A Test Structure for the Measurement and Characterization of Layout-Induced Transistor Variation

by

Albert Hsu Ting Chang

Bachelor of Science, Electrical Engineering and Computer Science,
University of California, Berkeley (2007)

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2009

Author .....
Department of Electrical Engineering and Computer Science
May 22, 2009

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

Accepted by .....
Terry P. Orlando
Chairman, Department Committee on Graduate Theses

A Test Structure for the Measurement and Characterization of Layout-Induced Transistor Variation

by

Albert Hsu Ting Chang

Submitted to the Department of Electrical Engineering and Computer Science
on May 22, 2009, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Transistor scaling has enabled us to design circuits with higher performance, lower cost, and higher density; billions of transistors can now be integrated onto a single die. However, this trend also magnifies the significance of device variability.

In this thesis, we focus on the study of layout-induced systematic variation. Specifically, we investigate how pattern densities can affect transistor behavior. Two pattern densities are chosen in our design: polysilicon density and shallow-trench isolation (STI) density. A test structure is designed to study the systematic spatial dependency between transistors in order to determine the impact of different variation sources on transistor characteristics and understand the radius of influence that defines the neighborhood of shapes which play a part in determining the transistor characteristics. A more accurate transistor model based on surrounding layout details can be built using these results.

The test structure is divided into six blocks, each having a different polysilicon density or STI density. A rapid change of pattern density between blocks is designed to emulate a step response for future modeling. The two pattern densities are chosen to reflect the introduction of new process technologies, such as strain engineering and rapid thermal annealing. The test structure is designed to have more than 260K devices under test (DUT). In addition to the changes in pattern density, the impact of transistor sizing, number of polysilicon fingers, finger spacing, and active area are also explored and studied in this thesis.

Two different test circuits are designed to perform the measurement. The first test circuit is designed to work with off-chip wafer probe testing equipment; the second test circuit is designed to have on-chip current measurement capabilities using a high dynamic range analog-to-digital convertor (ADC). The ADC has a dynamic range of over four orders of magnitude to measure currents from 50nA to 1mA. The test chip also implements a hierarchical design with a minimum amount of peripheral circuitry, so that most of the chip area is dedicated for the transistors under test.

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

Acknowledgments

The Massachusetts Institute of Technology (MIT) is where every science and engineering student wants to be. For the past two years, I have had the privilege to be a part of the MIT family and be surrounded by so many talented people. I am ever so grateful that I can go this far and receive my Master of Science degree after two years. I am indebted to many, many people. Without their support and help, this would have never been possible.

I would like to first express my thankfulness to my advisor, Duane Boning. Thank you for giving me the opportunity to join the group. Without your unwavering support and clear guidance, my stay at MIT would not have been as exciting and fruitful. You have always encouraged new ideas and been respectful of every decision I made throughout this process. You also presented me numerous opportunities to go to conferences, such as ISSCC, to come up with new research ideas and interact with industry people. These experiences are truly outstanding. I am very lucky to have you as my advisor. I would like to thank Professor Dimitri Antoniadis for many helpful discussions on transistor physics and new transistor models; I would like to thank Professor Hae-Seung Lee for the discussion on the current integrating ADC design.

I would also like to express my full gratitude to my family, especially my parents, Ben and I-Chen. Thank you for your unconditional love. You are the best mom and dad one can ever wish for. You've taught me so much in every aspect of my life, academic or non-academic. Thank you for always being there with me to share the cheers during my success and comfort me during my down times. It is your encouragement that makes me go this far. Without you, my accomplishment here would mean nothing to me. You deserve every part of this thesis as much as I do. I only hope I have lived up to your dreams and expectations. Thank you and I really love you. Arthur, thank you for being such a wonderful brother and always being on my side. Uncle Rong, thank you for helping me make some crucial decisions in my life, including coming to MIT. Yea-Lin and Pao-Shen, thank you for being my great

grandparents and the warmth and welcoming you always show when I come back home.

Many thanks to all my former and present fellow colleagues in the Statistical Metrology Group. All of you have given me so much, from countless hours of technical discussions to good company during lunch and dinner. I am so glad to be able to share many great moments with you. I would especially like to thank the circuit subgroup, Karthik Balakrishnan, Nigel Drego and Daihyun Lim. Karthik, you are a great colleague and a better friend. My life at MIT would be so boring without your friendship. Nigel and Daiyhun, thank you for your help with ideas regarding the design of my test chip.

Many thanks to all my friends at MIT. You have all made my life here truly fun and enjoyable. The laughs we shared, the parties we threw, the meals we cooked together were all great memories. I would like to thank Samuel Chang, Tsung-Yu Kao and Heechul Park in particular.

I would also like to thank all my friends outside of MIT. You have all enriched my life to make it so much more interesting. I would like to thank especially, Roxanne Chen, Ya-Lin Hsu, MonHsuan Wang, Jane Wang, Weilun Lee, and Haofang Chuang.

Last, I would like to acknowledge Kewei Zuo and Henry Lo from TSMC for their suggestions and tapeout of the chip. We gratefully acknowledge the support of TSMC for this research, and for their membership in the Microsystems Industrial Group of the MIT Microsystems Technology Laboratories (MTL).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In 1965, Gordon Moore noted that the number of transistors on a chip doubled every 18 to 24 months [10]; this observation, known as Moore's Law, has been the driving force in the semiconductor industry for the past few decades. In each technology generation, we are able to scale the minimum transistor dimension by a factor of 0.7 from the previous generation. Due to technology scaling, we are able to integrate two times more functions onto a chip without significantly increasing the cost of manufacturing. This trend significantly increases the rate of improvement in the electronics industry. For example, in 1946, the first general-purpose electronic computer called ENIAC[1] was invented; it took up 680 square feet of area, weighted 30 short tons,[2] and consumed more than 150kW of power to merely perform simple addition and subtraction operations. Nowadays, we can design a digital signal processing (DSP) microprocessor that is less than 2mm^2 in area and consumes less than 4μW in power to perform much more sophisticated and complex operations [88]. Scaling enables us to design more powerful electronics that consume orders of magnitude less power. In addition, scaling also enables us to integrate orders of magnitude more functionality.

The ability to consistently scale to smaller dimensions is one of the key factors in the success of the MOS transistors. Conventional scaling involves shrinking the gate oxide thickness, junction depth, gate pitch and source/drain extension. This is

[1]ENIAC stands for "Electronic Numerical Integrator And Computer."
[2]Short ton is a unit of weight equal to 2,000 pounds.

usually done by developing new lithography tools, masks, etch processes, photoresist materials and other process technology improvements [1]. As we continue to scale down to ever-smaller feature sizes, the development of new process technologies becomes more difficult. Roughly around the 180nm technology generation, feature sizes begin to approach the fundamental dimensions, such as atomic dimensions and light wavelengths. Process technology has difficulty keeping up with the pace of feature scaling, and many challenges need to be overcome before we can scale further into future generations. One of the most critical challenges is variation [5, 31].

Variation occurs when two nominally identical devices behave differently after fabrication. Designs can deviate significantly from simulations because of the presence of variation. For example, a 30% variation in operating frequency and a 5-10x variation in leakage power can occur in digital integrated circuits if variation problems are not appropriately handled and resolved [31]. Without careful analysis of the effects of variation, parametric variation can even result in non-functioning circuits in the worst case. The author in [35] shows that functional failure caused by performance variability can also occur in the design of high speed radio-frequency (RF) building blocks, if optimization techniques are not implemented during the design stage of the circuit. Since the behavior of the fabricated design can be so different in terms of power, performance or even functionality, variation is a big obstacle for engineers to overcome.

The remainder of this chapter is organized as follows. Section 1.1 presents a historical overview of process variation. Section 1.2 discusses different aspects of variation analysis. A discussion on the sources of variation, followed by the impact of these variation sources on transistor variability is presented in Section 1.3 and Section 1.4, respectively. These analyses will lead us to realize that understanding how layout impacts transistor characteristics is crucial for successful circuit design, as discussed in Section 1.5. Finally, we will outline the remainder of this thesis in Section 1.6.

1.1 Historical Overview of Process Variation

Though process variation is sometimes treated as a new challenge associated with technology scaling, the problem of variation has been studied for over 40 years. In 1961, Shockley analyzed the random fluctuation in junction breakdown [36]. In 1974, Schemmert and Zimmer presented a set of process parameters that can be tuned in order to minimize the threshold voltage sensitivity of a transistor, addressing systematic variation in transistor threshold voltage [37]. The first issue of the *IEEE Transactions on Semiconductor Manufacturing* appeared in 1988, aiming to encourage innovations in advanced process control, process yield analysis and optimization, and manufacturability. Process variation has always been a critical problem in semiconductor fabrication.

With the continued shrinking in the critical dimension of transistors, however, variation problems are increasing in significance. Finding ways to effectively control and reduce the problem of variation will remain a major challenge in the future.

1.2 Aspects of Variation

To understand variation, it is necessary to understand the taxonomies which are used to describe variability mechanisms from different perspectives. Variation can be discussed from the perspective of spatial scales, causes, behaviors, extraction techniques or proposed mitigating methods. Confusion may arise due to these different perspectives and taxonomies, used to explain variation. This section attempts to clarify some of these confusions.

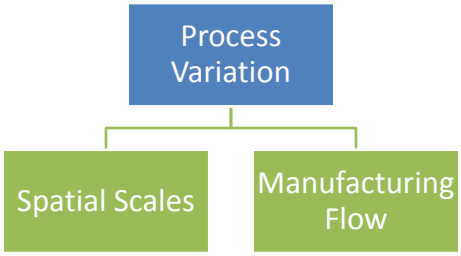

Figure 1-1: Different aspects of variation.

As shown in Figure 1-1, the first categorical decomposition of variation is based on its spatial perspective. In other words, we attempt to describe how variation sources manifest themselves on different spatial scales. The total variation can be separated into (1) lot-to-lot, (2) wafer-to-wafer, (3) die-to-die, and (4) within-die variation. Lot-to-lot, wafer-to-wafer, and die-to-die variation capture the differences in manufacturing conditions, such as temperature, pressure, or other controllable factors, between lots, wafers and dies, respectively. Process engineers worry about the variation problem from the top of the hierarchy, at the lot and wafer level. To mitigate lot-to-lot and wafer-to-wafer variability, they need to ensure better control and uniformity during the manufacturing process. A different temperature profile between wafers during the thermal annealing process, for example, could impact the diffusion of dopant atoms. This can introduce significant transistor threshold voltage variation between wafers [5].

On the other hand, circuit designers are more concerned with the die-to-die (inter-die) and within-die (intra-die) variation, since they either do not have access to the wafer-to-wafer or the lot-to-lot information, or they simply lump all sources of chip-to-chip variation together when dealing with process corners. Die-to-die variation occurs when the same transistor fabricated on different dies have different electrical properties. These are variations between identically designed structures on different dies. Inter-die variation can cause substantial shifts in performance between dies. For example, due to temperature non-uniformity on the wafer, transistors at different locations of the wafer can have different threshold voltages. Since the leakage power is exponentially dependent on the threshold voltage, 2-3% shift in the threshold voltage can result in dies having leakage power that differ by 5-10x.

Within-die (intra-die) variation occurs when nominally identical transistors on the same die have different electrical properties after fabrication. Intra-die variation can introduce significant offset and matching problems between transistors. While matching has long been a problem in analog circuit design, recently, digital circuit designers have also begun to worry about intra-die variation as scaling approaches atomic scales. In a 0.18μm CMOS process, for example, within-die variation in gate

critical dimension (CD) can affect the circuit performance by more than 25% [66]. In the worst case, variation can even result in non-functioning designs, significantly decreasing the yield of the circuit.

As noted in Figure 1-1, another way to categorize variation is based on the flow of semiconductor manufacturing process [9]. The entire semiconductor process can be divided into two general components: front-end and back-end. Front-end components usually refer to the processes that involve in fabricating transistors. Processes such as photolithography, ion implantation, polysilicon gate deposition and oxidation fall into this category. Back-end components usually refer to the processes that involve creating interconnects, and passive components, such as resistors and capacitors. Processes such as chemical mechanical polishing (CMP), copper electroplating and etching of metal wires fall into this category. Both front-end and back-end processing steps can introduce significant variations. Historically, front-end variation (or device variation) has contributed to about 90% of the total amount of performance variation in good designs [20]. Even though the front-end variation plays a major role in performance variation, back-end variation can be a key factor in areas such as on-chip clock distribution. The clock skew resulting from back-end variation can limit the maximum clock frequency [21, 22]. As we continue to scale, the back-end variation will continue to increase in its significance.

In order to improve manufacturing yield, it is necessary to form a systematic approach to mitigate the effects of variation. Therefore, it also makes sense to discuss variation in the context of its mitigation flow. Figure 1-2 below summarizes the flow: Cause → Behavior → Extraction → Solution. In the following sections, each step in the flow will be discussed in detail in the context of variation, to familiarize readers with that particular aspect of variation.

1.2.1 Cause: Intrinsic vs. Extrinsic

In terms of variation causes, we can categorize the sources of variation into two groups: intrinsic variation sources and extrinsic variation sources. Intrinsic variation sources are those which result from the atomic-level differences between transistors

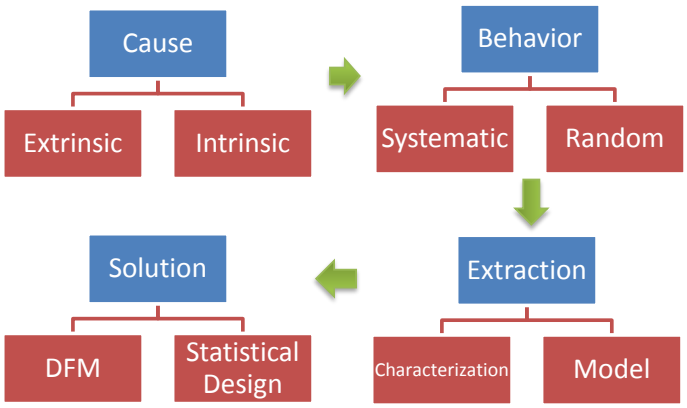

Figure 1-2: Studies of variation: Cause → Behavior → Extraction → Solution.

or structures, even when the layout geometry and the operating environment are the same. Instead of using classical physics, scaling to the atomic level will require us to explain device behaviors using the principles of quantum physics. Classical physics can no longer capture, for example, the probabilistic distribution of atoms inside the transistors. At this scale, because the dimension of the transistors is within one or two orders of magnitude of the lattice distance of silicon atoms, the precise location of each silicon atom becomes important and influences the overall characteristic of the transistor.

These stochastic variations between devices appear in threshold voltage, line-edge roughness (LER), film thickness, and energy level quantization. Transistor threshold voltage is directly related to both the dopant profiles and the number of dopants in the channel. Since the location of the dopants in the crystal silicon and the number of dopants in the channel are both governed by random processes, the threshold voltage will vary. Energy level quantization is also caused by intrinsic random variation. Randomness in the energy level will introduce electrical noise. The introduction of additional noise and the uncertainty in the noise level will make circuit design more difficult. For example, in the design of an analog-to-digital converter, precise noise level information is necessary to determine the resolution of the ADC. This also applies to most analog or radio-frequency designs where information about noise is essential. The final impact of stochastic variation on devices is in line-edge roughness.

The variation in the line-edge is quite noticeable when the device length scales below 100nm [13]. The typical edge roughness of a polysilicon gate is around 5nm, but the values could be much worse depending on how the polysilicon gate is formed. At the 180nm technology node, a variation of 5nm means a variation of less than 2%; however, in the 32nm technology regime, a variation of 5nm means a variation of more than 15%.

Extrinsic variation sources are those that result from an unintentional shift in manufacturing conditions. Unlike the problems associated with the fundamental atomistic scaling, these kinds of variation can usually be controlled by improving the manufacturing process. The traditional extrinsic variation sources are an unintentional shift in temperature, pressure, photoresist development, etching and other controllable conditions in the fab. Extrinsic variation sources are the main contributors to lot-to-lot, wafer-to-wafer and die-to-die variation, whereas within-die variation is made up of both intrinsic and extrinsic variation sources. In general, more variation occurs die-to-die than wafer-to-wafer within one lot. Historically, scaling has made process control much more difficult. While both the transistor critical dimension and the interconnect pitch are getting smaller, parts of the process technology have not been able to keep up with the pace of scaling. For example, the current lithography solutions, based on the 193nm light wavelength, have been the main lithography technology in use for more than a couple generations of scaling, due to the lack of affordable alternatives for new lithographic processes.

1.2.2 Behavior: Systematic vs. Random

Behaviors are the responses to the variation sources indicated in the previous section. There are two types of behavior: systematic behavior and random behavior. By definition, systematic variation describes the behavior of variation that can be decomposed in a methodical and organized manner, which can then be formulated by functional forms. Therefore, systematic variation is also called deterministic variation. Adequate modeling incorporating the systematic variation of the design can be used to predict design behavior. Designers can take advantage of this predictability to

design circuits accordingly and avoid designing for the worst-case. For example, temperature non-uniformity during the rapid thermal annealing (RTA) process is closely related to the non-uniformity of the polysilicon pattern density in the layout. We can either use a predictive model to estimate how much variation will be expected from the layout, or reduce the systematic variation by inserting dummy polysilicon to maintain a uniform pattern density across the die. Both solutions rely on the knowledge that RTA variation is systematic, rather than designing for the worst-case.

Another class of behavior falls into the category of random variation. Random variation is variation for which the designers do not have enough information to quantitatively or functionally relate to its origin of variation, and therefore are forced to design for the worse case. This means a large design margin must be incorporated to compensate for the worst case scenario. Designing for the worst case can waste resources that can potentially be used for performance improvement or energy reduction. Every effort should be made to understand this kind of variation better in order to minimize the design cost associated with accommodating it.

A variation that is referred to as systematic relies solely on the fact that designers can trace the origin of the variation back to a specific design parameter. In other words, nothing prevents a random variation source from becoming a systematic variation source if researchers can find a way to relate the variation source to a specific design parameter. Moreover, the same variation can be treated as either a systematic variation or a random variation, depending on who is defining it. From a circuit designer's point of view, a wafer-level temperature gradient induced variability is treated as random variation because, most of the time, circuit designers have neither wafer-level information nor do they have controls over the temperature gradient during processing steps. On the other hand, process engineers have full control and access to the wafer-level temperature information; thus, temperature induced variability at the wafer-level is treated as systematic variation by process engineers.

1.2.3 Extraction: Characterization vs. Modeling

Extraction of variation sources is the most essential step needed before designing mitigation techniques. In past technologies, a simple scribe line[3] which included a handful of transistors with different dimensions was sufficient to build an accurate model for circuit simulation. As transistor dimension and the gate pitch shrink due to scaling, interactions between the surrounding layout environment and the transistor characteristics are becoming significant. A simple scribe-line method is no longer sufficient to accurately capture the transistor characteristics in the more advanced technologies. To understand this point further, let us consider two transistors with the same dimensions being placed in different scribe lines: one scribe-line is surrounded with high shallow trench isolation (STI) pattern density and the other scribe-line is surrounded with low STI pattern density. The two transistors will experience different amount of stress from its surrounding area and thus will have different transistor characteristics. As a result, gathering the test results from any one of these transistors will not be sufficient for transistor or circuit modeling, of in-product circuits. Instead of modeling, designers are forced to "characterize" the performance of each particular circuit, since the usefulness of scribe-line information to other circuit designs is limited.

To develop a model, it is important to have a set of design of experiments to explore the effect of single or multiple parameters on the transistor characteristics that we would like to model. A simple design of experiments (DOE) may have elements where only one variable is varied at a time throughout the entire experiment, such that we can develop a model for the behavior of the design in response to that variable. Using our previous example, we can model the response to the changes in surrounding STI pattern density, but we cannot derive a complete transistor model which incorporates other layout-related variations. In order to build a more accurate and complete model, the DOE needs to include elements where multiple factors are simultaneously varied to see the interactions between different parameters. Again, without a proper DOE, the resulting measurements will only be the characterization of one particular circuit,

[3]An empty area between dies that can be used for the insertion of simple test structures and transistors for process monitoring.

and the information gathered will have limited usefulness for other circuits we might design later.

1.2.4 Solution: DFM vs. Statistical Design

Lastly, we will discuss different design approaches to mitigate variation. The first approach is called design for manufacturability (DFM). This design strategy takes advantages of the parts of design that are modelable (or in other words, systematic). For example, it is well-known that the transistor orientation impacts the fabricated channel length of the transistors. If matching is important, in a design, such as in the case of analog circuits, under a DFM strategy, designers will not use transistors with different orientations. The ultimate goal for engineers is to be able to approach all of the problems using DFM solutions. This requires the understanding and investigation of variation sources and ultimately the incorporation of these findings into modeling.

For the type of variation for which the source is either unknown or is truly random, we can use a second design approach called statistical design. This design approach follows the principle of better-than-the-worst-case-design. In the past, designing for the worst-case was common. For example, in a digital integrated circuit, in order to achieve a high yield, designers are forced to put large margins into their designs to ensure that the slowest logic path can still operate under the frequency constraint. However, it becomes exponentially more expensive to accommodate the slower tail of the distribution. According to the experiment in [14], to ensure that the slowest 5% of logic paths can still operate under the desired frequency, an extra 25% of leakage power is consumed. In other words, the amount of additional power spent to ensure the functionality of the slower logic paths has serious diminishing returns. The key concept of statistical design is not to lose too much performance accommodating a small percentage of circuits, but rather to make engineering tradeoffs between performance and statistical yield.

Post-fabrication testing is necessary if a statistical design approach is used during the design process. The main purpose is to find the dies that fall within the lower percentage of the statistical distribution curve, and either eliminate them or put them

into a lower performance bin. A second option is to use adaptive circuit techniques. Many common digital techniques, such as adaptive body biasing (ABB) and adaptive voltage scaling (AVS), have been developed to compensate for die-to-die or within-die variability. For example, dies with high standby leakage power can be improved by adaptively increasing the threshold voltage of the devices through the control of the voltage between the source and the body terminals of the transistor. In these cases, the tradeoff is between design complexity and area, and the resulting yield and performance.

1.3 Sources of Variation

In Section 1.2, we focused our discussion on different aspects of variation. It is equally important to understand the underlying physics that causes these variations and their ultimate impact on the transistor parameters. Therefore, this section will begin with an overview of the semiconductor manufacturing process, pinpoint the steps that cause significant variation, and review key details of those steps.

The semiconductor manufacturing process has grown ever so complicated. From generation to generation, new processing steps are continuously being added to the existing process; sometimes, even a complete revision of an older generation's process is required to accommodate the aggressive scaling of the critical dimensions in the modern era. Typically, more than one hundred individual steps are involved in the entire manufacturing procedure. Each fab is also slightly different from the others in terms of the equipments and processing steps involved during fabrication. Our goal is not to detail the specific aspects unique to the individual fab, but to highlight some of the most common steps in modern semiconductor manufacturing.

A typical fabrication process begins with the creation of the wafer. Polycrystalline silicon is melted in order to grow silicon crystals. The resulting silicon crystals are then sliced into individual wafers. In order to remove any scratches or impurities, one side or both sides of the wafer are polished to a mirror-like surface. The chips are built onto and into the wafer surface in the later steps. With the polished wafer,

the next step is to grow (or deposit) a silicon dioxide (S_iO_2) layer onto the wafer. S_iO_2 can either be grown onto the wafer by exposing the wafer to oxygen at very high temperatures, or it can be created by combining with oxygen in a process called chemical vapor deposition (CVD).

Photolithography, or lithography for short, is a process used to selectively remove and pattern different layers on the silicon surface. The wafer is first coated with a light-sensitive chemical called photoresist. A geometric pattern on the photomask is transferred through light exposure onto the photoresist layer on the wafer. This is similar to the way film is exposed to light to form a photographic image. After the pattern is formed on the resist by the lithography process, the resist patterns are washed away (or remain, depending on the type of resist). The material exposed below the resist, for example S_iO_2, is then etched away to form the desired pattern on the wafer. In a complex integrated circuit design, this process can be repeated more than twenty times in order to build up subsequent layers of various materials to form multiple layers of the circuit. For instance, polysilicon gate formation results from one of the lithography steps.

Subsequently, ion implantation and thermal annealing processes are used to place dopant atoms in the silicon substrate. In ion implantation, impurity ions are created and accelerated with high energies (ranging from keV to MeV) into the silicon lattice. These high energy ions break the original lattice structure of the silicon crystal; moreover, the ion itself will not necessarily reside at the lattice sites. Therefore, a follow-up annealing process is needed to repair the damage from ion implantation and also to activate the dopant atoms. Some common annealing processes are furnace annealing, rapid thermal annealing (RTA) and laser annealing. Other critical steps involved in the creation of transistors are the shallow trench isolation (STI) process, lightly doped drain implant process, sidewall spacer formation, contact formation, and local interconnect/via formation.

Variation occurs in all of the manufacturing steps described above. However, a number of these processing steps can be highlighted as major sources of variations: (1) photolithography, (2) ion implantation and thermal annealing, (3) etch, (4) STI

and sidewall spacer stress, and (5) chemical-mechanical polishing (CMP). Depending on the features being fabricated, each processing step affects transistors differently. Photolithography, etch and CMP variation would affect the physical fabricated dimensions of transistors, while ion implantation, thermal annealing and lattice stress would influence the internal molecular composition of the material making up the transistors. In the section below, we will identify the processing steps which most induce variations and also point out the transistor parameters that are most affected by them.

1.3.1 Photolithography

The photolithography process is used to project the design pattern from photomasks onto the actual wafer. In an ideal situation, it is best to use a light wavelength that is equal or shorter than the critical dimension (CD) in that technology. Due to the delayed introduction of lithography processes based on the 157nm wavelength of light, many of the recent technology generations have been forced to extend the older lithography based on 193nm light. As we continue to scale below 100nm, the lithography process cannot keep up with the aggressiveness of scaling, and a significant challenge has been that newer technologies continue to base their lithography process on wavelengths that are much longer than the CD.

Because of the longer wavelength used in the lithography process, image and shape distortion are inevitable. Shape distortion can be attributed to the low-pass filter behavior of the lithography process while trying to print smaller features than the light wavelengths. The low-pass filter characteristics can result in inaccuracy while resolving the high frequency components, such as corners or sharp turns, on the wafer. This inaccuracy translates into several major types of distortions: linewidth variation (proximity effect), line-end shortening and corner rounding. Due to the strong layout dependence, these kinds of variation are highly systematic.

The proximity effect refers to the strong dependence of the printed critical dimension on the surrounding layout. The closer the surrounding layout is, the more impact it is going to have on the printed dimension of the transistors around it. As we scale

from one generation to another, the minimum allowable pitch is getting smaller as shown in Figure 1-3. Therefore, more interaction between neighboring transistors is expected. That is one reason why scaling has increased the problem of variation. Line shortening refers to the reduction in line length while printing a rectangular structure. This is due to both the diffraction of light and photoresist diffusion. Corner rounding refers to the smoothing of a rectangular corner into a rounded corner, mainly due to the low-pass filter characteristics of the lithography process.

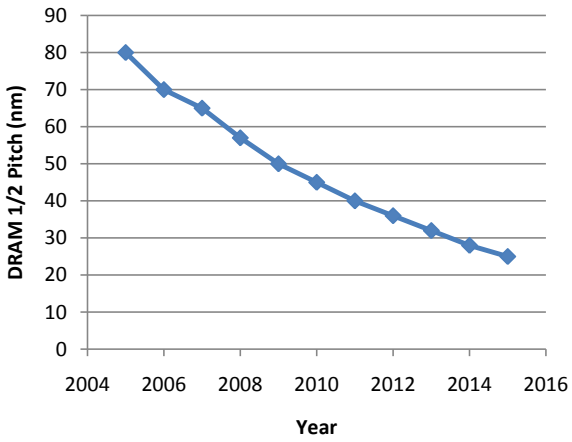

Figure 1-3: Minimum DRAM pitch size [1].

Another related phenomenon of the lithography process is line-edge roughness (LER). All of the previously discussed variations induced by the lithography process have been systematic. Though the amount of variation cannot be exactly predicted, we can still obtain consistent but approximate results, which can provide information about offsets in one direction or the other. Using corning rounding as an example, although the exact amount of rounding may not be obvious, we still expect rounding to make the printed dimension smaller than the original drawing and physical or empirical model based corrections can be calculated. LER, on the other hand, is considered to be random variation. In the past, since the critical dimension of transistor was orders of magnitude larger than the magnitude of the roughness, LER received little attention. However, as the devices have scaled to lengths below 50nm, the magnitude of the roughness does not scale with the device.

As a result, LER becomes a larger fraction of the gate length. As shown in

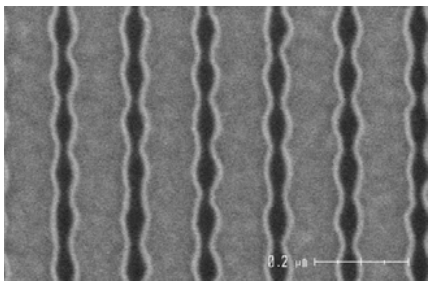

Figure 1-4: A top down CD-SEM image displays the magnitude of the roughness [2].

Figure 1-4, the line no longer resembles a clean straight line anymore, but looks significantly rough. The typical edge roughness remains on the order of 5nm, almost independent of the type of lithography used in either production or research [23]. The scaling of the lithographic features has made process control ever more difficult. The lithography process and line-edge roughness are major contributors to the critical dimension variation. They each contribute to systematic and random variation in the CD, respectively.

1.3.2 Etch

Similar to the photolithography process, the etching process has non-uniformities which also contribute to the linewidth variation. Etching induced variation manifests itself as a difference between the photoresist and the actual polysilicon dimension. Although both etching and photolithography result in CD variation, the two effects are caused by different physical reasons. Nevertheless, the amount of variations contributed by etch and photolithography are comparable and additive. We can classify etch induced variation into three categories: aspect-ratio-dependent etching, microloading and macroloading.

Aspect-ratio-dependent etch refers to the dependence of etch-induced variation on the feature size of the surrounding layout [24]. A common phenomenon seen in plasma etching is that smaller features usually etch at a slower rate compared to larger features. The exact etching process is complicated, but qualitatively, the feature size dependence of etching can be attributed to the changes in transport properties of

ions, etchants and sidewall inhibitors based on different feature sizes.

Microloading and macroloading are driven by the other physical mechanisms. The changes in the layout features can increase or decrease the density of the reactants. On a small scale, transistors designed to have the same dimension can be etched very differently, depending on the surrounding layout features due to microloading. Significant variation can be observed, especially when there is an abrupt change in local pattern density. On a large scale, the etching variation is determined by the average loading across the entire die or even the entire wafer instead of the local surrounding features. A wafer containing different types of design layout, such as SRAM, logic, and analog circuits, can experience significant etching variation due to macroloading.

The majority of the etch-related variation is systematic since it is highly layout-dependent. Etching and photolithography together are the two major contributors to both systematic and random critical dimension variation in transistor fabrication.

1.3.3 Ion Implantation and Thermal Annealing

As described previously, ion implantation and thermal annealing are ways to introduce dopant ions into the semiconductor material. This results in randomness, both in number and placement of dopant atoms in the channel, illustrated by a Poisson distribution as shown in Figure 1-5. Due to the scaling in transistor dimensions, the total number of dopant atoms required to be in the channel to achieve a certain level of doping concentration decreases from generation to generation. As a result, in the most advanced technology nodes of 45nm or even 32nm, the number of dopant atoms required is in the tens or low hundreds. Therefore, the variation in the number of dopants around a certain mean value increases significantly. As shown in Figure 1-5, a dopant profile in a 50nm n-mosfet is simulated by a Monte Carlo procedure. The source and drain are highly doped areas, while the dopants in the channel and body are very scarce. However, these few dopants underneath the channel are much more important than the dopants in the source/drain region, since their number and position determine the threshold voltage of the transistor. Due to Poisson statistics,

the uncertainty in the number of dopants in small transistors can be 5-10% of the total number of dopants. Since ion implantation and thermal annealing are the process steps which affect the number and the distribution of dopant atoms the most, we collectively call this problem random dopant fluctuation (RDF).

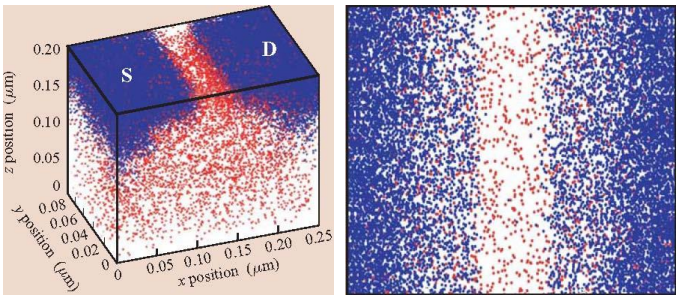

Figure 1-5: Randomly placed dopants in a 50-nm channel [3].

A few solutions have been proposed to alleviate the problem of RDF. In general, the dopant atoms closest to the surface of the channel will have the most impact on threshold voltage. The author in [27] is able to show that using a retrograde doping profile would significantly decrease the variation in threshold voltage. Retrograde doping is a way to dope the channel such that the high substrate doping is buried somewhat beneath the surface, leaving a lightly doped region close to the surface. Because there are not many dopants close to the surface, the amount of variation can be reduced significantly. In [32], the authors discuss the possibility of having fully depleted silicon on insulator[4] (SOI) without any dopants underneath the body. The threshold voltage for the fully depleted SOI is controlled by the gate-metal workfunctions. This can potentially alleviate the problem of dopant variations, but most of the potential transistor structures are not as scalable as the traditional bulk-CMOS technology. It is very difficult to replace CMOS using these new transistor structures, and understanding of dopant or other variations in scaled CMOS as well as in potential alternatives will continue to be important.

[4]Silicon on insulator technology (SOI) refers to the use of a layered silicon-insulator-silicon substrate in place of conventional silicon substrates in semiconductor manufacturing, especially microelectronics, to reduce parasitic device capacitance and thereby improving performance.

1.3.4 Lattice Stress

A more recent variation mechanism is caused by lattice stress. Up until about the 90nm node, the intrinsic device performance has steadily increased by about 17% per year, following an inverse gate-length ($1/L$) dependence. The steady increase in the intrinsic performance is enabled by the continuous increase in the intrinsic carrier velocity due to the scaling of gate-length [77]. However, from the 90nm technology generation onward, this intrinsic carrier property has stayed roughly constant, even with scaling. Additional innovations besides scaling are needed to further improve the intrinsic transistor performance, such as the use of strained silicon. Because mobility is a strong function of stress, by applying a physical stress on silicon lattice, we can increase the carrier mobility. This increase can lead to a higher saturation current and a higher switching speed for circuits. A tensile stress is desired for NMOS transistors to increase the mobility of electrons, and a compressive stress is desired for PMOS transistors to increase the mobility of holes. An increase in mobility of up to 60% has been reported by applying stress [78].

However, stress can also be introduced to the silicon lattice unintentionally. The mismatch in thermal expansion of different materials is one mechanism that can create unintentional stress. The use of shallow trench isolation (STI) is one example. During the oxidation step in the formation of STI, because of volume expansion, the neighboring transistors experience compressive stress. Compressive stress has a negative impact on the performance of NMOS transistors since it greatly decreases the electron mobility.

The strain-induced variability is also highly systematic since it depends on the layout of the transistor and its surrounding geometry. The size of the active area and the distance from the gate to the STI edge are especially important when dealing with stress. As the gate moves farther away from the STI edge, it will experience less compressive stress from the expansion of the dielectric material. Larger transistors also tend to be less sensitive to external stress. As the distance between transistors continues to decease, the channel gets closer to the STI edge; therefore, a significant

increase in unintentional stress on the channel is expected in future technologies. Design models which account for this systematic variation are needed to better predict transistor behavior.

The lattice stress can also influence transistors in other ways. Sheu and Zangenberg both investigated the stress effects on dopant diffusion in scaled MOSFETs [52, 53]. A model is provided by Zangenberg that relates stress to dopant diffusion. Liu explores the effect of STI strain on the thickness variation of silicon gate oxide [28]. One conclusion is that severe variation in the oxide thickness appears near the STI boundaries. This near-edge oxide region has a very weak dependence on the total oxidation time, compared to the oxide region near the center of the polysilicon gate.

1.3.5 Chemical-Mechanical Polishing

All of the variation sources described so far are part of the front-end process, which refers to the steps that create the actual transistors. Chemical-mechanical polishing (CMP) is used repeatedly in the back-end process, which refers to the steps that form the wiring and interconnect of the circuit. The goal of copper CMP is to completely remove the overburden copper or excess metal that is outside the trench area, leaving a flat surface that is coplanar with the surrounding dielectric. Polishing is an important step for modern IC fabrication processes. This planarity allows the entire metallization sequence to stack up multiple metal layers and eventually build up to the complete fabrication which can consist of more than ten metal layers.

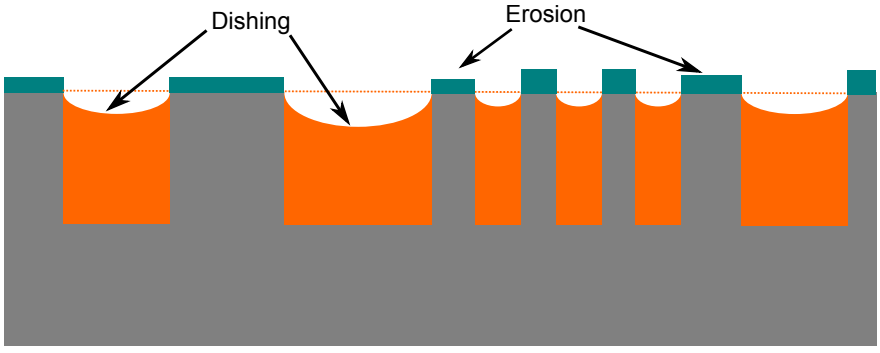

Figure 1-6: Pattern-dependent CMP variation due to dishing and erosion.

Unfortunately, the polishing procedure is not ideal as shown in Figure 1-6; it suffers from a significant amount of variation. Two kinds of variation are most common in the CMP process: dishing of copper and erosion of dielectric. Dishing refers the overpolishing of the features within the trench relative to the surface of the dielectric layer. Erosion refers to the removal of surrounding dielectric when it should not be removed. In general, larger features suffer more dishing than smaller features, but conversely, smaller features suffers more erosion compared to the larger features. For medium-sized-features, both dishing and erosion contribute to some degree of polishing variation. CMP-induced variability is highly systematic, since it relates directly to the feature size and layout pattern densities being polished.

1.4 Impact on Transistor Variability

In the previous section, we overviewed the major sources of variation in the IC fabrication process. It is equally important to understand the impact on transistor parameter variability from these variation sources. The total amount of variation, including mean shifts around μ_{total} and variances contributing to the total variance σ^2_{total}, in each of the transistor parameters can be decomposed into lot-to-lot ($L2L$), wafer-to-wafer ($W2W$), die-to-die ($D2D$), within-die (WID) and random (R) sources of variation.

$$\mu_{total} = \mu_{L2L} + \mu_{W2W} + \mu_{D2D} + \mu_{WD} + \mu_R \tag{1.1}$$

$$\sigma^2_{total} = \sigma^2_{L2L} + \sigma^2_{W2W} + \sigma^2_{D2D} + \sigma^2_{WD} + \sigma^2_R \tag{1.2}$$

This spatial decomposition has been discussed previously in Section 1.2. The total amount of mean-shift due to variation is the sum of the shifted means due to each of the individual variation components as depicted in Equation 1.1. Similarly, the total amount of variance due to variation is the sum of the variance of the individual variation components as depicted in Equation 1.2. The first four variation components, lot-to-lot, wafer-to-wafer, die-to-die, and within-die variation, are treated as system-

atic components in this equation, while the last component describes the random stochastic process. Each circuit parameter has a different proportion of variance contributed from each the components in the above equations, depending on the physics behind the fabrication steps. In the following sections, we associate the variation in transistor parameters with one or more of the previously discussed processing steps.

The equations below highlight some of the most important benchmarks in determining transistor performance. In these equations, W is the transistor width, L is the transistor gate length, μ is the mobility, V_T is the threshold voltage, and l' is the DIBL (drain-induced barrier lowering) factor. Equation 1.3 describes the saturation current which can be used to evaluate the drive strength of transistors. Equation 1.4 describes the leakage current, which can be used to evaluate the leakage power consumption during the idle stages. Equation 1.5 is the delay equation used in the ITRS roadmap to characterize transistor switching speed. Here, we can clearly see how each transistor parameter affects these key metrics. This will help us to relate the variation in transistor parameters directly to these performance metrics.

$$I_{sat} = \frac{1}{2} \cdot \frac{W}{L} \cdot \mu C_{ox} \cdot (V_{gs} - V_T)^{\alpha} \cdot (1 + \lambda V_{ds}) \tag{1.3}$$

$$I_{leak} = I_0 \cdot e^{\frac{-V_T}{kT/q}} \cdot e^{\frac{-V_{ds} exp(-L/l')}{kT/q}} \tag{1.4}$$

$$\tau = \frac{C_{load} \cdot V_{supply}}{\frac{1}{2} \cdot \frac{W}{L} \cdot \mu C_{ox} \cdot (V_{gs} - V_T)^2} \tag{1.5}$$

1.4.1 Transistor Dimension

The transistor dimension refers specifically to gate channel length and width. These are the two parameters that directly influence the transistor performance as shown in Equation 1.3. To increase the switching speed of circuits, it is often necessary to increase the on-current of the transistors. There are two options to increase the on-current: one is to increase the width, and the other is to decrease the length. However, increasing the width is, in general, not as effective as decreasing the length, since increasing the width will also increase the load capacitance. The switching speed

will not benefit by the same amount as the increase in width. On the other hand, decreasing the length, or the critical dimension, will not only increase the on-current of the transistor but also decrease the load capacitance. As a result, decreasing the channel length is a more effective way to improve the switching speed of transistors.

To design high performance circuits, the minimum channel length and larger-than-minimum channel widths are often chosen for transistor sizing. The width dimension is generally much larger than the length; therefore, variation in channel width has a negligible impact on the key performance parameters. A number of processing steps and modules can contribute to the overall variation in gate length. These factors include the wafer, mask, exposure system, etching process, spacer definition, source and drain implantation, and the environment during the manufacturing process. Of these factors, the primary variability sources for channel length are the steps within the photolithography process (systematic), plasma etch process (systematic), and line-edge roughness (random), as described previously in Section 1.3.

Saturation current has a linear dependence on channel length while the leakage current has an exponential dependence on channel length, as shown in Equation 1.4. The exponential dependence arises from a phenomenon called drain induced barrier lowering (DIBL), in which the threshold voltage of the transistor decreases with higher drain-to-source voltages. According to Equation 1.4, DIBL has a small impact when the transistor length, L, is large relative to the DIBL factor. As we continue to scale, however, the channel length becomes comparable to the DIBL factor. Therefore, a few percent variation in channel length can be greatly amplified in terms of its impact on leakage current variation. This translates directly into more power consumption. In this era where consumer products emphasize the importance of low-power operation, an order of magnitude increase in power consumption is unacceptable. Profit losses due to small production yield can occur as a result of channel length variation, and thus finding means to mitigate channel length variation is urgent.

1.4.2 Threshold Voltage

Threshold voltage is one of the most critical transistor parameters in CMOS technology. The strong nonlinearity at the threshold voltage allows transistors to behave like switches, turning on and off to perform logic operations. Because of this key feature, CMOS technology has been the dominant technology for many years. As shown in Equation 1.3-1.5, threshold voltage plays a major role in every important performance parameter. Not only does it determine the on-current of a transistor, but it also has an exponential effect on the leakage current. Threshold voltage variation, therefore, has always received a considerable amount of attention in the circuit design community.

$$2\phi_B = 2\frac{kT}{q}ln\frac{N_a}{n_i} \tag{1.6}$$

$$V_T = V_{fb} + 2\phi_B + \frac{\sqrt{qN_a 2\varepsilon_s}}{C_{ox}}(\sqrt{2\phi_B + V_{sb}} - \sqrt{2\phi_B}) \tag{1.7}$$

Threshold voltage is also one of the most difficult transistor parameters to control, due to a number of reasons. First, from Equation 1.7, we can see that the doping concentration, N_a, the flat-band voltage, V_{fb}, the oxide thickness, C_{ox}, and the band bending in the body, $2\phi_B$, are all part of the threshold voltage equation. We can therefore infer that many processing steps contribute to the variability in threshold voltage. Second, the variation behavior of threshold voltage is mostly random[5] due to random dopant fluctuation (RDF) in the ion implantation and thermal annealing steps. As a result, it is very difficult to develop mitigation techniques to control or reduce the variation in threshold voltage.

In 1989, Pelgrom investigated the matching properties of threshold voltage [17]. From his analysis, he concluded that the total mismatch P is normally distributed with zero mean as shown in Figure 1-7. The variance of the threshold voltage can be described by Equation 1.8.

[5]The systematic part of the variation is mostly due to DIBL; we consider that as part of the channel length variation.

$$\sigma^2_{V_{th}} = \frac{A^2_{V_{th}}}{W \cdot L} + S^2_{V_{th}} D^2 \tag{1.8}$$

This model is also called the Pelgrom model,[6] in which the variance of the threshold voltage is assumed to be inversely proportional to the transistor area, and proportional to the separation distance between two devices.

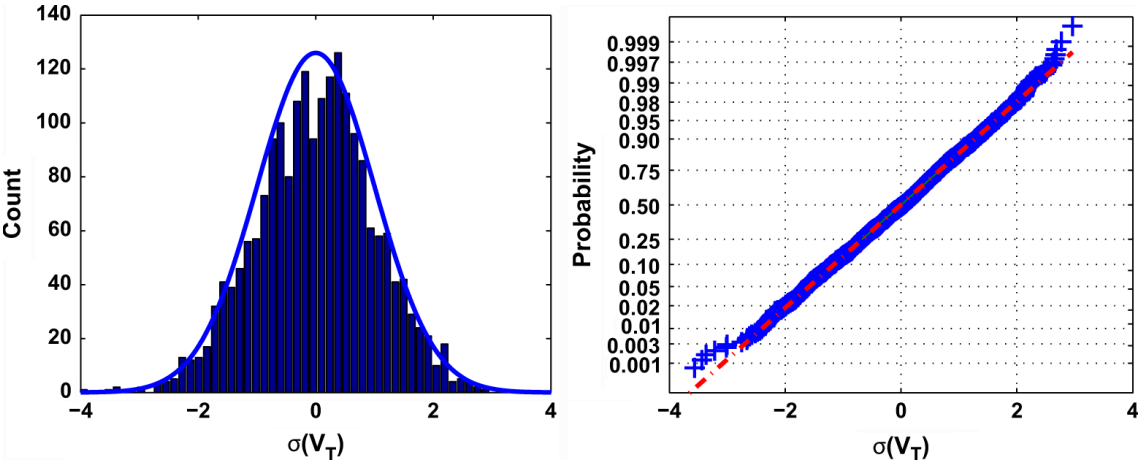

Figure 1-7: Normal distribution of threshold voltage [4].

The threshold voltage of a device largely depends on the dopant profiles, which can be specified by the total number of dopants and the location of each dopant. Frank reported a full 3D simulation of the MOSFET under the influence of random discrete doping [19]. The simulator is based on the drift-diffusion model. For each lattice site, the probability of a dopant atom presents at that location is calculated. The resulting dopant profiles under the channel, source, drain and gate of a MOSFET are then used to calculate the threshold voltage. Based on numerical simulation, Asenov is able to create an empirical model that captures the standard deviation of threshold voltage [18]:

$$\sigma^2_{V_{th}} = 3.19 \times 10^{-8} \frac{T_{ox} N_A^{0.4}}{\sqrt{W \cdot L}} \tag{1.9}$$

Threshold voltage variation greatly limits the ability to design accurate analog

[6]The variant of the Pelgrom model which describes the standard deviation of saturation current can be found in [71].

circuits, which generally require good matching. Circuits, such as amplifiers or comparators, can suffer from significant amounts of offset voltage. In digital circuits, threshold voltage variation is important in two respects. The first is in the design of digital memories. Variation can reduce the read and write margins of the memory circuits and result in non-functioning cells. The second relates to the significant impact of leakage current variations. A number of mitigation techniques have been used to compensate die-to-die variation due to threshold voltage. For example, Tschanz in [89] used body bias as a knob to adjust the threshold voltage to reduce variation in die frequency by a factor of seven. Techniques such as this can greatly minimize the die-to-die variation due to threshold voltage and significantly improve the yield of the circuit.

1.4.3 Mobility

Mobility is a proportionality constant that relates the drift velocity of electrons or holes to the electric field applied across a material. The relationship can be written as:

$$|v| = \mu \cdot E \tag{1.10}$$

$$\mu = \frac{q\tau_m}{m} \tag{1.11}$$

where $|v|$ is the drift velocity, μ is the mobility of electrons or holes, and E is the electric field applied across the material. The absolute value of drift velocity is used because the drift velocity of electrons and holes are in opposite directions. Equation 1.11 shows the relationship between carrier mobility and material properties, such as the mean free time between collisions, τ_m, and the effective mass, m. One way to improve carrier mobility is to increase the mean free time between collisions. However, an easier and more common way is to reduce the effective mass of the carriers by applying strain.

ITRS projects that the mobility enhancement factor due to strain will be approximately 1.8 for every technology generation up until the year 2022 [1]. This

requirement is due to an inability to improve carrier velocity further by merely scaling channel length. Therefore, in order to improve transistor performance, strain is necessary.

However, mobility improvement does not translate directly into an improvement in saturation current. The same report also shows that a saturation current multiplication factor of only 1.08 can be achieved by an 80% improvement in mobility. This raises two questions. First, why does an improvement in mobility not translate directly into an improvement in saturation current, as predicted by Equation 1.3? Second, if the change in mobility does not affect the saturation much, why is there a concern about mobility variation due to unintentional strain?

A change in mobility does not translate directly into an improvement in saturation current mainly due to a phenomenon called velocity saturation. Although drift velocity follows Equation 1.10, in a small device, once the electric field reaches beyond a certain limit, the electron and hole velocities will saturate at around 10^7 cm/s no matter how large the applied electric field is. In other words, the linear relationship between the electric field and the mobility will only hold in the triode region of transistor operation. While increasing the mobility will help to reach the saturation velocity faster, it will not contribute directly to the saturation current improvement.

Despite this, changes in mobility still play an important role in determining transistor performance. Transistor performance is determined by switching speed, which describes how fast a node can be charged from ground to the supply voltage and discharged back to ground. During the charging or discharging cycle, the transistor operates in the saturation regime for only a very small portion of the cycle, while for the majority of the cycle, the transistor operates in the linear region. Na in [79] shows that the discrepancy between a simulated model and the actual hardware measurement in logic delay can be resolved by defining a new parameter called I_{eff}, which takes into account the currents from different regions of operation. For this reason, the variation in mobility is still very important in determining transistor performance.

As described previously in Section 1.3.4, systematic unintentional strain can be introduced to the silicon lattice and create mismatches between transistors. The

nature of this systematic variation is highly dependent both on the size of active region and the distance to the edge of STI. Wils in [51] designed a dedicated set of test structures to study the drain current mismatching caused by STI stress. In Chapter 2, we will describe how our test structure design can also be used to study the effects of stress.

1.4.4 Gate Dielectric Film

The gate dielectric film separates the polysilicon gate channel from the silicon substrate. The thickness of this film impacts multiple transistor parameters of importance: gate oxide capacitance, gate tunneling current, mobility and threshold voltage. We will focus our discussion on gate oxide capacitance and gate tunneling current here because threshold voltage and mobility have already been discussed.

The film thickness scales down by approximately 30% every technology generation. This scaling is necessary in order to increase the controllability of the gate and thus minimize the short channel effect[7] of the scaled transistor. At the 65nm node, the oxide thickness approaches 10-12Å, and the ITRS projects that the oxide thickness will even scale down to 9Å in the future [1]. The silicon dioxide film is formed by a thermal oxidation process. Historically, this process had been well-controlled and the variation in the film thickness has not been a concern. As the thickness approaches atomic scales, however, the oxide thickness becomes so thin that the atomic-level probabilistic behavior starts to become significant. A thickness of 10Å corresponds to only about five atomic layers of S_iO_2. A change in thickness of just one atomic layer would correspond to a 20% variation in the oxide thickness. The control of this variation becomes very difficult because we are limited by the fundamental properties.

The effective gate capacitance is proportional to the oxide thickness. Any variation in gate capacitance directly translates to variation in transistor saturation current. In addition, the gate leakage current is exponentially dependent on the oxide thickness [29, 30]. Therefore, a variation in oxide thickness will have a large impact on the

[7] As the channel length shrinks, the drain terminal begins to fight with the gate terminal for the control of the channel. This phenomenon is called the short channel effect.

variation in gate leakage current. The gate leakage current can become comparable to or even greater than the channel leakage current once we scale to the 65nm node and beyond, without switching to a new gate oxide material. In general, NMOS transistors suffer significantly more gate leakage than PMOS transistors. This is because the effective mass of electrons is much less than the effective mass of holes, thus making the probability of tunneling much higher.

One solution to mitigate the variability problem of gate leakage due to dielectric film thickness variation is to use a thicker oxide film. It is beneficial to use a thicker dielectric film for two reasons. First, since the gate leakage current is exponentially dependent on the dielectric film thickness, using a thicker oxide film will greatly reduce its magnitude. Second, even though the amount of variation (∼1-2 atomic layers) will stay roughly constant, it will be a much smaller quantity in terms of the percentage of total larger thickness. However, we cannot simply increase the thickness of dielectric film while still using S_iO_2 as a dielectric material, since we want to maintain electrostatic integrity (EI).[8] As a result, industry has switched to using high-k dielectric materials such as H_fO_2 to replace S_iO_2.

1.5 Rise of the Layout Dependence

The previous sections have focused on understanding the sources of variation and their impact on transistor parameters, and ultimately, the transistor performance. We want to reinforce the key message that scaling has increased the amount of variability and raised many challenges for us to overcome. In the 1990s, the definition of a transistor was very simple. The critical dimension was defined by the overlap region between the polysilicon gate and the dielectric layer. This information was sufficient to determine complete transistor characteristics, since all the necessary information is contained in the local layout geometry of the design. Nowadays, process control

[8]The electrostatic integrity of a device reflects its resistance to parasitic two-dimensional (2-D) effects such as short channel effect and drain-induced barrier lowering (DIBL) [32].

and process independence[9] between transistors have been difficult to achieve due to scaling. Throughout the entire manufacturing process, different processing steps can introduce different amounts of variability. Processes such as photolithography, etch, chemical-mechanical polishing, ion implantation, thermal annealing, and strain engineering contribute most to the variation. The industry has explored varieties of processing, design and layout techniques to help mitigate the impact of variation.

In the case of photolithography, the industry has migrated to subwavelength lithography, using state-of-the-art resolution enhancement techniques (RET) such as optical proximity correction (OPC) and phase-shift masking (PSM) to improve the resolution of final fabricated printout. Even with these enhanced techniques, the gaps between the drawn and the final fabricated layout are still getting wider as technology scales. In the case of etch, tight design rules on the layout pattern of polysilicon, metals, and active area are in place at the design stage to avoid regions that can potentially cause failures during fabrication. In addition, post-layout dummy fill tools are used to reduce systematic variation [16]. In the case of thermal annealing, recent research has shown that non-uniformity in the polysilicon layer can translate into a non-uniformity in temperature profile during the annealing process [5]. This non-uniformity can significantly influence dopant diffusion and dopant activation process. A polysilicon dummy fill technique can be used in this case to improve the temperature profile, as shown in Figure 1-8.

As we continue to scale, more local layout features (or short-range features) and long-range layout patterns need to be incorporated into transistor modeling. The radius of influence that defines the neighborhood of shapes which play a part in determining the characteristics of a MOSFET also needs to be extracted. So far, there have been no reports on the magnitude of this parameter and also no indication of whether the radius of influence will increase or decrease with critical dimension scaling. Assuming that the radius of influence stays constant with scaling, the enclosed features within the area defined by the radius of influence will still increase just be-

[9]Process independence refers to a situation in which the local transistor layout geometry is not affected by its neighboring layout geometry.

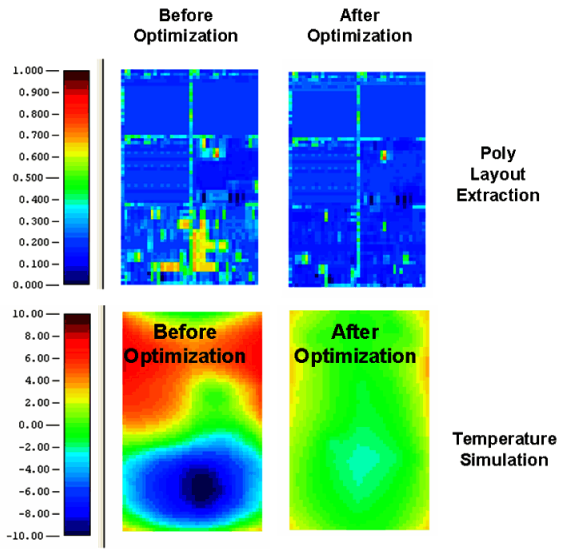

Figure 1-8: Polysilicon dummy fill optimization to improve temperature uniformity [5].

cause of scaling. This implies that approximately every 18 months, the number of surrounding features that must be taken into account when modeling a transistor doubles. Currently, the tools used to perform a parasitic extraction of a layout are quite slow. If we need to take into account more layout details in the future, the extraction of a simple design can take considerably more time compared to previous generation with a quadratic increase. The post-layout extraction simulation including these layout details can take even longer than the extraction time, making the entire design flow very time-consuming.

Finally, designers have to be aware of variation issues when designing not just analog circuits but all types of circuits. The traditional concerns in analog circuits, such as matching and offsets, are further aggravated with scaling. Problems such as frequency fluctuation and leakage power variation also limit digital circuit designs. Now, because of the radius of influence, different functional blocks on one die may interact with each other during the manufacturing process. Because different functional blocks (for example, logic vs. memory vs. ADC) tend to look very different in terms of layout, it can be difficult to enforce layout uniformity among them. Simulation of individual circuit blocks may no longer be enough to capture its behavior properly. A

complex simulation that takes into account all the long-range and short-range layout effects may be needed for accurate circuit simulation.

The trends highlight the increasing need to understand the layout-induced variation at the individual transistor level. This variability is closely related to many process steps, as discussed in the previous sections. The radius of influence is one of the parameters that needs to be extracted and studied to aid future modeling. This parameter and others will be the main focus of this thesis.

1.6 Thesis Organization

A robust circuit design relies greatly on a thorough understanding and characterization of process and layout pattern variations. In this chapter, we present a historical perspective on the studies of transistor variability. Although there is a tendency in the literature to view variation as a new challenge associated with CMOS technology, process variation has always been a vital concern throughout the history of semiconductor process engineering. The problem of process variation does, however, become more significant as technology scales. We describe process variation from various aspects according to cause, behavior, extraction techniques and mitigation solution. A brief overview of the semiconductor manufacturing procedure identifies the processing steps that contribute significantly more variation. We illustrate the impacts of these variation sources on the physical transistor parameters and how these parameters ultimately influence the transistor performance. Finally, from all the analysis and discussion above, we point out the rise of layout context dependence in transistor performance and also the importance of understanding and extracting the parameters related to the layout context. Layout related parameters, such as the radius of influence, can help us tremendously when it comes to transistor modeling.

As a result of the previous analysis, we will focus on the study of layout-induced systematic variation in this thesis. More specifically, we want to investigate how layout context can affect semiconductor processing steps, such as rapid thermal annealing (RTA), etching, and STI induced unintentional strain. Even though we mainly

concentrate our study on systematic variation, our analysis will not be limited to just systematic variation. Studies on random variation sources will also be done. In Chapter 2, we discuss a test structure that is designed to have different pattern density over different regions of the die in order to assist us with the study of layout induced variability. Different layout strategies including the micro-layout and macro-layout techniques will be discussed in detail. A step response emulated by a rapid change in pattern density is also designed in the layout for future modeling.

In Chapter 3, we will discuss our test circuit design. The test circuit is designed to perform the measurement on the test structure described in the previous chapter. We will first discuss a new accessing scheme that enables us to measure the complete I-V characteristics of individual transistors. This accessing scheme uses the minimum number of peripheral transistors (helpers) around the devices under test (DUT) compared to other designs found in the literature so far. It also enables us to maximize the ratio of DUTs to peripheral transistors for maximum usage of the die area. In order to perform the measurement, two different measurement schemes are designed. The first measurement scheme is direct probing, in which the fabricated dies are not required to be packaged. The measurement needs to be done on a voltage/current probing station, which is commonly accessible in most fabs. The second measurement scheme is on-chip voltage/current measurement. The fabricated dies are required to be packaged in this case. A printed circuit board and a programmed field-programmable gate array (FPGA) are required in order to assist the current measurement on-chip. We will also discuss a design of high dynamic range current measurement ADCs which can measure the current from 50nA to 1mA. A number of layout techniques to improve matching in sensitive analog circuits will also be mentioned and discussed. At last, we compare the two proposed measurement schemes and point out the pros and cons, particularly related to measurement speed and number of DUTs that can be tested in each approach.

Chapter 4 concludes this thesis with the evaluation of the overall contribution. This chapter will end with suggested future work, such as the test circuit design for on-chip parasitic capacitance measurement.

Chapter 2

Design of Test Structure

In the previous chapter, we described different aspects of variation, different sources of variation, and the impacts of these variation sources on transistor parameters. We then pointed out the trend that transistor performance will rely much more heavily on its layout context in the future compared to past CMOS technologies. In order to arrive at a quantitative understanding of these variation mechanisms and further model these in simulation tools, a careful design of an extraction technique is necessary.

Designing a test structure to perform measurement has always been the most common technique to help control, characterize, and model the behavior of transistors. Depending on the purpose of the test structure, it can either be simple or complex. For example, test structures designed for process monitoring purposes are usually simple. These kinds of test structures consist of simple dimensional changes between transistors and are usually placed in the scribe line on the wafer. Simple measurements are performed to determine parameters such as saturation current and resistivities of wires and vias to detect shifts in process conditions. Another kind of test structure can be used to perform a full I-V characterization for transistors in order to build a transistor model. It consists of a much richer collection of devices under test using different transistor dimensions, different surrounding layout geometries, different layout orientations, etc., to capture systematic variability effects.

This chapter is organized as follows. In Section 2.1, we motivate the importance of designing an effective test structure and discuss a few of the existing test structures

from the literature. We will analyze problems related to these test structures and ways to improve upon them. Then, we will shift our focus to the transistor parameters which we want to study and the reasons behind choosing these parameters. In Section 2.2, we present a comprehensive analysis on the layout strategy, which is divided into two parts: macro-layout strategy in Section 2.3 and micro-layout strategy in Section 2.4. Finally, we will summarize this chapter in Section 2.5.

2.1 Motivation

Designing an effective test structure is not a simple task. Failure to design an appropriate test structure can mislead us to conclude the existence of certain variation phenomena which may not exist or attribute a particular behavior to the wrong variation source. Difficulties usually arise from the fact that a large number of semiconductor processing steps influence the magnitude and specific behavior of variability mechanisms. Test structures need to be designed in such a way that it is still possible to distinguish between different variation sources.

The major figures of merit associated with test structure designs are applicability, genuineness, cost, count, and specificity.

1. Applicability is the most important figure of merit in test structure design. We want to make sure the information we obtain from a test structure can be applicable to other circuits which may be designed in the future. If the result obtained from a test structure consists only of the characterization of one particular test structure, this information cannot be used for general modeling as discussed in Section 1.2.3.

2. Design genuineness emphasizes the importance of careful test structure layout and design. Unintentional offsets in the test structure due to careless layout errors can occur. Researchers can mistakenly treat that as a real source of systematic variation when it is in fact only an offset created by the test structure itself.

3. The cost of the test structure consists of two parts: area and time. Area refers to the number of DUTs per unit area. Time refers to the amount of time needed to finish measurement on the test structure. In general, we want to lower the overall cost while still being able to efficiently gather enough information for design modeling.

4. The count refers to the amount of data we can collect from the test structure. Since we are doing statistical analysis on systematic and random variation, it is necessary to have an adequate number of samples so that the results will be statistically significant.

5. Some test structures provide a clear indication of variation sources such as the measurement of contact resistances [54], while many other test structures only provide a lumped parameter that contains several factors, as in the measurement of ring oscillator frequency [86]. Depending on the amount of detail we need to extract and the purpose of the experiment, different test structures can be designed to accommodate different degrees of specificity.

2.1.1 The Parameters

As mentioned in Section 1.5, the performance of a transistor increasingly depends on its layout context. As technology continues to scale, it is essential to comprehend and model the interaction between design and manufacturing steps on an individual transistor level (specificity). In this thesis, we will focus on the study of systematic process variation, especially layout-induced systematic variation. However, one key step is to determine which parameters are important. We can choose from different layout patterns and geometries; each layout practice is geared toward understanding one or more specific variation mechanisms. This thesis focuses on studying variation mechanisms associated with shallow trench isolation (STI), and etching and annealing processes. Detailed explanation of our choices of these specific variation mechanisms is presented in the following section.

STI: Carrier Velocity and Mobility

Variation associated with STI mainly refers to the variation caused by the strain induced by STI regions of neighboring transistors. This strain can change the internal crystalline structure of the silicon, and thereby either increase or decrease the effective mass of the carriers depending on the direction of the strain. The changes in the effective mass will ultimately affect both the mobility and velocity of the carriers inside the silicon. As discussed in Section 1.3.4, strain plays a key role in improving transistor performance. However, as technology scales, the distance between neighboring layout features becomes smaller; therefore, unintentional strain is unavoidable and will affect transistor performance.

In order to understand the impact of unintentional stress on transistor performance, it is first necessary to understand carrier velocity and mobility. Carrier velocity refers to the virtual source velocity; it is the velocity of carriers located in the MOSFET channel at the top of the barrier near the source (virtual source), denoted as v_{xo} in Figure 2-1. It has been the main driving force for improved transistor performance [82]. Surprisingly, however, no analysis on the variation of carrier velocity has been reported. Most reports are focused on other transistor parameters, such as threshold voltage [7, 56], saturation and leakage current [44, 51], and channel length [66] variation. Virtual source velocity had been extracted and studied before [80]. However, this report does not study the effect of variation; rather, the extracted parameters are averaged. Relatively little work has been done on the extraction of virtual source velocity and its variation analysis. Studying the more traditional transistor parameters does have an advantage in the sense that most of these parameters can easily fit into models such as BSIM [57]. However, virtual source velocity is a key parameter to examine because it directly reflects the trend in performance improvement from generation to generation.

Strained silicon technology is necessary since scaling itself can no longer improve the virtual source velocity. This is illustrated in Figure 2-1 and 2-2. Here, v_θ is the unidirectional thermal velocity (or ballistic velocity), B is the ballistic efficiency, r is

the reflection coefficient, and l is the scattering length.

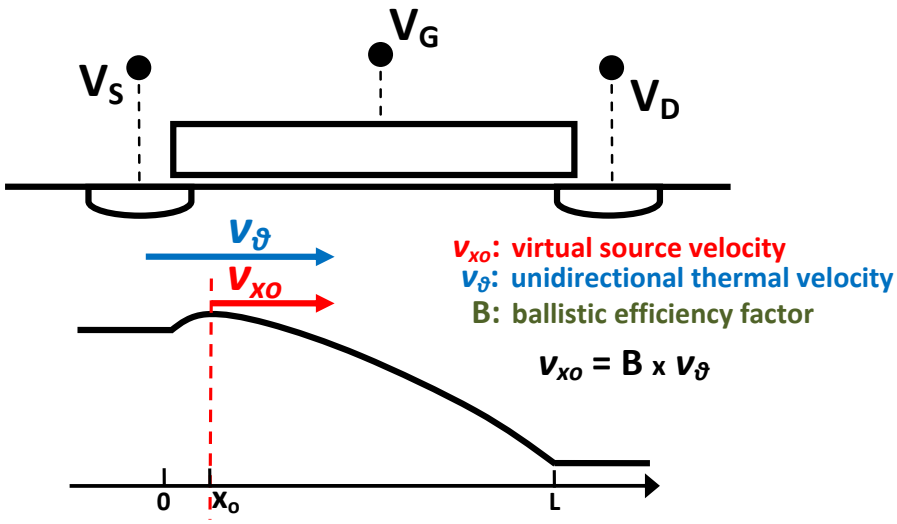

Figure 2-1: Virtual source velocity vs. thermal velocity. After [6].

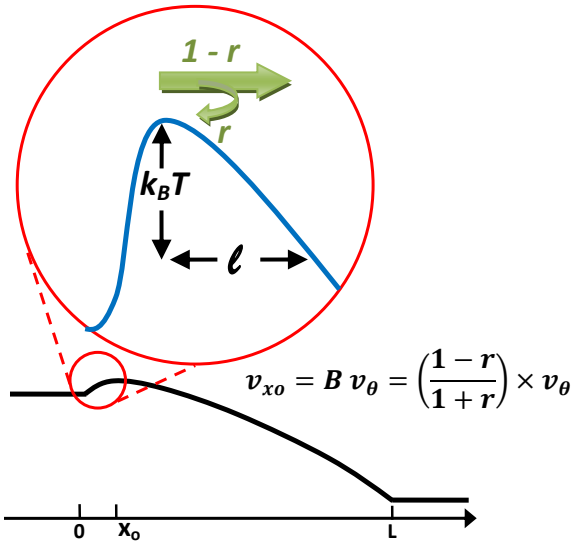

Figure 2-2: Ballistic efficiency. After [6].

In the past, virtual source velocity has increased due to increases in ballistic efficiency, B. As transistor channel length decreases, the reflection coefficient, r, decreases and carrier transport becomes more efficient. However, the ballistic efficiency saturates at around 60% for state-of-the-art silicon MOSFETs due to columbic scattering that results from the increased doping levels necessary to maintain electrostatic integrity. Therefore, in order to keep up with Moore's Law, it is necessary to increase

virtual source velocity by other means.

For gate lengths below 130nm, the saturation in ballistic efficiency has forced designers to look for different ways to increase the virtual source velocity. Looking at the equation presented in Figure 2-2, another way to improve the virtual source velocity is to improve the ballistic velocity itself. Ballistic velocity is a fundamental property of a material. Innovations in strain engineering, for example, can change the effective mass of carriers in a material and alter the ballistic velocity [85, 78]. Therefore, even though we cannot improve the ballistic efficiency further, we can still increase the virtual source velocity by increasing the ballistic velocity.

In addition to virtual source velocity, mobility is another critical transistor parameter which must be understood. Researchers in the modeling community have started to incorporate mobility variation into the widely used predictive technology model [73, 74]; the authors also emphasize the importance of understanding mobility variation. The ballistic velocity is inversely proportional to the square root of both effective mass and the density of states mass [81]. While mobility is inversely proportional to the effective mass of the carriers, we can expect a power law relationship between the ballistic velocity and the mobility. The mobility information can therefore be inferred based on the extracted virtual source velocity.

Looking at these new sets of parameters is vital for the understanding of performance improvement in future scaling. Variation in them directly correlates with the variation in performance. If our study shows that the increase in variability of these parameters is too high from one generation to another, we must assess whether or not it is worth moving to a new technology node, since performance and yield benefits may be degraded a lost altogether.

Etching and Thermal Annealing

Section 1.3.2 highlight the impacts of etch on transistor dimension variability. The variation mainly originates from the difference in surrounding feature geometries, feature patterns, and feature sizes.

Many studies have been done to investigate the variation in semiconductor man-

ufacturing due to etch. The dependence of the etch rate on layout patterns, for example, was even studied in 1992 [25]. The advent of advanced rapid thermal annealing (RTA) process has continued the need for such work. As technology advances, a faster annealing process is needed in order to form a shallower drain/source region to reduce the short-channel effect and maintain the same electrostatic integrity [68, 60]. This so-called "spike" RTA processing can be used to heat up the region of interest to the desired temperature in a very short period of time compared to more traditional annealing processes, such as flash-lamp anneal [62, 63]. However, since the length scale over which thermal equilibrium can be reached for a given time is in the millimeter range, the specific thermal profile created by device layout patterns may affect all transistors within that range. An investigation of the impact of RTA on process variation showed that device variation is highly correlated with layout pattern density [67].

Due to the use of this new annealing process, we believe it is necessary to design a new test structure that can be used to simultaneously study the dependences of etch and thermal anneal on transistor performance. Because both phenomena are a strong function of layout context, it will be helpful to study the interaction between the two. In this thesis, we want to examine two aspects of each of these variation sources: the range of influence and the magnitude of influence. One source can have longer range of influence but be smaller in magnitude, while the other source can have shorter range of influence but be larger in magnitude. To uncover these possible trends, we see the need to design an experiment.

2.1.2 The Need for a New Test Structure

The previous section illustrated the need to study the systematic variation induced by the STI as well as the etching and thermal annealing processing steps. Some test structures have been presented in the literature, but none of them is sufficient for studying systematic variation. Some of these approaches will be highlighted here.

Previous Test Structure for Etching & Thermal Annealing

In [67], a test structure is built to study the intra-die variation of the delay for a CMOS inverter in a 65nm technology, driven by millimeter-scale variations of RTA. A set of identical circuits is placed at different locations on the die and each set has a distinct pattern density. Different inverter delays are observed at each location on the die with only minor changes in pattern density. The authors attribute the observed systematic variation as resulting from the strong correlation between pattern density and local dopant activation using RTA. However, the interactions between the RTA and etch are not discussed and the methodology used is not suitable for a large number of devices.

Previous Test Structure for STI Strain

There are also many test structures designed to understand STI strain. In [40], the authors experiment with different transistor layouts to quantify the variations induced by stress. The amount of stress is controlled by the shape of the diffusion area and the density of the adjacent layout. Noticeable transistor performance changes are observed due to the changes in local layout shape and local layout density. Possible solutions to reduce the stress-induced performance variations, such as adjusting the shape of the transistor layouts, are also proposed in the paper.

The authors in [41, 42] utilize STI-induced stress to exploit possible ways to optimize transistor performance. They begin by building a model of the stress induced by different STI widths, which is then incorporated into the transistor mobilities in the BSIM model. TCAD process simulations are used for the generation and confirmation of the stress models. This optimization methodology is built into the standard cell generation flow implementing dummy diffusion placement [43]. The resulting optimization flow can increase the performance by 7% to 11% without any area penalty.

In these cases, the authors in [40] and [41, 42] only consider the influence of short-range layout shapes and features on transistors. The experiment and analysis for long-range effects need to be performed in order to build a more complete model.

A New Test Structure

The previously discussed test structure designs for both etching and thermal annealing and for STI strain are not sufficient for our purpose stated in Section 2.1.1. Our test structure attempts to correct these limitations and incorporate a number of new features, which are as follows:

1. *Clarity*: Many papers used the term "pattern density" but do not define it. We want to distinguish the difference between polysilicon pattern density and STI pattern density in our design of experiments to better separate the influences of each on device performance.

2. *Statistical Significance*: The test structure design needs to have more devices under test in order to increase the statistical significance of the result.

3. *Full-Range Length Scale*: In order to fully characterize the variation, it is necessary to understand both the short-range and the long-range layout pattern dependence. Some papers focus only on the influence of short-range layout features [38] while others focus only on long-range layout features [67]. Our test structure finds multiple ranges of influence and determines the magnitude of influence for each range.

4. *Individual Devices*: Our test structure will be designed to understand transistor characteristics on an individual transistor basis, rather than measuring a lumped parameter, such as the oscillator frequency.

2.2 Layout Strategy

Innovations in test structure design are needed to fulfill all of the targeted features listed in the previous section. In this section, we will discuss the test structure design procedure. The discussion will be divided into two parts: macro-layout strategies and micro-layout strategies. Comprehensive analysis on the choice of our final test structure layout is presented in the section below.

Before discussing the details of the actual layout strategy, it is important to understand the reasons for dividing the overall layout strategy into macro-layout and micro-layout strategies. The variation mechanisms caused by macro-layout and micro-layout are fundamentally different. At the macro-layout scale, the systematic device variation is determined by the overall surrounding layout features scaled by some appropriate weighting function. For example, two identical transistors can have very different device performance characteristics due to differences in the surrounding layout features. On the other hand, at the micro-layout scale, transistors having different device layout parameters while having the same neighboring layout are studied. It is necessary to distinguish these fundamentally different variation sources for modeling in the future. Keeping different sources as separated as possible can simplify modeling significantly.

2.3 Macro-Layout

Macro-layout strategy refers to a layout strategy on a large spatial scale. This strategy mainly concerns itself with average layout pattern and layout features over a large range of the die; it does not include the studies of individual transistor dimensions or layouts. One of the main focuses that belongs to macro-layout strategy in this thesis is the design of pattern density and the relative location of each pattern density region.

2.3.1 Effective Pattern Density and Filter Weighting Functions

We should recall from Section 2.1.2 that our goal is to build a test structure to examine the systematic variation induced by "STI strain" and "RTA and etching" on an individual transistor basis. For each transistor, in the macro sense, we want to be able to relate the performance with its effective pattern density. Depending on the type of systematic variation we want to investigate, the definition of effective pattern density can change accordingly. For example, in the RTA and etching case,

the definition of pattern density is the polysilicon pattern density.

Whether polysilicon or STI pattern density is considered, we need a formal way to calculate the effective pattern density. One of the most common ways to perform this calculation is to use a filter weighting function. One example of a weighting function is shown in Figure 2-3. The weighting function inherits two critical parameters, the range of influence and the magnitude of influence. In this particular example, the filter has a Gaussian shape, which implies that the surrounding layout features which are closer to the transistor are more highly weighted when calculating the effective pattern density than the layout features which are farther away. This distance dependence decays fast because of the nature of the Gaussian shape. The range of influence is determined by the standard deviation of the Gaussian filter; the magnitude of influence is determined by the magnitude of the Gaussian filter. In [26], Ouma presented some other possible filter weighting functions, such as square, cylindrical, and elliptic functions for the characterization of oxide chemical-mechanical polishing. In this case, it was found that the elliptical and Gaussian shapes have the best performance.

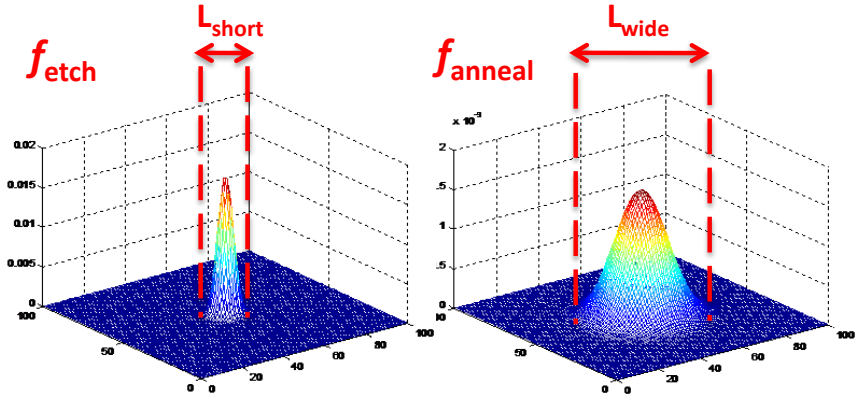

Figure 2-3: Sample filter weighting function: Gaussian-shape filter.

$$\rho_{etch} = f_{etch} \otimes f_{layout} \tag{2.1}$$

$$\rho_{anneal} = f_{anneal} \otimes f_{layout} \tag{2.2}$$

$$\Delta I_{dsat} = (a_0 + b_0 \cdot \rho_{etch} + c_0 \cdot \rho_{etch}^2) + (a_1 + b_1 \cdot \rho_{anneal} + c_1 \cdot \rho_{anneal}^2) \tag{2.3}$$

The goals of our macro-layout strategy are twofold: to identify the kind of filter weighting function which works best, and to find the range and the magnitude of influence for that particular filter weighting function. Figure 2-3 shows two Gaussian-shape filters. Equation 2.1-2.3 are the potential results we are looking for at the end of the experiment. f_{layout} is the layout pattern density, ρ_{etch} is the effective pattern density of etch, ρ_{anneal} is the effective pattern density of anneal, and ΔI_{dsat} is change in saturation current. Two Gaussian-shape filters with different filter characteristic length and different weighting magnitude are found. The etching-related filter characteristic length is shorter than the one invoked by annealing in this hypothetical case. The resulting pattern density is the convolution between the filters and the actual layout. A hypothetical quadratic current fitting equation is also provided in Equation 2.3.[1]

2.3.2 A Pattern Density Step Layout

In order to obtain the results we are looking for, we need to build a test structure that can accentuate the effect we want to see. The best input test structure to fully characterize a linear system is an impulse, as shown on the left side of Figure 2-4. An ideal impulse contains equal amount of frequency content over the entire spectrum; therefore, it is easy to see how the system responds to individual frequency components. However, it is impossible to implement a real impulse input in the layout as the width of the impulse cannot be designed to be infinitely small, and the maximum magnitude is limited as well (to 100%). As a result, in our layout, we use a pattern density step input, as shown on the right side of Figure 2-4, instead. Under some assumptions, the impulse response can be derived from the semi-one-dimensional step response, since the derivative of the ideal step response gives the impulse response. In order to obtain enough information to build a precise model, a high sampling rate along the step input and step response is also needed.

[1]The quadratic equation here is just an example; the final fitting equation does not have to be quadratic.

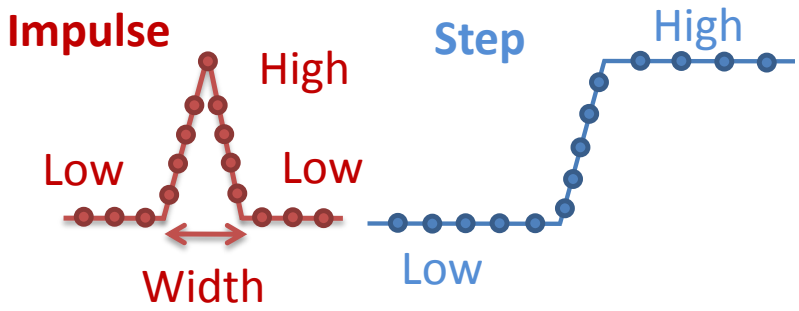

Figure 2-4: Step input vs. impulse input.

2.3.3 Assumptions and Challenges

After deciding to use a step input, the size of step region, the number of step regions and the height of the step input must be determined. To properly do this, a number of assumptions must be made to narrow down the potential possibilities for the test structure design.

The first assumption we need to make is shown in Figure 2-5. Point A and point B are both located in the center of a square layout region. The percentage represents either STI or polysilicon pattern density within that region. We can see that point A is in the region that has a 50% pattern density throughout, while point B is in the middle of an area that also has a 50% pattern density when computed using any symmetric filter weighting function. In our test structure design, we will not differentiate between the two cases, and thus assume the transistors in both locations behave the same.

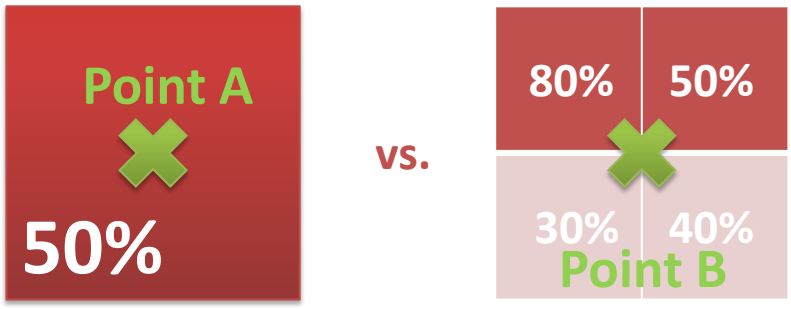

Figure 2-5: Design assumption I: point A and point B has the same effective pattern density.

The second assumption is that there are at most two dominant filter weighting

functions, one with a long characteristic length and the other one with a short characteristic length, as shown in Figure 2-3. We are not precluding the possibility of having more than two filter weighting functions, but rather the test structure cannot accommodate more than two filters that have very different characteristic lengths. Moreover, the test structure also cannot extract weighting functions whose characteristic lengths are much longer than the projected value. The assumption that there are just two filter lengths is based on literature studies. Altering polysilicon density, for example, will influence two processing steps: etching and thermal annealing processes. One has a long characteristic length and the other has a short characteristic length.

With these assumptions in mind, Figure 2-6 illustrates a key challenge in this test structure design. As an example, three square layout regions are shown. Each large region also contains a smaller square region of three different sizes. The outer red square represents the global (or long-range) pattern density and the inner green square represents the local (or short-range) pattern density.

Figure 2-6: Test structure design challenge: sizing of the inner green square.

Correctly choosing the size of the local pattern density region is a significant challenge. Using the previous assumptions, each device has two independent variables, long-range pattern density and short-range pattern density, which affect the transistor performance. In addition to the step response requirement, a good design of experiments[2] also needs to cover the design space indicated in Figure 2-7. For the same device type, we require all combinations of high and low short-range and long-range

[2]Design of experiments, or experimental design, is a structured, organized method for information gathering to determine the relationship between factors affecting a process.

pattern density. In Figure 2-6, if the local density region in green is much smaller than the short-range filter's characteristic length, then the real effective local density will include too much global density, resulting in no low local pattern density in the design of experiments. On the other hand, if the local density region in green is much larger than the short-range filter's characteristic length, then the real effective global density will include too much local density resulting in no high global pattern density in the design of experiments. It is important to find the balance between the two; however, it is difficult since we do not know the characteristic lengths of the filters in advance.

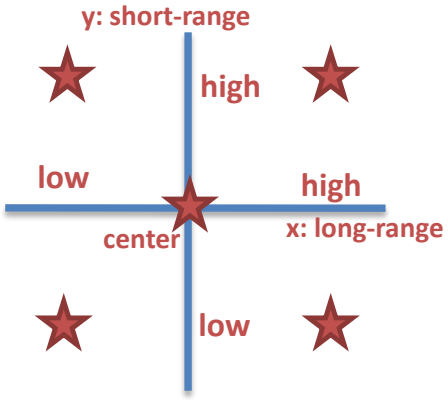

Figure 2-7: Good practice of design of experiments (DOE).

2.3.4 Test Structure Layout Proposals

Test Structure Proposal I

The first test structure proposal is shown in Figure 2-8. Due to the simplicity in this layout, it is fairly easy to extract the characteristic lengths of the filter. The design and layout of the test structure is also easy due to its regularity. The problem with this test structure layout is that the short-range and the long-range pattern density always increase in the same direction; therefore, there are no devices with low long-range pattern density and high short-range pattern density or vice versa. The second problem with this test structure design is that it is very area expensive. Each square needs to be at least as long as the longer characteristic length.

90%	10%	80%
10%	70%	20%
90%	30%	50%

Figure 2-8: Test structure layout proposal I.

Test Structure Proposal II

The second proposal, shown in Figure 2-9, has the same local density in the row direction and the same global density in the column direction. One advantage of this structure is that it has a rich combination of long-range and short-range pattern densities. The problem associated with this test structure is area overhead because of the large number of local pattern density regions used and a minimum distance required between of the two local pattern density regions. Moreover, it is difficult to design and lay out this test structure. It may not be necessary to have such a large number of regions to create a good design of experiments.

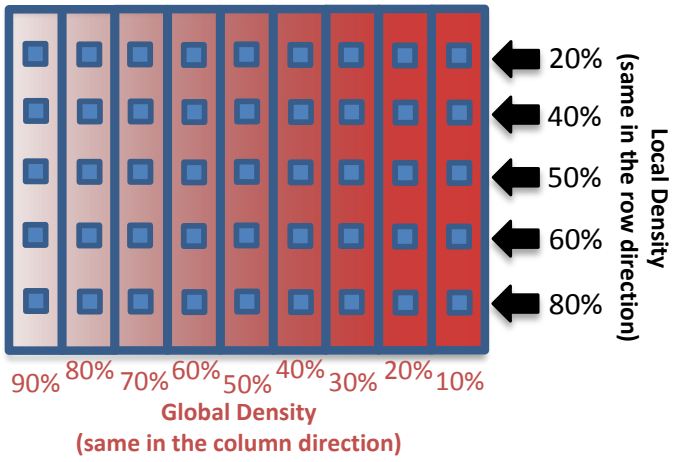

Figure 2-9: Test structure layout proposal II.

Test Structure Proposal III

The third proposal is shown on the left of Figure 2-10. As mentioned previously, it is difficult to design the size of the pattern density region because we do not know the characteristic length of the filter. This structure is designed to overcome that difficulty. Since we do not know the characteristic length, we continuously decrease the size of each region, counting on the fact that at least one of the regions will have the correct size. Even though all the regions are getting smaller, they are also getting closer to one another. However, we want the regions to get smaller, but also farther away from one another so nearby regions do not influence the calculation of long-range or short-range effective pattern density. Therefore, what we really need is shown on the right side of Figure 2-10.

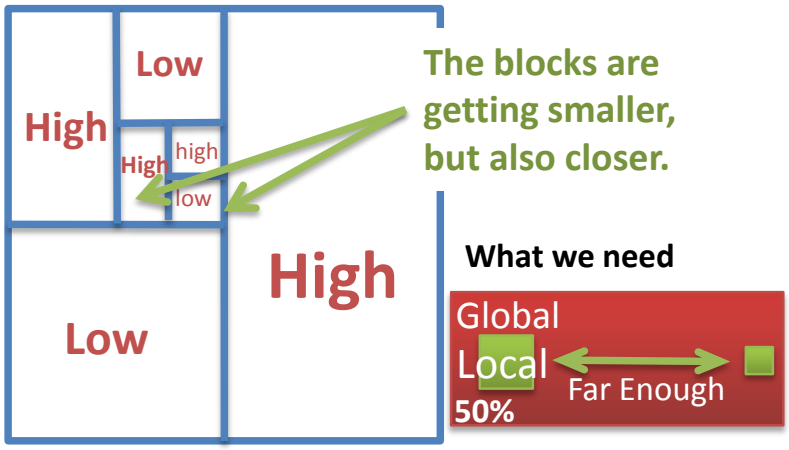

Figure 2-10: Test structure layout proposal III.

2.3.5 Our Final Test Structure, on the Macro-Scale

The major problems of all the previously proposed test structures are the amount of area required and the incapability of covering the entire design space for a good design of experiments. Our final proposal intends to resolve this problem. We begin with two simple regions as shown on the upper left corner of Figure 2-11. We then generate a plot of global pattern density versus local pattern density, for a large number of spatial points within the layout, as shown at right in Figure 2-11. We see that we are

able to cover most of the design space except for the lower right corner: high local pattern density and low global pattern density. In order to cover this design space, we add an additional block with a smaller high pattern density region inside a larger region with low pattern density.

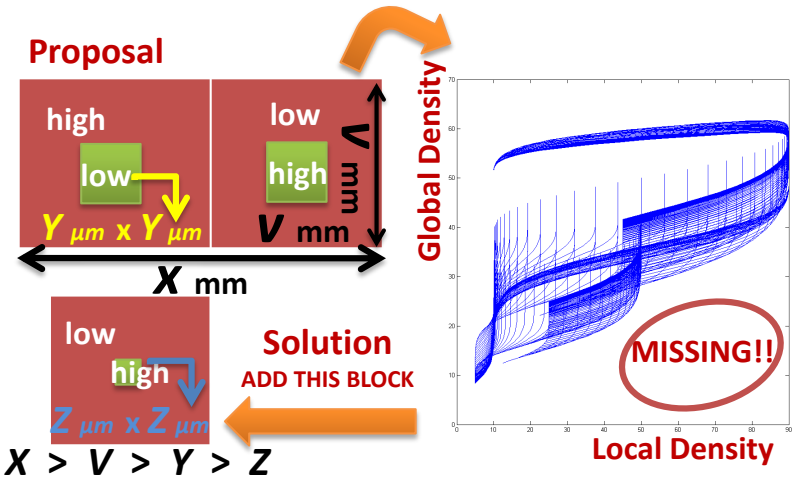

Figure 2-11: Proposed solution.

The final test structure layout is shown in Figure 2-12. The entire die is divided into six different regions: three for examining systematic variation due to polysilicon pattern density, and three for examining systematic variation due to STI pattern density. Step inputs are built into this test structure at specific locations around the die indicated by the blue arrows. Additional blocks in the middle row with small high pattern density regions inside a larger square are added for both polysilicon and for STI test parts to cover the remaining portion of the design space, as described earlier. Each region has the size of 1mm x 1mm. The larger square inside each region has the size of 400μm x 400μm, while the smaller square inside each region has the size of 100μm x 100μm. These sizes are our guesses on the characteristic lengths of the filters. We assume that 1mm is larger than the longer characteristic length and the shorter characteristic length is between 100μm and 400μm. This guess is based on our literature research.

In the ideal case, we would like to make the "high" pattern density as high as possible (probably close to 90 to 100%) and the "low" pattern density as low as

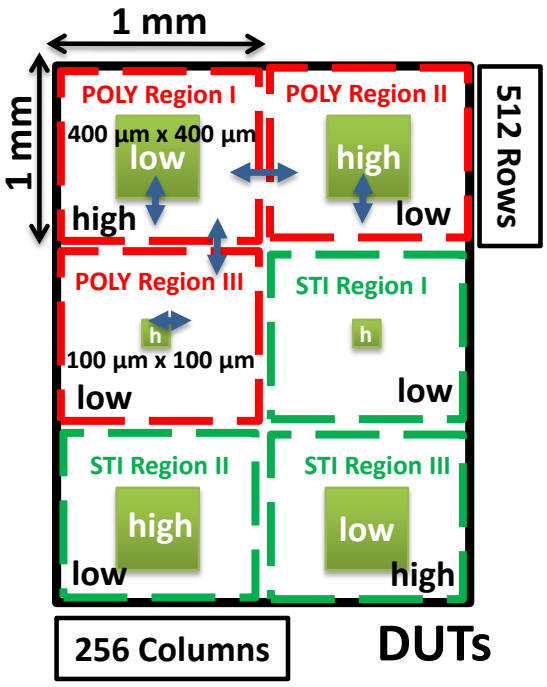

Figure 2-12: Final test structure on the macro-scale.

possible (probably close to 0 to 5%). Unfortunately, due to design rules and layout constraints, it is difficult to create regions having these densities. Therefore, in the "high" region, we are able to obtain about 60% pattern density, and in the "low" region, we are able to obtain about 10%.

2.4 Micro-Layout

Micro-layout strategy refers to layout dimensions and features of individual transistors. We explore the following design variables: transistor width, transistor length, active area size, number of polysilicon fingers, and finger spacing as shown in Figure 2-13. Different micro layout practices are used to accentuate one or more particular transistor variation sources.

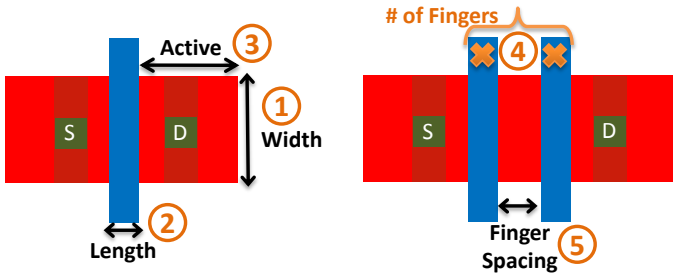

Figure 2-13: Micro-layout strategy.

2.4.1 Compromise Between Macro- and Micro-Layout

If each of the five layout parameters has two possible values, we will have a total of (2^5 =) 32 different combinations of unique transistor layout types. However, we do not intend to have this many types of transistor layout. Because our main purpose is to study the systematic effect of polysilicon and STI on transistor performance, it is important to have an entire spatial map so we can measure the performance on the same types of transistors at different locations of the die. Each location on the die is unique in terms of its long-range and short-range effective pattern density. As a result, in the ideal case, we would like to build a test structure that has only one micro-layout design type in order to accentuate the effect due to just the difference in macro-layout.

However, it is still important to see how the individual transistor layout change can affect transistor variability characteristics. One benefit of using different transistor dimensions is that it can also isolate different kinds of variation sources. For example, using a large transistor will help to diminish the effect of line-edge roughness on transistor performance.

In order to obtain benefits of both macro- and micro-layout design strategies, this design has a few variations of micro-layout design on top of the macro-layout design discussed in the previous section. A total of 10 different types of transistor layout are used in the test structure design, but there is not an equal number of each type. More minimum-sized transistors are used than the large transistors, to build a more complete spatial map for the transistors which are most often used in a real circuit

design.

Table 2.1 describes the 10 different transistor types used for micro-layout strategy. Each type has a different combination of the six layout parameters previously described. In this table, OD represents the distance between the polysilicon gate and the edge of the active area and PO represents the distance between two polysilicon fingers. Because all 10 combinations apply to both NMOS transistors and PMOS transistors, we have a total of 20 different device types. The reason for choosing each of these design variables as well as its value will be discussed in the next section.

DUT	**Width (nm)**	**Length (nm)**	**OD (nm)**	**# Fingers**	**PO (μm)**
1	200	60	195	1	N/A
2	400	60	195	1	N/A
3	200	60	500	1	N/A
4	200	180	195	1	N/A
5	200	180	195	2	220
6	200	180	195	3	220
7	1000	60	195	1	N/A
8	200	1000	195	1	N/A
9	1000	1000	295	1	N/A
10	200	180	195	2	500

Table 2.1: DUT design types.

2.4.2 Dimensional Differences

Different transistor sizes can be used to see how dimensional changes can influence the variability characteristics of a transistor. DUT1 is a reference transistor. DUT2 has a larger width; DUT4 has a larger length. A couple of other transistors are also used as monitoring transistors. For example, DUT7 has a very large width to accentuate the variation due to length; DUT8 has a very large length to accentuate the variation due to width; and DUT9 has both very large width and very large length to accentuate the variation effects due to other sources other than dimensional changes. DUT9 also can help to ensure that the variation sources of the smaller transistor are not purely due to photolithography variation because of the atomic dimensions needed to

be fabricated.[3] Therefore, we also refer to these larger size transistors as monitoring transistors since they are used to monitor a certain manufacturing process.

2.4.3 Number of Polysilicon Fingers

Different number of polysilicon fingers for some device types are used to study how the change in the number of polysilicon fingers can influence the variability characteristics of transistors. Transistors with the same effective total channel length but with different numbers of fingers can be used to examine a couple of variation effects. For instance, separating a transistor's polysilicon shape into multiple fingers can accentuate variation due to gate length variation.

Other effects such as stress can also be studied by changing the number of polysilicon fingers. Transistors with three separate polysilicon fingers may experience additional amounts of stress, introduced by the STI region outside the active area as compared to the transistor with only one longer polysilicon finger. This is because a polysilicon region closer to the edge of the active region generally experiences more stress as compared to one closer to the center of the active region. Transistors with three polysilicon fingers will have more finger areas that are near the edge than those with one polysilicon finger. Thus, even with the same effective total length, transistors with three polysilicon fingers may experience more stress.

For this case, DUT4 is the reference transistor with one polysilicon finger of 180nm. DUT5 has two 90nm polysilicon fingers for a total effective channel length of 180nm and DUT6 has three 60nm polysilicon fingers for a total effective channel length of 180nm.

2.4.4 Spacing between Fingers and Length of Active Region

Here, we examine two more parameters related to spacing: the spacing between polysilicon fingers and the spacing between a polysilicon finger and the active area edge (or the length of active region). The spacing between polysilicon fingers can be

[3]Described in more detail in Section 1.3.1.

used to study variation effects due to photolithography proximity [55]. The spacing between a polysilicon finger and the active area edge can be used to study the effect of stress induced by the STI region near the edge of the active region.

For the case of finger spacing, DUT5 is the reference transistor and for the case of active area length, DUT1 is the reference transistor. DUT10 increases the spacing between polysilicon fingers from 195nm in DUT5 to 500nm, and DUT3 increases the active region length from 195nm in DUT1 to 500nm.

2.4.5 DUT Layout Pattern and Number

As mentioned in Section 2.4.1, we do not have the same number of each DUT type listed in Table 2.1. We want to have more of the more commonly used transistors and fewer of the less commonly used transistors. In addition, some of the transistors are used as monitoring transistors. These monitoring transistors are not the main targets for modeling; as a result, they are not replicated many times.

Table 2.2 shows the layout pattern we use for this test structure. The number corresponds to the type of DUT described previously. From this pattern, we can see that there are many replicates of DUT1 and DUT2 compared to other DUT types. Because DUT7, DUT8, and DUT9 are used for monitoring purposes, they are not replicated much. This three-row block is repeated both horizontally and vertically to build the entire test structure.

Row	DUT Pattern							
1	1	1	3	2	1	4	2	2
2	7	5	6	8	10	5	6	9
3	1	1	3	2	1	4	2	2

Table 2.2: DUT layout pattern.

2.4.6 Using Micro-Layout to Achieve Macro-Layout

As described in Section 2.3, on a macro scale, our test structure is divided into six different regions to examine the systematic variation effect of both polysilicon and

STI pattern density. In order to create the pattern densities we need at the macro level, we need to start with designing a "test structure building block" on the micro level. The test structure building block is shown in Figure 2-14.

Each unit building block is 6μm wide and 8μm long. It consists of one NMOS transistor at the bottom and one PMOS transistor at the top. The empty area is to provide flexibility to change the pattern density to the desired pattern density in that region according to the specification of the macro design. Two examples are shown on the right side of the figure. The top example shows a low STI pattern density, and the bottom example shows a high polysilicon pattern density. For instance, in order to build a region of high polysilicon pattern density, all the unit building blocks in that region, regardless of the size of the DUTs, have to be laid out with the same high polysilicon pattern density.

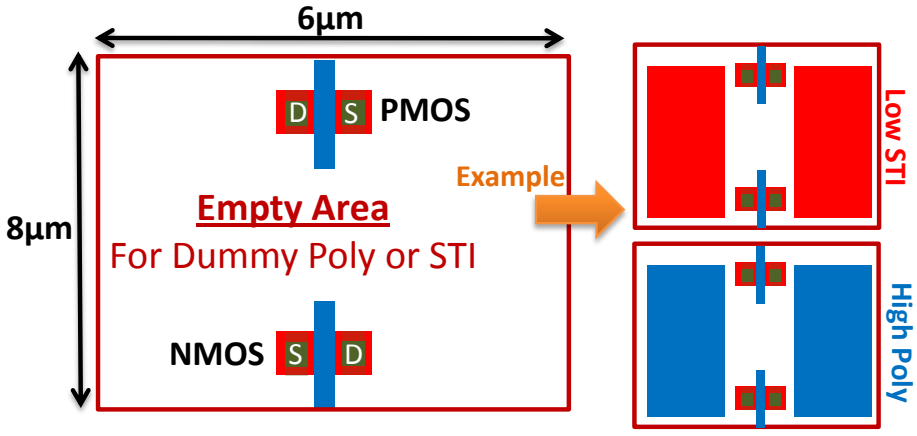

Figure 2-14: Test structure building block.

Since we have four different pattern density regions, high polysilicon density, low polysilicon density, high STI density and low STI density, and ten different kinds of DUTs, we will have a total of 40 different unit building blocks which will be used to build the entire test structure.

2.5 Summary

This chapter began with a motivation to study a set of transistor parameters which are becoming more important as transistor scaling continues, including virtual source velocity and mobility. We explained that it is necessary to observe these critical parameters because they are not only directly related to the steps of the new semiconductor manufacturing procedures, but they also directly capture the performance improvements from one technology generation to the next. We then presented a brief overview of the current test structure design in the literature and discussed why they are not sufficient.

A comprehensive analysis of our test structure design was presented. We began by explaining the fundamental differences between macro- and micro- test structure design. In the macro-layout discussion, we provided a definition of effective pattern density and discussed the essential layout features, such as a step input, which can accentuate the proper characteristics and help us extract this parameter more easily. Many test structures were proposed; we considered the pros and cons of each, and then we arrived at the final macro test structure design, which consists of six different pattern density layout regions and consumes a total area of 3mm x 2mm.

Since we are studying pattern density effect on the systematic variation of transistor, macro-layout is more important than micro-layout. However, we also discussed the benefits of experimenting with micro-layout strategy and the importance of not having too many variations of micro-layout. We explained the choices for each DUT type, and finally, we discussed how we can use micro-layout building blocks to achieve the pattern density we desire at the macro-level.

Chapter 3

Design of Test Circuits

Designing test circuits to study semiconductor device characteristics is becoming more important as transistors scale further. A growing number of possible mechanisms in the semiconductor manufacturing process contribute to the overall variability of transistor characteristics. Therefore, it becomes more difficult to predict the transistor behavior, particularly transistor variation effects, from one generation to another using basic scaling principles alone: it is almost necessary to design test circuits to extract the different variation sources that we are looking to model. In Chapter 2, we presented a test structure that we can use to study systematic variation due to pattern density changes in polysilicon and STI. In this chapter, we focus on building a test circuit that enables us to measure transistor current from the subthreshold region to the saturation region.

We will begin our discussion in Section 3.1 by motivating the need for a new test circuit to perform efficient measurements on the test structure designed in the previous chapter. Section 3.2 will then discuss a hierarchical transistor accessing scheme that can be used for both on-chip and off-chip current measurement schemes. An analysis of leakage mitigation techniques to ensure the accuracy of the current measurement is also provided. We will then explore different analog-to-digital converter (ADC) architectures in Section 3.3 and justify our final choice of the integrating ADC architecture in Section 3.4. A brief overview of the traditional integrating ADC architecture is presented to motivate the need for redesign in order to achieve high

dynamic range as well as offset and charge injection immunity. The design of the operational amplifier, comparator, current-steering DAC and switches, will be analyzed in Section 3.5. Section 3.6 overviews the algorithms for on-chip current measurement.

An alternative off-chip direct probing approach, which is compared with the on-chip ADC approach, is discussed in Section 3.7. The analysis of the pros and cons of the on-chip and off-chip measurement approaches are provided in Section 3.7.2. Finally, Section 3.8 summarizes this chapter.

3.1 Motivation

Our goal is to collect large amounts of information in order to increase the statistical significance of the data we obtain. The characterization also has to be done on an individual transistor basis. Therefore, we need to design a test circuit architecture that is able to measure every DUT on the test structure independently. The measurement must be done in a reasonable of amount time and with high enough accuracy for future modeling. In other words, we need a test circuit that is accurate, efficient and able to measure each transistor independently.

We divide the test circuit design into two parts. The first part is the architecture of arrangement. As described in Section 2.4.6, a unit building block consists of an NMOS DUT and a PMOS DUT. The entire test structure is built by tiling these unit building blocks together. The arrangement is not a trivial task. First, the placement of each DUT has to be dense enough to obtain the desired spatial resolution. Second, this arrangement has to ensure measurement independence between DUTs. For instance, when measuring transistor A, the leakage current coming from other transistors must be minimized to ensure the accuracy of the measurement on transistor A. A good architecture of arrangement can assure the independence between different transistor measurements.

The second part is the architecture of measurement, where key concerns are the accuracy and efficiency of the measurement. This discussion is partitioned into two parts: on-chip current measurement and off-chip current measurement. In our ap-

proach, we propose two different chip designs, one to enable early measurement of variation using wafer probing but with decreased spatial sampling, and a second to enable higher speed and more complete measurement of all DUTs using packaged chips. In our case, we thus need both on-chip and off-chip current measurements. Despite using different current measurement schemes, the two designs use the same test structure and architecture of arrangement. Using two different measurement schemes will also enable us to compare the outcome of the two and confirm that the conclusions drawn from both schemes are the same. No studies, which perform this comparison, have been found in the literature yet.

Due to parameters we want to extract and study, current measurements from the subthreshold regime all the way to the saturation regime are required. Over that range, the current can vary over four orders of magnitude, from 50nA to 1mA. This is one of the main challenges for the on-chip current measurement design. As a result, our discussion of the on-chip architecture of measurement will mainly focus on the circuit and architecture strategy to overcome this challenge. This challenge, on the other hand, is less of a problem for off-chip current measurement, since high precision test equipment (though expensive) is readily available. For off-chip current measurement, the main challenge is restriction on the number of probe pads available to the designer. The number of available probe pads is usually much smaller than the number of output pins on a packaged die.

In the section below, we will review a few test circuit designs in the literature to see why these test circuits are not sufficient for our purpose, to motivate further why we need to design a new test circuit.

3.1.1 Previous Test Circuit Design

Drego et al. design a dedicated test circuit to study threshold-voltage variation [7]. The current measurement is done in the subthreshold regime and at low values of V_{ds}, where the current is at least ten times more sensitive to V_T variation than it is to L variation. This allows the authors to separate the variation in V_T from the variation in L. A hierarchical access scheme, shown in Figure 3-1, divides the entire die into six

sections, each having 90 rows and 127 columns, and allows access to the individual transistor. An on-chip dual-slope ADC is used to perform the current measurement of around 70K transistors. All the analog measurements are done on-chip and only the digitized output comes off-chip.

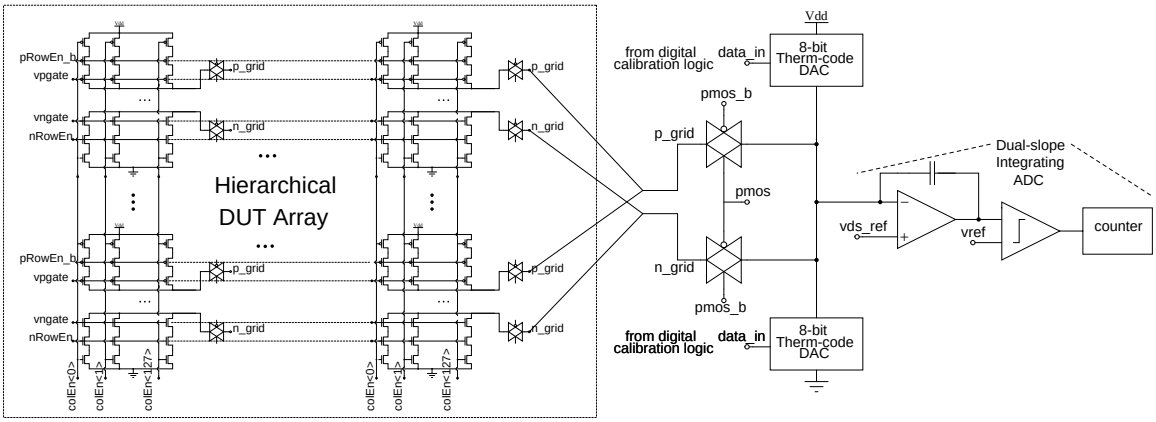

Figure 3-1: Hierarchical access scheme in [7].

However, our test circuit design is not a specialized test circuit for only threshold voltage extraction as in [7]. The range of current we need to measure is orders of magnitude higher than in the previous work, and therefore we cannot use the same architecture for our current measurement. Moreover, the hierarchical access scheme requires at least two additional transistors per DUT. This implies that the peripheral transistors will take up twice the area compared to the actual DUTs. Since we also need to ensure that the pattern density within a region stays constant, the layout of the peripheral transistors also needs to change according to the regions they are in. This can significantly increase the design complexity. Lastly, the peripheral transistors also introduce resistance and variation.

In [56], Agarwal et al. design a test structure for characterizing local device mismatches. Memory-like column and row decoders are implemented to access the desired DUT in the large array. Here, all the current measurements are performed by off-chip equipment. An indirect sensing of the drain voltage is necessary in order to measure the exact drain voltage applied on the DUTs, because there is no active circuit block, such as an operational amplifier, to force the node voltage. A memory-

like accessing scheme is adopted here, so fewer peripheral transistors are needed than in the hierarchical access scheme before. This results in a much denser layout of the DUTs. However, the test circuit proposed by Agarwal is designed mainly for off-chip voltage sensing and current measurements.

3.1.2 New Test Circuit Design Features

None of the previously designed test circuits is adequate. Our test circuit design will have the following features to overcome the limitations of other test circuit designs.

1. **Common test structure between two measurement schemes:** The architecture of arrangement we design for the test structure must be general enough that both on-chip and off-chip current measurement schemes can use it. This ensures an unbiased comparison between the conclusions drawn from each of the two measurement schemes.

2. **DUT independence:** It is important for us to study individual transistor characteristics. Each device under test has to be measured individually, independent of the other off-devices. Leakage mitigation is essential here to ensure that the off-current does not affect the measurement of the on-current.

3. **High DUT numbers and density:** In order to have statistically significant results, a large number of devices under test is required. It is also necessary to have a dense DUT layout which enables us to have a good spatial map of the systematic variation.

4. **Minimum peripheral transistors:** As much area as possible should be dedicated to the DUTs rather than to the periphery circuitry. The peripheral circuits may also introduce resistance and variation. Having fewer of them, possibly by sharing them between the DUTs, can improve the accuracy of our measurement.

5. **Separation between the peripheral transistors and DUTs:** Separation between the peripheral transistors and DUTs can simplify the overall layout. If

the layout of the peripheral transistors is far away from the layout of the DUT regions, then the specific layout pattern or density of peripheral circuits will not conflict with the pattern density of the DUTs.

6. **High dynamic range on-chip measurement:** The architecture for the on-chip current measurement must support a dynamic range from 50nA to 1mA. The area dedicated to the measurement circuits should also be minimized.

3.2 Hierarchical Accessing Scheme

In order to achieve the features described in the previous section, a new hierarchical accessing scheme is presented in Figure 3-2. The accessing scheme is analogous to that in memory design, where row and column enables select the DUT on which to perform a measurement. Row and column decoders convert the binary selection signals, and output row and column enable signals. For now, we assume there is only one type of transistor, either NMOS or PMOS: details on how we select between them will be provided in Section 3.2.4.

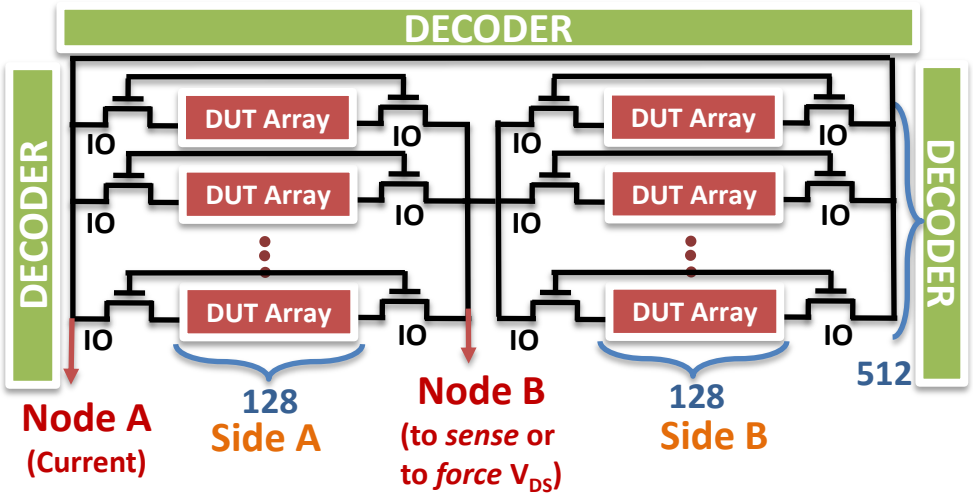

Figure 3-2: Our hierarchical accessing scheme.

Each DUT array consists of 128 DUTs placed in parallel, as shown in Figure 3-3. For NMOS transistors, the source of all the DUTs is connected to ground, and for PMOS transistors, the source of all the DUTs is connected to the supply voltage.

Two input-output (I/O) devices, acting as row-enabling switches, are placed across each DUT array. The gate of a DUT in an array is connected to the gates of all DUTs in other arrays which are in the same position as itself, but only within the same *side* of the test structure. Each of these gate connections forms one column, with a total of 256 columns. In additional to the 256 columns, we also have 512 rows, which allow for a total of 131,072 DUTs. Though there are two sides, *Side A* and *Side B* as shown in Figure 3-2, all of the DUT arrays are connected in parallel to one another in the test structure.

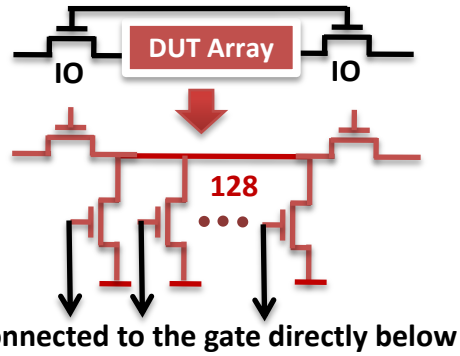

Figure 3-3: DUT array in the hierarchical accessing scheme.

The row enable signal is a normal digital select signal, with possible values of 0 or 1. Although all the DUT arrays are connected in parallel, the two different sides of the DUT array do not share the row enable signal. During each accessing cycle, only one DUT array is enabled, and within that DUT array, only one DUT is enabled. Therefore, we cannot share the decoders between *Side A* and *Side B*, since two DUT arrays cannot be enabled at the same time. As shown in Figure 3-2, we have two row decoders, one on each side of the test structure. On the other hand, the column enable signal is connected to the gate of the DUTs; therefore, it is not a digital decoder. Instead, it is an analog decoder that outputs the appropriate analog gate voltage to the DUT of interest, depending on the kinds of measurement to be performed. For the other off-state DUTs, either a negative voltage will be applied in the case of NMOS, or an above supply voltage will be applied in the case of PMOS, to minimize leakage.

For each DUT measurement, we apply the desired gate voltage to that DUT through the column enable signal, and apply the desired drain voltage indirectly through Node A or indirectly through Node B, depending on whether we are sensing or forcing.[1] All the other DUT arrays are off and all the other DUTs within the same array are off except the DUT to be measured. Then, we measure the current coming out of Node A. Node B can either be used to sense the drain voltage or force the drain voltage. In either case, there is no current coming out of Node B.

3.2.1 Forcing and Sensing Approaches

The design of the hierarchical accessing scheme is fairly flexible. To measure the individual transistor characteristics, we can either use the "forcing" approach or the "sensing" approach, as illustrated in Figure 3-4.

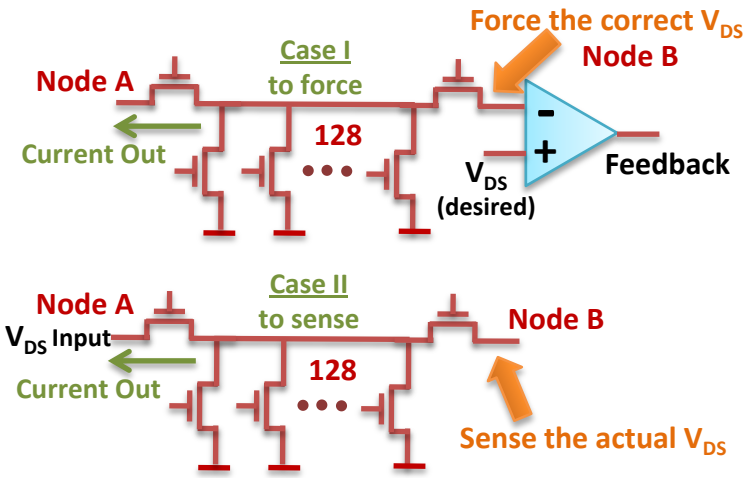

Figure 3-4: Forcing and sensing approaches.

In the forcing approach, Node B is forced to the desired V_{ds} by using an operational amplifier in a feedback configuration. The accuracy of the voltage at Node B depends on the bandwidth[2] of the amplifier. In the sensing approach, we apply a V_{ds} on Node A first. Since there is going to be voltage drop across the I/O switch next to Node A, we need to measure the voltage at Node B to determine the actual value of V_{ds}.

[1] The difference between sensing and forcing will be discussed in Section 3.2.1.
[2] The bandwidth dependence will be justified in Section 3.5.3.

This method is also called indirect voltage application since we do not know the exact drain voltage without an extra sensing step. The current measurement for both approaches is performed at Node A. Regardless of which method we use, there is no current flowing in or out of Node B.

3.2.2 Correct V_{ds}

Even though we know the voltage at Node B by either forcing it using an operational amplifier or by sensing it directly, we still cannot be sure that the voltage on Node B is the voltage that we are applying at the drain of the DUTs. Figure 3-5 explains some potential problems, such as the voltage drops across the wires, voltage drops across I/O switches, or the leakage current coming from the off transistors.

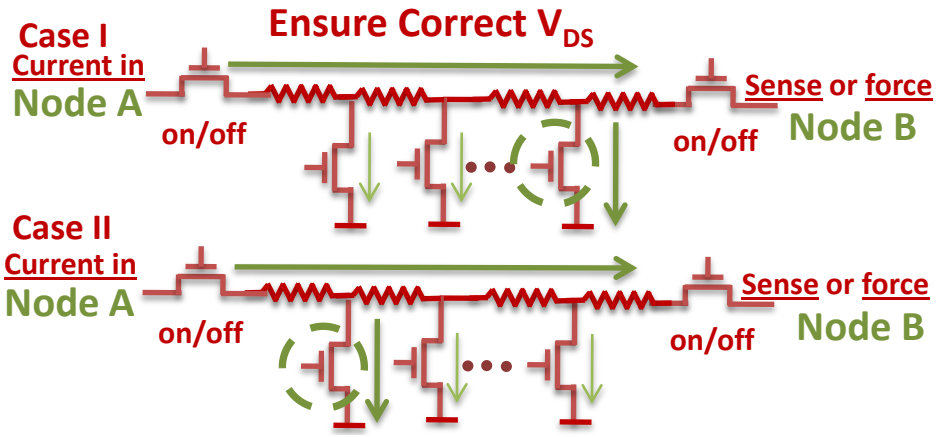

Figure 3-5: Ensuring the correct V_{ds}.

We divide the problem into two cases. In Case I, the DUT to be measured is located at the right edge of the DUT array. This case is fairly simple. Since there is no current flowing into or out of node B, all the current flow occurs to the left of the DUT. As a result, there is no voltage drop between the drain of the DUT and Node B.

In Case II, the DUT to be measured is located at the left edge of the DUT array. In this case, current flow does occur on the right side of the DUT. Thus, there will be a voltage drop between the DUT drain voltage and the voltage at node B. This voltage drop is unavoidable using this configuration. However, we can minimize the

magnitude of this voltage drop by using very wide interconnect to reduce the wire resistance and by minimizing the total current coming from the other DUTs.

There are two ways to control the amount of leakage current coming from other DUTs. The first is to apply a negative voltage at all of the gates that are in the off states to reduce the leakage current. The leakage current should decrease exponentially with a decrease in gate voltage. However, as the scaling of gate dielectric material continues, the gate leakage current plays a much more important role in the total amount of leakage current as compared to the channel leakage current. Therefore, the previous statement is no longer true. The channel leakage current exponentially decreases with the applied gate voltage, but the gate leakage current increases with a decrease in gate voltage. As shown in Figure 3-6, as we reduce the gate voltage, more than 80% of the total leakage current is contributed by the gate leakage current when the gate voltage is below ground.

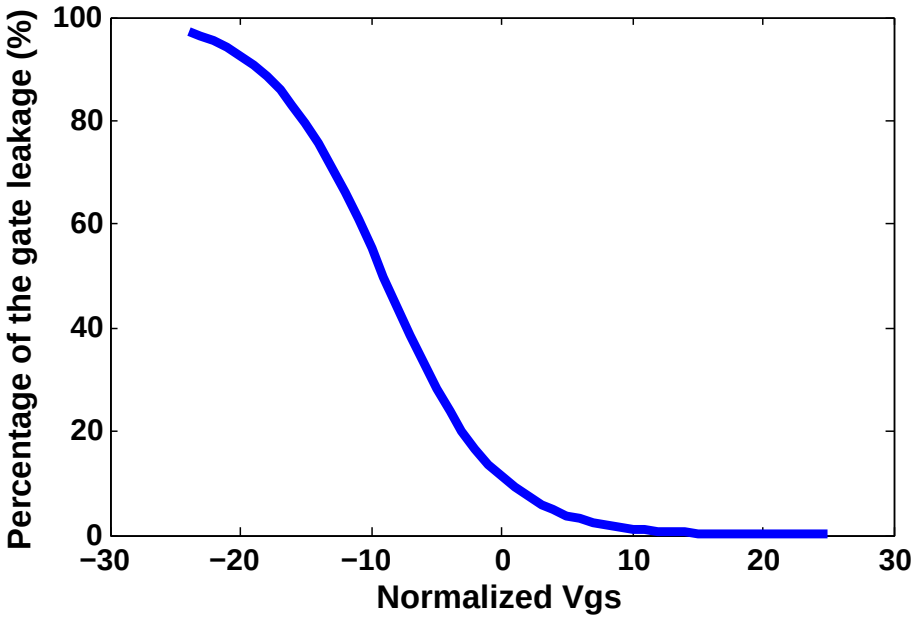

Figure 3-6: Contribution from the gate leakage current.

However, since the gate leakage current and channel leakage current change in different directions when $V_{gs} < 0$, there should be an optimal gate voltage that minimizes the overall leakage current from the DUT. Figure 3-7 shows that the optimal voltage does exist. Therefore, we cannot continue to decrease the gate voltage in-

definitely since the overall leakage will ultimately be dominated by the gate leakage component.

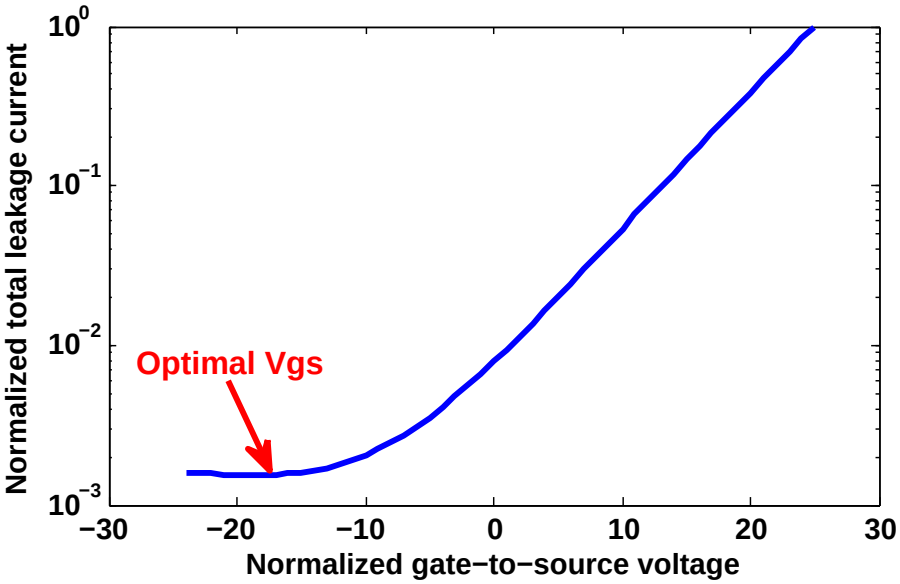

Figure 3-7: Optimal V_{gs} selection to minimize overall leakage.

The second way to reduce the leakage current is by reducing the number of DUTs per DUT array. However, in this case, the total number of DUTs in an array is not a limiting factor since the leakage current is only linearly dependent on the number of DUTs, while on the other hand, it is exponentially dependent on the gate voltage. The number of DUTs per DUT array is limited by the measurement accuracy, which will be discussed in the Section 3.2.3.

3.2.3 Leakage Mitigation

The current measurement of every DUT in the test structure is done through Node A on Figure 3-2. When measuring the current from one DUT, it is necessary to make sure that the leakage current from all the other DUTs within the same DUT array, and also from all the DUTs in other DUT arrays, is not significant compared to the minimum current we want to measure.

To ensure that the leakage current from other DUTs within the same DUT array does not influence the measurement accuracy, as described previously, we can find

an optimal gate voltage to minimize the overall leakage current from the DUTs. Additionally, we can also limit the number of DUTs within one array. 128 DUTs per array is chosen as the best number because the total leakage current is less than 0.5% of the minimum current, 50nA, that we want to measure, and it is a power of two, so it simplifies the binary decoder design.

To minimize the leakage current coming from the DUTs in other DUT arrays, we use two I/O devices as the row enable switches shown in Figure 3-3. By placing two I/O devices across the DUT array, the leakage current is determined by the I/O devices instead of the 128 DUTs inside that array. Therefore, the I/O switches can be used to limit the amount of leakage current coming from the DUT arrays. Because I/O devices are designed to work with higher supply voltages than nominal transistors, they have a much thicker gate dielectric layer to prevent oxide breakdown. Because of the thicker gate dielectric layer, the I/O devices also have much lower leakage current compared to normal transistors. The leakage ratio between the nominal transistors and the I/O devices for different values of V_{ds} is shown in Figure 3-8. The overall leakage mitigation can be improved by almost three orders of magnitude by using I/O devices.

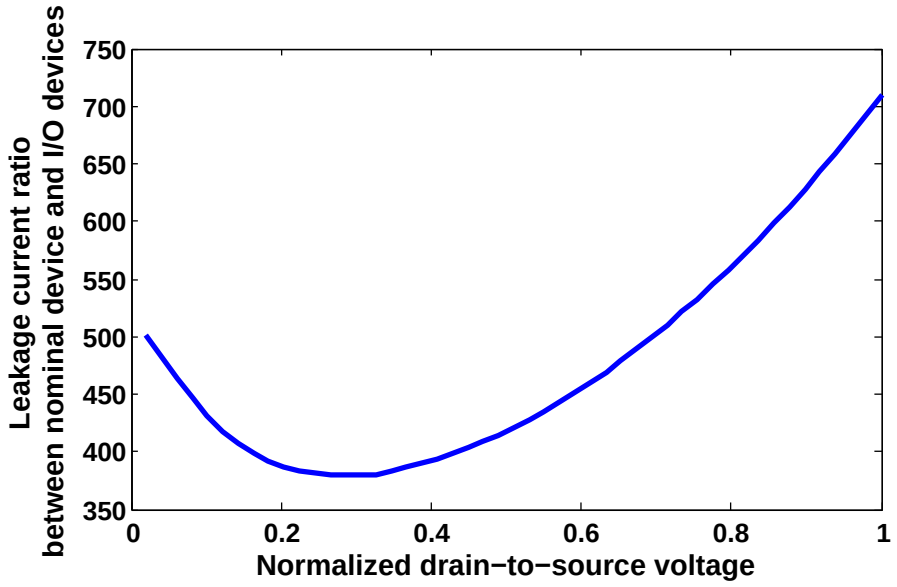

Figure 3-8: Leakage ratio of nominal devices to I/O devices.

The sizing of these I/O devices is also worth examining. Intuitively, in order to minimize the leakage current, we want to size the I/O switches to be as small as possible. The problem with sizing them so small is that the on-resistance of these devices will increase and the voltage drops across these devices during the current measurement will be too high. In the extreme case, the voltage drop can be so high that it goes over the supply voltage. This can cause oxide breakdown of the transistors in the measurement circuits. Finding the balance between leakage and the on-resistance is essential here.

3.2.4 NMOS and PMOS Transistors

Due to fundamental differences between NMOS and PMOS transistors, we cannot have a DUT array that contains both NMOS and PMOS devices. It has to either be all NMOS or all PMOS transistors. The reason is as follows. When the gate of a DUT in an array is connected to all the other gates in the same position of the other DUT arrays on the same side of the test structure, if some of the NMOS and PMOS gates are connected together, we can never turn them off simultaneously. This will significantly deteriorate our measurement accuracy.

To solve this problem, a new design unit block is shown in Figure 3-9. Instead of having just one DUT array, each block will have two DUT arrays: one for NMOS and one for PMOS. NMOS gate connections are shared among different blocks, and similarly, PMOS gate connections are shared among different blocks. We will not run into problems of not being able to turn off an NMOS and PMOS transistor at the same time because they are not wired together. With this design architecture, we will have twice the number of DUTs compared to before, with exactly half being NMOS DUTs and half being PMOS DUTs. Additional decoders are also required for the PMOS arrays in the test structure.

In this section, we describe a hierarchical accessing scheme for our architecture of arrangement. This accessing scheme allows us to measure all 131,072 NMOS transistors and 131,072 PMOS transistor on an individual basis. The design of leakage mitigation techniques ensures high current measurement accuracy and good confidence

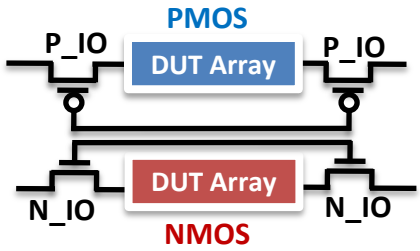

Figure 3-9: Co-existence of NMOS and PMOS transistors.

level on the applied drain voltage. This test structure is also able to accommodate two different measurement modes, forcing and sensing approach. The forcing approach is usually used by on-chip current measurement and the sensing approach is used by the direct probing measurement.

For every 128 DUTs, only two additional peripheral devices are needed. The ratio of the number of DUTs to peripheral devices is much higher than in many of the designs in the literature. The DUTs and peripheral devices also have very good separation, since all of the peripheral devices (the I/O switches) are on either side of the DUT array. This will simplify the pattern density layout practice significantly. All of the devices under test are placed in close proximity to assure good spatial resolution of our final extraction.

3.3 On-Chip Measurement Architecture

With the design of architecture of arrangement presented above, we will shift our discussion to the architecture of measurement. In order to perform on-chip current measurement, the main circuit building block that needs to be designed is the on-chip current measurement ADC. In evaluating the ADC architecture for our purpose, it is useful to understand the relevant figures of merit and requirements: (1) high dynamic range, (2) good precision, (3) good linearity and low gain errors, (4) low offsets, and (5) area-efficient. With these figures of merit in mind, we will briefly overview the common ADC architectures and justify the use of an integrating ADC architecture. Then, we will explain why the traditional integrating ADC architecture is unable to

meet the necessary specifications.

3.3.1 The Choice of Integrating ADCs

There is a very rich variety of ADC architectures available in the literature. A chart that summarizes a few common ADC architectures with respect to sampling frequency and resolution is shown in Figure 3-10.

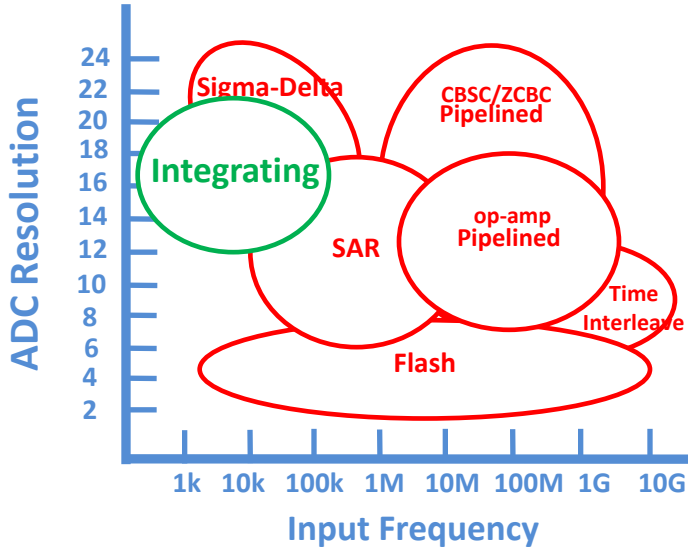

Figure 3-10: Common ADC architectures.

The flash ADC is also known as the parallel ADC, because it resolves all the bits at the same time. It is the fastest ADC architecture, but it is also area-inefficient and energy-inefficient. The resolution of the flash ADC is relatively small, but it is suitable for high-bandwidth applications. The pipelined ADC is the most popular ADC architecture for medium to high sampling rates, from a few megasamples per second to a few hundred megasamples per second. The resolution is in the range of 8-16 bits, which is much better than the resolution of the flash ADC. The traditional pipelined ADC design requires one operational amplifier per stage (or per bit). For high resolution, many operational amplifiers need to be designed, which can be very area expensive.

Successive-approximation-register (SAR) ADCs are most commonly used for medium

sampling frequencies with resolutions of 8-16 bits. The resolution is very similar to that of the pipelined ADC. The operation of this ADC is based on charge transfer and is therefore suitable for low-power applications such as portable devices. The sigma-delta ADC, also called an oversampling ADC, has a high sampling rate so it can perform noise shaping in the frequency domain. A very high resolution of 20 bits can be achieved using such an ADC architecture.

The integrating ADCs provide very good accuracy with relatively low sampling frequency. The architecture only requires a few simple building blocks. It provides very high accuracy, low gain errors and excellent noise rejection. The only drawback for this architecture is that it can only be used to sample slow-moving signals.

However, the integrating ADC is ideal for our purpose. Since we are performing DC measurement on transistor currents, we do not need an ADC that has high sampling frequency. The most important figures of merit for us are accuracy and area. The integrating ADC provides one of the best output resolutions of all the previously mentioned ADC architectures. In addition, it is also the most area efficient type of ADC.

3.3.2 Need for a New Integrating ADC

We cannot use a traditional integrating ADC architecture such as the one described in [59]. A few problems associated with this architecture need to be resolved before it can perform current measurements on our test structure. A sample integrating ADC architecture is presented in Figure 3-11.

In the figure, V_{in} is the input voltage that needs to be sampled, V_{ref} is the reference voltage, and C_1 is the charging capacitor. The required circuit blocks include an operational amplifier, a comparator, a counter, and control logic to output the control signals. A simplified timing diagram for dual-slope integrating operation is shown in Figure 3-12.

During phase I, the switch Φ_1 is closed. The operational amplifier is connected in a unity feedback configuration to reset V_x to the common mode voltage V_{CM}. During phase II, the switch Φ_1 is opened and the capacitor is charged for a fixed amount of

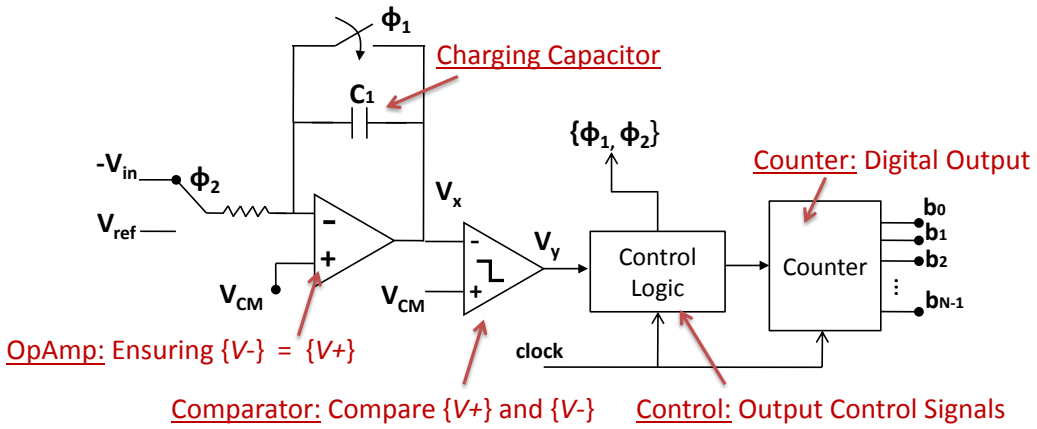

Figure 3-11: Traditional integrating ADC architecture.

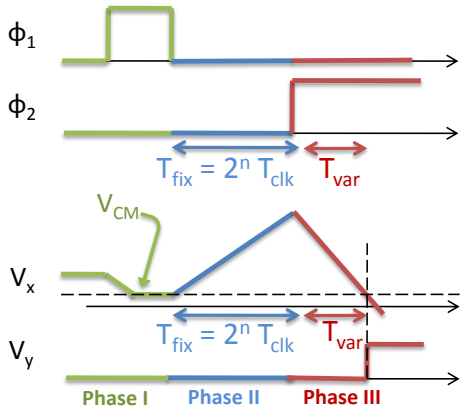

Figure 3-12: Timing diagram for traditional integration ADCs.

time. This time is equal to the time it takes for 2^N clock cycles, where N represents the number of bits of the ADC. Therefore, we have $T_{fix} = 2^N T_{clk}$, where T_{clk} is the period of the clock and T_{fix} is a fixed time that the capacitor is being charged for. At the end of the phase II, the voltage change in V_x can be written as in Equation 3.1. For larger input voltages, the increase in V_x will also be larger.

$$\Delta V_x = -\int_0^{T_{fix}} \frac{-V_{in}}{R \times C_1} = \frac{V_{in}}{R \times C_1} \times T_{fix} \tag{3.1}$$

During phase III, the capacitor is discharged for a variable amount of time, T_{var}, which depends on V_{in}. The counter is reset and the switch Φ_2 is connected to V_{ref}. Since the discharging current is always the same for any input voltage, for a larger input voltage, the charge built up on the capacitor is larger and it takes a longer time

to discharge, while for a smaller input voltage, the charge built up on the capacitor is smaller and it takes a shorter time to discharge. The counter stops counting when the inverter output switches from low to high. The final relationship between the input voltage and the reference voltage is shown in Equation 3.2.

$$V_{in} = \frac{V_{ref} \times T_{var}}{T_{fix}} \tag{3.2}$$

Using this measurement scheme, we see that the output voltage calculation does not depend on the absolute value of the passive components, R and C_1. Instead, it only depends on the reference voltage. Moreover, because of the dual-slope nature of the integrating operation, the conversion step also does not depend on the linearity of these passive components because the charging and discharging operations go through the same path. Dual-slope operation will help to cancel out the first-order non-linearity of the passive components.

Many of the other ADC architectures rely heavily on the matching and linearity properties of the passive components. For instance, the SAR and pipelined ADC rely on the matching between capacitors, and the flash ADC relies on the matching between resistors. It is very difficult to fabricate on-chip passive components with good matching and linearity. Therefore, these are great advantages of using the dual-slope integrating ADC. However, there are some drawbacks associated with this traditional architecture as it relates to our current measurement scheme.

Dynamic Range

One of the main goals of our test circuit is to be able to measure currents from 50nA to 1mA. This is a 14-bit current dynamic range. With the traditional integrating ADC architecture, in order to increase the bit resolution, according to Equation 3.5, we must either decrease the clock period, increase the capacitor value, or increase ΔV. ΔV is usually limited by the maximum voltage allowed in the technology to prevent oxide breakdown; it is difficult to increase this value. Therefore, the only tunable parameters are the clock frequency and capacitor value.

$$I \times \Delta T = C \times \Delta V \tag{3.3}$$

$$I \times T_{clk} \times 2^N = C \times \Delta V \tag{3.4}$$

$$2^N = \frac{C \times \Delta V}{T_{clk} \times I} \tag{3.5}$$

As an example, let us consider a situation where $\Delta V = 1V$, $T_{clk} = 10$ns (or frequency = 100MHz), $I = 14$mA, and $N = 14$ bits. This translates into a capacitor value of 164nF. For an on-chip capacitor to be this large, an area of approximately 10mm by 10mm is needed, which is definitely not suitable for any practical design. In addition, we do not want to increase the clock frequency because the current measurement will be more sensitive to noise and clock jitter.

Both increasing the capacitor value and increasing the clock speed will make the charging/discharging slope much slower with respect to the clock period. The amount of voltage increment within one clock cycle will become smaller. If this increment becomes smaller than the thermal noise of the circuit, the last few bits interpreted by the ADC would be too noisy to use. Therefore, using this architecture, we are also limited by the fundamental electronic noise.

Forcing V_{ds}

Another change that needs to be made to the traditional current integrating ADC architecture is shown in Figure 3-13. Instead of using input voltage sources and a resistor, an input current and a reference current source are used in the test structure. The positive terminal of the operational amplifier needs to be connected to the desired V_{ds} to force the drain voltage on the DUTs.

We can immediately see a problem after this replacement. During the reset stage, the output of the operational amplifier is reset to different voltages depending on the value of V_{ds} being used for the DUT measurements. This implies that ΔV can be larger or smaller between different measurements, which means that the resolution of the ADC can change over times as shown in Equation 3.5. Instead, we want ΔV to

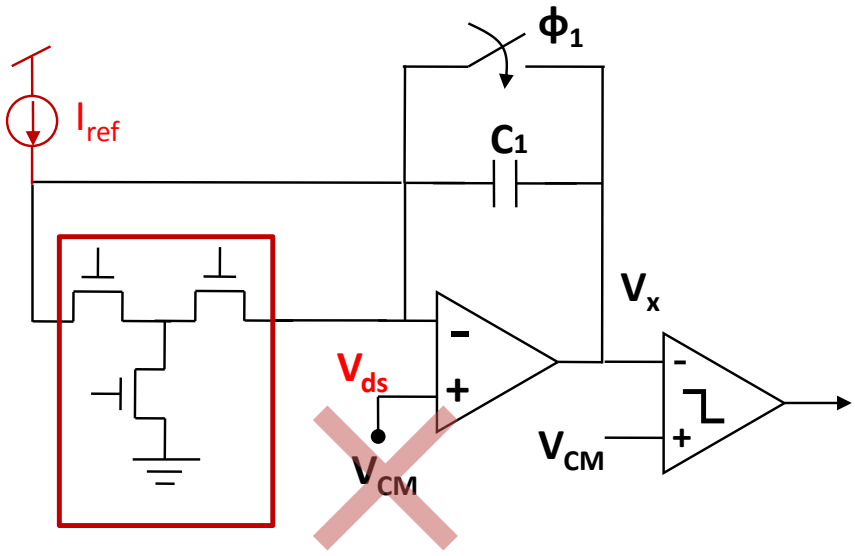

Figure 3-13: New switching scheme for V_{ds}.

be at its maximum and stay constant with time.

Offset and Charge Injection

The comparator in this design can have an internal offset voltage. The offset is the result of mismatch between the positive and the negative terminals of the comparator design and layout. Figure 3-14 shows the cases of both positive and negative comparator offset. This offset can cause the comparator to switch earlier or later depending on the sign of the offset. Both of these non-idealities will result in conversion errors.

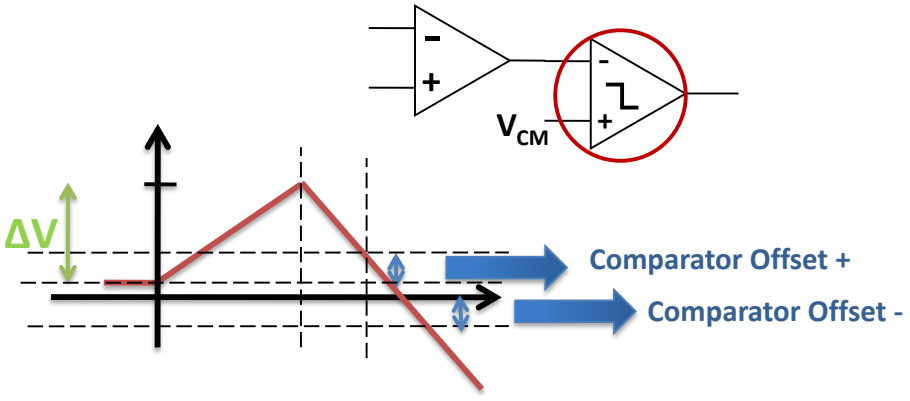

Figure 3-14: Comparator offset problem.

Another non-ideality comes in the form of charge injection from the switches.

As shown in Figure 3-15, a switch is a transistor. When the switch turns on, it stores charge on the gate overlap capacitance. When the switch turns off, the charge on these capacitors is injected onto the adjacent nodes. If the switch is directly connected to the input terminal of the comparator, it will directly affect the decision of the comparator.

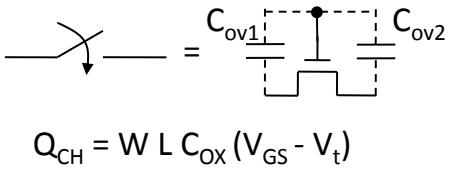

Figure 3-15: Switch charge injection.

There are two problems associated with the charge injection. First, the charge stored on the capacitors is not always the same: it depends on the V_{gs} across the switch transistor as shown in the equation in Figure 3-15. Since the source voltage of the switch depends on the V_{ds} applied to the DUT, this charge injection is signal dependent. For smaller values of V_{ds} applied to the DUT, there will be a larger V_{gs} across the switch and more charge injection. This can introduce unwanted non-linearity into the measurement result. Second, these switches are designed with I/O devices, so much larger overlap capacitances and larger charge injection are expected. Without careful offset mitigation techniques, the ADC resolution can be significantly degraded.

3.4 Redesigned Integrating ADC

A high dynamic range, a constant ΔV, and minimum offset and charge injection are the main challenges to be overcome in the traditional current integrating ADC design. These challenges inspire us to perform two major architectural changes.

3.4.1 High Accuracy vs. High Dynamic Range

The first architectural change is to overcome the lack of dynamic range in the traditional ADCs. Because the current we need to measure spans over 14 bits of dynamic range, we must either increase the capacitor size or increase the operating frequency, as analyzed in Section 3.3.2.

However, this calculation performed in Section 3.3.2 is for 14-bit accuracy, not 14-bit dynamic range. The real goal is a design that can provide a coarse absolute resolution in the measurement of high current range and a fine absolute resolution in the measurement of low current range. In other words, the relative resolution in all ranges should remain the same. For example, when measuring a 1mA current, we are indifferent to a change of 50nA because it is only 0.005% of the total current. However, when measuring a 50nA current, a change of 50nA would represent a change of 100% in the total current. The same 50nA has different implications for high and for low current measurements. Our goal for the ADC is to make sure that the current measurement is within 0.5% of the absolute current for all current ranges.

A current integrating ADC has very good resolution, but it is not ideal for measuring current of large absolute quantity. Because the current measurement is done by integrating the current onto a capacitor, to measure a large current value, we are required to use an unreasonably large on-chip capacitor. However, because we only need high dynamic range, we can split up the current measurement into two stages: the first stage, which performs a coarse measurement on the large current, and the second stage, which performs a fine measurement on small current. This is achieved by having a current steering digital-to-analog converter (DAC), as shown in Figure 3-16.

A 9-bit current steering DAC is added to perform measurement on large current (coarse measurement), so the original current integrating ADC only needs to measure currents that are smaller than the I_{LSB} of the DAC (fine measurement). The least significant bit of the DAC (I_{LSB}) is 2μA, and the most significant bit of the DAC (I_{MSB}) is 512μA. As indicated in the Equations 3.6-3.8, the current steering DAC

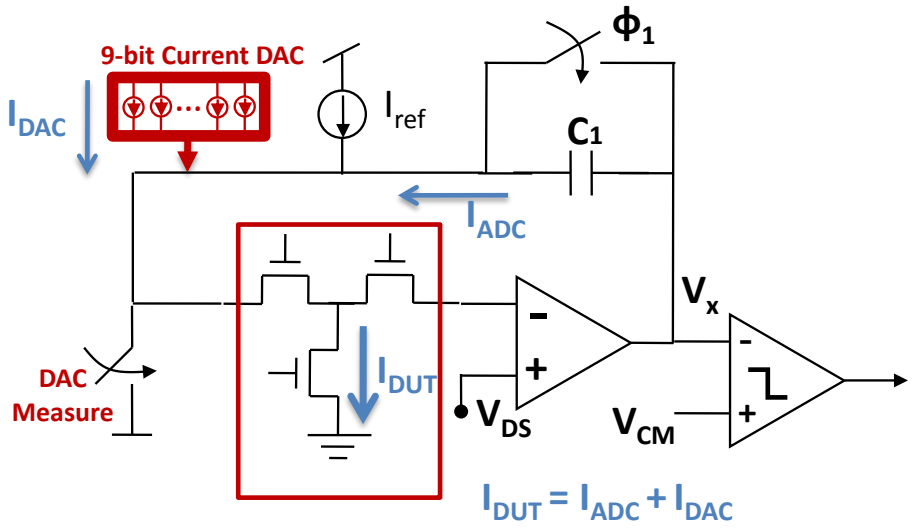

Figure 3-16: Current steering DAC to accommodate the dynamic range requirement.

needs to resolve currents from 2μA to 1024μA and the integrating ADC needs to resolve currents below 2μA.

$$I_{DUT} = I_{ADC} + I_{DAC} \tag{3.6}$$

$$2\mu A \leq I_{DAC} \leq 1024\mu A \tag{3.7}$$

$$I_{ADC} \leq 2\mu A \tag{3.8}$$

To perform a measurement on a DUT, the first step is to find the DAC supply current. This current has to be the maximum current out of 512 different DAC current values that is still smaller than the current that the DUT is drawing. Equation 3.9 explains this relationship. The DAC current we want to supply is $I_{DAC,(j)}$. Assuming we have good matching between the bits of the DAC for us to achieve low differential nonlinearity (DNL) error,[3] the difference between the DAC current and the DUT current should be less than one LSB of the DAC. This extra current is then measured by the current integrating ADC.

[3]Differential nonlinearity(DNL) is defined as the variation in analog step sizes away from 1 LSB. An ideal converter has its maximum differential nonlinearity of 0 for all digital values [59].

$$I_{DAC,(j)} \leq I_{DUT} \leq I_{DAC,(j+1)} \tag{3.9}$$

$$I_{DUT} - I_{DAC,(j)} \leq I_{ISB} = 2\mu A \tag{3.10}$$

Using this current measurement algorithm, both the absolute values of each DAC bit and the value of the reference current are needed to calculate the total current value. As shown in the lower left corner of Figure 3-16, an additional switch is added in order to measure the current at the beginning of the entire measurement. This current characterization only needs to done once for each chip. This must be done because there is no way to design an absolute value for any component on-chip. On-chip components can be designed to have good matching, but not for their absolute values.

3.4.2 Constant ΔV, Offset and Charge Injection Immunity

The second architectural change is designed to overcome the offset and charge injection problem, and the problems associated with a varying initial voltage at the charging node of the operational amplifier.

In order for the initial voltage at the charging node of the operation amplifier to be the same for different V_{ds} measurements, we add two additional switches in the original current integrating ADC architecture, as shown in Figure 3-17. During the reset phase, the left side of the capacitor is connected in a unity feedback loop with the operational amplifier and pre-charged to V_{ds}. However, the right side of the capacitor[4] is connected to a voltage, V_{LOW}, that remains the same for all measurements. This ensures that the initial charging voltage is always the same regardless of which V_{ds} is applied.

This architecture change can also help with the problems of offset and charge injection. In Figure 3-17, red circles are placed to next to the components that can

[4]The right side of the capacitor is connected in an unity feedback loop with the left side of the capacitor in the previous architecture during reset phase.

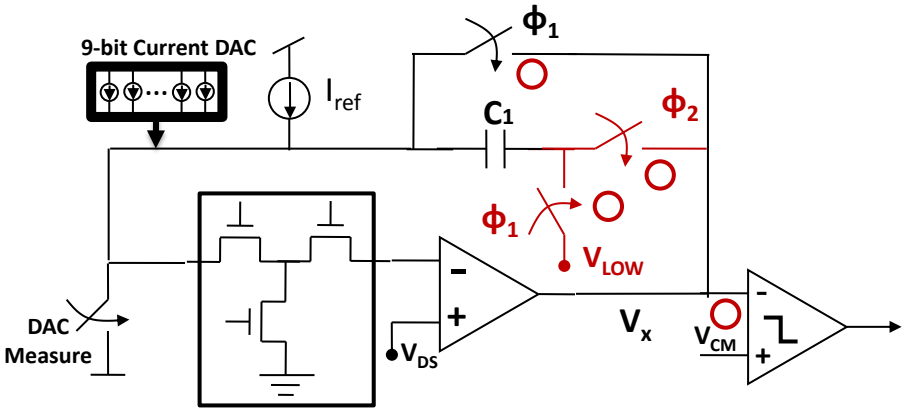

Figure 3-17: Offset and charge injection immunity design.

introduce either offset or charge injection. As discussed previously, because they are directly connected to the charging node of the operational amplifier, these effects can degrade the accuracy of the ADC. In the new proposed architecture, however, the output node of the operational amplifier is charged to a constant voltage, V_{LOW}, which is purposely designed to be lower than the common mode voltage, V_{CM}, as depicted in Figure 3-18. The comparator will change signs twice during the charging and discharging operation, once at the beginning when the voltage goes above V_{CM} and again at the end when the voltage goes below V_{CM}. The counter will only count during this period.

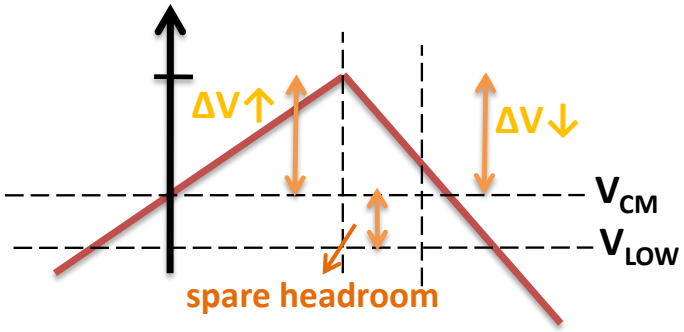

Figure 3-18: Constant ΔV for charing and discharging.

Using this new measurement scheme and architecture, we can resolve both the comparator offset and charge injection problems. The problem with the comparator offset voltage before is that it introduced offset between $\Delta V\uparrow$ and $\Delta V\downarrow$, but when we calculated the current, we assumed that they were the same. With this modification,

now the counter is only counting during the period of time when the comparator switches signs, so $\Delta V \uparrow$ and $\Delta V \downarrow$ are always guaranteed to be the same regardless of the presence of offset. The problem associated with charge injection can also be resolved by leaving enough headroom between V_{CM} and V_{LOW}, as indicated in Figure 3-18. As long as the charge injection is smaller than the amount of spare headroom, the overall conversion is not affected by the extra charge.

3.4.3 Final Integrating ADC Architecture

Figure 3-19 below shows the final integrating ADC architecture. This design has high dynamic range and correct V_{ds}, and is immune towards offset and charge injection. A new timing diagram explaining the operation of the new architecture is shown in Figure 3-20.

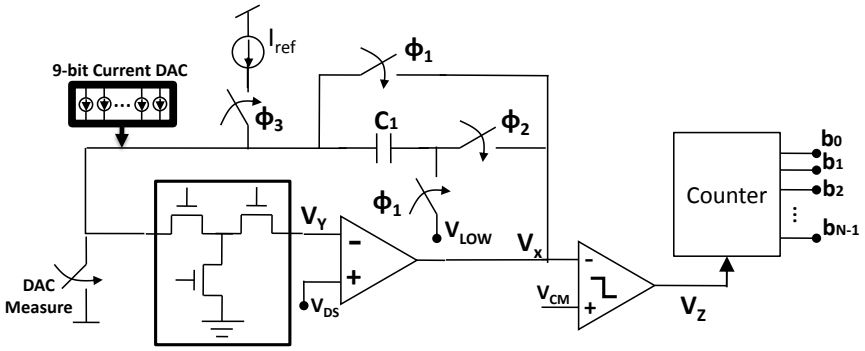

Figure 3-19: Final integrating ADC architecture.

During phase I, the two switches controlled by Φ_1 are closed. The left side of the capacitor plate is reset to V_{ds} and the right side of the capacitor plate is reset to V_{LOW}. During phase II, the switches controlled by Φ_1 are opened. Charge injection may occur at node V_X, but as long as it is smaller than the "spare headroom" we put in the design, it will not affect the ADC accuracy. During phase III, the DUT current begins to charge the capacitor. The counter will only start counting when the comparator hits its threshold for the first time. The counter will count a fixed amount of time, $2^N T_{clk}$, before resetting itself. In phase IV, the reference current, I_{ref}, will discharge the capacitor. The amount of time it takes to discharge V_X back

to the comparator threshold determines the final value of the current.

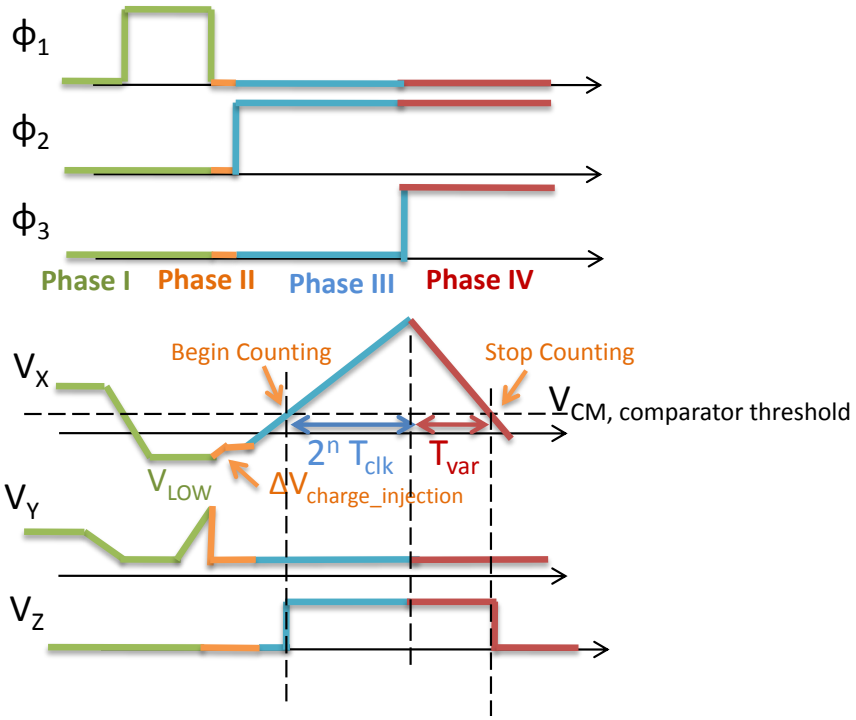

Figure 3-20: Timing diagram for the new integrating ADC architecture.

3.5 Circuit Components for the Redesigned Integrating ADC

In the section, we discuss the key circuit blocks in the newly-designed integrating ADC architecture. Each circuit component will be discussed in terms of its role in the integrating ADC, the specifications necessary to ensure the accuracy of conversion, and the design and layout challenges. The major circuit components we will focus on are the 9-bit current steering DAC, the comparator and the operational amplifier.

3.5.1 Current-Steering DAC

The current steering digital-to-analog converter (DAC) is designed in the integrating ADC to allow us to perform high dynamic range current measurement. Its main role is

to resolve the upper bits of the conversion (performing the coarse measurement) before the integrating ADC resolves the remaining lower bits (performing fine measurement). The figures of merit associated with the current steering DAC in this design are output impedance, matching, and leakage current.

Output Impedance

It is important for the DAC design to have high output impedance. As shown in Figure 3-19, one terminal of the DAC is connected to the supply voltage and the other terminal is connected to the left side of the capacitor plate. Between measurements, different voltages can be present across the DAC, since different values of V_{ds} are needed to apply to the DUTs in the measurements. For the same input DAC code, however, we want the output DAC current to remain the same regardless of the voltage across it. In other words, the DAC should behave like an ideal current source. However, in a real current source design, the linearity is greatly limited by the voltage headroom and the finite output impedance of the transistor.

In order to improve the DAC performance, I/O devices are used for the design. There are two benefits of using I/O devices. First, using I/O devices allows us to have much higher voltage headroom since the supply voltage for I/O devices is 2.5V, while for nominal transistors, it is only 1V in the 65nm technology node. Second, I/O devices are usually slower (smaller bandwidth), but they have much higher output impedance than the nominal transistors. Since the DAC current does not have to change very often, this presents a good tradeoff between slower speed and higher output impedance.

To improve the output impedance further, circuit techniques such as cascoding and feedback are also applied to the DAC design. Figure 3-21 shows the design of the unit DAC cell. Q_2 is cascoded on top of Q_1 to improve the output impedance of Q_1 by the gain of Q_2, $(1 + g_{m2} \cdot r_{o2})$. An additional feedback loop created by transistor Q_3 boosts the output impedance by another factor of $(1 + g_{m3} \cdot r_{o3})$, making the total small-signal output impedance approximately $r_{o1} \cdot (1 + g_{m2} \cdot r_{o2}) \cdot (1 + g_{m3} \cdot r_{o3})$. Simulation results of the output impedance for DAC1 (first bit of the DAC), DAC2

(second bit of the DAC) and DAC4 are shown in Figure 3-22.

For the purposes of the test circuit, the output voltage on the x-axis in Figure 3-22 is the same as V_{ds} with the addition of the voltage drops across the switch. For a nominal transistor, we are interested in values of V_{ds} between 0V and 1V. Assuming the voltage drop across the switch does not exceed 200mV, the output voltage is in the range of 0V to 1.2V. Within this range, the output current of DAC1, DAC2 and DAC4 change by less than 5×10^{-3} %, demonstrating very high output impedance of the DAC design.

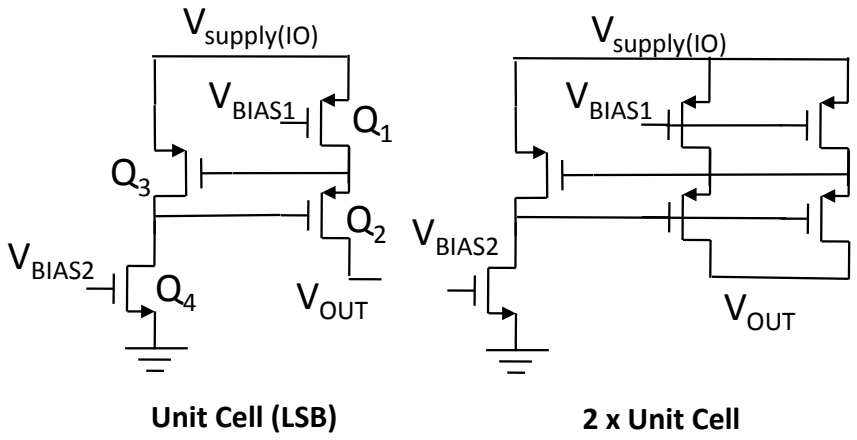

Figure 3-21: Current steering DAC cells.

Matching

Matching is another important figure of merit for the current steering DAC design. Poor matching in the DAC cell design is usually reflected in the differential nonlinearity (DNL) error. For an ideal converter, each analog step should be equal to 1 LSB. Due to mismatches between different bits of the converter, however, DNL can be much more or less than 1 LSB. As shown in Equation 3.8, the integrating ADC design is based on the assumption that the maximum current integrated onto the capacitor is less than 1 LSB of the DAC. If the DNL is significantly more than 1 LSB, the current is too large for the integrating ADC to resolve, which could cause significant problems in the design.

We use two approaches to improve the matching between DAC cells. The first

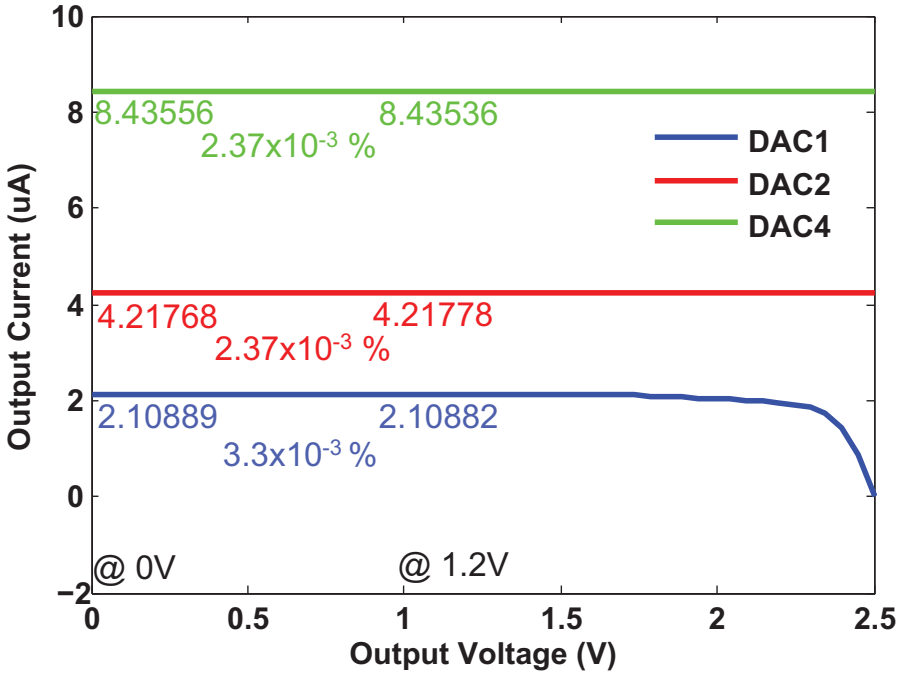

Figure 3-22: High output impedance of the DAC cells.

approach is to use the same unit DAC cell throughout the design. Figure 3-21 shows both DAC1 and DAC2, and instead of using transistors twice the width for DAC2, we use two unit DAC1 cells in parallel. We follow the same principle for the subsequent bits. Even though this will increase the layout area by a small amount, the improvement in matching is worth the tradeoff.

The second approach used to improve matching is by following good layout practice. One example of how much matching can be improved by good layout is shown in Figure 3-23, which shows the simulation results of post-layout extraction of the DAC. The plots on the left show the output current and the DNL of the original layout; the plots on the right shows the output current and the DNL of the revised layout. We can see that with poor DAC layout, the DNL can be as much as 9 LSB. With good layout practice, most of the DNL is well below 0.5 LSB.

The two layouts that generate these plots in Figure 3-23 are also shown in Figure 3-24. A comparison of the two layouts shows that in order to achieve good matching between the DAC cells, significantly more layout area is needed. The grey area in the figure represents the active area for I/O devices. The extra layout area comes

from the requirement that the active area enclosing the transistor must be larger than a certain minimum constraint to prevent STI stress from the edge. Moreover, the layout of NMOS and PMOS must be far away from each another.

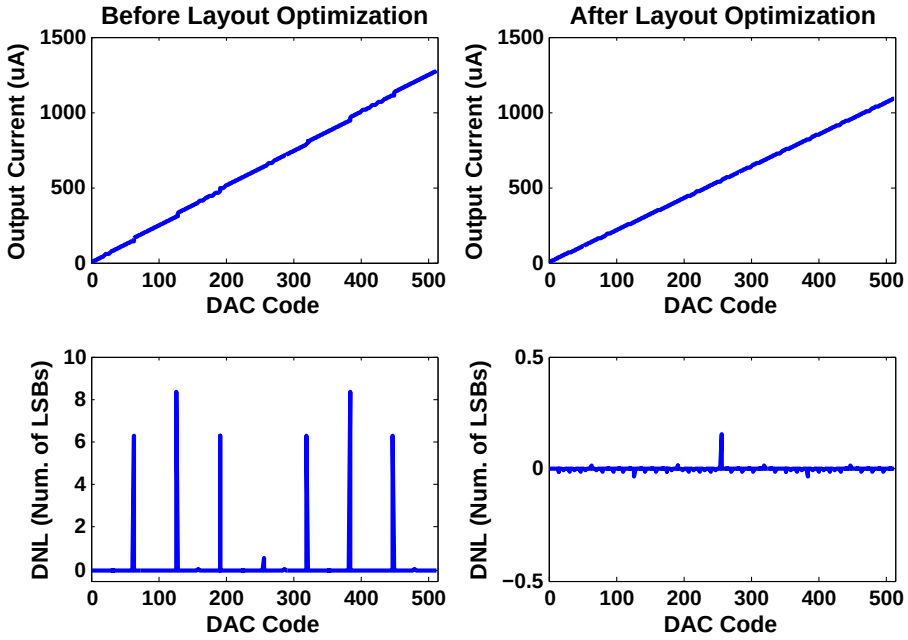

Figure 3-23: DNL before and after layout optimization.

DAC Leakage Current

Another problem worth mentioning is leakage current. When the DAC is turned off, we do not want any current leaking from the DAC to affect our measurement accuracy. Our DAC design has very low leakage during the off-state for two reasons: first, the I/O devices have much higher threshold voltage compared to nominal devices, and thus, much lower leakage current when the I/O devices are turned off; and second, cascading Q_1 and Q_2 introduces a stacking effect which reduces the leakage current [61]. Together, this makes the leakage current from the DAC insignificant compared to the current accuracy that the design needs.

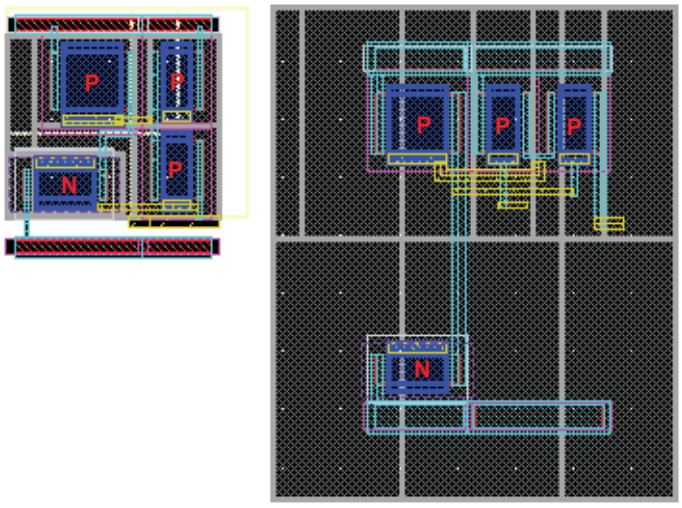

Figure 3-24: Layout optimization: before vs. after optimization.

3.5.2 Comparator

The comparator compares voltage on the charging node of the operational amplifier with the common mode voltage, V_{CM}, to determine when charging or discharging should occur. The output of the comparator acts as an enable signal to turn on or to turn off the counter. The main figures of merit associated with comparator design are offset, bandwidth and kickback noise. As described previously, our new proposed architecture is immune to any offset in the comparator; thus, our analysis will focus mainly on the latter two figures of merit.

Figure 3-25 shows the comparator architecture. It is comprised of two stages: a preamplification followed by a track-and-latch stage. The rationale behind this architecture is as follows. The preamplifier is used to amplify the difference between the input voltages, V_{in+} and V_{in-}. The gain factor of the preamplifier is typically from 2 to 10. The track-and-latch stage then amplifies the output of the preamplifier further during the track stage. In the latch stage, the positive feedback regenerates this difference at the output of the latch from the analog signal to the full-scale digital signal. The clock signal is used to reset the output of the latch to ground between each decision cycle. This is to ensure that no memory is carried over from one decision

cycle to another.[5]

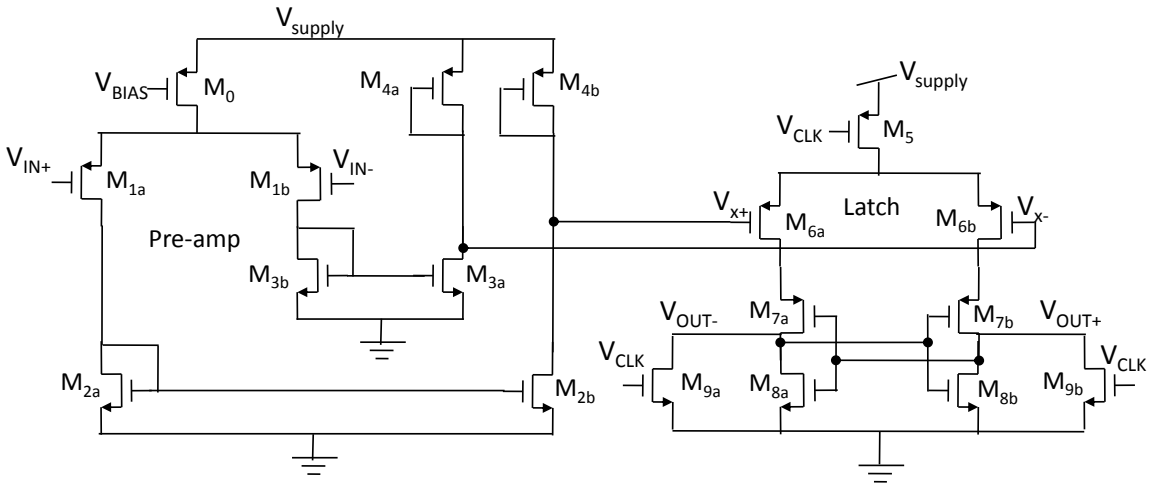

Figure 3-25: Comparator architecture.

Bandwidth

In a typical amplifier design, there is a tradeoff between gain and bandwidth. Depending on the configuration of the design, we can either have high speed or high gain, but not both at the same time when using a single stage design. In the preamplifier stage, we have very low gain in order to achieve the high bandwidth (or high speed) that we are looking for in our design. From the schematic of the preamplifier, we can see that the low gain (and high bandwidth) is achieved by having the output transistors M_{4a} and M_{4b} connected in a diode configuration. The gain is picked up by the track-and-latch stage using positive feedback.

Kickback Noise

Using a preamplifier not only benefits the design in terms of bandwidth, but also helps to minimize the effect of kickback noise. Kickback noise refers to the charge transfer either into or out of the input nodes when the latch stage goes from the track mode to latch mode. When the latch circuit regenerates the analog signal difference into a full-scale digital signal, it introduces a large instantaneous current spike. This

[5]Also referred as *hysteresis effect*.

current can couple back into the input nodes, V_{x+} and V_{x-}, through the parasitic capacitors C_{gs} and C_{gd} of the input transistors. This can cause large voltage glitches on the driving circuitry and significantly limit the accuracy of the comparator.

The preamplifier helps to minimize the effect of kickback noise in two ways. First, it serves as a buffer stage between the latch and the driving circuitry (in our case, the operational amplifier). The kickback noise will mostly appear on the nodes V_{x+} and V_{x-}, but not the input terminals V_{IN+} and V_{IN-}. Second, it provides low impedance looking into nodes V_{x+} and V_{x-}. Without the preamplifier, the latch input will be directly connected to the output of the operational amplifier, which has very high output impedance. In this design, however, the latch input is connected to the output of the preamplifier, which has relatively low output impedance. This can help minimize the voltage glitches induced by current coupling through the parasitic capacitors.

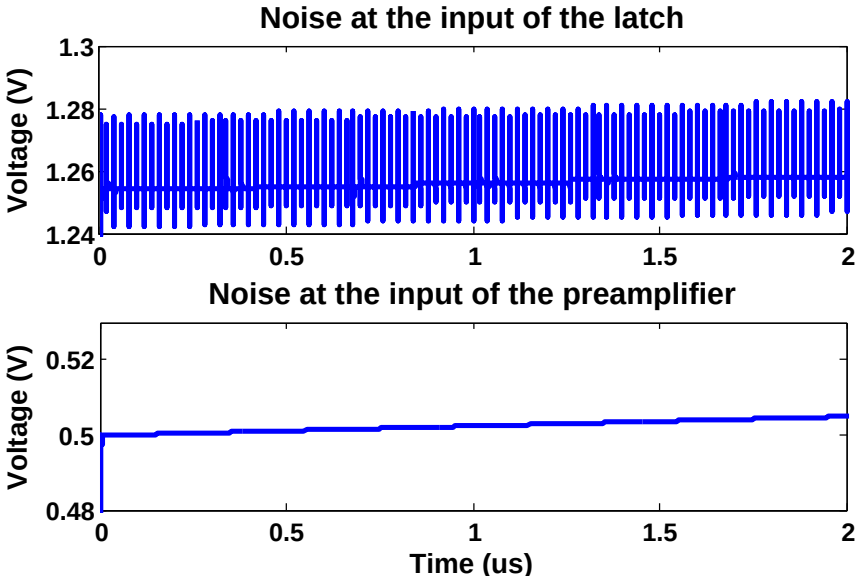

Figure 3-26: Kickback noise reduction with preamplifier.

Figure 3-26 shows the amount of kickback noise reduction. The top plot is the voltage at the input of the latch, and the bottom plot is the voltage at the input of the preamplifier, which is the overall input of the comparator, V_{IN}. We can see that the preamplifier does a very good job buffering out the noise coming from the

second-stage latch coming into the driving circuitry.

3.5.3 Operational Amplifier

The operational amplifier (opamp) is used in the current integrating ADC for two purposes. First, the opamp is put in a feedback configuration to ensure that throughout the entire charging or discharging cycle, the V_{ds} applied across the DUTs stays the same. Second, the opamp supplies the necessary current at the output to charge or discharge the capacitor.

The two-stage opamp architecture, shown in Figure 3-27, is designed for our integrating ADC. The first stage is a telescopic cascode amplifier, and the second stage is a common source amplifier. A nulling resistor and a compensation capacitor are also used to ensure the stability of the opamp.

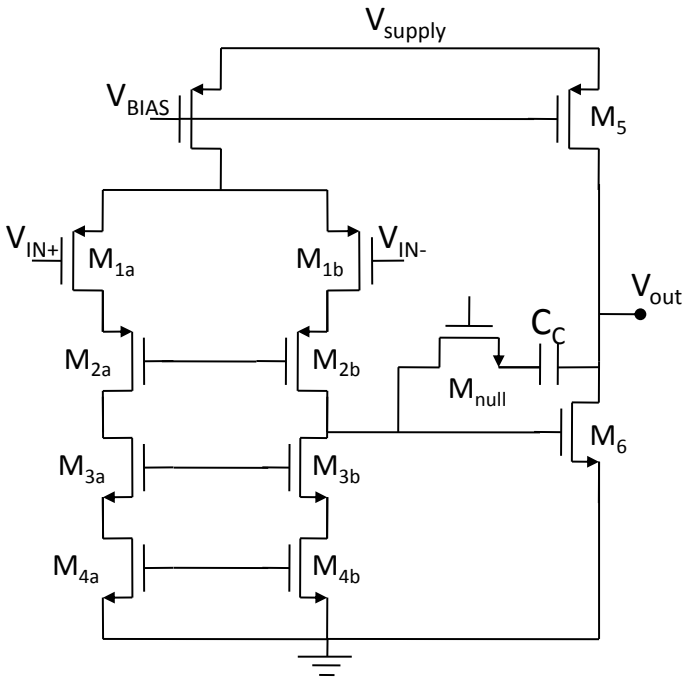

Figure 3-27: Operational amplifier architecture.

Minimize Dynamic ΔV_{IN}

As described previously, one of the main objectives for the opamp design is to minimize ΔV_{IN}, the difference between V_{in+} and V_{in-} during the charging cycles. It is a common misconception that in order to minimize this voltage difference, the only necessary figure of merit is the gain of the opamp. Equation 3.12 illustrates this misconception. A represents the DC gain of the opamp. According to this equation, as long as the gain is large enough, ΔV_{IN} should be minimized.

$$A \times (V_{in+} - V_{in-}) = V_{in-} \tag{3.11}$$

$$\Delta V_{IN} = \frac{1}{1+A} \times V_{in+} = \frac{1}{1+A} \times V_{ds} \tag{3.12}$$

The confusion arises from the fact that here, we are minimizing dynamic ΔV_{IN}, not the static ΔV_{IN}. Static ΔV_{IN} is the voltage difference between the positive and negative terminals when the opamp is in DC steady state, but dynamic ΔV_{IN} is the voltage difference between the positive and negative terminals when the opamp is still charging (or discharging) the output node. Intuitively, the dynamic ΔV_{IN} should have some time (or bandwidth) dependency. Before going into a formal derivation, we can guess that the dynamic ΔV_{IN} has the relationship described by Equation 3.13.

$$\Delta V_{IN} \approx \frac{dV_{out}}{dt} \times delay \tag{3.13}$$

Here, $\frac{dV_{out}}{dt}$ is the charging rate at the output of the opamp, and *delay* is the intrinsic delay of the opamp. The guess originates from the fact that the output of the opamp is changing, but the input of the opamp cannot respond infinitely fast to the change of output voltage. Thus, the voltage difference should be the product of the charging rate and the intrinsic delay. This delay of reaction time at the input should be proportional to the bandwidth of the opamp.

Figure 3-28 and Equation 3.15-3.18 explain this relationship formally. A two stage operational amplifier design can be divided into a current gain stage and a voltage

gain stage as shown in Figure 3-28. Here, C_c represents the compensation capacitor, $C_{feedback}$ represents the charging capacitor, g_m represents the transconductance and A represents the voltage gain.

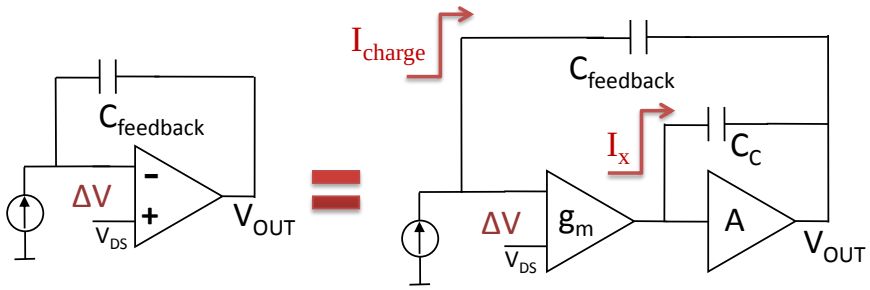

Figure 3-28: Minimizing dynamic ΔV_{IN}.

$$\frac{dV_{out}}{dt} = \frac{I_{charge}}{C_{feedback}} = \frac{I_x}{C_c} \tag{3.14}$$

$$I_x = \frac{C_c}{C_{feedback}} \times I_{charge} \tag{3.15}$$

$$I_x = g_m \times \Delta V_{IN} \tag{3.16}$$

$$\Delta V_{IN} = \frac{C_c}{C_{feedback}} \times \frac{I_{charge}}{g_m} = \frac{I_{charge}}{C_{feedback}} \times \frac{C_c}{g_m} \tag{3.17}$$

$$\Delta V_{IN} = \frac{I_{charge}}{C_{feedback}} \times \frac{1}{\omega_u} \tag{3.18}$$

The above derivation confirms our initial guess. The dynamic ΔV_{IN} is inversely proportional to the unity bandwidth, ω_u, of the two-stage amplifier, not the gain of the amplifier. Since the unity gain bandwidth is equal to the product of the gain and bandwidth, we cannot minimize ΔV_{IN} by trading off between gain and bandwidth. Unfortunately, the only way to improve the accuracy of ΔV_{IN} is by increasing power consumption.

Figure 3-29 shows that our design has a unity gain bandwidth of 270MHz. As a result, using the opamp design, the maximum dynamic ΔV_{IN} is only $\frac{I_{max}}{C_{charge}} \times \frac{1}{\omega_u} = \frac{2\mu A}{250pF} \times \frac{1}{2\times\pi\times 270MHz} = 4.7\mu V$. This is adequate because the minimum ΔV_{IN} of interest for the V_{ds} sweep is 100μV.

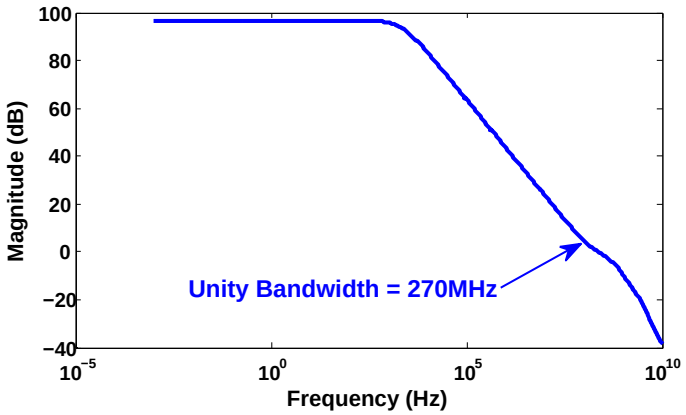

Figure 3-29: Bandwidth of the operational amplifier design.

3.6 Measurement Flow

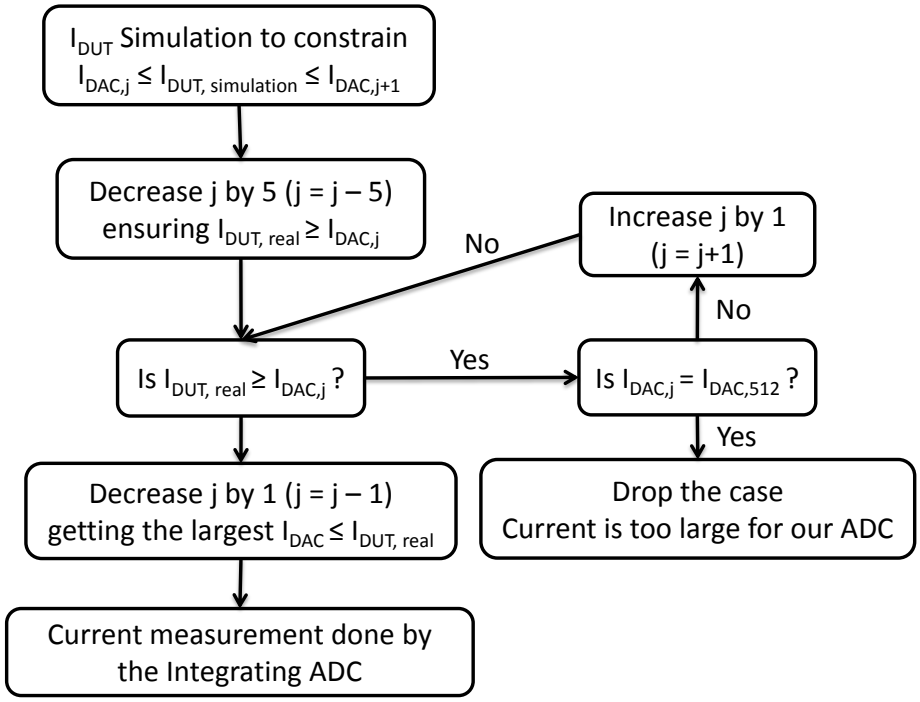

Figure 3-30: Measurement flow.

Figure 3-30 summarizes the steps which are needed to perform on-chip current measurement using this proposed architecture. These measurement steps need to be repeated for every measurement point (or every pair of V_{gs} and V_{ds}) of the DUTs. This may seem quite time-consuming since these calibration steps are required for

every measurement, but the tradeoff is expected when measuring current with such high dynamic range. However, it is worth noting that if the current to be measured is much smaller than the LSB of the DAC (for example, in the subthreshold regime of operation), no sweeping is necessary. Here is a detailed description of the calibration steps:

1. *Step 1*: Before performing the hardware measurement, the DUT current is bounded between $I_{DAC,j}$ and $I_{DAC,j+1}$ by simulation using the foundry transistor models. $I_{DAC,j}$ represents the j^{th} output of the current DAC and $I_{DAC,j+1}$ represents the $(j+1)^{th}$ output of the current DAC.

2. *Step 2*: The real DUT current may be larger or smaller than the simulated current due to variation. In order to ensure that $I_{DAC,j}$ is always smaller than the current drawn from the real DUT, $I_{DAC,real}$, we decrease the simulated j by 5.[6]

3. *Step 3*: The goal of this step is to find the largest I_{DAC} that is smaller than $I_{DAC,real}$. We first compare the magnitude of $I_{DAC,j}$ with the magnitude of $I_{DAC,real}$ using the integrating ADC part of the circuit. If $I_{DAC,real}$ is larger than $I_{DAC,j}$, then we check if $j = 512$. If j is equal to 512, that means we hit our measurement limit. The current is too large and we drop the case. If j is less than 512, then we increase j by 1 and repeat this loop. On the other hand, if $I_{DAC,real}$ is smaller than $I_{DAC,j}$, then we move to the next step.

4. *Step 4*: We find the largest current of DAC that is smaller than the current drawn by the DUT by decreasing j by 1. With this, we can move on to measure the current using the integrating ADC.

The counter in the ADC will provide us with the final digitized representation of the current we are measuring. In additional to the packaged die, a field-programmable gate array (FPGA), a printed circuit board (PCB), current sources, and voltage

[6] 5 is arbitrarily chosen in this case to ensure $I_{DAC,j}$ is always smaller than the current drawn from the real DUT. We can always increase this number if it is not sufficient for this purpose.

sources are also needed to assist the measurement procedure. The FPGA is programmed to provide the control signals necessary for the operation of the die. Signals such as clocks, resets and enables are supplied by the FPGA. Current sources are needed for the biasing circuitry and for the measurement of the DAC current. Voltage sources are needed for the power supply voltage. The PCB serves as the interface between the packaged die and the other off-chip components listed above.

3.7 Other Method: Direct Probing

Instead of using on-chip circuitry to perform current measurements on the proposed test structure, another approach is called direct probing. In this approach, the wafer does not have to be diced or packaged into different dies; instead, the measurement is done directly on the wafer. A probe card is usually built to interface between an electrical test system and the semiconductor wafer. The purpose of the probe card is to provide an electrical pathway between the test system and the circuits on the wafer to permit electrical measurement.

Probe cards are inserted into a wafer prober, inside which the wafer will be positioned to ensure that precise contact is made between the probe pads on the wafer and the probe tops or leads. The test voltage or current is then fed into the wafer through these probes and the measurement is done by the testers. Off-chip test equipment can usually be made with much higher precision than on-chip test circuitry. The main reason is that on-chip circuits normally have tight area constraints. For example, off-chip passive components can be made very precise while on-chip passive components cannot, mainly due to area budgets.

To demonstrate how precise current measurements can be done off chip, let us consider an off-chip current meter example, the Keithley 2400 sourcemeter. The datasheet shows that for a current range of 1mA, the resolution is 50nA with 0.034% measurement accuracy. This is for 14-bit accuracy, not for 14-bit dynamic range. This is a very high measurement accuracy and is quite hard to achieve on-chip.

3.7.1 Difference between the Two Approaches

For both the on-chip and direct probing approaches, we use the same test structure. The major difference between the two approaches are listed as follows.

Sensing, not Forcing

In Section 3.2.1, we implemented an architectural technique to allow us to use two different measurement approaches to measure the DUTs on the same test structure. For on-chip current measurement, we use the "forcing" approach to guarantee the correct V_{ds}. The forcing is done by putting an operational amplifier in a feedback configuration.

For direct probing, we use the "sensing" approach instead. Since we do not have an opamp to guarantee the voltage at *Node B* as shown in Figure 3-5, we apply a voltage on *Node A* first. We know that the voltage on *Node B* will not be the same as the voltage on *Node A* due to voltage drops across the I/O switches. The only way to know the voltage on *Node B* is through a direct measurement. That is why this method is also called an "indirect" application of V_{ds}, since we cannot directly apply the V_{ds} we want. The actual V_{ds} depends on the amount of current drawn by the DUT and the resistance of the I/O switches.

Probe Pad Limitation

The other difference between the two approaches is that in the direct probing approach, a set of probe pads needs to be designed and laid out with certain specifications. The dimension of each probe pad is in the range of 60μm x 60μm. These probe pads are quite large and can take up significant chip area. Due to the sizes of these probe pad, we are usually limited to having as few of them as possible.

Because of this limitation, some portion of the parallel design architecture implemented for speed purposes has to be replaced with a pipelined architecture. For example, in the previous architecture, all the binary selection signals of the decoders indicated in Figure 3-2 are loaded in parallel. As a result, we would need a total of

17 probe pads just for the 9-bit row decoder and 8-bit column decoder. To minimize the number of pads, a pipelined or serialized architecture to load in one bit at a time is proposed and designed. In terms of speed, this could be much slower compared to the previous parallel implementation, but this is the tradeoff needed to dedicate more area for the DUTs instead of the probe pads.

3.7.2 Direct Probing vs. On-Chip Current Measurement

This section discusses the advantage and disadvantage of direct probing versus on-chip current measurement approaches. Seven different aspects that commonly concern test circuit designers will be discussed and analyzed.

Time

Time can be discussed in two different aspects: design time and test time. Design time is the amount of time required to design a test structure and a test circuit; test time is the amount of time required to finish testing and extract the result.

In our proposed architecture, the direct probing and the on-chip current measurement approach share the same test structure; therefore, the design time for the test structure is the same for both approaches. However, the design for the test circuit is much more complicated in the case of on-chip current measurement compared to the case of direct probing. A complete on-chip current measurement system, including a current integrating ADC, a current steering DAC, an operational amplifier and a comparator, must be designed and integrated. As a result, on-chip current measurement can take much longer in design time.

In terms of testing time, the on-chip approach is much faster than direct probing. For on-chip measurement, the speed is limited by the speed of the test circuit; while for off-chip measurement, the speed is limited by response time of the tester. For our on-chip design, the maximum clock frequency is 50MHz and it takes roughly 2^{13} cycles on average to perform the charging and discharging. As a result, it takes about 160μs to complete one measurement. For the direct probing approach, the

tester has to first feed the row and column selecting signals to the test chip, all in series. Then, a voltage is applied and the current coming from the DUT is measured. Another voltage measurement is needed in order to obtain the correct V_{ds}.[7] The entire procedure takes a few milliseconds per measurement. The time is mainly due to the slow settling caused by large off-chip capacitance associated with the probes and wires. Therefore, the on-chip current measurement is about an order of magnitude faster than the direct probing approach in testing time. If a full set of measurements on all devices takes three days using the on-chip current measurement approach, it will take one month using the direct probing approach.

When considering the total time, it is important to note that the design time is a one-time consideration, but test time per DUT accumulates based on the number of measurements we want to perform. In order to have statistically significant results, enough data needs to be collected and analyzed. Test time usually has more weight on the total time compared to design time. As a result, on-chip current measurement has an advantage over direct probing in terms of time.

Cost and Resources

In terms of cost and resources, the direct probing approach requires a probe card and a tester. Each probe card design costs around $5,000. A good tester used in semiconductor foundries, such as in the Taiwan Semiconductor Manufacturing Company (TSMC), costs a few $100,000.

For the on-chip current measurement case, the total cost of packaging, FPGA, PCB, voltage source and current sources adds up to less than $10,000. The on-chip current measurement scheme is more cost effective, particularly if one must invest in new test equipment or resources.

Ease of Testing

In terms of ease of testing, direct probing requires an "indirect" application of V_{ds}: the approach has to apply a voltage first and then sense the actual voltage at the

[7] Section 3.7.1 describes why another voltage measurement is necessary.

drain node, due to the unknown voltage drop across the I/O switch, as described in Section 3.7.1. It is inconvenient to not be able to directly apply V_{ds}. For example, to extract the saturation current of a transistor, we want to apply $V_{gs} = V_{dd}$ and $V_{ds} = V_{dd}$. However, since we cannot be sure that V_{ds} is equal to V_{dd} due to this voltage drop, a couple of iterations are needed to find the saturation current. In the case of on-chip current measurement, we can directly force the drain voltage to be the voltage we want through an opamp feedback configuration. In this sense, the on-chip measurement is easier.

Due to the high dynamic range requirement, the on-chip current measurement must go through the measurement flow described in Section 3.6 in order to find the correct DAC current. In the case of direct probing, these steps are not necessary since the current can be measured directly through the off-chip tester. However, all of the steps described in the flow for on-chip current measurement can be fully automated through a control algorithm, which can be implemented by a FPGA. Thus, in terms of ease, both approaches are comparable.

On balance, it is easier to use on-chip current measurement because the drain voltage can be directly applied to the DUTs.

Amount of Data

The amount of data we collect determines the statistical significance of the results. Furthermore, the finer the spatial sampling (enabled by testing of more DUTs), the better the ability to estimate spatial variation effects we have. The number of measurements that can feasibly be gathered for each chip depends on the amount of time required for each DUT measurement. Since the measurement speed for on-chip current measurement is much faster than the measurement speed for direct probing, we can collect much more data using the on-chip current measurement scheme. Thus, the conclusion we derive using on-chip current measurement is expected to be more statistically significant and have better spatial resolution.

Scalability

In order to characterize systematic variation in semiconductor manufacturing, a test circuit and a test structure are designed to characterize how transistor performance changes with respect to a layout change. It would be useful if this test circuit and test structure can be reused to study the same variation effect in a newer technology without significant redesigns.

The direct probing approach is easily scalable from one technology to another since there is no analog circuit involved in the design. On the other hand, the on-chip current measurement scheme is based heavily on analog circuit design. Analog characteristics, such as output resistance, biasing points, and transconductance, do not scale linearly with the technology. A re-simulation or possibly re-design of key circuit components may be needed to ensure the test circuit still works under the new technology.

Therefore, in terms of scalability, it is easier to transfer the design using a direct probing approach than an on-chip current measurement.

Variety of Test Matrix

The off-chip direct probing approach is often limited by parasitics of passive components, such as the capacitances associated with the probes or the inductances associated with the bond wires. With these passive components, it is very difficult to perform high frequency measurements off-chip. Thus, off-chip measurement is limited to only DC or very low frequency. On the other hand, it is much easier to perform high frequency measurement on-chip and the measurement results can be converted to low frequency before coming off chip. Therefore, in terms of variety of test matrix, on-chip measurement can be both high frequency and low frequency, while off-chip measurement is limited to lower frequency.

Accuracy

Which measurement scheme is more accurate? This is the one of the key questions we want to answer using this thesis. No research has shown the comparison between the two approaches. The designs presented in this thesis are well-suited for this purpose, because we are using both approaches to measure the same test structure. The measurement difference between the two approaches is easily comparable.

Table 3.1 below summarizes and compares the two measurement approaches. Overall, we believe that on-chip current measurement is more advantageous compared to direct probing and justifies the additional design effort required.

Aspect	On-Chip	Direct Probing
Time	O	
Cost and Resources	O	
Ease of Testing	O	
Amount of Data	O	
Scalability		O
Variety of Test Matrix	O	
Accuracy	?	?
Winner	O	

Table 3.1: Comparison between on-chip current measurement and direct probing.

3.8 Summary

This chapter motivated the need for a new test circuit by pointing out the inadequacy in previous test circuit designs from the literature for the systematic variation we want to study. A list of new features including (1) a common test structure between different measurement schemes, (2) DUT independence, (3) a high DUT number and density, (4) minimum peripheral transistors, (5) separation of the peripheral transistors from DUTs, and (6) high dynamic range on-chip current measurement, were proposed as guidelines for our new test circuit design. The discussion was divided into the architecture of arrangement and the architecture of measurement.

In the discussion of the architecture of arrangement, a new hierarchical accessing

scheme was designed to allow us to access every individual transistor in the test structure. The transistor characteristics can be measured and extracted independent of other DUTs in the same test structure. Several architectural innovations, such as the creation of a DUT array, the arrangement of the DUT array, the number of transistors within the DUT array, the use of I/O devices as switches, and the application of negative gate voltages, allow us to precisely know the drain voltage of the DUTs and accurately measure the DUT current. The test structure design using this architecture of arrangement can be used by both measurement approaches, on-chip current measurement and off-chip direct probing. Being able to use the same test structure makes it easier to compare the results obtained using different approaches in the future. A highly dense layout with 131,072 NMOS transistors and 131,072 PMOS transistors are placed within the 2mm by 3mm test structure to ensure the statistical significance of the measurement and the spatial resolution required for future modeling.

The discussion of the architecture of measurement is divided into two parts: on-chip current measurement approach and direct probing approach. For the on-chip current measurement approach, the integrating ADC architecture is selected over all the other ADC architectures based on its high accuracy, its low offset and gain error, and its minimal chip area requirement. Although the integrating ADC can only be used to measure slow moving signals, it is desired for our design because the current measurement we are performing is at DC. In the new design, several architectural changes are made to the traditional integrating ADC to achieve high dynamic range current measurement, and immunity to comparator offset and switch charge injection. Each key analog circuit component is discussed in terms of its purpose, and its design and layout challenges in the integrating ADC.

In the current steering DAC design, the high output impedance is achieved by cascading and a local feedback network. A low DNL is achieved by matching between unit DAC cell and layout optimization. The comparator is designed using a preamplifier followed by a track-and-latch stage architecture. The low gain and low output impedance design of the preamplifier allows us to have high bandwidth and

low kicknoise in our design. Finally, in the operational amplifier design, we pointed out the common misconception that ΔV_{IN} depends only on the gain of the amplifier. An analysis was presented to show that ΔV_{IN} depends on the charging rate and the unity gain bandwidth of the operational amplifier. A measurement flow was also presented to summarize the steps we need in order to perform on-chip current measurement using the above proposed architecture.

The second approach we discussed is the direct probing approach. Instead of forcing the DUT voltage as in the case of on-chip current measurement, the drain voltage has to be indirectly applied and sensed. A limit on the number of probe pads is set to ensure the probe pads do not take up too much of the total die area. A comparison, in terms of seven different aspects of the measurement scheme, between the two approaches is also presented. Overall, the on-chip current measurement scheme is suggested to have more advantages than the direct probing measurement scheme.

Chapter 4

Thesis Contributions and Future Work

This thesis has demonstrated, from various aspects, the need for a new test structure and a new test circuit design to study the systematic variation due to STI and polysilicon pattern density. Beginning with the analysis of a few key transistor parameters, we were able to show that understanding variation in these parameters can determine our ability to design functional circuits and more importantly, our ability to continue the scaling trend in the future. A new test structure to accentuate the variation effect due to STI and polysilicon pattern density, and a new test circuit to measure current with the dynamic range over four orders of magnitude, are designed for this purpose. This concluding chapter summarizes this thesis and evaluates its contributions. Even though there are many advances in the design of the test structure and the test circuit, there is still much room for improvement. The last section provides suggestions for future work.

4.1 Thesis Summary

As process technology continues to scale, variation becomes a more important challenge to overcome in the near future. The increases in variation can be attributed to many sources. One example would be the decreasing distance between neighboring

transistors: more interaction between the surrounding layout features and the transistor characteristics is expected. Nearby transistors can influence each other's performance. Another example leading to the increase of variability is the introduction of new processing steps into the semiconductor manufacturing process. The introduction of RTA and stress engineering, for instance, can cause more layout-induced systematic variation. This kind of systematic variation is the focus of our thesis.

In Chapter 2, we motivated the need to analyze variation using a different set of transistor parameters, such as mobility and virtual source velocity, that are rarely viewed as subjects to study variation. These parameters can help designers to better understand the physics behind the manufacturing steps, and they also help reflect performance variation in terms of the scaling trend. STI and polysilicon pattern density were chosen as the main design parameters to study systematic variation.

Strain engineering is necessary to improve the transistor performance further once we scale below 65nm. STI can help introduce stress in strain engineering; however, it can also introduce unintentional stress that induces systematic variation between transistors. Mobility and virtual source velocity is highly correlated with the amount of stress experienced by the transistor. Since we want to study mobility and virtual source velocity to understand the scaling trend, STI pattern density was chosen as one of the design parameters. The introduction of a new RTA process can help to create very shallow junctions in order to maintain the electrostatic integrity of the scaled transistors. However, the uniformity of temperature profile is strongly affected by polysilicon layout. Non-uniform polysilicon layout can introduce systematic variation due to RTA. Since we want to better understand the RTA process, polysilicon pattern density is chosen as another parameter.

The key themes of this thesis are to study transistors on an individual basis, clearly define the term "pattern density," extract the long-range and short-range characteristic length of the systematic variation, and collect enough data to have statistically significant results. Many of the previous research in the literature failed to achieve one or more of the above goals; therefore, their results have limited impact and usefulness.

A new test structure and new test circuit are designed for this research purpose. Figure 4-1 below summarizes the flow of this thesis.

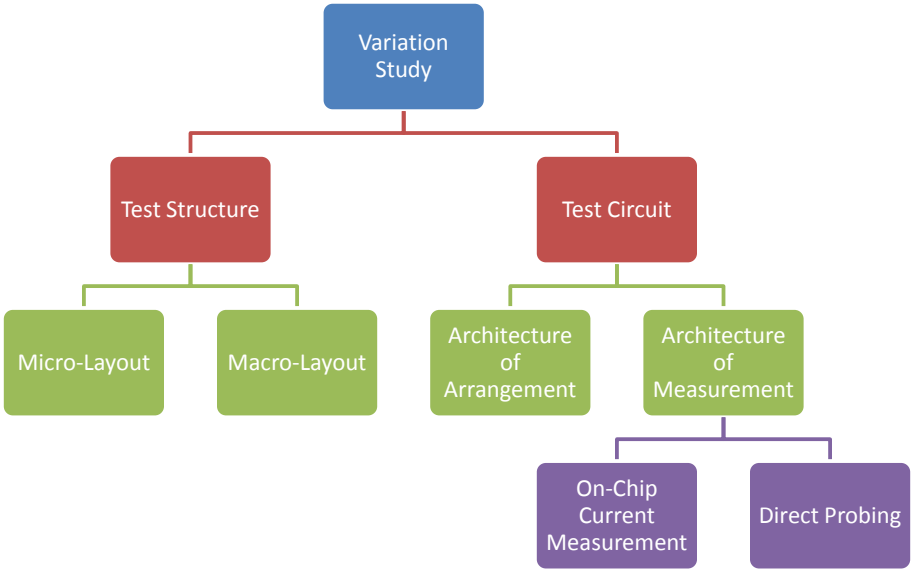

Figure 4-1: Thesis summary.

The test structure design is divided into macro- and micro-layout design. In macro-layout design, the test structure is divided into six different regions, each region with a unique STI or polysilicon pattern density. A careful design of experiments analysis is done to ensure we have all combinations of short-range and long-range pattern densities. A step input is also built into the test structure to accentuate the variation effect due to pattern density changes. In micro-layout design, different transistor dimensions, number of polysilicon fingers, spacing between fingers, and length of the active region are explored for our variation studies. The DUT layout pattern is designed such that we can obtain a good spatial resolution of the DUTs that are most commonly used in a design.

The test circuit design is divided into the architecture of arrangement and the architecture of measurement. In the architecture of arrangement design, a new hierarchical accessing scheme was designed to allow individual transistor measurement on the test structure with 131,072 NMOS transistors and 131,072 PMOS transistors. This test structure design has the minimum ratio of peripheral transistors to DUTs in the literature thus far. A number of architectural innovations, such as the cre-

ation of the DUT array, the arrangement of the DUT array, the number of transistors within the DUT array, the use of I/O devices as switches, and the optimization of gate voltage application, are adopted to enable us to mitigate leakage current from other unwanted DUTs. An accurate current measurement on every DUT can be done. Moreover, this test structure design has a dual usage: it can be used for both the on-chip current measurement and off-chip direct probing. This is an essential feature of the test structure to allow us to have a fair comparison between the results obtained from the two different measurement approaches.

For the on-chip current measurement approach, the integrating ADC architecture is selected over the other ADC architectures because of the high accuracy, low offset and gain errors, and small hardware requirement. A number of new architectural changes, such as the addition of current steering DAC and switches, the alternation of charging and discharging phases, and the changes of charging bias voltage, are designed and implemented into the traditional current integrating ADC for our test structure measurement. These changes allow us to have dynamic range of over four orders of magnitude, and comparator offset and switch charge injection immunity.

A high output impedance, low leakage, and low DNL current steering DAC is designed. The high impedance is necessary for the DAC design, to ensure that the DAC output current only changes due to the change of the DAC input codes, but does not change due to the change of the voltage applied across the DAC. Cascoding and local feedback configuration is designed into the unit DAC cell to increase the output impedance. Our simulation shows that the current change is less than 0.0035% in the range of voltage that we are interested in. Low leakage current during the off-state is achieved by using I/O devices and stacking effect. Low DNL and matching is achieved by only using the multiples of the LSB DAC cell to design the upper bits, and through layout optimization. We concluded that in order to have good matching, significant amount of area is needed.

A two-stage comparator is designed with a preamplifier stage followed by a track-and-latch stage. The low gain and low output impedance of the preamplifier design allow the overall comparator to have high bandwidth and low kickback noise, respec-

tively. A two-stage opamp is also designed with bandwidth of 270MHz. A ΔV_{IN} of less than 5μV can be achieved. We conclude that in order to improve the dynamic ΔV_{IN} of the opamp during the charging cycle, we have to increase the unity bandwidth, not just the gain. It is a common misconception to confuse between the static ΔV_{IN}, which only depends on the DC gain, and the dynamic ΔV_{IN}.

The direct probing approach is discussed and compared to the on-chip current measurement approach. Using the common figures of merit in test circuit design, we conclude that an on-chip current measurement scheme is more advantageous in terms of time, cost and resources, ease of testing, the amount of data, and variety of test matrix. On the other hand, the direct probing approach is more advantageous in terms of scalability to newer technology. We cannot yet compare the accuracy between the two and this is one of the main goals for the future work of this thesis.

The future goals of this project, following fabrication of our test chips, are to (1) extract the characteristic length of both STI and polysilicon, (2) study the systematic variation effect on a couple of key transistor parameters, including mobility and virtual source velocity, (3) compare the measurement results using on-chip current measurement and direct probing, and (4) model this result into the existing transistor models.

4.1.1 Thesis Contributions

Based on the summary above, the primary contributions of this thesis are:

- Propose a different set of transistor parameters, such as mobility and virtual source velocity, as a target for variation studies. These sets of transistor parameters have better physical meaning and they also help to reflect the scaling trend.

- Identify STI and polysilicon pattern density as the design parameters to study systematic variation. STI pattern density is chosen due to the emerging of strain engineering, and polysilicon pattern density is chosen due to the introduction of rapid thermal annealing processes.

- Design a test structure that includes macro- and micro-layout architecture. The macro-layout architecture can help us identify the characteristic length of both long-range and short-range influence; and the micro-layout architecture can help us study how systematic variation reacts to local parameter changes.

- Design an architecture of arrangement to allow individual transistor access for every DUT (131,072 NMOS and 131,072 PMOS) on the test structure. The design has the minimum ratio of peripheral transistors to DUTs in the literature thus far. This design can also be measured using either on-chip current measurement or direct probing.

- Design an on-chip current measurement circuit with dynamic range over four orders of magnitude.

- Provide a complete analysis and comparison between the on-chip current measurement and off-chip direct probing.

4.2 Future Work

In addition to measuring the transistors, and extracting the parameter variation and the characteristic length of the layout upon successful chip fabrication, other future work might help us improve the understanding of process variation further. In this section, we provide some thoughts on potential future projects.

4.2.1 Characteristic Length

The characteristic length of the layout is used for the calculation of the effective pattern density. One of the main challenges in designing the test structure is that we do not know the number of physical effects at work and the corresponding number of characteristic lengths that need to be estimated, and we do not know beforehand the value of each characteristic length. As we may recall, each region has the size of 1mm by 1mm. Within each region, there could be a large square with the dimension

of 400μm by 400μm or a small square with the dimension of 100μm by 100μm. As described in Section 2.3.3, using this macro-layout, we are assuming there are two characteristic lengths. The larger characteristic length is assumed to be within 1mm and the smaller characteristic length is less than 400μm.

Even though these assumptions are made based on the previous literature studies, it could be possible that the characteristic length is outside the range we expected. No matter if this initial guess is a good estimate of the real characteristic length or not, with the successful measurement of this test structure, this information can give us a better estimate of this characteristic length in the future designs.

4.2.2 Parasitic Capacitance

In this thesis, we analyze the systematic variation due to STI and polysilicon pattern density. All the measurements done here are DC measurement of current at different biasing points. In order to characterize transistor performance, a delay metric incorporating AC parameters, such as parasitic capacitance, is needed. An example performance metric proposed by [75] is shown below:

$$\tau = \frac{(1-\delta)V_{DD} - V_T + (C_f^* V_{DD}/C_{inv}L_G)}{(3-\delta)V_{DD}/4 - V_T}\frac{L_G}{\upsilon} \tag{4.1}$$

where C_f^* represents the equivalent gate fringing capacitance, with Miller effect taken into account, and υ represents the effective velocity of the carriers. From this equation, we can see that both velocity (DC) and fringing capacitance (AC) play a major role in determining transistor performance. As technology continues to scale, the gate to fringing capacitance becomes even more important in determining the transistor performance, because its magnitude can exceed that of the intrinsic gate capacitance below 90nm technology [76]. In order to give circuit designers an accurate assessment of the variation in delay performance, a study of systematic variation in the parasitic capacitance is needed.

Using a test structure similar to ours, it would be interesting to investigate how STI and polysilicon pattern density affect the systematic variation of parasitic fringing

capacitance. A hierarchical accessing scheme can be designed to extract the fringing capacitance of each DUT individually on the test structure. The only difference is that an AC measurement needs to be performed instead of the DC current measurement. However, doing an AC measurement is quite difficult using the same hierarchical accessing scheme, since there is no way to completely turn off a transistor AC-wise. The AC measurement current can leak through any parasitic capacitance of the off-DUTs.

A method called *charge-based capacitive measurement* (CBCM) can be used to measure transistor capacitance using charge based operation instead of doing AC measurements [91, 92]. The idea of CBCM can be extended in the design of a hierarchical accessing scheme to study variation effect of the fringing capacitance.

4.2.3 Modeling

After successful measurement and extraction of the systematic variation due to STI and polysilicon pattern density, a model needs to be built in order for this measurement result to be useful for circuit designers. A commonly used predictive technology model (PTM) built by [70] has started to include emerging physical effects, such as process variations and correlations among model parameters, in their model to accurately predict the characteristics of nanoscale CMOS technology. A potential project would be to incorporate the modeling result we obtain using this test structure into the widely used PTM.

Bibliography

[1] "International Technology Roadmap for Semiconductors 2007 Edition," tech. rep., International Technology Roadmap for Semconductors, 2007.

[2] C. Wang, K.-W. Choi, W.-E. Fu, R. Jones, D. Ho, C. Soles, E. Lin, W. li Wu, J. Clarke, J. Villarrubia, and B. Bunday, "Linewidth Roughness and Cross-sectional Measurements of Sub-50 nm Structures Using CD-SAXS and CD-SEM," pp. 142–147, IEEE/SEMI Advanced Semiconductor Manufacturing Conference, May 2008.

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

[4] N. Drego, A. Chandrakasan, and D. Boning, "Lack of Spatial Correlation in MOSFET Threshold Voltage Variation and Implications for Voltage Scaling," *IEEE Transactions on Semiconductor Manufacturing*, vol. 22, pp. 245–255, May 2009.

[5] K. Kuhn, "Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS," pp. 471–474, IEEE International Electron Devices Meeting, Dec. 2007.

[6] D. A. Antoniadis, "Technology Trends and Requirements of Future High Performance CMOS FETs." presented at the DARPA Review, July 2008.

[7] N. Drego, A. Chandrakasan, and D. Boning, "A Test-Structure to Efficiently Study Threshold-Voltage Variation in Large MOSFET Arrays," pp. 281–286, 8th International Symposium on Quality Electronic Design, March 2007.

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

[9] M. Orshansky, S. R. Nassif, and D. Boning, *Design for Manufacturability and Statistical Design: A Constructive Approach*. New York: Springer Science+Business Media, LLC, 1 ed., 2008.

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

[11] T. Mizuno, J. Okumtura, and A. Toriumi, "Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's," *IEEE Transactions on Electron Devices*, vol. 41, pp. 2216–2221, Nov. 1994.

[12] P. Stolk and D. Klaassen, "The effect of statistical dopant fluctuations on MOS device performance," pp. 627–630, IEEE International Electron Devices Meeting, Dec. 1996.

[13] P. Oldiges, Q. Lin, K. Petrillo, M. Sanchez, M. Ieong, and M. Hargrove, "Modeling line edge roughness effects in sub 100 nanometer gate length devices," pp. 131–134, International Conference on Simulation of Semiconductor Processes and Devices, 2000.

[14] M. Mani, A. Devgan, and M. Orshansky, "An efficient algorithm for statistical minimization of total power under timing yield constraints," pp. 309–314, Design Automation Conference, June 2005.

[15] S. Thompson, P. Packan, and M. Bohr, "MOS Scaling: Transistor Challenges for the 21st Century," *Intel Technology Journal Q398*, pp. 1–19, 1998.

[16] K. Kuhn, C. Kenyon, A. Kornfeld, M. Liu, A. Maheshwari, W.-K. Shih, S. Sivakumar, G. Taylor, P. VanDerVoorn, and K. Zawadzki, "Managing Process Variation in Intel's 45nm CMOS Technology," *Intel Technology Journal: Intel's 45nm CMOS Technology*, vol. 12, no. 02, pp. 92–110, 2008.

[17] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, pp. 1433–1439, Oct. 1989.

[18] A. Asenov, S. Kaya, and J. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations," *IEEE Transactions on Electron Devices*, vol. 49, pp. 112–119, Jan. 2002.

[19] D. Frank, R. Dennard, E. Nowak, P. Solomon, Y. Taur, and H.-S. P. Wong, "Device scaling limits of Si MOSFETs and their application dependencies," *Proceedings of the IEEE*, vol. 89, pp. 259–288, Mar. 2001.

[20] M. Orshansky, C. Spanos, and C. Hu, "Circuit Performance Variability Decomposition," pp. 10–13, Proceedings of Workshop on Statistical Metrology, 1999.

[21] V. Mehrotra, S. Nassif, D. Boning, and J. Chung, "Modeling the effects of manufacturing variation on high-speed microprocessor interconnect performance," pp. 767–770, IEEE International Electron Devices Meeting, Dec. 1998.

[22] P. Zarkesh-Ha, T. Mule, and J. Meindl, "Characterization and modeling of clock skew with process variations," pp. 441–444, IEEE Custom Integrated Circuits Conference, 1999.

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

[24] B. Morgan, C. M. Waits, and R. Ghodssi, "Compensated aspect ratio dependent etching (CARDE) using gray-scale technology," *Microelectron. Eng.*, vol. 77, no. 1, pp. 85–94, 2005.

[25] R. A. Gottscho, C. W. Jurgensen, and D. J. Vitkavage, "Microscopic uniformnity in plasma etching," *Journal of Vaccum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 10, pp. 2133–2147, Sept. 1992.

[26] D. Ouma, D. Boning, J. Chung, W. Easter, V. Saxena, S. Misra, and A. Crevasse, "Characterization and modeling of oxide chemical-mechanical polishing using planarization length and pattern density concepts," *IEEE Transactions on Semiconductor Manufacturing*, vol. 15, pp. 232–244, May 2002.

[27] D. Frank, Y. Taur, M. Ieong, and H.-S. Wong, "Monte Carlo modeling of threshold variation due to dopant fluctuations," pp. 171–172, Symposium on VLSI Circuits, 1999.

[28] 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," pp. 75–76, Symposium on VLSI Technology, 1999.

[29] H. Sasaki, M. Ono, T. Yoshitomi, T. Ohguro, S. Nakamura, M. Saito, and H. Iwai, "1.5 nm direct-tunneling gate oxide Si MOSFET's," *IEEE Transactions on Electron Devices*, vol. 43, pp. 1233–1242, Aug. 1996.

[30] C.-H. Choi, K.-Y. Nam, Z. Yu, and R. Dutton, "Impact of gate direct tunneling current on circuit performance: a simulation study," *IEEE Transactions on Electron Devices*, vol. 48, pp. 2823–2829, Dec. 2001.

[31] S. Borkar, "Designing reliable systems from unreliable components: the challenges of transistor variability and degradation," *IEEE Micro*, vol. 25, pp. 10–16, Nov.-Dec. 2005.

[32] T. Skotnicki, J. Hutchby, T.-J. King, H.-S. Wong, and F. Boeuf, "The end of CMOS scaling: toward the introduction of new materials and structural changes to improve MOSFET performance," *IEEE Circuits and Devices Magazine*, vol. 21, pp. 16–26, Jan.-Feb. 2005.

[33] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos, "Modeling within-die spatial correlation effects for process-design co-optimization," pp. 516–521, Sixth International Symposium on Quality of Electronic Design, March 2005.

[34] K. Gettings, *Study of CMOS Process Variation by Multiplexing Analog Characteristics*. Ph.D. Thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, June 2007.

[35] D. Lim, *Characterization of Process Variability and Robust Optimization of Analog Circuits*. Ph.D. Thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Sept. 2008.

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

[37] W. Schemmert and G. Zimmer, "Threshold-voltage Sensitivity of Ion-Implanted M.O.S Transistors due to Process Variation," *Electronics Letters*, vol. 10, pp. 151–152, May 1974.

[38] L.-T. Pang and B. Nikolic, "Impact of Layout on 90nm CMOS Process Parameter Fluctuations," pp. 69–70, Symposium on VLSI Circuits, 2006.

[39] H. Tsuno, K. Anzai, M. Matsumura, S. Minami, A. Honjo, H. Koike, Y. Hiura, A. Takeo, W. Fu, Y. Fukuzaki, M. Kanno, H. Ansai, and N. Nagashima, "Advanced Analysis and Modeling of MOSFET Characteristic Fluctuation Caused by Layout Variation," pp. 204–205, IEEE Symposium on VLSI Technology, June 2007.

[40] V. Moroz, L. Smith, X.-W. Lin, D. Pramanik, and G. Rollins, "Stress-aware design methodology," pp. 807–812, International Symposium on Quality Electronic Design, March 2006.

[41] A. Kahng, P. Sharma, and R. Topaloglu, "Exploiting STI stress for performance," pp. 83–90, IEEE/ACM International Conference on Computer-Aided Design, Nov. 2007.

[42] A. Kahng, P. Sharma, and R. Topaloglu, "Chip Optimization Through STI-Stress-Aware Placement Perturbations and Fill Insertion," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, pp. 1241–1252, July 2008.

[43] R. Topaloglu, "Standard Cell and Custom Circuit Optimization using Dummy Diffusions through STI Width Stress Effect Utilization," pp. 619–622, IEEE Custom Integrated Circuits Conference, Sept. 2007.

[44] G. Scott, J. Lutze, M. Rubin, F. Nouri, and M. Manley, "NMOS drive current reduction caused by transistor layout and trench isolation induced stress," pp. 827–830, IEEE International Electron Devices Meeting, 1999.

[45] Y.-M. Sheu, K.-W. Su, S.-J. Yang, H.-T. Chen, C.-C. Wang, M.-J. Chen, and S. Liu, "Modeling well edge proximity effect on highly-scaled MOSFETs," pp. 831–834, IEEE Custom Integrated Circuits Conference, Sept. 2005.

[46] S. Eneman, P. Verheyen, R. Rooyackers, F. Nouri, L. Washington, R. Degraeve, B. Kaczer, V. Moroz, A. De Keersgieter, R. Schreutelkamp, M. Kawaguchi, Y. Kim, A. Samoilov, L. Smith, P. Absil, K. De Meyer, M. Jurczak, and S. Biesemans, "Layout impact on the performance of a locally strained PMOSFET," pp. 22–23, Symposium on VLSI Technology, June 2005.

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

[48] C. Gallon, G. Reimbold, G. Ghibaudo, R. Bianchi, R. Gwoziecki, S. Orain, E. Robilliart, C. Raynaud, and H. Dansas, "Electrical analysis of mechanical stress induced by STI in short MOSFETs using externally applied stress," *IEEE Transactions on Electron Devices*, vol. 51, pp. 1254–1261, Aug. 2004.

[49] M. Miyamoto, H. Ohta, Y. Kumagai, Y. Sonobe, K. Ishibashi, and Y. Tainaka, "Impact of reducing STI-induced stress on layout dependence of MOSFET characteristics," *IEEE Transactions on Electron Devices*, vol. 51, pp. 440–443, March 2004.

[50] V. Moroz, G. Eneman, P. Verheyen, F. Nouri, L. Washington, L. Smith, M. Jurczak, D. Pramanik, and X. Xu, "The Impact of Layout on Stress-Enhanced Transistor Performance," pp. 143–146, International Conference on Simulation of Semiconductor Processes and Devices, Sept. 2005.

[51] N. Wils, H. Tuinhout, and M. Meijer, "Influence of STI stress on drain current matching in advanced CMOS," pp. 238–243, IEEE International Conference on Microelectronic Test Structures, March 2008.

[52] N. R. Zangenberg, J. Fage-Pedersen, J. L. Hansen, and A. N. Larsen, "Boron and phosphorus diffusion in strained and relaxed Si and SiGe," *Journal of Applied Physics*, vol. 94, no. 6, pp. 3883–3890, 2003.

[53] Y.-M. Sheu, S.-J. Yang, C.-C. Wang, C.-S. Chang, L.-P. Huang, T.-Y. Huang, M.-J. Chen, and C. Diaz, "Modeling mechanical stress effect on dopant diffusion in scaled MOSFETs," *IEEE Transactions on Electron Devices*, vol. 52, pp. 30–38, Jan. 2005.

[54] F. Liu and K. Agarwal, "A Test Structure for Assessing Individual Contact Resistance," pp. 201–204, IEEE International Conference on Microelectronic Test Structures, April 2009.

[55] K. Gonzalez-Valentin, "Extraction of variation sources due to layout practices," Masters Thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, June 2002.

[56] 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," pp. 67–68, Symposium on VLSI Circuits, 2006.

[57] "BSIM4 Manual," tech. rep., University of California, Berkeley, CA, 2005.

[58] "International Technology Roadmap for Semiconductors 2006 Edition," tech. rep., International Technology Roadmap for Semconductors, 2006.

[59] D. Johns and K. Martin, *Analog Integrated Circuit Design*. Wiley, 1996.

[60] Y. Taur and T. H. Ning, *Fundamental of Modern VLSI Devices*. Cambridge University Press, 1998.

[61] J. Rabaey and A. Chandrakasan and B. Nikolic, *Digital Integrated Circuits*. Prentice Hall, 2003.

[62] T. Gebel, M. Voelskow, W. Skorupa, G. Mannino, V. Privitera, F. Priolo, E. Napolitani, and A. Carnera, "Flash lamp annealing with millisecond pulses for ultra-shallow boron profiles in silicon," *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, vol. 186, no. 1-4, pp. 287–291, 2002.

[63] T. Ito, T. Iinuma, A. Murakoshi, H. Akutsu, K. Suguro, T. Arikado, K. Okumura, M. Yoshioka, T. Owada, Y. Imaoka, H. Murayama, and T. Kusuda, "Flash Lamp Anneal Technology for Effectively Activating Ion Implanted Si," pp. 182–183, International Conference on Solid State Devices and Materials, Sept. 2001.

[64] T. S. Cho, K. jae Lee, J. Kong, and A. Chandrakasan, "The design of a low power carbon nanotube chemical sensor system," pp. 84–89, ACM/IEEE Design Automation Conference, June 2008.

[65] K. Agarwal, S. Nassif, F. Liu, J. Hayes, and K. Nowka, "Rapid Characterization of Threshold Voltage Fluctuation in MOS Devices," pp. 74–77, IEEE International Conference on Microelectronic Test Structures, March 2007.

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

[67] I. Ahsan, N. Zamdmer, O. Glushchenkov, R. Logan, E. Nowak, H. Kimura, J. Zimmerman, G. Berg, J. Herman, E. Maciejewski, A. Chan, A. Azuma, S. Deshpande, B. Dirahoui, G. Freeman, A. Gabor, M. Gribelyuk, S. Huang, M. Kumar, K. Miyamoto, D. Mocuta, A. Mahorowala, E. Leobandung,

H. Utomo, and B. Walsh, "RTA-Driven Intra-Die Variations in Stage Delay, and Parametric Sensitivities for 65nm Technology," pp. 170–171, Symposium on VLSI Technology, 2006.

[68] P. Timans, W. Lerch, J. Niess, S. Paul, N. Acharya, and Z. Nenyei, "Challenges for ultra-shallow junction formation technologies beyond the 90 nm node," pp. 17–33, IEEE International Conference on Advanced Thermal Processing of Semiconductors, Sept. 2003.

[69] A. Shima, Y. Wang, S. Talwar, and A. Hiraiwa, "Ultra-shallow junction formation by non-melt laser spike annealing for 50-nm gate CMOS," pp. 174–175, Symposium on VLSI Technology, 2004. Digest of Technical Papers, June 2004.

[70] W. Zhao and Y. Cao, "New Generation of Predictive Technology Model for Sub-45 nm Early Design Exploration," *IEEE Transactions on Electron Devices*, vol. 53, pp. 2816–2823, Nov. 2006.

[71] V. Wang, K. Agarwal, S. Nassif, K. Nowka, and D. Markovic, "A Design Model for Random Process Variability," pp. 734–737, International Symposium on Quality Electronic Design, March 2008.

[72] D. White, D. Boning, S. Butler, and G. Barna, "Spatial characterization of wafer state using principal component analysis of optical emission spectra in plasma etch," *IEEE Transactions on Semiconductor Manufacturing*, vol. 10, pp. 52–61, Feb 1997.

[73] W. Zhao, Y. Cao, F. Liu, K. Agarwal, D. Acharyya, S. Nassif, and K. Nowka, "Rigorous extraction of process variations for 65nm CMOS design," pp. 89–92, European Solid State Circuits Conference, Sept. 2007.

[74] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm design exploration," pp. 6 pp.–590, International Symposium on Quality Electronic Design, March 2006.

[75] A. Khakifirooz and D. Antoniadis, "MOSFET Performance Scaling-Part I: Historical Trends," *IEEE Transactions on Electron Devices*, vol. 55, pp. 1391–1400, June 2008.

[76] A. Khakifirooz and D. Antoniadis, "MOSFET Performance Scaling-Part II: Future Directions," *Electron Devices, IEEE Transactions on*, vol. 55, pp. 1401–1408, June 2008.

[77] D. A. Antoniadis, I. Aberg, C. Chleirigh, O. M. Nayfeh, A. Khakifrooz, and J. L. Hoyt, "Continuous MOSFET performance increase with device scaling: The role of strain and channel material innovations," *IBM Journal of Research and Development*, vol. 50, pp. 363–376, July/Sept. 2006.

[78] J. Hoyt, H. Nayfeh, S. Eguchi, I. Aberg, G. Xia, T. Drake, E. Fitzgerald, and D. Antoniadis, "Strained silicon MOSFET technology," pp. 23–26, IEEE International Electron Devices Meeting, 2002.

[79] M. Na, E. Nowak, W. Haensch, and J. Cai, "The effective drive current in CMOS inverters," pp. 121–124, IEEE International Electron Devices Meeting, 2002.

[80] A. Lochtefeld, I. J. Djomehri, G. Samudra, and D. A. Antoniadis, "New Insights into carrier transport in n-MOSFETs," *IBM Journal of Research and Development*, vol. 46, pp. 347–357, March/May 2002.

[81] S. Takagi, "Re-examination of subband structure engineering in ultra-short channel MOSFETs under ballistic carrier transport," pp. 115–116, Symposium on VLSI Technology, June 2003.

[82] A. Khakifirooz and D. Antoniadis, "Transistor Performance Scaling: The Role of Virtual Source Velocity and Its Mobility Dependence," pp. 1–4, IEEE International Electron Devices Meeting, 2006. IEDM '06., Dec. 2006.

[83] A. Lochtefeld and D. Antoniadis, "On experimental determination of carrier velocity in deeply scaled NMOS: how close to the thermal limit?," *IEEE Electron Device Letters*, vol. 22, pp. 95–97, Feb. 2001.

[84] A. Lochtefeld and D. Antoniadis, "Investigating the relationship between electron mobility and velocity in deeply scaled NMOS via mechanical stress," *IEEE Electron Device Letters*, vol. 22, pp. 591–593, Dec. 2001.

[85] K. Rim, J. Chu, H. Chen, K. Jenkins, T. Kanarsky, K. Lee, A. Mocuta, H. Zhu, R. Roy, J. Newbury, J. Ott, K. Petrarca, P. Mooney, D. Lacey, S. Koester, K. Chan, D. Boyd, M. Ieong, and H.-S. Wong, "Characteristics and device design of sub-100 nm strained Si N- and PMOSFETs," pp. 98–99, Symposium on VLSI Technology, 2002.

[86] M. Bhushan, M. Ketchen, S. Polonsky, and A. Gattiker, "Ring oscillator based technique for measuring variability statistics," pp. 87–92, IEEE International Conference on Microelectronic Test Structures, March 2006.

[87] B. Ji, D. Pearson, I. Lauer, F. Stellari, D. Frank, L. Chang, and M. Ketchen, "Operational amplifier based test structure for transistor threshold voltage variation," pp. 3–7, IEEE International Conference on Microelectronic Test Structures, March 2008.

[88] N. Ickes, D. Finchelstein, and A. Chandrakasan, "A 10-pJ/instruction, 4-MIPS micropower DSP for sensor applications," pp. 289–292, IEEE Asian Solid-State Circuits Conference, Nov. 2008.

[89] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die

parameter variations on microprocessor frequency and leakage," *IEEE Journal of Solid-State Circuits*, vol. 37, pp. 1396–1402, Nov. 2002.

[90] Q. Yu, S. Zhibiao, C. Ting, and Z. Guohe, "A low kick back noise latched comparator for high speed folding and interpolating ADC," pp. 1973–1976, International Conference on Solid-State and Integrated-Circuit Technology, Oct. 2008.

[91] D. Sylvester, J. Chen, and C. Hu, "Investigation of interconnect capacitance characterization using charge-based capacitance measurement (CBCM) technique and three-dimensional simulation," *IEEE Journal of Solid-State Circuits*, vol. 33, pp. 449–453, Mar. 1998.

[92] Y.-W. Chang, H.-W. Chang, T.-C. Lu, Y.-C. King, W. Ting, J. Ku, and C.-Y. Lu, "Interconnect capacitance characterization using charge-injection-induced error-free (CIEF) charge-based capacitance measurement (CBCM)," *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, pp. 50–56, Feb. 2006.