High Frequency Techniques for Advanced MOS Device Characterization

By

Yun Wang

A Dissertation submitted to the

Graduate School- New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Graduate Program in Electrical and Computer Engineering

Written under the direction of

Professor Kin P. Cheung, Ph. D.

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2008

ABSTRACT OF THE DISSERATATION

High Frequency Techniques for Advanced MOS Device Characterization

by YUN WANG

Dissertation Director

Professor Kin P. Cheung

Rapid advances in the semiconductor industry have led to the proliferation of electric devices and information technology (IT). Integrated circuits(IC) based upon silicon MOSFET's have been used in virtually every electronic device produced today. The competitiveness of this huge market urges an increased device performance with lower cost. Over the past three decades, it is fulfilled by reducing transistor gate lengths and oxide thickness with each new generation of manufacturing technology. The leading edge CMOS technology is currently at the 45nm node with physical gate length at 18 nm and an equivalent gate oxide thickness (EOT) of 0.9 nm. However, as

the device is miniaturized into the nanometer-scale regime nowadays, some challenges abound. Some challenges are new, some are just getting tougher and most of them will continue to become even more difficult to deal with for future generations. It is the world-wide effort to meet these challenges for sustaining the rapid growth of the industry. In this thesis, we will address a few of these challenges and offer some new approaches to get around them. Specifically, we introduce a new measurement technique to solve the precision problem in C-V measurement based on Time domain Reflectrometry(TDR). We also use the combination of experiment and theory to resolve the defect depth-profiling ambiguity associated with charge pumping measurement. Moreover, we find a new mode in transistor degradation that will become much more serious as the transistor size shrinks further. All these results represent a major and important advance which is also timely to the IC industry.

Acknowledgements

I am grateful for this opportunity to acknowledge the people who have made this thesis proposal possible. First and foremost, I would like to thank my advisor Professor Kin.P. Cheung, who has introduced me to the world of research. It is an intellectually stimulating as well as exciting field. His endless enthusiasm and scientific knowledge continues to be a great inspiration to me. And I deeply appreciate the guidance and support he has given me over the years. In addition, I am very grateful for the collaborative opportunities I have with National Institutes of Standard and Technology (NIST). In particular, I must thank Dr. John Suehle, who is also one of my thesis committee members, for his kindness and help to provide me the facilities to do the NBTI experiment. Also the discussion with him has greatly increased the scope and importance of my thesis. I would also like to thank other committee members, those are, Dr.Lu, Dr.Sheng, for their continued guidance and encouragement. At the end, I am grateful to all my friends from Rutgers University, for being the surrogate family during the many years I stayed here and for their continued moral support. Finally, I am forever indebted to my father Dr. Wenbiao Wang , my mother Yunhong Lv and my wife Qiankun Sun for their understanding, endless patience and encouragement when it was most required.

Yun Wang

July 01, 2007

Table of Contents

List of Figures and Tables

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Appendix

Chapter 1
Introduction

The steady downscaling of transistor dimensions over the past two decades has been the main stimulus to the growth of silicon integrated circuits (ICs) and the information industry [1]. The driving force behind the relentless downscaling of the MOSFET in integrated circuit is, as always, cost per function. Smaller size means more transistors per unit area and a chip can either be smaller or do more [2-3]. At the same time, performance must also be improved. To achieve that, the gate dielectric thickness must shrink as well [4-5]. However, as the gate oxide gets thinner, challenges are faced as well. At state-of-the-art 1.2nm EOT MOS devices, the gate oxide is so thin that leakage current reaches over 800A/cm^2 [3]. Even though the high κ dielectrics used in the recent 45nm technology alleviates the leakage current, it is still believed that the leakage current will be higher in the future device because of the aggressive scaling down of oxide thickness. Such high leakage causes many problems. Power dissipation is clearly a serious issue. Even basic device characterizations are suffering.

As basic and important as C-V measurement, high precision is not achievable when the leakage is too high. Some advances such as multi-frequency approach have helped [6], but can only go so far [7-8]. Extensive investigations have been carried out to search for a simple and accurate C-V measurement technique [6-19]. The Time Domain Reflectometry (TDR) method introduced in this work is a novel solution to

this problem. With the TDR method, C-V measurement can be done with high precision and can be automated.

With power consumption problem associated with high gate leakage current, it is natural that the industry seeks alternative to the traditional SiO_2 gate dielectric [20-22]. Recently, high dielectric constant (high-κ) materials start as the gate oxide in 45 nm technology. However, the whole reason why silicon dominates the IC industry for many decades is that the thermally grown SiO_2 is an exceptionally good material for gate insulator. To replace SiO_2, the high-κ materials must also be low in defect density and can withstand high electrical stress. It represents a big challenge [12-14]. It is found that a high density of defects in these materials lead to undesired transport though the dielectrics and trapping-induced instabilities [23-25]. To point a direction for further manufacturing process improvement, it has become extremely important to characterize the defects and to study how new defects are created in them under electrical stress.

To that end, very few measurement techniques are at the disposal of scientists and engineers. Frequency dependent charge-pumping (FDCP) has emerged as the best candidate for the task [26-27]. However, even though the technique has been around for more than twenty years, the basic question of how deep does it probe has not been resolved yet. There are two camps of thought in the literature [28-29]. The differences in interpretation have led to an even more serious debate over whether new defects

can be created in the high-k materials under electrical stress. The implication in the reliability of high-κ material is obvious and the resolution of this debate becomes urgent. In this work, we resolve this debate using a combination of experimental and theoretical approaches. Our result provides the first clarification of this important question.

The reliability of high-κ gate dielectric is one of the many reliability issues that become more acute with scaled down MOSFETs. Another standout reliability issues in nanoscale MOSFET is the Negative Bias Temperature Instability (NBTI) of p-channel devices [30-33]. It has become much more worrisome because it gets worse rapidly when the gate oxide becomes thinner and nitrided oxide is used in the industry to combat boron penetration [34-36]. Pure NBTI for pMOS has already become the most serious reliability issue in current leading edge technology. The recent discovery that the NBTI degradation with a drain bias is even worse naturally pushes NBTI as the key reliability constraints for future device scaling [37-41]. The fact that the impact of drain bias increases with reduced channel length is even scarier. While it seems to be another issue associated with the ultrathin gate dielectric, the effect of the drain bias is not easy to explain.

Most of the studies of the drain bias effect on NBTI have been done at high enough drain voltage so that serious channel-hot-carrier (CHC) effect occurs [37-41]. Thus most explanations invoke either a NBTI enhanced CHC degradation or CHC

enhanced NBTI degradation. If that were the case, then the effect might not be so serious because the CHC effect for most advanced MOSFET is very minor at the operation voltage. However, another possible cause of this problem prompted us to look at this phenomenon at conditions that CHC is almost absent. We find that the drain bias enhanced NBTI is definitely a serious problem for advanced MOSFET in the nanoscale. Our results indicate that the cause comes from another nanoscale specific phenomenon, namely ballistic phonon effect [42-44]. While NBTI is exacerbated by the nanoscale gate dielectric, the drain bias effect, at least in our case, is also worsened by the nanoscale transistor channel length. This is a truly troublesome finding because it means a totally new mode of reliability degradation will exist. All temperature sensitive reliability issues of the transistor will be impacted and the impact will become more severe as the transistor shrinks further, particularly for high frequency operation.

Actually, high frequency application is another important area for these nanometer MOS devices to make their mark. Taking advantage of their excellent RF performance, these advanced MOS devices have been widely used for many microwave applications in areas of mobile communication and wireless networking [45-46]. At the same time, corresponding high frequency device measurement techniques are urgently demanded for properly characterizing their performances. However, a reliable high frequency measurement also requires a set of specific skills and instruments, which are not widely used in traditional device characterization. Absence

of careful consideration can easily cause measurement error and misinterpretation of results. For example, without specific design, when one launches a high speed pulse on a capacitor, some part of incident pulse is reflected back to source due to the impedance mismatch between the probe and capacitor. As a result, the actual voltage across the capacitor is much smaller than designated. These impedance mismatch issues are not important for traditional DC or low frequency measurements. However, as the measurement frequency gets higher enough, signal wavelength is reduced to be comparable with physical size of measurement components. Under that circumstance, all electrical components (such as cable, adaptor and probe) should be treated as a transmission line instead of a single wire and impedance match issue becomes a significant problem.

To get around this difficulty, we come up some solutions. In our study, we homely build a 50Ω terminated high-speed probe. It matches the source output impedance as well as the cable characteristic impedance. It is allowed to launch a high speed signal (<20GHz) at DUT with reflection loss less than 1%. Based on that, we are able to implement the charge pumping over 1GHz with square wave for the first time. The new results clear the debate over probe depth in high K defect characterization. This specific probe is also used in the setup to study the dynamic drain bias frequency dependent NBTI effect, which helps to identify the impact of our new mode of reliability degradation.

Moreover, any design of reliable high frequency measurement setup requires a tool to evaluate the setup for further improvement. Time domain Reflectromety(TDR) is such a system that is well known for its capability to detect impedance at time domain and identify the location of impedance mismatch. Besides its help to high frequency measurement, it also offers solutions to some challenges in advanced MOS device characterization. With a simple capture of time domain impedance profile of MOS capacitor, C-V characteristics, series resistance and overlap capacitor can be extracted with great accuracy. This new technique is free of difficulties that met by traditional C-V methods and can be applicable to routine industrial test procedure.

Chapter 2

Background and Literature Review

In chapter 1, we provide a brief introduction as well as the main frame of this thesis. In this chapter, we will give some background about this thesis topic. We will address recent developments of advanced thin oxide MOS device as well as emerging challenges in device characterization. Specifically, we will first show why the aggressive scaling down of oxide thickness is evitable in section 2.1. We will further address what the problem for thin oxide device is and how they affect the device characterization such as C-V measurement. C-V measurement is the first of three parts in this thesis work and will be discussed in detail in chapter 4 and 5. In order to prepare the readers with enough background, section 2.2 describes the principle of conventional C-V method and its difficulty in the application of advanced thin oxide MOS device.

Since the high κ dielectrics has been released to be used as gate oxide in the state of art 45 nm node technology, its reliability becomes a big concern. Section 2.3 will address recent advance and the unsolved question- where is the trap increase after the electrical stress. The answer must turn to the charge pumping measurement. Therefore, its principle and application to detect traps in the oxide is reviewed in section 2.4. Even though the measurement technique sounds simple, the interpretation is controversial. This is exactly the role of the second part of this thesis work –

answering the question: how deep does charge pumping probes.

Traps in the high κ dielectrics are one of the reliability concerns in the many. The current most serious one is negative bias temperature stability (NBTI). Section 2.5 will provides the basic background of NBTI. This knowledge is important for understanding third part of this thesis work – a new mode of NBTI.

The purpose of this chapter is to lay down enough knowledge or background for the reader to understand this thesis work. Those who are very familiar with related materials can skip this chapter and go directly to chapter 3.

2.1. Thin Oxide MOS Device

The continuous growth in integrated circuit(IC) density and speed is the heart of the rapid growth of electronics. The electronics industry is now the largest industry in terms of output as well as employment in many nations. It will be no doubt that this big industry will continue playing more important role in economic, social and even political development throughout the world. This importance is the motivation as well as a formidable driving force that urges the continued rise in IC integration density and speed.

The rise in circuit density and speed has been accompanied by the scaling of

MOSFET's to lower the cost per function and meanwhile increase the performance and functionality of the circuits. Metal-oxide-semiconductor field-effect transistor (MOSFET) is the most important and fundamental building block of very-large-scale-integrated (VLSI) circuits today in IC industry. Ideally, a MOSFET has high drive current (when the gate electrode is biased to turn the transistor on) and low leakage current (when the gate electrode is biased to turn the transistor off).

Reduced size for density requires a short channel length and small channel width. At the same time, high circuit speed is achieved by the reduction of the gate insulator oxide thickness which leads to an enhancement of the MOSFET's current drive capability as well as a better control of the short channel effects.

The transistor driving or saturation current I_{dsat} is an important parameter because it determines the time needed to charge and discharge the on chip capacitive loads. Thus, it impacts the product speed more than any other transistor parameter. This can be fulfilled by a short channel and high gate oxide field because the inversion layer charge density is proportional to the oxide field [4]. Practically, high drain current has been achieved by reducing the thickness of the gate dielectrics because I_{dsat} of a thin-oxide MOSFET can benefit more from channel length scaling, as shown in Figure 2.1 which is taken from Chenming Hu's paper [5]. As illustrated, scaling the channel length alone yields little increase in current from the 155 $\mathring{A}$ oxide thickness, far less than the textbook *1/L* dependence would predict. After all, even as *L*

approaches zero, I_{dsat} approaches a constant. In contrast, reducing the oxide thickness yields a considerable increase in I_{dsat} by increasing the oxide field and inversion charge density.

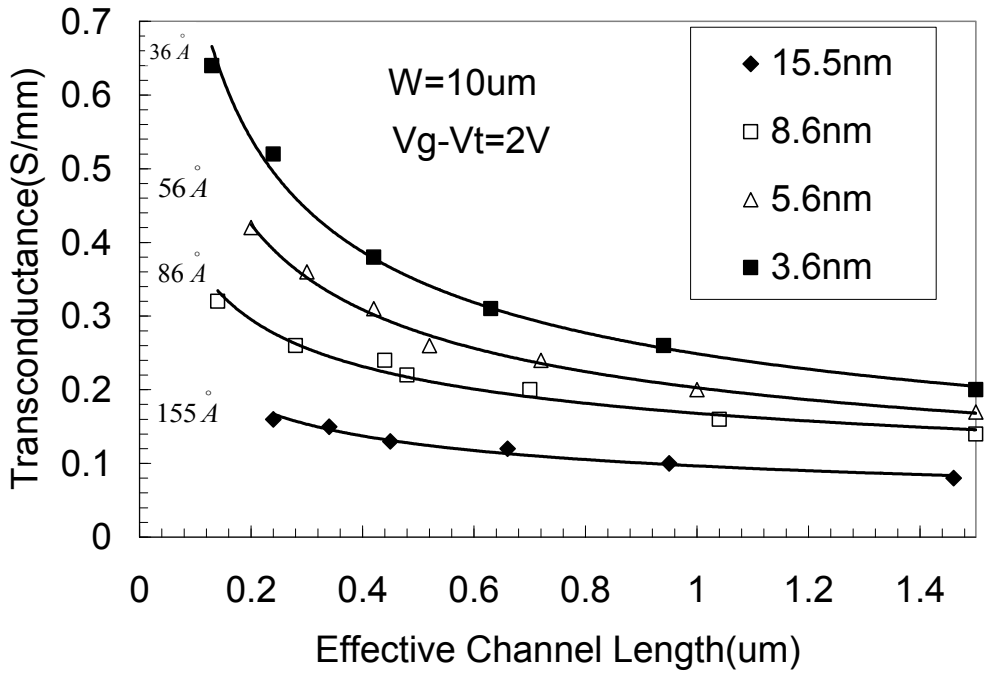


Figure 2.1 Experimental data reported by Chenming Hu [5] for MOSFET drain current versus effective channel length and gate oxide thickness. High oxide field not only increase MOSFET current, but also increase the benefit of shrinking the channel length. It is reported that the data is in good agreement with the theoretical prediction [4], although different from the textbook MOSFET model.

Consequently, motivated by the rise in circuit density and speed, advanced MOS device with thin gate oxide and short channel is highly demanded for pursuing device miniature and high driving current. Over the past three decades, transistor channel lengths and oxide thickness is reduced with each new generation of manufacturing technology. This trend is reflected in the International Technology Roadmap for Semiconductors (ITRS) [1]. Today the 45 nm channel length transistors are in mass

production with physical gate length down to 18 nm.

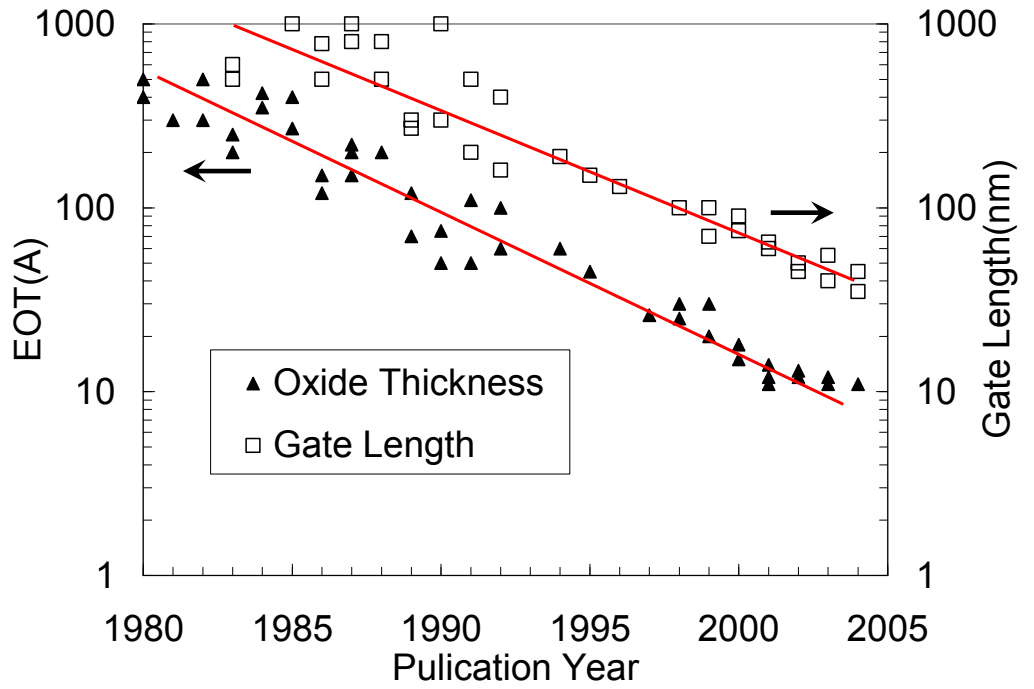


Figure 2.2 Scaling trend of gate oxide thickness over the past five technology generation from the 0.18um to 65nm node. Experimental data (L_{Gate}) and equivalent oxide thickness (EOT) are collected from device papers presented at conference. When only inversion oxide thickness (T_{Inv}) is provided in the literature, 8Å is subtracted from T_{Inv} to account for the additional capacitance component resulting from the poly depletion effect at the gate side and the quantum mechanical effect at the channel side. This figure is taken from reference [2].

Moreover, huge progress in gate oxide fabrication process has made it possible to realize ultra thin films with an acceptable thickness roughness and uniformity as well as a low defect density. As thin as 1.2nm EOT gate dielectrics are used in today's state-of-the-art complementary MOS (CMOS) technology, which corresponding to three to four layers of silicon atoms. Figure 2.2 illustrates the evolution of the technology nodes and gate oxide thickness as published over past two decades. It should be noticed that each update enhances the technology node every year.

2.2. C-V Measurement and its Difficulty

With dielectrics as thin as few atomic layers, this level of nanotechnology brings the challenges. When the oxide is thin, substantial direct tunneling current flows from the gate to the channel even under low voltage at operating conditions. More badly, this leakage current is found to increase exponentially with the decreasing oxide thickness [47]. At 1.2nm EOT MOS device, the gate leakage current density exceeds 800 A/cm^2 [3]. With further scaling down of oxide thickness, a tunneling current density over few kA/cm^2 is very possible.

Such a high leakage current has a significant impact on device characterization techniques. Even the basic measurement such as Capacitor-Voltage(C-V) is affected and loses its accuracy. Many efforts have been made to seek a reliable technique to obtain accurate C-V characteristics under this high leakage current situation [6-19]. Unfortunately, all methods reported so far have certain constrains and limitations. No satisfactory solution has been found yet. That's our motivation here to introduce this new simple accurate C-V measurement technique using the TDR method. It offers a possible solution to current difficulties suffering in industry now.

2.2.1. Standard C-V Characterization of MOS Capacitor

In MOS device characterization, MOS capacitor is a good test structure for its

simplicity. C-V characteristics refer to the capacitance(*C*) of MOS capacitor as the function of applied gate voltage (*V*). In addition to the capacitance values, a great deal of information about the MOS capacitor and oxide/semiconductor interface can be obtained. The C-V curve can be manipulated to extract the equivalent oxide thickness (EOT) [48-50] and effective mobility [51-52] from measured capacitance in the strong accumulation and inversion regime respectively. Besides that, interface traps [53-54]; substrate doping profiles [55-56] can be obtained as well. Therefore, accuracy is greatly demanded in C-V measurement because it is of significant importance for parameter extraction.

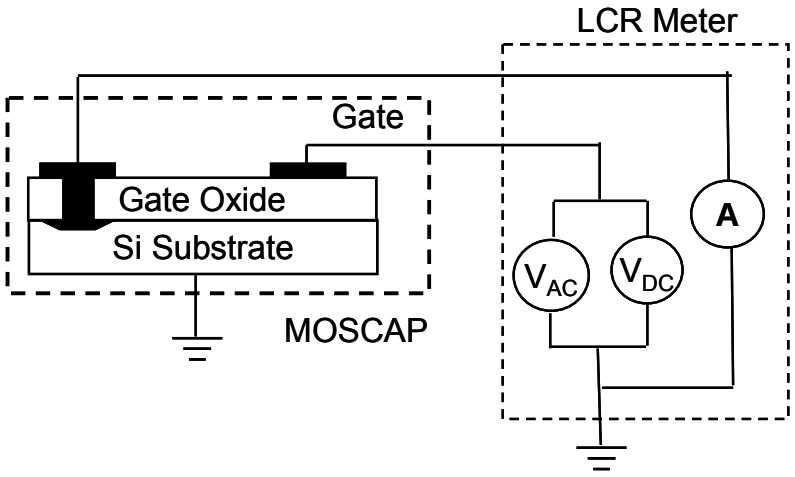


Figure 2.3 Illustration circuit connection of conventional standard C-V measurement set up using LCR meter. AC signal superimposed on a DC voltage is employed on gate and resulting AC current is measured at substrate electrode. This figure is taken from reference [53].

However, the available measurement techniques are not such wonderful for this important device characteristic. As shown in Figure 2.3, the existing method is typically fulfilled by applying a small AC signal (~10 to 50 mV) on top of the DC

bias across the structure and sensing the capacitive displacement current at the same frequency (75Hz to 1MHz). In order to do that, it is necessary to separate the AC current component (capacitive displace current) from DC component (leakage current). This is commonly performed using phase sensitive a LCR meter (such as HP 4285) or a lock in amplifier. Commonly, the phase and magnitude of the impedance is measured.

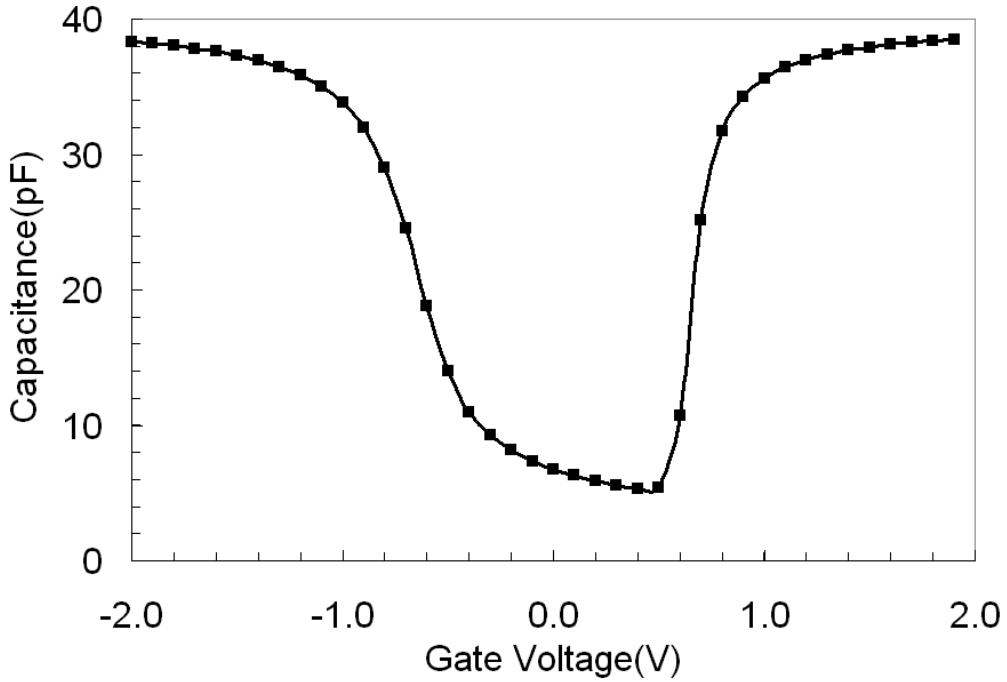


Figure 2.4 Simulation of ideal low frequency C-V using the CVC program [58]. At two ends of this voltage range, the shape of C-V is more flat and capacitance is high because either accumulation or inversion layer is formed. In the range of between, device is under depletion and capacitance is low.

Figure 2.4 shows ideally what a C-V curve would look like. This is a theoretical simulation of low frequency C-V using the CVC program developed by Professor John R. Hauser from North Carolina State University [58]. At very negative (below flat band voltage) gate bias, an accumulation layer of holes is formed at

oxide-semiconductor interface. While at very positive (above threshold voltage) gate bias, the electrons, which can respond to the slow change in low frequency capacitor voltage, builds an inversion layer at interface. At these two operation modes, the MOS capacitance is just oxide capacitance and exhibits itself as a flat line in C-V curve (Figure 2.4). When a gate voltage between the above two modes is applied, a space charge region is induced in the semiconductor. At this mode, the capacitance across depletion layer plays more important role and results in much smaller capacitance than oxide capacitance. Therefore, a decrease of capacitance is shown in C-V curve.

Figure 2.5 is the shape of expected C-V behavior. When the oxide gets thin, the existing measurement technique has problem in conveying results as expecting. Instead, it shows analogous behavior as shown in Figure 2.5. This is the C-V curve measured at EOT 1.3nm 1400um^2 n-channel HfO_2 MOSCAP with lock in amplifier. Same oxide thickness of SiO_2 is much worse and has difficult in accurate measurement. This part of data is extracted at 1KHz which is considered as high frequency. The low frequency C-V behavior obtained here is due to the transistor-like designed test structure to provide inversion charge. The detail description of the test structure will be done at chapter 4.

The most of important sign in this data is the deviation of inversion and accumulation (two ends) part of C-V curve from the expectation. Instead of pretty constant with voltage, the capacitance significantly decreases. Similar phenomena are observed by many other research groups [6, 9-13]. For example, Yang *et al.* found that the

capacitance of 1.7 nm MOS capacitor depends on the measurement frequency and decreases with increasing gate bias [6]. Ahmed *et al.* noticed a faster roll-off of capacitance at strong inversion with the increasing channel length for sub 2nm oxide devices [10].

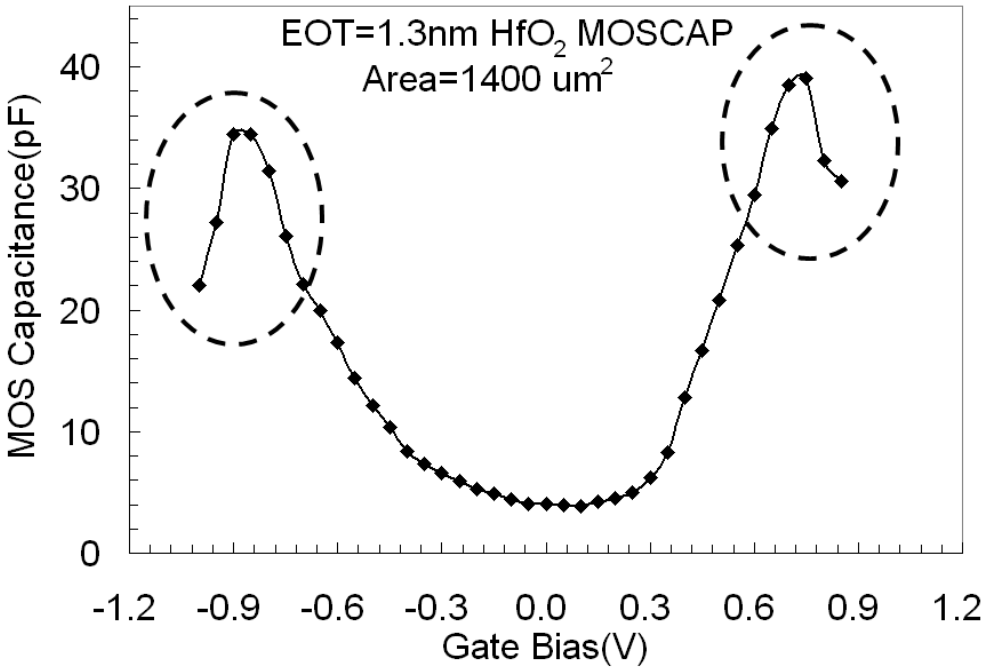


Figure 2.5 The measured C-V curve using conventional measurement techniques with lock in amplifier to detect phase and magnitude for EOT 1.3nm 1400um^2 n-channel HfO_2 MOSCAP. Same oxide thickness of SiO_2 is much worse and has difficult in accurate measurement. Note that the C-V is anomalous at inversion and accumulation with sharply decreases of capacitance.

2.2.2. Limitation of Conventional C-V Measurement

This anomalous behavior deprives the role of C-V as a basic characteristic for parameter extraction because the extraction like EOT and effective mobility is highly dependent on the accuracy of gate capacitance at accumulation and inversion region.

It is believed to come from the shortness of measurement technique, specifically, from the limitation to handle high leakage current and series resistance.

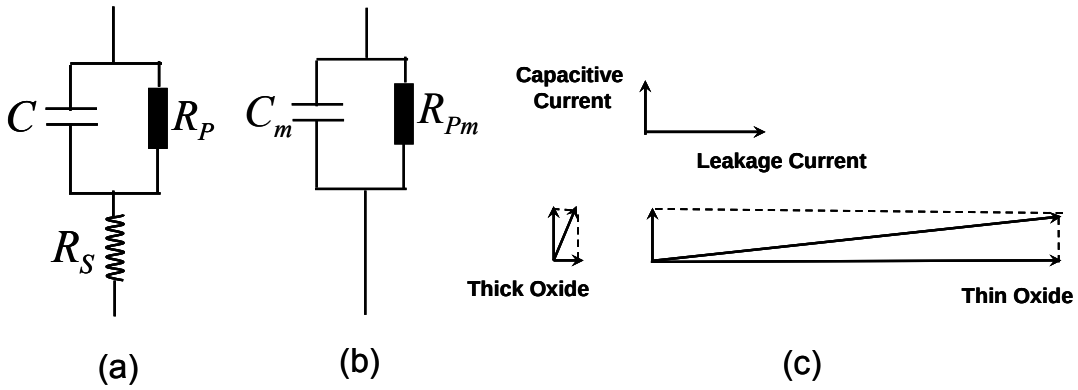


Figure 2.6 (a) three element equivalent circuits for leaky MOS capacitor, C is the measured capacitor, R_P represents the leakage current component, R_S is the series resistance. (b) Parallel circuit model used in conventional C-V measurement. It shows that the actual measurement yields a C_m (measured capacitance) in parallel with R_{Pm} (measured parallel resistance). (c) The current component in traditional C-V measurement, leakage current and capacitive current are 90 degrees out of phase. For the same capacitance with same amount of capacitive displacement current (the vertical component in the figure), the leakage current in thinner oxide is much larger. It may overwhelm the displacement current and the measurement technique will lose its accuracy.

To describe the exact reason, a simple three-element equivalent circuit, as shown in Figure 2.6(a), would be required to conceptually model the thin oxide MOS device [6]. In addition to measured capacitance C and series resistance R_S arising from the finite resistance of the source/drain contacts, inversion channel and gate material, a shunt resistance $R_P = \partial V_g / \partial I_g$ due to the leakage current is also included [59]. As the gate oxide gets thin, leakage current increases exponentially. As a result, shunt resistance R_P reduces sharply and becomes comparable in magnitude to the series resistance R_S. So both shunt and series resistances are required for thin oxide MOS device.

In the conventional C-V measurement for these leaky thin oxide capacitors, it typically uses parallel circuit model in Figure 2.6(c). Compared to more accurate three elements equivalent circuit, we have seen that we can no longer directly relate the measured capacitance value C_m with that of the dielectric film C because of the significant contribution of the series resistance R_S [7]. At the leaky MOS device, series resistance R_S and shunt resistance R_P are comparable so that the capacitance test voltage is divided between the MOS capacitor C and the series resistance R_S. Less magnitude of voltage across the MOS capacitor results in smaller measured capacitance. This capacitance degradation becomes more severe at high gate bias region where the leakage current is high. Finally, it causes huge gate capacitance attenuation observed in Figure 2.5.

2.2.3. Improvement and Limitation of C-V Technique for Thin Oxide

Being aware of the error in the conventional C-V measurement, many research groups have made efforts to find a correction procedure [6-19]. It can be seen from the three element equivalent circuit that, in order to obtain capacitance with high precision, the other two components R_S, R_P should also be determined accurately. However, from a single measurement of impedance phase and magnitude, only two of these three parameters can be ascertained. Therefore, one more relation is required to extract all these three unknown components. This problem can be solved by measuring capacitor at two different frequencies as proposed by Yang and Hu [6]. With an additional

measurement at a different frequency, all three parameters can be known.

Although, theoretically, this two-frequency model is correct, there are numerous limitations in practice. As illustrated in Figure 2.6(c), for the same capacitance with same amount of displacement current (the vertical component in the figure), the leakage current in thinner oxide is much larger. It may overwhelm the displacement current and instrument precision is mostly spent on accurate determination of the leakage current components.

As a result, a small error of the phase angle introduces a large error in the capacitance measurement. Nara *et al*. pointed out that the inevitable error in the measurement limits the usefulness of this method in ultra thin gate oxides [8]. Ghibaudo *et al*. showed that the measurement error caused by series resistance effect is amplified by the shunt resistance R_P and gets worse for thinner oxide [7]. Zhu *et al*. found that the dominant source of measurement error comes from R_S / R_P term, which becomes severe when oxide gets thinner [13].

To that end, many improvements have been made on the two frequency method to preserve its effectiveness [16-18]. For example, Luo et al. developed a four-element equivalent circuit which accounts for the parasitic capacitance [17]. Moreover, the finite channel resistance prompted some groups to use a distributed network to replace the three element model [10-13]. All these approaches share the same basic scheme –

the measurement of impedance is affected by both shunt and series resistance. Even more elaborated refinements and therefore complexities are introduced to cope with the increasing error associated with ultra thin gate oxide. This difficulty has led Teramoto et al. to try a new approach that relies on resonance [19]. As a result, all these difficulties have brought about a need for new measurement procedures which can account for both series and shunt resistance.

In chapter 3, we introduce a new simple accurate C-V measurement technique based on Time domain Reflectrometry (TDR). It completely solves the above problem and is very accurate even under leakage current as high as 5000A/cm^2. Besides that, it can also accurately offers series resistance and overlap capacitance simultaneously which will be demonstrated in detail in Chapter 4.

2.3. High κ Gate Dielectrics and its Reliability Issue

The difficulty in C-V measurement is not the only problem caused by such high leakage current. Power dissipation is another serious concern, even though leakage current is still negligible compared to transistor driving current. Typically, a standby power (when the transistor is off) within 100mW can be tolerated by today's high performance CMOS logic chips. With the typical 0.1 cm^2 active gate area per chip, the maximum tolerable gate leakage current would be of the order of 1 A/cm^2 for an operation voltage of 1 V.

Unfortunately, as the gate oxide is scaled down to 2nm, it has run into this off-power limitation [3]. To push oxide thickness further down, industry has introduced nitrogen in the growth of SiO_2 in 90/65 nm generation. In this way, a heavily nitrided SiON (oxynitride) films are formed to block boron diffusion reducing gate leakage. Besides the advance in the fabrication process, improvement is also made in circuit design area. More power tolerant chip-architecture IC designs as well as sophisticated power management schemes are developed. In recent logic circuit applications, these leaky advanced MOS devices are used in core of the chip for fast operation. The rest of functionality such as *I/O* voltage applications is implemented by thick oxide devices. In this way, the leaky gate active area can be reduced.

Although all these developments can help to somewhat extend the life of SiO_2 and push oxide thinner as to today's 1.2nm, it is clear that SiO_2 has approached its physical and electrical limits and leakage issue is the obstacle for further down scaling. To circumvent this impediment, it is natural that the industry seeks alternative gate dielectric as a replacement of SiO_2. Fortunately, they found one. In the recent released 45 nm technology by Intel, Hf- based materials is used as gate oxide with dielectrics 4~5 times larger than pure SiO_2. With these high-κ dielectrics, leakage current is 4~5 times lower than 65 nm technology because the physical thickness of the gate stack can be increased while maintaining the same capacitance.

Meanwhile, these high κ materials do not have such good interface quality as thermal

grown SiO_2. It is not surprising because SiO_2 has been used for more than 30 years. It offers so fantastic and tremendous important material and electrical properties, including a stable thermodynamic Si/SiO2 interface with low defect charge densities ($10^{10} cm^{-2} eV^{-1}$) as well as superior electrical insulation and interfacial bonding properties. Like other transition metal oxides, the high κ material exhibits a high density of intrinsic electron traps in the material. Moreover, the SiO_2 buffer layer, usually formed between high κ film and silicon substrate to improve interface quality, is not as good as thermal grown SiO_2. It also indicates large number of defects states.

Therefore, to correctly characterize these defects behavior under stress, it is of great importance to separate contributions from the traps in high κ film and in the interfacial SiO_2 layer. The results will help understanding the origin of these traps and point to the direction of further manufacturing process improvement. Among the very few measurement techniques available to fulfill this goal, the frequency dependent charge pumping (FDCP) is the best.

2.4. Frequency Dependent Charge Pumping(FDCP)

Among all the defect characterization methods, charge pumping (CP) technique [60-61] is proved to be the most successful and reliable one, because it is very simple to set up and shows, by far, the highest sensitivity ($\sim 1 \times 10^{9} cm^{-2} eV^{-1}$).

The basic experimental set-up to perform charge pumping measurements, as introduced by Brugler and Jespers [60], is illustrated in Figure 2.7 for the case of an n-channel MOS transistor. The gate of the MOSFET is connected to a pulse generator and a reverse bias is applied to the source and the drain junctions, while the substrate current is measured.

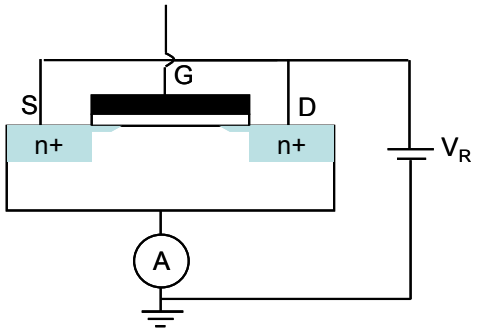


Figure 2.7 Basic experimental setup for charge pumping. Picture is taken from reference 60.

As illustrated in Figure 2.8, the basic idea of the charge pump technique is to rapidly switch a MOSFET from accumulation towards inversion and back forth. When the transistor is pulsed into inversion, electrons originating from the drain and source region get trapped in interface states (region C in figure). As the gate pulse is switched rapidly and drives the surface into accumulation, the mobile charge drifts back to the source and drain under the influence of the reverse bias, but the trapped electrons do not have sufficient time to get detrapped from the interface states (region D in figure). The trapped electrons will recombine with holes originating from the substrate (region E in figure). A similar process holds for the switching from accumulation (from region A to C). In this way, a net amount of charge is transferred ("pumped") from the bulk to the drain and source regions. By repeatedly switching

the gate voltage, a *RMS* average current can be measured at the substrate contact, which is directly proportional to the interface trap density, the transistor gate area, and the frequency of the gate pulses.

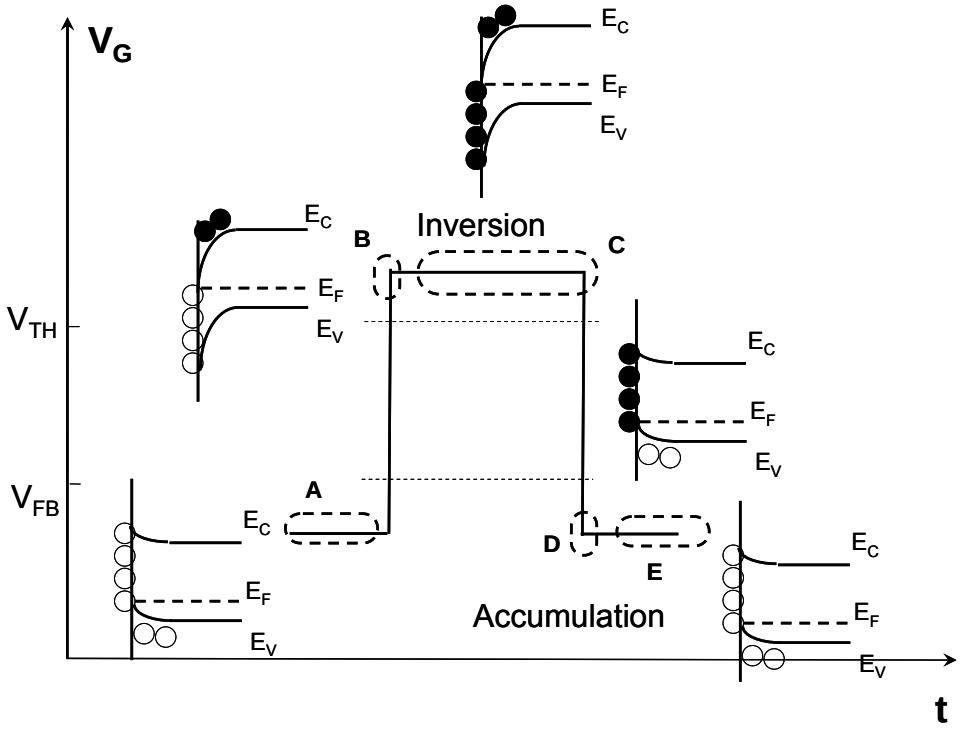


Figure 2.8 Illustration of physical mechanism of charge pumping. It relies on the application of a square waveform to the gate of the device which drives the device from accumulation into inversion and back to accumulation again. At region B, the device is just driven into inversion and interface states are still empty. At region C, after very short time of inversion, all the interface states are filled by the electrons from source/drain. At region D, accumulation just started and electrons are still trapped in interface states and have not enough time to escape it out. At region E, after certain time of accumulation, the interface states are emptied by the holes from substrate.. Solid circles represent electrons while empty ones are holes Solid circles in interface states means that they are filled with electrons while empty one means electrons are recombined by the holes. This picture is taken from reference 63.

This conventional charge pumping is mostly used to measure the interface state density. However, for the high κ dielectrics, besides interface states, large numbers of defects are also generated in the bulk of dielectrics. In addition, a lot of traps are also

presented at high κ/SiO_2 interface and bulk buffer SiO_2 because the quality of buffer SiO_2 layer is not so good. These oxide traps locate some distance away from the interface states and communicate with the semiconductor by electron tunneling from inversion layer or filled interface states (Figure 2.8). These slow oxide traps play an important role for the high κ dielectrics reliability such as threshold voltage shifts, mobility, transconductance degradation, bias temperature instability and dielectric breakdown [24]. Thus it becomes a big concern as well as the subject of an increasing interest

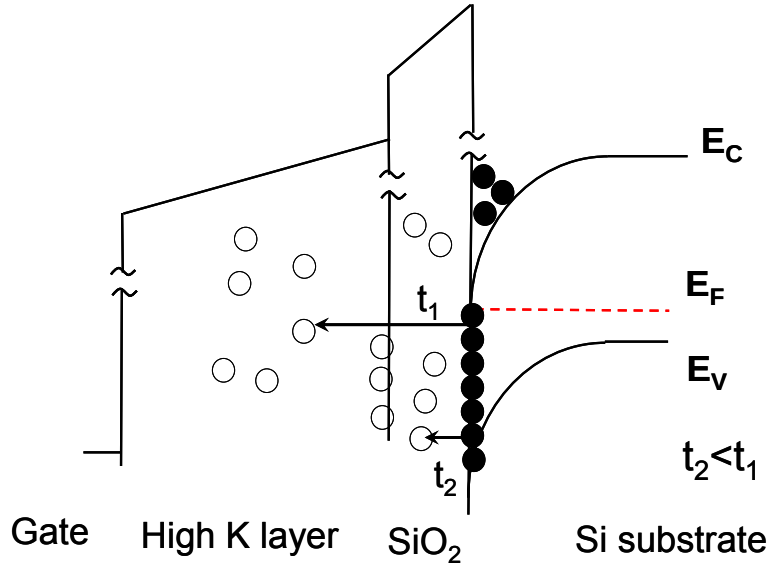


Figure 2.9 Illustration of filling defects in high K/SiO_2/Si MOS device. A large number of traps are generated in bulk of high K and buffer SiO_2 as well high K/SiO_2 interface. These traps away from interface can be filled by electron tunneling from filled interface states or inversion layer. It takes electrons longer time to tunnel and fill those traps that are far away. Solid circle represents electrons and empty ones are unfilled defects.

To that end, a new charge pumping technique, based on the evolution of the CP current as a function of frequency, has been developed to address depth profile of the electron traps. It is based on the mechanism that different locations of traps need

different time to fill. The traps locating further away from interface requires more time for electrons/holes to tunnel into and fill them (Figure 2.9). By varying the time allowed for electron tunneling the depth profile of traps can be obtained by detecting the trap density at a given time period. Experimentally, it is implemented by changing the gate square pulse frequency in charge pumping measurement.

When high frequency gate pulse is applied, only the interface traps are able to participate in the charge pumping process since there is insufficient time for electron to tunnel into the oxide traps further away. As the frequency is lowered, the applied gate pulse drives the device into inversion for longer time. Those traps within a tunneling distance appropriate for that particular frequency can also be filled and emptied by the tunneling in and out of electrons from interface states, leading to an increasing charge pumping current per cycle.

This frequency dependent charge-pumping (FDCP) method allows the assessment of spatial distributions of oxide traps. However, although this idea was reported by Declerck *et.al* more than 30 years ago [28], the theory behind it has only been improved recently. The probed depth in the dielectric as a function of CP frequency has become a controversial issue on the recent adaptation to study the defects in high κ gate dielectric stack. For the same experimental data, two different groups come up with two totally different conclusions [28, 29]. Who is right? How exactly deep does charge pumping probe? In chapter 5, we will use both experimental evidence and

theoretical model to answer this important question.

2.5. Negative Bias Temperature Instability(NBTI)

Besides the high κ reliability, Negative Bias Temperature Instability (NBTI) is another serious concern and becomes the number one reliability problem nowadays. NBTI is a degradation phenomenon in p-channel MOSFET, known since the late of 1960s on SiO_2 dielectrics. It has been observed that the application of negative gate bias on *p*-channel MOSFET causes instability of device behavior with time, such as an increase in the magnitude of threshold voltage (V_T) and a reduction in device driving current (I_{DSAT}) [30-32].

The involved physical mechanism is commonly admitted that under a constant gate voltage and an elevated temperature a build-up of positive charges occurs either at the interface Si/SiO2 or in the oxide layer leading to the reduction of MOSFET performances [34-35]. The kinetics of this effect is accelerated by temperature and the oxide electric field.

As shown in Figure 2.10, the NBTI induced V_T shift increases with stress time and could reach the level that hurts the logic function in the long term. Typically, for the modern devices operating at 1.2 V or below, V_T shifts on the order of 20-50 mV is considered to be very serious. The problem gets worse since this degradation are not uniform for pMOS and nMOS.

The name of "negative bias" suggests that it occurs primarily in *p*-channel MOSFETs with negative gate voltage bias and appears to be negligible for positive gate voltage and for either positive or negative gate voltages in *n*-channel MOSFETs. In MOS circuits, it occurs most commonly during the ''high'' state of *p*-channel MOSFETs inverter operation. As a result, greater V_T shift in one transistor of a matched CMOS pair can cause functional failure in analog circuits and logic circuits that are sensitive to parameter mismatch.

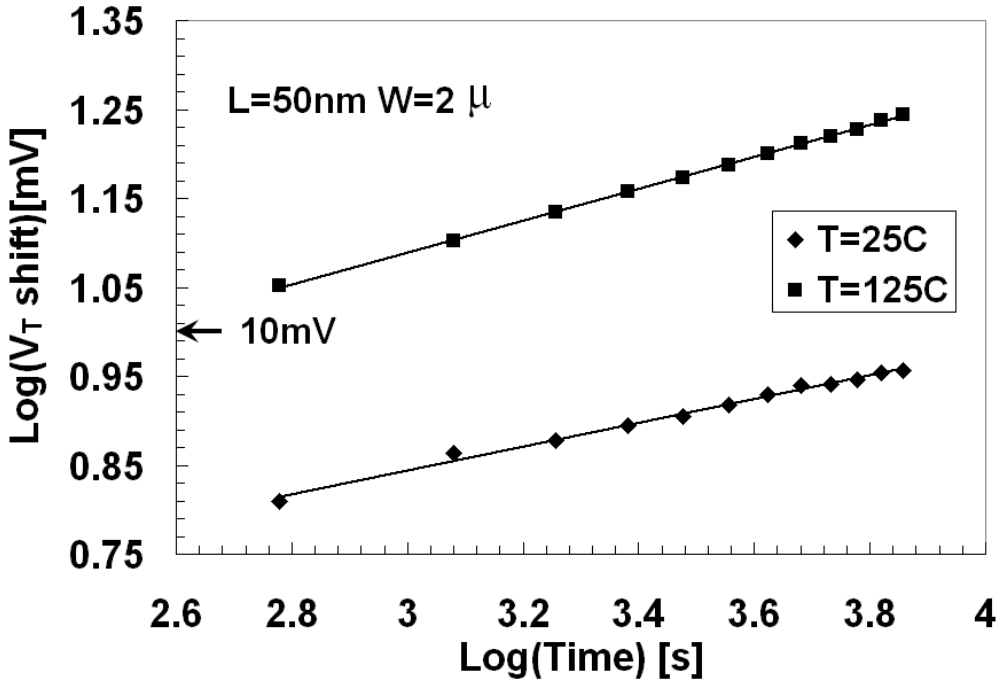


Figure 2.10 Measured NBTI induced V_T shift versus time plotted in standard log-log form. The device under test is an advanced MOSFET from 90 nm technology with 2 μm channel width and 50 nm physical gate lengths. Two different temperature conditions are used: room temperature (25 C) and elevated temperature (125 C). More VT shift is observed at higher temperature.

Moreover, the degradation is accelerated by temperature. As shown in Figure 2.10, V_T shift increases at higher temperature. Since most of device operates at elevated temperatures such as 125 C, this temperature dependent reliability issue becomes a

major reliability concern. With the recent relentless technology scaling down of oxide thickness, the concern becomes more serious because the degradation is also well known for its acceleration by increasing electrical field for thinner oxide [34].

In chapter 6, we will report a new mode of NBTI degradation – Ballistic Phonon Enhanced NBTI. It is a new finding and believed to be more serious for future transistors.

Chapter 3

C-V measurement I

Capacitance Extraction with Very High Leakage Current using Time Domain Method

In previous chapter, it explains the difficulty in the C-V measurement nowadays. In this chapter, we will solve the problem and present a new C-V technique. The reason to put this as the first of three parts in this thesis is C-V measurement is one of the first electrical tests for a new device. From the measurement, one can find out the equivalent oxide thickness, flat band voltage, substrate doping and so on. It is so basic and widely used that a reliable and accurate technique is demanded. The need becomes more urgent recently because the existing famous method runs into problem.

People have done extensive searching process to find a good candidate [6-19]. However, the result is not satisfactory because of the high standard of the new technique. This new technique must be simple but accurate. In addition, it can also be automated so that it can be integrated into the routine device measurement on the production line. Most important of all, it must have the strong ability to handle the accuracy problem from the leakage current for the advanced MOS device. Moreover, the leakage current will become higher for future device and this new technique should also be able to deal with that.

Fortunately, we find a good one and will be shown in this chapter in details. It is time domain reflectrometry(TDR) C-V method. It is based on a well known high frequency measurement technique called TDR with the commercial instrument available. It can provide accurate C-V result even under leakage current as high as 4000 A/cm^2, which is sufficient enough for at least next few generations of CMOS technology. This method offers a simple and high precision measurement technique and can be automated and implemented as a routine device characterization procedure.

3.1. Basic Principle

The basic principle of the new technique could be viewed as a step function with very fast rise time propagating down a transmission line. Upon encountering a capacitor, part of the power passes through but most of it is reflected. The shape of the reflected step function, intuitively, must contain information about the capacitor. Naturally, one cannot recover the full information of the capacitor without a proper reference. However, once a well designed reference is available, this method allows an accurate description of the capacitor. The idea of sending step voltage and measuring the reflected voltage is so called Time domain reflectrometry(TDR).

The step function generated by the TDR scope is very fast with 35ps rise time. It reached the device (capacitor) under test (DUT) through a bias-TEE (for DC bias) and a transmission line. Due to impedance mismatch between the transmission line and the capacitor, the step function is reflected back toward the scope which records it. By

studying the reflected voltage waveform, much information can be obtained on the nature of the load. Of course, a reference is necessary and chosen to be an open circuit (let the probe floating). Figure 3.1 shows the reflected waveforms from an open circuit reference, a MOS capacitor in depletion. For detail description and explanation of TDR theory, please refer to Appendix A.

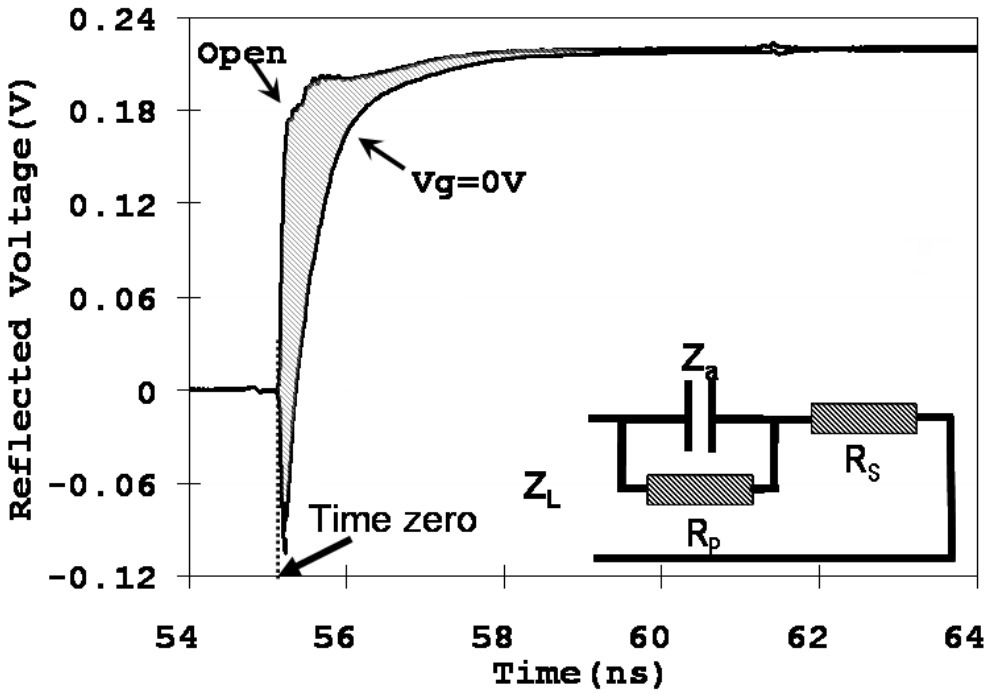


Figure 3.1 Reflected waveforms from open circuit (reference), MOS capacitor (HfO_2 gate dielectrics with EOT 1.2nm) at depletion (V_G = 0V). The shaded area represents the total stored charge in the depletion case. The insert is the equivalent circuit of the capacitor with thin oxide.

At any instant in time, the reflection coefficient is determined by the impedance mismatch:

$$\rho = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{3.1}$$

Where Z_0 is the impedance of the transmission line which is typically $50\,\Omega$ in high-speed measurements; Z_L is the impedance of the DUT. For an open circuit, Z_L

$=\infty$ and the reflection coefficient is 1. For the MOS capacitor, the equivalent circuit is in the insert of Figure 3.1. The impedance of a capacitor is a short circuit when the step function first arrives. As the capacitor charges up, it eventually becomes an open circuit. For the depletion case (V_g = 0V), leakage current is negligible and R_P can be removed from the equivalent circuit. At the end of charging, the signal reaches the same level as open circuit.

The charging behavior of the capacitor is a measure of the stored charge in the capacitor. Intuitively, the area enclosed (shaded area) by the open circuit waveform and the capacitor charging waveform represents the total stored charge at the end of the voltage step. Mathematically, it can be vigorously shown that

$$C = \frac{1}{2Z_0 V_{step}} \int_0^{\infty} \left(V_{Open}(t) - V_{DUT}(t) \right) dt \tag{3.2}$$

Where $V_{Open}(t)$ is the open circuit waveform; $V_{DUT}(t)$ is the waveform from the DUT (the capacitor) and V_{Step} is the height of the step function. The detail derivation of equation (3.2) is shown in Appendix B. The integral represents the enclosed area of the two waveforms which is in agreement with our intuition. With Equation (3.2), capacitance can be obtained by measuring the reflected waveform of the capacitor and the open circuit. Following this principle, we implement it experimentally and here are the basic procedures as well as setups.

3.2. Experimental Setup and Test Structure

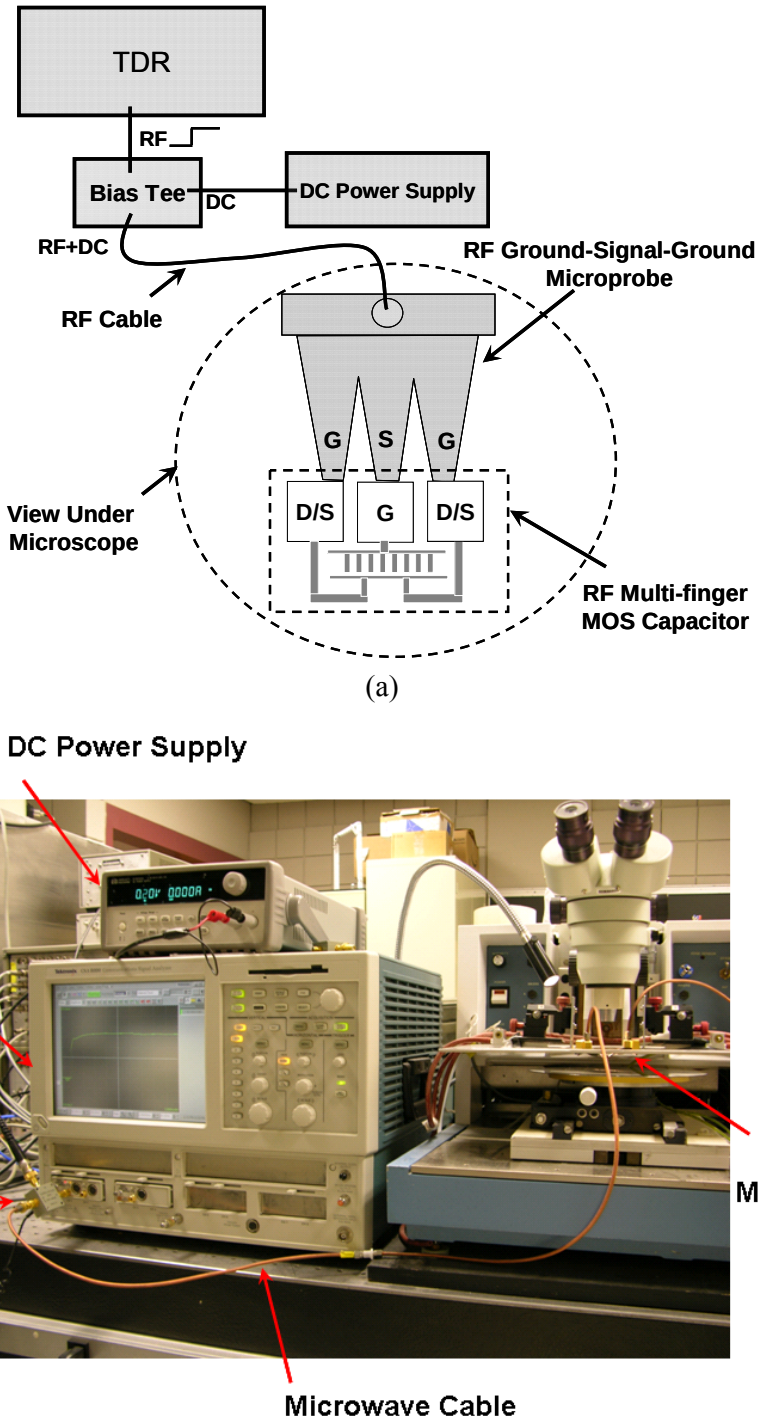


Figure 3.2(a) Basic illustration of building block of experimental setup in TDR based C-V measurement. (b) Picture of actual instruments and connections

Figure 3.2(a), (b) illustrates the block diagram and actual picture of experimental

setup for MOS capacitor C-V measurement using the TDR. A TDR scope which in our case is a Tektronix CSA8000 scope with 80E04 plug-in module is connected to the device-under-test (DUT) which is a RF-compatible MOS capacitor [14,16] through a Bias-TEE (Mini Circuit), a microwave cable and a microwave Ground-Signal-Ground probe (Cascade probe). A fast rise time (~35ps) step function with magnitude around 220mV is generated from the scope and reflection under different DC bias condition is recorded.

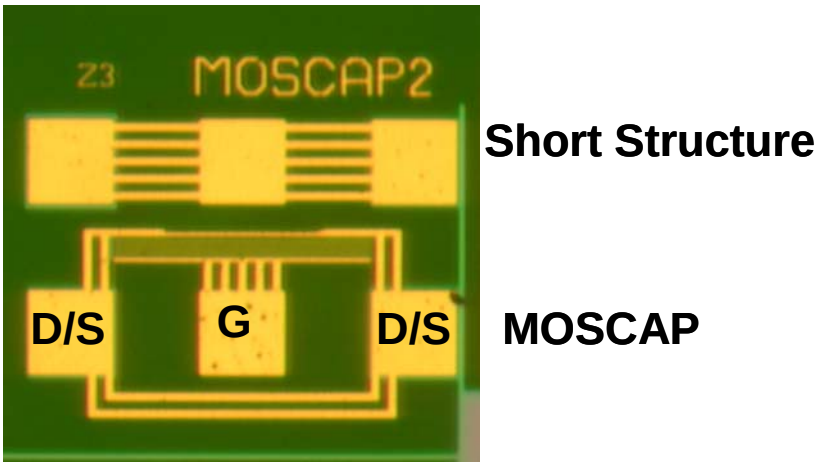


Figure 3.3 Picture of MOSCAP test structure, showing ground-signal-ground pads. It is taken under high magnification microscope

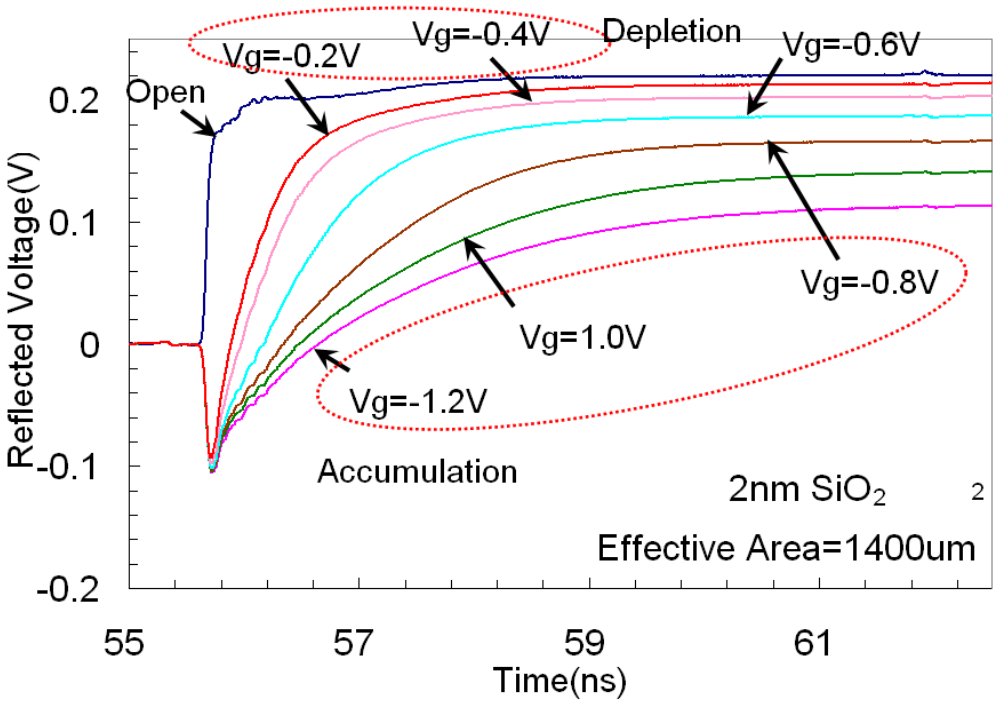
Our p-type substrate MOS capacitor test structure is supplied by the Sematech International Corp. Two wafers with same test structure but different type of gate dielectrics are used in this measurement. One is 2nm pure SiO_2 while the other is HfO_2 with 1.2 nm EOT HfO_2 (~1nm buffer SiO_2). This test structure is well designed. In this design, a large number of small capacitor elements are imbedded in a sea of highly doped material to lower the series resistance. The picture of a test structure taken under high magnification microscope is shown in Figure 3.3. The contact pad on the top surface at the wafer level is designed to permit one to use

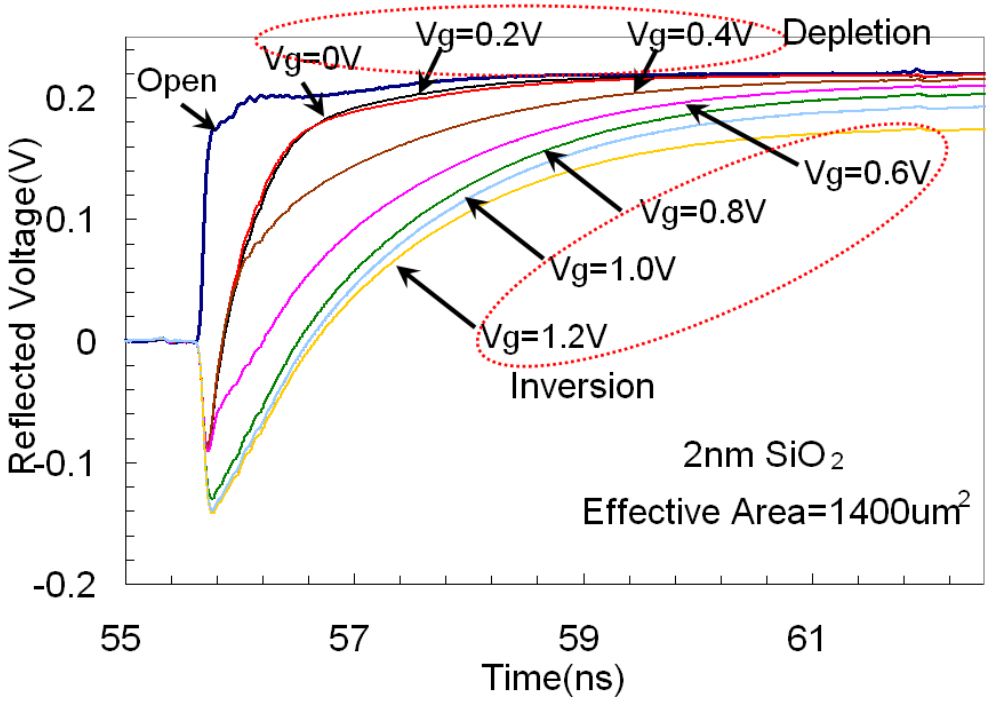
ground-signal-ground probe connection for high frequency measurement. Due to the physical limitation of device geometry associated with the process technology, the active area of the test structure is significantly distorted resulting in effective area around 1400 μm^2 [14]. Moreover, short circuit calibration structures are also fabricated on the same wafer.

3.3. Time Domain Response of Leaky MOS Capacitor

With the above setup and test structure, we obtained the reflected voltage of this MOS capacitor under different gate bias ranging from -1.2V to 1.2V as shown in Figure 3.4. Under this range of gate bias, this capacitor operates under accumulation, depletion and strong inversion respectively. As it can be seen in Figure 3.4, the basic shape drops down initially and rises up to final steady state. It is because the impedance of a capacitor initially behaves as a short circuit when the step function first arrives. And it eventually becomes an open circuit as the capacitor charges up.

Moreover, in some bias condition (such as V_G=1.2V and -1.2V), the steady state of capacitor does not reach the same level of the open circuit. In Figure 3.4, it can be also seen that the final state of capacitor at some gate bias when it finished the charging has the voltage level less than open circuit. That is because the test structure used in this experiment is thin oxide device with pretty large leakage current when applied voltage is high.


Vg=-0.2V
Vg=-0.4V
Depletion
Vg=-0.6V
Open
Vg=-0.8V
Vg=1.0V
Vg=-1.2V
Accumulation
2nm SiO2
Effective Area=1400um
Reflected Voltage(V)
0.2
0.1
0
-0.1
-0.2
55
57
59
61
Time(ns)


Vg=0V
Vg=0.2V
Vg=0.4V
Depletion
Open
Vg=0.6V
Vg=0.8V
Vg=1.0V
Vg=1.2V
Inversion
2nm SiO2
Effective Area=1400um2
Reflected Voltage(V)
0.2
0.1
0
-0.1
-0.2
55
57
59
61
Time(ns)

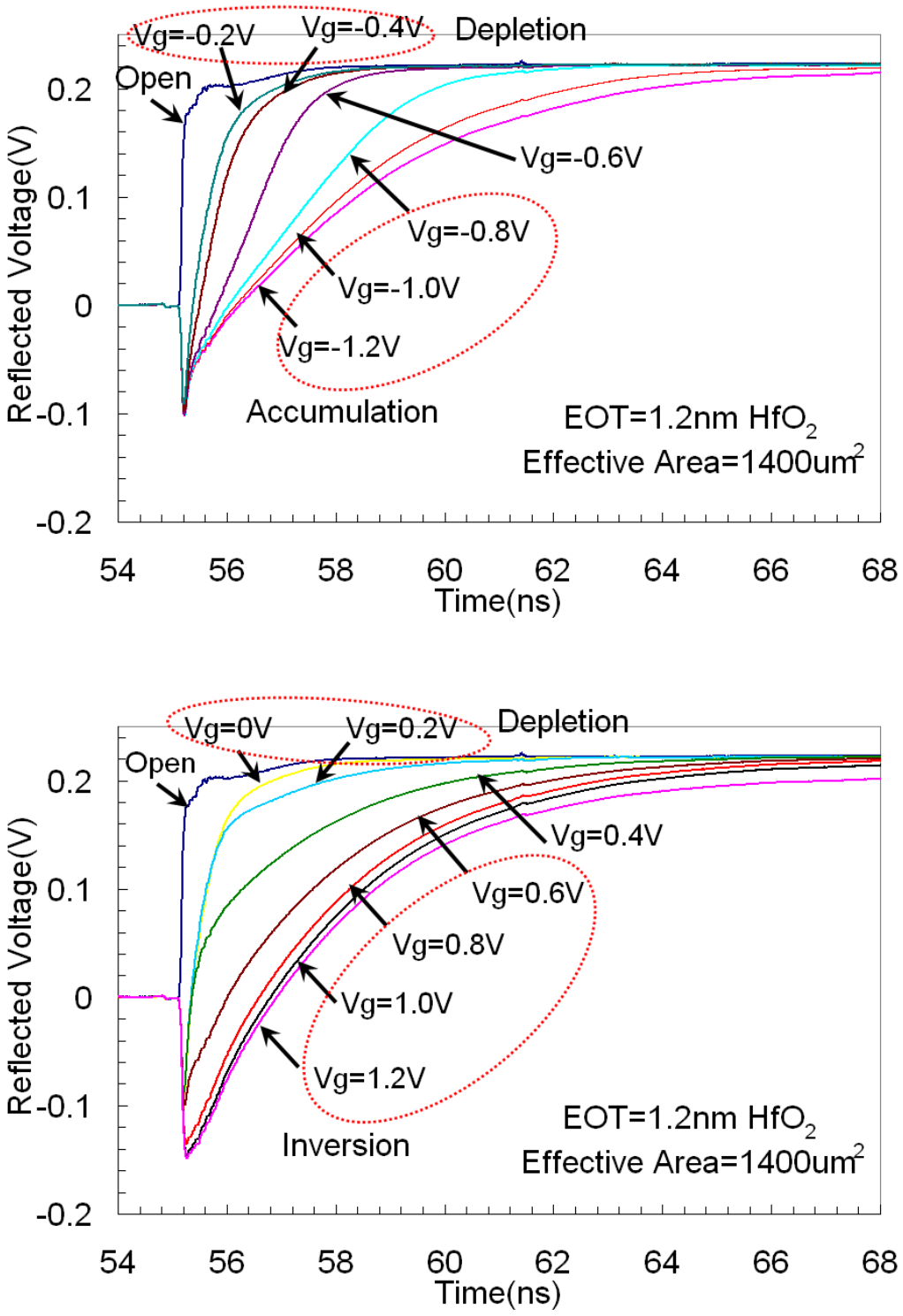


Figure 3.4 Time domain curve of MOS capacitor with 2nm SiO_2 and EOT 1.2nm HfO_2 gate dielectrics. Positive and negative gate voltage cases are plotted in a separated way.

As a result, some charges get lost through leakage current and lead to lower final

voltage even though the capacitor is fully charged up. The higher leakage current, the more charges are lost resulting in lower final voltage level. Since the leakage current in MOS capacitor is gate bias-dependent, the similar trend can also be observed in the final voltage of TDR reflected waveforms (Figure 3.4). From that, we can even extract the leakage current and shunt resistance. The results are shown in Figure 3.5.

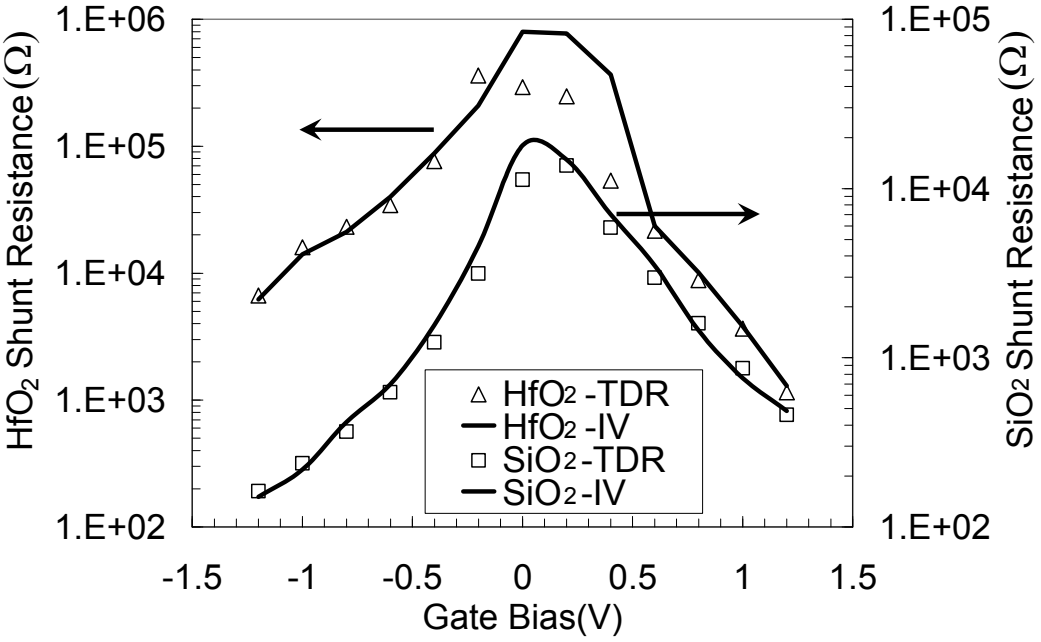


Figure 3.5 Comparison of shunt resistance obtained by new TDR method and traditional current voltage (*I-V*) measurement. The higher shunt resistance, the capacitor reaches close to open circuit at final steady state. As a result, the small noise causes deviation in extraction under high shunt resistance.

When the capacitor is fully charged, the impedance is $Z_L = R_P + R_S$. After we subtract the series resistance with the method that will be introduced in the chapter 4, we can obtain the extracted shunt resistance. Leakage current can also be obtained precisely with DC current-voltage measurement using the HP 4156A as shown in Figure 3.6. Figure 3.5 indicates that the shunt resistance extracted from both methods is basically

consistent under high leakage current (low shunt resistance). For low leakage case, the final steady voltage of reflection waveform is so close to open circuit that noise hurts the accuracy of extraction.

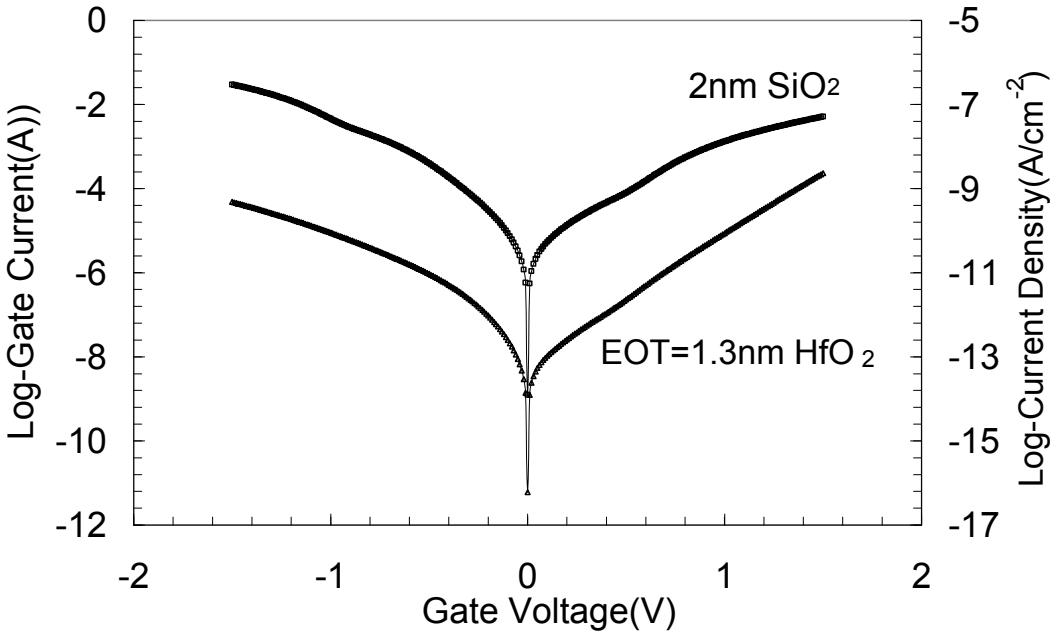


Figure 3.6 Leakage current as function of gate voltage is measured by HP 4156A. With the substrate/source/drain all grounded, gate current is obtained with the sweep of gate voltage. 2nm SiO_2 and EOT 1.2 nm HfO_2 MOS capacitor is measured respectively. After dividing the effective area 1400 μm^2, the current density is obtained.

3.4. Correction for Leakage Current

As shown in the Figure 3.4, for the MOS capacitor at inversion and accumulation, very high level of leakage current exists as the oxide is extremely thin. This is the source that lets the conventional method fail. In our new method, how could we deal with this problem?

It is found that this leakage current lowers the stored charge in the capacitor and manifests itself in time domain response as the shrink of enclosed area between the reflected waveform of the capacitor and open circuit. As a result, a correction procedure is required to account for that. In this case, the impedance Z_L when the capacitor is fully charge is equal to $R_P + R_S = R_0$. Since R_0 is not very large, the reflection coefficient is much less than 1 and the charging curve never reaches the open circuit level. It can be shown that in this case, the capacitance is given by Equation (3.3) as following:

$$C = \frac{1}{2Z_0 V_{step}} M \int_0^{\infty} \left[\left(\frac{R_0 - Z_0}{R_0 + Z_0} \right) V_{Open}(t) - V_{DUT}(t) \right] dt \qquad (3.3)$$

where $M = \dfrac{(R_0 + Z_0)^2}{R_P^2}$

This equation suggests that we can normalize the final level of the open circuit waveform to the final level of the reflected waveform from the capacitor. And then the capacitance can be extracted by integrating the enclosed area.

Figure 3.7 illustrate this normalization of the open circuit waveform and the resulting area enclosed. The enclosed area is obviously smaller than that of a leak-free capacitor. As shown in equation (3.3), the pre-integral factor M re-scales the enclosed area back to the actual area that would have been observed if there was no leakage current. Equation (3.3) can also be proved by the transmission theory and detail derivation can be referred in Appendix B.

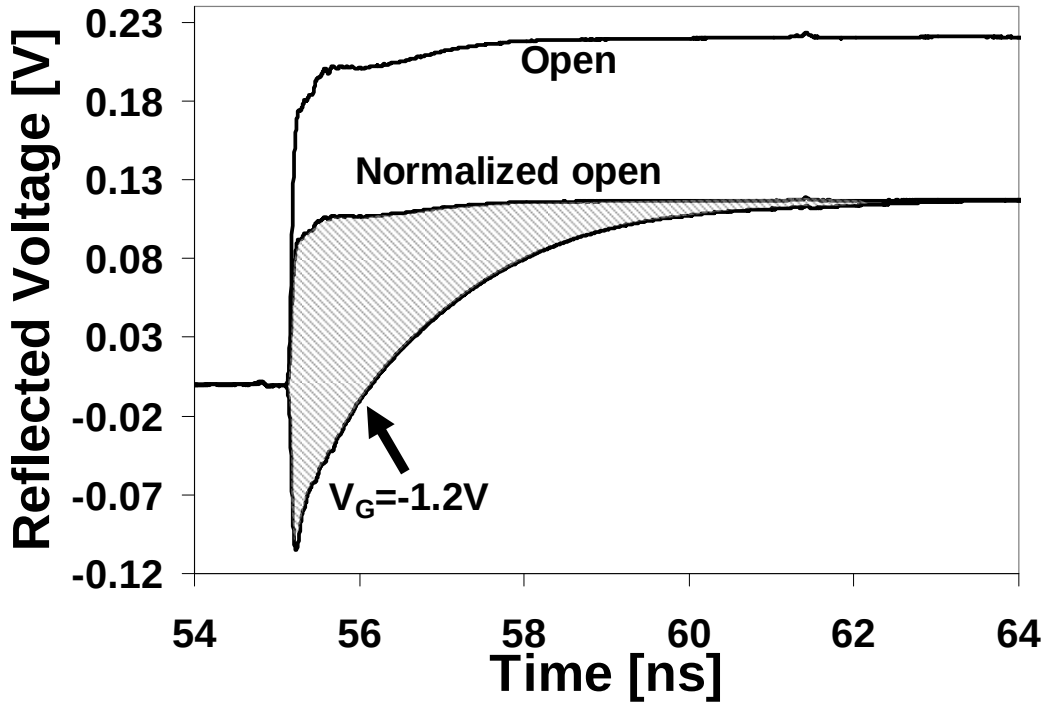


Figure 3.7 The reference waveform is normalized so that the final voltage level is scaled down to the final level of the leaky capacitor. The shaded area enclosed by the normalized reference and the leaky capacitor is the total stored charge in the leaky capacitor scaled down by a factor $M = \frac{(R_0 + Z_0)^2}{R_P^2}$.

3.5. Correction for Series Resistance

Besides the leakage current, series resistance is another source hurts the accuracy of C-V measurement. This can be easily avoided because our new method provides self-correction for series resistance, that is, no additional correction procedure is necessary. In the insert of Figure 3.1, we see that the series resistance R_S included in the equivalent circuit does not appear in equation (3.1). This can be understood by realizing that R_S not only slows down the charging process, but also modifies the effective impedance and therefore the reflected waveform's amplitude at every point

in time. This preserves the enclosed area which accounts for the change in stored charges due to the step voltage. This argument can also be proved mathematically as shown in Appendix B.

From the final voltage level of the reflected waveform, we can use equation (3.1) to calculate the value of R_0. To find R_P, we need to determine R_S. To do that, we note that at time zero the capacitor is a short circuit. The impedance at this point is simply R_S. If we have a perfect step function, then the reflection coefficient at time zero will give us R_S directly. In real situations, accurate extraction of R_S is still possible because the step function is accurately known from the open circuit waveform and the theoretical shape of the time dependent reflectivity curve is also known. The reflectivity curve can be recovered by fitting the experimental reflectivity data to the theoretical expression. The reflectivity at time zero can then be extracted by extrapolation. The procedure is straight forward but a number of potential error sources must be carefully dealt with. In chapter 4, we will demonstrate the exact procedure of implementing this idea to extract the series resistance. Here in a simple manner, we first show the C-V measurement without additional extraction of series resistance. More accurate but complicate C-V method taking account of series resistance as well as overlap capacitance will be shown in chapter 4.

When $R_P >> R_S$, the pre-integral fact M can be approximated as

$$M = \frac{(R_0 + Z_0)^2}{R_0^2} \tag{3.4}$$

By doing so we need not go through the trouble of extracting R_S and the TDR method becomes extremely simple. The resulting capacitance is an under estimation of the real value. Figure 3.8 shows the percent error as a function of R_S to R_P ratio. Typical R_S values are in the tens of Ohms. R_P needs to be at least 200 times larger for 1% accuracy. The approximation can be satisfied for most capacitors except for those with extremely high leakage.

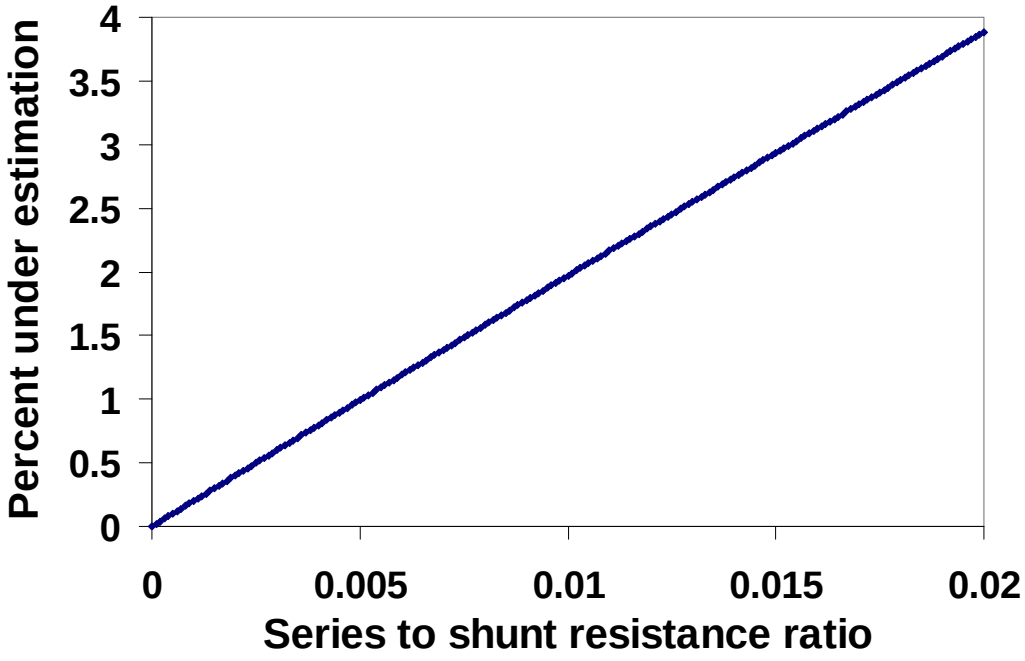


Figure 3.8 The plot of the percent underestimation of the capacitance when using equation (3.4) to approximate the pre-integral factor M as a function of the R_S to R_P ratio.

Figure 3.8 also serves to indicate how accurate R_S needs to be measured. If one replaces R_S with ΔR_S in the ratio (horizontal axis), the result is basically unchanged. Thus if we want 1% accuracy, then the ΔR_S to R_P ratio need to be less than 0.005. Since R_S is typically much smaller than R_P to begin with, even a rough estimate of R_S is enough to achieve high accuracy using equation (3.3). For example, if $R_S = 10\,\Omega$,

then 1% accuracy in CV measurement can be achieved with R_P as small as 20Ω, which is 4000A/cm^2 for our 1400 μm^2 capacitors.

3.6. Extracted C-V Characteristics

With the developed method and algorithm as discussed, we are now able to extract capacitance from time domain waveform in Figure 3.4. Figure 3.9(a) shows the C-V curve extracted using the TDR method on the MOS capacitor with 1.2 nm EOT HfO_2. Superimposed to the curve is the C-V curve measured using a lock-in amplifier. Similar to LCR meter, lock in amplifier (EG&G 5209) is basically a phase sensitive detector. Once a 1 KHz sinusoid AC signal with 10mV amplitude is superimposed on the DC and applied to the gate of MOS capacitor, it can extract capacitance out by measuring the out-of-phase current from the leakage in-phase current.

C-V characteristics of test capacitor with 2nm SiO_2 gate dielectrics are also extracted from the obtained time domain response. The result is shown in Figure 3.9(b). In this case, since the leakage current so high that the traditional method becomes inaccurate, no result is available for comparison with TDR method.

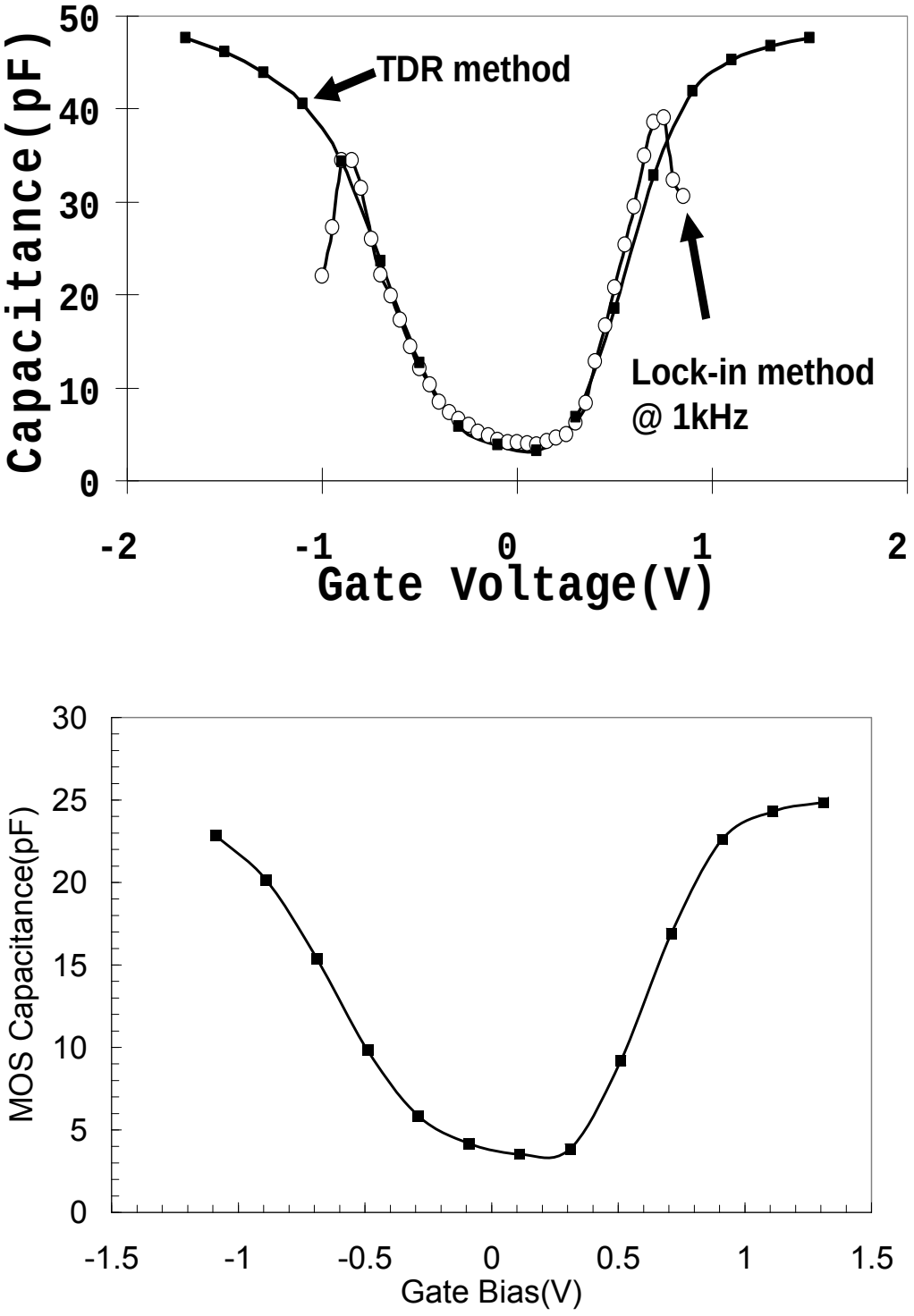


Figure 3.9 (a) Comparing the CV curves of a capacitor with a EOT=1.2nm high-κ dielectric stack measured with the new TDR method and the lock-in amplifier method. The high level of leakage at accumulation and inversion cause serious error in the lock-in amplifier method while the TDR method is not affected. (b) Extracted C-V curve of 2nm SiO_2 MOSCAP by TDR. In this case, the leakage current is so high that traditional C-V measurement can not do any reliable measurement.

As shown in Figure 3.9, the inversion layer response is observed when we are doing high frequency C-V measurement here. It is normal for our case because the MOS capacitor test structure is actually built as a transistor array with the source, drain and substrate/well all shorted together. The source and drain can generate enough minority carriers to follow the fast AC signal at gate. Clearly, the two methods agree well in region (-0.75V to 0.5V) where the leakage current is not very high. In regions with high leakage current, the TDR method produces the expected C-V curve while the lock-in method does not.

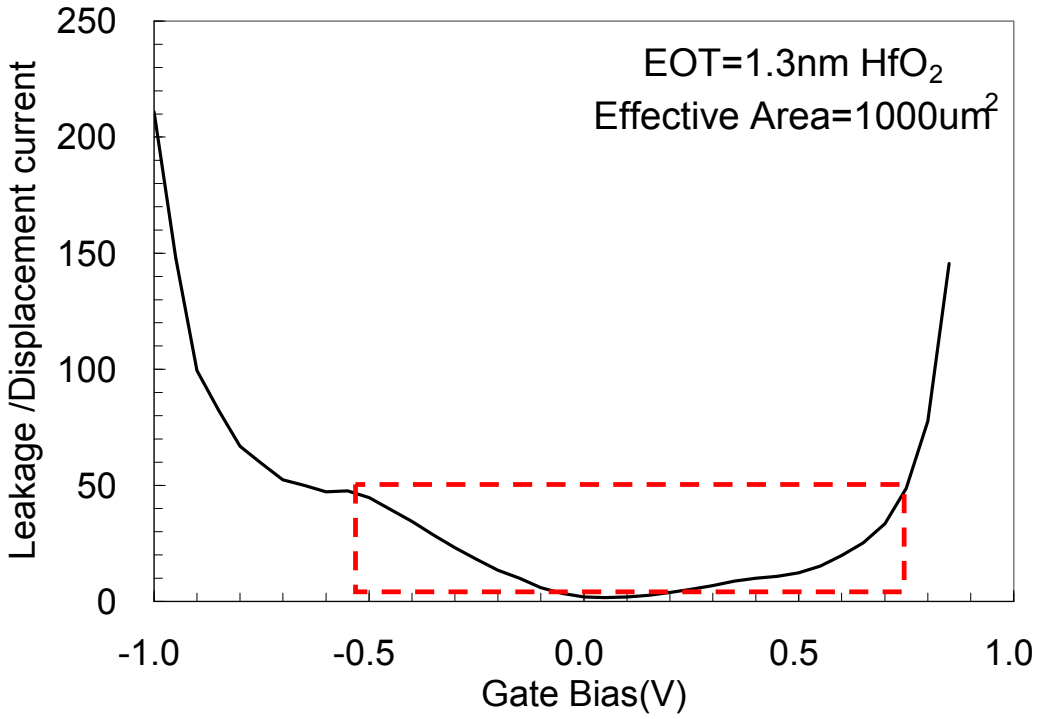


Figure 3.10 Dissipation factors which is the ratio of Leakage (in phase) current over displacement (out of phase) current in traditional C-V measurement of EOT 1.2 nm HfO_2. Within the dashed window, the lock in amplifier provides the result with less than 5% error.

This explanation can be further proven by the dissipation factors in the *y*-axis of Figure 3.10. Dissipation factors *D* are the ratio of leakage (in phase) current over

displacement (out of phase) current. It can relate to the measurement error with the following equation [7]:

$$\%error \approx e_0\sqrt{1+D^2} \approx 0.1\sqrt{1+D^2} \tag{3.5}$$

As a result, within the dashed window in Figure 3.10 where D<50, it can be estimated that the measurement error of this lock in amplifier method is within 5%. For the gate bias within this window, the capacitance extracted by TDR and lock in method basically agrees with each other. Beyond that window, the lock in amplifier method lost its accuracy (>5% error) and shows anomalous results. The result from TDR method is more reasonable.

One may ask the question, why we choose do conventional method at 1 KHz instead of any higher frequency such as 1MHz or even higher? As explained in chapter 2, the difficulty of C-V measurement involves both high leakage and finite series resistance. Long before leakage becomes a problem, the impact of finite series resistance on accurate C-V measurement has been recognized. In the 70s, standard C-V measurements were all done at 1MHz. However, this was changed in the 80s when the oxide thickness drops to near 100Å. The higher capacitance leads to lower impedance at high frequency. When the capacitor impedance is no longer much larger than the series resistance, the error introduced by series resistance can no longer be ignored and standard C-V measurements change the frequency to 100 kHz. As the gate oxide thickness shrinks further, even 100 kHz becomes a problem. However, leakage problem also becomes serious and it demands measurement to be done at

higher frequency. This unhappy trade off is really the dilemma of the C-V measurement problem.

When oxide gets ultra leaky, even the 1MHz measurement will not give accurate C-V curve. Since we already know that the C-V measurement using lock-in amplifier or LCR meter will have significant error at accumulation or inversion, we may as well go to low frequency to ensure that the measured capacitances at depletion are accurate. Note that we need at least something accurate to verify the new method. If we use the compromise of 100 kHz or 1MHz, no part of the C-V curve measured by the lock-in method is accurate. Using 1 kHz, we can at least check the depletion part of the C-V curve with confidence. This is critical because the new method must be proven accurate as a capacitance measurement before we can discuss its merit in handling high leakage situations such as accumulation and inversion.

3.7. Further Control Experiment to Test Accuracy

In the last section, we have demonstrated the principle of this new C-V method and it is proven to be successful in the application of leaky MOS capacitor. Then the next intuitive question will be: how accurate is this measurement, especially under the condition of very high leakage current? Except this new method, there is no other reliable capacitance measurement method available for very leaky capacitor as far as we know to this moment. Therefore, accuracy can not be tested straightforwardly by

direct comparison between two methods as shown in Figure 3.9. Even in Figure 3.9, conventional C-V can not be trusted at very leaky region. Then the accuracy of new method for high leakage application has still not been verified. Here, we design a series of control experiments to test the measurement accuracy. .

As we have discussed in chapter 2, in C-V measurement, the leaky MOS capacitor can be basically modeled as a simple three element equivalent circuit with an unleaky ceramic capacitor C, series resistance R_S, shunt resistance R_P as shown in the insert of Figure 3.1. Then we can build some test structures made by known ceramic capacitors and resistors to simulate the scenario of MOS capacitor. By testing the situation with different shunt resistance, we can know the capability of this new technique to handle leakage current.

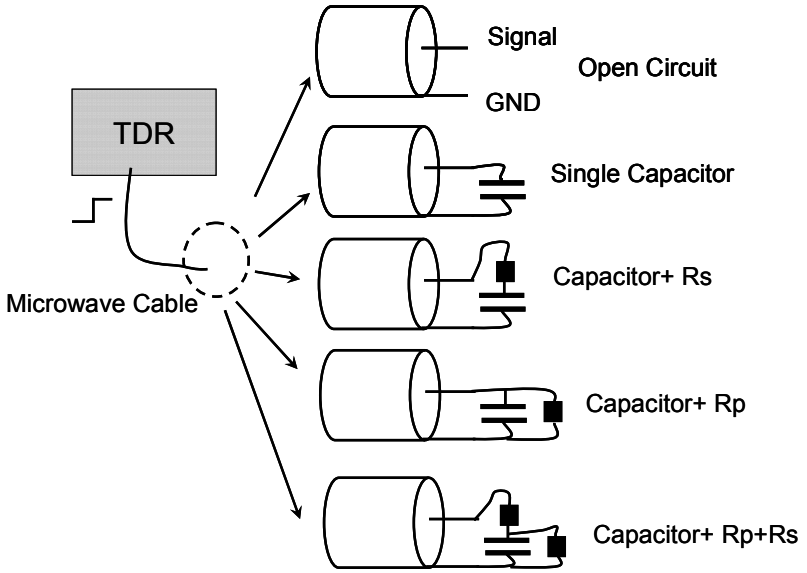


Figure 3.11 Demonstration of control experiment. It consists of TDR, microwave cable and load impedance. At the end of microwave cable, load impedance is built between the signal and ground pin. All components are made as small as possible and four cases of load impedances are built as shown. Open circuit is taken as a reference.

As shown in Figure 3.11, the control experiment consists of TDR, microwave cable and load impedance. At the end of microwave cable, the test structure built by resistor and capacitors is put between the signal and ground pin of the cable. Reference is also formed by letting the signal and ground pin open. Four different cases of load impedance are built to simulate two of three or fully three elements in the equivalent circuit. In each case, capacitance is extracted from reflected waveform obtained by TDR using the exact same procedure as C-V measurement. The capacitance is also obtained using lock in amplifier at 1 KHz as well. The error is evaluated by comparing the measured capacitance with the factory specified capacitance value with 1% tolerance. All these results are shown in Table 3.1-3.4.

Table 3.1 Control Experiment with ceramic capacitor as load only

Capacitor only	Obtained by Lock in Amp C_{Lock}	Extracted in TDR result C_{TDR}	Percentage of Error $(C-C_{TDR})/C$
220pF Capacitor	222pF	221pF	0.45%
22pF Capacitor	22.1pF	22.3pF	0.9%

Table 3.2 Control Experiment with ceramic capacitor and different series resistance as load

220pF Capacitor with different *Rs*	Obtained by Lock in Amp C_{Lock}	Extracted in TDR result C_{TDR}	Percentage of Error $(C-C_{TDR})/C$
R_S=5 Ω	222pF	221.2pF	0.54%
R_S=10 Ω	222pF	221.1pF	0.50%
R_S=20 Ω	222pF	221.4pF	0.63%
R_S=35 Ω	222pF	221.3pF	0.59%
R_S=43 Ω	222pF	221.2pF	0.54%

From the results, we can see that both TDR method and lock in amplifier method provides reliable measurement (<1% error) with the absence of shunt resistance (leakage current). When the load is only a simple capacitor (Table 3.1), both measurements overestimate the capacitance. This discrepancy may come from the 1%

tolerance of capacitance. When the series resistance is added into the load (Table 3.3), it has very little effect on the accuracy of both methods.

Table 3.3 Control experiment with ceramic capacitor with different shunt resistance as load

220pF Capacitor with different R_P	Obtained by Lock in Amp C_{Lock}	Extracted in TDR result C_{TDR}	Percentage of Error $(C-C_{TDR})/C$
R_P= 1KΩ	222pF	221.3pF	0.59%
R_P= 500Ω	212pF	221.5pF	0.68%
R_P= 220Ω	178pF	220.9pF	0.40%
R_P= 100Ω	120pF	221.6pF	0.72%
R_P= 47Ω	N/A	221.8pF	0.81%
R_P= 20Ω	N/A	217.7pF	1.04%
R_P= 10Ω	N/A	200.2pF	9.00%

Table 3.4 Control Experiment with ceramic capacitor with different shunt resistance and series resistance load

220pF Capacitor with different R_P and R_S	Obtained by Lock in Amp C_{Lock}	Extracted in TDR result C_{TDR}	Percentage of Error $(C-C_{TDR})/C$
R_S=43Ω R_P= 220Ω	185pF	221.0pF	0.45%
R_S=43Ω R_P= 100Ω	135pF	221.3pF	0.59%
R_S=10Ω , R_P= 47Ω	N/A	221.5pF	0.68%
R_S=35Ω , R_P= 47Ω	N/A	221.7pF	0.77%
R_S=5Ω , R_P= 20Ω	N/A	217.5pF	1.13%
R_S=20Ω , R_P= 20Ω	N/A	217.4pF	1.18%
R_S=10Ω , R_P= 10Ω	N/A	196.5pF	10.6%

*N/A: no reliable reading can be obtained under that situation

When the shunt resistance alone is added in (Table 3.3), leakage current affects the measurement accuracy. It can be seen that TDR can still provide accurate capacitance within 1% error for shunt resistance bigger than 20Ω, which is equivalent to 4000A/cm^2 leakage current density on 1400 μm^2 size MOS capacitor under 1V across the oxide. On the other end, for lock in amplifier method, significant error starts to be observed when the shunt resistance is just over 500Ω. After all, as shunt resistance is

further reduced, the leakage current is so high that eventually no reliable reading of capacitance can be obtained. From these comparisons, it is obvious that TDR method is much superior.

When both series resistance and shunt resistance are introduced into the test structure (Table 3.4), it is the closest situation to the leaky MOS capacitor. Compared to the case with the shunt resistance alone, the series resistance has not much effect on capacitance extraction of TDR method. TDR method still offers good result until shunt resistance is pushed below $20\,\Omega$. Moreover, the combination effect of series and shunt resistance greatly lowers the accuracy of traditional lock in method, consistent with the analysis done in chapter 2.

In summary, from these control experiments, it can be expected that the capacitance method based on TDR can maintain less than 1% accuracy under very high leakage current (4000A/cm^2 leakage current density on 1400 μm^2 size MOS capacitor under 1V across the oxide). Both series and shunt resistance effect can be well corrected. On the contrary, lock in method runs into trouble when R_P<500Ω. For the 65 nm technology, the reported leakage current is 800A/cm^2 which is equivalent to R_P~90 Ω with the size of test structure used here (1400 μm^2) and 1V applied bias. Obviously, the conventional can not hand that and it is fortunate to find this new TDR C-V method to provide a solution.

3.8. Suggestion on Future Improvement and Conclusion

Everything is two folded as the coin has two sides. Accompanying with so many fantastic merits, there are some disadvantages for this technique. Here are some suggestions on further work to improve these limitations.

First, TDR provides step function with ~0.2V magnitude. It is relatively too high for the applied DC gate bias ranging from -1.2V to 1.2V. It reduces the measurement sensitivity. Further improvement needs to lower the magnitude of step voltage. At the same time, the smaller input signal will in turn decrease the signal to noise ratio since the noise level does not change much. Actually, the noise can be reduced by taking the average of many repetitive captures. As far as not very high density of data points is required, it will not be a problem. Therefore, this disadvantage is limited by the available instrument rather than method itself. The instrument used in our experiment is from Tektronix CSA 8000 with TDR plug in. It is not designed for this C-V measurement purpose and the step voltage is fixed as ~0.2V. Recently, after we demonstrated our C-V method, Agilent shows the interest in further developing this technique and building a new instrument specifically for this purpose. If the agreement can be reached, the smaller step voltage is achievable.

Secondly, this TDR has maximum 2.5 μs time window. By assuming R-C charging time constant is one of fifth of time window and the resistance is 50Ω, it can be

estimated that the maximum capacitance it can extract is 10 nF. As to the low end, the rise time determines the minimum capacitance it can extract. Finite rise time itself is not a problem because all the artifacts have already been taken account by the open circuit reference. The error due to the rise time can be eliminated by calibration. However, when the capacitance is so small that the charging time of the capacitor is comparable with system rise time or even smaller, the resolution becomes a big problem. Under that circumstance, accurate measurement is questionable. TDR can generate 35 ps rise time step voltage. However, the step function is slowed by the low quality bias TEE and results in final 70 ps rise time pulse on test structure. In a conservative estimation, minimum 1.4 pF capacitance can be obtained. The only way to improve this issue is to find better bias TEE and instrument so that fast pulse can be achieved. In today's common large area MOS capacitor measurement application, this measurement range is good enough. Again, this disadvantage can be solved by a better instrument. It is exciting to see whether Agilent can build a better one that can greatly improve this method and put into commercial usage.

In summary, we introduce a new high-accuracy method to measure C-V on highly leaky MOS capacitors. This is a new application of a well established measurement technique, namely Time-Domain-Reflectometry (TDR). The effect of leakage current can be accurately corrected and the impact of series resistance can be eliminated. This method is simple to use and can be implemented as a routine device characterization procedure.

Chapter 4

C-V measurement II

Extraction of series resistor and overlap capacitance by time domain method

In chapter 3, we have demonstrated a simple and precise measurement technique to obtain the C-V characteristics of very leaky MOS device. This technique utilizes the Time domain Reflectrometry to generate the fast step voltage to the test device and monitor the reflected waveform. By analyzing the reflected waveform and comparing it to the response of calibration structure (open/short circuit), in theory, we can extraction many characteristics of the device under test. In chapter 3, we have demonstrated a way to extract capacitance. Is that all we can do? What else information can we learn about the device from the reflected waveform?

On the other end, as MOS device advances rapidly with thinner oxide and shorter channel, the design of test structure becomes more challenging. From both fabrication and design point of view, some parasitic is inevitable. For example, as the transistor-like test structure with source and drain has been widely used in C-V measurement nowadays, the overlap capacitance is introduced [10, 14-16, 64]. Moreover, the series resistance from substrate and source/drain junction is well known as an important source of an error for C-V measurement for many years [12, 65]. Therefore, simple and accurate methods are urgently required to extract these

“imperfections” so that the introduced error can be corrected.

However, although extensive efforts have been made, it is not easy task to find a simple and accurate way [66-73]. High leakage current brings a lot trouble and close out any possible solution. Moreover, the way to extract these parasitic is very different from the one used for C-V measurement and additional experimental setup or test structure is needed. For example, in split C-V measurement [66-73] which is usually used to extract overlap capacitance, small area (<100 μm^2) device is used and capacitive displacement current at source/drain as well as gate is measured respectively. But for C-V measurement by LCR meter, typically the source and drain is grounded and larger area (~1000 μm^2) device is chosen. Different setup and device increases the measurement complexities. More importantly, it also introduces more sources of error, such as variation between devices, contact resistance difference introduced by probe force, parasitic caused by difference between experimental setup and so on. Therefore, it will be ideal if there is a measurement technique that can not only offer the C-V characteristics but also provide the information about these parasitic. It sounds almost impossible for current situation that even finding a solution to extract each component alone is already difficult It will be a dream that one can find a way to hand all these components by a SINGLE measurement.

Fortunately, we find one, which is Time Reflectrometry (TDR) as we introduced in chapter 3. Besides its function of knowing the device C-V characteristics as we

showed in chapter 3, we will also demonstrate in this chapter that it can be manipulated to extract the series resistance and overlap capacitance, which are two main error sources. All the involved extraction procedures are just data analysis steps. It utilizes the same experimental reflected waveform from the C-V measurement. No additional step is added as to the experiment point of view. Besides its advantage of simplicity, these extraction methods are also proven to be accurate. It can further improve the accuracy of our new C-V measurement technique.

4.1. Source of Error in C-V Measurement- Series Resistance

Before we start discuss on how to correct the error introduced by device parasitic in C-V measurement, we need to identify what they are and where they come from. In chapter 2, we have shown the instrument error is amplified and limits the accuracy in the conventional method. Except that, it is also recognized that some parasitic components built in the MOS device itself can affect the accuracy as well. Among them, series resistance and overlap capacitance are two prominent error sources [12, 66-73].

In the MOS device, the series resistance is arising from the finite resistance of the source/drain contacts, inversion channel and gate material. With the accelerated scaling roadmaps in recently years, series resistance becomes having a much larger role for practical device performance, such as C-V characteristics.

Therefore, many efforts have been made to identify this important source of error [12,14-16]. Unfortunately, independent, accurate determination of series resistance is generally not possible. In conventional C-V measurement such as LCR parallel mode, the series resistance is buried in the measured result and cannot be separated. For thin oxide MOS device with low capacitor impedance and large leakage current, the effect of the series resistance is more significant. With thin oxide device, the high leakage current also forces one to measure C-V at high frequency (>100 KHz) recently instead of low frequency (~1 KHz) to increase the capacitive displacement current. It further lowers the capacitance impedance and magnifies the error of series resistance. Therefore, reliable extraction of the capacitance definitely requires the series resistance to be known accurately.

The multi-frequency method proposed by Yang *et al.* [6] allows one to extract, in theory, the true capacitance, the shunt resistance and the series resistance from the measured capacitance and dissipation factor from LCR meter. However, as shown in detail in chapter 2, this conventional method is no longer accurate as the oxide gets so thin. A better but simple measurement technique is urgently required. To that end, the method introduced in this chapter is such a solution that can accurately extract series resistance without knowing any other parameters first.

4.2. Source of Error in C-V measurement- Overlap Capacitance

Besides the concern for series resistance, the overlap capacitance is another big source of error due to the transistor-like test structure used in thin oxide device C-V measurement. As the gate oxide gets so thin, not only the C-V measurement technique itself becomes problematic but also the test structure needs to be designed carefully. If not, one will run into trouble. It has been experimental observed that only high frequency C–V curves are found no mater what frequency has been used in the C-V measurement on 2nm SiO_2 device[64]. As a result, it is impossible for one to get the correct inversion capacitance without carefully design.

On the other hand, since the normal operation of MOSFET is under strong inversion, the inversion gate capacitance is an important parameter to be extracted accurately. The extraction of oxide thickness also depends on accurate knowledge of the inversion capacitance. The lack of inversion behavior measured at the traditional MOS capacitor test structure [64] can no longer be used for inversion capacitance extraction.

Taking a n-channel MOS capacitor with traditional design as an example, to form the inversion state, the gate should be DC biased positively to generate the electrons in the surface of channel as shown in Figure 4.1(a). However, for the ultra-thin gate oxide, the generated electrons are directly tunneled into gate electrode [74]. As a

result, inversion layer can not be formed properly because electrons cannot be swept into the channel with the applied small AC signal as that for a thick gate oxide. That is the reason why no matter what frequency of input signal is applied to the gate, only high frequency C-V curve is observed experimentally.

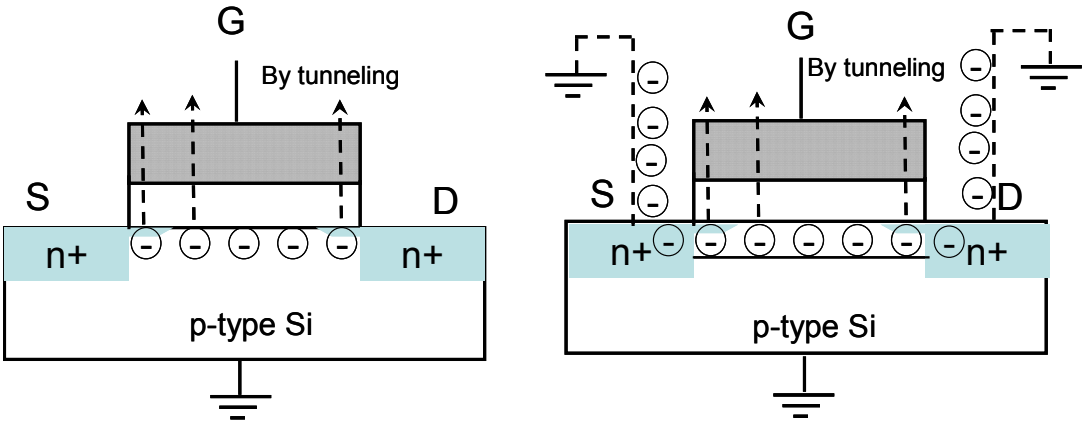


Figure 4.1(a) the distribution of electrons in strong inversion state of NMOSFET with ultrathin oxide and source/drain electrodes without connection to ground (or no source/drain). The generated electrons are tunneled into the gate electrode (b) Presence of source and drain continuously provides the electrons to form the inversion layer. Picture is taken from reference 64.

Therefore, to overcome this problem, one must find some sources that can provide the inversion charges. This source must have the ability to provide inversion charges unlimitedly so that it can tolerate the loss of charge due to carrier tunneling. The transistor's source and drain junction is the best candidate. The capacitor can be constructed like a transistor but with source, drain and substrate all tied together to ground as illustrated in Figure 4.1(b). In this way, the lost minority carriers are supplied from the ground through the source/drain regions, so that the inversion layer is built again. Therefore, this transistor-like capacitor structure basically solved the problem and becomes popular. Furthermore, at this time, the channel is very close to the ground through the high conduction of source and drain regions, the change of

electron numbers always keep up with the AC small-signal variation and lead to charges exchange with the inversion layer in step with the measurement signal. Thus, the low frequency C–V curves are found even it has been measured under 1 MHz high frequency signal [14, 16].

However, the presence of source and drain introduces error in C-V measurement and the capacitor must be designed carefully. In designing MOS capacitor with source and drain, the overlap capacitance is carefully minimized to reduce error. Additional error sources such as channel resistance induced non-uniform surface potential [15] must also be avoided using short channel length [14, 16]. As the channel length gets small with aggressive scaling, parasitic elements such as gate overlap and coupling capacitances at the source and drain ends of the channel is comparable to the gate-to-channel capacitance. It will play an important role in device characterization such as C-V characteristics.

Even though people are aware of the potential error due to overlap capacitance, the extraction of overlap capacitance is not easy when the leakage current gets high. The overlap capacitance is conventionally extracted by split C-V measurement at low frequency by LCR meter [66-74]. Basically, a small AC signal is applied to the gate of test structure- which is a transistor. The displacement current at source/drain junction and substrate is measured separately. This is where the name of “split” comes from. From the displacement current, the gate capacitance and overlap capacitance can be

extracted. Further extraction of effective channel length and mobility can also be worked out.

For thicker oxide device, this measurement technique has been successfully proven to be accurate. However, for most advanced device of interest nowadays, this tool is no longer useful any more. The leaky and short channel device poses a dramatic challenge. It is not surprising because this method evolutes from conventional C-V measurement. They share the same principle, instrument as well as the limitation. As a result, our TDR method becomes extremely valuable because it exactly offers such a fantastic solution to measure the overlap capacitance under high leakage.

Therefore, in this chapter, we will demonstrate that how it can be done and further show that it contributes a significant error to the C-V measurement. Once we can extract the overlap capacitance and series resistance, we can correct them on the measured C-V curve (chapter 3) and make it free of the influence of these parasitic.

4.3. Proof of Existence of Overlap Capacitor

Before we show the extraction procedure, it is better to identify the existence of overlap capacitance first. With the unique advantage of TDR, it can be identified in reflected voltage of MOS capacitor. According to TDR theory, the ideal capacitor initially behaves like a short circuit, then exponentially charges up and eventually becomes an open circuit. Figure 4.2 (dot line) shows the reflection from an ideal

inversion capacitor (zero leakage and zero series resistance) when the step function is also ideal.

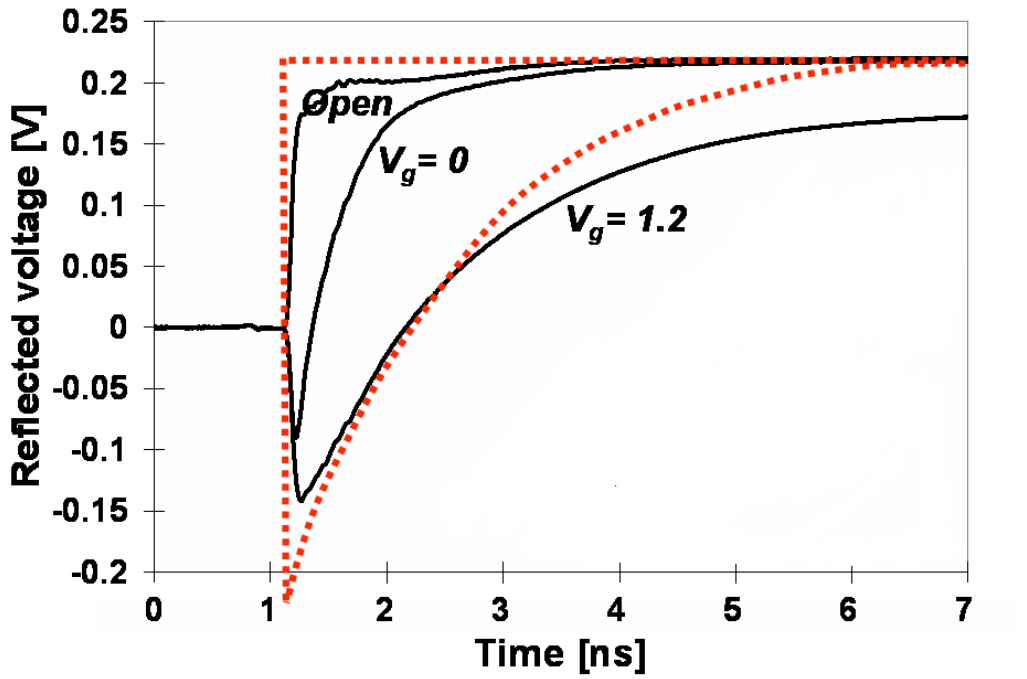


Figure 4.2 Reflected voltage waveforms for open circuit (reference), capacitor in depletion (V_G=0V) and capacitor in inversion (V_G=1.2V). Dotted lines illustrate the open circuit reflection of an ideal step function (zero rise-time) and the reflection of the ideal step function off an ideal inversion capacitor (zero leakage and zero series resistance)

Compared to the ideal case, the reflected voltage of 2 nm SiO_2 MOS capacitor at depletion and inversion bias (solid line in Figure 4.2) shows a similar trend but the shape is modified by the realistic artifacts such as series and shunt resistance. So it can be ascertained the behavior in both cases still looks like a simple capacitor charging curve.

On the other hand, the response of MOS capacitor at accumulation region shows something different. Figure 4.3 shows the reflected voltage waveforms for the same

capacitor under negative biases ranging from depletion to accumulation. The three waveforms in Figure 4.2 are also included for comparison. It is clear that under negative bias the waveforms are not simple capacitor charging curves. Instead, they look like two capacitors are being charged with the smaller one charging up much faster than the larger one. Such a separated charging of two capacitors cannot be observed in conventional C-V measurements. It is the unique ability of the TDR method.

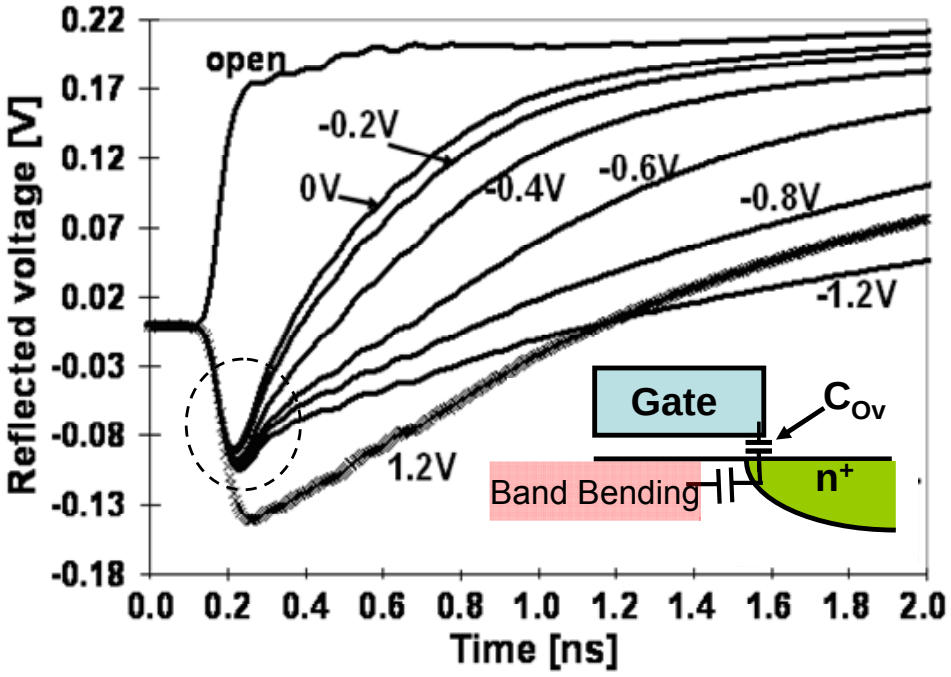


Figure 4.3 Reflected voltage curves from the SiO_2 capacitor. Only the negative bias curves are included to highlight the non-single time constant charging behavior. The curve from open circuit and from strong inversion is also included for reference. Insert: Illustration of two additional capacitors exist due to the presence of source and drain. One is the overlap capacitor C_{OV}, and the other is band-bending capacitor C_{Bb}.

Then the question comes as: what is the additional capacitance except the gate to

channel capacitance? Parasitic capacitance due to probe pad and cable are not possible candidates. That is because if they were we should see them in the reference waveform (open circuit) as well. An immediate suspect for the small capacitor is the overlap capacitor C_{OV} coming from the overlap region of gate to drain/source. However, an additional capacitor due to surface band bending also exists as shown in the insert of Figure 4.3. However, this capacitor is in parallel with the relatively small series resistance of the substrate contact and therefore can only contribute to a very small error. Thus the main source of the additional small capacitor is the overlap capacitor.

4.4. Accurate Model of MOSFET Including Overlap Capacitor

With the awareness of existence of overlap capacitance, the previous three element circuit model in C-V measurement should be modified. Figure 4.4 shows the cross section of MOS capacitor (or transistor). Besides the gate to channel capacitance C_{gc}, the parasitic capacitance C'_{OV} arising from the overlap region between gate and drain/source extension is also included. Because the carriers can tunnel through both the gate to channel and overlap region, shunt resistance R_{Pa} should be used to represent the leakage current at gate to channel region while shunt resistance R_{Pb} models the one from overlap region.

Furthermore, there are two types of series resistance. One is the series resistance (R'_{Sinv}) from source/drain junction. The other is the series resistance (R'_{Sacc}) coming

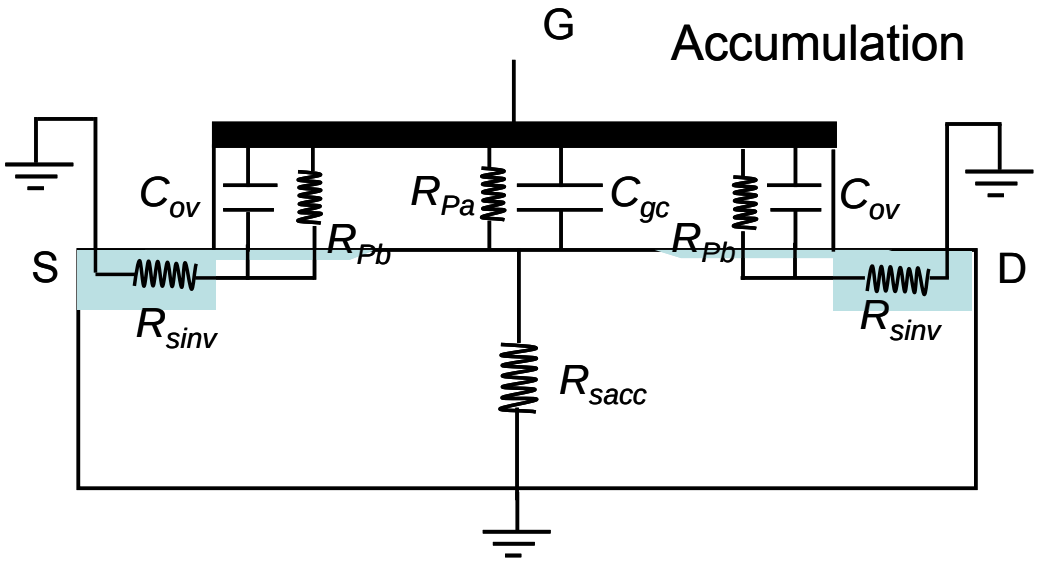
from substrate. For an n-MOSFET, drain/source is heavily doped (>10^{18}cm^{-3}) and has much larger doping concentration than substrate (~10^{16} cm^{-3}). Since the semiconductor conductivity is proportional to the carrier concentration, R'_{Sacc} from substrate is much larger than R'_{Sinv} from source/drain.

The equivalent circuit model is somewhat different between accumulation and inversion mode of MOSFET operation as shown in Figure 4.4(a) and (b) respectively. For an n-channel MOS device, when a negative gate bias is applied, the electrons at oxide/semiconductor interface are pushed away and an accumulation layer of holes is formed. These electrons can eventually reach the ground either through the substrate or through the source/drain junction. This stream of electrons is the capacitive charging current. According the flowing path, both series resistance R'_{Sinv} and R'_{Sacc} should be taken account into the circuit model as shown in Figure 4.4(a).

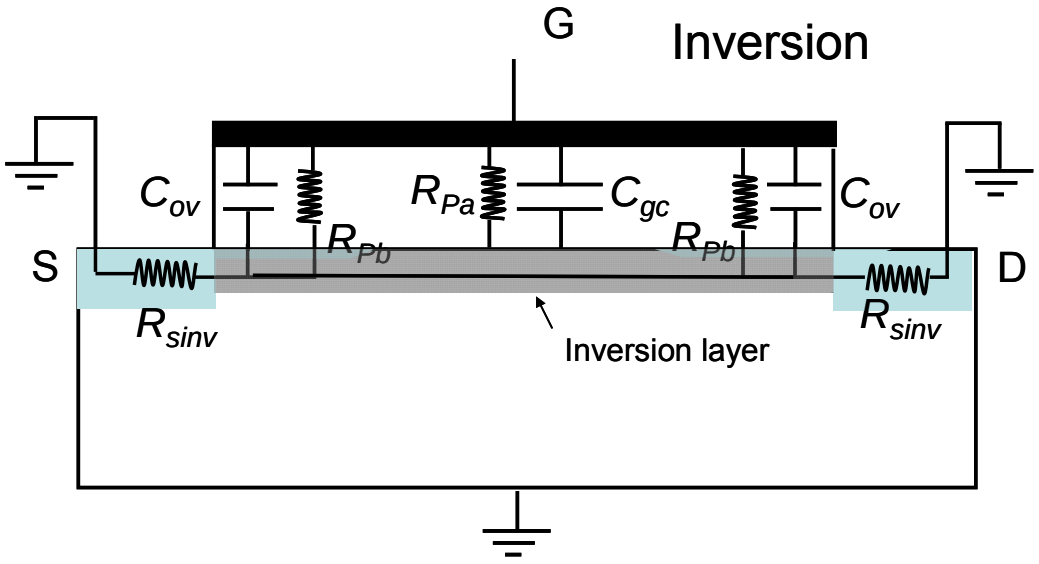
On the other hand, for the positive gate bias, electrons from source/drain form an inversion layer at oxide/semiconductor interface. Because the electrical field blocks the electrons to substrate, electrons can only arrive at the ground through source/drain. Therefore, in this case, only R'_{Sinv} is included and R'_{Sacc} is neglected as shown in Figure 4.4(b).

By assuming the symmetry of drain and source junction, two identical series resistance R'_{Sinv} , shunt resistance R'_{Pb} capacitance C'_{OV} can be simplified into one

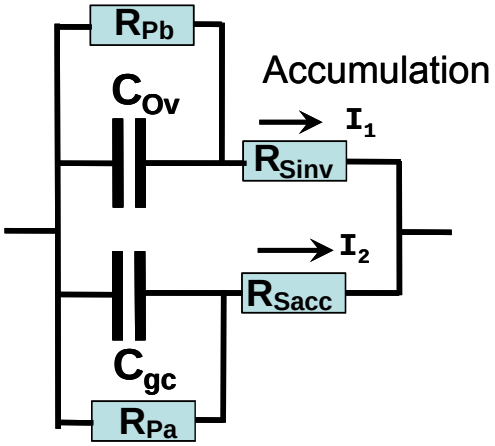
component as R_{Sinv}, R'_{Pb} and C_{OV} respectively (R_{Sinv}=0.5 R'_{Sinv}, R_{Pb}=0.5 R'_{Pb} and C_{ov} =2 C'_{OV}). After some simplifications, the equivalent circuit is illustrated in Figure 4.4(c) and (d). .



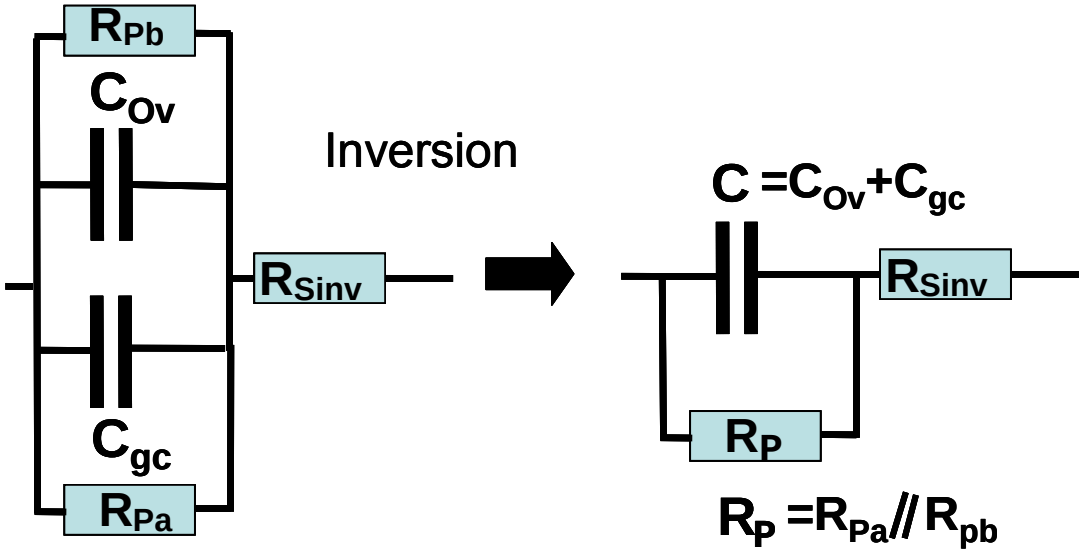
(a)



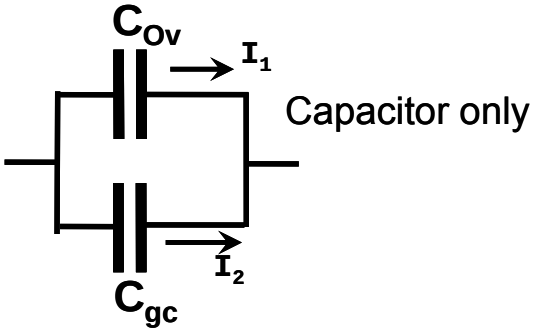
(b)



(c)



(d)



(e)

Figure 4.4, Cross section of the test capacitor (or transistor) with approximate circuit model with overlap capacitor included. R_{Sinv} is the series resistance to source/drain and R_{Sacc} is the series resistance to substrate. C_{OV} is the gate to drain/source overlap capacitance. C_{gc} is the gate to channel capacitance. (a) Circuit model of MOS capacitor under accumulation; (b) Circuit model of inversion case. The equivalent circuit after re-arrangement is also shown respectively in (c) and

(d). At accumulation, the charging process can be approximately divided into short time and long time region. Under each region, the equivalent circuit can be simplified as shown in(c). (e) The case without any external resistance, only overlap capacitance and gate to channel capacitance. I_1 and I_2 represent the current flow to overlap and gate to channel capacitance respectively.

Compared to the three elemental equivalent circuits as introduced in chapter 3, this new model is more accurate by taking account of overlap capacitance and series/shunt resistance at different region. This equivalent circuit is a complete model of the transistor-like test structure. In principle, it should be able to explain the observation of two different charging processes as shown in Figure 4.3. What exactly happened when a fast step pulse is applied to this equivalent circuit?

To analyze the behavior of this complicated system, it is better to start from a simple case – two capacitors in parallel without any external resistance. From the circuit theory, two capacitors in parallel is equivalent to one bigger capacitor and a simple capacitor charging will be expected. In another word, two capacitors are charging at the same rate. This can also be understood from the charging current point of view. As shown in figure 4.4(e), much more current flows to C_{gc} than C_{OV} due to its low impedance and larger capacitance. Although C_{gc} demands more charge due to its larger capacitance as well, it still can be kept charging at the same rate as C_{gc} with more charging current I_2. Once they are charged at same rate, the charging current is proportional to their capacitance:

$$\frac{I_1}{I_2} = \frac{C_{OV}}{C_{gc}} \tag{4.1}$$

However, the situation gets complicated when different external resistance is involved

as shown in figure 4.4(c), (d). Moreover, the involved mechanism will be different at accumulation and inversion case because they have different equivalent circuit. We need to analyze each case individually.

For the accumulation case, when the step voltage first reaches the device, both C_{OV} and C_{gc} behave like a short circuit. Therefore, at very beginning stage of charging, the amount of current flowing through each path is simply determined by the resistance in the circuit. Since R_{Sacc} from substrate is much larger than R_{Sinv} from source/drain, much more current is driven to charge C_{OV}. At the same time, due to its smaller capacitance, C_{OV} requires fewer carriers to finish charging. As a result, with more current and smaller capacitance, C_{OV} is charged up much faster than C_{gc} at beginning of the charging. Therefore, the initial capacitor charging process is dominated by the C_{OV}, consistent with the observation in Figure 4.3. After a while, since C_{OV} is charged up faster, its impedance also increases faster than the one of $C_{gc.}$ The ratio of charging current on C_{OV} and C_{gc} is reduced down and eventually reached the steady state: $I_1/I_2 = C_{ov}/C_{gc}$. From that moment on, the two capacitors keep charging at the same rate. This is the steady condition for the current ratio between these two capacitors. That is because any small relative increase of the current on one capacitor will result in faster charging and faster impedance increase. And then it will in turn drive the current ratio back to this steady state. In a summary, at accumulation, the overlap capacitor dominates the initial charging and eventually two capacitors are charged at the same rate. This is the reason for the observation of two capacitor charging phenomenon at

accumulation as shown in Figure 4.3.

For the inversion case, the formation of inversion layer and electrical field forces all the electrons flow into source and drain junction. Different from the accumulation case, the unbalanced series resistance situation does not exist in this case. The amount of current flowing to C_{gc} and C_{ov} is simply determined by their impedance. It is simply two capacitors in parallel and can be modeled as a single capacitor $C_T = C_{gc} + C_{OV}$ as shown in Figure 4.4(d).Therefore, the two capacitors are charged at the same rate all the time. It is also consistent with the observation in Figure 4.3 that the two charging processing phenomenon disappears in inversion region

4.5. Basic Principle of Series Resistance Extraction

Now we have device equivalent circuit, incident step voltage from open circuit, capacitor reflected voltage. In principle, we should be able to extract all the parameters in this circuit. We first start with the series resistance R_{Sacc} and $R_{Sinv.}$

Using our TDR technique, the series resistance can be extracted very accurately with a simple manner. Series resistance can be obtained by fitting the capacitor charging curve and calculating the impedance at time zero. With the incoming step function, the MOS capacitor behaves like a short circuit at first (time zero) and then like an open circuit when fully charged. In other words, at time zero ($t = 0$) the three-element

equivalent circuit is reduced to just the series resistance. This fact immediately suggests that the series resistance can be directly determined from the reflectivity ρ at time zero using

$$\rho = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{R_S - Z_0}{R_S + Z_0}\bigg|_{t=0} \tag{4.2}$$

If the step function were ideal (zero rise-time), the above simple method would give the true series resistance. With step function that is less than ideal, the reflectivity at time zero is masked by the rise-time, and must be extracted by extrapolating the capacitor charging curve to time zero as shown in Figure 4.5.

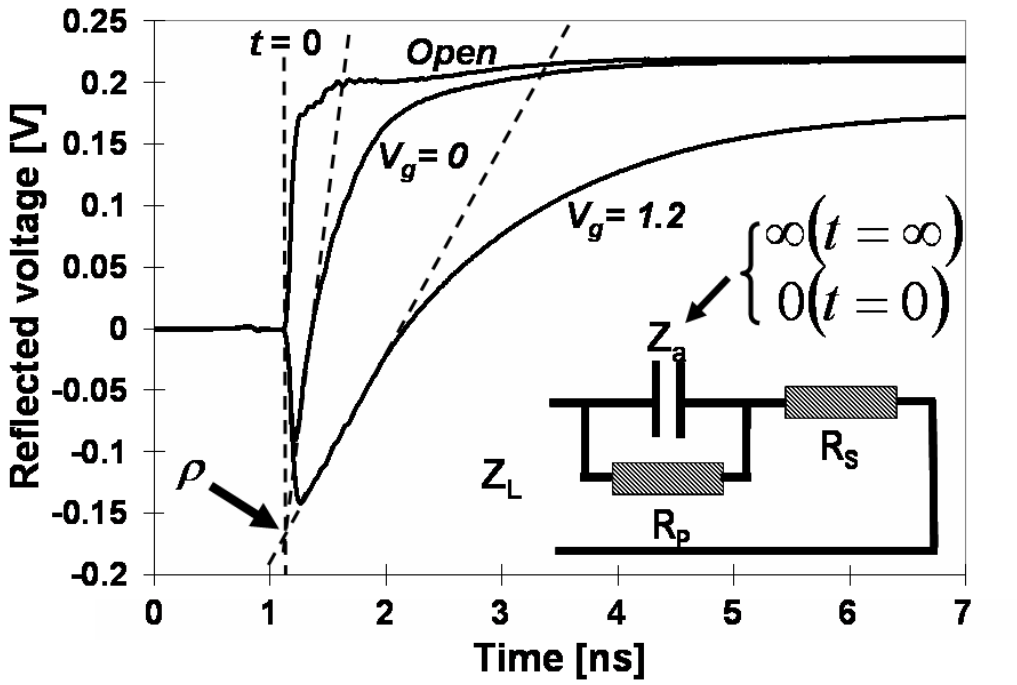


Figure 4.5 Reflected waveforms from open circuit (reference), MOS capacitor at depletion ($V_G = 0$) and at inversion ($V_G = 1.2V$).the dotted lines are extrapolations of the charging curves toward time zero (also marked by a dotted line) to extract the time zero reflected voltage that can be used to calculate the time zero reflectivity.

From Figure 4.5, it is clear that the step function not only has non-zero rise-time, but also has non-ideal wave shape. This wave shape inevitably affects the direct extrapolation of the charging curve to time zero. Figure 4.6 illustrates this problem with a measured and a simulated reflected voltage curve with low and high series resistance respectively. The non-ideal wave shape of the step function creates structure in the reflected voltage curves. Moreover, the effect is more prominent for the high series resistance case. Since the imperfection of the step function typically happens at early time where the extrapolation is sensitive, this can degrade the accuracy of series resistance extraction.

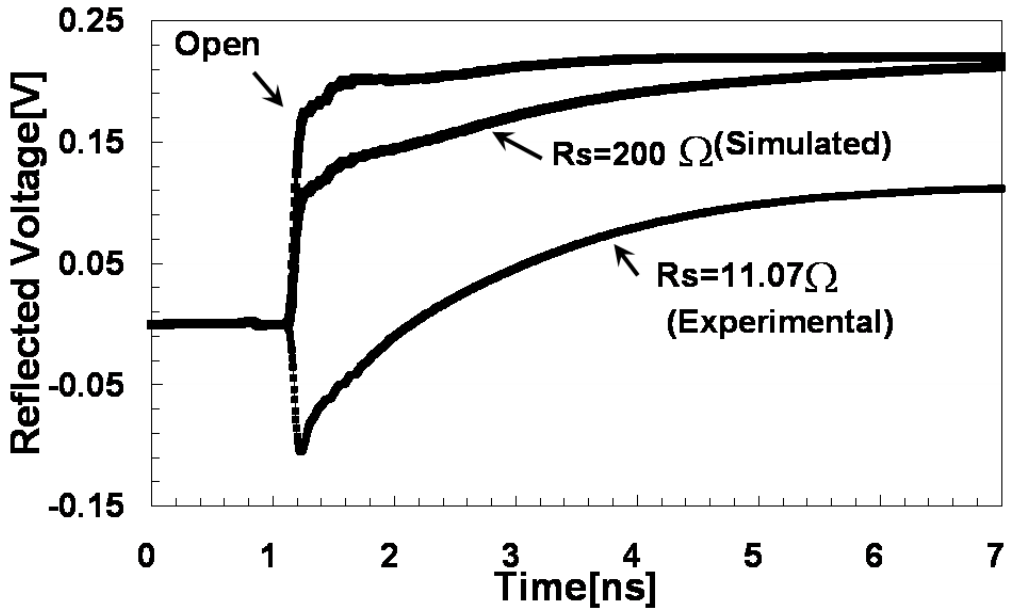


Figure 4.6 measured reflected voltage curve for a low series resistant case and a simulated curve using ideal capacitor similar in size of the measured curve but with much higher series resistance.

To reduce the effect of the non-ideal step function, we divide the charging curve with the step function to recover the reflectivity curve which is theoretically, for the three-element model, given by equation (A.18) in Appendix A:

$$\rho(t) = a + b\exp(-t/\tau) \tag{4.3}$$

where $a = \frac{R_P + R_S - Z_0}{R_P + R_S + Z_0}, b = -\frac{2Z_0 R_P}{(R_S + Z_0)(R_P + R_S + Z_0)}, \tau = \frac{R_P + Z_0 + R_S}{R_P (R_S + Z_0) C_L}$

We can do this because the reflected voltage waveform is the product, not the convolution, of the incident waveform (open circuit reference) and the reflectivity curve. The parameters of equation (4.3) can be determined by fitting to the reflectivity data. We typically avoid using the reflectivity data at early time before the incident waveform (open circuit reference) reaches 90% of its full voltage in our fitting. We choose to stop the fitting range at the 63 percent point of the final steady state of reflected voltage waveform of the capacitor.

The purpose is to cover the rapidly changing section of the curve without being skewed by excessive data points from the slowly varying or flat sections. Since the form of the equation is known, the extraction of the three unknowns is highly dependable. The resulting equation with known constants is then used to generate the reflectivity at time zero as shown in Figure 4.7. With the reflectivity at time zero, the series resistance is given by

$$R_S = Z_0 \frac{1 + \rho(0)}{1 - \rho(0)} \tag{4.4}$$

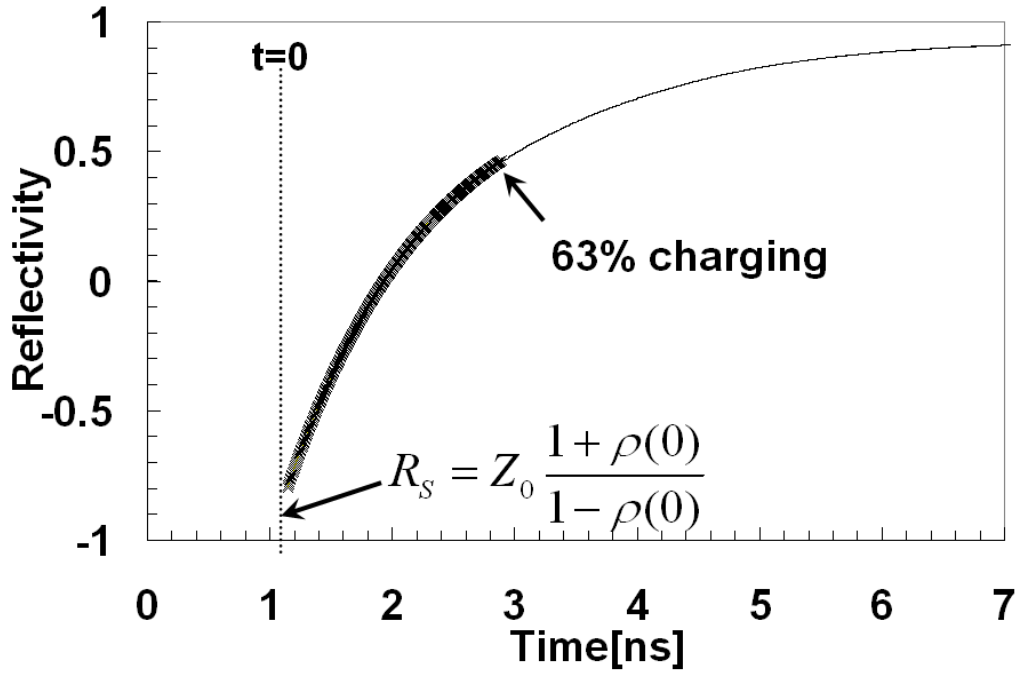


Figure 4.7 Reflectivity curve calculated by dividing the reflected voltage curve (charging curve) by the open circuit reflection. The X curve is the fitting result extrapolated to time zero.

4.6. Extraction of Series Resistance

With this principle, we can extract the series resistance from the obtained the time domain response of 2nm SiO_2 MOS capacitor (Figure 3.4 in chapter 3). In inversion, a single capacitor charging behavior is involved. Therefore, the fitting becomes quite straightforward. The section chosen for fitting starts from the 90% of the incident open circuit waveform where the capacitor charging takes over the rise trend of step voltage as shown in Figure 4.8. We also set 63% charging point as the end of fitting section. Then we calculate the reflectivity of this specified fitting section of reflected voltage and fit it with an exponential function (Equation (4.3)).

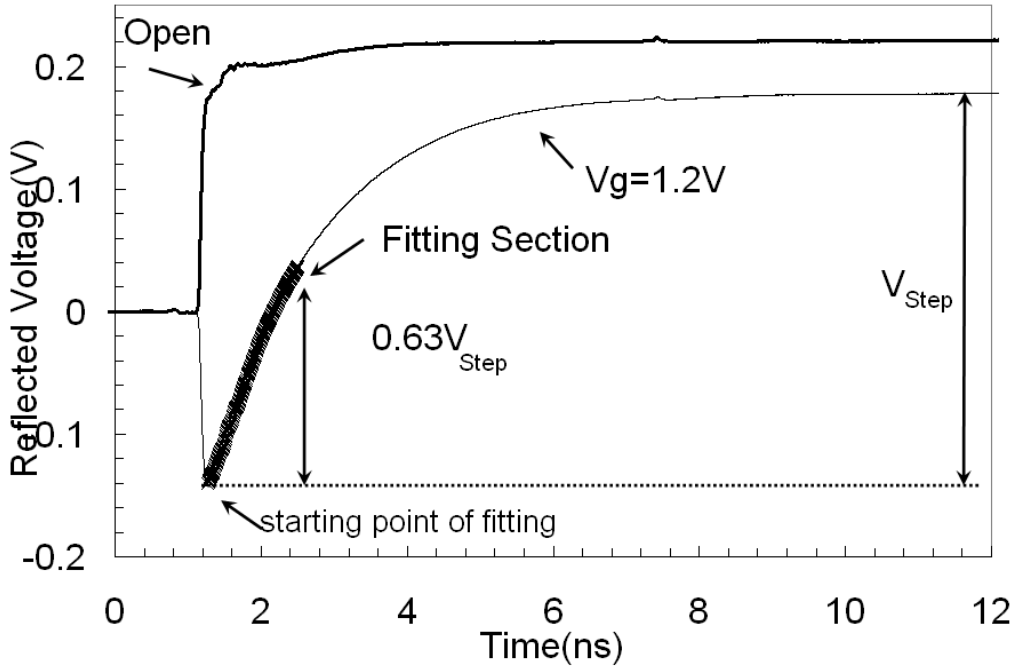


Figure 4.8 The fitting section is chosen from the reflected voltage of MOS capacitor at V_g=1.2V (strong inversion region). X is the fitting section. It starts from 90% of the incident open circuit waveform where the capacitor charging takes over the rise trend of step voltage. The section ends at the point with 63% charging.

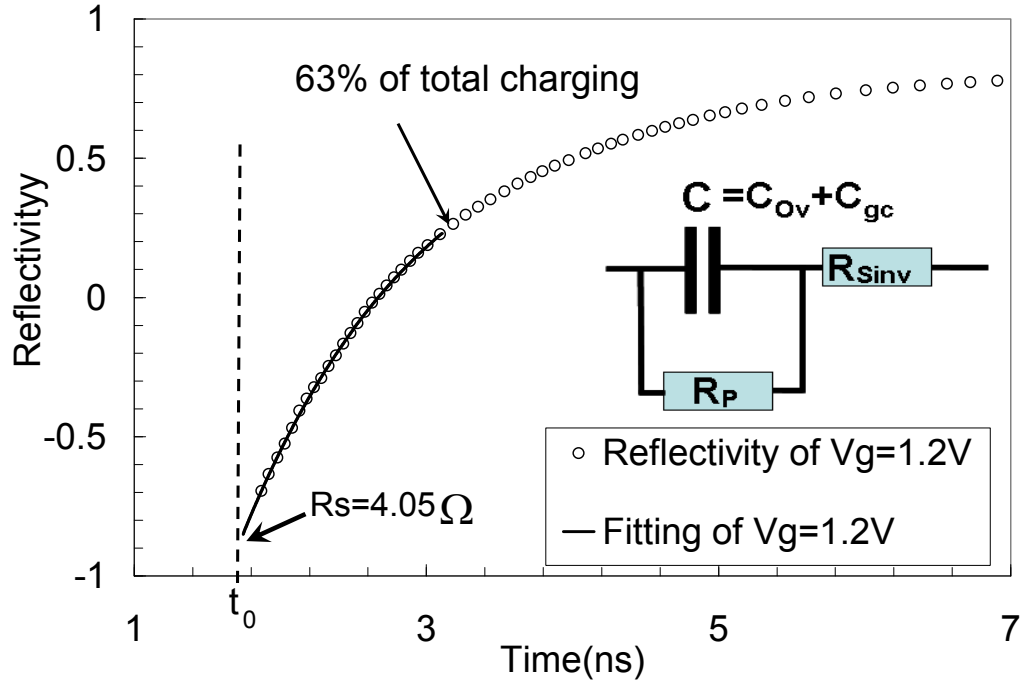


Figure 4.9 Extraction of series resistance of MOS capacitor at V_g=1.2V (strong inversion region). Insert: equivalent circuit for inversion case. The reflectivity extracted from TDR is plotted with

reduced density points to compare it with the fitting curve. R_S=4.05 Ω is extracted from the time zero. At that moment, the capacitor behaves like a short circuit and the impedance is R_S.

Figure 4.9 shows the fitting of reflectivity of MOS capacitor at V_G=1.2V (strong inversion). Obviously, it is a fairly good fit. From that, the series resistance R_{Sinv}=4.05 Ω can be extracted by extrapolating the fitting function to time zero as shown in Figure 4.9. It is also recognized that extracted R_{Sinv} is source/drain junction series resistance.

For the accumulation case such as V_G=-1.2V, the series resistance extraction is much more complicated because of the two capacitor charging process: both C_{gc} and C_{OV} get involved in the charging at any moment of time. This process is so complicated that the equation for modeling is too long to be used. An approximation is a must.

The initial part of the charging is dominated by an overlap capacitor. It can be approximated as a single capacitor charging and this part of data can be used as a fitting section. This fitting section is extremely close to the time zero and we can get high confidence of extraction because the data extension length is short for extrapolation. Figure 4.10 illustrated the fitting of reflectivity of MOS capacitor under V_G=-1.2V (strong accumulation) at short time region. Similar to the series resistance extraction procedure, we take time when the open circuit reaches its 90% of steady state voltage as the starting point of our fitting section. The end point is chosen as the time when C_{OV} loses its domination role and C_{gc} starts to take over the charging. This

transition point is determined as the time when the largest slope change happens in the reflectivity curve. As shown in Figure 4.10, 3.15Ω series resistance is extracted at accumulation case. By shorting the capacitance in the equivalent circuit in this case (insert of Figure 4.10), we can recognize that the extracted 3.15Ω series resistance is the parallel resistance of series resistance from substrate (R_{Sacc}) and source/drain junction (R_{Sinv}). Since R_{Sinv}=4.05Ω as extracted from inversion case, then R_{Sacc} itself will be 14.2Ω.

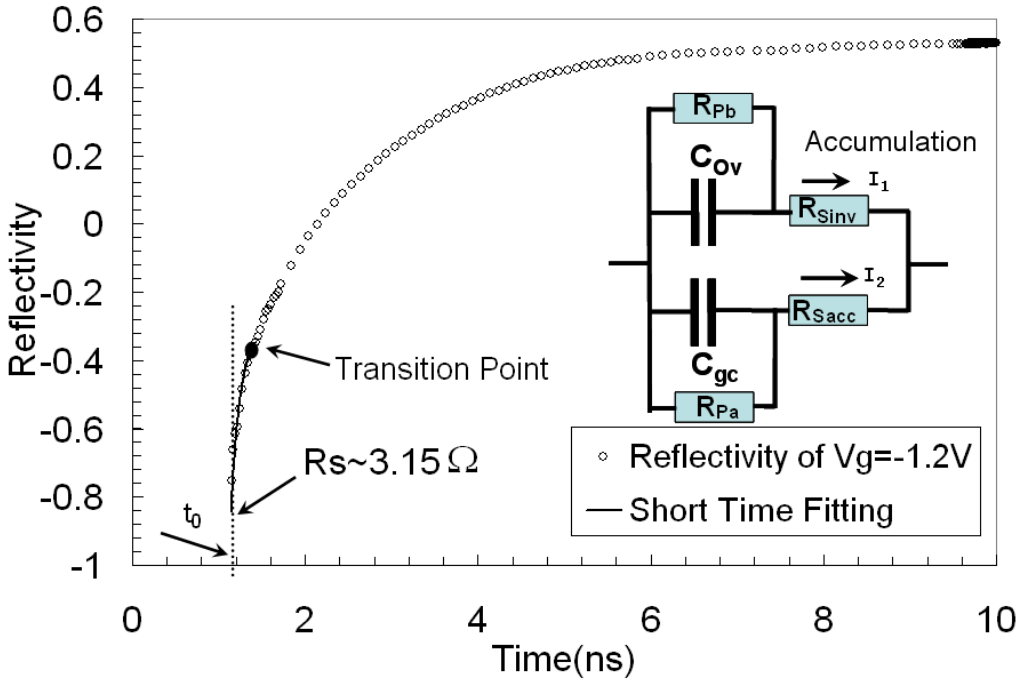


Figure 4.10 Extraction of series resistance of MOS capacitor at V_g=-1.2V (strong accumulation). Insert: equivalent circuit for accumulation short time situation. The reflectivity extracted from TDR is plotted with reduced density points to compare it with the fitting curve. Short time region of reflectivity curve is chosen to be fitted with exponential curve. *Rs*=3.15Ω is extracted from the time zero.

We can also apply the same extraction procedure to the rest of time domain behavior of 2 nm SiO_2 MOS capacitor. We use the reflected voltage waveform as we shown in

Figure 3.4. For convenience, we re-plot here with the same time scale as all the other figures used in this chapter.

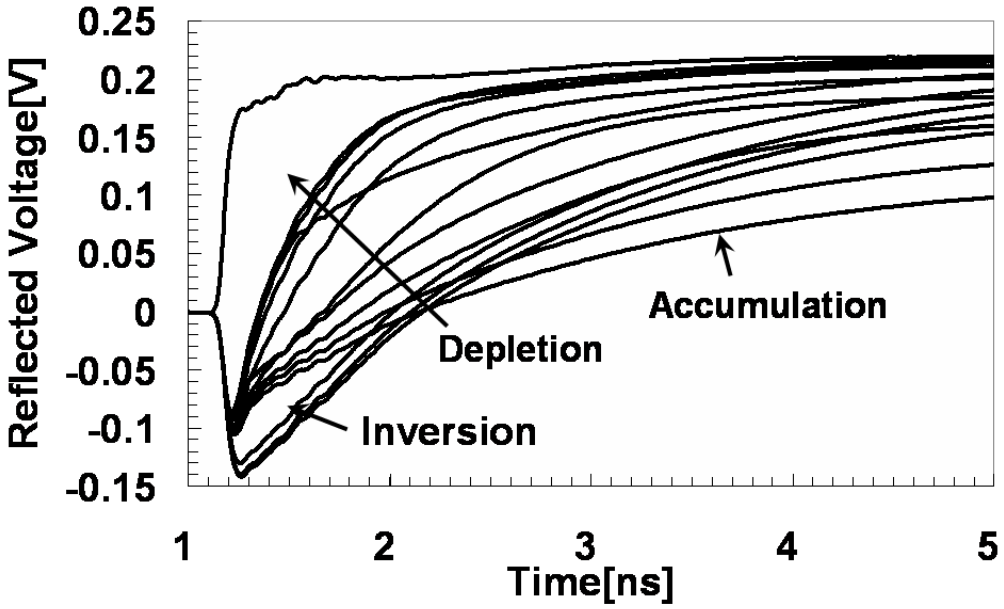


Figure 4.11, The as measured reflected voltage curve (charging curve) for all the bias conditions (accumulation, depletion and inversion). It is a re-plot of figure 3.5 with different time scale.

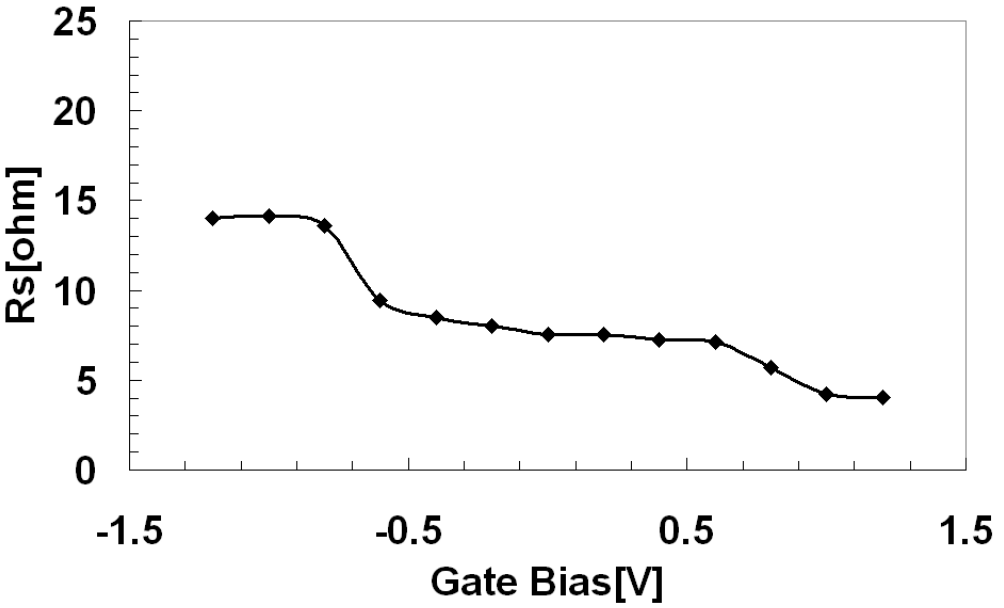


Figure 4.12 Extracted series resistance as the function of gate bias for the SiO_2 capacitor with TiN gate. In accumulation case, the substrate resistance is used as the series resistance.

Figure 4.12 shows the extracted series resistance as the function of gate bias. We use the series resistance extracted from long time region if there are two charging process involved. We can see that the series resistance is higher for accumulation than for inversion. This is expected because current must flow through the substrate for accumulation. In the case between accumulation and inversion, the series resistances have values also somewhere in between.

4.7. Time Zero Determination and Related Error

In the above series extraction procedure, it is noticed that a time zero is required. A natural question is how to determine time zero when the step function is not ideal? As a systematic approach, we use the steep rising edge of the step function and extrapolate it to the base of the function as time zero. The process is shown in Figure 4.13. It also indicates the error (shaded area) introduced by using this method to determine time zero. The shaded area represents the total charge already flowed into the capacitor at time zero as defined by our method. This is an error because the capacitor acts likes a short circuit only when it has not been charged to any degree.

To see how much time zero error does this represent, we need to keep in mind that we are seeking the time zero of an ideal step function. For the ideal step function and the associated reflection waveform, the total charge flowed into the capacitor per unit time is the difference between the magnitude of the two waveforms (dotted lines in

Figure 4.13). Thus the small amount of charge in the shaded area is translated into an extremely small time zero error. Fortunately, it is found that this error only introduces a small shift in time domain. It is even smaller than the timing jitter associated with the TDR instrument, which is about 10ps. From the slope of the reflectivity curve near time zero, we can estimate that even 10ps jitter will only contribute to 1% error in the extracted reflectivity.

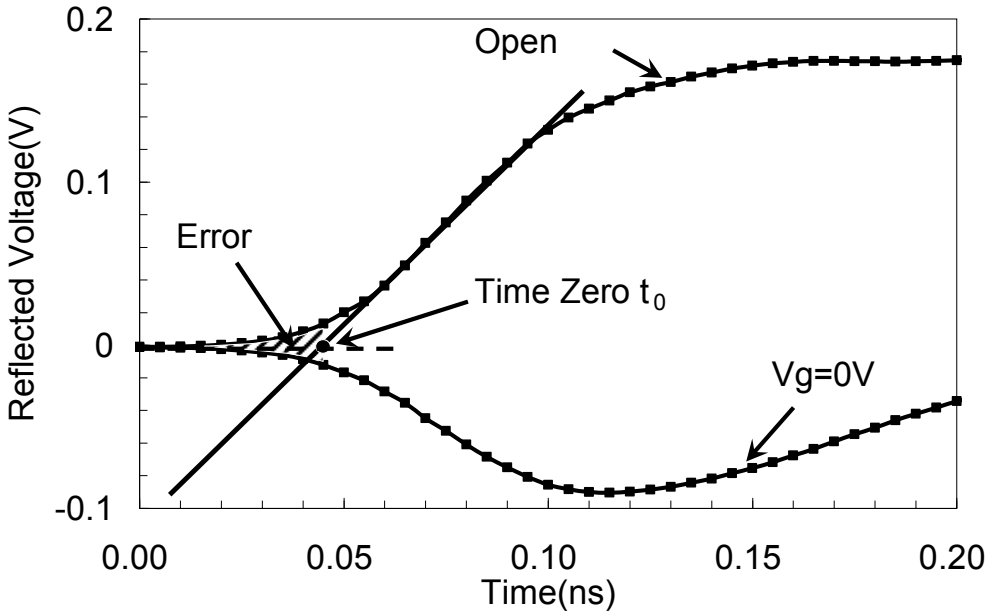


Figure 4.13 Reflected voltage curve expanded in time scale showing how time zero is determined. The shaded area represents charged already flowed into the capacitor at time zero and therefore becomes an error.

It is instructive to see how is 1% error in series resistance extraction affect the measurement of the capacitance using the TDR method. The error in capacitance extraction is about twice the $\Delta R_S/(R_S + R_P)$. Since R_P is larger than R_S in almost all cases, the effect on capacitance extraction accuracy is less than 1%.

4.8. Extraction of Shunt Resistance R_{Pa} and R_{Pb}

4.8.1 Basic Principle

Besides the series resistance, the overlap capacitance can be extracted with high precision as well. We have known the equivalent circuit with overlap capacitance (re-plotted in Figure 4.14). Since the circuit can be simulated, overlap capacitance can be extracted by finding the best simulation to fit experimental result. To get better confidence, it is better to know all the other components first. In section 4.6-4.8, we have extracted the series resistance as R_{Sinv}=4.05 Ω and R_{Siacc}=14.2 Ω. We are still missing the shunt resistance of channel region R_{Pa} and overlap region $R_{Pb.}$

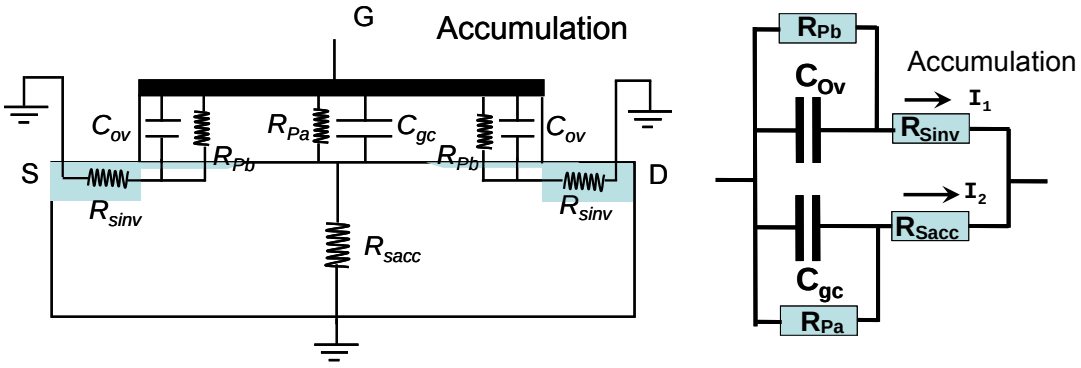


Figure 4.14 Equivalent circuit of MOS Device at accumulation.

During this calculation, the oxide field difference at overlap and channel region must be taken care because the leakage current is very oxide field dependent. Figure 4.15(a) (b) illustrates the band diagram of both regions. At overlap region, any additional carrier injection at the gate can hardly alter the surface potential because the heavily

doped (10^{20} cm^{-3}) source/drain junction behaves like a metal. The Fermi-level is pinned at the top of valence band E_V and the full applied gate voltage (-1.2V in this case) is applied to oxide. On the other hand, at the gate to channel region, the band bends at surface with gate bias. With -0.65V flat band voltage known from the C-V curve, there is only -0.55V dropped across the oxide with -1.2V gate bias.

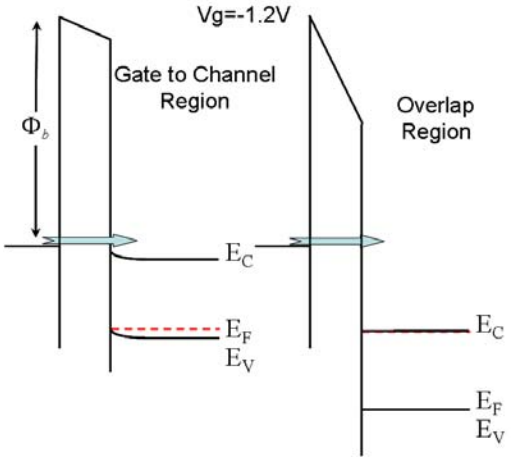


Figure 4.15 the band diagram of test MOS capacitor with Vg=-1.2V (strong accumulation) at two regions: channel and overlap region. Full gate bias is dropped at the overlap region due to the high doping source/drain junction. With -0.65V flat band voltage known from the C-V curve, there is only -0.55V dropped across the oxide at gate to channel area.

The oxide field difference sheds a light on a possible way to separate the current from these two regions and facilitates the solution. Since the leakage current exponentially increases with oxide field, it is possible that leakage current from one region is much higher than the other. It is possible for us to identify a range in leakage current as gate voltage (I_g-V_g) curve where the current is completely dominated by the overlap region. Then we can fit this part of current with a known tunneling function [75] and then know the current from channel region by first extrapolating the fitting and subtracting it from the total current. Finally, we can pin down the values of R_{Pa} and $R_{Pb.}$.

4.8.2 Tunneling Current Model

To do the fitting, we need a known tunneling function. We take the equation of direct tunneling current from K. F. Schuegraf and C. Hu's paper [21] as:

$$J_{DT} = C\left(\frac{V_{OX}}{t_{OX}}\right)^2 \exp\left(\frac{-B\left(1-\left(1-V_{OX}/\Phi_B\right)^{3/2}\right)}{V_{OX}/t_{OX}}\right) \tag{4.5}$$

Where V_{OX} is the voltage dropped across the oxide. t_{OX} is oxide thickness and Φ_B is the barrier height at metal gate side(as shown in Figure 4.15). B is the function of barrier height Φ_B as:

$$B = \frac{8\pi\sqrt{2m_{OX}}\Phi_B^{3/2}}{3\hbar q} \tag{4.6}$$

The equation (4.5) is simple to use but not adequate. It is shown that the simulated gate current based on this equation does not approach as V_g goes to zero. In addition, it does not fit the experimental very well in the sub-1V region [75-76]. The authors further improve their model in the later paper [76] and point out the WKB approximation might not hold very well for the ultra-thin oxide. A new and better model is proposed and the equation (4.5) is modified as:

$$J_g = \frac{q^3}{8\pi h\Phi_B\varepsilon_{OX}} \cdot C(V_g, V_{OX}, t_{OX}, \Phi_B) \cdot \exp\left(\frac{-8\pi \cdot t_{OX}\sqrt{2m_{OX}}\Phi_B^{3/2}}{3hq|V_{OX}|}\left[1-\left(1-\frac{|V_{OX}|}{t_{OX}}\right)^{3/2}\right]\right) \tag{4.7}$$

With

$$C(V_g, V_{OX}, t_{OX}, \Phi_B) = \exp\left[\frac{20}{\Phi_B}\cdot\left(\frac{|V_{OX}|-\Phi_B}{\Phi_{B0}}+1\right)^{\alpha}\cdot\left(1-\frac{|V_{OX}|}{\Phi_B}\right)\right]\cdot\left(\frac{V_g}{t_{OX}}\right)\cdot N \tag{4.8}$$

Equation (4.7) and (4.8) is rather complicated to use. Many additional equation is required to know the fitting parameter as α, N. Most important of all, the situation in their model is quite different from ours. One of the main reasons for proposing this complicated model is to take account of the quantization effect in the semiconductor. This quantization is created when band bends. For our case in the overlap region with source/drain junction is highly doped, there is hardly any band bending and no quantization effect. Obviously, the equation 4.7 and 4.8 is not quite applicable to our case.

Then we go back to equation 4.5. The main reason that it fails in simulation is the usage of WKB approximation, which is questionable when the oxide gets so thin. But the basic form of this equation is still correct as compared to equation 4.7. We can take this form as:

$$J_{DT} = C\left(\frac{V_{OX}}{t_{OX}}\right)^{\eta} \exp\left(\frac{-B\left(1-\left(1-V_{OX}/\Phi_B\right)^{3/2}\right)}{V_{OX}/t_{OX}}\right) \tag{4.9}$$

But the parameter needs to be revised. The previous definition of *B* as the function of the barrier height Φ_B may not hold. The simulation of gate leakage current for ultrathin oxide is not an easy task, especially at low gate bias region. Instead of struggling with the exact solution, we set all the parameter *B, C,* Φ_B,t_{OX} as fitting variable parameters and let the fitting decide the best one set. Moreover, in the overlap region with highly doped source/drain, the voltage across the oxide V_{OX} will be equal to applied external gate bias.

4.8.3 Determination of Shunt Resistance with Fitting

Besides the tunneling function, we still need to know which part of data is best for fitting. From the band diagram, gate bias ranging from 0V to flat band voltage (-0.65V) is the best choice. Under that condition, the channel is under depletion and most of the applied gate bias is on the bend bending. Leakage current from the channel is very negligible since very little voltage drops across the oxide. Leakage current is mostly dominated from overlap region because full gate voltage drops across from the overlap region.

It is reasonable for us to take an approximation that the leakage current in this gate bias range is from the overlap region alone. Figure 4.16 shows the total gate leakage current for 2 nm SiO_2, 1400 μm^2 MOS capacitor. With our assumption, we can fit the measured leakage current at gate bias between 0V to -0.65V with the tunneling function.

Figure 4.16 shows the fitting result. Obviously, it is a pretty good fit. From the fitting, we can know the function of leakage current component from overlap region. Then after subtracting from the experimental measured total leakage current, we can extract the leakage current component from gate to channel region alone as illustrated in Figure 4.16. Then, we can convert it to the shunt resistance by diving the gate bias over the leakage current as $R_p=V_g/I_g$. At the condition of strong accumulation with

V_G=-1.2V, the shunt resistance of overlap and gate to channel region will be R_{Pa}=192 Ω and R_{Pb}=240 Ω respectively.

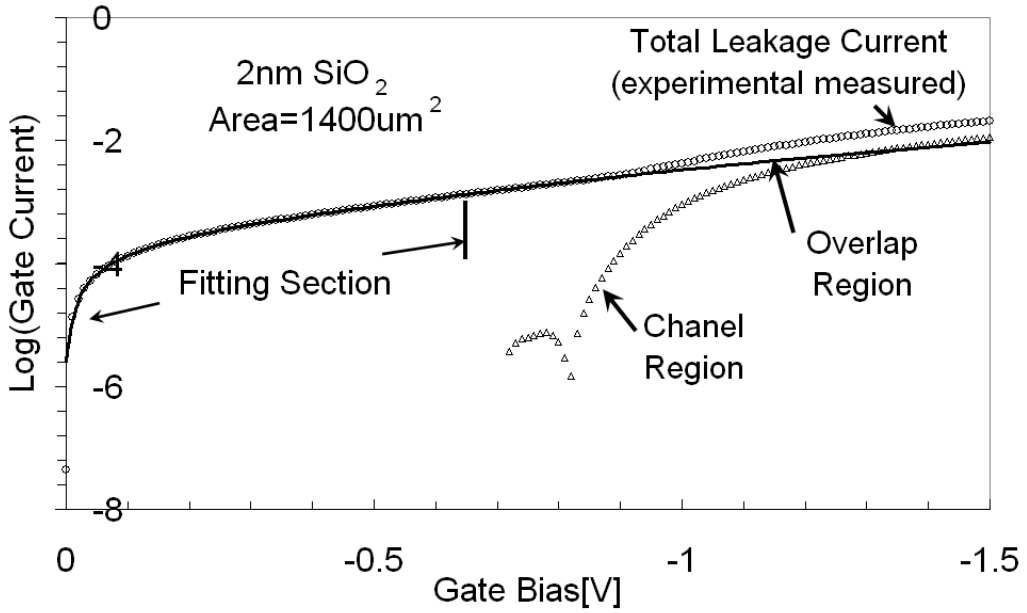


Figure 4.16 Fitting of measured leakage current with known tunneling function. The leakage current within fitting section is assumed to come from overlap region only. The extrapolation of fitted curve allows us to extract the leakage current component from overlap and channel region at accumulation (V_G=-1.2V).

4.9. Better Calibration Structure

From the previous sections, we have extracted all the resistive components. In order to extract the overlap capacitance, we still need to convert the measured reflected voltage $V_{Ref}(t)$ into reflectivity. Actually, the measured reflected voltage waveform $V_{Ref}(t)$ is the product of the incident step waveform $V_{Inc}(t)$ and the reflectivity $\rho(t)$ as a function of time. Therefore, we can get the reflectivity by dividing the reflected waveform with the incident waveform as:

$$\rho(t) = \frac{V_{Ref}(t)}{V_{Inc}(t)} \tag{4.10}$$

The incident waveform $V_{Inc}(t)$ can be obtained from the response of calibration structure, which ideally should be the structure including all the other parasitic except the device under test. For convenience, open circuit can be considered. It is done by lifting the probe up and includes all the parasitic up to the probe tip.

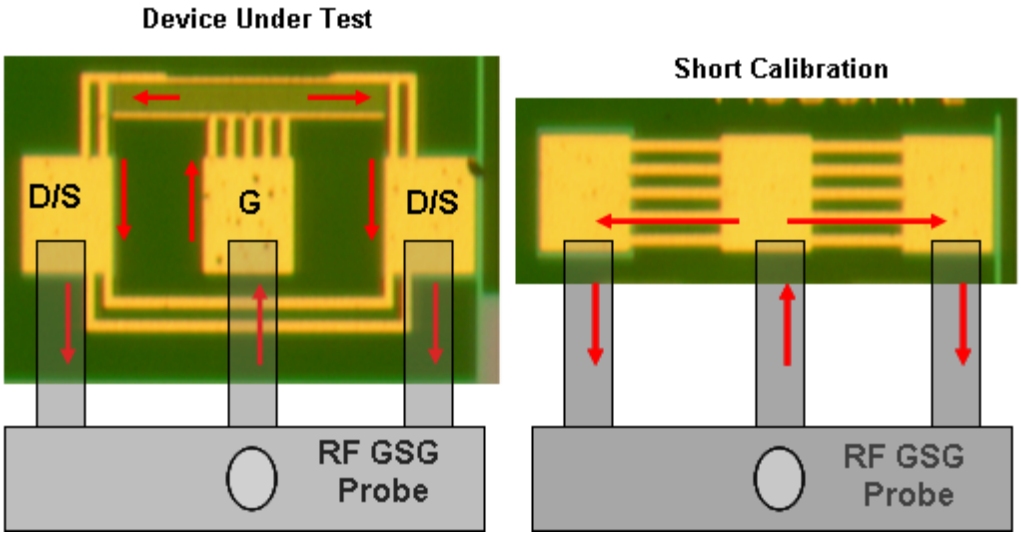


Figure 4.17 Picture of device under test(left) and short calibration structure(right). The arrow represents for the current path during the test. Obviously, there is an addition length for the current to flow through in the device under test.

However, as shown in Figure 4.17, besides the capacitor under test, the test structure also has the probe pad and small extension length. These additional parasitic do not belong to the device under test (DUT) but are not included by open calibration structure. They only affect very high frequency parts of data or very beginning part of data in time domain. It might not be critical to our previous C-V measurement or series resistance extraction, which is either start fitting after the rise part or more focus on the long term overall part of data. It will become a problem for our overlap

capacitance extraction, which is mainly focus on the fitting of initial part of capacitor charging with the need of as many data points as possible.

An improvement can be made by using the short calibration structure as shown in the right of Figure 4.17. It takes It is better than the open circuit with taking account of the probe contact and probe pad. However, it is still not good enough. As shown in the red arrow in Figure 4.17 which is the direction of current flow, there is an additional length in the device (left) compared to the short calibration structure (right). It is well known that any length in the TDR will result in the time shift between DUT reflected voltage and reference which is from calibration structure. Any misalignment in data points greatly affects the reflectivity, especially the initial part.

Therefore, an accurate calibration structure is extremely important for the high frequency test such as TDR. Since there is not any better calibration structure available on the wafer provided by Sematech, we have to make our own. We can take one sample of DUT and intentionally hard break down by stressing sufficiently high voltage. The hard breakdown creates a short path at the two electrodes of the capacitor. This “broken” device can be taken as our calibration structure. Since this calibration structure is modified from DUT, it preserves all the features except the short at two electrodes.

To be careful, we should make sure this hard breakdown offers good short. We can

confirm it by measuring the I-V curve (solid line in Figure 4.18) and extracting the resistance as 4.5 Ω. For the short calibration structure itself (Figure 4.17) which contains probe tip, contact pad and the rest of system, the 1.6 Ω resistance is extracted from the I-V(dashed line in Figure 4.18). Therefore, there is an additional 2.9 Ω for our "created short", which is very possible the series resistance. It also matches our previous extracted value as 3.15 Ω. It confirms that this "created short" does not introduce any additional resistance and is indeed a good short. The evidence of matched result from I-V and TDR also in turn confirms the correctness of our series resistance extraction.

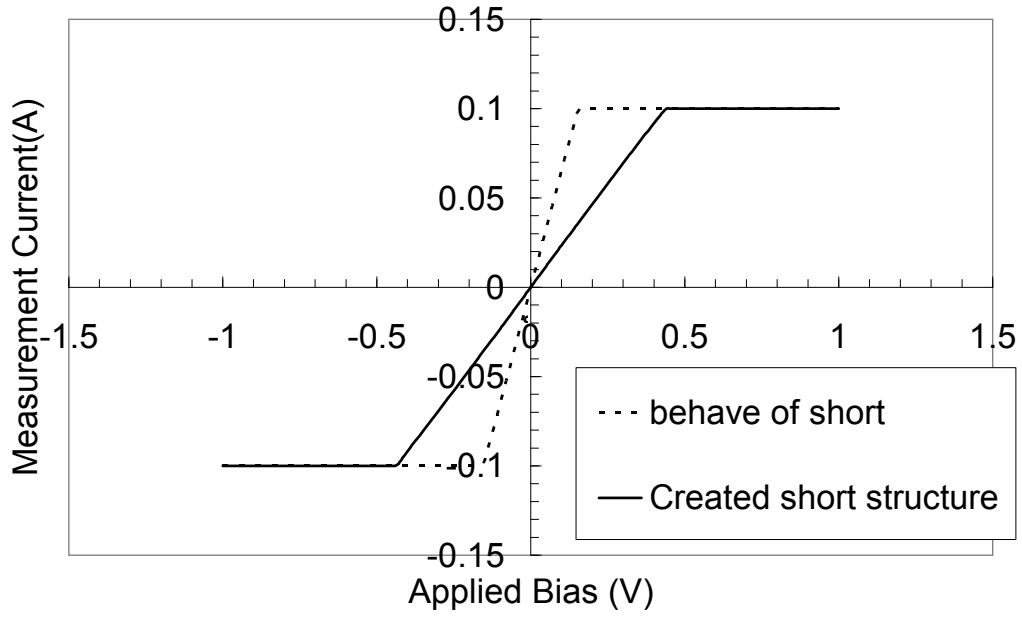


Figure 4.18 I-V of short calibration structure and "created short" calibration structure by internationally hard break down DUT.

The solid line in Figure 4.19 shows the measured the reflected waveform of this "created short" calibration structure. The step down behavior is because the reflection of the short is the mirror of incident voltage. To be used as a reference, we need to flip

it and result is shown as the dashed line in Figure 4.19.

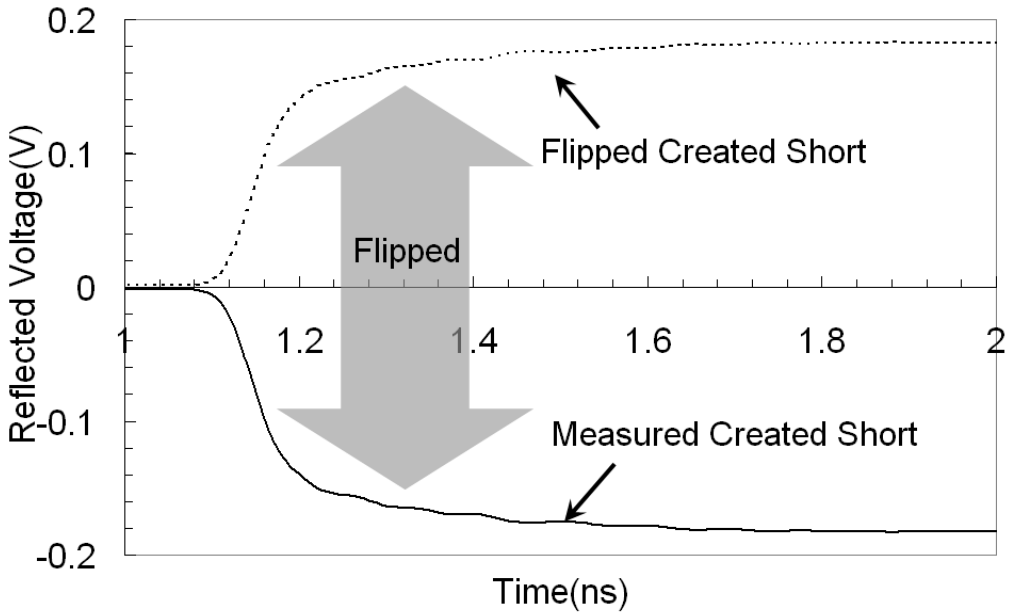


Figure 4.19 The measured Reflected waveform of created short (solid line). It needs to be flipped first (dashed line) and then can be used as reference.

Taking it as reference, we can get the reflectivity by dividing reflected waveform of DUT on it. The result is shown in Figure 4.20. For comparison, the reflectivity using open or short as calibration structure is also included. Obviously, the result from different calibration structure is quite different at initial time part of data. The result of "created short" is more accurate because it includes all of the parasitic while open or short does not.

Moreover, it is interesting to observe a big bump very close to the time zero t_0 - the time point when the step first arrives. The bump in TDR represents an inductor. Opposite to capacitor, inductor behaves like an open circuit or impedance increase when the fast step voltage arrives. After carefully examining our DUT, we found that there is indeed an inductance. It arises from the current loop flowing from the center

probe to the other two as illustrated in Figure 4.17. This inductance exists both in the calibration structure and DUT. In principle, it should be cancelled out by the division. However, the action of flip before division causes the problem. Instead of cancellation, the inductance behavior gets enhanced after flip then division. Obviously, this is not what we wanted and correction procedure is required. It is very important because it messes up the initial part of the data and makes the fitting extremely hard.

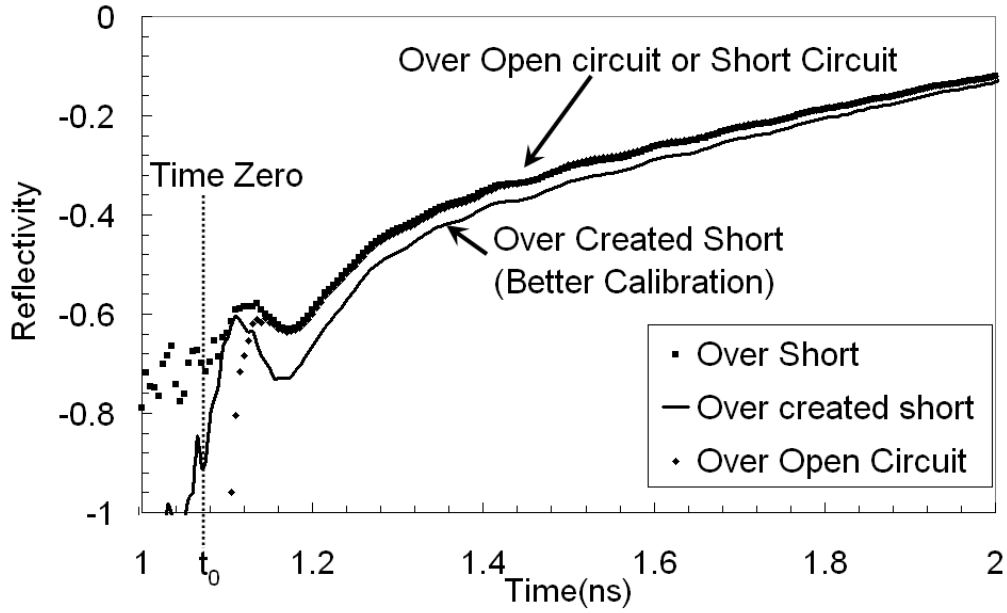


Figure 4.20 The reflectivity obtained by dividing the measured reflected voltage of DUT over the one of flipped created short. The reflectivity extracted when the open or short circuit as calibration structure is also included for comparison.

To fix that, we can first find out the additional impedance coming from the inductance and then subtract it from the impedance of DUT. Here are the basic procedures:

1. Find out the reflectivity of created short calibration by dividing measured reflected voltage of created short over the open circuit.

2. Convert the reflectivity to impedance as shown in Figure 4.21.
3. Get the inductance by subtracting the initial sudden increase with the impedance at steady state. The inductance does not play much role to open circuit due to its sufficient large impedance. However, this inductance becomes significant for short due to its extremely small impedance. By dividing short over open, the inductance behavior is preserved and behaves itself as the sudden increase of impedance at the beginning of time scale.
4. Convert the reflectivity of DUT into impedance
5. Subtract the inductance impedance from the one of DUT
6. Covert the corrected DUT impedance back to reflectivity as shown in 4.22.

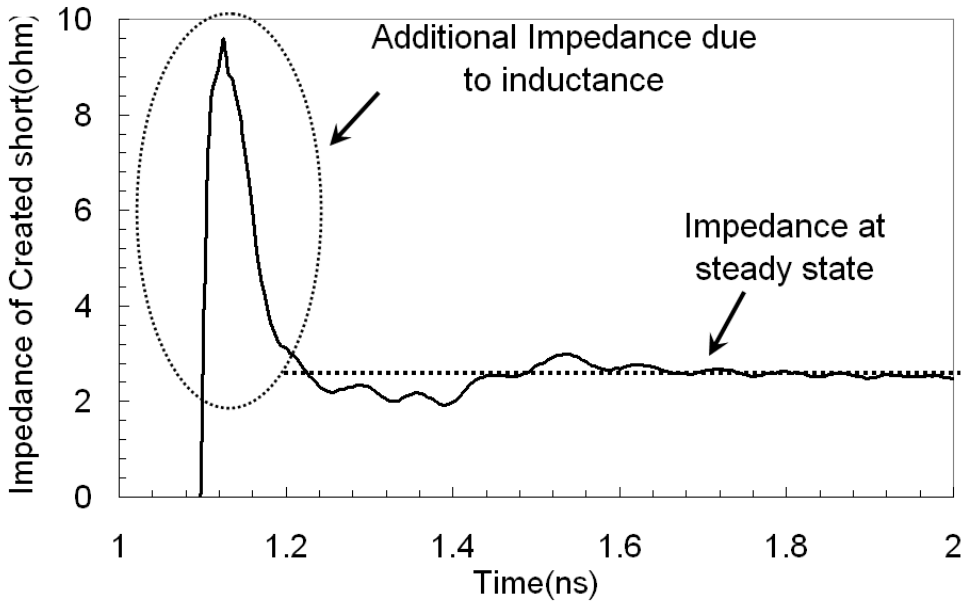


Figure 4.21 Impedance of created short in time domain. The peak in this impedance comes from the inductance. It can be used to subtract from the impedance of DUT and then correct the effect of parasitic inductance.

Figure 4.22 shows the reflectivity of DUT after corrected with inductance. The disappearance of bump indicates the success of correction procedure. This reflectivity

is extracted with a better calibration structure and free of the inductance. As a result, it is the most accurate and becomes the one used to do the fitting and then extraction of the overlap capacitance.

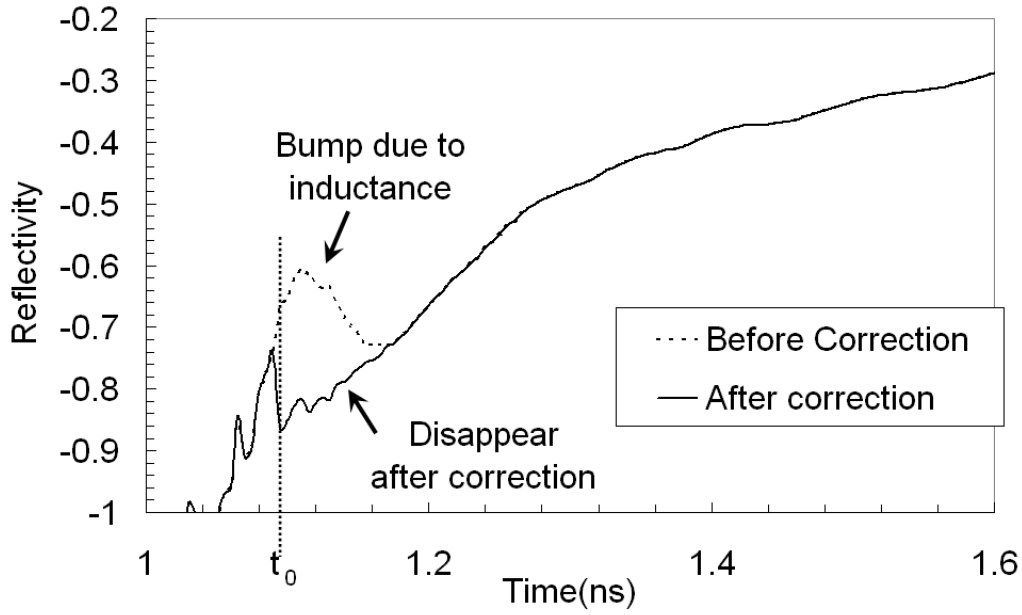


Figure 4.22 The reflectivity of DUT after corrected with inductance. Obviously, the disappearance of bump indicates the success of correction procedure.

4.10. Overlap Capacitance Extraction

In order to extract the overlap capacitance by fitting, we still need a program that can simulate the reflectivity. With the equivalent circuit, we can calculate the reflectivity by impedance from circuit. The exact analytical expression of reflectivity can be worked out by inverse Laplace transform. From equation (A.11), the detected voltage by TDR will be:

$$V_{TDR}(t) = L^{-1}\left[\frac{V_{Step}}{s}\left(\frac{Z_L - Z_0}{Z_L + Z_0}\right)e^{\frac{2ls}{v_p}}\right] + V_{Step}U(t) \tag{4.11}$$

Therefore, the reflectivity will be:

$$\rho(t) = \frac{V_{Ref}(t)}{V_{Step}} = L^{-1}\left[\frac{1}{s}\rho(s)\right] = L^{-1}\left[\frac{1}{s}\left(\frac{Z_L(s) - Z_0}{Z_L(s) + Z_0}\right)\right] \tag{4.12}$$

Equation (4.12) tells us that we can obtain the reflectivity at time domain by inverse Laplace transformation of load impedance and incident signal. For this case, the load impedance for the equivalent circuit in our case for the equivalent shown is:

$$Z_L(s) = \frac{1}{\dfrac{1}{R_{s,acc} + \dfrac{R_{Pa}}{1 + s \cdot R_{Pa} \cdot C_{gc}}} + \dfrac{1}{R_{s,Inv} + \dfrac{R_{Pb}}{1 + s \cdot R_{Pb} \cdot C_{ov}}}} \tag{4.13}$$

The idea to obtain the reflectivity is very straightforward. However, the calculation is a little bit complicated and can not be accomplished by hand. We turn to mathematic software and make a small program to do that. The codes are:

```
h[s_]:=1/s;                                  //incident step voltage
m[s_]:=1/(Rsinv+Rpb/(1+Rpb· Cov· s));        //Impedance of overlap
                                               capacitance path
n[s_]:=1/(Rsacc+Rpa/(1+ Rpa · Cgc· s));      //impedance of gate to channel
                                               capacitance path
g[s_]:=1/(n[s]+m[s]);                        //load impedance is the parallel
                                               one of above two
l[s_]:=(g[s]-50)/(g[s]+50);                  //Convert load impedance to
                                               reflectivity in frequency
                                               domain
y[s_]:=h[s] · l[s];                          //output is load impedance with
                                               incident signal
y[s]
InverseLaplaceTransform[y[s],s,t]            //convert to reflectivity in time
                                               domain
```

Here is the outcome expression for reflectivity:

To illustrate the basic behavior of this complicated function, we put some number in. Here we use $R_{s,acc}$=R_b=14.2Ω, $R_{s,inv}$=R_a=4.05Ω , C_{OV}=2pF C_{gc}=20pF. The result will be reduced to:

$$\rho(t) = 1 - 0.271 \cdot e^{(-0.032t)} - 1.589 \cdot e^{(-0.00073t)} \tag{4.14}$$

Obviously, it shows that there are two exponential function involved. The one with small time constant rises up fast at beginning and then is overwhelmed by the other one with longer time constant. It matches the two capacitor charging picture. It is indeed observed in the simulation as shown in Figure 4.23. By comparison with the behavior of single capacitor charging, the long term behavior is more close to the behavior of single 22pF capacitor. That is expected as well because two capacitors is charging at the same rate on long term and the equivalent capacitance is the sum of these two (20pF+2pF=22pF).

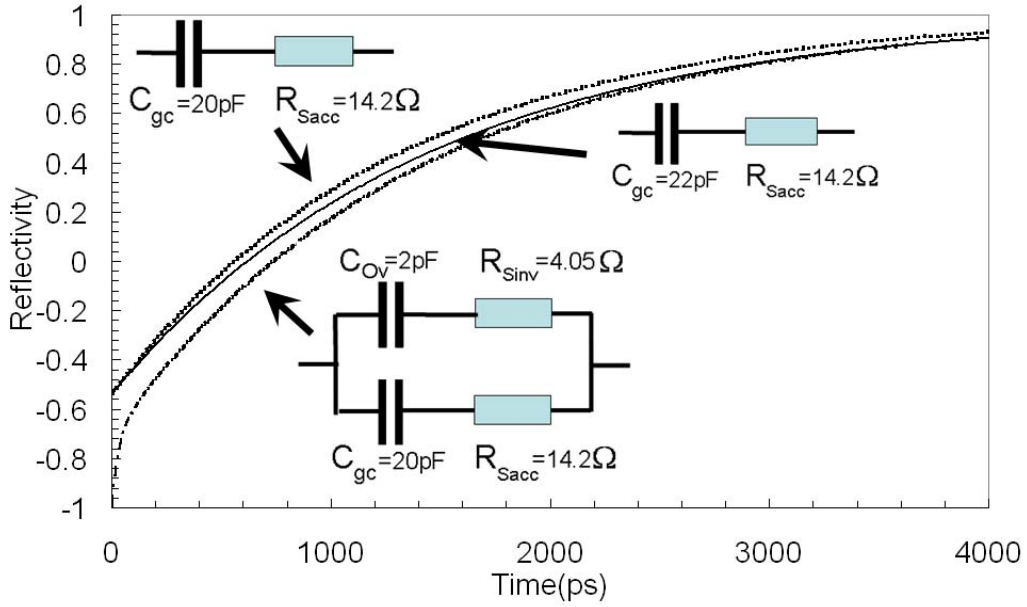


Figure 4.23 Simulation of reflectivity of equivalent circuit with $R_{s,ac}$=14.2Ω, $R_{s,inv}$ =4.05Ω , C_{OV} =2pF C_{gc}=20pF. For comparison, a single 20pF or 22pF capacitor charging is also shown. On the long term, the two capacitors charge at the same rate and close to a 22pF capacitor (sum of those two).

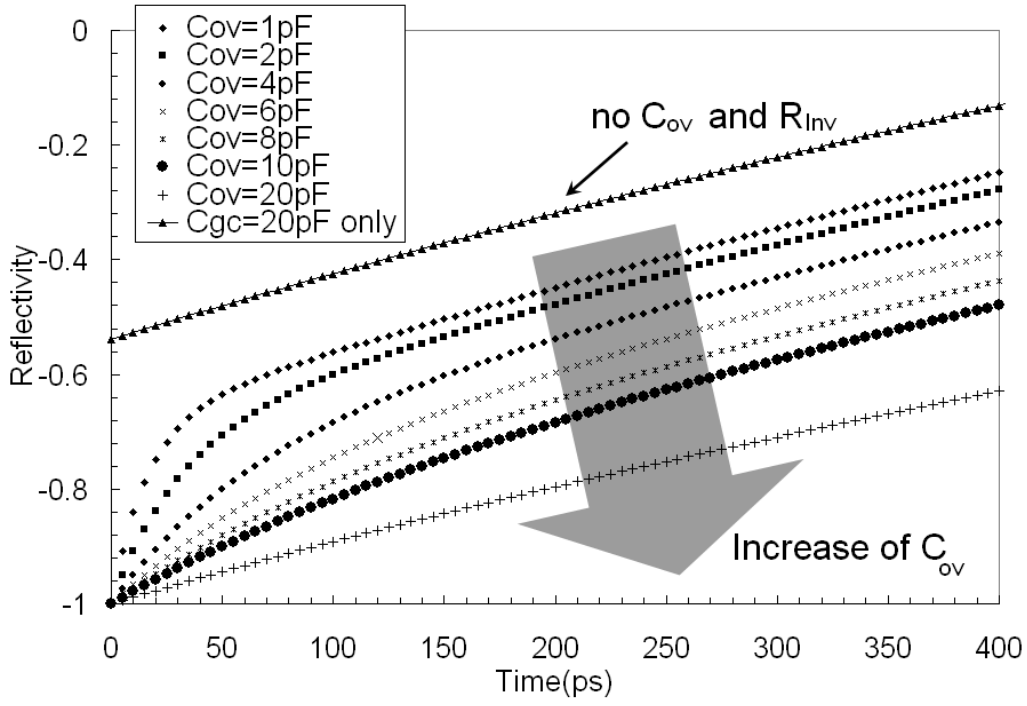


Figure 4.24 Simulation of reflectivity with different overlap capacitance. Smaller overlap capacitance has more distinct charging time with channel capacitance. Therefore, shaper transition is observed.

This simulation program also enables us to study the influence of overlap capacitance on reflectivity. Figure 4.25 shows the simulation of reflectivity with different overlap capacitance. The case without any overlap capacitance is also shown. The overlap capacitance dramatically affects the initial part of the MOS device charging. With smaller overlap capacitance, the charging time between two capacitors is more distinct and the transition between two charging process is sharper. This data also suggest us that we can extract the overlap capacitance by comparing the simulation result with experimental observed reflectivity.

With all the efforts made in previous sections, we are ready to extract the overlap capacitance. First review what we have:

- Total capacitance at -1.2V from C-V characteristics of chapter 3 as $C_T=C_{gc}+C_{OV}$=22pF.
- Series resistane R_{SInv}=4.05Ω , R_{Sacc}=14.2 Ω (from section 4.6)
- Shunt resistance R_{Pa}=192Ω , R_{Pb}=240 Ω (from section 4.8)
- Improved experimental reflectivity of DUT at V_G=-1.2V (from section 4.9)
- An available simulation program able to calculate the reflectivity once all the parameter of the equivalent is known. (This beginning of this section)

With all these available parameters, there is only one unknown: the area ratio of overlap to channel region $f=C_{ov}/C_{gc}$. To extract f and therefore the overlap capacitance, we simply find the best fit to the reflectivity curve extract from the measured reflection waveform. Figure 4.25 shows the resulting best fit of reflectivity at strong

accumulation (*Vg*=-1.2V) with *f=0.142*. In addition to the excellent fit, the result for fitting to a single capacitor is also shown. Clearly, the two capacitors charging process is necessary to explain the experimental data.

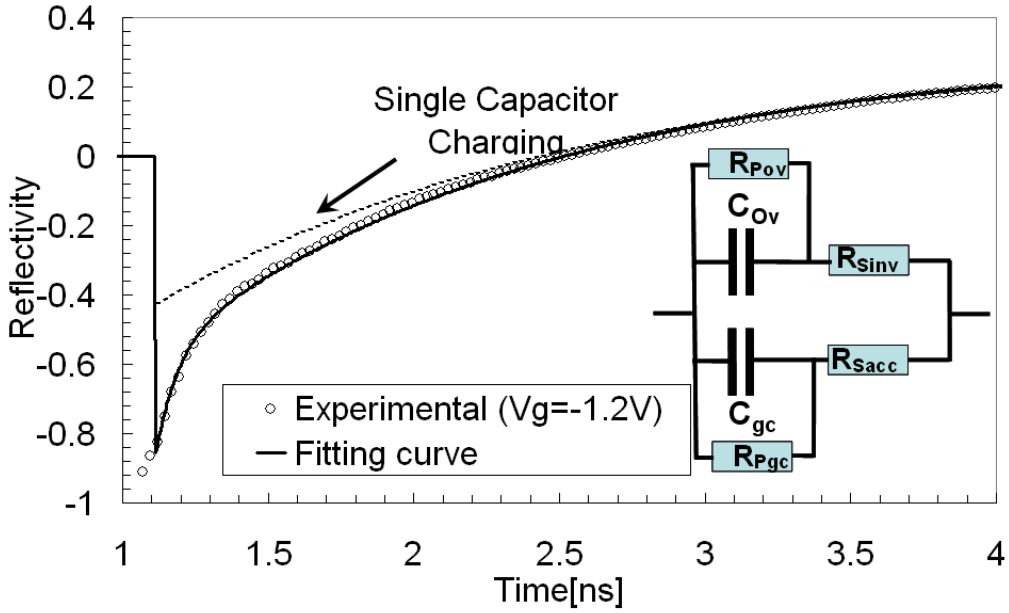


Figure 4.25 The comparison of the fitting curve with experimental reflectivity curve of the capacitor at V_G=-1.2V. A simulation with a single capacitor charging is also included to show that it provides a poor fit. Insert: the equivalent circuit of the oxide capacitor with overlap capacitor.

With 23 pF total capacitance found in chapter 3 under this test condition, the overlap capacitance works out to be 3.266pF. We can now remove the overlap capacitance from the measured C-V curve of the capacitor in Figure 3.9 from chapter 3 to obtain the channel capacitances. The result is shown in Figure 4.26. It should be emphasized that the error is not due to poor test structure design. A number of factors constraints the design space [14, 16]. It is not obvious that significant reduction of the overlap capacitance is possible.

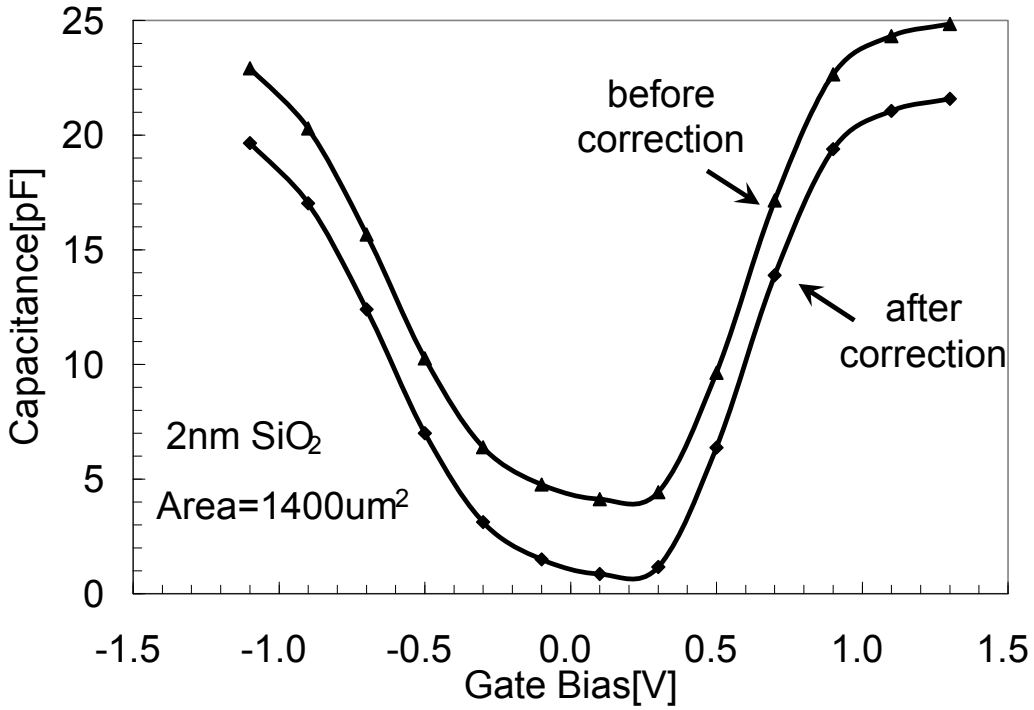


Figure 4.26 The measured C-V curve for the 2 nm SiO_2 capacitor with 1400 μm^2 area (upper curve) and the corrected C-V curve (lower curve) after the removal of the overlap capacitance.

4.11. Accuracy of Overlap Capacitance Extraction

In overlap capacitance extraction, we perform fitting of the equivalent circuit to the measured reflectivity data. Since we independently determine 5 out of the six parameters of the equivalent circuit and the accuracy of their determination has already been discussed above, we assume them to be accurate. In other words, in this analysis we ignore error propagation and focus on accessing the accuracy of the area ratio extraction that has direct bearing on the accuracy of the overlap capacitance extraction. The main task is therefore to determine how well can the fitting process pin down the value of the area ratio *f*. The goodness of fit (R^2 square value) is a good tool for this purpose.

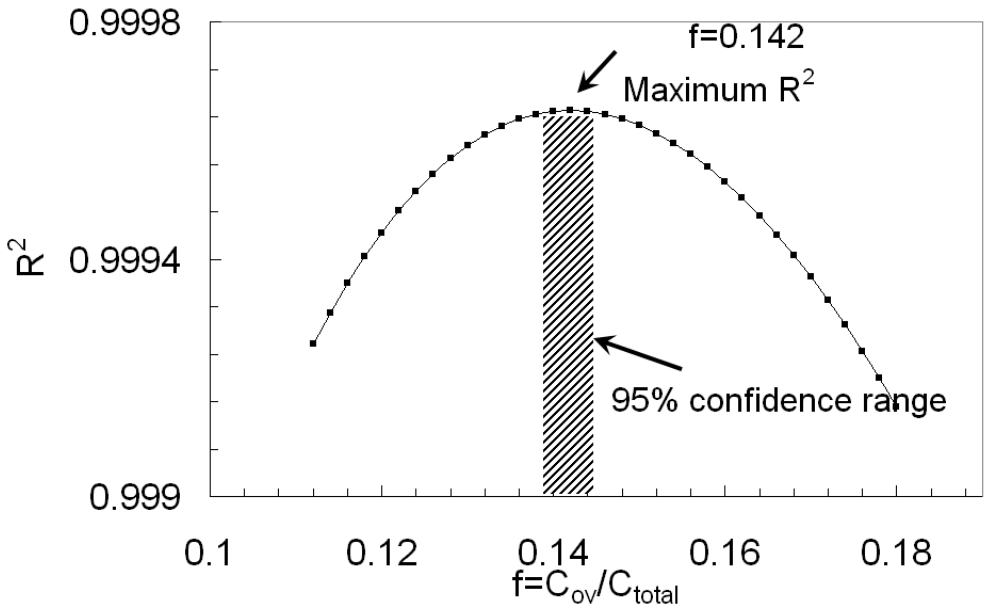


Figure 4.27 The R^2 value of the fitting with different ratio of overlap capacitance to total. The best fit is at maximum R^2 condition with f = 0.142. The error is the range of f that is less than 95% R^2 confidence interval.

The R^2 square value for different f is plotted in Figure 4.27. The best fit (maximum R^2) is f = 0.142. The 95% confidence value is f =0.142±0.003, or 2%. The uncertainty (95% confident) will of course depend on each capacitor design. We expect that the uncertainty will decrease with larger value of f. While it is more desirable to design transistors with lower f, the trend in practical reality is the opposite. Thus we expect the overlap capacitance extraction to remain highly accurate as CMOS technology continues to advance. It is instructive to note that the 2% error is referring to the area ratio, not the actual overlap capacitance. The total capacitance can be determined to better than 1%. The small 2% error in separating the channel capacitance and overlap capacitance essentially leaves the determination of the overlap capacitance with the same (percentage) accuracy as the total capacitance.

4.12. Conclusion

In this chapter, we described a highly accurate method of independently extracting series resistance from TDR measurement. The ability to measure the series resistance accurately at inversion raises the possibility of measuring the series resistance of a transistor - a long standing challenge. If this could be done, the transistor effective channel length measurement will be greatly simplified. The hardware used in our experiment requires at least 1 pF to perform the reliable measurement. With the best equipment available today, the capacitance can be 5 times smaller. However, even 0.2pF is very large for transistor. Only transistor with multi-finger gate can have such large capacitance. In future, if such test structure is available, we can try to extend the application of this new simple method to series resistance extraction in transistor.

We also show that by introducing source and drain to the MOS capacitor to support the extraction of inversion capacitance, the overlap capacitance can contribute to significant error to the measured capacitance, particularly in the depletion region. We have demonstrated that overlap capacitance can be extracted by TDR in a very simple and accurate manner. Finally, very accurate capacitance measurement can be done by combining the extracted series resistance and overlap capacitance with the TDR method of capacitance extraction.

Chapter 5

Frequency Dependent Charge Pumping, How deep it probes

We have introduced a new time domain reflectrometry C-V method in the last two chapters. It only solves one of many challenges in the advanced MOS device characterization. The appropriate evaluation of device reliability is another. It requires a good knowledge of the traps on the interface or in the bulk oxide, such as how the traps are distributed and how many of them. This demand becomes more urgent especially when high κ dielectrics have been recently used as the gate oxide in the 45nm CMOS technology. These new dielectrics from transitional metals are full of defects spreading not only at the interface but also in the oxide bulk. In addition, there are also a lot of defects on the buffer SiO_2, which is grown between high κ and Si-substrate to improve the interface quality. Therefore, for the reliability concern, it is important to find out which part (high κ or buffer SiO_2) initiates the failure. This is a very important question because it points out the focus of future improvement [20-22]. A good way to approach the answer is to stress the device and watch the increase of trap density. The technique to measure spatial distribution of trap density in the bulk is frequency dependent charge pumping. Different from conventional charge pumping measurement, it changes the frequency of input gate pulse allowing the access of traps in the oxide. The big question in this method is how to convert the frequency to location of the traps- how deep it can probe at certain frequency? Because this technique is only available electrical test method to know the spatial

distribution of traps, this question is critical. The lack of answer on this question leads to the debate in the literature [28, 29]. For the same piece of experiment result, the interpretation becomes very different from group to group. It further leads the confusion of who is responsible for reliability- high κ stack or buffer SiO_2.

In this chapter, we will demonstrate our experiment evidence to address the answer of how deep the frequency dependent charge pumping probes. Specifically, we use the GHz charge pumping experimental data to show the interface filling time is less than 0.7 ns. It is very different from the previous theoretical calculation. Therefore, a new model is proposed to predict the relation of probe depth versus charge pumping frequency. The conclusion is that charge pumping can probe as deep as into high κ stack.

5.1. Frequency Dependent Charge Pumping

Charge pumping has been widely used to study interface traps in the Si/SiO_2 system for more than thirty years [60-61]. It is well known for its simplicity and high sensitivity for trap characterization. When performing CP measurement, the source and drain are usually applied with a small reverse bias. Periodic pulses are applied to the gate and drive the surface into inversion and accumulation periodically. An average CP current measured from substrate is attributed to the recombination of trapped electrons and holes. By measuring the substrate current, an estimate of the

mean value of the interface-state density over the energy range swept by the gate pulse can be obtained.

In the Si/SiO_2 system, most of the defects is located at the Si/SiO_2 interface and very few of them is in the bulk oxide. However, the story gets changed once the high κ dielectrics come up. With many of the defects in the oxide, the charge pumping is no longer probing the truly interface traps. During the time when the surface is driven to inversion or accumulation for sufficient time in the CP measurement, the electrons/holes are able to fill the traps in the bulk by tunneling. The longer time and lower charge pumping frequency it has, the further traps from the interface can be filled. Therefore, one can know the trap density at different distance from the interface by making the measurement with different charge pumping frequency. This is so called frequency dependent charge pumping [7].

In order to know the spatial distribution of the traps, one needs to convert the charge pumping frequency to the distance from the interface. Figure 5.1 shows the relationship. It shows the time of carrier at inversion/accumulation in CP, which is half of the period for the symmetric gate pulse, versus the probe distance from the interface. In this case, dual layer of dielectrics is used – 1 nm buffer SiO_2 and 3 nm HfO_2 (high κ). As shown in the insert of this figure, the relationship can be derived from the two-step filling mechanism [62, 63]. The electron/holes first fill the interface states and then tunnel into the bulk traps from the filled interface states. Therefore, the

relationship of time and depth will start at the time when the interface states are all filled. As a result, the interface filling time is the starting point (t_1 as shown in the figure).

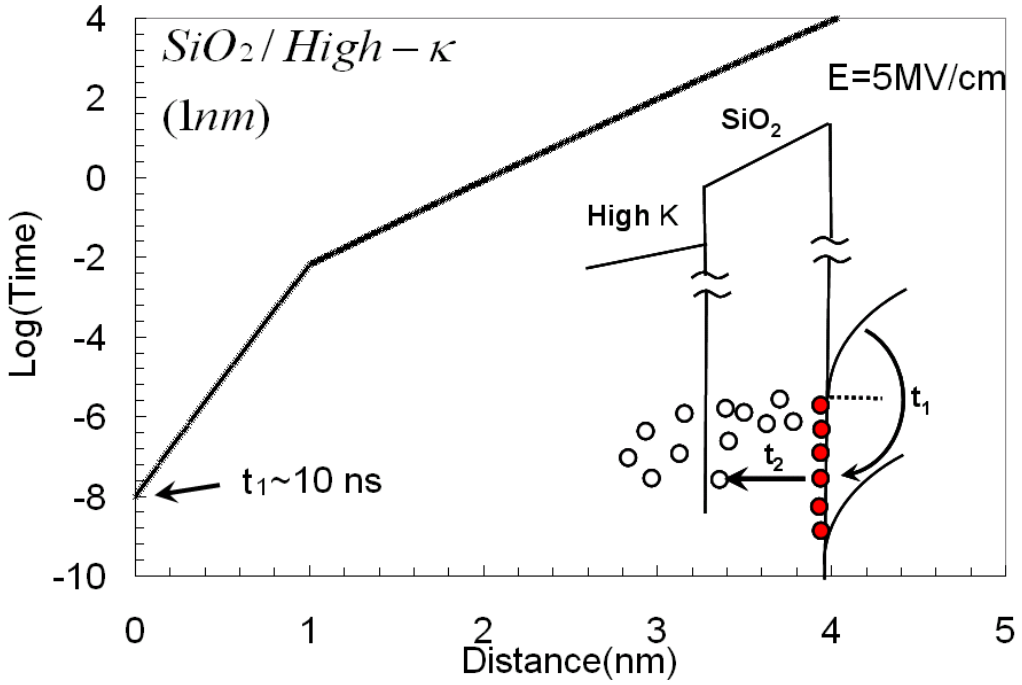


Figure 5.1 The relationship of the time in inversion/accumulation in charge pumping versus the probing distance from interface. The longer time it has, the deeper from the interface can be probed. The relationship can be obtained from the two-step filling model and tunneling probability calculation. The mechanism is shown in the insert. The electrons/holes will fill the interface traps first and then tunneling into the traps in the bulk from the filled interface traps. Therefore, the probe depth starting at the time when the interface is filled which is interface filling time (t_1 as shown in the figure). The distance will keep increasing with more inversion time allowed. The slope this function is determined by the tunneling probability which is determined by the material parameter of the dielectrics. That is reason why the slope is different at SiO_2 and high κ.

With longer inversion time and lower CP frequency, the electrons/holes have more time to fill traps further from interface. This is the tunneling process and the rate of filling is simply determined by the tunneling probability. Therefore, the slope of the time-depth relationship in the figure depends on the material prosperities of dielectrics

which are important in the calculation of tunneling probability. For the detail calculation, one can refer to the Appendix C. Since SiO_2 and high κ is very different materials, the slope of them is not the same. High κ has lower barrier height (1.5eV compared to 3.1eV of SiO_2). It is relatively easier for tunneling and electrons can reach further traps for the same amount of time.

5.2. Controversy in Frequency Dependent Charge Pumping

This relationship of time-depth is not consistent from groups to groups [28, 29]. It further leads to the debate where the increased traps come from after the electrical stress. As we shown, the slope of this relationship depends on the material prosperities such as barrier height, effective mass. It is more or less consistent and there is not much question on that. The controversy really is the starting time - interface filling time. One camp in the literature [29] uses the 10 ns as starting time and derives the relationship with two step filling model as shown in figure 5.1. They conclude that CP is probing the buffer SiO_2 for the typical frequency range from 100Hz to 1 MHz.

Different from the pervious shallow camp (SP), the other camp uses 66ps as the starting time and concludes that it can probe deep into high K for the same frequency range [28]. The time-depth relationship used by this deep camp (DP) is shown in figure 5.2. They assumes a direct filling process, that is, the electrons/holes tunnel and fill the bulk traps directly from inversion layer instead of filling the interface first. This is so called tunneling front model [77] as illustrated in the insert of the figure. In

this case, the starting time is the escape time for the electrons from the inversion layer. It is 66ps as determined by Lunstrom [78].

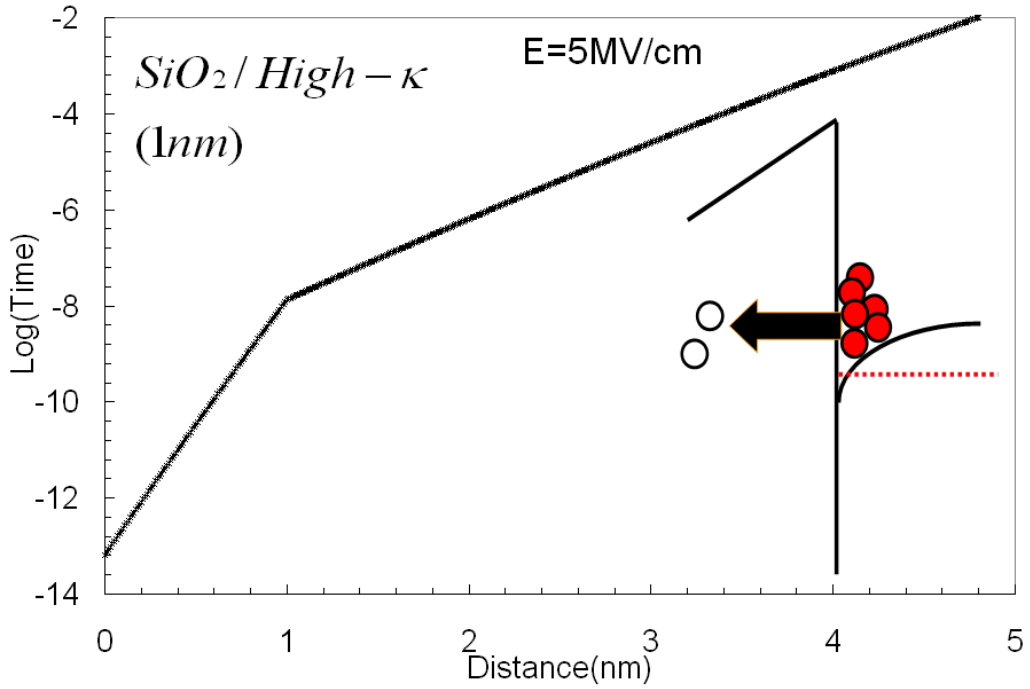


Figure 5.2 The relationship of the time in inversion/accumulation in charge pumping versus the probing distance from interface. Different from figure 5.1, the starting time is 66ps instead of 10ns. This value is from the tunneling front model as shown in the insert. The electrons/holes tunnels to the bulk traps directly instead of filling the interface states first in the two step filling process.

Now, we have two camps with different starting time and then different conclusion even having same experimental data. Who is right? This controversy has been around for a while and no answer has been made so far. We are the first one to provide a good answer with enough evidence. It will be shown step by step in the rest of this chapter.

From previous discussion, one may notice that the underlying mechanism to explain the bulk traps filling is quite different from these two camps. One uses a two step filling while the other assumes a direct filling. We believe the two step filling process

from the shallow camp is more appropriate for charge pumping. That is because the CP itself is also a two step process – electron and hole filling. Only the traps filled both by electron and holes can be counted as a net recombination current. Moreover, the filling is executed by a tunneling process- an electron moving between two states at the same energy level. If it is a direct tunneling from inversion layer, those traps can only be filled by electrons but not holes because there are very few holes in the inversion layer. The only traps can be both electron and hole filling is those located at the same energy level of the interface states. It can only be done by letting the electron/hole fill the interface states first and tunnel into bulk traps. It is exactly the two step filling process.

After recognizing the two step filling model as the responsible mechanism, the starting time should be the interface filling time. In the shallow camp, they assume this time is 10 ns. It originates from the theoretical calculation by Maneglia *et.al* [27, 79-86]. However, it has never been experimentally proved. Obviously, there is a need to find out some way to confirm this number. We are the one to carry out this important procedure.

5.3. Basic Principle of Finding Interface Trap Filling Time

The way to find the interface filling time experimentally is very simple – enough high frequency charge pumping measurement. Traditionally, CP measurement is typically

carried out up to 1MHz. In these measurements, the time during which the channel is in inversion/accumulation is long enough to allow all the carriers to be trapped. When assuming very few bulk traps and all the traps are located at interface, then detected trap density is frequency independent as long as the magnitude of gate pulse is unchanged.

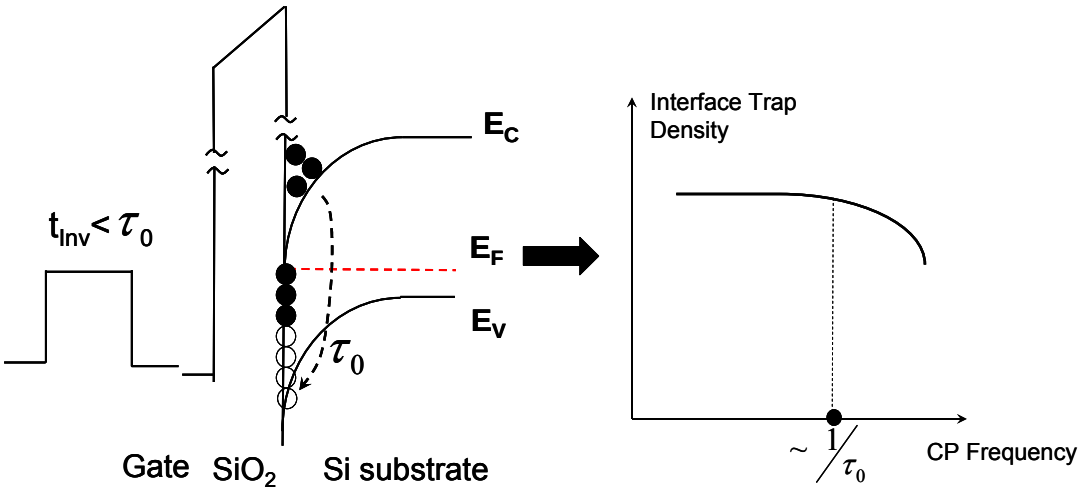


Figure 5.3 Basic principle of finding interface filling time τ_0. Fast gate pulse with inversion/accumulation time less than τ_0 is applied. Not all of all of the interface traps are filled and result in attenuation in trap density beyond certain frequency. Interface filling time can be identified by find the frequency where the trap density starts to decrease.

However, as high frequency gate pulse is used, it switches the MOSFET between accumulation and inversion so fast that the filling of interface states might not follow. At certain point when the inversion/accumulation time becomes smaller than interface trap filling time, the interface states may no longer be completely filled and the detected trap density starts to decrease (Figure 5.3). In this way, the interface filling time can be determined by finding the frequency when the attenuation of trap density starts.

Maneglia *et al*'s calculated the interface filling time as 10 ns, which means that all the interface states are filled within 10ns. To prove or disapprove that, a CP experiment beyond 50 MHz is required. With inversion/accumulation time less than 10 ns (half period of 50 MHz symmetric gate pulse), a reduced trap density will start to be observed around 50 MHz if their argument is true. Otherwise, there will be constant trap density up to 50 MHz. The exact interface filling time can be found by keeping increasing the CP frequency and identifying the point where the trap density starts decreasing.

5.4. High Frequency CP Experiment Setup

High frequency CP measurement is not easy but rather challenging. Even though 50 MHz CP experiment is thought to be enough for confirming Maneglia *et al*'s calculation, it is necessary to extend the frequency high enough such as beyond 1GHz to actually find out the interface filling time. One can see the reason later and even 1GHz might not be enough. However, it has already pushed to the limit of experiment. As far as we know, no reliable charge pumping beyond 1GHz with square gate pulse has been reported. The main difficulty lies on how to ensure the fidelity of the waveform applied to the gate. This difficulty leads Sasse and Schmitz [87] to try the sinusoidal gate bias instead of a pulse waveform with fast rise and fall time. The frequency can be beyond 1GHz but the extraction of trap density requires complex fitting that undermines the level of confidence on the out come. In this chapter, we

report GHz charge-pumping experiment using pulse waveform with very fast rise and fall time. To ensure that the pulse waveform at the gate is faithfully reproduced, we used a home-built, 50Ω terminated probe that is good (verified) up to 20GHz.

In this study, nMOS transistors from a 0.16μ technology with 2.4 nm gate oxide thickness were used in the experiment. As shown in Figure 5.4, both source and drain were biased at 0.1V. The average substrate current was measured at 0V by HP4156A. Gate bias was switched between 1V (inversion) and -1.2V (accumulation) repetitively. At frequency up to 50 MHz, we used a square wave generator with ~1ns rise and fall time. For higher frequency, we use a pulse generator with low duty cycle (1MHz pulse rate), leading to extremely asymmetric square wave.

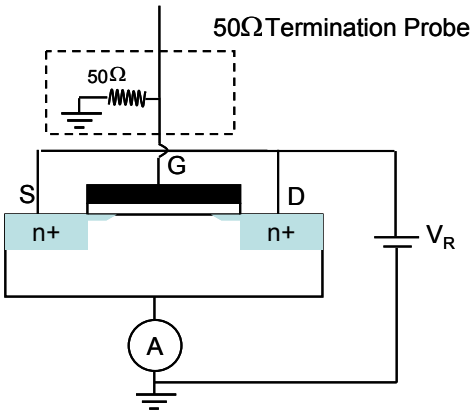
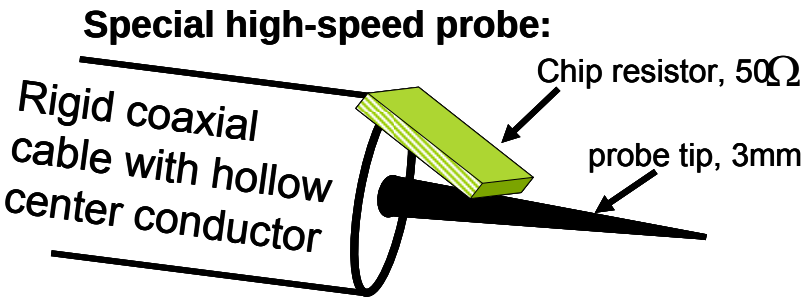


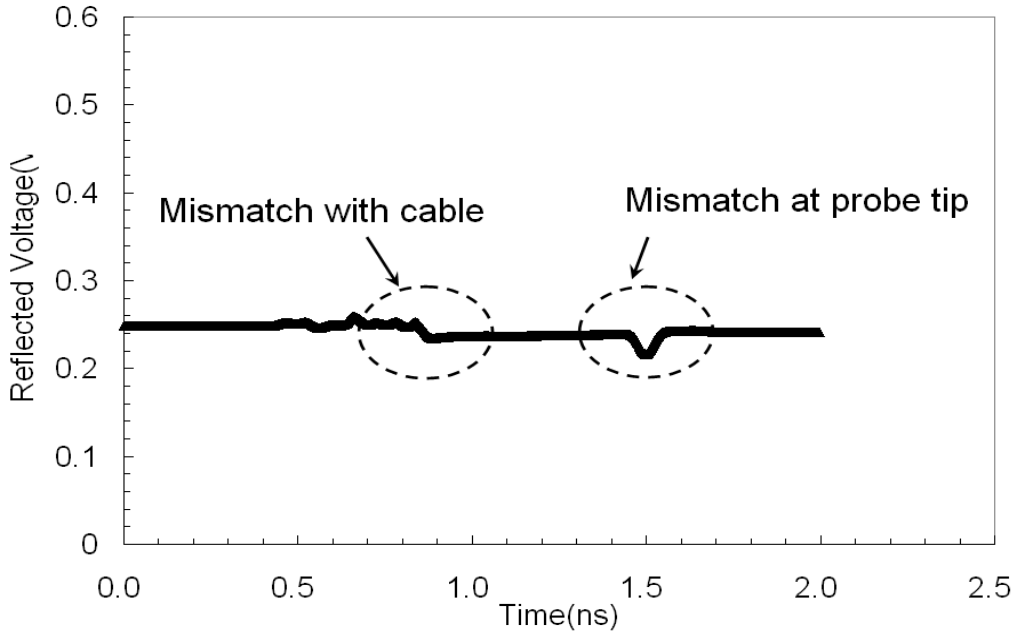
Figure 5.4 Experimental setup diagram of charge pumping measurement, gate of MOSFET is probed by specific 50Ω termination probe

It is also noticed that the novelty in this setup is the introduction of specific 50Ω termination probe. It is one of the critical parts to offer the ability to extend CP measurement beyond 1GHz. At frequencies above 10 MHz, the wavelength of the gate voltage signal is in the same order of magnitude as the length of the measurement

cables used in the setup. All the circuits should be treated as high frequency feature – transmission line theory and impedance matches. The 50Ω input resistance at gate probe is exactly the key solution to ensure the impedance match between gate and output of the pulse.



(a)



(b)

Figure 5.5 (a) A home-made high-speed single probe with 50Ω termination. The probe allows reflection-free application of high-frequency bias up to 20GHz for device characterization (b) TDR characteristics of this probe with probe arm

As illustrated in the Figure 5.5(a), a 1 mm^2 size 50Ω chip resistor leans close to the probe tip and connects the probe (central conductor passing the signal) to the outside shield (ground). We use conductive silver paint to allow the chip resistor to make a

good contact. With this 50Ω input resistance, the impedance match ensures that the input pulse reach the gate without any attenuation. It can be verified by the help of Time Domain Reflectrometry (TDR). The principle and usage of TDR has been well explained in chapter 3 and Appendix A. It is well known for its advantage of detecting the location of impedance mismatch.

Figure 5.5(b) shows the detected reflected voltage measured by TDR when it is connected to our home-made probe with 50Ω chip resistor. Flat line in this figure means all the incident voltage are transmitted and any change means some part of input gets reflected or partially transmitted with attenuation. Figure 5.5(b) indicates that most of the incident waveform transmits without attenuation. The first very small observed bump is due to the connection between probe arm and microwave cable. The other dip is from the probe tip. These small irregularities are not worrisome. The half time width of the dip indicates that it takes around 50 ps for the signal to complete a single trip. As a result, this 1% reflection will not be considered as an important factor for the measurement within 20 GHz.

5.5. Charge Pumping up to 50MHz

We first carry out CP measurement with symmetric square wave up to 50MHz. The test structure is 2.5 um^2 nMOSFETs from 0.16um technology with 2.4nm high quality pure thermal grown SiO_2. Its picture is shown in Figure 5.6. It has three transistor

bunched together with independent gate/drain but same well/source. Its high quality thermal SiO_2 is well known for its low density bulk traps. Therefore, the detected traps are truly interface states.

Figure 5.6 A test structure of MOSFET having three transistors in bunch with independent gate and drain.

In this study, a symmetric square wave with 1ns rise/fall time is applied to the gate. While keeping the base level voltage of gate pulse as -1.2V (accumulation), three top level gate bias conditions are chosen as 0.7V (depletion), 1.0V and 1.3V (strong inversion) respectively. We record down the charge pumping current as the function of frequency from 1Hz to 50MHz. The result is shown in Figure 5.7.

It is observed that CP current linearly increases with frequency. It can be explained by equation (5.1) which is the measured CP current:

$$I_{CP} = fQ_{CP} = fqA_G \bar{D}_{it} \Delta E \qquad (5.1)$$

Where q is the electron charge, f is the input gate pulse frequency, A_G is the channel area of the transistor (cm^2), $\bar{D}_{it}$ is the mean interface trap density over the energy range ΔE swept by gate pulse.

Equation (5.1) indicates a linear relation of charge pumping current I_{cp} with frequency f if all the other parameters are constant, consistent with observation. Moreover, in

Figure 5.7, a higher CP current is observed for higher top level voltage in gate pulse. It can be also explained by equation (5.1). With higher top level of gate pulse, Fermi level can sweep a larger energy range ΔE of interface states resulting in higher CP current $I_{CP.}$

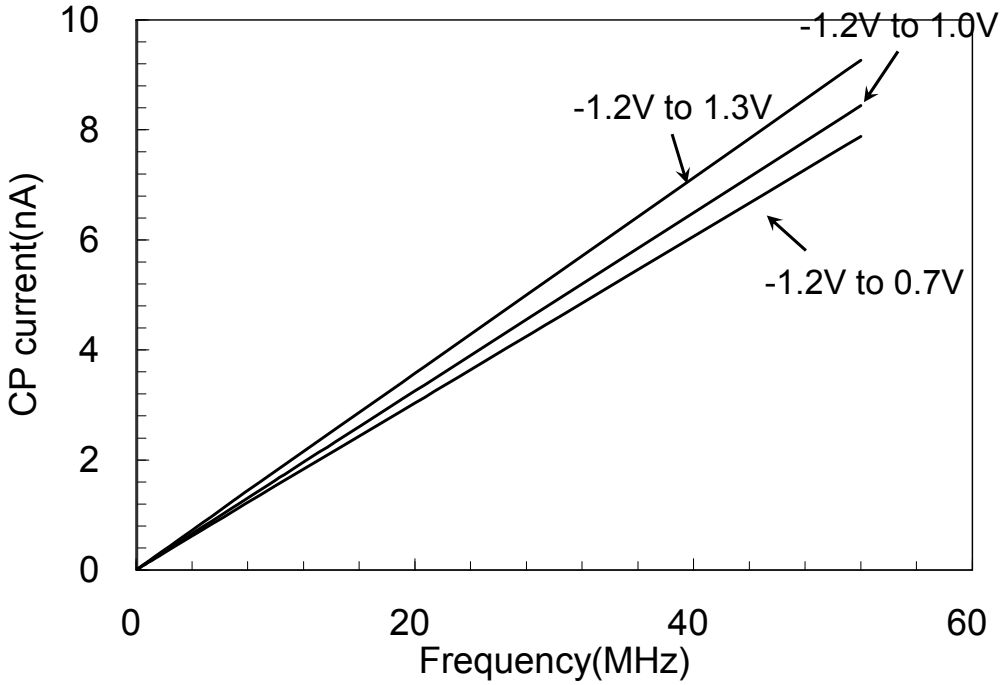


Figure 5.7 measured charge pumping current as a function of frequency up to 50MHz. CP current linearly increase with higher frequency. Comparing the case with various gate bias, higher top level gate bias results in higher CP current.

Equation (5.1) also shows that the trap density can be obtained by normalizing the detected CP current to the frequency. Figure 5.8 shows the trap density obtained from the result in Figure 5.7. For the frequency range from 100kHz to 50MHz, the measured trap density is basically constant. The lack of frequency dependent suggests that bulk oxide trap is negligible as expected in production quality thermal oxide. All the measured traps are interface states. The deviation below 100kHz is likely due to

leakage current [88].

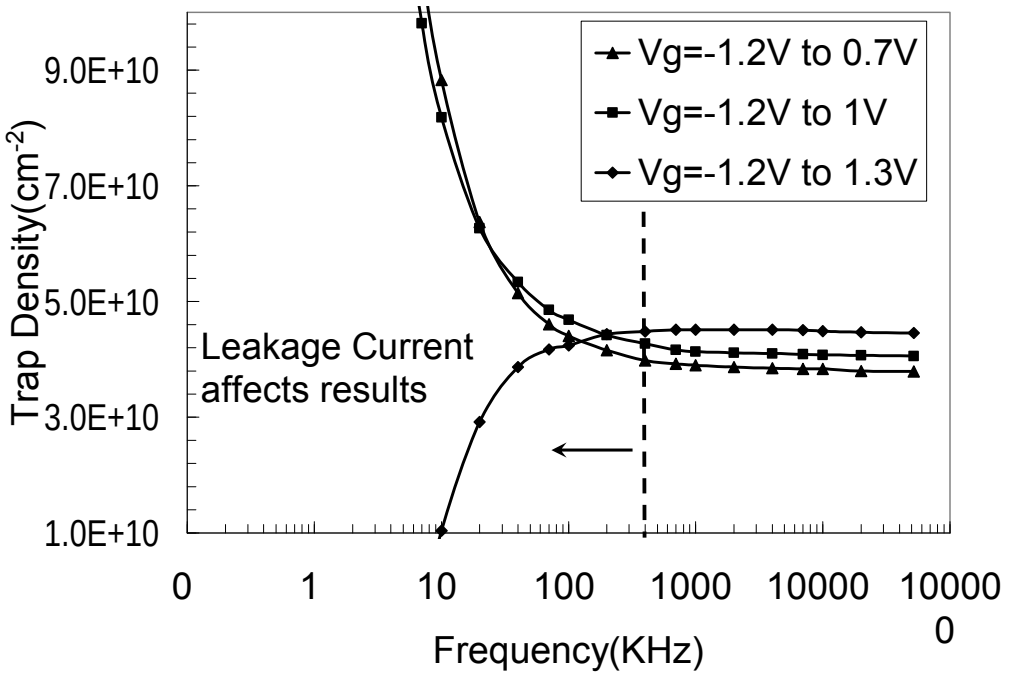


Figure 5.8 The measured trap density as a function of charge-pumping frequency using square wave applied to the gate. From 100 kHz to 50MHz, the trap density is basically the same. Below 100 kHz, the result may be affected by gate leakage current.

The leakage current can also be seen in the behavior of CP current at low frequency range (1Hz~10 KHz) as shown in Figure 5.9. At the frequencies beyond 1 KHz, a CP current increases with frequency as expected. However, the constant CP current is observed below 1 KHz. This behavior is believed to come from the effect of leakage current.

The detected substrate current I_{SUB} in CP actually consists of two components: the recombination current of trapped holes I_{CP} and DC leakage current flowing to substrate $I_{Leakage}$. This can be expressed as:

$$I_{SUB} = I_{CP} + I_{Leakage} = AD_{it}f + I_{Leakage} \quad (5.2)$$

The recombination current I_{CP} is linearly dependent on frequency while the leakage current $I_{Leakage}$ is not. As the frequency gets lower, the I_{CP} is smaller. Once I_{CP} is much smaller than $I_{Leakage}$, the detected total substrate current is dominated by the leakage current which is constant with frequency.

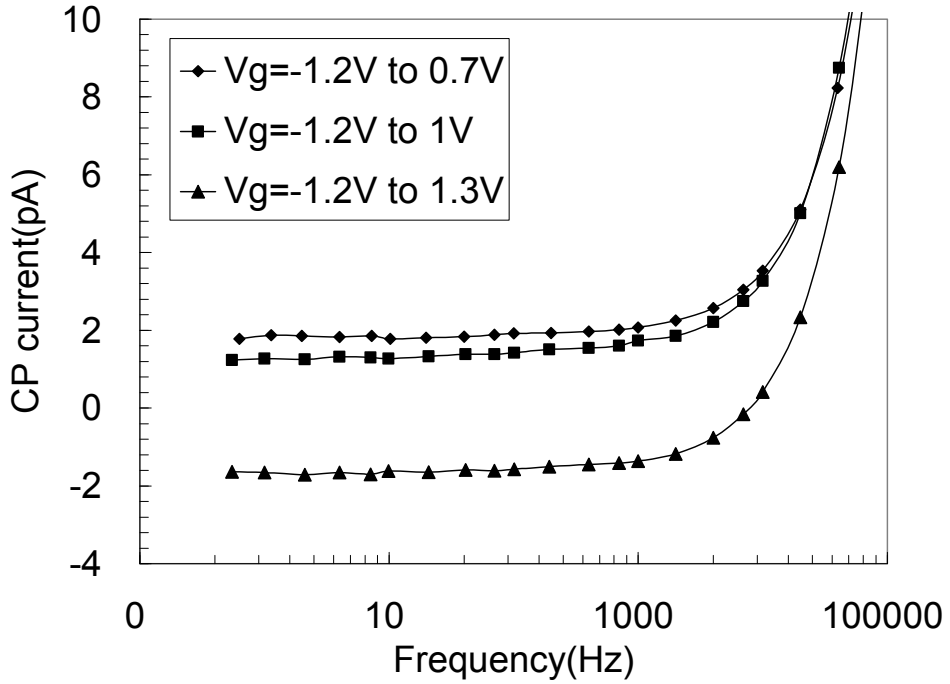


Figure 5.9 Plot of measured charge pumping current as the function of frequency from 1 Hz to 100KHz.Below 1 KHz, the CP current is overwhelmed by the DC leakage current and exhibits frequency independent behavior.

This leakage current can be measured in a very easy manner. We use the exact same experiment setup but replace the square wave gate pulse with a constant DC bias. The measured current at substrate current will be the leakage current. We take the average of 100 repetitive measurements to reduce the random noise. Figure 5.10 shows the measured substrate current under constant gate bias of -1.2V, 0.7V, 1V, 1.3V

respectively. These values of gate biases are used in the previous charge pumping measurement.

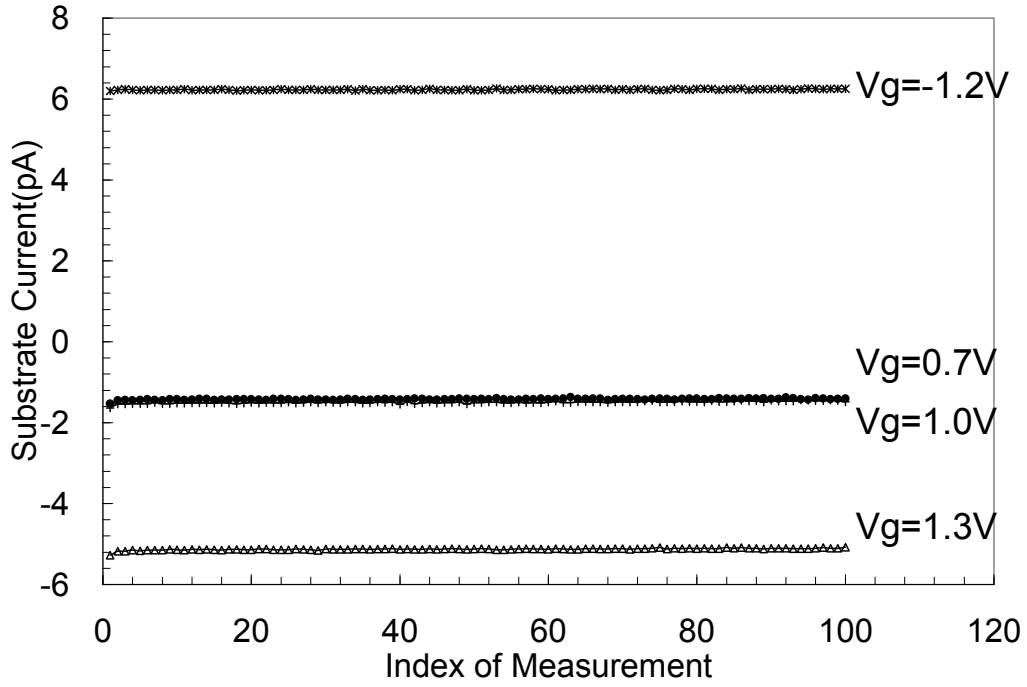
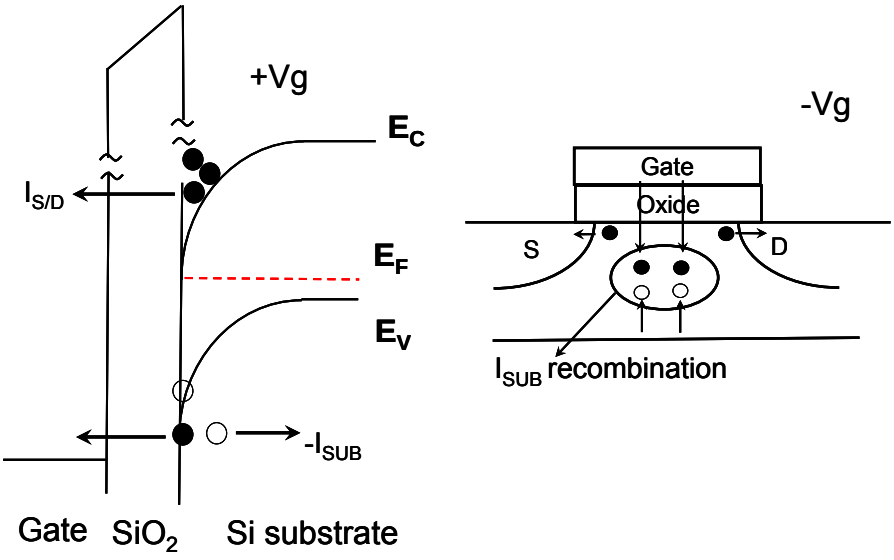


Figure 5.10 The measured substrate current using the same setup except applying a constant DC voltage on gate instead of a square wave.

In Figure 5.10, a 5~6 pA current flows to substrate at either very positive (V_g=1.3V) or very negative gate bias (V_g=-1.2V). The responding mechanisms are shown in Figure 5.11(a) and (b). During negative gate voltages, the gate injected electrons recombine with the holes at p-type substrate and contribute to positive substrate current $+I_{SUB}$. On the other hand, at a positive gate bias, the electrons tunnel from the valence band of the substrate to the conduction band of the gate electrode through the ultra-thin gate oxide. As a result, the remaining holes at valence band flow into the substrate and contribute to a negative current $-I_{SUB}$. Obviously, the sign of this current predicted in this model is consistent with the observation in Figure 5.10.



(a) (b)

Figure 5.11 The schematic diagram for various substrate current components. (a) The recombination process at a negative gate bias. (b) Valence electron tunneling at a positive gate bias. The picture is taken from reference [88].

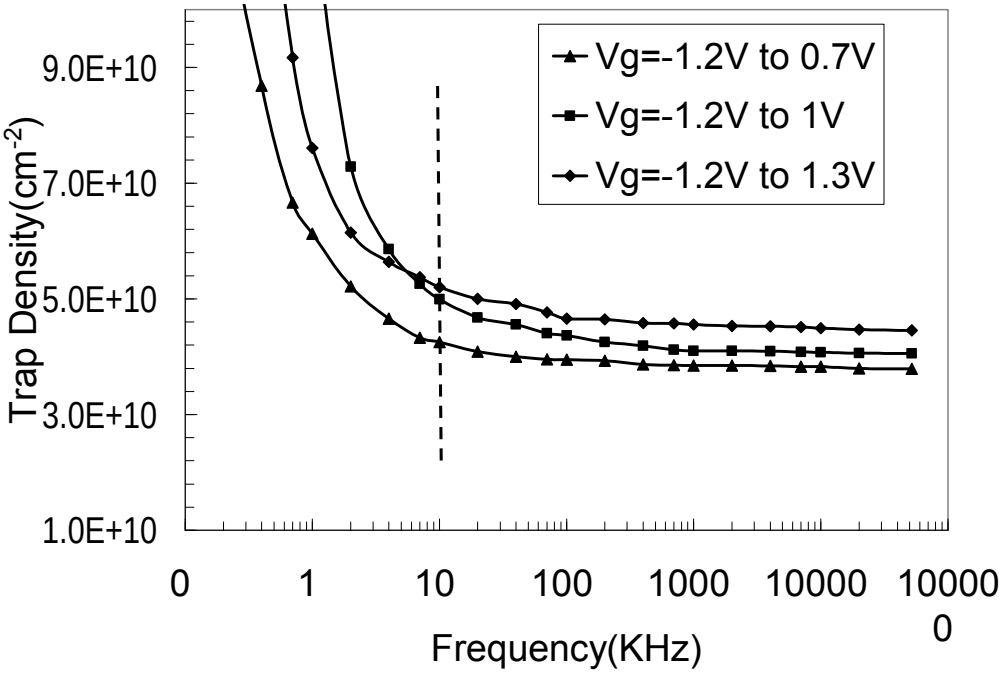


Figure 5.12 The measured trap density as a function of charge-pumping frequency after correcting the contribution of leakage current. It extends the constant trap density to frequency as low as 10 KHz. Below 10 KHz, CP current is overwhelmed by leakage current and correction becomes problematic.

Since a symmetric square wave is applied to the gate in CP measurement, the net

substrate leakage current will be the average of substrate current at top and bottom gate voltage. Therefore, we can subtract the leakage current component and get the current component truly from interface states. Then a more reliable trap density can be obtained shown in Figure 5.12. Trap density is constant between 10 KHz and 50MHz. Below 10 KHz, the recombination current is overwhelmed by leakage current and even correction with subtraction becomes problematic.

5.6. GHz Charge Pumping Results

Constant trap density indicates that interface filling can always follow gate pulse up to 50MHz. To actually pin down the interface filling time, higher frequency CP measurement is a must. In the experiment, we use a short pulse generator with 1MHz repetition rate and <150ps rise/fall time, pulse durations from 300ps to 4ns. The asymmetric waveform ensures that the CP current is either trap filling (positive pulse: short inversion time) or unfilling (negative pulse: short accumulation time) limited.

Since all the traps are at interface, in every gate signal cycle, the channel undergoes less than 4ns at inversion followed by 1us at accumulation as shown in the insert of Figure 5.13. During this short inversion time, the electrons have a chance to fill the interface traps. The following 1us at accumulation is sufficient long enough to make sure that all the electrons trapped in the interface states can be recombined by the holes coming from substrate.

Figure 5.13 shows the measured trap density as a function of the pulse duration. The trap density remains constant down to ~0.7ns. The trap density is in perfect agreement with the value measured at 1MHz using square wave (4.03E+10 versus 4.05E+10). Combining the square wave results from previous section, the measured trap density remains constant from 1.4GHz (0.7ns) all the way down to 100Hz.

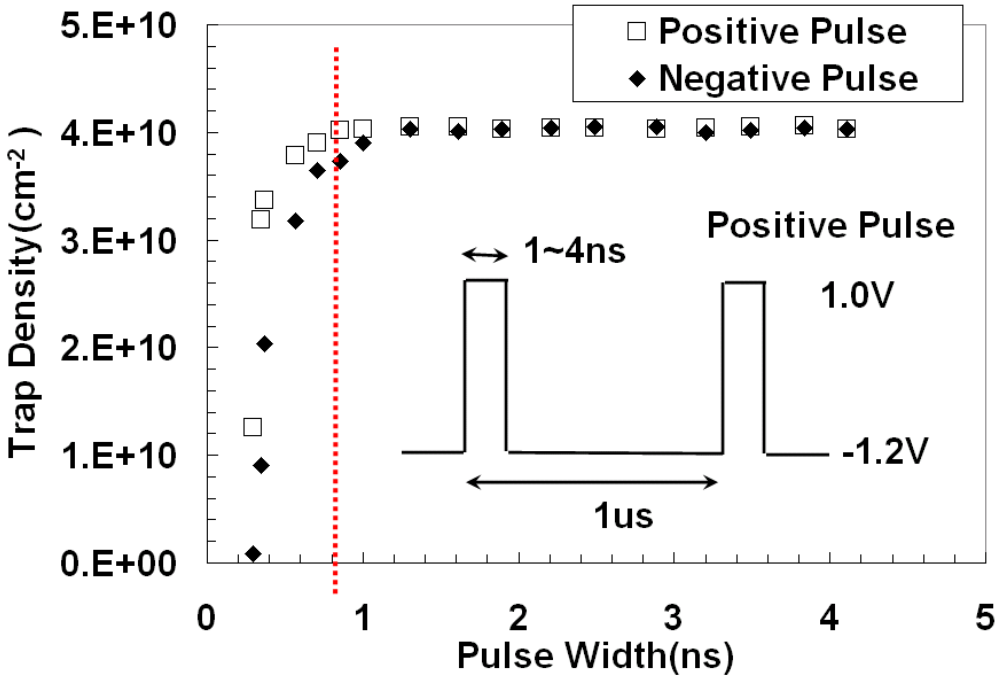


Figure 5.13 Measured trap density as a function of pulse duration is an asymmetric charge-pumping experiment. The repetition rate is 1MHz. The trap density is independent of pulse duration until below 600ps at which point the pulse shape becomes uncertain.

The negative-going pulse shape is shown in Figure 5.14. The measurement is limited by the oscilloscope which is also the source of the ringing. As the pulse drops below 700 ps, the pulse height and width becomes uncertain. Thus the measured trap densities below 700ps are not reliable. Thus the sharp drop in measured trap density in Figure 5.13 may or may not be real. Since both positive going pulse and negative

going pulse cases showed sharp drop at roughly the same pulse duration, we can say that both the interface filling time by electron/hole are less than or equal to 0.7ns.

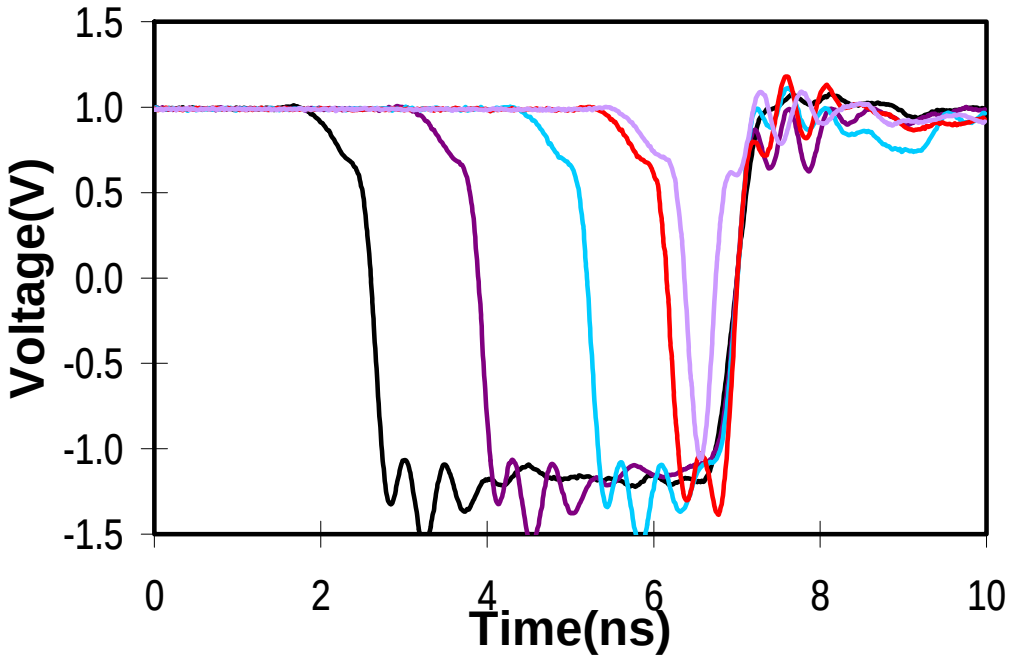


Figure 5.14 The negative going pulse shape as a function of pulse duration. The rise/fall time and the ringing is due to the limitation of the oscilloscope

5.7. Theoretical Model

From previous couple section, we have shown the GHz charge pumping experimental results and concluded that the interface filling time must be shorter than 0.7 ns. It is not consistent with the 10 ns filling time theoretical prediction made by Maneglia *et al*. It is necessary to review the model they use and find out the reason why there is a disagreement. The Maneglia's model is one of the two existing models used by the two camps [28, 29] as we introduced in the beginning of this chapter. To find out the correct CP time-depth relationship to interpret the spatial distribution, it is very

necessary to review in detail the existing two theoretical models. Then, based on our first-hand new experimental results, we can take a new look at these two models and find out who is right or neither of them is correct and a model is needed.

In the CP measurement, interface traps get filled by capturing the electrons from the inversion layer where a large number of free electrons are available. As a result, a recombination process is involved. To characterize this process, the conventional charge pumping model introduces a concept of capture/emission time constant [60, 61]. They calculate these time constants based on SRH theory [89, 90].

As the oxide traps become a subject of interest, the model has been extended by including the tunneling to near-interface oxide traps to interpret the relation between probe distance into the dielectric and the CP frequency [91, 92]. The calculation of probe depth is based on a tunneling-front (TF) model introduced by Heiman *et al* [77], which assumes that oxide traps are filled by carriers in the inversion or accumulation layer through the direct tunneling (DT).

For traps located away from the interface, the fill-time increases exponentially from an initial time constant τ_0 which is the mean time between electron-phonon collisions. Lundstrom *et al.* found τ_0 to be 6.6×10^{-14}s [78]. It appears that the DP camp (refer to section 2.5) based their depth calculation on this model [25, 27, 93, 94]. A depth versus time which is half of a CP cycle for this model is simulated and shown in

Figure 5.13(a)-(c). This simulation is based on our newly derivation of tunneling front model. For details, please refer to Appendix C.

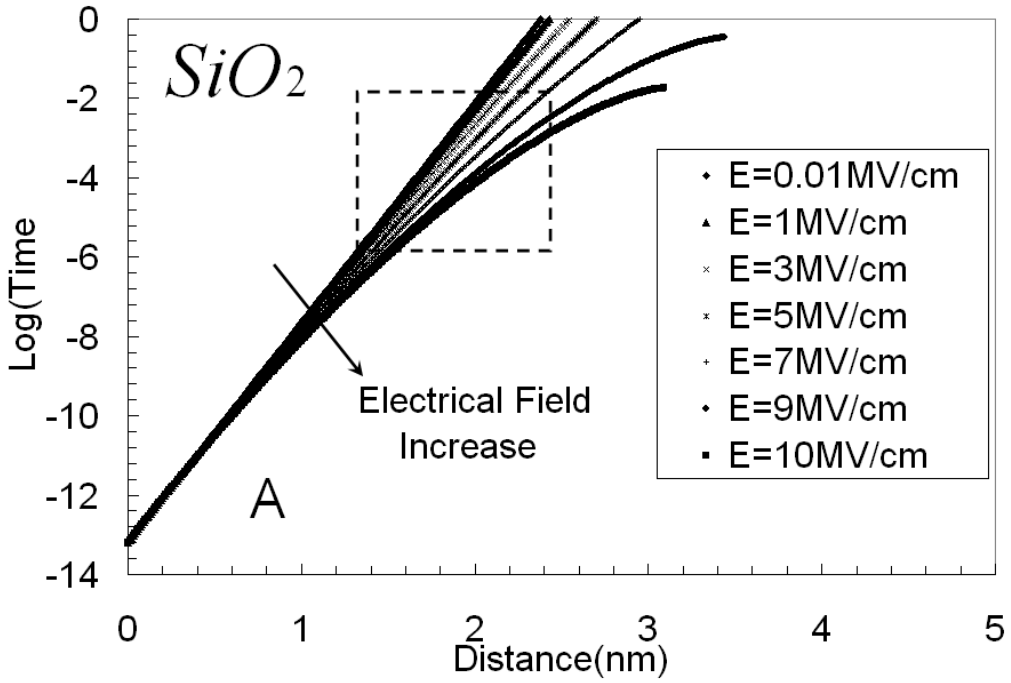
The tunneling front model developed in early work [77] neglects the effect of oxide electric field and a square barrier shape is assumed. However, as the oxide gets thinner, the transistor operation voltage doesn't scale down at the same rate as gate oxide thickness. As a result, the oxide field becomes very high (~6 MV/cm) even at normal operation condition. Therefore, the effect of oxide electric field can not be neglected any more. The barrier shape should be assumed as a triangular shape instead of a square. The demands from modern MOS device motivate us to revise this tunneling front model by taking the oxide field into account.

Based on this new model as discussed in Appendix C, we also develop a simulation program to calculate the probe depth as a function of time period/frequency. It can be applied to single layer with pure SiO_2 or dual layer with high κ and buffer SiO_2. Of course, the conversion from time to probe depth also depends on dielectric parameter values such as the effective tunneling electron mass in SiO_2 or HfO_2 ($m_{SiO_2,e}$, $m_{HfO_2,e}$), effective electron barrier height at SiO_2 or HfO_2 side (Φ_{SiO_2}, Φ_{HfO_2}) and interface filling time (τ_0). Except the different interface filling time, all the following simulations use the same material parameter as listed in Table 5.1.

Table 5.1 Dielectric parameter used in probe depth simulation

Φ_{SiO_2} (eV)	Φ_{HfO_2} (eV)	$m_{SiO_2,e}$	$m_{HfO_2,e}$
3.1	1.5	$0.5m_0$	$0.15m_0$

These parameters are commonly accepted values used in the simulation of pure SiO_2 or HfO_2.Figure 5.15(a)-(c) shows the probe depth as a function of frequency, which is used by deep camp group. For the typical frequency range of 1MHz to 100Hz (dashed window), the CP is probing deep into the high- κ layer even with 1 nm of bottom oxide.



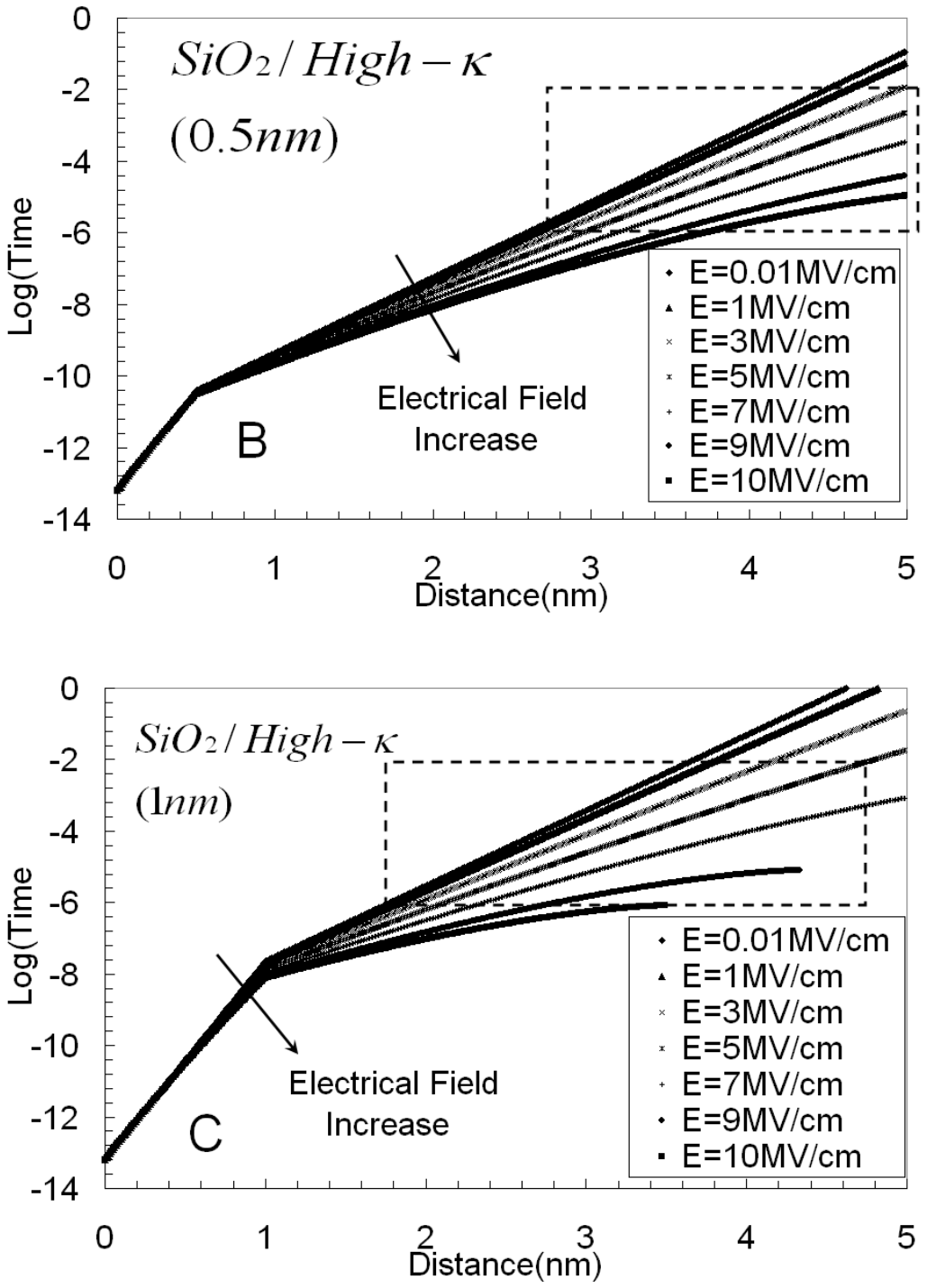
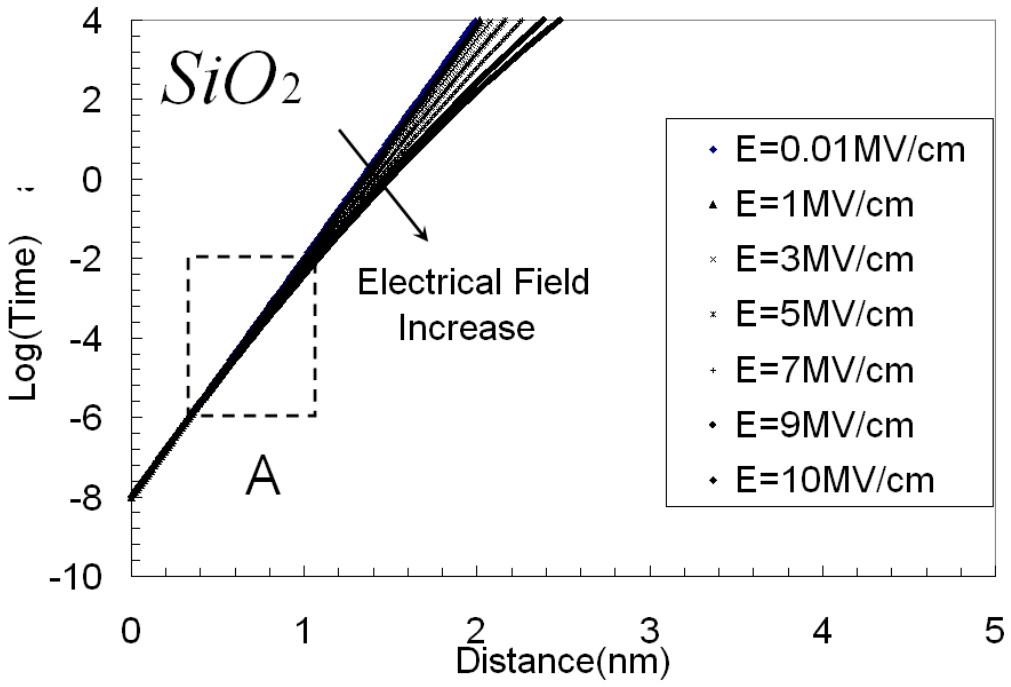
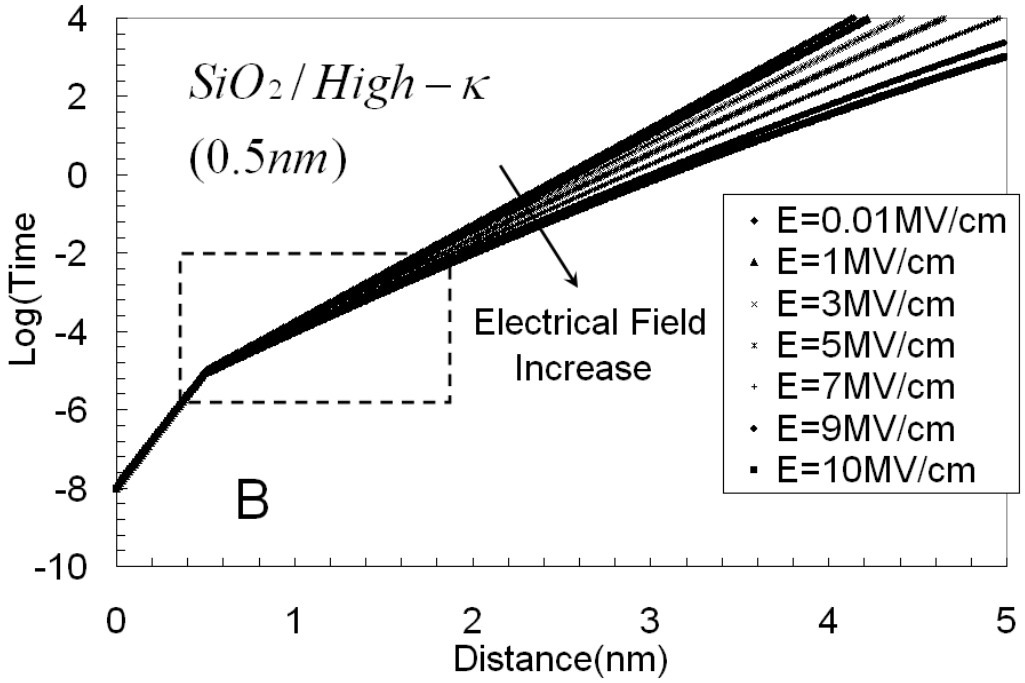


Figure 5.15, The distance into the dielectric probed by charge-pumping as a function of time which is half of a charge-pumping cycle. Calculation is based on the tunneling front model with τ_0=6.6x10^{-14}s. (a) Pure SiO2, (b) High-k dielectric with 0.5nm SiO2 bottom layer, (c) High-k dielectric with 1nm SiO2 bottom layer. The dotted boxes enclose the frequency range of 100Hz to 1MHz and the depth range covered along the 5MV/cm electric field line. The band offset between SiO2 and high-k layer is assumed to be 1.6eV. The effective mass in SiO2 and high-k material is 0.5 and 0.15m$_0$ respectively.

Besides the above model, Paulsen *et al*. also pointed out that CP is a two-step process [62, 63]. In their model, the carriers in conduction or valence bands first communicate with the fast surface states located at the interface through the Shockley- Read-Hall (SRH) process. Then the carriers tunnel into or out of the oxide traps located at some distance away from the interface elastically. Thus the τ_0 must be the interface trap-filling time, not the one given by Lundstrom [78]. Maneglia *et al.* estimated this interface trap-filling time as ~10ns [27, 79-86]. The SP camp based their depth calculation on this model [29]. A depth versus time for this model is shown in Figure 5.16. The set of curves basically shifted in time scale by the τ_0 difference. As it can be seen, with 1nm bottom oxide, CP can never reach the high- κ layer.





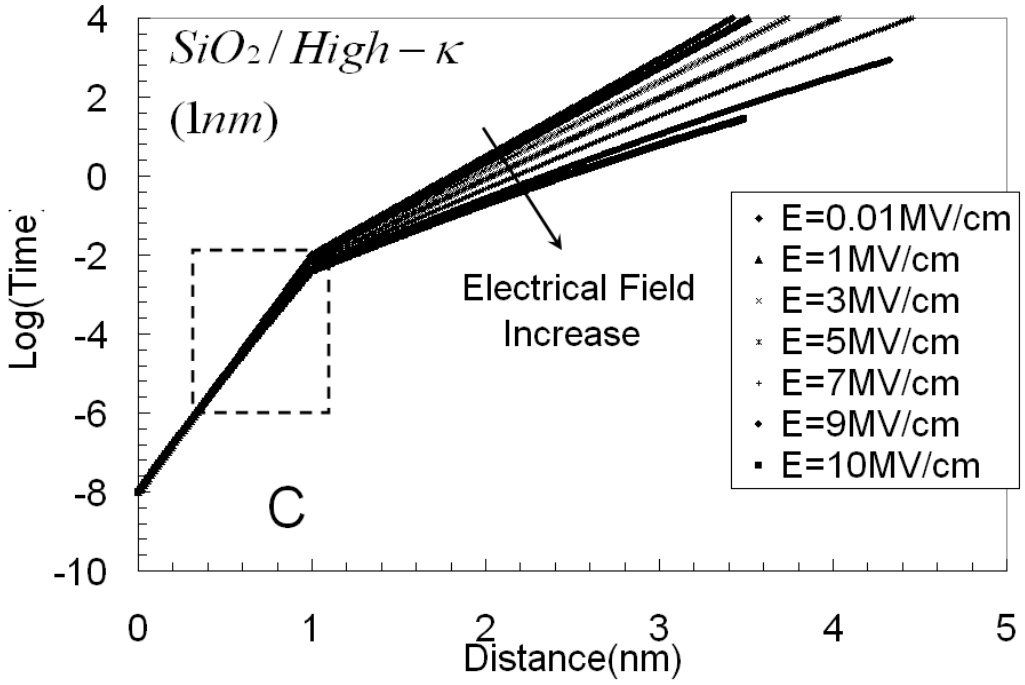


Figure 5.16, Similar to Figure 5.13 except the model is the two step CP model with τ_0=10ns. Compare to Figure 5.15, the main effect is a shift in the time axis equal to the difference in the τ_0 value. As a result, for the same frequency range, the depth probed is much shallower.

5.8. Examining Existing Theoretical Model

As a result, there are two groups with different models therefore different conclusion. The DP camp thought the bulk traps are filled by electrons from inversion directly and CP can probe into high κ. On the other hand, the SP camp thought there is a two step filling process involved and CP can only probe into buffer SiO_2. Compared to direct filling assumed by DP camp, the two-step process model is more plausible since it is well known that the thermal SRH process is very efficient. And the continuous energy distribution of fast surface states at the interface is a common characteristic of silicon-silicon-dioxide interface. The two-step process model is also supported by the dependence of the interface state density and near-interface trap density on the *l/f* noise [95-96]. However, the data from DP camp strongly suggests that they are indeed probing deep into the high- κ dielectric layer [28]. It can only be reconciled this by a much shorter interface trap-filling time.

Our GHz experimental data seems support this much shorter interface filling time. It is much shorter than the one used by SP camp (10ns). The origin of the 10ns trap-filling time is from a classical scattering treatment [27], which can be traced back to Shockley-Read and Hall [67, 68]. In the SRH theory, it says that a defect in the forbidden gap can capture electrons from conduction band and holes from valence band. At the same time, these capture carriers can also emitted back to bands. The emission and capture must reach equilibrium and only those defects at the mid-gap

have the best efficiency therefore referred as the recombination center. The capture time for electron τ_n and hole τ_p can be calculated by following equation:

$$\tau_n = \frac{1}{\sigma_n v_{th} n} \quad , \qquad \tau_p = \frac{1}{\sigma_p v_{th} p} \tag{5.3}$$

Where σ_n , σ_p is the electron/hole capture cross section, v_{th} is the thermal velocity of the electron, n,p is electron/hole density available to be captured. In the Maneglia *et al*'s calculation, they got 10 ns hole capture time (τ_p=10ns) with a thermal velocity v_{th}~10^7 cm/s, capture cross-section σ_p of the size of an atom (~ $5\text{x}10^{-16}\text{cm}^{-2}$) when the hole density p is ~$2\text{x}10^{16}/\text{cm}^3$. Then they concluded that this is the time to fill all the interface states.

After carefully examining of their model, we found some assumption used in this calculation might not be solid. We first examine what exactly the interface states are and how they get involved in the charge pumping. There are two kinds of interface states [97]. One is the silicon dangling bonds or P_b centers that is well-known to be amphoteric and have energies about 0.25 eV below and above the midgap. The other is the residual U-shaped continuum whose origin is not completely clear. In charge-pumping experiments, the measured interface states are normally assumed to be P_b centers.

For amphoteric interface defects, trap-filling in each half cycle of CP involves two steps. Starting from accumulation, the traps are filled with holes. After the sudden switch to inversion, the defects first capture an electron to become neutral and then

capture another electron to become negative. Similarly, in the reverse half cycle, the trap first capture a hole to become neutral and then another hole to become positive. Due to Columbic attraction, one can expect that the first step is much faster than the second step. Thus the neutral defect capturing an electron or a hole is the rate limiting step.

Therefore, the fillings of interface states are indeed a carrier capture process. It is acceptable to draw the statement by taking the capture time as interface filling time. Maneglia *et al* did so but they took the wrong value of capture cross section in their calculation. They took the capture cross section as $5x10^{-16}cm^{-2}$ without marking where they come from. This is why their value can not match our experimental results.

Actually, capture cross section for defects at the SiO_2-Si interface has been studied by many groups over the years using a variety of measurement methods [98-109]. The reported value varies from less than $10^{-18}cm^2$ to greater than $10^{-14}cm^2$ for electrons as well as holes. Such large variation cannot be attributed to sample quality control. More likely, the problem lies with the experiments.

Charge-pumping is one of the commonly used techniques. Capture cross section is obtained from measured emission rate using an expression from detailed balance [89], the condition when the emission and capture rate are equal:

$$e_n = \sigma_n v_{th} N_c \exp\left(-\left(E_c - E_t\right)/kT\right) \tag{5.4}$$

Where σ_n is the electron capture cross section, v_{th} is the thermal velocity of the electron, N_c is the effective density of states of the conduction band, E_c and E_t are the energy of the conduction band edge and the defect respectively. Equation (5.4) is valid only at equilibrium. However, most situations encountered in the measurement are non-equilibrium, easy for emission than capture. This fact is perhaps one of the reasons why the reported cross section varies from experiment to experiment.

In our GHz CP experiment, we sweep the channel from accumulation to inversion. We pin down the capture time of interface traps by varying the CP frequency. So our method is a direct measurement of capture time. Our result of less than 0.7ns capture time has higher confidence level than all the previous ones. With 0.7ns, using equation (5.3), we can calculate the capture cross section as (v_{th}=2x10^{7}cm/s and n~10^{17}cm^{-3}) ~10^{-15}cm^2 or larger for both electron and hole. On the one hand, this value is within the range reported in the literature. On the other, this is a large capture cross section that defies explanation.

In CP, only defects that capture electron and hole alternatively can contribute to measured current. Only when the fill time is too short to fill the defect deepest in the band gap will the measured charge-pumping current starts to decrease. Thus the interface filling time must refer to the capture time of the deepest trap.

The generally accepted mechanism for nonradiative capture of carriers by deep traps

is the multi-phonon emission mechanism [110-112]. In this model, the trap depth is deep only upon the capture of a carrier due to lattice relaxation to accommodate the polarization. Before capture, the neutral trap has a shallow depth so that its excited vibronic state has finite probability of crossing the conduction (valence) band edge to capture an electron (hole) [111]. The capture cross section is determined by the barrier of this crossing E_b and is given by:

$$\sigma = \sigma_\infty \exp\left(-E_b / kT\right) \tag{5.5}$$

Equation (5.5) predicts that the capture cross section can only be ~10^{-15}cm^2 or smaller. While our ~10^{-15}cm^2 or larger result can arguably be within range, the value of E_b must be practically zero. That is to say, the neutral trap must be right at the band edge. This is not the nature of P_b center. A neutral P_b center is exactly mid-gap and the capture of a charge only makes the trap shallower, not deeper. Thus the multi-phonon emission model is ill suited to explain carrier capture at the interface.

5.9. New Cascade Filling Model

As a result, there is a need for new model to explain our extraordinary larger cross section and less than 0.7 ns capture time. All the previous models consider the case of a single trap. With many traps distribute at different energy level of the interface, this direct capture electrons from conduction band might not be the only way to fill traps. The equation (5.3) used in Maneglia *et al*'s calculation is taken from SRH theory. An electron can be first captured by a shallow trap with higher probability and then jump

to the one located at deeper energy level. This multi-step jumping process is called cascade filling process. The process is illustrated in figure 5.17. Assuming there is a quasi continuum of interface states across the band gap, electrons make many transitions to reach the bottom of the band gap. Smaller jumps happen faster and larger jumps happen slower.

Actually, we are not the first one to propose this cascade filling idea. Lax introduced the cascade model in which carrier is captured by a high density of excited states of the defect [113]. However, it was pointed out that deep traps do not have the required excited states [111,112] and therefore the cascade model cannot apply.

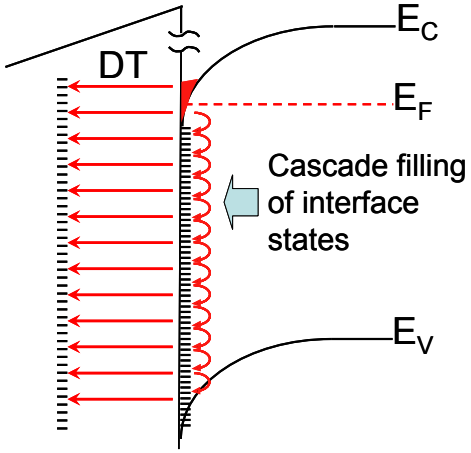


Figure 5.17, Illustration of the cascade interface trap-filling model of charge-pumping. Electrons make many transitions to reach the bottom of the band gap. Smaller jumps happen faster and larger jumps happen slower. Nature automatically optimizes to produce the shortest trap-filling time.

Here we introduce a new cascade model to explain the large capture cross section and extreme short capture time of the neutral P_b centers. Instead of excited states associated with the center, we argue that there is a large number of energy states associated with the P_b center due to strained Si-Si bonds.

As shown in Figure 5.18 at Si/SiO_2 interface, the mismatch between SiO_2 and Si lattice creates the misalignment of atoms at the interface. The P_b center refers to the bonding at the interface with oxygen atom deficiency as circled in figure 5.18(a). For atom circled in the figure, it has three Si-Si bond while the other is "dangling"- no bonding and free electron/hole available. This is so called dangling bond or P_b center. It is the microscopic picture of the interface states.

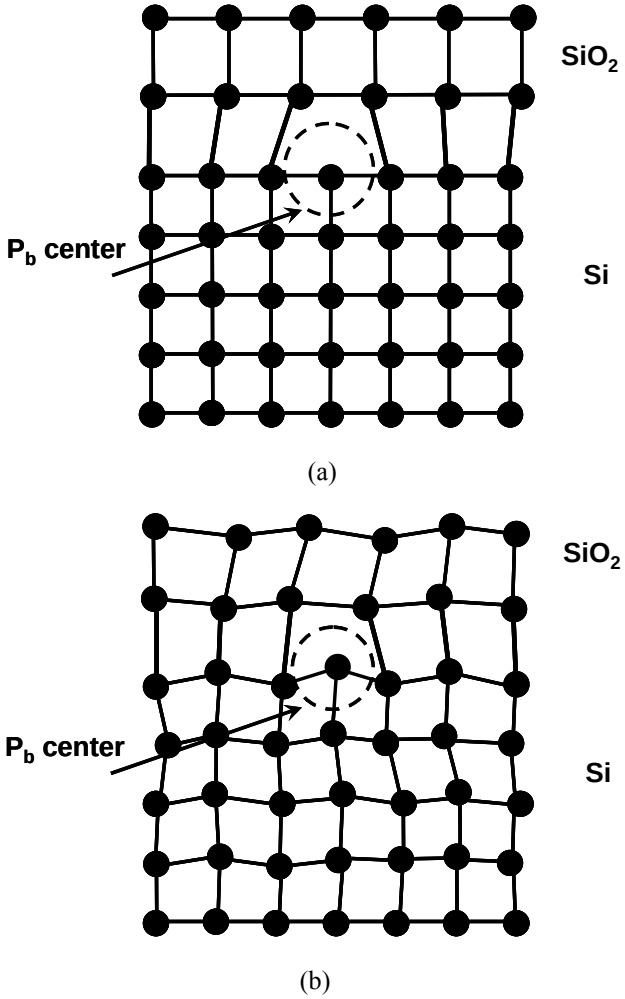


Figure 5.18 Illustration of what is the Pb center and how the strains are formed. The strain can be transferred and affect the bonding between atoms in three to four layers in three dimension.

When this dangling bond is formed, it destroyed the sp^3 hybridization of the silicon atom. As shown in figure 5.18(b), the three remaining Si-Si back bonds can no longer maintain the tetrahedral geometry and are strained [114]. It is reasonable to expect that each one of these back bond will transfer some of that strain to another layer of three bonds through the shared silicon atoms. This process will propagate until the strain is smaller than the thermal fluctuation. Molecular orbital theory says that weaker bonds has smaller energy split between the bonding and anti-bonding states. Note that each weakened bond introduce two energy states, one closer to the conduction band and one closer to the valence band. If we assume that strained bonds are weaken and that the reduction in bond energy is proportional to strain, a large number of energy levels are introduced in the band gap in the immediate surrounding of the P_b center as shown in Figure 5.19. The further it is from the primary defect, the higher the density of energy states and the closer they are to the band edge.

Theoretical calculations, when including the surrounding atoms, do suggest that localized energy states in addition to the trap state are introduced by deep traps [115, 116]. However, full accounting of the surrounding atoms in the lattice is difficult. Laughlin *et al.* [117] used essentially the same picture of neighboring strain bonds to explain the origin of the U-shaped residual interface state continuum. While the strained bonds associated with P_b centers may not account for all the states in the U-shaped continuum (there are other possible reasons [118]), it qualitatively has the right distribution.

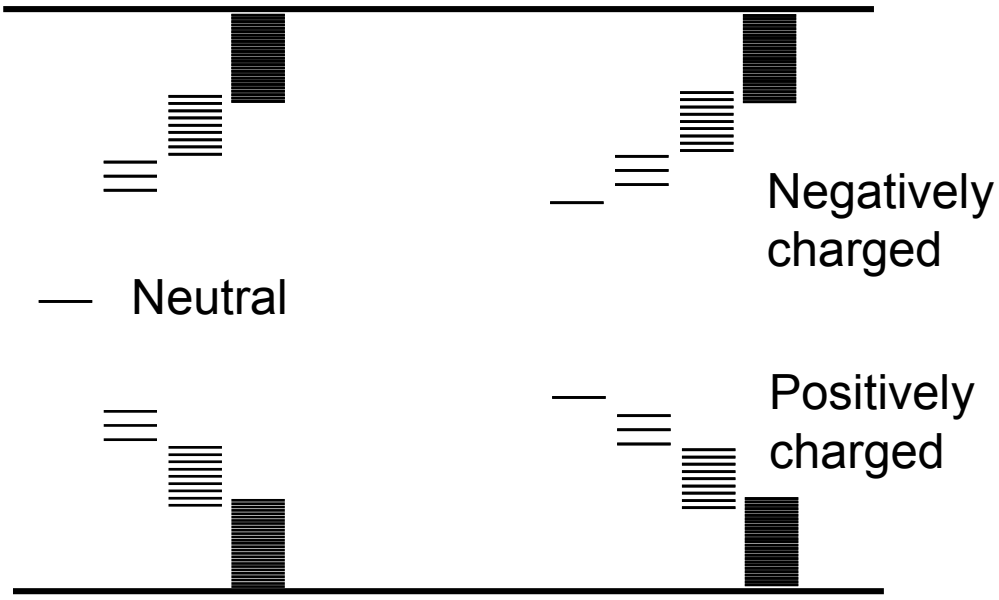


Figure 5.19, Illustration of the strained-bond energy levels in the silicon band gap. The number of strained bond increases three folds every layer away from the primary defect (P_b center). Also illustrated are the energy levels of the three charge states of the P_b center. As the charge state changes, the strained bond energy levels also change.

Figure 5.19 illustrates our idea of the energy levels associated with the three charge states of the P_b center. With these strained states, cascade capture becomes possible. Note that once the carrier cascade down a few steps of the energy ladder, it is captured because the re-emission probability becomes very low.

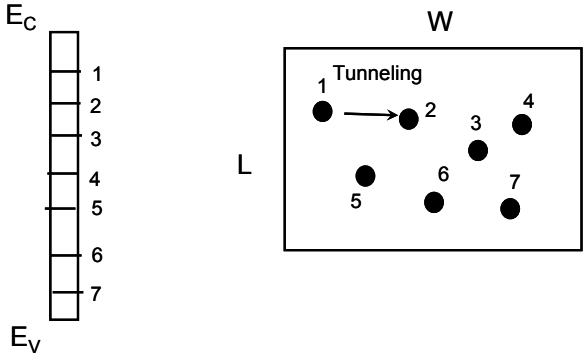


Figure 5.20 Illustration of interface trap distribution that can be considered to be continuously distributed in energy and over the transistor area.

Note that all these energy levels can be produced by a single defect. This is very important. In our cascade model, electrons make transition between traps at different energy levels. If those traps are not from a single defect, electrons also need to jump between defects with different locations. Therefore, there should have an additional tunneling process involved as shown in Figure 5.20. The overall transition probability should be modified by the tunneling probability of electrons. The average distance between defects is around 100 nm with commonly interface state density of 10^{10} cm^{-2} and assumption of equal space trap distribution. This is extremely long distance for a tunneling and the probability is extremely low. Therefore, an unrealistic long capture time (>1s) will be predicted, which is obviously not consistent with our GHz charge pumping experiment. Fortunately, the possibility of these many energy level traps can be from a single defect make our cascade filling model more complete.

5.10. How deep does FDCP probe

With our new cascade model, we can calculate the interface filling time. The total filling time should be the product of time in each jump and the number of step. Each step is an electron capture process. During the capture, a phonon must be emitted due to energy conservation. The capture time is therefore the inverse of collision frequency modified by an exponential factor related to the phonon energy:

$$\tau = \tau_t e^{-\Delta E/kT} \tag{5.6}$$

Equation (5.6) calculates time for a single transition from state to state. τ_t is the

inverse of collision frequency and $e^{-\Delta E/kT}$ is the probability to get a phonon with right energy ΔE. To fill the traps at the far end of the band gap, the required time will be extremely long because phonons with band gap energy are extremely rare. This difficulty is perhaps the reason why the problem of interface trap-filling time has never been explicitly treated.

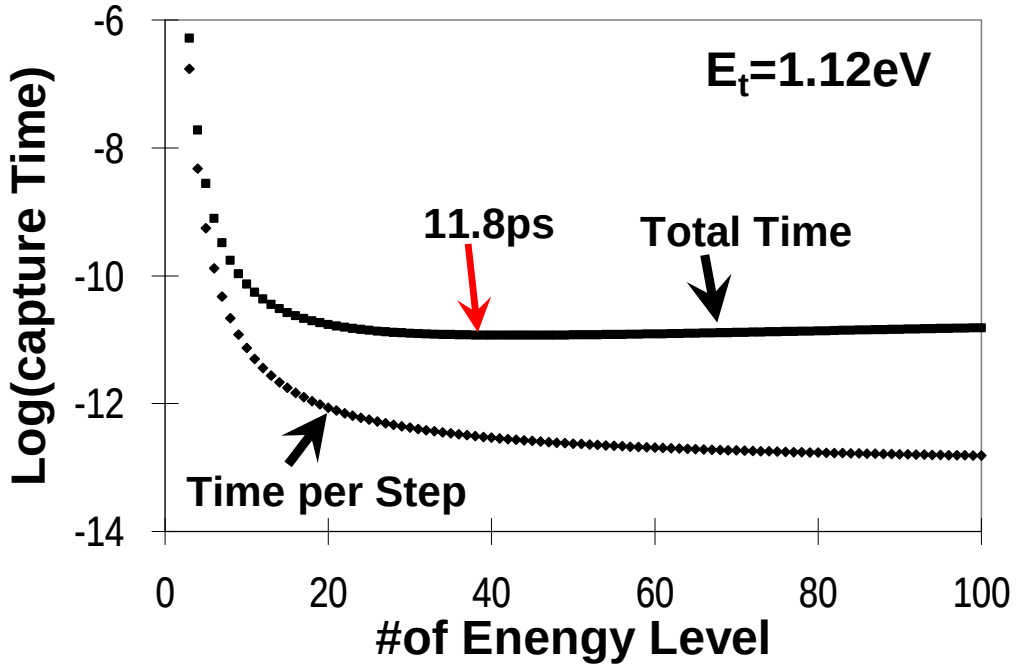


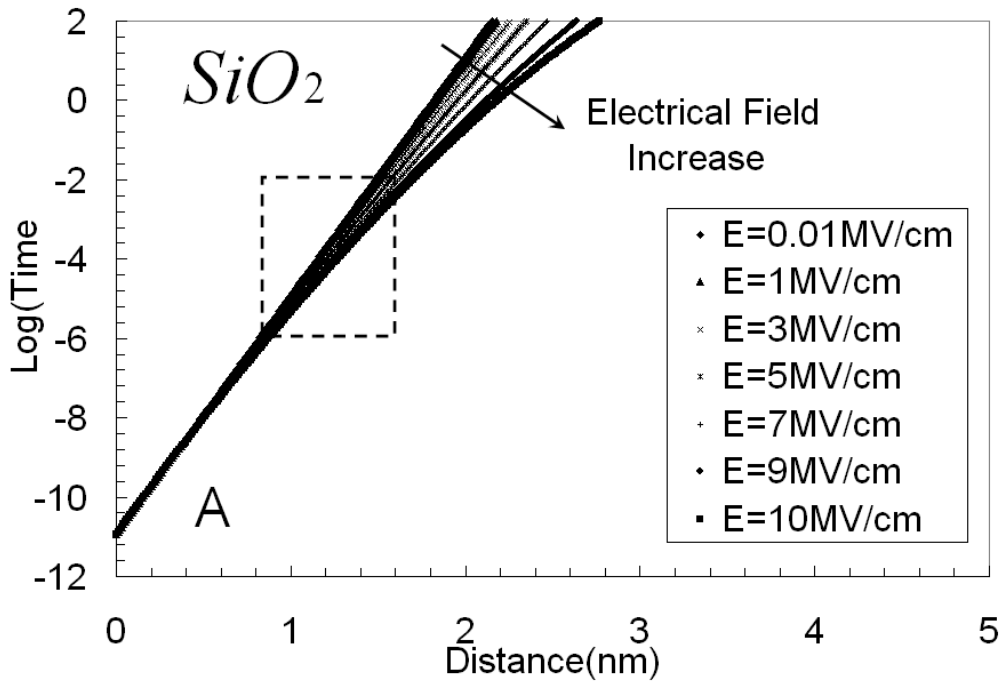
Figure 5.21, The per step time and the total time required to cross the whole band gap (assuming equal energy steps) are plotted as a function of number of steps (energy levels) needed to cross the band gap. A broad minimum of 11.8ps total time is evident.

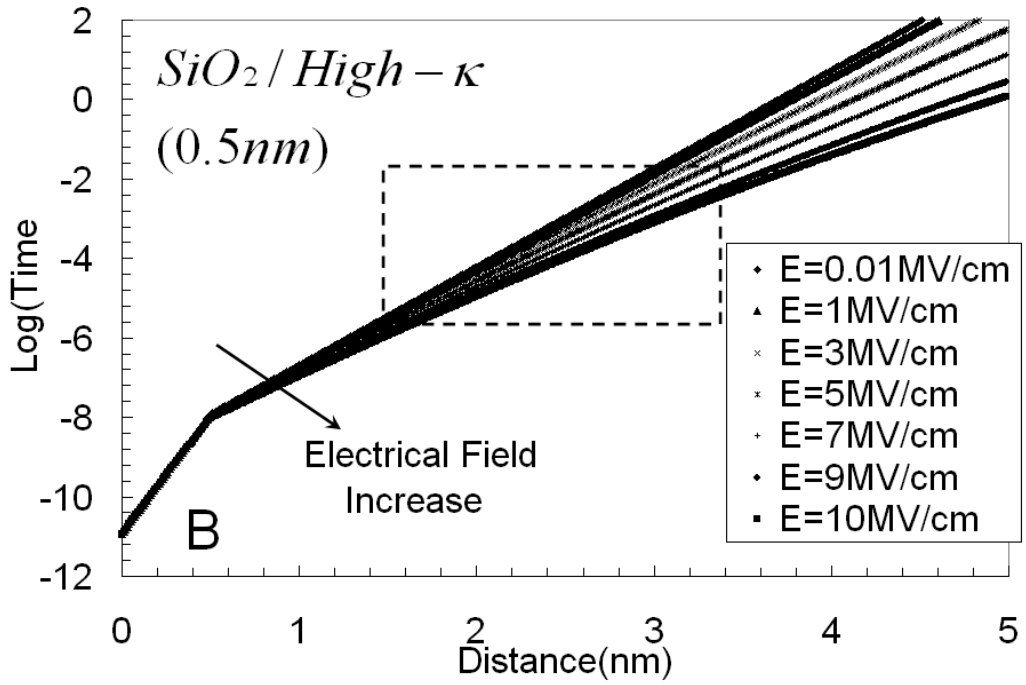
In that case, instead of the unrealistic classical collision frequency of ~10^8/s, we use Lundstrom's result [78] of τ_t~66 ps which is the inverse of the attempt frequency for each transition from state to state. The total interface filling time t_c is the number of steps N multiplying the time for each transition τ as shown in Equation (5.7):

$$t_c = N\tau = \frac{E_g}{\Delta E}\tau_t \cdot e^{-\Delta E/kT} \tag{5.7}$$

Nature automatically optimizes to produce the shortest trap-filling time. A self optimization process is introduced by varying the number of cascade steps needed to cross the band gap to find the most probable interface trap-filling time, which is 11.8ps as shown in Figure 5.21.

With a ~10ps interface trap-filling time, the depth versus time behavior is in between the DP and SP camps but closer to the DP camp as shown in Figure 5.22. This suggests that the DP camp has been probing deep into the high-κ layer in their FDCP experiment, but not quite as deep as they thought.





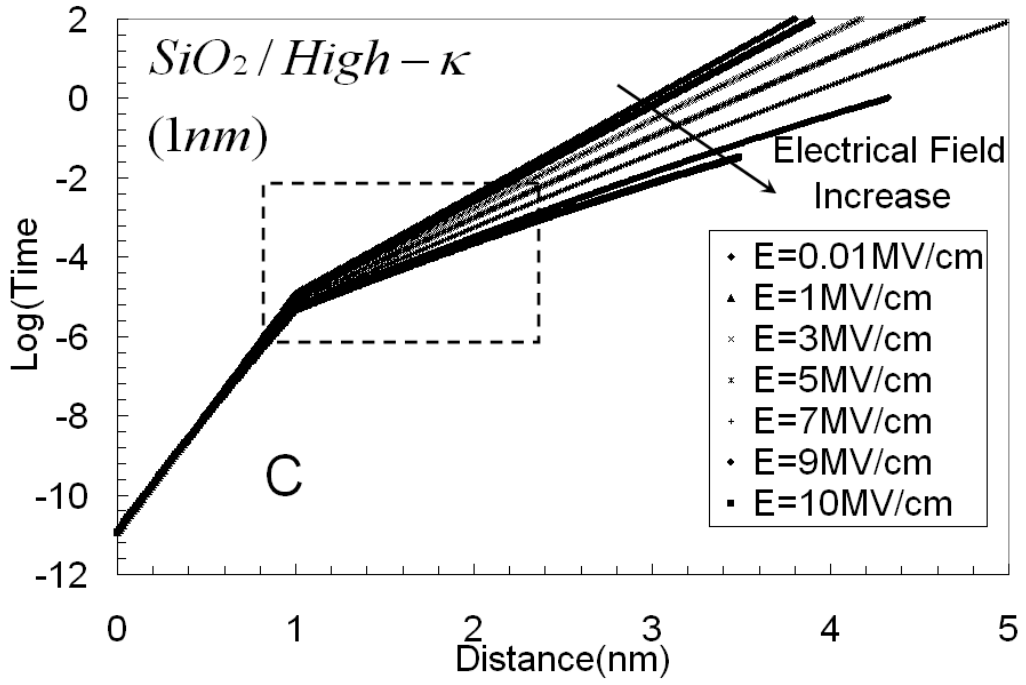


Figure 5.22, Similar as Figures 5.13 and 5.14 except that the two-step CP model with a new τ_0=10ps is used. The result is fairly similar to those in Figure 5.13 with a little shallower depth for the common frequency range.

5.11. Conclusion and Suggestion on Future Work

In summary, we have carried out a reliable charge pumping experiment beyond 1GHz with square wave for the first time. With that result, we draw a conclusion that the interface filling time must be less than 0.7 ns. This finding has resolved the probe depth controversy in the FDCP measurement. Based on our new experimental observation, we also introduce a self-optimized trap filling model to predict a more reasonable relation of CP probe depth as a function of frequency. This work is significant and timely to the current concern of bulk trap study in high κ. With our new model, the study of defect generation in high- κ dielectric is on a more solid ground.

Moreover, there are few suggestions on future extension of this work. For example, , it is possible to apply this new model and characterize the spatial trap distribution of some high κ samples. Further more, one can compare the trap distribution in the similar test structures with SiO_2 and high κ to study the defect generation mechanism.

The proposed new theoretical model can be further improved as well. For example, it is possible to calculate the exact energy level distribution of traps introduced by the bonding. Even though there is some similar type work in the literature, the careful treatment of this reconstruction problem is always a wonderful thing. It will make this model more convincing.

Chapter 6

Ballistic Phonon Enhanced NBTI

The defect generation in high-κ is one of the reliability concerns. Another reliability degradation in nano-scale CMOS technology is Negative Bias Temperature Instability (NBTI). The general conclusion of NBTI is: when a PMOSFET is stressed with negative bias and elevated temperature, threshold voltage is linearly increased with stress time. The amount of shift is believed to increase with higher electrical field and temperature. It is a very serious concern and may cause the performance degradation (e.g. timing or power) or even an unrecoverable malfunction in fabricated chip during operation. As a result, it is becoming more important to characterize its effect. Under the normal test methodology, device is stressed under negative gate bias with other electrodes grounded. However, many circuit application require transistors to operate at high gate drain bias in addition to a high gate bias, especially for analog and RF circuits. Therefore, it is critical to understand the effect of drain bias on NBTI degradation mechanism and its impact on the circuit reliability at the operation condition. There are several reports on this topic [37-41,119]. Many groups have studied with long channel device and drain bias as high as gate bias, which NBTI is greatly accelerated under channel hot carrier (CHC) effect[38,39,119]. Recently, there are two different groups reported NBTI with various drain bias [37,41]. They observed suppressed NBTI at low drain bias and turn around behavior at high drain bias. However, we observed the drain bias enhanced NBTI even at low drain bias

condition. In this chapter, we will show our experimental results as well as the explanation. It is a new mode of NBTI and can only be explained by local hot spot phenomenon. Specially, a localized hot spot due to the accumulation of high density of high energy phonon is formed at the drain with the size of 10-20 nm and temperature over 400K. Since NBTI is very temperature sensitive, this induced high temperature by drain bias becomes very worrisome. It is a mechanism unique for nano-scale transistors and expected to become more serious for future generation of device with shorter channel and high frequency operation.

6.1. Basic Experimental Setup and Details

Our drain bias dependent NBTI experiment setup is illustrated in Figure 6.1.Different from conventional setup with grounded drain, a drain bias is applied in our experiment. Moreover, we don't heat the wafer and work it under room temperature. We study cases with either constant drain bias (static) or an AC pulse on drain (dynamic). It is also noticed that a 50Ω terminated probe is used on the drain side. It is exact the same probe used in GHz charge pumping experiment in chapter 5. It is designed here to minimize the reflection.

With high frequency drain pulse, a significant part of it might be reflected because of the impedance mismatch between the drain input of transistor and transmission line. With this probe, the mounted 50Ω chip resistor is in parallel with the channel resistance from transistor, which is around 1 KΩ at stress condition (V_g=-2V,

V_d=-1V). Therefore, the overall input impedance at the drain is close to 50 Ω matched to the transmission line.

Besides drain voltage, the gate is stressed with -2V for certain time with the source and substrate grounded. The stress is interrupted at fixed interval for measurement. Figure 6.1 shows the diagram of this "stress-sense" process. Three full I_S-V_G measurements are taken at 20 second interval between each measurement. The average of extracted threshold voltage from three measurements is recorded.

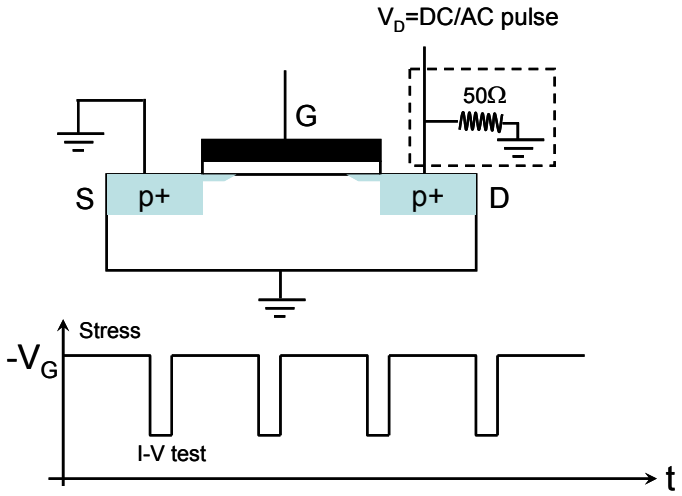


Figure 6.1 The experimental set up for NBTI study with drain bias. The gate is stressed and interrupted after certain time to measure the device I_D-V_G characteristics. The source and substrate are grounded while either a dynamic or static voltage is applied to drain. Here a 50Ω terminated probe is used at drain to minimize the reflection of high speed signal.

Typically, threshold voltage is extracted by measuring the drain current (I_D) as a function of gate voltage (V_G) instead of I_S-V_G here. In our study, because this special probe affects the impedance of the drain as well as the drain current, we measure the source current I_S instead. By sweeping the gate voltage from -0.8V to 0 V, the threshold voltage V_T can be extracted from I_S-V_G with -0.6V drain voltage and

source/substrate grounded. Therefore, the V_T measured here is the saturation threshold voltage $V_{T,SAT}$. This is a very important experimental detail because the slope of V_T shift in NBTI literature is very different between linear ($V_{T,LIN}$) and saturation($V_{T,SAT}$).

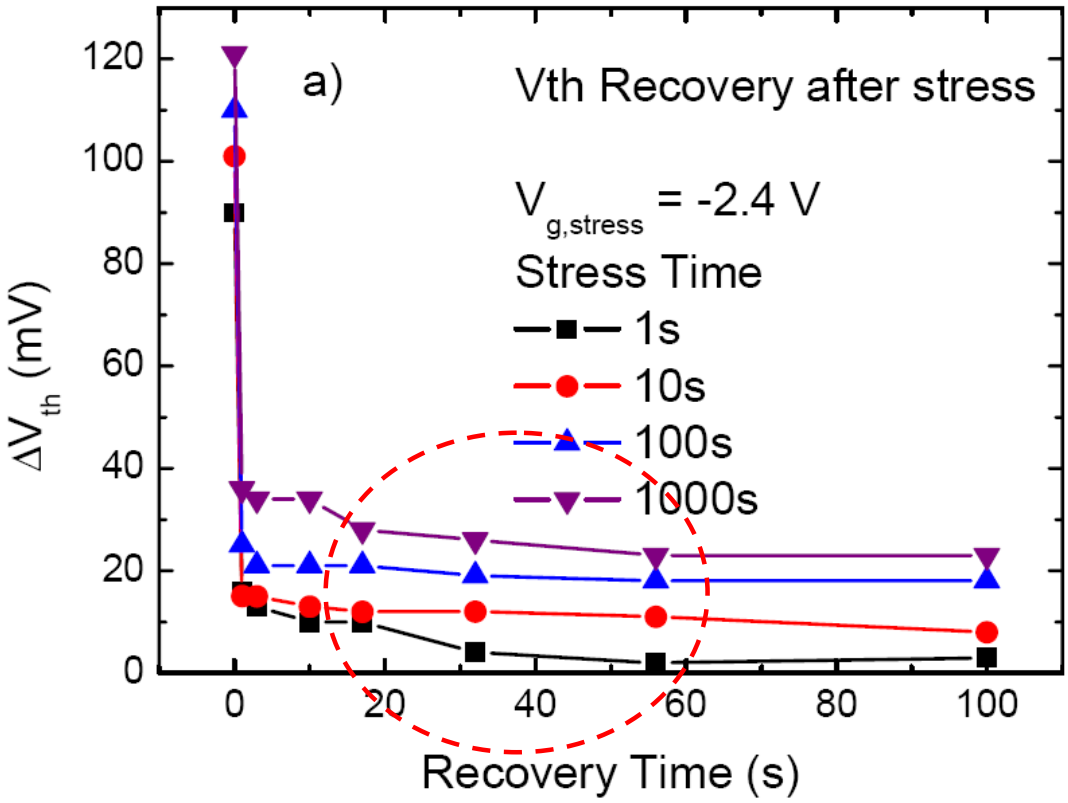


Figure 6.2 After stress is removed, the majority of ΔVth recovers in a very short time. Even after 1000 sec stress, >60% of the ΔV_{th} recovers within 1 sec after the stress is removed. This figure is taken from reference [120].

Furthermore, there is another very critical measurement detail - delay time between the end of stress and first *I-V* measurement. During this time, the stress voltage is over and measurement has not been taken. The created defects and V_T shift during the stress can be recovered by the hole de-trapping, which is proportional to delay time. Therefore, the extracted V_T shift is highly affected by the delay time and measurement

methodology. Figure 6.2 shows the percentage of recovery versus the time after stress. The figure is taken from reference [120]. It is noticed that the recovery process happens very fast initially and becomes relative steady after 10-20 seconds. Although many efforts have been made to shorten the delay time down to 1 μs, the initial fast recovery is still not very clear.

Since our propose is to compare the V_T shift with and without drain bias, we should put the focus on the measurement repeatability. We use relative slow V_T measurement. The delay time between the end of stress and our first *I-V* measurement is 20 second. At that point, the initial fast recovery is largely over. Further more, we take the average of *I-V* measured at every 20 seconds further reducing the effect of recovery. The fact that the difference between three V_T measured at 20 sec, 40 sec and 60 sec after stress is very small (within the noise level) suggests that the recovery is very negligible. Taking the average of these three measurements further enhances our measurement repeatability.

6.2. Drain Bias Dependent NBTI

Besides the drain bias NBTI experiment using the above setup, we also carry out normal NBTI without drain bias for comparison. In this study, EOT 1.6 nm SiON P-MOSFET from 90nm technology node is used as the test structure. Figure 6.3 is the standard log-log plot of our results showing V_T shift versus time. It is done on the

device with 2 μm channel width and 50nm physical gate length. Four traces are exhibited in the figure. Two of them are normal NBTI (without drain bias) at two different temperatures, which are 25C (room temperature) and 125 C respectively. Clearly, the result can be fitted with a power-law function with the exponent of 0.137 and 0.179. It is in good agreement with the published NBTI exponent around 0.1~0.18 for SiON [121-122].

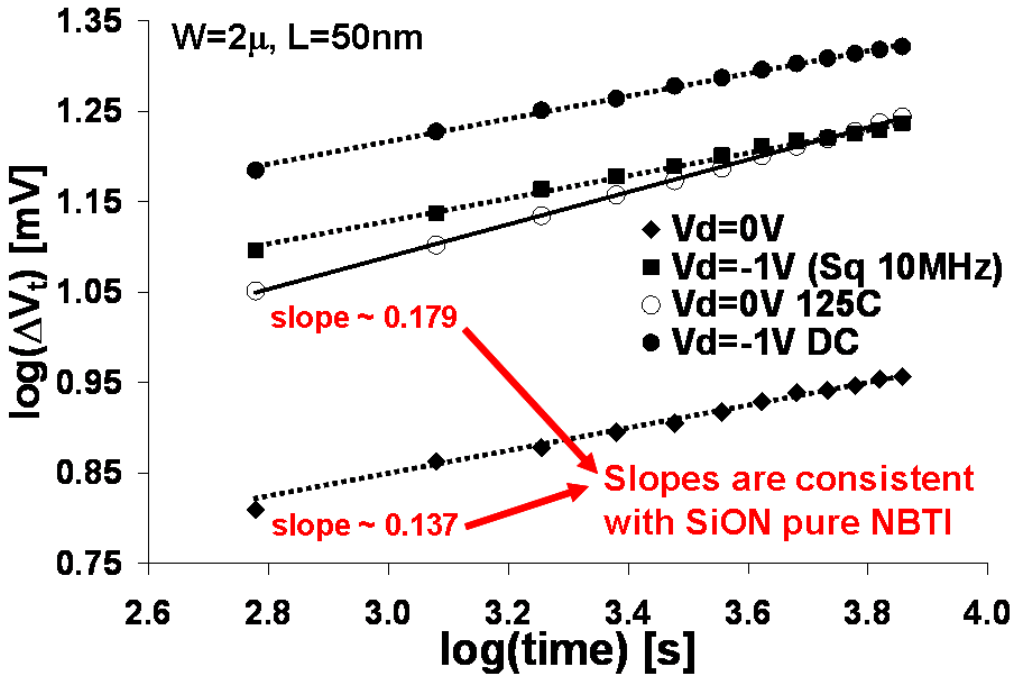


Figure 6.3 Log-log plot of our NBTI results showing V_T shift versus time. Four traces are shown here. Two of them are normal NBTI(without drain bias) at different temperature 0C and 125C. The other two are NBTI with drain bias but room temperature. Both DC voltage and a pulse with 50% duty cycle drain bias condition are studied.

This time dependence can be explained by a widely accepted reaction-diffusion(R-D) model [123-125]. It believes that diffusing hydrogen species is generated during NBTI stress and recombines at the silicon/oxide interface. Thus charged interface

states are created and will be responsible for the threshold voltage shift. In this model, power law dependence (t^n) down to $t^{1/4}$ is predicted and agrees very well with the experimental observation (~0.2-0.3) in the early NBTI studies on thicker oxide [124-125]. For thinner oxide, the theoretical model itself becomes rather illusive and complicated. But in general, it is reported that the slope is around 0.1-0.18 for SiON dielectrics. Clearly, our data is pretty consistent. Moreover, our data also suggests a temperature dependent exponent- higher slope for higher temperature. It is also observed by other groups before [121-122].

The success of our normal NBTI gains us the confidence and offers a way to check our measurement system. The other two curves in Figure 6.3 are the results of NBTI with drain bias but room temperature. Two drain bias conditions are used: -1V DC voltage and a pulse with 50% duty cycle. Compared to normal NBTI at 125 C, even though under the room temperature, NBTI with -1V drain bias has much more degradation or V_T shift. Even for drain bias with 50% duty cycle, there is more or comparable V_T shift than normal high temperature NBTI. These are very important findings. It is well known that NBTI is sensitive to temperature- more V_T shift for higher temperature. The fact that NBTI with drain bias but room temperature has higher degradation than normal NBTI (without drain bias) at elevated temperature suggests that drain bias must have some significant impact to overcome the external strong temperature effect.

This mystery will be taken away its mask step by step in the following sections. First, it is necessary to show that our observation of drain bias dependent is new, different from the one previously reported.

6.3. Compared to Literature

The drain bias dependent NBTI experiment is not new. Many groups have reported many experimental results on this subject [37-41,119]. Some of important conclusion is briefly summarized here. Most of previous drain bias dependent NBTI experiments are done with long channel device and drain bias as high as gate bias where channel-hot-carrier (CHC) induced degradation is strong. Under high drain bias with influence of high lateral fields in short-channel MOSFETs, carriers in the channel can gain sufficient energy to surmount the energy barrier or tunnel into the oxide. This impact ionization process leads to injection of a gate current into oxide and subsequently to the trap generation causing the device degradation. This is so called channel hot carrier (CHC) effect. It happens at high drain bias condition and its degradation has higher slope than NBTI.

Indeed, these previous experiments of NBTI with drain bias were carried out at the maximum CHC stress conditions ($V_g=V_d$). Therefore, the general conclusion is that the enhanced degradation is due to CHC effect dominating over the NBTI effect. There were differences in interpretations, but they were either CHC interacting with

NBTI [37, 38] or CHC and NBTI coexist [119]. Either way, the real cause of the enhanced degradation was clearly the existence of a strong CHC effect.

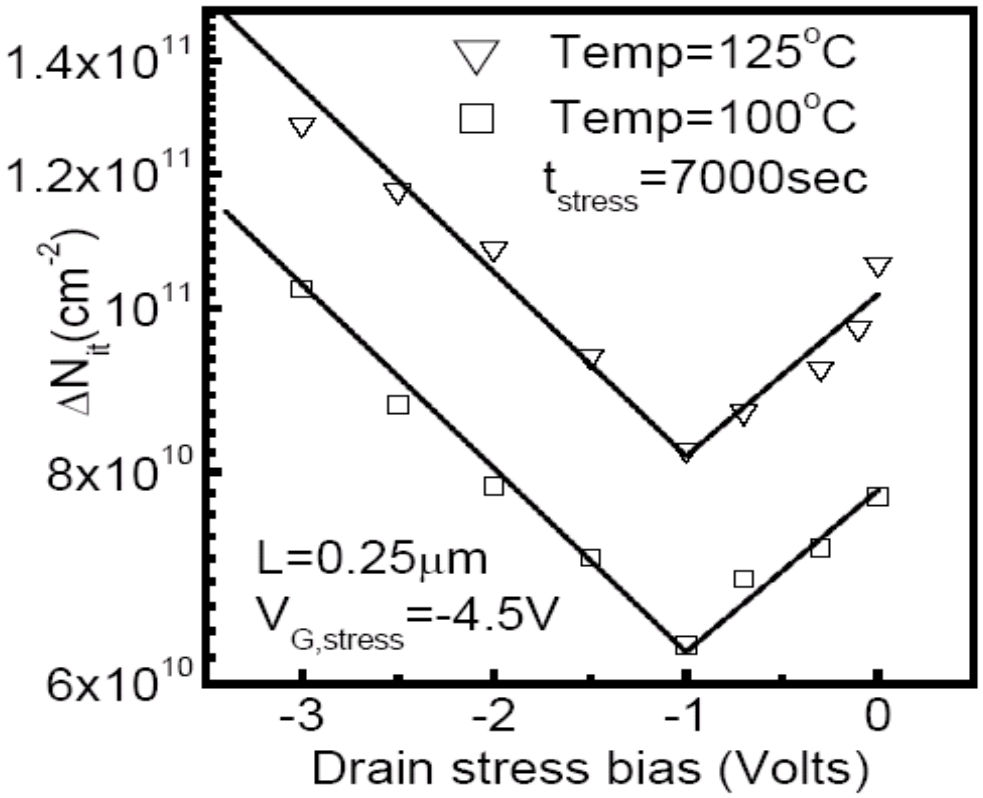


Figure 6.4 The reported NBTI result with various drain bias. The turn around behavior is happened at V_D=-1V. The data is taken from reference [37].

More recently, two different groups reported NBTI with full range of drain bias [37, 40-41]. One of the results [37] is taken and shown in Figure 6.4. It was observed that as drain bias increases from zero to -1V, NBTI degradation was ***suppressed*** progressively. However, moving beyond -1V drain bias, a turn around is observed and NBTI degradation is enhanced beyond certain drain bias. This observation is consistent with the CHC explanation. At drain bias below -1V, the population of hot holes energetic enough to overcome the oxide barrier is so low that CHC effect is

negligible. Since drain bias also reduces the vertical field (oxide field) in the channel, a reduction in NBTI degradation is reasonable. As drain bias increases, CHC increases quickly and eventually dominates NBTI, leading to accelerated degradation.

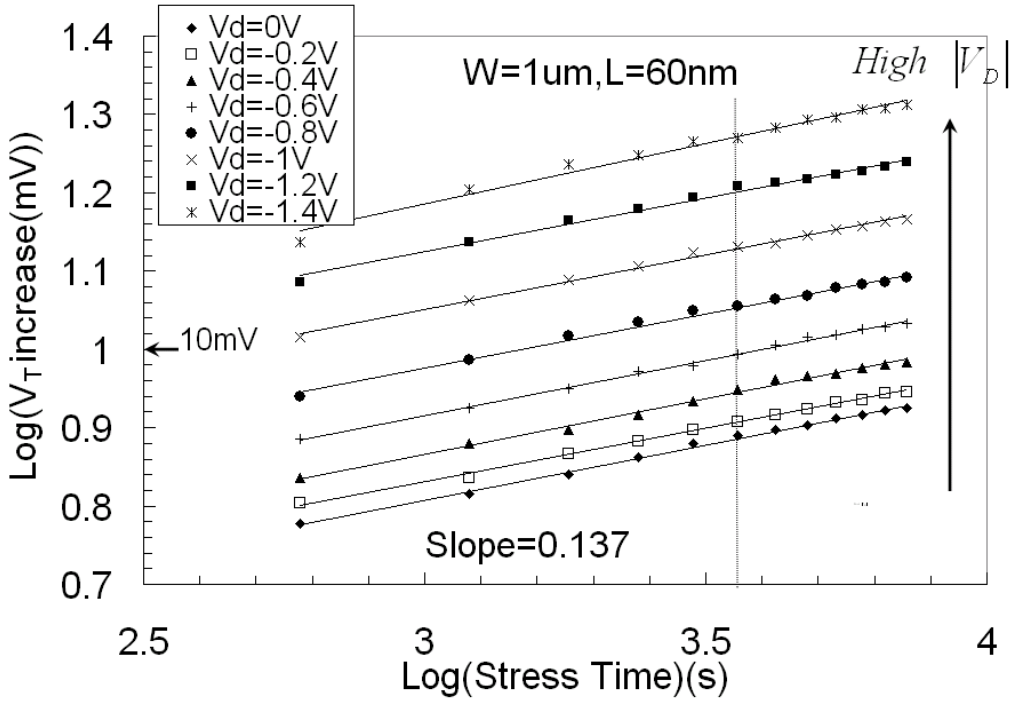


Figure 6.5 NBTI degradation in the customary log-log plot of threshold voltage shift versus stress time for various drain bias from zero (conventional NBTI condition) to -1.4V at room temperature. A monotonic increase in degradation with increasing drain bias is evident. Note the identical slope (0.137) for all the degradation trends. All transistors have 60nm physical gate length.

The reported turn around point is at roughly -1V, or when the drain bias starts to be larger than the silicon band gap. At that point, the strongest suppression of NBTI was observed [37, 40-41]. Contrary to these observations, our data showed only accelerated degradation even at very low drain bias. Figure 6.5 shows the accelerated degradation for various drain biases (at room temperature). Notice the DC voltage from 0V to -1.4 V with 0.2 V voltage step is used as drain bias. Most of the drain bias is in the range of 0V to -1V where the CHC is very negligible. Clearly, for the

condition with higher drain bias, there is always more degradation instead of turn around behavior reported before. Even at drain bias as low as -0.2V, clear acceleration of degradation is evidenced. Clearly, the entire observed trend is opposite to previous results in the drain bias range of 0 to -1V. This provides strong evidence that our observation is new, not the same phenomenon as previously reported.

6.4. Rule Out Possible CHC

Before finding the explanation of our new drain bias dependent NBTI experimental data, it is more carefully to discount the remote possibility that we have a set of transistors that are particularly prone to CHC effect. Here is some evidence to rule out that CHC is not the mechanism.

Evidence I – Poor Fit with Exponential Function

While our data shows enhanced degradation at drain bias as low as -0.2V, literature data shows that the enhanced degradation only happens with at least -2V drain bias [37, 40-41]. Since CHC originates from impact ionization and impact ionization depends exponentially on drain bias voltage, the change in CHC effect from -2V to 0V is very large. Figure 6.6 shows impact ionization rate (measured as substrate current) as a function of drain bias at 2V gate bias for 0.2μm n-channel transistor. (Data extracted from [126]).

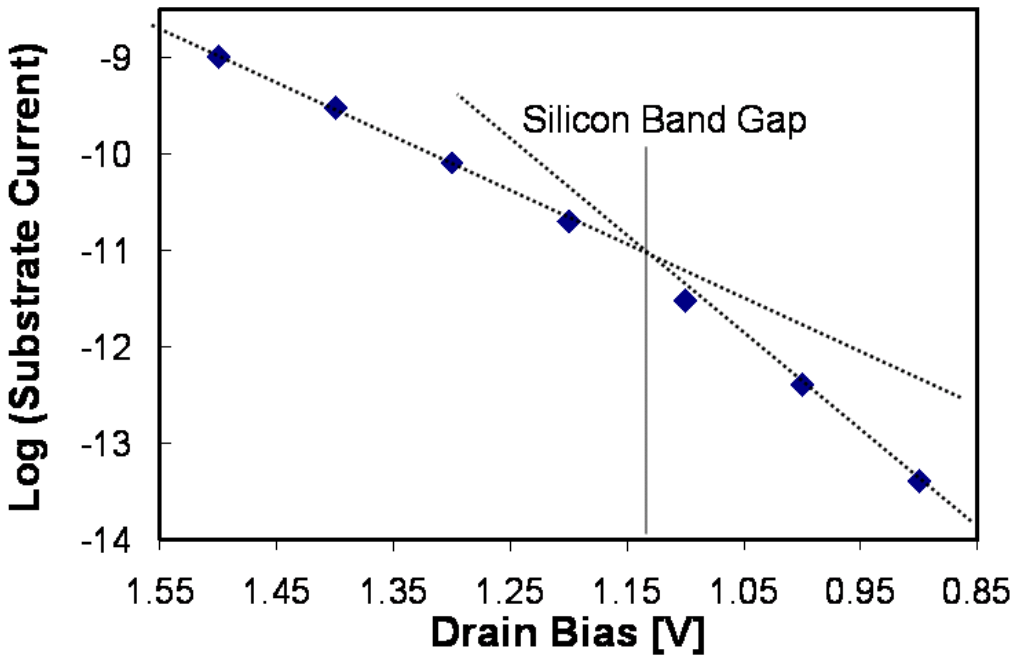


Figure 6.6 measured substrate current due to impact ionization as a function of drain bias at 2 V gate bias. For drain bias larger than silicon band gap energy, the impact ionization rate changes at roughly six orders of magnitude per volt. For drain bias less than band gap energy, the rate changes roughly ten orders of in magnitude per volt. Data extracted from [126].

For above band gap drain bias, impact ionization rate change by roughly six orders of magnitude per volt. Below band gap, the change increases to ten orders of magnitude per volt. Thus the CHC effect at -2V can be as much as 14 orders of magnitude larger than that at -0.2V. It is also clear that in the drain bias range of Figure 6.5, CHC effect is completely negligible and the observed enhancement in NBTI has nothing to do with it.

Further evidences supporting our conclusion that our observed enhanced degradation is not due to CHC effect can be found in the data itself. If the CHC is the responsible mechanism, the degradation is proportional to impact ionization rate, which is

exponentially dependence on drain voltage. Figure 6.7 plots the degradation after 3,600 seconds of stress as a function of drain bias.

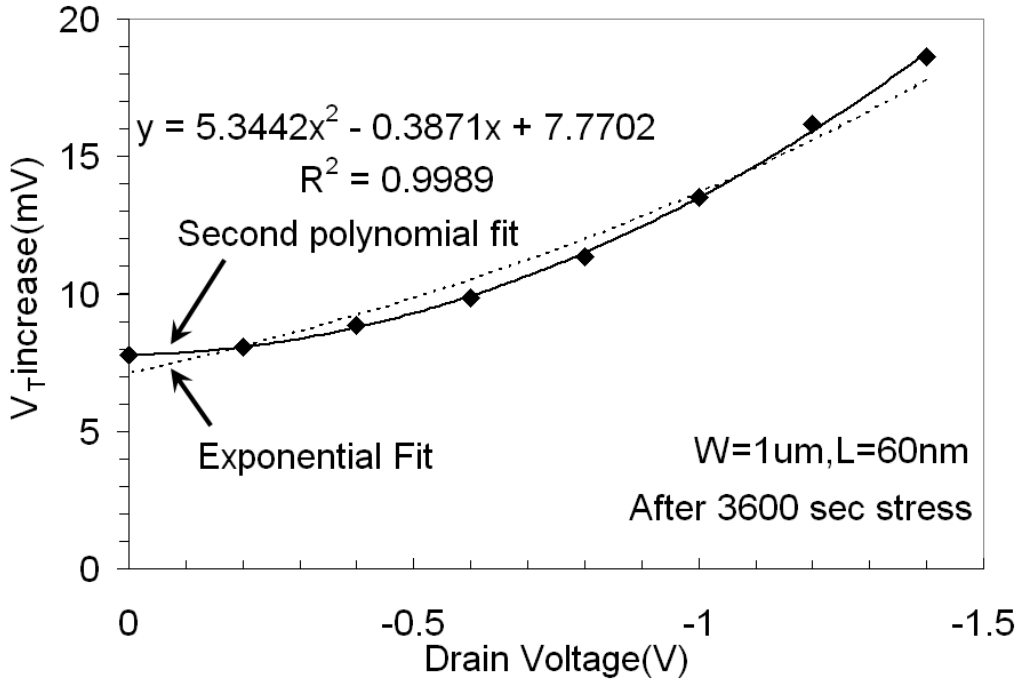


Figure 6.7, Threshold voltage increase as a function of stress drain bias after 3,600 second stress with V_G=-2V. The experimental results are very poorly fitted by exponential function (dashed line). It can only be fitted by second order polynomial curve (solid line).

Had our enhancement of degradation been due to CHC, we should expect the increase in degradation depends on drain bias exponentially. Attempts to fit the data in Figure 6.7 to an exponential function (dashed line in the figure) completely failed, arguing against significant contribution from CHC effect.

Evidence II – Constant slope in NBTI with various V_D

From our experimental data, there is another evidence to support that CHC is not the right explanation. In Figure 6.5, the slope is constant for various drain bias. It is well

known that while the signature of NBTI and CHC may be similar, the power-law exponent or slope for the CHC is twice as high as that for NBTI [127]. If CHC plays an important role like in all previous studies, one expects the power law exponent to at least increase from pure NBTI value if not double. Moreover, if there is additional degradation contributed by CHC at higher drain bias, the slope will increase with drain bias. Data in Figure 6.5 clearly show that the power law exponent remains the same with or without drain bias, even though the degradation is significantly increased with drain bias. It is a very clear indication of negligible CHC.

Evidence III – Data from NBTI with square wave V_D

Another proof can be found from the square wave drain bias data. Figure 6.8 shows the percentage of threshold voltage increase as a function of stress time for L=100nm (60nm physical gate length) pMOSFET under -2V gate bias and various drain bias including DC at 0V and at -1V, 10MHz square wave with 0V to -0.5V and 0V to -1V swing.

Instead of standard log-log plot, this data is drawn in a linear scale. Point A in this figure is the case with -1V DC drain bias at 10,000 second while point B is the condition with 50% duty cycle square wave drain bias from 0V to -1V swing but double stress time (20,000 seconds). Therefore, the total stress time for these two points with -1V drain bias on is the same. NBTI is known to have relaxation behavior

while CHC is not. The induced degradation by CHC only depends on the total stress time with drain bias on. Therefore, if CHC dominates the increase in degradation, then one expects the same degradation at these two points. Obviously, this is not the case shown in Figure 6.8.

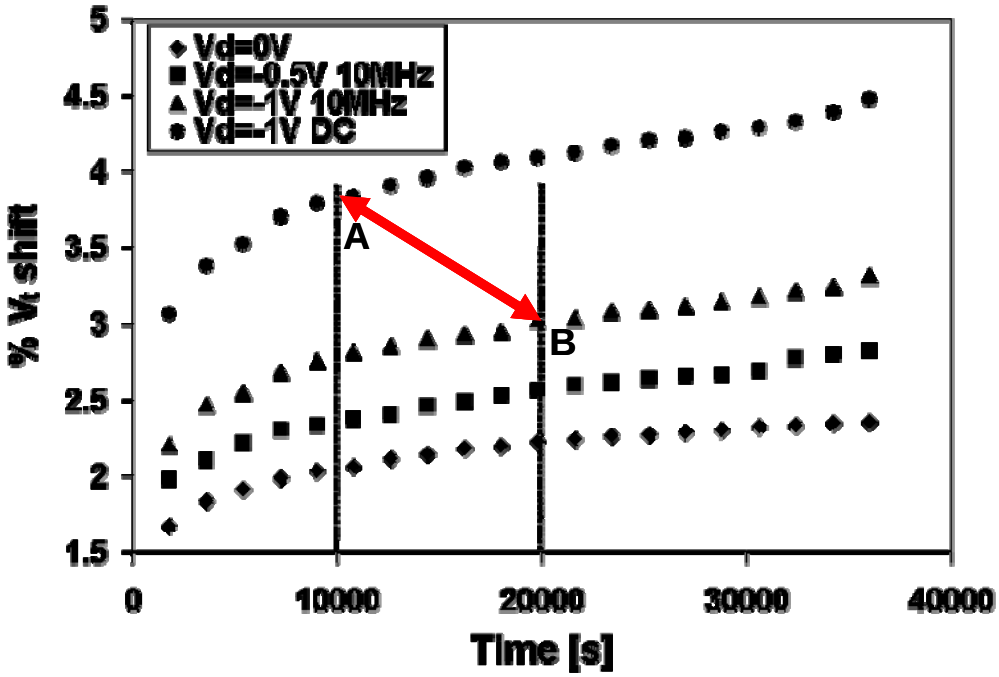


Figure 6.8, Percentage of threshold voltage increase as a function of stress time in linear scale for 100nm pMOSFET under -2V gate bias and various drain bias including DC at 0V and at -1V, 10MHz square wave with 0V to -0.5V and 0V to -1V swing. All stresses are done at room temperature. Point A and B in this figure have the same stress time with drain bias on but different degradation.

With all the above evidences, we can safely conclude that our result is obtained under condition that the CHC effect is completely negligible and that the drain bias enhanced NBTI degradation observed in our experiment is different from all previous similar studies.

6.5. Possible Explanation- Temperature from Localized Hot Spot

Let us briefly review what we have from the last couple sections. It is observed (Figure 6.3) that NBTI degradation with -1V drain bias but room temperature is larger than normal NBTI without drain bias but high temperature. It is a new finding and very different from the one previously reported. Historically, drain bias dependent NBTI is believed to relate with channel hot carrier (CHC). After carefully compared to literature, our observation is very different- enhanced instead of suppressed degradation at low drain bias. Moreover, from the data itself, there is sufficient evidence to support that CHC is not the proper explanation. Then a natural question will be: what could be the cause?

Clearly, the answer must be associated with the drain bias. NBTI is well known to be dependent on both temperature and electrical field. It is even more sensitive with temperature than electrical field. The only effect of drain bias is the reduction of electrical field leading to suppressed NBTI. Clearly, it is not the root responsible for our enhanced NBTI result. The only possibility left is the drain bias induced channel heating leading to high temperature than expected. Moreover, Figure 6.3 indicates that the degradation with drain bias is more than the one without it but at 125 C higher temperature. Therefore, the drain bias must introduce sufficient high temperature, higher than 125 C. Is that possible to happen?

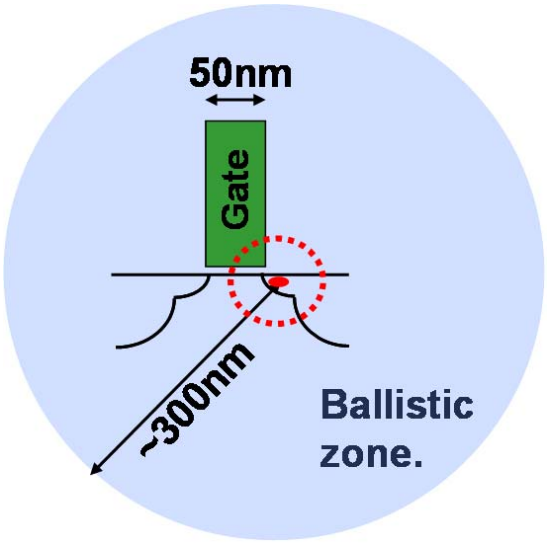


Figure 6.9 The drain bias induced a heating source (red point in the figure) and localized hot spot (dashed circle) leading to high temperature in the channel.

After careful literature search, it is found that there is only one possible explanation- hot spot phenomenon [42-44]. The basic idea is explained here. Figure 6.9 shows the heat dissipation around the drain junction for 50nm gate length transistor. The red point in the figure represents the heating source introduced by the applied drain bias. According to the classical picture, the heat is dissipated through diffusion from high temperature region (heating source at drain end) to the rest of cold region (channel and source end). The diffusion equation by Fourier law is widely used to handle the temperature in this case and only predicts very negligible temperature rise in the channel (the dashed line in Figure 6.10).

However, this type of equation only validates when the dimension for concern is much larger than the phonon mean free path (MFP). For silicon, the MFP at 300K is around 300nm [41]. In this experiment, device with 50 nm physical gate length is

used. It is much smaller than the MFP. Therefore, the classical Fourier equation becomes inappropriate to handle the problem. That is because the phonon becomes ballistic within a mean-free-path distance from the point of creation. Therefore, their distribution is not in thermal equilibrium which breaks the assumption of Fourier law.

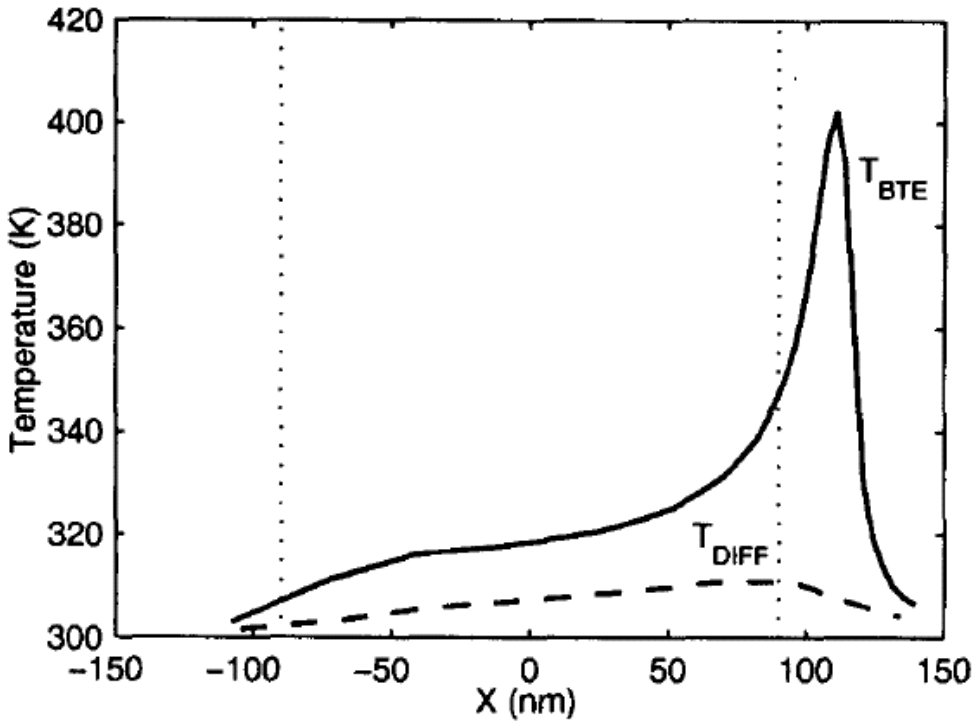


Figure 6.10 Temperature distributions along the channel of a 180 nm MOSFET. The solid line represents the BTE solution averaged over the channel depth and the dashed line the diffusion theory prediction. The vertical dotted lines represent the metallurgical source and drain junctions, respectively. This figure is taken from [42].

Realizing the invalidations of conventional diffusion method, theoreticians have made a lot effort and used the Bolzman Transport Equation (BTE) to treat this problem [42-43]. Figure 6.10 taken from reference [42] shows the simulation of temperature distribution along the channel of a 180 nm MOSFET. The solid line represents the BTE solution averaged over the channel depth and the dashed line the diffusion theory prediction. When modeled with BTE, it is found that a “temperature” rise of

over 100C resulted at 10 ~20 nm away from the heating source as shown in Figure 6.10. This is so called hot spot phenomenon. This big discrepancy in the predicted temperature has been used successfully by Sverdrup *et al.* [44] to explain the observed second breakdown phenomenon in an ESD experiment. In other words, the predicted high "temperature" led to thermal runaway just like regular high temperature.

For our case, the drain bias enhanced NBTI could be due to the effect of temperature- the temperature from hot spot. NBTI is very sensitive to temperature and it has higher degradation for higher temperature. When a drain bias is applied, the hot spot is formed leading high temperature in the channel. As shown in Figure 6.3, even under room temperature, -1V drain bias may heat up the channel to very high temperature and lead to more degradation than the one at 125C without drain bias. In this case, the effective channel temperature with 1V drain bias is higher than 125C. We use the word "effective" because the highly localized nature of the hot-spot. The temperature is expected to vary drastically across the length of the channel. This highly non-uniform temperature and highly non-uniform oxide field across the channel may be the reason why the drain bias induced degradation does not have the same slope as the conventional NBTI at high temperature.

This temperature explanation is also consistent with drain bias enhanced NBTI degradation data even at very low drain bias. Since the temperature originates from hot spot introduced by applied drain bias, it must be proportional to the input power

density of heating source. Higher drain bias will have higher drain current resulting in higher power density and higher temperature. It is also believed that this temperature should be strong functional dependent on drain bias because both voltage and current increase with higher drain bias.

Moreover, NBTI is more sensitive to temperature than electrical field. It is very possible that this strong temperature effect from hot spot overwhelms the one from electrical field. That is the reason why the degradation with drain bias gets enhanced even though the drain bias lowers the electrical field. For the previous drain bias dependent NBTI experiments in the literature [37-41,119], long channel transistors are used. The size of the hot spot is relatively small compared to the channel length. Therefore, only very small part of the channel is affected and average temperature rises is very negligible. That is the reason why the suppressed NBTI due to the reduced electrical field is observed in previous experiments at low drain bias. For our case, with a much shorter channel device, a monotonic but accelerating increase in degradation is observed for all drain bias. This is in stark contrast with previous reports where degradation was suppressed at low drain bias due to reduced oxide field in the channel. This suggests that the high temperature generated by the hot-spot is more than compensating the reduced oxide field.

Even though number of groups reported the predicted hot spot phenomenon [43-44,128-131], the magnitude and size of the hot spot varies drastically from model

to model [43-44,128-131]. The direct measurement of temperature is extremely difficult within this kind of spatial resolution [131-133]. Therefore, our data with the signature of temperature from hot spot is exciting because it offers the experimental evidence for the first time.

6.6. Physics of Localized Hot Spot Formation

It has been shown that our NBTI data can be explained by temperature effect generated by the applied drain bias. This temperature is very different from the traditional drain junction heating picture. Actually, drain junction heating is well known heat dissipation problem in high-performance chips such as CPU. Its thermal modeling is based on heat diffusion equation with continuum mechanics. The temperature is calculated through the amount of dissipated energy. When millions of transistors are working simultaneously, they dump a lot of power (up to 100 Watts). As a result, a lot of heat is generated easily leading to high temperature. That is the purpose of putting the cooling fan on the top CPU to dissipate the heat away.

However, when only a single transistor is working, it creates very little power. For the transistor used in our NBTI experiment with V_d=-1V and around 1mA drain current, the power is only 1mW. For a transistor operating with such small power dissipation, there is a very little, if any, temperature rise at the drain junction. Figure 6.11 taken from reference [134] shows the channel temperature variation versus the power

supply for a 0.12um gate length buried channel n-MOSFET SOI device. Both measured and simulation by diffusion equation is shown in the figure. Notice there is less than couple degrees of temperature rise at 1mW DC power.

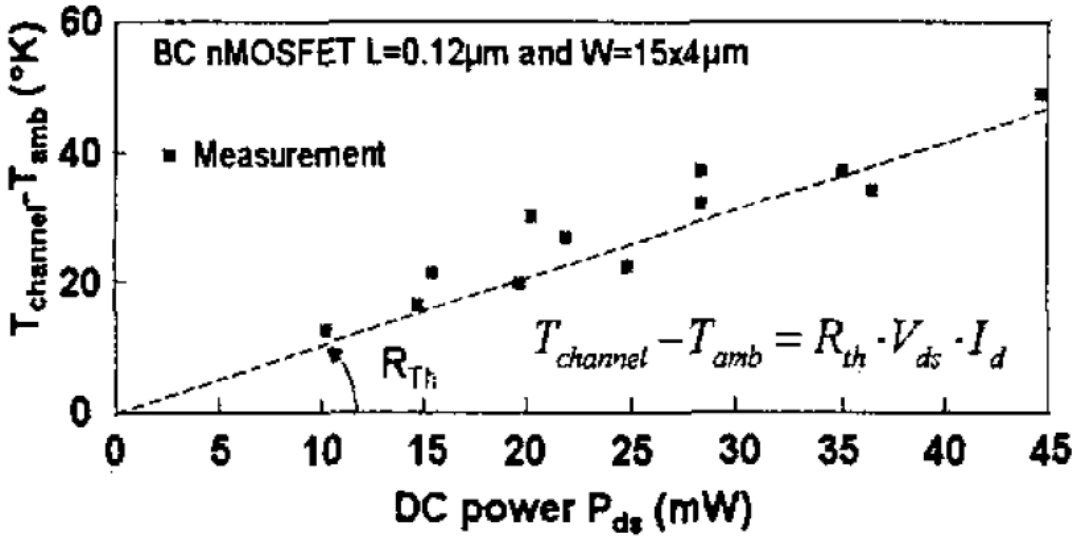


Figure 6.11 Channel Temperature variation versus the power supply for a 0.12um gate length buried channel n-MOSFET SOI device. Figure is taken from reference [134].

Moreover, this piece of data is collected on the buried channel SOI device which has much poorer thermal conductivity than the surface channel MOSFET from 90 nm technologies used in our experiment. Therefore, according to this traditional drain heating picture based on heat diffusion, there is very negligible temperature rise. Obviously, it can not explain the significant observed NBTI degradation.

Actually, we are looking at nano-scale picture and its corresponding phenomena-localized hot spot. Then, from the physical point of view, how the hot spot is formed? The hot-spot phenomenon is theoretically predicted as a result of two nanoscale effects: Nano source effect when the heat source is much smaller than the phonon

scattering mean-free-path (MFP); Ballistic phonon bottleneck effect at distance very close to the heat source. The phonon MFP in silicon at room temperature is ~300nm. When a MOSFET is turned on, the drain junction becomes a heat source that is much smaller than the phonon MFP. With a physical gate length of a few tens of nanometers, the entire transistor channel is at distance to the heat source much smaller than the phonon MFP. Thus we expect to have hot-spot due to both effects.

Figure 6.12 illustrates the physics and the expected time scale involved in the hot-spot phenomenon. When the drain bias is applied, the nano-scale transistor is turned on. Driven by the driven bias, the holes accelerate through the channel from the source to the drain. When reaching the destination, these holes can transfer their accumulated energy to the drain junction by creating predominantly optical phonons. These optical phonons can relax through conversion to acoustic phonons. At high power density, the generation rate exceeds the relaxation rate. Therefore, there is a high density of optical phonons accumulated.

Moreover, these optical phonons have low group velocity (sometimes referred to as non-propagating) and thus ineffective in conducting the energy (heat) away from the source. Therefore, at the region close to the heating source (drain), a highly localized, high density of high energy optical phonon is accumulated, which can be characterized as high temperature. This is the phonon bottleneck effect.

The conversion of optical phonon tends to create energetic acoustic phonons. Within the ballistic zone, the collision rate with the cooler phonons from the surrounding is limited, slowing down the establishment of thermal equilibrium. A non-equilibrium distribution that favors more energetic phonon represents a higher temperature. This is the nano source effect.

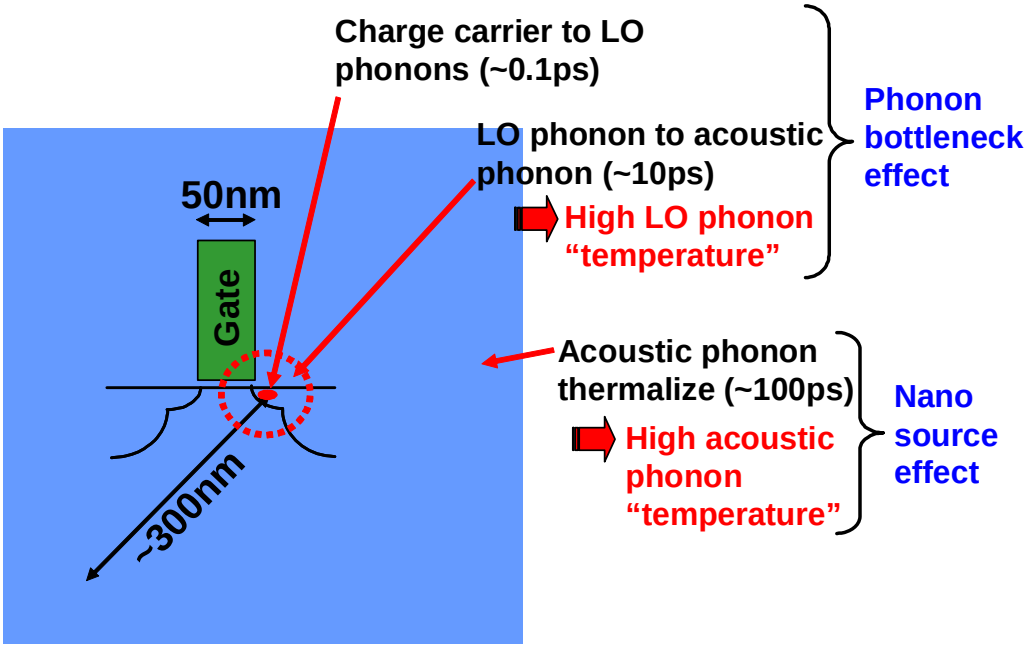


Figure 6.12 Illustration of the physics and time scale involved in the hot-spot phenomenon. The solid dot inside the drain region is the nano sized heat source. The broken circle is the localized hot-spot (the size varies from model to model). The entire transistor is small comparing to the phonon scattering mean-free-path (MFP).

6.7. Support of "Thermal Effect" - Channel Length Dependent NBTI

Besides the drain bias dependent NBTI data, it is better to have more evidence to support this new hot spot phenomenon. In what follows we will have certain expectation and the experimental results should be indeed consistent. If the enhanced

degradation is due to drain bias introduced “hot spot”, then it must increase with the drain current density. We can examine this effect by investigating the channel length dependent degradation. As the channel length is reduced, two factors accelerate the NBTI degradation.

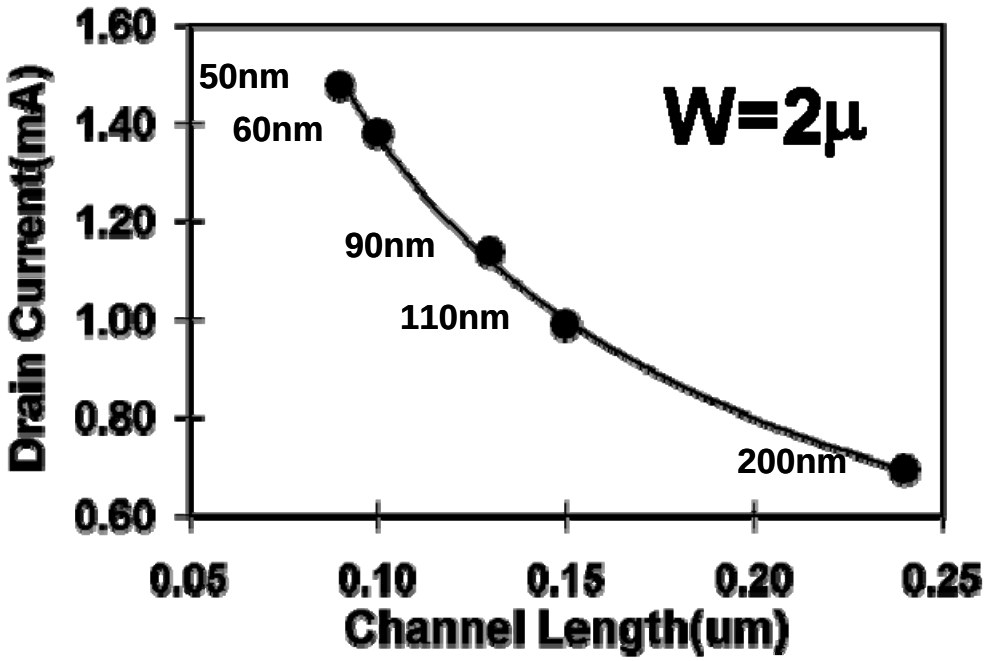


Figure 6.13, Drain current and therefore drain current density (channel width fixed) as a function of channel length. The X axis shows the designed channel length while the figure is marked with actual physical gate length.

Under the same gate and drain bias, shorter channel length leads to higher drain current density. Figures 6.13 show the drain current and therefore drain current density (channel width fixed) as a function of channel length at the stress condition (V_G=-2V, V_D=-1V). The X axis shows the designed channel length while the figure is marked with actual physical gate length. As clearly shown in figure 6.13, at fixed stress condition, higher drain current flows through the channel for short channel device. With the same applied drain voltage, higher drain current results in more input

power density of heating source.

If the temperature rise in the channel comes from the hot spot introduced by drain bias, it will increase with higher input power density for short channel device. It has been theoretically proposed and experimentally supported by Sverdrup *et.al* that the average lattice temperature rise within the drain junction hot spot is proportional to the input power density [43,44]. They also provided a simple formula for estimation of rise in temperature,

$$\Delta T = \frac{Q' \Lambda^2}{3 A_{eff} k_s} \tag{6.1}$$

Where Q' is the input power per device width(the *I-V* product, *I* being the current per unit width), Λ is the acoustic phonon mean free path, k_s the thermal conductivity of bulk silicon and A_{eff} the effective hot spot area. For device with the same channel width, according to this equation, higher input power density due to the higher drain current in shorter channel device results in higher temperature rise in the channel. Thus, more severe NBTI degradation is expected.

Secondly, in shorter channel device, the local hot spot covers relatively more percentage of the channel region. Boltzmann Transport Equation (BTE) simulations (Figure 6.14, taken from Svserdrup *et.al*'s work [43]) show that for the 100nm device, the hot spot region, defined as the area covering the temperature profile down to half of the peak temperature, covers around 33% of the channel. On the other hand, only 20% of the 240nm length channel is covered according to this simulation. More

percentage of the channel in the hot region results in higher channel temperature.

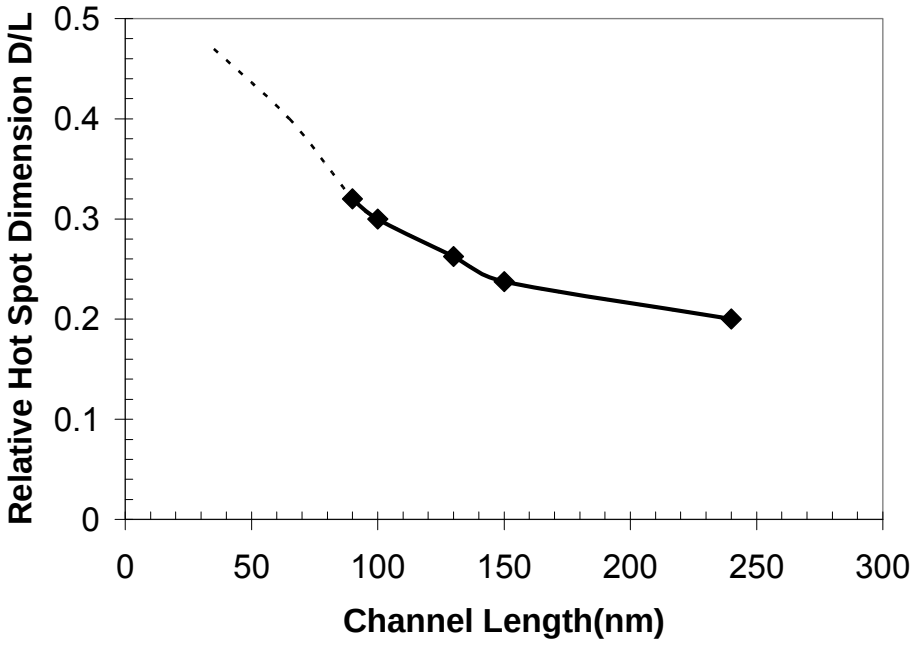


Figure 6.14 Ratio of characteristic hot spot dimension (D, along device channel) to channel length (L) of the test structure used to study channel length dependent NBTI. At shorter channel device, the size of local hot spot covers more percentage of the channel. The figure is taken from [43].

In summary, increased NBTI degradation will be expected in shorter channel devices because of higher temperature due to higher input power and more percentage of the channel being covered by the hot spot. This is exactly what we have observed in channel length dependent NBTI experiments. Figure 6.15 shows the channel length dependent degradation of NBTI under room temperature. This experiment is carried out on the device with fixed channel width (W=2 μm) but different channel lengths from 100nm to 240nm. For each device, both conventional NBTI (0V drain bias) and NBTI with 10MHz square wave with 0V to -1V swing is done.

Two sets of measurement conditions are used. Some data in this set are measured by

interrupting the stress gate voltage every 30 minutes. In order to save measurement time, some data were collected with more frequent interruption (10 minutes instead of 30 minutes) for measurement and shorter total stress time. The different interrupt frequency affects relaxation as well as the slope of the log-log plots. These data with more frequently interruption tend to have higher slope due to the relaxation phenomenon.

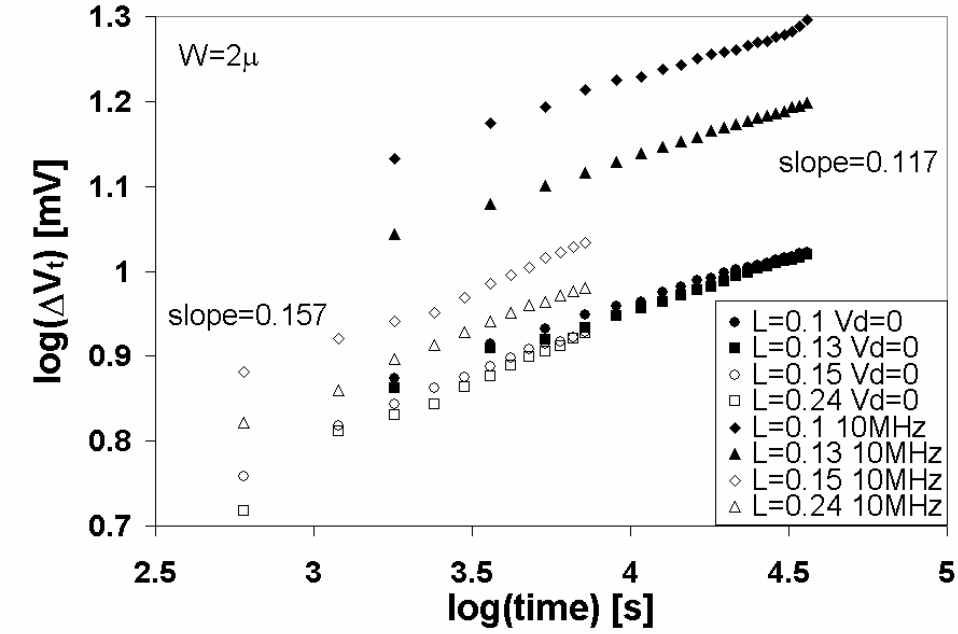


Figure 6.15 NBTI (room temperature) as a function of channel length. Two sets of measurement conditions are used. One interrupts the stress at every 30 minuets interval. The other at 10 minutes interval. The effect of more frequency interruption is a steeper slope in the log-log plot. Four different channel lengths are used. The drain bias is either 0V or 10MHz square wave with -1V amplitude. Other than the slope change due to changes in measurement interval, the 0V drain bias has no channel dependent degradation. A strong channel length effect is evident with drain bias

Other than the slope change due to changes in measurement interval, the basic trend is very clear. For conventional NBTI with 0V drain bias, there is no channel dependent degradation. It is clear that a strong channel length dependent degradation occurs only

when there is a drain bias presence. Shorter channel length results in a larger enhanced degradation, completely consistent with expectation from a “hot” spot interpretation.

6.8. Support of “Thermal Effect” - Channel Width Dependent NBTI

Next we look at the effect of channel width on the enhanced degradation. Although drain current increases linearly with channel width as shown in Figure 6.16, the drain current density remains the same.

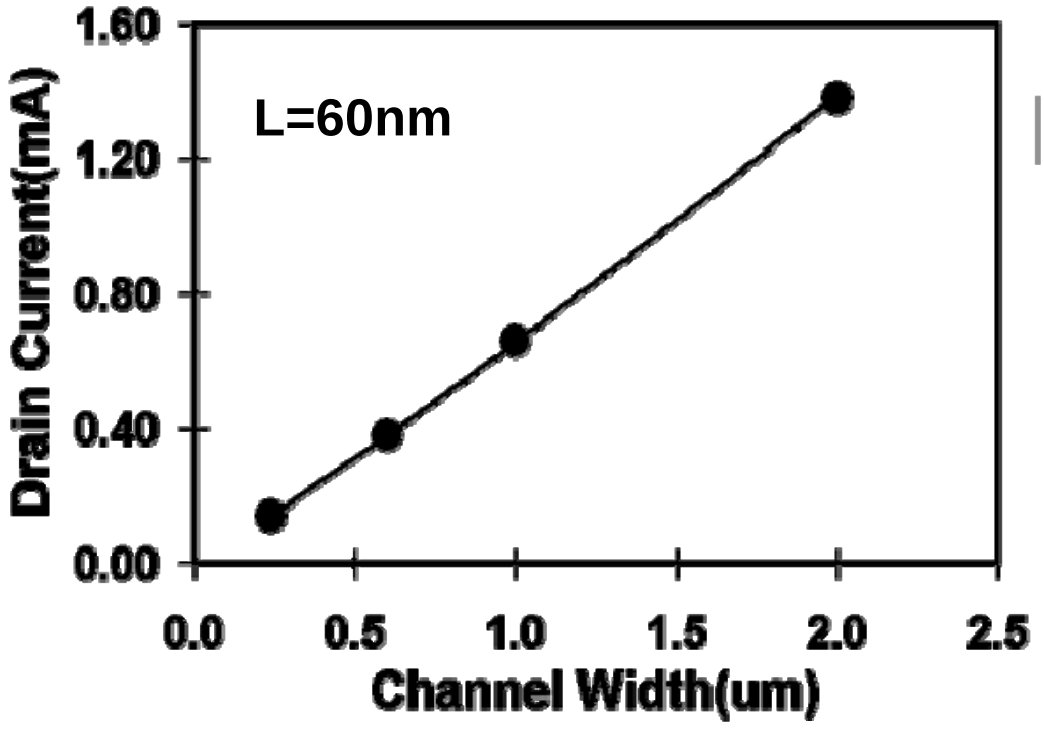


Figure 6.16 drain current as a function of different channel width from 0.25 μm to 2μm

At first glance, one would expect channel width to have no effects. However, if the enhanced degradation is indeed a "thermal" effect, then the shape of the thermal source can affect the resulting "temperature" at points near the source. This near-field effect predicts that the temperature will be higher for wider channel because narrow channel devices behaves closer to a point source while wider channel devices behave more like a line source.

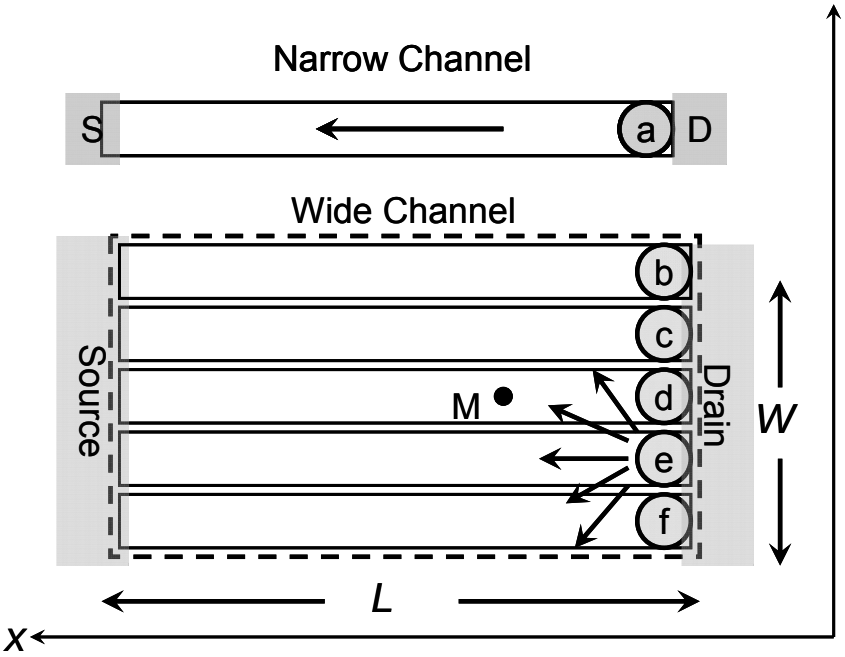


Figure 6.17 Illustration of heat dissipation for a wider channel device. It is the top view of the channel. Wider channel can be modeled as the integration of narrower channel. In a sufficient narrow channel device, the hot spot at drain can be modeled as a point heating source, where the heat can only dissipate along x-direction and get confined at y direction due to the narrow channel width. For a wide channel device, the point heating source becomes a line source or superposition of point sources. The heat dissipation in wider channel is 3-D case. So the point *M* in the channel can experience the temperature rise from the influence of all point heating sources from point *b* to *f*.

As the channel width increases, heat diffusion increasingly changes from a 3-D problem into a 2-D problem. As shown in Figure 6.17, in a narrow channel device, the local hot spot at drain junction (point *a*) is a point heating source. The heat can only

dissipate along the channel length direction (*x*-direction) but is confined by the limited channel width. On the other hand, the wide channel device can be modeled as the integration of many identical narrow channel devices. When many same hot objects without any heat interaction are put together, the overall temperature will be unchanged. Then the wide channel device is expected to have the same channel temperature as each narrow ones if no interaction is assumed.

However, in a wide channel device, heat dissipates not only along the channel length (*x*-direction) but also along channel width direction(*y*-direction). Therefore, any point in the channel (such as point *M* in Figure 6.17) not only absorbs the heat from heating source point *d* but also is able to get heat from all other heating sources (point *b* to *f*). The actual temperature rise at point *M* should be the superposition of the effect of all heating sources (point *b* to *f*). Therefore, the wide channel device with more point heating sources is expected to experience higher channel temperature leading to more severe degradation.

This is exactly what is observed in channel width dependent NBTI experiment. Figure 6.18 shows the channel width effect in pure NBTI (zero drain bias) as well as in enhanced NBTI with square wave drain bias with 0V to -1V swing. Four test structures with different channel width from 2 μm to 0.24 μm but fixed channel length (L=100 nm with 60 nm physical gate length) are used. A weak channel width effect may exist without drain bias but it is so small that one can easily make the opposite

conclusion. When a drain bias is presence, however, the channel width effect is much stronger, consistent with the "thermal" interpretation.

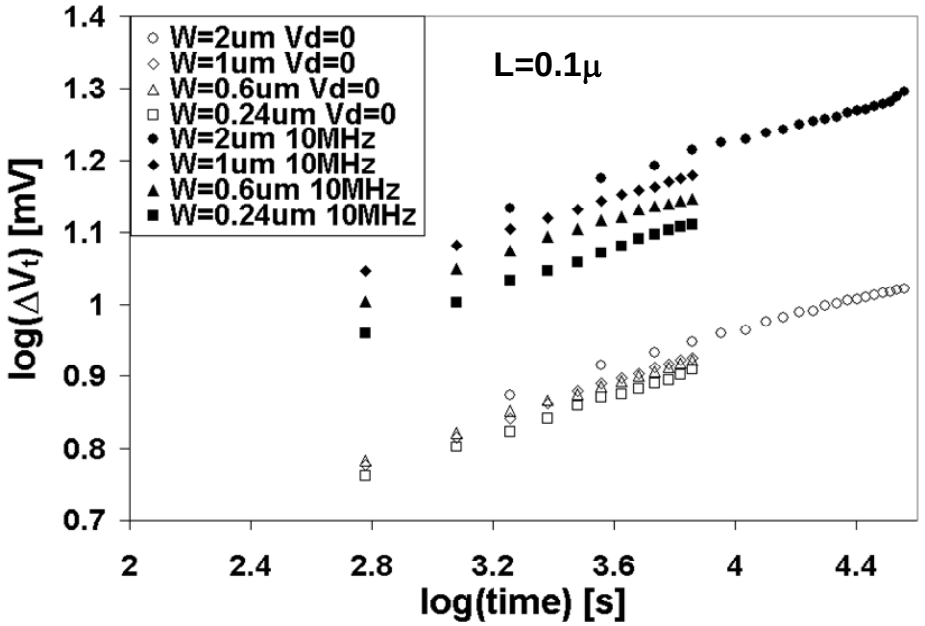


Figure 6.18 NBTI (room temperature) as a function of channel width with/without drain bias. A very weak channel width effect may exist in pure NBTI case (0V drain bias) while a much stronger channel width effect is more evident with the 10MHz and -1V amplitude square wave drain bias.

Comparing the channel width effect of Figure 6.18 and the channel length effect of Figure 6.15, the channel width effect is clearly weaker. This is also consistent with the "thermal" picture because the near-field effect in "temperature" is a secondary effect and therefore much smaller.

Furthermore, even our narrowest channel device is not really a point source because the width is already ~5 times the channel length. In another words, the channel width we are working on is much bigger than the size of hot spot. Even though more point

heating sources are included as the channel gets wider, the influence on the overall channel temperature is still very limited. For example, in Figure 6.17, the point heating source at the edge of channel (such as point *b*) has little impact on the temperature rise at point *M* since the distance between them is rather big compared to the size of hot spot. Thus we cannot demonstrate the full effect of changing from a 3-D heating problem (point source) to a 2-D heating problem (line source).

6.9. Support of "Thermal Effect" - Drain Bias Frequency Dependent NBTI

An even more powerful proof of the "thermal" effect is drain bias frequency dependent degradation enhancement. As discussed earlier that the presence of a drain bias should actually suppress NBTI by reducing the vertical field near the drain end. In the absence of CHC or "hot" spot effect, the reduced field leads to a reduced NBTI degradation. Thus when the drain bias is a square wave, the NBTI degradation should be smaller during the ON cycle and bigger during the OFF cycle, assuming that the "temperature" remains constant.

However, the "temperature" does vary during the full cycle of the drain bias and the variation depends on the duration of the cycle or the frequency of the drain bias. Figure 6.19 shows an illustration of how drain bias frequency affects the "temperature" evolution during a drain bias cycle. At lower frequency, the ON cycle reaches higher "temperature" because it has more time to build up and the OFF cycle

drops to lower “temperature” because it has more time to cool off. As frequency increases, the “temperature” swing reduces and eventually converges to just the average temperature.

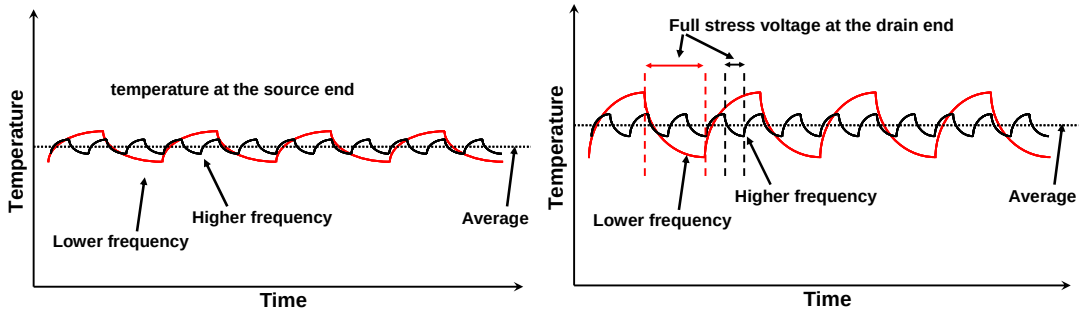


Figure 6.19, Temperature transients at the source end and at the drain end for two different frequencies. It shows a larger swing at lower frequency due to more energy is deposited per cycle. The average temperature at the source end is lower due to larger distance from the heat source. During the drain voltage on time, the channel to gate voltage is lower and NBTI decrease significantly. So the biggest effect is during the off time when the full electric field is at present. For lower frequency, the off time temperature is lower and the NBTI is smaller.

One should keep in mind that “temperature” is not uniform in the channel and the size of the swing is different in different part of the channel. The mechanisms responsible for NBTI at source and drain are somewhat different.

At the source end, being further from the drain, heat has to dissipate from the hot spot at drain to raise the temperature at source. So the temperature modulation at source is damped down and the average temperature is also lower (Figure 6.19). Since NBTI is very sensitive to temperature, the lower average temperature could make degradation at the source end much smaller than drain even though it is being stressed at full gate bias at all time.

While at the drain end, the situation becomes a little bit complicated. The drain bias not only introduces the hot spot leading to high temperature but also reduce the oxide field. Both temperature and oxide field effect is playing a role. During the ON cycle, although the “temperature” is higher, the vertical field is lower. During the OFF cycle, even though the “temperature” is lower, the full vertical field is present. Thus, in either ON or OFF cycle, there is opposing force at work. As long as they don’t end up canceling each other exactly, a frequency effect can be expected. If the “temperature” swing is small, then the vertical field is more important and we should expect an increase in degradation with drain bias frequency. Since the “temperature” swing decreases with frequency, the experiment must be carried out at high frequency range. (Note that at low frequency, the analysis is more complex because each cycle is long and the power-law behavior of degradation must be accounted for.)

To apply a high frequency drain bias, an important issue to avoid is ringing due to impedance mismatch. Any ringing will change the effective drain bias voltage and completely overwhelm the frequency effect. We used the same home-made terminated probe as used in GHz charge pumping experiment.

Figure 6.20 shows the drain bias frequency effect on the enhanced NBTI degradation on test structure with 50nm physical channel length and 2um channel width. Drain bias at all frequency has fixed amplitude of -1V. The modulations were either square wave at lower frequency or sine wave at higher frequency. The square waves were symmetric and therefore have 50% duty cycle. A frequency dependent degradation is

clearly seen. As shown in the figure, 250MHz sine wave and 50MHz square wave produced almost identical result. Thus we can think of the sine wave as square wave at a factor of 5 lower in frequency. With that in mind, we see that the degradation increases monotonically with frequency but show sign of saturation, as expected from the "temperature" swing convergence.

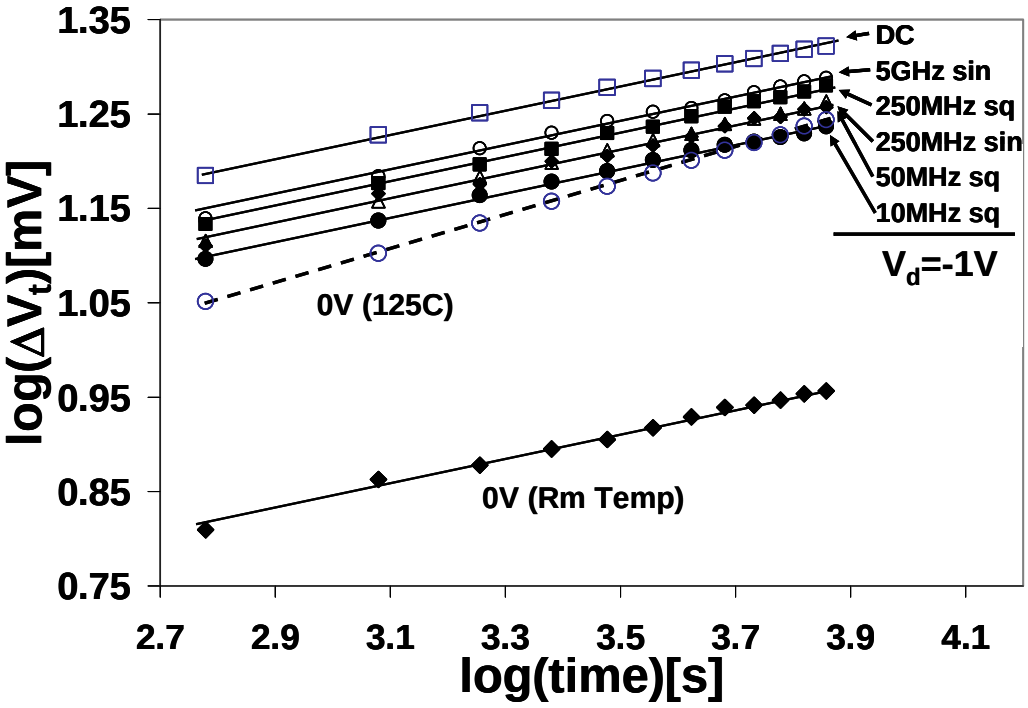


Figure 6.20, Room temperature NBTI degradation at -1V drain bias with various modulation frequencies. Also shown are the conventional NBTI (zero drain bias) degradations at both room temperature and 125C. The trend lines are there to high-light each set of data. Only one trend line is used for both the 250MHz sine wave and the 50MHz square wave data sets because they basically overlap each other. A saturation trend is evident as frequency increases. All transistors have 50nm physical gate length. Gate bias was -2V.

From Figure 6.20, we observe that the drain bias effect is a much stronger effect than the frequency effect. One way to explain this is that the hot-spot phenomenon creates the high temperature that produced the main effect. This effect is too fast to have frequency dependent observables. As mention before, the temperature must be high

enough that it produces a large increase in degradation even when the oxide field is reduced. The temperature profile in the channel as a result of the hot-spot is shown in Figure 6.21 (a).

The acoustic phonons will eventually thermalize and convention heat diffusion will take over. While the diffusion equation does not apply to the region smaller than the phonon MFP, it is not difficult to see that a temperature peak to exist at the origin (heat source). It is this temperature peak that is producing the frequency dependent degradation. The temperature profile in the channel as a result of this slower effect is shown in Figure 6.21(b).

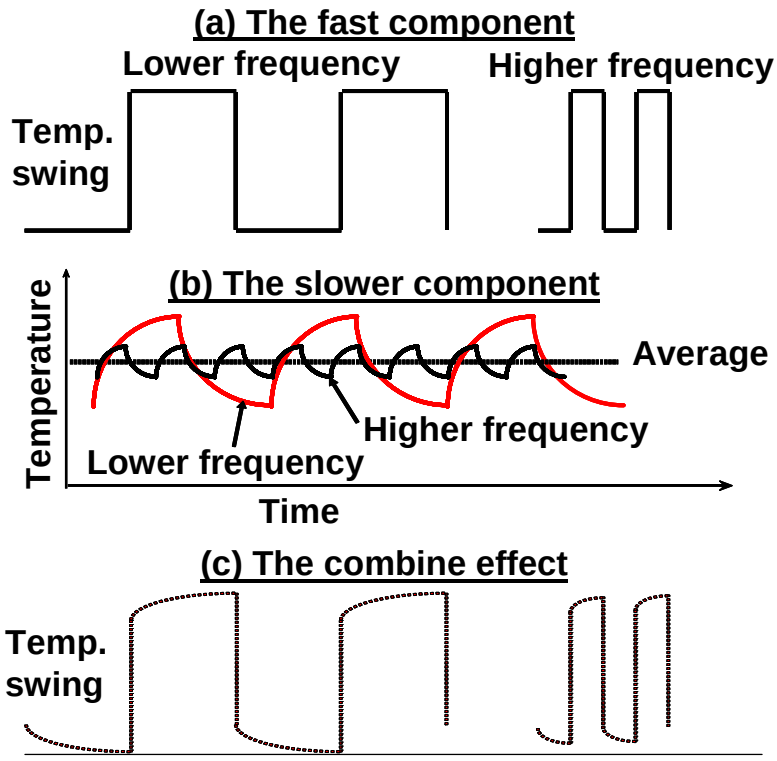


Figure 6.21 With drain voltage modulated by a square wave. The channel temperature is modulated as well. (a) The fast component of temperature due to the hot-spot effect. The temperature response is so fast that it follows faithfully the drain bias. (b) The slow component of temperature due to heat diffusion. The lower the frequency, the more time for temperature build up

during the ON cycle and more time for cooling during the OFF cycle. The temperature swing is larger for lower temperature. (c) The combined fast and slow temperature profile. The hot-spot effect is much larger than the heat diffusion effect.

Notice that the temperature swing is larger for lower frequency in Figure 6.21(b). During the ON cycle, channel temperature is higher at lower frequency. During the OFF cycle, the channel temperature is higher for higher drain bias frequency. If the temperature effect is larger than the oxide field effect, one expects a reverse dependent on frequency. Since we observed higher degradation at higher frequency, the diffusion controlled temperature peak must be small enough that the oxide field plays a dominant role.

Combining the hot-spot effect and the diffusion effect, the channel temperature profile takes on the form shown in Figure 6.21(c). It should be emphasized that temperature from the hot-spot effect is working against the oxide field reduction while the diffusion temperature is not affected by the oxide field (OFF cycle). So the actual temperature difference between the two effects is larger than seems to be suggested by data in Figure 6.21.

We like to emphasize that the drain bias frequency effect is not the AC (or dynamic) NBTI effect often reported in the literature [135-136]. In AC NBTI study, it is the gate voltage that is modulated. In our experiment, the gate voltage was fixed at -2V DC.

6.10. How high is the "temperature"?

With the channel length, channel width and drain bias frequency effects, we have strong supports for the "hot" spot interpretation of our observed drain bias enhanced NBTI degradation. The remaining question is how "hot" the "hot" spot is. Since we expect the "temperature" to be highly non-uniform across the channel, all we can do is to estimate the effective temperature by comparing the degradation to pure NBTI at elevated temperature.

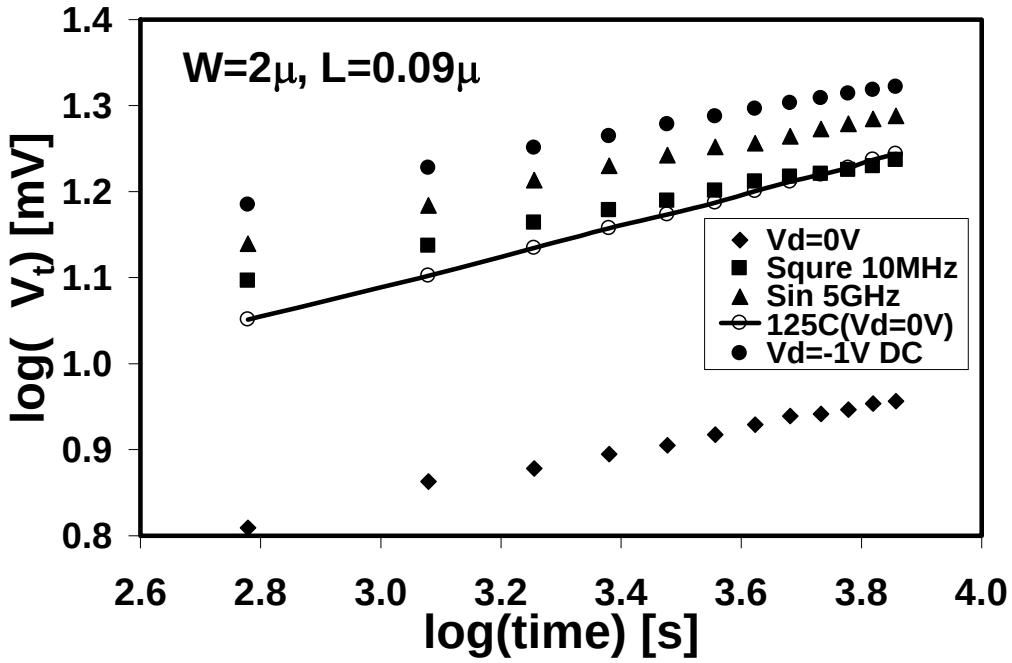


Figure 6.22 NBTI with and without drain bias. Two pure NBTI cases are shown, one at room temperature and the other at 125 C. The 125 C pure NBTI produces degradation similar to the 10MHz square wave drain bias case with a small difference of slope.

Figure 6.22 shows the NBTI degradation under various drain bias, including pure NBTI at 125C. Other than a small slope difference, the 125C pure NBTI curve is

similar to that of the 10MHz -1V square wave drain bias. From this bench mark, we can estimate that the effective “temperature” in the channel for the -1V DC drain bias case reaches between 150C to 175C, or a “temperature” rise between 130C to 155C.

The experimental result suggests, however, the drain junction has an effective “temperature” rise far exceeding 100C. Since we expect the “temperature” at the drain junction to be higher than the estimated effective temperature, the drain junction may have reached a “temperature” well over 200C. It should be noted that the heat diffusion equation is valid in the rest of the system, up to near the boundary of the ballistic zone. Thus the “temperature” must drop to less than a few degrees above room temperature shortly beyond the boundary of the ballistic zone. This is why the “hot” spot as defined by full width at half maximum is smaller than the ballistic zone.

6.11. Conclusion and Suggestion on Future Work

In summary, we have experimentally shown that the “hot” spot formed within the ballistic phonon zone of the drain junction can thermally enhance temperature sensitive transistor degradation modes such as NBTI. Although drain bias enhanced NBTI degradation was reported in the literature before, what we observed is a completely new phenomenon that has nothing to do with the CHC effect found to be responsible for all previously reported results. Since there is no other known possible source of drain bias related degradation, the drain bias accelerated degradation

represents a strong support to the hot spot phenomenon. We show that although phonon distribution is not thermal within the ballistic zone, thermal-like consequence still results. This newly discovered reliability issue is expected to become much more serious as the transistor size shrinks further and operation frequency increases. The phenomenon is quantum mechanic in origin and is entirely nano-scale specific.

Meanwhile, even though there is a lot of study in the hot spot phenomenon. The size of the hot spot has significant variation depending on the model. The model that predicts a highly localized hot spot also produces the highest temperature for the hot spot [43]. Our data seems to support a rather high local temperature model and therefore a highly localized hot spot. On the other hand, the hot spot is clearly extending into the transistor channel or we would not observe the accelerated NBTI degradation. However, our data cannot tell how large the hot spot is.

At the same time, even though the existence of numerous theoretical predication, there are not much experimental evidence. Our observed accelerated degradation of nano-scale transistor provides solid evidence that a localized hot spot with high temperature exist around a nano-scale heat source. We demonstrate the use of nano-scale transistor as a probe for heat conduction in nano-scale. Since the transistors can be made with high level of control and readily available from many advanced IC companies, our approach offers a practical mean to the fundamental study of heat conduction in the nano-scale.

Since this is the first piece of work characterizing this effect on device reliability in a detail manner, it is worth the effort for any follow up. As the 45 nm technology is in production today, it is more interesting to see the effect of this new reliability mode on shorter channel test structure. Moreover, it is also worthwhile to study the NBTI recovery as a function of relaxation time to explain the slope difference in results.

Chapter 7
Conclusion

The goal of this thesis has been to offer some solutions for the challenges met by advanced MOS device characterization. Specifically, we have introduced a high-accuracy method to measure C-V in the presence of high leakage current. It is based on a well established measurement technique, namely Time Domain Reflectometry (TDR). Because this new method is a radical departure to the basic scheme of traditional method, the effect of leakage current can be accurately corrected and precise C-V can be easily obtained. As an additional advantage, this new method also offers a simple and independent way to extract the series resistance as well as overlap capacitance with high precision. It can be expected that this simple measurement technique can be implemented as a routine device characterization procedure.

Besides this fantastic TDR method, we also use the combination of experiment and theory to resolve the probe depth controversy in the frequency dependent charge pumping (FDCP) measurements. We have reported the first charge pumping (CP) experiment beyond 1GHz with square wave and confirmed that the interface trap-filling time must be less than 0.7 ns. Based on this new experimental observation, we have introduced a new self-optimized trap filling model to predict a more reasonable relation of CP probe depth as a function of frequency. With our new model,

the study of defect generation in high-k dielectric is on a more solid ground.

Moreover, as the reliability becomes more important with device scaling down, we have found a new mode of reliability concern- ballistic phonon effect. It is a specific nanometer scale phenomena and has been studied in device reliability here for the first time. Using 90nm MOSFET test structure, we have observed that it induced channel heating and greatly enhanced the negative bias temperature instability (NBTI) degradation when the drain bias is applied. In addition, its effects with different channel length, channel width and drain bias frequency have also been thoroughly characterized. The conclusion is very worrisome. It implies that this new mode of reliability will become much more serious as channel length decreases.

Reference

[1] *International Technology Roadmap for Semiconductors (ITRS),* 2005 Edition [http.www.itrs.net], Semiconductor Industry Association.

[2] B. H. Lee, J. Oh, H. H. Tseng, R. Jammy, and H. Huff, "Gate stack technology for nanoscale devices," *Materials Today*, vol. 9, pp. 32-40, 2006.

[3] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S. H. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H. J. C. Wann, S. J. Wind, and H. S. Wong, "CMOS scaling into the nanometer regime," *Proceedings of the IEEE*, vol. 85, pp. 486-504, 1997.

[4] J. E. Chung, M.-C. Jeng, J. E. Moon, P.-K. Ko, and C. Hu, "Performance and reliability design issues for deep-submicron MOSFETs," *IEEE Transactions on Electron Devices*, vol. 38, pp. 545-554, 1991.

[5] C. Hu, "Future CMOS scaling and reliability," *Proceedings of the IEEE*, vol. 81, pp. 682-689, 1993.

[6] K. J. Yang and C. Hu, "MOS capacitance measurements for high-leakage thin dielectrics," *IEEE Transactions on Electron Devices*, vol. 46, pp. 1500-1501, 1999.

[7] G. Ghibaudo and R. Clerc, "Characterization and modeling issues in MOS structures with ultra thin oxides," *Proceedings, 24th International Conference on Microelectronics*, Vol. 1, pp103-114, 2004

[8] A. Nara, N. Yasuda, H. Satake, and A. Toriumi, "Applicability limits of the two-frequency capacitance measurement technique for the thickness extraction of ultrathin gate oxide," *IEEE Transactions on Semiconductor Manufacturing,* vol. 15, pp. 209-213, 2002.

[9] W. K. Henson, K. Z. Ahmed, E. M. Vogel, J. R. Hauser, J. J. Wortman, R. D. Venables, M. Xu, and D. Venables, "Estimating oxide thickness of tunnel oxides down to 1.4 nm using conventional capacitance-voltage measurements on MOS capacitors," *IEEE Electron Device Letters*, vol. 20, pp. 179-81, 1999.

[10] K. Ahmed, E. Ibok, G. C.-F. Yeap, Q. Xiang, B. Ogle, J. J. Wortman, and J. R. Hauser, "Impact of tunnel currents and channel resistance on the characterization of channel inversion layer charge and polysilicon-gate depletion of sub-20-angstrom gate oxide

MOSFET's," *IEEE Transactions on Electron Devices*, vol. 46, pp. 1650-1655, 1999.

[11] D. W. Barlage, J. T. O'Keeffe, J. T. Kavalieros, M. M. Nguyen, and R. S. Chau, "Inversion MOS capacitance extraction for high-leakage dielectrics using a transmission line equivalent circuit," *Electron Device Letters, IEEE*, vol. 21, pp. 454-456, 2000.

[12] C.H. Choi, J.-S. Goo, T.-Y. Oh, Z. Yu, R. W. Dutton, A. Bayoumi, M. Cao, P. Vande Voorde, D. Vook, and C. H. Diaz, "MOS C-V characterization of ultrathin gate oxide thickness (1.3-1.8 nm)," *IEEE Electron Device Letters*, vol. 20, pp. 292-294, 1999.

[13] W. J. Zhu, M. Khare, J. Snare, P. R. Varekamp, S. H. Ku, P. Agnello, T. C. Chen, and T. P. Ma, "Thickness measurement of ultra-thin gate dielectrics under inversion condition," *International Symposium on VLSI Technology, Systems, and Applications. Proceedings of Technical Papers, 18-20 April 2001*, pp. 212-215, 2001.

[14] G. A. Brown, "Capacitance characterization in integrated circuit development: the intimate relationship of test structure design, equivalent circuit and measurement methodology," *Proceedings of the International Conference on Microelectronic Test Structures (ICMTS)*, pp. 213-217, 2005.

[15] D.Rideau, P. Scheer, D. Roy, G. Gouget, M. Minondo, A. Juge, (2003). “Series resistance estimation and C-V measurements on ultra thin oxide MOS capacitors.” *IEEE International Conference on Microelectronic Test Structures* (ICMTS), pp191-196. 2003

[16] J. Schmitz, F. N. Cubaynest, R. J. Havens, R. De Kort, A. J. Scholten, and L. F. Tiemeijer, "Test structure design considerations for RF-CV measurements on leaky dielectrics," *IEEE International Conference on Microelectronic Test Structures*, pp. 181-185, 2003.

[17] Z. Luo and T. P. Ma, "A new method to extract EOT of ultrathin gate dielectric with high leakage current," *IEEE Electron Device Letters,*, vol. 25, pp. 655-657, 2004.

[18] H. T . Lue, C. Y. Liu, and T. Y. Tseng, "An improved two-frequency method of capacitance measurement for SrTiO3 as high-k gate dielectric," *IEEE Electron Device Letters*, vol. 23, pp. 553-555, 2002.

[19] A. Teramoto, R. Kuroda, M. Komura, K. Watanabe, S. Sugawa, and T. Ohmi, "Capacitance-Voltage Measurement Method for Ultrathin Gate Dielectrics Using LC Resonance Circuit," *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, pp. 43-49, 2006.

[20] B. H. Lee, L. Kang, W.J. Qi, R. Nieh, Y. Jeon, K. Onishi, and J. C. Lee, "Ultrathin hafnium oxide with low leakage and excellent reliability for alternative gate dielectric application," *IEEE International Electron Devices Meeting*, pp. 133-136, 1999.

[21] W. Qi, R. Nieh, B. H. Lee, L. Kang, Y. Jeon, K. Onishi, T. Ngai, S. Banerjee, and J. C. Lee, "MOSCAP and MOSFET characteristics using ZrO2 gate dielectric deposited directly on Si," *IEEE International Electron Devices Meeting*, pp. 145-148, 1999.

[22] L. Manchanda, W. H. Lee, J. E. Bower, F. H. Baumann, W. L. Brown, C. J. Case, R. C. Keller, Y. O. Kim, E. J. Laskowski, M. D. Morris, R. L. Opila, P. J. Silverman, T. W. Sorsch, and G. R. Weber, "Gate quality doped high K films for CMOS beyond 100 nm: 3-10 nm Al_2O_3 with low leakage and low interface states," *IEEE International Electron Devices Meeting*, pp. 605-608, 1998.

[23] R. M. Wallace, "Challenges for the characterization and integration of high-K dielectrics," *Applied Surface Science*, vol. 231-232, pp. 543-551, 2004.

[24] G. Ribes, J. Mitard, M. Denais, S. Bruyere, F. Monsieur, C. Parthasarathy, E. Vincent, and G. Ghibaudo, "Review on high-k dielectrics reliability issues," *IEEE Transactions on Device and Materials Reliability*, vol. 5, pp. 5-19, 2005.

[25] A. Kerber, E. Cartier, L. Pantisano, M. Rosmeulen, R. Degraeve, T. Kauerauf, G. Groeseneken, H. E. Maes, and U. Schwalke, "Characterization of the Vt-instability in SiO2/HfO2 gate dielectrics," *IEEE International Reliability Physics Symposium,*, pp. 41-45, 2003.

[26] M. Declercq and P. Jespers, "Analysis of interface properties in MOS transistors by means of charge pumping measurements," *P.Rev HF Acta Techn Belgium*, vol. 9, pp. 244-53, 1974.

[27] Y. Maneglia, D. Bauza, and T. Ouisse, "Extraction of the slow oxide trap density in MOS transistors using the charge pumping method," *International Semiconductor Conference, CAS'95 Proceedings*, pp. 155-158, 1995.

[28] R. Degraeve, A. Kerber, P. Roussell, E. Cartier, T. Kauerauf, L. Pantisano, and G. Groeseneken, "Effect of bulk trap density on HfO2 reliability and yield," *IEEE International Electron Devices Meeting 2003,* pp. 38-5, 2003.

[29] C. D. Young, S. Nadkarni, D. Heh, H. R. Harris, R. Choi, J. J. Peterson, J. H. Sim, S. A. Krishnan, J. Barnett, E. Vogel, B. H. Lee, P. Zeitzoff, G. A. Brown, and G. Bersuker,

"Detection of electron trap generation due to constant voltage stress on high-k gate stacks," *International Reliability Physics Symposium*, pp. 169-173, 2006.

[30] Y. Miura and Y. Matukura, "Investigation of silicon–silicon dioxide interface using MOS structure.," *Japanese Journal of Applied Physics*, vol. 5, pp. 180, 1966.

[31] A. Goetzberger and H. E. Nigh, "Surface charge after annealing of Al–SiO2–Si structures under bias," *Proceedings of IEEE*, vol. 54, pp. 154, 1966.

[32] A. Goetzberger and A. D. Lopez, "On the formation of surface states during stress aging of thermal Si–SiO2 interfaces," *Journal of Electrochemistry Society*, vol. 20, pp. 90-96, 1973.

[33] B. E. Deal, M. Sklar, A. S. Grove, and E. H. Snow, "Characteristics of the surface-state charge (Qss) of thermally oxidized silicon," *J Electrochem Soc*, vol. 114, pp. 266-274, 1967.

[34] D. K. Schroder and J. A. Babcock, "Negative bias temperature instability - Road to cross in deep submicron silicon semiconductor manufacturing," *Journal of Applied Physics*, vol. 94, pp. 1-18, 2003.

[35] J. H. Stathis and S. Zafar, "The negative bias temperature instability in MOS devices - A review," *Microelectronics Reliability*, vol. 46, pp. 270-286, 2006.

[36] N. Kimizuka, T. Yamamoto, T. Mogami, K. Yamaguchi, K. Imai, and T. Horiuchi, "The impact of bias temperature instability for direct-tunneling ultra-thin gate oxide on MOSFET scaling," *Symposium on VLSI Technology. Digest of Technical Papers,* pp. 73-74 ,1999.

[37] N. K. Jha and V. R. Rao, "A new oxide trap-assisted NBTI degradation model," *IEEE Electron Device Letters*, vol. 26, pp. 687-689, 2005.

[38] H. Aono, E. Murakami, K. Okuyama, K. Makabe, K. Kuroda, K. Watanabe, H. Ozaki, K. Yanagisawa, K. Kubota, and Y. Ohji, "NBT-induced hot carrier (HC) effect: positive feedback mechanism in p-MOSFET's degradation," *40th Annual Reliability Physics Symposium Proceedings*, pp. 79-85, 2002.

[39] B. S. Doyle, B. J. Fishbein, and K. R. Mistry, "NBTI-enhanced hot carrier damage in p-channel MOSFETs," *International Electron Devices Meeting*, pp. 529-532, 1991.

[40] P. Chaparala and D. Brisbin, "Impact of NBTI and HCI on PMOSFET threshold voltage drift," *Microelectronics Reliability*, vol. 45, pp. 13-18, 2005.

[41] P. Chaparala, J. Shibley, and P. Lim, "Threshold voltage drift in PMOSFETS due to NBTI and HCI," *IEEE International Integrated Reliability Workshop Final Report*, pp. 95-97,

2000.

[42] Y. S. Ju and K. E. Goodson, "Phonon scattering in silicon films with thickness of order 100 nm," *Applied Physics Letters*, vol. 74, pp. 3005-7, 1999.

[43] E. Pop, K. Banerjee, P. Sverdrup, R. Dutton, and K. Goodson, "Localized heating effects and scaling of sub-0.18 micron CMOS devices," *IEEE International Electron Devices Meeting.*, pp. 677-680, 2001.

[44] P. G. Sverdrup, K. Banerjee, C. Dai, W.-k. Shih, R. W. Dutton, and K. E. Goodson, "Sub-continuum thermal simulations of deep sub-micron devices under ESD conditions," *IEEE International Conference on Simulation of Semiconductor Processes and Devices*, pp. 54-57, 2000.

[45] Z. Liu, X. Lei, Y. Xuan, J. Wei, Z. Chen, L. Liu, and Z. Li, "The key technologies in silicon based microwave and RF MEMS device fabrication," *IEEE International Conference on Microwave and Millimeter Wave Technology Proceedings*, pp. 1-6, 2004.

[46] P. Sen, N. Srirattana, A. Raghavan, and J. Laskar, "Analysis of device scaling towards the performance enhancement of Si-MOSFET RF amplifiers," *IEEE 13th European Gallium Arsenide and other Compound Semiconductors Application Symposium*, pp. 221-224, 2006.

[47] J. Maserjian, “Tunneling in thin MOS structures,” *Journal of Vacumm Science Technology,* vol. 11, no. 6, pp. 996–1003, 1974.

[48] S.H. Lo, D. A. Buchanan, Y. Taur, and W. Wang, "Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFET's," *IEEE Electron Device Letter* ,vol. 18, pp. 209-211, 1997.

[49] W. K. Henson, K. Z. Ahmed, E. M. Vogel, J. R. Hauser, J. J. Wortman, R. D. Venables, M. Xu, and D. Venables, "Estimating oxide thickness of tunnel oxides down to 1.4 nm using conventional capacitance-voltage measurements on MOS capacitors," *IEEE Electron Device Letter* ,vol. 20, pp. 179-181, 1999.

[50] C. Bowen, C. L. Fernando, G. Klimeck, A. Chatterjee, D. Blanks, R. Lake, J. Hu, J. Davis, M. Kulkarni, S. Hattangady, and I.-C. Chen, "Physical oxide thickness extraction and verification using quantum mechanical simulation," *IEEE Electron Devices Meeting*, pp.869-872, 1997.

[51] C. L. Huang, J. V. Faricelli, and N. D. Arora, "A new technique for measuring MOSFET

inversion layer mobility," *IEEE Transaction on. Electron Device,* vol. 40, pp. 1134-1141, 1993.

[52] F. Lime, C. Guiducci, R. Clerc, G. Ghibaudo, C. Leroux, and T. Ernst, "Characterization of effective mobility by split C(V) technique in N-MOSFETs with ultra-thin gate oxides," *Solid-State Electronics*, vol. 47, pp. 1147-1153, 2003.

[53] E. A. Fogels and C. A. T. Salama, "Characterization of surface states at the Si-SiO2 interface using the quasi-static technique," *Journal of Electrochemstry Society,* vol. 47, pp. 2002-2007, 1971.

[54] A. Koukab, A. Bath, and E. Losson, "An improved high frequency C-V method for interface state analysis on MIS structures," *Solid-State Electronics*, vol. 41, pp. 635-639, 1997.

[55] A. F. Yaremchuk, "New interpretation of C-V measurements for determining the concentration profile in a semiconductor," *Applied Physics A (Material. Science. Processing)*, vol. 73, pp. 503-509, 2001.

[56] A. Pirovano, A. L. Lacaita, A. Pacelli, and A. Benvenuti, "Novel low-temperature C-V technique for MOS doping profile determination near the Si/SiO2 interface," *IEEE Transaction on. Electron Device,* vol. 48, pp. 750-756, 2001

[57] Y. Okawa, H. Norimatsu, H. Suto, and M. Takayanagi, "The negative capacitance effect on the C-V measurement of ultra thin gate dielectrics induced by the stray capacitance of the measurement system," *International Conference on Microelectronic Test Structures*, 2003.

[58] N. Yang, W. K. Henson, J. R. Hauser,J J. Wortman, "*Modeling study of ultrathin gate oxides using direct tunneling current and Capacitance–Voltage measurements in MOS devices*", *IEEE Transaction on. Electron Device,* vol. 46, no.7,pp. 1464-1471, 1999

[59] M. Depas, B. Vermeire, P. W. Mertens, R. L. Van Meirhaeghe, and M. M. Heyns, "Determination of tunneling parameters in ultra-thin oxide layer poly-Si/SiO_2/Si structures," *Solid-State Electron.,* vol. 38, no. 8, pp. 1465–1471, 1995.

[60] J. S. Brugler and P. G. A. Jespers, "Charge pumping in MOS devices," *IEEE Transactions on Electron Devices*, vol. 16, pp. 297-302, 1969.

[61] G. Groeseneken, H. E. Maes, N. Beltran, and R. F. De Keersmaecker, "A reliable approach to charge-pumping measurements in MOS transistors," *IEEE Transactions on Electron Devices*, vol. 31, pp. 42-53, 1984.

[62] R. E. Paulsen, R. R. Siergiej, M. L. French, and M. H. White, "Observation of near-interface oxide traps with the charge-pumping technique," *IEEE Electron Device Letters*, vol. 13, pp. 627-629, 1992.

[63] R. E. Paulsen and M. H. White, "Theory and application of charge pumping for the characterization of Si-SiO2 interface and near-interface oxide traps,", *IEEE Transactions on Electron Devices*, vol. 41, pp. 1213-1216, 1994.

[64] Y. L. Hsu, Y. K. Fang, F. C. Tsao, F. J. Kuo, and Y. Ho, "Modeling of abnormal capacitance-voltage characteristics observed in MOS transistor with ultra-thin gate oxide," *Solid-State Electronics*, vol. 46, no.11,pp. 1941-1943, Nov. 2002.

[65] J. Schmitz, F. N. Cubaynes, R. J. Havens, R. de Kort, A. J. Scholten, and L. F. Tiemeijer, "RF capacitance-voltage characterization of MOSFETs with high leakage dielectrics," *IEEE Electron Device Lett.*, vol. 24, no.1, pp. 37-39, Jan. 2003.

[66] J. Koomen, "Investigation of the MOST channel conductance in weak inversion," *Solid-State Electronics*, vol. 16, pp. 801-10, 1973.

[67] S. Mileusnic, M. Zivanov, and P. Habas, "MOS transistors characterization by split C-V method," *IEEE CAS 2001 Proceedings. International Semiconductor Conference*, vol. vol.2, pp. 503-506, 2001.

[68] F. Lime, C. Guiducci, R. Clerc, G. Ghibaudo, C. Leroux, and T. Ernst, "Characterization of effective mobility by split C(V) technique in N-MOSFETs with ultra-thin gate oxides," *Solid-State Electronics*, vol. 47, pp. 1147-1153, 2003.

[69] P. M. Zeitzoff, C. D. Young, G. A. Brown, and Y. Kim, "Correcting effective mobility measurements for the presence of significant gate leakage current," *Electron Device Letters, IEEE*, vol. 24, pp. 275-277, 2003.

[70] K. Romanjek, F. Andrieu, T. Ernst, and G. Ghibaudo, "Improved split C-V method for effective mobility extraction in sub-0.1um Si MOSFETs," *IEEE Electron Device Letters,* vol. 25, pp. 583-585, 2004.

[71] K. Romanjek, F. Andrieu, T. Ernst, and G. Ghibaudo, "Characterization of the effective mobility by split C-V technique in sub 0.1 μm Si and SiGe PMOSFETs," *Solid-State Electronics*, vol. 49, pp. 721-6, 2005.

[72] V. Kilchytska, D. Lederer, P. Simon, N. Collaert, J.-P. Raskin, and D. Flandre, "Revised

split C-V technique for mobility investigation in advanced devices," *IEEE International SOI Conference,* , vol. 2005, pp. 110-111, 2005.

[73] V. Kilchytska, D. Lederer, N. Collaert, J.-P. Raskin, and D. Flandre, "Accurate effective mobility extraction by split C-V technique in SOI MOSFETs: Suppression of the influence of floating-body effects," *IEEE Electron Device Letters*, vol. 26, pp. 749-751, 2005.

[74] K. F. Schuegraf and C. Hu, "Hole injection oxide breakdown model for very low voltage lifetime extrapolation," *Proceedings of IEEE International Reliability Physics Symposium*, pp. 7-12, 1993.

[75] C.H. Choi, K.H. Oh, J.S. Goo, Z. Yu, and R. W. Dutton, "Direct tunneling current model for circuit simulation," *IEEE International Electron Devices Meeting*, pp. 735-738, 1999.

[76] W. C. Lee and C. Hu, "Modeling CMOS tunneling currents through ultrathin gate oxide due to conduction- and valence-band electron and hole tunneling," *IEEE Transactions on Electron Devices*, vol. 48, pp. 1366-1373, 2001.

[77] P. Hermans, J. WIitters, G. Groeseneken, and H. E. Maes, "Analysis of the Charge Pumping Technique and Its Application for the Evaluation of MOSFET Degradation," *IEEE Transactions on Electron Devices*, vol. 36, pp. 1318-1335, 1989.

[78] I. Lundstrom and C. Svensson, "Tunneling to traps in insulators," *Journal of Applied Physics*, vol. 43, pp. 5045-5047, 1972.

[79] D. Bauza and G. Ghibaudo, "Analytical study of the contribution of fast and slow oxide traps to the charge pumping current in MOS structures," *Solid-State Electronics*, vol. 39, pp. 563-570, 1996.

[80] Y. Maneglia and D. Bauza, "Extraction of slow oxide trap concentration profiles in metal--oxide--semiconductor transistors using the charge pumping method," *Journal of Applied Physics*, vol. 79, pp. 4187-4192, 1996.

[81] D. Bauza and Y. Maneglia, "In-depth exploration of Si_SiO2 interface traps in MOS transistors using the charge pumping technique," *IEEE Transactions on Electron Devices*, vol. 44, pp. 2262-2266, 1997.

[82] Y. Maneglia and D. Bauza, "Study of the near Si_SiO2 interface trap layer using the charge pumping technique," *International Semiconductor Conference, CAS'97 Proceedings*, vol. 1, pp. 135-138 vol.1, 1997.

[83] Y. Maneglia and D. Bauza, "Extraction of the Si_SiO2 interface trap layer parameters in MOS transistors using a new charge pumping analysis," *Proceedings of the International Conference on Microelectronic Test Structures*, pp. 201-205, 1998.

[84] Y. Maneglia and D. Bauza, "Evolution of the Si_SiO2 interface trap characteristics with Fowler-Nordheim injection," *Proceedings of the International Conference on Microelectronic Test Structures*, pp. 117-120, 1999.

[85] D. Bauza, "Extraction of Si-SiO2 interface trap densities in MOS structures with ultrathin oxides," *IEEE Electron Device Letters,* vol. 23, pp. 658-660, 2002.

[86] D. Bauza, "Extraction of Si-SiO2 Interface Trap Densities in MOSFET's with Oxides Down to 1.3 nm Thick," *Solid-State Device Research Conference, Proceeding of the 32nd European*, pp. 231-234, 2002.

[87] G. T. Sasse, J. Schmitz, ”Charge Pumping at Radio Frequencies - Methodology, Trap Response and Application.” *IEEE International Reliability Physics Symposium*(IRPS), pp627-628, 2006,

[88] H. C. Lai, N. K. Zous, W. J. Tsai, T. C. Lu, T. Wang, Y. C. King, and S. Pan, "Reliable extraction of interface states from charge pumping method in ultra-thin gate oxide MOSFET's," *IEEE International Conference on Microelectronic Test Structures*, pp. 99-102, 2003.

[89] W. Shockley and W. T. Read, "Statistics of the Recombinations of Holes and Electrons," *Physical Review*, vol. 87, pp. 835 - 842, 1952.

[90] R. N. Hall, "Electron-Hole Recombination in Germanium," *Physical Review*, vol. 87, pp. 387 - 387, 1952.

[91] W. V. Backensto and C. R. Viswanathan, "The utilization of charge pumping techniques to evaluate the energy and spatial distribution of interface states of an MOS transistor," *International Electron Devices Meeting*, pp. 287-91, 1976.

[92] T. J. Russell, C. L. Wilson, and M. Gaitan, "Determination of the spatial variation of interface trapped charge using short-channel MOSFET's," *IEEE Transactions on Electron Devices*, vol. 30, pp. 1662-1671, 1983.

[93] A. Kerber, E. Cartier, L. Pantisano, R. Degraeve, G. Groeseneken, H. E. Maes, and U. Schwalke, "Charge trapping in SiO2/HfO2 gate dielectrics: comparison between

charge-pumping and pulsed Id-Vg," *13th Biennial Conference on Insulating Films on Semiconductors, Microelectronic Engineering*, vol. 72, pp. 267-272, 2004.

[94] A. Kerber, E. Cartier, L. Pantisano, R. Degraeve, T. Kauerauf, Y. Kim, A. Hou, G. Groeseneken, H. E. Maes, and U. Schwalke, "Origin of the threshold voltage instability in SiO2/HfO2 dual layer gate dielectrics," *IEEE Electron Device Letters*, vol. 24, pp. 87-89, 2003.

[95] M. H. Tsai and T. P. Ma, "1/f noise in hot-carrier damaged MOSFET's: effects of oxide charge and interface traps," *IEEE Electron Device Letters*, vol. 14, pp. 256-258, 1993.

[96] D. M. Fleetwood, P. S. Winokur, R. A. Reber, Jr., T. L. Meisenheimer, J. R. Schwank, M. R. Shaneyfelt, and L. C. Riewe, "Effects of oxide traps, interface traps, and `border traps' on metal-oxide-semiconductor devices," *Journal of Applied Physics*, vol. 73, pp. 5058-5074, 1993.

[97] C. R. Helms, E. H. Poindexter "The silicon-silicon dioxide system: Its microstructure and imperfections." *Reports on Progress in Physics*, vol. 57, no. 8, pp. 791-852, 1994.

[98] L. L. Rosier, "Surface state and surface recombination velocity characteristics of Si-SiO2 interfaces." *IEEE Transactions on Electron Devices*, vol. 13, pp. 260-268 1966.

[99] M. Schulz, N. M. Johnson, "Transient capacitance measurements of hole emission from interface states in MOS structures." *Applied Physics Letters*, vol. 13, pp.622-625, 1977.

[100] N. M. Johnson, "Energy-resolved DLTS measurement of interface states in MIS structures." *Applied Physics Letters*, vol. 34, pp.802-804, 1979.

[101] U. Cilingiroglu, "A pulsed interface-probing technique for MOS interface characterization at mid-gap levels." *IEEE Transactions on Electron Devices*, vol. 35, pp. 2391-2396 1988.

[102] D. Vuillaume, R. Bouchakour, M. Jourdain, J. C. Bourgoin, "Capture cross section of Si-SiO2 interface states generated during electron injection." *Applied Physics Letters*, vol. 55, pp.153-155, 1989.

[103] D. Goguenheim, D. Vuillaume, G. Vincent, N. M. Johnson, "Accurate measurements of capture cross sections of semiconductor insulator interface states by a trap-filling experiment: The charge-potential feedback effect." *Journal of Applied Physics*, vol. 68, pp.1104-1113, 1990.

[104] N. S. Saks, M. G. Ancona, "Determination of interface trap capture cross sections using

three-level charge pumping." *IEEE Electron Device Letters*, vol. 11, pp. 339-341, 1990.

[105] L. Vishnubhotla, W. Chen, T. P. Ma, "ac conductance measurements on radiation-damaged (100) Si-SiO2 interface after defect transformation." *Applied Physics Letters*, vol. 57, pp.1778-1780, 1990.

[106] G. Van den bosch, G. V. Groeseneken, P. Heremans, H. E. Maes, "Spectroscopic charge pumping - A new procedure for measuring interface trap distributions on MOS transistors." *IEEE Transactions on Electron Devices*, vol. 38, pp. 1820-1831, 1991.

[107] M. Kejhar, "Double-pulse charge pumping in MOSFETs." *IEEE Electron Device Letters*, vol. 13, pp. 344-346, 1992.

[108] L. Militaru, P. Masson, G. Guegan, "Three level charge pumping on a single interface trap." *IEEE Electron Device Letters*, vol. 23, pp. 94-96, 2002.

[109] L. Wang, A. Neugroschel, "Method for determination of carrier capture cross-sections at Si-SiO2 interface." *Electronics Letters*, vol. 40, pp. 148-149, 2004.

[110] G. Rickayzen, "On the Theory of the Thermal Capture of Electrons in Semi-Conductors." *Proceeding of Royal Society. Series A, Mathematical and Physical Sciences*, vol. 241, pp.480-494,1957

[111] C. H. Henry, D. V. Lang, "Nonradiative capture and recombination by multiphonon emission in GaAs and GaP." *Physics Review B*, vol.15, pp.989-1016, 1977.

[112] M. Lax, H. J. Carmichael, W. J. Shugard, "Nonadiabatic formulation for radiationless transitions induced by classical lattice vibrations." *Physics Review B*, vol.26, pp.3547-3558, 1982.

[113] M. Lax, "Cascade Capture of Electrons in Solids." *Physics Review*, vol. 119, pp.1502-1523, 1960.

[114] A. H. Edwards, "Theory of the Pb center at the 111 Si-SiO2 interface." *Physics Review B* vol.36, pp.9638-9648, 1987.

[115] R. P. Messmer, G. D. Watkins, "Molecular-Orbital Treatment for Deep Levels in Semiconductors - Substitutional Nitrogen and the Lattice Vacancy in Diamond." *Physics Review B*, vol.7, pp.2568-2590, 1973.

[116] M. Jaros, S. Brand, "Localized defects in III-V semiconductors." *Physics Review B*, vol. 14, pp.4494-4505, 1976.

[117] R. B. Laughlin, J. D. Joannopoulos, D. J. Chadi, "Theory of the electronic structure of the Si-SiO2 interface." *Physics Review B*, vol. 21, pp.5733-5744, 1980.

[118] T. Sakurai, T. Sugano, "Theory of continuously distributed trap states at Si-SiO2 interfaces." *Journal of Applied Physics*, vol.52**,** pp.2889-2896, 1981.

[119] G. La Rosa, F. Guarin, S. Rauch, A. Acovic, J. Lukaitis, and E. Crabbe, "NBTI-channel hot carrier effects in PMOSFETs in advanced CMOS technologies," *IEEE International Reliability Physics Symposium*, pp. 282-286, 1997.

[120] C. Shen, M.-F. Li, C. E. Foo, T. Yang, D. M.Huang, A. Yap, G. S. Samudra, Y.-C, Yeo, "Characterization and Physical Origin of Fast Vth Transient in NBTI of pMOSFETs with SiON Dielectric," *International Electron Devices Meeting*, 2006

[121] B. Kaczer, V. Arkhipov, M. Jurczak, and G. Groeseneken, "Negative bias temperature instability (NBTI) in SiO2 and SiON gate dielectrics understood through disorder-controlled kinetics," *14th Biennial Conference on Insulating Films on Semiconductors*, vol. 80, pp. 122-125, 2005.

[122] B. Kaczer, V. Arkhipov, R. Degraeve, N. Collaert, G. Groeseneken, and M. Goodwin, "Temperature dependence of the negative bias temperature instability in the framework of dispersive transport," *Applied Physics Letters*, vol. 86, pp. 143506, 2005.

[123] K. O. Jeppson and C. M. Svensson, "Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices," *Journal of Applied Physics*, vol. 48, pp. 2004-2014, 1977.

[124] M. A. Alam and S. Mahapatra, "A comprehensive model of PMOS NBTI degradation," *Microelectronics Reliability*, vol. 45, pp. 71-81, 2005.

[125] S. Ogawa and N. Shiono, "Generalized diffusion-reaction model for the low-field charge-buildup instability at the Si-SiO2 interface," *Physical Review B (Condensed Matter)*, vol. 51, pp. 4218-4230, 1995.

[126] A. Kottantharayil, "Low Voltage Hot-Carrier Issues in Deep sub-micron Metal-Oxide-Semiconductor Field-Effect-Transistors." *Ph.D. Thesis*, Depart. Electronics & Information Technology, University of Munich., 2002

[127] D. Varghese, S. Mahapatra, M. A. Alam, "Hole energy dependent interface trap generation in MOSFET Si-SiO2 interface." *IEEE Electron Device Letters*, vol. 26, pp. 572-574, 2005.

[128] J. Lai and A. Majumdar, "Concurrent thermal and electrical modeling of sub-micrometer silicon devices," *Journal of Applied Physics*, vol. 79, pp. 7353-7361, 1996.

[129] S. V. J. Narumanchi, J. Y. Murthy, and C. H. Amon, "Boltzmann transport equation-based thermal modeling approaches for hotspots in microelectronics," *Heat and Mass Transfer*, vol. 42, pp. 478-491, 2006.

[130] R. Yang, G. Chen, M. Laroche, and Y. Taur, "Simulation of Nanoscale Multidimensional Transient Heat Conduction Problems Using Ballistic-Diffusive Equations and Phonon Boltzmann Equation," *Journal of Heat Transfer*, vol. 127, pp. 298-306, 2005.

[131] P.Sverdrup, S.Sinha, M.Asheghi, S.Uma and K.E.Goodson, "Measurement of ballistic phonon conduction near hotspots in silicon," *Applied Physics Letters*, vol. 86, pp. 3331-3333, 2001

[132] D. G. Cahill, K. Goodson, and A. Majumdar, "Thermometry and Thermal Transport in Micro/Nanoscale Solid-State Devices and Structures," *Journal of Heat Transfer*, vol. 124, pp. 223-241, 2002.

[133] S. Sinha, E. Pop, R. W. Dutton, and K. E. Goodson, "Non-Equilibrium Phonon Distributions in Sub-100 nm Silicon Transistors," *Journal of Heat Transfer*, vol. 128, pp. 638-647, 2006.

[134] M. Reyboz, R. Daviot, O. Rozeau, P. Martin, and M. Paccaud, "Compact modeling of the self heating effect in 120 nm multifinger body-contacted SOI MOSFET for RF circuits," *IEEE International SOI Conference*, pp. 159-161, 2004.

[135] S. Zhu, A. Nakajima, T. Ohashi, H. Miyake, "Enhancement of BTI degradation in pMOSFETs under high-frequency bipolar gate bias" *IEEE Electron Device Letters*, vol. 26, pp. 387-389, 2005.

[136] S. S. Tan, T. P. Chen, L. Chan, "Dynamic NBTI lifetime model for inverter-like waveform" *Microelectronics Reliability*, vol. 45, pp. 1115-1118, 2005.

Appendix A

Introduction of Time Domain Reflectrometry(TDR)

A.1. Principle of TDR

We are familiar with echoes that occur when sound encounters a change in impedance, such as a wall. In electrical system, a similar phenomenon occurs when electrical energy traveling in a transmission line encounters a change in impedance. Any change in the impedance of the transmission line causes a reflection with amplitude proportional to the magnitude of the impedance change.

Time domain Reflectometry (TDR) refers to the measurement of the reflection of a fast step function from an unknown device relative to that of known standard impedance. The amount of energy reflected is a function of the transmitted energy and the magnitude of the impedance discontinuity. The time lapse between transmitting and the returning signal can be used to pinpoint the location of the impedance discontinuity. By studying the reflected voltage waveform, much information can be obtained on the nature of the load.

A TDR can be constructed from a pulse generator and an oscilloscope. Purposely built instruments offer a single, easier-to-use, and generally high-performance package. Figure A.1 shows a simplified block diagram of a typical TDR instrument connected

to unknown load impedance. This contains a high-speed pulse generator, high-bandwidth oscilloscope and reference $50\,\Omega$ instrument TDR output resistance. The sampler represents the receiver and is used to capture the reflected response as an input to the oscilloscope. It monitors the voltage at the front panel connector of the TDR. When the pulse generator sends out the pulse, the oscilloscope display shows the rising as it leaves the TDR unit. Some time later, the rising pulse edge reaches the device under test. Any resulting reflections travel back toward the pulse generator. When they arrive at the input to the TDR, the oscilloscope displays the reflection. Thus, the oscilloscope display shows the round-trip time between the incident edge and the reflection.

All TDR impedance measurements are based on the ratio of transmitted voltage to reflect voltage .It is defined a reflection coefficient ρ as the reflected signal amplitude divided by the incident signal amplitude. At instant of time, ρ is determined by the impedance mismatch between the TDR reference characteristic impedance and the load impedance. Given known impedance and a measured reflection coefficient, the unknown impedance that caused the reflection can be calculated from the following equation

$$\rho = \frac{V_{\mathrm{Re}\,flected}}{V_{Incident}} = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{A.1}$$

$$Z_L = Z_0 \frac{1+\rho}{1-\rho} \tag{A.2}$$

The equation (A.1) and (A.2) can be used to calculated the DUT load impedance (Z_L),

where Z_0 represents the output impedance of the TDR, which it is 50 Ω in most practical situations. The load impedance can be determined by calculating the value of ρ.

The pulse generator launches a traveling wave into the reference coaxial cable at node A which typically has $Z_0 = 50\Omega$ characteristic impedance. Here we will assume that the pulse generator produces a step function pulse as shown in Figure A.1. This step-like traveling wave pulse propagates through the cable at a velocity, V_P and arrives at the far end (Node B) after a time T_D.

$$T_D = \frac{l}{V_P}, V_P = \frac{c}{\sqrt{\varepsilon}} \qquad \text{(A. 3)}$$

Where c is the speed of light and ε is the relative dielectric constant of the transmission line. The step-like wave traveling to the right is designated as V_{Inc}.

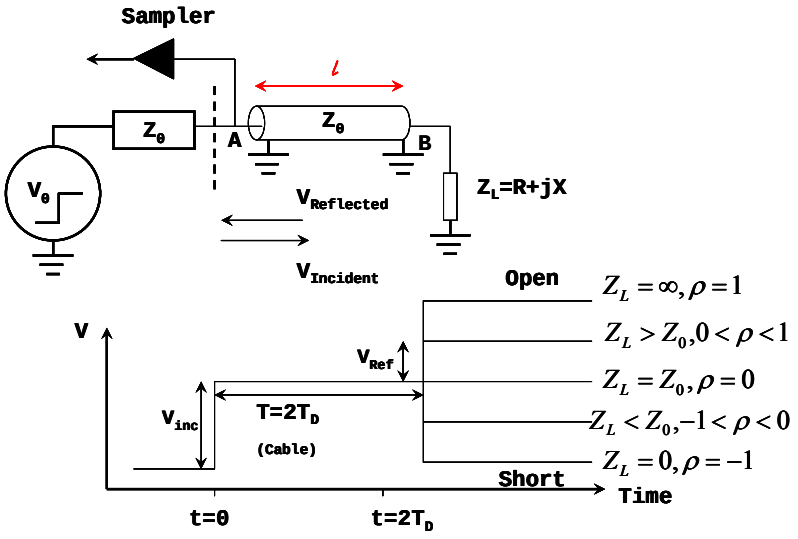


Figure A.1 Basic TDR setup and TDR waveforms with resistive terminations

If the terminating impedance, Z_L matches the transmission line characteristic

impedance, Z_0 i.e., $Z_L = Z_0$ then the TDR pulse is perfectly absorbed. However, if Z_L is not equal to Z_0, some of the incident pulse energy will be reflected back as an echo towards the left as a new traveling wave $V_{\mathrm{Re}f}$. This reflected pulse $V_{\mathrm{Re}f}$ will arrive back at the TDR test port (node A) at time $t = 2 \times T_D$. The critical importance of the reference cable is now seen. It provides a separation in time between the TDR observation at node A of the incident test signal V_{Inc} and the reflected signal $V_{\mathrm{Re}f}$. The waveform observed at node A is the algebraic sum of the test pulse and any returning echoes from impedance discontinuities, except that the echoes are delayed in time by the two-way travel time $2 \times T_D$. Thus, an examination of the time delay and wave shape of the echoes present on node A allows us to determine the location and nature of discontinuities within the transmission line and/or mismatched terminations to the line.

Ideally the length of the reference cable would be chosen such that any wave shape irregularities in the test pulse will have been damped out before $2 \times T_D$ when the first reflections arrive at node A. Figure A.1 shows the TDR waveforms observed for various resistive terminations Z_L. If $Z_L = Z_0$, then no reflection occurs, and the TDR waveform displayed on the oscilloscope is a flat line for $t > 2 \times T_D$. If Z_L is greater than Z_0, then a positive step is observed. If Z_L is less than Z_0, a negative step is observed. The actual value of Z_L may be calculated from the size of these steps by Equation (A.1) and (A. 2). The amplitude of the incident step, V_{Inc} and the reflected pulse, $V_{\mathrm{Re}f}$ are defined as shown in Figure A.1.

A.2. Analytical Expression of TDR Voltage Response

Using TDR to characterize and derive load impedance has been around for some time. As a working understanding of the underlying principles is of utmost importance, the formulas are being derived in this section using Lap lace transformation analysis and range of validity of these formulas are described. Further enhancement for TDR responses for complex loads are also added in this section. To proceed with the analysis in the time domain, a sensible approach would be to use Lap lace transformation. Consider the system as described in Figure A.1, a step pulse generator with step amplitude of V_0 sends a step pulse along the transmission line of length *l* to complex load Z_L. Assume the transmission line to be lossless.

At the load end

$$Z_L(s) = \frac{V_L(s)}{I_L(s)} \tag{A.4}$$

Where

$$V_L(s) = V_{Inc}(s) + V_{\mathrm{Re}f}(s) \tag{A.5}$$

$$I_L(s) = \frac{1}{Z_0}\left(V_{Inc}(s) - V_{\mathrm{Re}f}(s)\right) \tag{A.6}$$

Where $V_{Inc}(s)$ = Incident voltage wave or, $V_{\mathrm{Re}f}(s)$ = Reflected voltage wave, Z_0 is the characteristic impedance of transmission line. From equations (A.4), (A.5) and (A.6)

$$V_{\mathrm{Re}f}(s) = V_{Inc}(s)\left(\frac{Z_L - Z_0}{Z_L + Z_0}\right) \tag{A.7}$$

For a step source of $V_0/2$ (the generator impedance and the transmission line impedance form a divide by 2 voltage divider), the reflected voltage waveform at load end is

$$V_{\mathrm{Re}f}(s) = \frac{V_0}{2s}\left(\frac{Z_L - Z_0}{Z_L + Z_0}\right) \tag{A.8}$$

At the TDR generator end (node A), the transformed voltage at generator end is the sum of initial step voltage and the time-delayed reflected voltage after a round trip delay:

$$V_{TDR}(s) = V_{\mathrm{Re}f}(s)e^{\frac{2ls}{v}} + \frac{1}{2s}V_0 \tag{A.9}$$

In the time domain:

$$V_{TDR}(t) = V_{\mathrm{Re}f}(t - T_D)U(t - T_D) + \frac{V_0}{2}U(t) \tag{A.10}$$

Or

$$V_{TDR}(t) = L^{-1}\left[\frac{V_{Step}}{s}\left(\frac{Z_L - Z_0}{Z_L + Z_0}\right)e^{\frac{2ls}{v_p}}\right] + V_{Step}U(t)$$

$$\text{With } V_{Step} = \frac{V_0}{2}, U(t) = \begin{cases} 0 & t < 0 \\ 1 & t > 0 \end{cases} \tag{A.11}$$

Hence by studying the waveform of equation (A.11), we could in principle be able to determine the equivalent circuit of the load connected at the transmission line end of a TDR system. Although we could in theory estimate the values of circuit parameters of the unknown load by examining the characteristics and shape of the reflection, this direct method is not desirable as the exact waveform of the reflection depends upon the bandwidth and sampling rate of the oscilloscope chosen. An indirect method of determining inductance and capacitance of reactive load using time domain extraction

method is discussed in later section.

A.3. TDR Voltage Response of Complex Load

Reactive components can also be measured using TDR. Figure A.2 shows the simulated TDR waveforms for an ideal 2pF capacitor. Recall that the fast step rise time contains the high frequencies, while the flat top of the step contains the low frequency components.

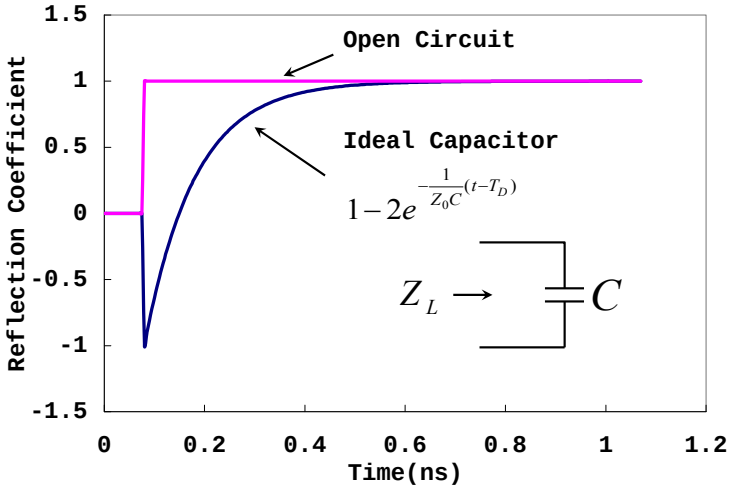


Figure A.2 Simulated TDR response of Ideal 2pF capacitor

The capacitor's impedance, $Z_L = \frac{1}{j\omega C}$, initially appears as a short circuit to the fast rising edge of the TDR step pulse. Thus, we initially see ρ of -1. Later in time, the capacitor appears as an open circuit to the flat top portion of the step pulse, i.e., the low frequency and DC portion. Therefore, the final TDR value is ρ of 1. The waveform connecting these two end points is an exponential. An inductor performs exactly opposite.

For capacitive load of capacitance C, $Z_L = \frac{1}{sC}$, Put it into the expression in Equation (A.10) and work out the Laplace transform we can get the TDR voltage response of a ideal capacitor as the load:

$$V_{TDR}(t) = L^{-1}\left[\frac{V_{Step}}{s}\left(\frac{\frac{1}{sC} - Z_0}{\frac{1}{sC} + Z_0}\right)e^{\frac{2ls}{v_p}}\right] + V_{Step}U(t)$$

$$\Rightarrow V_{TDR}(t) = V_{Step}\left[1 - 2e^{-\frac{1}{Z_0 C}(t-T_D)}\right]U(t-2T_D) + V_{Step}U(t) \qquad \text{(A.12)}$$

The analysis for different case of complex loads is similar to the previous cases. Shown below are the results.

- Series R-C Load: $Z_L = R_S + 1/sC$

$$V_{TDR}(t) = V_{Step}\left[1 - \frac{2Z_0}{R_S + Z_0}e^{-\frac{1}{(R_S+Z_0)C}(t-2T_D)}\right]U(t-2T_D) + V_{Step}U(t) \qquad \text{(A.13)}$$

- Shunt R-C Load : $Z_L = \frac{R_P}{1 + sC \cdot R_P}$

$$V_{TDR}(t) = V_{Step}\left[\frac{R_P - Z_0}{R_P + Z_0} - \frac{2R_P}{(R_P + Z_0)}e^{-\frac{(R_P+Z_0)}{R_P Z_0 C}(t-2T_D)}\right]U(t-2T_D) + V_{Step}U(t) \qquad \text{(A.14)}$$

- Series R-L Load: $Z_L = R_S + sL$

$$V_{TDR}(t) = V_{Step}\left[\frac{R_S - Z_0}{R_S + Z_0} + \frac{2Z_0}{R_S + Z_0}e^{-\frac{(R_S+Z_0)}{L}(t-2T_D)}\right]U(t-2T_D) + V_{Step}U(t) \qquad \text{(A.15)}$$

- Shunt R-L Load: $Z_L = \frac{sR_P L}{1 + sL \cdot R_P}$

$$V_{TDR}(t) = V_{Step}\left[-1 + \frac{2Z_0}{R_S + Z_0}e^{-\frac{R_S Z_0}{(R_S+Z_0)L}(t-2T_D)}\right]U(t-2T_D) + V_{Step}U(t) \qquad \text{(A.16)}$$

- Series R-L-C Load: $Z_L = R_S + sL + \frac{1}{sC}$

$$V_{TDR}(t) = V_{Step}\left[\begin{array}{l} 1 + A\cdot\exp\left(-\dfrac{(R_S+Z_0)+\sqrt{(R_S+Z_0)^2-4\cdot\dfrac{L}{C}}}{2L}\right) \\ + B\cdot\exp\left(-\dfrac{(R_S+Z_0)-\sqrt{(R_S+Z_0)^2-4\cdot\dfrac{L}{C}}}{2L}\right) \end{array}\right]\cdot U(t-2T_D) + V_{Step}U(t)$$

(A.17a)

Where

$$A = 2Z_0\cdot\frac{(R_S+Z_0)-\sqrt{(R_S+Z_0)^2-\dfrac{4L}{C}}}{\left[(R_S+Z_0)\left((R_S+Z_0)-\sqrt{(R_S+Z_0)^2-\dfrac{4L}{C}}\right)-\dfrac{4L}{C}\right]} \qquad \text{(A.17b)}$$

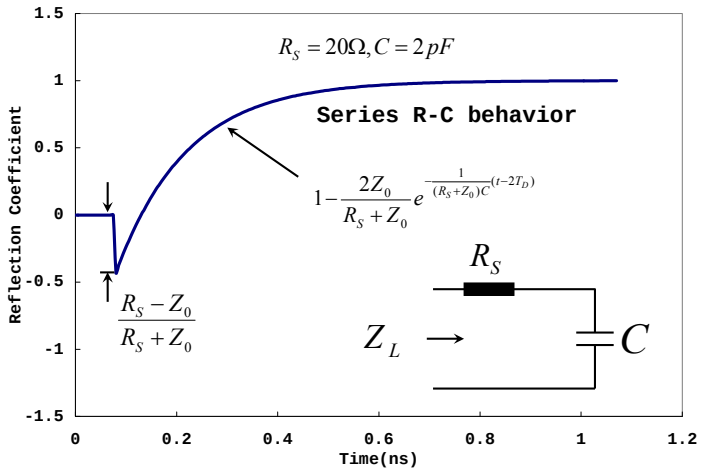
$$B = 2Z_0\cdot\frac{(R_S+Z_0)+\sqrt{(R_S+Z_0)^2-\dfrac{4L}{C}}}{\left[(R_S+Z_0)\left((R_S+Z_0)+\sqrt{(R_S+Z_0)^2-\dfrac{4L}{C}}\right)-\dfrac{4L}{C}\right]} \qquad \text{(A.17c)}$$

- Series/Shunt R-C Load: $Z_L = R_S + R_P \,//\, \dfrac{1}{sC}$

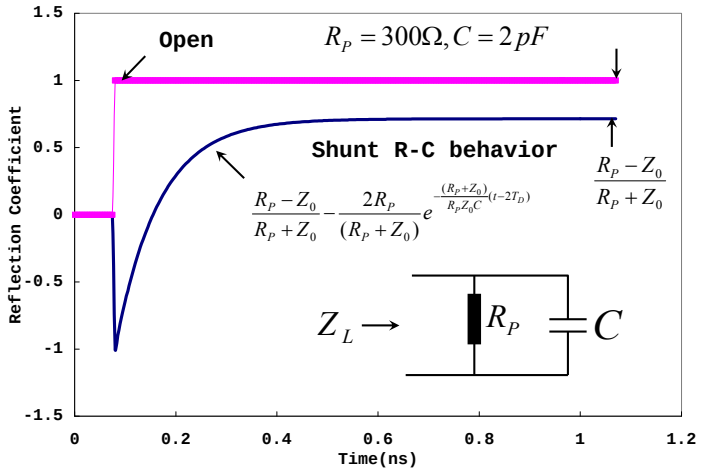
$$V_{TDR}(t) = V_{Step}\left[\frac{R_P+R_S-Z_0}{R_P+R_S+Z_0} - \frac{2R_PZ_0}{(R_S+Z_0)(R_P+R_S+Z_0)}e^{-\frac{(R_P+R_S+Z_0)}{R_PC(R_S+Z_0)}(t-2T_D)}\right]U(t-2T_D) + V_{Step}U(t)$$

(A.18)

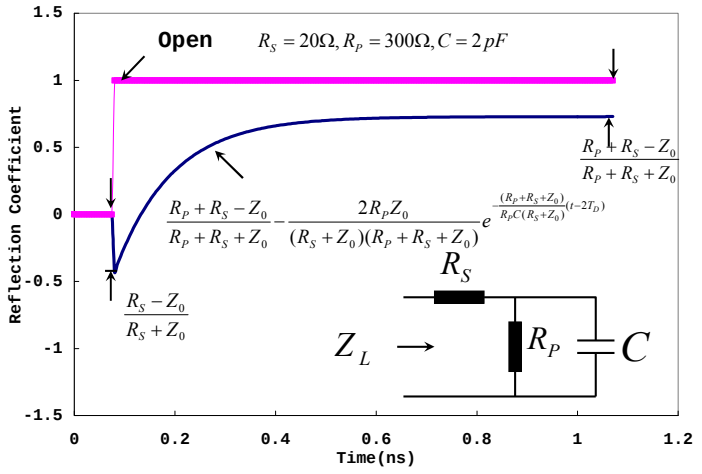
The functions of equations (A.13), (A.14), (A.18) are plotted in Figure A.3(a) to Figure A.3(c) for fictitious $R_S = 20\Omega, R_P = 300\Omega, C = 2\,pF$.



$$\tau = (R_S + Z_0)C$$



$$\tau = \left(\frac{R_P \cdot Z_0}{R_P + Z_0} \right) C$$



$$\tau = \left(\frac{(R_S + Z_0) \cdot R_P}{R_S + Z_0 + R_P} \right) C$$

Figure A.3(a) Series R-C load TDR response(b) Shunt R-C load TDR response (c) Series/Shunt R-C load TDR response

A.4. Measurement Factors

While the fundamental concepts behind TDR are relatively simple and modern instruments take care of most of the math, to properly characterize and interpret TDR response characteristics as related to the DUT (device under test), the real-life factors that affect TDR resolution in the laboratory must be understood.

In general modern instruments make it relatively easy to make impedance measurements near the reference impedance (usually 50Ω) with about 2% accuracy ($\pm 1\Omega$). Achieving higher accuracy or making measurement farther from the reference impedance requires more care. Insufficient resolution will produce misleading results that can miss details of impedance and velocity profiles. This is common when items to be measured are very closely spaced or when the impedance discontinuities are electrically very short. The following list covers a few key considers in making accurate and repeatable impedance measurements.

A.4.1. System Rise Time

The TDR and TDT waveforms shown previously were all for the ideal case when the test pulse was a zero rise time step function. In the real world, the finite rise time of both the pulse generator and the oscilloscope will distort these waveforms and limit the resolution. The rise time determines the smallest impedance discontinuity that the

TDR instrument can measure. If a discontinuity is small respect to the system rise time, the reflection will not accurately represent the impedance of the discontinuity. In extreme cases, the discontinuity may effectively disappear.

The general rule of thumb is that the reflection from two narrowly spaced discontinuities will be indistinguishable if they are separated by less than half the rise time. Subsequently, the TDR can only accurately resolve structures that are electrically long compared to the rise time. System Rise time (T_{rise}) is characterized by the fall or rise time of the reflected edge from an ideal short or open at the probe tip.

$$TDR_{\mathrm{Re}\,solution} \geq \frac{T_{rise}}{2} \tag{A.19}$$

System rise time is the combined rise time of the pulse generator, the oscilloscope and the interconnect between the TDR and DUT.

$$T_{rise} = \sqrt{\left(T_{r,stepgen}\right)^2 + \left(T_{r,Oscilloscope}\right)^2 + \left(T_{r,Interconnect}\right)^2} \tag{A.20}$$

For "Large" inductors or capacitors, their time constants will be much greater than the system rise time. In these cases the TDT and TDR displays will appear to be the same as those shown previously for the ideal zero rise time case. For "Tiny" inductors or capacitors with time constants much less than the system rise time, they will not produce visible reflections on TDR displays and thus cannot be measured.

For "Small" inductors or capacitors that have time constants of the same order of magnitude as the system rise time, useful measurements can still be made. The

waveforms will, however, be considerably different from the ideal cases previously shown in Figure A.3. Their peaks will never reach ρ of +1 or -1.Here we simulate the TDR reflected voltage of 10pF capacitor under 50ps, 100ps, 150ps rise time 1V magnitude incident step signal as shown Figure A.4. Compared to the ideal step signal, the rise time will modify the shape of initial reflection of capacitor. At the time incident step signal arrives at the capacitor, the capacitor will behave like a short circuit and reflect the incident signal. This initial reflection of a short behavior of the capacitor is actually an imaging process. It reflects the entire feature in the incident signal including the rising part and shows itself as a falling part in the response of capacitor. Then with time going on, the capacitor will charge up and reflection voltage will increase exponentially. Due to the rise time of the incident signal, the charging process of the capacitor will be overwhelmed by the falling part. At certain time, the charge process will take it over and it is where the minimum reflection happens.

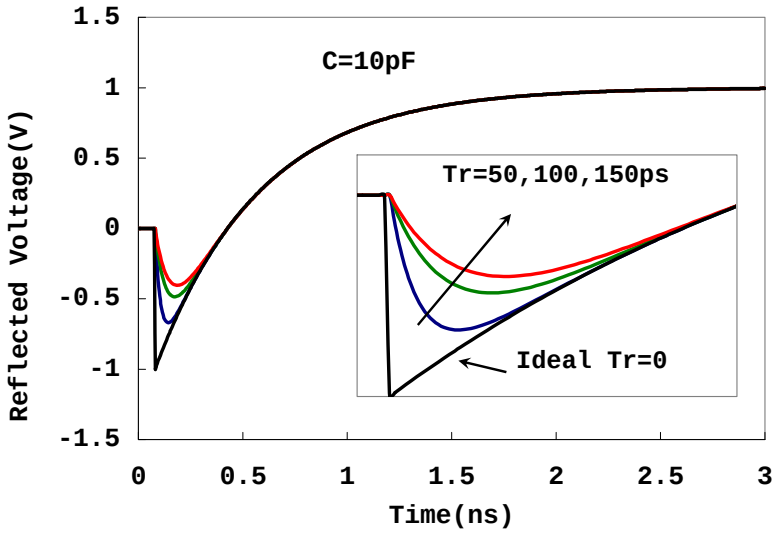


Figure A.4 TDR response of capacitor with different rise of incident step voltage

A.4.2. Reference Impedance

All TDR measurements are relative-they compare unknown impedance to known impedance. The accuracy of the results depends directly on the accuracy of the reference impedance. Any error in the reference impedance translates to error in the measured impedance. Thus, high-quality reference impedance is required for high-quality measurements.

A.4.3. System Noise

System noise can reduce the accuracy and repeatability of the measurements. Fortunately, modern instruments usually provide convenient on-board signal averaging that can significantly reduce noise and improve the repeatability of the measurements. Also, remember that signal averaging takes time. The rate at which the instrument can acquire waveform and perform averaging can have a significant measurement throughput. A slow acquisition rate is particularly detrimental to averaging since many waveforms must be required.

A.4.4. Cable/probe/connector Losses

The cable that connects the TDR unit to the DUT not only degrades the system rise time, but can cause other aberrations in the system response that add to measurement

error. Always use the shortest high quality cable possible to connect to the test structure. Cables and connectors between the step source, the DUT, and the oscilloscope can significantly affect measurement results. Impedance mismatches and imperfect connectors add reflections to the actual signal being measured. These can distort the signal and make it difficult to determine which reflections are from the DUT and which are from other sources.

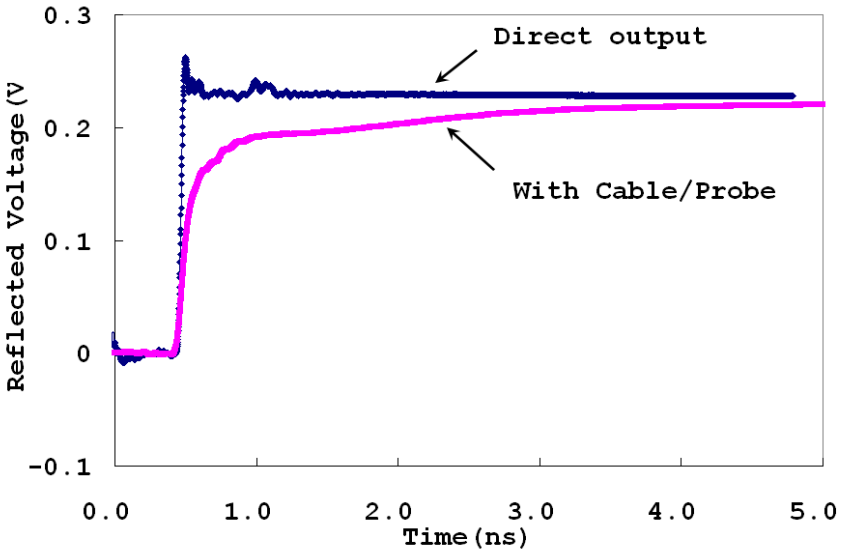


Figure A.5 cables and probe effect on TDR measurements

In addition, cables are imperfect conductors that become more imperfect as frequency increases. Cable losses, which increase at higher frequencies, increase the rise time of edges and cause the edges to droop as they approach their final value. Figure A.5 illustrates how cables and connectors affect TDR measurements. The upper waveform is the reflection of a step directly output from TDR without any cables and connections. As it is illustrated, cable loss yields the rounded transition of the step and

increase the rise time significantly. Calibration through normalization can correct the measured data, resulting in the better result.

Appendix B

Derivation of TDR Capacitance

In Appendix A, we describe the principle this new TDR C-V measurement technique in a physical explanation. Moreover, the extraction procedure can also be proved by the basic electromagnetic theory and derivation. In section we will further confirm the result by theoretical deviation.

From Appendix A. Equation (A.11), we know that in time domain:

$$V_{TDR}(t) = L^{-1}\left[\frac{V_{Step}}{s}\left(\frac{Z_L - Z_0}{Z_L + Z_0}\right)e^{\frac{2ls}{v_p}}\right] + V_{Step}U(t)$$

$$\text{With } V_{Step} = \frac{V_0}{2}, U(t) = \begin{cases} 0 & t < 0 \\ 1 & t > 0 \end{cases} \qquad \text{(B.1)}$$

Hence by studying the waveform of equation (B.1), we could in theory be able to estimate the values of circuit parameters of the unknown load by examining the characteristics and shape of the reflection. However, this direct method is not desirable as the exact waveform of the reflection depends upon the bandwidth and sampling rate of the oscilloscope chosen. An indirect method of determining inductance and capacitance of reactive load using time domain extraction method is suggested here.

Although the MOS capacitor is modeled as a three-element circuit model, we will start from the simplest case which is only a single ideal capacitor to illustrate the

mathematical derivation.

B.1. Ideal Capacitor Load

For an ideal capacitive load *C*, the impedance is $Z_L = \frac{1}{sC}$. Performing integration on both sides of equation (A.12) (Appendix A) with respect to time from zero to infinity:

$$\int_0^\infty V_{TDR}(t)dt = \int_{2T_D}^\infty \left(V_{Step} - 2V_{Step} e^{-\frac{1}{Z_0 C}(t-2T_D)} \right) dt + \int_0^\infty V_{Step} dt$$

$$\int_0^\infty V_{TDR}(t)dt = -2V_{Step} Z_0 C + \int_0^\infty V_{Step} U(t-2T_D)dt + \int_0^\infty V_{Step} dt$$

Solve for the capacitance C,

$$C = \frac{1}{2Z_0 V_{Step}} \int_0^\infty \left\{ V_{Step} \left[\left(U(t) + U(t-2T_D) \right) \right] - V_{TDR}(t) \right\} dt \tag{B.2}$$

In the discussion of appendix A.1, we mentioned that the detected voltage is the superposition of the incident voltage waveform and the reflected voltage waveform delayed by the round trip travel time $2 \times T_D$. Open circuit load will result in the incident voltage fully reflected and mathematically it can be expressed as:

$$V_{Open}(t) = V_{Step} \left[\left(U(t) + U(t-2T_D) \right) \right] \tag{B.3a}$$

Finally we get

$$C = \frac{1}{2Z_0 V_{Step}} \int_0^\infty \left(V_{Open}(t) - V_{TDR}(t) \right) dt \tag{B.3b}$$

Here $V_{Open}(t)$ is the reflected voltage when the load is replaced by an open circuit. Equation (B.3b) serves as the basis for measuring capacitance of a capacitive load at the end of the transmission line in time domain.

B.2. Series R-C load

For series R-C load of capacitance C and series resistor R_S, $Z_L = R_S + 1/sC$. Perform the integration on Equation (A.13) we get

$$\int_0^\infty V_{TDR}(t)dt = \int_{2T_D}^\infty \left[V_{Step} - V_{Step} \frac{2Z_0}{R_S + Z_0} e^{-\frac{1}{(R_S+Z_0)C}(t-2T_D)} \right] dt + \int_0^\infty V_{Step} dt \qquad \text{(B.4 a)}$$

Solving equation (B.2a) for C:

$$\int_0^\infty V_{TDR}(t)dt = -2V_{Step}Z_0C + \int_0^\infty V_{Step}\left[U(t) + U(t-2T_D)\right]dt$$

$$C = \frac{1}{2Z_0V_{Step}} \int_0^\infty \left[V_{Open}(t) - V_{TDR}(t)\right]dt \qquad \text{(B.4 b)}$$

Equation (B.4b) suggests that although the series resistor R_S changes the time domain behavior, it will not affect the expression of capacitance extraction. Physically it can be understood by realizing that while the R_S slows down the charging process, the capacitor will still be fully charged eventually. As we integrate all the enclosed area, all stored charges are accounted for. This characteristic is a major advantage of the TDR method for capacitance extraction.

B.3. Shunt R-C load

For a shunt R-C load with capacitance C and shunt resistor R_P, $Z_L = \frac{R_P}{1 + sC \cdot R_P}$.

From equation (A.14), as the capacitor finishes the charging, the steady state voltage will be less than the steady state value of the open circuit due to leakage current:

$$V_{TDR}(t=\infty)=\frac{R_P-Z_0}{R_P+Z_0}\cdot V_{Step}U(t-2T_D)+V_{Step}U(t) \tag{B.5a}$$

Perform the integral of Equation (A.14), solve the capacitor as:

$$\int_0^\infty V_{TDR}(t)dt=\int_{2T_D}^\infty\left(V_{Step}\cdot\frac{R_P-Z_0}{R_P+Z_0}-\frac{2R_P}{(R_P+Z_0)}\cdot V_{Step}\cdot e^{-\frac{(R_P+Z_0)}{R_PZ_0C}(t-2T_D)}\right)dt+\int_0^\infty V_{Step}dt \tag{B.5b}$$

$$\int_0^\infty V_{TDR}(t)dt=2V_{Step}\frac{R_P^2Z_0C}{(R_P+Z_0)^2}+\int_0^\infty\left(V_{Step}\cdot\frac{R_P-Z_0}{R_P+Z_0}U(t-2T_D)+V_{Step}U(t)\right)dt$$

$$C=\frac{1}{2Z_0V_{Step}}\left(\frac{R_P+Z_0}{R_P}\right)^2\int_0^\infty\left\{V_{Step}\left[\left(\frac{R_P-Z_0}{R_P+Z_0}\right)U(t-2T_D)+U(t)\right]-V_{TDR}(t)\right\}dt$$

$$C=\frac{1}{2Z_0V_{Step}}\left(\frac{R_P+Z_0}{R_P}\right)^2\int_0^\infty\left\{\left(\frac{R_P-Z_0}{R_P+Z_0}\right)V_{Open}(t)-V_{TDR}(t)\right\}dt \tag{B.5c}$$

B.4. Series/Shunt R-C model

As mentioned earlier, the equivalent circuit of ultra thin MOS capacitor is a three-element model with both series resistance R_S and shunt resistance R_P. In this case we have $Z_L=R_S+R_P//\frac{1}{sC}$. With the knowledge of previous example of capacitance extraction, we can handle this case easily. Perform the integral from time zero to infinity on both sides of equation (A.18) we get:

$$\int_0^\infty V_{TDR}(t)dt=\int_{2T_D}^\nabla V_{Step}\left[\frac{R_P+R_S-Z_0}{R_P+R_S+Z_0}-\frac{2R_PZ_0}{(R_S+Z_0)(R_P+R_S+Z_0)}e^{-\frac{(R_P+R_S+Z_0)}{R_PC(R_S+Z_0)}(t-2T_D)}\right]dt+\int_0^\infty V_{Step}dt$$

$$\int_0^\infty V_{TDR}(t)dt=2V_{Step}\frac{R_P^2Z_0C}{(R_P+R_S+Z_0)^2}+\int_0^\infty\left(V_{Step}\cdot\frac{R_P+R_S-Z_0}{R_P+R_S+Z_0}U(t-2T_D)+V_{Step}U(t)\right)dt$$

$$C = \frac{1}{2Z_0 V_{Step}} \left(\frac{R_P + R_S + Z_0}{R_P} \right)^2 \int_0^\infty \left\{ V_{Step} \left[\left(\frac{R_P + R_S - Z_0}{R_P + R_S + Z_0} \right) U(t - 2T_D) + U(t) \right] - V_{TDR}(t) \right\} dt$$

$$C = \frac{1}{2Z_0 V_{Step}} \left(\frac{R_P + R_S + Z_0}{R_P} \right)^2 \int_0^\infty \left\{ \left(\frac{R_P + R_S - Z_0}{R_P + R_S + Z_0} \right) V_{Open}(t) - V_{TDR}(t) \right\} dt \qquad \text{(B.6)}$$

From circuit point of view, all the previous three cases of load is simply the three-element model under specific condition. For example, series R-C model is simply taken $R_P = \infty$ and shunt R-C model is just $R_S = 0$. So the final expression of MOS capacitor extraction will be reshaped as:

$$C = \frac{1}{2Z_0 V_{step}} M \int_0^\infty \left[\left(\frac{R_0 - Z_0}{R_0 + Z_0} \right) V_{Open}(t) - V_{DUT}(t) \right] dt \qquad \text{(B.7)}$$

where $R_0 = R_P + R_S$, $M = \dfrac{(R_0 + Z_0)^2}{R_P^2}$

Appendix C

Derivation of Tunneling Front Model

C.1. Physical Model

We use n-channel MOS capacitor under strong inversion to discuss the our new derived tunneling front model. The high field results in a large density of electrons ($10^{13}cm^{-2}$) which are confined to a narrow layer at the Si-SiO_2 interface (2-D electron gas). The strong confinement further quantizes the energy levels leading to the formation of sub-bands. Within a semi-classical picture, the electrons residing in the discrete localized states of inversion layer of a MOS structure, bounces back and forth in the confining potential well and can escape by tunneling to the traps in the oxide through the thin dielectric barrier.

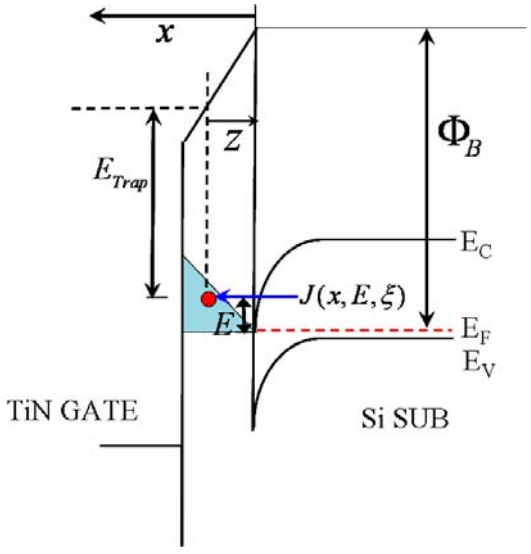


Figure C.1 Energy diagram of n-channel MOS capacitor at inversion. Electrons at inversion have

certain probability to tunnel into the traps located in the oxide.

C.2. Tunneling Probability for Single Dielectrics

The electrical field ξ in the oxide is related to the charge induced across the capacitor by the field over the permittivity according to Gauss' law:

$$\xi = \frac{1}{\varepsilon_{OX}} \int_{V_{FB}}^{V} C dV \tag{C.1}$$

The tunneling transmission probability of an electron in the substrate with energy E to the trap states located at distance z in oxide under the electrical field ξ is $T(E,z,\xi)$. It can be calculated using the appropriate wavevector $k(E,x,\xi)$ from WKB approximation:

$$T(E,z,\xi) = \exp\left\{-2\int_0^z k(E,x,\xi)dx\right\} \tag{C.2}$$

and $k(E,x,\xi)$ is given by the energy E of the tunneling electrons and the barrier height $E_{OX}(x)$ at specific distance.

$$k(E,x,\xi) = \sqrt{-\frac{2m}{\hbar^2}\left[E - E_{OX}(x)\right]} \tag{C.3}$$

where $E_{OX}(x)$ is related to the barrier height at the Si/SiO_2 interface Φ_B by

$$E_{OX}(x) = \Phi_B - q\xi x \tag{C.4}$$

If we assume $E < \Phi_B - q\xi x$ (direct tunneling), from equations (C.2), (C.3) and (C.4), the tunneling probability $T(E,z,\xi)$ becomes

$$T(E,z,\xi) = \exp\left\{-2\int_0^z \sqrt{-\frac{2m_{OX}}{\hbar^2}\left[E - E_{OX}(x)\right]}dx\right\}$$

$$= \exp\left\{-2\int_0^z \sqrt{-\frac{2m_{OX}}{\hbar^2}\left[E - \Phi_B + q\xi x\right]}dx\right\}$$

After the integral, we can get

$$T(E,z,\xi) = \exp\left[-\frac{4\sqrt{2m_{OX}}}{3\hbar q\xi}\left\{\left(\Phi_B - E\right)^{3/2} - \left(\Phi_B - E - q\xi z\right)^{3/2}\right\}\right] \qquad \text{(C.5)}$$

Equation (C.5) is the tunneling probability of a single electron in the Si substrate with energy E above the conduction band edge to an empty energy state located at distance z inside the oxide under the electric field ξ.

C.3. Tunneling Probability for Dual Layer Dielectrics

Electrons first tunnel from the Fermi level E_F in the Si substrate through the interfacial SiO_2 layer and the high-k layer to traps located at an energy level E_{Trap} as shown. The voltage drop across the dielectric stack is divided between the two layers determined by the ratio of their equivalent oxide thickness (EOT). We can write:

$$\frac{\varepsilon_{OX} t_{HK}}{\varepsilon_{HK} t_{OX}} = \frac{\left(EOT\right)_{High-K}}{\left(EOT\right)_{SiO_2}}$$

and

$$V_{OX} = V\left(\frac{\varepsilon_{OX} t_{HK}}{\varepsilon_{HK} t_{OX}} + 1\right)^{-1} = \left(\frac{\varepsilon_{OX} t_{HK}}{\varepsilon_{HK} t_{OX}} + 1\right)^{-1}\left(V_G - V_{FB}\right)$$

$$V_{HK} = V\left(\frac{\varepsilon_{HK} t_{OX}}{\varepsilon_{OX} t_{HK}} + 1\right)^{-1} = \left(\frac{\varepsilon_{HK} t_{OX}}{\varepsilon_{OX} t_{HK}} + 1\right)^{-1}\left(V_G - V_{FB}\right) \qquad \text{(C.6)}$$

V is the applied voltage, i.e. $V = V_G - V_{FB}$.

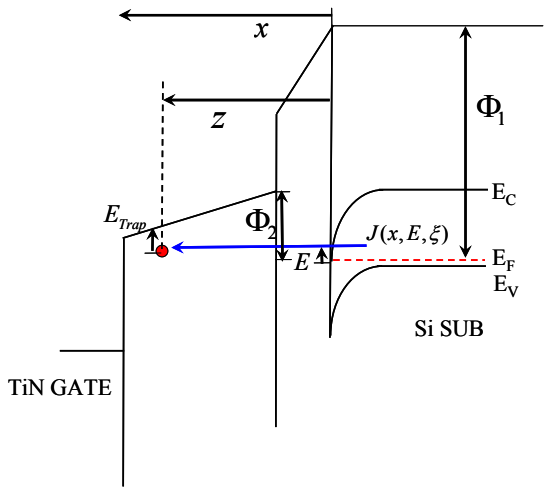


Figure C.2 Energy band diagram of dual layer with SiO_2/HfO_2 as gate dielectrics. Electrons at inversion can tunnel into the traps located in the bulk of high K layer.

Depending on the value of V, the tunneling mechanism can be either direct (DT) or Fowler-Nordheim (FN), or a combination of both. For substrate injection, there are four distinct tunneling regimes defined by the applied voltages.

a) Direct tunneling in both oxide and high-k dielectric layer

$$\Phi_1 - E > qV_{OX} \quad \text{and} \quad \Phi_2 - E > qV_{HK} + qV_{OX} \tag{C.7}$$

b) Direct tunneling in the oxide and FN tunneling in high-k layer

$$\Phi_1 - E > qV_{OX} \text{ and } \quad qV_{OX} < \Phi_2 - E < qV_{HK} + qV_{OX} \tag{C.8}$$

c) Direct tunneling in the oxide and conduction in high-k layer

$$\Phi_1 - E > qV_{OX} \quad \text{and} \quad \Phi_2 - E < qV_{OX} \tag{C.9}$$

d) FN tunneling in oxide layer, $\Phi_1 - E < qV_{OX}$ (C.10)

Cases (c) and (d) will not be considered in this work.

For cases (a) and (b) we can write the expression of the transmission probabilities for a single electron tunneling into an empty trap state located in the high-k layer as a

function of incident energy, distance and electrical field:

For $z > t_{OX}$

$$T_{HK}(E,z,\frac{V_{HK}}{t_{HK}}) = T(E,t_{OX},\frac{V_{OX}}{t_{OX}})\exp\left[-\frac{4t_{HK}\sqrt{2m_{HK}}}{3\hbar q V_{HK}}\left\{(\Phi_2 - E)^{3/2} - \left[\Phi_2 - qV_{OX} - E - q\frac{V_{HK}}{t_{HK}}(z - t_{OX})\right]^{3/2}\right\}\right] \tag{C.11}$$

where $T(E,t_{OX},\frac{V_{OX}}{t_{OX}})$ is the probability of tunneling through the SiO_2 layer that is given in equation (C.5).

C.4. Tunneling Current Density

When the final state is always available to receive electrons (ignoring final state effect), the tunneling current from the inversion layer can be calculated as the number of electrons in the inversion layer divided by the average time an electron spends in the inversion layer (lifetime $t(E_{ij})$). The inverse lifetime is given by the impact frequency with which a particle localized in the well hits the permeable wall multiplied by the probability $T(E,z,\xi)$ to escape the well (tunneling):

$$\frac{1}{t(E_{ij})} = f(E_{ij})\cdot T(E_{ij},z,\xi) \tag{C.12}$$

Thus the current density is given by:

$$J(E,z,\xi) = q\sum_{i,j}\frac{N_{ij}}{t(E_{ij})} \tag{C.13}$$

where the sum runs over each energy valley (i) and subband (j).

Since the contribution of electrons at high energy level to the total tunneling current is very small compared to the lowest energy level E_0 (~0.2eV above E_C due to strong confinement, we can neglect the escape time $t(E_{ij})$ variance with energy level and replace it with the life time at energy level E_0, so that

$$J(E,z,\xi) = qN_e f(E)\cdot T(E,z,\xi)$$
$$\text{with } N_e = \sum_{i,j} N_{ij} \qquad \text{(C.14)}$$

For the lowest energy level, the impact frequency $f(E)$ can be estimated as

$$f(E) = \frac{E_0}{h} \sim 10^{14}\, s^{-1} \qquad \text{(C.15)}$$

Note that equation (C.15) did not come from the *E=hv* formula for photons. It happens to look that way because the rest of the factors combined to have a numerical value of roughly 1.

When there is no guarantee that a final state exist for every electron attempting to tunnel, the tunneling current must reduce. Realizing that only when a suitable empty final state exist can an electron tunnel, the tunneling current can be obtained by replacing the density of electrons to the density of available final states:

$$J_T(z,\xi) = q\cdot N_{trap}(z,E_{Trap})\cdot\{1-\rho(z,E)\}\cdot f(E)\cdot T(E,z,\xi) \qquad \text{(C.16)}$$

where $\rho(z,E)$ is the occupancy fraction.

Conceptually, equation (C.16) may seem strange. It says that tunneling current is independent of available electrons as long as its density is larger than the final state density. For example, if there are 10 empty final states, they will be filled at the same

rate whether the available electron for tunneling is 20, 50 or 200. This may seem counter intuitive until one realizes that tunneling requires the energy of the initial and final state to match. By exclusion principle, there can only be one electron in the initial state that matches the final state. All the rest of the excess electrons cannot be classified as “available for tunneling”.

A further issue to consider is the localized nature of the defect states. Some researchers introduce a capture cross section to account for this highly localized nature of the final state. The idea is appealing. Since the energy states at the receiving electrode are delocalized (assuming the crystal size is comparable to the electrode size), the probability of an electron to find it is much higher than the highly localized defect state. On the other hand, the electrons in the inversion layer occupy delocalized state and each electron is everywhere across the surface. Thus it has no problem interacting with a defect state that is highly localized. Thus adding a cross section term in the tunneling current equation is really inappropriate.

C.5. Tunneling Front Model (Field Free Case)

Tunneling front model was first proposed by Manzini *et.al* to explain the tunneling discharge of holes trapped in thin oxide by avalanche injection. [Manzini *et al.* INFOS 83]. It was found that the time dependence of the flatband voltage shifts after avalanche injection of holes could be modeled by the motion of a tunneling front

moving through the oxide with distance/time dependence.

Consider the problem of an electron with energy E impinging upon a fixed barrier height Φ_B (without any electrical field). Assume $E < \Phi_B$, the tunneling probability will exponentially decay with the distance into the classical forbidden region as:

$$T(E, z, \xi) = \exp(-2kx) \tag{C.17}$$

where k is the wavevector of tunneling electron in the classical forbidden region, which is related to the barrier height facing the tunneling electron.

$$k = \sqrt{\frac{2m_{OX}}{\hbar^2}\Phi_B} \tag{C.18}$$

For typical situation with $E < \Phi_B$, the tunneling probability decays steeply as a function of distance into the oxide. Coupled with the effect on occupancy, it is not hard to see that at any given time the distribution of tunneling rate into a uniformly distributed trap density would sharply peak at a certain distance - Traps that are at shorter distance are mostly filled and the traps that are further away have very low tunneling probability. A sharp boundary of occupied and unoccupied trap state thus exists and this boundary move further into the oxide logarithmically with time:

$$z_m(t) = \frac{1}{2k}\ln\left(\frac{t}{\tau_0}\right) \tag{C.19}$$

This is the tunneling front model. All the terms in the equation are known except τ_0 and that is the key parameter that determines how deep the tunneling front reaches at any given time.

C.6. Tunneling Front Model with Electrical Field

Before we turn our attention to the assessment of τ_0, we need to address the presence of an applied field because almost all the experiment involves a high electric field. Using a trapezoidal approximation for the tunneling barrier (neglecting the image potential and space charge effects due to the trapped electrons), Oldham *et al*. finds that a modified wavevector is all that is required [Oldham *et al.* *IEEE Tran. Nuclear Science,* 33, 1986]

$$k(\xi) = k_0 \sqrt{1 - \frac{q\xi z_0}{\Phi_B}} \tag{C.20}$$

where ξ is the electrical field in the oxide.

For electron traps within 3 to 4 nm from the Si/SiO_2 interface, which is the approximate range sampled by many charge pumping study, equation (C.20) predicts a decrease in k of ~25 percent between zero field and an oxide field 4MV/cm.

Oldham *et al.'s* simple approach ignored the fact that with the existence of electrical field, the tunneling probability is no longer a pure exponential decay with the distance (see equation (C.5)). To accurately assess $z_m(t)$, we must start with equation (C.5) and the result is:

For $z < t_{OX}$

$$z_m(t) = \frac{\Phi_B - E}{q\xi} - \frac{1}{q\xi}\left[\frac{3\hbar q\xi}{4\sqrt{2m_{OX}}} \ln\left(\frac{t}{\tau_0}\right) + \left(\Phi_B - E\right)^{3/2}\right]^{2/3} \tag{C.21}$$

For $z > t_{OX}$ (dual layer stack)

$$z_m(t) = t_{OX} + \frac{t_{HK}}{qV_{HK}}\left\{\Phi_2 - qV_{OX} - E - \left[\left(\Phi_2 - E\right)^{3/2} + \frac{3\hbar qV_{HK}}{4t_{HK}\sqrt{2m_{HK}}}\ln\left(\frac{t \times T\left(E, t_{OX}, \frac{V_{OX}}{t_{OX}}\right)}{\tau_0}\right)\right]^{2/3}\right\} \tag{C.22}$$

Again, all the terms are known except τ_0.

Curriculum Vita

YUN WANG

Education

- **Ph.D candidate**, Department of Electrical and Computer Engineering 01/2003– present
  Rutgers, the State University of New Jersey, New Brunswick, New Jersey

- **Bachelor of Science**, Department of Electrical Engineering 09/1998– 06/2002
  Nanjing University, Nanjing,China

Working Experience

- **Project Engineer**, Full time, Jingzi Electronic Device Company, Suzhou, China 06/2002–01/2003
- **Project Engineer**, Summer Intern, Fujitsu Software Co.Ltd, Nanjing, China 05/2000–09/2000

Publications

- ***Y. Wang***, V. Lee, K. P. Cheung, **"**Frequency dependent charge-pumping, how deep it probes?" *Presented at IEEE International Electron Device Meeting (IEDM), 2006*
- ***Y. Wang***, K. P. Cheung, A. Oates, P. Mason, "Ballistic phonon enhanced NBTI", *Presented at IEEE International Reliability Physics Symposium (IRPS),2007*
- ***Y. Wang***, K. P. Cheung, R. Choi, G. A. Brown, B.H. Lee, "Time Domain Reflectometry for Capacitance-Voltage Measurement with Very High Leakage Current," *IEEE Electron Device Letters,* vol.28, pp. 51-53, Jan.2007
- ***Y. Wang***, K. P. Cheung, R. Choi, G. A. Brown, B.H. Lee, "Accurate series resistance extraction from capacitor using Time-Domain-Reflectometry," *IEEE Electron Device Letters,* vol.28, pp. 279-281, Apr.2007
- ***Y. Wang***, K. P. Cheung, R. Choi, G. A. Brown, B.H. Lee, "Error and correction in capacitance-voltage measurement due to the presence of source and drain," *IEEE Electron Device Letters,* vol.28, pp. 640-642 July 2007
- ***Y. Wang***, K. P. Cheung, K. Sheng, and C. S. Pai, "New low-cost MEMS capacitive pressure sensor concept," *Presented at SPIE*, vol. 5592, pp. 313-319, 2005.
- K. P. Cheung, D. Hits, and ***Y. Wang***, "Electron trap distribution in thin oxide after high-field stress," *Applied Physics Letters*, vol. 86, pp. 102905-1, 2005.
- K. P. Cheung, R. Grover, ***Y. Wang***, C. Gurkovich, G. Wang, and J. Scheinbeim, "Substrate effect on the thickness of spin-coated ultrathin polymer film," *Applied Physics Letters*, vol. 87, pp. 214103-1, 2005.
- K. P. Cheung, ***Y. Wang,*** and C. S. Pai, "Laser closing of window as a novel wafer-level hermetic packaging technology," *IEEE Proceedings. 55th Electronic Components and Technology Conference,* vol. 1, pp. 566-71, 2005
- ***Y. Wang***, K. P. Cheung, R. Choi, G. A. Brown, B.H. Lee, "An Accurate Capacitance-Voltage Measurement Method for Highly Leaky Devices", submitted to IEEE Transaction of Electron Device.

Curriculum Vita

YUN WANG

Education

- **Ph.D candidate**, Department of Electrical and Computer Engineering 01/2003– present
  Rutgers, the State University of New Jersey, New Brunswick, New Jersey

- **Bachelor of Science**, Department of Electrical Engineering 09/1998– 06/2002
  Nanjing University, Nanjing,China

Working Experience

- **Project Engineer**, Full time, Jingzi Electronic Device Company, Suzhou, China 06/2002–01/2003
- **Project Engineer**, Summer Intern, Fujitsu Software Co.Ltd, Nanjing, China 05/2000–09/2000

Publications

- ***Y. Wang***, V. Lee, K. P. Cheung, **"**Frequency dependent charge-pumping, how deep it probes?" *Presented at IEEE International Electron Device Meeting (IEDM), 2006*
- ***Y. Wang***, K. P. Cheung, A. Oates, P. Mason, "Ballistic phonon enhanced NBTI", *Presented at IEEE International Reliability Physics Symposium (IRPS),2007*
- ***Y. Wang***, K. P. Cheung, R. Choi, G. A. Brown, B.H. Lee, "Time Domain Reflectometry for Capacitance-Voltage Measurement with Very High Leakage Current," *IEEE Electron Device Letters,* vol.28, pp. 51-53, Jan.2007
- ***Y. Wang***, K. P. Cheung, R. Choi, G. A. Brown, B.H. Lee, "Accurate series resistance extraction from capacitor using Time-Domain-Reflectometry," *IEEE Electron Device Letters,* vol.28, pp. 279-281, Apr.2007
- ***Y. Wang***, K. P. Cheung, R. Choi, G. A. Brown, B.H. Lee, "Error and correction in capacitance-voltage measurement due to the presence of source and drain," *IEEE Electron Device Letters,* vol.28, pp. 640-642 July 2007
- ***Y. Wang***, K. P. Cheung, K. Sheng, and C. S. Pai, "New low-cost MEMS capacitive pressure sensor concept," *Presented at SPIE*, vol. 5592, pp. 313-319, 2005.
- K. P. Cheung, D. Hits, and ***Y. Wang***, "Electron trap distribution in thin oxide after high-field stress," *Applied Physics Letters*, vol. 86, pp. 102905-1, 2005.
- K. P. Cheung, R. Grover, ***Y. Wang***, C. Gurkovich, G. Wang, and J. Scheinbeim, "Substrate effect on the thickness of spin-coated ultrathin polymer film," *Applied Physics Letters*, vol. 87, pp. 214103-1, 2005.
- K. P. Cheung, ***Y. Wang,*** and C. S. Pai, "Laser closing of window as a novel wafer-level hermetic packaging technology," *IEEE Proceedings. 55th Electronic Components and Technology Conference,* vol. 1, pp. 566-71, 2005
- ***Y. Wang***, K. P. Cheung, R. Choi, G. A. Brown, B.H. Lee, "An Accurate Capacitance-Voltage Measurement Method for Highly Leaky Devices", submitted to IEEE Transaction of Electron Device.