at t_i and t_{i+1} is similar, and RTN levels appear as data clusters along the $I_D(t_i) = I_D(t_{i+1})$ diagonal of the TLP. This approach, however, has been impractical for analyzing large amounts of RTN data because of the lack of an unambiguous way to identify RTN levels, particularly in the presence of white and $1/f$ noise in the measured data.

In order to overcome these limitations, an enhanced TLP data analysis technique



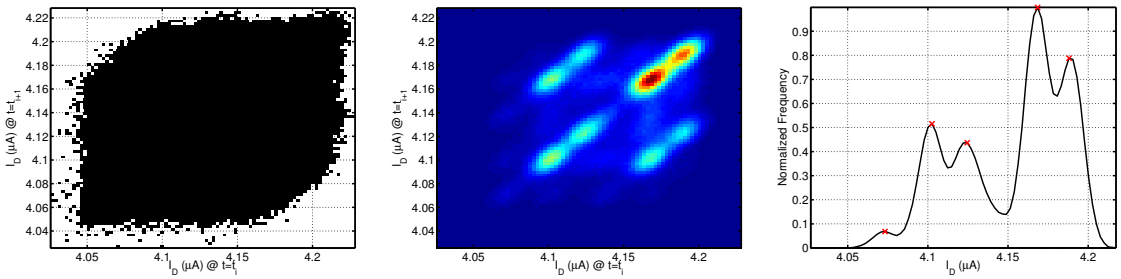
(a) (b) (c)

Figure 5.4: Comparison between (a) standard and (b) enhanced TLP for a single-trap RTN signal; the enhanced TLP diagonal along with detected RTN levels is shown in (c).



(a) (b) (c)

Figure 5.5: Comparison between (a) standard and (b) enhanced TLP for a triple-trap RTN signal; the enhanced TLP diagonal along with detected RTN levels is shown in (c).



(a) (b) (c)

Figure 5.6: Comparison between (a) standard and (b) enhanced TLP for another triple-trap RTN signal; the enhanced TLP diagonal along with detected RTN levels is shown in (c); it should be noted that while only five RNT levels are detected, this is enough evidence for the presence of three RTN traps

is developed. The new approach aims to automate detection of individual RTN levels even in the presence of additional noise, making the analysis of large statistical data sets feasible. In this case, we record the frequency with which each point of the TLP is occupied, transforming the TLP into a two-dimensional histogram of $I_D(t_i)$ vs $I_D(t_{i+1})$ with a bin size equal to one I_{LSB}. A comparison between using a standard and an enhanced TLP is shown in Fig. 5.4, Fig. 5.5, and Fig. 5.6, where three different RTN signals are analyzed. Analyzing the frequency with which a data point from the lag scatter plot is occupied makes the detection of distinct RTN levels possible even in cases where the amount of noise present makes the standard TLP approach impractical. These enhanced TLPs make possible the extraction of the number of traps as well as the trap amplitudes from RTN waveforms.

Extraction of Number of Traps, N_T

The first step in the extraction of the number of active traps in a device, N_T, is the detection of the number of distinct RTN levels present in the measured signal. For this purpose, the focus is shifted to the diagonal of the enhanced TLP, shown in Fig. 5.4(c), Fig. 5.5(c), and Fig. 5.6(c). Similarly to the approach described in [109], RTN levels are identified along the line $I_D(t_i) = I_D(t_{i+1})$. However, the diagonal of the enhanced TLP also gives information about the frequency, with which each point is occupied, and the number of local maxima extracted along the enhanced TLP diagonal represents the number of detected RTN levels, N_L. Local maxima are generally well-defined, due to the large number of samples in each RTN waveform, and can be easily extracted using a threshold-based peak detection algorithm.

In order to extract the number of RTN traps present, the following relationship is used [109]:

$$N_T = ceil(\log_2(N_L)) \tag{5.2}$$

where $ceil(x)$ represents the ceiling function, which rounds up x to the nearest integer value. Eq. 5.2 is based on the assumption that the effects of multiple traps are additive, so that the superposition of two traps results in four RTN levels and the superposition of three

traps results in eight RTN levels, and so on. The $ceil(x)$ function is used to get around the fact that not all combinations of RTN trap occupancy may be present in a measured RTN waveform.

It should be noted that if the RTN signal is considered to be represented by a mixture of Gaussians with different means, where each Gaussian represents a noisy RTN level, then a histogram of the RTN signal should give similar information to that acquired using the diagonal of the enhanced TLP, while at the same time being more computationally efficient to extract. This is certainly the case, as can be seen in Fig. 5.7, where the enhanced TLP diagonal and the histogram of the three-trap signal from Fig. 5.5 are plotted together for comparison. While the two curves carry essentially the same information, the enhanced TLP diagonal shows a sharper separation between the individual peaks, and unlike the simple histogram, allows all of the different RTN levels to be extracted accurately. Intuitively, this is due to the fact that the diagonal of the enhanced TLP represents a histogram of the points where the RTN signal lingered for two consecutive time steps, and as such, is expected filter out some of the noise present in the RTN signal. In fact, a more computationally efficient extraction of the TLP diagonal by constructing a histogram of the points occupied by the RTN signal for two consecutive time steps can be performed. This approach offers a good balance between accuracy and computational efficiency, which can be a concern when a large volume of data is processed.

Extraction of Single-Trap Amplitude, ΔI_D

The first step in extracting single-trap RTN amplitudes from measured data is to extract the changes in the measured current, ΔI_D, due to a single trap. The most straight-forward approach is to extract ΔI_D from RTN measurements where only a single trap has been observed, as suggested in [105]. Since the two peaks in the TLP diagonal represent the two RTN levels in a single-trap RTN waveform, the distance between the two peaks gives ΔI_D, as shown in Fig. 5.8. Therefore, the diagonal of the enhanced TLP can be leveraged not only in the detection of the number of traps, but also in the extraction of individual trap

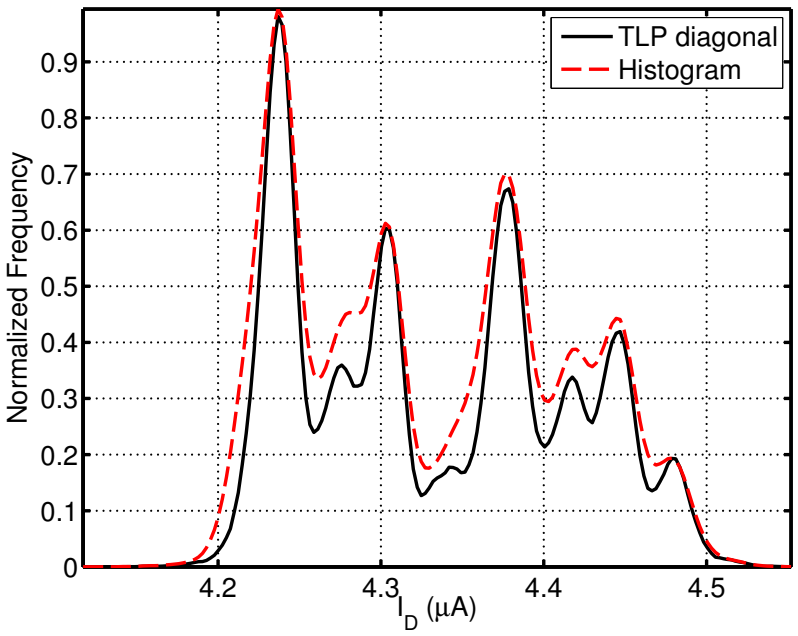
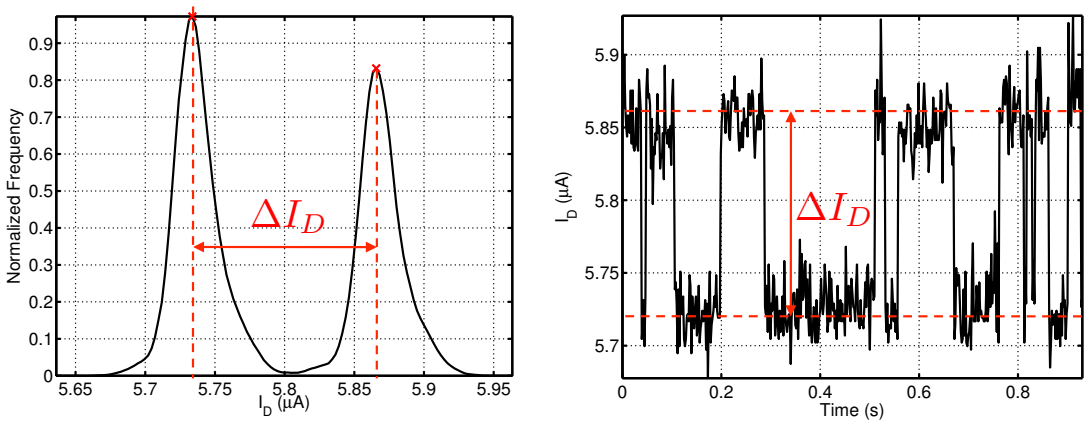


Figure 5.7: Comparison between the diagonal of the enhanced TLP and a histogram of the RTN waveform; both curves carry essentially the same information, but the diagonal of the enhanced TLP shows a better peak separation, making level detection more accurate.

amplitudes from noisy RTN measurements.

One concern about the approach described above is that it limits the number of extracted single-trap amplitudes to measurements based only on single-trap RTN signals. This can be problematic, especially if single-trap waveforms are rare in the studied population. In order to extract a larger number of single-trap amplitudes, the enhanced TLP diagonal can be used to extract multiple ΔI_D measurements from multi-trap RTN waveforms. In particular, since the effects of individual RTN traps are assumed to be additive and are modeled as such, multi-trap RTN signals are considered to be the superposition of multiple single-trap RTN signals. Since each trap has its own characteristic capture and emission times, presumably independent of any other traps in the same device, RTN levels due to different traps can be distinguished. The relative heights of the peaks in the diagonal of the enhanced TLP give an indication of the relative frequency, with which each trap is occupied. Therefore, the distance between the two highest peaks indicates the amplitude, $\Delta I_{D,1}$, associated with the dominant RTN trap; the distance between the highest peak



(a) (b)

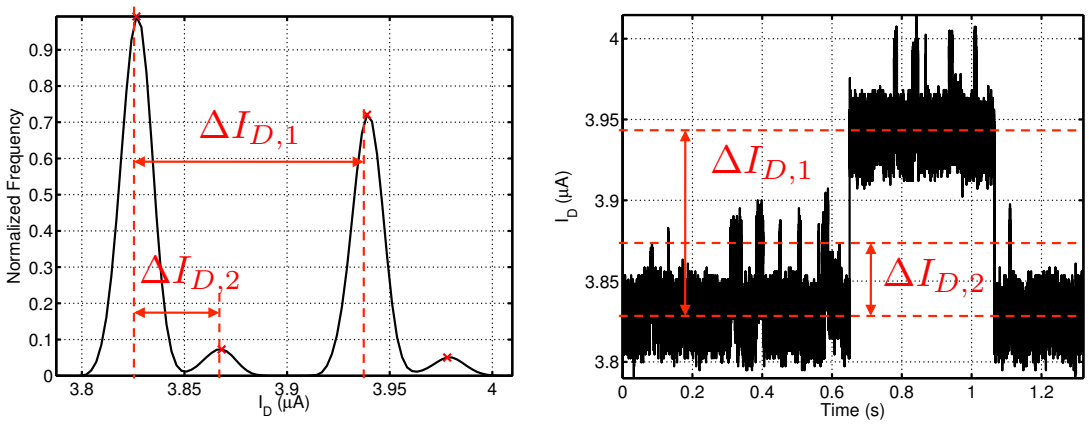
Figure 5.8: Extraction of ΔI_D from the distance between peaks in the enhanced TLP diagonal is shown in (a); the corresponding measured RTN waveform is shown in (b)

and the third highest peak indicates the amplitude, $\Delta I_{D,2}$, of the second dominant trap measured while the device is in its preferred state with regards to the first dominant trap.

This concept is illustrated in the case of a two-trap RTN signal in Fig. 5.9. However, it applies to any multi-trap RTN waveform. While extending the same logic in an attempt to extract more than two single-trap amplitudes from multi-trap signals becomes challenging, the increase in the number of samples gathered by just extracting two individual RTN amplitudes from each multi-trap RTN waveform is sufficient for the purposes of studying the statistics of individual trap amplitudes.

Extraction of ΔV_{th} from ΔI_D

In order to develop a robust statistical model for RTN, it is important to have an understanding of the dominant mechanism that governs the modulation of the drain current. As discussed in Section 5.2, two potential mechanisms have been proposed – effective mobility fluctuation and number of carrier fluctuation inducing a modulation of the flat-band/threshold voltage. If Eq. 5.1 is considered, then fluctuations in the mobility, μ, and the threshold voltage, V_{th}, are expected to have different effects on the drain current, I_D, as the gate-to-source voltage, V_{GS}, is swept.



(a) (b)

Figure 5.9: Extraction of two independent ΔI_D measurements from a single two-trap RTN waveform; measurements along the TLP diagonal are shown in (a) and the corresponding measured RTN waveform is shown in (b).

In particular, mobility fluctuation would result in

$$\Delta I_D = \Delta\mu C'_{ox}\frac{W}{L}\left(V_{GS} - V_{th} - \frac{V_{DS}}{2}\right)V_{DS}, \tag{5.3}$$

which would scale linearly with increasing V_{GS}. In order to avoid any potential secondary dependance on the other variables, Eq. 5.3 can be divided by Eq. 5.1, to get

$$\frac{\Delta I_D}{I_D} = \frac{\Delta\mu}{\mu}. \tag{5.4}$$

Therefore, if it is assumed that the relative change in mobility due to a trapped charge remains constant throughout the measured bias range, then the measured quantity $\frac{\Delta I_D}{I_D}$ should remain constant as well.

On the other hand, if the random telegraph noise is better described by a fluctuation in V_{th} due to a modulation of the number of carriers in the channel, then according to Eq. 5.1

$$\Delta I_D = \mu C'_{ox}\frac{W}{L}V_{DS}\Delta V_{th}. \tag{5.5}$$

Once again, assuming that the channel mobility, μ, remains relatively constant throughout the bias range of interest, ΔI_D should also remain constant for a constant ΔV_{th}.

In order to avoid any issues associated with IR drops across the DUT array, and

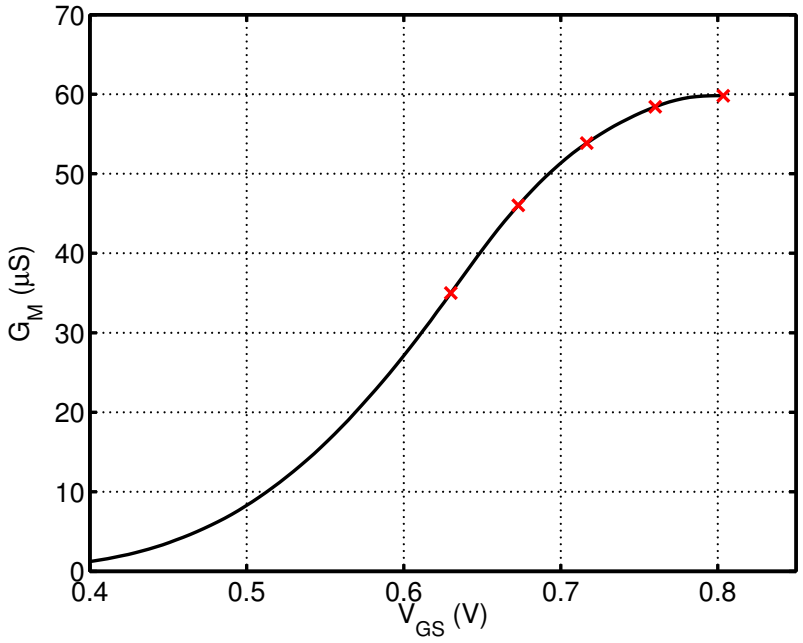


Figure 5.10: G_M extracted from a sweep of I_D vs V_{GS} for $V_{DS} = 50\ mV$; red crosses mark G_M values at the bias points where RTN measurements are taken.

more importantly, the dependence of μ on V_{GS}, Eq. 5.5 can be expressed as

$$\Delta V_{th} = \frac{\Delta I_D}{G_M} \tag{5.6}$$

where G_M is given by

$$G_M \equiv \frac{\partial I_D}{\partial V_{GS}} = \mu C'_{ox} \frac{W}{L} V_{DS}. \tag{5.7}$$

G_M can be measured directly by performing an I-V sweep of I_D as a function of V_{GS} for a V_{DS} nominally set to 50 mV, and the values of G_M at the bias points where the RTN measurements are performed can be extracted as shown in Fig. 5.10. Any variations of V_{DS} and μ as a function of V_{GS} cancel out in the extraction of V_{th} using Eq. 5.6, as they affect the measured G_M and the measured ΔI_D in the same way. Therefore, if the cause of RTN is purely a modulation in the number of carriers, then the extracted values for ΔV_{th} using Eq. 5.6 should remain constant across V_{GS}.

Fig. 5.11 shows $\frac{\Delta I_D}{I_D}$ and ΔV_{th} extracted from the same set of representative single-trap NMOS and PMOS RTN signals as a function of V_{GS} and V_{SG}, respectively; all of the plotted quantities are normalized with respect to their average values across the mea-

surement bias range in order to facilitate easy comparison. A dependance on V_{GS}/V_{SG} is observed in either case, which indicates that RTN fluctuations cannot be perfectly modeled by either a fluctuation in the mobility, $\Delta\mu$, or a fluctuation in the threshold voltage, ΔV_{th}. However, modeling the RTN amplitude as a ΔV_{th} effect results in considerably less variation across the bias range, which implies that the modulation of carriers in the channel accounts for most of the observed ΔI_D fluctuations, and the mobility modulation due to Coulomb scattering as a result of trapped carriers at the Si/SiO_2 interface can be considered a secondary effect. Consequently, from this point on, RTN is analyzed and modeled as a modulation of the threshold voltage, mainly in an attempt to facilitate data analysis and interpretation, as decoupling the individual contributions of $\Delta\mu$ and ΔV_{th} to ΔI_D is difficult. It should be noted that this is the approach taken in most of the literature, especially in cases where measurement data is presented and analyzed [91, 100, 105].

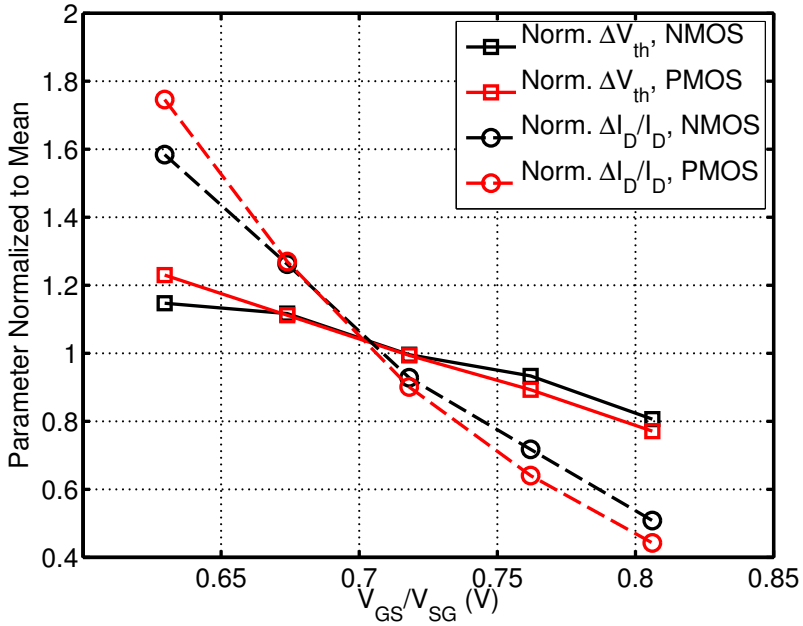


Figure 5.11: Plots of the amplitude of a single-trap RTN waveform interpreted as ΔV_{th} and $\frac{\Delta I_D}{I_D}$ for representative NMOS and PMOS devices as a function of V_{GS} and V_{SG}, respectively; the extracted parameters are normalized with respect to their mean values across the bias range for easier comparison.

Extraction of Characteristic Capture and Emission Times, τ_c and τ_e

While the enhanced TLP enables the characterization of the overall magnitude of an RTN waveform, it gives little information regarding the time-domain characteristics of the signal. The relative frequencies associated with peak heights in the TLP diagonal give an indication of the relative occupancy of each RTN state. However, in order to extract the actual characteristic capture and emission times associated with RTN traps, a different approach is needed. While the the superimposed noise in measured RTN waveforms makes direct time-domain analysis difficult, hidden Markov models (HMMs) [111] can be used to extract idealized RTN waveforms that can then be analyzed with regards to their time-domain characteristics.

HMMs have recently been widely adopted in the study and modeling of RTN in MOS devices [104, 112–114]. A single-trap RTN waveform can be modeled as a two-state Markov chain, where the two states are obscured by superimposed Gaussian noise. As a result, a wide range of readily available tools for the study of HMMs can be applied to analyze the measured RTN signals. In particular, the Baum-Welch algorithm [115] can be used to estimate the parameters of the hidden Markov chain using a log-likelihood estimation-maximization (EM) approach, and the Viterbi algorithm [116] can be used to extract the most likely path based on the measured data and extracted HMM parameters.

An example of noisy measured RTN data and the corresponding extracted ideal HMM waveform are shown in Fig. 5.12. The ideal waveform can be used to extract the probability distributions of capture and emission times, which are expected to be exponential, with a probability density function (PDF) given by

$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}. \tag{5.8}$$

The characteristic RTN capture and emission times, τ_c and τ_e, respectively, can be extracted by fitting Eq. 5.8 to the measured PDFs, as shown in Fig. 5.13.

Fig. 5.14 shows plots of τ_c and τ_e across bias for an NMOS and a PMOS single-trap RTN waveform. In both cases, a bistable trap, generally considered rare [97], is observed.

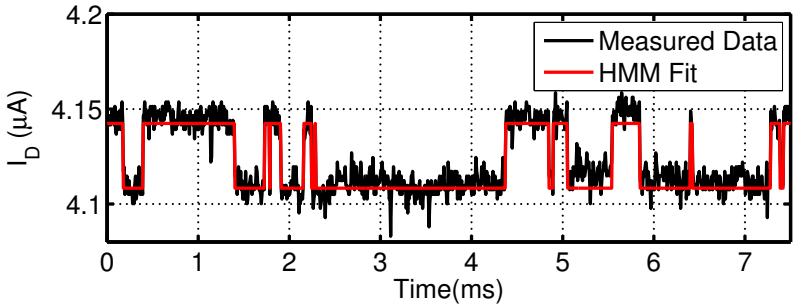
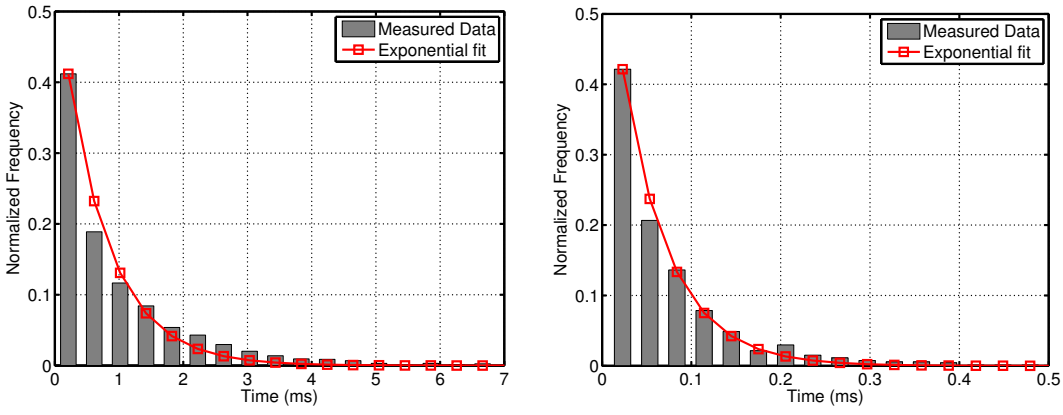


Figure 5.12: Measured RTN data and extracted ideal RTN waveform using the Viterbi algorithm.



(a) (b)

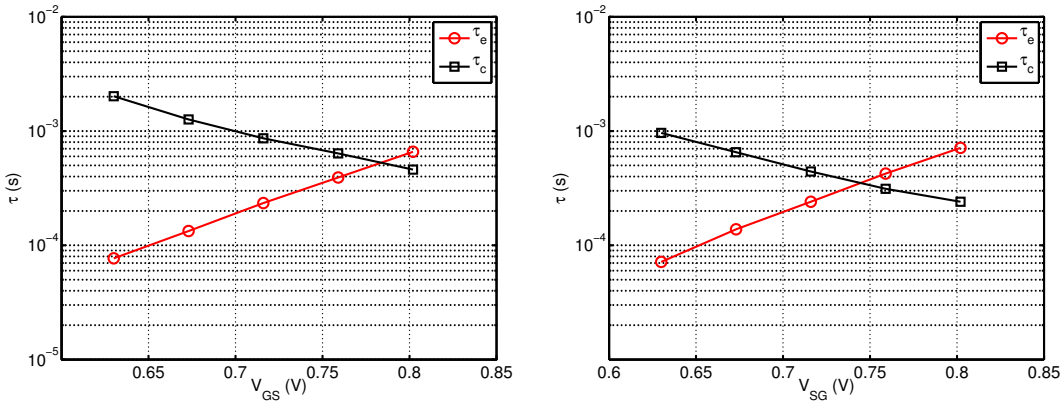
Figure 5.13: Extracting characteristic (a) capture and (b) emission times by fitting to an exponential PDF.

A bistable trap is characterized by capture and emission time constants that vary with bias. The characteristic capture time decreases as bias increases and more carriers are present in the channel. This is explained by the fact that as the number of carriers in the channel increases, the likelihood that a trap at the channel/oxide interface will become occupied by one of these carriers grows. The characteristic emission time of the bistable traps, on the other hand, increases as the number of carriers in the channel increases. This is indicative of a Coulomb interaction between the trapped carrier and carriers in the channel, as the presence of more carriers of the same polarity in the channel makes it more difficult for the trapped charge to be released. Such an interpretation supports the hypothesis that mobility degradation due to Coulomb scattering partially contributes to the overall RTN magnitude. In fact, the work of Miki, et. al. [112] shows that a statistically larger overall $\Delta I_D/I_D$ is observed in the case of bistable traps, as would be expected if mobility degradation is added on top the the primary threshold voltage effect.

While bistable traps are generally more prevalent in the studied population, neutral traps, such as the one shown in Fig. 5.15 are also present. Similarly to bistable traps, these traps exhibit a characteristic capture time which decreases as the number of carriers in the channel increases. However, the observed characteristic emission time is independent of gate bias, which is consistent with a trap situated deeper inside the gate oxide, where the Coulomb interaction between the trapped charge and carriers in the channel is limited.

5.4 Statistical Modeling of RTN

In order to accurately model the overall amplitude variations in I_D due to RTN, a comprehensive statistical model that encompasses the combined effects of amplitude variations and variations in the number of traps is needed. One approach is to separately model the statistics of the number of traps, N_T, and the statistics of single-trap amplitudes, ΔV_{th}, and then combine the two to construct a compact model for the prediction of the overall variation in I_D. This approach allows each of the two statistics to be observed independently as a function of the device dimensions in an attempt to uncover the basic mechanisms behind



(a) (b)

Figure 5.14: Characteristic capture and emission times for a bistable trap observed in (a) NMOS and (b) PMOS device; bistable traps are characterized by capture times, which decrease as the number of carriers in the channel increases, and emission times which increase as the number of carriers in the channel increases.

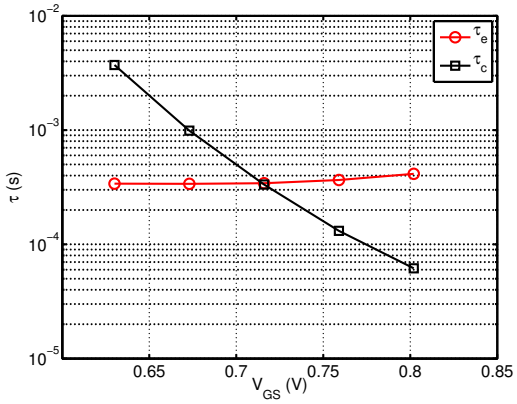


Figure 5.15: A neutral trap observed in an NMOS device; neutral traps are more rare than bistable traps in the studied population, and are characterized by capture times, which decrease as the number of carriers in the channel increases, and emission times which are independent of the number carriers in the channel.

the scaling trends associated with RTN.

5.4.1 Statistics of Number of Traps, N_T

The statistics of N_T have widely been reported in the literature to follow a Poisson probability distribution [101, 105, 109, 117]. While generally no theoretical basis is given for this interpretation, intuitively, a Poisson distribution is well-suited for the modeling of discrete random events that occur within a fixed area with a given average rate and independently of one another. As such, the Poisson distribution should lend itself well to the modeling of the random occurrence of potential traps along the channel/oxide interface of FETs.

The probability density function (PDF) of the Poisson distribution expressed in the context of predicting N_T is given by

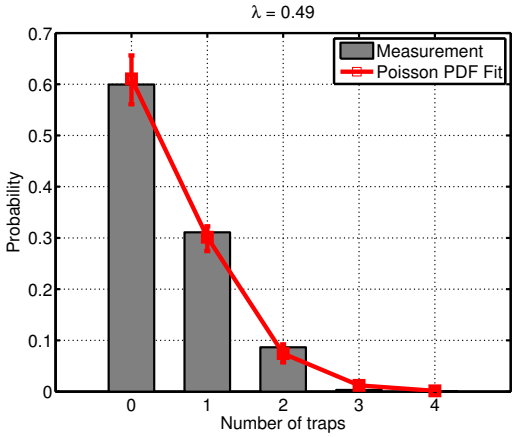
$$f_T(N_T;\lambda) = \frac{\lambda^{N_T} e^{-\lambda}}{N_T!}, \tag{5.9}$$

where λ represents the population mean of N_T. λ is the only parameter describing the Poisson distribution and is, therefore, the parameter of interest to be extracted from the measured distributions of N_T.

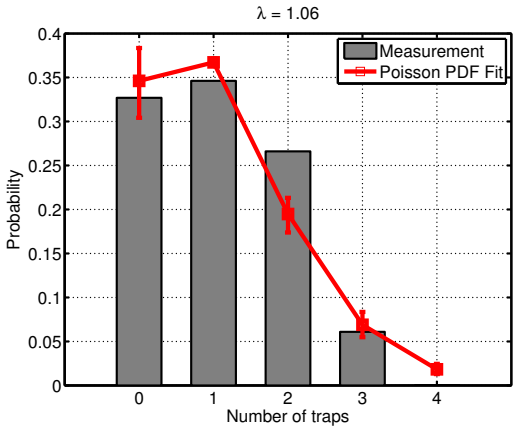
Fig. 5.16 shows a number of examples of measured PDFs for N_T across the studied population along with the corresponding fits to Eq. 5.9. As expected, the Poisson distribution gives an accurate representation of the statistics of N_T.

Fig. 5.17 shows λ plotted across the measurement bias range for all measured device types. When comparing NMOS to PMOS devices, in every instance, the PMOS devices exhibit a higher average number of traps, which is consistent with the results reported in [109]. Additionally, regular- (RVT) and high-V_{th} (HVT) devices exhibit approximately the same number of traps, which is also consistent with [109].

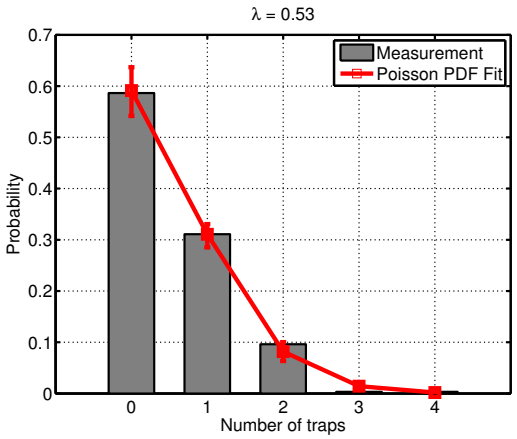
What is more intriguing, however, is the scaling behavior of λ with device dimensions, W and L. Based on data from both NMOS and PMOS measurements, it appears that λ is largely independent of W, and is inversely proportional to L. This point is exemplified in Fig. 5.18, where the average λ across bias is plotted agains the inverse of the effective gate length, $1/(L - \Delta L)$.



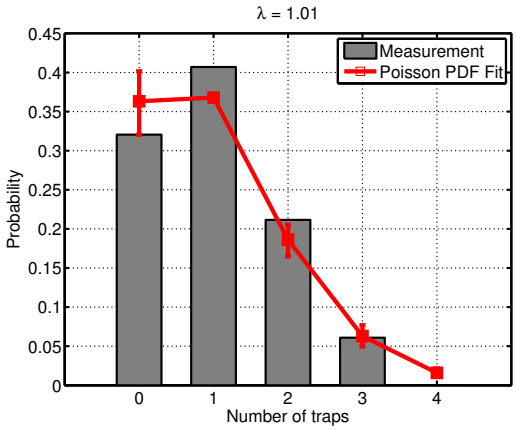
(a) RVT NMOS, $W/L = 0.6/0.04\ \mu m$, $\lambda = 0.49$



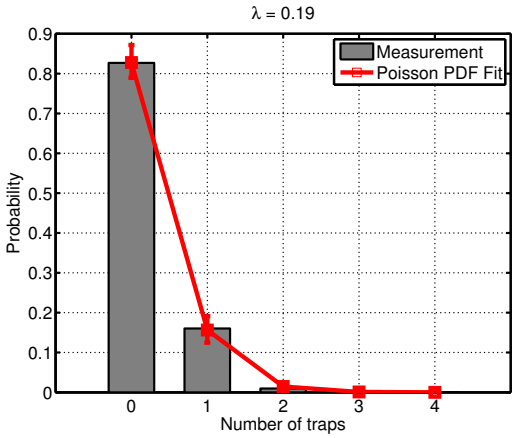
(b) RVT PMOS, $W/L = 0.6/0.04\ \mu m$, $\lambda = 1.06$



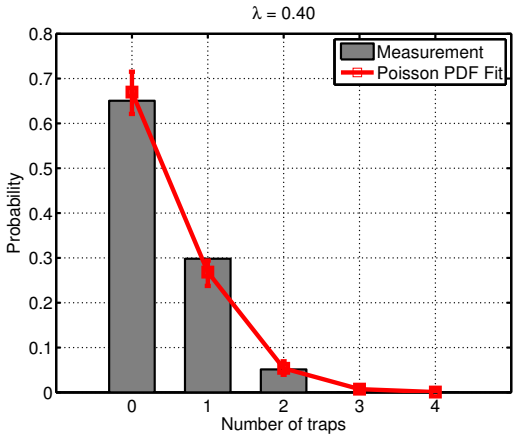
(c) HVT NMOS, $W/L = 0.6/0.04\ \mu m$, $\lambda = 0.53$



(d) HVT PMOS, $W/L = 0.6/0.04\ \mu m$, $\lambda = 1.01$

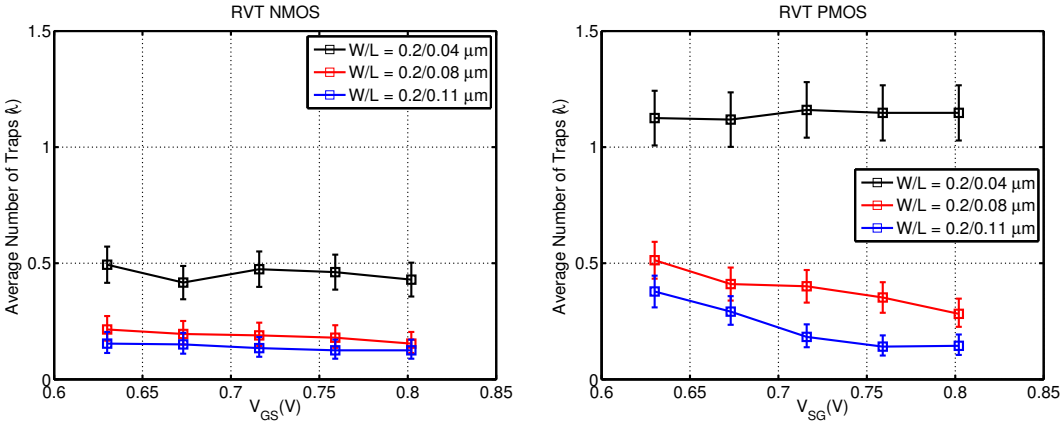


(e) RVT NMOS, $W/L = 0.2/0.08\ \mu m$, $\lambda = 0.19$



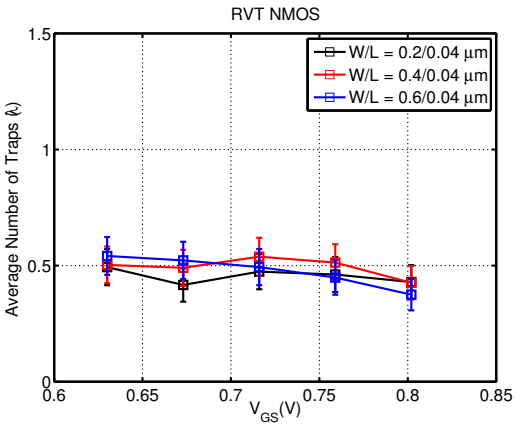
(f) RVT PMOS, $W/L = 0.2/0.08\ \mu m$, $\lambda = 0.40$

Figure 5.16: Measured PDFs for N_T and the corresponding Poisson fits for a number of different device types and sizes at mid-bias, along with the extracted values of λ; in each case the Poisson PDF (Eq.5.9) fits the measured data well.

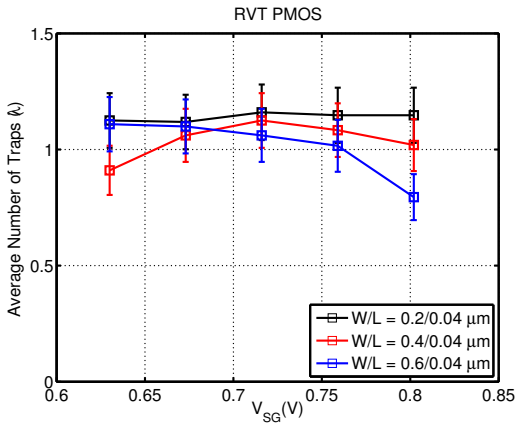


(a) RVT NMOS, min W

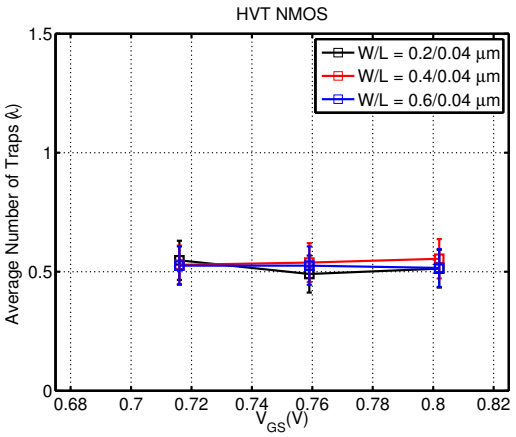
(b) RVT PMOS, min W



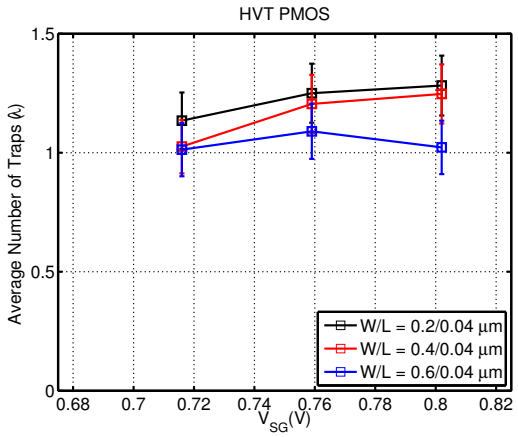
(c) RVT NMOS, min L



(d) RVT PMOS, min L



(e) HVT NMOS, min L



(f) HVT PMOS, min L

Figure 5.17: Extracted values of λ across bias for RVT minimum width devices, RVT minimum length devices, and HVT minimum length devices; λ appears to be relatively independent of bias and doping, and to scale inversely to L, but remain independent of W.

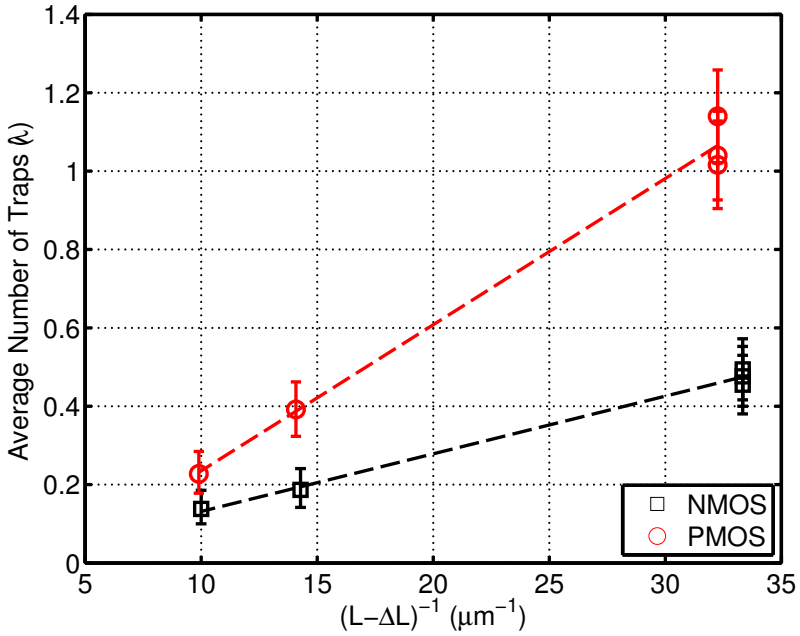


Figure 5.18: A plot of the mean λ measured across bias vs. the inverse of the effective channel length given by $1/(L - \Delta L)$.

Theoretical analysis [105] predicts that the average number of traps should be proportional to the area under the gate, given by WL, rather than scale inversely with the effective length. However, such analysis ignores the issue of observability, and in particular, the fact that RTN is not observed in large-area devices due to reduced single-trap amplitude and the tendency of multiple RTN signals to combine and form $1/f$ noise [97]. As a result, while the actual number of traps present in a device can grow proportionally to the gate area, it is still possible for the average number of observed RTN traps, λ, to scale inversely with the effective channel length, as observed here.

5.4.2 Statistics of Single-Trap Amplitude Fluctuations, ΔV_{th}

Log-Normal vs. Exponential Distribution of ΔV_{th}

Accurately characterizing the statistical distribution of single-trap RTN amplitude fluctuations is vital for constructing an accurate compact model for the prediction of overall ΔV_{th} fluctuations due to RTN. While results in the literature, both based on device simulation and experimental measurements, indicate that the distribution of single-trap ΔV_{th}

is skewed and exhibits a fat tail, there is disagreement on which distribution captures the statistical effects best. In particular, two distributions are considered: the exponential distribution [17, 92, 105, 109], given by

$$f_e(\Delta V_{th}; \sigma_e) = \frac{1}{\sigma_e} e^{-\frac{\Delta V_{th}}{\sigma_e}}, \tag{5.10}$$

where σ_e is a parameter, which represents the population mean, and the log-normal distribution [103, 117], given by

$$f_l(\Delta V_{th}; V_{th0}, \sigma_l) = \frac{1}{\sigma_l \Delta V_{th} \sqrt{2\pi}} e^{-\frac{1}{2\sigma_l^2}(\ln \Delta V_{th} - \ln V_{th0})^2}, \tag{5.11}$$

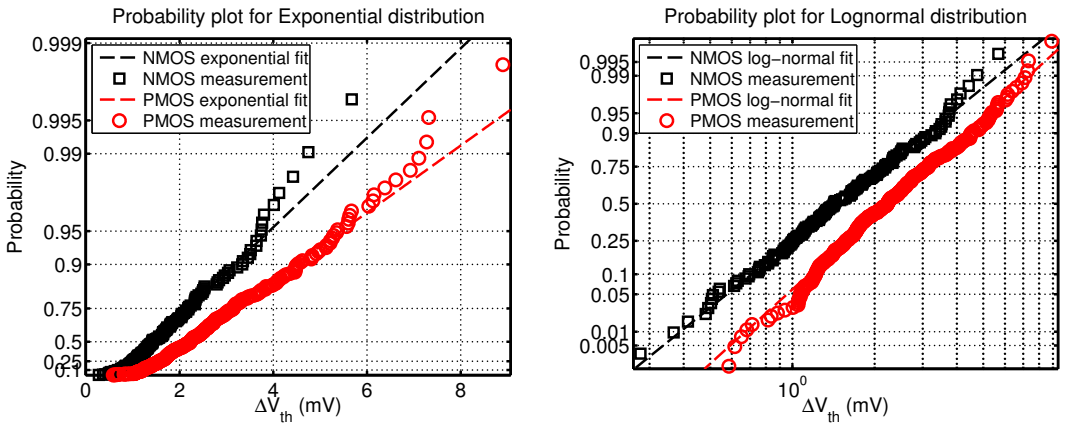
where σ_l is a dimensionless parameter representing the lognormal shape, and V_{th0} is given by

$$V_{th0} = e^{\mu}, \tag{5.12}$$

with μ representing the mean of the distribution of $\ln(\Delta V_{th})$.

While the exponential distribution is more commonly used for the modeling of the statistics of single-trap RTN amplitude, the log-normal distribution yields a better fit to the measured data in this work. Fig. 5.19 shows examples of representative single-trap ΔV_{th} distributions for an NMOS and a PMOS device fit to both an exponential and a log-normal distribution. Both distributions appear to offer acceptable fits to the measured data, but in either case, the log-normal distribution is better at modeling the tails of the observed statistical distribution. In particular, compared to the log-normal fit, the exponential fit over-predicts both the low end and the high end of the measured distribution.

The choice of the log-normal distribution as discussed above is largely made on empirical grounds. Fig. 5.19 demonstrates that the log-normal distribution simply fits the data better then the exponential distribution. However, Sonoda, et. al. [117] also offer some theoretical basis as to why the log-normal distribution might be an appropriate choice. The argument is that as a result of the random distribution of dopant atoms in the channel, traps with different positions along the channel have a different impact on the overall V_{th}



(a) Exponential (b) Log-Normal

Figure 5.19: Representative measured ΔV_{th} distributions for a $W/L = 0.4\ \mu m/0.04\ \mu m$ NMOS and PMOS device fitted using (a) an exponential PDF and (b) a log-normal PDF (dashed lines represent ideal fits); in either case, a log-normal PDF more accurately models the tails of the measured distribution.

of the device, based on the difference in the underlying local dopant profile. In particular, in the absence of random dopant fluctuations (RDF), ΔV_{th} is given by

$$\Delta V_{th} = \frac{q}{C'_{ox} W_{eff} L_{eff}}, \tag{5.13}$$

where q is an elementary charge, C'_{ox} is the gate oxide per unit area, and $W_{eff}L_{eff}$ is the effective area under the gate. However, once RDF effects are included in the derivation, an additional term accounting for the effect of uneven V_{th} distribution across the channel needs to be added, which results in

$$\Delta V_{th} = \frac{q}{C'_{ox} W_{eff} L_{eff}} e^{\frac{q}{nkT}(V_{th} - V_{th,j})}, \tag{5.14}$$

where n is the sub-threshold factor, k is the Boltzman constant, T is the absolute temperature, and $V_{th,j}$ is the threshold voltage in the vicinity of the oxide trap. If it is assumed that $V_{th,j}$ is normally distributed, then ΔV_{th} is expected to have a log-normal distribution by definition, since according to Eq. 5.14, $\ln(\Delta V_{th}) \propto V_{th,j}$. For more details on the derivation above, as well as numerical simulation results demonstrating its validity, the reader is referred to [117].

$E[\Delta V_{th}]$ Across Bias, Geometry, and Device Type

In order to gain some understanding of the statistical behavior of ΔV_{th} due to individual traps, it is instructive to consider how the expected value of ΔV_{th}, given by

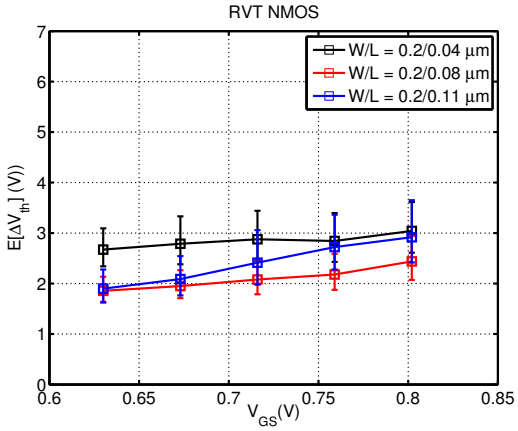
$$E[\Delta V_{th}] = V_{th0} e^{\sigma^2/2}. \tag{5.15}$$

behaves across bias, geometry, and device type. Fig. 5.20 shows the measured results.

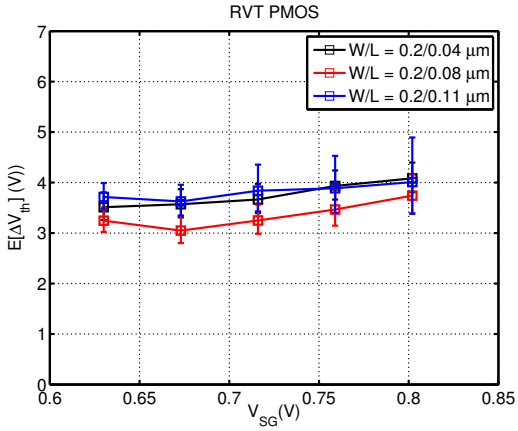
$E[\Delta V_{th}]$ is essentially constant across bias, which shows that modeling RTN as a ΔV_{th} effect is well-justified. Even if there is still some uncertainty as to whether this is the best representation from a theoretical standpoint, from a purely practical standpoint, the low bias sensitivity makes the resulting model significantly less complex. Similar insensitivity of $E[\Delta V_{th}]$ to bias conditions has also been reported in [103]. PMOS devices exhibit higher single-trap amplitudes as compared to NMOS devices, which is also reported in [105].

In terms of area dependance, $E[\Delta V_{th}]$ is inversely related to both W and L, as Eq. 5.14 suggests. However, the inverse relationship is not purely linear. Such an observation is also made in [92], where $E[\Delta V_{th}]$ is said to be proportional to $W^{-1}L^{-0.5}$. The stronger dependance on W is due to the so-called "percolation effect", which states that the increased field along the STI edge of the device causes charge carrier concentration to be higher along the length of the device. As W is decreased, a larger portion of the overall I_D is contributed by current flowing close to the STI edge of the device. As a result, a larger effective ΔV_{th} variation can be expected on average due to the comparatively larger contribution of traps along the percolation paths. The current-crowding effects giving rise to greater ΔV_{th} fluctuations due to traps close to the gate edges along the length of the device are also discussed in [105].

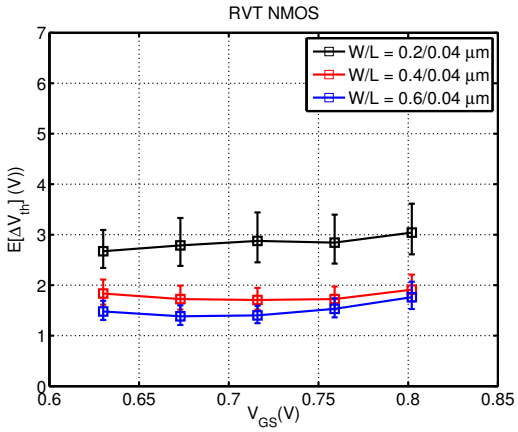
Fig. 5.21 shows that a proportionality of $E[\Delta V_{th}]$ to $W^{-1}L^{-0.5}$ reasonably fits the measured data for both NMOS and PMOS devices. The only exception is devices with $L = 0.11\ \mu m$, which seem to reverse the trend. However, a very small average number of traps is observed in these devices, as discussed in Section 5.4.1, and it is possible that the



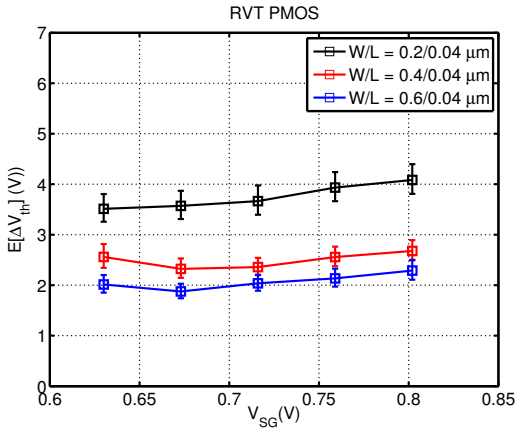
(a) RVT NMOS, min W



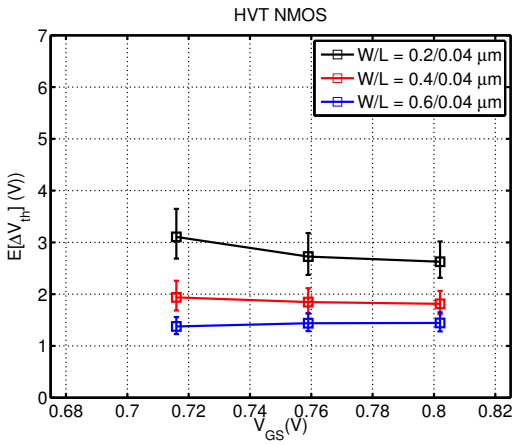
(b) RVT PMOS, min W



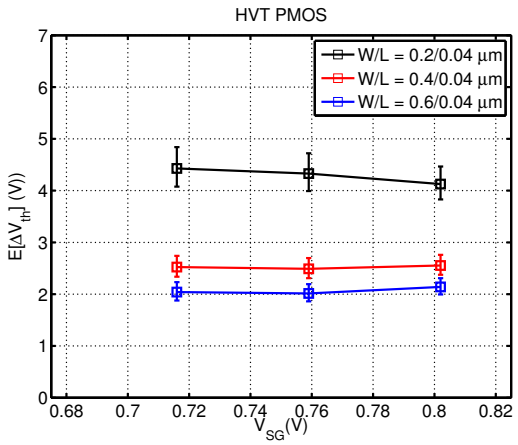
(c) RVT NMOS, min L



(d) RVT PMOS, min L



(e) HVT NMOS, min L



(f) HVT PMOS, min L

Figure 5.20: Extracted values of $E[\Delta V_{th}]$ across bias for RVT minimum width devices, RVT minimum length devices, and HVT minimum length devices; $E[\Delta V_{th}]$ appears to be relatively independent of bias and doping, and to scale inversely with both W and L, with a stronger coupling to W then to L.

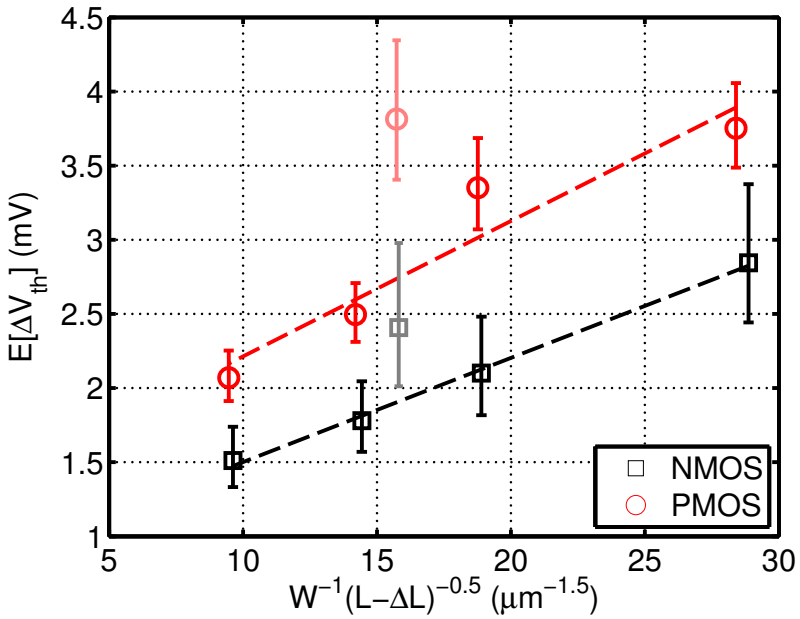


Figure 5.21: A plot of the mean $E[\Delta V_{th}]$ measured across bias vs. $W^{-1}(L-\Delta L)^{-0.5}$; dashed lines indicate trend lines; faded data points correspond to devices with $L = 0.11\ \mu m$, which do not follow the observed trend, possibly due to an insufficient number of samples.

statistics of single-trap amplitudes are inaccurate simply because of the smaller available sample set.

5.4.3 Complex CDF Model for Overall RTN Fluctuations

The final step of constructing a statistical model for overall ΔV_{th} fluctuations (referred to, from this point on, as ΔV_{th}^{*} to distinguish them from single-trap ΔV_{th} fluctuations) is to combine the statistics of number of traps, N_T, with the statistics of single-trap amplitudes, ΔV_{th}, into one comprehensive statistical model. Assuming that the effects of individual traps are additive, i.e. individual trapping and de-trapping events are independent of one another and act in superposition, then the PDF for n observed traps can be derived using the successive convolution of n single-trap distributions [105]. Using the log-normal PDF to model the statistics of single-trap amplitude (Eq. 5.11), we can express the PDF of a system of n traps as

$$f_{l,n}(\Delta V_{th}; V_{th0}, \sigma_l, n) = \int_{-\infty}^{\infty} f_{l,n}(\Delta V_{th} - u; V_{th0}, \sigma_l, n-1) f_l(u; V_{th0}, \sigma_l)\, \mathrm{d}u. \qquad (5.16)$$

The relative contribution of an RTN system with n observed traps to the overall distribution of ΔV_{th}^* can be derived from the Poisson distribution of N_T given by Eq. 5.9 as

$$a_n = P(N_T = n) = \frac{\lambda^n e^{-\lambda}}{n!}. \tag{5.17}$$

Finally, the two statistics can be combined by multiplying each a_n coefficient by the corresponding $f_{l,n}(\Delta V_{th}; V_{th0}, \sigma_l, n)$; a delta function, $\delta_0(x)$, is used to represent the distribution of devices with no traps. The products are summed up as n goes to infinity to give $f_c(\Delta V_{th}^*; V_{th0}, \sigma_l, \lambda)$, the overall PDF of ΔV_{th}^*, as

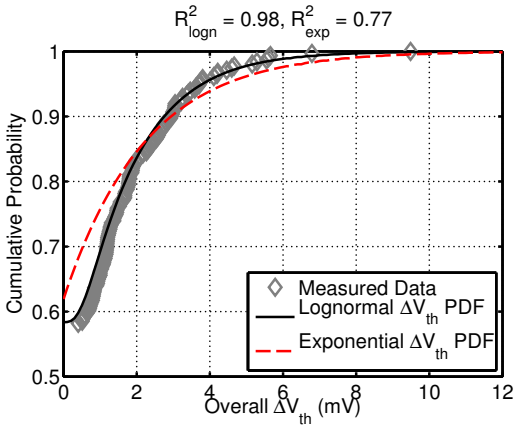
$$f_c(\Delta V_{th}^*; V_{th0}, \sigma_l, \lambda) = a_0 \delta_0(\Delta V_{th}^*) + \sum_{i=1}^{\infty} a_i f_{l,n}(\Delta V_{th}^*; V_{th0}, \sigma_l, i). \tag{5.18}$$

Eq. 5.18 can be used to derive the cumulative distribution function (CDF) of ΔV_{th}^* as given by

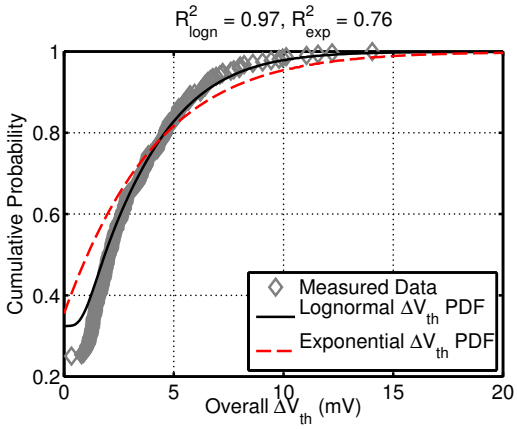
$$F_c(\Delta V_{th}^*; V_{th0}, \sigma_l, \lambda) = \int_0^{\Delta V_{th}^*} f_c(x; V_{th0}, \sigma_l, \lambda)\, \mathrm{d}x. \tag{5.19}$$

Fig. 5.22 shows example CDFs derived using parameter values extracted from the measured distributions of N_T and ΔV_{th}, as described in Section 5.4.1 and Section 5.4.2, respectively. In order to underscore the importance of using a log-normal distribution to model ΔV_{th} (Eq. 5.11), in contrast to an exponential one (Eq. 5.10), as proposed in [105], both treatments are considered. The modeled CDFs are compared to the actual measured ΔV_{th}^* distributions. R^2 values are calculated in order to help evaluate the goodness of the fit, where R^2 is the coefficient of determination; the closer the value of R^2 is to 1, the better the fit.

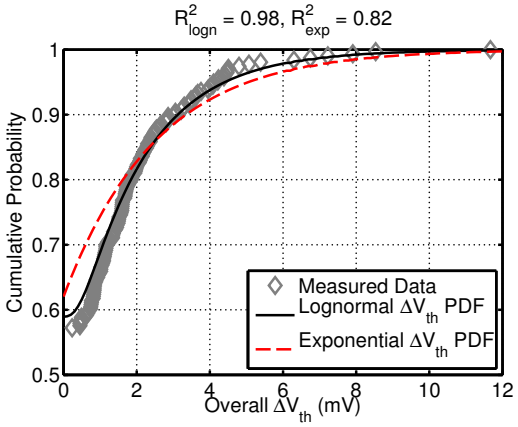
In all cases, when a log-normal PDF is used to model the single-trap ΔV_{th} (solid black lines in Fig. 5.22), the fits are excellent, with a mean R^2 value of 0.97 across all samples. On the other hand, using an exponential PDF for ΔV_{th} (dashed red lines in Fig. 5.22) results in much poorer fits to the measured data, with a mean R^2 of 0.73 across all samples. This comparison once again demonstrates that the statistics of single-trap amplitude are better modeled by a log-normal distribution, and that the exponential distribution, otherwise popular in the literature, does an inferior job at modeling measured data, especially when considering the tails of the distribution.



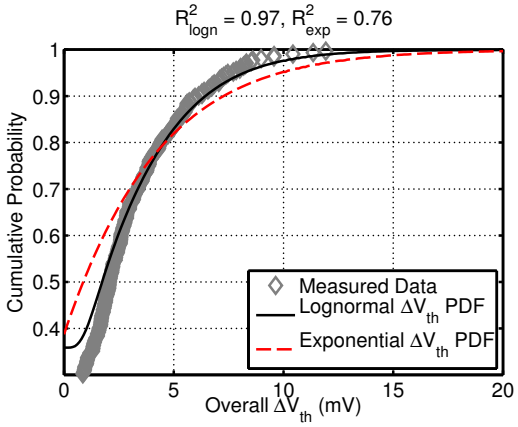
(a) RVT NMOS, $W/L = 0.4\ \mu m/0.04\ \mu m$



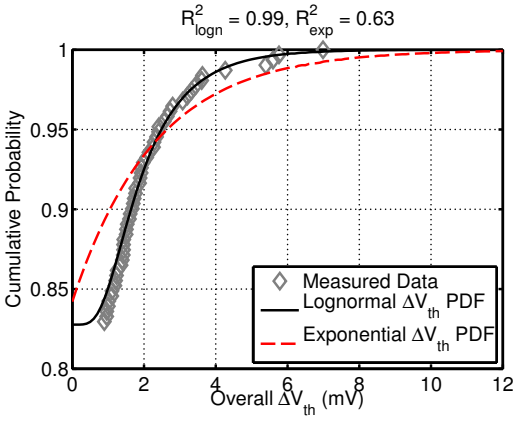
(b) RVT PMOS, $W/L = 0.4\ \mu m/0.04\ \mu m$



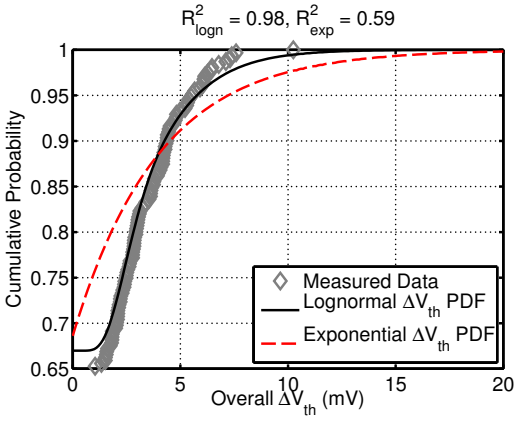
(c) HVT NMOS, $W/L = 0.4\ \mu m/0.04\ \mu m$



(d) HVT PMOS, $W/L = 0.4\ \mu m/0.04\ \mu m$

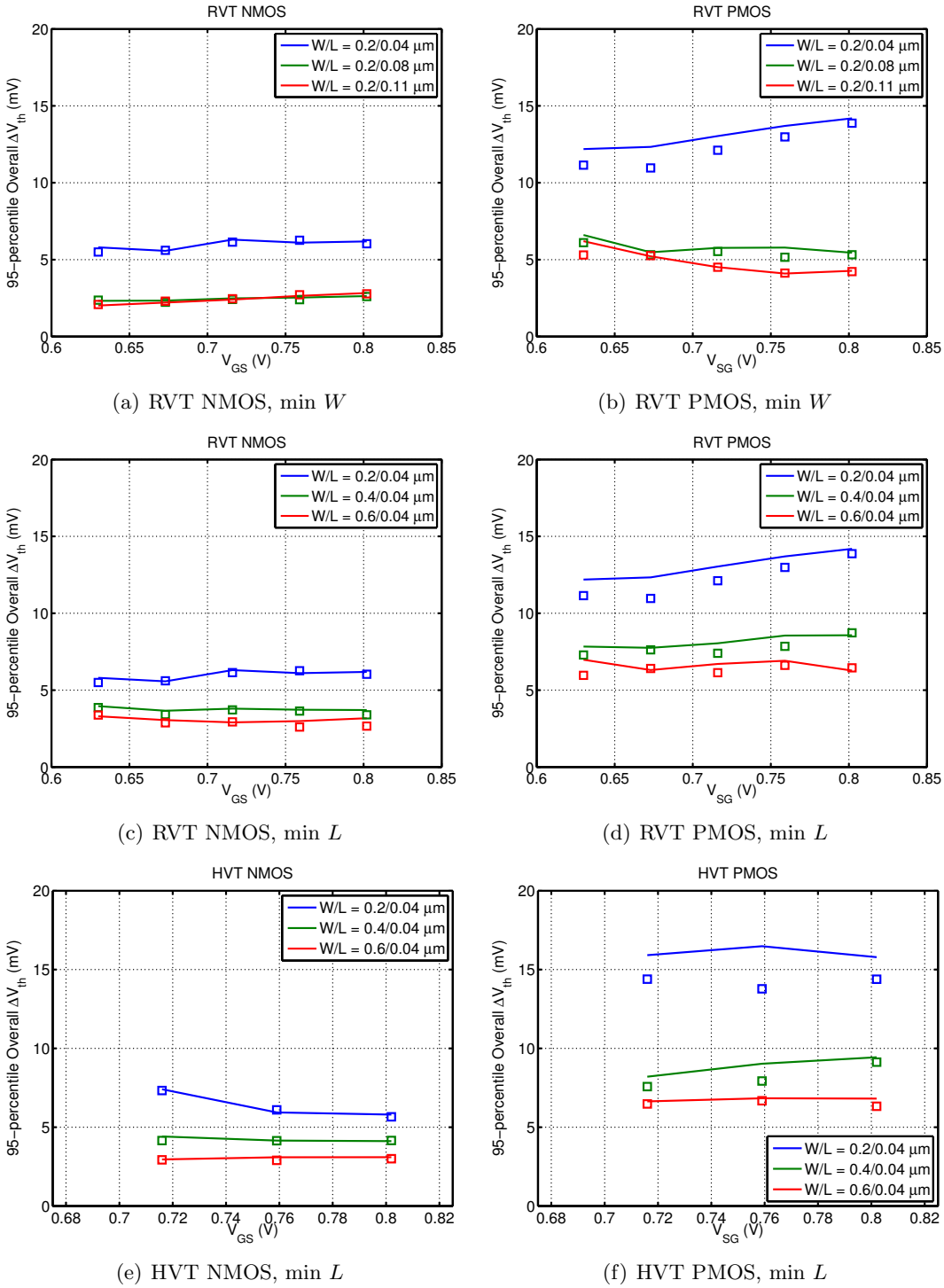


(e) RVT NMOS, $W/L = 0.2\ \mu m/0.08\ \mu m$



(f) RVT PMOS, $W/L = 0.2\ \mu m/0.08\ \mu m$

Figure 5.22: Representative fits of measured overall ΔV_{th}^* to Eq. 5.19 across device type and geometry using a log-normal single-trap amplitude PDF (solid black line) and an exponential single-trap amplitude PDF (dashed red line); R^2 values quoted for both cases above the individual graphs show that using a log-normal PDF yields a considerably better fit to the measured data.



(a) RVT NMOS, min W

(b) RVT PMOS, min W

(c) RVT NMOS, min L

(d) RVT PMOS, min L

(e) HVT NMOS, min L

(f) HVT PMOS, min L

Figure 5.23: 95-percentile measurements (squares) and predictions (solid lines) for the overall ΔV_{th} across device type, bias, and geometry; excellent agreement between measurement and prediction even in the tail of the distribution is demonstrated.

The fact that the measured CDFs of ΔV_{th}^{*} fit well the estimated CDFs calculated based on extracted parameters for N_T and ΔV_{th} is a solid proof that the overall statistical model developed in this work gives an accurate representation of the statistical behavior of RTN and can be used to predict the total impact of RTN with high confidence even at the tails of the distribution. In order to further underscore this point, Fig. 5.23 shows a comparison between measured and predicted ΔV_{th}^{*} across the entire sample set at the 95-percentile level. In all cases, the agreement between prediction and measurement is excellent.

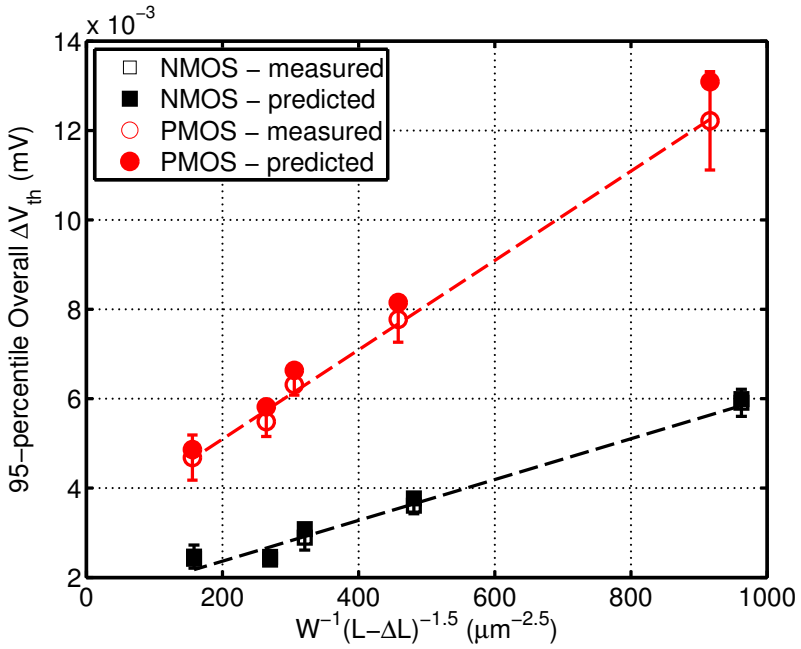


Figure 5.24: 95-percentile measurements (unfilled markers) and predictions (filled markers) for the overall ΔV_{th}^{*} from NMOS and PMOS devices; excellent agreement between measurement and prediction even in the tail of the distribution is demonstrated and a scaling trend inversely proportional to $W(L-\Delta L)^{1.5}$ is observed.

A scaling trend of the 95-percentile of ΔV_{th}^{*} proportional to $W^{-1}(L-\Delta L)^{-1.5}$ is observed in Fig. 5.24. This can be traced to the scaling of the number of traps, N_T, with $(L-\Delta L)^{-1}$, as shown in Fig. 5.18, and the scaling of ΔV_{th} with $W^{-1}(L-\Delta L)^{-0.5}$, as shown in Fig. 5.21. The fact that the tail of the overall ΔV_{th}^{*} scales with $W^{-1}(L-\Delta L)^{-1.5}$ indicates that the comparative impact of RTN on V_{th} is expected to worsen with device scaling in relation to the effect of random dopant fluctuations (RDF), where the tails of the

distribution scale with $W^{-0.5}(L-\Delta L)^{-0.5}$ [39], as discussed in Chapter 4.

Comparing the overall RTN magnitude in NMOS and PMOS devices, PMOS devices tend to exhibit considerably higher RTN, which can be traced to both a higher number of observed traps and a larger single-trap ΔV_{th}. This result is interesting, since in terms of the RDF effects on V_{th}, PMOS devices generally exhibit less variation than NMOS devices [57]. Therefore, it would be expected that as the device dimensions scale with new technology nodes and the comparative effect of RTN grows, PMOS devices would be affected more dramatically than NMOS devices.

5.5 Conclusion

The on-chip variability characterization system of Chapter 3 is successfully used for direct time-domain measurements of random telegraph noise is small-area devices. A procedure for the automated extraction of RTN parameters from large volumes of measured data is developed and verified. The statistics of number of traps, N_T, and single-trap amplitudes, ΔV_{th}, are studied across device polarity, bias, and gate area. A Poisson distribution is used to model N_T and a log-normal distribution is used to model ΔV_{th}. The scaling of the two sets of statistics across gate dimensions is discussed; the expected value of N_T is shown to scale with $(L-\Delta L)^{-1}$, whereas the expected value of ΔV_{th} is shown to scale with $W^{-1}(L-\Delta L)^{-0.5}$. The statistics of the two RTN parameters are combined in a compact RTN probabilistic model representing the statistics of the overall V_{th} fluctuations due to RTN. This model is demonstrated to give accurate predictions of the tails of the measured RTN distributions at the 95-percentile level, which scale with $W^{-1}(L-\Delta L)^{-1.5}$. A comparison between NMOS and PMOS devices shows that PMOS devices exhibit both a higher average number of traps and a larger average single-trap ΔV_{th} amplitude, leading to a comparatively larger overall impact of RTN.

Chapter 6

Conclusion

The design of an on-chip system for transistor variability characterization implemented in a 45-nm low-power bulk CMOS process has been presented. Complete on-chip system integration has been achieved, including both capacitance-voltage (C-V) and current-voltage (I-V) characterization capability for devices with circuit-representative geometries. The functionality of the system has been demonstrated through a detailed study of random and systematic quasi-static device variability using a novel combined C-V/I-V characterization methodology. Additionally, the effects of random telegraph noise (RTN) have been studied through direct measurement of time-domain current waveforms using the on-chip system. A compact statistical model for predicting the overall impact of RTN on the performance of small-area devices has been developed and verified.

6.1 Summary of contributions

This thesis contains the following original contributions:

- The first fully-integrated on-chip combined C-V/I-V characterization system is presented; the system functionality is validated in the study of quasi-static device variability and statistics of random telegraph noise.
- A leakage- and parasitics-insensitive charge-based capacitance measurement (CBCM)

technique capable of C-V characterization of circuit-representative 45-nm CMOS devices with atto-Farad resolution is developed, implemented, and validated.

- The variation in the gate-to-channel capacitance of circuit-representative devices in a 45-nm process across gate dimensions is reported for the first time; the variation at this scale is shown to be dominated by line-edge-roughness (LER) rather than variations in the gate oxide; LER statistical parameters are extracted from C-V data for the first time.

- Correlation between information gathered based on C-V and I-V measurements on the same set of devices is used to identify a systematic gradient in the effective channel length (L_{eff}) across the reticle; such analysis is uniquely enabled by the presented combined C-V/I-V characterization methodology.

- A fully automated methodology for the extraction of number of traps and individual trap amplitudes from time-domain RTN measurements using an enhanced time-lag-plot (TLP) is presented and verified; statistics of both parameters are modeled across device polarity and geometry.

- A compact model for the statistical prediction of overall RTN amplitude is shown to yield more accurate predictions if a log-normal rather than exponential distribution is used for modeling the distribution of single-trap amplitudes.

Several peer-reviewed publications have resulted from the original contributions contained in this work. These include:

- S. Realov, W. McLaughlin, and K. L. Shepard, "On-chip transistor characterization arrays with digital interfaces for variability characterization," *Proceedings of the 2009 IEEE International Symposium on Quality Electronic Design (ISQED)*, September 2009, pp. 167-171.

- S. Realov and K. L. Shepard, "Random telegraph noise in 45-nm CMOS: Analysis

using an on-chip test and measurement system," *Proceedings of the 2010 IEEE International Electron Devices Meeting (IEDM)*, December 2010, pp. 28.2.1-28.2.4.

- S. Realov and K. L. Shepard, "On-chip combined C-V/I-V transistor characterization system in 45-nm CMOS," *Proceedings of the IEEE 2011 Symposium on VLSI Circuits (VLSIC)*, June 2011, pp. 218-219.

6.2 Future Work

The basic approach to modeling variability in the quasi-static electrical device characteristics presented in Chapter 4 can be extended to cover a complete industry-standard BSIM device model. I-V device characterization can be extended beyond the linear region of operation to cover variability in all regions of operation. The C-V characterization methodology can be extended to decouple the gate capacitance into its individual components, which can then be characterized across device bias. Additionally, the system can be used to characterize variability in the back-end-of-line (BEOL) electrical performance, including device contact resistance and capacitive coupling of metal interconnect, for a truly comprehensive study of the impact of variability on circuit performance. Ring-oscillator structures can be used to verify the relationship between variation observed in the electrical parameters of individual devices and variation in the performance of circuits implemented in the characterized technology; integration on the same chip would ensure tight coupling between the two characterization structures.

In the context of RTN characterization, the system can be used to develop a stress-based characterization approach and compare the results to those using the direct measurement approach. This stress-based approach can be used to examine the effects of bias-temperature-instability (BTI) and the relationship of BTI to RTN. These results can also be coupled with results from ring-oscillator test structures where effects of device stress are measured as variations in the oscillation frequency.

Bibliography

[1] J. Lilienfeld, "Method and Apparatus for Controlling Electric Currents," no. 1745175, Jan. 1930.

[2] D. Kahng, "Electric Field Controlled Semiconductor Device," no. 3102230, May 1960.

[3] R. Noyce, "Semiconductor Device-and-Lead Structure," no. 2981877, Jul. 1959.

[4] F. Wanlass and C. Sah, "Nanowatt logic using field-effect metal-oxide semiconductor triodes," in *Solid-State Circuits Conference. Digest of Technical Papers. 1963 IEEE International*, 1963, pp. 32–33.

[5] G. Moore, "Cramming More Components Onto Integrated Circuits," in *Proceedings of the IEEE*, 1998, pp. 82–85.

[6] Wgsimon. Transistor Count and Moore's Law. [Online]. Available: http://en.wikipedia.org/wiki/File:Transistor_Count_and_Moore's_Law_-_2011.svg

[7] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, Oct. 1974.

[8] P. Chatterjee, W. Hunter, T. Holloway, and Y. Lin, "The impact of scaling laws on the choice of n-channel or p-channel for MOS VLSI," *Electron Device Letters, IEEE*, vol. 1, no. 10, pp. 220–223, 1980.

[9] T. Fischer, S. Arekapudi, E. Busta, C. Dietz, M. Golden, S. Hilker, A. Horiuchi, K. Hurd, D. Johnson, H. McIntyre, S. Naffziger, J. Vinh, J. White, and K. Wilcox, "Design solutions for the Bulldozer 32nm SOI 2-core processor module in an 8-core CPU," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, 2011, pp. 78–80.

[10] M. Yuffe, E. Knoll, M. Mehalel, J. Shor, and T. Kurts, "A fully integrated multi-CPU, GPU and memory controller 32nm processor," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, 2011, pp. 264–266.

[11] W. Shockley, "Problems related to p-n junctions in silicon," *Solid-State Electronics*, vol. 2, pp. 35–60, Jan. 1961.

[12] R. Keyes, "Effect of randomness in the distribution of impurity ions on FET thresholds in integrated electronics," *Solid-State Circuits, IEEE Journal of*, vol. 10, no. 4, pp. 245–247, 1975.

[13] K. Bernstein, D. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer, "High-performance CMOS variability in the 65-nm regime and beyond," *IBM Journal of Research and Development*, vol. 50, no. 4, pp. 433–449, 2006.

[14] D. K. Schroder and J. A. Babcock, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing," *Journal of Applied Physics*, vol. 94, no. 1, pp. 1–18, Jul. 2003.

[15] V. Reddy, A. Krishnan, A. Marshall, J. Rodriguez, S. Natarajan, T. Rost, and S. Krishnan, "Impact of negative bias temperature instability on digital circuit reliability," *Microelectronics Reliability*, vol. 45, no. 1, pp. 31–38, 2005.

[16] S. Mahapatra, P. Bharathkumar, and M. A. Alam, "Investigation and Modeling of Interface and Bulk Trap Generation During Negative Bias Temperature Instability of p-MOSFETs," *IEEE Transactions on Electron Devices*, vol. 51, pp. 1371–1379, Sep. 2004.

[17] K. Fukuda, Y. Shimizu, K. Amemiya, M. Kamoshida, and C. Hu, "Random telegraph noise in flash memories-model and technology scaling," *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pp. 169–172, 2007.

[18] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, P. Roussel, and M. Nelhiebel, "Recent advances in understanding the bias temperature instability," *Electron Devices Meeting (IEDM), 2010 IEEE International*, p. 4, 2010.

[19] K. Ralls, W. Skocpol, L. Jackel, R. Howard, L. Fetter, R. Epworth, and D. Tennant, "Discrete Resistance Switching in Submicrometer Silicon Inversion Layers: Individual Interface Traps and Low-Frequency (1/f?) Noise," *Physical Review Letters*, vol. 52, no. 3, pp. 228–231, Jan. 1984.

[20] S. K. Saha, "Modeling Process Variability in Scaled CMOS Technology," *IEEE Design & Test of Computers*, vol. 27, no. 2, pp. 8–16, 2010.

[21] L.-T. Pang and B. Nikolic, "Measurements and Analysis of Process Variability in 90 nm CMOS," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 5, pp. 1655–1663, 2009.

[22] M. Orshansky, L. Milor, and C. Hu, "Characterization of spatial intrafield gate CD variability, its impact on circuit performance, and spatial mask-level correction," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 17, no. 1, pp. 2–11, 2004.

[23] H. Aikawa, E. Morifuji, T. Sanuki, T. Sawada, S. Kyoh, A. Sakata, M. Ohta, H. Yoshimura, T. Nakayama, M. Iwai, and F. Matsuoka, "2008 Symposium on VLSI Technology," in *2008 Symposium on VLSI Technology*. IEEE, 2008, pp. 90–91.

[24] N. Wils, H. Tuinhout, and M. Meijer, "Characterization of STI Edge Effects on CMOS Variability," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 22, no. 1, pp. 59–65, 2009.

[25] R. Salem, A. ElMously, H. Eissa, M. Dessouky, and M. Anis, "A DFM tool for analyzing lithography and stress effects on standard cells and critical path performance in 45nm digital designs," *Design and Test Workshop (IDT), 2010 5th International*, pp. 13–17, 2010.

[26] A. Bansal, A. Singhee, E. Acar, and G. Costrini, "Electrical monitoring of gate and active area mask misalignment error," *VLSI Circuits (VLSIC), 2011 Symposium on*, pp. 220–221, 2011.

[27] K. Kuhn, C. Kenyon, A. Kornfeld, M. Liu, A. Maheshwari, W. Shih, S. Sivakumar, G. Taylor, P. VanDerVoorn, and K. Zawadski, "Managing Process Variation in Intel's 45nm CMOS Technology," *Intel Technology Journal*, vol. 12, no. 01, pp. 93–109, Feb. 2008.

[28] R. Heald and P. Wang, "Variability in sub-100nm SRAM designs," in *Computer Aided Design, 2004. ICCAD-2004. IEEE/ACM International Conference on*, 2004, pp. 347–352.

[29] P. Zuber, M. Miranda, M. Bardon, S. Cosemans, P. Roussel, P. Dobrovolny, T. Chiarella, N. Horiguchi, A. Mercha, T. Hoffmann, D. Verkest, and S. Biesemans, "Variability and technology aware SRAM Product yield maximization," *VLSI Technology (VLSIT), 2011 Symposium on*, pp. 222–223, 2011.

[30] C. Mezzomo, A. Bajolet, A. Cathignol, R. Di Frenza, and G. Ghibaudo, "Characterization and Modeling of Transistor Variability in Advanced CMOS Technologies," *Electron Devices, IEEE Transactions on*, vol. 58, no. 8, pp. 2235–2248, 2011.

[31] Y. Cheng and E. A. Sullivan, "Effect of Coulomb scattering on silicon surface mobility," *Journal of Applied Physics*, vol. 45, no. 1, pp. 187–192, 1974.

[32] W. Zhao, F. Liu, K. Agarwal, D. Acharyya, S. R. Nassif, K. Nowka, and Y. Cao, "Rigorous extraction of process variations for 65-nm CMOS design," *IEEE Transactions on Semiconductor Manufacturing*, vol. 22, no. 1, pp. 196–203, 2009.

[33] A. Asenov, S. Kaya, and A. R. Brown, "Intrinsic parameter fluctuations in decananometer mosfets introduced by gate line edge roughness," *IEEE Transactions on Electron Devices*, vol. 50, no. 5, pp. 1254–1260, May 2003.

[34] A. Diebold, "The ITRS metrology roadmap," *Semiconductor Device Research Symposium, 2009. ISDRS '09. International*, pp. 1–2, 2009.

[35] S. Pidin, T. Mori, K. Inoue, S. Fukuta, N. Itoh, E. Mutoh, K. Obkoshi, R. Nakamura, K. Kobayashi, K. Kawamura, T. Saiki, S. Fukuyama, S. Satoh, M. Kase, and K. Hashimoto, "IEDM Technical Digest. IEEE International Electron Devices Meeting, 2004." in *IEDM Technical Digest. IEEE International Electron Devices Meeting, 2004.* IEEE, 2004, pp. 213–216.

[36] R. Keyes, "Explaining strain [in silicon]," *Circuits and Devices Magazine, IEEE*, vol. 18, no. 5, pp. 36–39, 2002.

[37] X. Wang, B. Cheng, S. Roy, and A. Asenov, "Simulation of strain enhanced variability in nMOSFETs," in *Ultimate Integration of Silicon, 2008. ULIS 2008. 9th International Conference on*, 2008, pp. 89–92.

[38] C. Liu, F. Baumann, A. Ghetti, H. Vuong, C. Chang, K. Cheung, J. Colonell, W. Lai, E. Lloyd, J. Miner, C. Pai, H. Vaidya, R. Liu, and J. Clemens, "Severe thickness variation of sub-3 nm gate oxide due to Si surface faceting, poly-Si intrusion, and corner stress," *VLSI Technology, 1999. Digest of Technical Papers. 1999 Symposium on*, pp. 75–76, 1999.

[39] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *Solid-State Circuits, IEEE Journal of*, vol. 24, no. 5, pp. 1433–1439, 1989.

[40] P. G. Drennan and C. C. McAndrew, "Understanding MOSFET mismatch for analog design," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 3, pp. 450–456, Mar. 2003.

[41] B. Nikolic and L.-T. Pang, "Measurements and analysis of process variability in 90nm CMOS," in *Solid-State and Integrated Circuit Technology, 2006. ICSICT '06. 8th International Conference on*, 2006, pp. 505–508.

[42] M. Bhushan, A. Gattiker, M. Ketchen, and K. Das, "Ring oscillators for CMOS process tuning and variability control," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 19, no. 1, pp. 10–18, 2006.

[43] J.-H. Park, L.-T. Pang, K. Duong, and B. Nikolic, "Fixed- and variable-length ring oscillators for variability characterization in 45nm CMOS," in *Custom Integrated Circuits Conference, 2009. CICC '09. IEEE*, 2009, pp. 519–522.

[44] Y.-Y. Chen, C.-T. Lin, J.-N. Lee, and C.-F. Wu, "Monitoring gate and interconnect delay variations by using ring oscillators," *VLSI Design, Automation and Test (VLSI-DAT), 2011 International Symposium on*, pp. 1–4, 2011.

[45] T. Iizuka and K. Asada, "An all-digital on-chip PMOS and NMOS process variability monitor utilizing shared buffer ring and ring oscillator," *Design and Diagnostics of Electronic Circuits & Systems (DDECS), 2011 IEEE 14th International Symposium on*, pp. 115–120, 2011.

[46] R. Rao, K. Jenkins, and J.-J. Kim, "A Local Random Variability Detector With Complete Digital On-Chip Measurement Circuitry," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 9, pp. 2616–2623, 2009.

[47] K. Balakrishnan, K. Jenkins, and D. Boning, "A simple array-based test structure for the AC variability characterization of MOSFETs," *Quality Electronic Design (ISQED), 2011 12th International Symposium on*, pp. 1–6, 2011.

[48] L.-T. Pang and B. Nikolic, "Measurement and analysis of variability in 45nm strained-Si CMOS technology," in *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, 2008, pp. 129–132.

[49] A. Asenov, R. Balasubramaniam, A. R. Brown, J. Davies, and S. Saini, "Random telegraph signal amplitudes in sub 100 nm (decanano) MOSFETs: a 3D 'Atomistic' simulation study," *Electron Devices Meeting, 2000. IEDM Technical Digest. International*, pp. 279–282, 2000.

[50] A. R. Brown, A. Asenov, and J. Watling, "Intrinsic fluctuations in sub 10-nm double-gate MOSFETs introduced by discreteness of charge and matter," *IEEE Transactions On Nanotechnology*, vol. 1, no. 4, pp. 195–200, Dec. 2002.

[51] B. Cheng, S. Roy, and A. Asenov, "The impact of random doping effects on CMOS SRAM cell," in *Solid-State Circuits Conference, 2004. ESSCIRC 2004. Proceeding of the 30th European*, 2004, pp. 219–222.

[52] A. R. Brown and A. Asenov, "Capacitance fluctuations in bulk MOSFETs due to random discrete dopants," *Journal of Computational Electronics*, vol. 7, no. 3, pp. 115–118, Jan. 2008.

[53] S. Roy, A. Brown, C. Millar, and A. Asenov, "Evaluation of statistical variability in 32 and 22nm technology generation LSTP MOSFETs," *Solid-State Electronics*, vol. 53, pp. 767–772, 2009.

[54] M. Bukhori, S. Roy, and A. Asenov, "Simulation of Statistical Aspects of Charge Trapping and Related Degradation in Bulk MOSFETs in the Presence of Random Discrete Dopants," *Electron Devices, IEEE Transactions on*, vol. 57, no. 4, pp. 795–803, 2010.

[55] L. Leunissen, W. Lawrence, and M. Ercken, "Line edge roughness: experimental results related to a two-parameter model," *Microelectronic engineering*, vol. 73, pp. 265–270, 2004.

[56] K. Inoue, F. Yano, A. Nishida, and H. Takamizawa, "Dopant distributions in n-MOSFET structure observed by atom probe tomography," *Ultramicroscopy*, vol. 109, pp. 1479–1484, 2009.

[57] H. Takamizawa, Y. Shimizu, K. Inoue, T. Toyama, N. Okada, M. Kato, H. Uchida, F. Yano, A. Nishida, T. Mogami, and Y. Nagai, "Origin of characteristic variability in metal-oxide-semiconductor field-effect transistors revealed by three-dimensional atom imaging," *Applied Physics Letters*, vol. 99, no. 13, pp. 133 502–133 502, 2011.

[58] A. Prakash, Y. Kim, K. Uram, R. Finch, P. Coutu, M. Passaro, K. Davis, A. Lafond, G. May, M. Russel, M. Spinelli, B. Nehrer, H. Lam, M. Lei, S. Chiah, L. Yang, C. Wang, P. Aggrawal, H. Ye, T. Tjhie, H. Wang, N. Taraka, G. Chia, C. Cheng, P. Long, C. Leong, Y. Teo, G. Wong, and S. Lian, "Characterization of 90 nm SOI SRAM Single Cell Failure by Nano Probing Technique and TCAD Simulation," *Integrated Circuits, 2007. ISIC '07. International Symposium on*, pp. 252–254, 2007.

[59] S. O'uchi, T. Matsukawa, T. Nakagawa, K. Endo, Y. Liu, T. Sekigawa, J. Tsukada, Y. Ishikawa, H. Yamauchi, K. Ishii, E. Suzuki, H. Koike, K. Sakamoto, and M. Masahara, "Characterization of metal-gate FinFET variability based on measurements and compact model analyses," *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–4, 2008.

[60] T. Mizuno, J. Okamura, and A. Toriumi, "Experimental study of threshold voltage fluctuations using an 8k MOSFET's array," *VLSI Technology, 1993. Digest of Technical Papers. 1993 Symposium on*, pp. 41–42, 1993.

[61] 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," *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, pp. 67–68, 2006.

[62] 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," *Electron Devices, IEEE Transactions on*, vol. 55, no. 1, pp. 131–144, 2008.

[63] T. Sato, H. Ueyama, N. Nakayama, and K. Masu, "Accurate Array-Based Measurement for Subthreshold-Current of MOS Transistors," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 11, pp. 2977–2986, 2009.

[64] Z. Guo, A. Carlson, L.-T. Pang, K. Duong, T.-J. K. Liu, and B. Nikolic, "Large-Scale SRAM Variability Characterization in 45 nm CMOS," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 11, pp. 3174–3192, 2009.

[65] T. Hiramoto, M. Suzuki, X. Song, K. Shimizu, T. Saraya, A. Nishida, T. Tsunomura, S. Kamohara, K. Takeuchi, and T. Mogami, "Direct Measurement of Correlation Between SRAM Noise Margin and Individual Cell Transistor Variability by Using Device Matrix Array," *Electron Devices, IEEE Transactions on*, vol. 58, no. 8, pp. 2249–2256, 2011.

[66] V. Wang and K. Shepard, "On-chip transistor characterisation arrays for variability analysis," *Electronics Letters*, vol. 43, no. 15, pp. 806–807, 2007.

[67] N. D. Arora and L. Song, "Atto-Farad Measurement and Modeling of On-Chip Coupling Capacitance," *IEEE Electron Device Letters*, vol. 25, no. 2, pp. 92–94, Feb. 2004.

[68] E. Baruch, S. Shperber, R. Levy, Y. Weizman, J. Fridburg, and R. Marks, "A simple system for on-die measurement of atto-Farad capacitance," in *Microelectronic Test Structures (ICMTS), 2011 IEEE International Conference on*, 2011, pp. 19–21.

[69] Y.-W. Chang, H.-W. Chang, T.-C. Lu, Y.-C. King, K.-C. Chen, and C.-Y. Lu, "Combining a Novel Charge-Based Capacitance Measurement (CBCM) Technique and Split – Method to Specifically Characterize the STI Stress Effect Along the Width Direction of MOSFET Devices," *Electron Device Letters, IEEE*, vol. 29, no. 6, pp. 641–644, 2008.

[70] H. Zhao, S. Rustagi, N. Singh, F.-J. Ma, G. Samudra, K. Budhaaraju, S. Manhas, C. Tung, G. Lo, G. Baccarani, and D. Kwong, "Sub-femto-farad capacitance-voltage characteristics of single channel gate-all-around nano wire transistors for electrical characterization of carrier transport," *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–4, 2008.

[71] S. Polonsky, P. Solomon, J.-h. Liao, L. Medina, and M. Ketchen, "Front-end-of-line quadrature-clocked voltage-dependent capacitance measurement," in *Microelectronic Test Structures (ICMTS), 2011 IEEE International Conference on*, 2011, pp. 4–7.

[72] K. Tsuji, K. Terada, R. Kikuchi, T. Tsunomura, A. Nishida, and T. Mogami, "Evaluation of MOSFET C-V curve variation using test structure for charge-based capacitance measurement," in *Microelectronic Test Structures (ICMTS), 2011 IEEE International Conference on*, 2011, pp. 8–12.

[73] N. H. E. Weste, D. Harris, and A. Banerjee, *CMOS VLSI Design: A Circuits And Systems Perspective*, 3rd ed. Pearson Education, Sep. 2006.

[74] A. S. Sedra and K. C. Smith, *Microelectronic circuits*. Oxford University Press, USA, 1998.

[75] A. Mutoh and S. Nitta, "Noise immunity characteristics of dual-slope integrating analog-digital converters," *Electromagnetic Compatibility, 1999 International Symposium on*, pp. 622–625, 1999.

[76] B. Razavi, *Design of Analog CMOS Integrated Circuits*, 2002nd ed. Tata McGraw-Hill.

[77] G. D. Boreman, *Basic electro-optics for electrical engineers*. SPIE-International Society for Optical Engineering, 1998.

[78] XEM3010 User's Manual. [Online]. Available: http://www.opalkelly.com/library/XEM3010-UM.pdf

[79] Agilent Technologies 81133A and 81134A 3.35 GHz Pulse Pattern Generators. [Online]. Available: http://www.home.agilent.com/agilent/product.jspx?nid=-536902258.536882009.00&cc=US&lc=eng

[80] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures." *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.

[81] S. Realov and K. Shepard, "Random telegraph noise in 45-nm CMOS: Analysis using an on-chip test and measurement system," *Electron Devices Meeting (IEDM), 2010 IEEE International*, pp. 28.2.1–28.2.4, 2010.

[82] A. Ortiz-Conde, F. Sanchez, J. Liou, A. Cerdeira, M. Estrada, and Y. Yue, "A review of recent MOSFET threshold voltage extraction methods," *Microelectronics Reliability*, vol. 42, pp. 583–596, 2002.

[83] G. Niu, J. Cressler, S. Mathew, and S. Subbanna, "A total resistance slope-based effective channel mobility extraction method for deep submicrometer CMOS technology," *Electron Devices, IEEE Transactions on*, vol. 46, no. 9, pp. 1912–1914, 1999.

[84] J. Kim, J. Lee, I. Song, Y. Yun, J. D. Lee, B.-G. Park, and H. Shin, "Accurate Extraction of Effective Channel Length and Source/Drain Series Resistance in Ultrashort-Channel MOSFETs by Iteration Method," *IEEE Transactions on Electron Devices*, vol. 55, no. 10, pp. 2779–2784, 2008.

[85] J. Kim, M. Choi, and S. Lee, "Accuracy Analysis of Extraction Methods for Effective Channel Length in Deep-Submicron MOSFETs," *Journal of Semiconductor Technology and Science*, vol. 11, no. 2, p. 131, 2011.

[86] D. Fleury, A. Cros, K. Romanjek, D. Roy, F. Perrier, B. Dumont, H. Brut, and G. Ghibaudo, "Automatic Extraction Methodology for Accurate Measurements of Effective Channel Length on 65-nm MOSFET Technology and Below," *IEEE Transactions on Semiconductor Manufacturing*, vol. 21, no. 4, pp. 504–512, 2008.

[87] J. Koomen, "Investigation of the MOST channel conductance in weak inversion," *Solid-State Electronics*, vol. 16, pp. 801–810, 1973.

[88] Y. Tsividis and C. McAndrew, *Operation and modeling of the MOS transistor*. Oxford Univ Pr, Sep. 2010.

[89] N. Tega, H. Miki, F. Pagette, D. Frank, A. Ray, M. Rooks, W. Haensch, and K. Torii, "Increasing threshold voltage variation due to random telegraph noise in FETs as gate lengths scale to 20 nm," *VLSI Technology, 2009 Symposium on*, pp. 50–51, 2009.

[90] J. Campbell, L. Yu, K. Cheung, J. Qin, J. Suehle, A. Oates, and K. Sheng, "Large random telegraph noise in sub-threshold operation of nano-scale nMOSFETs," in *IC Design and Technology, 2009. ICICDT '09. IEEE International Conference on*, 2009, pp. 17–20.

[91] N. Tega, H. Miki, T. Osabe, A. Kotabe, K. Otsuga, H. Kurata, S. Kamohara, K. Tokami, Y. Ikeda, and R. Yamada, "Anomalously large threshold voltage fluctuation by complex random telegraph signal in floating gate flash memory," *Electron Devices Meeting, 2006. IEDM'06. International*, pp. 1–4, 2006.

[92] A. Ghetti, C. Compagnoni, F. Biancardi, A. Lacaita, S. Beltrami, L. Chiavarone, A. Spinelli, and A. Visconti, "Scaling trends for random telegraph noise in deca-nanometer Flash memories," *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–4, 2008.

[93] G. Pearson and B. Sawyer, "Silicon P-N Junction Alloy Diodes," in *Proceedings of the IRE*, 1952, pp. 1348–1351.

[94] D. Wolf and E. Holler, "Bistable Current Fluctuations in ReverseBiased pn Junctions of Germanium," *Journal of Applied Physics*, vol. 38, no. 1, pp. 189–192, 1967.

[95] P. Lauritzen, "Noise due to generation and recombination of carriers in p-n junction transition regions," *Electron Devices, IEEE Transactions on*, vol. 15, no. 10, pp. 770–776, 1968.

[96] S. Hsu, "Characterization of burst noise in silicon devices," *Solid-State Electronics*, vol. 12, pp. 867–878, Nov. 1969.

[97] H. H. Mueller and M. Schulz, "Individual interface traps at the $Si-SiO_2$ interface," *Journal of Materials Science: Materials in Electronics*, vol. 6, no. 2, pp. 65–74, 1995.

[98] C. Leyris, F. Martinez, M. Valenza, A. Hoffmann, J. C. Vildeuil, and F. Roy, "Impact of Random Telegraph Signal in CMOS Image Sensors for Low-Light Levels," in *Solid-State Circuits Conference, 2006. ESSCIRC 2006. Proceedings of the 32nd European*, 2006, pp. 376–379.

[99] M. Deen, S. Majumder, O. Marinov, and M. El-Desouki, "Random telegraph signal noise in CMOS active pixel sensors," in *Noise and Fluctuations (ICNF), 2011 21st International Conference on*, 2011, pp. 208–211.

[100] P. Fantini, A. Ghetti, A. Marinoni, G. Ghidini, A. Visconti, and A. Marmiroli, "Giant Random Telegraph Signals in Nanoscale Floating-Gate Devices," *Electron Device Letters, IEEE*, vol. 28, no. 12, pp. 1114–1116, 2007.

[101] A. Ghetti, C. Compagnoni, A. Spinelli, and A. Visconti, "Comprehensive Analysis of Random Telegraph Noise Instability and Its Scaling in Deca–Nanometer Flash Memories," *Electron Devices, IEEE Transactions on*, vol. 56, no. 8, pp. 1746–1752, 2009.

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

[103] N. Tega, H. Miki, Z. Ren, C. D'Emic, Y. Zhu, D. Frank, J. Cai, M. Guillorn, D.-G. Park, W. Haensch, and K. Torii, "Reduction of random telegraph noise in High-κ / metal-gate stacks for 22 nm generation FETs," *Electron Devices Meeting (IEDM), 2009 IEEE International*, pp. 32.4.1–32.4.4, 2009.

[104] K. Ito, T. Matsumoto, S. Nishizawa, H. Sunagawa, K. Kobayashi, and H. Onodera, "Modeling of Random Telegraph Noise under circuit operation — Simulation and measurement of RTN-induced delay fluctuation," *Quality Electronic Design (ISQED), 2011 12th International Symposium on*, pp. 1–6, 2011.

[105] K. Takeuchi, T. Nagumo, S. Yokogawa, K. Imai, and Y. Hayashi, "Single-charge-based modeling of transistor characteristics fluctuations based on statistical measurement of RTN amplitude," *VLSI Technology, 2009 Symposium on*, pp. 54–55, 2009.

[106] T. B. Watkins, "1/f Noise in Germanium Devices," in *Proceedings of the Physical Society*, Jan. 1959, pp. 59–68.

[107] F. N. Hooge, J. Kedzia, and L. K. J. Vandamme, "Boundary scattering and 1/f noise," *Journal of Applied Physics*, vol. 50, no. 12, pp. 8087–8089, 1979.

[108] K. Hung, P. Ko, C. Hu, and Y. Cheng, "A unified model for the flicker noise in metal-oxide-semiconductor field-effect transistors," *Electron Devices, IEEE Transactions on*, vol. 37, no. 3, pp. 654–665, 1990.

[109] T. Nagumo, K. Takeuchi, S. Yokogawa, K. Imai, and Y. Hayashi, "New analysis methods for comprehensive understanding of Random Telegraph Noise," *Electron Devices Meeting (IEDM), 2009 IEEE International*, pp. 1–4, 2009.

[110] S. O. Toh, Y. Tsukamoto, Z. Guo, L. Jones, T.-J. K. Liu, and B. Nikolic, "Impact of random telegraph signals on Vmin in 45nm SRAM," *Electron Devices Meeting (IEDM), 2009 IEEE International*, pp. 1–4, 2009.

[111] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989, pp. 257–286.

[112] H. Miki, M. Yamaoka, N. Tega, Z. Ren, M. Kobayashi, C. D'Emic, Y. Zhu, D. Frank, M. Guillorn, D. Park, W. Haensch, and K. Torii, "Understanding short-term BTI behavior through comprehensive observation of gate-voltage dependence of RTN in highly scaled high-κ / metal-gate pFETs," *VLSI Technology (VLSIT), 2011 Symposium on*, pp. 148–149, 2011.

[113] C. Monzio Compagnoni, R. Gusmeroli, A. Spinelli, A. Lacaita, M. Bonanomi, and A. Visconti, "Statistical Model for Random Telegraph Noise in Flash Memories," *Electron Devices, IEEE Transactions on*, vol. 55, no. 1, pp. 388–395, 2008.

[114] A. Chimenton, C. Zambelli, and P. Olivo, "A New Methodology for Two-Level Random-Telegraph-Noise Identification and Statistical Analysis," *Electron Device Letters, IEEE*, vol. 31, no. 6, pp. 612–614, 2010.

[115] L. Baum, T. Petrie, and G. Soules, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[116] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[117] K. Sonoda, K. Ishikawa, T. Eimori, and O. Tsuchiya, "Discrete Dopant Effects on Statistical Variation of Random Telegraph Signal Magnitude," *Electron Devices, IEEE Transactions on*, vol. 54, no. 8, pp. 1918–1925, 2007.